

Theodosia Prodromou *Editor*

# Big Data in Education: Pedagogy and Research

# Policy Implications of Research in Education

Volume 13

## Series Editors

Prof. Stephen L. Jacobson, State University of New York, University at Buffalo – State University of New York, Buffalo, NY, USA

Paul W. Miller, University of Greenwich, London, UK

## Editorial Board

Prof. Helen Gunter, University of Manchester, Manchester, UK

Prof. Stephan Huber, Institute for the Management and Economics, University of Teacher Education Central, Zug, Zug, Switzerland

Prof. Jonathan Jansen, University of the Free State, Bloemfontein, South Africa

Prof. Karen Seashore Louis, Educational Policy and Admin, University of Minnesota, Minneapolis, MN, USA

Dr. Guri Skedsmo, University of Oslo, Oslo, Norway

Prof. Allan Walker, Ctr, c/o Anthon Chu Yan Kit, Hong Kong Institute of Education, Tai Po, New Territories, Hong Kong

In education, as in other fields, there are often significant gaps between research knowledge and current policy and practice. While there are many reasons for this gap, one that stands out is that policy-makers and practitioners may simply not know about important research findings because these findings are not published in forums aimed at them. Another reason is that policy-makers and educational authorities may tend to apply only those findings that agree with and legitimate their preferred policies. Yet we hear often the mantra that policy and practice should be research based and informed by evidence. This claim relates to the interplay between the social realities of science, politics and educational practice and draws attention to knowledge production and application, processes of implementation, change and innovation. However, there are often different interests involved, different knowledge domains, political and economic interests, and legitimate questions can be raised with regard to what counts as 'research', what counts as 'evidence', who should define it, what are their implications for policy, and what kind of actions should consequently be taken to improve education for children and youth.

Please contact Astrid Noordermeer at [Astrid.Noordermeer@springer.com](mailto:Astrid.Noordermeer@springer.com) if you wish to discuss a book proposal.


More information about this series at <http://www.springer.com/series/11212>

Theodosia Prodromou  
Editor

# Big Data in Education: Pedagogy and Research

 Springer

*Editor*

Theodosia Prodromou   
University of New England  
Armidale, NSW, Australia

ISSN 2543-0289

ISSN 2543-0297 (electronic)

Policy Implications of Research in Education

ISBN 978-3-030-76840-9

ISBN 978-3-030-76841-6 (eBook)

<https://doi.org/10.1007/978-3-030-76841-6>

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Introduction

Data is essential for people, corporations, governments, and others to make decisions about the future. In the past, collecting data was costly and time consuming. The emergence of the Internet as a unified global platform for digital connectivity has provided many diverse new sources of human- and machine-generated data. These sources, often called “Big Data,” include commercial transactions, remote imagery, sensor measurements, geospatial positioning, web content, and online user activity. The strategy used by various global governmental agencies is to systematically combine complementary information derived from Big Data sources and traditional data sets, in order to create a richer, more dynamic, and better focused statistical picture of the issue under investigation. This is intended not only to reduce the cost and time-to-market of existing statistical products, but also to deliver innovative solutions that meet the evolving information needs of statistical consumers, generating new economic value.

Recent open data initiatives around the world (United Nations, 2014) are changing the data landscape for statistical analysis, the ways in which data is captured and presented, and the necessary level of statistical literacy to analyses and interpret data for future decision making. The advent of Big Data accentuates the need to enable citizens to develop statistical skills, thinking and reasoning needed for representing, integrating and exploring complex information from diverse sources in a Big Data era.

In education, Big Data are a powerful tool that can be used to inform, engage, and great opportunities for students, teachers and policy makers. Large amounts of educational data are captured and generated every day from different sources and in different formats in schools and higher education. Of particular interest for Big Data in education are two categories: (a) Educational Data related to administrative, educational, and quality-improvement processes and procedures; (b) data produced for and from students’ use and interaction with learning management systems (LMSs), online learning platforms, learning material and activities, course information consisting of learning objectives, syllabuses, results of examination, evaluations of students, and other materials. Moreover, the pandemic is accelerating the long-term

trend in education of putting more education online, which leads to the creation of even more massive data sets. But what can we do with those data and how can we use them to make more informed decisions to facilitate achievement?

## **Description and Purpose of Work**

This book discusses how Big Data can be implemented in educational settings and research, using empirical data. It suggests both best practices and areas in which to invest future research and development, in line with its broader vision of supporting “informed and increased use of statistics” for representing, integrating, and exploring complex information from diverse sources in the big-data era. Research in this field is important and offers guidance to researchers who are seeking suitable topics to explore. Research about the skills that data practitioners (data analysts, data managers, statisticians, and data consumers) use would provide insights into the statistical skills, thinking, and reasoning they use and the skills needed to work with Big Data. Moreover, research about Big Data integration in educational settings could provide a concise reference for policymakers, who must make critical decisions regarding funding and applications. There is a need for research into the implications of the Big Data revolution for statistics education and research. Such research is of particular importance as the biggest leaps forward in the next several decades—in business, and society at large—will come from insights gained through understanding the vast quantities of open data being collected by government and non-governmental organisations. While data collection moves forward, it is often without a concomitant investment in developing practices for the use of Big Data. In response, educators are driven to ask how to develop data knowledge and data literacy to benefit from these new resources.

## **Primary Audience of the Book**

This book is for educators interested in using Big Data. It includes ideas for students, researchers, practitioners, and administrators interested in using Big Data as well as scholars in the areas of education, primarily in statistics and mathematics education; curriculum developers; forward-thinking policy makers such as DARPA; national bureaus of statistics, such as the Australian Bureau of Statistics; data-driven journalists; Tertiary education data scientists, statisticians, and mathematics researchers and lecturers; postgraduate students and advanced undergraduate students in the field of Statistics Education, Mathematics Education and Data science; and computer developers.

## **The Organization of the Book**

The book begins with a section on theoretical and ethical concerns. It then moves on to a section addressing questions of teaching people to use Big Data more effectively. The third section is about using Big Data in education/systems that use Big Data in education. The final section deals with policy/curriculum development.

### ***Theoretical and Ethical Concerns***

This section about theoretical and ethical concerns includes a chapter by Jacques Raubenheimer examining the challenges, pitfalls, and opportunities for Big Data in academic research. Raubenheimer discusses Big Data as a product of the computer era, enabling the knowledge economy, in which academic researchers are key players, although researchers have been slow to adopt Big Data as a source for academic enquiry. He explains that this may be in part because Big Data are curated by commercial or governmental entities, not by researchers, and discusses several challenges facing researchers, including those associated with the size of the data, the development and growth of data sources, and the temporal changes in large data sets. Further challenges are that Big Data are gathered for purposes other than research, making their fit-for-purpose problematic; that Big Data may easily lead to overfitting and spuriousness; and that there are biases inherent to Big Data. Linkage of data sets always remains problematic. The chapter discusses difficulties in generalizing Big Data results, and considers new ethical problems raised, and others that might be obviated. Big Data offer many opportunities, allowing researchers to study previously inaccessible problems, with previously inconceivable sources of data. Although Big Data can overcome some of the challenges of small data studies, Big Data studies will not supplant small data studies—these should work in concert, leading to real-world translation that can have a lasting impact.

### ***Teaching People to Use Big Data Effectively***

The second section of the book is about how to teach people to use Big Data effectively.

Chapter 2, by Peter Howley, Kevin Wang, and Ayse Aysin Bilgin discusses Big Data in early education. Noting that an individual's aspirations and occupational futures may be shaped at a much earlier age than previously considered, they consider how to bring Big Data to young students. This chapter provides introductory approaches for engaging early learners in the Big Data arena. It articulates by example the broad application of the field, and provides relatively straightforward yet appealing interactive activities, and support resources to develop rudimentary but



essential concepts and processes for each of the 4–7 years and 8–18 years age groups. It presents a combination of activities and freely available software to support data exploration.

The third chapter, by Ayse Aysin Bilgin and Peter Howley, focuses on Big Data in a Master of Applied Statistics unit. It discusses how to prepare graduates of statistics courses for their future work by providing opportunities to learn how to deal with big, complicated and complex data sets. This chapter provides a case study on the inclusion of Big Data into a unit about data mining. The short history of the unit, topics of study, learning outcomes, assessment tasks, and how students were included in decision making for their projects is described before presenting two examples of student projects.

Chapter 4, by Markus Zwick and Sabine Köhler, focuses on the role of National Statistical Institutes (NSIs) in upskilling of statisticians by providing opportunities for work and training at the NSIs, providing competences and expertise to combine complex data sets such as satellite or mobile phone data and use future official statistics products. The chapter focuses on the cooperation with academic, public, and private sector entities to further develop statistical skills. One focus of these considerations is the European Master of Official Statistics (EMOS), a knowledge network of official data producers and universities. The approaches and experiences of the Federal Statistical Office of Germany (Destatis) are described as well.

### *Using Big Data to Improve Teaching*

Chapter 5, by Theodosia Prodromou and Ron Kennett, considers, with an example, opportunities and challenges of analytics in education. It discusses how the information quality (InfoQ) framework can be applied in order to enhance students' conceptual understanding and other learning outcomes, and applies the framework to an example concerning academic performance of university students pursuing a Bachelor and Master Degree Programme in Education. The rationale is to provide information regarding the students' performance and their actions on the online learning platform. It investigates how the day of assignment submission affects the grade of the students and how this information provides unit coordinators with feedback to evaluate and review the unit through the lens of best practices.

The sixth chapter, by Masha Etkind, Theodosia Prodromou, and Uri Shafrir, discusses how a specific pedagogical tool—meaning equivalence reusable learning objects (MERLO)—generates Big Data that can be used to assist teaching. A MERLO-based pedagogy generates Big Data for each student in large undergraduate classes, include scores of weekly quizzes, mid-term tests, and final exams. These data show details of the evolution of deep understanding of each concept in the course, from weekly formative MERLO quizzes that reveal individual students' conceptual strengths and weakness. This allows the instructor to suggest individual corrective measures with Interactive Concept Discovery in the course digital

Knowledge Repository, that are conducted and discussed by individual students with their peers, and enhance learning outcomes.

The seventh chapter by William Billingsley and Peter Fletcher, examines the opportunities for student-facing learning analytics in authentic tasks using authentic tools. By employing professional tools, the authors designed environments that allowed students to work on realistic open-ended problems while gathering data on the strategies and practices the students used in the creation process. In some fields, such as software engineering, professional and open source projects gather this sort of data, and those same tools allow collection of student data, allowing exploration of whether students are adopting strategies that experts find to be successful. The authors considered this as the goal of developing cognitive apprenticeships, supported by smart technology, that use more authentic environments.

Chapter 8, by Belinda A. Chiera, Małgorzata Korolkiewicz, and Lisa Schultz, aims to harness data in order to gain a better understanding of students' online engagement patterns in a standard university setting that combines face-to-face learning with online resources and learning environments. In the context of first-year University students, it is particularly vital to understand student patterns of online engagement. The data for the research conducted here has been sourced from a large Australian university. Patterns of student engagement are explored using a variety of data visualisations and statistical analysis techniques to develop an understanding of student engagement with course resources over the duration of a semester.

### *Educational Systems that Use Big Data*

Chapter 9, by Dr. Georgios Papaioannou, reviews Big Data in museums, considering both theoretical perceptions and relevant practices in current Big Data projects in museums. While schools and libraries are the most obvious institutions in the educational system, museums, too, can be viewed as primarily educational institutions, given that museums often explicitly acknowledge an educational role (e.g., The Smithsonian Institution website says it is “the world’s largest museum, education, and research complex,” <https://www.si.edu/about>).

For museums, too, Big Data, data analytics and data mining processes comprise relatively new fields of research and inquiry, but are expected to become core disciplines as data input increases and digital data collections from museums' everyday activities and operations become larger and larger. The data considered for this study, include rich elaborated databases of museum objects and services, data related to museum visits, and the deluge of visitor-generated data on the internet and social media. The author presents connections between Big Data in museums and Big Data in libraries, arts, and humanities. He also offers several case studies from museums operating Big Data projects, which illustrate problems, prerequisites, and potentials.

## ***Policy/Curriculum Development***

Chapter 10, by Carlos Eduardo Ferreira Monteiro, Maria Niedja Pereira Martins, and Theodosia Prodromou, discusses and compares explicit and implicit prescriptions related to the use of Big Data in statistics education in the basic education curricula of Brazil and Australia. The study used a qualitative documentary approach, examining official governmental policy guidelines and curricular programs for education. For the Brazilian context, curriculum documents were collected from 27 Brazilian states. Forty-five documents were analysed in order to identify Big Data potentialities in the teaching of statistics for the secondary level of schooling. The analysis of the Australian curricula documents is of particular interest as the development of statistics education is a major area of research interest in Australia, and this has influenced recent Australian curricular reform. Therefore, the chapter examines Brazilian and Australian contexts in order to identify limits and possibilities for utilization of Big Data in school curriculum. It concludes with suggestions for the development and inclusion of knowledge and practices for teaching Big Data in statistics, as well as a call for adding these elements to teacher-education programs, as current programs do not seem to explore these concerns.

# Contents

## Part I Theoretical and Ethical Concerns

<b>1</b>	<b>Big Data in Academic Research: Challenges, Pitfalls, and Opportunities</b> . . . . .	<b>3</b>
	Jacques Raubenheimer	

## Part II Teaching People to Use Big Data Effectively

<b>2</b>	<b>Big Data for Early Learners</b> . . . . .	<b>41</b>
	Peter Howley, Kevin Wang, and Ayse Aysin Bilgin	
<b>3</b>	<b>Using Big Data in a Master of Applied Statistics Unit</b> . . . . .	<b>65</b>
	Ayse Aysin Bilgin and Peter Howley	
<b>4</b>	<b>Statistical Education as Part of the Digital Transformation of Statistical Offices</b> . . . . .	<b>89</b>
	Markus Zwick and Sabine Köhler	

## Part III Using Big Data to Improve Teaching

<b>5</b>	<b>Big Data, Analytics and Education: Challenges, Opportunities and an Example from a Large University Unit</b> . . . . .	<b>103</b>
	Ron S. Kenett and Theodosia Prodromou	
<b>6</b>	<b>Enhancing Learning Outcomes with ‘Big Data’ from Pedagogy for Conceptual Thinking with Meaning Equivalence Reusable Learning Objects (MERLO) and Interactive Concept Discovery (INCOD)</b> . . . . .	<b>125</b>
	Masha Etkind, Theodosia Prodromou, and Uri Shafrir	
<b>7</b>	<b>Employing Authentic Analytics for More Authentic Tasks</b> . . . . .	<b>141</b>
	William Billingsley and Peter Fletcher	

**8 Learning from Learning Analytics: How Much Do We Know About Patterns of Student Engagement? . . . . . 163**  
Belinda A. Chiera, Małgorzata W. Korolkiewicz,  
and Lisa J. Schultz

**Part IV Educational Systems That Use Big Data**

**9 Museum Big Data: Perceptions and Practices . . . . . 201**  
Georgios Papaioannou

**10 Analysing Aspects of Brazilian Curricula for Teaching Statistics Involving Big Data . . . . . 217**  
Carlos Eduardo Ferreira Monteiro, Maria Niedja Pereira Martins,  
and Theodosia Prodromou

**Part V Conclusion**

**11 Concluding Comments . . . . . 239**  
Theodosia Prodromou

**Part I**  
**Theoretical and Ethical Concerns**

# Chapter 1

## Big Data in Academic Research: Challenges, Pitfalls, and Opportunities



Jacques Raubenheimer

**Abstract** Big Data are a product of the computer era, enabling the knowledge economy, in which academic researchers are key players, although researchers have been slow to adopt Big Data as a source for academic enquiry. This may be in part because Big Data are curated by commercial or governmental entities, not by researchers. Big Data present several challenges to researchers, including those associated with the size of the data, the development and growth of data sources, and the temporal changes in large data sets. Further challenges are that Big Data are gathered for purposes other than research, making their fit-for-purpose problematic; that Big Data may easily lead to overfitting and spuriousness; and the biases inherent to Big Data. Linkage of data sets always remains problematic. Big Data results are hard to generalize, and working with Big Data may raise new ethical problems, even while obviating old ethical concerns. Nonetheless, Big Data offer many opportunities, allowing researchers to study previously inaccessible problems, with previously inconceivable sources of data. Although Big Data overcome some of the challenges of small data studies, Big Data studies will not supplant small data studies—these should work in concert, leading to real-world translation that can have a lasting impact.

**Keywords** Big data · Volume, velocity, variety · Dimensionality · Overfitting · Ethics · Google Trends · Facebook · Crimson Hexagon · Social media

---

J. Raubenheimer (✉)  
University of Sydney, Sydney, Australia  
e-mail: [jacques.raubenheimer@sydney.edu.au](mailto:jacques.raubenheimer@sydney.edu.au)

© Springer Nature Switzerland AG 2021  
T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy  
Implications of Research in Education 13,  
[https://doi.org/10.1007/978-3-030-76841-6\\_1](https://doi.org/10.1007/978-3-030-76841-6_1)

## Introduction

The concept of “Big Data” may very well be relative. Certainly, if the term had existed at the time, the library of Alexandria would have been described as a Big Data warehouse—it was purported to contain “all the books in the world” (Charles, 1913). In fact, its closest modern equivalent, the Library of Congress, has holdings for which the complete catalogue itself counts as a big data set even by today’s standards, being in excess of 164 million items (Library of Congress, 2017).

Regardless of how we define *big data*, one thing is certain of the twenty-first century—we are no longer in an industrial era, where wealth and progress stem primarily from manufacturing. We are now solidly embedded in the knowledge era, a concept formalized and developed by Peter Drucker, as far back as the 1950s (see, for example, chapters 4 and 5 of Drucker, 1957). He realized that, in this new age, knowledge would become the new capital, and that education would undergo a complete revolution in terms of its role in society (Turriago-Hoyos et al., 2016; Wartzman, 2014). This concept may be extended, though, in that knowledge is always *about* something, and since knowledge is the skill of interpreting and using information, knowledge requires information. Or, more to the point, *knowledge requires data*. Thus, having data does not mean having knowledge, but one cannot gain knowledge without data. And so, the true capital of the twenty-first century is not knowledge, but data, and those with knowledge deal in data. Of course, industrial production—and hence manufacturing—will always remain. Just as industrialization did not obviate agriculture, but rather profoundly changed agriculture, especially through the introduction of mechanization to farming and the subsequent creation of super-farms, so also the knowledge era already has profoundly altered manufacturing, and will continue to do so (Thomas & McSharry, 2015).

The primary commodity of the twenty-first century is thus data. And the more data one has, the more one can benefit—provided, of course, the data can be used. Big Data, then, is what it is all about at the moment. And yet, the idea of exactly what constitutes *Big Data* remains a moving target. As will be pointed out, though, the necessity of a precise definition is becoming more of a moot point as time progresses—the real issue is no longer the size of a data set, but its scope and its utility. As Stephens-Davidowitz notes (Stephens-Davidowitz, 2017, p. 62): “The Big Data revolution is less about collecting more and more data. It is about collecting the right data.”

At present, Big Data is a concept that appears to have much more traction in business than in academia. Early academic forays into the world of Big Data appeared to be mostly meta-study—for example, defining what constitutes Big Data, methods for working with Big Data, etc. (Youtie et al., 2017). It has taken longer for academics to come to the realization that Big Data itself can be used as a data source for answering their research questions, so much so that Stephens-Davidowitz could still lament in 2017 (p. 274) that “the majority of academics have ignored the data explosion caused by the digital age.” Now, however, we are on the cusp of an explosion of research using Big Data (Harlow & Oswald, 2016; King,



2014; Lazer, Pentland, et al., 2014b). This chapter seeks to provide an assessment of the progress made, and the changes that are to come, showing how the wider world’s embrace of Big Data impacts researchers and their research studies, across the world, and across disciplines.

## A Brief Overview of the Historical Development of Big Data

Big Data would not be possible without computers, which have, thus far (Courtland, 2015), followed what is commonly called Moore’s law (Moore, 1965)—that the computing power of silicone semiconductor chips being manufactured would double every 2 years—an exponential growth in computing power. Without these ever more powerful computers, we would not have the tools with which to do the massive number of calculations needed to analyse Big Data.

However, a complementary process, more often ignored, which is vital for the very concept of Big Data, is that before data can be analysed by a computer, they need to be stored electronically. Big Data would not have been possible without digitization. Thankfully, due to a slew of new technologies which have appeared over the years, the tools for data storage have grown in capacity. Figure 1.1 shows the storage capacity, in bytes, of various data storage tools over the years (using information from Anthony, 2011; Data Recovery Group, 2011; Pingdom, 2008). The lines show how select devices (e.g., hard drives) have grown in capacity since

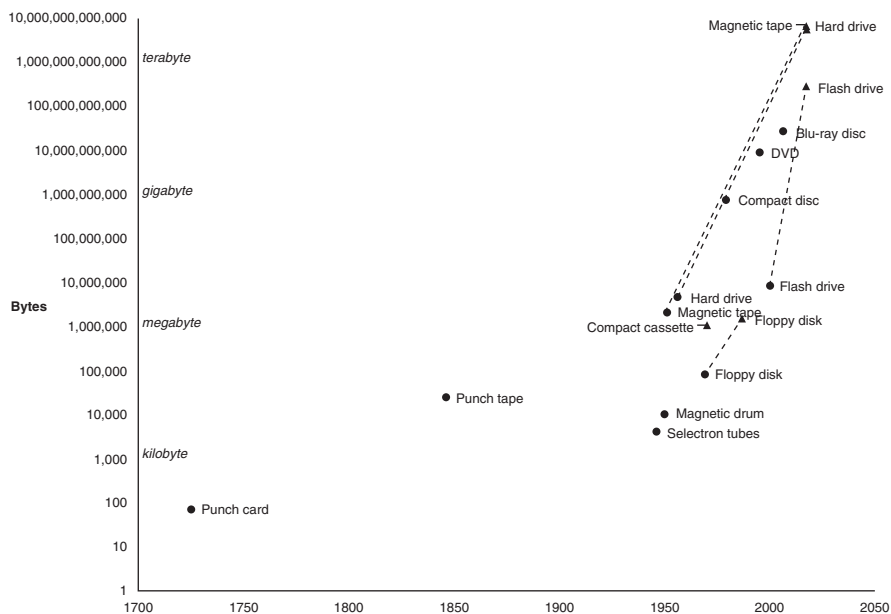


Fig. 1.1 Storage capacity of various data storage devices over time

their initial introduction. What this shows is that the storage capacity of individual digital storage devices has grown massively over the years, but what it fails to show is the rate at which these devices have been manufactured. While it is reported that IBM manufactured 10,000,000 punch cards daily as early as 1937 (Anthony, 2011), some of the more esoteric devices were never manufactured in huge quantities. Today, however, hard drives, flash drives, and optical disks are being manufactured by the hundreds of millions—estimates (e.g., The Statistics Portal, 2018) are for half a billion hard drives alone being shipped annually, although that covers devices over the full range of storage capacities, making estimates of the total storage difficult. Nonetheless, Coughlin (2015) has estimated that the annual storage volume created by hard drive manufacture alone is in the vicinity of 540 exabytes.

Hilbert has studied the rate at which the volume of digitized information is growing, and has made two important observations. First, he defines the current era as the digital era, estimating its inception in 2002, the year in which, by his calculations, more data were stored in digital format than in analog (Hilbert, 2012a, p. 9). Second, his calculation of the total amount of storage space (combining the missing element of the amount of storage devices with their capacity; Hilbert, 2012a, 2012b, 2014, Hilbert & López, 2011, 2012a, 2012b) shows that the amount of available storage is, at least for the present, and probably into the foreseeable future, growing faster than the amount of data being generated (see also Lesk, 1997).

It should also be noted that this growth in the amount of digitized information stems not only from the capture and storage of *new* digital information, but also from the digitization of *existing* analog data. Many examples can be given, but mentioning the digitization of performance recordings (e.g., video recordings, vinyl records) and the scanning and often the optical character recognition of print material (e.g., Google Books, academic journals digitizing old print issues) should suffice.

It is, then, this combination of the ever-increasing computing power, and the ever-expanding storage capacity, together with the increasing tendency to capture and store information, past and present, digitally, that has enabled the Big Data revolution.

However, if the Big Data revolution has been enabled by advances in computing power and storage capacity, researchers should realize that we have entered a new phase, viz., the Big Data explosion. The rapid proliferation of Internet-connected devices, together with the advent of cloud computing (i.e., data being sent directly to, and stored on, the Internet), giving rise to the Internet of Things (IoT), means that potentially anything, and perhaps more intimidatingly, everything, can be become a data-capture device. New companies are being started, offering services solely for the purpose of collecting data. Data itself has become a commodity, and the new economy is an economy of data—Big Data (Thomas & McSharry, 2015).

Massive data sets are literally being generated everywhere, and even academic researchers themselves are involved in creating a massive set of data through their scholarly publications. The questions of how many academic publications are in existence and how many are published per year are complicated by the fact that there are numerous academic publishers involved, and several large (and

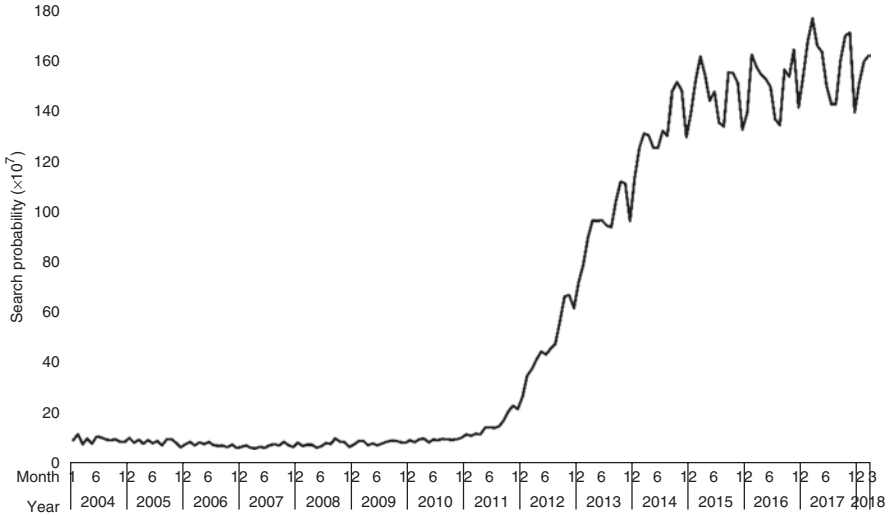
competing) cataloguing services. Nonetheless, various estimates (Björk et al., 2009; Jinha, 2010; Larsen & von Ins, 2010; Mabe & Amin, 2001; Ware & Mabe, 2009, 2012, 2015) put the annual growth rate in publications at 3% or above (even as high as 8% for certain fields), so that it was estimated that at the end of 2009, there were over 50 million academic publications in existence (Jinha, 2010). Using these estimates, and an absolute minimum growth rate of 3%, we can safely estimate that at the very latest, 2019 will be the year in which two million new journal articles are published, although Plume and van Weijen (2014) estimated 2.4 million new journal articles in 2013 already.

### *Where Did the Concept of “Big Data” Originate?*

The coining of the phrase “Big Data” is commonly (e.g., Lohr, 2013) attributed to Mashey (cf. Mashey, 1998), who used the term in various presentations and discussions in the 1990s (but see also Dontha, 2017). The concept itself was perhaps first succinctly articulated by Laney (2001). Although he did not use the term “Big Data,” per se, he discussed the data challenges large enterprises would face going into the twenty-first century, and described the three dimensions of data as volume, velocity, and variety. These so-called “3 Vs” of Big Data have become a catchphrase. These concepts are still important, especially to academic researchers wishing to work with Big Data, and will be elaborated in the section on the challenges of Big Data to academic research below.

In essence, however, an awareness had long been arising that information workers (including researchers) were being exposed to, and required to work with, ever more voluminous sets of data, and that, apart from the “tools” (the aforementioned computing power and storage capacity) required, special techniques and approaches would be required to make sense of this information (cf. Press, 2013 for a good historical overview).

Interestingly, one of the exciting new sources of Big Data for academic researchers which will be discussed below, Google Trends, can give us a good understanding of how interest in the topic has grown. Figure 1.2 shows a plot of 2040 samples (giving a margin of error of 2.17%) of worldwide search volume from the Google Trends Extended Health API, using a custom-designed data extraction tool developed by this researcher (Raubenheimer, 2021). Curiously, despite the term first being used in the 1990s, and appearing in an academic publication in 1997 (Cox & Ellsworth, 1997), the topic only really gained interest, in terms of it being something people started searching for on the Internet, towards the middle of 2011. This is further corroborated by Youtie, Porter, and Huang’s (2017, p. 67) review which “did not find a substantial number of social science Big Data articles until 2012.” From there, however, it showed a sharp rise in interest, and has reached something of a plateau in the last 2 years, so that it can accurately be called a buzzword.



**Fig. 1.2** Google Trends searches for Big Data (Topic), 2004–2017

### *The Custodians of Big Data*

To date, the bulk of work on the analysis of Big Data has been conducted in the commercial/industrial sector, where companies gather data and then analyse it for some financial advantage. It is no secret that Google and Facebook, as well as other large technology companies like Apple and Microsoft, use their huge volumes of data to create immense profits. These companies often offer a service for free, with their reward being the data provided by users. They are then able to leverage that data to generate immense profits, most commonly through the sale of targeted advertising (Hachman, 2015; Holland, 2017; Monnappa, 2017; Mullin, 2012). But the list of companies using their data in this way is by no means confined to these more well-established tech companies. Further names in different industries, like Uber, Airbnb, etc., are all alike in their use of large volumes of data for generating profit. And, of course, governments themselves have established processes by which they can amass large volumes of data on their citizens and their behaviour, sometimes contentiously so, under the premise that these data can be used for the betterment of the citizenry. Transport for London, as an example, has shared its London Tube data (gathered via its Oyster card system) with researchers, so that it could use the results of their analyses to increase its own efficiency (cf. Badger, 2014; Chiang et al., 2017; “London’s Tube and Big Data: Underground movement,” 2013; Sari Aslama et al., 2015). In a similar fashion, Strava, a commercial company which will be discussed below in the section on ethical considerations, has released its Metro data (Gordon, 2014) to “[partner] with departments of transportation and city planning groups to plan, measure and improve infrastructure for bicyclists and pedestrians.” Large data sets are thus available, and although not purpose-built for academic research, these data can be used for academic research.

## ***Big Data as a Source for Academic Research***

Traditionally, researchers have had to gather their own data. Of course, researchers have long used pre-existing data (e.g., a health researcher using hospital records as a source of data) in their research as well, and so the concept of using data not initially gathered for the purposes of research is not new. However, the difference with Big Data is that while, in more conventional research, researchers might choose between collecting their own data or using already collected data (as a matter of necessity or of convenience), with Big Data, most researchers will, at least for the foreseeable future, have neither the means nor the resources to amass large data sets for their own purposes—they will necessarily be reliant on external sources for large data sets. (There are, of course, exceptions to this, such as the mapping of the human genome, or large volumes of data collected by massive telescopes, but for the majority of researchers, this holds true.)

Thankfully, several sources of Big Data are now being opened up to researchers, often in the hope that by allowing researchers to work with the data, the data owners may gain insights they might not otherwise have come by, perhaps as a consequence of the lack of motivation, time, or ability to extract such insights from their own data. In a sense, this is the best form of collaboration possible. Big companies and government institutions amass data, and share them with researchers, who then provide valuable information back to the data owners. A prime example of this is how Transport for London has shared its London Tube data with academic researchers mentioned above. Researchers are, then, to use Drucker's term, becoming some of the best knowledge workers. Their reward is provided up front—they are given access to data, and the product of their labour is knowledge, through which both they (in their academic careers) and the data custodians benefit.

## **Challenges of Big Data for Academic Research**

At the best of times, research can be challenging. Working with Big Data is no different, although the domain does present some unique challenges.

### ***The Big-Ness of Big Data***

Perhaps the best place to start discussing the challenges which Big Data presents to academic researchers is to return to the 3 Vs which Laney (2001) listed as business enterprise challenges. These three points have come to be viewed as the “classical” defining characteristics of Big Data (e.g., Dave, 2013), even though others have sought to elaborate on the concept (e.g., DeVan, 2016; IBM, n.d.-a, n.d.-b; Jain, 2016; Marr, 2014; McNulty, 2014; “Why the 3V's are not sufficient to describe big

data,” 2015), and as such, this chapter will retain its focus on the original three points. Given that data may be defined as “big” when they meet *at least* the conditions of the 3 Vs, it would do well to consider each of these as a challenge to researchers wishing to use Big Data for their studies.

*Volume* refers to the fact that data sets are becoming massive, even more so as data are seen as being inherently valuable, leading to the situation where new data are added, but old data are not discarded. Lesk (1997) noted that our capability to store information has outpaced the already massive amounts of information being stored, with the paradoxical outcome that no data will have to be lost (simply everything can be stored), while most data will not be used (at least not by humans). This presents a twofold challenge to researchers: Firstly, researchers simply will have to learn to use the technology required to deal with voluminous data. As an example, social media produces data sets that are truly massive (Moe & Schweidel, 2017)—Crimson Hexagon allows its clients (which includes researchers) access to over a trillion items of social media data (Crimson Hexagon, 2018; Moriarity, 2017).

The volume of Big Data present many practical challenges for analysis, as detailed by Jacobs (2009). This means researchers will have to use more powerful computers or networked high-performance computing facilities (Doctorow, 2008). At the very least, it necessitates working in a 64-bit, as opposed to a 32-bit environment (e.g., 32-bit programs cannot handle data files larger than 4 Gb, which can easily be found with voluminous data sets). It requires that researchers learn to modify their programming techniques to optimize performance with large data sets (e.g., when using an inefficient sorting method with a typical research data set, the result might be a time difference of a second or two, but with a massive data set, the result of an inefficient sorting algorithm could be hours of additional computing time). Stephens-Davidowitz (2017, p. 180) casually mentions that “one day I decided to download Wikipedia. (You can do that sort of thing nowadays.)” Given that Wikipedia (2018) itself estimates the size of its text-only dumps at 58 Gb (when uncompressed—the actual compressed download file is a mere 14 Gb!), this is a clear example of the volume of data that one may have to work with.

Secondly, and more importantly, the true challenge of volume, given Lesk’s conclusions mentioned above, is not the computational challenge of dealing with millions or even billions of records of data per se, but rather the challenge of finding the information we need amidst a growing amount of irrelevant and redundant data. The size of the needle is not changing, but the size of the haystack is growing exponentially (Taleb, 2013). Stephens-Davidowitz (2017, p. 21) notes that “The size of a dataset... is frequently overrated.... You don’t always need a ton of data to find important insights. You need the right data.... Most important... *you have to ask the right questions*” (emphasis added). To meet this challenge, researchers will have to become innovative and learn to think outside of the box, to use another management catchphrase. As will be discussed below, working with Big Data requires rethinking even the questions we want to ask, so that we can ask the questions that are suited to what the data can give us, instead of banging our heads against a wall, trying to force Big Data to give us answers it cannot yield.

Laney's (2001) second concept of *velocity* refers to the fact that data sets are growing continuously. An analysis of data today will include information not available yesterday. A good example of this is Google Trends data: Firstly, the data are updated in near real time, so that an investigation into the popularity of searches will continuously provide new (never ending) information. As a side note, it should be pointed out, especially in further discussions of Google Trends below, that even Google scientists contended that the data may not be suitable for the forecasting of future events, but may be useful for contemporaneous forecasting (cf. Choi & Varian, 2009, 2012). Furthermore, even for historic data, Google samples 10–15% of all searches, and resamples it each day, so that repeating an analysis of the same search terms for the same time frame on successive days will provide a researcher with a sample of samplings for that query (Matsa et al., 2017). To meet this challenge, researchers will have to accept the time-boundedness and the immediacy of their work. Time is an enormous confounder in any research, and even more so with Big Data, because the data may change considerably over the longer time frames common in large data sets. More will be said about this when discussing the variety of data, but the challenge of velocity is that researchers will, in many instances, have to think about how their studies can be tailored to make use of new data as it arrives, often in pursuit of making predictions based on real-time data. This is no simple task, no matter how much data one has at one's disposal.

Laney (2001) was convinced, rightly, that the largest challenge to data management was the *variety* of data. Primarily, this refers to the fact that data come in so many different forms. Disparate data sets are “incompatible... non-aligned... and inconsistent.” However, this extends beyond just incompatibilities between types of data, and even incompatibilities between data sets of the same type, to inconsistencies with the same sets of data over time—data definitions, variables captured, and even data location may change. A good example of this are the data provided by the United States National Highway Traffic Safety Administration's Fatality Analysis Reporting System (<http://www.nhtsa.gov/FARS>). The user guide (National Highway Traffic Safety Administration, 2016) explains how variables were added, removed, and changed, over the years (e.g., accidents due to cell phone usage were obviously not part of the data recorded in 1975). Also, the user guide (NHTSA, 2016, pp. 520–585) details the ways in which the very structure of the FARS data sets was changed profoundly in 2010 so as to improve its compatibility with the National Automotive Sampling System General Estimates System data.

Perhaps the best comment about the variety inherent in Big Data comes from Lazer, Kennedy, King, and Vespignani (Lazer, Kennedy, et al., 2014a, p. 1204), who, in discussing the failure of Google Flu Trends to maintain its predictive accuracy beyond the original model development (as discussed below), note that,

A more likely culprit is changes made by Google's search algorithm itself. The Google search algorithm is not a static entity—the company is constantly testing and improving search. For example, the official Google search blog reported 86 changes in June and July 2012 alone.... Search patterns are the result of thousands of decisions made by the company's programmers in various subunits and by millions of consumers worldwide.

The variety inherent in Big Data means that researchers will not be able to work with Big Data and stay armchair theorists. They will have to get their hands dirty, and become data wranglers, able to beat unruly data into shape, as it were (of course, without altering or fabricating data!), struggling with incompatibilities in various data sets, and developing a clear understanding of the intricacies and nuances of their data, so that they can wield it appropriately.

### ***Big Data are Not Research Data***

One of the biggest problems faced by researchers wishing to use Big Data for their own research work stems from the fact that the data are often not gathered for the sake of research, but rather for the specific purposes of the data custodians. Of course, as has been mentioned, this is not a new challenge, but is faced by any researcher wishing to use externally collected data for their work. Big Data, however, exacerbates this problem. Essentially, Big Data provide us with masses of information but the information may not be ideally suited to our research questions. Lazer, Kennedy, King, and Vespignani (Lazer, Kennedy, et al., 2014a, p. 1203) note that “the core challenge is that most Big Data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.” Lazer et al. (Lazer, Pentland, et al., 2014b, p. 1203) then warn against “Big Data hubris” which they define as “the often implicit assumption that Big Data are a substitute for, rather than a supplement to, traditional data collection and analysis.” Stephens-Davidowitz (2017), after noting earlier (p. 245) that there are things that cannot be done with Big Data, and things that ought not be done with Big Data, points out that Big Data do not supersede, but rather complement, conventional research. He notes (p. 256) that having access to Big Data for our research “... does not mean that we can just throw data at any question. And Big Data does not eliminate the need for all the other ways humans have developed over the millennia to understand the world. They complement each other.” This means that Big Data do not spell the end of small data. In fact, many times, research questions can only be answered successfully when insights from Big Data are supplemented with insights from small data, such as combining Big Data analysis results with small survey results. This combination of disparate data sources must be done with care, however. The process is well illustrated by Mellon (2013) who demonstrated the use of survey data to validate Google Trends search data and concludes (pp. 289–290):

Rather than replacing the traditional survey, using Internet data may require researchers to do more survey work, particularly in hard to reach areas. Paradoxically researchers may only be able to unlock the potential of the Internet data being generated today, in retrospect, when they have enough data generated with conventional survey tools to test the validity of search data.



An extension of this issue (that Big Data are not created for specific research problems) is the issue of dimensionality, or overfitting (Clarke & Chien, 2017). A problem may arise when our dependent variables are themselves not massive, such that many independent variables are used to predict dependent variables of limited scope (Stephens-Davidowitz, 2017, p. 247), confounding analysis and prediction. Big Data tend very easily to gather information about too many variables but not enough subjects. This may not be immediately apparent, as we naturally assume that Big Data contains information about many people, but when our dependent variable is a rare condition, even vast data sets may easily contain more variables than subjects (Salsburg, 2017, pp. 69–77; Taleb, 2013). A classic example of this is the Google Flu issue, as clearly indicated by Lazer, Pentland, et al. (2014b), p. 1203) when they explain that “essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points.... The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high.” The history of Google Flu is as follows: Building on work of Johnson et al. (2004) and Eysenbach (2006), Polgreen et al. (2008) found that they could use Yahoo! search queries related to flu, to predict the incidence of flu. This was swiftly followed by Ginsberg et al., who, after an announcement on the Google blog (Ginsberg & Mohebbi, 2008), followed with a paper in *Nature* (Ginsberg et al., 2009) showing similar results from Google searches. This led to the creation of the Google Flu Trends web page. Initial results seemed to show that the algorithm worked reasonably well (Araz et al., 2014; Carneiro & Mylonakis, 2009; Dugas et al., 2012, 2013; Malik et al., 2011; Valdivia et al., 2010; Wilson et al., 2009). But over time, this ability degraded (Butler, 2013; Lazer, Kennedy, et al., 2014a; Olson et al., 2013), nor did it perform well when tested against retrospective data (Ortiz et al., 2011), requiring adjustments to the algorithm (Cook et al., 2011; Copeland et al., 2013). The project was labelled as a failure (Butler, 2013; Fung, 2014; Lazer & Kennedy, 2015; Salzberg, 2014; Walsh, 2014), and eventually the project was shuttered (The Flu Trends Team, 2015). The correlation they found for one season did not hold well across many seasons. This means that researchers using Big Data for forecasting would do well to use out-of-sample predictions to further validate their findings, and the insights from Big Data might best be understood when validated with small data. It should be noted, though, that right from the outset, and even noted by its critics, Google Flu Trends was best seen as a way to supplement, not supplant, the conventional epidemic surveillance systems (Lazer, Kennedy, et al., 2014a, p. 1204; Madrigal, 2014; Valdivia et al., 2010, p. 5; Wilson et al., 2009, p. 4).

A last implication of the fact that Big Data are not gathered explicitly for research purposes, is that the quality of the data may be a double-edged sword. On the one hand, the data may be surprisingly free of conventional biases that confound research, such as placebo effects, Hawthorne effects, and a litany of other biases researchers have to deal with. But at the same time, the data may present new biases of their own, biases, perhaps, that researchers are less accustomed to dealing with. So, for example, Stephens-Davidowitz (2017, p. 111) notes that “Google [data] can display a bias towards unseemly thoughts,” and researchers using social media data,

need to be aware of, and account for, the biases inherent in these data (Moe & Schweidel, 2017). Most data gathered from the Internet contains an inherent bias in that it is restricted to Internet users, whose demographics still do not adequately represent the world population, or any country's population, even though Internet usage is growing, and Internet access is being considered as fundamental to the exercise of basic human rights (Human Rights Council of the United Nations, 2016; La Rue, 2011). But even more subtle in its effect, and hence more difficult to control for, is that different demographics of Internet users use the Internet in different ways, with, for example, different age groups showing preferences for different social media networks (Nanji, 2017). Furthermore, a variety of "socioeconomic, political, cultural, social, and technological factors" all help shape Internet access at the national level (Fuchs, 2008).

### *Data Linkage*

Although this could have been mentioned when discussing the variety of Big Data above, this topic is complex enough to warrant a brief discussion. A surprisingly common problem encountered when working with Big Data stems from the previously discussed fact that large data sets are most commonly created for commercial reasons, and are not bespoke research data sets. When researchers plan a conventional study, they go to great pains to ensure that their data collection encompasses all the variables needed to answer the research questions. However, it is not unusual for researchers using Big Data to require variables from multiple, disparate data sets in order to answer their research questions. The only way this can be done is to link the data in some way, so that every record for each individual in one data set is successfully matched with every record for that individual in each of the other data sets. This quickly becomes a very complex exercise, as, for example, the identifying information for each individual may differ in the various data sets. A variety of data linkage methodologies have arisen to deal with this problem (Bohensky et al., 2010; Churches & Christen, 2004; Clark, 2004; Dusetzina et al., 2014; Fellegi & Sunter, 1969; Ong et al., 2014; Tromp et al., 2011; Winkler, 1993, 2014), but the level of specialisation required means that the exercise is often out of reach of the individual researcher. The obstacles are many: Data have to be obtained from multiple custodians, each of which may have their own ethical requirements as to the usage of their data. There are ethical issues around the linkage of data (e.g., personal identification may be impossible in the individual data sets, but could become possible in the linked data; or, conversely, de-identification for ethical purposes may make the data difficult to link). The exercise is typically laborious, time-consuming, costly, and requires significant computing power—even when specialist service providers are used to link the data, time delays are significant, and costs are high. While government institutions (such as census bureaus) may have the personnel and skills to perform these functions themselves, academic institutions typically do not. Thankfully, collaborative efforts may often arise, which can result in the creation of

data linkage services, such as the NSW Centre for Health Record Linkage (CHeReL) in Australia. Examining their guide (Centre for Health Record Linkage, n.d.) gives one a clear understanding of the process involved. The implication for researchers, though, is that the linking of data requires a considerable amount of time and ethical approval, and involves significant cost. This is, however, the reality of attempting to work with data of this nature.

Furthermore, working to analyse these linked data involves a significant step up in terms of the complexity of the statistical programming required. Small data studies very commonly use data sets where the entire study utilizes only a single data set with a 1:1 participant-to-record ratio. The most complex extension of that is longitudinal studies where multiple records are collected for each individual, but still, typically, within a single data set, or with separate data sets for each time point, but with each time-bound data set containing a single record for each participant. Working with these data remains a fairly simple matter. Big Data studies, however, are very often longitudinal (it takes time to amass so much data), and routinely entail working with multiple 1-to-many or many-to-many participant-to-record data sets. The complexity of the way in which records from these data sets are linked, then, grows in order of magnitude for each additional data set, and techniques are needed for determining temporal start and end points for each individual, extracting first and/or last records for each individual from any number of data sets, aggregating records within, and even between, data sets, and identifying individuals with more or less than a certain number of records. The problem of missing data and its underlying reasons also grows in complexity, and the statistical techniques needed to analyse these data also become more complex.

### ***Big Data May Still Not Deliver Generalizable Results***

Even though it is hoped that the results we derive from Big Data studies will lead to translational research, the very nature of Big Data presents a fundamental obstacle to this. The ideal flow of events with conventional medical studies is that when researchers perform clinical trials that show the benefits of certain interventions (e.g., health screenings), and these results are confirmed in further studies, the evidence would be sufficient to recommend the intervention to everyone in the risk category, even though we know that not all at-risk individuals will truly be afflicted over time. But the results of these studies are generalized to the population nonetheless.

However, generalizability is one of the largest challenges of Big Data. Big Data analyses are observational, and can, as an example, indicate increased odds that certain individuals might display a large degree of suicidal ideation. But on data sets of millions of individuals, even with highly significant odds, that would leave a large number (in real terms) of individuals identified as false positives. Even if Big Data analysis results are generalizable to populations at large, they are not specifiable to individuals in particular. The point is that while Big Data insights can help

us understand dynamics in populations at large, we should still refrain from making inferences about individuals in particular. As Agatha Christie's character Sir Henry Clithering notes in "The Four Suspects" (Christie, 1932): "It isn't really guilt that is important—it's innocence. That's the thing that nobody will realize.... Some people may go through a lifetime crushed by the weight of a suspicion that is really unjustified." This can perhaps best be understood by means of an illustration: In the short span of August 2017 to January 2018, two women, one in Australia (cf. Chillingworth, 2018) and one in the U.S. (cf. Hitt, 2018) searched the Internet for information on how to murder their partners. While it is apparent, from these, and other cases (e.g., Deery, 2013; also mentioned in Stephens-Davidowitz, 2017, p. 266) that would-be murderers actually are brazen enough to search for tips on the Internet, and while Big Data allow us to immediately flag any individual carrying out such searches, it would be wrong to immediately prosecute, or even place under surveillance, any individual carrying out such searches (à la *Minority Report* [Spielberg, 2002], as Stephens-Davidowitz, 2017, p. 266] intimates). If we ever were to do that, then fiction novelists (researching their next murder mystery), journalists (researching articles about murderers who google their intended methods), and even academic researchers (investigating the uses of Big Data) also might be unfairly accused.

### ***Spurious Correlations, Sub-Populations, and Other Confounders***

The results from the analyses of Big Data are, simply by nature of their size (and thus the large diversity of their underlying populations), particularly susceptible to the two extremes of delivering spurious correlations on the one hand (as already discussed on the issue of dimensionality above), and being bedeviled by numerous confounders on the other. Academics working with Big Data will need a good level of statistical literacy in order to make the right inferences (and only the right inferences) from their data (Prodromou & Dunne, 2017). These skills, as well as the necessary software programming skills for analysing these data, need to be incorporated into the curricula of university degrees (Puang-ngern et al., 2017).

The study of suicide and the Internet serves as a good example. Several studies appear to have found evidence of a correlation between suicide rates and Internet searches related to suicide and self-harm (mostly analysed using Google Trends): Yang et al. (2011) in Taipei; Hagihara et al. (2012) in Japan; Gunn and Lester (2013) and Ma-Kellams et al. (2016) in the US; Arora, Stuckler, and McKee (2016, in certain subgroups of the population), and Kristoufek et al. (2016) in the UK; and Song et al. (2014) in South Korea. Other ancillary findings also support this idea: Ayers et al. (2013) found evidence of seasonality in searches related to suicide (and other conditions). Biddle et al. (2012, 2016) and Gunnell et al. (2012) examined the growth of suicide-related information on the Internet. Bragazzi (2013) found a strong correlation between Google Trends searches and non-suicidal self-injury.

However, the results are not as clear-cut as they may seem at first. McCarthy (2010) found that the correlation between completed suicides and Internet searches is inverse for the population as a whole, but strongly positive for youth, suggesting that Google Trends may only serve as a predictor for certain demographics. Even one of the aforementioned studies—Ma-Kellams et al. (2016, p. 483, emphasis added)—noted the following:

By capturing ecological variation in suicide-directed Internet behavior, Google search data can be more useful than random-survey-sampling techniques because Google search data capture a broader proportion of the population in an unobtrusive way. This system is *not meant to be a replacement for traditional surveillance mechanisms* or to supplant the need for other forms of suicide surveillance. Rather, data on notable increases in suicide search activity can be used to prevent suicide by indicating where the needs are for closer suicide surveillance... or by reinforcing the need... for more sensitive programs among particular cultural groups.

Furthermore, Fond et al. (2015) found a relationship between suicide searches and depression (but not bipolar) searches, but noted that search volumes could be influenced heavily by external factors, such as media reports. They concluded (p. 917) that “Google Trends® cannot be used as an effective tool for monitoring potential suicide epidemic peaks, because of the lack of precision of existing tools and data provided. However, this tool can be further developed and offers great potential.” The strong influence of media reports on search volume was corroborated by Page et al. (2011) in Australia, who noted that Google Trends could not serve as a predictor on its own. They did note the difficulty of comparing the online data, which is available in near-real time, to official mortality data, which has a significant delay. Additionally, Solano et al. (2016) could not find any predictive ability of Internet searches for completed suicides in Italy (by contrast, the searches tended to lag behind completed suicides), nor could Sueki (2011) in Japan. Bruckner et al. (2014) also reported mixed results, and concluded (it should be noted that their comment also touches on the previously mentioned issue of extrapolating from the general to the individual):

We caution against inferring an individual’s Internet search behavior, depressive symptoms, and suicide risk from our population-level tests. Temporal fluctuations in Google searches are indicators of a population-level interest, which cannot shed light on individual differences in suicide risk. In addition, our findings of a positive relation between Internet search volume for “suicide and depression” and completed suicide in the same month cannot definitively establish that suicide-related searches precede suicide completion. Nevertheless, our findings indicate that current surveillance efforts for suicide may benefit from real-time use of Google-based Internet searches for depression.

Mars et al. (2015) performed a very instructive small data study. Using survey responses in the UK, they found that adolescents with a history of self-harm were significantly more likely to have searched for information on self-harm on the Internet (although no causality could be inferred). But their study also showed that suicide-searchers additionally use the Internet to search for help, and thus making inferences based on searching for information about suicide alone fails to account adequately for the reality of a very complex issue.

Perhaps the most instructive application of Google Trends suicide data as an example of the utility of Big Data, is this: Noting the association of Internet searches with suicide, both Kristoufek et al. (2016) and Parker et al. (2017) found that they could use Google Trends data to predict suicide rates accurately enough for policy decision-making purposes, while conventional data are slow to be made available (e.g., coroner's statistics that could be up to 2 years behind). But this indicates a limited scope for the application of this data. Finally, in the most rigorous of the studies, Tran et al. (2017) performed a multi-country investigation with strict methodology to account for spurious effects in the time series analysis, and found that the associations found did not follow any discernible pattern, and were in line with the Type-I error rate for the expected number of significant results, indicating that Google Trends data may well be too afflicted by spurious correlations to accurately predict suicide rates.

Results from the analysis of Big Data, then, need to be interpreted with caution. While Big Data may provide useful insights, the application of these results should not be overstated and the susceptibility to spurious effects which may be present in the data purely because of the size and nature of the data set, must always be acknowledged, and used to temper hubris about what researchers find.

### ***Big Data Gathering Is Not Passive***

Reference will be made below to the fact that Big Data are sometimes gathered in the form of non-consensual (or at least, unconsciously consensual) randomized trials. What is relevant here is to raise the very large temporal confounder of how Big Data gathering actually changes the behaviour of individuals. Vespignani (2009, p. 428) discusses the implications of this:

An interesting and ethically challenging aspect of predicting and managing the unfolding of catastrophic events in techno-social networks is the system's adaptation to predictions when they are made publicly available. Social behaviors react and adapt to knowledge of predictions. Contrary to what happens in physical systems, the predictions themselves are part of the system dynamic. In addition, predictions may point to unethical control and anticipation strategies favoring specific demographic sectors of the society. Finally, the risk of erroneous predictions may lead to costly or unethical social control mechanisms with no actual benefits.

On a more practical level, the implication for researchers is that the nature of their data may change over time. Lazer and Kennedy (2015), discussing Google Flu Trends, point out how changes over time to the way the Google search page provides additional search terms (e.g., in its autosuggest feature) and recommends certain searches, may cause people to favour the suggested terms instead of the term they had intended entering, which results in temporal changes to the actual data collected. Google itself, ironically, may have been the greatest culprit for the failure of Google Flu Trends, because Google has not only allowed people to search the Internet for content more efficiently, it has also changed the very way in which we

search (and, it can be argued, continues to do so). On a different level, it can additionally be argued (although less easy to prove substantively), that in the relatively short lifespan of Internet searching—although Google was not the first search engine, none can argue that it came to define Internet searching, and the recent 20-year anniversary of the incorporation of Google (September 4, 1998, see Brin & Page, 1998)—the number of people conducting searches, the things people were willing to search for, and the very way in which people searched the Internet has changed, making any comparisons over longer time periods fraught with difficulty.

### *Ethical Considerations with Big Data*

Big Data offers unique ethical complications to the researcher. The issue it appears to resolve is anonymity: When data from millions of individuals are examined in bulk, no finding can be traced back to a specific individual. But Big Data are sometimes not truly anonymous. A typical example of this is found when researchers analyse social media posts, where identifying information may still abound (as users typically do not adequately control their privacy settings). Thus Big Data may contain very personal, very identifying information. And this remains an ethical problem. A clear instance of this has been demonstrated recently in the Strava global heat map. Strava is a software company offering a mobile app and a website interface; it manufactures no physical products. The company gathers geolocation information from users of its training app on their mobile devices, a service that exemplifies data collection in the Big Data era. While the app offers a free service to millions of people worldwide (or enhanced services for a fee), there can be no doubt that the value Strava derives is from the data it harvests from the app users. Strava has used those data, amongst others to release a global heat map (<https://labs.strava.com/heatmap>)—a very high-tech global visualization of the exercise routes and patterns of their users (Robb, 2017). And while, in theory, the map does not allow identification, in practice, it does. As a hypothetical example, if only one person in a rural area has used Strava, then anyone viewing that area on the heat map would be able to identify that person's routes. This may be less problematic than the fact that a student of the International Security Studies programme at the Australian National University was able to identify a surprising amount of detail about US military bases in Syria and other conflict zones, using the Strava heat map (Bogle, 2018). Strava has since corrected this, and allowed users to opt out of having their data included in the heat map, and has suppressed very low volume activities. Thus, academics working with data provided by a commercial data custodian may easily run into ethical issues not foreseen at the inception of their studies.

A further ethical problem is that research is meant to be translational. We do research with the intention that it will, at the very least, make the world a slightly better place. But the results of Big Data analyses may lead to unintended consequences. This is especially so when attempts are made to relate the general findings derived from a Big Data analysis to individuals, something which, it was pointed

out above, should not be done, based purely on the nature of the analysis and the data, let alone the ethical considerations. Nonetheless, Stephens-Davidowitz (2017, pp. 260–1) discusses the possibilities of a world in which inferences like this might be attempted, and notes that “better prediction can lead to subtler and more nefarious discrimination.” In the world of online business, this could easily lead to data wars, where businesses use Big Data to extract maximum profits from consumers using strategies like targeted marketing and differential pricing, while consumers will fight back using Big Data for price comparison, service review, and information sharing. This is a struggle in which academic researchers using those data could easily become embroiled.

A further ethical issue with Big Data stems, once again, from the fact that Big Data are ordinarily not gathered for the purposes of research: The individuals whose data are contained in the data set were not given the opportunity to consent for their data to be used in the research study—such as the people who searched for information on committing suicide discussed above, or people who proved to be thought leaders in fanning the flames of xenophobia in the Citizen Research Centre (2017) study. This is indeed a problematic area. The Google Trends data on suicide have been anonymized, and cannot lead to individual identification. Social media data, on the other hand, contain whatever identifying information is allowed by the individual’s privacy settings, which means, in practice, that many people who either do not understand how to, or do not care to, set privacy settings have their personal information available to any entity (research or corporate) who scrapes that information from the social networking site. Admittedly, the terms of service of the various social networking sites—which, it should be noted, most people never read (Bakos et al., 2014; Obar & Oelof-Hirsch, 2016)—work along the general principle that whatever is not set to be private is considered to be public domain. And it is also so that the trade-off for using a social media site for free is that the site owners are given access to one’s data (cf. Hachman, 2015), to do with pretty much as they please (within reasonable limits), which may include sharing that data with researchers, or using the data for research of their own. And so, the analysis of both Google Trends data and social media data is well within what is permitted by the terms of service of the sites on which the data were gathered. But around the fringes, some of this research definitely traverses into ethically grey areas.

What is even more frightening, though, and is definitely on the fringe of the grey ethical areas, and sometimes well beyond, is that users of these sites often find themselves being actively experimented upon. The simplest form of this is what is known in the information technology world as A/B testing. Essentially, A/B tests are elementary randomized controlled trials, and they are being performed on users of Internet sites continuously, so much so that Facebook reportedly performs over 1000 A/B tests *daily* (Stephens-Davidowitz, 2017, p. 211). Given Facebook’s user base of over a billion people, this means that Facebook is actively experimenting on at least tens of millions, if not hundreds of millions of people, daily. It must be said that many of the tests are innocuous, such as determining which colour or font of a control users respond to best (of course, it should be noted that the ultimate aim of much of this testing is to maximize user engagement with the site, which, when



stripped of its technobabble, means making the site as addictive as possible). However, this experimentation can take on a decidedly darker and more sinister colour: After only two studies (Coviello et al., 2014; Ugander et al., 2012), in which they did not account fully for the issue of dimensionality in their data, Facebook engineers believed they had sufficient evidence of the way interactions on their platform affected the emotional states of users. On the basis of this belief, they embarked on actual and purposeful experimental manipulation of the emotional state of millions of their users (Kramer et al., 2014), totally unbeknownst to those users, and without any consent other than the generic consent given in the Facebook terms of use. This elicited a storm of criticism (e.g., Blue, 2014; Vaughan-Nichols, 2014), and eventually garnered an editorial expression of concern (Verma, 2014). The truly terrifying element from this is that the editorial statement notes that “because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell’s Human Research Protection Program.” What this means, for all practical purposes, is that as a private company, Facebook staffers may still be pursuing research of this nature, although they are likely to be less transparent about it and less likely to publish their findings in academic journals, because they feel that it “may not have justified all of this anxiety” (Kramer, 2014).

Recent events from the first half of 2018 have revealed just how easy it can be for academic researchers to become embroiled in the ethics minefield surrounding big data. Against the backdrop of an ongoing investigation into how the Facebook platform was possibly manipulated to influence the outcome of the US election (Entous et al., 2018), a Cambridge University psychology researcher established an outfit named “Global Science Research” through which he developed an application that conducted an online personality test. He purportedly received funding from a company known as Cambridge Analytica to pay Facebook users a nominal fee for taking the personality test as part of his data gathering strategy (Lewis & Carrie Wong, 2018; Ortutay et al., 2018). Unbeknownst, however, to those Facebook users, was the fact that his application was also harvesting both their data, and all the data of their friends, so that an eventually estimated 87 million or more Facebook users’ data was harvested (Hern, 2018; Ortutay, 2018), and turned over to Cambridge Analytica as part of the funding agreement, while Facebook had reportedly turned a blind eye to this kind of data harvesting from its platform (Lewis, 2018). The researcher claimed that he had used the full \$800,000 paid by Cambridge Analytica to pay the initial 270,000 test takers (from whose wider Facebook networks the remaining data were harvested), indicating that he had not profited at all, but that his “motivation was to get a dataset [he] could do research on” (Ortutay et al., 2018). Cambridge Analytica, for its part, had apparently used these data in providing marketing services for politicians in a number of elections, including the aforementioned US election (Ortutay et al., 2018).

The matter caused even graver concern when it was revealed that Facebook itself was, in fact, harvesting not only the data of its users, but even the data of non-users (who obviously had not consented to such data being collected) who, for example,

had visited pages on other websites (i.e., not belonging to Facebook) but which contained Facebook Like or Share buttons (Ingram, 2018).

Cambridge Analytica and Global Science Research eventually shut down over the controversy (Reuters, 2018). Facebook has, under pressure from various directions, committed itself to labelling politically-paid advertising content (Leathern, 2018), and removed the “Trending news” feature from its pages in June 2018. However, the biggest outcome of this, for academic researchers, is that Facebook announced a new research initiative through which it hopes to empower academic researchers to independently set the agenda for, and work on solutions towards, misinformation, political interference, and other abuse of its platform (Schrage & Ginsberg, 2018). Importantly, the initiative contains stringent guidelines to counter “the threat presented by the recent misuse of Facebook data, including by an academic associated with Cambridge Analytica” and is built on King and Persily’s (2018) model for industry-academic collaboration, an important work for any academic researcher wishing to work with social media Big Data. As will be seen below, Gary King is at the very forefront of creating opportunities for academic researchers to work with Big Data in the social sciences.

In summary, while Big Data may present wonderful opportunities, researchers would do well to consider carefully the full ethical implications of the work they do and the data they work with. A consideration of all the challenges posed by Big Data for academic research may prove disheartening, and lead one to wonder whether working with Big Data is worth the effort. It would do well, then, to consider some of the opportunities provided by Big Data.

## Opportunities of Big Data for Academic Research

Big Data present us with many opportunities. We should avoid the “Big Data hubris” (Lazer, Pentland, et al., 2014b) which could beguile us into thinking that Big Data research could supplant existing “small data” research, but neither should we devalue insights gained from Big Data. Big Data will allow us information that can either supplement traditional survey results, providing additional nuance, or information that may contradict survey results. Even this is good, because it will lead us to the important question of which results should be believed, the deliberation of which should lead us to a clearer understanding of the reality we are attempting to uncover. In their article discussing one of the most widely reported failures of Big Data analysis (Google Flu Trends), Lazer, Pentland, et al. (2014b, p. 1205) conclude by pointing out that “Big Data offer enormous possibilities for understanding human interactions at a societal scale, with rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables.” Despite all the complications surrounding Big Data, Big Data still offer incredible opportunities that deserve mention.

First and foremost, new frontiers are opening up for academic researchers. Big Data often allow us to discover heretofore unknown things. In fact, Lazer, Pentland,

et al. (2014b, p. 1205) encourage scientists to “use Big Data to understand the unknown.” It should be noted that one of Lazer’s co-authors is Gary King, of the Institute for Quantitative Social Science (IQSS) at Harvard University, and also one of the founders of Crimson Hexagon (now Brandwatch), an exciting company for researchers—especially researchers in the social sciences—wanting to explore the Big Data world of social media. The work done at the IQSS (e.g., Hopkins & King, 2010) was spun off into a commercial entity—Crimson Hexagon (Breese, 2015; King, 2011)—which now holds several patents for text mining and sentiment analysis (e.g., US20160189171 A1, 2016, US 9483544 B2, 2012, US20160070748 A1, 2017). The company monitors all major social media and blogging networks, continuously harvesting their data into a database that now contains over a trillion social media posts (Crimson Hexagon, 2018; Moriarity, 2017), which can be analysed using their ForSight™ tool. The company offers its services for a fee, and its major clients are corporate conglomerates who use its real-time data gathering capabilities for brand management and campaign tracking. However, the tool is also being used for academic research (Breese, 2015), and when one considers the cost of figuratively reinventing the wheel (King, 2014) and trying to access the same information, let alone develop an alternative to their patented methods, the fee pales in comparison (although there are alternatives—Chan et al., 2017). The tool has been used in a number of compelling studies, such as a study of how people deal with crisis events (Bair, 2016), public discourse on nanotechnology (Runge et al., 2013), encryption (Addawood et al., 2017), net neutrality (Faris et al., 2015), nuclear energy (Kim et al., 2016; Li et al., 2016), xenophobia (Citizen Research Centre, 2017), food prices (UN Global Pulse, 2014), fracking (Hopke & Simis, 2017a, b), and even as a surveillance tool for eye disease (Deiner et al., 2016) and suicide (Ueda et al., 2017). Big Data, and Crimson Hexagon (now Brandwatch) in particular, are opening up the frontier of social media to the world of the academic researcher.

These new frontiers indicate that Big Data allow us—compel us—to study heretofore inaccessible research questions. Big Data can perhaps best be employed to answer questions where conventional methods are failing (although the caveat may be that the failure of conventional methods may only become apparent when their results are contrasted with those obtained from the analysis of Big Data). However, it is not only that Big Data bring new insights to old questions. Far rather, academics wishing to employ Big Data in their research should be fundamentally rethinking the questions they need to be studying with those data (e.g., Vespignani, 2009). An insight into this comes from one of the early pioneers of academic research using Big Data, Raj Chetty. He notes (in Stephens-Davidowitz, 2017, p. 173): “Big Data is not just about doing the same thing you would have done with surveys, except with more data. Big Data really should allow you to use completely different designs than what you would have with a survey.” Big Data have the ability to upend even very established modes of research. Even medicine, with its strong focus on clinical trials as a gold standard, and evidence-based medicine (but see Smith & Pell, 2003 and the ensuing debate), will see changes resulting from Big Data. Ayers et al. (2014) provide compelling arguments for ways in which behavioural medicine,

powered by Internet Big Data, can open new frontiers in medical science (see Emery et al., 2014 for a good example).

However, Big Data are not only redefining our research questions, but are also redefining what constitutes our research data, opening up endless new possibilities. Stephens-Davidowitz (2017, p. 103) notes that “These days, a data scientist must not limit herself to a narrow or traditional view of data.... Everything is data!” An example of a surprising use of data is the attempt to predict economic growth with night light visible from space (Henderson et al., 2008). In fact, it may be this very redefinition of what we are using as data that compels us to revise the questions we should be asking of that data. But, this redefinition of data also extends to other areas. Stephens-Davidowitz refers to Big Data as “honest data” (2017, pp. 54, 105–163). His enthusiasm may have been dampened if he had googled “Fitbit cheat,” but the truth of the matter is that conventional survey methodology can be (although is not necessarily) significantly biased (e.g., Fish & Russell, 2017; Kamenetz, 2014; Przybylski, 2016; Robinson-Cimpian, 2014). However, while Big Data can give surprisingly unpretentious data, it may also be affected by large amounts of noise and bias (as Stephens-Davidowitz does acknowledge when he discusses social media data, p.150–3). The task of the researcher is to bear this in mind when considering the nature of the data. For example, Fitbit general motion data may be more biased than Fitbit exercise data, and data gathered from devices only activated and used during exercise (e.g., Garmin fitness watches or Strava apps) may be even more unbiased.

Although it was mentioned that research translation can be a challenge when using Big Data, it should be mentioned that when the right data are used to investigate the right questions, Big Data may precisely present opportunities for translation that were not available through small data studies. Big Data *can* give us insights that lead to translation. Stephens-Davidowitz (2017, p. 163) notes that “collecting rich data on the world’s problems is the first step towards fixing them.” So it is hoped that some of the most intangible problems which have always failed analysis by conventional research methods, might find new solutions through the use of Big Data. The examples of how Big Data are being used to improve and transform transportation networks, mentioned earlier, show how Big Data can lead to real-world changes.

One of the biggest advantages of Big Data is that they finally overcome one of the Achilles’ heels of so many research studies (but only in some areas of research, it warrants adding): Sample size. The sheer volume of data means that, if dealing carefully with the issue of dimensionality mentioned above, researchers may still find sufficient data points in Big Data to study even exceptionally rare phenomena. Stephens-Davidowitz (2017, p. 171) makes the simple observation that “you need a lot of observations in a dataset in order to be able to zoom in with clarity on one small subset of that data,” and then goes on to contend (p. 197) that “the world is complicated.... These connections and relationships... cannot be traced with tiny surveys or traditional data methods. The world, quite simply, is too complex and too rich for little data.” Big Data, by the very nature of their volume, contain a depth of cases and a breadth of variables that allow researchers to tackle problems they

would ordinarily have struggled to do with conventional data gathering. Inasmuch as Big Data do not allow us to investigate certain problems, which are best left to conventional small data studies, so also Big Data open up new questions and new areas of research, simply by virtue of their volume, which small data studies cannot investigate.

Following from the issue of large samples, Big Data allow us to find enough data on people with rare conditions or in rare situations, and allow us to find enough individuals who can act as controls matched to those subjects on a wide variety of criteria. When one considers that each additional matching criterion exponentially reduces the number of available controls, it quickly becomes clear why Big Data allow matching when small data fail. It must be said, however, that researchers have devised techniques to compensate for this problem in small data studies (D'agostino, 1998; Fogarty & Small, 2016; Rassen et al., 2011; Roberts et al., 2015; Rosenbaum, 1987, 1989; Schneeweiss, 2006), and Big Data do not make these techniques obsolete, but can complement them beautifully.

Thus, while, as discussed above, many Internet companies (for better or worse) have turned their users into non-consenting (or—considering that those users probably did not read the terms and conditions as mentioned above—unconsciously consenting) research participants in the Internet equivalent of randomized trials (A/B testing), Big Data, because they so often incorporate time components, do allow for the examination of ex post facto conditions, where the size of the resulting groups in the data set are sufficiently powered to compensate for the design. This is a huge boon for academic researchers, where such data are extremely difficult to obtain otherwise.

## Some Examples of the Application of Big Data

Many examples have already been provided in this chapter of the application of Big Data. However, some final examples will be provided, not to produce an encyclopaedia of Big Data sources, but rather as examples to stimulate readers to find their own.

The first example involves deep-vein thrombosis (DVT). Although clinical trials can be used to study factors like the risk for, and possible protective measures against, DVT associated with long-haul flights (Scurr et al., 2001), Big Data studies using record linkage (Kelman et al., 2003) can also provide similar insights into this risk for DVT.

The second example relates to prescription data. Many countries have started making routinely collected prescription data available for research, and the Nordic countries are definitely at the forefront of this (Furu et al., 2010; Lester, 2009; Wettermark et al., 2013). However, the Australian government has also made a sample of the de-identified prescription data of 10% of its population available to researchers, subject to a number of controls and provisions (Mellish et al., 2015; Paige et al., 2015; Pearson et al., 2015). Using this data, researchers were able to

show the massive impact on the medicine adherence of the general population brought about by a single television documentary program (Schaffer et al., 2015).

### ***Openly Available Sources of Data***

Perhaps, rather than saying that “Everything is data!” (Stephens-Davidowitz, 2017, p. 103), we should say that “Data are everywhere!” Researchers can look to many sources for data, some of which have been mentioned in this chapter already. But many, many more exist. Petabytes of data are being generated and stored, and much of that can be accessed for academic research. The United States government has committed to releasing massive amounts of data through its open data initiative (<https://www.data.gov>), and the European Union (<http://data.europa.eu/euodp/en/data>) and the UK (<http://data.gov.uk>) have similar initiatives. For example, the US has released 125 years of health data (<https://www.healthdata.gov/content/about>) for researchers and others to use in the interests of improving health outcomes for all, and the United Kingdom’s NHS also provides similar data (<https://digital.nhs.uk/data-and-information/data-collections-and-data-sets>). Many countries make large amounts of census data available (e.g., <https://www.census.gov/data.html>). Data really can be found everywhere. All we need do is look, find, and ask.

In closing, two large portals for *finding* data that need to be mentioned are those provided by Google (<https://www.google.com/publicdata/directory>) and Amazon (<https://aws.amazon.com/data-sets>), where researchers can search for a wide array of different existing data sources which they might profitably employ in their research work.

### **Conclusion**

This chapter has demonstrated that the world of Big Data is not purely the preserve of industry, but that academic researchers can, and should, be using Big Data for their studies. This is not to say that researchers have not, but rather, that a critical mass of researchers extending the possibilities of what can be done with Big Data has not yet been accumulated. This may partly be ascribed to the challenges presented by working with Big Data. But these challenges are now better understood, and thus easier to account for. Additionally, the obstacles presented by these challenges are also being eroded as the dual forces of technology and methodology become more robust at working with voluminous data sets that may require some form of linking. It was also shown that one of the most important characteristics of Big Data used for academic research is precisely the fact that Big Data sets are typically not gathered for academic research, but obtained via a data custodian who has accumulated the data for other purposes. This impacts every aspect of working with those data, from accessing it, to linking and analysing it, through to the ethical

implications of working with such data. Furthermore, the opportunities which Big Data present to academic research are vast. Researchers should realise that Big Data will not replace traditional small data research methods, and that the opportunities presented by Big Data are dependent on the context of the researcher, the population and the data—in some contexts, Big Data may offer great advantage, but not in others. Ideally, Big Data and small data will become complementary tools in the armouries of academic researchers, which they may draw upon variably, and in concert if need be, as the situation demands.

**Acknowledgments** This project is partially funded by the National Health and Medical Research Council (NHMRC) through the Translational Australian Clinical Toxicology Program (TACT) (grant ID1055176).

## References

- Addawood, A., Schneider, J., & Bashier, M. (2017). Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th international conference on Social Media & Society*. Association for Computing Machinery. <https://doi.org/10.1145/3097286.3097288>
- Anthony, S. (2011). *The history of computer storage (slideshow)*. Retrieved January 19, 2011, from <https://www.extremetech.com/computing/90156-the-history-of-computer-storage-slideshow>
- Araz, O. M., Bentley, D., & Muelleman, R. L. (2014). Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. *American Journal of Emergency Medicine*, 32(9), 1016–1023. <https://doi.org/10.1016/j.ajem.2014.05.052>
- Arora, V. S., Stuckler, D., & McKee, M. (2016). Tracking search engine queries for suicide in the United Kingdom, 2004–2013. *Public Health*, 137, 147–153. <https://doi.org/10.1016/j.puhe.2015.10.015>
- Ayers, J. W., Althouse, B. M., Allem, J. P., Rosenquist, J. N., & Ford, D. E. (2013). Seasonality in seeking mental health information on Google. *American Journal of Preventive Medicine*, 44(5). <https://doi.org/10.1016/j.amepre.2013.01.012>
- Ayers, J. W., Althouse, B. M., & Dredze, M. (2014). Could behavioral medicine lead the web data revolution? *Journal of the American Medical Association*, 311(14), 1399–1400. <https://doi.org/10.1001/jama.2014.1505>
- Badger, E. (2014). *10 Years of London Underground ridership data in one map*. Retrieved January 26, 2018, from <https://www.citylab.com/transportation/2014/01/10-years-london-underground-ridership-data-one-map/8156/>
- Bair, A. R. (2016). *From crisis to crisis: A Big Data, antenarrative analysis of how social media users make meaning during and after crisis events*. Utah State University. <http://digitalcommons.usu.edu/etd/5045/>
- Bakos, Y., Marotta-Wurgler, F., & Trossen, D. R. (2014). Does anyone read the fine print? Consumer attention to standard form contracts. *The Journal of Legal Studies*, 43(1). <https://www.journals.uchicago.edu/doi/abs/10.1086/674424>
- Biddle, L., Derges, J., Mars, B., Heron, J., Donovan, J. L., Potokar, J., Piper, M., Wyllie, C., & Gunnell, D. (2016). Suicide and the Internet: Changes in the accessibility of suicide-related information between 2007 and 2014. *Journal of Affective Disorders*, 190, 370–375. <https://doi.org/10.1016/j.jad.2015.10.028>
- Biddle, L., Gunnell, D., Owen-Smith, A., Potokar, J., Longson, D., Hawton, K., Kapur, N., & Donovan, J. (2012). Information sources used by the suicidal to inform choice of method. *Journal of Affective Disorders*, 136(3), 702–709. <https://doi.org/10.1016/j.jad.2011.10.004>

- Björk, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research*, 14(1). <http://informationr.net/ir/14-1/paper391.html>
- Blue, V. (2014). *Facebook: Unethical, untrustworthy, and now downright harmful*. Retrieved July 1, 2014, from <http://www.zdnet.com/facebook-unethical-untrustworthy-and-now-downright-harmful>
- Bogle, A. (2018). *Strava just published details about secret military bases, and an Australian was the first to know*. Retrieved January 29, 2018, from <http://www.abc.net.au/news/science/2018-01-29/strava-heat-map-shows-military-bases-and-supply-routes/9369490>
- Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., & Brand, C. A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Services Research*, 10, 1–7. <https://doi.org/10.1186/1472-6963-10-346>.
- Bragazzi, N. L. (2013). A Google trends-based approach for monitoring NSSI. *Psychology Research and Behavior Management*, 7, 1–8. <https://doi.org/10.2147/PRBM.S44084>.
- Breese, E. B. (2015). When marketers and academics share a research platform: The story of crimson hexagon. *Journal of Applied Social Science*, 10(1), 3–7. <https://doi.org/10.1177/1936724415569953>
- Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1/7), 107–117. <https://doi.org/10.1.1.109.4049>.
- Bruckner, T. A., McClure, C., & Kim, Y. (2014). Google searches for suicide and risk of suicide. *Psychiatric Services*, 65(2), 271–272. <https://doi.org/10.1176/appi.ps.201300211>.
- Butler, D. (2013). When Google got flu wrong. *Nature*, 494(7436), 155–156. <https://doi.org/10.1038/494155a>.
- Carneiro, H. A., & Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10), 1557–1564. <https://doi.org/10.1086/630200>.
- Centre for Health Record Linkage. (n.d.). *Guide to health record linkage services*. Sydney. [http://www.nss.gov.au/nss/home.nsf/0/e2d861c453d7b7f6ca25756700191b53/\\$FILE/CHeReL\\_Guide\\_version\\_1.3.pdf](http://www.nss.gov.au/nss/home.nsf/0/e2d861c453d7b7f6ca25756700191b53/$FILE/CHeReL_Guide_version_1.3.pdf).
- Chan, M. S., Morales, A., Farhadloo, M., Palmer, R. P., & Albarracín, D. (2017). Harvesting and harnessing social media data for psychological research. In H. Blanton (Ed.), *Social psychological research methods*. Taylor & Francis.
- Charles, R. H. (1913). *The letter of Aristaeus to Philocrates*. Retrieved January 17, 2018, from <http://www.attalus.org/translate/aristaeus1.html#9>
- Chiang, L.-Y., Crockett, R., Johnson, I., & O’Keefe, A. (2017). Passenger flow in the tube. Worcester Polytechnic Institute, . [http://wp.wpi.edu/london/files/2017/06/Tube\\_IQP\\_E17\\_Final.pdf](http://wp.wpi.edu/london/files/2017/06/Tube_IQP_E17_Final.pdf)
- Chillingworth, B. (2018). *Woman charged over death of NSW farmer allegedly searched “murder” before his death*. Retrieved January 26, 2018, from <http://www.smh.com.au/nsw/woman-charged-over-death-of-nsw-farmer-allegedly-searched-murder-before-his-death-20180124-h0o1ts.html>
- Choi, H., & Varian, H. (2009). *Official Google research blog: Predicting the present with Google Trends*. <http://googleresearch.blogspot.com/2009/04/predicting-present-with-google-trends.html>
- Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *Economic Record*, 88(SUPPL.1), 2–9. <https://doi.org/10.1111/j.1475-4932.2012.00809.x>.
- Christie, A. (1932). *The thirteen problems*. Ulverscroft.
- Churches, T., & Christen, P. (2004). Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making*, 4, 1–17. <https://doi.org/10.1186/1472-6947-4-9>.
- Citizen Research Centre. (2017). *Xenophobia on Social Media in SA, 2011–2017: Anatomy of an Incident: Violence in Gauteng and the “March against Immigrants.”* <http://citizenresearchcentre.org/2017/03/15/xenophobia-on-social-media-in-south-africa/>
- Clark, D. E. (2004). Practical introduction to record linkage for injury research. *Injury Prevention*, 10(3), 186–191. <https://doi.org/10.1136/ip.2003.004580>.



- Clarke, F., & Chien, C.-H. (2017). Visualising Big Data for official statistics: The ABS experience. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and Big Data* (pp. 224–252). : IGI Global. <https://doi.org/10.4018/978-1-5225-2512-7.ch009>.
- Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing Google Flu Trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic. *PLoS One*, 6(8), 1–8. <https://doi.org/10.1371/journal.pone.0023610>.
- Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). Google Disease Trends: An update. <http://research.google.com/pubs/archive/41763.pdf>
- Coughlin, T. (2015). *HDD annual unit shipments increase in 2014*. Retrieved January 22, 2018, from <https://www.forbes.com/sites/tomcoughlin/2015/01/29/hdd-annual-unit-shipments-increase-in-2014>
- Courtland, R. (2015). *Gordon Moore: The man whose name means progress: The visionary engineer reflects on 50 years of Moore's Law*. Retrieved September 15, 2016, from <http://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress>
- Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. *PLoS One*, 9(3), 1–6. <https://doi.org/10.1371/journal.pone.0090315>.
- Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In *Proceedings of the 8th IEEE Visualization '97 Conference* (pp. 235–244).
- Crimson Hexagon. (2018). *Data library: A trillion posts can answer a lot of questions*. Retrieved January 17, 2018, from <https://www.crimsonhexagon.com/data-library/>
- D'agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17, 2265–2281. [https://doi.org/10.1002/\(SICI\)1097-0258\(19981015\)17:19<2265::AID-SIM918>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B)
- Data Recovery Group. (2011). *Data storage history and future*. Retrieved January 19, 2018, from <http://www.datarecoverygroup.com/articles/data-storage-history-and-future>
- Dave, P. (2013). *Big Data – what is Big Data – 3 Vs of Big Data – volume, velocity and variety – day 2 of 21*. Retrieved January 25, 2018, from <https://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/>
- Deery, S. (2013). *Ex-girlfriend Adriana Donato's murder planned for weeks by boyfriend using Google searches*. Retrieved January 29, 2018, from <http://www.heraldsun.com.au/exgirlfriend-adriana-donatos-murder-planned-for-weeks-by-boyfriend-using-google-searches/news-story/4ce0d80e6a5582c3782befbd601508cc>
- Deiner, M. S., Lietman, T. M., McLeod, S. D., Chodosh, J., & Porco, T. C. (2016). Surveillance tools emerging from search engines and social media data for determining eye disease patterns. *JAMA Ophthalmology*, 134(9), 1024. <https://doi.org/10.1001/jamaophthalmol.2016.2267>.
- DeVan, A. (2016). *The 7 V's of Big Data*. Retrieved January 25, 2018, from <https://www.impactradius.com/blog/7-vs-big-data/>
- Doctorow, C. (2008). Big Data: Welcome to the petacentre. *Nature*, 455(7209), 16–21. <https://doi.org/10.1038/455016a>
- Dontha, R. (2017). *The origins of Big Data*. Retrieved January 17, 2018, from <https://www.kdnuggets.com/2017/02/origins-big-data.html>
- Drucker, P. (1957). *The landmarks of tomorrow*. Heinemann.
- Dugas, A. F., Hsieh, Y. H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., Gaydos, C. A., Perl, T. M., & Rothman, R. E. (2012). Google Flu Trends: Correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases*, 54(4), 463–469. <https://doi.org/10.1093/cid/cir883>.
- Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., & Rothman, R. E. (2013). Influenza forecasting with Google Flu Trends. *PLoS One*, 8(2). <https://doi.org/10.1371/journal.pone.0056176>.

- Dusetzina, S. B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L., & Carpenter, W. R. (2014). *Linking data for health services research: A framework and instructional guide*. Rockville. [https://doi.org/AHRQ No.14-EHC033](https://doi.org/AHRQ%20No.14-EHC033).
- Emery, S. L., Szczypka, G., Abril, E. P., Kim, Y., & Vera, L. (2014). Are you scared yet? Evaluating fear appeal messages in tweets about the tips campaign. *Journal of Communication*, *64*(2), 278–295. <https://doi.org/10.1111/jcom.12083>.
- Entous, A., Dwoskin, E., & Timberg, C. (2018). *Obama tried to give Zuckerberg a wake-up call over fake news on Facebook*. Retrieved June 1, 2018, from [https://www.washingtonpost.com/business/economy/obama-tried-to-give-zuckerberg-a-wake-up-call-over-fake-news-on-facebook/2017/09/24/15d19b12-ddac-4ad5-ac6e-ef909e1c1284\\_story.html](https://www.washingtonpost.com/business/economy/obama-tried-to-give-zuckerberg-a-wake-up-call-over-fake-news-on-facebook/2017/09/24/15d19b12-ddac-4ad5-ac6e-ef909e1c1284_story.html)
- Eysenbach, G. (2006). Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *Proceedings of the American Medical Informatics Association Annual Symposium 2006* (pp. 244–8). <https://doi.org/PMC1839505>
- Faris, R., Roberts, H., Etling, B., Othman, D., & Benkler, Y. (2015). Score another one for the Internet? The role of the networked public sphere in the U.S. net neutrality policy debate. *SSRN Electronic Journal*, *1*, 0–34. <https://doi.org/10.2139/ssrn.2563761>.
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. <https://doi.org/10.1080/01621459.1969.10501049>.
- Fish, J. N., & Russell, S. T. (2017). Have mischievous responders misidentified sexual minority youth disparities in the National Longitudinal Study of adolescent to adult health? *Archives of Sexual Behavior*, 1–15. <https://doi.org/10.1007/s10508-017-0993-6>.
- Fogarty, C. B., & Small, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association*, *111*(516), 1820–1830. <https://doi.org/10.1080/01621459.2015.1120675>.
- Fond, G., Gaman, A., Brunel, L., Haffen, E., & Llorca, P. M. (2015). Google trends®: Ready for real-time suicide prevention or just a Zeta-Jones effect? An exploratory study. *Psychiatry Research*, *228*(3), 913–917. <https://doi.org/10.1016/j.psychres.2015.04.022>.
- Fuchs, C. (2008). The role of income inequality in a multivariate cross-national analysis of the digital divide. *Social Science Computer Review*, *27*(1), 41–58. <https://doi.org/10.1177/0894439308321628>.
- Fung, K. (2014). *Google Flu Trends' failure shows Good Data > Big Data*. Retrieved January 25, 2018, from <https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data>
- Furu, K., Wettermark, B., Andersen, M., Martikainen, J. E., Almarsdottir, A. B., & Sørensen, H. T. (2010). The Nordic countries as a cohort for pharmacoepidemiological research. *Basic & Clinical Pharmacology & Toxicology*, *106*(2), 86–94. <https://doi.org/10.1111/j.1742-7843.2009.00494.x>.
- Ginsberg, J., & Mohebbi, M. H. (2008). *Tracking Flu Trends*. Retrieved March 5, 2018, from <https://googleblog.blogspot.com.au/2008/11/tracking-flu-trends.html>
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. <https://doi.org/10.1038/nature07634>.
- Gordon, M. (2014). *What is Strava Metro?* Retrieved March 2, 2018, from <https://support.strava.com/hc/en-us/articles/216918877-What-is-Strava-Metro>
- Gunn, J. F., & Lester, D. (2013). Using Google searches on the Internet to monitor suicidal behavior. *Journal of Affective Disorders*, *148*(2–3), 411–412. <https://doi.org/10.1016/j.jad.2012.11.004>.
- Gunnell, D., Bennewith, O., Kapur, N., Simkin, S., Cooper, J., & Hawton, K. (2012). The use of the Internet by people who die by suicide in England: A cross sectional study. *Journal of Affective Disorders*, *141*(2–3), 480–483. <https://doi.org/10.1016/j.jad.2012.04.015>.
- Hachman, M. (2015). *The price of free: How Apple, Facebook, Microsoft and Google sell you to advertisers*. Retrieved January 26, 2018, from <https://www.pcworld.com/article/2986988/privacy/the-price-of-free-how-apple-facebook-microsoft-and-google-sell-you-to-advertisers.html>

- Hagihara, A., Miyazaki, S., & Abe, T. (2012). Internet suicide searches and the incidence of suicide in young people in Japan. *European Archives of Psychiatry and Clinical Neuroscience*, 262(1), 39–46. <https://doi.org/10.1007/s00406-011-0212-8>.
- Harlow, L. L., & Oswald, F. L. (2016). Big Data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4), 447–457. <https://doi.org/10.1037/met0000120>.
- Henderson, V., Storeygard, A., & Weil, D. (2008). Measuring economic growth from outer space. *American Economic Review*, 102(2), 994–1028. <https://doi.org/10.1257/aer.102.2.994>.
- Hern, A. (2018). *Far more than 87m Facebook users had data compromised, MPs told*. Retrieved June 4, 2018, from <https://www.theguardian.com/uk-news/2018/apr/17/facebook-users-data-compromised-far-more-than-87m-mps-told-cambridge-analytica>
- Hilbert, M. (2012a). How much information is there in the “information society”? *Significance*, 9(4), 8–12. <https://doi.org/10.1111/j.1740-9713.2012.00584.x>.
- Hilbert, M. (2012b). How to measure “how much information”? Theoretical, methodological, and statistical challenges for the social sciences. *International Journal of Communication*, 6(1), 1042–1055. <http://ijoc.org/index.php/ijoc/article/view/1318/746>
- Hilbert, M. (2014). What is the content of the world’s technologically mediated information and communication capacity: How much text, image, audio, and video? *Information Society*, 30(2), 127–143. <https://doi.org/10.1080/01972243.2013.873748>.
- Hilbert, M., & López, P. (2011). The world’s technological capacity to store, communicate, and compute information. *Science*, 332(April), 60–65. <https://doi.org/10.1126/science.1200970>.
- Hilbert, M., & López, P. (2012a). How to measure the world’s technological capacity to communicate, store, and compute information, Part I: Results and scope. *International Journal of Communication*, 6(1), 956–979. <http://ijoc.org/index.php/ijoc/article/view/1562/742>
- Hilbert, M., & López, P. (2012b). How to measure the world’s technological capacity to communicate, store, and compute information, Part II: Measurement unit and conclusions. *International Journal of Communication*, 6, 936–955. <http://ijoc.org/index.php/ijoc/article/view/1563/741>
- Hitt, C. (2018). Woman accused of murdering husband searched for “how to kill someone and not get caught” online. Retrieved January 26, 2018, from <http://www.nydailynews.com/news/crime/woman-web-searched-pin-murder-article-1.3753079>
- Holland, T. (2017). *How Facebook and Google changed the advertising game*. Retrieved January 26, 2018, from <https://theconversation.com/how-facebook-and-google-changed-the-advertising-game-70050>
- Hopke, J. E., & Simis, M. (2017a). Discourse over a contested technology on Twitter: A case study of hydraulic fracturing. *Public Understanding of Science*, 26(1), 105–120. <https://doi.org/10.1177/0963662515607725>.
- Hopke, J. E., & Simis, M. (2017b). Response to “word choice as political speech”: Hydraulic fracturing is a partisan issue. *Public Understanding of Science*, 26(1), 124–126. <https://doi.org/10.1177/0963662516643621>.
- Hopkins, D., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1), 229–247. <https://doi.org/10.1111/j.1540-5907.2009.00428.x>.
- Human Rights Council of the United Nations. (2016). *The promotion, protection and enjoyment of human rights on the Internet*, Pub. L. No. A/HRC/32/L.20 (2016). [https://www.article19.org/data/files/Internet\\_Statement\\_Adopted.pdf](https://www.article19.org/data/files/Internet_Statement_Adopted.pdf).
- IBM. (n.d.-a). *Extracting business value from the 4 V’s of big data*. Retrieved January 25, 2018, from <http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data>.
- IBM. (n.d.-b). *Infographic: The four V’s of Big Data*. Retrieved January 25, 2018, from <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>.
- Ingram, D. (2018). *Facebook fuels broad privacy debate by tracking non-users*. Retrieved April 16, 2018, from <https://www.reuters.com/article/us-facebook-privacy-tracking/facebook-fuels-broad-privacy-debate-by-tracking-non-users-idUSKBN1HM0DR>
- Jacobs, A. (2009). The pathologies of Big Data. *Queue*, 7(6), 36–44. <https://doi.org/10.1145/1563821.1563874>.

- Jain, A. (2016). *The 5 Vs of Big Data – Watson health perspectives*. Retrieved January 25, 2018, from <https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/%0A>
- Jinha, A. E. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3), 258–263. <https://doi.org/10.1087/20100308>.
- Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W., Olszewski, R. T., Dowling, J., & Barnas, G. (2004). Analysis of web access logs for surveillance of influenza. *Studies in Health Technology and Informations*, 107(2), 1202–1206. <https://doi.org/10.3233/978-1-60750-949-3-1202>.
- Kamenetz, A. (2014). “Mischievous responders” confound research on teens. Retrieved August 12, 2017, from <https://www.npr.org/sections/ed/2014/05/22/313166161/mischievous-responders-confound-research-on-teens>
- Kelman, C. W., Kortt, M. A., Becker, N. G., Li, Z., Mathews, J. D., Guest, C. S., & Holman, C. D. J. (2003). Deep vein thrombosis and air travel: Record linkage study. *BMJ (Clinical Research Ed.)*, 327(7423), 1072. <https://doi.org/10.1136/bmj.327.7423.1072>.
- Kim, J., Brossard, D., Scheufele, D. A., & Xenos, M. (2016). “Shared” information in the age of Big Data. *Journalism & Mass Communication Quarterly*, 93(2), 430–445. <https://doi.org/10.1177/1077699016640715>.
- King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018), 719–721. <https://doi.org/10.1126/science.1197872>.
- King, G. (2014). Restructuring the social sciences: Reflections from Harvard’s Institute for Quantitative Social Science. *PS: Political Science & Politics*, 47(01), 165–172. <https://doi.org/10.1017/S1049096513001534>.
- King, G., & Persily, N. (2018). *A new model for industry-academic partnerships*. <http://j.mp/2q1IQpH>
- Kramer, A. D. I. (2014). *Facebook post by A Kramer*. Retrieved July 1, 2014, from <https://www.facebook.com/akramer/posts/10152987150867796>
- Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences USA*, 111(24), 8788–8790. <https://doi.org/10.1073/pnas.1412469111>.
- Kristoufek, L., Moat, H. S., & Preis, T. (2016). Estimating suicide occurrence statistics using Google trends. *EPJ Data Science*, 5(1), 32. <https://doi.org/10.1140/epjds/s13688-016-0094-0>.
- La Rue, F. (2011). *Report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression*. Frank La Rue. [http://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27\\_en.pdf](http://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf)
- Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *Application Delivery Strategies*, 949(February 2001), 4. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>
- Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3), 575–603. <https://doi.org/10.1007/s11192-010-0202-z>.
- Lazer, D., & Kennedy, R. (2015). *What we can learn from the epic failure of Google Flu Trends*. Retrieved January 25, 2018, from <https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/>
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014a). The parable of Google Flu: Traps in Big Data analysis. *Science*, 343(6176), 1203–1205. <https://doi.org/10.1126/science.1248506>.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2014b). Computational social science. *Science*, 323, 721–723. <https://doi.org/10.1126/science.1169410>.
- Leathern, R. (2018). *Shining a light on ads with political content*. Retrieved June 2, 2018, from <https://newsroom.fb.com/news/2018/05/ads-with-political-content/>
- Lesk, M. (1997). *How much information is there in the world?* Retrieved December 8, 2017, from <http://www.lesk.com/mlesk/ksg97/ksg.html>

- Lester, D. (2009). The Nordic prescription databases as a resource for pharmacoepidemiological research—A literature review. *Clinical Neuropsychiatry*, 6(5), 188–191. <https://doi.org/10.1002/pds>.
- Lewis, P. (2018). “Utterly horrifying”: Ex-Facebook insider says covert data harvesting was routine. Retrieved March 21, 2018, from <https://www.theguardian.com/news/2018/mar/20/facebook-data-cambridge-analytica-sandy-parakilas>
- Lewis, P., & Carrie Wong, J. (2018). Facebook employs psychologist whose firm sold data to Cambridge Analytica. Retrieved June 4, 2018, from <https://www.theguardian.com/news/2018/mar/18/facebook-cambridge-analytica-joseph-chancellor-gsr>
- Li, N., Akin, H., Yi-Fan, L. S., Brossard, D., Xenos, M., & Scheufele, D. A. (2016). Tweeting disaster: An analysis of online discourse about nuclear power in the wake of the Fukushima Daiichi nuclear accident. *Journal of Science Communication*, 15(5), 1–20. [https://jcom.sissa.it/archive/15/05/JCOM\\_1505\\_2016\\_A02](https://jcom.sissa.it/archive/15/05/JCOM_1505_2016_A02)
- Library of Congress. (2017). *General information*. Retrieved January 17, 2018, from <https://www.loc.gov/about/general-information/>
- Lohr, S. (2013, February 1). The origins of “Big Data”: An etymological detective story. *The New York Times*. <https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/>
- London’s Tube and Big Data: Underground movement. (2013). Retrieved January 26, 2018, from <https://www.newscientist.com/article/in415-londons-tube-and-big-data-underground-movement/>
- Ma-Kellams, C., Or, F., Baek, J. H., & Kawachi, I. (2016). Rethinking suicide surveillance: Google search data and self-reported suicidality differentially estimate completed suicide risk. *Clinical Psychological Science*, 4(3), 480–484. <https://doi.org/10.1177/2167702615593475>.
- Mabe, M., & Amin, M. (2001). Growth dynamics of scholarly and scientific journals. *Scientometrics*, 51(1), 147–162. <https://link.springer.com/article/10.1023/A:1010520913124>
- Madrigal, A. C. (2014). *In defense of Google Flu Trends*. Retrieved January 25, 2018, from <https://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/>
- Malik, M. T., Gumel, A., Thompson, L. H., Strome, T., & Mahmud, S. M. (2011). “Google Flu Trends” and emergency department triage data predicted the 2009 pandemic H1N1 waves in Manitoba. *Canadian Journal of Public Health*, 102(4), 294–297. <http://www.jstor.org/stable/41995614>
- Marr, B. (2014). *Big Data—The 5V everyone must know*. Retrieved January 25, 2018, from <https://www.slideshare.net/BernardMarr/140228-big-data-volume-velocity-variety-varacity-value%0A>
- Mars, B., Heron, J., Biddle, L., Donovan, J. L., Holley, R., Piper, M., Potokar, J., Wyllie, C., & Gunnell, D. (2015). Exposure to, and searching for, information about suicide and self-harm on the Internet: Prevalence and predictors in a population based cohort of young adults. *Journal of Affective Disorders*, 185, 239–245. <https://doi.org/10.1016/j.jad.2015.06.001>.
- Mashey, J. R. (1998). *Big Data and the next wave of infraStress*. University of California, Berkeley. [http://static.usenix.org/event/usenix99/invited\\_talks/mashey.pdf](http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf)
- Matsa, K. E., Mitchell, A., & Stocking, G. (2017). *Methodology*. Retrieved February 26, 2018, from <http://www.journalism.org/2017/04/27/google-flint-methodology/>
- McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of Affective Disorders*, 122(3), 277–279. <https://doi.org/10.1016/j.jad.2009.08.015>.
- McNulty, E. (2014). *Understanding Big Data: The seven V’s*. Retrieved January 25, 2018, from <http://dataconomy.com/2014/05/seven-vs-big-data/>
- Mellish, L., Karanges, E. A., Litchfield, M. J., Schaffer, A. L., Blanch, B., Daniels, B. J., Segrave, A., & Pearson, S.-A. (2015). The Australian pharmaceutical benefits scheme data collection: A practical guide for researchers. *BMC Research Notes*, 8(1), 634. <https://doi.org/10.1186/s13104-015-1616-8>.
- Mellon, J. (2013). Where and when can we use Google trends to measure issue salience? *PS: Political Science & Politics*, 46(02), 280–290. <https://doi.org/10.1017/S1049096513000279>.

- Moe, W. W., & Schweidel, D. A. (2017). Opportunities for innovation in social media analytics. *Journal of Product Innovation Management*, 34(5), 697–702. <https://doi.org/10.1111/jpim.12405>.
- Monnappa, A. (2017). *How Facebook is using Big Data – The good, the bad, and the ugly*. Retrieved January 26, 2018, from <https://www.simplilearn.com/how-facebook-is-using-big-data-article>
- Moore, G. E. (1965). Cramping more components onto integrated circuits. *Electronics*, 38(8), 114–117. <https://doi.org/10.1109/N-SSC.2006.4785860>.
- Moriarty, M. (2017). *15 Things you can learn from 1 trillion Posts: How 1,000,000,000,000 posts can change the world we live in*. Retrieved January 17, 2018, from <https://www.crimsonhexagon.com/blog/15-things-you-can-learn-from-1-trillion-posts/>
- Mullin, J. (2012). *How much do Google and Facebook profit from your data?* Retrieved January 26, 2018, from <https://arstechnica.com/tech-policy/2012/10/how-much-do-google-and-facebook-profit-from-your-data/>
- Nanji, A. (2017). *The most popular social networks with Millennials, Gen X and Baby Boomers*. Retrieved March 2, 2018, from <https://www.marketingprofs.com/charts/2017/31792/the-most-popular-social-networks-with-millennials-gen-x-and-baby-boomers>
- National Highway Traffic Safety Administration. (2016). *Fatality Analysis Reporting System (FARS): Analytical User's Manual 1975–2015* (No. DOT HS 812 315). [ftp://ftp.nhtsa.dot.gov/FARS/FARS-DOC/Analytical User Guide/USERGUIDE-2015.pdf](ftp://ftp.nhtsa.dot.gov/FARS/FARS-DOC/Analytical%20User%20Guide/USERGUIDE-2015.pdf).
- Obar, J. A., & Oelof-Hirsch, A. (2016). The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. In *The 44th Research Conference on Communication, Information and Internet Policy 2016*. <https://doi.org/10.2139/ssrn.2757465>
- Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9(10). <https://doi.org/10.1371/journal.pcbi.1003256>.
- Ong, T. C., Mannino, M. V., Schilling, L. M., & Kahn, M. G. (2014). Improving record linkage performance in the presence of missing linkage data. *Journal of Biomedical Informatics*, 52, 43–54. <https://doi.org/10.1016/j.jbi.2014.01.016>.
- Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., & Goss, C. H. (2011). Monitoring influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends. *PLoS One*, 6(4), 2–10. <https://doi.org/10.1371/journal.pone.0018687>.
- Ortutay, B. (2018). *Facebook scandal affected more users than thought: Up to 87M*. Retrieved April 5, 2018, from <https://www.apnews.com/e0e0df2083fe40c0b0ad10ff1946f041>
- Ortutay, B., Kirka, D., & Katz, G. (2018). *Facebook's Zuckerberg apologizes for 'major breach of trust'*. Retrieved March 22, 2018, from <https://apnews.com/c8f615be9523421998b4fcc16374ff37>
- Page, A., Chang, S.-S., & Gunnell, D. (2011). Surveillance of Australian suicidal behaviour using the Internet? *Australian and New Zealand Journal of Psychiatry*, 45(12), 1020–1022. <https://doi.org/10.3109/00048674.2011.623660>.
- Paige, E., Kemp-Casey, A., Korda, R., & Banks, E. (2015). Using Australian Pharmaceutical Benefits Scheme data for pharmacoepidemiological research: Challenges and approaches. *Public Health Research & Practice*, 25(4), 1–6. <https://doi.org/10.17061/phrp2541546>.
- Parker, J., Cuthbertson, C., Loveridge, S., Skidmore, M., & Dyar, W. (2017). Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google trends data. *Journal of Affective Disorders*, 213(November 2016), 9–15. <https://doi.org/10.1016/j.jad.2016.10.038>
- Pearson, S.-A., Pesa, N., Langton, J. M., Drew, A., Faedo, M., & Robertson, J. (2015). Studies using Australia's Pharmaceutical Benefits Scheme data for pharmacoepidemiological research: A systematic review of the published literature (1987–2013). *Pharmacoepidemiology and Drug Safety*, 24, 447–455. <https://doi.org/10.1002/pds.3756>.
- Pingdom. (2008). *The history of computer data storage, in pictures*. Retrieved January 19, 2018, from <http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/>

- Plume, A., & van Weijen, D. (2014). *Publish or perish? The rise of the fractional author...* *Research Trends*, 38. <https://www.researchtrends.com/issue-38-september-2014/publish-or-perish-the-rise-of-the-fractional-author/>
- Polgreen, P. M., Chen, Y., Pennock, D. M., & Nelson, F. D. (2008). Using Internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11), 1443–1448. <https://doi.org/10.1086/593098>.
- Press, G. (2013). *A very short history of Big Data*. Retrieved December 8, 2017, from <https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1eaca84c65a1>
- Prodromou, T., & Dunne, T. (2017). Data visualisation and statistics education in the future. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and Big Data* (pp. 1–28). : IGI Global. <https://doi.org/10.4018/978-1-5225-2512-7.ch001>.
- Przybylski, A. K. (2016). Mischievous responding in internet gaming disorder research. *PeerJ*, 4, e2401. <https://doi.org/10.7717/peerj.2401>.
- Puang-ngern, B., Bilgin, A. A., & Kyng, T. J. (2017). Comparison of graduates' and academics' perceptions of the skills required for Big Data analysis: Statistics education in the age of Big Data. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and Big Data* (pp. 126–152). : IGI Global. <https://doi.org/10.4018/978-1-5225-2512-7.ch006>.
- Rassen, J. A., Glynn, R. J., Brookhart, M. A., & Schneeweiss, S. (2011). Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American Journal of Epidemiology*, 173(12), 1404–1413. <https://doi.org/10.1093/aje/kwr001>.
- Raubenheimer, J. E. (2019). Google Trends Extraction Tool. <https://doi.org/10.5281/zenodo.2620618>
- Raubenheimer, J. E. (2021). Google Trends Extraction Tool for Google Trends Extended for Health data. *Software Impacts*, 8, 100060. <https://doi.org/10.1016/j.simpa.2021.100060>
- Reuters. (2018). Cambridge Analytica and British parent shut down after Facebook scandal. Retrieved May 3, 2018, from <https://www.reuters.com/article/us-facebook-privacy/cambridge-analytica-and-british-parent-shut-down-after-facebook-scandal-idUSKBN1132L7>
- Robb, D. (2017). *The Global Heatmap, now 6x hotter*. Retrieved January 1, 2018, from <https://medium.com/strava-engineering/the-global-heatmap-now-6x-hotter-23fc01d301de>
- Roberts, M. E., Stewart, B. M., & Nielsen, R. (2015). *Matching methods for high-dimensional data with applications to text*. <http://www.margaretroberts.net/wp-content/uploads/2015/07/textmatching.pdf>
- Robinson-Cimpian, J. P. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher*, 43(4), 171–185. <https://doi.org/10.3102/0013189X14534297>.
- Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1), 13–26. <http://www.jstor.org/stable/2336017>
- Rosenbaum, P. R. (1989). Sensitivity analysis for matched observational studies with many ordered treatments. *Scandinavian Journal of Statistics*, 16(3), 227–236. <http://www.jstor.org/stable/4616136>
- Runge, K. K., Yeo, S. K., Cacciatore, M., Scheufele, D. A., Brossard, D., Xenos, M., Anderson, A., Choi, D. H., Kim, J., Li, N., Liang, X., Stubbings, M., & Su, L. Y. F. (2013). Tweeting nano: How public discourses about nanotechnology develop in social media environments. *Journal of Nanoparticle Research*, 15(1). <https://doi.org/10.1007/s11051-012-1381-8>.
- Salsburg, D. S. (2017). *Errors, blunders, and lies: How to tell the difference*. CRC Press.
- Salzberg, S. (2014). *Why Google Flu is a failure*. Retrieved January 25, 2018, from <https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/#42fed4945535>
- Sari Aslama, N., Cheshire, B. J., & Cheng, T. (2015). *Big Data analysis of population flow between TfL oyster and bicycle hire networks in London*. University College London. [http://leeds.gis-ruk.org/abstracts/GISRUK2015\\_submission\\_92.pdf](http://leeds.gis-ruk.org/abstracts/GISRUK2015_submission_92.pdf)
- Schaffer, A. L., Buckley, N. A., Dobbins, T. A., Banks, E., & Pearson, S.-A. (2015). The crux of the matter: Did the ABC's catalyst program change statin use in Australia? *Medical Journal of Australia*, 11(11), 591–595. <https://doi.org/10.5694/mja15.0010>.

- Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety*, 15, 291–303. <https://doi.org/10.1002/pds.1200>.
- Schrage, E., & Ginsberg, D. (2018). *Facebook launches new initiative to help scholars assess social media's impact on elections*. Retrieved June 2, 2018, from <https://newsroom.fb.com/news/2018/04/new-elections-initiative/>
- Scurr, J. H., Machin, S. J., Bailey-King, S., Mackie, I. J., McDonald, S., & Coleridge Smith, P. D. (2001). Frequency and prevention of symptomless deep vein thrombosis in long-haul flights: A randomised trial. *Lancet*, 357, 1485–1489. <https://www.thelancet.com/journals/lancet/article/PIIS0140673600046456/abstract>
- Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ (Clinical Research Ed.)*, 327(7429), 1459–1461. <https://doi.org/10.1177/154510970400300401>.
- Solano, P., Ustulin, M., Pizzorno, E., Vichi, M., Pompili, M., Serafini, G., & Amore, M. (2016). A Google-based approach for monitoring suicide risk. *Psychiatry Research*, 246, 581–586. <https://doi.org/10.1016/J.PSYCHRES.2016.10.030>.
- Song, T. M., Song, J., An, J. Y., Hayman, L. L., & Woo, J. M. (2014). Psychological and social factors affecting Internet searches on suicide in Korea: A Big Data analysis of Google search trends. *Yonsei Medical Journal*, 55(1), 254–263. <https://doi.org/10.3349/ymj.2014.55.1.254>.
- Spielberg, S. (2002). *Minority Report*. USA: Twentieth Century Fox. <http://www.imdb.com/title/tt0181689>
- Stephens-Davidowitz, S. (2017). *Everybody lies*. HarperCollins.
- Sueki, H. (2011). Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: Data from 2004 to 2009 in Japan. *Psychiatry and Clinical Neurosciences*, 65(4), 392–394. <https://doi.org/10.1111/j.1440-1819.2011.02216.x>.
- Taleb, N. (2013). *Beware the big errors of "Big Data."* Retrieved December 8, 2017, from <https://www.wired.com/2013/02/big-data-means-big-errors-people/>
- The Flu Trends Team. (2015). *The next chapter for Flu Trends*. Retrieved January 25, 2018, from <https://research.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html>
- The Statistics Portal. (2018). *Global shipments of hard disk drives (HDD) from 4th quarter 2010 to 3rd quarter 2017 (in millions)*. Retrieved January 22, 2018, from <https://www.statista.com/statistics/275336/global-shipment-figures-for-hard-disk-drives-from-4th-quarter-2010/>
- Thomas, R., & McSharry, P. (2015). *Big Data revolution: What farmers, doctors and insurance agents teach us about discovering Big Data patterns*. John Wiley & Sons.
- Tran, U. S., Andel, R., Niederkrotenthaler, T., Till, B., Ajdacic-Gross, V., & Voracek, M. (2017). Low validity of Google trends for behavioral forecasting of national suicide rates. *PLoS One*, 12(8), 1–26. <https://doi.org/10.1371/journal.pone.0183149>.
- Tromp, M., Ravelli, A. C., Bonsel, G. J., Hasman, A., & Reitsma, J. B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology*, 64(5), 565–572. <https://doi.org/10.1016/j.jclinepi.2010.05.008>.
- Turriago-Hoyos, A., Thoene, U., & Arjoon, S. (2016). Knowledge workers and virtues in Peter Drucker's management theory. *SAGE Open*, 6(1). <https://doi.org/10.1177/2158244016639631>.
- Ueda, M., Mori, K., Matsubayashi, T., & Sawada, Y. (2017). Tweeting celebrity suicides: Users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. *Social Science and Medicine*, 189, 158–166. <https://doi.org/10.1016/j.socscimed.2017.06.032>.
- Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences USA*, 109(16), 5962–5966. <https://doi.org/10.1073/pnas.1116502109>.
- UN Global Pulse. (2014). *Mining Indonesian tweets to understand food price crises*. Jakarta. <https://www.unglobalpulse.org/projects/social-media-social-protection-indonesia>
- Valdivia, A., Lopez-Alcalde, J., Vicente, M., Pichiule, M., Ruiz, M., & Ordobas, M. (2010). Monitoring influenza activity in Europe with Google Flu Trends: Comparison with the findings



- of sentinel physician networks—Results for 2009–10. *Euro Surveillance*, 15(29), 1–6. <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19621>
- Vaughan-Nichols, S. (2014). *We're all just lab rats in Facebook's laboratory*. Retrieved January 25, 2018, from <http://www.zdnet.com/article/were-all-just-lab-rats-in-facebooks-laboratory/>
- Verma, I. M. (2014). Editorial expression of concern and correction. *Proceedings of the National Academy of Sciences USA*, 111(29), 10779. [www.pnas.org/cgi/doi/10.1073/pnas.1412469111](http://www.pnas.org/cgi/doi/10.1073/pnas.1412469111).
- Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939), 425–428. <https://doi.org/10.1126/science.1171990>.
- Walsh, B. (2014). *Google's Flu project shows the failings of Big Data*. Retrieved January 25, 2018, from <http://time.com/23782/google-flu-trends-big-data-problems/>
- Ware, M., & Mabe, M. (2009). The STM report: An overview of scientific and scholarly journal publishing. <http://www.markwareconsulting.com/institutional-repositories/the-stm-report-an-overview-of-scientific-and-scholarly-journal-publishing/>
- Ware, M., & Mabe, M. (2012). The STM report: An overview of scientific and scholarly journal publishing (3rd ed). [http://www.stm-assoc.org/2012\\_12\\_11\\_STM\\_Report\\_2012.pdf](http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf)
- Ware, M., & Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing (4th ed). [http://www.stm-assoc.org/2015\\_02\\_20\\_STM\\_Report\\_2015.pdf](http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf)
- Wartzman, R. (2014). *What Peter Drucker knew about 2020*. Retrieved January 23, 2018, from <https://hbr.org/2014/10/what-peter-drucker-knew-about-2020>
- Wettermark, B., Zoëga, H., Furu, K., Korhonen, M., Hallas, J., Nørgaard, M., Almarsdóttir, A. B., Andersen, M., Andersson Sundell, K., Bergman, U., Helin-Salmivaara, A., Hoffmann, M., Kieler, H., Martikainen, J. E., Mortensen, M., Petzold, M., Wallach-Kildemoes, H., Wallin, C., & Sørensen, H. (2013). The Nordic prescription databases as a resource for pharmacoepidemiological research—A literature review. *Pharmacoepidemiology and Drug Safety*, 22(7), 691–699. <https://doi.org/10.1002/pds.3457>.
- Why the 3V's are not sufficient to describe Big Data. (2015). Retrieved January 25, 2018, from <https://datafloq.com/read/3vs-sufficient-describe-big-data/166>
- Wikipedia. (2018). *Wikipedia: Database download*. Retrieved January 31, 2018, from [wikipedia.org/wiki/Wikipedia:Database\\_download](http://wikipedia.org/wiki/Wikipedia:Database_download).
- Wilson, N., Mason, K., Tobias, M., Peacey, M., Huang, Q. S., & Baker, M. (2009). Interpreting “Google Flu Trends” data for pandemic H1N1 Influenza: The New Zealand experience. *Euro Surveillance*, 14(44), 1–3. <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19386>
- Winkler, W. E. (1993). *Matching and record linkage*. US Census Bureau – Research Reports. <https://www.census.gov/srd/papers/pdf/r93-8.pdf>
- Winkler, W. E. (2014). Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(5), 313–325. <https://doi.org/10.1002/wics.1317>.
- Yang, A. C., Tsai, S. J., Huang, N. E., & Peng, C. K. (2011). Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004–2009. *Journal of Affective Disorders*, 132(1–2), 179–184. <https://doi.org/10.1016/j.jad.2011.01.019>.
- Youtie, J., Porter, A. L., & Huang, Y. (2017). Early social science research about Big Data. *Science and Public Policy*, 44(1), scw021. <https://doi.org/10.1093/scipol/scw021>.

**Jacques Raubenheimer** completed a PhD in Research Psychology at the University of the Free State, South Africa, then worked as a research consultant, before beginning a career as a lecturer in biostatistics in 2008. He moved from South Africa to Australia at the start of 2017 to work as a research fellow with the NHMRC-funded Translational Australian Clinical Toxicology programme. His work focuses on analysing one of the largest cohorts of poisoning admissions data in the world (with data from 2001 to the present), and working with national pharmaceutical dispensing datasets from the Australian population. His current aim is to establish methods for using internet search- and social media data for the monitoring and nowcasting of trends in drug interest, particularly novel substances and substances slipping under the radar of conventional regulation, such as Human Enhancement Drugs (HEDs).

**Part II**  
**Teaching People to Use Big Data**  
**Effectively**

## Chapter 2

# Big Data for Early Learners



Peter Howley, Kevin Wang, and Ayse Aysin Bilgin

**Abstract** There has been an explosion in the availability of digital data and an ever-increasing need for analysts in the big data arena; the gap between demand and supply is widening. Big data, data science and the field of statistics are pervasive; every discipline across industry, business, government and education is increasingly generating copious data and combining data from multiple sources. Future generations will need to be familiar with data and aspects of investigation and communication so they may participate as socially responsible citizens in the data-rich society. Concurrently, research reveals that an individual's aspirations and occupational futures may be shaped at a much earlier age than previously considered.

This chapter provides introductory approaches for engaging early learners in the big data arena. It articulates by example the broad application of the field, and provides relatively straight forward yet appealing interactive activities, and support resources, to develop rudimentary but essential concepts and processes for each of the 4–7 years and 8–18 years age groups. A combination of activities and freely available software to support data exploration such as iNZight is drawn upon to incite interest and appeal.

**Keywords** Early childhood · Primary education · Secondary education · Aspirations · Big Data · Data science · Statistics

---

P. Howley (✉)

School of Mathematical and Physical Sciences/Statistics, The University of Newcastle,  
Callaghan, NSW, Australia

e-mail: [peter.howley@newcastle.edu.au](mailto:peter.howley@newcastle.edu.au)

K. Wang

Centre for Epidemiology and Evidence, NSW Ministry of Health, St Leonards, Australia

e-mail: [Kevin.Wang1@health.nsw.gov.au](mailto:Kevin.Wang1@health.nsw.gov.au)

A. A. Bilgin

Department of Statistics, Macquarie University, Macquarie Park, NSW, Australia

e-mail: [ayse.bilgin@mq.edu.au](mailto:ayse.bilgin@mq.edu.au)

© Springer Nature Switzerland AG 2021

T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy

Implications of Research in Education 13,

[https://doi.org/10.1007/978-3-030-76841-6\\_2](https://doi.org/10.1007/978-3-030-76841-6_2)

## Introduction

It's never too early to learn, and this is certainly the case for the world of big data. We are increasingly exposed to copious data from multiple sources that shape our decisions and the decisions of those around us. There is, in parallel, an increasing motivation to prepare the next generation for this data-rich future, by engaging them at early ages to appreciate and feel familiar with data exploration and concepts important in data investigation. Affording this opportunity to students provides them a critical intellectual tool for engagement and participation as socially responsible citizens.

There are reports identifying the relative shortfall of adequately skilled individuals in the new big data age (Manyika et al., 2011; Puang-Ngern et al., 2017). To fill the increasing need for such big data skills, students must engage with and become more greatly interested in the analytical, statistical and machine-learning aspects.

The interpretation of data is ubiquitous in the knowledge society and consequently statistical literacy assumes an increasingly prevalent role in education at all levels. However, due to common perceptions of statistics and the general appeal, or lack thereof, of the subject, extra efforts to overcome initial hurdles to gain student interest or engagement is required (Howley, 2008, 2009). Statistical anxiety is increasingly a major issue which has been identified as a deterrent for young people to access and succeed in university degrees (Onwuegbuzie & Wilson, 2003; Williams, 2010).

This chapter explores introductory approaches for engaging early learners in the Big Data arena. Relatively simple interactive activities, and support resources, to develop rudimentary but essential concepts and processes are provided for the 4–7 years and 8–18 years age groups.

The methods utilise aspects of active learning, making it memorable, Vygotsky's zone of proximal learning, breaking down barriers, scaffolded building upon students' intuitions and abilities, and making the field accessible and appealing (Berk & Winsler, 1995; Bryce, 2005; Howley, 2008; Sowey, 1995, 2001). Whilst Big Data alludes to not only large but additionally complex data sets, and the combining of such across multiple sources, its exploratory basis is formed upon certain key principles that are rarely expressly taught in schools. The latter more often focus on the mechanics of calculations than developing the particular conceptual understanding.

Exploring data, especially tangible and contextualised data from areas of interest to the specific audience, and building upon the intuitive knowledge of young learners, are examples of key aspects that will facilitate a trajectory for embracing big data. Importantly, the chapter begins with some contextualisation of the field, clarifying the broad application of big data, statistics and data science in practice. It then provides some very accessible and achievable methods for introducing data exploration and revealing patterns or information, from the perspective of statistics. It focuses on conceptual development of data exploration, investigation and turning data into information. A combination of activities and freely available software to

support data exploration such as *iNZight* is drawn upon to incite interest and appeal. The chapter aims to inspire the reader to create additional methods for introducing the topic to early learners and for ways of educating about and within the field.

Creating a mindset that supports a trajectory of exploration, creativity and investigation is key to supporting the development of future generations able to combine their field of interest with big data needs. With research revealing that an individual's aspirations and occupational futures may be shaped at a much earlier age than previously considered (Gore et al., 2015), the need to introduce ideas early and often becomes all the more urgent.

## Key Resources Referenced Within the Chapter

This chapter references two key supporting resources:

1. *iNZight*
2. *StatsTuneUp*

### About *iNZight*

*iNZight* is a free online software that was initially designed for engaging school students and teachers interested in quick, easy and appropriate data explorations, particularly the creation of visual displays. It has extended to include multivariate graphs, time series and generalised linear modelling (including modelling of data from complex data sets) and continues to evolve. It is available from [www.stat.auckland.ac.nz/~wild/iNZight/getinzight.php](http://www.stat.auckland.ac.nz/~wild/iNZight/getinzight.php).

The initiative is led by Professor Chris Wild, Department of Statistics, University of Auckland and continues to be developed thanks to his ongoing support team, [www.stat.auckland.ac.nz/~wild/iNZight/about/team/](http://www.stat.auckland.ac.nz/~wild/iNZight/about/team/).

Getting involved with data is a critically important step towards developing one's understanding of big data, statistics and data science. This software facilitates the introduction needed by many, instantly producing appropriate graphs for example rather than the user needing to specify such graphs. Seeing data come to life visually is invaluable as part of exploring, interpreting and communicating. This chapter introduces the software package along with some tutorial exercises that facilitate data exploration.

## About *StatsTuneUp*

This is a free online set of resources primarily designed to benefit teenage and mature students, and *teachers* of young, teenage and mature students, available from [www.statstuneup.com.au](http://www.statstuneup.com.au). They provide a friendly, practical and simple audio-visual animated introduction to fundamental statistical concepts and methods, aimed at increasing the appeal and accessibility of the field.

The creation of these resources was led by then Prof Peter Howley, The University of Newcastle [www.newcastle.edu.au/profile/peter-howley](http://www.newcastle.edu.au/profile/peter-howley). Each of the 12 animated videos is only a few minutes long and provides a short and supportive overview of a key statistical concept or method. Associated practice questions (with answers) and additional information are also provided. These resources complement the novice reader's self-development and their preparation of activities for stimulating young minds. The *iNZight* worksheets presented later in this chapter reference the *StatsTuneUp* videos and resources.

## Why Big Data, Statistics and Data Science?

*“Data is the new oil”*

The term “*big data*” is generally considered as involving large and complex data sets, and the combining of these across multiple sources, where traditional data storage and processing methods may prove inadequate. One question for educators is how to engage young or inexperienced students to enter into this complex domain. As with any field, there are some fundamental principles and concepts that provide the foundations for the processes, data exploration and logic utilised within big data. This chapter focusses on these, in particular through the lens of statistics, and at an introductory level developing conceptual understanding.

Overwhelmingly data has become a key commodity. Investigations associated with data interrogation and analysis are arguably a mix of science, coding and creativity. Accordingly applications are found across all fields. Some brief examples include:

- market researchers, financial institutions and telecommunication system organisations which aim to identify patterns in consumer behaviour and thus engage in targeted marketing in order to improve returns on advertising investments;
- forensic anthropologists attempting to profile individuals, to determine a person's height or sex based on bones recovered, or to identify what happened based on blood splatter at crime scenes;
- data linkage systems such as the Centre for Health Record Linkage, see [www.cherel.org.au](http://www.cherel.org.au), used by public health professionals, policy makers and investigators who need to monitor and understand national health (from routine hospital use to potential outbreaks), assess the effectiveness of health policies and interventions, and develop supportive and resilient health care systems;

- genetics and DNA sequencing and nanotechnologies which yield copious amounts of data to be assessed for ever-increasing knowledge and development for mankind.

Retail organisations' loyalty programs enable the collation of data about customers and their purchasing behaviours; for example, their demographics and when and what they buy. From such data, customer behaviours or purchasing patterns (whether it be day of week, types of purchases, and so forth) are identified. Customers are categorised (or described) by their purchasing habits and marketing strategies may be devised to better target individuals (or individualise marketing strategies). The linking of information either across multiple customers and/or from varying sources (other organisations' loyalty programs for example) further builds the 'database' from which customer behaviours are essentially categorised. Many people don't realize that when they sign up for loyalty programs or provide information when making purchases, that they are usually giving the organisation permission to provide or on-sell their personal data to other organisations.

This is similarly the case when browsing the web. An internet user will likely notice certain advertisements more consistently appearing on their browser. This is 'the robots' at work – that is the algorithms developed to collate information about your browsing behaviours, which is based on you showing interest in (tapping on links for) certain types of information (for restaurants, hotels, cars, etc.) or making online purchases. These bits of information help form an understanding of the 'type' of person you are, what you have shown interest in and hence what 'additional information' (items or services for sale) may pique your interest.

Credit card fraud-prevention processes and algorithms operate in a similar vein, by understanding customer behaviours (purchasing patterns) to establish a baseline with which to then compare ongoing purchases, and ultimately identify a purchase that is unlike previous patterns. An extreme example would be for someone who used their credit card consistently for grocery shopping only in Toowoomba, Australia, to then have purchases of a diamond ring in New York, USA, several large screen televisions in Africa and movie tickets in London all within a few hours of their grocery shop that week.

Big Data is also very helpful in public health surveillance. By linking ambulance, emergency department and admitted patient (hospitalisation) data, a comprehensive understanding of 'patient journeys' in the health system is developed. This helps governments and policy makers understand the strengths and weaknesses of the system and where resources may be better directed.

As a peek into the future, consider automated car taxiing services that have learnt your primary times of need, and thus have cars in the area in preparation for your call – no longer will you have to wait as long for the service. And of course, they are in an interconnected network of information, knowing not only your preferences but those around you, and hence traffic routes more commonly taken by others. The combining of information from all of these sources supports the optimisation of the service – this is another example of Big Data systems at work.

So we understand how data is a commodity, and pervasive.

Throughout all of these examples, there is one key underlying aspect, namely the collection of data to reveal associations and, in many cases, patterns, or categories, of behaviour. More broadly and simply, it is about the establishment of key information to describe a situation from amongst the sea of data.

## **Stimulating Young Minds – Making it Accessible and Appealing**

Whilst the above helps contextualise the value of statistics, data science and big data, this chapter aims to discuss strategies/techniques/methods/exercises to develop interest and ability in novice readers and teachers of younger audiences, such that the field feels increasingly accessible, and their interest to know more and explore is piqued, either overtly or unwittingly so.

So, how do we help stimulate interest and set a trajectory of learning within the statistics, data science and big data arena from an early age? Nurturing the inquiring minds of the young is important. Enabling them to participate, take the lead, explore, investigate whilst providing the appropriate level of scaffolding to support learning and development are key.

The remainder of this chapter focuses on two activities designed on such a premise, as well as describing the use of *iNZight* and *StatsTuneUp* as supporting resources. The first activity was demonstrated by the first co-author with a pre-school audience and would suit audiences who are aged 4 to 7 years. The second was created by the first co-author and is currently engaging hundreds of Grade 3 to 12 students (aged 8 to 18) annually in Australia across all disciplines, and thousands of Grade 3 to 12 students as part of an international collaboration.

## **Making the Connection**

The following provides a practical example which is aimed for presentation to school students 4 to 7 years of age. It is based on a presentation given to a group of pre-schoolers. It relies upon analogous situations to the ‘large and complex data’ and ‘teasing out and communicating key information’ surrounding statistics and big data – on a scale accessible to the young. The approaches adopted are aligned with those in Howley (2008), “*Keeping it real, keeping them interested and keeping in their minds*”. Readers interested in developing approaches for teaching the field to more mature audiences, or better understanding some elements of the field themselves, may like to consider that article, and those it references. Instilling certain conceptual knowledge and behaviours is the key at such young ages to unlocking the skills and setting young students on a trajectory towards engaging with big data.



The description of the activity and the responses from the early childhood class are punctuated with statements, situated within a rectangle, of a key message or lesson/skill being developed. The activity is described from the first-person perspective (the chapter's first co-author), identifying the key actions and words spoken.

### ***A Bag of Leaves Brings Statistics and Data Science to the Classroom***

The aim of the visit to the pre-school class was to describe aspects of my job as a statistician and university professor of statistics, and help those in attendance become aware of some valuable underlying concepts in the field. Before reading ahead, I'd recommend pausing to consider how you may approach such, or how you perceive the profession yourself or may describe the profession and its practice to even a more mature audience.

Armed with a clear plastic snap-lock bag of items from my garden I ventured into the pre-school classroom.

#### **Key Message (Aspect Being Developed)**

Enable the audience to participate often and early – thus increasing the appeal and perceived accessibility.

After a brief introduction, I began with some questions about whether they knew any teachers (knowing that they all did of course, their pre-school teachers) as part of inspiring their participation. All hands went up of course. I proceeded to ask if anyone had heard of 'big school' (as they call it) and then about how you go on to learn more things at primary school and high school, punctuating these questions (and their responses) with the news that there are teachers at each of these levels of education (not using that last phrasing of course).

I then mentioned how some people want to learn more after completing school and asked if anyone knew of where adults may go for such additional learning. To my delight, university was one of the responses, with many others saying they know of university...a great segue for me, and where I was heading. I told them that I am a teacher, at university.

From there, I mentioned that as well as being a teacher at university, I am also a statistician. I asked if they could say 'statistician'. They did, and better than many adults! "(Why do adults want to say 'statistician' (missing the second 's')? But I digress.)"

So, what is it that I do as a statistician, I asked rhetorically. I proceeded to ask the children whether anyone had ever gone to a hospital, or had to take medicine. All hands went up, as expected, and some wanted to share their experience too. I asked whether the medicine made them feel better and they all said that it had. So,

where was I heading with this? I then described how people need to test medicines, to see if they will work before we all can use them, and that I work with people in hospitals to do such testing – this is part of my job as a statistician! I was alluding to my work with health professionals designing and conducting clinical trials, but with simpler terminology, so the children may connect and understand.

### Aside

Videos of experts from NASA, NSW Health, Radiation Oncology, Physical Education, Hunter Valley Coal Chain and other areas describing how and where statistics is used in practice, are available at [www.ssapostercomp.info/#videos](http://www.ssapostercomp.info/#videos). Each is about 3–5 minutes. They may be useful for students or those interested in a professional’s perspective of statistics in practice.

I described how my work as a statistician involves me working with other people to help them investigate, collect information, explore and come up with answers about things like whether a medicine will work (referring to clinical trials), or what are the best ways to teach (referring to the testing of pedagogical approaches), and even where we may expect to find various types of animals (ecology and statistics), or how to best use solar panels to drive an electric car (environmental sciences – identifying the best angle for efficiency, whilst accounting for resistance). The additional information in parentheses was not mentioned in the class, but is provided here to help recognise contexts.

### Aside

Some examples of articles exemplifying the afore-mentioned contexts are provided in the reference section at the end of the chapter. See Paterson et al. (2017), Howley et al. (2016), Reynolds et al. (2015), Howley and Roberts (2020).

### Key Message (Aspect Being Developed)

Describing a set of data is a fundamental skill required within statistics, data science and the big data world. On first view of a data set, it isn’t easy to describe patterns of distributions, and many jump to erroneous conclusions or descriptions based on initial observations.

So, on to the bag of leaves....

I held aloft my snap-lock bag (see Fig. 2.1) and asked how they would describe what I had in this bag. Unsurprisingly, given the size and number of them visible in the packet, the response was “leaves”. I could have teased out whether they are all the same size, colour or texture, but at this point, the key wasn’t to further describe what they saw, but rather to ultimately help them value the importance of *exploration*.



**Fig. 2.1** A snap-lock bag of leaves



**Fig. 2.2** The bag was turned around to reveal the contents

I then turned the bag around to reveal the contents as shown in Fig. 2.2. When asked can they describe the contents of the bag further, there were descriptions such as “pink and green leaves”. This seemed to be better describing the contents, and considering their view, it was understandable.

**Key Message (Aspect Being Developed)**

Exploration is the key. A scaffolded approach towards this, which maintains audience connection, and develops increasing levels of understanding and knowledge, is valuable. Enabling a soft realisation of the error of one’s ways can be helpful – rather than simply being told of facts.



**Fig. 2.3** The contents of the bag were emptied



**Fig. 2.4** The emptied contents were spread out

I then asked “Should we open the bag and explore?” This was met with great enthusiasm for doing so. I proceeded to empty the contents of the bag, and spread them out (see Figs. 2.3 and 2.4). I asked if it still seemed that there were just green and orange leaves, which the class confirmed.

I suggested we should keep exploring, and I further spread out the contents, as per Figs. 2.5 and 2.6. Aha! What have we here? Is it just a bag of leaves?

There are rocks and leaves!

#### **Key Message (Aspect Being Developed)**

Describing the whole – categorising and describing key characteristics are important.



**Fig. 2.5** The emptied contents were further spread out



**Fig. 2.6** The contents were spread further again

So now, I could begin to ask questions to identify underlying similarities and difference, and thus patterns or categories. I asked if the leaves seemed to all be the same type, teasing out responses of how they differed. Can we describe the bag better? I asked are there different sized leaves, are there different types of leaves, and rocks?

We discussed how we could sort them (categorise them) and as we did, I showed them a Pareto chart (see Fig. 2.7) – of course I didn't call it that. I informed the class how statisticians like to create pictures, or graphs, or charts, which explain things.

As I showed them the Pareto chart, I informed them that the graph tells us how many of each item (each leaf or rock, as per the pictures at the bottom) we have, and the taller the colour (bar) the more of that leaf or rock there is. I could then ask them which leaf they think there were most of, pointing to each one on the chart and they could then identify the one which had the tallest bar. I could ask them which one there was the least of in the bag.

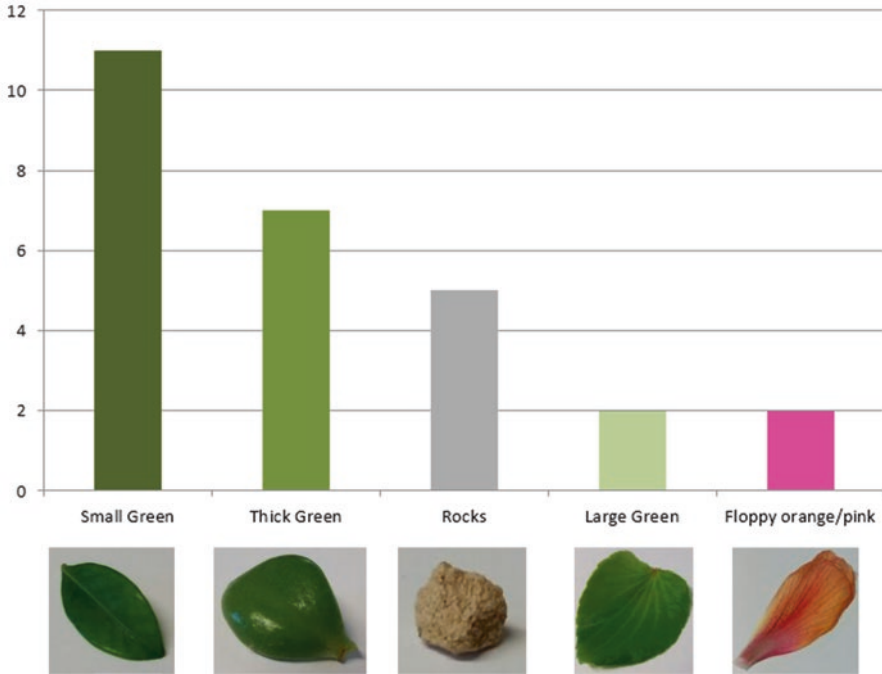


Fig. 2.7 Pareto chart

I then explained how this is a key part of my job as a statistician, namely, trying to best describe what is before me, through exploration, looking for groups of similar things (categorising), comparing groups and identifying patterns, and, importantly, being able to communicate key information simply.

There are many other aspects which could be considered. For example, how the categories vary or indeed how it's perhaps starting to give a picture of my garden, particularly the various types of plants and items in my garden. This of course raises another aspect that may be discussed, even from the beginning, of how I have taken some things from my garden (a sample) and for them to then try and develop a picture of what my garden may look like, or even how they may draw the garden. Doing so could really tease out ideas about what proportion of the garden may have certain coloured leaves or rocks, and so forth, based on how many were in the bag. This ties in nicely with inference – a key aspect of analysing data, and part of the Foundation Year Content within Australia's National Mathematics Curriculum.

To help consolidate the discussions, before I left I provided the class with a bag of uninflated balloons, informing the teachers and class that they may like to consider how to best describe and group (categorise) the many coloured balloons. I suggested that they may like to consider which one appears most, or least, frequently and see whether there were many or few colours, whether all balloons were the same, in shape, size, or texture. There are many ways to categorise the objects, and this sets the platform for introducing the many 'measures of interest' (random variables) we consider as we develop our understanding of statistics and exploring data.

The above activity supports the Australian National and State Curricula for Mathematics.

The Statistics and Probability strand within the Early Stage 1 (also referred to as Kindergarten – primarily encompassing those aged 5 to 6) and Stage 1 (Year 1 and 2 of school – primarily encompassing those aged 6 to 7) student cohorts in the Australian state of New South Wales, for example, has intended student outcomes which include the use of pictorial representations to support conclusions and the ability to represent and interpret data displays (Early Stage 1), and the ability to organise and display data (Stage 1). Within the Early Stage 1 cohort, students will organise and group objects into simple data displays, and compare the sizes of groups (NSW Mathematics Curriculum – Early Stage 1, <https://syllabus.nesa.nsw.edu.au/mathematics/mathematics-k10/content/634/>). Within the Stage 1 cohort, students will interpret information presented in data displays and describe data, communicating a least or most popular category, and other relative comparisons of categories identifying whether there are more or less of certain categories.

**Exercises** The following exercises are intended for a novice audience; new teachers endeavouring to develop statistical education plans and self-studying students.

1. Have you an example of where a first inspection may produce quite different decisions than those after unpacking or exploring the situation more deeply?
2. In the leaves activity described above, how else could the context be used? Inference (the use of a sample to infer about the wider population from which the sample was selected) is very important.
  - (a) Consider how the content of the bag may be used to conjure images of the garden from which it was obtained.
  - (b) Discuss whether the way the bag contents were chosen, and the number of such items, may influence how reliable our image may be, or how close it is to the truth the image may be.
  - (c) Consider having each child attempt to describe the garden based on the contents of the bag (or one that you create) – see if there is consistency or great variation in the descriptions.
  - (d) Question your audience to see what information may be required to help building a reliable picture of the garden from which the sample was obtained.
    - (i) See what such questions may inspire – perhaps indications of the types of plants, trees, leaves, terrain in the yard.
    - (ii) Perhaps have two bags, one selected from your garden and one selected from your neighbour's (or friend's) garden. Commence discussions that compare the contents of the bags from two yards – is your yard like mine? How do they differ? How big a bag do we need – how small is too small? How similar is similar enough? Use the separate descriptions of each bag to help compare and contrast.

## The X + STATS Initiative: What's Your X?

The *X + STATS* initiative is an example of a program designed to engage students of 8 to 18 years in statistical reasoning.

Big data, data science and statistics are omnipresent; however, introducing a field outside of an appealing context, or from a seemingly contrived context may decrease its appeal. The often isolated and seemingly disparately taught statistical tools and methods are rarely brought together in school, despite the value of such integration, both in helping students understand the material and in motivating them. Utilising and building upon the interests of the audience, enabling them to recognise the need and value of a new field or skill, and facilitating activities which enable them to feel the additional learning is within their grasp are valuable motivators of learning.

It has been said that the best way to teach the sciences inspirationally is to teach it the way it is practised. To this end, a national initiative known as the National Schools' Poster Competition ([www.ssapostercomp.info](http://www.ssapostercomp.info)) was created; it has been delivered annually in Australia since 2014. This *X + STATS* initiative is an example of a program designed to engage students of 8 to 18 years in statistical reasoning. It is fun, educational and engages students in project-based learning. It also dovetails biennially into an international initiative (the International Statistical Literacy Project, <http://iase-web.org/islp/>).

Check out the 1-minute overview video and 6-minute testimonials video at <https://ssapostercomp.info/#resources>.

The National Schools' Poster Competition (NSPC) develops students' critical thinking, investigation, collaboration, communication, reporting and creative skills, and their awareness of how statistics, quantitative and scientific skills can help us better understand the world, no matter the field of interest. It engages students of all disciplines and taps into their interests, demonstrating the interdisciplinary nature of investigations, and developing their data-oriented skills. It also maps to and addresses National and State Boards of Studies' curriculum outcomes including but not limited to 'Data Collection and Representation', 'Single Variable Data Analysis' and 'Bivariate Data Analysis'.

Piloted in the Hunter Region, Australia, in 2014 and expanding nationally in 2015, the project-based learning (PBL) activity involves teams of 2 to 5 school students creating an informative poster presentation (akin to a conference poster) addressing a practical question on an area of interest to the students, from *any field* (the *X* in '*X + STATS*'), and utilising the collection, presentation and interpretation of data. Significantly, the students lead the discussion, identifying topics of interest to them and potential investigations they may like to undertake. The 4-minute video 'How to Run the PBL Activity' available from [www.ssapostercomp.info](http://www.ssapostercomp.info) provides a quick yet detailed overview for conducting the activity.

Tapping into what students like, or are interested in (the *X*), and surreptitiously helping them to realise they not only need but can utilise statistical tools and methods, helps bring statistics to life and makes it accessible and appealing. Allowing



them to not only see the need for learning an area, but to enable them to value the area and recognise that additional learning as accessible are key motivators of learning.

Perhaps students are interested in music and also psychology; they may be interested for example in how music affects concentration. Perhaps the loudness of the music, or the type of music, may affect how well someone can concentrate on the road, much like mobile phone use when driving is illegal due to the effects on concentration. There are always apps that can be used to test concentration, whilst varying the type of background music.

Blind taste-testing of varying brands to test individual preferences, considering the deterioration of teeth in varying liquids, or testing whether there is an association between a person's gender or cultural background and a particular interest are but a few of the thousands of investigations undertaken and submitted to-date. The list of posters of past winners and honourable mentions for inspiration are available at [www.ssapostercomp.info/winners.shtml](http://www.ssapostercomp.info/winners.shtml). However, allowing the students to come up with the topics is important. Past topics have included investigations surrounding social media, wildlife, aerodynamics, cultural studies, sea life, transport, the environment, and many more. Teachers and students have commented on the value and interest in this project-based learning activity (the 6-minute testimonials video available at right of screen at [www.ssapostercomp.info](http://www.ssapostercomp.info)). Significantly, the activity helps consolidate the many often disparately taught statistical tools and methods for a practical purpose.

Winners may be forwarded to the biennial International Statistical Literacy Project (ISLP), with which it is aligned, albeit with many key distinctions. One of the unique aspects of the national competition is the strategy of building an Australia-wide network of mentors to attend schools to facilitate the activity's delivery and engage students with discussions of the practical importance of statistics. With an increased number of out-of-field teachers in secondary schools, and the ever-increasing demands upon teachers, providing mentors and support mechanisms contributes to the competition's success.

Mentors are trained and supported nationally, guided by the national coordinator and the resources he has produced surrounding how to engage teachers and students and facilitate the activity. Mentors include undergraduates with experience in statistics, pre-service mathematics teachers, current and retired industry practitioners and practicing professionals.

Teachers', students' and industry reflections have highlighted the twenty-first-century nature of the initiative – the very focus of what the general capabilities (GCs) of the state and national curriculum embody. “*General capabilities encompass the knowledge, skills, attitudes and behaviours to assist students to live and work successfully in the 21st century*”, NESAS. These reflections explicitly or implicitly covered GCs; the nature of the activity itself addressing key elements identified for each GC by NESAS. The flexibility of the activity, driven by students' interests and facilitated by teachers and mentors, allows for all of the GCs to be addressed:

- GC1 – Critical and creative thinking: “*combining the skills and knowledge they have acquired in their study of a number of different topics, within a new context*”, NESAS. Students devise the area of focus, research questions and hypotheses, and how to conduct and communicate the results of their investigations;

- GC2 – Ethical understanding: “*Examples include: collecting, displaying and interpreting data; ...exploring the importance of fair comparison and interrogating ... sources*”, NESA. Topics themselves may also surround ethical considerations.
- GC3 – Information and communication technology capability: “*Students can use their ICT capability to perform calculations; draw graphs; collect, manage, analyse and interpret data...*”, NESA. Posters may be developed electronically, incorporating graphs and data analyses, with submissions online.
- GC4 – Intercultural understanding: “*Students can apply mathematical thinking to identify and resolve issues related to living with diversity*”, NESA. The 2014 Stage 6 winner, who examined representations of ethnicities on television (<https://ssapostercomp.info/winners.shtml#tab2014>), was one of many considering cultural diversity.
- GC5 – Literacy: “*from graphs to complex data displays*”, NESA. This surrounds the effective and succinct communication of key information.
- GC6 – Numeracy: “*apply mathematical understanding and skills in context, ... and in real-world scenarios*”, NESA. The activity is introduced as a smaller-scale version of what happens in practice.
- GC7 – Personal and Social Capability “*opportunities for initiative-taking, decision-making, communicating their processes and findings, and working independently and collaboratively*”, NESA. The fundamental aspect of the national poster competition, presenting to a real audience, and working collaboratively in teams of 2 to 5 students.

The activity, as noted through teacher, student and industry reflections, enables teachers to also address the other areas NESA identifies as “important learning”, namely “Civics and citizenship”; “difference and diversity” and “work and enterprise”. A collaboration with Environmental Science and Sustainability colleagues and the Wollotuka Institute at The University of Newcastle, Australia (Howley and Roberts, 2020), extended the national poster competition to remote areas. This additionally enabled teachers to address two of NESA’s three *cross-curriculum priorities*: Aboriginal and Torres Strait Islander histories and cultures; and Sustainability.

This initiative enhances authentic student learning and engagement and addresses a lacuna in the education system surrounding the teaching of statistics: the failure to integrate statistics into curricula in a holistic and practical fashion. It enables teachers to address curriculum requirements succinctly and holistically. Importantly, the initiative not only assists those currently teaching, but additionally begins the early embedding of statistical thinking and concepts through authentic learning and engagement.

## iNZight

To assist with the above NSPC activity, or any data-based investigation, and the development of interest in and connection with the big data, data science and statistics fields more widely, we have developed a series of short (mostly 1-page, up to max 6-page) *iNZight* worksheets.

Key worksheet features:

- Consistent clear format (bullet-pointed)
  - Key *iNZight* skills addressed
  - Relevant *StatsTuneUp* videos to consider
  - Concise and specific stepwise instructions
  - Side notes – additional information in rectangles if required
  - Troubleshooting
- Visual support – screenshots, arrows

These worksheets cover how to install *iNZight*, import data, explore and visually present data in *iNZight*. They are available at <https://ssapostercomp.info/#additional>.

**Exercises** These exercises are intended for big data novices; this may be teachers wishing to engage their students in data exploration and visualisation, or students themselves who are self-learning.

The exercises utilise the National Health and Nutrition Examination Survey data (**nhanes 2009–2012 dataset**). The survey is a program of studies designed to assess the health and nutritional status of adults and children in the United States (<https://www.cdc.gov/nchs/nhanes/index.htm>). A modified version of the **nhanes 2009–2012 dataset** is available within *iNZight*. It is in a form that is suitable for educational purposes. For further details about this modified version, please refer to [https://www.stat.auckland.ac.nz/~wild/data/FutureLearn/Docs/1.6b%20NHANES%202009\\_2012.pdf](https://www.stat.auckland.ac.nz/~wild/data/FutureLearn/Docs/1.6b%20NHANES%202009_2012.pdf)

Solutions which include how the resulting output should look and be interpreted are provided after the exercises. It is recommended that the *iNZight* worksheets described above are attempted before testing yourself with this exercise.

1. Open **nhanes2009\_2012** in *iNZight*.
2. How many females were included in this dataset (hint: drag the Gender variable into Variable 1, then click on the Get Summary button)?
3. Consider the total annual gross household income (HHIncome). Is this an ordinal or a nominal variable? Why?
4. Graph HHIncome. Is the default ordering of the bars appropriate? Re-order if needed.
5. Compare income groups by gender. How would you describe the distribution of salaries for each of the males and female groups relative to one another? Are the proportions of males and females similar across the salary groupings?
6. Remove all variables from the drop-down list. Put Weight into Variable 1. Comment on the graph.
7. Remove Weight from Variable 1 and replace it with Height. Comment on the graph.
8. What explanation could you provide for the visual displays and associated descriptions arising from the output in Questions 6 and 7?

(Hint: could it be affected by another variable? Try keeping either Height or Weight in Variable 1, and add a variable that you think may affect them into Variable 2).

### Solutions to above Exercises

1. Open **nhanes2009\_2012** in *iNZight*.
2. How many females were included in this dataset (hint: drag the Gender variable into Variable 1, then click on the Get Summary button)?  
*A: 4574 females*
3. Consider the total annual gross household income (HHIncome), is this an ordinal or a nominal variable? Why?  
*A: HHIncome is an ordinal variable, because income groups has a natural logical order (low income groups through to high income groups).*
4. Graph HHIncome, is the default ordering of the bars appropriate? Re-order if needed.  
*A: you need to re-order income group 5000–9999 as demonstrated in the following figure (after re-ordering) (Fig. 2.8):*
5. Compare income groups by gender. How would you describe the distribution of salaries for each of the males and female groups relative to one another? Are the proportions of males and females similar across the salary groupings?

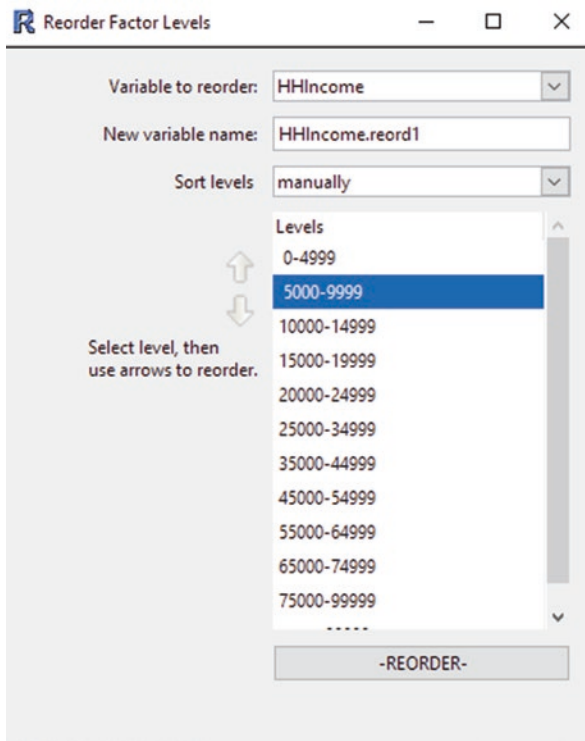


Fig. 2.8 Reordering factor levels

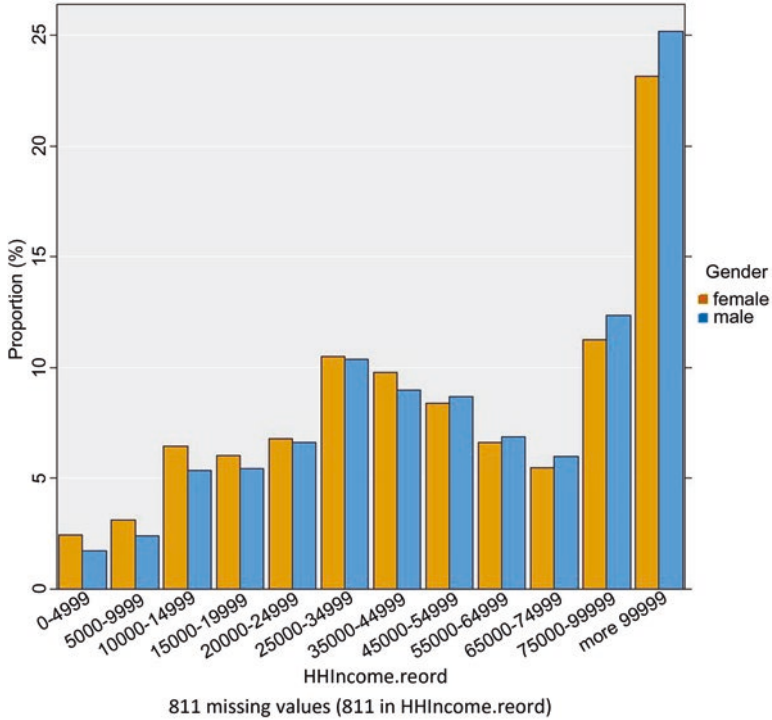


Fig. 2.9 Distribution of income by gender

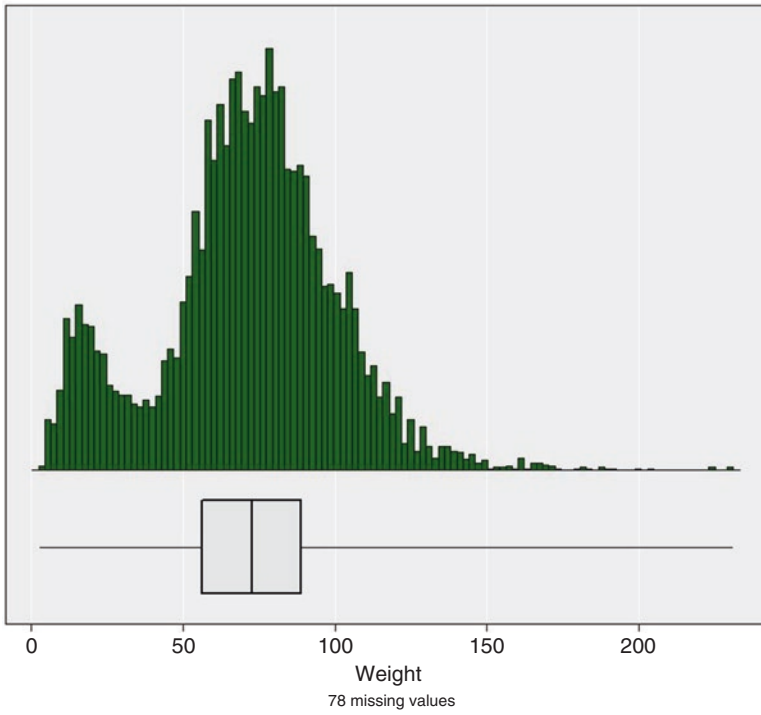
A: The distribution of females (orange bars) are slightly shifted towards the lower end of the scale compared with the distribution of males (blue bars). Notice how the proportions of females with less than \$45,000 annual income are higher than the proportions of the males, however we see greater proportions of males than females with incomes greater than \$45,000 (Fig. 2.9).

- Remove all variables from the drop-down list. Put Weight into Variable 1. Comment on the graph.

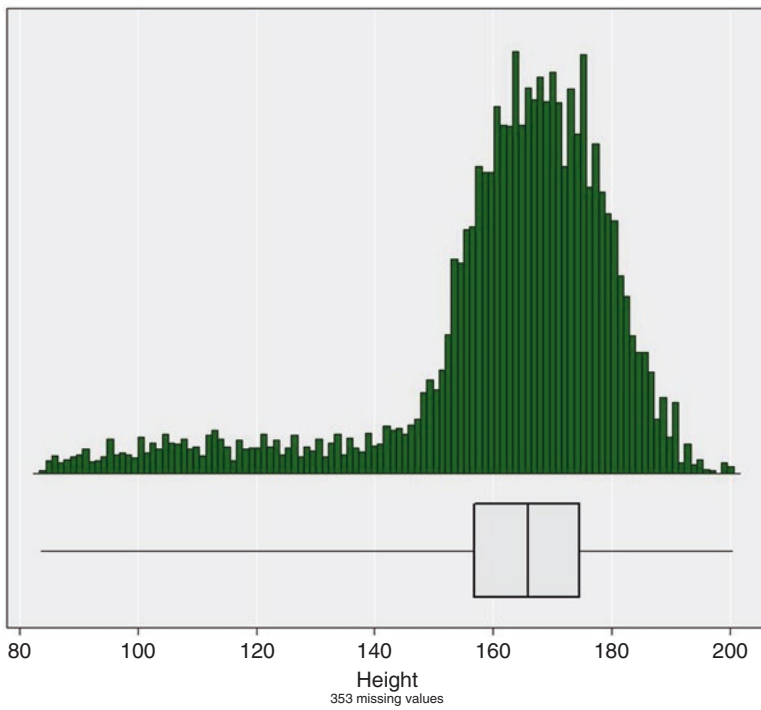
A: It appears there are two underlying distributions, one which falls between 0 to ~ 40 kg, and one from 40 kg onwards. The two distinct peaks (and associated decreasing frequencies of data either side of these peaks) indicate this (Fig. 2.10).

- Remove Weight from Variable 1 and replace it with Height. Comment on the graph.

A: The graph shows what is known as a skewed to the left distribution, a peak nearer the upper end of the scale and a long tail (of decreasing frequencies extending out to the left) at the lower end of the scale from about 140 cm. You may like to consider the StatsTuneUp videos to learn more about these visual displays (Fig. 2.11).



**Fig. 2.10** Distribution of weight



**Fig. 2.11** Distribution of height

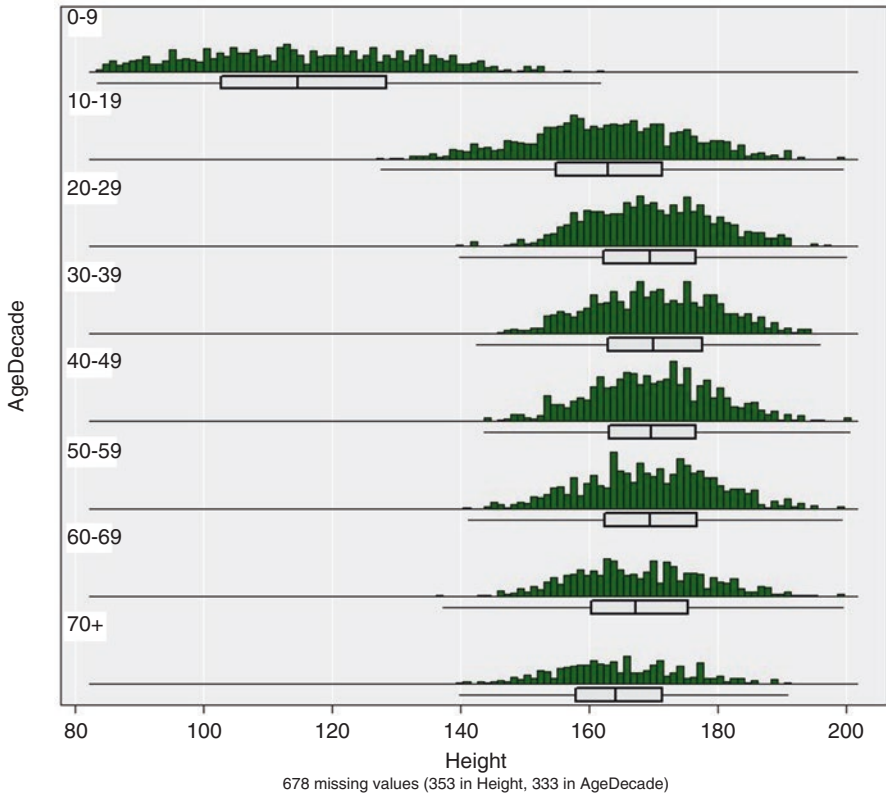


Fig. 2.12 Distribution of height by age decade

8. What explanation -- visual displays and associated descriptions arising from the output in Questions 6 and 7?

(Hint: could it be affected by another variable? Try keeping either Height or Weight in Variable 1, and add a variable that you think may affect them into Variable 2).

A: *Weight and height are highly dependent on age. Before adulthood, a person is also growing but once a person reaches adulthood, he/she will stop growing (at least in height!) (Fig. 2.12).*

### Concluding Remarks

Breaking down barriers to learning is a critical step required for many students. These barriers may arise from many reasons, for example through a lack of prior knowledge, anxieties surrounding the very nature of the field, ignorance of the field’s relevance, or apathy. Enthusing individuals, utilising aspects within their

grasp to enable them to achieve a higher level of engagement and learning is an invaluable approach for ultimately inspiring a sense of accessibility and appeal that overcomes such barriers. Significantly, the wide application of statistics and big data, and a holistic integration of the often disparately taught aspects, must be leveraged to help create enthusiasm and improved learning. The success and nature of the initiatives described in this chapter, will hopefully inspire both researchers to pursue this area, and for teachers to adopt and adapt them for their own learning and teaching requirements.

The adage ‘it takes a village to raise a child’ is apt for education. A key desired outcome of student education is the development of an ultimately productive contributor or member of the workplace. Accordingly, industry, business and government, along with higher education personnel must proactively support early childhood, primary and secondary educators. An *arm’s length* approach of expecting primary and secondary teachers to upskill alone to the point of piquing students’ interests in big data and statistics and driving students to further studies in the discipline, is fraught. Activities such as the National Schools Poster Competition and the CSIRO’s STEM Professionals-in-Schools program with which it collaborates ([www.csiro.au/en/Education/Programs/STEM-Professionals-in-Schools](http://www.csiro.au/en/Education/Programs/STEM-Professionals-in-Schools)) are designed to contribute to the required support mechanisms. The challenge is to ensure that it is not left to the few to continue this charge, rather that all recognise the supply chain is only as strong as its weakest link, and thus similarly unite in this charge towards addressing the big data age.

## References

- Berk, L.E. & Winsler, A. (1995). *Scaffolding children’s learning: Vygotsky and early childhood education*. National Association for the Education of Young Children.
- Bryce, G. R. (2005). Developing Tomorrow’s statisticians. *Journal of Statistics Education [Online]*, 13(1) <http://www.amstat.org/publications/jse/v13n1/bryce.html>
- Gore, J., Holmes, K., Smith, M., et al. (2015). *Australian Educational Researcher*, 42, 155. <https://doi.org/10.1007/s13384-015-0172-5>
- Howley, P., Reynolds, R., & Southgate, E. (2016). The teaching discipline doesn’t matter? An assessment of preservice teachers’ perception of the value of professional experience in attaining teacher competencies. *Australian Journal of Teacher Education*, 41(10). <https://doi.org/10.14221/ajte.2016v41n10.3>
- Howley, P. & Roberts, T. (2020). Engaging school students and educators with the practice of statistics. *Statistics Education Research Journal*.
- Howley, P. P. (2008). Keeping it real, keeping them interested and keeping it in their minds. *Journal of Statistics Education*, 16(1), 1–16. Online at [www.amstat.org/publications/jse/v16n1/howley.html](http://www.amstat.org/publications/jse/v16n1/howley.html)
- Howley, P. P. (2009). Winning them over. In David Allingham (Ed.), *ASEARC: Proceedings of the Third Annual ASEARC Research Conference*, (pp. 1–4). : Applied Statistics Education and Research Collaboration (ASEARC). Online at <http://www.uow.edu.au/informatics/math/research/groups/asearc/2009rescon/index.html>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global



- Institute. Online at <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>
- Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments – a comprehensive review of the literature. *Teaching in Higher Education*, 8(2), 195–209.
- Paterson, M. A., Smart, C. E. M., Lopez, P. E., Howley, P., McElduff, P., Attia, J., Morbey, C., & King, B. R. (2017). Dose dependent effect of dietary protein in type 1 diabetes mellitus. *Diabetic Medicine*. <https://doi.org/10.1111/dme.13347>
- Puang-Ngern, B., Bilgin, A., & Kyng, T. J. (2017). Comparison of graduates' and academics' perceptions of the skills required for big data analysis. In T. Prodrômou (Ed.), *Visualization and statistical literacy for open and big data* (pp. 126–152). IGI Global. <https://doi.org/10.4018/978-1-5225-2512-7.ch006>
- Reynolds, R., Howley, P., Southgate, E., & Brown, J. (2015). Just add hours? An assessment of pre-service teachers' perception of the value of professional experience in attaining teacher competencies. *Asia-Pacific Journal of Teacher Education*, 44(5). <https://doi.org/10.1080/01359866X.2015.1086971>
- Sowey, E. R. (1995). Teaching statistics: Making it memorable. *Journal of Statistics Education [Online]*, 3(2) <http://www.amstat.org/publications/jse/v3n2/sowey.html>
- Sowey, E. R. (2001). Striking demonstrations in teaching statistics. *Journal of Statistics Education [Online]*, 9(1), 2001. <http://www.amstat.org/publications/jse/v9n1/sowey.html>
- Williams, A. S. (2010). Statistics anxiety and instructor immediacy. *Journal of Statistics Education*, 18(2), 1–18.

**Professor Peter Howley** [www.newcastle.edu.au/profile/peter-howley](http://www.newcastle.edu.au/profile/peter-howley)

PhD(Statistics), BMath(Hons), AStat(Statistical Society of Australia)

Professor of Statistics and Assistant Dean (Outreach & Engagement) for The University of Newcastle's Faculty of Science, Peter is nationally and internationally recognised as a leader and award winning academic and practitioner in Statistics and STEM education and initiatives.

Peter is the National Chair of Statistical Education (Statistical Society of Australia (SSA)) and an elected member of the International Statistical Institute (ISI).

He sits on National and International Advisory Boards for STEM, Statistics and Data Science initiatives, is an expert advisor to State and National Educational Boards, National Science and Engineering Challenge Director, inaugural 2018 Science and Technology Australia STEM Ambassador and an International Editorial Boards member (and Guest Editor, *Statistical Education Research Journal Special Edition* 2020).

Peter leads cross-disciplinary school and community engagement initiatives including the National Schools Poster Competition (engaging over 1000 students annually in teams who: undertake a small-scale industry-style investigation on any topic of their choosing; participate in the collection and analysis of data; and report upon their project in poster format, which they submit online). He has established associated national and local support networks and collaborations uniting schools, academics, tertiary, government, industry and national organisations for which he won the ISI Best Cooperative Project Award (2017) and SSA's National Services Award (2018) for leadership in statistical education.

He has received over two million dollars in grants surrounding education, health, school and community engagement, and interdisciplinary collaborations in STEM, Environmental Sciences, Health and Education and in 2015 was awarded an Australian Office for Learning and Teaching Citation for 'Outstanding Contributions to Student Learning'.

**Kevin Wang** MSc(Hon), AStat

Kevin Wang is an Accredited Statistician (Statistical Society of Australia) and a biostatistician with over 10 years of experience in the health sector. He is currently working at the Centre for Epidemiology and Evidence, NSW Ministry of Health. He works with many “big” administrative datasets, such as admitted patient, emergency department and perinatal data. He has a particular interest in the epidemiology of infectious diseases and is currently completing his Master of Public Health at the University of New South Wales. Prior to becoming a public health professional, he worked in health management consulting (KPMG) and the pharmaceutical industry (Eli Lilly).

Kevin is also passionate about statistical education for early learners. He has been a mentor in CSIRO’s STEM Professionals in Schools and National Schools Poster Competition programmes, where he has been mentoring school students (ranging from Years 5 to 10) through activity based learning, since 2014. He enjoys working in partnerships with the teachers and students, and watching the students getting inspired by mathematics and statistics.

**A/Prof Ayse Aysin Bilgin** BEng., MBA, MMaths, PhD, PostGradDip HE (L&T), MHE (L&M)

Ayse is an Associate Professor in the Department of Mathematics and Statistics at Macquarie University. She is the President-Elect of the International Association for Statistical Education (IASE) and an elected Member of the International Statistical Institute (ISI). Ayse’s research explores applications of statistics in health sciences and learning and teaching in statistics.

Ayse was a recipient of several learning and teaching awards for her outstanding contributions to student learning such as an Australian Learning and Teaching Council Citation for ‘Outstanding Contributions to Student Learning’, a Macquarie University VC Citation, ATEM Award Community Engagement (Highly Commended); Higher Education Award for Employability; joint-recipient of Australian Awards for University Teaching (AAUT). Her pedagogical research interests are focused on statistics education, such as learning approaches in statistics, work integrated learning in statistics, impact of learning spaces to students’ learning. Recently she was awarded “Excellence in research: Five Future-shaping Priorities (Healthy People)” by Macquarie University and she was a member of research team which were the finalist of the Eureka Prize.

# Chapter 3

## Using Big Data in a Master of Applied Statistics Unit



Ayse Aysin Bilgin and Peter Howley

**Abstract** There have been various discussions in the academic community, especially by Statistics Educators about Big Data in various fora. These include online discussion forums, round table discussions (such as World Statistics Congress), and during conference presentations (such as IASE satellite(s), IASE round table(s) and OZCOTS). Is it statistics or computing or both? However one defines it, Big Data is valued by industry, governments, and NGOs alike all around the world. Recent job advertisements are increasingly desiring skills in Big Data analysis, instead of statistical skills. Therefore, it is important to prepare graduates of our courses for their future work by providing opportunities to learn how to deal with big, complicated and complex data sets. This chapter provides a case study on the inclusion of Big Data into a Master of Applied Statistics unit, called data mining. The short history of the unit, topics of study, learning outcomes, assessment tasks, and how students were included in decision making for their projects is described before presenting two case studies from students' projects.

**Keywords** Statistics education research · Big data · Authentic assessments

### Introduction

With the explosion of data collections and data storage in the recent years, it became clear that new methodologies and possibly skill sets are required for analysis of such large and complex data sets, usually called “Big Data”. Although, there is increasing focus on computer scientists and computing methodologies to be used

---

A. A. Bilgin (✉)

Department of Statistics, Macquarie University, Macquarie Park, NSW, Australia

e-mail: [ayse.bilgin@mq.edu.au](mailto:ayse.bilgin@mq.edu.au)

P. Howley

School of Mathematical and Physical Sciences/Statistics, The University of Newcastle, Callaghan, NSW, Australia

e-mail: [peter.howley@newcastle.edu.au](mailto:peter.howley@newcastle.edu.au)

© Springer Nature Switzerland AG 2021

T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy

Implications of Research in Education 13,

[https://doi.org/10.1007/978-3-030-76841-6\\_3](https://doi.org/10.1007/978-3-030-76841-6_3)

for data mining, dredging, or assessing correlations within data, existing and modified statistical methods continue to be invaluable within the big data space. It is reported by many people that there is a shortage of statisticians/data scientists/Big Data analysts in many countries. It is claimed that “statistical analysis and statistical software skills” were the most important skills required for Big Data analysis (Puang-Ngern et al., 2017) which is also supported by Lockhart (2018), “... these classical statistical ideas continue to have a crucial role to play in keeping data analysis honest, efficient, and effective” (p. 4) and “that statistical ideas remain relevant in the face of massive, complex, dynamic data.” (p. 6).

In a 2017 blog, a Toronto-based statistical consultant, Macnaughton (2017) asked “Do we and our methods belong in our own small tent at the fair? Or are we a key part – sometimes the center ring – of the big tent data-science culture? What is the best name for our beautiful field?” The debate of whether statistics should change its name to reflect what a statistician does, flares up time to time between statisticians and statistics educators, and Macnaughton’s essay certainly created more debate (such as in the Australia and New Zealand statistics mailing list discussions in 2017). Some of the suggested names fit well with the current climate of workforce, such as “data scientist” however some others may be less palatable to many, such as “datatics” similar to “dietetics” or “dataticist” similar to “economist”.

In October 2017, the then-president of the International Statistical Institute (ISI), Professor Helen McGillivray (2017) wrote that while she was reading an article on an airplane magazine related to ‘big data’ and ‘big data analytics’ job opportunities, she noticed the word “statistics” was not used even once, despite the person quoted in the article being a statistician. In the same message, she also wrote that “I also witnessed beliefs of some at the first UN World Data Forum that statistical literacy and data literacy are different, and that the former is ‘smaller’ than the latter.” It is clear, as a discipline, statistics has a public relations problem. The general public, even the well-educated members aren’t really sure what a statistician does. Would it be possible to better inform the population at large about what a statistician does, so that we do not need a name change? It’s unlikely statistics will change its name; however, educating people about the profession is crucial if we are going to thrive as a discipline in the twenty-first century. We also must educate our students to be aware of the place of their statistical skills in the “big data” era as future statisticians, so that they are able to secure employment after graduation.

An artist paints a picture by using his/her palette of colours and various shapes and sizes of brushes, transferring their mind’s eye into a creation for others to behold. Similarly, a statistician solves problems by using multiple techniques (similar to the palette of colours used by an artist) and different kinds of software (similar to brushes used by an artist) to solve problem(s). However, it is not sufficient to learn statistical techniques and software applications to be a statistician. It is important to be curious, investigative, creative, and a problem solver (Allen, 2018). Therefore as educators of statistics, we need to awaken the curiosity and creativity of the students as well as providing them opportunities to improve their

critical-thinking and problem-solving skills. One way of improving critical-thinking and problem-solving skills with increased student engagement is authentic assessments, which also provide the best feedback for the teachers/educators about their students' understanding or misunderstandings of the unit content (Bidgood, 2010; Chance, 1997; MacGillivray, 2016; Smucker & Bailer, 2015). The statistics education literature provides many examples from undergraduate statistics courses, however not many for postgraduate (master's) courses, nor about the place of "Big Data" in statistics education. Therefore, the aim of this chapter is to show how "Big Data" was incorporated into a Master of Applied Statistics unit and how it was possible to do that within an authentic assessment framework.

## Background

The Master of Applied Statistics (MAppStat) degree has been offered at Macquarie University in Sydney, Australia since 1983. Until 2015, the length of the program was 1 year if the applicant held a Bachelor degree with a Major in Statistics, slightly longer for an applicant from other disciplines, since they had to complete a qualifying program, such as a Graduate Certificate in Applied Statistics before admission to the MAppStat. The original program was designed to educate graduates so that they can find jobs as statisticians in government departments, such as the Australian Bureau of Statistics (ABS); research organisations, such as Commonwealth Scientific and Industrial Research Organisation (CSIRO); medical institutions, such as hospitals; and area health services and industry. Development of computing skills was an integral part of the program from the beginning in 1983. More or less the aims of the program and the units contributing to the program remained the same; however, the contents and software tools within the units have changed over time. The number of core units in the program also changed and the program became more flexible in their delivery mode. Currently all of the units are available externally to suit students studying part-time while working full-time. Internal units are mostly offered in late afternoons with a starting time of 6 pm and ending time of 10 pm.

Recently, the Australian Qualification Framework (AQF) identified requirements for postgraduate degrees at level 9 to remain accredited 'level 9' (Masters Level) programs in Australian higher education sector. AQF states "the purpose of the Master's Degree (Coursework) is to qualify individuals who apply an advanced body of knowledge in a range of contexts for professional practice or scholarship and as a pathway for further learning" (AQF, 2018). To comply, in 2015 the MAppStat program, which is coursework Master's degree, became a 2-year degree, since then there have been no changes. The earlier qualifying program is integrated into the MAppStat, which needs to be completed during the first semester (Macquarie University Handbook, 2018). The MAppStat program can be completed within a year and a half (full-time study) for applicants who hold a Bachelor degree with a Major in Statistics or similar disciplines such as Actuarial Studies.

Since 2015, the structure of the program includes four core units, which are equivalent to qualifying program type units (Macquarie University Handbook, 2018). These units are Mathematical Modelling, Introductory Statistics, Applied Statistics, and Introduction to Probability. If an applicant already has a Major in Statistics, s/he does not need to study these units and is given Recognition of Prior Learning (RPL) credits for these units. In addition to qualifying type of units, there are five required (core) units. These units are Statistical Theory, Generalized Linear Models, Multivariate Analysis, Modern Computational Statistical Methods, and Statistics Project. The remaining 7 units are electives that need to be selected from a list of statistics, actuarial studies, or biostatistics units. The Data Mining unit, which incorporates Big Data into its content, is one of these elective units.

The data mining unit was first offered in 2007 as an elective unit to MAppStat students. As early as 2008, when it was noticed by students of other programs, we had students from Actuarial Studies, Commerce, Information Technology, non-award students (who choose to study a unit instead of a degree at the university), and exchange students from other countries enrolling into the units. Their enrollment depended on their previous studies and experience. They needed approval to be able to enroll into data mining unit to ensure that they were not academically at risk of failure. In 2013, a new degree program, the Master of Research, was created at Macquarie University to replace existing Honours programs so that graduates would have a pathway to a PhD, similar to European students. With the creation of this new degree, students are required to complete 3 years undergraduate and 2 years postgraduate, where 1 year is coursework, and the second year is a research thesis before they are admitted to any PhD program. The data mining unit was also included into the Master of Research degree program as an elective unit.

In 2017, the MAppStat degree became the first postgraduate degree in the Faculty of Science and Engineering at Macquarie University with all of its units available externally. That was the first year that the data mining unit was offered externally, as well as remaining available internally. The difference between internal and external offerings of the unit is that students who enroll into external unit do not need to attend classes, however they still need to attend an invigilated exam at the end of the semester.

Data mining requires a combination of disciplines to contribute to its development. The main contributing disciplines are statistics, computing, machine learning, and artificial intelligence. In many universities, such a unit is taught very theoretically in mathematics or computing departments. At Macquarie University, the unit was developed and delivered by academics within the Department of Statistics. Therefore the statistical principles and statistical thinking played and still play an important role on how the unit is delivered and assessed.

From the beginning (in 2007), the intention was to make the unit externally available. Therefore, as well as using a commercial software package within the unit, a free software, R (R Core Team, 2017), was also introduced. Later, RStudio (RStudio Team, 2016) was also included in the unit. The commercial software changed over

time. During the most recent delivery of the unit, IBM SPSS Modeler 18.0 (IBM, 2016) was used. This software was made available to students within the university computer laboratories and through a virtual lab, called iLab (Macquarie University Wiki, 2018). Through iLab students were able to access to IBM SPSS Modeler through the internet at no cost. However, sometimes due to (bad) internet connection, this software can have efficiency problems. This might be why most of the students choose either coming to campus to use the IBM SPSS Modeler or they only use R and RStudio for all assessment tasks.

During the earlier years, almost none of the students knew R or similar command-line statistical software languages. Because of the integration of R into some of our undergraduate units, the number of students requiring extra help to learn R declined. However, since some students come from other disciplines, there is still a need to teach students the intricacies of how to write R scripts, how to find useful packages for their data analysis.

## Student Demographics

The number of students choosing the data mining elective unit was comparable to other elective units in the MAppStat program and increased over time except in 2015 and 2016 (Fig. 3.1). Excluding 2015 and 2016, the data mining unit was annually the most popular elective unit having the highest numbers of students. At the

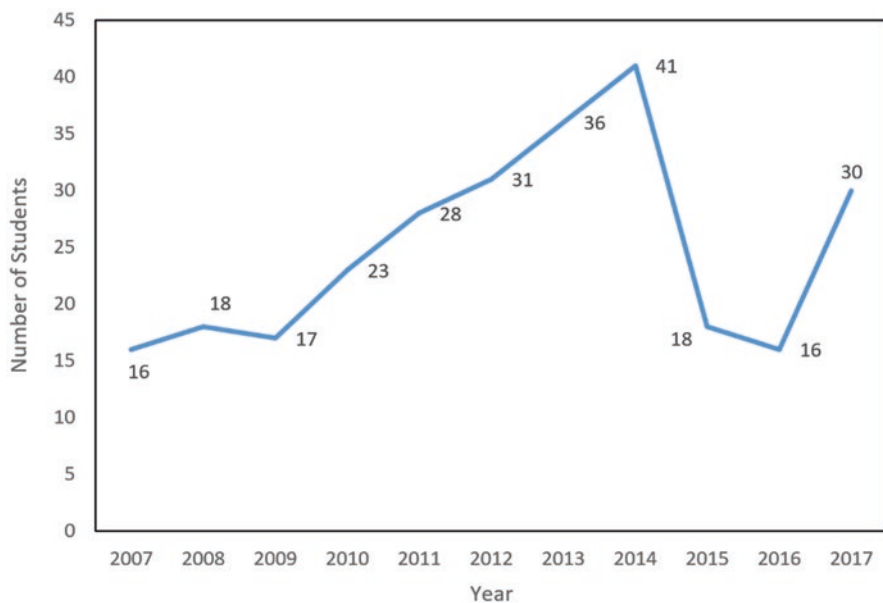


Fig. 3.1 Number of students in data mining unit

beginning of 2018, the number of students (37 internal, 17 external) in this unit was the highest of all master level statistics units, except for the compulsory (core) Statistical Theory unit.

In 2017, there were 30 students enrolled in data mining unit. More than two-thirds of students were MAppStat students (70%). The other students were studying towards Master of Research (10%), Master of Information Technology (10%) or other degree programs (10%) such as Master of Commerce and Master of Banking and Finance. The majority of the students were male (70%) and local (73%), i.e., domestic postgraduate fee paying. The gender distributions for local and international students were similar: for each female student in the class, there were two male students.

The data mining unit was first available externally in 2017. In this initial external offering of the unit, one-third of the students enrolled as external students, the remaining were internal students. Most of the external students were working full-time and studying part-time. Despite some of them living locally to the University (living in Sydney) and classes being offered after working hours, they chose to study externally. They would only come on campus for the final exam.

## Graduate Capabilities and Learning Outcomes

Macquarie University identified the following graduate capabilities for the post-graduate students which need to be covered in a program such as MAppStat but not necessarily in each unit. The graduate capabilities are (1) *critical, analytical, and integrative thinking*, (2) *effective communication*, (3) *discipline knowledge and skills*, (4) *research and problem solving capabilities*, (5) *engaged and responsible, active and ethical citizens*, and (6) *capable of professional and personal judgement and initiative*. These graduate capabilities were considered before writing out the learning outcomes for the unit.

The overarching learning outcome of the unit is to provide students an opportunity to gain an extensive understanding of the principles and concepts of data mining methods and their applications (Data Mining Unit Guide, 2018). Big Data is brought into the unit through the assessment tasks where students are using the data mining techniques to analyse big data sets. The second important learning outcome of the unit is demonstrating level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data (data preprocessing) and presentation of results of data mining and modelling to possible users. There are also some topic-specific learning outcomes such as examining and comparing the differences between different decision trees and interpreting sophisticated decision tree models for decision makers by writing a professional data mining report.

Each graduate capability was addressed by multiple, if not all, learning outcomes. Each assessment, including formative assessments, is designed to include skills related to each graduate capability. All learning outcomes were assessed at least once by a formative assessment task followed by a summative assessment task



to be able to provide opportunities for students to benefit from the feedback provided to their formative assessments. The full list of learning outcomes, mapping of learning outcomes to graduate capabilities, and assessment tasks can be found in the Data Mining Unit Guide (2018).

## Topics

The topics covered in the unit changed over time but the two-part approach to topics remained the same. The first part of the unit concentrates on descriptive data mining similar to Tukey's (1977) *Exploratory Data Analysis* (EDA). The descriptive data mining takes a few weeks which also includes introducing students to the software packages; students are required to complete tutorial exercises each week using either R and RStudio or IBM SPSS Modeler. The topics in this part of the unit include simple five-number summaries, graphical displays, crosstabulations, summary tables, creation of new variables such as normalising the variables and creating categorical variables based on continuous variables, detecting outliers and missing data and dealing with them, and visualisation. The emphasis of these exercises is on process, in particular thinking like a detective searching for clues to solve a crime, or a statistician (data detective) understanding the variables to be able to answer research questions. In many data mining projects, the research questions are unknown. An analyst is given a data set to analyse without any questions or vaguely posed questions. Therefore, the data preprocessing step helps students to formulate data mining questions based on their understanding of the data set. What students learn in this part is also tied to the CRoss Industry Standard Process for Data Mining (CRISP-DM; Chapman et al., 2000).

The CRISP-DM process model includes six phases that resembles Deming's quality improvement Plan, Do, Check and Act (PDCA) cycle (Deming, 1986, p. 88). The six phases of CRISP-DM fit together in a cyclical process to address the main steps in any data mining project (Fig. 3.2). The phases of CRISP-DM are translated into statistical language by the lecturer to emphasise the importance of statistical thinking for any data mining project, especially for the students coming from other disciplines.

The first step, *business understanding* is similar to *formulating a research question* in statistics. Without a question, we may not realise we have the answers. The second step, *data understanding* can be considered as *recognising the required data and/or variables*, which may come from the same or different sources. The importance of having whatever can be gathered when the project started instead of trying to add variables or data later in any project is emphasised, so that when students start working on their projects they are not disadvantaged by not including required variables for their analysis. At this stage, the research question(s) can be modified based on the available data. The third step, *data preparation* is about *getting the data ready for data mining*, in other words, *data preprocessing* which includes dealing with missing data, outliers, recoding variables, categorising continuous



**Fig. 3.2** The six phases of CRISP-DM (Chapman et al., 2000, p. 10)

variables, and re-categorising categorical variables. The fourth step is *modeling*. Students are encouraged to consider as many models as possible for their data set, especially when they are using classification techniques for their second project. This way, they are able to identify the best model from the ones that they have applied to their data set to answer their research question, which is the fifth step of the CRISP-DM cycle, *evaluation* or in statistical terms, deciding which *model is the best for the data and the question at hand*. The final step of CRISP-DM, *deployment* or *using the model for day-to-day business activities*, cannot be addressed in the unit, however students are encouraged to consider this aspect when they are writing their project reports so that they can convince the management to act on their conclusions. They do that by including at least a paragraph into their conclusion to provide examples of how their results can be used in day-to-day activities.

During the first couple weeks, the need for data mining and the similarities and differences between data mining and statistics is discussed based on the literature (Hand, 1999; Sharp Sight Labs, 2016).

The topics covered under the undirected knowledge discovery (UDKD) – unsupervised learning (Han & Kamber, 2006; Hastie et al., 2009) – are market basket

analysis (Berry & Linoff, 2004, pp. 287–320; Hahsler et al., 2018), K-means and hierarchical cluster analysis (Maechler et al., 2017; R Core Team, 2017; Suzuki & Shimodaira, 2015), and Kohonen networks (self-organizing maps, SOM; Wehrens & Buydens, 2007). Since the majority of the students come with a major in statistics in their undergraduate studies, they would have studied K-means and hierarchical clustering, however they would not have learnt Kohonen networks and market basket analysis. At least 3 weeks of the semester is dedicated to UDKD techniques, their applications, and related software tools.

The directed knowledge discovery (DKD) – supervised learning – topics are wider than the topics for UDKD (Han & Kamber, 2006; Hastie et al., 2009). However, since some of the techniques are covered in other MAppStat units such as logistic regression and generalized linear models, a subset of modeling is covered in this unit. Also, student input into what to cover is sought during the first 2 weeks of the semester to adopt the unit topics to students' expectations. During the last offering of the unit, the covered topics for DKD included various decision trees such as C 5.0 decision tree, Classification and Regression Tree (C&RT or CART; Ripley, 2016; Therneau et al., 2017), Neural Networks (NN; Fritsch & Guenther, 2016; Venables & Ripley, 2002), Support Vector Machines (SVM; Meyer et al., 2017), Naïve Bayes Classifier (Meyer et al., 2017), and Random Forests (RF). As well as specifics of the classification techniques, various impurity measures for decision trees are studied so that students' understandings of the decision trees are deeper. Finally, various model-evaluation metrics are explored and their weaknesses and strengths are discussed.

Although the underlying algorithms are explored, the topics in this unit are covered with a focus on the applications of the methodologies and the interpretations of the results of the modeling since the unit is designed for a master's program on applied statistics.

A guest lecturer from industry or government is invited to give a talk to students in week 13 before the revision takes place. The guest lecturers come from industries such as banking, real estate, and education. They present how they are using various data mining techniques in their workplace. Students see these presentations as an encouragement to prepare for the final exam and future. After all, they are told that what they learnt in this unit has an application in the workplace; it is not something they can forget about after they sit the final exam.

The revision starts with a seemingly unrelated figure (Fig. 3.3). This is deliberate. The figure is given because there can be more than one explanation for any figure and analysis depending on which angle you look at them. Therefore students are encouraged to explain what they think along with what their answer is to a specific question in the final exam. In other words, they are told, there might not be one right answer for each question in the final exam. Students are encouraged to go through their own assessment submissions, both formative and summative and consider the feedback provided to them while preparing for the final exam.

**Fig. 3.3** What do you see?  
A young girl or an old  
woman?



### Assessments in the Data Mining Unit

The assessments in the unit are designed to provide ongoing feedback to students through formative assessment tasks (weekly tutorial exercise submissions and participation) throughout the semester. These assessments feed into summative assessment tasks. The summative assessment tasks are two projects, their related assessments, and a final exam. Both of the projects can be completed using big data sets if the students choose to do so.

Assessment of student learning and learning outcomes they achieved is a crucial part when authentic, real world problems are at the core of assessment tasks. However, at first it might be hard to set authentic assessment tasks, because it is difficult to identify data sources or potential partners and potential projects that are suitable for a unit of study. In addition, creating rubrics for marking and grading for such assessments are harder. But when a database of data sources and partners are identified and rubrics are set, setting up authentic assessment tasks becomes easier and does not require a lot of time. However, it is documented that the marking and grading such tasks take much longer time than standard assessments where the lecturer poses questions based on a specific data set with expected answers (Bilgin et al., 2017).

As well as learning the analysis and software applications, students are required to become fluent explaining the results of their analysis to possible audience who might not have any knowledge of statistics, in other words, students are required to write their results in everyday language instead of statistical jargon. Students usually identify this as the hardest part of the unit.

## ***Formative Assessments***

Weekly tutorial exercises are designed as formative assessment tasks. They go hand-in-hand with weekly topics covered in the lectures. They are presented to students usually with the following headings: (a) *the required resources for the tutorial exercise*, such as downloading an R package, a reading or a dataset; (b) *what to do*, specifies the expectations such as what kind of problems needs to be solved; (c) *what to read*; and (d) *what to write* for the tutorial exercise submission.

Each week students are given one or two readings to expand their understanding of what was covered during the lectures and tutorials. To enable students to better concentrate on their readings, they are given short questions to answer based on their reading(s) as part of the weekly tutorial exercises. For example, in the second week, students are recommended to read the first chapters of *Statistical Analysis with Missing Data* by Little and Rubin (2014) and *Outliers in Statistical Data* by Barnett and Lewis (1994). They are asked to identify the two most important things from Little and Rubin (2014) and the two most useful things from Barnett and Lewis (1994) and then write no more than 100 words for each reading.

The feedback to these exercises are given as a class feedback highlighting the possible mistakes students made, and sharing the best student(s) solutions and/or a model solution in a written format. Students are encouraged to talk about their previous submission(s) in the light of the class feedback they have received with their peers (only possible for internal students) and/or reflect on their learning that week by using an online discussion forum (both internal and external students).

## ***Summative Assessments***

Two projects and their related assessments form the in-semester summative assessment tasks, along with participation. The invigilated final exam at the end of the semester contributes 50% towards the final mark.

The first project is the UDKD project with 15% weight towards the final mark. It is more structured with a simulated data set, where students are asked to analyse using cluster analysis and market basket analysis. It is an individual project, and, due to the size of the data set, both number of observations and the number variables, as well as the way students preprocess the data (i.e., deal with missing values, outliers, inconsistent values, and creation of new variables) and identify research questions, each student report could be different within a given semester. Students are given a document to describe the project, expected sections of the project report, how CRISP-DM steps are to be presented within their project report, and a marking

rubric. In addition, the three best reports from each of the previous 3 years, a total of nine projects, are shared with the students within the learning management system (LMS). If students decide to use another data set, they can do so, after discussion with the lecturer so that she can assess the suitability of the data set for the project.

Students are required to submit their report, the R scripts, IBM SPSS Modeler streams and any other software log files. In addition to these, a self-assessment is submitted by each student by answering a few questions like “*Which part of this project was most difficult for you?*”, “*What would have been helpful to overcome your difficulties?*” and “*Reflect on one of the employability skills you gained by completing this project (no more than 250 words)*”.

Individual feedback to this project is given by using the marking rubric (for internal and external students) as well as face-to-face during a tutorial (internal students only). The top three students are asked permission to share their projects anonymously. If they decline, the invitation is extended to next in line. So far, students have been happy to share their projects. Students are encouraged to learn from their peers in the same semester by reading and reflecting on the top three projects shared by the lecturer within LMS. As a final feedback, the distribution of the marks for each part of the project and the distribution of the total marks for the project is provided to students so that they can see where they are sitting compared to their peers.

The second project is the DKD project with a total of 30% weight towards the final mark. This is a group project for the internal students and an individual project for external students. External students are encouraged to form their own groups, however so far only once a group has been formed. Due to the assessment policy of the university, 15% of the marks are allocated to individual contributions and 15% to group work, if groups are formed.

This project starts with forming the groups in week 3 of the semester. Students are given a list of web pages as possible data sources, however they can also use data from other sources or from their work. In week 4, a project plan is submitted so that suitability of the data set and project aims can be confirmed. The project plan has no marks attached to it but it needs to be submitted (0% summative assessment). With the feedback provided either or both written and oral, students modify their project plan and resubmit it. This is similar to a contract. If they want to modify it, they need to inform the lecturer and get her approval.

Each week during the tutorial, students are encouraged to use their project data set as part of tutorial exercise. This way they can get feedback to different parts of their final project throughout the semester. However, there have not been many students taking advantage of this opportunity. Two weeks before the submission of the report, a draft report is submitted (5% worth of total mark for the unit). Written feedback is provided to each draft report. The quality and quantity of the feedback depends on the draft report, the more details are provided in the draft, the better feedback can be given.

Similar to the UDKD project, students are required to submit their (group) project report, the R scripts, IBM SPSS Modeler streams, and any other software log files. In addition to these, the group-work students, must report their experience of

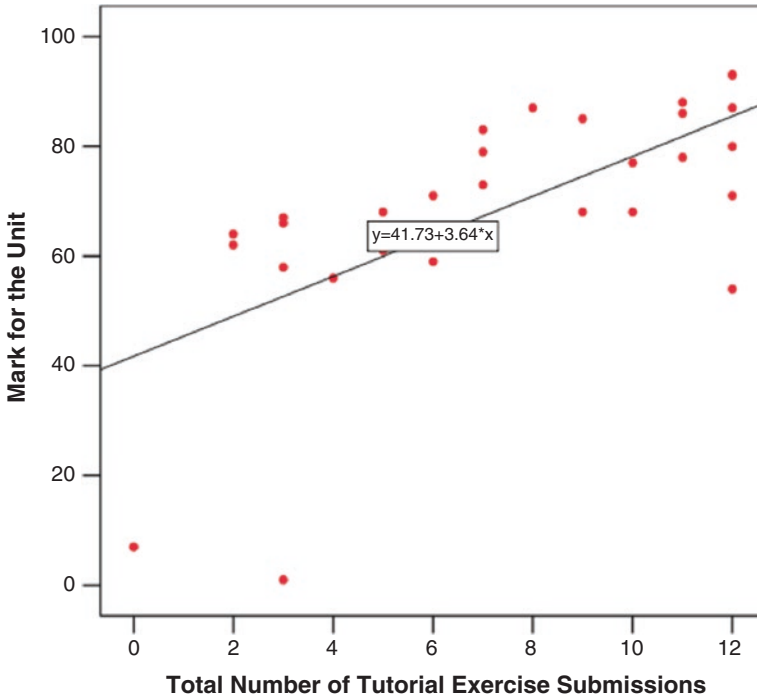
working in the group for this project and peer and self-assessments within the LMS, individually and confidentially. The experience of working in a group submission is tailored around a few questions like “*Did you find it easy or hard to work in a group? Why?*”, “*What kind of help would have been useful to you to improve the group dynamics?*”, and “*Is there any part of the report that you did not contribute at all? If so, why?*”. For the peer and self-assessment, students are asked to rate themselves and their group members on logistics (Did they attend meetings? Did they meet deadlines?), leadership, group dynamics (no disruption to group responsibilities), intellectual contribution, research, writing, and editing. As a last resort to inform the lecturer about any problems with the group or group members, they are given an open ended question “*Is there anything else that you would like to mention?*”

The feedback for this project is given in the same fashion as the UDKD project. In addition, each member of the group might not share the same mark for the group part of the project due to peer and self-assessments. If this is the case, a written feedback by the lecturer is provided to the student(s) where the reasons for the modification of the mark is explained and the student(s) is given a chance to meet with the lecturer to discuss this alteration of the mark. So far, none of the students complained about their reduced marks; if they said anything it was that they did not contribute to group work as much as they could.

An invigilated three-hour final exam is the last summative assessment. Students are allowed to bring one A4 page, hand-written or printed, to the exam. Almost all of the formulae used in the unit are provided to students as the last page of the exam paper. The formulae page of the exam is available to the students from week 1 of the semester within the LMS, so it is no surprise in the exam. Although students are not allowed to take the exam paper from the exam room, previous semester exams are available to them in the university library. Some of the questions in the exam are unique to each student such as “*In your own words, write two advantages and two disadvantages of xxx*”, “*Why would it be a good idea “never to believe that the data set is clean and ready for analysis”? Provide at least five points and explain in your own words by giving examples from your Directed Knowledge Discovery project.*” and “*As a data mining expert, state your advice for improving a classification model’s prediction accuracy for a rarely observed outcome category.*”.

### ***Alignment of Formative and Summative Assessments***

The relationship between the number of tutorial exercise submissions and the final mark students achieved in 2017 cohort is positive and linear (Fig. 3.4), where 43% variability of the final mark is explained with the total number of tutorial exercise submissions. In addition the total number of tutorial exercise submissions are significantly correlated with the UDKD project mark ( $r(30) = 0.68, p < 0.001$ ), the DKD project mark ( $r(30) = 0.64, p < 0.001$ ), and the final exam mark ( $r(30) = 0.57,$



**Fig. 3.4** The relationship between formative and summative assessments

$p = 0.001$ ). These are clear indications of alignment of the formative and summative assessments.

Similar significant positive correlations are observed between the UDKD project mark and the final exam mark ( $r(30) = 0.69$ ,  $p < 0.001$ ), and between the DKD project mark and the final exam mark ( $r(30) = 0.73$ ,  $p < 0.001$ ). These results support that in-semester assessments and end of the semester assessment are aligned.

## Case Studies – Students’ Directed Knowledge Discovery Projects

Students are given responsibility to choose a data set for their DKD project. To aid them on their journey to find a suitable data set, they are provided with a list of data sources such as the Australian Government linkable de-identified 10% sample of Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS) (<https://data.gov.au/dataset/mbs-sample-10pct-1984-gz>), the UC Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>), and the Kaggle Data Science Competitions (<https://www.kaggle.com/datasets>). The search for a data set starts as soon as the groups are formed in Week 3. Some students are not



happy that they have to find a data set, and complain that they had to spend (waste) too much time to find a data set. Instead of seeing it as “freedom to choose your own data set”, this small group of students see it “as a punishment”. However the majority of the students are happy that they can choose a data set that is interesting to them. Some choose to work on a data set related to their work or a data set from their workplace.

The DKD project groups are named each year based on a theme. Since almost one-third of the students were international students, Australian-specific themes have been chosen to educate students in Australian culture and knowledge. Some of the recent themes for the groups names were Australian Prime Ministers, Australian food (i.e., lamingtons, pavlova, fairy bread), capital cities of Australian states, and Aboriginal town names in New South Wales, Australia, such as Wagga Wagga, Kurri Kurri, and Mogil Mogil (Thorpe, 1935).

The groups are formed either randomly by using the birth month of the students or based on some characteristics of the students, for example, their gender (at least one female in each group) to ensure gender diversity, or the degree they were enrolled in, to ensure student strengths and knowledge can be diversified to enable peer learning.

The DKD project is the most important learning and (authentic) assessment task for the data mining unit. It is designed to address all graduate capabilities. For example, “*engaged and responsible, active and ethical citizens*” graduate capability is addressed when students make decisions to include/exclude variables/observations for their project data set; and when they consider the implications of their analysis (and results) for the local or the global society.

### ***Case Study 1: Predicting Airbnb Availability***

A group of students decided to analyse a “listing” dataset from Airbnb (2017) in 2017. Given the increased popularity of Airbnb as an alternative to traditional holiday accommodation all around the world, and Airbnb’s recent plans to move into long-term rental business (Business Insider, 2017; Zaleski, 2017), the students aimed to identify characteristics of an Airbnb listing for high/low availability for Sydney market by using a publicly available data set (Airbnb Sydney, 2016).

The original data set had 95 variables, 24,038 observations. The variables included text fields where the property was described such as a description “*Bright and airy 2 extra large bedroom 2 floor town house in Warriewood sleeping 4-6 people. Open plan living downstairs with living, dining and kitchen opening out onto a small private garden with BBQ and table and chairs, plus trampoline. Main bedroom with ensuite and balcony. 2nd bedroom with bunkbed, cot bed and sofa bed if required. Close to parks, beaches and large shopping centre.*” and house rules “*We have a NO party policy. NO smoking. NO pets. Minimum stay 5 days.*” Some variables were about the host, such as name, location, the duration of Airbnb offering by that host and how responsive the host was in the past. Then more property

related variables were listed such as longitude, latitude of the property, type of the property (i.e., townhouse, apartment, or cabin), how many people can be accommodated, number of bedrooms, bathrooms, etc. Price-related variables included daily price for the property for one person and extra charges for extra guests, cleaning, etc. Cancellation-related variables and review-related variables were also part of the data set. Finally, there were a few variables about the availability of the property. After data preprocessing there were 30 variables and 24,038 observations.

Students developed a number of classification models by using the training data set. The models were tested on the evaluation and test data sets to identify the best model for future predictions of low/high availability of the Sydney Airbnb listed properties. The classification models included decision trees (C5.0, CART, CHAID, QUEST), neural networks, support vector machines (SVM; with Radial Basis Function), and naïve Bayes classifier (NBC). IBM SPSS Modeler was used for all decision tree models and NN, while R and RStudio were used for data preprocessing, SVM, and NBC. Both programs were used for visualisations. The submitted project report had 97 pages with 12 pages of report and an extensive appendix with preprocessing related graphs and details of the models.

Students concluded that the most reliable model was the C 5.0 Decision tree created in IBM SPSS Modeler with the accuracy of “high availability” prediction is around 72–76% and of “low availability” prediction is around 63–66% among training, validation, and test data sets. They also identified *the number of reviews* was the most important predictor variable, however, they also noted that importance level of many predictors were close to each other, therefore they advised the use of all the predictors for future analysis of long-term rental market predictions. Their project was summarised and presented to class with a poster similar to a conference poster (Fig. 3.5).

## ***Case Study 2: Predicting the Popularity of Mashable Online Posts***

The immersion of social media in the daily lives of many people, coupled with online media and news providers in the recent decade, has been a threat to newspapers. The traditional daily newspapers were able to survive not just because of profiting from their sales but because of their advertising revenues. Most digital media are freely available therefore they can't make any money from their readers. An avenue for them to make money, similar to traditional media, is advertisements. Mashable is a global, multi-platform media and entertainment company with specific news sites for Asia, Australia, UK, France, and India (Mashable, 2018). A group of students in 2016 used the data collected from Mashable and applied data mining techniques to identify features of articles that were most likely to be shared on social media.



## PREDICTING AIRBNB AVAILABILITY - EVALUATING THE BUSINESS POTENTIAL FOR LONG TERM STAYS

Group: Mogil Mogil

### INTRODUCTION

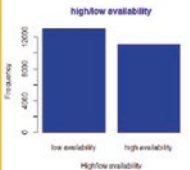
- Airbnb is a global online renting platform allowing home owners to earn extra income leasing their properties. The Airbnb phenomena is quickly disrupting traditional rental and tourist industry operations.
- Recently, the company has been considering expanding its operations into the long-term rental market. This study hopes to assess the viability of such a venture.
- Availability condition of a property is one of the most important factors in assessing the potential of an expansion into long term stays parallel with the existing Airbnb market.
- Objectives:** Identify factors that influence the availability condition of a property and create predictive models using data mining techniques to predict the high/low availability condition of an Airbnb property.

### DATA DESCRIPTION

- The original Airbnb dataset taken from insideairbnb.com is compiled at April 3<sup>rd</sup> 2017, which has a total of 95 variables and 24038 rows.
- Information contained about properties, host identifications, locations, prices, availabilities and descriptions of the properties.
- About 42 variables, containing text descriptions of hosts, places, URL web links, calendars or some have significantly huge amount of missing data, were removed from further analysis.

### DATA PREPROCESSING

- No duplications of entries.
- Number of outliers < 3% of the dataset. So, keep for analysis.
- Variables with >30% missing values are removed.
- Remaining variables (missing data <1%) are imputed with the mode value.
- Variables about host identity verification, accuracy of information, property information and reviews scores will be kept for analysis.
- Variable created (output):** "high\_low\_availability" – "high availability" (> 90 days available in a year) and "low availability" (≤ 90 days available in a year).
- Final Dataset:** 30 variables and 24038 rows of observations.




**Image 1:** Visual plots of high/low availability and number of reviews

### METHODOLOGY

- Split final dataset into training (60%), evaluation (20%), and test set (20%) with random seed 123 in R.
- SPSS Modeler 18 and R are the two statistical tools used for modelling.
- Many classifiers have been applied in building models from the training data: Decision Trees (C5.0, CART, CHAID, QUEST), Neural Networks (MLP and RBF), Support Vector Machines (RBF), and Naive Bayes (no laplace and laplace).

- Each model was evaluated based on the lift, level and stability of the prediction accuracy rates in high/low availability (see Table 1).

Models	Prediction	Training		Evaluation		Test	
		Accuracy	Lift	Accuracy	Lift	Accuracy	Lift
C5.0 Expert (1:2 Cost Matrix) No Window Attribute	high availability	73.06	1.43	71.90	1.35	72.90	1.39
	low availability	65.09	1.41	63.32	1.35	65.00	1.37
CART (1:1 Cost Matrix)	high availability	60.02	1.31	57.76	1.27	59.66	1.28
	low availability	65.85	1.22	65.67	1.20	64.73	1.21
CHAID (1:1 Cost Matrix)	high availability	57.15	1.43	54.34	1.35	55.34	1.37
	low availability	74.52	1.24	72.15	1.21	72.12	1.21
QUEST (1:2 Cost Matrix)	high availability	61.58	1.34	61.44	1.30	61.80	1.31
	low availability	65.07	1.24	65.32	1.24	65.35	1.24
Neural Networks (MLP)	high availability	55.79	1.43	55.14	1.38	55.80	1.41
	low availability	75.05	1.23	73.48	1.22	74.29	1.23
Support Vector Machine (C=100, Gamma=0.001)	high availability	56.40	1.55	53.50	1.45	53.52	1.45
	low availability	80.41	1.26	77.91	1.23	77.80	1.22
Naive Bayes - No laplace	high availability	66.09	1.45	63.77	1.36	62.71	1.36
	low availability	61.75	1.14	59.45	1.12	60.23	1.11

**Table 1:** Comparisons of the best models chosen.

### RESULTS

- The C5.0 Expert-No Window (1:2 cost matrix) was the most reliable model as it balances the trade-off between the accuracy of high and low availability (both have high accuracy in overall) and the lift values achieved are generally high. The model does not overfit the results as the prediction accuracy among partitioned sets were stable.
- Important predictors:**

#### Predictor Importance

Target: high\_low\_availability



**Image 2:** Predictor importance of C5.0-Expert-No Window

- The decision trees and rules from the chosen model were found significant. There are some key insights we can generalize; Firstly, the number of reviews seem likely to have significant impact on the availability condition of a place. Secondly, the importance of levels among predictors seem to be similar. This would indicate that the prediction model generally intakes similar portions of information extracted for each variable.

### CONCLUSIONS

- Even though the chosen model generates better results compared to others, it is hard to justify with our findings a business venture into the long-term market as the differences in predictor importance levels among variables were not significant.
- There is general consensus between models that the most influential factor is the number of reviews, to be expected as those people with more reviews are likely make their places available for more nights per year.
- Limitations:** (1) lack of techniques in mining text variables, (2) missing values, (3) short timeframe analyzed (only the last 365 days), and (4) R and SPSS are quite slow in handling big data.
- Suggestions:** (1) Text mining techniques recommended, (2) advanced method in handling and imputing missing values, (3) extended timeframe might be considered, and (4) better statistical tools in analyzing and processing big data.

Fig. 3.5 Group Mogil Mogil’s DKD poster for predicting Airbnb availability

The original data set had 60 variables, 39,644 observations (published articles on Mashable within a specific timeframe). The variables in the data set included a set of variables for describing the article content, including title and content length; a set of variables for describing the metadata, including keywords included and their popularity; binary variables indicating the channel on which the article was

published; a set of variables regarding the time, such as day of week, and days since published; and a number of subjective indicators of the polarity and sentiment of the article. The outcome variable derived from the number of shares of an online article as popular (6000 or more shares on social media) or unpopular (less than 6000 shares on social media). After preprocessing, the number of variables was reduced to 47, while the number of observations remained the same. The original data set was sourced from the UCI Machine Learning Repository (Dua & Karra Taniskidou, 2017; Fernandes et al., 2015).

Classification techniques applied to the Mashable data included decision trees (C&RT, CHAID, QUEST, C5.0 in IBM SPSS Modeler and tree in R), random forest (in R), NN (in IBM SPSS Modeler), SVM (in R), Generalised Linear Model (GLM) (in R), and NBC (in R). Each model was developed using the same training data set (60% of the data set) and tested on testing (20% of the data set) and evaluation (20% of the data set) data sets. The report had 37 pages, where 12 pages of it was the report and remaining were the additional information.

The best performing model on the evaluation set was the Random Forest with an overall accuracy of 55% (lift of 1.92) in predicting popular articles and 74% (lift of 1.05) in predicting unpopular articles. The most predictive factors were the average shares of the keywords included in the articles and the average shares of any linked articles. However students warned the readers by stating that *“Our finding across all models indicates that it is very difficult to accurately predict article popularity in advance though there are some factors which can increase the likelihood that an article is popular in general terms.”*

It is important to note here that GLM was not covered in the data mining unit. Only an introductory RF information was provided to students since they wanted to know about it towards the end of the semester. This group of students invested their time to master RF and apply what they had learnt in another unit (GLM) to their project. Their project poster can be seen in Fig. 3.6.

## Conclusion

In addition to providing technical data mining knowledge, this unit provided students an opportunity to apply their knowledge to authentic, complex Big Data. The selection of a data set for the projects was left to students, albeit they needed to get an approval from the lecturer. The unit was also designed to help students to learn and/or improve work related soft skills throughout the semester. The importance of language, words to specify which skills is required for a specific job or job application is emphasised throughout the unit, especially after the lecturer had an encounter with a big telecommunication company representative in Australia. While this

# Predicting the popularity of Mashable posts

## What is the secret to attracting high online traffic?

Using the data collected from Mashable's website we attempt to unlock the secret of attracting high online traffic

### INTRODUCTION

Mashable, a digital media website like many other free web-based content provider, relies heavily on selling banner advertisement and advertorial to generate revenue. It is important that their articles generate traffic and are shared by viewers on their Facebook or Twitter accounts. Using the data collected from Mashable's website we attempt to unlock the secret of attracting high online traffic.

#### The objectives:

1. Identify factors that contribute to high sharing of an article
2. Create a model using data mining technique to predict the popularity of an article before it is published.

Online content providers, e-commerce marketers and webhosts could use the findings of this study to optimize their online content, improve their online traffic forecast and enhance their website resources management.

### DATA DESCRIPTION

#### Original Dataset<sup>1</sup>

- 39,644 observations (published articles on Mashable) and 60 independent variables
- Target variable: popularity of an online article measured by number of shares

#### Data Pre-processing

- No missing data and no extreme outliers
- Collapsing of mutually exclusive variables to one variable
- Extreme and impossible entries / perfectly correlated variables are excluded or set to median

#### Features used in Predictions

- Direct article attributes e.g. number of words in the content
- Relationship with other articles e.g. number of total links
- Subjectivity and polarity

#### Final Dataset

- 39,643 observations and 47 independent variables
- Target variable: popularity of an online article – 6,000 or more shares

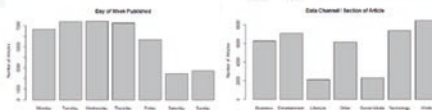


Image 1: Histograms of variables "Day of Week" and "Data Channel"

### METHODOLOGY

Data is split into three samples as 60% training, 20% testing and 20% evaluation sets.

Multiple classifiers are applied to the training data and parameter optimized against the test set using grid search where applicable.

Methods with parameter choice:

- Decision Trees (C&RT, QUEST, tree, C5, aimed for about 10 terminal nodes)
- Naive Bayes Classifier
- Support Vector Machine (cost 100, gamma .001)
- Generalised Linear Model (forward stepwise selection)

Method	Popular		Unpopular		Overall AUC
	Acc.	Lift	Acc.	Lift	
Random Forest	55%	1.92	74%	1.05	71%
SVM	55%	1.91	72%	1.05	72%
Best Tree	52%	1.91	75%	1.00	69%
GLM	52%	1.75	73%	0.99	67%
Naive Bayes	50%	1.7	73%	0.96	66%

Table 1: Best model decided by highest accuracy and lift by class.

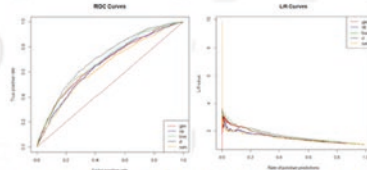


Image 2: Graphical comparison of ROC and Lift Curves for applied models

### RESULTS

The SVM was the best average classifier across all possible levels of the ROC curve (trade-off between true positive and true-negative rates) as indicated by AUC values.

However, as our focus is on popularity which has a 10% natural response rate, when weighting positive accuracy as 5:1 importance the random forest outperforms all other methods and has higher lift at the desired popular accuracy of above 50%.

### CONCLUSION

It is very difficult to accurately predict article popularity as there are a huge amount of unobserved variables including the author popularity and audience type.

In general, to make more popular articles:

1. Choose the **most popular keywords**
2. **Link to other popular articles** within your article
3. **Always have a description** in your article, even if it's just a video

#### Recommendations:

- Provide writers publishing guides on what features, topics are popular
- Predict popular articles and optimise web presence
- Optimise shares by deliberate self references
- Include Author information, time stamp of date published and data collection date in future data for analysis
- Future analysis on Mashable reader demographics
- Compare model to a channel specific model

# Mashable

Group Sydney

<sup>1</sup>Original Data Source: UCI Machine Learning Repository. 2015. *Online News Popularity Data Set*.

[ONLINE] Available at: <http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

Fig. 3.6 Group Sydney's DKD poster for predicting the popularity of mashable online posts

telecommunication company required the applicants to have a strong background in a broad spectrum of machine-learning algorithms, including unsupervised and supervised learnings and a variant of neural network models, they were seeking someone with a PhD or Master's degree in Computer Science or Electrical Engineering. Clearly the MAppStat students would have the required skills, however by specifying the degrees in the advertisement, our students were excluded from the possibility of applying for this specific job. This is in line with what Ridgway (2016) was warning us about when he was writing about Big Data and its place in statistics education "*there are threats to the whole notion of 'statistician' – computer science skills are essential for accessing, storing and processing big data; traditional methods developed by statisticians for data storage, access and analysis cannot be applied.*" Even if the statistical methods can be applied, the employers and so-called "data scientists" might not be aware of them. The assumption is people with "Data Science", "Business Analytics", "Computing", or similar degrees would be good for Big Data analysis but not someone with a statistics degree. Statistics somehow considered very separate from Big Data in the eyes of many people but not by many statisticians.

The necessity to include Big Data analysis techniques into curriculum is highlighted by Ridgway (2016), however, just educating our students is not sufficient, we also need to educate the public, especially employers about the place of statistics and statisticians in the world of Big Data. As Lockhart (2018) argued: "huge new computing resources do not put an end to the need for careful modelling, for honest assessment of uncertainty, or for good experimental design." We all heard "Garbage in, Garbage Out" or as the computer scientists use it "GIGO". Unfortunately, at times it feels like only statisticians and statistics educators care about the quality of the data and importance of the understanding the relationship between the independent and dependent variables.

Cook and Forzani (2018) are hopeful that "the appeal to big data has, we think, been overhyped". Big Data analysis is important for example for biology and medicine due to nature of the data collected in these fields (complex, large data sets), however without understanding the relationship between various biological processes, new theories cannot be developed (Coveney & Dougherty, 2016). Statistical literacy along with the expertise in the area from which the data are coming are important qualities to be successful in the Big Data era (Hewson & MacGillivray, 2016). Just skills in dealing with Big Data are not sufficient. Therefore there might be some light at the end of the Big Data tunnel. As Thamm (2017) also wrote in LinkedIn, "Big Data is dead. Data is "Just Data," regardless of quantity, structure, or speed.", just a mindset and education on data and data analysis is enough for successful data science projects. The future of statistics as a discipline might be bright, if we equip our students with Big Data analysis skills to ride through these hard times.

**Acknowledgments** The first author would like to acknowledge the contributions of late Associate Professor Julian Leslie to the unit development and his endless support for the students and everyone around him.

The authors also would like to thank the students for being engaged and explored more than what was covered in the unit. Without the students' enthusiasm and willingness to learn, it would not have been possible to achieve the unit learning outcomes. Their opinions are valuable.

## References

- Airbnb Sydney. (2016). "Listing" data set from Inside Airbnb. Retrieved from: <http://data.insideairbnb.com/australia/nsw/sydney/2016-12-04/data/listings.csv.gz>
- Airbnb. (2017). "Listing" data set from Inside Airbnb. Retrieved from: <http://insideairbnb.com/get-the-data.html>
- Allen, J. (2018). Who wants to be a statistician? An analysis of ACT-tested public school students. *The American Statistician*. <https://doi.org/10.1080/00031305.2017.1419143>.
- Australian Qualifications Framework (AQF). (2018). Retrieved from: <https://www.aqf.edu.au/aqf-qualifications>
- Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques for marketing sales and customer relationship management* (2nd ed.). Wiley.
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Wiley.
- Bidgood, P. (2010). Relating assessment to the real world. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective*. Wiley. <https://doi.org/10.1002/9780470710470.ch12>
- Bilgin, A. A., Rowe, A., & Clark, L. (2017). Academic workload implications of assessing student learning in WIL. *Asia-Pacific Journal of Cooperative Education (APJCE)*, 18(2), 167–183.
- Business Insider. (2017). *Airbnb might be getting into the long-term rental business*. Retrieved from: <http://www.businessinsider.com/airbnb-might-be-getting-into-the-long-term-rental-business-2017-3?IR=T>
- Chance, B. L. (1997). Experiences with authentic assessment techniques in an introductory statistics course. *Journal of Statistics Education*, 5(3). <https://doi.org/10.1080/10691898.1997.11910596>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 step-by-step data mining guide*. The CRISP-DM consortium. Retrieved from: <https://www.the-modeling-agency.com/crisp-dm.pdf>
- Cook, R. D., & Forzani, L. (2018). Big data and partial least-squares prediction. *The Canadian Journal of Statistics*, 46(1), 62–78.
- Coveney, P. & Dougherty, E. R. (2016). Big data has not revolutionised medicine: We need big theory alongside it. *The Conversation*. Retrieved from: <https://theconversation.com/big-data-has-not-revolutionised-medicine-we-need-big-theory-alongside-it-55356>
- Data Mining Unit Guide. (2018). Retrieved from: [http://unitguides.mq.edu.au/unit\\_offerings/92104/unit\\_guide](http://unitguides.mq.edu.au/unit_offerings/92104/unit_guide)
- Deming, W. E. (1986). *Out of the crisis*. MIT Press.
- Dua, D., & Karra Taniskidou, E. (2017). *UCI machine learning repository*. : University of California, School of Information and Computer Science. Retrieved from: <http://archive.ics.uci.edu/ml>
- Fernandes, K., Vinagre, P., & Cortez, P. (2015). *A proactive intelligent decision support system for predicting the popularity of online news*. In Proceedings of the 17th EPIA 2015 – Portuguese conference on artificial intelligence, September, Coimbra, Portugal. [https://doi.org/10.1007/978-3-319-23485-4\\_53](https://doi.org/10.1007/978-3-319-23485-4_53).

- Fritsch, S., & Guenther, F. (2016). *Neuralnet: Training of neural networks*. R package version 1.33. Retrieved from: <https://CRAN.R-project.org/package=neuralnet>
- Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2018). *Arules: Mining association rules and frequent itemsets*. R package version 1.6–0. Retrieved from: <https://CRAN.R-project.org/package=arules>
- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann Publishers.
- Hand, D. J. (1999). Statistics and data mining: Intersecting disciplines. *SIGKDD Explorations*, 1(1), 16–19.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Hewson, P., & MacGillivray, H. (2016). Big data and little people. *Teaching Statistics*, 38(3), 75–76.
- IBM Corporation. (2016). *IBM SPSS modeler for windows, version 18.0*. IBM Corp.
- Little, R. J. A., & Rubin, D. B. (2014). *Statistical analysis with missing data* (2nd ed.). Wiley.
- Lockhart, R. (2018). Special issue on big data and the statistical sciences: Guest editor's introduction. *The Canadian Journal of Statistics*, 46(1), 4–9.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2017). *Cluster: Cluster analysis basics and extensions*. R package version 2.0.6. Retrieved from <https://cran.r-project.org/web/packages/cluster/index.html>
- MacGillivray, H. (2016). Improving educational connections with authentic statistical practice and real complex data. *Teaching Statistics*, 38(2), 37–39.
- MacGillivray, H. (2017). President's message – October 2017. *The International Statistical Institute (ISI) October 2017 News*. Retrieved from: <https://www.isi-web.org/images/news/2017-10PM.pdf>
- Macnaughton, D. B. (2017). Should we change the name of the field of statistics to “data science”? *Statsjobs blog*. Retrieved from: <https://www.statsjobs.com/change-name-field-statistics-data-science/>
- Macquarie University Handbook. (2018). *Master of Applied Statistics*. Retrieved from: <http://handbook.mq.edu.au/2018/DegreesDiplomas/PGDegree/Master+of+Applied+Statistics>
- Macquarie University Wiki. (2018). *iLab: Macquarie University's personal computer laboratory on the Internet*. Retrieved from: <https://wiki.mq.edu.au/display/iLab/About>
- Mashable. (2018). Mashable. Retrieved from: <https://mashable.com/>
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien*. R package version 1.6–8. Retrieved from: <https://CRAN.R-project.org/package=e1071>
- Puang-Ngern, B., Bilgin, A., & Kyng, T. J. (2017). Comparison of graduates' and academics' perceptions of the skills required for big data analysis. In T. Prodromou (Ed.), *Visualization and statistical literacy for open and big data* (pp. 126–152). IGI Global.
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: <https://www.R-project.org/>.
- Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, 84(3), 528–549.
- Ripley, B. (2016). *Tree: Classification and regression trees*. R package version 1.0–37. Retrieved from: <https://CRAN.R-project.org/package=tree>
- RStudio Team. (2016). *RStudio: Integrated development for R*. Boston, MA: RStudio. Retrieved from: <http://www.rstudio.com/>.
- Sharp Sight Labs. (2016). *What's the difference between machine learning, statistics, and data mining?* Retrieved from: <https://www.r-bloggers.com/whats-the-difference-between-machine-learning-statistics-and-data-mining/>
- Smucker, B. J., & Bailer, A. J. (2015). Beyond normal: Preparing undergraduates for the work force in a statistical consulting capstone. *The American Statistician*, 69(4), 300–306. <https://doi.org/10.1080/00031305.2015.1077731>



- Suzuki, R., & Shimodaira, H. (2015). *Pvclust: Hierarchical clustering with p-values via multiscale bootstrap resampling*. R package version 2.0–0. Retrieved from: <https://CRAN.R-project.org/package=pvclust>
- Thamm, A. (2017). Big Data is dead. Data is “Just Data,” regardless of quantity, structure, or speed. *LinkedIn*. Retrieved from: <https://www.linkedin.com/pulse/big-data-dead-j-ust-regardless-quantity-structure-speed-thamm/>
- Therneau, T., Atkinson, B., & Ripley, B. (2017). *Rpart: Recursive partitioning and regression trees*. R package version 4.1–11. Retrieved from: <https://CRAN.R-project.org/package=rpart>
- Thorpe, W. W. (1935). *List of New South Wales aboriginal place names and their meaning* (3rd ed.). The Australian Museum.
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.
- Wehrens, R., & Buydens, L. M. C. (2007). Self- and super-organising maps in R: The Kohonen package. *Journal of Statistical Software*, 21(5), 1–19.
- Zaleski, O. (2017). *Airbnb said to be exploring expansion in long-term home rentals*. Retrieved from: <https://www.smh.com.au/business/airbnb-said-to-be-exploring-expansion-in-longterm-home-rentals-20170309-gutx9o.html>

**Ayse** is an Associate Professor in the Department of Mathematics and Statistics at Macquarie University. She is the President-Elect of the International Association for Statistical Education (IASE) and an elected Member of the International Statistical Institute (ISI). Ayse’s research explores applications of statistics in health sciences and learning and teaching in statistics.

Ayse was a recipient of several learning and teaching awards for her outstanding contributions to student learning such as an Australian Learning and Teaching Council Citation for ‘Outstanding Contributions to Student Learning’, a Macquarie University VC Citation, ATEM Award Community Engagement (Highly Commended); Higher Education Award for Employability; joint-recipient of Australian Awards for University Teaching (AAUT). Her pedagogical research interests are focused on statistics education, such as learning approaches in statistics, work integrated learning in statistics, impact of learning spaces to students’ learning. Recently she was awarded “Excellence in research: Five Future-shaping Priorities (Healthy People)” by Macquarie University and she was a member of research team which were the finalist of the Eureka Prize.

**Peter Howley** Professor of Statistics and Assistant Dean (Outreach & Engagement) for The University of Newcastle’s Faculty of Science, Peter is nationally and internationally recognised as a leader and award winning academic and practitioner in Statistics and STEM education and initiatives.

Peter is the National Chair of Statistical Education (Statistical Society of Australia (SSA)) and an elected member of the International Statistical Institute (ISI).

He sits on National and International Advisory Boards for STEM, Statistics and Data Science initiatives, is an expert advisor to State and National Educational Boards, National Science and Engineering Challenge Director, inaugural 2018 Science and Technology Australia STEM Ambassador and an International Editorial Boards member (and Guest Editor, *Statistical Education Research Journal Special Edition 2020*).

Peter leads cross-disciplinary school and community engagement initiatives including the National Schools Poster Competition (engaging over 1000 students annually in teams who: undertake a small-scale industry-style investigation on any topic of their choosing; participate in the collection and analysis of data; and report upon their project in poster format, which they submit online). He has established associated national and local support networks and collaborations uniting schools, academics, tertiary, government, industry and national organisations for which he won the ISI Best Cooperative Project Award (2017) and SSA’s National Services Award (2018) for leadership in statistical education.

He has received over two million dollars in grants surrounding education, health, school and community engagement, and interdisciplinary collaborations in STEM, Environmental Sciences, Health and Education.

# Chapter 4

## Statistical Education as Part of the Digital Transformation of Statistical Offices



Markus Zwick and Sabine Köhler

**Abstract** National Statistical Institutes agencies are faced with growing opportunities and challenges from the ongoing digital revolution. This chapter considers the cooperation among academic, public, and private sector entities to further develop statistical skills to make effective use of the newly available data. Such cooperation offers opportunities for better understanding and learning from each other about best practices. This chapter specifically focuses on the European Master of Official Statistics (EMOS), a degree program to prepare students for roles in data analysis and also a knowledge network of official data producers and universities working together to develop best practices and skilled analysts. It also describes the approaches and experiences of the Federal Statistical Office of Germany (Destatis).

**Keywords** Statistics education · Digital transformation · Official statistics · National Statistical Institutes (NSI)

### Introduction – The Digital Revolution Means Opportunities and Challenges for Statistical Offices

The opportunities and challenges faced by official statistics are growing with the digital revolution. In general, official statistics will have to be produced faster, be more precise for small groups, and involve lower burdens for the respondents. To achieve that, statistical offices have to be transformed into modern information providers, otherwise private data producers will provide part of what we currently call official statistics. That would be a big loss for society because private data producers will not offer information as a high-quality public good to the society, to politicians, or to the research community. The digital transformation of statistical offices is more than reaching the goal of being competitive in the information market. It will

---

M. Zwick (✉) · S. Köhler  
Statistisches Bundesamt/Destatis, Wiesbaden, Germany  
e-mail: [Markus.Zwick@destatis.de](mailto:Markus.Zwick@destatis.de); [sabine.koehler@destatis.de](mailto:sabine.koehler@destatis.de)

be a necessary step to maintain an independent information infrastructure for modern democracies.

In order to steer this digital transformation process, well-considered concepts will be necessary. That means using new digital data sources, combining this new kind of data with traditional data like administrative and survey data as well as developing new and more tailor made official statistical products for the different user groups. The whole statistical production process is to be reconsidered, beginning with the Generic Statistical Business Process Model used by most of the National Statistical Institutes (NSI).<sup>1</sup>

For realising the digital transformation process, a lot of new skills and knowledge are essential. Therefore one part of the transformation strategy aspects is human resources management. Especially with the permanent growth of accessible digital data, commonly denoted as Big Data, the requirements regarding competences of data producers as well as data analysts are changing. Analysing or using satellite or mobile phone data needs different skills than those needed to do a survey. For future official statistics products, NSIs require both competences as well as expertise to combine both kinds of data.

Most NSIs in Europe have a mix of staff members with master and bachelor degrees, coming from different academic fields. In particular, those with bachelor degrees often have had only introductory courses in statistics at university. Master and bachelor programmes should include more aspects of new digital data sources. Furthermore, introductory courses in statistics have to be further developed. Introductory courses as well as masters' courses in statistics also reach students who may not necessarily go on to work as a data producer; many may well become data users. Introductory courses should be the focus of statistical literacy programmes run by NSIs (Forbes et al., 2011). Therefore one solution for NSIs could be to strengthen the cooperation with universities on academic programmes.

Another important part of the digital transformation process of NSIs will be permanent internal training. Better educated newcomers are one side of the coin – the other side is the challenge of having a lot of staff members inside the NSIs with skills that are no longer valid. Further educating this experienced staff is a big task for the human resources units inside the NSIs. For official data producers, these questions of personnel development and training are essential (UNECE, 2013).

The following section, Section “Upskilling of statisticians via internal training courses”, will deal with the upskilling of statisticians via permanent internal training inside the NSIs. Section “Cooperation with academic, public, and private entities to further develop statistical skills” focuses on the cooperation with academic, public, and private sector entities to further develop statistical skills. One focus of these considerations is the European Master of Official Statistics (EMOS), a knowledge network of official data producers and universities. In each part, the approaches and experiences of the Federal Statistical Office of Germany (Destatis) are described as well.

---

<sup>1</sup>For the Generic Statistical Business Process Model see <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>

## Upskilling of Statisticians Via Internal Training Courses

### *Training Must Include Modern Subjects in Order to Get the Right Upskilling of Staff*

Modern statistics call for a skilled pool of statistical professionals to face an increasing number of challenges in the production of high-quality statistics (Cameron et al., 2017). Official statisticians need a lot of new competences in addition to classic statistical skills. New skills are needed in the context of projects on web scraping and on the use of scanner data in price statistics, on the internet-based data collection of labour market statistics or on the analysis of mobile phone data. Machine learning is a mix of new and traditional methods. Often the methods used are traditional, like cluster or discriminant analysis. A new phenomenon is the combination with informatics. Other approaches like record linkage or statistical matching are also used today, but only for a few applications. With new digital data, linking new kinds of data with survey and administrative data will be an important issue in the future.

In the long run, nearly all statistical processes will be affected by the digital changes (including confidentiality measures to mask statistical data and the dissemination of data). New methods of sample surveys will be used and combined with model-based instruments like small-area estimation techniques. The use of these instruments has to be assisted by adequate training courses. Also the use of imputation techniques needs to be trained in order to guarantee a good quality of statistics.

All in all, in the context of the digital revolution the following competences can be mentioned and are to be treated in internal training programmes (Köhler & Zwick, 2017):

- Analytical expertise (structured and unstructured data)
- IT skills (including programming, machine learning, and imputation techniques)
- Management skills (including project management, process management, quality management, risk management, and change management)
- Communication and data visualisation skills

The following section describes approaches used by Destatis and in the European Statistical System to ‘skill up’ the staff in this regard.

### *Training at Destatis Is in Line with its Digital Agenda*

In order to face the challenges of the digital transformation, Destatis drafted a digital agenda in 2017 summarising a lot of internal transformation measures referring to all kinds of statistical processes. In this context, a continuous increase and development of skills and competences is planned. Therefore, the training programme

2018 of Destatis is entitled ‘Agile working in a digital environment’. On the one hand, training courses are offered on ‘Machine Learning’, the ‘Use of Geographic Information Systems (GIS) ’; ‘Imputation Techniques’, ‘Small Area Estimation Methods’ and ‘The use of new digital data (including Big Data and Internet of Things)’.

Training measures for internal staff also have to take into account that agile working will help statistical offices to be more flexible in the digital world. In the future, division of labour and more flexible work will become increasingly important. In this context, virtual teams or teams working together for a limited time will be needed. Therefore, Destatis offers training courses on agile working and on agile frameworks for work management (like SCRUM, which has already been used in the IT sector for a long time). SCRUM certificates can be obtained as well.

### ***The European Statistical Training Programme (ESTP) for NSIs Inside the ESS***

Also the European Statistical Training Programme (ESTP) takes into account the challenges of the digital revolution. It provides European statisticians with a wide selection of training courses, workshops, and seminars, greatly enhancing their professional qualifications.<sup>2</sup> In addition, when statistical experts come together, they get to know each other, have opportunities to share their work experience and exchange best professional practices. An interesting and relevant training programme can also play an important role in attracting and retaining new staff members. Small countries in particular often do not have enough participants to offer special courses by their own. In this way the ESTP also shares knowledge across the ESS.

The ESTP courses cover well-established domains, such as methodology, quality, and dissemination, as well as sectoral statistics, such as for economic and social domains. To make sure that the ESTP offer reflects the professional needs of today, the 2018 programme also includes courses on big data, visualisation, and social data mining.

Special Big Data ESTP courses in 2018 are:

- Introduction to Big Data and its tools
- Big Data sources – Web, Social Media and text analyses
- Social data mining
- Advanced Big Data sources– Mobile phones and other sensors

---

<sup>2</sup>For the ESTP programme see <http://ec.europa.eu/eurostat/web/european-statistical-system/training-programme-estp>

Meanwhile many of the ESTP courses are influenced by the question of integrating new digital data sources into official statistics. Therefore ESTP courses include a lot of Big Data content, so this issue isn't only offered by the special courses.

Commonly, ESTP courses run 3 to 4 days each and are placed in one of the offices in an ESS country. Trainers are professionals from the NSIs and depend to the topic from outside. Currently Statistics Netherlands (CBS) as well as the statistical office of Italy ISTAT offer most of the Special Big Data ESTP courses. In particular, CBS, with the Center for Big Data Statistics (CBDS), is one of most important knowledge sources for using Big Data for Official Statistics in the world. With the ESTP courses, the other NSIs in the ESS are able to participate in and benefit from the research of the CBDS.

### ***Competence and Training Profiles and Opportunities to Study at Universities Help Destatis with Systematic Upskilling***

In order to guarantee systematic upskilling of the competences and skills of all staff members, Destatis is currently drafting competence and training profiles for all colleagues. These include competence profiles on digital aspects described in Section [“Training must include modern subjects in order to get the right upskilling of staff”](#). The whole staff is systematically trained in order to get the competences needed (that means academic and non-academic staff). Also the ESTP courses mentioned can be attended by interested staff members.

In addition to internal and external training measures, Destatis offers qualified staff members the opportunity to study at universities in order to obtain a bachelor's or master's degree (especially in the field of IT).

In the long run both approaches can help to increase the competences of a statistical office. All in all, lifelong learning becomes more and more essential in the digital world.

### ***E-Learning Can Help to Make Training more Flexible and Attractive***

Innovative training methods like webinars, learning videos, and e-learning can help to upskill statisticians, pupils, and students. It makes training more flexible as regards the place and time of learning. In this regard the approach of Destatis is a step in the right direction which can be helpful also at international level.

Destatis, together with the statistical offices of the German Federal states (regional offices), developed a large e-learning programme on official statistics in 2016/17. Half of the e-learning modules are about organizational and legal aspects and the other half are about statistical and methodological subjects. There were

authors from all statistical offices, with the Federal Statistical Office accounting for most of them. Users of the programme come from statistical offices and universities as well. This can help to enrich the curricula of universities regarding the characterisation and the content of official statistics.

In order to use the German e-Learning on official statistics internationally, large parts of it will be translated into English by Eurostat. So in the long run all interested people will have access to the programme via the internet.

In addition, the training unit of Destatis is testing the possibilities of producing and of using learning videos, webinars, and other innovative learning methods at the moment. Starting in 2019 these techniques will enrich the training possibilities of the statisticians. In the long run, external users also will have the opportunity to use these products.

## **Cooperation with Academic, Public, and Private Entities to Further Develop Statistical Skills**

As discussed in Section “[Upskilling of statisticians via internal training courses](#)”, for producing high-quality official statistics in the digital age, a lot of new skills are necessary.

For a single NSI, it would not be possible to manage the digital transformation change alone. Therefore, statistical offices around the world are working together (at bilateral as well as multilateral level). For Destatis, for example, the cooperation with the German-speaking NSIs of Austria and Switzerland (D-A-CH) is an important network.

Multilateral cooperation involves developing concepts for using new digital data sources in the production of official statistics and developing concepts for managing human resources. In Europe, the UNECE and the European Statistical System (ESS) are the institutions which steer this process.

Additionally, cooperation with universities and research institutes is very important. Current experience indicates that NSIs’ cooperation with private-sector entities is getting more and more valuable, too.

The following three sections highlight some examples of cooperation with academic, public, and private-sector entities which help develop statistical skills – starting with the description of the European Master in Official Statistics (EMOS).

## *The European Master in Official Statistics (EMOS) as a Way to Combine Training and Recruitment Activities*

EMOS was launched by Eurostat and the ESS in 2014 to connect producers of official statistics and academia at the European level.<sup>3</sup> As a joint project of universities and data producers, EMOS aims at including official statistics in statistical curricula, meeting training and recruitment needs in the ESS and constituting a future recruitment pool of highly educated professional statisticians for the ESS and the European System of Central Banks (ESCB). After two calls for applications, the EMOS network comprises 23 programmes in 15 countries and collaborating partners in statistical offices. The network builds on existing, nationally accredited masters' programmes which are in line with the EMOS learning outcomes. EMOS-labelled masters' programmes also collaborate actively with the National Statistical Institutes or other producers of official statistics for relevant master's thesis topics and internships in the sphere of official statistics. Currently the third EMOS call is running to further expand the EMOS network.

EMOS is very valuable for both NSIs and universities. For NSIs, EMOS offers a lot of different opportunities. First of all, NSIs get direct contact to students, which means they get contact to potential new academic staff. In the competition to recruit well educated academic staff, statistical offices are in a weak position. New digital skills are well paid in the information market. Considerably higher starting salaries are paid by the big players such as Google, Amazon, and others. Therefore innovative concepts are necessary. The opportunity for offering internships and topics as well as support for master's theses in running masters' programmes is an innovative concept. Furthermore, the NSIs participating in EMOS are in a position to influence the content of the curricula. In this way, statistical offices are able to steer the content of education provided to their potential newcomers.

But EMOS is more than a joint academic programme. EMOS offers a lot of opportunities to use the masters' courses for training inside the statistical offices. The EMOS webinars are a tool that a lot of NSIs are using for [further education](#). For all participating NSIs, the EMOS network is an opportunity to get up-to-date knowledge from the universities and to work closely together with potential academic newcomers. Furthermore some staff members of the NSIs are also teachers in the EMOS courses. These lecturers often use their experience and material also for statistical courses inside the statistical offices.

Destatis is very active in the Europe-wide EMOS network and maintains diverse work contacts with the German speaking EMOS masters' programmes. The EMOS curriculum foresees an internship as well as a master's thesis with a strong link to applied statistics. Destatis offers internships for EMOS students and supports students during their master's thesis work. In addition to Destatis, some other official data producers are part of the EMOS network in Germany. The German Central Bank, the Federal Employment Agency, and some statistical offices of the German

---

<sup>3</sup>For EMOS see [https://ec.europa.eu/eurostat/cros/content/what-emos\\_en](https://ec.europa.eu/eurostat/cros/content/what-emos_en)



Federal states (regional offices) are offering support to the EMOS network. A yearly meeting and common initiatives like summer schools tie the German-speaking network.

Five German universities offer EMOS Master programmes:

- Free University Berlin, Master of Science in Statistics
- Technical University Dortmund, Master in Statistics
- University of Bamberg, Master of Science in Survey Statistics
- University of Munich (LMU), Master of Science in Statistics in Economic and Social Sciences
- University of Trier, Master of Science in Survey Statistics

In addition to the German universities, the Austrian Johann Kepler University, Linz offers an EMOS master programme, too, which is also attractive for German students.

For Destatis, EMOS is an opportunity to establish a more intensive network with universities in Germany and across Europe. Cooperation with the academic field has a long tradition in Germany. For more than 10 years Destatis has offered bachelor's and master's courses as an education service for German universities (Hochgürtel & Zwick, 2010). The LMU Munich and TU Dortmund included these courses in their Master programmes at an early stage. Later they were well prepared to fulfil the EMOS requirements and to get the EMOS label. Furthermore, Destatis offered the programme at Goethe University, Frankfurt as well as at the University Freiburg.

All in all, it was and it is helpful for NSIs to be in touch with academics. In this way, NSIs are able to acquire newcomers and educate them with the necessary statistical knowledge. Furthermore cooperation with universities is suitable to get a permanent flow 'of state of the art research'. This is why Destatis has been part of the EMOS initiative from the beginning and has invested resources to develop EMOS together with European universities and Eurostat.

Beside the EMOS master's programmes, single EMOS courses are also an instrument for offering statistical training inside the statistical offices. EMOS webinars are well used for internal training in a lot of NSIs, including in Germany.<sup>4</sup> Most of the EMOS materials are [accessible](#) for interested persons.

Internal training courses and the ESTP are influenced by EMOS but not harmonized. An ESS degree for professional statisticians, such as the 'Graduate Statistician' of the Royal Statistical Society,<sup>5</sup> could be one solution to integrate the current independent education initiatives. There have been first discussions inside the Federation of European National Statistical Societies (FenStatS) to establish a

---

<sup>4</sup>For the EMOS webinar programme in the winter term 17/18 see [https://ec.europa.eu/eurostat/cros/content/emos-webinars-february-june-2017\\_en](https://ec.europa.eu/eurostat/cros/content/emos-webinars-february-june-2017_en)

<sup>5</sup>For the RSS statistician see [http://www.rss.org.uk/RSS/pro\\_dev/pro\\_awards/gradstat/RSS/pro\\_dev/pro\\_awards/Graduate\\_statistician/Graduate\\_Statistician.aspx?hkey=3751895f-02e8-4359-a20b-2a8548cca371](http://www.rss.org.uk/RSS/pro_dev/pro_awards/gradstat/RSS/pro_dev/pro_awards/Graduate_statistician/Graduate_Statistician.aspx?hkey=3751895f-02e8-4359-a20b-2a8548cca371)

curriculum for a European Statistician and to use the different education activities inside the ESS. But these discussions have just started.

### ***Enhanced Cooperation with Universities Can Help to Integrate New Digital Data Sources into Official Statistics***

Since 2016, Destatis has established some new kinds of cooperation with universities as well as with public and private-sector entities. This cooperation grows with the intensifying work on the integration of new digital data sources into official statistics. These activities are more scientific rather than production-oriented. In this constellation, more intensive cooperation with the scientific community is logical. For the topics ‘using of mobile phone data’, ‘integration of satellite data’ and ‘micro simulation’, Destatis has installed doctoral (PhD) positions for academic newcomers where candidates require a PhD as a qualification. Each of these positions is supported by a responsible person at Destatis and a professor from a collaborating university. The contracts run for 4 years. PhD students work at Destatis most of their time. Every 7 weeks they work at the supporting university for 1 week. In this way, the PhD students have the opportunity to participate in the university programmes and to be part of the PhD network at the university, while also bringing knowledge from the universities into the statistical office.

### ***Cooperation with Private Firms Is Important for Staying Competitive***

Another kind of cooperation is connected with the challenges of getting access to privately held data. Access to privately held data is an important issue for statistical offices. There is an intensive discussion inside the EU and also in the individual member states on how to use data which are generated by private enterprises. Some initiatives are on the way to establish European and national regulations on using this data for official purposes like official statistics. Without clear legal regulations which allow NSIs access to privately held data for statistical purposes, it won't be possible to use new digital data sources for official statistics. Otherwise the production of official statistics products is dependent on the willingness of companies and current market situations. This discussion includes the question of resources. If private enterprises have to submit these data for official purposes, the companies will have the burden of data preparation. At this stage, we as NSIs have to decide what we would like do by ourselves and what we should buy in as a service. For handling a big amount of new digital data like mobile phone signals or satellite measurements, the appropriate hardware, software, and new skills are needed. Data scientists with the necessary skills and experience to do this work are rare at the job

market. They can only be hired for a very high salary, which official institutions often cannot pay. These are issues discussed under the title ‘Trusted Smart Statistics’ in the European Statistical System.

More time is needed to find solutions for common agreements how to handle privately held data in combination with official data. Nevertheless, NSIs are already working with data generated by private enterprises. Some cooperative projects have been established across Europe. This is a way for NSIs to have first access to these new data sources. In the long run it will not be enough to have access to privately held data for producing official statistics only in the context of pilot projects. But for feasibility studies, these cooperative efforts are sufficient.

Current experiences show that private companies are also very interested in cooperating with professional data producers and in learning more about their own data. Often the new opportunities of analysing the digital ‘data’ are not the core business of the enterprises. Therefore, this kind of cooperation could be a win-win situation where both NSIs and private enterprises benefit.

## Conclusions

In the age of digitization, official statistics as neutral, independent, and high-quality statistics become more and more important. With the digital revolution, a lot of new information producers have appeared. On the one hand, some of them offer statistical products in classic fields of official statistics like price statistics (e.g., scanner data). Generally, this is a positive development.

On the other hand, private companies are not able to produce public statistics independently and with the same quality as NSIs because it would be too expensive for them and too difficult to get a return on the investment from the market. Furthermore, only NSIs have access to administrative data and the whole benefit of using new digital data is their combination with survey and administrative data.

That means, from our point of view, that statistical offices are necessary and will always be necessary, even in these digital times. But for making NSIs really digital, a lot of transformational challenges have to be solved and processes be modified. Cooperation as well as research and education are answers to these challenges. Internal training and cooperation with universities like EMOS are highly valuable for NSIs and universities.

Not long from now we will see new forms of cooperation between NSIs, academia, and private-sector entities. All of these have to gain experience with the digital transformation and it will benefit all to do that in cooperation. These kinds of cooperation also offer opportunities for better understanding and learning from each other about best practices.

## References

- Cameron, C., Iosua, E., Parry, M., Richards, R., & Jaye, C. (2017). More than just numbers: Challenges for professional statisticians. *Statistics Education Research Journal*, 16(2), 362–375.
- Forbes, S., Camden, M., Pihama, N., Bucknall, P., & Pfannkuch, M. (2011). Official statistics and statistical literacy: They need each other. *Statistical Journal of the IAOS*, 27(3/4), 113–128.
- Hochgürtel, T., & Zwick, M. (2010). *The lecture series “Economic statistics. Data production and data analysis in official statistics.”* ICOTS8. [http://iase-web.org/documents/papers/icots8/ICOTS8\\_10E2\\_HOCHGUERTEL.pdf](http://iase-web.org/documents/papers/icots8/ICOTS8_10E2_HOCHGUERTEL.pdf)
- Köhler, S., & Zwick, M. (2017). *Statistical education in times of big data: Note by the German Federal Statistical Office (Destatis)*. In UNECE conference of European Statisticians. [https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2017/CES\\_26\\_E\\_Next\\_generation.pdf](https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2017/CES_26_E_Next_generation.pdf)
- United Nations Economic Commission for Europe (UNECE). (2013). *Human resources management and training: Compilation of good practices in statistical offices*. [http://www.unece.org/fileadmin/DAM/stats/publications/HRMT\\_w\\_cover\\_resized.pdf](http://www.unece.org/fileadmin/DAM/stats/publications/HRMT_w_cover_resized.pdf)

**Markus Zwick** is Head of Section at the Institute for Research and Development in Official Statistics of the Federal Statistical Office in Germany (Destatis). He is responsible for research cooperation's and strategies. Furthermore he is engaged in different national and European projects concerning the implementation of Big Data into official statistics. Markus Zwick is been working at Destatis since 1996. He started his career as junior researcher at the tax statistics unit. From 2001 until 2012 he developed and led the Research Data Centre of Destatis. From 2012 until 2016 he was responsible for the Projekt 'European Master in Official Statistics' as Seconded National Expert at Eurostat.

He is Honorary Professor for Statistics at the Goethe University Frankfurt and Editor-in-Chief for 'ASTA- Wirtschafts- und Sozialstatistisches Archiv' one of the both journals of the German Statistical Society.

**Sabine Köhler** is Head of Section 'Education and Training' of the Federal Statistical Office in Germany (Destatis). She is in charge of all training matters, including the common training programmes with the statistical offices of the German federal states at national level and at European level (ESTP)). She is been working at Destatis since 1991, before that she worked in the administration of the J. W. Goethe University Frankfurt a. M. for 5 years. In 1991 she started her career at Destatis as Assistant Head of the Section 'National coordination and strategic programme planning'. From 1993 until 2000 she was Head of Section 'Regional statistics/Anonymisation of microdata' and from 2000 until 2009 Head of the Section 'European and International coordination'. In addition, in 2007 she was Head of the German EU Council Presidency Team (Statistics).

**Sabine Köhler is member of the German Statistical Association.**

**Part III**  
**Using Big Data to Improve Teaching**

# Chapter 5

## Big Data, Analytics and Education: Challenges, Opportunities and an Example from a Large University Unit



Ron S. Kenett and Theodosia Prodromou 

**Abstract** The challenge in educational technology (EdTech) is to apply modern analytics to educational data in order to derive information. Information quality (InfoQ) has been proposed by Kenett and Shmueli as a framework for assessing the quality of information generated by empirical studies by using specific empirical methods such as regression models, analysis of variance or predictive analytics. InfoQ is determined by eight dimensions: 1) Data Resolution, 2) Data Structure, 3) Data Integration 4) Temporal Relevance, 5) Chronology of Data and Goal, 6) Generalizability, 7) Operationalization and 8) Communication.

The chapter considers, with an example, opportunities and challenges of analytics in education. Among other topics, it discusses how the InfoQ framework can be applied in order to achieve conceptual understanding and other learning outcomes, and applies the framework to an example concerning academic performance of university students pursuing a Bachelor and Master Degree Programme in Education. The rationale is to provide information regarding the students' performance and their actions on the online learning platform. It investigates how the day of assignment submission affect the grade of the students and we predicted the day of the week for assignment submission, by each student. The results revealed that students received highest grades on Wednesdays and Thursdays and lowest grades on Sunday. It is predicted that when students submit assignments on Sunday, their grades are lower. Days with the highest grades were Thursday, for the first assignment, and Tuesday for the second assignment and final score. The results of the case study provided the unit coordinator with feedback to evaluate and review the unit through the lens of best practices.

---

R. S. Kenett (✉)

The KPA Group and the Samuel Neaman Institute, Technion, Haifa, Israel  
e-mail: [ron@kpa-group.com](mailto:ron@kpa-group.com)

T. Prodromou

University of New England, Armidale, Australia  
e-mail: [theodosia.prodromou@une.edu.au](mailto:theodosia.prodromou@une.edu.au)

© Springer Nature Switzerland AG 2021

T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy Implications of Research in Education 13,  
[https://doi.org/10.1007/978-3-030-76841-6\\_5](https://doi.org/10.1007/978-3-030-76841-6_5)

103

**Keywords** Big data · Learning analytics (LA) · Educational data · Information quality (InfoQ) · University education · Formative assessment · Educational technology (EdTech)

## Introduction

In higher education today, there are many online courses (Huang et al., 2014), and in-service teacher education programmes (Chen et al., 2009). These online university courses are generating a huge amount of data pertaining to students in electronic form. It becomes imperative to effectively transform this massive collection of data into knowledge which will assist teachers, administrators and policy makers with decision making. Furthermore, it becomes imperative for educators, researchers, practitioners and technology designers to seek ways to foster learners' engagement in these online learning environments, support students' development, and understand and improve the effectiveness of the design of the teaching practices and teaching materials that will subsequently inform the design and delivery of the online course.

Owing to the digitisation of academic processes and the rise of learning analytics (Siemens, 2013), it becomes increasingly appealing and feasible to harness students' online actions and discourse data for gaining insights that can aid learners and instructors with sense-making and decision making (Dawson, 2010; Wise et al., 2014). Learning analytics (LA) also portrays the future of online courses by predicting students' dropouts, validating pedagogical interventions, and closing the instructional design cycle.

Inspired by the social perspective of learning, and students' interaction and engagement with the teaching materials provided by teaching practices, this study analyses the performance of university students based on "dynamic" learning traces of students' interaction with the online learning environment and its resources.

## Literature Review

Definitions of learning analytics vary. Generally speaking, LA are defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Long & Siemens, 2011, p. 34).

Junco and Clem (2015) and Xing et al., (2015) define LA as the collection and analysis of data for the prediction of educational outcomes for the purpose of tailoring education. Drachsler and Kalz (2016) and Rubel and Jones (2016) define LA as a means to assist educators examine, understand, and support students' study behaviours and change their learning environment.

Learning analytics (LA) and educational data mining (EDM) are closely associated research areas (Viberg et al., 2018) because their overall purpose is to understand how students learn and support research and practice in education based on the analysis of large-scale educational data (Siemens & Baker, 2012).

There are several differences between LA and EDM (Siemens & Baker, 2012). On the one hand, EDM is an emerging interdisciplinary research that is concerned with developing methods for exploring, researching and applying computerized methods focusing on automated discovery to identify patterns of an enormous volume of data. EDM methods have a strong focus on the analysis of individual components and relationships amongst them. On the other hand, LA use models that are concerned with the “measurement, collection, analysis, and reporting of data about learners and their context, for the purposes of understanding and optimizing learning and the environments in which it occurs” (Siemens & Gašević, 2012, p. 1). Hence, LA aims to understand better complex educational systems as wholes in order to be able to better inform instructors and students (Viberg et al., 2018).

The emerging field of learning analytics may be exploited to:

- Measure key indicators of students’ performance
- Support students’ development
- Understand and improve the effectiveness of teaching practices and improve learning outcomes
- Inform decisions and strategies about the design and delivery of the course.

In this study, we are particularly interested in applying modern analytics to educational data in order to analyse the performance of university students and study students’ activity online that would inform decisions about the re-design of the university unit.

## Research Background

Although the LA research field in higher educational settings is in its infancy, there have already been a number of literature reviews (Avella et al., 2016; Clow, 2013; Ferguson & Clow, 2017; Ihantola et al., 2016; Leitner et al., 2017; Sin & Muthu, 2015) and research studies. The numerous research studies focus explicitly on using different theories to explain different aspects of LA, such as, learning and knowledge outcomes (Kovanović et al., 2015), or learning and knowledge outcomes (Meshulam et al., 2020; Popoola et al., 2018). Others propose study of the performance of university students (Asif et al., 2017) with the use of LA or analysis of learners’ footprints in online learning environments and their personal data, interaction data, and academic information as the necessary sources for LA (Ferguson, 2012).

While the theoretical development of the field is still in its infancy, a few field-specific theories have been developed and applied. Van Barneveld, Arnold, and Campbell (2012), for example, developed the conceptual framework of analytics in education that was used by Agudo-Peregrina, Iglesias-Pradas, Conde-González, and



Hernández-García (2014). Shum and Crick (2012) developed a framework of dispositional LA to investigate the predictions of learning dispositions, outcomes of formative assessments, and other system-generated data in modelling students' performance.

Other research literature focuses on analysing data with the purpose of proposing theoretical models. For example, de Freitas, Gibson, Du Plessis, Halloran, Williams, Ambrose, et al. (2015) propose a foundational LA model of higher education, that focuses on the dynamic interaction of stakeholders using data supported by visual analytics. The *Let's Talk Learning Analytics* framework of issues important to the institutional implementation of LA was suggested by West, Heath, and Huijser (2016). Other pedagogical models explicated analytics interventions based on integration, diversity, agency, reflection, parity, and dialogue (Wise et al., 2014) or codes of ethics based on input and feedback from practitioners across the LA spectrum (Lang et al., 2018). Chen, Chang, Quyang, and Zhou (2017) monitored and devised student-facing social LA and an analytics toolkit in a post-secondary educational setting to foster students' discussion in online learning platforms.

Notwithstanding, we cannot overlook the emerging complexity inherent in the landscape of higher education that creates barriers that prevent data from being used systematically and effectively despite increasing interest in employing learning analytics to increase the quality of teaching and learning (Bichsel, 2012). Bichsel (2012) further claimed that “data quality, data ownership, data access, and data standardization along with organizational culture, and expertise available to implement learning analytics” (p. 15) need to be further investigated and analysed. The challenge of interpreting LA by university teachers requires them to have a statistical background. Teachers without an appropriate statistical background found it impossible to make sense of the LA and translate analytics back into teaching actions by drawing connections between the data and their teaching practices (Jørnø & Gynther, 2018).

The challenge of analysing, interpreting, and making-sense of LA is the foremost barrier to creating value from LA by translating them back into practices of higher education. We follow Kenett and Shmueli (2014, 2016) who proposed the InfoQ framework that considers the issue of translating LA and translating them back to actions related to the teaching of a university course.

## InfoQ Framework

The information quality (InfoQ) framework (Kenett & Shmueli, 2014, 2016) can be applied in order to achieve conceptual understanding and other learning outcomes as defined in (Kenett & Shmueli, 2014, 2016) as the potential of a dataset to achieve a specific (scientific or practical) goal using a given empirical analysis method. Although InfoQ is different from data quality and analysis quality, it is dependent on these components and on the relationship between them. The purpose of the

InfoQ framework is to survey statistical methods for increasing InfoQ at the study design and post-data collection stages.

Technically, the definition of InfoQ is the derived utility ( $U$ ) from an application of a statistical or data analytic model ( $f$ ), to a data set ( $X$ ), given the research goal ( $g$ ). This can be written algebraically as:  $\text{InfoQ}(f, X, g, U) = U(f(X|g))$ .

Kenett and Shmueli (2014) propose eight dimensions in order to assess InfoQ or information quality:

1. *Data Resolution* refers to the measurement scale and level of aggregation of the data. The measurement scale of the data is evaluated with regards to its suitability to the goal, the analysis methods to be used, and the required resolution of  $U$ . Taking into consideration the original recorded scale, its adequacy should be evaluated. The level of aggregation of the data must also be adequate in relation to the goal and to the task at hand.
2. *Data Structure* relates to the structured quantitative data in different forms (e.g., cross-sectional, time series, and network data), as well as unstructured data, non-numerical data (e.g., semantic data, audio, video, text, text with hyperlinks). The InfoQ-level of a certain data type depends on the goal of the statistical investigation. Missing values and corrupted data need to be removed or handled by imputation, data recovery or other methods, depending on the goal. Wrong values might be treated as missing values when the task is to estimate a population parameter, such as in surveys where participants intentionally enter wrong answers.
3. *Data Integration* Data is often spread out across multiple data sources. Hence, properly identifying the different relevant sources, collecting the relevant data, and integrating the data, directly affects information quality.
4. *Temporal Relevance*. A data set contains information collected during a certain time window. The process of deriving knowledge from data can be placed on a timeline that involves the data collection, data analysis and study deployment periods including the temporal gaps between these periods. The degree of relevance of the data in that time window to the current goal at hand must be assessed and is crucial to attain sustainable control of the process or the product.
5. *Chronology of Data and Goal*. The selection of variables to collect, the temporal relation amongst the variables and their meaning in the context of the *goal* critically affect InfoQ. Depending on the nature of the goal, the chronology of the data can support the goal to different degrees.
6. *Generalizability*. There are two types of generalizability: statistical and scientific. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations.
7. *Operationalization* is about deriving concrete actions from the information provided by a study.
8. *Communication*: If the information does not reach the right person at the right time in a clear and understandable way, then the quality of information becomes poor.

The InfoQ framework is applied to a case study in order to achieve conceptual understanding and other learning outcomes.

## Research Goal and Questions

This chapter presents information about academic performance of university students pursuing a Bachelor and Master Degree Programme in Education. The rationale is to provide information regarding the students' performance and their actions on the online learning platform.

Since teaching online is still emerging, we shall present research questions, but will not propose specific hypotheses. Our first question is:

- How does the day of assignment submission affect the grade of the students?

There are several factors that could impact on the grade of the students such as the teaching materials, video recordings, online video lectures, interaction of the students with the unit coordinator, suitable technological tools, interactivity of online video lectures, learners' attention span (Geri et al., 2017), time allocated by students for individual study, etc.

It is important to point out that this study considers only the individual student because there is neither collaboration nor interaction with the other students regarding the submission day of the assignments, and no communication with the unit co-ordinator. The students work individually according to their own learning style, pace, and family/family obligations, and they submit their assignments before the due date. The unit coordinator planned the submission date to be Monday, so the students could spend the weekend to work on their assignments if needed. However, one may wonder if Monday is indeed the most suitable day for the due date of the assignment or if the unit coordinator can predict the day of the week for assignment submission by each student based on each student's engagement with the teaching materials of the online platform every day of the week. Since, those issues are crucial for the unit coordinator's reflective strategies in order to evaluate and review the unit through the lens of best practices, the second question that we examine is:

- What is the predicted day of the week for assignment submission, by each student?

The analysis of Moodle data in a course with student assignments can be evaluated to assess the impact of the day of delivery and predict student grades.

The goal is to help the unit co-ordinator and study programme directors improve the programme and redesign the online unit. The derived utility is enhanced educational outcomes and more effective educational processes.

## The Case Study

### *Context and Participants*

The research context was an undergraduate and postgraduate online course offered at a large public university in Australia in a 12-week trimester. This study explored students' participation during the online course; their online activity, including completion of online activities, and their interaction with their peers and unit co-ordinator; and their performance. The design of the course was informed by socio-cultural perspectives of learning. Online, asynchronous discussion about online activities and the completion of those activities was a central component throughout this course. Students participated in weekly discussions, in which they posted reflections on weekly online tutorial activities and interacted with peers, the instructor of their group, and the unit co-ordinator. They used Moodle, an online learning platform for the duration of the trimester. Moodle provided discussion forums within the unit, community sites, and messaging, as well as providing access to online video conferencing that supported academic interactions within and outside of formal teaching.

When posting on Moodle, students were required to post their reflections on three weekly tutorial activities and also provide constructive feedback in response to a peer's post and to add material to supplement the peer's work.

Moodle provides teaching resources, video-recordings of lectures, two-hour video-recordings of online tutorial sessions which ran once per week, recordings of video-based communications amongst students and the unit-coordinator, activities, and announcements for the unit. Moreover, Moodle facilitates the submission of assignments.

Participants were undergraduate and postgraduate primary teachers who came from a range of courses including Bachelor of Education (Early Childhood and Primary), [Bachelor of Education \(K-12 Teaching\)](#), [Associate Degree in Teaching \(Pacific Focus\)](#), Bachelor of Special Education (Primary), Bachelor of Special Inclusive Education (Primary), Bachelor of Early Childhood Education, and [Master of Teaching \(Primary\)](#). Nearly three-quarters of them were undergraduate teachers and one-sixth of them were international or ESL students.

Participants were taught by the same instructor who was also the single unit co-ordinator and one of the investigators of this study. The unit was attended by 251 students using a total of 176,186 recorded web transactions in 2017. The unit covers the content and pedagogy of teaching Year 5 and Year 6 school students in the strands of Data, Chance, Patterns, and Algebra, and later Number. The university students were expected to become able to demonstrate their personal content knowledge in these strands, discuss associated teaching strategies, use the understandings to create developmental learning sequences, and demonstrate an understanding of the importance of statistical literacy and numeracy. The expected learning outcomes of the unit are as follows:

- demonstrate knowledge and understanding of the concepts, substance and structure of the content and teaching strategies of the teaching of the mathematics in Data, Chance, Algebra and Number, necessary for interpreting student assessment data and evaluate student learning for the purposes of keeping accurate and reliable records of student achievement and modifying teaching practices (GTSD 2.1.1, 5.3.1, 5.4.1, 5.5.1);
- demonstrate knowledge and understanding of strategies for differentiating teaching to meet specific learning needs across the full range of abilities, and for planning and preparing outcomes-based mathematics programs for years K-12 (GTSD 1.5.1);
- know and understand numeracy teaching strategies and their application in teaching mathematics to provide children with opportunities to build on their existing mathematical understandings and numeracy (GTSD 2.5.1);
- demonstrate knowledge of a range of resources including ICT to engage students in their learning (GTSD 3.4.1); and
- demonstrate an understanding of working with data and basic statistical literacy including planning, gathering and applying data to solve problems in an effective learning and teaching sequence (GTSD 2.2.1)

Before the beginning of the trimester, the unit-coordinator uploaded a study schedule that guided the students with weekly instructions about the expected readings from the set textbook, the video recordings they were expected to watch, and the tasks they were expected to perform.

Moreover, the unit co-ordinator ran a two-hour Zoom session every week to provide guidance students during their engagement with solving mathematical problems with or without the use of technological tools. After solving those problems, a discussion followed about the integration of those mathematical exercises in teaching practices and how they fit into educational taxonomies (e.g., the taxonomy of educational objectives of Bloom et al., 1956; the taxonomy for learning, teaching, and assessing; Bloom's Taxonomy Revised by Anderson & Krathwohl, 2001; Six Facets of Understanding, Wiggins & McTighe, 1998; the SOLO [Structure of Observed Learning Outcomes] taxonomy by Biggs & Collis, 1982; and Cooperative Learning Principles of Mills & Cottell, 1998), and practical classroom techniques (e.g., use of hands-on activities).

There were four forums on the Moodle online learning platform. One forum, titled "general discussion", was for students to discuss any topics of general interest that were related to the students' study of the unit. Two forums, "Forum questions about assignment one" and "Forum questions of assignment two", provided students with opportunities to post their questions and discuss with peers and the unit co-ordinator specific aspects of the assignments. The fourth forum was the virtual space where students could solve their weekly tutorial activities, receive feedback on their work, and provide constructive feedback to the workings of their peers.

The students were required to complete two assignments and submit them via Moodle. In week 6, students were required to submit assignment one, and in week 11 students were required to submit assignment two. Both assignments were due by the end of the day, Monday, and both were written tasks of 2000 words' length. The first assignment had a probability and statistics focus and was concerned with statistical literacy. The second focused on number, which is taught the last two years of primary school, as well as patterns and algebra. Although the two assignments counted 50% towards the overall mark of the unit, all students were required to

attend and participate in at least 80% of the online weekly tutorial activities by their forum postings. Students not meeting the latter requirement received an NC (Compulsory Fail) grade for the unit.

Completion of the mathematical exercises of the diagnostic quizzes did not count towards the overall grade of the unit. It simply provided students with feedback about their understanding, conceptual gaps, and difficulties with certain mathematical concepts.

## Data Sources

To address the research questions related to students' actions on Moodle and use of the teaching materials, we analysed student access to the Moodle learning platform and the specific features recorded in data logs.

The data logs recorded the sequence of time-stamped events and actions that occur in the learning platform, such as users' clicks in the learning platform, time and duration of their engagement with specific readings, tutorial activities, time spent listening to a lecture recording, messages exchanged between different users, and students' posting on Moodle.

## Data Analysis

The data was analysed using the JMP version 14.2 software ([www.jmp.com](http://www.jmp.com)). Twenty-four (24) types of activities were logged on the course web site with 10% being labelled "assignments" (see Fig. 5.1).

An analysis of the assignment-related activity by the day of the week reveals some interesting patterns (Fig. 5.2). Days with significantly higher activity were Tuesdays (3030 versus the expected 2756 under a day-independent assumption). Wednesday and Thursday also had some surges in activity; however, these were moderate compared to Tuesdays. Friday and Saturday showed a reduction in activity (1334 versus an expected 1626 and 1882 versus an expected 2187, respectively), which suggests Fridays and Saturdays are less dedicated to assignment-related activity. The discrepancies between actual and expected counts under independence are assessed using the chi square cell contribution.

Following the assessment of activity regarding assignments, we integrated the grades of the participating students with the web activity related to assignments.

The distribution of the grades of 173 students and some summary statistics are displayed in Fig. 5.3.

The scatterplot of the three grades are showing a correlation reflecting the ability of students (see Fig. 5.4).

Following data integration by student name, an analysis of means was performed to assess average student grades by day of the week when assignments were

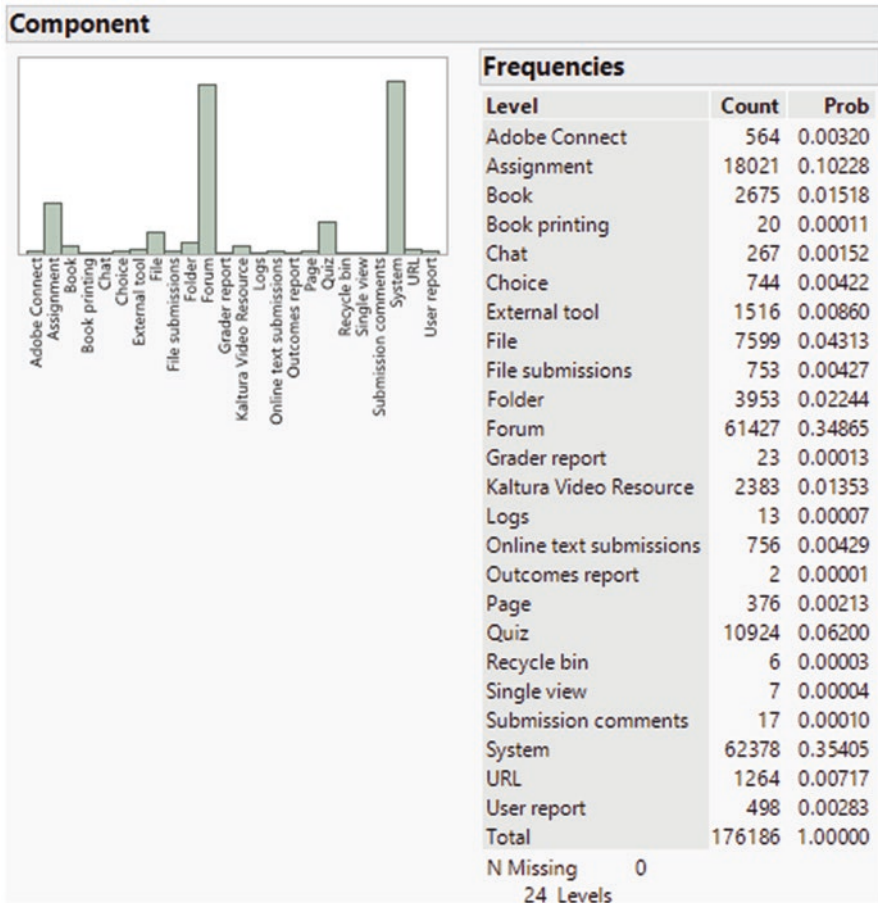


Fig. 5.1 Descriptive statistics of online activities

submitted (Fig. 5.5), Wednesday and Thursdays were with highest grades, Sunday with the lowest grades.

The typical assignment submission day (the average day), by student is shown in Fig. 5.6.

A decision tree analysis of students, by day of assignment submission, provides, for every student, a predicted day for submission of assignment.

Based on these predictions, we can compare student grades by day. Sunday is a day where students who submit assignments as a “preferred day” are getting lower grades.

The analysis performed is based on pair comparisons using student’s *t* corrected for multiple comparisons. The lower disjoint circles indicate a significant difference highlighting the lowest grade days (Fig. 5.7).

**Fig. 5.2** Assignment related activity, by the day of the week

Contingency Table			
Assignment			
Count	0	1	Total
Total %			
Col %			
Row %			
Expected			
Cell Chi^2			
1	21708 12.32 13.72 89.65 21736.4 0.0371	2505 1.42 13.90 10.35 2476.6 0.3256	24213 13.74
2	35801 20.32 22.64 89.70 35828.7 0.0215	4110 2.33 22.81 10.30 4082.25 0.1886	39911 22.65
3	23923 13.58 15.13 88.76 24196.1 3.0834	3030 1.72 16.81 11.24 2756.86 27.0619	26953 15.30
4	21092 11.97 13.34 89.27 21210.3 0.6602	2535 1.44 14.07 10.73 2416.66 5.7946	23627 13.41
5	21566 12.24 13.64 89.15 21716.6 1.0451	2625 1.49 14.57 10.85 2474.35 9.1721	24191 13.73
6	14568 8.27 9.21 91.61 14275.5 5.9940	1334 0.76 7.40 8.39 1626.52 52.6078	15902 9.03
7	19507 11.07 12.33 91.20 19201.2 4.8686	1882 1.07 10.44 8.80 2187.75 42.7306	21389 12.14
Total	158165 89.77	18021 10.23	176186



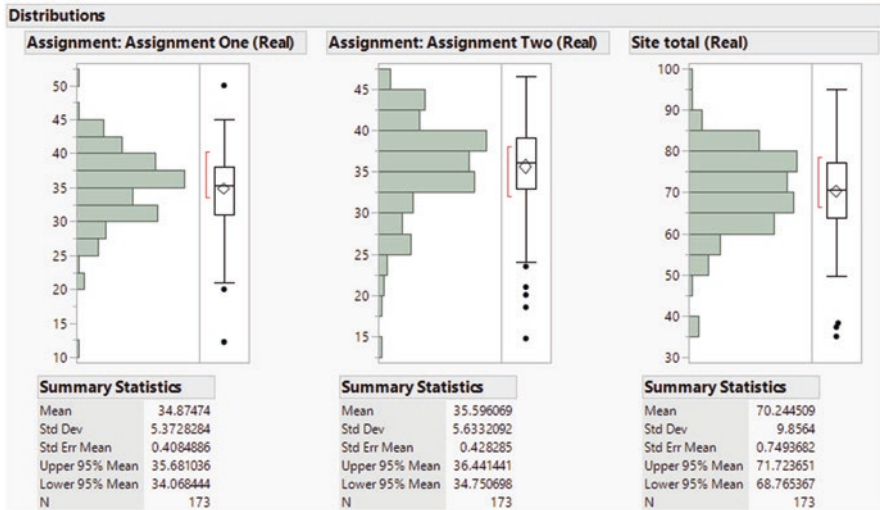


Fig. 5.3 Distribution of the grades of 173 students in Assignment one, Assignment two and the final grades

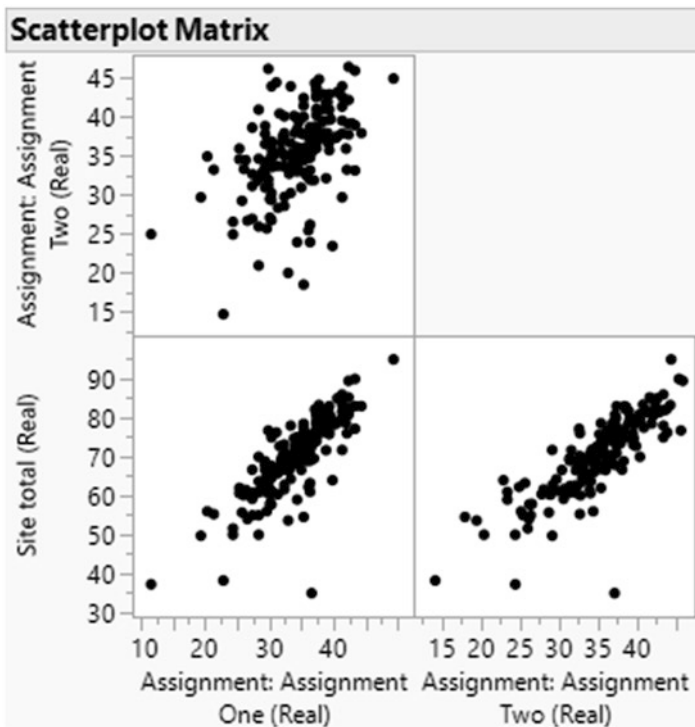
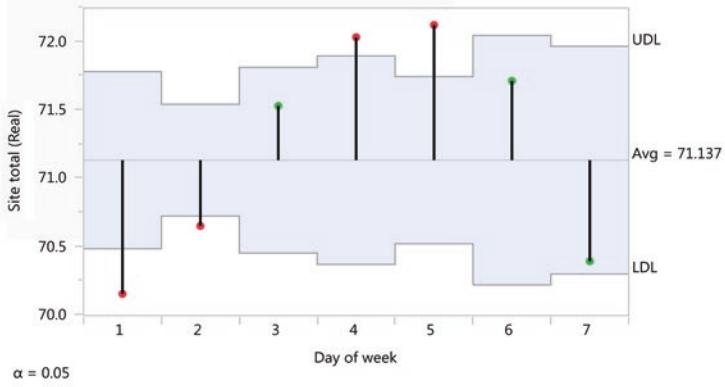
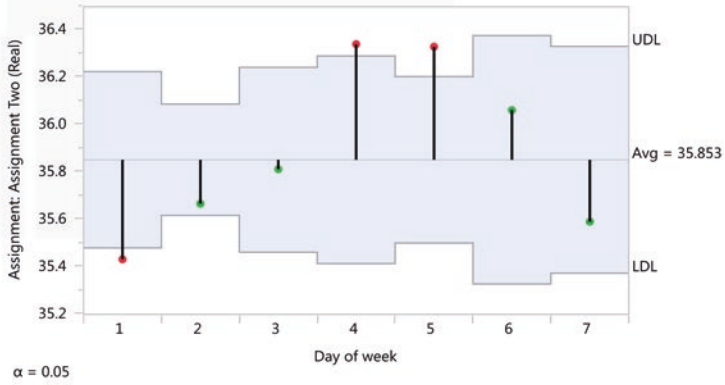


Fig. 5.4 Scatterplot of the three grades (assignment one, assignment two, and final grade)

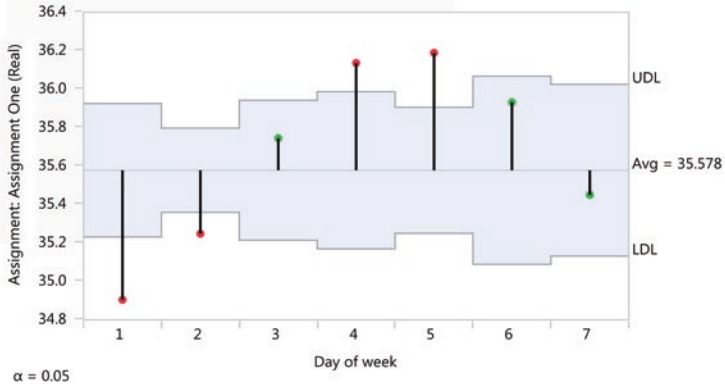
**Analysis of Means**



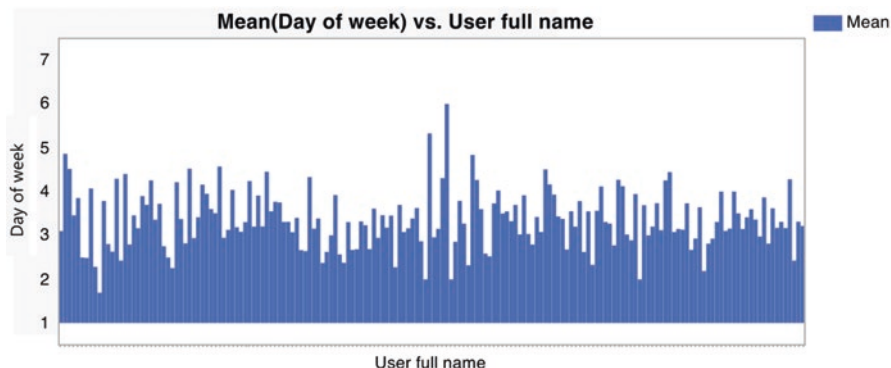
**Analysis of Means**



**Analysis of Means**



**Fig. 5.5** Analysis of means of student grades by average day of submission



**Fig. 5.6** Prediction of day of the week for assignment submission, by student. (Names have been omitted from the figure. Each bar represents a single student)

The days with lowest high grades are highlighted in Fig. 5.8. It is Friday for the first assignment and Sunday for the second assignment and the final score.

Finally, we used this data to fit a least squared model to the final grade using the grades of the first two assignments and the day of the week of the assignment. The summary statistics are presented in Fig. 5.9.

The observed versus predicted is shown in Fig. 5.10. We observe a cluster of points with actual grades around 35 when the predicted grades from the model are about 75. All the points in the cluster correspond to one student who delivered most assignments on Monday but unusually few assignments on Tuesday.

The analysis of Moodle data in a course with student assignments can be evaluated to assess the impact of the day of delivery and predict student grades. This analysis provides the teacher insight on factors affecting grades and a way to highlight unusual grades that should be double checked. Table 5.1 evaluates the analysis using the 8 InfoQ dimensions. The goals, data, analysis and utility of the analysis are described above.

The analysis we presented adequately covers five of the information quality dimensions listed in Table 5.1, considering the goals stated in the section on Research Goals and Questions. Three dimensions have not been considered: Chronology of Data and Goal, Operationalization and Communication. Our case study is an ex-ante analysis to demonstrate the type of information learning analytics can provide. To close the loop one needs to take a holistic approach and consider in the loop, course designers, education technology experts and instructors. The findings in this study need to be communicated to these three communities so that they can act upon it. The course designers can consider the assignment content and the full picture enveloping the student such as other courses and other assignments given. In this way, a course is not considered in isolation but as part of an integrated picture affecting the student educational experience and performance. The education technology experts can find ways to alert the instructor on assignment delivery patterns and the high predicted value of the two considered assignments on the final

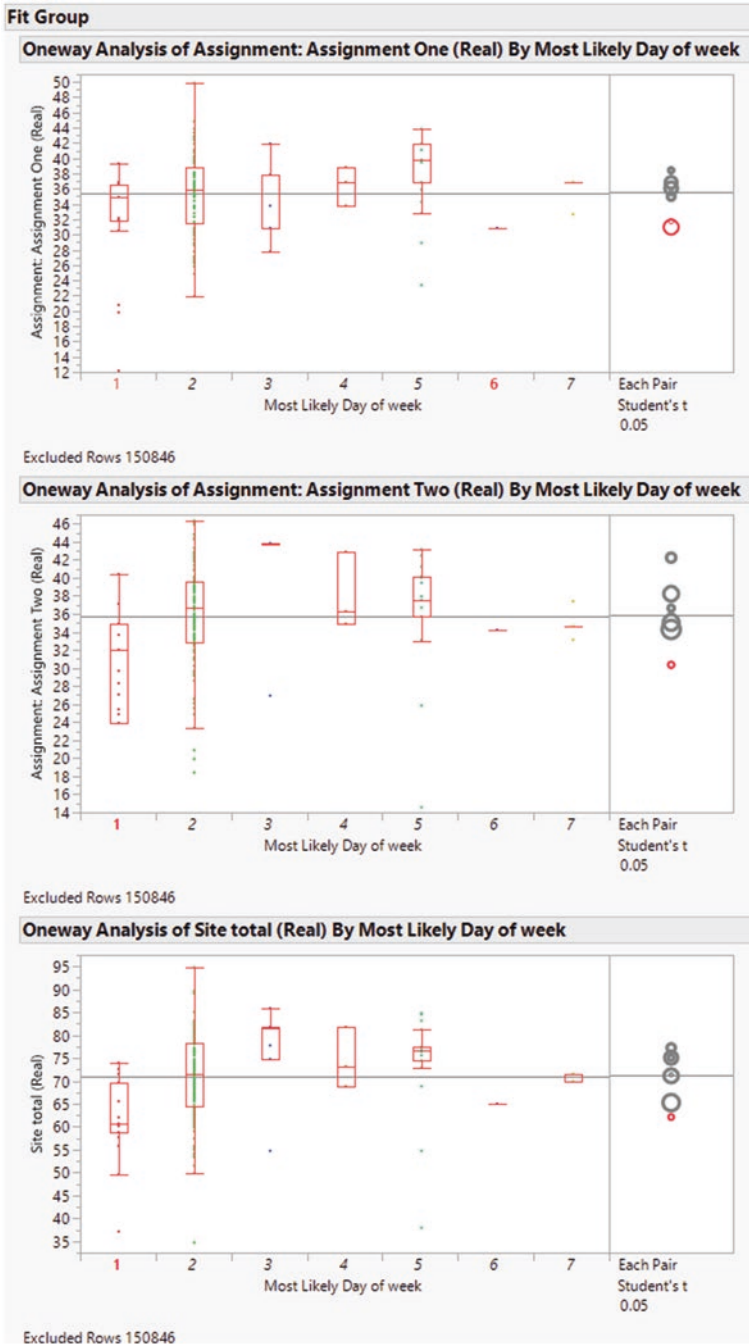
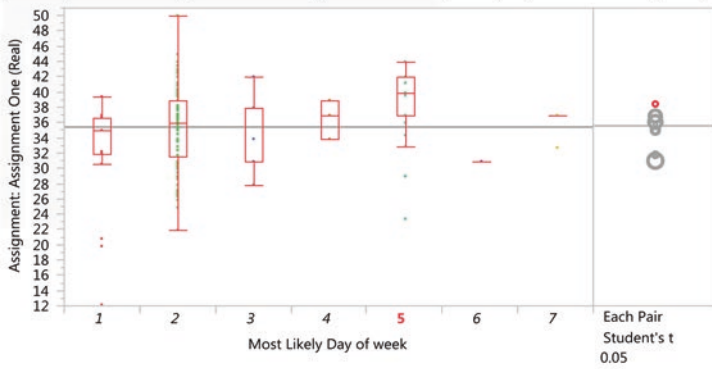
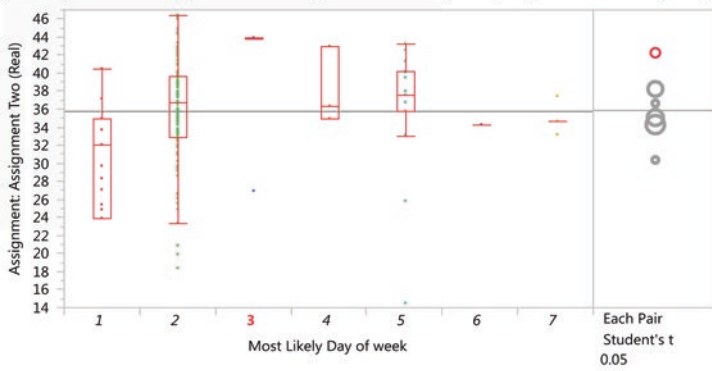


Fig. 5.7 Days with lowest grades

One-way Analysis of Assignment: Assignment One (Real) By Most Likely Day of week



One-way Analysis of Assignment: Assignment Two (Real) By Most Likely Day of week



One-way Analysis of Site total (Real) By Most Likely Day of week

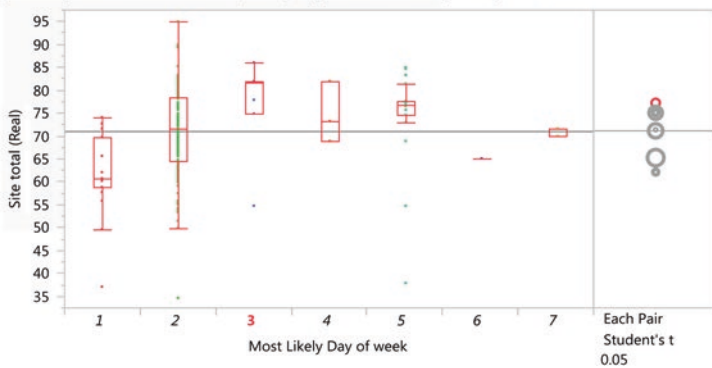


Fig. 5.8 Days with highest grades

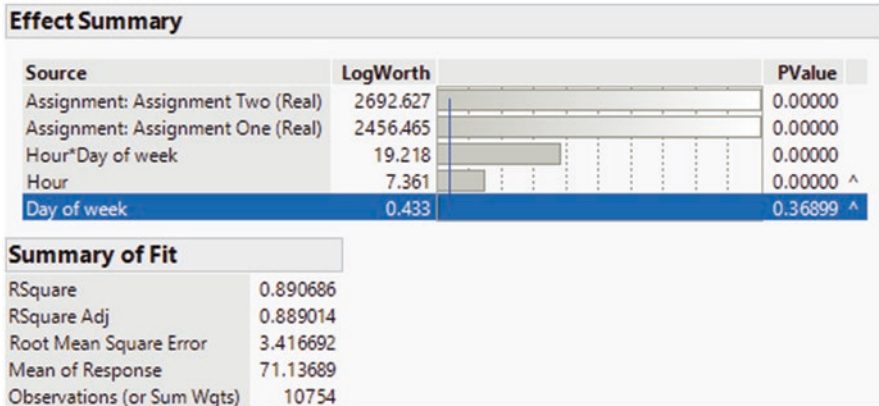


Fig. 5.9 Effect of assignment one and two grades, submission day of the week, hour of submission, and their interaction on final grade

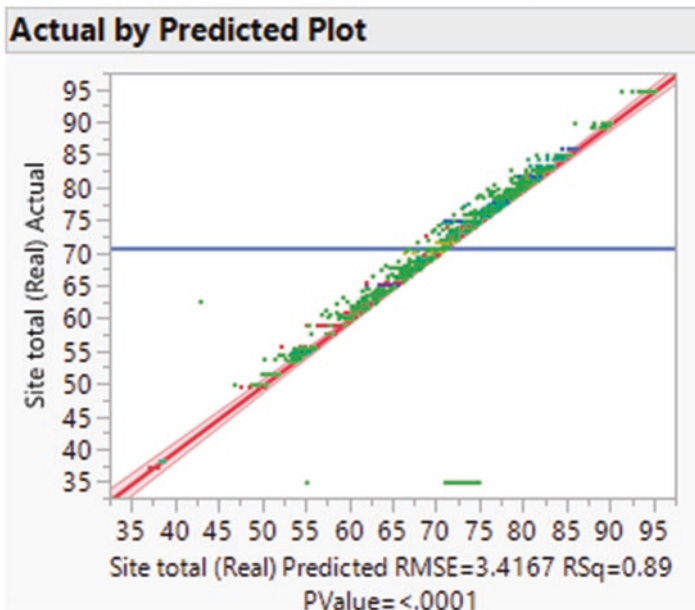


Fig. 5.10 Observed versus predicted final grade

score. This overlap could perhaps lead the instructor to consider for a final score a problem-based assignment, thereby expanding the educational experience of the student. If one can predict with accuracy the score a student gets in an exam, perhaps that exam is redundant. In addition, this makes visible unusual scores like the outlying student in Fig. 5.10, allowing a focused investigation into these special circumstances. Finally, the analysis we conducted should be communicated to the

**Table 5.1** Evaluation of the analysis using the 8 InfoQ dimensions

Dimension	Questions
1. Data resolution	Is the data granularity adequate for the intended job? Yes. The data are by student and activity and therefore adequate for the study goals
2. Data structure	Is it possible to use data from different sources that reflect on the problem at hand? Indeed. The data consists of both activity and grades.
3. Data integration	How is data from different sources integrated? Integration is by student ID.
4. Temporal relevance	Does the time gap between data collection and analysis cause any concern? No. data were available as they were generated. Students had real-time access to use data for their learning during the course, and teachers and researchers were able to use data to plan adaptation of the course.
5. Chronology of data and goal	Are the analytic findings communicated to the right persons in a timely manner? This has not been considered here.
6. Generalizability	Can general conclusions be derived beyond what was explicitly studied? The effect of day of the week might be significant to other courses.
7. Operationalization	Are there any stated action item recommendations derived from the study? This has not been considered here.
8. Communication	Are findings properly communicated to the intended audience? This has not been considered here.

instructor with both a short-term and long-term perspective, Short term, they can initiate interventions on the class they are currently teaching. Long term, they can review pedagogical approaches and assignments for enhanced educational experience. In addition, the instructor can share the analysis of the course with others and form a reference group for benchmarking and identification of best practices.

We did not quantitatively score the eight information quality dimensions or offer an aggregated score. This can be done of course (see Kenett & Shmueli, 2016) but we wanted to show the value of the information quality assessment, even at a qualitative level.

## Conclusions

This study examined student activity by day of the week on the Moodle learning platform.

The results revealed that the highest activity was on Tuesdays and Wednesdays, and Thursday also had some surges in activity. There was a reduction in activity on Friday and Saturday. When integrating the grades of the participating students with the web activity related to assignments, the results showed that the grades of assignment one, the grades of assignment two, and the total grade of the two assignments show a correlation reflecting the ability of the students. Furthermore, we investigated the average student grades by day of the week when assignments were submitted. The results revealed that students received highest grades on Wednesdays and Thursdays and lowest grades on Sunday. Finally, we analysed each student by

day of assignment submission and we obtained for every student, a predicted day for submission of assignment. The results showed that when students submit assignments on Sunday, their grades are lower. Days with the highest grades were Thursday, for the first assignment, and Tuesday for the second assignment and final score.

## Future Research

Future research should seek further best practices to find out the effect of essential elements (e.g., including online video lectures, teaching materials, video recordings, online video lectures, interaction of the students with the unit coordinator, suitable technological tools, interactivity of online video lectures materials) on students' performance.

The analytics of educational activity can be organised into four groups:

1. Monitoring: Tracking activities as the course is delivered
2. Diagnostics: Providing possible reasons for behavioural patterns or student performance
3. Prognostics: Predicting the impact of behavioural patterns on student outcomes
4. Prescriptive analysis: Providing explicit guidance on how to handle specific situations.

Further research aiming at improving monitoring, diagnostic, prognostic and prescriptive analytics in education is needed.

One promising research direction for increasing the students' performance by day, would be to replicate this research study but use a different day of the week as the due date for submission. If the new submission day is associated with better outcomes, the unit coordinator will set that day as the due date for submitting future assignments.

Other issues that could be studied using online data might include the study of a seemingly linear online learning, which is supposedly based on serial viewing of online video lectures, reading particular chapters from the set textbook, solving tutorial activities etc. (as per the study schedule made available by the unit-coordinator before the beginning of the trimester). Nevertheless, much learning occurs ad-hoc, as people encounter a problem they want to solve and look for the relevant information online. If the relevant video lecture or instructional video foster students' performance, it would be interesting to examine whether those factors influence students' performance. Overall, we plan to increase our understanding about whether all the teaching and learning materials and their basic characteristics have an impact on the assignment submission day and grade.



## References

- Agudi-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior*, *31*, 542–550.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*. <https://doi.org/10.1016/j.compedu.2017.05.007>
- Avella, J., Kebritchi, M., Nunn, S., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning*, *20*(2), 13–29.
- Bichsel, J. (2012). *Analytics in higher education: Benefits, barriers, progress, and recommendations*. EDUCAUSE Center for Applied Research.
- Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. David McKay.
- Chen, Y., Chen, N.-S., & Tsai, C.-C. (2009). The use of online synchronous discussion for web-based professional development for teachers. *Computers & Education*, *53*(4), 1155–1166.
- Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683–695.
- Dawson, S. (2010). "Seeing" the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, *41*(5), 736–752.
- De Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., Dunwell, I., et al. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *British Journal of Educational Technology*, *46*(6), 1175–1188.
- Drachsler, H., & Kalz, M. (2016). The MOOC and learning analytics innovative cycle (MOLAC): A reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning*, *32*, 281–290.
- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6), 304–317.
- Ferguson, R., & Clow, D. (2017). Where is evidence? A call to action for learning analytics. In *LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference, ACM International Conference Proceeding Series* (pp. 56–65). ACM.
- Geri, N., Winer, A., & Zaks, B. (2017). A learning analytics approach for evaluating the impact of inter-activity in online video lectures on the attention span of students. *Interdisciplinary Journal of e-Skills and Lifelong Learning*, *13*, 215–228. <https://doi.org/10.28945/3875>
- Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums. In *Proceedings of the First ACM Conference on Learning @ Scale Conference* (pp. 117–126). ACM.
- Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S., Isohanni, E., et al. (2016). Educational data mining and learning analytics in programming: Literature review and case studies. In *Proceedings of the 2015 ITiCSE working group reports* (pp. 41–63). ACM.
- Jørnø, R. L., & Gynther, K. (2018). What constitutes an 'actionable insight' in learning analytics? *Journal of Learning Analytics*, *5*(3), 198–221.
- Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, *27*, 54–63.
- Kenett, R. S., & Shumeli, G. (2014). On information quality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(1), 3–38. <https://doi.org/10.1111/rssa.12007>
- Kenett, R. S., & Shumeli, G. (2016). *On information quality: The potential of data and analytics to generate knowledge*. Wiley.

- Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R., & Hatala, M. M. (2015). Penetrating the black box of time-on-task estimation. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 184–193). ACM.
- Lang, C., Macfadyen, L., Slade, S., Prinsloo, P., & Sclater, N. (2018). The complexities of developing a personal code of ethics for learning analytics practitioners: Implications for institutions and the field. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 436–440). ACM.
- Leitner, P., Khallil, M., & Ebner, M. (2017). Learning analytics in higher education – A literature review. In A. Peña-Ayala (Ed.), *Learning analytics: Fundamentals, applications, and trends: A view of the current state of the art to enhance e-learning* (pp. 1–23). Springer.
- Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), 341–340.
- Meshulam, M., Hasenfratz, L., Hilman, H., Liu, Y.-L., Nguyen, M., Norman, K. A., & Haon, U. (2020). Think like an expert: Neural alignment predicts understanding in students taking an introduction to computer science course. *bioRxiv*. <https://doi.org/10.1101/2020.05.05.079384>
- Mills, B., & Cottell, P. G. (1998). *Cooperative learning for higher education faculty*. Oryx Press.
- Popoola, S. J., Atayero, A. A., Badejo, J. A., John, T. M., Odukoya, J. A., & Omole, D. O. (2018). Learning analytics for smart campus: Data on academic performances of engineering undergraduates in Nigerian private university. *Data in Brief*, 17, 76–94.
- Rubel, A., & Jones, K. (2016). Students privacy in learning analytics: An information ethics perspective. *The Information Society*, 32(2), 143–159.
- Shum, S. B., & Crick, R. D. (2012). Learning dispositions and transferable competencies: Pedagogy, modelling and learning analytics. In *LAK '12: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 92–10). LAK.
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *The American Behavioral Scientist*, 57(10), 1380–1400.
- Siemens, G., & Baker, R. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the second international conference on learning analytics & knowledge* (pp. 252–254). ACM.
- Siemens, G., & Gašević, D. (2012). Special issue on learning and knowledge analytics. *Educational Technology & Society*, 15(3), 1–163.
- Sin, K., & Muthu, L. (2015). Application of Big Data in education data mining and learning analytics – A literature review. *ICTAC Journal of Soft Computing*, 5(4), 1035–1049.
- Van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. *Educause Learning Initiative*, 1(1), 1–11.
- Viberg, O., Hatakka, M., Balter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110.
- West, D., Health, D., & Huijser, H. (2016). Let's talk learning analytics: A framework for implementation in relation to student retention. *Online Learning*, 20(2), 30–50.
- Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Association for Supervision and Curriculum Development.
- Wise, A. F., Hausknecht, S. N., & Zhao, Y. (2014). Attending to others' posts in asynchronous discussions: Learners' online "listening" and its relationship to speaking. *International Journal of Computer-Supported Collaborative Learning*, 9(2), 185–209.
- Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, 47, 168–181.

**Professor Ron S. Kenett** is Chairman of the [KPA Group](#), Israel, Senior Research Fellow at the [Samuel Neaman Institute](#), Technion, Haifa, Israel. abd Research Professor at the University of Turing, Italy. He is an applied statistician combining expertise in academic, consulting and business domains. He developed, with colleagues an approach to assess conceptual understanding of

students based on meaning equivalence reusable learning objects. Ron is Past President of the *Israel Statistical Association* (ISA) and of the *European Network for Business and Industrial Statistics* (ENBIS). He authored and co-authored over 250 papers and 14 books on topics such as data science, industrial statistics, biostatistics, healthcare, customer surveys, multivariate quality control, risk management and information quality. He is member of the Israeli national council for statistics overseeing the work of the Israel Central Bureau of Statistics. The KPA Group he founded in 1994, is a leading Israeli firm focused on generating insights through analytics. He is editor in chief of Wiley's [StatsRef](#), serves on the editorial board of several international journals and was awarded the [2013 Greenfield Medal](#) by the Royal Statistical Society and, in 2018, the [Box Medal](#) by the European Network for Business and Industrial Statistics. Ron holds a BSc in Mathematics (with first class honors) from Imperial College, London University and a PhD in Mathematics from the Weizmann Institute of Science, Rehovot, Israel.

**Dr Theodosia Prodromou** is a senior lecturer of Mathematics Education at the University of New England in Australia. She has published numerous journal articles, book chapters and edited/ authored books on a wide range of subjects, including the use of big data in educational settings, the relationship between technology and mathematical thinking, the integration of digital technologies in the teaching and learning of mathematics, STEM education, augmented reality in Educational settings, professional development of mathematics teachers, and statistical thinking including perceptions of probability and chance. Her most recent edited/authored books were, *Augmented reality in educational settings* (2020) and *Primary and Middle Years Mathematics: Teaching Developmentally* (2019).

# Chapter 6

## Enhancing Learning Outcomes with ‘Big Data’ from Pedagogy for Conceptual Thinking with Meaning Equivalence Reusable Learning Objects (MERLO) and Interactive Concept Discovery (INCOD)



Masha Etkind, Theodosia Prodromou , and Uri Shafrir

**Abstract** Learning outcomes of traditional pedagogy focus on memory of facts, problem-solving procedures, and multiple-choice or true/false questions. Pedagogy for conceptual thinking focuses on higher-order thinking skills, exploration of equivalence of meaning among ideas, and relationships between issues that denote commonality of meaning across representations. MERLO learning assessments capture these important aspects of conceptual thinking. In this chapter, we will look at how MERLO generates big data that can be used to assist teaching. ‘Big data’ for each student in large undergraduate class, include scores of MERLO CIFD weekly quizzes, mid-term tests, and final exams in individual courses. This data shows details of the evolution of deep understanding of each concept in the course, from weekly formative MERLO quizzes that reveal individual students’ conceptual strengths and weakness. This allows the instructor to suggest individual corrective measures with Interactive Concept Discovery (InCoD) in the course digital Knowledge Repository (KR), that are conducted and discussed by individual students with their peers, and enhance learning outcomes.

---

M. Etkind  
Ryerson University, Toronto, Canada

T. Prodromou  
University of New England, Armidale, NSW, Australia  
e-mail: [theodosia.prodromou@une.edu.au](mailto:theodosia.prodromou@une.edu.au)

U. Shafrir (✉)  
University of Toronto, Toronto, Canada  
e-mail: [uri.shafrir@utoronto.ca](mailto:uri.shafrir@utoronto.ca)

**Keywords** Meaning equivalence reusable learning objects (MERLO) · Big data · Interactive concept discovery (InCoD) · Knowledge repository (KR) · Formative interactive cooperative diagnostic (CIFD) assessment

## Big Data and Learning Analytics

Big data are characterised by the 4 Vs: volume, velocity, variety and veracity.

Volume refers to the size of the data. Many different data sources generate several terabytes of data. The International Data Corporation “IDC has released a report on the ever-growing datasphere, what it calls the collective world’s data, and just like the recent [Cisco study](#), the numbers are staggering. IDC predicts that the collective sum of the world’s data will grow from 33 zettabytes this year to a 175ZB by 2025, for a compounded annual growth rate of 61 percent” (Patricio, 2018).

Velocity refers to how fast the data is being produced. For example, our smart phones generate large amounts of data during use, such as the frequency of incoming data from many SMS messages or videos being watched on YouTube, not to mention data from various types of sensors embedded in the phone. These data can be text data captured by the mobile phone combined with data in other formats such as audio, video images, photos, geolocation data, etc.

Variety means that the data can comprise of a variety of forms (Twitter feeds, videos watched on YouTube, pictures, audio files, MRI images, web pages, web logs, etc.), which can be unstructured data. This implies that it cannot be easily stored in a relational format and that we have to use non-traditional methods to analyse or query it.

Veracity refers to the trustworthiness of the data and taking into consideration that there are inherent discrepancies in all data collected.

All of the above four characteristics of data are driving for alternate, non-traditional solutions that help us analyse and make sense of big data. New data-processing technologies allow analysis of large quantities of data that could not be easily gathered or analyzed in the past. One example of those solutions is an evolving trend in educational informatics, called *learning analytics*. Learning analytics uses big data, for measurement, collection, analysis and reporting of information about learners—for example, engagement and achievement—aiming to understand and optimise learning and the environment in which it occurs (Lockyer et al., 2013; p. 1441).

Learning analytics often focuses on technological issues, such as digital content delivery and interactive learning support. Its primary focus is on “the development, testing and evaluation of the educational content over small form factors devices in order to provide an interactive form of learning support... The metadata presentation forms the learning analytics system, which will serve as the performance benchmark for the educators, technology developers, education administrators and stakeholders” (Abhyankar & Ganapathy, 2014; p. 345).

In a typical educational setting, data are available that can be used to monitor and analyse different aspects of students’ learning, including behaviour, achievement, assessment, etc. Learning analytics “dashboards” include data on learners’ classroom activities and a variety of variables such as time spent to complete an activity, social interaction, document and tool use, artefacts produced, exercise results/quizzes, etc.. The above variables are recorded by tracking learners’ traditional learning – related activities. We will look at MERLO, which generates big data. In this paper, we will look at one pedagogical system–MERLO–and *the big data this system generates* that can be used to assist teaching.

## What Is MERLO?

Higher-order conceptual thinking skills are now recognized as a cornerstone of effective learning, ways of thinking that explore patterns of meaning in the context of a conceptual framework (Bransford et al., 2004). Meaning equivalence is a construct that denotes common meaning across representations: a polymorphous – one-to-many – transformation of meaning. Meaning Equivalence Reusable Learning Objects (MERLO) is a pedagogical tool for both teaching and assessment, by asking students to sort and map important concepts through exemplary target statements of particular conceptual situations, and relevant statements that may – or may not – share that same meaning. MERLO assessment items for different concepts, combined into a large database for a course of study, can provide valuable information about students’ learning patterns (Etkind et al., 2010, 2016) (Fig. 6.1).

Each node in such MERLO database is an item family, anchored by a Target Statement (TS) that describes a conceptual situation and encodes different features of an important concept; and also includes other statements that may – or may not – share equivalence-of-meaning with TS. Collectively, these item families encode the conceptual mapping that covers the full content of a course (a particular content area within a discipline). Statements in the four quadrants of the template, namely, Q1; Q2; Q3; and Q4; are thematically sorted by their relation to the Target Statement that anchors the particular node (item family). They are classified by two sorting criteria: Surface Similarity (SS) to the target; and Meaning Equivalence (ME) with the target. ‘Surface similarity’ refers to similar modes of representation, for example, use of natural language, or use of a graph, and similar elements, such as similar or the same words in different orders as the target statement. By ‘meaning equivalence’ we mean that a majority in a community that shares a sublanguage (Kittredge, 1983) with a controlled vocabulary (e.g., statistics) would likely agree that the meaning of the statement being sorted is equivalent to the meaning of the target statement. Each MERLO assessment item contains five unmarked statements: an unmarked Target Statement plus four additional (unmarked) statements from quadrants Q2; Q3; and Q4. Our experience has shown that inclusion of statements from quadrant Q1 makes a MERLO assessment item too easy, because it gives away the shared meaning due to the valence-match between surface similarity and meaning

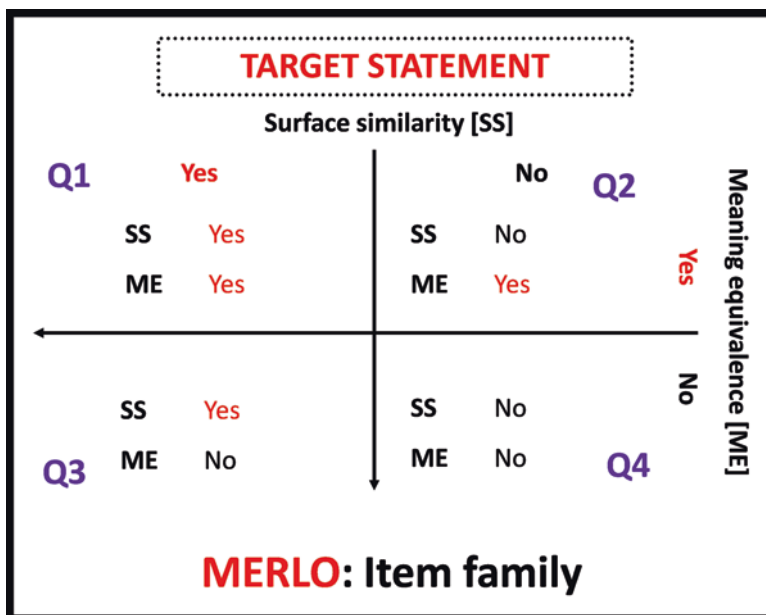


Fig. 6.1 Template for constructing an item-family in MERLO

equivalence, a strong indicator of shared meaning between Q1 and the target statement. Therefore, Q1 statements are excluded from MERLO assessment items. Task instructions for MERLO assessment are:

- At least two out of these five statements – but possibly more than two – share equivalence-of-meaning. (1) Mark all statements – but only those – that share equivalence-of-meaning.
- (2) Write down briefly the reasons for making these decisions.

There are two scores for each MERLO item, namely, *recognition score* and *production score*. The learner is first asked to carry out a *recognition task* in a situation where the particular target statement is not marked, namely, features of the concept to be compared are not made explicit. In order to perform this task, a learner needs to begin by decoding and recognizing the meaning of each of the five statements in the set. This decoding process is carried out, typically, by analyzing concepts that define the ‘meaning’ of each statement. Successful analysis of all the statements requires deep understanding of the conceptual content of the specific domain. MERLO item format requires both rule inference and rule application in a similar way to the solution of analogical reasoning items. Once the learner marks those statements that – in their opinion – share equivalence-of-meaning, they formulate and briefly describe in writing the concept/idea/criteria they had in mind when making these decisions (*production task*). Figure 6.2 is an example of MERLO assessment item (mathematics/functions; Independent Learning Center (ILC), of TVOntario). Its 5 representations include text, equations, and diagrams; at least 2 two of these representations share equivalence-of-meaning (Ripley et al. 2004).

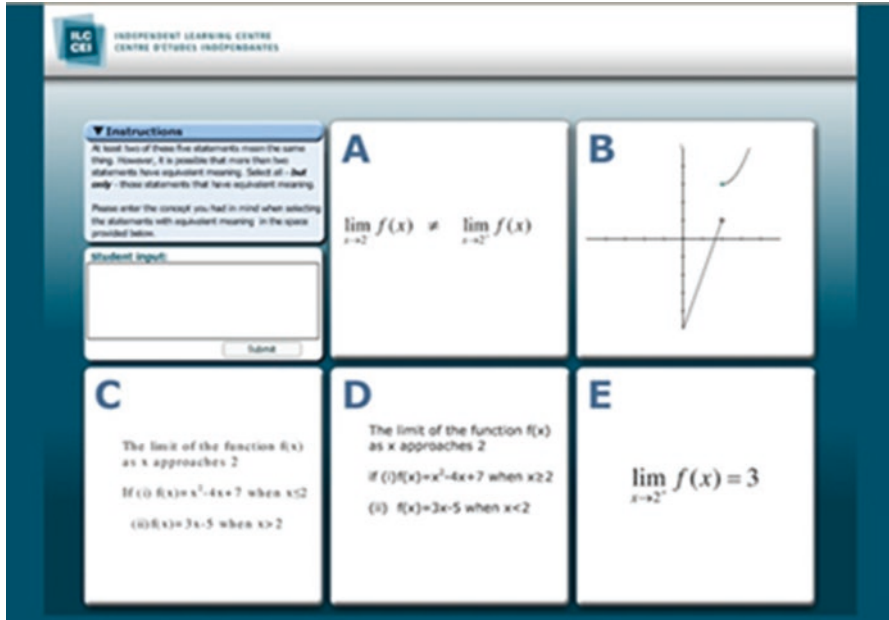


Fig. 6.2 Example of MERLO item (mathematics/functions). In this example, A, B and D share equivalence of meaning

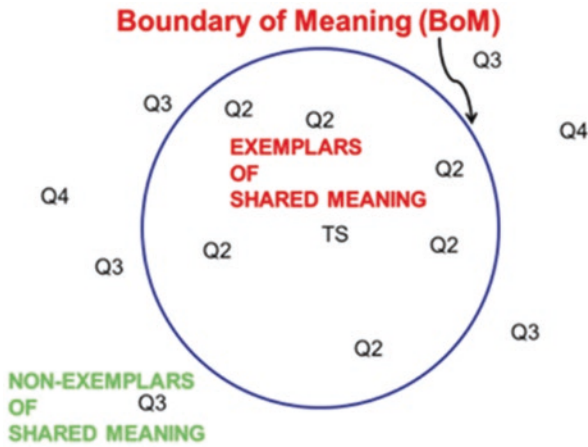
### Diagnostics of Misconceptions with Boundary of Meaning (BOM)

Recognition and production scores provide clear and reliable evidence for diagnosing misconceptions, and clues for remediation. Given a community of specialists that share a sublanguage, and a Target Statement that encodes a particular conceptual situation; then ‘Boundary of Meaning’ (BoM) is defined as the boundary between two mutually exclusive semantic spaces in the sublanguage:

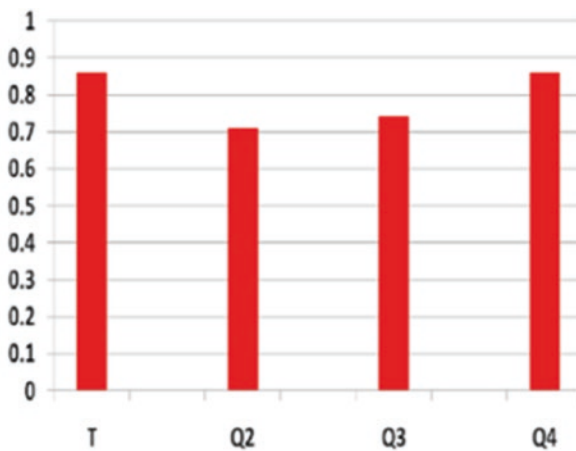
- a semantic space that contains *only representations that share equivalence-of-meaning* with the Target Statement.
- a semantic space that contains *only representations that do not share equivalence-of-meaning* with the Target Statement.

Figure 6.3 illustrates a practical way to demarcate the BoM for a particular Target Statement in a knowledge domain. Specific conceptual comprehension deficits are traced if the learner’s response to a MERLO assessment item with five unmarked statements, includes choosing Q3 statements; as well as if the learner does not choose Q2 statements. Therefore, specific comprehension deficits can be traced as depressed recognition scores on quadrants Q2 and Q3, due to the mismatch between the valence of surface similarity and meaning equivalence (see Fig. 6.1 above). Figure 6.4 shows mean class scores for TS; Q2; Q3; and Q4; of 12 MERLO





**Fig. 6.3** Boundary of Meaning (BoM) of a particular TS include Q2 statements; and exclude Q3 statements and Q4 statements



**Fig. 6.4** MERLO final exam mean recognition scores for T and quadrants Q2, Q3, and Q4; (physics/magnetism), Lycee Physico-Technical High School, St. Petersburg (2003)

assessment items in the final exam of a grade 11 physics/magnetism course at Russian Academy of Sciences, Ioffe Physico-Technical Institute, Lycee Physico-Technical High School, St. Petersburg (2003). The proportions of correct scores for each statement type for each student were calculated for Target Statement TS, and quadrants Q2, Q3, and Q4; for example:

$$Q2_{\text{score}} = \frac{\text{Number of } \textit{correctly chosen} Q2 \text{ statements in the twelve MERLO statements}}{\text{Total number of } Q2 \text{ statements in the twelve MERLO assessment items}}$$

$$Q3_{\text{score}} = \frac{\text{Number of } \textit{correctly unchosen} Q3 \text{ statements in the twelve MERLO statements}}{\text{Total number of } Q3 \text{ statements in the twelve MERLO assessment items}}$$

The U-shaped form of these results reveal reduced scores in Q2 and Q3 representations. However, the interpretations of these two reduced scores are very different:

- *Reduced score on Q2* indicates that the learner *fails to choose* (include within the Boundary of Meaning (BoM) of the concept) certain statements that do share equivalence-of-meaning (but do not share surface similarity) with the target; such depressed Q2 score signals *an over-restrictive (too exclusive)* understanding of the meaning of the concept.
- *Reduced score on Q3* indicates that the learner did choose (*fails to exclude* from the Boundary of Meaning (BoM) of the concept) certain statements that do not share equivalence of-meaning (but that do share surface similarity) with the target; this depressed Q3 score signals *an under-restrictive (too inclusive)* understanding of the meaning of the concept.

Therefore, the U-shaped form of these results reveal certain *deficiencies of both over-restrictive and under-restrictive understanding of the meaning of the concept.*

## Weekly Quizzes with Merlo Formative Assessments

Pedagogy for conceptual thinking was designed to motivate and engage learners, and encourage cooperation. It includes weekly classroom MERLO *Formative Interactive Cooperative Diagnostic (CIFD)* assessment with a MERLO item for one important concept, and provides learners with opportunities to cooperate through discussions in small groups; then make their own decisions and send their individual responses to the instructor’s computer via mobile communication devices; followed by class discussion. MERLO formative assessment takes about 20 minutes, and includes the following 4 steps:

- *Step 1: Peer cooperation and small group discussion* (approx. 3–5 min), includes: PowerPoint projection of the MERLO item; students discuss the item in small groups (turn back/sideways to discuss with those seated next to them).
- *Step 2: Individual response* (3 min); each student sends their individual *recognition response* through a personal communication device (smartphone/tablet/clicker), marking at least 2 out of 5 statements (A; B; C; D; E) in the multi-semiotic MERLO item that, in their opinion, share equivalence-of-meaning; then sends a *production response*, briefly describing the reasons they had in mind when making the above recognition decisions.

- *Step 3: Feedback and class discussion of students' production responses* (3–5 min), including projection of the MERLO item with several production responses sent by individual students, and the instructor's description of the MERLO conceptual context; class discussion and comparison of the various individual production responses.
- *Step 4: Feedback and class discussion of students' recognition responses* (3–5 min), including projection of the MERLO item, showing the correct recognition responses; class discussion and comparison of the various individual recognition responses.

The weekly MERLO CIFD assessment provides learners with opportunities to cooperate with their peers in the context of the multi-semiotic conceptual situation under consideration, and to present to the group alternative responses. In weekly MERLO CIFD assessments, members of a group listen to a brief presentation by one of their peers. The presentation is often accompanied by engaging arguments that call attention to different aspects of multi-semiotic, meaning-equivalent representations in the conceptual situation being considered.

These weekly assessments also provide students with opportunities to develop habits of conceptual thinking, and demonstrate the effectiveness of the pedagogy for conceptual thinking and peer cooperation in motivating and engaging students. This is particularly important in large undergraduate classes.

Individual scores from the weekly MERLO CIFD assessments help the teachers to identify students' conceptual issues and misunderstandings and remedy conceptual issues in their future teaching practices. MERLO assessment items are also involved in mid-term and final exams, additionally to other assessment types such as essays and multiple-choice quizzes.

## **Interactive Concept Discovery (InCoD)**

In addition to weekly MERLO formative quizzes, pedagogy for conceptual thinking encourages learners to conduct weekly Interactive Concept Discovery (InCoD) searches in the course Knowledge Repository (KR) that includes learners' annotating and indexing of a collection of relevant found documents related to course content (e-books; journals; e-databases; reports; etc.), commentaries on other learners' annotations, recognition of equivalent-of meaning of two or more different representations in different sign systems in MERLO items, and production of written descriptions of the reasons and rationale for making these judgments. Hence, the KR—a set of unstructured data – is another form of big data involved in MERLO-based pedagogy.

InCoD is a novel semantic search learning tool, developed, tested and implemented, based on Concept Parsing Algorithms (CPA; Shafrir & Etkind, 2005, 2018). It is an intuitive, interactive procedure that allows a learner to search large digital databases (eJournals; eBooks; databases; eArchives), and to discover the building

blocks of a concept within a particular context of the course knowledge domain, namely, co-occurring subordinate concepts and relations. InCoD constructs concept maps that clearly identify not only the conceptual content of important concepts in course material, but also its internal conceptual structure – hierarchical and lateral relations among concepts and their building blocks.

In the following paragraphs, we describe how big data are used in education and how students manage big data of the KR. Following the initial concordance/search, the learner may notice that in several locations/sentences, another concept consistently appears in close proximity to the searched concept. By clicking on the second concept, the learner activates a subsequence search of co-occurrence of both the initial concept and the second concept, aimed at discovering consistent co-occurrence with the initial concept. The learner begins the process of Interactive Concept Discovery by conducting semantic search (concordance) of Key Word In Context (KWIC). In each successive iteration, the learner can read and save found relevant documents online; annotate and evaluate the degree of relevance of a particular document to the specific conceptual content under consideration; and construct graphical representations of links between concepts.

A comprehensive record of a learner’s sequence of such inquiries permits for a detailed reconstruction of the learning episodes generated by InCoD over time. The InCoD data reveal the learner’s consistency of ‘drilling-down’ for discovering deeper building blocks of the particular concept, as well as the temporal evolution of outcomes of the learning sequence. This big data digital record is an authentic, evidence-based demonstration of acquisition of mastery of knowledge that can be used as a springboard for follow-up classroom and chat-room discussions. It provides a credible record to the individual’s learning process and learning outcomes (Shafir et al., 2006).

Interactive Concept Discovery makes available to the learner all the different locations of sentences that include the searched keyword (name of concept) in the digital documents in the course Knowledge Repository, likely written by different authors with different points-of-view and different examples that contain a particular searched concept. Following the initial concordance/search, the learner may notice that in several locations/sentences, another concept consistently appears in close proximity to the searched concept. By clicking on the second concept, the learner activates a subsequence search of co-occurrence of both the initial concept and the second concept, aimed at discovering consistent co-occurrence with the initial concept. *Learner Individual Index* records data from these learner’s activities, including:

- Alphabetic indexing by name of concept
- Document ID/page
- Ranking (on scale of 1–5) of degree of relevance to course content
- Annotations, including: brief summary of the specific conceptual context; followed by learner’s comments, tags, and links

Clicking on a particular entry in the Individual Index provides the date of creation of the entry by the learner, plus complete details of content. A Learner

Individual Index is available to the instructor, and tracks the learner's progress in mastering and conceptual understanding of the documents in the Knowledge Repository, and the specific learning outcomes accumulated throughout the course.

The *Concepedia* (Conceptual Encyclopedia) is a weekly aggregation of all Individual Indexes of all learners in the class, in the context of the course knowledge domain, and is accessible to the instructor and to all students in the course. It also includes learners' commentaries on other learners' annotations. The *Concepedia* enhances individuals' reputations as cooperators who contribute to the public good, and reflects the cumulative process-learning-curve of the class.

## Big Data Evidence-Based Learning Analytics for Merlo and INCOD

Detailed data of teaching and learning processes and outcomes are collected, analyzed, updated weekly, and are available to the learners and to the course teacher through learning analytics (Shafirir & Etkind, 2006; Shafirir & Kenett, 2016). In particular, the evidence-based MERLO Learning Analytics collects, documents, analyses, and reports time-marked data that describes continuous learning processes and outcomes at both learner and teacher levels and from the point of view of educational institutions. These learning processes follow a pedagogy for conceptual thinking and peer collaboration (Etkind et al., 2016) that includes Interactive Concept Discovery (InCoD) and Meaning Equivalence Reusable Learning Objects (MERLO) assessments.

More precisely, the learning-related data that are analysed include documented, detailed results of explorations of conceptual content through semantic searches in KR. It involves annotating and indexing of relevant documents found by the learner, commentaries on other learners' annotations, recognition of equivalence-of-meaning of two or more different representations in different sign systems in MERLO items, and written descriptions and justifications for making these judgments.

Each individual student's 'big data' profile includes:

- MERLO weekly formative CIFD assessments for specific concepts, as well as MERLO summative assessment items in mid-term tests, final exams in individual courses, mean MERLO CIFD scores in weekly assessments, as well as mid-term tests, and final exams.
- These MERLO scores identify level of understanding, as well as specific deficits in conceptual understanding of course content expressed as lower mean Q2 and Q3 class scores in conceptual understanding of specific course content; as well as teacher's recommended corrective interventions for individual learners.
- Weekly evolution of the individual indexes of InCoD semantic searches by each of the students in the course, including annotations; as well as weekly updated *Concepديات* for different courses.

These data are collected continuously during the learning-process, across many classrooms, all learning and teaching activities throughout the semester and the academic year. The data are continuously available through Concept Science Evidence-Based MERLO Learning Analytics to the learners and to the teachers. Moreover, these learning analytics are available to the academic institutions, subject to strict privacy procedures that protect individual student identity and privacy.

Figure 6.5 is an example of MERLO weekly formative assessment in course ‘ASC 406 – Ideas, Technologies and Precedents III’, third year course on modern and contemporary architecture focused on ‘the cultural, political and technological forces shaping the architecture of western civilization in the 19th and 20th centuries’. Table 6.1 shows example of double-median-split of Z-scores of the MERLO assessment item for the concept ‘modular’, shown in Fig. 6.5. It includes the following data columns: Student number; raw recognition score (range 0–5); Z-recognition score; raw production score (0–5); Z-production score; written description of the concept by each individual student. The color of the data numbers in the columns for Z-Recognition and Z-Production are determined by the median split for each of these variables:

An example of big data from an Interactive Concept Discovery (InCoD) at University of Toronto Libraries, for the course ‘ASC 406 – Ideas, Technologies and

Course: ASC406 Ideas Technology and Precedents III  
 Topic: Architecture and Modernity  
 Meaning Equivalence MERLO Quiz #7 March 16 2017




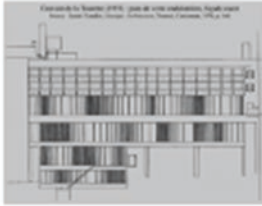

<p>At least two – possibly more - of the five images share the same meaning or concept. Please mark all those statements – but only those – That share equivalent meaning.</p> <p>Formulate the shared concept</p>	<p>1</p> 	<p>2</p> 
<p>3</p> 	<p>4</p> 	<p>5</p> 

Fig. 6.5 Weekly MERLO quiz for concept ‘modular’

**Table 6.1** Individual scores of double-median-split of MERLO weekly quiz for the (un-named) concept ‘modular’ (sample of 15 students).

Student #	Recogn	Z-Recogn	Product	Z-Product	Individual student's description of concept
80	5	1.24	5	1.61	Le Corbusier's ideas of modularity
81	3	-1.01	3	-1.05	Corbusier's universal system of proportion
82	4	0.12	4	0.28	Le Corbusier and his modular form of measure
83	5	1.24	4	0.28	modular and new architecture
84	4	0.12	4	0.28	Le Corbusier's "Le Modulor" proportional system
85	4	0.12	3	-1.05	beauty in proportions
86	5	1.24	5	1.61	Le Corbusier's idea of Modularity
87	4	0.12	4	0.28	Le Corbusier's proportionality and modularity.
88	4	0.12	4	0.28	Designing with anthropometric scale and the modular with respect to Le Corbusier.
89	4	0.12	3	-1.05	Application of the Modular
90	4	0.12	4	0.28	Le corbusiers modular theory regarding scale and proportion
91	3	-1.01	4	0.28	proportioned and modular systems
92	3	-1.01	4	0.28	The modular
93	5	1.24	5	1.61	modular and the universal application of Le Corbusier's new architecture forms
94	5	1.24	5	1.61	the modular: theory and application
95	4	0.12	4	0.28	Le Corbusier's theory of the modular and anthropometric scale of proportions.

**Table 6.2** Search results of concept/keyword ‘modern architecture’

1	Photographic modern architecture: Inside ‘the new deep’ by Zimmerman, Claire The Journal of Architecture, ISSN 1360-2365, 07/2017, Volume 22, Issue 5, pp. 968–991
2	Utopia and modern architecture? By Nathaniel Coleman Architectural Research Quarterly, ISSN 1359-1355, 12/2012, Volume 16, Issue 4, pp. 339–348
3	Modern architecture of World’s libraries by Vatin, Nikolay and Gamayunova, Olga Advanced Materials Research, ISSN 1662-8985, 12/2014, Volume 1065–1069, pp. 2622–2625
4	Ageing of modern architecture by Darya Paramonova Project Baikal, ISSN 2309-3072, 05/2011, Volume 8, Issue 28, pp. 44–45
5	Preservation, design and modern architecture: The challenges ahead by Prudon, Theodore Journal of Architectural Conservation, ISSN 1355-6207, 05/2017, Volume 23, Issue 1–2, p. 27–35
6	The Memory of Inhabiting Modern Architecture: Villa Portales, 1955–2010 by Forray, Rosanna and Márquez, Francisca Latin American Perspectives, ISSN 0094-582X, 5/2017, Volume 44, Issue 3, pp. 64–82
7	Relationship between modern architecture and interior space by Wang, Peng Advanced Materials Research, ISSN 1662-8985, 12/2014, Volume 1065–1069, pp. 1593–1596
8	Color composition features in modern architecture by Serra, Juan and García, Ángela and Torres, Ana and Llopis, Jorge Color Research & Application, ISSN 0361-2317, 04/2012, Volume 37, Issue 2, pp. 126–133
9	Glass, the fundamental material of modern architecture by Le Corbusier and Paul Stirton and Tim Benton West 86th: A Journal of Decorative Arts, Design History, and Material Culture, ISSN 2153-5531, 09/2012, Volume 19, Issue 2, pp. 282–308

*Precedents III*’, using a Key Word In Context (KWIC) search, with the following options: ‘Advanced Article Search’ for Subject Terms ‘*modern architecture*’ (scholarly journals; full text online) produces the results shown in Table 6.2 that include a list of relevant references that comprise the Knowledge Repository (KR) for the course.

**Table 6.3** Student’s individual index

Document	Discovered co-occurrence of concept/keyword	Pages
1	Abstraction	974, 984, 986
2	Orthodox	340, 345
3	Modern library	2622, 2624
4	Transition	28, 29, 34
5	Collective	66, 67, 69, 71, 72, 77, 80
6	Virtual	1593, 1594, 1596
7	Chromatic	126, 127, 128, 131
8	Glass tiles Glass wall	284, 285, 286, 289 285, 286, 287, 290, 294, 295, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306

Students’ individual searches of each of these references result in discovery of co-occurrence of concepts/keywords related to ‘*modern architecture*’, as shown in Table 6.3. It allows individual students to create and develop their own individual indexes; to write annotations for specific discovered co-occurring concepts; and to rank the degree of relevance to the course content.

## Conclusion

This chapter demonstrates advantages of learning in the context of big data generated by modern information technologies while using the MERLO pedagogical tool for learning and formative assessment. A recent OECD review provides strong evidence for the important role of formative assessment in enhancing students’ learning outcomes (OECD, 2013, p. 145). The present chapter is about enhancing learning outcomes and deep understanding of concepts through interactive learning and methods for assessing such understanding. In the digital age, technology is within reach of everyone, and the focus on depth of understanding is gaining importance and urgency. The wide range of experience in MERLO and InCoD tools demonstrate their universality and reflects their large potential in future research. Each individual student’s big data profile includes time-marked data that describes continuous learning processes and outcomes at both individual learner and individual teacher levels, useful for both individuals and educational institutions.

Implementation, testing, and validation, since 2002, of pedagogy for conceptual thinking (Shafir & Etkind, 2018; Shafir & Kenett, 2016), lend support to the following conclusions:

- Weekly MERLO formative assessments enhance peer cooperation and conceptual thinking.
- Weekly Interactive Concept Discovery (InCoD) provides individual learners with different points of view on conceptual issues relevant to the course content, written by different authors.



- Pedagogy for conceptual thinking motivates and engages students; this is particularly evident - and important - in large undergraduate classes.
- Conceptual thinking is learnable.
- Pedagogy for conceptual thinking, when implemented as a regular part of the instructional methodology, replicates the above pattern of results and the following enhanced learning outcomes: applying knowledge and skills in central subject areas and analysing, reasoning, and communicating effectively as they examine, interpret, and solve problems

## References

- Abhyankar, K., & Ganapathy, S. (2014). Technology-enhanced learning analytics system design for engineering education. *International Journal of Information and Education Technology*, 4(4), 345–350. <https://doi.org/10.7763/IJJET.2014.V4.427>
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (2004). *How people learn: Brain, mind, experience, and school (expanded edition)*. National Academy Press.
- Etkind, M., Kenett, R. S., & Shafir, U. (2010). *The evidence-based management of learning: Diagnosis and development of conceptual thinking with meaning equivalence reusable learning objects (MERLO)*. In Proceedings of the 8th International Conference on Teaching Statistics (ICOTS8). ICOTS.
- Etkind, M., Kenett, R. S., & Shafir, U. (2016). Learning in the digital age with meaning equivalence reusable learning objects (MERLO). In E. Railean, G. Walker, A. Elçi, & L. Jackson (Eds.), *Handbook of research on applied learning: Theory and design in modern education* (Vol. 1, pp. 310–333). IGI Global.
- Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *The American Behavioral Scientist*, 57(10), 1439–1459. <https://doi.org/10.1177/0002764213479367>
- Kittredge, R. I. (1983). Sematic processing of texts in restricted sublanguages. In N. J. Cercone (Ed.), *Computational Linguistics* (pp. 45–58). Pergamon Press.
- OECD. (2013). *Student assessment: Putting the learner at the centre*. In Synergies for better learning: An International perspective on evaluation and assessment. OECD Publishing. <https://doi.org/10.1787/9789264190658-7-en>
- Patricio, A. (2018). *IDC: Expect 175 zettabytes of data worldwide by 2025*. Retrieved October 29, 2019, from <https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html>
- Ripley, C., Etkind, M., & Shafir, U. (2004). *MERLO design studio workshop designer's handbook: Teaching for deep comprehension with meaning equivalence reusable learning objects*. In G. Rice & D. Baker (Eds.), Proceedings of the International Society for Exploring Teaching and Learning (ISETL) conference. ISETL.
- Shafir, U., & Etkind, M. (2005). *Concept parsing algorithms: Mapping the conceptual content of disciplines* (version 11.0, January, 2005). PARCEP.
- Shafir, U., & Etkind, M. (2006). eLearning for depth in the semantic web. *British Journal for Educational Technology*, 37(3), 425–444. <https://doi.org/10.1111/j.1467-8535.2006.00614.x>
- Shafir, U., & Etkind, M. (2018). *Concept parsing algorithms (cpa) for textual analysis and discovery: Emerging research and opportunities*. IGI Global.
- Shafir, U., Etkind, M., & Treviranus, J. (2006). eLearning tools for ePortfolios. In *Handbook of Research on ePortfolios* (Chapter xx, pp. 206–216).

Shafir, U., & Kenett, R. S. (2016). Concept science evidence-based MERLO learning analytics. In E. Railean, G. Walker, A. Elçi, & L. Jackson (Eds.), *Handbook of research on applied learning: Theory and design in modern education* (Vol. 1, pp. 334–357). IGI Global.

**Masha Etkind** is a professor in the Department of Architectural Science at Ryerson University. She teaches design, theory and history of architecture. Her work and recent research focus on heritage conservation, language and roots of living architecture, as well as on Pedagogy for Conceptual Thinking in architectural education. Etkind received a professional degree in architecture from St. Petersburg University of Architecture and Engineering, and Masters in Architecture from the University of Toronto. Etkind is member of the Royal Architectural Institute of Canada.

“**Dr Theodosia Prodmou** is a senior lecturer of Mathematics Education at the University of New England in Australia. She has published numerous journal articles, book chapters and edited/ authored books on a wide range of subjects, including the use of big data in educational settings, the relationship between technology and mathematical thinking, the integration of digital technologies in the teaching and learning of mathematics, STEM education, augmented reality in Educational settings, professional development of mathematics teachers, and statistical thinking including perceptions of probability and chance. Her most recent edited/authored books were, *Augmented reality in educational settings* (2020) and *Primary and Middle Years Mathematics: Teaching Developmentally* (2019).”

**Uri Shafir** is an associate professor in Department of Human Development and Applied Psychology at Ontario Institute for Studies in Education of University of Toronto. His recent research focuses on concept parsing algorithms in various knowledge domains and on instructional methodologies for deep comprehension of conceptual content that enhance learning outcomes. Shafir received doctorate in mathematical sciences from the University of California at Los Angeles and doctorate in developmental psychology from York University, Toronto. Shafir was founder and director of the Institute of Planetary and Space Science at Tel-Aviv University; before moving to the University of Toronto he was Adjunct Professor at the University of Wisconsin and at Columbia University.

# Chapter 7

## Employing Authentic Analytics for More Authentic Tasks



William Billingsley and Peter Fletcher

**Abstract** Although teaching systems and other infrastructure collect large amounts of information that can give insights into the learning behaviour of students, much of this data tends to be concerned with secondary aspects – for example, when students accessed resources or how they performed in inauthentic tasks such as adaptive quizzes. Whereas, when a design task, a higher-order thinking task, or a long-form writing task is given, students typically do their thinking and research outside of the learning environment and only the submitted product is available for analysis. In this chapter, we examine the opportunities for student-facing learning analytics in authentic tasks using authentic tools. By employing professional tools, we can design environments that allow students to work on realistic open-ended problems while gathering data on the strategies and practices they use in the creation process. In some fields, such as software engineering, professional and open source projects gather this sort of data, and those same tools allow collection of student data, allowing us to explore whether students are adopting strategies that experts find to be successful. We see this as the goal of developing cognitive apprenticeships, supported by smart technology, that use more authentic environments. We suggest this trend is the coming together of three strands of development in education: rich learning environments, learning analytics, and authentic tasks.

**Keywords** Learning analytics · Authentic tasks · Cognitive apprenticeships · Learning environments · Authentic tasks

---

W. Billingsley (✉) · P. Fletcher  
University of New England, Armidale, NSW, Australia  
e-mail: [wbillin@une.edu.au](mailto:wbillin@une.edu.au); [pfetch2@une.edu.au](mailto:pfetch2@une.edu.au)

© Springer Nature Switzerland AG 2021  
T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy  
Implications of Research in Education 13,  
[https://doi.org/10.1007/978-3-030-76841-6\\_7](https://doi.org/10.1007/978-3-030-76841-6_7)

## Introduction

Although teaching systems and other infrastructure collect large amounts of information that can give insights into the learning behavior of students, currently most of this data tends to be concerned with secondary aspects – for example, when students accessed resources, or how they performed in simplified assessment tasks such as adaptive quizzes. Whereas, when a design task, a higher-order thinking task, or a long-form writing task is given, students typically do their thinking and research outside of the learning environment and only the submitted product is available for analysis.

For an educator, this can create a tension between the authenticity of the task and the richness of the data that is collected. The task could be altered to facilitate collecting richer data about the students' working practices, for instance by requiring work to be done in a specialised and instrumented interface designed for education, or by requiring the student to submit intermediate stages of work as well as the finished product. However, this would place artificial constraints on the students' manner of working that do not match their intended future professional practice: real work is not performed through educational interfaces. Specialised educational interfaces also face the challenge that they can be expensive to develop and maintain, especially when they need to keep pace with the kinds of functionality that is provided in professional tools.

At this point, we should pause to consider what we mean by a “more authentic task”. Authentic tasks, activities, and assessment have commonly been discussed in the education and educational technology literature since around the 1990s (Herrington et al., 2006; Montgomery, 2002; Wiggins, 1989, 1990). Typically, these articles observe that educational tasks are often shaped around what is easy to assess and omit the challenges of ambiguity and multi-facetedness that graduates will have to encounter in professional work. The authors propose adopting aspects of professional work into assessment tasks, to promote higher order thinking in challenging environments on contextualised problems.

However, there is rightly some criticism of the use of the phrase “authentic tasks”. The tasks set in education are not genuinely authentic professional tasks – they are undertaken in an educational setting and set for a particular educational purpose. Petraglia (1998) describes this as *preauthentication* – “the attempt to make learning materials and environments correspond to the real world prior to the learner's interaction with them”. Our own perspective is that this is not necessarily problematic. The educator is attempting to design a learning experience, and borrowing aspects of professional tasks (be it problem, context, environment, or tools) can be a pragmatic approach for adding depth and higher-order thinking to a problem, or simply for lessening the apparent gap between educational work and professional work. Accordingly, we use the phrase “more authentic tasks” in the title in two ways – to refer to tasks where educators explicitly borrow elements of professional work, and to argue how the use of professional analytics tools can bring tasks closer to professional practice.

Professional practice in more and more fields has become computer-mediated and many professional systems capture and record work-in-progress and include automated analysis. So, the opportunity is arising for educators to leverage professional tools as data sources. This enables tasks to be designed where students work not just on authentic problems but also in an authentic environment. This provides richer data and analytics that regularly inform real life work practice.

For example, for the last 6 years, we have taught a collaborative software engineering course that uses some authentic collaboration tools from the software industry: distributed version control, ticket management, and continuous integration (Billingsley & Steel, 2013; Billingsley et al., 2019; Süß & Billingsley, 2012). The version control system keeps a recorded history of every edit every student has committed to the code, as opposed to just the final version of the product that is available at the due date. The automated build system compiles each version of the code and runs tests, many of which are written by the students themselves, that indicate which changes worked and which did not. “Static analysis” tools can gauge the overall code quality and identify areas of the codebase that contain code that, in best practice, should be revisited. Professional teams use these systems to be able to track the development of their code, work in parallel on different parts of the code, and revert changes where necessary. For our course, the systems also give us the information we need to identify what and when each student has contributed to the codebase.

The task is tailored to students’ learning and the data collection is authentic, using the tools of industry. Conveniently, there is also a wealth of data available from professional and open source projects, providing the opportunity to compare students’ coding behaviour with expert and professional ways of working.

Even in less technical fields than software development, there is a proliferation of computer-supported collaborative work systems being adopted into professional practice. For example, it is common for media writers to draft their work in a content management system with a publication workflow, and collaborative report writing is often conducted over file synchronization services such as Dropbox or collaborative editors such as Google Docs. As soon as an educational task becomes computer-supported cooperative work, the communication need of the underlying system to send edits to other participants becomes an opportunity to gather revisions of students’ work, and collect more authentic data concerning students’ production practices.

In this chapter, we look at the trajectory of how three strands of educational innovation are converging to provide a platform to support rich authentic learning. We first consider the rich models and analysis of students’ learning that have taken place in tailored educational environments, such as intelligent tutoring systems. As these have evolved, they have developed techniques for analysis that can be applied to data coming from more authentic environments. We then discuss how learning analytics grew from mining the log data of learning environments to consider more diverse sources of data so that deeper questions about students’ learning experiences could be examined. The third strand we then describe is how progress in teaching pedagogies has called for greater incorporation of professional skills and authentic learning experiences in courses. We suggest the progress towards rich

analytics of authentic practice can be considered as the evolving-together of these three strands.

## Rich Modelling in Tailored Environments

If our goal is to assess and model authentic student tasks, it may be convenient to review a selected history of rich interactive modelling in tailored environments. From the early days of artificial intelligence (AI), computer scientists have endeavoured to design systems that can be students' partners in work and model their progress through problem-based learning tasks. Initially this involved hand-crafting individual AI models and specialised environments for each kind of exercise.

### *Reactive Learning Environments and Cognitive Apprenticeship*

In the 1970s, John Seely Brown proposed *Reactive Learning Environments* – intelligent learning environments that would encourage students to “explore ideas, create conjectures or hypotheses about a situation, and then to receive detailed feedback as to the logical validity of [their] ideas” (Brown et al., 1976). Brown was interested in the thought processes that students go through when engaged in lab-work for electronics – particularly, he was interested in *troubleshooting*. When a student is faced with a circuit that has an unknown fault, they have to apply their knowledge and understanding of electronics to form ideas about what could be wrong, and generate strategies for what measurements to take in order to test their hypotheses and refine their ideas.

Brown created three versions of an AI environment, SOPHIE (Brown et al., 1975), that could simulate this experience through dialogue with a student. The student (and SOPHIE) were given a model of an electronic circuit, where one of the electronic components had a simulated fault. The student could ask SOPHIE via a textual interface, to make measurements on the circuit and could propose hypotheses for what the fault was. SOPHIE modelled not only the circuit, but also the students' measurements, and was able to respond on whether the student's hypothesis was logical based on what they had measured, as well as whether it was correct. An excerpt from a session is reproduced in Fig. 7.1.

SOPHIE is an early example of a *mixed initiative* system, in that it supports the principle that as well as the system being able to ask the student questions, the student should be able to ask the system questions. This requires a very detailed model of the problem – rather than trace the student through known steps, the system has to be able to react to the student's chosen strategy. It also, despite being a task carried out in text dialogue, subscribes closely to the idea that learning is situated in its use – learning through guided experiences to develop cognitive and metacognitive (rather than physical) skills. Brown, with colleagues, would later expand on this

**Student** : REPLACE Q6  
**Sophie** : I am going to ask you some questions about how Q6 is faulted. Are there any junctions shorted?  
**Student** : I DON'T KNOW  
**Sophie** : Are any terminals open?  
**Student** : YES  
**Sophie** : Which terminals are open?  
**Student** : THE BASE  
**Sophie** : Q6 is not faulted in that manner. Q6 not replaced.  
**Student** : IS IT POSSIBLE THAT Q6 HAS AN OPEN BASE  
**Sophie** : Is the complete fault mode EB/EC/BC = OP/OK/OP what you intended?  
**Student** : SURE  
**Sophie** : When the settings were: CC=1.0, CR=HIGH, LOAD=20, VC=1.0, VR=HIGH, that fault conflicts with the following measurement. You observed that the output voltage was 11.7 volts. If Q6 had fault mode EB/EC/BC = OP/OK/OP it would be 23.9 volts. In a working circuit it is 19.9 volts.

**Fig. 7.1** An extract from a SOPHIE session, as reported in Brown et al. (1974)

theme to propose *cognitive apprenticeships* (Brown et al., 1989; Collins et al., 1989) as a model of education. Lajoie and Lesgold adopted a similar approach to how they viewed SHERLOCK, which they developed as a coached practice environment in electronics diagnosis for Air Force technicians, basing it on a detailed cognitive task analysis of the problem (Lajoie & Lesgold, 1992).

We have chosen to introduce one of the earliest systems to highlight that this goal of intelligent support for situated learning was established very early on in the history of intelligent learning environments. Also to situate and acknowledge the nature of these systems owing to their emphasis on pseudo-authentic tasks and student initiative, even if not conducted in authentic environments.

## ***Model Tracing***

From the 1980s, research on Intelligent Tutoring Systems (ITSs) began to emphasise building detailed models of students' knowledge as well as the problem they are working on. As the student works through problems and examples, the tutor attempts to estimate their understanding of the domain – a process called *model tracing*. This student model is constantly updated as the tutor observes the student working through problems, typically as a Bayesian network. A pedagogical model built into the program refers to the student model when deciding what advice to give and what exercise to select for the student next.

Intelligent Tutoring Systems research has frequently been inspired by Bloom's observation that one-to-one tutoring is significantly more effective than classroom instruction (Bloom, 1984). Many papers in ITS research have measured the learning

gains of particular systems, and VanLehn (2011) found in his review that ITSs can produce similar learning gains to human tutors – also finding that human tutors often have limited skills in assessing students’ misconceptions and rarely employ sophisticated pedagogical strategies.

The model-tracing approach was popularised by the Cognitive Tutors from Carnegie Mellon University (Anderson et al., 1995; Corbett, 2001). These tutors were based on Anderson’s ACT-R model of cognition, which separated *declarative* knowledge from *procedural* knowledge (Anderson, 1993). A Cognitive Tutor would hold a set of production rules describing how an ideal student would progress through the problem, as well as a set of “buggy rules” representing common misconceptions. The tutor would hold a model of the probability that the student understood each rule – typically in a Bayesian network – and this would be updated by observing the student’s actions and progress as they worked through each problem. Model tracing tutors were developed for algebra (Ritter & Koedinger, 1997), geometry (Anderson et al., 1985), physics (VanLehn et al., 2005), and programming (Corbett & Trask, 2000) amongst other topics. ActiveMath (Melis et al., 2005) combined model-tracing tutoring with a detailed ontology of mathematical concepts, to create a more comprehensive learning environment.

A key issue with this approach, however, is that model tracing requires a complete, or at least comprehensive, set of production rules for how a student should consider which step to take next. In the traditional rendition of a model-tracing tutor, the production rules needed to be hand-crafted, making them labour and time intensive and only amenable to “tame” problems where the appropriate production rules are known. More recently, ITS research began to examine how these models can be constructed from data. Initially, example-tracing tutors (Aleven et al., 2009, 2016) were introduced, which allowed a non-programmer to develop the model for an ITS by demonstrating appropriate solutions to it.

Conversely, if solutions can be demonstrated by the tutor, then they can also be generated from examples of student work. Research has looked into how the model can be generated from large quantities of data from students working on exercises – using machine learning approaches. For example, Barnes and Stamper (2010) used a Markov Decision Process to establish solution paths in a logic tutor, and Deep Knowledge Tracing (Piech et al., 2015) applies each student’s actions as input to a recurrent neural network, allowing the network to learn an appropriate student model and success factors from exposure to large quantities of data. This student-data approach has recently been a popular topic of analysis in the artificial intelligence community, as researchers examine the depth and predictive power of the student models that can be learnt (Wilson et al., 2016; Xiong et al., 2016). This data-driven approach also affords new opportunities – whereas manually derived models could only feasibly be applied to relatively small problems, deep knowledge tracing can learn success factors for more open-ended tasks such as computer programming (Wang et al., 2017).

Potentially, this ability to apply model tracing to more open tasks affords a link back to the more lab-like environments that Brown originally considered. In many courses, situated learning is expressed in the use of virtual (simulated or remote)



laboratories. These are an environment where an authentic task is made digital so that it can be improved for educational and logistical purposes. Experiments can be run more quickly (Zacharia et al., 2008), measurement error can be removed (Pyatt & Sims, 2012), and phenomena can be made more easily observable (Chiu et al., 2015; Deslauriers & Wieman, 2011). Simulated (as opposed to remote) labs also allow students to be less constrained by the practical limitations of reality – for example, being able to pause the simulation while considering a change to make – and consequently Ma and Nickerson (2006) found simulations to be more focused on design than comparable remote laboratories.

In these more open and exploratory environments, the question arises of how to identify whether a particular student’s activities are being productive. This is an area of recent research in the artificial intelligence education community, typically involving searching within logs of student activity within the laboratory for common sequence patterns (Perez et al., 2017; Venant et al., 2017) – a process with some conceptual similarity to finding solution paths. Intelligent tutoring systems have also begun to explore questions beyond just solving the problem. Janning et al. (2016) propose ways to infer the perceived difficulty of tasks from the logs of student interactions within the ITS. McCarthy et al. (2018) consider whether metacognitive prompts help or hinder student performance. San Pedro et al. (2014) consider the relationship between student affect and carelessness in the exercises.

As these ITS environments evolve to consider more open-ended questions and aspects of the learning beyond just the solution to the problem, they are accruing capabilities that will be useful for more authentic tasks.

### ***Intelligent Books: A Project Reflection***

As this chapter partly forms a design argument for a future model of education, it may support the reader to describe two of our own project experiences developing intelligent learning environments. In the early 2000s we developed “Intelligent Books” which comprised a suite of technologies supporting adaptive materials, that provided students an educational experience where the AI acted as their partner rather than their assessor (Billingsley, 2020; Billingsley & Robinson, 2009). One was a project for constructing and analysing electronic circuits (Billingsley & Billingsley, 2004) and another was a project for constructing proofs in discrete mathematics (Billingsley & Robinson, 2007b).

These projects were directly inspired by Brown’s work on SOPHIE – students would work hand-in-hand with AI models, interacting with the problem via diagrams in the browser, and with the AI able to animate and explain the consequences of students’ choices. As the intelligent books were mediated over the Internet, it was possible for a student to interact with the AI via diagrams, and for the book to adapt based on students’ interactions (Billingsley & Robinson, 2009).

As we were working on authentic and open-ended problems – for example, constructing a mathematical proof is an equivalent task to writing a computer

program— rather than try to model the steps of the problem, we drew on the concept of Constraint-Based Tutors (Mitrovic, 2012; Mitrovic et al., 2001). These allow the student to progress through the task in any manner they like, so long as at any given step the student has not broken one of the rules of the domain. This is very similar to how a compiler checks the code of a computer program that is being written – it has no concept of whether the programmer is trying the right strategy, but only of whether the program in its current state has errors. In our exercises for number theory proofs, the AI we used to model the domain was an authentic proof assistant used by researchers – Isabelle/HOL (Nipkow et al., 2002). Our learning environment performed automatic translation between a student representation of the problem, which used our own blocks-based proof programming environments (Billingsley & Robinson, 2005) – and a formal model that was checked by Isabelle.

However, these were tailored environments created for education, rather than authentic environments. We also found that there are cost-benefit trade-offs, in that rich models for particular kinds of exercise (for example, designing a circuit) support only a narrow range of exercises, making supporting an entire course arduous. We also found that many on-paper tutorial problems in discrete mathematics are asked using English prose, rather than mathematical notation, and the challenge of translating that prose into a more formal mathematical argument is one of the skills that is being assessed. This made these questions problematic to set in the tile-based proof environment, as the prose needed to be translated into a formal mathematical proposition before the question could be asked. In response, we experimented with approaches that used less formal modelling to support a wider array of questions (Billingsley & Robinson, 2007a).

## The Rise of Learning Analytics

From an engineer's perspective, software systems lend themselves to analytics. In the majority of new systems, even before data scientists have had time to consider what data they might wish to collect for analysis, the programmers and product managers have already inserted logging and analytics to understand usage patterns and to assist with debugging. So pervasive can the culture of data collection be that sometimes even seemingly trivial questions are resolved through analytics – in a famously extreme example, when engineers at Google were unable to choose between two shades of the colour blue for search links, they resolved the argument by testing 41 different shades on sections of the public to see which performed better (Hobson, 2009).

From a technical perspective, it seems inevitable that educational technology would deliver a wealth of learning analytics data before educationalists could decide what questions to ask. A few years ago, colleagues at an Australian university beginning to offer MOOCs through one of the major providers received weekly dumps of the clickstream data from their courses. The first question posed was how to explore this data, even before considering what questions to ask of it.

The field of learning analytics emerged out of the earlier field of educational data mining (Romero & Ventura, 2007) – Ferguson (2012) places the emergence of learning analytics as a distinct field in 2010. Educational Data Mining had significant overlaps with the ITS community. Baker and Yacef (2009) surveyed the state of educational data mining as the editorial introduction to the first issue of the *Journal of Educational Data Mining*, and are authors of the third (Merceron & Yacef, 2003) and fifth (Baker et al., 2004) most-cited papers in the earlier review by Romero and Ventura (2007) – both of which were on educational data mining in intelligent tutoring systems.

As learning analytics has emerged as a distinct field, it appears to have gravitated towards larger sources of data. Colvin et al. (2015) and Mah (2016) considered the large pools of data that reside within the learning management system (LMS), lecture capture, and other university-wide systems. Pardos et al. (2017) demonstrated using clickstream navigational data in a Massively Open On-line Course (MOOC) to learn student behaviours and recommend the next steps that students should take. Gibson et al. (2016) set up a task where students submitted reflections on their learning throughout the term, and performed content analysis on them. This has the benefit of directly encouraging students to reflect on their learning while also collecting data on their metacognition.

Social network analysis has frequently been used to gain insights into students based on their discussions and interactions (Dawson et al., 2011; De Liddo et al., 2011). Extending this concept, Pardo (2013) created an augmented social graph that also includes interactions between students and course resources. However, these larger data streams tend not to be directly about the work students are performing but about surrounding aspects, such as their discussions or accesses of useful resources. Where researchers have explored the use of learning analytics from outside the LMS, it has tended to be with a view to bringing events back into central systems. Pardo and Kloos (2011) propose a central server that can listen for events from external tools. Bakharia et al. (2016) describe their experiences using The Experience API (xAPI) to bring data from social media into the Connected Learning Analytics Toolkit (Kitto et al., 2016). While these enable outside experiences to be studied, there still appears to be a focus on the conversation rather than the work, and comparison between less successful and more successful students. This is in contrast to the comparison between students and practitioners as was the case in the cognitive apprenticeship models. Khousa et al. (2015) proposed cognitive apprenticeship based on social learning analytics, but this is through engendering communities of practice where a human mentor can provide apprenticeship rather than through building apprenticeship into the analytics themselves.

## **Authentic Learning: An Opportunity for Authentic Analytics**

In recent years, there have been increasing calls for more authentic tasks and assessment in university education. These calls have been coming from many directions.

In computer science and information technology, employers consistently call out universities to improve employability – for example in the most recent Australian Employer Satisfaction Survey (QILT, 2017), employability skills was the lowest ranked category, with only 85% satisfaction, compared to 95% for students' technical skills. This gap between the relatively structured environment of studying computing and the essential complexity and difficulty of real-world engineering has long been recognised. In his famous essay *No Silver Bullet*, Brooks (1987) examines the essential complexities of real-world technology development, arguing that software entities are more complex for their size than any other human construct, and yet they are intangible, vaguely defined, constantly changing, and constantly required to conform to other arbitrary systems they must integrate with. This was 30 years ago, and since then software has only become more complex.

Brooks suggested part of the key may lie in training great designers, and from the 1990s, computing courses have sought to adopt techniques from design disciplines. Studio pedagogies have long held pride of place in design disciplines such as architecture and planning (Dutton, 1987; Long, 2012). In these, students work on authentic problems in a design community, observing and critiquing each other's work in progress under the guidance of skilled designers. Schön (1984, 1987), in analysing what makes the studio approach effective, identified it as a way of developing reflective practitioners who could enter the "swampy lowlands" of messy problems that are not theoretically solvable.

Returning to the history of tutoring systems, educationalists have found that constructivist learning and inquiry-based learning (Alfieri et al., 2011; Furtak et al., 2012) inherently have learning advantages over classroom or instructivist teaching models. Brown, the proposer of Reactive Learning Environments, has called for building reflective learning communities around practice – creating shared, distributed, reflective practicums (Seely Brown & Adler, 2008).

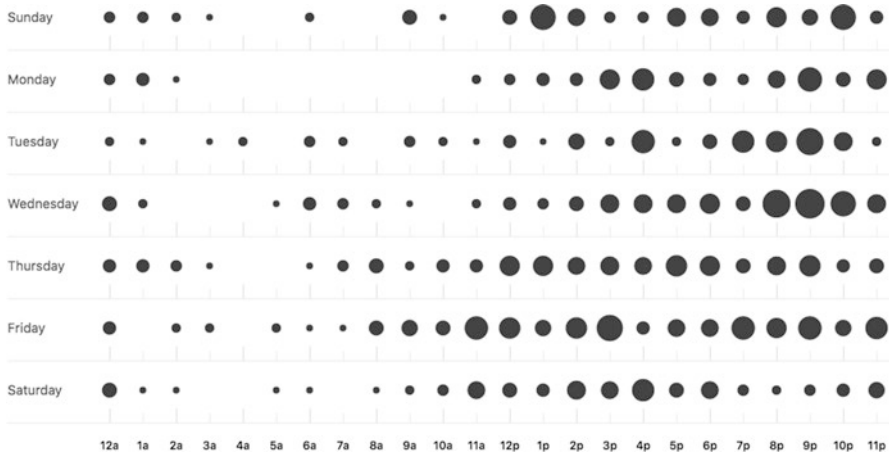
In the 1990s, Tomayko (1991, 1996) proposed the first studio software development course, and since then computer science courses around the world have adopted variations on this approach (Docherty et al., 2001; Hendrix et al., 2010; Hundhausen et al., 2008; Nurkkala & Brandle, 2011; Root et al., 2008). Recently, this has been the space we have been working in – designing a collaborative software engineering studio in which teams of students work on different features of a common class-wide project (Billingsley & Steel, 2013, 2014; Süß & Billingsley, 2012).

The studio pedagogies we use in software engineering do not just use authentic tasks, but also use authentic tools for the work students undertake. These systems themselves inherently produce large quantities of data that is directly about the production of the work. To collaborate with each other, the students use a professional distributed version control system (called git), which means that we have not only their finished work, but also every change they committed to their code on the way. Owing to this being a professional system, there are thousands of repositories available on professional sites such as GitHub containing this data for expert professional practice. In addition to considering analytics from a student-to-student perspective concerning what enables a student to pass the course, this professional

data allows us to make comparisons between student and expert practice. Potentially, we can make use of analytics to inform us whether we are achieving the kind of cognitive apprenticeship that we might seek in setting these authentic challenges.

One analysis we undertook on the course was to examine when the students were making their changes to code during the week. We were interested in whether the students faced some of the challenges that global software engineering teams face – those of dealing with collaborators who are distant from them both physically and in terms of having differing schedules (Billingsley et al., 2019). One data visualisation that the system natively produced is the punchcard of when commits are taking place, for example the one shown in Fig. 7.2. It was immediately visible that the students’ commits were dispersed across the entire week, whereas those for professional teams tend to bunch around the working hours of the working days of the week. This occurred in both an on-campus version of the course at The University of Queensland, where students would all be in the same timezone, as well as a predominantly distance education offering at the University of New England (where some students would not be). This suggested that even in the on-campus offering, students faced some challenges from “temporal distance” due to their differing study schedules, although their exact nature was different to the situation of professional teams.

The version control history also provided us with a sense of students’ branching behaviour – how often they are performing individual changes that make their copy of the code diverge from that of their peers, and how often they are integrating their changes and merging them with those of their peers. Figure 7.3 shows a fragment of the standard visualisation of this behaviour – in an actual project, there may be thousands of nodes in the graph, so it would be too large to display.



**Fig. 7.2** Commit punchcard from distributed version control (git). Despite students effectively being part-time developers, their commits are distributed across more than a full working week (Billingsley et al., 2019)



**Fig. 7.3** The version history graph shows when different branches of development diverged and merged. Typically, in a software development toolchain, many of the merges into the master branch will have been automatically tested and analysed, providing snapshots of the quality of the code at points in time

Mining the version history of students' work is relatively common. Piech et al. (2012) instrumented a development environment so that it automatically captured snapshots of students' work in an introductory computer science class. This allowed program similarity measures to generate a model of the different paths students use to solve the problem. Among professional projects, Rodríguez-Bustos and Aponte (2012) examined how the activities of contributors to Mozilla's projects changed when they changed their version control system. However, it seems to be rare to link the analytics of expert behaviour in professional and open source repositories, with analytics of student behaviour in projects for learning, in order to examine whether the task is causing students to take up industry-relevant approaches.

Authentic development toolchains can also provide some of the benefits of automated tutoring systems. Earlier in the chapter, we likened the Intelligent Book sending proof lines to Isabelle/HOL for checking to the operation of a compiler looking for errors in a program. In the software development course, we have not only a compiler but also a "continuous integration" server that is constantly testing the students' code, and static analysis tools that can search for common mistakes and poor practice.

As well as being useful to inform and guide students, having access to this data also allows us to think about the assessment task in different ways. In the afore-described software development course, our assessment task requires students to work consistently throughout the term rather than just to submit work by a due date. The rationale for this is that if students were only contributing work at the last minute, other students would not have the opportunity to work with their code, so a regular cadence of contribution is part of being a good colleague.

The availability of the data also influences how we can consider equity in group-work. With traditional group projects, it is sometimes necessary to ask group members to evaluate the contribution of their peers, using systems such as WebPA (Lodding et al., 2009). In practice, we have observed that for social reasons students can be reluctant to give their colleagues low scores for contribution – these are the classmates they will be studying with in future courses as well – and consequently groups often report that their contributions were equal when they were not. Because the studio course generates concrete data on which students have been contributing to the codebase, we can reverse this situation, and instead only require

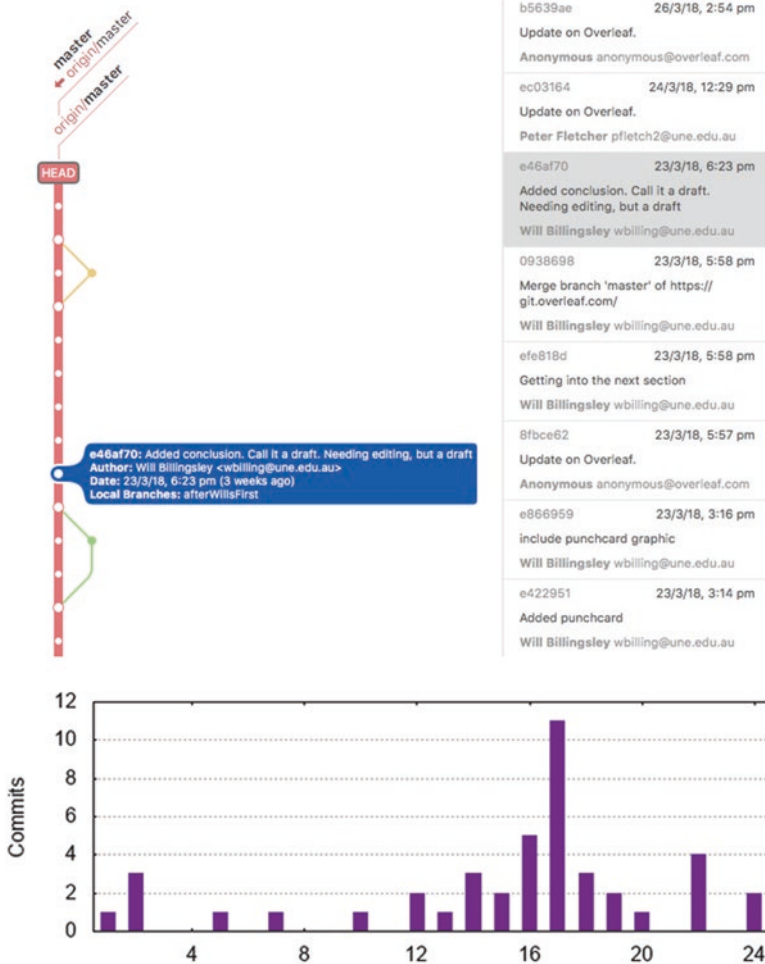
groups to tell us of occasions when a group member's contribution has been greater than it would appear from the version control log – for example, if a student has taken on a visual design or communication role in the group that is less focused on writing code. As in this situation, we are asking students to reward their peers, rather than to report under contribution, we find students appear to be more willing to give us this information.

## Beyond Computing

Computing and engineering are often early targets for education technology innovations – if only because the participants in those fields innately have the technical capability to create them. However, the working practices of almost every other field are also becoming computer-mediated, and this affords opportunities to access analytics of work in progress.

As a trivial example, this chapter was written using Overleaf (Perkel, 2014), which is an online collaborative environment for writing scientific papers that inter-operates with the git version control system. Consequently, the full version history of the chapter has been kept and can reveal details of how we wrote it. Figure 7.4 shows visualisations of this chapter's history from two common applications for git: GitUp and GitStats. From these visualisations, we can see that the commit graph is linear – there are very few branches in the graph that would indicate the authors working on different parts of the paper in parallel. The change log, shown in the right of the figure, records comments on what changes were made, as well as who made them and when. In this log, some of these comments just indicate that the chapter was edited in Overleaf's web interface, while others contain text comments from editing using desktop tools. The chart of commits by time-of-day shows that most of the writing occurred between the hours of noon and 8 pm, with some unfortunate periods of overnight work. If we had included the chart of commits by week, it would reveal that editing occurred in three concentrated periods, and that there was a 2 week gap between the first draft being completed and when it was revised in earnest. A chart of commits by day of the week, meanwhile, would reveal one author's Monday-to-Wednesday teaching schedule, as few edits occur on those days.

While those are simple characteristics of timing, we could also identify from the version history whether particular sections of the chapter were edited frequently, which might indicate which parts we found hardest to phrase. We might also expect to see that the nature of the edits changes as the chapter develops – from introducing and moving whole paragraphs and sections in the early revisions, to more fine-grained modifications to the phrasing in existing paragraphs. By passing each revision through simple natural language processing tools, we would be able to explore how language use evolves with the document – for example whether long sentences are identified and broken up, and whether overused words or phrases are found and eliminated.



**Fig. 7.4** Part of the version history of the writing of this chapter, visualised using GitUp and GitStats. In the commit graph in the top left, we can see that with two authors, the development has been mostly linear, with only short sections of writing in parallel. In the top right, we can see the times, authors, and comments of each change when it was committed, including those that are Anonymous from using Overleaf without logging in. From the data in the version log, GitStats can show us many different statistics – in this case, at what time of day most changes were made

If this data is already available for the writing of academics, there is little reason why it should not be available for students’ work. Online collaborative editors, such as Google Docs, keep a revision history as part of how they communicate changes between each user’s browser. WordPress, an open source content management system that may be useful for journalists, similarly keeps a version history of all its posts, which is made available via a web API. Even the simple act of saving a



document into DropBox or another cloud file storage service imbues that item with a revision history via the cloud service's API.

Having access to students' editing history makes it possible to give live feedback on the students' production process as well as their content. For example, identifying thrashing behaviours where a student edits the same paragraph over and over again. Plagiarism could also be considered as an invalid production method – Schneider et al. (2017) proposed such a system for software development, using logs from an instrumented development environment to detect when students were copying and then modifying code rather than writing it from scratch.

As well as versioning, most modern tools support professional workflows, either natively or through plugins. Using workflow tools has the twin advantages that they provide scaffolding, showing and leading students through professional ways of working, and they also provide data on the students' engagement with those processes. For example, WordPress supports many different workflow plugins for teams' content production and publishing – including stages of review and tagging. This makes visible to students what the expected production process is, and also collects data on when reviews and tagging were performed. Similarly, Invision, a tool for visual prototyping of user interfaces, builds in support for users to interact with and comment directly on designs over the web – in an interaction design course this may be useful as a way of ascertaining whether a student has tested their design with users, and how they revised their design in response. However, Invision are still creating their API to access this data programmatically.

## Conclusion

We foresee a future where intelligent tutoring, rich analytics, and authentic practice are regularly combined in education. Individually these strands, as they have developed, have begun to examine students' metacognition and strategies when working on open-ended problems.

When we ask students to carry out work using professional tools, we gain access to rich sources of data about how they produce their work. As modern professional systems predominantly tend to be designed with good APIs and plugin ecosystems, we also gain access to workflow integrations and plugins that can guide students into more professional practices, and to analysis tools that can help assess the quality of students' work. By analysing successive versions of the student's work, we can discover which aspects they are working on when – for example, which errors or issues in their work do they decide to address immediately, and which issues do they only revisit when the work is nearer completion. This provides us with a lens by which to examine the strategies that students employ, and the potential to compare them with the strategies that experts employ.

The key challenge for the educator is in designing the task, choosing the environment and the components to connect. There are a number of aspects to consider. The first aspect we would recommend considering is which parts of the task

to exclude. A professional work environment is complex and does not necessarily fit into a term-sized box, so for students' tasks, it requires shaping. In our opinion, it is important to decide what is out of scope at each stage. For example, in our second-year undergraduate collaborative software development course, we do not include a real customer – teaching students how to work with each other is enough for now, and working with a real customer is held back for a subsequent course.

Once the task has been scoped, the second question we find helpful to consider is how it unfolds. As the work itself will take place in a connected environment of professional systems, students have to come to grips with the ecosystem. Typically, these systems are designed for professionals, rather than novices, and often can be large and complex. So, it becomes necessary to introduce them gradually and in well-scaffolded ways. This can be likened to how in computer games, new game mechanics are unlocked as the player progresses through the story.

When it then comes to choosing which systems to deploy in the course, we find there are two facets to consider. The first is how representative the system is of professional practice: is it a product that many real teams would employ, or at least does it work along similar lines? The second facet is how easily it can be integrated into the ecosystem of the course? How easy is it to connect to existing analytics? This could be through providing a version history in a format that is well understood and can be processed by open source analysis tools, or by providing an API that other systems can hook into, or by producing “hidden analytics” – analytics included in the product that gain a new and stronger purpose in an educational context. We also find it is important to consider how identity management will work between the Learning Management System (LMS) and the professional tools, especially for cloud-based systems. While it is possible just to ask students to enter their cloud login details and student number into a form, many of them will make mistakes, or fill in the form before they have signed up to the service only to discover their chosen username is already taken.

To conclude by relating back to the reactive learning environments of the 1970s, we still find ourselves in a situation where we are carefully designing an environment that will allow students to work on realistic open-ended problems and gather data on their working processes. However, instead of having to create the entire environment, educators strategically select as many as possible from the authentic tools of the profession. As we connect these to automated analysis and advice, we can begin to realise the goal of cognitive apprenticeships, supported by smart technology, in more authentic environments.

## References

- Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, 19(2), 105–154.
- Aleven, V., McLaren, B. M., Sewall, J., van Velsen, M., Popescu, O., Demi, S., Ringenberg, M., & Koedinger, K. R. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education*, 26(1), 224–269.
- Alferi, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1–18.
- Anderson, J. R. (1993). *Rules of the mind*. Erlbaum.
- Anderson, J. R., Boyle, C. F., & Yost, G. (1985). The geometry tutor. In *Proceedings of the 9th international joint conference on artificial intelligence – Volume 1, IJCAI'85* (pp. 1–7). Morgan Kaufman.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.
- Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems* (pp. 531–540). Springer.
- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.
- Bakharia, A., Kitto, K., Pardo, A., Gašević, D., & Dawson, S. (2016). Recipe for success: Lessons learnt from using xAPI within the Connected Learning Analytics Toolkit. In *Proceedings of the sixth international conference on learning analytics & knowledge, LAK '16* (pp. 378–382). ACM.
- Barnes, T., & Stamper, J. (2010). Automatic hint generation for logic proof tutoring using historical data. *Journal of Educational Technology & Society*, 13(1), 3–12.
- Billingsley, W. (2020). Revisiting the intelligent book: Towards seamless intelligent content and continuously deployed courses. In S. Gregory, S. Warburton, & M. Parkes (Eds.), *ASCILITE's first virtual conference* (pp. 230–240). Proceedings ASCILITE 2020 in Armidale.
- Billingsley, W., & Billingsley, J. (2004). The animation of simulations and tutorial clients for online teaching. In *Proceedings of the 15th annual conference for the Australasian Association for Engineering Education and the 10th Australasian women in engineering forum, Toowoomba, Australia* (pp. 532–540). Australasian Association for Engineering Education.
- Billingsley, W., & Robinson, P. (2005). Towards an intelligent textbook for discrete mathematics. In *Proceedings of the 2005 international conference on active media technology, AMT2005* (pp. 291–296). IEEE Press.
- Billingsley, W., & Robinson, P. (2007a). Searching questions, informal modelling, and massively multiple choice. In S. Wheeler & N. Whitton (Eds.), *Beyond control: Learning technology for the social network generation. Research proceedings of the 14th Association for Learning Technology conference, ALT-C 2007* (pp. 159–168). The Association for Learning Technology.
- Billingsley, W., & Robinson, P. (2007b). Student proof exercises using MathsTiles and Isabelle/HOL in an intelligent book. *Journal of Automated Reasoning*, 39(2), 181–218.
- Billingsley, W., & Robinson, P. (2009). Intelligent books – Combining reactive learning exercises with extensible and adaptive content in an open-access web application. In C. Mourlas, N. Tsianos, & P. Germanakos (Eds.), *Cognitive and emotional processes in web-based education: Integrating human factors and personalization* (pp. 229–244). IGI Global.
- Billingsley, W., & Steel, J. (2013). A comparison of two iterations of a software studio course based on continuous integration. In *Proceedings of the 18th ACM conference on innovation and technology in computer science education, ITiCSE '13* (pp. 213–218). ACM.
- Billingsley, W., & Steel, J. R. H. (2014). Towards a supercollaborative software engineering MOOC. In *Companion proceedings of the 36th international conference on software engineering – ICSE companion 2014* (pp. 283–286). ACM.

- Billingsley, W., Torbay, R., Fletcher, P. R., Thomas, R. N., Steel, J. R., & Süß, J. G. (2019). Taking a studio course in distributed software engineering from a large local cohort to a small global cohort. *ACM Transactions on Computing Education*, 19(2). <https://doi.org/10.1145/3218284>
- Bloom, B. (1984). The two sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13, 4–15.
- Brooks, F. P. (1987). No silver bullet. *IEEE Computer*, 20(4), 10–19.
- Brown, J. S., Burton, R. R., & Bell, A. G. (1975). Sophie: A step toward creating a reactive learning environment. *International Journal of Man-Machine Studies*, 7(5), 675–696.
- Brown, J. S., Burton, R. R., Bell, A. G., & Bobrow, R. J. (1974). Sophie: A sophisticated instructional environment. In *Technical report AFHRL-TR-74-93*. Bolt Beranek and Newman Inc.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42.
- Brown, J. S., Rubinstein, R., & Burton, R. (1976). *Reactive learning environment for computer assisted electronics instruction*. Technical report AFHRL-TR-76-68. Bolt Beranek and Newman Inc.
- Chiu, J. L., DeJaegher, C. J., & Chao, J. (2015). The effects of augmented virtual science laboratories on middle school students' understanding of gas properties. *Computers & Education*, 85, 59–73.
- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–491). Laurence Erlbaum Associates.
- Colvin, C., Rogers, T., Wade, A., Dawson, S., Gašević, D., Buckingham Shum, S., & Fisher, J. (2015). *Student retention and learning analytics: A snapshot of Australian practices and a framework for advancement*. Australian Office for Learning and Teaching.
- Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In *User modeling 2001: 8th international conference, UM 2001, Sonthofen, Germany* (pp. 137–147). Springer.
- Corbett, A., & Trask, H. (2000). Instructional interventions in computer-based tutoring: Differential impact on learning time and accuracy. In *CHI '00: Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 97–104). ACM Press.
- Dawson, S., Tan, J. P.-L., & McWilliam, E. (2011). Measuring creative potential: Using social network analysis to monitor a learners' creative capacity. *Australasian Journal of Educational Technology*, 27(6), 924–942.
- De Liddo, A., Shum, S. B., Quinto, I., Bachler, M., & Cannavacciuolo, L. (2011). Discourse-centric learning analytics. In *Proceedings of the 1st international conference on learning analytics and knowledge, LAK '11* (pp. 23–33). ACM.
- Deslauriers, L., & Wieman, C. (2011). Learning and retention of quantum concepts with different teaching methods. *Physical Review Special Topics – Physics Education Research*, 7(1), 010101.
- Docherty, M., Sutton, P., Brereton, M., & Kaplan, S. (2001). An innovative design and studio-based CS degree. In *Proceedings of the thirty-second SIGCSE technical symposium on computer science education, SIGCSE '01* (pp. 233–237). ACM.
- Dutton, T. A. (1987). Design and studio pedagogy. *Journal of Architectural Education*, 41(1), 16–25.
- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5–6), 304–317.
- Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, 82(3), 300–329.
- Gibson, A., Kitto, K., & Bruza, P. (2016). Towards the discovery of learner metacognition from reflective writing. *Journal of Learning Analytics*, 3(2), 22–36.
- Hendrix, D., Myneni, L., Narayanan, H., & Ross, M. (2010). Implementing studio-based learning in CS2. In *Proceedings of the 41st ACM technical symposium on computer science education* (pp. 505–509). ACM.
- Herrington, J., Reeves, T. C., & Oliver, R. (2006). Authentic tasks online: A synergy among learner, task, and technology. *Distance Education*, 27(2), 233–247.

- Hobson, L. M. (2009, March). Putting a bolder face on Google. *New York Times*.
- Hundhausen, C. D., Narayanan, N. H., & Crosby, M. E. (2008). Exploring studio-based instructional models for computing education. *ACM SIGCSE Bulletin*, 40(1), 392.
- Janning, R., Schatten, C., & Schmidt-Thieme, L. (2016). Perceived task-difficulty recognition from log-file information for the use in adaptive intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26(3), 855–876.
- Khousa, E. A., Atif, Y., & Masud, M. M. (2015). A social learning analytics approach to cognitive apprenticeship. *Smart Learning Environments*, 2(1), 14.
- Kitto, K., Bakharia, A., Lupton, M., Mallet, D., Banks, J., Bruza, P., Pardo, A., Shum, S. B., Dawson, S., Gašević, D., et al. (2016). The Connected Learning Analytics Toolkit. In *Proceedings of the sixth international conference on learning analytics & knowledge, LAK '16* (pp. 548–549). ACM.
- Lajoie, S. P., & Lesgold, A. M. (1992). Dynamic assessment of proficiency for solving procedural knowledge tasks. *Educational Psychologist*, 27(3), 365–384.
- Loddington, S., Pond, K., Wilkinson, N., & Willmot, P. (2009). A case study of the development of WebPA: An online peer-moderated marking tool. *British Journal of Educational Technology*, 40(2), 329–341.
- Long, J. G. (2012). State of the studio: Revisiting the potential of studio pedagogy in U.S.-based planning programs. *Journal of Planning Education and Research*, 32(4), 431–448.
- Ma, J., & Nickerson, J. V. (2006). Hands-on, simulated, and remote laboratories: A comparative literature review. *ACM Computing Surveys*, 38(3), 1–24.
- Mah, D.-K. (2016). Learning analytics and digital badges: Potential impact on student retention in higher education. *Technology, Knowledge and Learning*, 21(3), 285–305.
- McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload!: Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education*, 28(3), 420–438.
- Melis, E., Kärger, P., & Homik, M. (2005). Interactive concept mapping in ActiveMath. In J. N. Haake, U. Lucke, & D. Tavangarian (Eds.), *Delfi 2005: 3. Deutsche eLearning Fachtagung Informatik, volume 66 of LNI* (pp. 247–258). Rostock.
- Merceron, A., & Yacef, K. (2003). A web-based tutoring tool with mining facilities to improve learning and teaching. In *Proceedings of the 11th international conference on artificial intelligence in education* (pp. 201–208). IOS Press.
- Mitrovic, A. (2012). Fifteen years of constraint-based tutors: What we have achieved and where we are going. *User Modeling and User-Adapted Interaction*, 22(1–2), 39–72.
- Mitrovic, A., Mayo, M., Suraweera, P., & Martin, B. (2001). Constraint-based tutors: A success story. In *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 931–940). Springer.
- Montgomery, K. (2002). Authentic tasks and rubrics: Going beyond traditional assessments in college teaching. *College Teaching*, 50(1), 34–40.
- Nipkow, T., Paulson, L. C., & Wenzel, M. (2002). *Isabelle/HOL: A proof assistant for higher-order logic, volume 2283 of LNCS*. Springer.
- Nurkkala, T., & Brandle, S. (2011). Software studio: Teaching professional software engineering. In *Proceedings of the 42nd ACM technical symposium on computer science education, SIGCSE '11* (pp. 153–158). ACM.
- Pardo, A. (2013). Social learning graphs: Combining social network graphs and analytics to represent learning experiences. *International Journal of Social Media and Interactive Learning Environments*, 1(1), 43–58.
- Pardo, A., & Kloos, C. D. (2011). Stepping out of the box: Towards analytics outside the learning management system. In *Proceedings of the 1st international conference on learning analytics and knowledge, LAK '11* (pp. 163–167). ACM.
- Pardos, Z. A., Tang, S., Davis, D., & Le, C. V. (2017). Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework. In *Proceedings of the fourth (2017) ACM conference on learning @ scale* (pp. 23–32). ACM.

- Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D., Yee, N., & Roll, I. (2017). Identifying productive inquiry in virtual labs using sequence mining. In *International conference on artificial intelligence in education* (pp. 287–298). Springer.
- Perkel, J. M. (2014). Scientific writing: The online cooperative. *Nature News*, *514*(7520), 127–128.
- Petraglia, J. (1998). The real world on a short leash: The (mis)application of constructivism to the design of educational technology. *Educational Technology Research and Development*, *46*(3), 53–65.
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)* (pp. 505–513). Cambridge, MA, USA: MIT Press.
- Piech, C., Sahami, M., Koller, D., Cooper, S., & Blikstein, P. (2012). Modeling how students learn to program. In *Proceedings of the 43rd ACM technical symposium on computer science education* (pp. 153–160). ACM.
- Pyatt, K., & Sims, R. (2012). Virtual and physical experimentation in inquiry-based science labs: Attitudes, performance and access. *Journal of Science Education and Technology*, *21*(1), 133–147.
- QILT. (2017). *2017 employer satisfaction survey: National report*. Technical report. The Social Research Centre.
- Ritter, S., & Koedinger, K. R. (1997). An architecture for plug-in tutoring agents. *Journal of Artificial Intelligence in Education*, *7*, 315–347.
- Rodriguez-Bustos, C., & Aponte, J. (2012). How distributed version control systems impact open source software projects. In *Mining software repositories (MSR), 2012 9th IEEE working conference* (pp. 36–39). IEEE Press.
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.
- Root, D., Rosso-Llopart, M., & Taran, G. (2008). Exporting studio: Critical issues to successfully adopt the software studio concept. In *21st conference on software engineering education and training, CSEET '08* (pp. 41–50). <https://doi.org/10.1109/CSEET.2008.21>
- San Pedro, M. O. Z., Baker, R. S. J. D., & Rodrigo, M. M. T. (2014). Carelessness and affect in an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education*, *24*(2), 189–210.
- Schneider, J., Bernstein, A., vom Brocke, J., Damevski, K., & Shepherd, D. (2017). Detecting plagiarism based on the creation process. *IEEE Transactions on Learning Technologies*, *11*(3), 348–361.
- Schön, D. A. (1984). The architectural studio as an exemplar of education for reflection-in-action. *Journal of Architectural Education*, *38*(1), 2–9.
- Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey Bass.
- Seely Brown, J., & Adler, R. (2008). Open education, the long tail, and learning 2.0. *Educause Review*, *43*(1), 16–20.
- Süß, J. G. & Billingsley, W. (2012). Using continuous integration of code and content to teach software engineering with limited resources. In *Proceedings of the 34th International Conference on Software Engineering, ICSE2012* (pp. 1175–1184). doi: <https://doi.org/10.1109/ICSE.2012.6227025>.
- Tomayko, J. E. (1991). Teaching software development in a studio environment. *ACM SIGCSE Bulletin*, *23*(1), 300–303.
- Tomayko, J. E. (1996). Carnegie Mellon's software development studio: A five year retrospective. In *Proceedings of the 9th conference on software engineering education* (pp. 119–129). IEEE.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education, 15*, 147–204.
- Venant, R., Sharma, K., Vidal, P., Dillenbourg, P., & Broisin, J. (2017). Using sequential pattern mining to explore learners' behaviors and evaluate their correlation with performance in inquiry-based learning. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data driven approaches in digital education* (pp. 286–299). Springer.
- Wang, L., Sy, A., Liu, L., & Piech, C. (2017). Deep knowledge tracing on programming exercises. In *Proceedings of the fourth (2017) ACM conference on learning @ scale, L@S '17* (pp. 201–204). ACM.
- Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703–713.
- Wiggins, G. (1990). *The case for authentic assessment*. ERIC digest. Washington, DC: Office of Educational Research and Improvement. (ERIC document reproduction service no. ED328611).
- Wilson, K. H., Karklin, Y., Han, B., & Ekanadham, C. (2016). Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. *arXiv:1604.02336*.
- Xiong, X., Zhao, S., Van Inwegen, E., & Beck, J. (2016). Going deeper with deep knowledge tracing. In *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016* (pp. 545–550). International Educational Data Mining Society.
- Zacharia, Z. C., Olympiou, G., & Papaevripidou, M. (2008). Effects of experimenting with physical and virtual manipulatives on students' conceptual understanding in heat and temperature. *Journal of Research in Science Teaching, 45*(9), 1021–1035.

**William Billingsley** is a Lecturer in Computer Science at the University of New England (UNE). He has been exploring how to create technology that can enable smarter and more social education since the early 2000s. Originally trained as a software engineer, he undertook his PhD as part of the *Intelligent Book* project – a joint project between the University of Cambridge and Massachusetts Institute of Technology that considered what learning would be like if the artificial intelligence was your partner instead of your marker. He created one of the first blocks-based programming languages for mathematical proofs and explored how artificial intelligence and students can think together. From 2009, he worked as a senior research engineer with NICTA, Australia's centre of excellence for ICT research. His projects were about connecting data and multi-modal analysis to gain a better understanding of collaboration and communication. While working with NICTA, he began teaching software engineering topics with the University of Queensland. Together with colleagues, he developed the first “supercollaborative” software studio class, in which teams of students collaborate on a class-wide project (in 1 year, 200 students working together on the same project). In 2015, he moved to UNE, where he coordinates the university's undergraduate computer science degree, creates interactive outreach activities for school students, teaches software development and interaction design topics, and has brought studio collaboration to the university's online students.

**Peter Fletcher** is a Lecturer in ICT Education. He teaches into both undergraduate and post-graduate programmes in the School of Education, at the University of New England (UNE), Australia. Prior to working at UNE, Peter was a secondary teacher with a teaching background in Computing Studies, Science and Mathematics. Peter fondly looks back on his days teaching in rural New South Wales and brings many of these experiences to his current research directions and teaching philosophy. These directions include his research interests in Learning Design, Learning Management Systems, e-learning competencies, assessment, information literacy and the integration of ICT across the curriculum. Peter teaches ICT Education to pre-service and in-service teachers within primary, secondary, and post-graduate programs. In recognition for his teaching, Peter was jointly awarded a *Citation for Outstanding Contributions to Student Learning*, 2016 Australian Awards for University Teaching. Peter's teaching and research is based on the constructivist theory

of knowledge, which assumes that knowledge is created in the mind of the learner. His driving passion involves the search for alternatives to the classical lecture technique that has dominated the science and mathematics classroom. Peter's goal is to understand and overcome the constraints teachers face that interfere with creating an effective classroom environment which integrates ICT, that is consistent with what they believe is the optimum way for students to learn.



## Chapter 8

# Learning from Learning Analytics: How Much Do We Know About Patterns of Student Engagement?



Belinda A. Chiera, Małgorzata W. Korolkiewicz, and Lisa J. Schultz

**Abstract** Learning analytics is a rapidly growing field of research to inform the process of using data to improve learning and teaching. A common focus is to explore and develop strategies to address student learning and engagement issues at the institutional level. Typically, the data are captured by virtual learning environment systems and often complemented with student demographic information. At present, there is no standardised, holistic approach to data storage and collection across learning institutions, resulting in student learning data sets that are very large, incomplete and non-homogeneous.

In this chapter, we aim to harness these data in order to gain a better understanding of students' online engagement patterns specifically in a standard university setting that combines face-to-face learning with online resources and learning environments. Moreover, in the context of first-year University students, it is particularly vital to understand student patterns of online engagement. The data for the research conducted here has been sourced from a large Australian university. Patterns of student engagement are explored using a variety of data visualisations and statistical analysis techniques to develop an understanding of student engagement with course resources over the duration of a semester.

**Keywords** Learning analytics · University · Online engagement · Patterns

---

B. A. Chiera (✉) · M. W. Korolkiewicz · L. J. Schultz  
University of South Australia, Adelaide, SA, Australia  
e-mail: [Belinda.Chiera@unisa.edu.au](mailto:Belinda.Chiera@unisa.edu.au); [malgorzata.korolkiewicz@unisa.edu.au](mailto:malgorzata.korolkiewicz@unisa.edu.au);  
[Lisa.Schultz@unisa.edu.au](mailto:Lisa.Schultz@unisa.edu.au)

© Springer Nature Switzerland AG 2021  
T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy  
Implications of Research in Education 13,  
[https://doi.org/10.1007/978-3-030-76841-6\\_8](https://doi.org/10.1007/978-3-030-76841-6_8)

## Introduction

Learning analytics is a rapidly growing field of research to inform the process of using data to improve learning and teaching. According to the definition proposed by the Society for Learning Analytics Research (SoLAR), learning analytics refers to the measurement, collection, analysis, and reporting of data about the progress of learners and the contexts in which learning takes place (Siemens & Gasevic, 2012). Typically, the data are captured by virtual learning environment systems and complemented with student demographic information. Although learning analytics in higher education is still at relatively early stages of development, it is expected to make significant contributions in a number of areas, including quality assurance, improvement of retention rates, identifying students at risk, assessment of outcomes among distinct student sub-groups, and adaptive learning (Sclater et al., 2016).

The widespread adoption by universities of learning management systems to create virtual learning environments has fundamentally changed how students engage with their studies. There is no single definition of student engagement, however students' effort, involvement in learning activities, motivation to learn, and consequently their academic achievement are typically considered (see for example Beer et al., 2010; Henrie et al., 2015). Virtual learning environments offer a unique opportunity to capture aspects of students' engagement in a rich data source providing information about changes in learner behaviour. In this chapter, we study different aspects of students' online activity to identify how and when they engage with online course resources.

To date, student online engagement has been explored with a predominant focus on Massive Open Online Courses (MOOCs), with latent engagement patterns examined in Ramesh et al. (2014), connections with performance explored in Coffrin et al. (2014) and Phan et al. (2015), and learner profiles studied in Kizilcec et al. (2013) and Ferguson and Clow (2015), just to name a few. There is however comparatively less emphasis on student patterns of engagement in the standard university setting that combines face-to-face learning with online resources and learning environments. Studies that do exist (see for example Beer et al., 2010 or Henrie et al., 2015) indicate a vital link between student engagement online and a successful learning experience overall. In the context of first-year University students, it therefore seems to be particularly vital to understand student patterns of online engagement as early experiences are likely to set the foundations for meaningful interaction with the University learning environment for the duration of their study.

The data considered here have been sourced from a large Australian university and encompass all student records across the 2016 academic year. The data combine student demographic information and weekly counts of visits to online course resources, used as a proxy for student online engagement. The weekly course click-count data provide a fine granularity of online engagement data and reflect the heterogeneous design and delivery of online resources in courses from distinct learning

areas. For the purposes of this chapter, we focus on a specific first-year course selected from the health area and examine click-count data for its virtual learning environment for semester one 2016. Patterns of student engagement are explored using data visualisations and statistical analysis techniques to develop an understanding of student behaviour, including how students engage with course resources over the duration of a semester. As in studies dedicated to student engagement patterns in MOOCs (see for example Hung & Zhang, 2008; Kizilcec et al., 2013; Hughes & Dobbins, 2015), we apply data mining techniques to analyse weekly virtual learning environment visits, in particular a hybrid hierarchical  $k$ -means clustering approach.

In light of existing studies and our own experience with virtual learning environments, we aim to gain insight into the following questions:

- How do levels of access to online resources vary over the course of a semester?
- How do student characteristics influence their engagement with online resources over the course of a semester?
- What role does assessment play in guiding the use of online resources?

The layout of this chapter is as follows. In section “[Course Data](#)”, the student engagement data used for the analysis is described, and the results of the analysis are presented in section “[Results](#)”. In particular, the visualisations of patterns in engagement behaviours are presented and discussed in section “[Patterns of Student Engagement with Course Virtual Environment](#)”. This is then followed by a statistical analysis of the weekly engagement data using hybrid hierarchical  $k$ -means clustering across all students section “[Clustering Weekly Engagement Patterns](#)”, with the analysis extended to clustering between student cohorts in section “[Clustering Weekly Engagement Patterns by Cohort Classification](#)”. Conclusions are then drawn in section “[Conclusions](#)” on the patterns of student engagement identified in this chapter.

## Course Data

The course selected for analysis is a first-year biology subject dedicated to the study of the structure and function of the human body. Our data consist of counts based on online learning platform logs for semester one in 2016 and captures the online activity of 745 individuals. There were both internal and external (online) offerings in the course, with the majority (75%) of students enrolled to study internally. The student cohort was predominantly female (85%) and 66% of students were under 21 years of age. To reflect their status as recent high school graduates, these students who were under 21 years of age are hereafter referred to as ‘school leavers’, with the remaining 34% of students who were over 21 years old referred to as ‘mature age’ students. It is worth noting that mature age students were the majority in the external offering of the course (66%), in contrast to the internal offering, where the

majority were school leavers (76%). The gender split in the internal and external offering was similar to that in the course cohort overall.

The course online learning environment was created within the university Moodle-based platform and included a range of resources and activities. The course syllabus was divided into weekly topics and organised into ‘book’ resources populated with an overview, readings, lecture notes, exercises, lecture recordings, ‘test your knowledge’ formative quizzes and additional resources. There was also a social forum created for each course offering. Course assessment included both formative and summative online quizzes as well as in-person tests and exams. Our analysis of student engagement with the virtual learning environment for this course is based on course site visits as well as student interactions with the ‘book’ resource and the ‘forum’ activity.

One of the limitations of our data is that for a selection of cases where final grades were not recorded, we are not able to distinguish between students who have unenrolled from the course and students who stopped engaging online however remained enrolled in the course. We are also not able to tell whether the recorded counts are simply multiple instances of the course website or a particular resource being opened in a student’s browser or the number of student’s clicks through the sections of the website. We are thus not able to meaningfully interpret the size of the counts or infer the exact nature of the student’s engagement with the virtual learning environment. For those reasons, in what follows we mainly focus on general patterns only of student’s online activity.

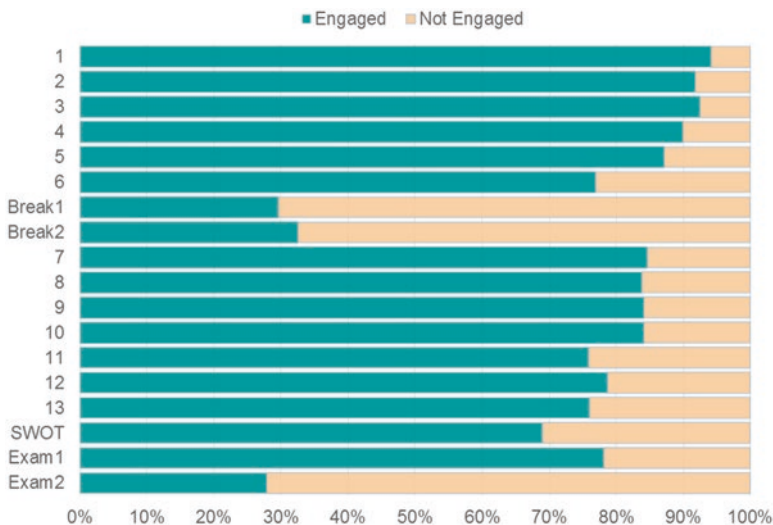
## Results

Results presented in this section are based on weekly counts of clicks or ‘visits’ to the course website as well as specific Moodle resources and activities that exist within it. We focus on the main teaching weeks plus two ‘mid-break’ weeks, the end-of-semester examination period and the dedicated revision week separating the teaching weeks and the examination period, known as the ‘swotvac’ week. Based on our own teaching experience, we expected assessment to be an important factor in determining when and how students engage with online learning resources. In the course we have selected for analysis, the major pieces of assessment were a mid-test scheduled for teaching Week 5, a series of six online quizzes in Weeks 3–4 and 7–10, a practical examination in Week 12, as well as a final examination. In addition to summative assessment, there were also a number of formative online quizzes and so one can reasonably expect the timing of assessment to be strongly reflected in student engagement with the course website.

## *Patterns of Student Engagement with Course Virtual Environment*

The overall pattern of student engagement is depicted in Fig. 8.1. Engagement with the course website during a week was defined as visiting the course website at least once. In general, the proportion of students visiting the course website was quite high at 70–80% for most weeks. It is however immediately apparent that students engaged with course online resources predominantly in the teaching weeks. The proportion of students visiting the course website during the mid-break and the second exam week was approximately 30% compared to over 90% in Week 1. Interestingly, the proportion of students visiting the course website during the first exam week was very high at 80%. Although we are not able to verify this, we expect the likely reason for such a high level of interaction with the online learning environment to be that the final examination was scheduled early in the second exam week. We also note a tendency in the student cohort to engage progressively less as the semester unfolded, with fewer visits to the course website in weeks without summative assessment, for example in Week 6 following the mid-test or in Weeks 11–13, despite the fact that summative practical examination was scheduled in Week 12.

In order to better understand the weekly patterns of visits to the course website, we also examine the proportions of students who engaged with the course website by study mode (internal versus external), gender (female versus male) and age (school leavers versus mature age students). From Figs. 8.2 and 8.3 there appears to be a very similar pattern of engagement across weeks for internal female and male students, with one noteworthy difference, namely much higher proportion of male



**Fig. 8.1** Student engagement with course virtual environment

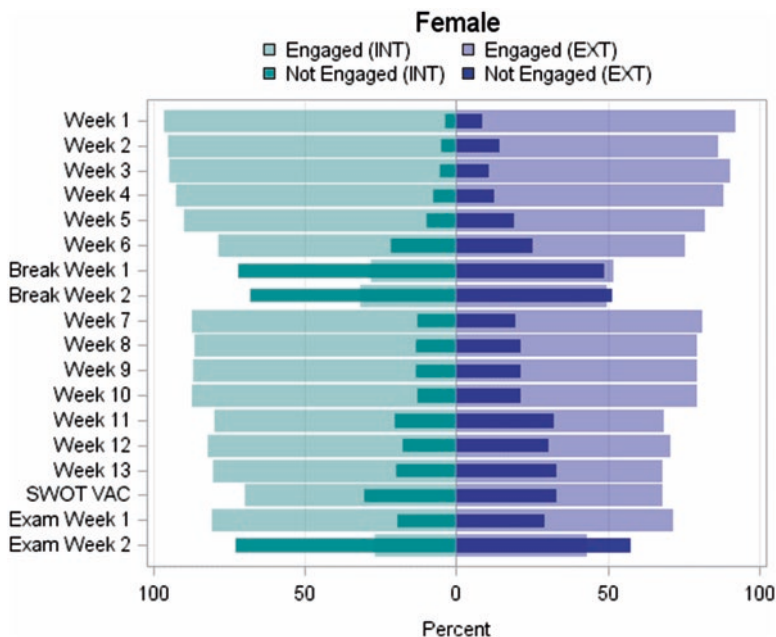


Fig. 8.2 Student engagement with course virtual environment by study mode among female students

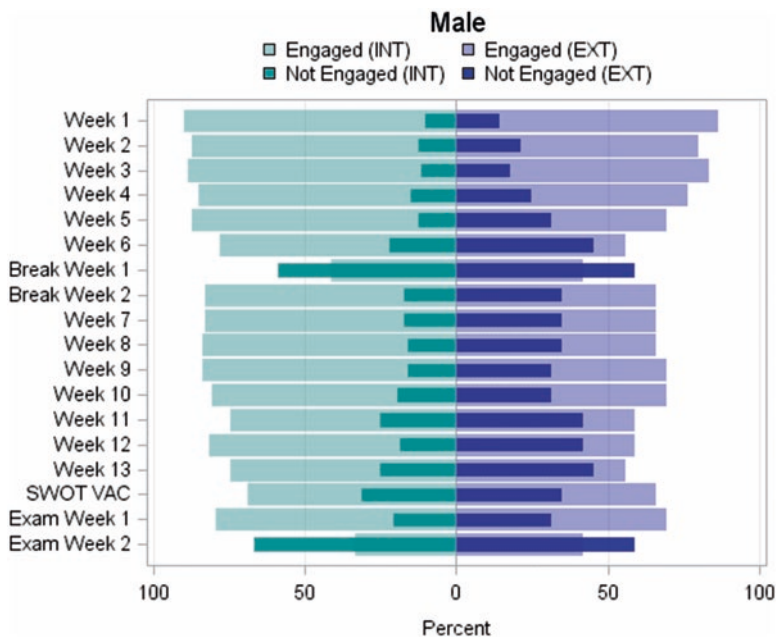


Fig. 8.3 Student engagement with course virtual environment by study mode among male students

students visiting the course website during the second week of the mid-break. In contrast, the proportion of female students visiting the course website was higher in the first few weeks of the semester. Comparing weekly proportions for internal and external female students we see a very similar pattern, however the proportion of female external students visiting the website is generally slightly lower.

This is particularly the case after Week 10 when there are no more summative online quizzes. The only exceptions are the mid-break and the second exam week when a much higher proportion of female external students visited the website. Finally, male external students appear to stop visiting the course website much sooner in the semester. In short external students, in particular male external students, tended to disengage from the course sooner than their internal or female counterparts. Based on Figs. 8.4 and 8.5 we can also say that generally, a greater proportion of mature age students continued to visit the course website throughout the semester. There was little difference in engagement pattern of mature age students between the internal and external offering, with the exception of the mid-break. Approximately 60% of mature age external students continued to visit the course website during the break, compared with 30% of mature age internal students. For school leavers we note a higher proportion of students progressively disengaging from the course website, and likely from the course in general, particularly in the external offering.

The frequency with which students engaged with the course website each week was then explored, with the results shown in Fig. 8.6. Note that this analysis focussed only on students who accessed the course site at least once in a given week. From

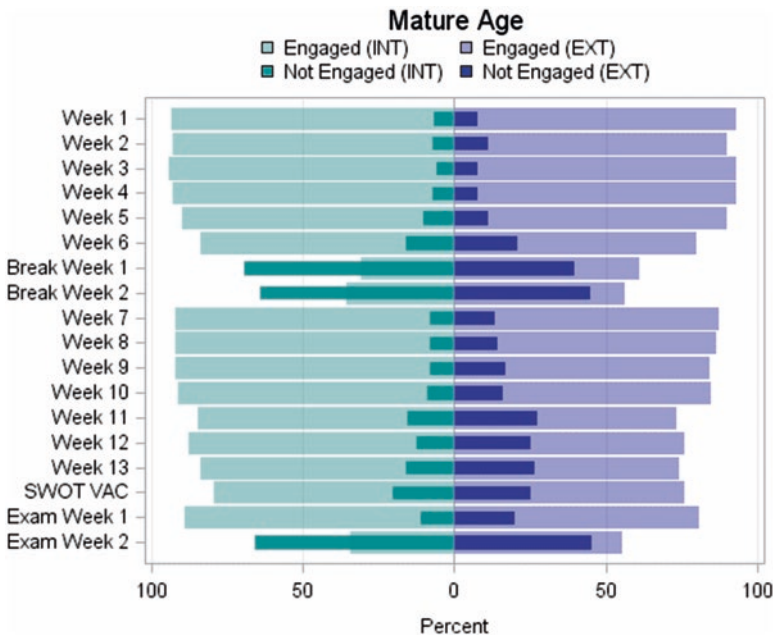
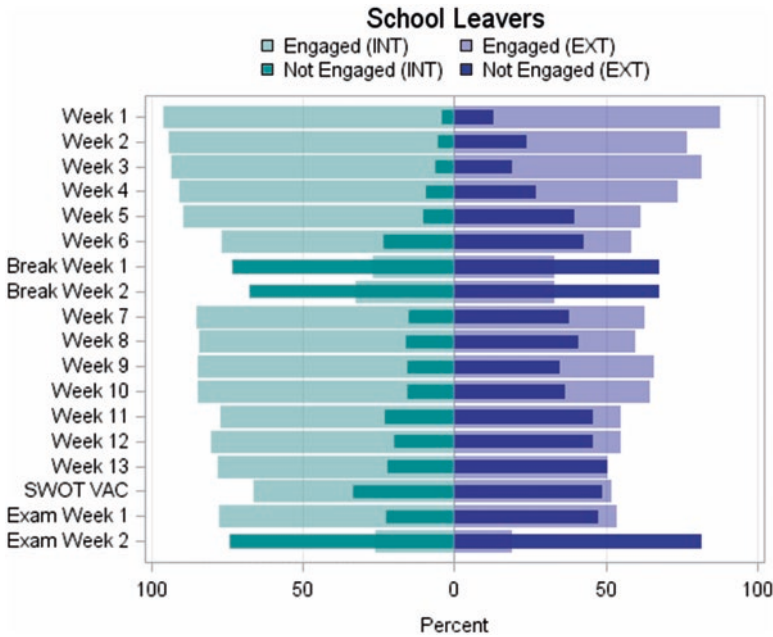
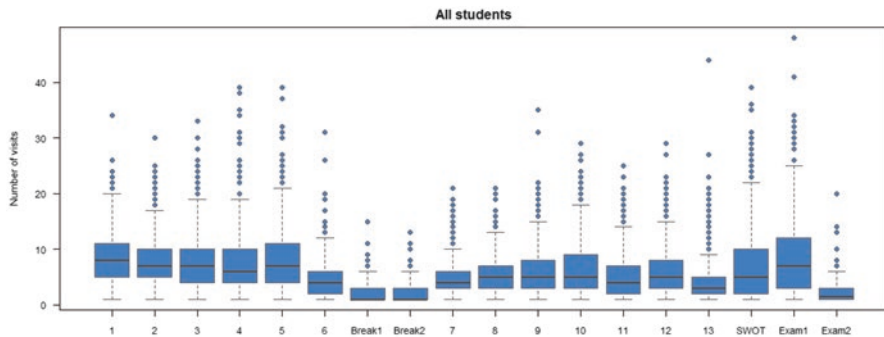


Fig. 8.4 Student engagement with course virtual environment by study mode and age category



**Fig. 8.5** Student engagement with course virtual environment by study mode and age category



**Fig. 8.6** Weekly course site visits for all students

Fig. 8.6 it can be seen that there are a selection of students with unusually high counts of course site visits across all teaching weeks, with some students accessing the course website more than 40 times in the space of a week. For the early teaching weeks, namely Weeks 1–5, the distribution of engagement counts is relatively consistent in terms of the typical number of course site visits and the associated variability in engagement between students. Overall engagement counts appear to decrease notably in Week 6 however and remain low across each of the two mid-semester break weeks. Following the mid-semester teaching break, course site visits become relatively high in variability.



Figure 8.6 also shows engagement counts for visits to the course website are higher during the first week of the exam period. This is suspected to be a reflection of an increased level of revision by the students in the lead-up to their final exam for the course.

Among students who engaged with the course site visit, the weekly frequency with which engagement occurred was then explored by the students' study mode, gender and age. Figure 8.7 shows the distribution of weekly engagement counts by gender, from which it can be seen that patterns of engagement frequency do not appear to differ substantially between male and female students.

Figure 8.8 shows the weekly engagement counts for internal and external students. The overall temporal pattern in weekly engagement appears to be similar

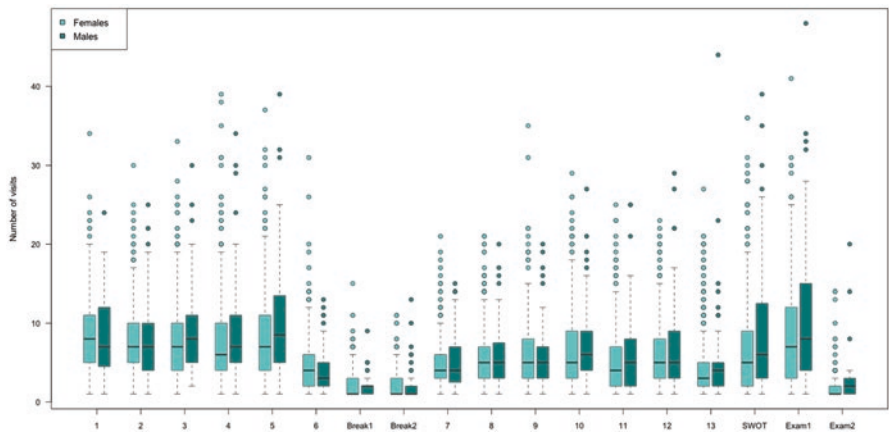


Fig. 8.7 Weekly course site visits by gender

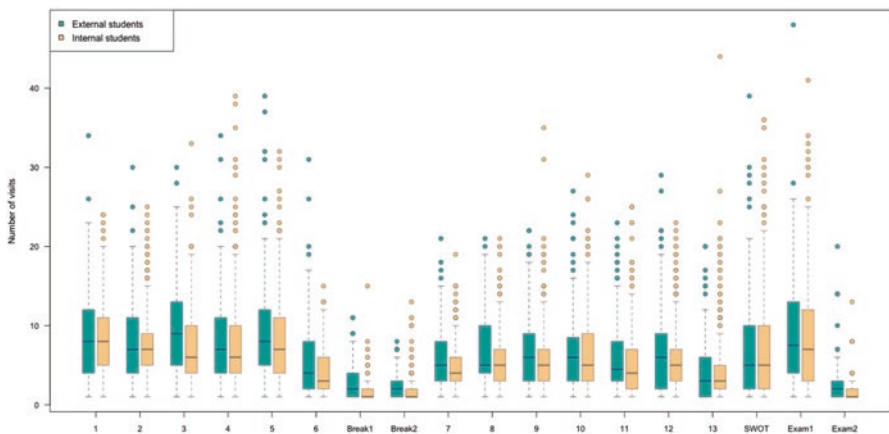


Fig. 8.8 Weekly course site visits by study mode

between the two different study modes, with the exception that students who are externally enrolled typically have higher engagement counts for course site visits than their internal counterparts, as expected. The external students also appear to be more consistent in their engagement behaviour than students who are internally enrolled.

The engagement behaviours of mature age and school leaver students are comparatively shown in Fig. 8.9. Mature age students are typically more engaged than their school leaver counterparts, with greater variability observed in the course site counts between mature age students than school leavers. This trend in engagement patterns persists when separately considering internal and external students by age group (Figs. 8.10 and 8.11). Among the internal students, the mature age and school

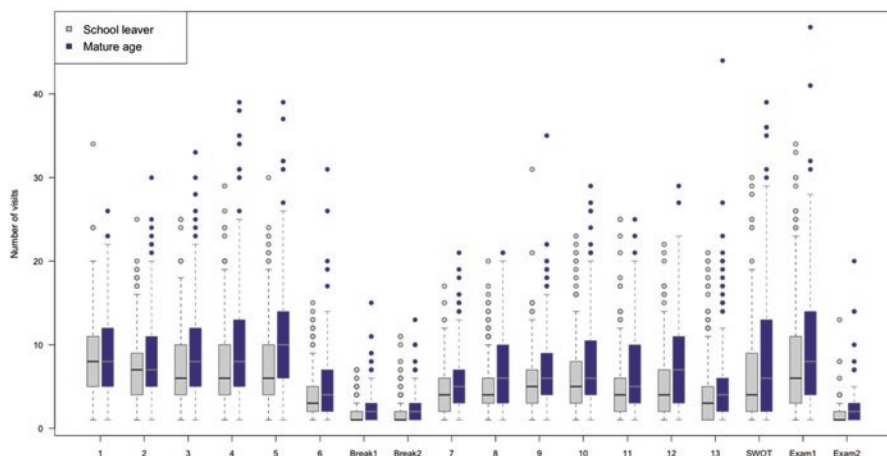


Fig. 8.9 Weekly course site visits for by age group

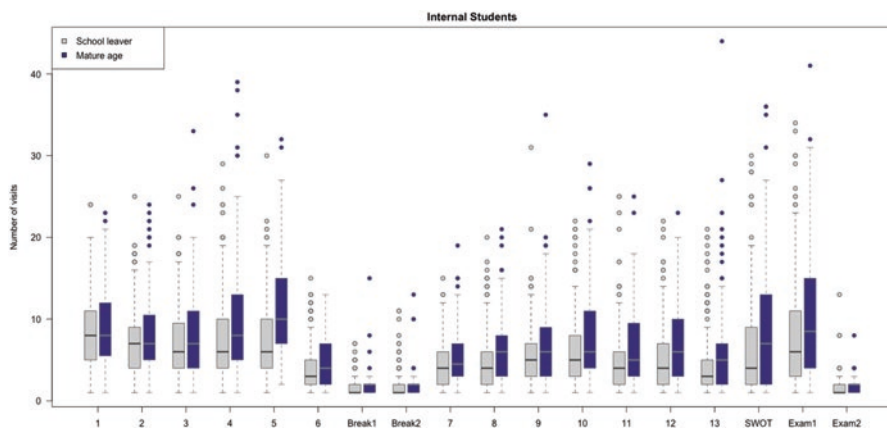


Fig. 8.10 Weekly course site visits for the internal students by age group

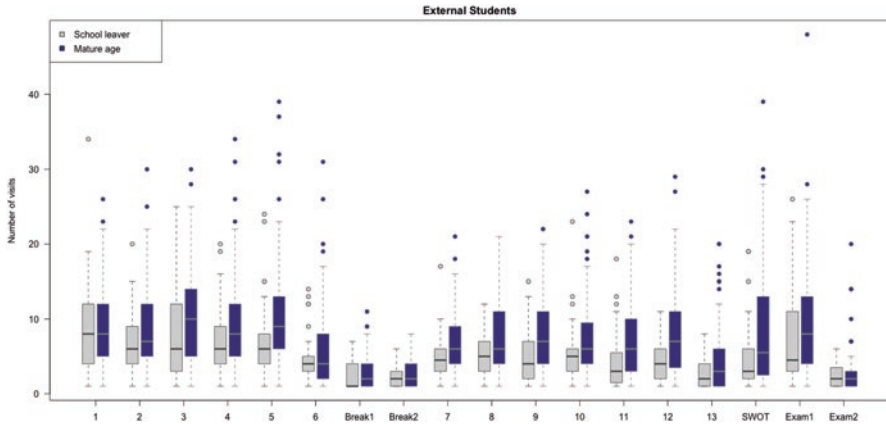


Fig. 8.11 Weekly course site visits for the external students by age group

leaver students have similar engagement behaviour for the early teaching weeks with the mature age students typically having higher engagement from approximately Week 5 onwards. By contrast, external mature age students are consistently higher in course site visits than external school leaver students across the entire semester.

To gain further insight into the nature of student demographics and the effect they may have on engagement with course online resources, a treemap of online course site visits was constructed using student gender and study mode as the grouping variables (Fig. 8.12).

Each of the four grouping variable combinations is represented with a rectangular block outlined in black, the size of which is determined by the number of students in that group. Each block is further divided into tiles representing weeks, with the size of each tile driven by the number of course visits for the corresponding week. The tiles are then coloured according to the average number of course site visits.

From Fig. 8.12 it can be seen that the blocks for the internally and externally enrolled females are considerably larger than the blocks for their male counterparts, reflecting the gender composition of the student cohort. Additionally, the blocks for the internal students are consistently larger in size than for the external students, again in keeping with the student enrolment distribution for this course.

Figure 8.12 also shows that Weeks 1–5 and Week 17 (i.e., exam week 1) occupy the greatest areas in the treemap for internal and external females as well as internal male students. This indicates that these are the teaching weeks for which these students are most actively accessing the online course site. In contrast, external male students appear to peak later in the teaching semester, with Weeks 1, 3, 16 and 17 as the most significant weeks for engagement. This indicates that external male students tend to increase their engagement significantly in the weeks immediately preceding the exam, with comparatively low engagement early in the semester.

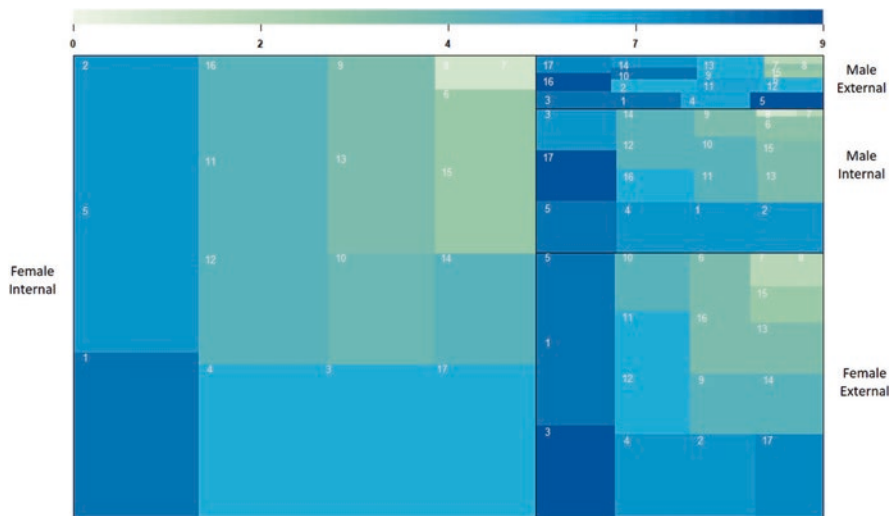
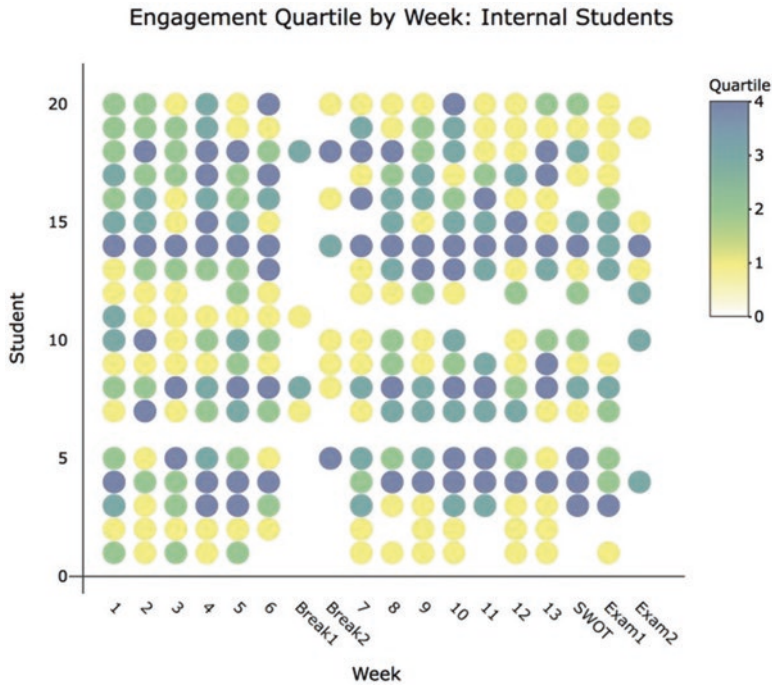


Fig. 8.12 Treemap of the weekly course site visits by student age, gender, and study mode

In terms of tile colour, Fig. 8.12 shows the typical number of course site visits per student appears to be highest in the first 5 weeks in the semester across all blocks. Among the internal female students, Week 1 corresponds to the greatest typical number of visits whereas Week 3 had the greatest number of visits for the external female students. In contrast, male students typically increased their engagement in the weeks leading up to the examination period, namely Week 16 for the external students and Week 17 for the internal students. Patterns of engagement therefore appear to be strongly associated with gender and study mode.

Additional insight into engagement patterns can be gained by categorising weekly student engagement counts into quartiles such that the lowest 25% of students are classified as being in Quartile 1, the lowest quartile of engagement, and the top 25% are classified as Quartile 4. This provides an opportunity to analyse student engagement trajectories across the teaching semester. As an example, Figs. 8.13 and 8.14 show the engagement quartiles trajectories for a random sample of 20 internal students and 20 external students, respectively. Within each figure, the light yellow circles in a given week indicate that the corresponding student was in the lowest quartile of engagement, Quartile 1, whereas the darker shaded circles show students in the highest quartile of engagement, Quartile 4, for that week. Gaps in the engagement trajectories indicate that the students did not engage at all that week. It can be seen that there are notably more yellow circles in the sample of internal students as compared to the external students, indicating these students were more often less engaged with the online course resources than their external peers. Figures 8.13 and 8.14 show that students with low engagement tended to remain in the lower quartiles of engagement across the entire semester while students in the highest quartile of engagement tended to have consistently high engagement. This is particularly



**Fig. 8.13** Engagement quartile trajectories for the internally enrolled students across the semester

apparent among the external students (Fig. 8.14), with the exception of the mid-semester break and the exam weeks where engagement generally decreased.

We now turn our attention to students who visited the course website and examine their patterns of engagement with Moodle books containing weekly topic resources. Overall proportions of students visiting topic resources each week are depicted in Fig. 8.15. As was the case with course site visits, a very high proportion of students (over 90%) were visiting topic resources in the first few weeks, with the proportion decreasing progressively but staying relatively higher in weeks with summative assessment (Weeks 3–5 and then 7–10). Further, the proportion of students visiting the course website who then navigated to topic resources was quite high (around 80%) in weeks leading up to the final examination.

Among students who engaged with the weekly course topic resources, the frequency with which students accessed course resources is shown in Fig. 8.16. Engagement frequency was typically greatest in the early teaching weeks, with notably fewer visits to the topic resources during the mid-semester break and Week 13. As was seen with the frequency of course site visits (Fig. 8.6), the frequency with which students accessed topic resources was greatest during the swotvac week and the first week of the exam when students are revising for their final examination. Engagement patterns according to study mode, gender, and age (not shown) were very similar.

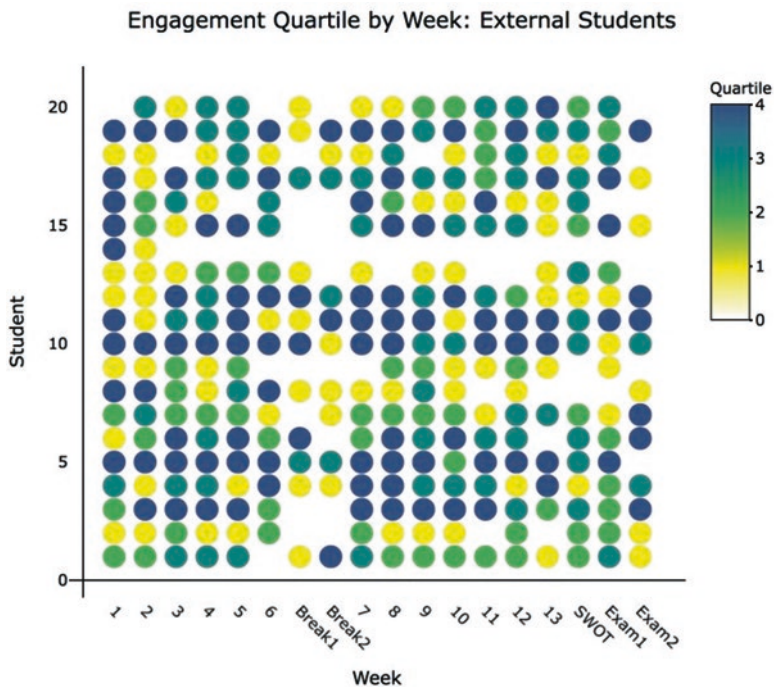


Fig. 8.14 Engagement quartile trajectories for the externally enrolled students across the semester

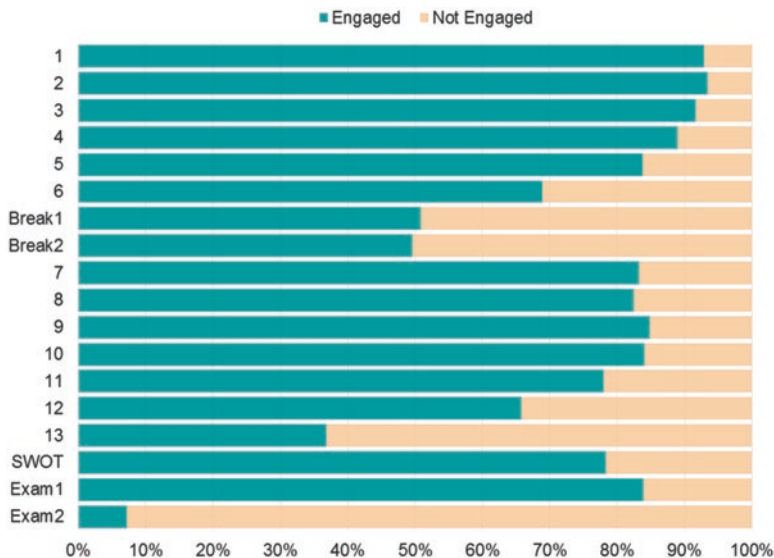


Fig. 8.15 Student engagement with course topic resources

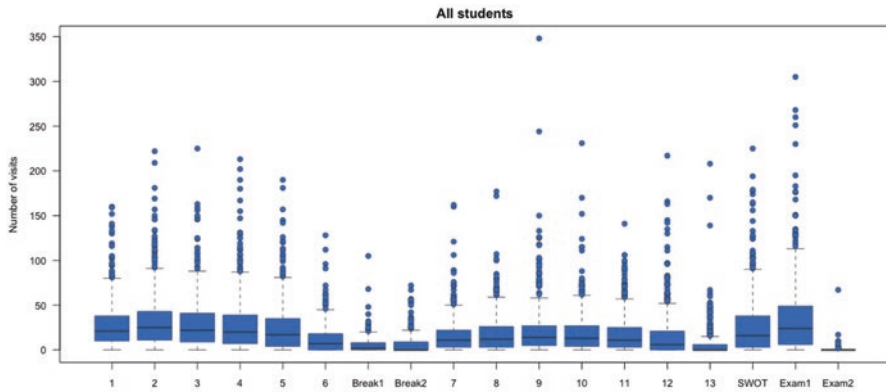


Fig. 8.16 Student engagement with course topic resources by week

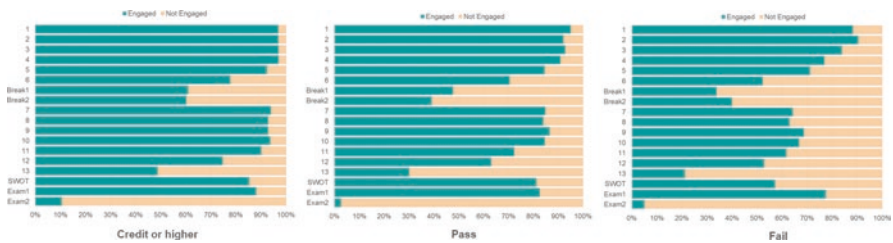


Fig. 8.17 Student engagement with course topic resources by final grade

When students’ final grades in the course were taken into account, notable differences were observed in the patterns of online engagement. Figure 8.17 shows the proportions of students engaging with topic resources broken down by final grades. Here, final grades have been divided into three categories, namely ‘Credit or higher’, ‘Pass’ and ‘Fail’. The general pattern of a high proportion of students visiting resource pages in the early weeks and weeks with scheduled summative assessment is apparent for all three grade categories. Unsurprisingly, the group with the most consistent pattern of engagement are students who eventually earned at least a Credit for the course. In contrast, students who failed the course tended to disengage from course resources as the weeks progressed, many seemingly returning right before the final exam.

This can also be seen in the weekly engagement frequencies for each grade bracket, shown in Figs. 8.18, 8.19 and 8.20. Students who received a credit or higher as their final grade had consistently higher engagement than students who received a passing grade. Moreover, students who failed the course overall had the lowest engagement with course topic resources each week.

It is also informative to consider how often students engage with weekly resources according to study mode, gender and age. Median visits to topic resources across all weeks are shown in Fig. 8.21. Key observations are as follows:

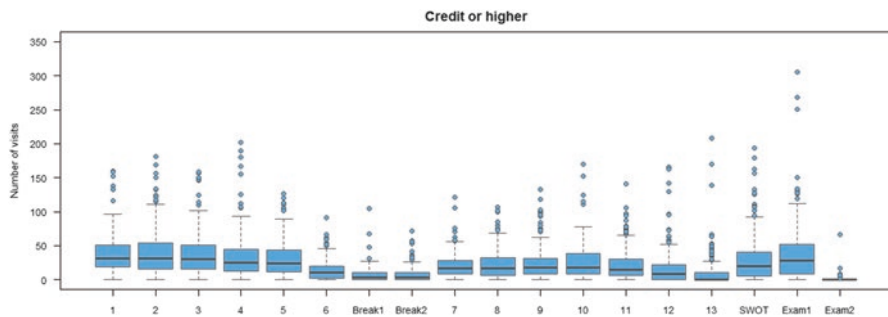


Fig. 8.18 Weekly engagement with topic resources for students who received a credit or higher as their final course grade

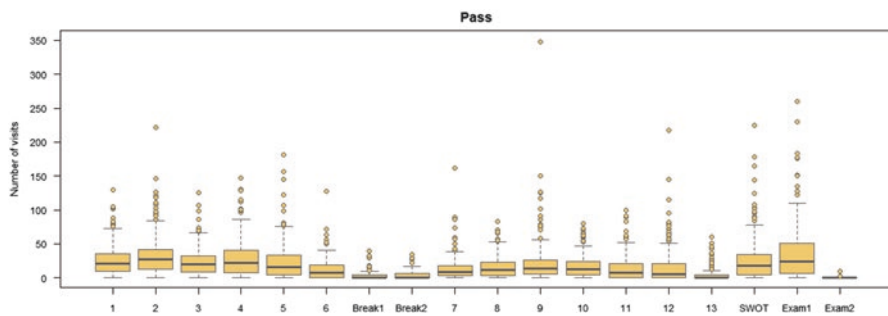


Fig. 8.19 Weekly engagement with topic resources for students who received a pass as their final course grade

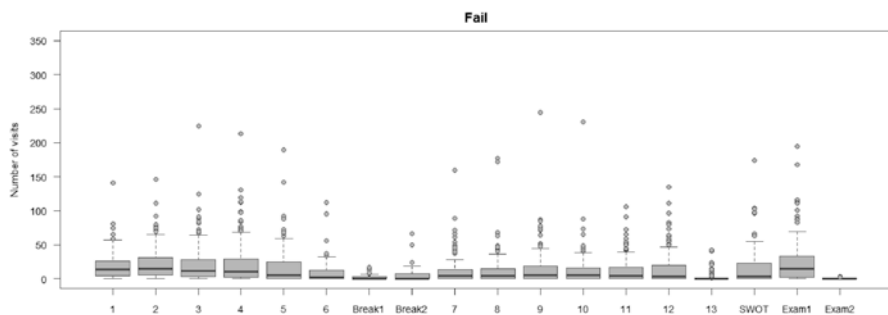


Fig. 8.20 Weekly engagement with topic resources for students who received a fail as their final course grade



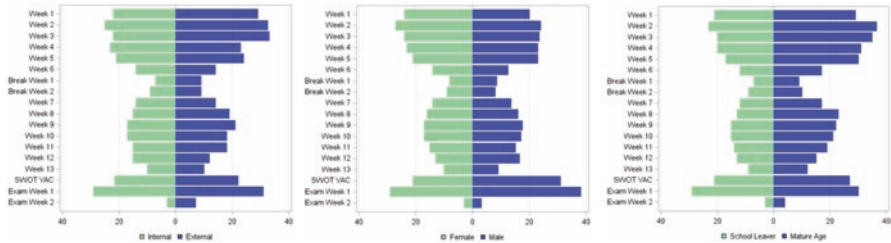


Fig. 8.21 Median visits to topic resources by study mode, gender, and age

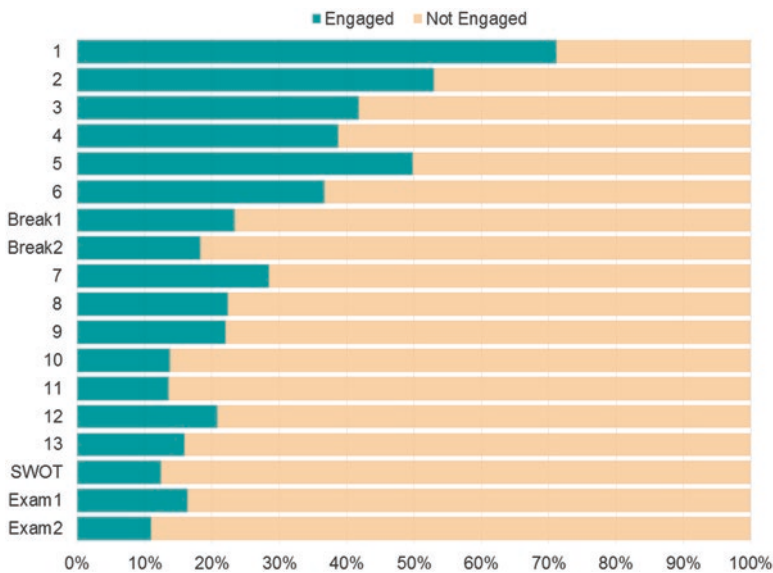


Fig. 8.22 Student engagement with social forum

- In most weeks, a typical student visited topic resources 20–30 times;
- External students tended to visit course resources more frequently than internal students, particularly in the first 3 weeks of the semester;
- Female and male students visited topic resources at the same rate in all weeks except for swotvac and the first exam week. Male students tended to be more active in those weeks (30–40 visits for a typical male student compared with 20–30 visits for a female student);
- Mature age students tended to visit the topic resources more often than school leavers, particularly during the first five teaching weeks.

Finally, we consider patterns of engagement with course social forums. Overall, forums did not appear to generate much student activity as the median number of visits per week to the forums was typically between 2 and 4. As illustrated in Fig. 8.22, the proportion of students visiting the forums was decreasing quite

dramatically from week to week, starting at approximately 75% in Week 1 and going down to about 10% in the second exam week. The week that stands out the most is Week 5 as the proportion of students visiting the forums in that week (50%) was higher than immediately before and after (40%). This increase in forum engagement may have been due to the mid-test being held in Week 5.

With regards to the number of times students accessed the forum, Fig. 8.23 also shows that the frequency of forum access decreased substantially as the semester progressed. There are a number of students however who accessed the forum an unusually large number of times each week relative to their peers. This suggests that despite generally low levels of engagement, there are some students who appear to benefit from the social forum. Overall however, Figs. 8.22 and 8.23 show that the proportion of students accessing the forum decreases over time with the students who do access the forum progressively reducing their level of engagement with this course resource over the semester. A similar pattern was observed when study mode, gender, and age were taken into account.

As in the case of visits to course resources, we also examine patterns of engagement with forums according to the final grade earned for the course, see Fig. 8.24. Again, for students who eventually failed the course, the proportion visiting forums

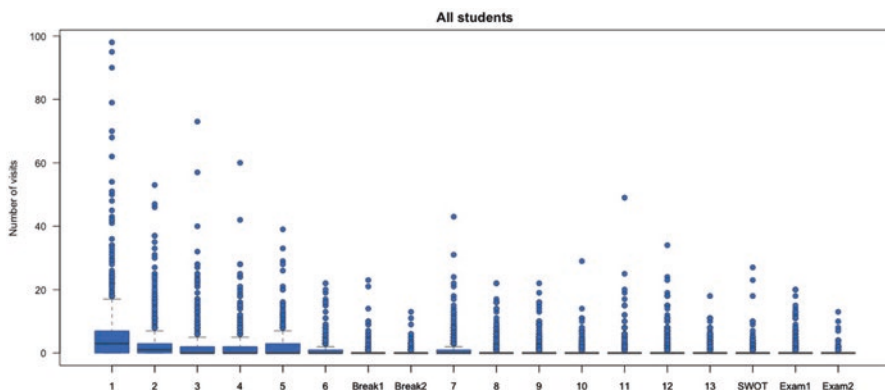


Fig. 8.23 Weekly student engagement frequency with the social forum over the course of the semester

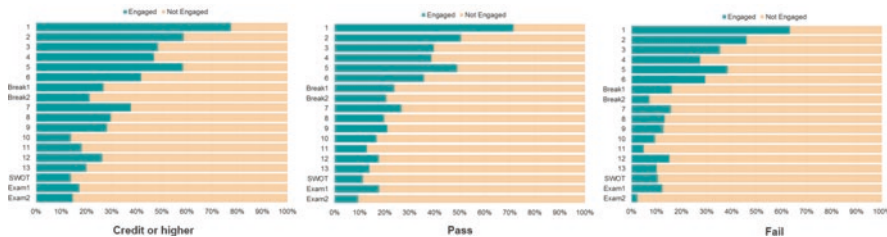


Fig. 8.24 Student engagement with social forum by final grade

was consistently much lower than for the other two grade categories, and decreasing at a faster rate from week to week following the mid-break. Similar patterns can be seen for the weekly frequency of engagement among students who did engage with the social forum each week, shown in Figs. 8.25, 8.26 and 8.27.

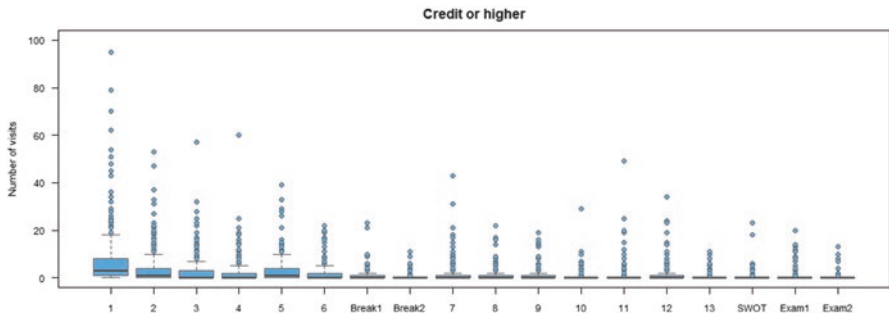


Fig. 8.25 Weekly engagement frequency with the social forum for students who received a credit or higher as their final grade

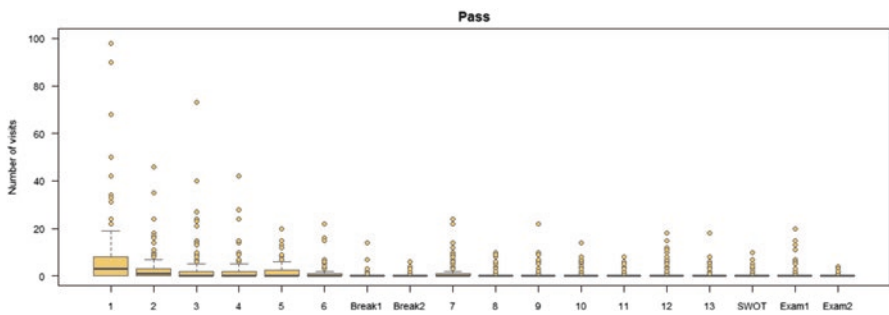


Fig. 8.26 Weekly engagement frequency with the social forum for students who received a pass as their final grade

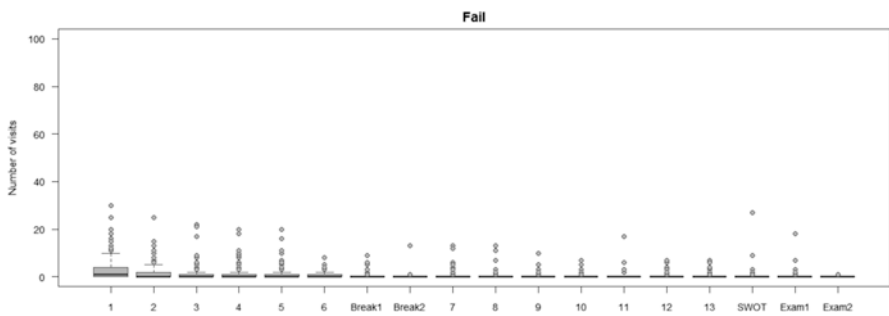


Fig. 8.27 Weekly engagement frequency with the social forum for students who received a fail as their final grade

## *Clustering Weekly Engagement Patterns*

To further investigate the patterns indicated by the initial exploratory analysis, clustering was employed as a means to uncover potential structure within the weekly engagement data. Given the available demographic information about the students, it is natural to investigate whether intrinsic groupings can be detected across the weekly engagement.

Of the available clustering methods, the majority can be divided as belonging to either the group of *hierarchical* or *partitional* clustering methods. Hierarchical clustering is preferred when there is a perceived hierarchy in the data and is preferred for smaller data sets. Partitional clustering, which assumes there are central features around which the data can be clustered in non-overlapping partitions with the number of centres defined a priori, however boasts the advantage of being computationally inexpensive.

In this study the sample size was sufficiently small to allow for the investigation of clusters using agglomerative hierarchical clustering techniques as a first step before trialling partitional clustering methods. With regards to hierarchical clustering, it is not unreasonable to expect the presence of a hierarchical structure across the weekly engagement data, as from the exploratory analysis in section “[Patterns of Student Engagement with Course Virtual Environment](#)” it is apparent that student patterns of interaction with the course site and resources varied across the semester. In order to determine the best level of interpretability of the clusters we also investigated the use of partitional clustering via the  $k$ -means algorithm as well as a hybrid hierarchical  $k$ -means clustering approach as a way to overcome the limitations of both clustering methodologies.

For the hierarchical clustering we used a linkage algorithm and explored the use of single, complete, and average linkage, as well as Ward’s distance to minimise within-cluster variance. Suitability of the clusters produced was evaluated by taking into account the cophenetic distances of each method, as well as cluster interpretability. For the partitional  $k$ -means algorithm, we assessed the number of clusters  $k$  to request a priori, by considering the output of the “elbow” method (Thorndike, 1953) combined with the silhouette method (Rousseeuw, 1987) to achieve an appropriate balance between cluster cohesion and separation. Typical values of  $k$  ranged between 2 and 4 and all values of  $k$  within the recommended ranges were trialled here. We also explored the use of a hybrid hierarchical  $k$ -means clustering method (Hasan & Duan, 2015).

We ultimately selected the hybrid hierarchical  $k$ -means approach in order to capture the expected hierarchy in weekly engagement patterns. This approach also overcomes the limitation of the  $k$ -means algorithm, namely identifying a priori the number of clusters in the data. An additional attractive advantage of retaining a hierarchical clustering approach is that a complete hierarchy of clusters are generated and as such is considered more informative than  $k$ -means (Hasan & Duan, 2015) while  $k$ -means is often preferred as it produces tighter clusters. For the hierarchical clustering method, we trialled the use of single-, complete-, and

average-linkage as well as Ward’s method to minimise within-cluster variance. Ward’s method produced the most interpretable clusters in this case, with groupings that reflected learning events across the semester and for this reason the results of using Ward’s method are presented in what follows.

Weekly engagement as measured by the number of visits to the course website is shown in Fig. 8.28, as well as visits to topic resources and the forum (Figs. 8.29 and 8.30 respectively). As the clustering method is a hybrid hierarchical *k*-means algorithm, the resulting output is the traditional dendrogram arising from hierarchical clustering with the groupings formed along the dendrograms corresponding to clusters generated by the *k*-means algorithm.

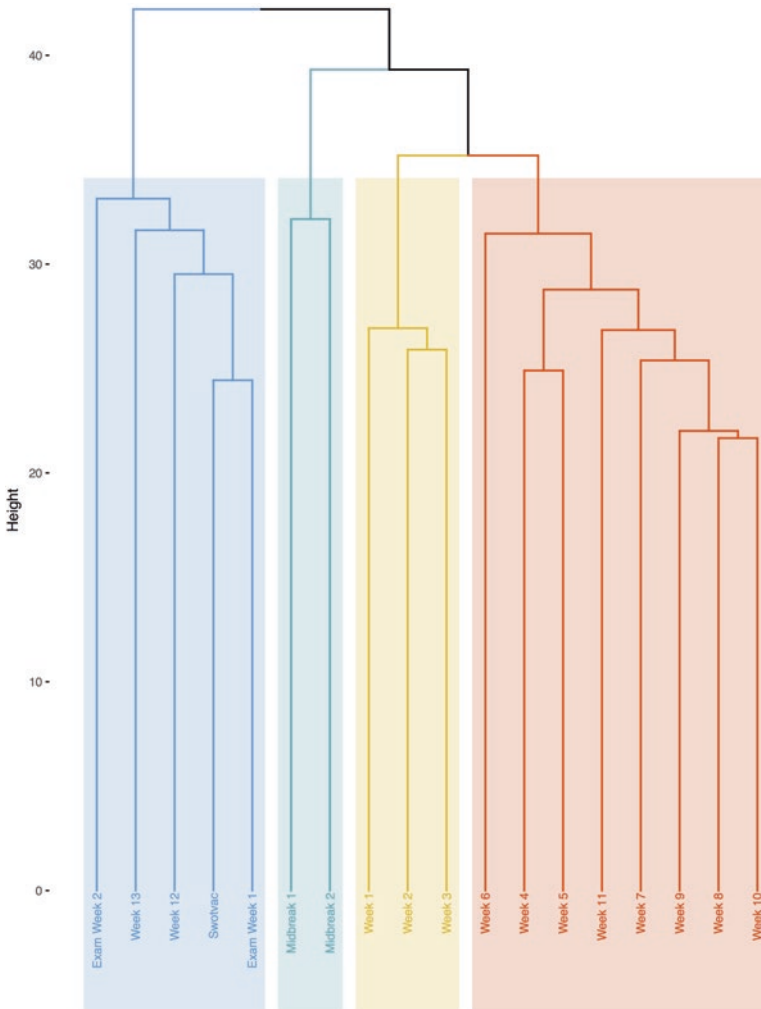
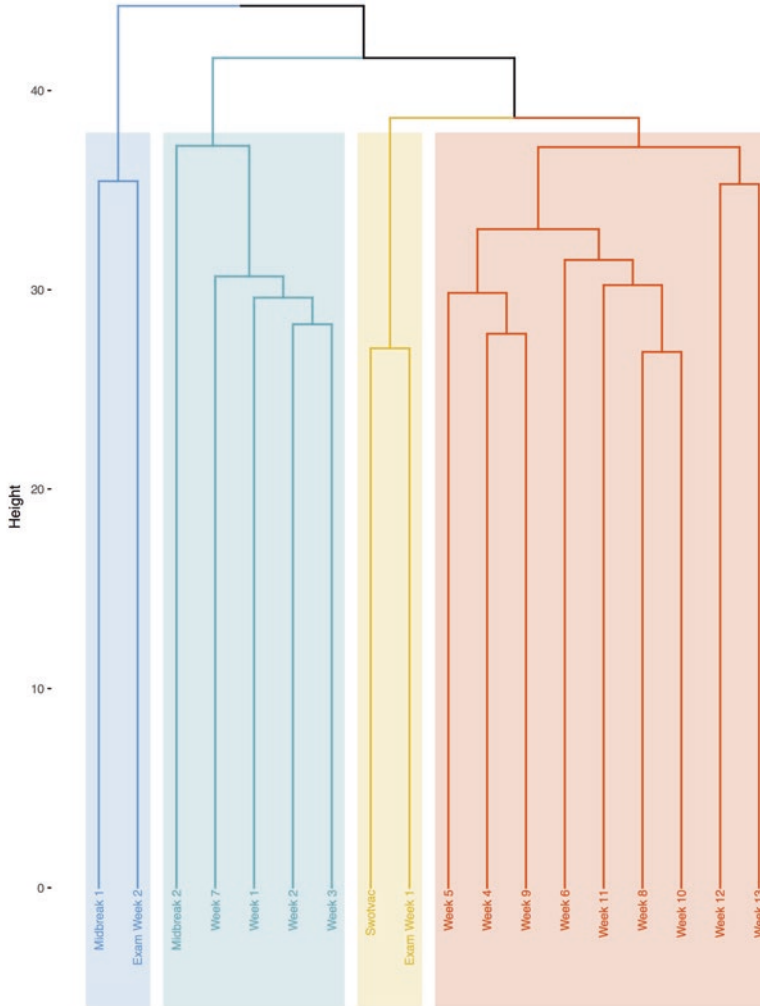


Fig. 8.28 Weekly course site visits



**Fig. 8.29** Weekly visits to topic resources

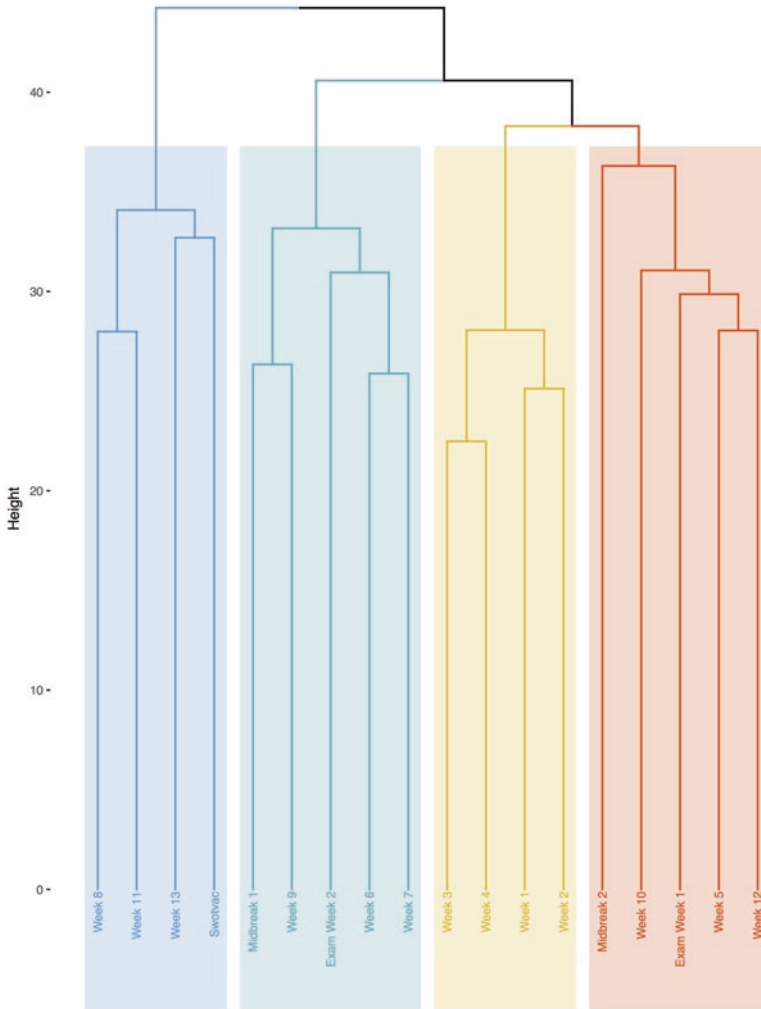
From Fig. 8.28 the pattern that emerges emulates the distinct periods a semester will naturally fall in to as well as the specific assessment structure for the selected course.

Specifically, Weeks 1–3 (second cluster from the right) form an individual component of behaviour as might be expected with students anecdotally being motivated within the first few weeks of semester; the second component (rightmost cluster) indicates the in-between behaviours that in part reflect the assessment structure of the course. Specifically, the students had a major assessment piece in Week 5 while Week 6 was a quiet week from an instructional viewpoint. From the clustering we can see Week 6 forms an unusual week in terms of classification as it is the

leftmost member of the cluster in the dendrogram which typically indicates the outlier position in a data set. Similarly, Week 11 forms an outlier as well in that there are no assessment items in that particular week. On the other hand Weeks 3 and 4, as well as Weeks 7–10 all contain online quizzes. The mid-break cluster is intuitive in its grouping as it is natural students would access the course website differently during this time as there is no instruction taking place. It is interesting to note that Week 12 falls in the fourth cluster (first cluster on the left) and collectively with the run-up to the exam period. Week 12 marks a practical exam for the students before exam review commences (Week 13) and then the final exam period. In particular, Swotvac and Exam Week 1 are closest together within this larger block which would reflect preparation for the final exam itself.

When considering visits to topic resources (Fig. 8.29) a somewhat different pattern emerges. In the leftmost cluster, we can see Midbreak 1 and Exam Week 2 have been grouped together – these are the two non-instructional/non-assessment times in the course and it is natural they will appear together in a cluster. On the other hand, it is reasonable to expect visits to topic resources during Midbreak 2 as students had an assessment item in Week 7 and it is possible they were focused on quiz preparation. In these weeks, visits to topic resources matched the pattern of visits in Weeks 1–3, while Swotvac and Exam Week 1 (second cluster to the right) formed their own grouping which would reflect final exam preparation via the online topic resources. The remainder of the weeks (rightmost cluster) behave in a similar fashion to one another. There are sub-groupings within this cluster however the overall generic visitation behaviour is collectively distinguishable from the other three clusters.

Forum visits (Fig. 8.30) tell a different story once again. The leftmost cluster contains groupings specifically in Weeks 8 and 11, then Weeks 13 and Swotvac. Without knowledge of the content of the quiz it could be possible that week 8 indicated a challenging topic, however with Week 11 it could be postulated the upcoming practical exam (Week 12) was the driver for forum interaction. The second cluster from the left appears to contain the low-activity weeks of the course from the perspective of instruction and/or assessment. The cluster second from the right seems to have captured the forum visits over the first month of the course. The standout cluster in this case is the rightmost cluster in which both Weeks 5 and 12 are grouped together and both contained a large assessment item and Exam Week 1 also joined this grouping, indicative of the upcoming final examination.



**Fig. 8.30** Weekly forum visits

### ***Clustering Weekly Engagement Patterns by Cohort Classification***

In this section we further investigate clusters for student engagement on a finer level of detail by classifying students by:

- Gender;
- Student age;
- study mode; and
- Grade category based on the final grade achieved in course.



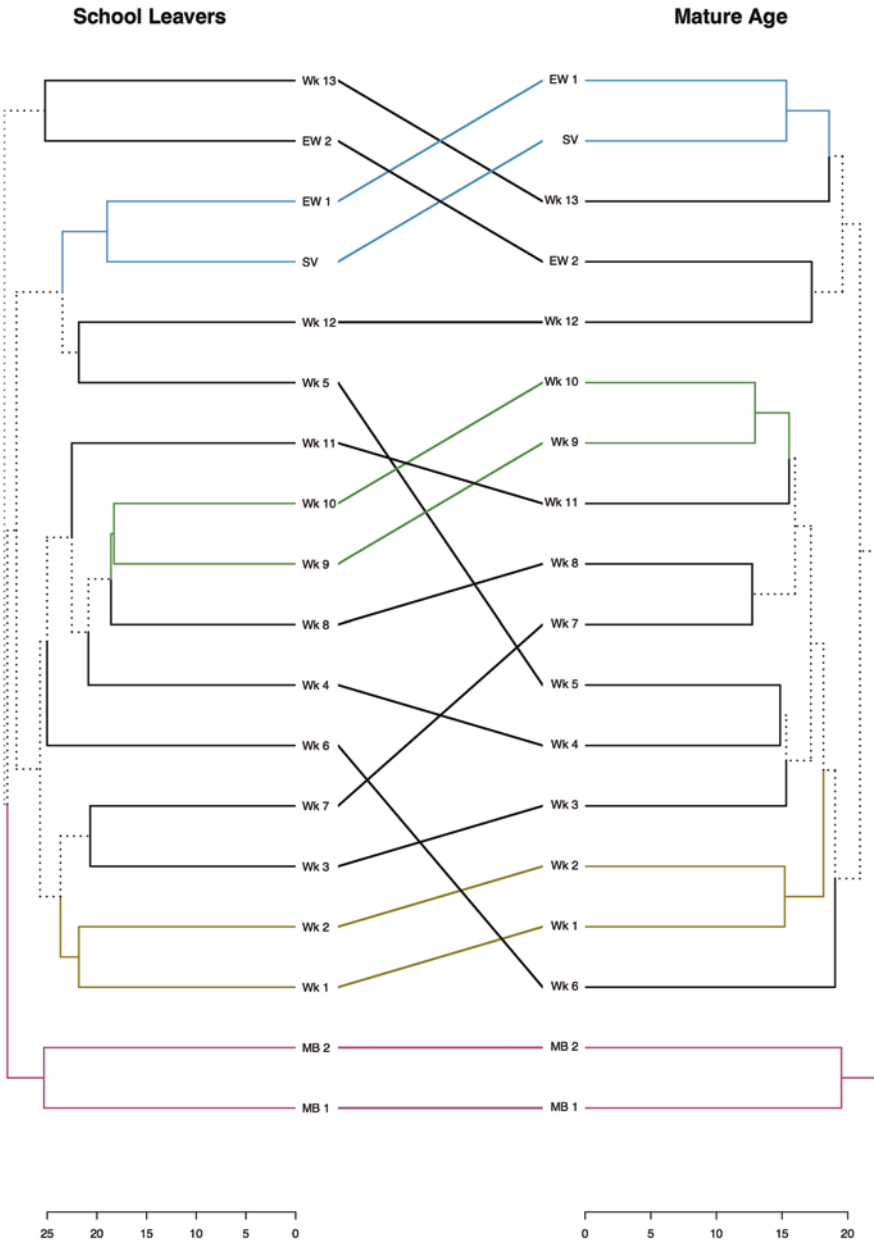
To compare clustering patterns according to these classifications we use tanglegrams capturing student engagement in terms of course site visits, topic resources visits, and forum visits. For student age, we again categorised students as being either school leavers or mature age; study mode reflected whether the student took the course internally or externally, while grade category allowed for three possibilities namely that a student achieved a final grade of credit or higher, any other type of pass or a fail grade for the course.

The tanglegram shown in Fig. 8.31 depicts the difference in course site visit behaviour when comparing school leavers and mature age students. The dashed lines on the outermost edges of the dendrograms indicate unique nodes whereas solid lines indicate common subtrees. Thus when comparing school leavers to mature age students, starting from the top of the tanglegram we can see both tanglegrams displayed similar behaviour in terms of engagement with the course site in SwotVac (SV) and Exam Week 1 (EW1) whereas behaviours differed in Week 13 (Wk 13) and Exam Week 2 (EW2) as these nodes appear in different subtrees of their respective tanglegrams. Specifically, for school leavers visits to the course site were similar in Week 13 (Wk13) and Exam Week 2 (EW2) whereas for mature age students visits to the course site in Week 13 were closer in behaviour to the SwotVac and Exam Week 1 behaviours, while Exam Week 2 was closer to the visitation behaviour in Week 12 (Wk 12). The behaviours in Weeks 9 and 10 (Wk 9 and Wk 10 respectively) are identical in both age groups, however Weeks 7 (Wk 7) and 8 (Wk 8) differ considerably between the two groups. A similar interpretation carries through to the bottom of the tanglegram where the two-week mid-break behaviours (MB1 and MB2) are identical in this case. It is particularly interesting to note the behaviour around the final exam for which school leavers behave similarly in SwotVac and Exam Week 1 whereas from the clustering it appears mature age students also include Week 13 as part of the exam preparation behaviour.

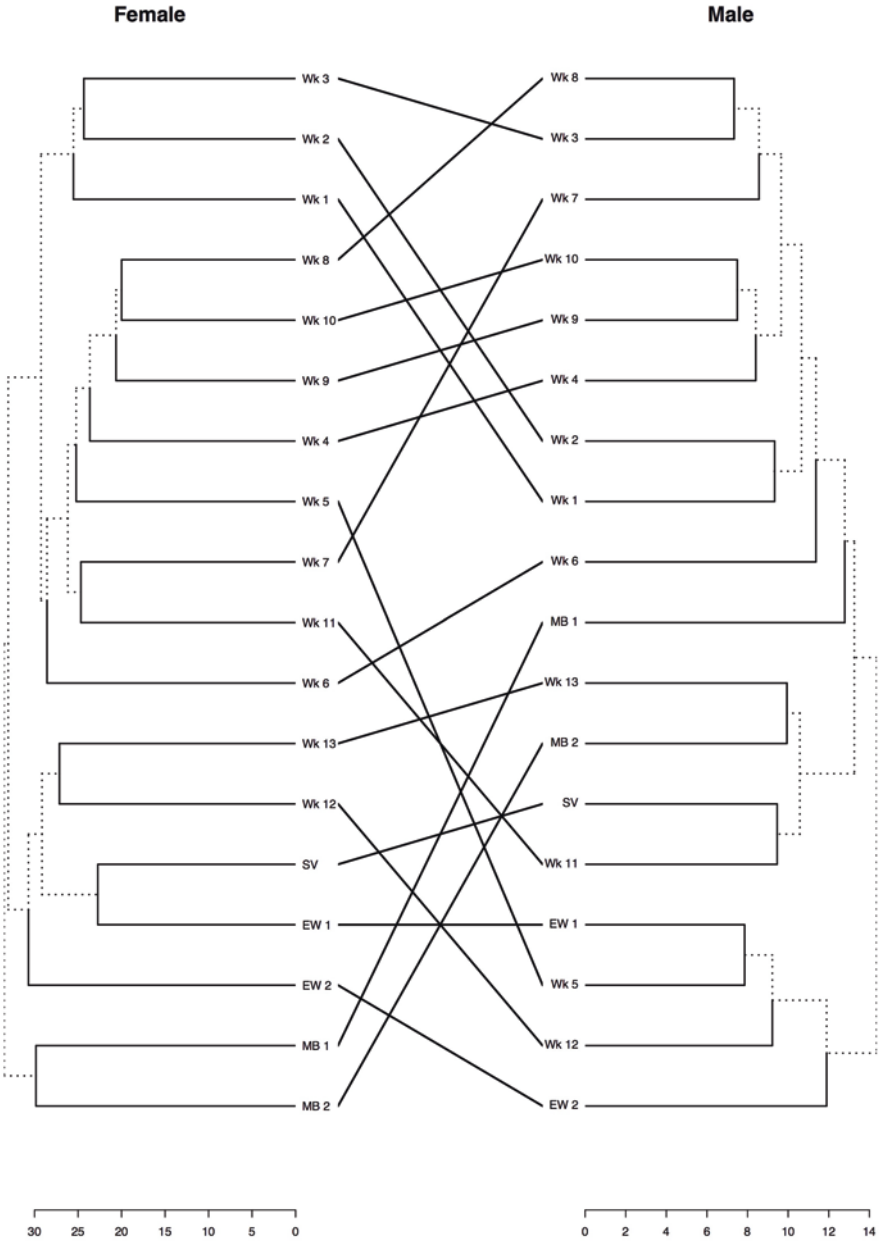
As an interesting contrast to the tanglegram in Fig. 8.31 is the tanglegram for course site visit behaviour by gender (Fig. 8.32). Recalling that dashed lines in the tanglegram indicate unique nodes, it appears that there is no commonality of behaviour in engagement, as measured via course site visitation, between males and females throughout the semester. Starting from the top of the dendrogram, females tend to treat the first 3 weeks of semester equally in terms of course visitation whereas male students treat Weeks 1 and 2 similarly but by Week 3 engagement behaviour has shifted. At the other end of the semester when comparing exam preparation via visits to the course site, female students treat SwotVac and Exam Week 1 with equal importance whereas male students treat Exam Week 1 in the same vein as Week 5 – for which there was a major assessment item — however do not include SwotVac as part of that preparation.

Again there is very little commonality when comparing students by study mode (Fig. 8.33). Course site visit behaviour tends to behave similarly in SwotVac and Exam Week 1 across both study modes. For external students however, their behaviour in Week 13 is included as part of this exam period, whereas for internal students Week 12 is treated similarly to the exam period. This could be due to the practical examination that takes place in this week. On the other hand course site visit

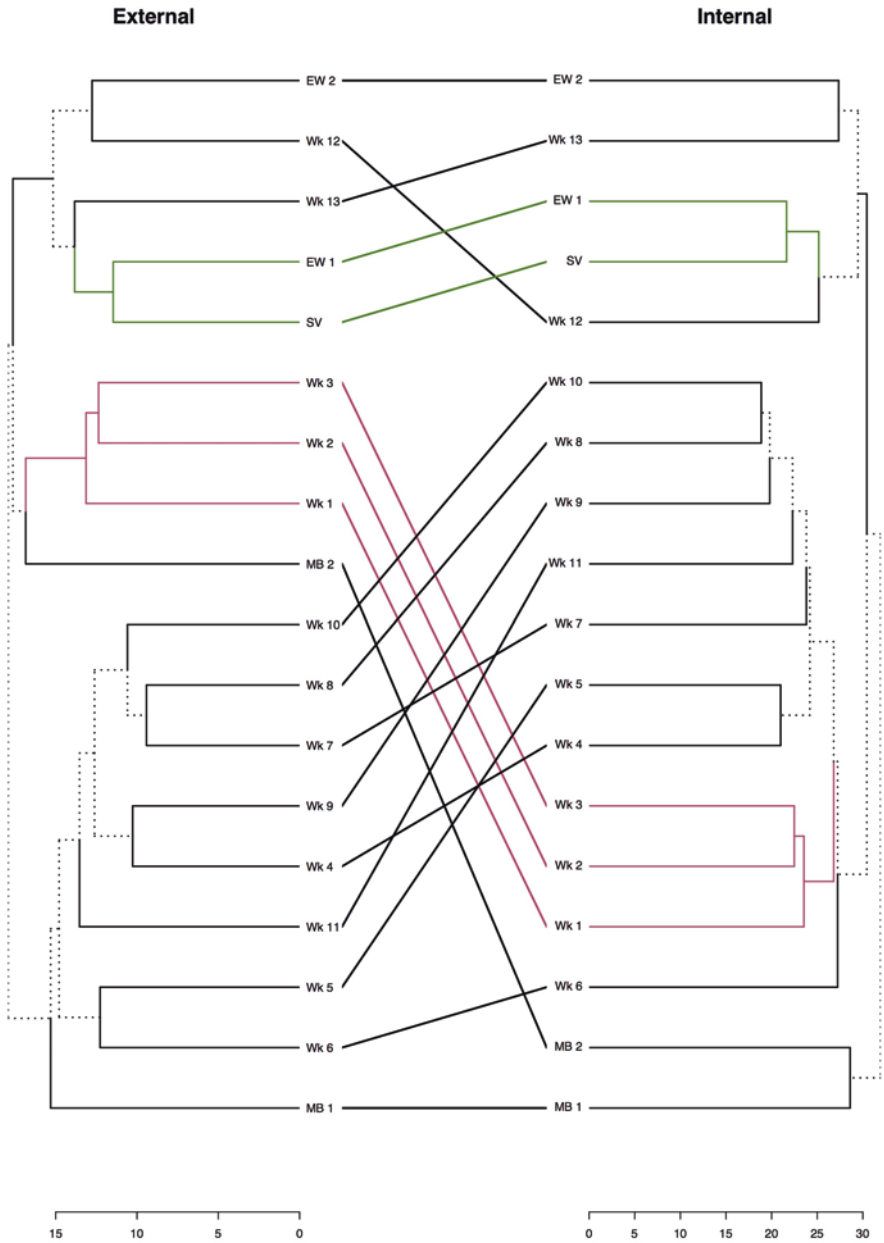
behaviour is identical in the first 3 weeks of semester (Wk1–Wk3), in Exam Week 2 (EW2) and in the first week of the mid-term break (MB1).



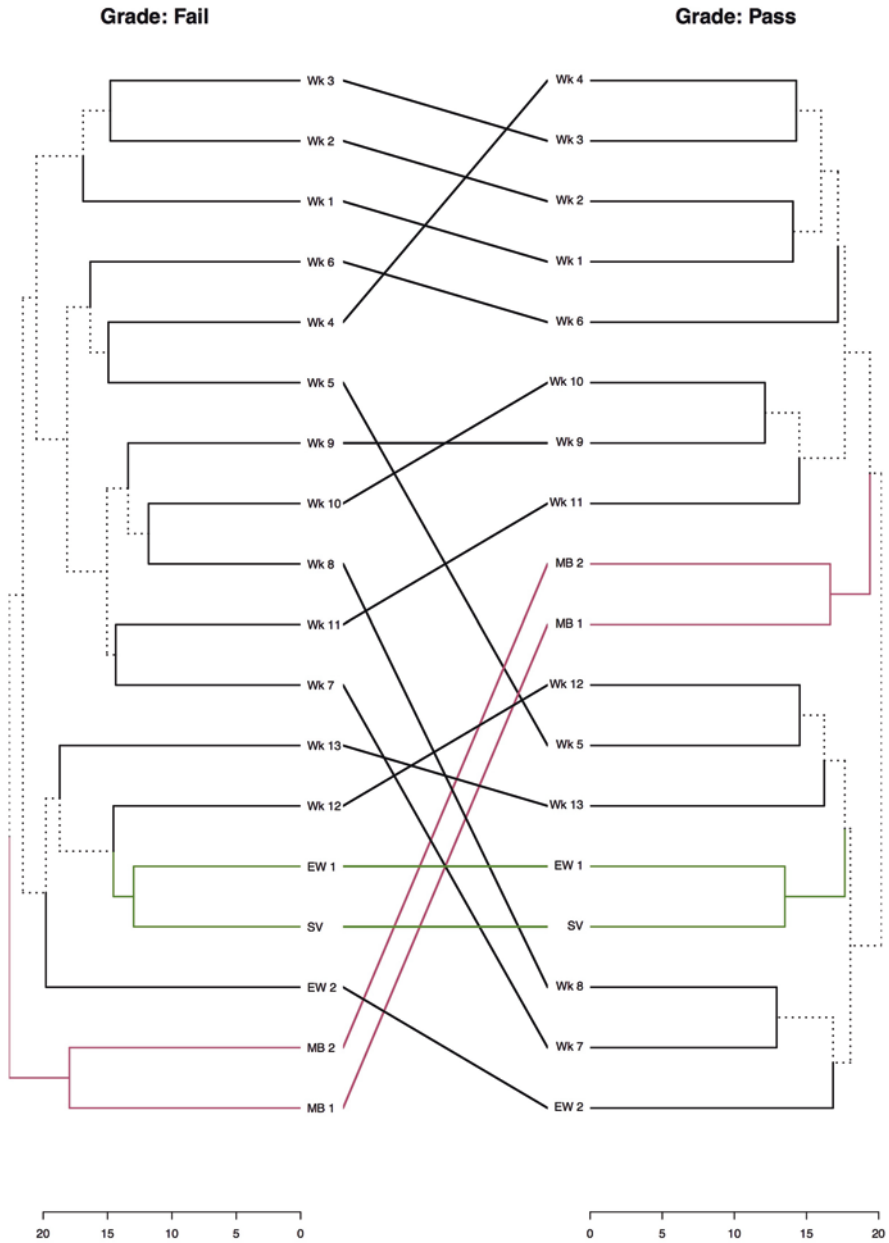
**Fig. 8.31** Tanglegram of course site visits by age. Dashed lines indicate unique nodes in the dendrogram



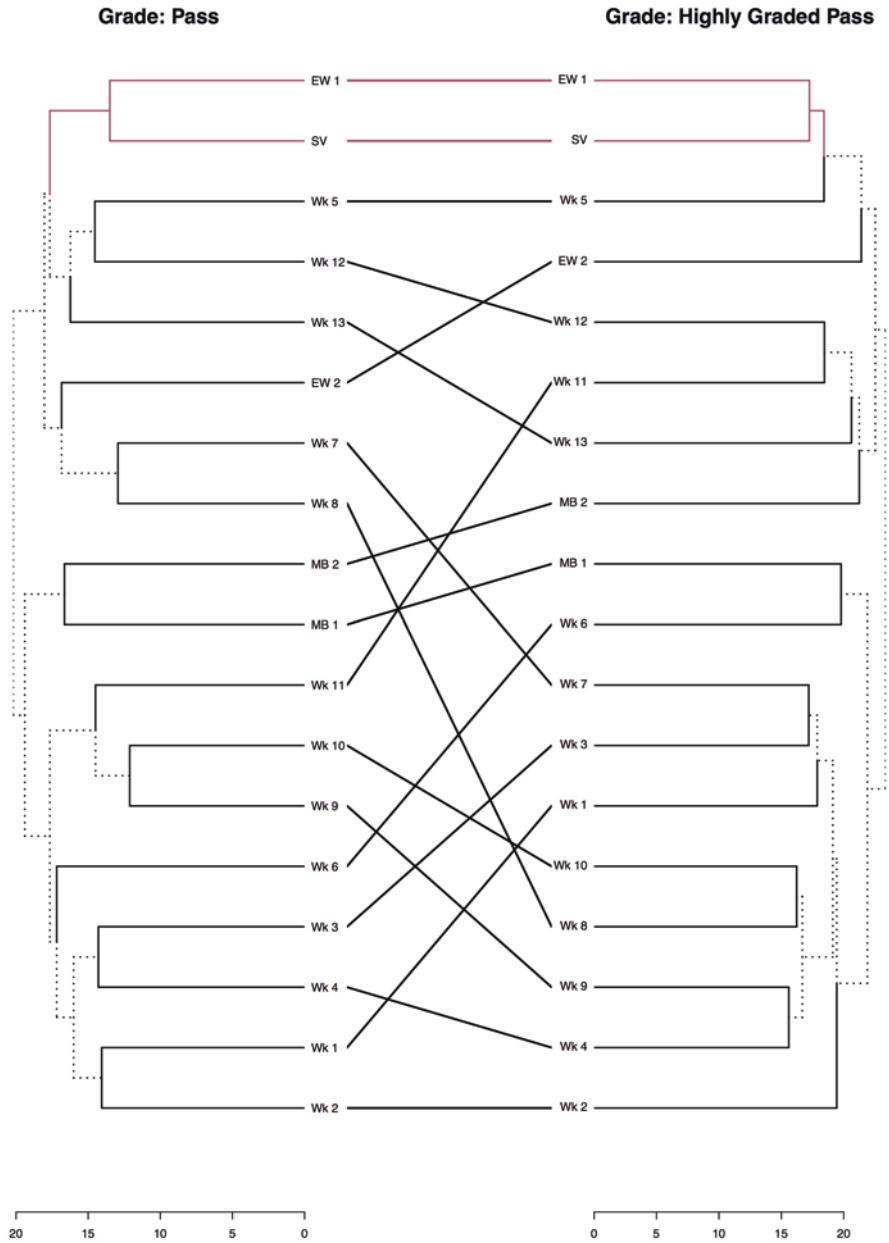
**Fig. 8.32** Tanglegram of course site visits by gender. Dashed lines indicate unique nodes in the dendrogram



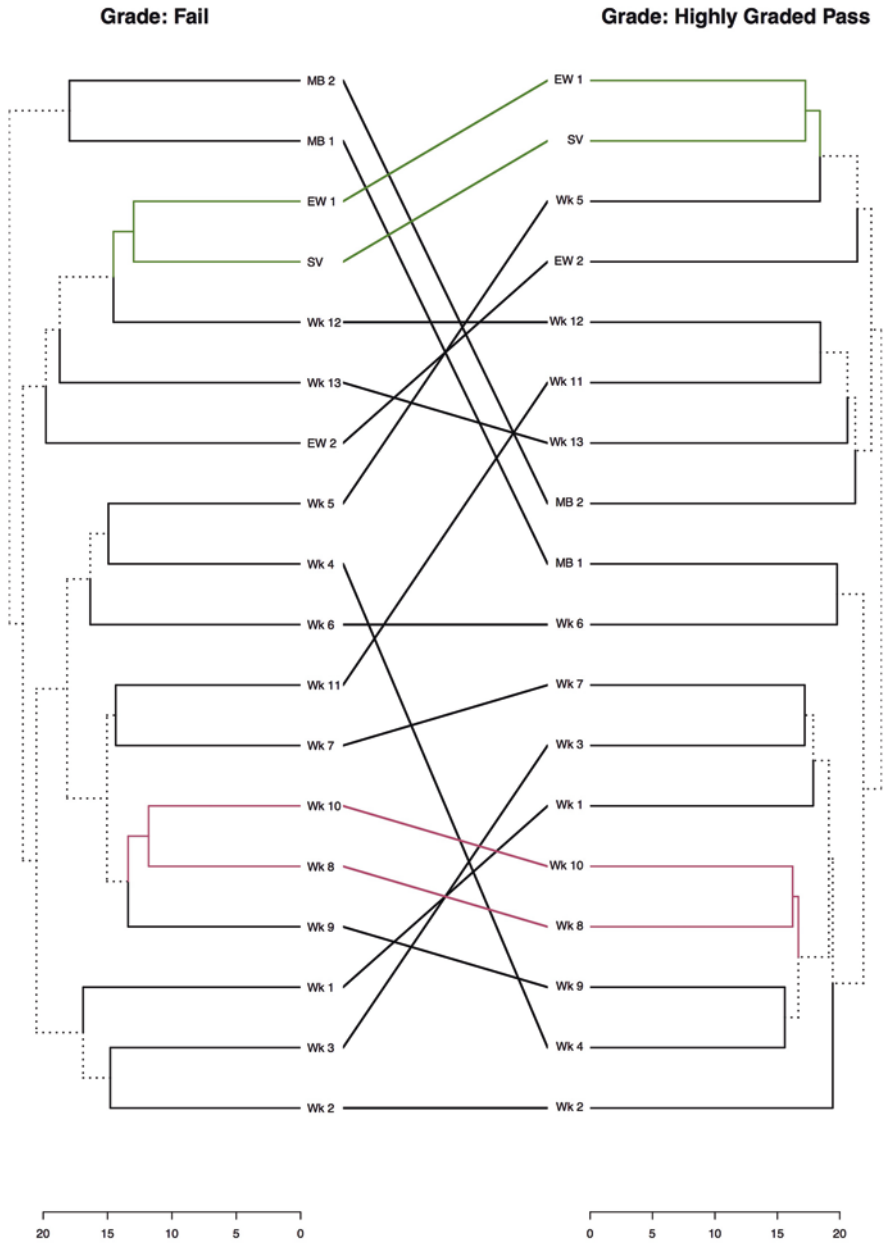
**Fig. 8.33** Tanglegram of course site visits by study mode. Dashed lines indicate unique nodes in the dendrogram



**Fig. 8.34** Tanglegram of course site visits based on students who failed versus those who passed although did not achieve a highly graded pass. Dashed lines indicate unique nodes in the dendrogram



**Fig. 8.35** Tanglegram of course site visits based on students who passed versus those who achieved a highly graded pass. Dashed lines indicate unique nodes in the dendrogram



**Fig. 8.36** Tanglegram of course site visits based on students who failed versus those who achieved a highly graded pass. Dashed lines indicate unique nodes in the dendrogram

The final set of tanglegrams are presented in Figs. 8.34, 8.35 and 8.36 and depict the differences in course site visits when considering students by the final grade achieved in the course. It is interesting to note that all three groups of students treated SwotVac and Exam Week 1 with equal importance. Students who failed or passed both treated the 2 week mid-term break in a similar fashion. However, students who achieved a highly graded pass distinguished between these 2 weeks by treating the first week of the mid-term break similarly to Week 6. Recalling there was a major assessment item in Week 5, students achieving a highly graded pass treated this week with equal importance to SwotVac and Exam Week 1. In contrast, students achieving a pass treated Week 5 similarly to Week 12 (practical examination) while students who failed did not view Week 5 as a major assessment week in that the tanglegram (Fig. 8.34) shows these students treat this week in a similar vein to Week 4, in which there was no assessment of any kind.

Comparing students who achieved a pass and those who achieved a highly graded pass (Fig. 8.35) those achieving a higher grade treated Week 12 as part of a larger block spanning Weeks 11–13 indicating preparation for the practical examination in Week 12, whereas students who passed with a lower grade displayed different behaviour in Week 11 and behaved as though Weeks 12 and 13 were similar. Students who failed compared to those with a highly graded pass treated Week 12 at a similar level of importance as SwotVac and Exam Week 1 (Fig. 8.36) however the lead-up preparation in Week 11 is missing; in this case students who failed behaved similarly in Weeks 7 and 11 – the third week of online assessment (Week 7) and the week after the final online assessment.

Investigation of engagement behaviour when viewing visits to topic resources and the forum indicated similar differences between the cohort classifiers used here. For space considerations these tanglegrams are shown here, however it is noted that these behaviours permeated across the course site irrespective of the resources utilised.

## Conclusions

In this chapter, patterns of engagement were explored for three components of the online course environment, namely the course website, weekly topic resources, and the social forum. Engagement patterns were relatively consistent across the three online components, with the exception of the social forum, where engagement levels declined substantially over the course of the semester.

Across each course component, there is a selection of students with unusually high levels of weekly engagement. Students appear to engage most frequently during the early semester teaching weeks and the weeks leading up to the examination period. Internal and external students differ in their patterns of engagement, with external students consistently engaging at a higher frequency than their internal counterparts.



It was also seen that male and female students do not appear to differ substantially in their engagement behaviour with the exception that a higher proportion of male external students tend to disengage from the course earlier in the semester than their internal counterparts. Mature age students were found to typically have a higher frequency of access for course resources than school leaver students, with greater variability in their engagement counts per week.

When considering engagement patterns by final course grade, it was found that students who receive a credit or higher are engaging with the online course resources at a much greater frequency than students who receive a passing grade. Students who failed the course overall had considerably lower levels of engagement throughout the semester. Another distinction was that students achieving a credit or higher also viewed earlier major assessment items in the same vein as the final exam at the end of the exam period.

With regards to study mode, it was determined that for internally enrolled students and externally enrolled female students, the weeks with the greatest engagement are Weeks 1–5 and the first week of the exam period. For external male students however, the key teaching weeks were Weeks 1 and 3 and the late-semester weeks for SwotVac and the examination period.

When considering the engagement trajectories of individuals over the course of the semester, it was found that high-engaging students tend to remain highly engaged whereas students in the lowest quartile of engagement tend to maintain relatively low engagement each week. Moreover, internal students are more likely to have lower online engagement levels than their external counterparts.

What we are able to learn from learning analytics appears to depend crucially on the structure of a virtual learning environment as well as decisions made by online engagement data custodians. There is a potential disconnect between what data is available and what data is useful, making learning from learning analytics challenging. For the course on which analyses in this chapter are based, the main difficulty lay with the interpretation of weekly counts which were obtained from the virtual learning environment access logs. Without more detailed information of how student activities contributed to these counts, we were able to only gain some insights in relation to student engagement. That being said, based on our analysis, we can observe that external and mature age students appear to engage with virtual learning environments differently. A higher proportion of external students tend to disengage from the online resources, particularly if they are male and school leavers. On the other hand, once they engage, mature age students appear to maintain strong online presence throughout the semester. Our analysis also leads us to conclude that assessment, both formative and summative, is one of the deciding factors whether a first year student engages with online course resources. This suggests the need for differentiation of online support resources and advice given to students on what contributes to study success.

## References

- Beer, C., Clark, K., & Jones, D. (2010). Indicators of engagement. In C. Steel, M. Keppell, P. Gerbic, & S. Housego (Eds.), *Curriculum, technology & transformation for an unknown future. Proceedings ASCILITE Sydney 2010* (p. 75–76).
- Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. In *Proceedings of the fourth international conference on learning analytics and knowledge LAK'14* (pp. 83–92). ACM.
- Ferguson, R., & Clow, D. (2015). Examining engagement: Analysing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the fifth international conference on learning analytics and knowledge LAK'15* (pp. 51–58). ACM.
- Hasan, M., & Duan, Z.-H. (2015). Hierarchical k-means: A hybrid clustering algorithm and its application to study gene expression in lung adenocarcinoma. In H. Arabnia & Q. Tran (Eds.), *Emerging trends in computational biology, bioinformatics, and systems biology* (pp. 51–67). Elsevier.
- Henrie, C., Bodily, R., Manwaring, K., & Graham, C. (2015). Exploring intensive longitudinal measures of student engagement in blended learning. *The International Review of Research in Open and Distance Learning*, 16(3), 131–155.
- Hughes, G., & Dobbins, C. (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). *Research and Practice in Technology Enhanced Learning*, 10(10). <https://doi.org/10.1186/s41039-015-0007-z>
- Hung, J.-L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*, 4(4), 426–437.
- Kizilcec, R., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge LAK'13* (pp. 170–179). ACM.
- Phan, T., McNeil, S., & Robin, B. (2015). Students' patterns of engagement and course performance in a massive open online course. *Computers & Education*, 95, 36–44. <https://doi.org/10.1016/j.compedu-2015.11.015>
- Ramesh, A., Goldwasser, D., Huang, B., Daume, H., III, & Getoor, L. (2014). Learning latent engagement patterns of students in online courses. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence* (pp. 1272–1278). AAAI Press.
- Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Slater, N., Peasgood, A., & Mullan, J. (2016). *Learning analytics in higher education. A review of UK and international practice*. Retrieved from <https://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v3.pdf>
- Siemens, G., & Gasevic, D. (2012). Guest editorial – Learning and knowledge analytics. *Educational Technology & Society*, 15(3), 1–2.
- Thorndike, R. (1953). Who belongs in the family? *Psychometrika*, 18(4), 267–276.

**Belinda A. Chiera** is a senior lecturer in Statistics and the Program Director for Service Teaching at the School of IT & Mathematical Sciences, University of South Australia. Belinda has a wide and varied range of interests including social network analysis, modelling of high-dimensional networks and enjoys uncovering insights hidden in large databases. She also works on a number of projects with a particular focus on the student wellbeing and data analytics in support of student learning and academic success.

**Malgorzata W. Korolkiewicz** is a senior lecturer in Mathematical Sciences and the Program Director for Data Science at the School of IT & Mathematical Sciences, University of South

Australia. Her research interests include mathematical finance, risk management, environmental modelling, as well as data analytics and challenges posed by Big Data. She is particularly interested in developing students' statistical thinking and providing them with the skills needed to derive insights from data. She is a high-quality university teacher who in 2011 was awarded the UniSA Award for Teaching Excellence.

**Lisa J. Schultz** is a Research Associate in the Centre for Industrial and Applied Mathematics at the University of South Australia. She was awarded a PhD in mathematics and statistics in 2016 and has a broad range of research interests, including environmental modelling, student health and well-being, outlier analysis, and data analytics. She has a keen interest in uncovering the hidden stories in large datasets through statistical analysis and is a member of the Society for Learning Analytics Research as well as the International Association for University Health and Wellbeing.

**Part IV**  
**Educational Systems That Use Big Data**

# Chapter 9

## Museum Big Data: Perceptions and Practices



Georgios Papaioannou

**Abstract** This chapter reviews Big Data in museums, considering both theoretical perceptions and relevant practices in current Big Data projects in museums. Big Data, data analytics and data mining processes comprise relatively new fields of research and inquiry. They are expected to become core disciplines as data input increases and digital data collections from museums' everyday activities and operations become larger and larger. The data include rich elaborated databases of museum objects and services, data related to museum visits, and the deluge of visitor-generated data on the internet and social media. Big Data in museums focus on developing methods and techniques to identify new and uncover hidden, important and useful information, patterns, clusters and relationships among large sets of museum data. After reviewing the notion and characteristics of Big Data, we present connections between Big Data in museums and Big Data in libraries, arts, and humanities. We also offer several case studies from museums operating Big Data projects, which illustrate problems, prerequisites and potentials.

**Keywords** Big Data · Museums · Data analytics · Data mining · Libraries · Arts · Humanities

### Introduction

The twenty-first century is the century of data explosion. Although the data are not new and unfamiliar to society, what is new in this emerging era and area is their exceptional excess. Technology and digital evolution led to new smart products and devices capable of capturing, controlling, creating, and disseminating a massive amount of information with the help of embedded network sensors. Communication, as well as knowledge sharing, became more and more digitized through social

---

G. Papaioannou (✉)  
University College London in Qatar, Doha, Qatar  
e-mail: [gpapaioa@ionio.gr](mailto:gpapaioa@ionio.gr); [g.papaioannou@ucl.ac.uk](mailto:g.papaioannou@ucl.ac.uk)

© Springer Nature Switzerland AG 2021  
T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy Implications of Research in Education 13,  
[https://doi.org/10.1007/978-3-030-76841-6\\_9](https://doi.org/10.1007/978-3-030-76841-6_9)

201

media, smart devices, and products, which led to the appearance of the so-called cyber society. As a result, the notion of Big Data comes to describe a new generation of data characterized by extended volume, velocity, and variety. Big Data, together with analytics and data mining, are valuable tools and drivers for improving products, understanding the behaviour of clients and consumers, developing new products and services, data visualisation, and lately the Internet of Things (Dey et al., 2018, Krishna Prasad, 2018; Prodrinou, 2017). In some companies such as Google, Amazon, and Facebook, extensive datasets are already used to navigate competitive policy and strategy.

Within this context, the cultural sector, including museums, faces a time of exploration and updating. Museums are multi-character cultural institutions with activities and operations ranging from information and conservation/preservation to dissemination, education, and fun. Museums produce data on and with their museum objects (cataloguing, curating, conserving, digitising), their operational activities (daily functions, communications, management, research, marketing, other operations), and their visitors (museum visits, other activities, educational activities, visitors' involvement/input and feedback, and social media). In the following sections, we explore the meaning of Big Data and their links to museums and specific projects related to museums and Big Data. We aim to describe the presence of Big Data in the world of museums.

## **Big Data: Definitions and Thoughts**

Big Data implies so much data that it can't be handled traditionally, whatever 'traditionally' may mean in the context of an organisation. In the last decade, and especially the last 5 years, many attempts to define Big Data have appeared in the literature (Beyer & Laney, 2012; De Mauro et al., 2015; Zikopoulos et al., 2012; Mayer-Schönberger & Cukier, 2013), especially since Big Data models and techniques have been put into practice to help decision-making processes.

In the twenty-first century, the enormous amount of electronic data produced, the advancement of technology and the ever-increasing involvement of people with computers have created the need for increased speed of collection, processing, and distribution of electronic digital data. A simple and standard set of characteristics used to describe Big Data is the 3Vs scheme: Volume, Velocity, Variety (McAfee & Brynjolfsson, 2012). This scheme refers to the continuously increasing speed, scale, and diversity in data and data sources. Big Data's sources are widespread, cover a wide range of electronic data production and imaging, and are extensively used and produced by people for their daily information needs (GPS, social networking, electronic communication, etc.). They are also characterized by flexibility as well as speed in terms of production and dissemination. To use Reinhalter and Wittmann's example (2014), Google Flu Trends and their Big Data website have the ability to provide information about flu-related illnesses much faster than the disease control centers, such as the U.S. Centers for Disease Control and Prevention (CDC). As

Gobble (2013) reported, 90% of the world's data were produced in between the years 2010 and 2012, which is partly due to the increasing number of people owning and/or using personal computers and smartphones.

To reach a definition of Big Data and an understanding of research issues associated with it, De Mauro et al. (2015) compared fifteen (15) proposed definitions of Big Data in published research work. They also attempted a synthesis to arrive in interesting and summative conclusions. Starting from the 3Vs, they pointed out the informational characteristics of Big Data. To make use of Big Data, they described a need for targeted technologies and appropriate analytical methods related to data analytics and data mining. The outcomes of putting the above into practice, refer to the creation of value (economic, social and/or other), which is Big Data's impact on companies, institutions, and society. Here is the overall De Mauro et al. (2015) Big Data definition in 27 words: "*Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value*". This definition covers the four main themes associated with Big Data: Information, Technologies, Methods, and Impact.

To elaborate on the Information theme further, we need to stress the extensive digitisation processes and projects in the last decades, which has resulted in the production and availability of immense amounts of digital data. Digitization gained great popularity. Mass digitization projects converted conventional physical data into digital collections. Libraries created digital copies of their books, archives digitized their documents, and museums introduced databases of digital images, text, even 3D models of museum objects.

Intense digitisation has led to a life full of (or even dependent on) data and technology, a 'datafied' life (van Dijck, 2014). Big data and datafication involve processes of data management that reveal patterns unseen (or difficult to see) in analog or digital format. New software, new devices (sensors, advanced computers, mobile phones, RFID tags, etc.), the fast internet, the development of Social Media sharing and commenting, and lately the Internet of Things (Atzori et al., 2010; Evans, 2011), comprise parts of the Big Data universe.

Without technology supporting their potentials, Big Data collections are just static collections. Increased memory storage in multiple interconnected machines and strong computational performance are required to support Big Data endeavours and complex processing operations. It is the data distribution, handling, transmission, and querying that demand particular technological advances, such as the open-source Big Data software Apache Hadoop and its components: Common, HDFS, YARN, and MapReduce. Hadoop is considered a significant point in the development of Big Data technologies (Ghemawat et al., 2003; Xiong et al., 2013). Also, the ability to store more data on smaller devices is a practice that requires ongoing and costly research, given the velocity of data growth.

There is a list of available methods, most of them deriving from statistical methods, to be used for Big Data (Manyika et al., 2011). Methods of processing beyond traditional statistical techniques are needed, as well as trained professionals with appropriate skills and understanding.

The Impact of Big Data in our lives covers a wide range of subjects and phenomena. One may wonder whether our age can be seen as the “Age of Big Data”, to use Lohr’s words (2012), and whether Big Data have a “human face”, to cite Smolan and Erwitte (2012). Big Data can certainly be used beyond trading, advertising, and internet marketing in many practices beneficial to humanity. On the positive side, Big Data have a socially positive impact, as they can produce patterns and predict infections, outbreaks, epidemics, unemployment, and inflation (De Mauro et al., 2015, p. 100). Other examples include data analytics from large urban centers on pollution, waste, traffic management, energy and water consumption, and even the measurement of crime. On the level of the individual, Big Data can be used on a computer or smartphone for a number of uses, such as creating a diet plan, exercise, and generally improving the quality of human life by self-tracking and monitoring health and schedules (Jacobs, 2012). On the negative spectrum, issues of privacy, abuse of power, unfair competition, and conversion of data into a competitive advantage must be addressed. In the cultural sector, Big Data have started to enter the world of the arts and humanities, as well as libraries, archives and museums.

## **Big Data in Museums, Libraries, Arts, and Humanities: Connections and Characteristics**

In the field of museums, as in the field of arts and humanities, the notion of Big Data is still in an early stage. When referring to Big Data in these fields, people think of mass digitisation processes and large datasets. People may even be surprised because museums, libraries, arts, and humanities used to be considered solidly on one side of the old bipolarity of humanities versus science. Today, in the era of Big Data and datafication, this is not the case anymore. Big Data and Big Data techniques in arts, humanities, and museums can offer new perspectives, potentials, and paths towards new and upgraded ways for operation, value, and creation.

The link between Big Data and the arts and humanities is bi-directional. On the one hand, digital evolution and Big Data have an impact on arts and humanities in their institutional management and performance assessment. On the other hand, these digital transformations need to abide by the long-established mechanisms in which arts and humanities work. Also, the emphasis on the aesthetic dimensions of practices in museums, arts, and humanities have to be respected.

The connection between museums, arts, and humanities and Big Data has specific challenges to address. Big Data related to arts and cultural organizations offering means of quantitative and qualitative assessment and data-driven decision making to deal with their multi-faceted challenges and concerns such as audience management and development. Therefore, alliances and collaborations between arts and humanities and computer science need to be established with a universal language and communication code. This link between Big Data and arts and



humanities is not just a computational issue. It is also about how Big Data are used to construct organizational, cultural, and contextual narratives.

An example comes from modern libraries, especially academic ones, which need to produce Data Management Plans to attract funding. Librarians today are asked to create and provide data management services so that funding is feasible. Besides, libraries have the obligation to host and maintain data produced by users, students, and researchers. Also, libraries need to be familiar with digital data editing and dissemination processes, to know external sources, and to understand and provide guidance concerning intellectual property and copyright (Gordon-Murnane, 2012; Heidon, 2011). Big Data are expected to transform operations and services in data repositories within libraries. Also, to initiate new and relevant policies in data management and preservation (Schwartz, 2013), and to provide career opportunities to librarians and information scientists (Huwe, 2012; Reinhalter & Wittmann, 2014).

Museums, libraries, and archives share common aims and objectives as memory institutions. We can, thus, assume that the above relates to Big Data for museums as well, and that, therefore, museums, the arts, and humanities comprise areas where Big Data can play an important role. There is a need for interfaces between museums, libraries, archives, arts/humanities, and computer science so that there is a common path to digital transformation. Relevant publications have started appearing, including Milosch et al. (2018) and Gilmore et al. (2018) in the volume edited by Schiuma and Carlucci *Big Data in the Arts and Humanities: Theory and Practice* (2018).

## Museum Big Data Projects: Some Examples

We discuss here individual museum projects that relate to Big Data, aiming to show directions that the museum world has taken in the last few years. We have chosen projects for which some published information is available. These are the *Pen Project* (New York USA), the *Culture Counts/Culture Metric Project* (Victoria, Australia and Manchester, UK), the *DMA Friends Project* (Dallas, USA), the *storytelling API* (Kansas, USA), the MET API (New York, USA) and the *MBD Qatar Project* (Doha, Qatar).

### ***The Pen Project (Cooper-Hewitt, Smithsonian Design Museum, New York, USA)***

In 2011, the curatorial team of the Cooper-Hewitt, Smithsonian Design Museum had the opportunity to rethink the museum during a complete renovation and extension of the building. They were able to imagine the museum they wanted (Chan, 2015a). The main idea was to enlarge their public beyond the usual professional

designers already coming to the museum (Wright, 2017, p. 122). They launched a few projects focused on interactivity with the help of digital technologies.

Towards achieving these goals, a Digital & Emerging Media department was created with Sebastian Chan hired as director, later replaced by Micah Walter. They expanded the team and changed their name to Cooper Hewitt Lab while broadening the museum's digital presence (Chan, 2015a). They developed a Collection Browser to give access to online collections through an Application Programming Interface (API). The API was thought of as the core of all digital projects (Chan, 2014). Secondly, they worked with the company Diller, Scofidio + Renfro (DS + R), recruited as exhibition architects, and a company called "Local Projects" as media designers.

'The Pen' itself is a digital stylus given to each visitor and paired with their entrance ticket. It contains a RFID tag (Radio Frequency Identification) reacting to NFC tags (Near Field Communication) placed on label rails. It will then allow the visitor to "collect" objects putting the pen in contact with the label. The objects are memorised through the API. The visitor can retrieve the collected objects inside the museum on giant tablets and gain access to metadata. It will also allow visitors to play around with the form and material of the object, and draw new designs. These designs can then be used to create a wallpaper in a dedicated room (the so-called *Immersion Room*), where the designs are projected on the surrounding walls. Visitors will be able to retrieve and print the collected objects and designs via a URL they receive at the ticket entrance (Chan, 2015a; Cooper Hewitt et al., 2014; Dale, 2016; Ferro, 2014; Walter, 2015, 2016a; Wright, 2017, pp. 121–124).

'The Pen' was issued by DS + R and realised by the company 'Sistelnetworks'. Bloomberg Philanthropies have covered the entire development and implementation costs through their Bloomberg Connects program (Chan, 2015a).

'The Pen' was launched in March 2015. Five months later, the Cooper-Hewitt Lab published their first results on their blog, showing very high participation: 94.05% of visitors used 'the Pen'. This percentage is significantly higher than other experiments done in the past (Chan, 2015b). One year later, 'the Pen' experience was still consistent. They deployed the same technology in their temporary exhibition *Beauty-Cooper Hewitt Design Triennial*, which created a peak in collecting objects with 'the Pen' (Walter, 2016b).

What made the success of 'the Pen'? The opportunity to look at the entire museum and rethink, together with the funding by Bloomberg Philanthropies, offered ideal conditions and gave a rare chance to the museum to reinvent. Moreover, the project secured its development by having a dedicated team inside the museum. It allowed bug corrections, structure modifications (if required), and constant experiments. Actions were published in the museum's blog and in research papers (e.g., Walter, 2015, 2016a). They are also sharing data with an open-source spirit for non-profit researchers and developers, data they do not have the time to exploit (Walter 2016b, c). We expect to see research done with these Big Data shortly.

## ***Culture Counts and Culture Metric (Victoria, Australia & Manchester, UK)***

Culture counts is a digital application and web portal system that collects and measures data, providing libraries and arts organisations with standardised metrics submitted to ‘the three peers.’ i.e., professionals of the sector, self-evaluation, and the public. The project has been funded by the Department of Culture and the Arts (DCA) of Western Australia (WA) and by the National Portfolio of the Arts Council England (ACE). The development of the metrics first occurred in 2010 in Western Australia, co-produced by the DCA and the cultural community. Based on the work of John Holden on Cultural Value (Holden, 2006), their aim (Department of Culture and the Arts & Knell, 2014, p. 6) was to evaluate:

- the intrinsic value – how a person is emotionally affected by the arts,
- the instrumental value – economic and social outcomes, and
- the institutional value – “the contribution of culture to producing a democratic and well-functioning society.

An app offered the metrics and two questions on gender and age. The public could download the app during or after the event. Peers and self-assessors completed the questionnaire by themselves.

In Australia, the following metrics have been submitted (Department of Culture and the Arts & Knell, 2014, p. 12):

1. Relevance (Currency): “It had something to say about today’s world” (self, peer, and public).
2. Captivation: “It was absorbing and held my attention” (self, peer, and public)
3. Originality: “It was ground-breaking” (self and peer)
4. Distinctiveness (Originality): “It was different from things I’ve experienced before” (public)
5. Excellence – global: “It was amongst the best of its type in the world” (self and peer)
6. Excellence – national: “It was amongst the best of its type in Australia” (self, peer and public)
7. Risk: “The artists/curators really challenged themselves with this” (self, peer, and public)
8. Rigour: “It seemed well thought through and put together” (self, peer, and public)

In Manchester in 2013, the metrics were modified to nine (9) metrics for self, peer and public assessment and five (5) specific to self and peer assessment (Bunting & Knell, 2014):

1. Presentation: “It was well produced and presented”.
2. Distinctiveness: “It was different from things I’ve experienced before”.

3. Rigour: “It was well thought through and put together”.
4. Relevance: “It has something to say about the world in which we live”.
5. Challenge: “It was thought provoking”.
6. Captivation: “It was absorbing and held my attention”.
7. Meaning: “It meant something to me personally”.
8. Enthusiasm: “I would come to something like this again”.
9. Local impact: “It is important that it’s happening here”.

- 
1. Concept: “It was an interesting idea”.
  2. Risk: “The artists/curators really challenged themselves with this work”.
  3. Originality: “It was ground-breaking”.
  4. Excellence (national): “It is amongst the best of its type in the UK”.
  5. Excellence (global): “It is amongst the best of its type in the world”.

The tests were launched in Manchester in 2013, at thirteen (13) cultural institutions (museums as well as theatres, cinemas, or music orchestras), and over 60 events since then. It has been considered a success (Arvanitis et al., 2016; Bunting & Knell, 2014; Gilmore et al., 2017, 2018).

Before the launch, the expectations were to produce performance measurements, to explore the cultural audience, and to be able to make data-driven decisions. The last part was not effective, as the survey occurred after the event. After the first trial in Manchester, it was found that the public viewed the metrics as being ‘too dry’ to express what they felt during the show or exhibition. It was then suggested to include open-ended questions like “What three words best described how you felt about the work?” (Bunting & Knell, 2014, p. 59). The first report also suggested the need to modify the self-assessment before evaluating the organisers’ expectations on the public response rather than their appraisal. And, as reported by Gilmore et al. (2017, p. 289), there was bias, which was introduced by the difference of profile (age, gender, social background) between the respondents and the public, as participation was voluntary.

However, the app was evaluated as user-friendly and provided reports ready to satisfy all stakeholders (Arvanitis et al., 2016; Bunting & Knell, 2014; Gilmore et al., 2018). Standardised metrics open the possibility for users to benchmark their institutions. Since its use in Australia and Manchester, the National Portfolio of the ACE founded a free trial on 150 institutions, and Culture Counts have been explored in Singapore and Canada. Research publications and results are expected.

### ***The DMA Friends Project (the Dallas Museum of Art, USA)***

In 2012, Robert Stein joined the Dallas Museum of Art (DMA). Stein has a background in computer science and has worked in the museum sector since 2006 (Stein, 2015). Bruce Wyman has a similar profile, working in both digital technologies and

museums (About Bruce Wyman, Museums and the Web, n.d.). The *DMA Friends Project* was launched on January 21, 2013, after extensive work on the engagement of the audience with art conducted from 2005 to 2009 and published in the book *Ignite the Power of Art: Advancing Visitor Engagement in Museums* (Pitman & Hirzy, 2010). The project came with free entrance and free membership to broaden and diversify the audience (Stein & Wyman, 2013). The aim was to measure the audience and the performance of the museum. The process consists of an on-site enrolment giving access to a member card with a specific barcode. The sign-in can be on site on the iPad or the visitor's smartphone, and the programme suggested completing different activities redeemed with digital badges. The system has been inspired by commercial affiliate programs and has offered rewards (like free parking or free exhibition ticket) in exchange for the points earned by participation in proposed activities (Stein & Wyman, 2013, 2014).

The *DMA Friends Project* is based on *BadgeOS*, a free plugin to WordPress (BadgeOS, n.d.) which has been modified with the help of LearningTimes. They designed the whole system to be flexible and therefore chose open-source systems as much as possible. The interface is web-based design for smartphones, and they have used *Twilio*, an API supplying an SMS interface to communicate with older cell phones (Stein & Wyman, 2013).

After 1 year of running, nearly 50,000 visitors had joined the programme, 343,000 badges had been awarded, and nearly 12,000 rewards collected. The results show that they have been able to encourage people to come back and engage with the museum. They have also generated more than 21.8 million records annually (Stein & Wyman, 2014).

These data were used for decision making with the help of three tools designed for reports and statistical analysis: *Chart.io*, *Tableau*, and some *Python* modules. With the charts, the museum administration was able to react to an apparent lack of interest in activities by deploying staff to welcome visitors and use teaching skills. Modification of the Friends' behavior was also attested through specific incitation (Stein & Wyman, 2014). With personal data collected, museum administration has developed an interactive map of the subscribers at the scale of the city and measured the representability of the population. They have also conducted quantitative and qualitative questionnaires.

The project received a National Leadership Grant from the Institute for Museum and Library. It was expected to expand to the Denver Art Museum (DAM), the Los Angeles County Museum of Art (LACMA), and the Minneapolis Institute of Arts (MIA; Stein & Wyman, 2014), but, according to the DMA website (Dallas Museum of Art, 2018), the *DMA Friends Project* ended on December 31, 2017. Publications on this are still pending.

### ***The API of the Spencer Museum of Art (University of Kansas, USA)***

This project is part of the ‘Integrated Arts Research Initiative’ (IARI), which aims to stimulate transdisciplinary research between science and the Spencer Museum of Art. The Andrew W. Mellon Foundation has funded it from 2016 to 2019 (Integrated Arts Research Initiative, Spencer Museum of Art, n.d.). James Miller, a professor of computer science, received a fellowship in spring 2017 to pursue his research on visualisation and what he called *storytelling* (Miller, n.d.). With Robert Hickerson, the museum’s database manager and archivist, they have been working together to create an API over a database of 45,000 objects. This API allows users to create new stories using the database’s data on the museum’s objects (Davis, 2017; Lynch, 2017). The API enables specific queries on the objects and the metadata, and offers interactive answers, showing the objects and their place of production (Davis, 2017). The results of the queries create new stories. An example is a statement by Joey Orr, curator for research, who stated: “one of our initial storytelling ideas relates to studying the impact of immigrants and the art they created after settling in the U.S.” (Lane, 2017).

### ***The MET API (Metropolitan Museum of Art, New York, USA)***

The Metropolitan Museum of Art (MET) has developed its API to manage the almost half-million object pages in its database. According to Liz Filardi, Senior Product Manager for collections in the Digital Department, who has cited Sebastian Chan, Director of the Cooper-Hewitt Lab at that time, MET has been developing the API to be at the core of the digitalised MET. This API is not yet accessible by the public. Museum staff has received relevant training, since this API constitutes a huge investment in the museum’s data management future, including the MET’s website and apps (Filardi, 2018). Interestingly, MET has had a partnership with Google BigQuery to give public access to parts of the database (200,000 objects). Sara Robinson, Developer Advocate on Google’s Cloud Platform team, also uploaded all images to Cloud Vision API and stored the results in a BigQuery table. This action allowed web detection of the items as well as other options, such as sorting items by colour (Robinson, 2017a, b).

### ***The Museum Big Data (MBD) Qatar Project (University College London in Qatar, Doha, Qatar)***

This project started in late 2017 and is ongoing. It is on museum Big Data and data mining, using data related to museums in Qatar as a case study. It has been acronymised as MBD Qatar and has secured funding for 2 years. It refers to museums in Qatar and aims to help towards developing Big Data and Data Mining methods and techniques on museum datasets, to add to a policy document on Big Data and the museums in Qatar, to initiate at University College London in Qatar (UCL Qatar) a research team on Museum/Cultural Heritage Big Data, and to explore links to information research schemes related to Social Media cultural dataset-producing processes in Qatar.

The project has so far worked on sentiment analysis for museums in Qatar to analyse feedback. It uses a specially developed dynamic data dashboard representing market benchmarks from direct and indirect competition. As a partner to this study/project, TycheTech Ltd., after processing provided datasets (public and/or custom), has undertaken the design and development to deliver a dynamic data dashboard. The result has been the *SENSIOM*, a dynamic multi-functional data dashboard on Museum Big Data to serve research needs (i.e., Negative/Neutral/Positive sentiment analysis, analysis on visitors' views on different museum-related topics, such as price of tickets, queue, other assets and issues), benchmarked against other museums and/or cultural institutions and hotspots, etc. (Papaioannou, 2018).

### **Conclusions: Towards a Fourth V (for Visitors)**

Museum Big Data, as Big Data in libraries, archives, and the arts and humanities, are still in their infancy, with an emphasis so far on digitisation-derived information rather than technologies, methods, and impact. There are a few Museum Big Data projects, some of them still ongoing, that contribute towards advancing Big Data influence and potentials in the museum world.

Projects presented here have common characteristics as well as diversity in terms of starting points, aims, objectives and reported outcomes. Their overall common ground is that museum Big Data tend to address and inform visitors' and the public's needs, demands, views, and perceptions. We can, therefore, argue that Museum Big Data boast an extra V, standing for visitors, since the three Vs of volume, velocity and variety in museum settings are related to museum visitors.

Diversity is another characteristic of museum Big Data projects, since they deal with a variety of issues including data generation and creation (Pen Project), evaluation (Culture Counts and Metrics project), services for visitors (the DMA Friends Project), storytelling (API Kansas Project), data management (the API MET Project), and sentiment analysis (the MBD Qatar Project). In this early stage, diversity is expected, since museum Big Data projects are still of exploratory character

both in terms of specific aims and objectives as well as for other uses and potentials. To this end, some projects have a dedicated computer science team involved.

All museum Big Data projects presented here seem to work towards models to be tested and used in the future. Culture Metrics have gone a step further by initiating a model in Australia, which was further elaborated and tested in Manchester. Museum Big Data projects are diverse and of exploratory character. There is also an emerging pattern towards evaluating aspects of museums' operations and experiences (Culture Counts, DMA, and UCL Qatar) and another one towards data generation and management (Pen Project, API Kansas, and API MET Projects). Moreover, the MBD Qatar project aims to produce a Museum Big Data Policy for the framework in which museums in Qatar operate, while APIs in Kansas and MET can be further tested in other museum contexts. Common patterns, policies, and applications of proposed models seem to be the next steps for research and practice. It is also interesting that, despite the availability and sharing of technical issues and possibilities, data and results are yet to be available from these projects, which is expected to be solved soon.

## References

- About Bruce Wyman, Museums and the Web.* (n.d.). Retrieved March 27, 2018, from <https://www.museumsandtheweb.com/author/bwyman/>
- Arvanitis, K., Gilmore, A., Florack, F., & Zuanni, C. (2016). Data culture and organisational practice. *Museums and the Web 2016*. Retrieved March 27, 2018, from <https://mw2016.museumsandtheweb.com/paper/data-culture-and-organisational-practice/>
- Atzori, L., Iera, A., & Morabito, G. (2010). *The Internet of Things: A survey*. Retrieved from <http://www.ic.unicamp.br/~bit/mo809/seminarios/Ciro-IoT/suporte/surveyIoT.pdf>
- BadgeOS. (n.d.). *BadgeOS*. Retrieved March 27, 2018, from <https://badgeos.org/>
- Beyer, M. A., & Laney, D. (2012). *The importance of "Big Data": A definition*. Gartner report (pp. 1–9). Gartner.
- Bunting, C., & Knell, J. (2014). *Measuring quality in the cultural sector*. Retrieved March 27, 2018, from <https://www.artscouncil.org.uk/publication/measuring-quality-cultural-sector>
- Chan, S. (2014). *The API at the center of the museum*. Cooper Hewitt Labs. Retrieved from <https://labs.cooperhewitt.org/2014/the-api-at-the-center-of-the-museum/>
- Chan, S. (2015a). Strategies against architecture: Interactive media and transformative technology at Cooper Hewitt. *Museums and the Web 2015*. Retrieved March 27, 2018, from <https://mw2015.museumsandtheweb.com/paper/strategies-against-architecture-interactive-media-and-transformative-technology-at-cooper-hewitt/>
- Chan, S. (2015b). *5 months with the Pen: Data, data, data*. Cooper Hewitt Labs. Retrieved from <https://labs.cooperhewitt.org/2015/5-months-with-the-pen/>
- Cooper Hewitt, Smithsonian Design Museum. (2014). *The new Cooper Hewitt experience*. Cooper Hewitt Smithsonian Design Museum. Retrieved March 27, 2018, from <https://www.cooperhewitt.org/new-experience/>
- Dale, B. (2016). The Cooper-Hewitt's connected Pen brings visitors back, online. *Observer*. Retrieved March 27, 2018, from <http://observer.com/2016/02/20-percent-of-this-smithsonian-museums-connected-pen-users-come-back-to-its-site/>
- Dallas Museum of Art. (2018). *DMA friends*. Retrieved March 27, 2018, from <https://www.dma.org/visit/dma-friends>



- Davis, A. (2017). *New interactive program to help museum visitors use data to explore artworks*—IEEE—The Institute. Retrieved March 27, 2018, from <http://theinstitute.ieee.org/technology-topics/big-data/new-interactive-program-to-help-museum-visitors-use-data-to-explore-artworks>
- De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings*, 1644, 97.
- Department of Culture and Arts & Knell, J. (2014). *Public value measurement framework – Measuring the quality of the arts*. Department of Culture and the Arts, Government of Western Australia.
- Dey, N., Hassanien, A. E., Bhatt, C., Ashour, A. S., & Satapathy, A. C. (Eds.). (2018). *Internet of Things and Big Data analytics towards next-generation intelligence*. Springer.
- Evans, D. (2011). *The Internet of Things – How the next evolution of the internet is changing everything*. Retrieved March 27, 2018, from [https://www.cisco.com/c/dam/en\\_us/about/ac79/docs/innov/IoT\\_IBSG\\_0411FINAL.pdf](https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf)
- Ferro, S. (2014). *The 5 coolest things about the revamped Smithsonian Design Museum*. Co.Design. Retrieved March 27, 2018, from <https://www.fastcodesign.com/3039370/the-5-coolest-things-about-the-revamped-smithsonian-design-museum>
- Filardi, L. (2018). *How does the met maintain half a million web pages devoted to art?* The Metropolitan Museum of Art. Retrieved March 27, 2018, from <https://www.metmuseum.org/blogs/collection-insights/2018/met-collection-api>
- Ghemawat, S., Gobiuff, H., & Leung, S. T. (2003). The Google File System. *SOSP '03 proceedings of the nineteenth ACM symposium on operating systems principles* (pp. 29–43). Retrieved March 27, 2018, from <https://static.googleusercontent.com/media/research.google.com/ell/archive/gfs-sosp2003.pdf>
- Gilmore, A., Arvanitis, K., & Albert, A. (2018). “Never mind the quality, feel the width”: Big Data for quality and performance evaluation in the arts and cultural sector and the case of “culture metrics”. In G. Schiuma & D. Carlucci (Eds.), *Big Data in the arts and humanities: Theory and practice* (pp. 27–40). CRC Press.
- Gilmore, A., Glow, H., & Johanson, K. (2017). Accounting for quality: Arts evaluation, public value and the case of “culture counts”. *Cultural Trends*, 26(4), 282–294.
- Gobble, M. A. (2013). Big Data: The next big thing in innovation research. *Technology Management*, 56(1), 64. Retrieved March 27, 2018, from [www.gale.cengage.com/AcademicOneFile/](http://www.gale.cengage.com/AcademicOneFile/)
- Gordon-Murnane, L. (2012). Big Data: A big opportunity for librarians. *Online*, 36(5), 30–34.
- Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, 51(7–8), 662–672.
- Holden, J. (2006). *Cultural value and the crisis of legitimacy: Why culture needs a democratic mandate*. Demos.
- Huwe, T. K. (2012). Big data, big future. *Computers in Libraries*, 32(5), 20–22.
- Integrated Arts Research Initiative, Spencer Museum of Art*. (n.d.). Retrieved March 27, 2018, from <https://spencerart.ku.edu/integrated-arts-research-initiative>
- Jacobs, A. J. (2012). Quantifying myself. In *The human face of big data* (pp. 54–57). Against All Odds Productions.
- Krishna Prasad, A. V. (Ed.). (2018). *Exploring the convergence of Big Data and the Internet of Things*. IGI Global.
- Lane, K. (2017). *Data visualization and storytelling around museum collections using APIs – DZone Big Data*. Dzone.Com. Retrieved March 27, 2018, from <https://dzone.com/articles/data-visualization-and-storytelling-around-museum>
- Lohr, S. (2012). The age of Big Data. *The New York Times*. Retrieved March 27, 2018, from <http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html>
- Lynch, B. (2017). *Computer scientist to boost interactivity of Spencer Museum of Art’s history and holdings*. The University of Kansas. Retrieved March 27, 2018, from <https://news.ku.edu/2017/05/25/computer-scientist-boost-interactivity-spencer-museum-art%E2%80%99s-history-and-holdings>

- Manyika, J., Chui, M., Brown, B., & Bughin, J. (2011). *Big Data: The next frontier for innovation, competition, and productivity*. Retrieved March 27, 2018, from [https://www.mckinsey.com/~lmedia/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI\\_big\\_data\\_exec\\_summary.ashx](https://www.mckinsey.com/~lmedia/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx)
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- McAfee, A., & Brynjolfsson, E. (2012). Big Data: The management revolution. *Harvard Business Review*, 90(10), 59–68.
- Miller, J. (n.d.). *Spencer Museum of Art*. Retrieved March 27, 2018, from <https://spencerart.ku.edu/iari/2017/jim-miller>
- Milosch, J., Kurz, M. J., Jansen, G. J., Hull, A., & Marciano, R. (2018). Museums, archives, and universities: Structuring future connections with Big Data. In G. Schiuma & D. Carlucci (Eds.), *Big Data in the arts and humanities: Theory and practice* (pp. 159–172). CRC Press.
- Papaioannou, G. (2018). Museum (big) data mining in Qatar: Researching and developing methods, techniques, and a policy. *Museum and the Web 2018*. Retrieved March 27, 2018, from <https://mw18.mwconf.org/proposal/museum-big-data-mining-in-qatar-researching-and-developing-methods-techniques-and-a-policy/>
- Pitman, B., & Hirzy, E. (2010). *Ignite the power of art: Advancing visitor engagement in museums*. Dallas Museum of Art.
- Prodromou, T. (Ed.). (2017). *Data visualisation and statistical literacy for open and Big Data*. IGI Global.
- Reinhalter, L., & Wittmann, R. J. (2014). The library: Big Data's boomtown. *The Serials Librarian: From the Printed Page to the Digital Age*, 67(4), 363–372.
- Robinson, S. (2017a). Processing a large dataset in less than 100 lines of Node.js with Async.Queue. *Medium*. Retrieved March 27, 2018, from <https://medium.com/the-node-js-collection/processing-a-large-dataset-in-less-than-100-lines-of-node-js-with-async-queue-9766a78fa088>
- Robinson, S. (2017b). *When art meets big data: Analyzing 200,000 items from the Met collection in BigQuery*. Google Cloud Big Data and Machine Learning Blog. Google Cloud Platform. Retrieved March 27, 2018, from <https://cloud.google.com/blog/big-data/2017/08/when-art-meets-big-data-analyzing-200000-items-from-the-met-collection-in-bigquery>
- Schiuma, G., & Carlucci, D. (Eds.). (2018). *Big data in the arts and humanities: Theory and practice*. CRC Press.
- Schwartz, M. (2013). What government big data may mean for libraries. *Library Journal*, 1. Retrieved March 27, 2018, from <https://lj.libraryjournal.com/2013/05/oai/what-governmental-big-data-may-mean-for-libraries/>
- Smolan, R., & Erwit, J. (2012). *The human face of big data*. Against All Odds Productions.
- Stein, R. (2015). *About Robert Stein*. Retrieved March 27, 2018, from <http://rjstein.com/about/>
- Stein, R., & Wyman, B. (2013). Nurturing engagement: How technology and business model alignment can transform visitor participation in the museum. *Museums and the Web, 2013*. Retrieved March 27, 2018, from <https://mw2013.museumsandtheweb.com/paper/nurturing-engagement/>
- Stein, R., & Wyman, B. (2014). Seeing the forest and the trees: How engagement analytics can help museums connect to audiences at scale. *Museums and the Web, 2014*. Retrieved March 27, 2018, from <https://mw2014.museumsandtheweb.com/paper/seeing-the-forest-and-the-trees-how-engagement-analytics-can-help-museums-connect-to-audiences-at-scale/>
- van Dijk, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197–208. Retrieved from. <https://doi.org/10.24908/ss.v12i2.4776>
- Walter, M. (2015). *The data won't save us*. Retrieved March 27, 2018, from <https://uncategorized.micahwalter.com/the-data-won-t-save-us-4f57dbd5a3ec#ncqjwuug>
- Walter, M. (2016a). *Building analytics at Cooper Hewitt*. GitBook. Retrieved from <https://www.gitbook.com/book/micahwalter/building-analytics-at-cooper-hewitt/details>
- Walter, M. (2016b). *A very happy & open birthday for the Pen*. Cooper Hewitt Labs. Retrieved March 27, 2018, from <https://labs.cooperhewitt.org/2016/a-very-happy-open-birthday-for-the-pen/>

- Walter, M. (2016c). *The-Pen-Data: Open data generated by Cooper Hewitt's Pen*. Cooper Hewitt, Smithsonian Design Museum. Retrieved March 27, 2018, from <https://github.com/cooperhewitt/the-pen-data>
- Wright, L. (2017). New frontiers in the visitor experience. In A. Hossaini & N. Blankenberg (Eds.), *Manual of digital museum planning* (pp. 109–128). Rowman & Littlefield.
- Xiong, W., Yu, Z., Bei, Z., Zhao, J., Zhang, F., Zou, Y., Bai, X., Li, Y., & Xu, C. (2013). A characterization of big data benchmarks. In *2013 IEEE international conference on big data* (pp. 118–125). IEEE.
- Zikopoulos, P. C., Eaton, C., Deroos, D., Deutsch, T., & Lapis, G. C. (2012). *Understanding Big Data. Analytics for enterprise class Hadoop and streaming data*. McGraw-Hill. Retrieved March 27, 2018, from <https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/II11025E.pdf>

**Dr. Georgios Papaioannou** is an Associate Professor in Museum Studies at University College London in Qatar, MA in Museum and Gallery Practice. He has studied classics, archaeology, cultural heritage and IT in Greece, the U.K. (UCL and King's) and in Spain. His research interests lie on applications of new technologies for cultural heritage, archeology and education, including digitisation, data management, virtual reality, augmented reality and mobile applications. He has focused lately on Big Data for museums and cultural heritage, and he organised the first International Conference on Museum Big Data (Doha, 2019). He directs and participates in internationally-funded (of budgets exceeding €2 million) research work and projects of cultural/museum works in the Mediterranean and the Arab world (Greece, Cyprus, Syria, Jordan, Oman, Qatar), where he has founded and organized museums and museum exhibitions, including the Museum at the Lowest Place on Earth in the Dead Sea, Jordan. He has received awards for his research, teaching and administrative work. He has taught at universities in Greece, UK, Australia, France, Cyprus and Qatar, including teaching in open and distance education settings via new technologies. He has authored books, papers and conference presentations, including textbooks and alternative teaching materials in the methodology of open and distance learning. He is a Senior Fellow of the Higher Education Academy (UK), the Secretary General of the Hellenic Studies Society of Near East (HSNES), a member of the International Council of Museums (ICOM), and a member of the Pool of Experts of the European Museum Academy.

# Chapter 10

## Analysing Aspects of Brazilian Curricula for Teaching Statistics Involving Big Data



Carlos Eduardo Ferreira Monteiro, Maria Niedja Pereira Martins,  
and Theodosia Prodromou 

**Abstract** Recent studies in statistical education have discussed the potential in using statistical data called *big data* – information from large databases that are frequently produced and have relative accessibility. However, the prescribed school curricula are not always in line with most recent recommendations of statistics education scholars, and implemented curricula also do not fully utilize elements that contribute to the learning of statistics, such as technological artefacts. This chapter discusses and compares explicit and implicit prescriptions related to the use of big data in statistics education in the basic education curricula of Brazil and Australia. The study used a qualitative documentary approach. The documentary artefacts used included guidelines and curricular programs available online. For the Brazilian context, the research was conducted online, collecting curriculum documents from 27 Brazilian states. Forty-five documents were selected and analysed in order to identify potentialities in the teaching of statistics for the secondary level of schooling. The content analysis indicated that only some curricular documents address the issue of big data, and even those documents did not explicitly refer to *big data* or *open data*. However, the documents analysed did discuss the use of Information and Communication Technologies. Such recommendations are still discussed in a general way and more focused on data handling. The analysis of the Australian curricula documents is of particular interest as the development of statistics education is a major area of research interest in Australia, and this has influenced recent Australian curricular reform. Therefore, we intend to examine Brazilian and Australian contexts in order to identify limits and possibilities for utilization of big data in school curriculum. The chapter concludes with suggestions for the development and inclusion of knowledge and practices for teaching big data in sta-

---

C. E. F. Monteiro · M. N. P. Martins (✉)  
Federal University of Pernambuco, Recife, Brazil  
e-mail: [carlos.fmonteiro@ufpe.br](mailto:carlos.fmonteiro@ufpe.br)

T. Prodromou  
University of New England, Armidale, NSW, Australia  
e-mail: [theodosia.prodromou@une.edu.au](mailto:theodosia.prodromou@une.edu.au)

tistics, as well as a call for adding these elements to teacher-education programs, as current programs do not seem to explore these concerns.

**Keywords** Statistics education · Big data · Basic education curricula

## Introduction

Statistical knowledge is essential for critical reflective and participatory citizenship (Carvalho & Solomon, 2012), therefore statistics education is an important area to prepare citizens to understand statistical data critically. Contemporary information technologies produce vast amounts of data about social processes and practices, and citizens need to be able to make sense of these large and changing data sets, called *big data*, in order to make informed decisions.

Zeelenberg and Braaksma (2017) classify big data as large datasets related to social activities which are not covered by official statistics. These authors emphasize some characteristics of big data. Firstly, big data often comes from sources in which the populations are not well-defined, as well as from sources based on survey, census, or administrative data. Secondly, big data may be highly volatile and selective because the population to which it refers may change from day to day, which produces non-stationary time-series. Additionally, some big data sets do not have linking variables which would allow relating them to other datasets or population frames. These limitations increase the possibility of error in statistical results. In order to minimize possible bias, which may be done by combining big data with data from other sources, it is important to have what can be called *big data* literacy (Franseshey & Monteiro, 2018).

Big data literacy should be part of basic school curricula. At the secondary and tertiary school levels, concerns for big data are not based on how to manage a lot of data, but in finding ways to prepare students to deal with this type of data. Ainley et al. (2015) state that “data are not big because of the size of the data file, but because they belong to a new class of data that differ in structure and source from traditional data that have inspired institutional changes in how we learn from data” (p. 409). Therefore, thinking of ways to help students learn from the data necessarily requires that the school curricula emphasize the most relevant content for this new context, as well as the strategies and materials most significant for this teaching.

This chapter analyses school curriculum documents in Australia and Brazil to see whether and how they are including or taking consideration of these new sources of data, as well the new demands on statistics literacy.

## Australian Curriculum

The Australian curriculum in Mathematics is being developed progressively by the Australian Curriculum, Assessment, and Reporting Authority (ACARA, 2020a, b). The Australian curriculum is developed at the national level, so all schools across Australia are following the same plan. It was designed as a developmental sequence of learning from foundation to Year 10 (age 5–6 to age 15–16; ACARA, 2020a, b) and was originally developed during 2008–2010, and then revised in 2014. It describes what is to be taught and the quality of learning expected of the students as they proceed through school. It acknowledges the importance of disciplinary knowledge, skills, and understanding in mathematics alongside general across-curriculum priorities, which are Aboriginal and Torres Strait Islander histories and cultures, Asia and Australia's engagement with Asia, and sustainability.

The Australian curriculum in mathematics includes the foundation to year 10 (F-10) Australian curriculum in mathematics and the senior secondary subjects of essential mathematics, general mathematics, mathematical methods, and specialist mathematics. The F-10 Australian Curriculum in mathematics is designed to provide students with essential mathematical skills and knowledge in number and algebra, measurement and geometry, and statistics and probability. It develops the numeracy capabilities that all students need in their everyday professional and civic life. The F-10 Australian Curriculum also includes the fundamentals on which mathematical specialities and professional applications of mathematics are built.

In the Australian curriculum, students from their early years in primary school are expected to learn to pose questions and collect categorical or numerical data by observation or survey. They are taught to construct displays, including column graphs, dot plots and tables, appropriate for data type, with and without the use of digital technologies. They also learn how to describe and interpret different data sets in context. By year 10, students are expected to calculate and interpret the mean and standard deviation of data and use those to compare numerical data sets. They are expected to use the standard deviation to describe the spread of data and use information technologies to investigate bivariate numerical data sets. They are also required to understand variation in data, when it is appropriate to use a straight line to describe a relationship, as well as different techniques for finding a line of best fit. In addition to the F-10 curriculum, the Australian Curriculum, Assessment, and Reporting Authority (ACARA, 2020a, b) developed a senior secondary curriculum for mathematics, which specifies the content and achievement standards for advanced students.

In addition to the F-10 curriculum, ACARA developed a senior secondary curriculum for mathematics, which specifies the content and achievement standards for advanced students.

The senior secondary Australian Curriculum for mathematics is organised into four units. Each unit is designed to be taught in one half of a school year, in approximately 50–60 h, including assessments and examinations. The last two units, which

are about statistics, are cognitively more challenging than the first two, which cover more general subject matter.

State and territory curriculum, assessment, and certification authorities take responsibility over the structure and organisation of their senior secondary courses and decide about the integration of the senior Secondary Curriculum content and achievement standards into their courses, including assessment, certification and the attendant quality assurance mechanisms.

Each of these authorities acts according to its respective legislation and the policy framework of its state government and Board. The state and territory authorities also determine assessment and certification specifications for their courses and any additional information, guidelines and rules to satisfy local requirements, along with advice on entry and exit points and credit for completed study.

The Australian curriculum does not explicitly discuss big data, but it advocates that students who study essential mathematics be expected to learn a range of skills related to use and understanding of big data. The curriculum in essential mathematics requires students to represent categorical data in tables and column graphs, as well as to compare and analyse data, and interpret a range of statistical information. The curricula also calls for students to analyse data obtained from practical everyday situations, investigate the procedure for conducting a census including the advantages and the disadvantages of conducting a census (e.g., comparing samples). The students are asked to understand that the purpose of sampling is to provide an estimate of population values when a census is not used, and investigate the various types of samples (e.g., systematic samples, self-selected samples, simple random samples). They are also taught how to identify the target population to be surveyed and investigate questionnaire design principles, including, use of simple language, unambiguous questions, consideration of number of selections, issues of privacy and ethics, and freedom from bias.

The Australian curriculum also requires students to describe the faults in the data collection process and the sources of error in the surveys (e.g., sampling error and measurement error). Additionally, it requires students to investigate the possible misrepresentation of the results of a survey, including examples of media misrepresentations of surveys because of misunderstanding the procedure or the reliability of generalising the findings of a survey to the entire population.

The Australian essential mathematics curriculum requires students to describe the patterns and features of bivariate data in bivariate scatterplots and the association between two numerical variables in terms of (positive/negative) direction, (linear/non-linear) form and (strong/moderate/weak strength). The students also need to identify the dependent and independent variables and find the line of best fit by eye or by the use of technology, interpret relationships in terms of the variables, find the correlation coefficients with the use of technology and use the line of best fit to make predictions, both by interpolation and extrapolation, including recognising the dangers of extrapolation and distinguishing between causality and correlation through examples.

National statistics agencies like the Australian Bureau of Statistics (ABS) produce statistical products that help with data visualisation and the analysis of big data

with the use of analytical platforms created by the ABS (e.g., the Graphically Linked Information Discovery Environment; GLIDE; Clarke & Chien, 2017; Prodromou, 2017). The aim of ABS is to combine complementary information derived from big data sources and traditional data sets (ABS, 2014, 2015), in order to deliver innovative solutions that meet the evolving information needs of Australian students and citizens. The ABS strives to have a transformative role in assisting a variety of users including students for their growth of extracting, interpreting, critically evaluating statistical information, and communicating analytical insights from complex statistical data. Furthermore, “it leverages a groundswell of support for open data in government, industry and academia, as a way of making public services more efficient, generating new economic value, and reducing the distance between governments and citizens” (Clarke & Chien, 2017, p. 225). The Bureau of Communications research (ABS, 2015) mentions the importance of open government data in Australia that change the data landscape for statistical analysis, the way data is presented by visualisations, and the necessary skills required for the analysis. To a variety of data users, the ABS introduces applications of advanced visualisation approaches to support the analysis of big data and the inductive reasoning in both exploratory and explanatory analysis for complex problems and data.

In summary, the Australian curriculum shows interest in many issues related to big data. Although it does not explicitly discuss big data, it advocates approaches and methods that would equip students with big data concepts and enriched them with a range of skills and techniques that are useful for understating big data. Whereas, the following 45 state curricular documents in Brazil selected for the analysis of aspects of big data, with their emphasis on ways of approaching big data, provide another example of a curriculum that is in line with recent pedagogical theory of statistical literacy.

## Brazilian National Guidelines

The Brazilian educational system is organized in three levels: national, state, and municipal. The Federal government publishes curricular documents which give orientation to educational networks of the 27 states. Each state is comprised of municipalities which also have the autonomy to publish specific curricular documents. There are approximately 5600 municipalities in Brazil.

The National Curriculum Parameters (PCNs – Parâmetros Curriculares Nacionais) suggested the introduction of statistics from early school years (primary school) as part of the curriculum contents of Mathematics. The first PCN document was published in 1997 by the Brazilian Ministry of Education (MEC).

Other documents were published by MEC in order to suggest curricular contents for the final years of the fundamental school level, from Year 6 until Year 9 (Ministério da Educação, 1998), and for middle school, called “Ensino Médio” (Ministério da Educação, 1999, 2002).



In 2006, the Federal Brazilian government published curricular guidelines for the middle school level (Ministério da Educação, 2006).<sup>1</sup>

The PCNs and Curricular Guidelines do not have the status of the standard official curriculum for fundamental school (12 years of schooling). The states and municipal governments are responsible for the details and operational actions which implement the suggestions for local public-school networks.<sup>2</sup> Private schools have more independence to adapt their own programmes in order to achieve the curricular prescriptions.

PCNs emphasise that the teaching of mathematics and statistics should be based on problem-solving approaches which challenge pupils to elaborate different types of reasoning and conceptual processes using available technological tools. They also indicate three types of teaching content: conceptual, procedural, and attitudinal (Ministério da Educação, 1997).

One of the innovations of the PCNs was to introduce data handling as one of four mathematics content blocks alongside other blocks, which address number and operations; space and shapes; and measurement. At the time of PCNs' first publication, statistical concepts were not taught in the first years of fundamental school. Therefore, the inclusion of data handling demanded important changes in teacher education programmes and textbook contents. Teachers were faced with a great challenge: teaching a topic which they had never taught before or learnt in their pre-service or in-service teacher education (Monteiro et al., 2000).

The PCNs suggest different contents for two teaching cycles in order to approach specificities of the ways in which school curricula are organised in different parts of the country.<sup>3</sup> The contents vary slightly considering that in the second cycle the objectives are wider than in the first cycle.

- (first Cycle) *Conceptual and procedural*: Reading and interpretation of data presented in images; Collect data, and construct personal ways to organize and communicate collected data; Interpretation and elaboration of lists, tables, and bar graphs. *Attitudinal*: Interest to learn, interpret, and produce messages which use graphical representations.
  - (second Cycle) *Conceptual and procedural*: Collection, organization, and description of data; Reading and interpretation of data presented in lists, tables, diagrams, and graphs; Construction of graphs and tables based on data from media, scientific, and other types of written texts, understanding the utilities of graphs as a global presentation of data and highlighting the relevant aspects. *Attitudinal*: Interest in analysing all meaningful elements of a graphical representation; avoiding partial and premature interpretations

(adapted from Ministério da Educação, 1997).

<sup>1</sup>In 2013, the MEC published a document which systemised all school curricular guidelines on General National Curricular Guidelines for Education (Diretrizes Curriculares Nacionais Gerais da Educação Básica) (Ministério da Educação, 2013).

<sup>2</sup>There are only two types of Fundamental schools in Brazil: public non-fee-paying schools, which are supported by the government, and private fee-paying schools.

<sup>3</sup>In Brazil, schools have autonomy to organise the initial periods of Fundamental Teaching in cycles which do not have the length of a school calendar year.

The PCNs recommend that the teaching of graphing should be associated with other mathematical and statistical concepts and notions, other school subjects, and pupils' daily lives. According to the PCNs, the main aim is to teach pupils beyond the level of knowing how to read and interpret graphical representations, therefore they need to be able to describe and interpret their real-world experiences using statistical knowledge. Teachers should stimulate pupils to ask questions, make relationships, build justifications, and develop an investigative spirit.

Although PCNs suggest that statistical knowledge is important for interpreting and critically evaluating data presented in daily life, there is not much discussion of how such knowledge should be developed. The document exemplifies topics of interest to children, such as: their birthdays, grandparents' nationalities, and football teams. However, it does not establish explicit connections between those examples and a problem-solving perspective. In addition, the evaluation criteria presented in the PCNs do not seem to approach the complexity of the aim to enable pupils to interpret data critically.

In this paper, we aimed to identify whether these curricular guidelines have explicit or implicit prescriptions related to the use of big data or open data in teaching statistics. In the next sections, we summarize the main results from our analysis of national guidelines and state curricular documents to teach statistics.

## State Curricular Documents

Brazil is composed of 27 states (counting the Federal district of Brasilia). These units have relative autonomy of government, legislation, and Tax collection or Census data collection. The political-administrative organization of Brazil allows the state governments to construct documents that guide the teaching in each state, and thus each Brazilian state has a secretary of education and a state government department responsible for managing and implementing the curriculum in the state educational network. In some states, more than one current document containing curricular guidelines was found, and for these states, the multiple documents were all considered in this analysis.

A summary of the documents<sup>4</sup> found, as well as those selected and their respective publication years can be seen in Table 10.1:

According to the data presented in Table 10.1, 96 state curricular documents were found. However, only 45 of them presented content and learning expectations for statistics and, therefore, were those selected for the analysis of big data aspects in this study.

In addition, we were not able to access documents from 4 states (Acre, Pará, Rio Grande do Norte, and Roraima). The state with oldest document was Goiás, with a

---

<sup>4</sup>The research was made between 11/05/2017 at 11/05/2017. The documents are available on the respective webpages of the state of education secretariats.

**Table 10.1** States, their respective documents, and year of publication

Name of States	Number of documents found on the websites of state secretaries.	Number of documents selected	Year of publication of the documents analysed
Acre	00	00	–
Alagoas	05	03	2010–2012–2014
Amapá	02	01	2009
Amazonas	05	01	2012
Bahia	09	03	2013–2015–2015
Ceará	04	01	2009
Distrito Federal	03	02	2015–2015
Espírito Santo	04	03	2017–2017–2017
Goiás	03	01	2007
Maranhão	01	01	2014
Mato Grosso	02	01	2010
Mato Grosso do Sul	04	02	2012
Minas Gerais	02	02	2008–2014
Pará	01	01	2010 (normative resolution)
Paraíba	09	01	2010
Paraná	03	02	2008–2011
Pernambuco	08	04	2012–2013–2015–2015
Piauí	04	02	2013–2013
Rio de Janeiro	05	05	2012
Rio Grande do Norte	02	00	–
Rio Grande do Sul	07	03	2009–2009–2016
Rondônia	02	02	2012–2013
Roraima	00	00	–
Santa Catarina	04	01	2014
São Paulo	05	01	2012
Sergipe	01	01	2011
Tocantins	01	01	2008
Total	96	45	–

curricular proposal in force since 2007; and the most recent document, published in 2017, was from the state of Espírito Santo.

In the states of Brazil, the mathematics curriculum is generally subdivided into 4 main axes; one of them is labelled as *data handling*, and it also includes content about combinatorics and probability.

The federal states and municipalities **share** responsibility for fundamental school (school years 1–9), while the federal states **have sole** responsibility for high school (10–13). As a result, there are many official curricular documents in each Brazilian

**Table 10.2** Division of the Brazilian educational system

Ages	School year	Nomenclature	Level of Education	Classification
0–3 years old	I, II, III	Nursery school	Child education	Basic Education
4 and 5 years old	IV, V	Preschool		
6–10 years old	1–5°	Early years (1° cycle)	Fundamental School Level	
11–14 years old	6–9°	Final years (2° cycle)		
15–17 years old	10–12°	High School (3° cycle)	Middle School	

state for the final years of fundamental school. However, the large number of municipalities in the country led us to focus our investigation on the documents only formulated by the states, as presented in Table 10.2.

## Methods

This study is based on an interpretative paradigm for data collection and analysis. We used a documentary approach to identify explicit and implicit elements related to big data in the Brazilian curriculum of the last 4 years of basic schooling, namely, 9th, 10th, 11th and 12th school years.

To examine the data, we used content analysis, which is a range of techniques aimed at identifying the contents of messages using systematic procedures (Bardin, 1977). Minayo (2004) indicates five techniques of content analysis: relations analysis, expression analysis, enunciation analysis, evaluation analysis, and thematic analysis. In thematic analysis, the central concept of a message is the theme and can be presented with a word or sentence. The theme, in turn, can be understood as the unit of signification that liberates itself from a text according to the theory that guides the researcher’s reading (Bardin, 1977).

## Word Frequencies in the Documents

With NVivo 11, we performed a search for words and their frequencies in all 45 state documents presented in Table 10.1. The NVivo resource-counting feature was used to aid in the floating reading step of the material, in order to direct our search to the excerpts of the documents were more associated with *big data* terms.

NVivo generated a list of the 500 most frequent words in the texts of 45 documents. Of these, 26 were classified as potentially being related to *big data*. The words identified and their frequencies in the documents are listed in Table 10.3.

Among the terms most used by documents and highlighted as associated with the context of Big Data were the words “representation”, “representations”, and “to

**Table 10.3** Words related to big data in state curricular documents (N = 45)

Word	Frequency	%
Representação/representações/representar (Representation/to represent)	2613	9.49%
Análise/analisar (Analysis/to analyse)	2583	9.38%
Sistema/sistemas (system/systems)	2453	8.91%
Organização (Organization)	2244	8.15%
Interpretar/interpretação (Interpret/interpretation)	1886	6.85%
Informações/informação (Information)	1843	6.69%
Recursos (Resources)	1820	6.61%
Dados (Data)	1530	5.55%
Tecnologia/tecnologias (Technology/technologies)	1429	5.19%
Gráficos (Graphs)	1258	4.57%
Cotidiano (Daily Life)	1137	4.13%
Realidade (Reality)	1102	4.00%
Comunicação (Communication)	981	3.56%
Pesquisa (Research)	916	3.32%
Procedimentos (Procedures)	859	3.12%
Fenômenos (Phenomena)	811	2.94%
Técnicas (Techniques)	521	1.89%
Busca (Search)	516	1.87%
Tratamento (Handling)	495	1.78%
Total	27,518	100%

represent” (verb), which occurred a total of 2613 times in the documents reviewed. We also found the terms “analysis” and “to analyze” (verb) with 2583 entries made and “system” or “systems”, present in documents 2453 times. On the other hand, terms used less, but still among the 500 most frequent words were “techniques” with 521 mentions, “search”, with 516, and “Treatment”, with 495.

We identified that three terms: “Data”, “Technology”, and “Research”, which are more explicitly related to Big Data, appeared in the documents 1530, 1429 and 916 times, respectively. These words are associated with Big Data, because it is generated and managed by technologies and specific data analysis software. For this reason, a search was made in the documents to identify the contexts in which these words were used.

From this, it was possible to relate the frequency of those words to each document belonging to the different federal states. Table 10.4 lists the distribution of these three words in the documents of the various states.

It can be seen that the word “dados”, appears most frequently in the official documents of the states of Pernambuco (464), Rio Grande do Sul (127) and Bahia (97). On the other hand, the states of Pará (2), Santa Catarina (6), as well as the Distrito Federal (3) included the fewest uses of that word in their documents. Regarding the word “tecnologia”, the documents that used those terms most were those of the states of Amazonas (179), Rondônia (145) and Bahia (103), while we

**Table 10.4** Distribution of words related to Big Data found in 45 State documents examined in relation to their States belonging

Name of states	Words found		
	Dados (Data)	Tecnologia (Technology)	Pesquisa (Research)
Alagoas	33	28	64
Amapá	18	36	41
Amazonas	45	179	30
Bahia	97	103	66
Ceará	31	85	12
Distrito Federal	3	63	20
Espírito Santo	76	75	74
Goiás	68	47	19
Maranhão	9	90	31
Mato Grosso	43	65	39
Mato Grosso do Sul	81	86	41
Minas Gerais	87	9	4
Pará	2	25	3
Paraíba	37	32	61
Paraná	26	24	15
Pernambuco	464	11	121
Piauí	79	53	87
Rio de Janeiro	8	2	0
Rio Grande do Sul	127	19	32
Rondônia	71	145	93
Santa Catarina	6	38	14
São Paulo	14	32	2
Sergipe	28	92	18
Tocantins	77	71	27
Total	1530	1429	916

saw that the states of Rio de Janeiro (2) and of Minas Gerais (9) used that word least frequently in their documents. The third and last selected word– “pesquisa” – was used most frequently in the documents of Pernambuco (121), Rondônia (93) and Piauí (87), and least frequently in the documents of the states of Rio de Janeiro (0), São Paulo (2) and Pará (3).

The frequency of the found terms should be considered in the context of the amount of documents that each state published. In the case of the state of Pernambuco, 4 documents were considered for analysis, and it was a state with high frequencies of the terms “Data” and “Research”. On the other hand, the states of Rio de Janeiro and of Pará, that presented low frequencies of those words, were those for which we only found one valid document, for example, the state of Pará only presented one normative resolution.

Furthermore, it can be noted that the Brazilian states that used the most frequently listed words in Table 10.4 are more concentrated in the north and northeast of the country, while the states that used the words less were distributed in other

Brazilian areas. This observed trend is relevant, since the states of the north and northeast of Brazil are those with low rates of development of human income, income per capita and Gross Domestic Product – GDP- compared to the rest of the country. That relationship is worth mentioning because the investments for the placement of technological tools in the schools in areas with smaller incomes can also be smaller.

## Contextual Analysis of the Words “Data”, “Research”, and “Technology” in the Documents

In relation to contexts of the word “data”, of the 1530 mentions of the word in the documents, 250 excerpts containing sentences or even complete paragraphs were identified in the guidelines for mathematics and other sciences for the levels 9th, 10th, 11th and 12th in the documents. These fragments surrounding the use of the word “data” are associated with statistics and probability content. In these excerpts, we identified expectations for students to: carry out data collection; access statistical data; represent and interpret data in graphs and tables; analyse data sets; access databases; and construct representations using technological resources.

These documents all address the need for data collection activities not only in mathematics, but also in other sciences, such as geography, chemistry, and biology. Some of these passages include explicit discussion of the use of media and other technological resources that could support the search for information. Although there is significant use of these words in several of the documents, the discussion does not offer any deep or substantial discussion of the themes or ideas of big data. They do, however, make explicit mention of sites or platforms that provide access to big data.”

In excerpts of the guidelines related to mathematics, and specifically to probability and statistics, the use of the term “data” was mainly associated with the terms “data selection”, “data collection”, “data set analysis”, and “construction of tables and graphs starting from data”. These expressions, which are related to the stages of the treatment of information and statistics, appear primarily in sections that describe the learning expectations for students. Some examples are given in the sections excerpts that follow:

To infer information starting from data displayed in tables or graphs. To use data presented in tables or graphs in the resolution of problems. Data collection and presentation in graphs. *Knowledge and Expectations for the 3rd year of the Middle School*. (Secretaria da Educação, Espírito Santo, 2017, without page).

To build tables of frequencies and to represent statistical data graphically using several resources. *Expectations for the 9th year of the Fundamental School*. (Secretaria da Educação, Ceará, 2009, p. 221).

Identify relevant data on a given situation-problem to seek possible resolutions; for example, in a situation-problem of data presented through tables, graphs, technical specifications, to recognize relevant information from a given subject needed for problem solu-

tion. To recognize the relevant information for a given subject that she seeks to understand. Expectations for Middle School. (Secretaria da Educação, Ceará, 2009, p. 94).

Only a few sections in the documents appeared to be explicitly concerned with big data. Some directly discussed the topic, but did not propose any activities that specified sources of data. Moreover, a number of these discussions were not made as part of an explicit discussion of statistics or probability. An example of this can be seen from the curricular proposal for middle school from the State of Amazonas. With regard to the teaching of biology, the document presents an example of activity that could be developed by students of 11th grade and whose goal is discussing fingerprinting as a crime detection tool. In order to contextualize the activity, the paper presents the following clarification:

The criminologists take samples of prints at the crime scene and compare them with others taken from the suspect. Although the classificatory system was developed in 1899, it is still widely used. However, fingerprints are now reproduced digitally, and recorded in a huge database, which can be used to find matches quickly, a procedure that used to require a small army of researchers. (Secretaria da Educação, Amazonas, 2012, p. 182).

Despite bringing a context that takes into account the processes of collecting digital data, the suggestions for activities proposed in the document shortly after this clarification did not involve the development of research access to the database. These and other passages found seem to suggest that the documents contextualize their practices taking into account the need for access to data and to the research. However, few clarifications are made more explicit about the access to technological databases in the methodological descriptions of the guidelines here investigated.

Thus, we note that the suggestions for data collection exercises focus on field work, access to documents, and carrying out research on the spot or by access to collections with census results and other finite data. In this context, we did not find suggestions for search and analysis of data from platforms that capture a lot of information:

Observing the spot data related to environmental problems like: garbage, sewage, water treatment, occupation of the springs pollution of urban rivers of cities. *Middle school, 3<sup>o</sup> grade – Biology*. (Secretaria da Educação, Amazonas, 2012, p. 117).

In the case of the stages of data analysis and construction of statistical representations, we found guidelines regarding the use of technological resources to assist in the implementation of these steps with the students. This is an excerpt from the document that most emphasized the use of technological resources:

In this context, the correct use of calculators and computers is essential. Those instruments are important because they allow the approach to problems with actual data, which can be resolved quicker, while the student may have the opportunity to become familiar with hardware and specific software to analyse statistics and probability. *Basic Education*. (Secretaria da Educação, Mato Grosso, 2010, p. 149).

In spite of the few explicit recommendations about big data, the specific expectations for learning to be achieved by students, have the potential for working with the big data. That is because the assumptions presented in support of the introduction of



statistics on these curricula agree with the approach of this area with the transformations in the world of the technology and with the large amount of information to which students have access. These reasons, for instance, are stated in the curriculum of the final years in the state of Minas Gerais:

In today's world, every day, every individual receives a large amount of information and often uses statistical techniques to correlate data and, from these, draw conclusions. In addition, other areas of knowledge, such as biology, physics, chemistry, geography, among others, make use, constantly, of statistical language. Thus, we see the teaching of statistics assuming an instrumental-operational role in integrating studies across various disciplines and even between different topics with in mathematics itself. *Curriculum of the final years of Fundamental School*. (Secretaria da Educação, Minas Gerais, 2014, p. 57).

Concerning the use of the word “research”, we have seen that despite the 916 appearances of this term; of the three crucial terms, it appeared the least in the math curriculums for the grades 9–12.

From a second search about the word context, we selected 80 excerpts with phrases and paragraphs that are associated with the term. Even with little direct use of this word in these school years, many documents refer to the idea of teaching research as a pedagogical approach to basic schooling. For example, the proposal of the Distrito Federal states:

Thus, research should be one of the principles of everyday school life, both in practice of teaching, amplifying the Freirean concept of teacher as researcher, and in the routine of students, providing a new way of looking at events around them, developing the capacity of thinking, and enjoying the new knowledge. *Middle School*. (Secretaria da Educação, Distrito Federal, 2015, p. 17).

This assumption is present in every document that we reviewed from all the different Brazilian states, since each discusses, in its own way, the importance of doing research in school. Thus, some sections related to the word “research” included clarifications about principles of curriculum proposals, about various approaches to research projects and, finally, about the practice of research in statistics. We saw, however, that the use of this term was more associated with other sciences in the documents reviewed. Therefore, research was less mentioned in the sections where the teaching of statistics and probability was discussed.

However, the use of research as a teaching approach in statistics and mathematics seems to be generally little addressed in the resources to support teaching activities. Silva and Guimarães (2011) conducted an investigation in two collections of textbooks adopted by the Brazilian National Textbook Plan and concluded that the textbooks do not propose complete research situations, and that the research approaches discussed are quite fragmented in science and math activities.

In addition to the curricular guidelines, textbooks are important features to work with the math and statistics content in the classroom. In Brazil, where teacher-training programs lack discussion of pedagogy in statistics (Kataoka et al., 2011), having texts and other documents that mention more detailed aspects of statistical pedagogy that integrates issues of how to carry out research could enrich teaching practices in schools.

In the fragments that mentioned the word “research” in the guidelines for the teaching of statistics in grades 9, 10, 11, and 12, we observed that some documents provided more explicit guidance for the use of this approach in the classroom in those school levels.

As an example, the Pernambuco guidelines for the classroom learning objectives for these 4 years of schooling include the practice of statistical activities from research:

To carry out research, including all its stages (planning, selection of samples, elaboration and application of collection instruments, organization and representation of the data, interpretation, critical analysis and popularization of the results); To select an appropriate sample for a certain research project. *Guidelines for the classroom- 11th grade*. (Secretaria da Educação, Pernambuco, 2013, p. 79).

The guidance offered for the achievement of these objectives involves the selection of samples, connection to other areas of knowledge, exploration of concepts and statistical ideas, and the use of software and other resources in the steps of analysis and representation of data. A similar recommendation can be found in the guidelines of Rio Grande do Sul state, in an activity described in the teaching of samples, frequencies, and statistical representations, in grade 10 of High School, teachers are instructed:

To request that the students carry out research with their friends in school. Initially, to present to the students some themes that they could study: a) the average bodyweight of students in the school, taking into consideration that there are many students; b) the favourite soft drink(s) of the students in the school; c) the favourite television programs of high school students (A, B, C, D, E). In pairs or quartets, students must choose one of the themes presented or any other of interest. Each theme should have a background and issues related to it. (Secretaria da Educação, Rio Grande do Sul, 2009, p. 277)

With regard to the proposed activity, additional details are provided to help spark the interest of the students, and there is also a discussion of statistical concepts that could be utilised in the development of this activity. In this document, the research suggestions do not focus on big data contexts, since the document offers suggestions in which the collection of data occurs primarily through contact with individuals and access to books and other documents. In spite of the importance of that task type for the understanding of statistics’ usefulness in everyday concepts, we believe that due to the increase of the importance of big data in society, activities that also incorporate tasks with large data sets could be valuable to students in the final years of fundamental school and in middle school.

With respect to the word “technology”, in general, the documents recommend the use of technologies in high school, so the incorporation of technological resources is an explicit goal at that stage of education, as well as in the final years of fundamental school. In fact, according to the law of Guidelines and Bases for National Education (Lei de Diretrizes e Bases da Educação Nacional – LDB) Law n°. 9.394/96, art. 26, the curricula of fundamental school and middle school should have a common national base and a diversified part in order to contemplate the needs of each region. However, both parts must be planned to create a coherent and

continuous curricular proposal for Information and Communication Technologies (ICT) for pre-school through high school.

This characteristic is found in the documents reviewed. As an example, we have an excerpt of the Maranhão state curriculum guidelines, that emphasizes the symbiosis of the curriculum, the school context, and information and communication technologies.

Curricular development inevitably runs through the new relation with the knowledge that Information and Communication Technologies (ICT) opportunize and potentialize, articulating the school with other spaces producing knowledge, provoking substantial changes inside the school and building a collaborative culture. The deep transformations in the universe of the media impel the school, for the role it plays, to understand the technological culture, taking advantage of the characteristics of the different communication and information vehicles, and aiming at improving the teaching-learning processes and, consequently, quality standards. (Secretaria da Educação, Maranhão, 2014, p. 82)

Regarding the practices related to mathematics, statistics, and probability, we find a smaller number of indications of the word “technology”. In these cases, guidance about the use of technological resources is also generally offered. In the curricular guidelines of the Bahia state, for example, some recommendations are made for the use of technologies in teaching mathematics and it is made clear that the technological tools favour new forms of data presentation and analysis in the teaching of statistics.

A learning environment conducive to working with these areas is the use of media and information technologies that enable new processes for presenting and analyzing important data regarding social and economic conditions. Such an approach favors the exercise of citizenship, directly linked to the understanding of numerical data that allow a conscious and informed action. (Secretaria da Educação, Bahia, 2013, p. 132)

Certain documents choose as an expectation of learning some competences associated with the use and understanding of technology in mathematics. In the Mato Grosso proposal, for example, among the various teaching objectives for the 2nd and 3rd cycle of learning in mathematics, are:

Use technology as a means of researching some scientific and mathematical knowledge. (Secretaria da Educação, Mato Grosso, 2010, p. 23)

Argue and reason critically, using technology as a means of research in relation to scientific and mathematical subjects. (Secretaria da Educação, Mato Grosso, 2010, p. 33)

Based on these excerpts, it can be stated that the curricula analysed here offer particular attention to the use of technology in the teaching of mathematics and statistics in the final years of fundamental school and in middle school. The guidelines, however, have a more generalist character since only some few documents discuss in any detail how and what technological resources could be used in the tasks in statistics.

## Conclusions

The review of documents in Australia and Brazil shows that some national and state-level school districts are moving towards more modern views of mathematics and statistics pedagogy, but these views are not yet universally accepted. The Australian curriculum (ACARA) and applications of advanced visualisation approaches by the ABS for the analysis of complex problems and big data and 45 state curricular documents in Brazil selected for the analysis of big data aspects in this study, with their emphasis on ways of approaching big data, provide good examples for curricula that are in line with recent pedagogical theory of statistical literacy.

Realising this emergence and significance of the Big data era, it is crucial for teachers and policy makers to update curricula by integrating computer-based applications of advanced visualisation approaches—such as GLIDE introduced by ABS—that would support the analysis of big data. Teachers should follow up on that by providing relevant instructions to the students about the use of the advanced visualisation approaches and guide them to understand the analysis of big data and inductive reasoning in both exploratory and explanatory analysis for complex problems and data. Further, support for connecting statistical concepts with advanced visualisation approaches and bridging informal and formal understandings, is required to impart necessary knowledge to students.

As researchers and educators, we suggest three pedagogical goals that would enable curricula to impart updated knowledge to middle school and secondary students. These three goals are focused explicitly to meet our overall goal of preparing statistically literate students for this era's data revolution (Prodromou & Dunne, 2017).

Goal 1 is concerned with providing students with strong foundations in data visualisation and basic mechanical processes that can be used to derive meaningful insights of big data, for example, running through several data sets to look for meaningful correlations between each other. The suggested approach to accomplish goal one, is to equip students with big data concepts and use applications of advanced visualisation approaches—similar to GLIDE—that would support the analysis of big data.

Goal 2 is concerned with motivating students about the potential of big data to solve problems suitable to local or global contexts and the impact of big data. This could be achieved by providing students with examples of real world problems that are governed by big data and the noteworthy impact of big data on the target real world problems. For example, big data could help women overcome fertility issues by tracking critical fertility signs, such as menstrual cycles, morning temperatures, weight, stress levels, etc.

Goal 3 is concerned with preparing students to meet growing requirements for making sense of and analysing big data. Goal 3 could be accomplished by enriching students with the latest educational tools and techniques for analysis of big data that

enhance the use of a range of skills for making sense of certain characteristics big data.

Teaching curriculum for big-data-centric pedagogy requires changes and integration (at least) of the suggested approaches and methods in teaching practices. We believe that global cooperation is required in meeting possible challenges, for example, limits on resources or lack of availability of big data that is suitable to local contexts. Furthermore, educators should emphasise comprehensive coverage for curriculum, integrating in their practices the teaching of basic programming methods, tools, statistical skills, algorithms, and tools for network access.

Although big data isn't explicit in the curricula we discussed in this chapter, if you're interested in teaching or developing policy or researching big data, these goals and their respective suggestions do give possibilities for building big-data-centric pedagogy within the bounds of the current curricula (or at least mostly within those bounds, with a few exceptions).

## References

- ABS. (2014). *Big data for informed decisions*. Australian Bureau of Statistics. Unpublished.
- ABS. (2015). *Special report: Big data plays a big role in the future of statistics*. Annual Report, 2014–15, Australian Bureau of Statistics. Retrieved from <http://www.abs.gov.au/>
- Ainley, J., Gould, R., & Pratt, D. (2015). Learning to reason from samples: Commentary from the perspectives of task design and the emergence of “big data”. *Educational Studies in Mathematics*, 88(3), 405–412.
- Australian Curriculum and Assessment Reporting Authority (ACARA). (2020a). *F-10 Curriculum*. [www.australiancurriculum.edu.au/mathematics/curriculum/f-10](http://www.australiancurriculum.edu.au/mathematics/curriculum/f-10) (viewed May 2020).
- Australian Curriculum and Assessment Reporting Authority (ACARA). (2020b). *Senior-secondary-curriculum*. [www.australiancurriculum.edu.au/senior-secondary-curriculum/mathematics](http://www.australiancurriculum.edu.au/senior-secondary-curriculum/mathematics) (viewed May 2020).
- Bardin, L. (1977). *Análise de conteúdo*. edições 70.
- Carvalho, C., & Solomon, Y. (2012). Supporting statistical literacy: What do culturally relevant/realistic tasks show us about the nature of pupil engagement with statistics? *International Journal of Educational Research*, 55, 57–65.
- Clarke, F., & Chien, C. H. (2017). Big data in official statistics. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and big data* (pp. 224(pp2)). : IGI Global.
- Franershey, K., & Monteiro, C. (2018). *Big data literacy*. In International conference on teaching statistics, 10., 2018, Kyoto, Japan. Proceedings... Voorburg. The Netherlands: International Statistical Institute.
- Kataoka, V. Y. et al. (2011, July). A Educação Estatística no Ensino Fundamental II em Lavras, Minas Gerais, Brasil: Avaliação e intervenção. *Revista latinoamericana de investigación en Matemática educativa*, México, 14(2), 233–263. Available in: <http://redalyc.uaemex.mx/src/inicio/artpdfred.jsp?icve=33519238005>
- Ministério da Educação. (1997). *Parâmetros curriculares nacionais para o ensino fundamental*. Brasília, MEC/SEF.
- Ministério da Educação. (1998). *Introdução. Ensino fundamental*. Brasília, MEC/SEF.
- Ministério da Educação. (1999). Secretaria de Educação Média e Tecnológica. *Parâmetros curriculares nacionais: Ensino Médio*. Brasília: Ministério da Educação, 360p.

- Ministério da Educação. (2002). Secretaria de Educação Média e Tecnológica. *Pcns+ Ensino Médio: orientações educacionais complementares aos Parâmetros Curriculares Nacionais. Ciências da Natureza, Matemática e suas Tecnologias*. Brasília: MEC, SEMTEC, 144p.
- Ministério da Educação. (2006). *Orientações curriculares para o Ensino Médio: Ciências da Natureza, Matemática e suas Tecnologias*, 2. Brasília: SEB.
- Ministério da Educação. (2013). *Diretrizes Curriculares Nacionais Gerais da Educação Básica*. Brasília, secretaria de educação básica. Secretaria de educação continuada, alfabetização, diversidade e inclusão. Secretaria de educação profissional e tecnológica. Conselho nacional da educação. Câmara nacional de educação básica, 562 p.
- Minayo, M. C. S. (2004). *O desafio do conhecimento: pesquisa qualitativa em saúde* (8th ed.). hucitec.
- Monteiro, C., Selva, A., & Ferreira, J. (2000). *Tratamento de informações: investigando o processo de interpretação de gráficos*. In Proceedings of the 52a. reunião anual da sociedade para o progresso da ciência. Brasília, Brazil.
- Prodromou, T. (2017). *Data visualization and statistical literacy for open and big data*. IGI Global.
- Prodromou, T., & Dunne, T. (2017). Statistical literacy in data revolution era: Building blocks and instructional dilemmas. *Statistics Education Research Journal*, 16(1), 38–43.
- Silva, E., & Guimarães, G. (2011). *A pesquisa em livros didáticos dos anos iniciais do ensino fundamental*. In Conferência Interamericana de Educação Matemática. Anais do XIII CIAEM, Recife, UFPE.
- Secretaria da Educação, Espírito Santo. (2017). *Ensino Médio Regular*. Secretaria da Educação.
- Secretaria da Educação, Goiás. (2007). *Currículo em debate: Reorientação curricular do 1º ao 9º ano, expectativas de aprendizagem - convite à reflexão e à ação*. Secretaria de Educação, Goiânia, 228p.
- Secretaria da Educação, Ceará. (2009). *Metodologias de Apoio: matrizes curriculares para ensino médio*. Secretaria da Educação, Fortaleza, 156p.
- Secretaria da Educação, Amazonas. (2012). *Proposta Curricular para o Ensino Médio*. Secretaria da Educação, Manaus, 202p.
- Secretaria da Educação, Mato Grosso. (2010). *Orientações Curriculares: área de Ciências da Natureza e Matemática – Educação Básica*. Cuiabá, 166p.
- Secretaria da Educação, Bahia. (2013). *Orientações curriculares e subsídios didáticos para a organização do trabalho pedagógico no ensino fundamental de nove anos* Secretaria da Educação, Salvador, 177p.
- Secretaria da Educação, Maranhão. (2014). *Diretrizes Curriculares*. Secretaria de Educação, São Luís, 107p.
- Secretaria da Educação, Minas Gerais. (2014). *Currículo Básico Comum dos Ensino Fundamental – Anos finais*. Ouro Preto, 60p.
- Secretaria da Educação, Distrito Federal. (2015). *Currículo em Movimento da Educação Básica – Ensino Médio*. Brasília, 82p.
- Secretaria da Educação, Pernambuco. (2013). *Parâmetros na sala de aula – Matemática – Ensino Fundamental e Médio*. Recife, 209p.
- Secretaria da Educação, Rio Grande do Sul. (2009). *Referencial Curricular – Matemática – Ensino Médio*, 316p.
- Zeelenberg, K., & Braaksma, B. (2017). Big data in official statistics. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and big data* (pp. 274–296). IGI Global.

**Carlos Eduardo Ferreira Monteiro** Professor at Department of Educational Psychology at of the Centre of Education (CE) of The Federal University of Pernambuco (UFPE). PhD in Mathematics Education at The University of Warwick (2005) and Master Cognitive Psychology at UFPE (1998). Since 1995 has been involved in teacher education courses mainly related to Educational Psychology and Mathematics and Statistics Education. Since 2008 I am lecturer and supervisor of PhD and Master students of the Post-Graduate Programme in Mathematics and

Technological Education (Edumatec) <https://www.ufpe.br/ppgedumatec> I have been involved in national and international research projects, including periods as visiting post-doctoral fellow at The University of Leicester, England (2007), The University of Lisbon, Portugal (2014–2015), The Federal University of Rio Grande do Sul – UFRGS (2017), Brussels Free University – VUB, Belgium (2012; 2019–2020). I am currently leader of two research groups: GPEMCE associated with teaching and learning processes of mathematics in field/rural education, GPEME that conducts research projects on mathematics and statistics education. My research projects are on psychology of mathematics education, statistics education, critical sense in statistics, statistical literacy, big data, sociocultural aspects involved in the teaching and learning of mathematics and statistics, field/rural education, mathematics and statistics teacher education. My main publications are related to these topics.

**Maria Niedja Pereira Martins** Pedagogical advisor of education secretary of Pernambuco state. PhD in Education with emphasis in Educational Psychology at University of Lisbon, Portugal (2018), mathematical and technological education master degree at the Federal University of Pernambuco – UFPE, Brazil (2014), and graduated in Pedagogy at UFPE (2011). Conducts research with emphasis on the processes of teaching and learning statistics. Member of GPEMCE that a research group that investigates teaching and learning processes on mathematics education in the contexts of Rural Education, and member of research group GPEME that conducts studies on mathematics and statistics education. The main research themes are linked to the following topics: statistics teaching and learning; statistical literacy; mathematics and statistics teacher education in the early years of elementary school, the utilization of technology to teach and learn statistics, attitude scale towards statistics. I have experience in teaching at early years of elementary school. At university level, I have taught courses associate with distance education theory, supervision of student teacher internship, general didactics, theoretical fundamentals of education; teacher education, and distance learning education. I am currently involved technical and pedagogical advising to Pernambuco state program called child literacy, whose focus is literacy in mathematics and Portuguese Language of children at elementary school initial grades.

**“Dr Theodosia Prodrμου** is a senior lecturer of Mathematics Education at the University of New England in Australia. She has published numerous journal articles, book chapters and edited/ authored books on a wide range of subjects, including the use of big data in educational settings, the relationship between technology and mathematical thinking, the integration of digital technologies in the teaching and learning of mathematics, STEM education, augmented reality in Educational settings, professional development of mathematics teachers, and statistical thinking including perceptions of probability and chance. Her most recent edited/authored books were, *Augmented reality in educational settings* (2020) and *Primary and Middle Years Mathematics: Teaching Developmentally* (2019).”

# **Part V**

## **Conclusion**



# Chapter 11

## Concluding Comments



Theodosia Prodromou 

Big Data arose from the growing use of interconnected computational systems. As the Internet exploded during the 1990s, so, too, did the amounts of data generated, and with these massive new data sets, supported by increasing computation and storage power, came new problems and opportunities. Confronted with so much data, traditional methods for data analysis were no longer sufficiently effective or efficient, and new techniques began to develop to deal with this new, often unstructured data. Because Big Data is relatively new, so, too, is its study. In this new field, there is a wide range of issues to be studied, and many disparate views and divergent approaches. Researchers are just beginning to scratch the surface of the vast areas for research and examination, and each step forward offers many new questions to be answered. The research in this book offers a view into many different questions about Big Data, and how and why it can be used in education and educational research.

As Big Data becomes a common part of the fabric of our world, education and research are challenged to use these data to improve educational and research systems, and also are tasked with teaching coming generations to deal with Big Data both effectively and ethically. Big Data presents many opportunities and concerns, as can be seen in the range of offerings in this book, some of which are concerned different ways of teaching students to use Big Data (Chaps. 2 and 3), others with using Big Data to teach more effectively (Chaps. 5, 6, 7 and 8), and others with questions about how to use educational systems to respond to Big Data (Chap. 4), how Big Data could be integrated into educational systems (Chap. 9), and the extent to which Big Data is being adopted into educational systems (Chap. 10). These various studies show a field of research in which the work is relatively new, and in which new questions and possibilities are opening.

---

T. Prodromou (✉)  
University of New England, Armidale, NSW, Australia  
e-mail: [theodosia.prodromou@une.edu.au](mailto:theodosia.prodromou@une.edu.au)

In Chap. 1, we started with some of the big questions faced in world of Big Data and discussion of where academic researchers can reach for a wide array of different existing sources of data that they might profitably employ in their research work. Author Jacques Raubenheimer observed the need for a critical mass of researchers extending the possibilities of what can be done with Big Data because of the challenges presented, and discussed these challenges along with the corresponding dangers. He also noted how the dual forces of technology and methodology are eroding some problems when working with voluminous data sets that may require some form of linking, but that these new abilities are coupled with potential ethical dangers.

He discusses important characteristics of Big Data sets that are used for academic research but are typically obtained for other purposes and not for academic research. This impacts every aspect of working with those data, from accessing it, to linking and analysing it, through to the ethical implications of working with such data. Chapter 1 concludes that Researchers should realise that Big Data will not replace traditional small data research methods, and that the opportunities presented by Big Data are dependent on the context of the researcher, the population and the data – in some contexts, Big Data may offer great advantage, but not in others. Ideally, Big Data and small data will become complementary tools in the armouries of academic researchers, which they may draw upon variably, and in concert if need be, as the situation demands.

In Chap. 2, Peter Howley, Kevin Wang, and Ayse Aysin Bilgin focus on the critical step of breaking barriers to learning in order to reach students. They discuss how to leverage the wide application of statistics and big data, and a holistic integration of the often disparately taught aspects, in order to create enthusiasm and improved learning. They present the nature of the initiatives described in this chapter on hopes of inspiring researchers to pursue this area, and for teachers to adopt and adapt them for their own learning and teaching requirements.

They argue that industry, business, and government, along with higher education personnel must proactively support early childhood, primary, and secondary educators to achieve in the future the development of an ultimately productive citizen or member of the workplace. An arm's length approach of expecting primary and secondary teachers to upskill alone to the point of piquing students' interests in big data and statistics and driving students to further studies in the discipline, is fraught. Activities such as the National Schools Poster Competition and the CSIRO's STEM Professionals-in-Schools program with which it collaborates ([www.csiro.au/en/Education/Programs/STEM-Professionals-in-Schools](http://www.csiro.au/en/Education/Programs/STEM-Professionals-in-Schools)) are designed to contribute to the required support mechanisms. The challenge is to ensure that it is not left to the few to continue this charge, rather that all recognise the supply chain is only as strong as its weakest link, and thus similarly unite in this charge towards addressing the Big Data age.

Another look into educational for Big Data, Chap. 3 shifts the focus to graduate studies in statistics. Ayse Aysin Bilgin and Peter Howley provide a case study on the inclusion of Big Data into a Master of Applied Statistics unit, called data mining. They discuss the short history of the unit, topics of study, learning outcomes,

assessment tasks, and how students provided with an opportunity to apply their knowledge to authentic, complex Big Data. In addition, they consider how students were included in decision making for their projects when they selected a data set for the projects provided that they get an approval from the lecturer. Bilgin and Howley stressed the aim of the design of the unit to help students to learn and/or improve work-related soft skills needed in practice. Although the MAppStat students would have the required skills, they need the language used by the workforce to specify which skills are required for a specific job or job applications. For example, a big telecommunication company in Australia required the applicants to have a strong background in a broad spectrum of machine-learning algorithms, including unsupervised and supervised learnings and a variant of neural network models, they were seeking someone with a PhD or Master's degree in Computer Science or Electrical Engineering. Although the MAppStat students would have the required skills, because the degrees were specified in the advertisement, the MAppStat students were excluded from applying for this specific job. This raises a discussion about Big Data and its place on statistics education. Even if the statistical methods can be applied, the employers and so-called "data scientists" might not be aware of them. The assumption is people with "Data Science", "Business Analytics", "Computing", or similar degrees would be good for Big Data analysis but not someone with a statistics degree. Statistics is somehow considered very separate from Big Data in the eyes of many people, though not by many statisticians. The authors stress the importance to include Big Data analysis techniques into curriculum, as highlighted by Ridgway (2016), and make the public, especially employers, aware of the place of statistics and statisticians in the world of Big Data. The authors argue that statistical literacy along with the expertise in the area from which the data are coming are important qualities to be successful in the Big Data era (Hewson & MacGillivray, 2016). Hence, it is of paramount importance to equip students who study statistics with Big Data analysis skills for successful data science projects.

Chapter 4 discusses how, with the digital revolution, private companies produce new information and some of them offer statistical products in classic fields of official statistics like price statistics (e.g., scanner data). Although, this is a positive development, these companies are not able to produce public statistics independently and with the same quality as the statistical offices that produces neutral, independent, and high-quality public statistics. Markus Zwick and Sabine Köhler explain that for those companies, it would be too expensive and difficult to get a return on the investment from the market. Moreover, national statistical offices are the only ones who have access to administrative data. Those reasons reinforce the importance of the national statistical offices. A discussion follows about making national statistical institutes digital and the cooperation between NSIs, academia, and private-sector entities.

Chapter 5 considers an example of the use of Learning Analytics in education – using Big Data to understand student performance – to explore the opportunities and challenges of analytics in education. More specifically, it examined university students' activity on the Moodle learning platform to gain insights into relationships between students' performance and their actions on the online learning platform.

The chapter investigates how the day of assignment submission relates to the grades received by the students and we predicted the day of the week for assignment submission, by each student. Based on the data analyses the case study, the unit coordinator was provided with feedback to evaluate and review the unit through the lens of best practices. Future research should seek further best practices to find out how the design of essential elements of the unit affects students' academic performance. Amongst these elements could be online video lectures, recordings of Zoom sessions, teaching materials, video recordings, interaction of the students with the unit coordinator, suitable technological tools, interactivity of online video lectures materials.

Chapter 6 further considers the place of Big Data learning analytics generated from a new, computer-based implementation using the Meaning Equivalence Reusable Learning Objects (MERLO) pedagogical tool for learning and formative assessment. MERLO learning assessments capture these important aspects of conceptual thinking of students and reveal individual students' conceptual strengths and weakness. Each individual student's Big Data profile includes time-marked data that describes continuous learning processes and outcomes at both individual learner and individual teacher levels, useful for both individuals and educational institutions.

Chapter 6 concludes that the implementation, testing, and validation, since 2002, of pedagogy for conceptual thinking (Shafir & Etkind, 2018; Shafir & Kenett, 2016) showed that the MERLO formative assessments enhance peer cooperation and the conceptual thinking and Interactive Concept Discovery (InCoD) provides individual learners with different points of view on conceptual issues relevant to the course content, written by different authors. The study of the learning analytics data generated from use of the MERLO and InCoD tools help the authors to draw conclusions about the pedagogy for conceptual thinking in large undergraduate classes. The wide range of experience in MERLO and InCoD tools demonstrate their universality and reflects their large potential in future research.

Chapter 7 also examined questions of learning analytics. William Billingsley and Peter Fletcher discuss the use of professional tools, specifically version-control systems, in designing environments that provide students with opportunities to work on realistic open-ended problems. As part of their operation, such systems gather data about the students using the systems, and the data allow teachers and researchers to investigate the strategies that the students apply when they work, and eventually compare those strategies with the strategies that professionals employ.

The authors suggest that the future is where intelligent tutoring, rich analytics, and authentic practice will be regularly combined in education. However, they discuss a number of aspects that need to be considered which are involved in the design of the task by the educator, the choice of the environment, and the components to connect. After the discussion of those aspects, authors consider how a task unfolds once it has been scoped. The authors discuss the connected environment of professional systems where the work itself will take place and the necessity to introduce students to the ecosystem gradually and in well-scaffolded ways. This discussion is followed by a discussion about the facets that are important to consider when it then comes to choosing which systems to deploy in the course. The authors also stress

the importance of examining how identity management will work between the Learning Management System (LMS) and the professional tools, especially for cloud-based systems. The chapter concludes by discussing the goal of developing cognitive apprenticeships that are supported by smart technology and that use authentic environments, and argue that could be achieved when the three strands of rich learning environments, learning analytics, and authentic tasks come together.

Chapter 8 is a further exploration of Learning Analytics and the use of Big Data as an educational tool. Authors Belinda A. Chiera, Małgorzata W. Korolkiewicz and Lisa J. Schultz focus on exploring and developing strategies to address student learning and online engagement patterns in a standard university setting that combines face-to-face learning with online resources and learning environments. In this chapter, patterns of engagement were explored for three components of the online course environment, namely the course website, weekly topic resources, and the social forum. For example, they note that in the context of first-year University students, it is particularly vital to understand student patterns of online engagement in order develop an understanding of student engagement with course resources over the duration of a semester. The information gained from the use of the learning analytics appears to depend crucially on the structure of a virtual learning environment as well as decisions made by online engagement data custodians. However, the authors discuss how learning from learning analytics could become challenging because there a potential disconnect between what data is available and what data is useful.

The authors noted that the main difficulties lay with the interpretation of weekly data which were obtained from the virtual learning environment access logs. They argued that it is important to have more detailed information of how student activities or other factors contributed to these weekly counts, so a wider picture of student engagement will be achieved. Moreover, a conclusion stems from data analysis refers to assessment, both formative and summative that is one of the deciding factors whether a first year student engages with online course resources. Hence, it is needed to differentiate the online support resources and the advice given to students on what contributes to study.

Chapter 9 explores the use of Big Data in museums, taking into consideration both theoretical perceptions and relevant practices in current Big Data projects in museums, especially educational projects intended to make museum resources more available and useful to students and researchers. Museum Big Data, as Big Data in libraries, archives, and the arts and humanities, are still in their infancy, with an emphasis so far on digitisation-derived information rather than technologies, methods, and impact. The data, data analytics, and data mining processes are expected to be used and become core disciplines as data input increases and digital data collections from museums' everyday activities and operations become larger and larger.

The chapter explores a few Museum Big Data projects, some of them still ongoing, that contribute towards advancing Big Data influence and potentials in the museum world. The projects presented in Chap. 9 have common characteristics as well as diversity in terms of starting points, aims, objectives and reported outcomes.

Their overall common ground is that museum Big Data tend to address and inform visitors' and the public's needs, potential demands, views, and perceptions.

The author, Georgios Papaioannou, argues that Museum Big Data boast an extra V, standing for visitors, since the three Vs of volume, velocity, and variety in museum settings are related to museum visitors. There is a discussion about "diversity" as another characteristic of museum Big Data projects since they are still of exploratory character both in terms of specific aims and objectives as well as for other uses and potentials. To this end, some projects have a dedicated computer science team involved. All museum Big Data projects presented in Chap. 9 seem to work towards models to be tested and used in the future. Common patterns, policies, and applications of proposed models seem to be the next steps for research and practice.

Chapter 10 takes a more policy-centric view. Authors Carlos Eduardo Ferreira Monteiro, Maria Niedja Pereira Martins, and Theodosia Prodromou look at the place of Big Data in national educational curricula. Specifically, using a qualitative documentary approach, they discuss and compare the integration of big data in statistics education in the basic education curricula of Brazil and Australia.

The evidence from Australia and Brazil shows that some national and state-level school districts are moving towards more modern views of mathematics and statistics pedagogy, but these views are not yet universally accepted. Realising this emergence and significance of the Big Data era, it is crucial for teachers and policy makers to update curricula by integrating computer-based applications of advanced visualisation approaches – such as GLIDE introduced that would support the analysis of big data. The authors discuss the need for teachers to provide relevant instructions to the students about the use of the advanced visualisation approaches and guide them to understand the analysis of big data and inductive reasoning in both exploratory and explanatory analysis for complex problems and data. Further, support for connecting statistical concepts with advanced visualisation approaches and bridging informal and formal understandings, is required to impart necessary knowledge to students.

The authors suggest three pedagogical goals that would enable curricula to impart updated knowledge to middle school and secondary students. These three goals are focused explicitly to meet our overall goal of preparing statistically literate students for the big data era revolution (Prodromou & Dunne, 2017).

Chapter 10 continues to emphasise that the teaching of a curriculum for Big-Data-centric pedagogy requires changes and integration (at least) of appropriate approaches and methods in teaching practices. Mathematics/statistics educators should comprehensively cover the curriculum of Big Data, integrating in their practices the teaching of basic programming methods, tools, statistical skills, algorithms, and tools for network access. Additionally, global cooperation is required in meeting possible challenges, for example, limitations on resources or lack of availability of big data that is suitable to local contexts.

This book covers a wide range of people, but it only begins to scratch the surface of possibilities for Big Data in education. In particular, it's worth noting that the chapters of this book were written before the COVID-19 Pandemic that led to a

dramatic expansion of online learning and teaching, and thus miss insights that might have followed from having a larger proportion of educational opportunities online.

Looking to the future, it might also be interesting to see research on some areas not included in this book. In this conclusion chapter, I would like to ask all researchers to invest some time and energy into thinking about possible futures about providing students of all levels (from primary to tertiary education) with opportunities to apply their knowledge to authentic complex Big Data and use analytics to gain insights in students' metacognitive strategies, providing reasons for behaviour patterns or academic performance.

It is an important role of research, whatever the focus, to explore avenues beyond those already possible, but in the specific case of Big Data, learning analytics and education, this is an imperative necessity. Several examples (Chaps. 3, 5, 6, 7 and 8) in the contributions to this book show how fascinating can be the results of taking such a position for both the design of educational tasks, assessments, and components of online course environments, and the design of their use based on the analysis of data collected from participants' actions on online learning platforms, which reflect the effect of essential elements (e.g., online video lectures, teaching materials, video recordings, authenticity and engagement of the students with the tasks, online video lectures, interaction of the students with the unit coordinator, suitable technological tools, interactivity of online video lectures materials) on students' performance.

In fact, the aforementioned research aims are visions of the possible roles of Big Data in educational pedagogy and research that this book tries to build. The many facets of research and design work that contribute to our understanding of learning and teaching processes on online platforms, with the derived Big Data from and for educational activities, can be organised into four groups as per Kenett and Prodromou's chapter (monitoring, diagnostics, prognostics and prescriptive analysis) and improving our understanding of these four uses in education can further the evolution of mathematics education.

I would like to add that it is not enough to propose a critical and insightful reflection analysing what has been achieved before the coronavirus pandemic, from which the research of this book was drawn. In order to develop a vision and show possible ways for making it reality, we should also make this vision widely accessible beyond the community of researchers to all those who are professionally interested in education and whose contribution and support is needed. This review of "Big Data in Education: Pedagogy and Research" aims at making possible productive exchanges and collaboration between all those who are part of the interdisciplinary communities that use Big Data and data analytics in their professional educational practices and settings. The diversity of those communities and educational settings reflects the richness and diversity of expertise and the affordances of accumulating Big Data and analytics from rich, diverse resources. This book aims at being a source of insights for those communities. This is all the more important regarding the theme of this book for which competencies and creativity will be considered to manage Learning Management System (LMS) or museum data

management (the API MET Project) or sentiment analysis (the MBD Qatar Project) or metacognitive behavioural analysis and other professional tools. I am confident that the researchers and education will face this challenge successfully.

One particular challenge that necessitates the use of Big Data that offer new and powerful tools for supporting and accompanying the professional development of teachers, is the development of collaboration. From this point of view, it would be insightful when all the teams (statistics/mathematics teachers, computer scientists, etc.) working in this area are aware of the dynamics of their team and acknowledge that the quality of Mathematics/statistics educators' preparation and their professional development are the key of any possible evolution of the role of Big Data in education.

Not only does Big Data present an intense complexity, but it also requires complex skills and tools to be integrated into teaching practices. However, such complexity is not related to the complex systems that have the capacity to auto-organise and structure in bottom-up processes under favourable conditions. Sets of Big Data are too large or complex to be dealt with by traditional data-processing application software and cannot be organised in bottom-up processes.

Data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Therefore, Big Data often includes quantities of data that exceed the capacity of traditional software to process within an acceptable time. Current usage of the term Big Data tends to refer to the use of predictive analytics, user behaviour analytics (Chaps. 5, 6, 7 and 8), or certain other advanced data analytics methods that extract value from big data (Chaps. 5 and 6), and seldom to a particular size of data set. Analysis of data sets can find new correlations to "spot characteristics of students' performance trends" (Chaps. 5, 6, 7 and 8), spot which date is more appropriate to submit assignments (Chap. 5), identify students' interaction with the activities of a learning platform, and so on. Scientists, business executives, medical practitioners, advertising, and governments alike regularly meet difficulties with large data sets when dealing with healthcare analytics, environmental research, geographic information, urban informatics, and business informatics. With patience, coherent evolutions of technology in the long run can be achieved. Also with regards to complexity, Big Data and the evolution of the data ecosystem bring new challenges related to cooperation and collaboration work.

Finally, in this book it was emphasised that global cooperation is required in meeting possible challenges. Digital technologies can support and foster cooperation and collaborative work between students or between Mathematics/statistics educators and also between teachers and researchers. The consequences for such collaboration that this can have on students' learning processes, on the evolution of teachers' practices is certainly one essential technological evolution that educational research has to systematically explore in the future. As mentioned above, most of this space is still nearly terra incognita for researchers and educators. There is great excitement about the power that Big Data that can provide us with predictive analytics, user-behaviour analytics, and especially the affordances of Big Data that enable us windows for learning how and what our students do to learn while



interacting with activities on learning platforms, sending sms with their mobiles, or otherwise engaging in learning activities.

But we should also be careful. Cooperation and collaborative work when dealing with big data does not necessarily imply better learning, or better teaching. There is an avalanche of information, data, programming methods, tools, statistical skills, algorithms, tools for network access, and possibilities of connections, and only through careful work to organise, and process this avalanche, can we transform it into useful knowledge and avoid being overwhelmed.

## References

- Hewson, P., & MacGillivray, H. (2016). Big data and little people. *Teaching Statistics*, 38(3), 75–76.
- Prodromou, T., & Dunne, T. (2017). Statistical literacy in data revolution era: Building blocks and instructional dilemmas. *Statistics Education Research Journal*, 16(1), 38–43.
- Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, 84(3), 528–549.
- Shafir, U., & Etkind, M. (2018). *Concept parsing algorithms (CPA) for textual analysis and discovery: Emerging research and opportunities*. IGI Global.
- Shafir, U., & Kenett, R. S. (2016). Concept science evidence-based MERLO learning analytics. In E. Raïlean, G. Walker, A. Elçi, & L. Jackson (Eds.), *Handbook of research on applied learning: Theory and design in modern education* (Vol. 1, pp. 334–357). IGI Global.

**Dr. Theodosia Prodromou** is a senior lecturer of Mathematics Education at the University of New England in Australia. She has published numerous journal articles, book chapters and edited/ authored books on a wide range of subjects, including the use of big data in educational settings, the relationship between technology and mathematical thinking, the integration of digital technologies in the teaching and learning of mathematics, STEM education, augmented reality in Educational settings, professional development of mathematics teachers, and statistical thinking including perceptions of probability and chance. Her most recent edited/authored books were, *Augmented reality in educational settings* (2020) and *Primary and Middle Years Mathematics: Teaching Developmentally* (2019).