Chapter 3
A Big Data Primer

Judith J. Warren

Abstract The aim of this chapter is to describe the history of big data and its characteristics—variety, velocity, and volume—and to serve as a big data primer. Many organizations are using big data to improve their operations and/or create new products and services. Methods for generating data, how data is sensed, and then stored, in other words data collection, will be described. Mobile and internet technologies have transformed data collection for these companies and new sources are emerging at an unheard of speed. Due to the explosion of data, the teams needed to manage the data have evolved to include data scientists, domain experts, computer scientists, visualization experts, and more. The ideas of intellectual property are also changing. Who owns the data, the products generated from the data, and applications of the data? Challenges and tools for data analytics and data visualization of big data will be described, thus, setting the foundation for the rest of the book.

Keywords Big data • Data science • Data scientist • Data visualization • Digitization • Datafication • Privacy risks • Hadoop • NoSQL • Internet of Things

3.1 What Is Big Data?

"Big Data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, or organizations, the relationship between citizens and governments, and more" (Mayer-Schonberger and Cukier 2013). Big data occurs when the size of the data becomes the major concern for the data analyst. New methods of data

J.J. Warren, Ph.D., R.N., F.A.A.N., F.A.C.M.I. (✉)
University of Kansas School of Nursing, Kansas City, KS, USA

Warren Associates, LLC, Plattsmouth, NE, USA
e-mail: jjwarren@live.com

C.W. Delaney et al. (eds.), *Big Data-Enabled Nursing*, Health Informatics,
DOI 10.1007/978-3-319-53300-1_3

collection have evolved where the data input becomes passive. The data can come from social network posts, web server logs, traffic flow sensors, satellite imagery, audio streams, online searches, online purchases, banking transactions, music downloads, uploaded photographs and videos, web page content, scans of documents, GPS location information, telemetry from machines and equipment, financial market data, medical telemetry, online gaming, athletic shoes, and many more. The volume of data is so large, so fast, and so distributed that it cannot be moved. With big data, the processing capacity of a traditional database is exceeded (Dumbill 2012a, b). Fortunately, new methods of storage, access, processing, and analysis have been developed.

Big data has transformed how we analyze information and how we make meaning in our world. Three major shifts in our thinking occur while dealing with big data (Mayer-Schonberger and Cukier 2013). The first shift is about sampling data from all the data or the population for analysis to understand our world. With big data, we no longer need to sample from the population. We can collect, store and analyze the population. Due to innovations in computer memory storage, server design, and new software approaches, we can analyze all the data collected about a topic rather than be forced to only look at a sample. For all of analytic history, from the cave paintings in Lascaux to record the movements of animal herds (Encyclopedia of Stone Age Art 2016) to the cuneiform tablets used to record harvest and grain sales (Mark 2011) to statistical formulas from the 1700s and 1800s used to describe behavior (Stigler 1990), we were only able to collect, record, and analyze a sample of the population. Collecting data was a manual process and thus very labor intensive and expensive. [See Box 3.1].

Determining a sample and collecting a limited data set was the answer to this labor-intensive data collection process. Statistical sampling helped to select a representative set of data and to control for error in measurement. Now the innovations in computer science, data science, and data visualization create an opportunity to analyze all the data. As Mayer-Schonberger and Cukier summarize, n = all (2013). Analyzing all the data facilitates exploring subcategories/submarkets—to see the variations within the

Box 3.1 The Domesday Book of 1086

William the Conqueror mandated a tally of English people, land, and property to know what he ruled and how to assess taxes. Scribes were sent across England to interview and collect information about his subjects. It took years to collect and analyze the data. It was the first major census of its kind and served to document and datafy people's rights to property and land, and the ability to give military service. The book was used to award titles and land to worthy individuals. It has been used over a thousand years to settle disputes. It was last used in a British Court in 1966. The United Kingdom's National Archives have datafied the Domesday Book (National Archives 2016) and have digitized it so that it may be searched (Domesday Book Online 2013).

population. Google used their large database based on millions of Google searches and the Centers for Disease Control's (CDC) flu outbreak database to develop an algorithm that could predict flu outbreaks in near real-time (Google Flu Trends 2014). Google collected 50 million common search terms and compared with CDC data of spread of seasonal flu between 2003 and 2008. They processed 450 million different math models using machine learning to create an algorithm. Then Google compared the algorithm's prediction against actual flu outbreaks in 2007 and 2008. The model, comprised of 45 search terms, was used to create real time flu reporting in 2009. Google searches are powered by the big data effort to find connections between web pages and the search engine. As the searches are conducted and new connections are made, data is created. Amazon uses big data to recommend books and products to their customers. The data is collected every time a customer searches and purchases new books and products. These companies use big data to understand behaviors and make predictions about future behavior. This increased sophistication in the analysis and use of that data created the foundation of data science (Chartier 2014). Data science is based on computer science, statistics, and machine learning-based algorithms. With the advent of the Internet and the Internet of Things, data is collected as a byproduct of people seeking online services that is recorded as digitized behavior to be analyzed. Digitization is so pervasive that in 2010, 98% of the United States economy was impacted by digitization (Manyika et al. 2015).

The second shift in thought created by big data is the ability to embrace the messiness of data, to eliminate the need to be perfect without error (Mayer-Schonberger and Cukier 2013). Having the population of data means that the need for exactness in a sample lessens. With less error from sampling, we can accept more measurement error. Big data varies in quality since collection is not supervised nor controlled. Data is generated by online clicks, computerized sensors, likes and rankings by people, smart phone use, or perhaps credit card use. The messiness is managed through the sheer volume of data—the population (n = all). Data is also distributed among numerous data warehouses and servers. Bringing the distributed data together for analysis has its own challenges with exactness. Combining different types of data from different sources causes inconsistency due to different formatting structures. Cleaning this messiness in the data has led to evolution of a new role in big data—the data wrangler.

The third shift in thought created by big data is to move to thinking only about correlation, not causality (Mayer-Schonberger and Cukier 2013). Yet, mankind has a need to understand the world and jumps to thinking in causal terms to satisfy this need. In big data, the gold is in the patterns and correlations that can lead to novel and valuable insights. The use of big data and data science doesn't reveal WHY something is happening, but reveals THAT something is happening. Our creative need to combine data sets and to use all the data to create new algorithms for understanding the data leads us away from thinking of a dataset developed for a single purpose to thinking about what does the value of this dataset have by itself and in combination with other datasets (Chartier 2014). Data becomes a reusable resource, not a static collection point in time. A note of caution about correlation: very large data sets can lead to ridiculous correlations. Interpretation

of results needs to be investigated by a domain expert to insure an analysis that truly leads to knowledge and insight. The focus on correlation creates data-driven decisions instead of hypothesis-drive decisions.

3.1.1 Datafication and Digitization

To understand the innovation of big data, data itself needs to be explored. What makes data? The making of data occurred when man first measured and recorded a phenomenon. Early man in Mesopotamia counted grain production, recorded its sale, and analyzed it to calculate taxes owed to the king. So to datafy a phenomenon is to measure it and put it into a quantified format so that it can be tabulated and analyzed. Datafication made it possible to record human activity so that the activity can be replicated, predicted, and planned. Modern examples of datafication are email and social media where relationships, experiences, and moods are recorded. The purpose of the Internet of Things is to datafy everyday things.

With the advent of computers, we can also digitize our data. Digitization turns analog information into a format that computers can read, store, and process. To accomplish this, data is converted into the zeros and ones of binary code. For example, a scanned document is datafied but once it is processed by optical character recognition (OCR), it becomes digitized. The Gartner reports 4.9 billion connected things are currently in use in 2015 and by 2020, 25 billion connected things will be in use (Gartner 2014). These connected objects will have a "digital voice" and the ability to create and deliver a stream of data reflecting their status and their environment. This disruptive innovation radically changes value proposition, creates new services and usage scenarios, and drives new business models. The analysis of this big data will change the way we see our world.

3.1.2 Resources for Evaluating Big Data Technology

With the disruptive changes of big data, new products and services are needed for the storage, retrieval, and analysis of big data. Fortunately, companies are creating reports that list the services available and their penetration in the marketplace. Consumers new to the field should study these reports before investing in new servers, software, and consultants. Both Gartner and Forrester have rated the products of companies engaged in big data hardware, software, and consulting services. Both of these consulting firms provide a service to consumers by providing information on the status of big data as a new trend that is making an impact in industry.

Gartner has rated big data on their Magic Quadrant (2016a, b). The magic Quadrant is a two-by-two matrix with axes rating the ability to execute and the completeness of vision. The quadrants where the products are rated depict the challengers, leaders, niche players, and visionaries. The quadrant gives a view of market

competitors and how well they are functioning. Critical Capabilities is a deeper dive into the Magic Quadrant (Gartner 2016a). The next tool is the famous Gartner Hype Cycle (2016c). The axes are visibility vs. maturity. The graph formed depicts what is readily available and what is still a dream, thus informing of where the hype and adoption are for the trend. The five sections of the graph or the lifecycle of the trend are technology trigger, peak of inflated expectations, trough of disillusionment, slope of enlightenment, and plateau of productivity. Big data has a Hype Cycle of its own that breaks out the components and technologies of big data (GilPress 2012).

Forrester rates products on the Forrester Wave (Forrester Research 2016; Gualtieri and Curran 2015; Gualtieri et al. 2016; Yuhanna 2015). The Wave is a graph with two axes, current offering vs. strategy. Market presence is then plotted in the graph using concentric circles (waves) to show vendor penetration in the market. This information helps the customer to select the product best for their purpose.

3.2 The V's: Volume, Variety, Velocity

Big data is characterized by the "Three V's." The three V's can be used to understand the different aspects of the data that comprise big data and the software platforms needed to explore and analyze big data. Some experts will add a fourth "V", value. Big data is focused on building data products of value to solve real world problems.

3.2.1 Volume

Data volume is quantified by a unit of storage that holds a single character, or one byte. One byte is composed of eight bits. One bit is a single binary digit (1 or 0). Table 3.1 depicts the names and amounts of memory storage. In 2012, the digital

Table 3.1 Names and amounts of memory storage

Name	Symbol	Binary measurement	Decimal measurement	Number of bytes	Equal to
Kilobyte	KB	2^{10}	10^3	1024	1024 bytes
Megabyte	MB	2^{20}	10^6	1,048,576	1024 KB
Gigabyte	GB	2^{30}	10^9	1,073,741,824	1024 MB
Terabyte	TB	2^{40}	10^{12}	1,099,511,627,776	1024 GB
Petabyte	PB	2^{50}	10^{15}	1,125,899,906,842,624	1024 TB
Exabyte	EB	2^{60}	10^{18}	1,152,921,504,606,846,976	1024 PB
Zettabyte	ZB	2^{70}	10^{21}	1,180,591,620,717,411,303,424	1024 EB
Yottabyte	YB	2^{80}	10^{24}	1,208,925,819,614,629,174,706,176	1024 ZB

universe consisted of one trillion gigabytes (1 zettabyte). This amount will double every two years and, by 2020, will consist of 40 trillion gigabytes (40 zettabytes or 5200 gigabytes per person) (Mearian 2012).

As data storage has become cheaper, as predicted by Moore's Law, the ability to keep everything has become a principle for information technology (Moore 2016). In fact, it is sometimes easier and cheaper to keep everything than it is to identify and keep the data of current interest. Big data is demonstrating that the reuse and analysis of all data and the combinations of data can lead to new insights and new data products that were previously not imagined. Some examples will demonstrate the volume of data that exists in big data initiatives:

- Google processes 24 petabytes of data per day. A volume that is thousands of times the quantity of all printed material in the US Library of Congress (Gunelius 2014)
- Facebook users upload 300 million new photos every hour; the like button or comment is used three billion times a day (Chan 2012)
- YouTube has over one billion users who watch hundreds of millions of hours of video per day (YouTube 2016).
- Twitter has over 100 million users log in per day; with over 500 million tweets per day Twitter Usage Statistics 2016).
- IBM estimates that 2.5 quintillion bytes of data (2.3 trillion gigabytes) is created daily; 90% has been created in the last two years (IBM 2015, 2016).

First, different approaches to storing these very large data sets have made big data possible. The foremost tool is Hadoop that efficiently stores and processes large quantities of data. Hadoop's unique capabilities support new ways of thinking about how we use data and analytics to explore the data. Hadoop is an open-source distributed data storage and analysis platform that can be used on large clusters of servers. Hadoop uses Google's MapReduce algorithm to divide a large query into multiple smaller queries. MapReduce then sends those queries (the Map) to different processing nodes and then combines (the Reduce) those results back into one query. Hadoop also uses YARN (Yet Another Resource Negotiator) and HDFS (Hadoop Distributed File System) to complete its processing foundation (Miner 2016). YARN is a management system that keeps track of CPU, RAM, and disk space and insures that processing runs smoothly. HDFS is a file system that stores data on multiple computers or servers. The design of HDFS facilitates a high throughput and scalable processing of data. Hadoop also refers to a set of tools that enhance the storage and analytic components: Hive, Pig, Spark, and HBase are the common ones (Apache Software Foundation 2016). Hive is a SQL-like query language for use in Hadoop. Pig is also a query language optimized for use with MapReduce. While Spark is a framework for general purpose cluster computing, HBase is a data store that runs on top of the Hadoop distributed file storage system and is known as a NoSQL database. NoSQL databases are used when the volume of data exceeds the capacity of a relational database. To be able to engage in big data work, it is essential that these tools are understood by the entire big data team (Grus 2015).

3.2.2 Variety

The data in big data is characterized by its variety (Dumbill 2012a, b). The data is not ordered, due to its source or collection strategy, and it is not ready for processing (characteristics of structured data in a relational database). Even the data sources are highly diverse: text data from social networks, images, or raw data from a sensor. Big data is known as messy data with error and inconsistency abounding. The processing of big data uses this unstructured data and extracts ordered meaning. Over 80% of data is unstructured or structured in different formats. Initially, data input was very structured, mostly using spreadsheets and data bases, and collected in a way for analytics software to process. Now, data input has changed dramatically due to technological innovation and the interconnectedness of the Internet. Data can be text from emails, texting, tweets, postings, and documents. Data can come from sensors in cars, athletic shoes, bridge stress, mobile phones, pressure readings, number of stairs climbed, or blood glucose levels. Data from financial transactions such as stock purchases, credit cards, and grocery purchases with bar codes. Location data is recorded via the global position satellites (GPS) residing on our smart phones know where they are and communicate this to the owners of the software. Videos and photographs are digitized and uploaded to a variety of locations. Digitized music and speech are shared across many platforms. Mouse clicks are recorded for every Internet and program use (think of the number of times you are asked if the program can use your location). Hadoop and its family of software products have been created to explore these different unstructured data types without the rigidity required by traditional spreadsheet and database processing.

3.2.3 Velocity

The Internet and mobile devices have increased the flow of data to users. Data flows into systems and is processed in batch, periodic, near real time, or real time (Soubra 2012). Before big data, companies usually analyzed their data using batch processing. This strategy worked when data was coming in at a slow rate. With new data sources, such as social media and mobile devices, the data input speed picks up and batch processing no longer satisfies the customer. So as the need for near real time or real time data processing increases, new ways of handling the data velocity come into play. However, it is not just the velocity of incoming data, but the importance of how quickly the data can be processed, analyzed and returned to the consumer who is making a data-driven decision. This feedback loop is critical in big data. The company that can shorten this loop has a big competitive advantage. Key-value stores and columnar databases (also known as NoSQL databases) that are optimized for the fast retrieval of precomputed information have been developed to satisfy this need. This family of NoSQL databases was created for when relational databases are unable to handle the volume and velocity of the data.

3.3 Data Science

3.3.1 What Is Data Science?

The phrase data science is linked with big data and is the analysis portion of the innovation. While there is no widely accepted definition of data science, several experts have made an effort. Loukides (2012) says that using data isn't, by itself, data science. Data science is using data to **create a data application** that acquires the value from the data itself and creates more data or a **data product**. Data science combines math, programming, and scientific instinct. Dumbill says that big data and data science create "the challenges of massive data flows, and the erosion of hierarchy and boundaries, will lead us to the statistical approaches, systems thinking and machine learning we need to cope with the future we're inventing" (2012b, p. 17). Conway defines data science using a Venn diagram consisting of three overlapping circles. The circles are math (linear algebra) and statistical knowledge, hacking skills (computer science), and substantive expertise (domain expertise). The intersection between hacking skills and math knowledge is machine learning. The intersection between math knowledge and expertise is traditional research. The intersection between expertise and hacking skills is a danger zone (i.e., knowing enough to be dangerous and to misinterpret the results). Data science resides at the center of all the intersections (Conway 2010). O'Neil and Schutt add the following skills to their description of data science: computer science, math, statistics, machine learning, domain expertise, communication and presentation skills, and data visualization (2014). Yet, the American Society of Statistics weigh in on data science by saying it is the technical extension of statistics and not a separate discipline (O'Neil and Schutt 2014). The key points in thinking about data science, especially in arguing for a separateness from statistics, are mathematics and statistical knowledge, computer science knowledge, and domain knowledge. A further distinction about data science is that the product of engaging in data science is creating a **data product** that feeds data back into the system for another iteration of analysis, a practical endeavor not traditional research. A more formal definition of data science proposed by O'Neil and Schutt is, "a set of best practices used in tech companies, working within a broad space of problems that could be solved with data" (2014, p. 351).

3.3.2 The Data Science Process

The data science process closely parallels the scientific process while including a feedback loop. Each step of the process feeds to the next one but also has feedback loops. First, the real world exists and creates data. Second the data is collected. Third, the data is processed. In the fourth step data cleaning occurs and feeds into machine learning/algorithms, statistical models, and communication/visualization/reports. The fifth step is exploratory analysis but also feeds back into data

collection. The sixth step is creating models with machine learning, algorithms, and statistics but also feeds into building a data product. The seventh step is to communicate the results, develop data visualizations, write reports and feeds back into decision making about the data. The eighth step is to build a data product. This data product is then released into the real world, thus closing the overall feedback loop.

A data scientist collects data from a multitude of big data sources as described in the previous section on Variety. However, the data scientist needs to have thought about the problem of interest and determine what kind of data is needed to find solutions for the problem or to gain insight into the problem. This is the step that uses Hadoop and its associated toolbox-HDFS, MapReduce, YARN, and others. This data is unprocessed and cleaning it for analysis consumes about 80% of the data scientist's time (Trifacta 2015). Programming tools, such as Python, R, SQL, are used to get the data ready for analysis. This cleaning and formatting process is called data munging, wrangling, joining, or scraping the data from the distributed databases (Provost and Fawcett 2013; Rattenbury et al. 2015). Common tools for this process, other than programming language, are Beautiful Soup, XML parsers, and machine learning techniques. Quality of the data must also be assessed, especially handling missing data and incongruence of data. Natural language processing tools may be used for this activity. Once the data is in a desired format, then analysis, interpretation, and decision-making using the data can occur.

The data scientist can then begin to explore the data using data visualization and sense-making of the data (the human expertise). The beginning step, keeping in mind the problem of interest, in working with the data is to conduct an exploratory data analysis (EDA; Tukey 1977). Graphing the data helps to visualize what the data is representing. The analyst creates scatterplots and histograms from different perspectives to get "a feel" for the data (Jones 2014). The graphs will help to know how and what probability distributions (curves plotted on an x and y axes) to calculate as the data is explored (remember to look at correlation and not causality). EDA may reveal a need for more data, so this becomes an iterative process. Experience determines when to stop and proceed to the next step. A firm grasp of linear algebra is essential in this step.

Next, use the data to "fit a model" using the parameters or variables that have been discovered (this uses statistical knowledge). Caution, do not overfit the model (a danger zone event described by Conway 2010). The model is then optimized using one of the two preferred programming languages in data science: Python and R. Python is usually preferred by those whose strength is in computer science, while R is preferred by statisticians. MapReduce may also be used at this step. Algorithms, from statistics, used to design the model may be linear regression, Naïve Bayes, k-nearest neighbor, clustering, and so forth. The algorithm selection is determined by the problem being solved: classification, cluster, prediction, or description. Machine learning may also be used at this point in analysis and uses approaches from computer science. Machine learning leads to data products that contain image recognition, speech recognition, ranking, recommendations, and personalization of content.

The next step is to interpret and visualize the results (data visualization will be discussed later in the chapter). Communication is the key activity in this step.

Informal and formal reports are written and given. Presentations are made to customers and stakeholders about the implications and interpretations of the data. Presentation skills are critical. Remember in presenting complex data, a picture is worth a thousand words. Numerous tables of numbers and scatterplots confuse and obscure meaning for the customer. A visual designer can be a valuable member to design new data visualization approaches or infographics (Knaflic 2015).

The final step is to create a data product from the analysis of the data and return it to the world of raw data. Well known products are spam filters, search ranking algorithms, or recommendation systems. A data product may focus on health by collecting data and returning health recommendation to the individual. Research productivity may be communicated through publications, citations of work, and names of researchers following your work as ResearchGate endeavors to do (ResearchGate is an online community of nine million researchers; 2016). As these products are used in the big data world, they contribute to ongoing data resources. The data science process creates a feedback loop. It is this process that makes data science unique and distinct from statistics.

3.4 Visualizing the Data

Data visualization has always been important for its ability to show at a glance very complicated relationships and insights. Data represents the real world but it is only a snapshot covering a point in time or a single time series. Visualization is an abstraction of the data and represents its variability, uncertainty, and context in a way that the human brain can apprehend (Yau 2013). Data visualization occurs prominently in three steps of the data science process: step four data cleaning, step five exploratory data analysis, and step seven communication (O'Neil and Schutt 2014). Graphing and plotting the data in step four depicts outliers and anomalies that a data wrangler may want to explore to see if there are issues with the data. The issues could be with format, missing data, or inconsistencies. During exploratory data analysis, the graphing of data may demonstrate insights, inconsistencies, or the need for more data. In step seven the results of the data science project are communicated requiring more complex graphs that depict multiple variables.

In designing a visualization of data, there are four components to consider (Yau 2013). The first component is the use of visual cues to encode the data in the visualization. The major cues are shape, color, size, and placement in the visualization. The second component is selecting the appropriate coordinate system. There are three main systems from which to choose: a Cartesian system (x and y axes), a polar system (points are on a radius at an angle, Nightingale used this in her graphic of British solder deaths in the Crimean War), or a geographic system (maps, longitude, latitude). The third component is the use of scale defined by mathematical functions. The most common scales are numeric, categorical and time. The last component is the context that helps to understand the who, what, where, when, and why of

the data. Data must be interpreted in context and the visualization must demonstrate this context to the viewer. Doing data visualization well is understanding that the task is to map the data to a geometry and color thus creating a representation of the data. The viewer of the data visualization must be able to go back and forth between what the visual is and what it represents—to see the pattern in the data.

A good visual designer and data scientist follow a process to develop the visualization (Yau 2013). As with the analysis process the team must have some questions to guide the visualization process. First the data collected and cleaned must be graphed to enable the team to know what kind of data they have. Tools, such as Excel, R (though R is a programming language, it can generate graphics as well), Tableau, or SAS, can be used to describe the data with scatterplots, bar charts, line graphs, pie charts, polar graphs, treemaps, or other basic ways to display data. This step must be continued until the team "knows" the data they have. The second step is to determine what you want to know about the data. What story do you want to tell with the data? The third step is to determine the appropriate visualization method. The nature of the data and the models used in analysis will guide this step. The data must be visualized with these assumptions and the previous four components of visualization design in mind. The last step is to look at the visualization and determine if it makes sense. This step may take many iterations until the visualizations convey the meaning of the data in an intuitive way to the viewer or customer of the analysis. Using a sound, reproducible process for creating the visualization insures that the complexity and art of creating representations of data are accurate and understood.

Three of the most well-known data visualizations of all times are Nightingale's Mortality in the Crimean War (ims5 2008; Yau 2013), Minard's Napoleon March on Moscow (Sandberg 2013; Tufte 1983); and Rosling's Gapminder (2008; Tableau 2016). These visualizations depict multiple variables and their interrelationships. They demonstrate well thought-out strategies for depicting data using more than simple graphs. Nightingale invented the coxcomb graph to depict the causes of death in the Crimean war. The graph displays time, preventable deaths, deaths from wounds, and death from other causes. Her graph is said to be the second best graph ever drawn (Tableau 2016). Minard depicted Napoleon's march on Russia by displaying geography, time, temperature, the course and direction of the army, and the number of troops remaining. He reduced numerous tables and charts into one graphic that Tufte (1983) called the best statistical graph ever drawn. Rosling's bubble graph depicts the interaction between time, income per person, country, and life expectancy. The great data visualizations go beyond the basic graphing approaches to depict complex relationships within the data (Tufte 1990, 1997).

3.5 Big Data Is a Team Sport

Doing data science requires a team as no one person can have the all the skills needed to collect, clean, analyze, model, visualize, and communicate the data. Teams need to have technical expertise in a discipline, curiosity with a need to

understand a problem, the ability to tell a story with data and to communicate effectively, and the ability to view a problem from different perspectives (Patil 2011). When pulling together a team consider the following people: programmers with skill in Python, R, and other query languages; database managers who can deploy and manage Hadoop and other NoSQL databases; information technology (IT) professionals who know how to manage servers, build data pipelines, data security, and other IT hardware; software engineers who know how to implement machine learning and develop applications; data wranglers who know how to clean and transform the data; visual designers who know how to depict data that tell a story and to use visualization software; scientists who are well versed in crafting questions and searching for answers; statisticians who are well versed in developing models, designing experiments, and creating algorithms; informaticians who understand data engineering; and experts in the domain being explored (O'Neil and Schutt 2014).

The team must determine who has which skills and how to collaborate and enhance these skills to create the best data product for the organization. The organizational culture must be one that supports and embraces data science to its fullest for the greatest success (Anderson 2015; Patil and Mason 2015). As organizations begin to use data science in their product development initiatives, a certain level of data science maturity is required. Guerra and Borne have identified ten signs of a mature data science capability (2016). A mature data science organization makes all data available to their teams; access is critical and silos are not allowed. An agile approach drives the methodology for data product development (The Agile Movement 2008). Crowd sourcing and collaboration are leveraged and promoted. A rigorous scientific methodology is followed to insure sound problem solving and decision making. Diverse team members are recruited and given the freedom to explore; they are not micromanaged. The teams and the organization ask the right questions and search for the next question of interest. They celebrate a fast-fail collaborative culture that encourages the iterative nature of data science. The teams show insights through illustrations and storytelling than encourage asking "what if" questions that require more that simple scatterplots and bar charts. Teams build proof of values, not proof of concepts. Developing proof of value focuses on value leading to solving the unknowns, not just that it is a good thing to do. Finally, the organization promulgates data science as a way of doing things, not a thing to do. Data science drives all functions in the organization and shifts how organizations operate.

As with any team that develops products, intellectual property is a key concern. When data is used to develop data products, the ethics of data ownership and privacy become critical issues. Traditional data governance approaches and privacy laws and regulations don't completely guide practice when big data is ubiquitous and practically free. With big data, no one organization owns all the data they need. New models of collaboration and data sharing are emerging. As these models evolve, new questions emerge about data ownership, especially if it

is collected as a byproduct of conducting business—banking, buying groceries, searching the Internet, or engaging in relationships on social media. Ownership of this type of data may not always be clear. Nor is it clear who can use and reuse the data. Numerous questions arise from exploring this grey area. If the data is generated from a transaction, who controls and owns that? Who owns the clicks generated from cruising the Internet? To add to this confusion, consumers are now wanting to control or prevent collection of the data they generate—the privacy issue (Pentland 2012). There have been situations where individuals have been re-identified from anonymized data. The White House, in response to this concern, has drafted a Consumer Privacy Bill of Rights Act (2015). The draft acknowledges a "rapid growth in the volume and variety of personal data being generated, collected, stored, and analyzed." Though the use of big data has the potential to create knowledge, increase technological innovation, and improve economic growth, big data has the potential to harm individual privacy and freedom. The bill urges that laws must keep current with technology and business innovation. As the practice of using big data and data science becomes more mainstream, the ethical issues and their solutions will appear. The Data Science Association has a Code of Conduct for their members (2016). This Code speaks to conflict of interest, data and evidence quality, and confidentially of the data. For a good discussion of the ethics of big data, read Martin's (2015) article on the ethical issues of the big data industry.

3.6 Conclusion

In January 2009, Hal Varian, the chief economist at Google, said in an interview that, "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—is going to be a hugely important skill in the next decades…" (Yau 2009). Varian says this skill is also important for elementary school, high school and college kids because data is ubiquitous and free. The ability to understand that data and extract value from it is now a scarce commodity. Varian believes that being a data scientist and working with big data will be a 'sexiest' job around. The Internet of Things has created disruption in the way we think about data as it is coming at us from everywhere and interconnected. In a period of combinatorial innovation, we must use the components of software, protocols, languages, capabilities to create totally new inventions. Remember, however, that big data is not the solution. Patterns and clues can be found in the data, yet have no meaning nor usefulness. The key to success is to decide what problem you want to solve, then use big data and data science to help solve the problem and meet your goals (Dumbill 2012). Finally, big data is just one more step in the continuation of mankind's ancient quest to measure, record, and analyze the world (Mayer-Schonberger and Cukier 2013).

Case Study 3.1: Big Data Resources—A Learning Module

Judith J. Warren and E. LaVerne Manos

Abstract This case study is a compilation of resources for a learner to explore to gain beginning knowledge and skill in big data, data science, and data visualization. The resources focus on acquiring knowledge through books, white papers, videos, conferences, and online learning opportunities. There are also resources for learning about the hardware and software needed to engage in big data.

Keywords Big data • Data science • Data visualization • Data wrangling • Hadoop/mapreduce • Data analytics • Data scientist • Data science teams • Volume/variety/velocity of data sets • Data products

3.1.1 Introduction

The volume, variety, and velocity of big data exceed the volume of datasets common in health care research and operations. New technologies created to manage and analyze big data are being developed and tested at a rapid rate. This life cycle process is happening so fast that it is difficult to learn the technology and approaches much less keep up on the latest innovations. The phases of this life cycle are development, testing, discarding, testing, adopting, combining, using and discarding/reworking. These phases transpire in swift iterative cycles, and data scientists who utilize the tools work with a toolbox composed of well-developed software to niche software designed for specific uses, many of which are open source.

Today we are overwhelmed with an unprecedented amount of information and data. Big data comes from all kinds of sources: global positioning devices (GPS), loyalty shopping cards, online searches and selections, genomic information, traffic and weather information, health data from all sorts of personal devices (person generated health data), as well as data created from healthcare during inpatient and outpatient visits. Data is collected every second of every day. These types of data, including unstructured raw data, have been used in other industries to understand their business and create new products. Healthcare has been slower to adopt the use of big data in this way. The 2013 report by McKinsey Global Institute proposes that the effective use of big data in healthcare could create large value for the healthcare industry, over $300 billion every year (Kayyali B, Knott D, Van Kuiken, S. The big-data revolution in US health care: Accelerating value and innovation. Mc Kinsey & Company. 2013. Accessed at http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care).

J.J. Warren, Ph.D., R.N., F.A.A.N., F.A.C.M.I. (✉)
University of Kansas School of Nursing, Kansas City, KS, USA

Warren Associates, LLC, Plattsmouth, NE, USA

E. LaVerne Manos, D.N.P., R.N.-B.C.
University of Kansas School of Nursing, Kansas City, KS, USA

The effective use of big data requires a data science approach to find and analyze subsets of data that administrators, clinicians, and researchers will find usable. Unlike a relational database where writing a query is fairly straight forward, gathering data from multiple data stores/warehouses of big data is much more complex. The ability to manage an incoming data stream of extraordinary volume, velocity and variety of data requires the expertise of a team. This case study provides a beginning resource for learning about big data, data science, and data visualization.

3.1.2 Resources for Big Data

As big data has caught the imagination of corporations and health care, the resources have exploded and most are readily available on the Internet. The following resources have been selected for learners who are just beginning their exploration of big data and a few that will stretch their knowledge towards competence. As you do your own searches, you will find many more. This listing will get you into the field and Internet space to find more resources that fit your learning style.

3.1.2.1 Big Data Conferences

Conferences are good places to explore a new field or gain more understanding of a field with which you have expertise. Networking is key at these events and can link you to others for future project work. These are just the tip of the iceberg of conferences, so enjoy looking for new ones near you.

1. In 2013, the University of Minnesota School of Nursing convened the first conference called Nursing Knowledge: Big Data Science, http://www.nursing.umn.edu/icnp/center-projects/big-data/index.htm. The first conference was invitational and explored the potential of big data for the improvement of patient outcomes as the result of nursing care. The conference was so successful, it has been held annually and been open to all registrants. Nursing Knowledge: Big Data Science is a working conference with many workgroups creating projects that are making an impact in Nursing research, education, and practice.
2. "Big Data 2 Knowledge" hosted by the National Institutes of Health (NIH) also has conferences, training sessions, and webinars. These events are geared towards creating a research cohort that is expert in big data and data analytics.
3. The Strata + Hadoop World Big Data conference is a meeting where business decision makers, strategists, architects, developers, and analysts gather to discuss big data and data science. At the conference you explore big data and hear what is emerging in the industry (http://conferences.oreilly.com/strata/hadoop-big-data-ca). O'Reilly Media and others put on this conference and afterwards post all the presentations to their web site. So even if you can't attend, you can hear about cutting-edge big data.

3.1.2.2 Big Data Books and Articles

A tried and traditional way to learn about any knowledge is through books, journal articles, and white papers. The following are basic references to get you started in the big data initiative.

1. Anderson C. Creating a Data-Driven Organization. Sebastopol, CA: O'Reilly Media; 2015. http://shop.oreilly.com/product/0636920035848.do
2. Betts R, Hugg, J. Fast Data: Smart and at Scale. Sebastopol, CA: O'Reilly Media; 2015. https://voltdb.com/blog/introducing-fast-data-smart-and-scale-voltdbs-new-recipes-ebook
3. Brennan PF, Bakken S. Nursing needs big data and big data needs nursing. Journal of Nursing Scholarship, 2015;47: 477–484.
4. Chartier, T. Big Data: How Data Analytics Is Transforming the World. Chantilly, VA: The Great Courses. 2014. (includes video lectures). (http://www.thegreat-courses.com/courses/big-data-how-data-analytics-is-transforming-the-world.html)
5. Davenport T, Dyche J. Big data in big companies. 2013. http://www.sas.com/reg/gen/corp/2266746. Accessed 15 Dec 2015.
6. Mayer-Schonberger V, Cukier K. Big Data: A revolution that will transform how we live, work, and think. New York: Hought Mifflin Harcourt Publishing. 2013.
7. O'Reilly Radar Team. Planning for big data: A CIO's handbook to the changing data landscape. Sebastopol, CA: O'Reilly Media. 2012. http://www.oreilly.com/data/free/planning-for-big-data.csp
8. O'Reilly Team. Big data now. Sebastopol, CA: O'Reilly Media. 2012.
9. Patil DJ, Mason H. Data driven: Creating a data culture. Sebastopol, CA: O'Reilly Media. 2015. http://datasciencereport.com/2015/07/31/free-ebook-data-driven-creating-a-data-culture-by-chief-data-scientists-dj-patil-hilary-mason/#.Vp6x33n2bL8

3.1.2.3 Big Data Videos

For those who need to see and hear, videos are great. Below are some from YouTube and other websites. Don't forget to look at Tim Chartier's work listed in the Books section. Great Courses combine expert faculty, a book, and video lectures. Explore TED Talks for more information about Big Data.

1. Big Data Tutorials and TED Talks, http://www.analyticsvidhya.com/blog/2015/07/big-data-analytics-youtube-ted-resources/
2. Kenneth Cukier: Big data is better data, https://www.youtube.com/watch?v=8pHzROP1D-w
3. The Secret Life of Big Data | Intel, https://www.youtube.com/watch?v=CNoi-XqwJnA (a good overview of the history of Big Data, a must watch)
4. What is Big Data? https://www.youtube.com/watch?v=c4BwefH5Ve8

5. What is BIG DATA? BIG DATA Tutorial for Beginners, https://www.youtube.com/watch?v=2NLyIqU-xwg
6. What Is Apache Hadoop? http://hadoop.apache.org/
7. What is Big Data and Hadoop? https://www.youtube.com/watch?v=FHVuRxJpiwI
8. What Does The Internet of Things Mean? https://www.youtube.com/watch?v=Q3ur8wzzhBU
9. MapReduce, https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

3.1.2.4 Big Data Web Sites

Many companies, with a web site, provide information, free books, white papers, tutorials, and free trial software. These sites are a rich resource. Gartner and Forrester are companies that evaluate and rate emerging companies and products in the big data industry.

1. IBM Big Data and Analytics platform, now known as IBM Watson Foundations, http://www.ibmbigdataanalytics.com
2. Forrester Wave, https://www.forrester.com/The+Forrester+Wave+Big+Data+Hadoop+Distributions+Q1+2016/fulltext/-/E-res121574#AST1022630
3. Gartner, http://www.gartner.com/technology/research/methodologies/research_mq.jsp
   (a) Magic Quadrant
   (b) HypeCycle
   (c) Critical Capabilities
4. Intel Processors, http://www.intel.com/content/www/us/en/homepage.html
   (a) The Butterfly Dress, https://www.youtube.com/watch?v=6ELuq3CzJys (a bit of fun with data and technology)
   (b) 50th Anniversary of Moore's Law, http://newsroom.intel.com/docs/DOC-6429 (if you are in informatics, you must know about Moore's Law)
   (c) How Intel Gave Stephen Hawking his Voice, http://www.wired.com/2015/01/intel-gave-stephen-hawking-voice; https://www.youtube.com/watch?v=JA0AZUj2lOs
5. Kaggle, www.kaggle.com
6. O'Reilly Media, https://www.oreilly.com/topics/data
7. SAS, http://www.sas.com/en_us/insights/big-data.html
8. VoltDB, https://voltdb.com
9. Yuhanna, N. (August 3, 2015). The Forrester Wave: In-Memory Database Platforms, Q3 2015. http://go.sap.com/docs/download/2015/08/4481ad9e-3a7c-0010-82c7-eda71af511fa.pdf
10. Zaloni. http://www.zaloni.com/health-and-life-sciences

3.1.3 Resources for Data Science

Data science is composed of data wrangling and data analysis. Data wrangling is the process of cleaning and mapping data from one "raw" form into another format. Then algorithms can be applied to make sense of big data. The following resources have been selected for learners who are just beginning their exploration of data science and a few that will stretch their knowledge towards competence. As you do your own searches, you will find many more. This listing will get you into the field and Internet space to find more resources that fit your learning style.

3.1.3.1 Data Science Conferences

Conferences are good places to explore a new field or gain more understanding of a field with which you have expertise. Networking is key at these events and can link you to others for future project work.

1. "Big Data 2 Knowledge" hosted by the National Institutes of Health (NIH) also has conference, training sessions, and webinars. These events are geared towards creating a research cohort that is expert in Big Data and Data Analytics.
2. The Data Science Conference, http://www.thedatascienceconference.com.

3.1.3.2 Data Science Books and Articles

A tried and traditional way to learn about any knowledge is through books, journal articles, and white papers. The following are basic references to get you started in Data Science.

1. Ghavami, PK. Clinical Intelligence: The big data analytics revolution in healthcare: A framework for clinical and business intelligence. CreateSpace Independent Publishing Platform. 2014.
2. Grus, J. Data science from scratch: First principles with python. Sebastopol, CA: O'Reilly Media. 2015.
3. Gualtieri M, Curran R. The Forrester Wave: Big data predictive analytics solutions, Q2, 2015. April 1, 2015. https://www.sas.com/content/dam/SAS/en_us/doc/analystreport/forrester-wave-predictive-analytics-106811.pdf
4. Janert, PK. Data analysis with open source tools: A hands-on guide for programmers and data scientists. Sebastopol, CA: O'Reilly Media. 2010.
5. Loukides, M. What is data science? Sebastopol, CA: O'Reilly Media. 2012.
6. Marconi K, Lehmann H. Big data and health analytics. Boca Raton, FL: CRC Press. 2015.
7. O'Neil C, Schutt R. Doing data science: Straight talk from the frontline. Sebastopol, CA: O'Reilly Media. 2015.
8. Optum. Getting from big data to good data: Creating a foundation for actionable analytics. 2015. https://www.optum.com/content/dam/optum/CMOSpark%20

Hub%20Resources/White%20Papers/OPT_WhitePaper_ClinicalAnalytics_ONLINE_031414.pdf
9. Patil DJ. Building data science teams: The skills, tools, and perspectives behind great data science groups. Sebastopol, CA: O'Reilly Media. 2011.
10. Provost F, Fawcett T. Data science for business: What you need to know about data mining and data-analytic thinking. Sebastopol, CA: O'Reilly Media. 2013.
11. Rattenbury T, Hellerstein JM, Heer J, Kandel S. Data wrangling: Techniques and concepts for agile analysts. Sebastopol, CA: O'Reilly Media. 2015.
12. Tailor K. The patient revolution: How big data and analytics are transforming the health care experience. Hoboken, NJ: John Wiley and Sons. 2016.
13. Trifacta. Six Core Data Wrangling Activities. 2015. https://www.trifacta.com/wp-content/uploads/2015/11/six-core-data-wrangling-activities-ebook.pdf. Accessed 15 Jan 2016.

3.1.3.3 Data Science Videos

For those who need to see and hear, videos are great. Below are some from YouTube and other websites. Explore TED Talks for more information about Big Data.

1. Analytics 2013—Keynote—Jim Goodnight, SAS, https://www.youtube.com/watch?v=AEI0fBQYJ1c
2. Big Data Analytics: The Revolution Has Just Begun, https://www.youtube.com/watch?v=ceeiUAmbfZk
3. Building Data Science Teams, https://www.youtube.com/watch?v=98NrsLE6ot4
4. Deep Learning: Intelligence from Big Data, https://www.youtube.com/watch?v=czLI3oLDe8M
5. The Future of Data Science—Data Science @ Stanford, https://www.youtube.com/watch?v=hxXIJnjC_HI
6. The Patient Revolution: How Big Data and Analytics Are Transforming the Health Care Experience, https://www.youtube.com/watch?v=oDztVSDUbxo

3.1.3.4 Data Science Web Sites

Many companies, with a web site, provide information, free books, white papers, tutorials, and free trial software. These sites are a rich resource.

1. Alteryx, http://www.alteryx.com.
2. Data Science at NIH, https://datascience.nih.gov/bd2k
3. IBM, http://www.ibmbigdataanalytics.com.
4. Kaggle-the Home of Data Science, https://www.kaggle.com
5. Python Programming Language, https://www.python.org/
6. R Programming language, https://www.r-project.org/about.html
7. SAS, https://www.sas.com/en_us/home.html.
8. Trifacta, https://www.trifacta.com/support.

3.1.4 Resources for Data Visualization

Data visualization is the third part of big data. Humans can absorb more data when it is depicted in images or graphs. The following resources have been selected for learners who are just beginning their exploration of data visualization and a few that will stretch their knowledge towards competence. As you do your own searches, you will find many more. This listing will get you into the field and Internet space to find more resources that fit your learning style.

3.1.4.1 Data Visualization Conferences

Conferences are good places to explore a new field or gain more understanding of a field with which you have expertise. Networking is key at these events and can link you to others for future project work. Most conferences on big data and data science include presentations on data visualization.

3.1.4.2 Data Visualization Books and Articles

A tried and traditional way to learn about any knowledge is through books, journal articles, and white papers. The following are basic references to get you started in the data visualization.

1. Beegel J. Infographics for dummies. Hoboken, NJ: John Wiley & Sons. 2014.
2. Few S. Now you see it: Simple visualization techniques for quantitative analysis. Oakland, CA: Analytics Press. 2009.
3. Harris RL. Information graphics: A comprehensive reference. Atlanta, GA: Management Graphics. 1996.
4. Jones B. Communicating data with tableau: Designing, developing, and delivering data visualization. Sebastopol, CA: O'Reilly Media. 2014. http://cdn.oreillystatic.com/oreilly/booksamplers/9781449372026_sampler.pdf
5. Knaflic CN. Storytelling with data: A data visualization guide for business professionals. Hoboken, NJ: John Wiley & Sons. 2015.
6. Tufte ER. Envisioning information. Cheshire, CN: Graphics Press. 1990.
7. Tufte ER. The visual display of quantitative information. Cheshire, CN: Graphics Press. 1983. (This is the classic text in visualization.)
8. Tufte ER. Visual explanations: Images and quantities, evidence and narrative. Cheshire, CN: Graphics Press. 1997.
9. Yau N. Data points: Visualization that means something. Indianapolis, IN: John Wiley & Sons; 2013.
10. Yau N. Visualize this: The FlowingData guide to design, visualization, and statistics. Indianapolis, IN: John Wiley & Sons; 2011.

3.1.4.3 Data Visualization Videos

For those who need to see and hear, videos are great. Below are some from YouTube and other websites. Explore TED Talks for more information about Big Data.

1. The beauty of data visualization, https://www.youtube.com/watch?v=5Zg-C8AAIGg
2. The best stats you've ever seen, https://www.youtube.com/watch?v=usdJgEwMinM
3. Designing Data Visualizations, https://www.youtube.com/watch?v=lTAeMU2XI4U
4. The Future of Data Visualization, https://www.youtube.com/watch?v=vc1bq0qIKoA
5. Introduction to Data Visualization, https://www.youtube.com/watch?v=XIgjTuDGXYY

3.1.4.4 Data Visualization Web Sites

Many companies, with a web site, provide information, free books, white papers, tutorials, and free software.

1. FlowingData, https://flowingdata.com.
2. SAS, http://www.sas.com/en_us/home.html
   (a) Data visualization and why it is important, http://www.sas.com/en_us/insights/big-data/data-visualization.html
3. Tableau, http://www.tableau.com/
   (a) Tableau. (2015). The 5 Most Influential Data Visualizations of All Time. http://www.tableau.com/top-5-most-influential-data-visualizations (note Florence Nightingale is the number two graph)
   (b) Visual Analysis Best Practices: Simple Techniques for Making Every Data Visualization Useful and Beautiful, http://get.tableau.com/asset/10-tips-to-create-useful-beautiful-visualizations.html
4. Trifacta, https://www.trifacta.com

3.1.5 Organizations of Interest

As the field of big data, data science and data visualization evolve, professional organizations will be formed. Listservs and blogs will be created. Academia will offer courses and degree programs. Certification and accreditation organizations will help to establish quality programs and individual performance. The following are just a sampling of what exists.

3.1.5.1 Professional Associations

Professionals will form professional organizations as they define their discipline. The organizations provide a forum for discussing practice, competencies, education, and the future.

1. American Statistics Association, http://www.amstat.org/
2. American Association of Big Data Professionals, https://aabdp.org/
   (a) Offers certification in various Big Data roles, https://aabdp.org/certifications.html
3. Data Science Association, http://www.datascienceassn.org/
4. Digital Analytics Association, http://www.digitalanalyticsassociation.org/

3.1.5.2 Listservs: A Sampling

Most web sites, organizations, industry, and publishers have listservs. This is a very efficient way to keep up with what is happening in these areas. The listserv is pushed to your email and enables you to see the latest thoughts, conferences, books, and software an industry that is evolving rapidly.

1. 10 Data Science Newsletters To Subscribe To, https://datascience.berkeley.edu/10-data-science-newsletters-subscribe
2. Information Management, http://www.information-management.com/news/big-data-analytics/Big-Data-Scientist-Careers-10026908-1.html
3. O'Reilly Data Newsletter, http://www.oreilly.com/data/newsletter.html. Sign up to get the latest information about Big Data, Data Analytics, Data Visualization, and Conferences.

3.1.5.3 Certificates and Training: A Sampling

As jobs in these fields become more widely available, the demand for these skills will grow. Online education and formal degrees will become important for employers to consider. Certification may make a difference for employment.

1. Data Science at Coursera, https://www.coursera.org/specializations/jhu-data-science
2. Data at Coursera, https://www.coursera.org/specializations/big-dataQ
3. SAS Certification program, http://support.sas.com/certify/index.html
4. MIT Professional Education, https://mitprofessionalx.mit.edu/about
5. R Programming, https://www.coursera.org/learn/r-programming

3.1.5.4 Degree Programs: A Sampling

Degree programs are proliferating as the demand for big data professionals and data scientists increases. It will be important to select well before investing time and money into the programs. Always look for programs that are accredited. The University/College must be accredited by the US Department of Education. Even the department/school they reside in must be accredited by the appropriate accreditor. Accreditation assures the quality of the education.

1. 23 Great Schools with Master's Programs in Data Science, http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science
2. Carnegie Mellon University, http://www.cmu.edu/graduate/data-science/
3. Harvard, http://online-learning.harvard.edu/course/big-data-analytics
4. List of Graduate Programs in Big Data & Data Science, http://www.amstat.org/education/bigdata.cfm
5. Map of University Programs in Big Data Analytics, http://data-informed.com/bigdata_university_map/
6. Northwestern Kellogg School of Management, http://www.kellogg.northwestern.edu/execed/programs/bigdata.aspx?gclid=CLTa_Jf5u8oCFYVFaQodCpwHag

3.1.6 Assessment of Competencies

Teachers and students have used Bloom's Taxonomy to create objectives that specify what is to be learned. The levels of Bloom can also be used to guide evaluation of the attainment of these objectives by the student. In 2002, Bloom's was revised to reflect cognitive processes as well as knowledge attainment (http://www.unco.edu/cetl/sir/stating_outcome/documents/Krathwohl.pdf). The new taxonomic hierarchy is as follows (Krathwohl, 2002, p215):

1. "Remember—retrieving relevant knowledge from long-term memory
2. Understand—determining the meaning of information
3. Apply—using a procedure in a given situation
4. Analyze—breaking material into its constituent parts and detecting the relationships between the parts and the whole
5. Evaluate—making judgements based on criteria
6. Create—putting elements together to form a coherent whole or make a product."

For big data and data science assignments the graduate student should be able to master the levels of "remember, understand, and apply" by engaging with the above resources. Objective assessments, in the form of tests, can then be used to determine

mastery. Performance assessments are used to evaluate the achievement of the higher levels of Bloom-- analyze, evaluate and create. Performance assessments are conducted by experts and faculty through the use of case studies, simulations, projects, presentations, or portfolios.

3.1.7 Learning Activities

The following are several learning activities designed to help you apply the knowledge and skills learned from the above resources. The Bloom level for each activity is listed.

1. Conduct a web search on HADOOP and data warehouses. What did you learn about big data? What are the issues in storing and accessing data that has volume, velocity, and variety? Define Oozie, PIG, Zookeeper, Hive, MapReduce, and Spark. How are they used in big data initiatives? (Bloom level—Understand)
2. A good source of data to practice wrangling, analysis and visualization is DATA.gov, http://www.data.gov. Download a file and then one of the free trial software packages and try different things. Trifacta lets you work on data wrangling. Excel can help with analysis. Tableau can help with visualization. Other sources of data are

   (a) https://r-dir.com/reference/datasets,
   (b) https://www.kaggle.com/datasets and
   (c) http://www.pewresearch.org/data/download-datasets.(Bloom level—Apply)

3. Take a data set and graph the data five different ways, e.g. scatter plot, histogram, radar chart, or other types of graphs. What insight did you get looking at the graphs? What analytic questions do you have that you would like to pursue based on the graphs? Were the graphs consistent? Was there one that represented the data best and why? (Bloom level—Analyze)
4. Keep a log of data that you personally generate through online use, mobile devices, smart phones, email, music, videos, pictures, financial transactions, and fitness/health apps. What format is this data in? Conduct an exploratory data analysis. Visualize the results several ways. Evaluate the visualizations using Yau's (2013) four components: visual cues, coordinate system, scale, and context. (Bloom level—Evaluate)
5. Create a list of keywords and a glossary for a document using Python. Download Python 3.4.4.msi (https://www.python.org/downmoads) and numpy-1.11.0.zip (http://www.numpy.org). Select a document and save it as a'.txt' file (if the name of the file contains a /U, then replace that with //U so the name will parse; Python uses /U as a code). Develop a Python script to determine word frequency in the document (http://programminghistorian.org/lessons/counting-frequencies). Wrangle the data so that only words are left ad remove stop words. From the remaining list select keywords and glossary words. (Bloom level—Create)

3.1.8 Guidance for Learners and Faculty Using the Module

This case study has provided learning resources for faculty and students to learn about big data, data science, and data visualization. The best strategy is to select some of the resources that best match your learning style—visual, audio, and tactile—and interact with them first. You may also want to use various search engines to search for other information about big data, data science, and data visualization. All online resources were accessed in January or February 2016. Download some programs and data and explore the process of wrangling, analysis and visualization.

References

The Agile Movement. 2008, Oct 23. http://agilemethodology.org. Accessed 25 Jan 2016.
Anderson C. Creating a data-driven organization. Sebastopol, CA: O'Reilly Media; 2015. http://shop.oreilly.com/product/0636920035848.do
Apache Software Foundation. Welcome to Apache Hadoop. 2016. http://hadoop.apache.org. Accessed 15 Jan 2016.
Chan C. What Facebook deals with every day: 2.7 billion likes, 300 million photos uploaded and 5—terabytes of data. 2012, Aug 22. http://gizmodo.com/5937143/what-facebook-deals-with-everyday-27-billion-likes-300-million-photos-uploaded-and-500-terabytes-of-data. Accessed 18 Jan 2016.
Chartier T. Big data: how data analytics is transforming the world. Chantilly, VA: The Great Courses (includes video lectures); 2014. http://www.thegreatcourses.com/courses/big-data-how-data-analytics-is-transforming-the-world.html
Conway D. The data science Venn diagram. 2010, Sept 30. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram. Accessed 30 Jan 2016.
Data Science Association. Code of conduct. 2016. http://www.datascienceassn.org/code-of-conduct.html. Accessed 15 Apr 2016.
The Domesday Book Online. 2013. http://www.domesdaybook.co.uk. Accessed 10 Jan 2016.
Dumbill E. What is big data? In: O'Reilly Team, editor. Big data now. Sebastopol, CA: O'Reilly Media; 2012a. p. 3–10.
Dumbill E. Why big data is big: the digital nervous system. In: O'Reilly Team, editor. Big data now. Sebastopol, CA: O'Reilly Media; 2012b. p. 15–7.
Encyclopedia of Stone Age Art: Lascaux Cave Paintings. 2016. http://www.visual-arts-cork.com/prehistoric/lascaux-cave-paintings.htm. Accessed 10 Jan 2016.
Forrester Research. 2016. https://www.forrester.com/home. Accessed 6 Jan 2016.
Gartner. Gartner says 4.9 billion connected "things" will be in use in 2015. 2014, Nov 11. http://www.gartner.com/newsroom/id/2905717. Accessed 10 Jan 2016.
Gartner. Gartner magic quadrant. 2016a. http://www.gartner.com/technology/research/methodologies/research_mq.jsp. Accessed 5 Jan 2016.
Gartner. Gartner critical capabilities. 2016b. http://www.gartner.com/technology/research/methodologies/research_critcap.jsp. Accessed 5 Jan 2016.
Gartner. Gartner hype cycle. 2016c. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp. Accessed 5 Jan 2016.
GilPress. Gartner's hype cycle for big data. 2012, Oct. https://whatsthebigdata.com/2012/08/16/gartners-hype-cycle-for-big-data. Accessed 5 Jan 2016.
Google Flu Trends. 2014. https://www.google.org/flutrends/about. Accessed 15 Jan 2016.

Grus J. Data science from scratch: first principles with python. Sebastopol, CA: O'Reilly Media; 2015.
Gualtieri M, Curran R. The Forrester Wave: big data predictive analytics solutions, Q2, 2015. 2015, Apr 1. https://www.sas.com/content/dam/SAS/en_us/doc/analystreport/forrester-wave-predictive-analytics-106811.pdf. Accessed 18 Jan 2016.
Gualtieri M, Yuhanna N, Kisker H, Curran, R, Purcell B, Christakis S, Warrier S, Izzi M. The Forrester Wave™: big data Hadoop distributions, Q1 2016. 2016, Jan 19. https://www.forrester.com/report/The+Forrester+Wave+Big+Data+Hadoop+Distributions+Q1+2016/-/E-RES121574#AST1022630, Accessed 25 Jan 2016.
Guerra P, Borne K. Ten signs of data science maturity. Sebastpol, CA: O'Reilly Media; 2016.
Gunelius S. The data explosion in 2014 minute by minute—infographic. 2014, Jul 12. http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic. Accessed 18 Jan 2016.
IBM. The four V's of big data. 2015. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg?cm_mc_uid=24189083104014574569048&cm_mc_sid_50200000=1457456904
IBM. What is big data? 2016. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html. Accessed 18 Jan 2016.
Ims5. Nightingale's coxcombs. 2008, May 11. http://understandinguncertainty.org/coxcombs. Accessed 5 Feb 2016.
Jones B. Communicating data with tableau: designing, developing, and delivering data visualization. Sebastopol, CA: O'Reilly Media; 2014. http://cdn.oreillystatic.com/oreilly/booksamplers/9781449372026_sampler.pdf
Knaflic CN. Storytelling with data: a data visualization guide for business professionals. Hoboken, NJ: Wiley; 2015.
Loukides M. What is data science? Sebastopol, CA: O'Reilly Media; 2012.
Manyika J, Ramaswamy S, Khanna S, Sazzazin, H, Pinkus, G, Sethupathy G, Yaffe A. Digital America: a tale of the haves and have-mores. 2015, Dec. http://www.mckinsey.com/industries/high-tech/our-insights/digital-america-a-tale-of-the-haves-and-have-mores. Accessed 15 Apr 2016.
Mark JJ. Cuneiform. In: Ancient history encyclopedia. 2011. http://www.ancient.eu/cuneiform. Accessed 10 Jan 2016.
Martin KE. Ethical issues in the big data industry. MIS Q Exec. 2015;14(2):67–85.
Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. New York: Hought Mifflin Harcourt Publishing; 2013.
Mearian L. By 2020, There will be 5,200 GB of data for every person on Earth. Computer World. 2012, Dec 11. http://www.computerworld.com/article/2493701/data-center/by-2020--there-will-be-5-200-gb-of-data-for-every-person-on-earth.html. Accessed 20 Feb 2016.
Miner D. Hadoop: what you need to know. Sebastopol, CA: O'Reilly Media; 2016.
Moore GE. Moore's Law. 2016. http://www.mooreslaw.org. Accessed 18 Apr 2016.
National Archives: Domesday book. 2016. http://www.nationalarchives.gov.uk/museum/item.asp?item_id=1. Accessed 10 Jan 2016.
O'Neil C, Schutt R. Doing data science: straight talk from the frontline. Sebastopol, CA: O'Reilly Media; 2014.
Patil DJ, Mason H. Data driven: creating a data culture. Sebastopol, CA: O'Reilly Media; 2015. http://datasciencereport.com/2015/07/31/free-ebook-data-driven-creating-a-data-culture-by-chief-data-scientists-dj-patil-hilary-mason/#.Vp6x33n2bL8
Patil DJ. Building data science teams: the skills, tools, and perspectives behind great data science groups. Sebastopol, CA: O'Reilly Media; 2011.
Pentland AS. Big data's biggest obstacles. Harvard Business Review Insight Center Report. The promise and challenge of big data supplement. 2012, Oct 2. p. 17–8.
Provost F, Fawcett T. Data science for business: what you need to know about data mining and data-analytic thinking. Sebastopol, CA: O'Reilly Media; 2013.

Rattenbury T, Hellerstein JM, Heer J, Kandel S. Data wrangling: techniques and concepts for agile analysts. Sebastopol, CA: O'Reilly Media; 2015.
ResearchGate. About us. 2016. https://www.researchgate.net/about. Accessed 30 Jan 2016.
Rosling H. Wealth and health of nations. 2008. http://www.gapminder.org/world. Accessed 25 Mar 2016.
Sandberg M. DataViz history: Charles Minard's flow map of Napoleon's Russian campaign of 1812. 2013, May 26. https://datavizblog.com/2013/05/26/dataviz-history-charles-minards-flow-map-of-napoleons-russian-campaign-of-1812-part-5. Accessed 25 Mar 2016.
Soubra D. The 3 Vs that define big data. 2012, Jul 5. http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data. Accessed 10 Jan 2016.
Stigler SM. The history of statistics: The measurement of uncertainty before 1900. Cambridge, MA: Belknap Press of Harvard University Press; 1990.
Tableau. The 5 most influential data visualizations of all time. 2016. http://www.tableau.com/top-5-most-influential-data-visualizations. Accessed 15 Jan 2016.
Trifacta. Six core data wrangling activities. 2015. https://www.trifacta.com/wp-content/uploads/2015/11/six-core-data-wrangling-activities-ebook.pdf. Accessed 10 Jan 2016.
Tufte ER. The visual display of quantitative information. Cheshire, CN: Graphcs Press; 1983.
Tufte ER. Envisioning information. Cheshire, CN: Graphics Press; 1990.
Tufte ER. Visual explanations: images and quantities, evidence and narrative. Cheshire, CN: Graphics Press; 1997.
Tukey JW. Exploratory data analysis. Boston: Addison-Wesley; 1977.
Twitter Usage Statistics. 2016. http://www.internetlivestats.com/twitter-statistics. Accessed 18 Jan 2016.
The Whitehouse. Draft consumer privacy bill of rights act. 2015. https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf. Accessed 30 Jan 2016.
Yau N. Google's chief economist Hal Varian on statistics and data. Jan 2009. https://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data. Accessed 5 Jan 2016.
Yau N. Data points: visualization that means something. Indianapolis, IN: Wiley; 2013.
YouTube. Statistics. 2016. https://www.youtube.com/yt/press/statistics.html. Accessed 18 Jan 2016.
Yuhanna N. The Forrester wave: in-memory database platforms, Q3. 2015, Aug 3. http://go.sap.com/docs/download/2015/08/4481ad9e-3a7c-0010-82c7-eda71af511fa.pdf