Chapter 6
The Transparency of Big Data, Data Harvesting and Digital Twins

Stefan Kendzierskyj, Hamid Jahankhani, Arshad Jamal, and Jaime Ibarra Jimenez

Abstract Computer storage and cloud computing has become more powerful with multiple algorithms running complex data analysis looking at intelligence trends, user behaviour, profiling and ways to make use of these outputs. Added with the artificial intelligence (AI) interaction has meant a new and dynamic method to create models forging analysis to be more clinical, proficient and continually seeking more improvement with the self-learning and intelligent programming of machine learning (ML). In the healthcare sector there is deep interest in collecting, curating the data and making the best use of silo'd data through methods such as blockchain. This can then lead to a multitude of innovations such as precision based medicine, targeting individual variability in genes, their environment, etc. It also means that big data analytics in healthcare is evolving into providing these insights from very large data sets and improving outcomes while reducing costs and inefficiencies. However, there also are some ethical impacts in the process of Digital Twins which can lead to segmentation and discrimination. Or perhaps the data that is automatically collected from healthcare sensors in IoMT and what type of governance are they scrutinized to. It is clear that data is the most important asset of not just an organisation but also to the individual and why the General Data Protection Regulation (GDPR) has taken an important stance in data protection by design and default, that all organisations needs to follow. This chapter aims to highlight some of the concerns.

Keywords Data Harvesting · Digital Twin · Big data · Transparency · eHealth · Healthcare · Social media

S. Kendzierskyj · H. Jahankhani (✉) · A. Jamal · J. I. Jimenez
London Campus, Northumbria University, London, UK
e-mail: Stefan.Kendzierskyj@northumbria.ac.uk; hamid.jahankhani@northumbria.ac.uk; arshad.Jamal@northumbria.ac.uk; jaime.jimenez@northumbria.ac.uk

H. Jahankhani et al. (eds.), *Blockchain and Clinical Trial*,
Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-030-11289-9_6

6.1 Introduction

The term Big Data has been around since the 1990s and describes the large volume, veracity and variety of data – structured, semi- structured or unstructured and may be too large to be handled by traditional databases, software technologies and methodologies. Approximately 80% of the data being processed daily today is unstructured where the data can come in any shape or form, so not containing a set record format. Examples of these can be documents, digital pictures, social media feeds, etc. Structured data is therefore arranged with data fields side-by-side in fixed lengths and contains a sequence of attributes and spreadsheets are a good example of this.

Due to some of the advances in processing power, storage, speed, types of devices, etc., has meant the quantity of data produced every day is at 2.5 quintillion bytes and with IoT accelerating will have a direct effect to more large data volume increase; Marr (2018). In the last 2 years 90% of the data in the world today was generated. IDC did some research for predictions by 2025 and some staggering numbers given as globally, by then, will be generating 163 zettabytes of data a year (1 zettabyte is equivalent to 1 trillion gigabytes), Cave (2017) and that is ten times the amount of data as currently is being created. Other IDC research ties in with the IoT explosion as mentioning the average connected person will interact with smart devices approximately 4800 times a day which equates to every 18 sec there is an interaction, Cave (2017). This points to a lot of behavioural analytics being generated on individuals and there are concern over this and its use cases from these analytics.

The primary goals of Big Data analytics are to help businesses unravel their data to make more evidenced based business decisions that can be critical. Ng et al. (2015) indicated that as the volume and variety of healthcare related data continues to grow, the analysis and use of this data will increasingly depend on the ability to appropriately collect, curate and integrate disparate data from many different sources. This means that Big Data analytics in healthcare is evolving into providing insight from very large data sets and improving outcomes while reducing costs.

Looking through past years and the advancement in data mining technology it’s clear that many organisations have made monetary and time investments into how to format metadata requirements and curate data and apply what is priority, etc. To remain competitive business executives needed to adopt the new technologies and techniques emerging due to Big Data. But as Big Data has become high profile for companies, and this could be competition, kudos, technological advancement, etc., there hasn’t been much done in the way of protecting the rights of the individual who owns the content of the data and what can be done to the data especially around the area of consent and this has led to a lot of people being concerned about how their data is accessed, what purposes, identity leakage and so on. These concerns, plus increasing data breaches, laid the foundation process for GDPR as it puts emphasis on organisations to proactively secure data and gain the consent of individuals before processing can commence and apply a lifecycle approach to data protection by design and default.

6.2 Big Data and Healthcare Impact

Data has and is still regarded as the most important asset of any organisation. Early adopters such as organisations like Google, Apple and Facebook have taken advantage of what data has to offer over the years especially when there was a realisation of what potential it holds around data mining which added value to the companies. Healthcare is a sector that can benefit from the use case of Big Data as it is a complex industry with many moving parts and with patients at the heart of it. Better health profiles and predictive modelling can be created to give much more precision based medicine with high rates of success in diagnosis, treatment and prevention. It is clear that patients need benefit from these better and improved outcomes and one way to do this is by having records and information digitalised in a way that can be easily analysed for any patterns, trends and preventative analysis. The predictive modelling tackles the complicated understanding around the biology of a disease so Big Data plays a part to aggregate lots of data components. As can patient health tracking on various vital health statistics through smart device and IoT. Characteristics can be monitored seamlessly and without relying on any patient memory call-back as all data is captured and automatically updated to cloud based systems. This does present another complex issue of data security and proposals of other methods such as blockchain can be introduced. Another benefit to having this type of patient tracking is keeping patients out of hospital since they can be monitored remotely in this way and this increase in analytics can have more positive behavioural effects on the patients as they figure out and interpret the importance of the smart wearable technology.

Fundamentally, the advent of the age of Big Data poses opportunities and challenges for industry. Previously unavailable formats of data can now be saved, retrieved and processed (but curation comes at a cost as to know what priority should be saved). Essentially data is being generated in a growing number of ways and therefore the use of traditional transaction databases have been supplemented by multimedia content, social media, and countless types of sensors. Big Data use is increasing rapidly, in that the world is changing and becoming an ever more digital space – compared to a few years ago, and today a lot is managed and shared online. The data that has been collected from smartphones, computers, devices, social media platforms such as Twitter, Facebook, Instagram, is then subsequently analysed, transmitted and reports driven on this data. Many observations are already drawn that this is only just the beginning; the evidence is overwhelming that this will increase, and data will be collected on just about everything. If data is compared to what was collected since the beginning of time up until the end of 2000; it will be significantly less than what is now collected in a minute (Marr 2018). Likely it is impossible to stop this journey of data accumulation since so much of the world is not connected and automatically analysed with sophisticated, artificial intelligence and machine learning algorithms.

6.2.1 Data Harvesting and Mining

A huge amount of data and wealth of information is regularly generated about our lives with or without knowing. A digital footprint is established that stays forever and mostly cannot be erased. Whilst individuals think on obvious places such as credit card, banking, purchases, social interactions etc., that builds this picture of preferences and routines, there is the aspect of data mining that is going on in the background that is more cause for concern. Forrester undertook research and output of a report in 2014, *Big Data's Big Meaning for Marketing*, and some highlights discussed by Kramer (2015) with regards to personal data protection, financial liabilities and ethical dilemmas. Methods of protecting individual's identity may not go far enough in the case of identity protection in the data mining process. Forrester outlines how Netflix released data after believing it had anonymised the data, but University of Texas researchers were able to identity Netflix users for anonymous reviews, but by knowing some parameters such as movies rented then it was possible to reverse-engineer the data and find out all viewing history; Pepitone (2010).

Whilst many would be in favour of healthcare providers mining data to ensure best placed precision based healthcare, where data mining is used to predict health needs; it might still have questions raised. This sounds good but could raise ethical questions on privacy invasion. An example of Carolinas Healthcare System who manages 900 care centres and purchase data collected from credit card purchases, store loyalty programs, etc., to allow identification of high-risk patients in attempts to intervene prevention on any health issues developing, Kramer (2015). This identification by medical practitioners would enable gaining insight into patients' lifestyles and habits. A risk score is used so doctors can see flagged up issues. The data is collected from credit card purchases, store loyalty programs, and other public records. In theory, medical practitioners can learn more about their patients—and their patients' lifestyles—from their shopping habits than from brief, or sometimes non-existent, consultations. Although the data doesn't yet identify individual purchases, it does provide a risk score doctors can use to highlight potential problems. The issue could become a more trust based issue between medical providers and patients if the data mining intrudes into the privacy and questions even healthy patients about their habits and digital footprints they leave. Or it may not take too long before insurance companies also start to review this mined data and risk score and that influences the service a patient receives, or worst case is refused if deemed too high a risk.

Or perhaps the case of Target, a retail organisation, that through a number of factors was able to identify and assign shoppers with a pregnancy prediction score (due to the array of 25 products when analysed together) and estimate a birth delivery due date to a small window. It allowed Target to provide coupons to the specific stages of pregnancy and highlighted a case of a dad who discovered his teen daughter was pregnant because Target mined her purchased data and sent her ads for baby products, Hill (2012).

6.2.2 Social Media Data Misuse

The widespread success of online social networking sites (OSNS) such as Facebook is a tempting resource for businesses engaged in electronic commerce.

Using personal information, willingly shared between online friends' networks, OSNS appear to be a natural extension of current advertising strategies such as word-of-mouth and viral marketing. However, the use of OSNS data for business marketing purposes has provoked outrage amongst social network users and highlighted issues over privacy. Within such environments, OSNS users disclose information that would be potentially rich sources of data mining for commercial organisations because it includes information that can personally identify an individual in rich detail (Krishnamurthy and Wills 2010). Such 'personally rich' information includes attributes such as name, location (city), telephone numbers, email addresses, photos, interests and purchases etc. This rich online social network data together with electronic word-of-mouth (eWOM) communications of OSNS users represents a tempting resource for viral and word-of-mouth marketing unlike other online and offline data which has to be prepared before systematically explored for patterns of use meaningful to commercial organisations (Kohavi et al. 2002; Zhang et al. 2011). A qualitative investigation of 861 blog comments from 715 individual online users were collected during the launch of Beacon, an unsuccessful third party marketing initiative by Facebook. Results show that business integrity, transparency of data use, user control, automatic disclosure and data leakage were key privacy concerns posing significant challenges to using business analytics in online social networks. However, attempts to leverage personal information and eWOM communications for commercial gain have provoked outrage amongst OSNS users because of privacy concerns. Privacy concerns of online social network users include use of personal information by unknown others for potential harmful purposes (e.g. by sexual predators), use and selling of personal information without notice and consent, access of personal information by unwanted audiences (Young and Quan-Haase 2009), involuntary disclosure of personal information, damaged reputation because of rumours and gossips, unwanted contact and harassment or stalking, third party use of personal information, and identity theft (Boyd and Ellison 2008). Consequently, privacy concerns challenge the classic thinking outlined by Kohavi and Provost (2010) that online (social) environments are particularly suitable domains for data mining because of the rich and large volume of data publicly available. Rather, issues of privacy concerns have emerged that overshadow the commercial potential of OSNS data (Hoadley et al. 2010) and highlight the boundaries of acceptance and use of business analytics in social networks. Privacy concerns have emerged as a critical factor determining the willingness, or not, of internet users to divulge personal information to online companies. Many studies have used 'privacy concern' construct to understand privacy in online contexts. Therefore 'privacy concern' has become a central construct to study privacy in information systems research. Likewise, it is a useful construct for business analytics because it provides theoretical guidance in defining and measuring privacy-related issues in the context of mining social network data for business marketing.

Facebook's personalised marketing tool "Beacon" was initially withdrawn by the Facebook due to users' backlash because of privacy concerns and ultimately shut down due to settlement of a lawsuit of $9.5 million. What should have been a successful innovation, however, was damaged and ultimately withdrawn because the nature and form of privacy concerns in OSNS was poorly understood.

6.3 Digital Twins in Healthcare: Transparency, Ethical Implications and Security Concerns

The term "Digital Twin" is an emerging engineering paradigm, which can occur in healthcare data-driven practices such as delivering personalized 3D printing as prosthesis for a surgery or dedicated manufacturing pieces for building or vehicle maintenance purposes, as examples. On the other hand, it involves conceptual and ethical implications, but what is really the concept of "Digital Twin"? It basically means the connection between the physical and digital world; the ability to visualise in computers, mobile devices or even in holographic projections what we are accustomed to watch daily. Looking back years ago, the film IronMan, and those scenarios where Tony Stark was able to manipulate and visualise the design of the new chemical element required in order to stop poisoning his body from using palladium for his arc reactor.

The Digital Twin takes the concept of Cyber-Physical Systems (CPS) to a higher level. This is because within an organisation, their assets can get digitalised (i.e., artefacts, devices, processes) and people can understand their behaviour, extract data in real-time in a 24/7 basis because of the implemented sensors across the asset, measuring physical values (e.g., voltage, pressure, amount of sugar). It requires to take advantage of Big Data analytics along with cloud computing systems in order to process tons of data in just seconds, making probabilistic approaches, math algorithms and establishing best options based on the calculations made. With all these features, organisations can make decisions and assess their results firstly over the digitalised assets extracting pros and cons, along with their associated risks prior to its implementation in the physical environment. This term stands for a paradigm, where individual physical assets are paired with digital models reflecting its status in a dynamic manner. The concept of Digital Twin has been applied by NASA for the development and monitoring of aerospace vehicles in order to last longer and tolerate extreme conditions compared to the Earth.

Modern engineering has provided a great support for the evolution of medicine. The establishment of mathematical models of patients, processing tons of "biodata" leading it to precise and effective medical interventions. Nowadays we count with supercomputers able to read molecular data making it possible to build personalised models, complemented by a continuous health and lifestyle tracking, resulting possibly in the "digital" representation of a patient – a "virtual patient". Therefore, the Digital Twin is an instrument showing the impact analysis of cutting-edge engineering solutions on core topics within healthcare such as health, disease, preventative care, and enhancement. It is claimed that many technical universities are training and preparing students in clinical technology, whilst doctors are working alongside with engineers from a wide range of backgrounds to enhance the functionality of modern medicine. Engineering standpoint and innovation drives a debate regarding human enhancement such as replacing broken parts of the body using 3D printed implants, arguing the possible; to enhance the human body with new capabilities. For instance, neural implants used for visual prosthetics addressed to blind people; however, it can lead towards capabilities that can get likely assessed beyond what is speculated as "normal" human sight providing access to parts that are considered normally inaccessible under the electromagnetic spectrum. The idea of digitalising molecular and physiological structure of people in order to deter whether the person is in healthy condition, estimate potential disease based on daily monitoring, measuring physical values such as heartbeat, blood pressure, sugar levels, etc., in order to elaborate the adequate medical prescription. In fact, it has been proven the efficacy of an approach done to pick the most appropriate drug for cancer treatment besides chemotherapy.

The concept of Digital Twin is used in industry to monitor the performance of artefacts and pieces of machinery in order to perform preventative maintenance. In fact, digitalisation of individual artefacts is simple because it is based on the instrumentation of electronic sensors placed across the artefact and besides, artefacts have an unique shape after its manufacturing, making easier the instrumentation. In healthcare otherwise, the human structure is more complex because of the constant molecular and physiological changes throughout their lives, making it complex to extract precise molecular data even though it is available the usage of wearable devices for medical purposes. Unfortunately, Digital Twin is still far from real and currently modern engineering approaches have reached digital models of genetic, biochemical, physiological and behavioural features of individuals. Therefore, the concept of Digital Twin offers a reliable instrument addressed to the impact analysis in healthcare because the usage of probabilistic models of individuals for customised medicines supports the engineering of a healthy condition and the advantage of big data to represent either a person or artefact.

To implement the concept of Digital Twin, it is necessary to differentiate heterogeneity when acquiring data over one's life time because in medicine the declaration of "healthy" or "normal" is done based on a population after clinical trials or following a pattern based on international medical committees. With the "digitalisation" of patients nonetheless, it must be required therefore a sharper statistical model in order to deter the declaration of normal or healthy status, and likewise of disease status and susceptibilities. An approach of Digital Twin in healthcare should rely on a detailed status of a healthy individual rather than basing it on diseased status records. In medicine, the declaration of healthy can reach to the state of "symptomless illness" and the biodata processed using probabilistic and statistical models within the concept of Digital Twin can allow doctors to infer the possibility of developing diseases. However, the engineering prototype coming along with Digital Twins brings into debate within medicine the optimal declaration of healthy – normal – status, carrying the question whether certain human features should be optimised or enhanced. Therefore, it is essential for decision making in healthcare to have the distinction between therapy, preventative care and enhancement, depending the condition of the patient.

The main purposes of enhancement in modern engineering actions are addressed to either restoring the functionality of a system or its modification. Digital Twins change the stereotype of existing engineering thanks to the elevated transparency of the status and performance of an artefact including the centrality of each one. Compared to medicine for instance the individual approach will impact the differentiation between therapy and enhancement because the declaration of normal or healthy status is often based on group or population statistics. However, with the concept of Digital Twin, individualisation is essential for its optimal functionality in healthcare. In addition, establishing an accurate digital model of a person would not be based on instrumentation for better decisions during healthcare interventions, but will also be part of the patient's identity. Digital Twins may therefore make doctors to review again what therapy should be considered when deploying personalised medicine. Digital Twins implies moral issues as well. For instance, depending on whether medical interventions are considered as daily treatment, therapy or enhancement, it can lead to different conclusions depending on what extent, which conditions and the public costs covered by the healthcare system.

Digital Twins brings great features and a significant contribution for the deployment of the hospital of future, because it can give a detailed account of molecular, physiological, phenotypic and lifestyle of people. It is considered currently an interesting conceptual tool which is worth to understand the technological trend in medicine along with a reflection on its future implementation, thereby the need of understanding the categories of health, disease and enhancement. Whilst the processing of data involves several biological aspects, along with behavioural data involving personality, manners, stress levels and also lifestyle of patients such as diet regimes, whether the patient does exercise, smoking, alcohol, etc. However, it is worth the analysis of some possible ethical and social implications of this trend. It is mandatory to understand that currently human beings are already using enhancement techniques. For instance, people can improve their lifestyle by performing exercise, a customised diet to increase muscular mass or stamina. The introduction of wearable devices and Wearable Body Area Networks (WBAN) allows real-time monitoring in a 24/7 manner delivering full-time support to the person in order to make fast decisions. A better lifestyle obtained by training and diet schemes might have the same results as the enhancements obtained by pharmaceutical treatment. Other arguments that could likely occur is the fear that human enhancement technologies might lead to people separation, having a disruptive effect on the current democracy, or the higher payments to afford this service as well because personalised medicine will increase the cost at individual level compared to traditional treatment.

In addition, Digital Twins can support the industry in the deployment of 3D printed organs used for prosthesis implementations such as replacing bones which are part of the spine given by a herniated disc or a heart transplant with the data extracted from the characteristics of the human being. A research from Lee Cronin from the University of Glasgow has demonstrated that chemical synthesis is possible using 3d printers in order to produce drugs to improve their lifestyle. The features mentioned bring concerns in terms of data privacy because the processing of tons of data measured in real-time along with the extraction of information from Electronic Patient Health Records (ePHR) and the data processed from wearable devices, makes the healthcare market prone to data theft and tampering. For instance, the modification of data while printing pills in order to synthesise illegal drugs such as cannabis or marijuana even though in some countries is considered legal. Data theft under a Digital Twin would lead to dedicated terrorist attacks terminating the life of the person by following different social engineering attacks. Therefore, there is importance of the need for strong governance frameworks and mechanisms addressed to ensure transparency on how Digital Twins are being used ensuring data privacy, integrity and availability, along with the protection of humans' rights and distribution of benefits given by the population's personal biological information.

6.4 Conclusions

Industry 4.0 (the fourth industrial revolution) sees the fast moving advances of the Internet of Things (IoT), big data, artificial intelligence (AI), more interaction in cyber-physical systems (CPS) and scalable cloud computing. Predictions by Gartner of more than 20 billion devices connected to IoT by 2020 are well understood, but more importantly will mean a huge volume of data will be generated.

Effects of data breaches, identity theft are widely known to take place in healthcare, but the more concerning impacts are how the data analysis or data harvesting are being utilised with numerous recent examples such as Cambridge Analytica, Google DeepMind App project with the NHS, and so on. Whilst it is understood big data analysis is needed and progressive to the requirements of Industry 4.0 there should, in tandem, be sanity checks on how the governance is developed on these more innovative technologies and their impacts studied.

References

Boyd D, Ellison NB (2008) Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 13:210–230

Cave A (2017) What will we do when the world's data hits 163 Zettabytes in 2025? Forbes. Available at: https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025/#25ed8bc9349a. Accessed 10 Nov 2018

Hill K (2012) How target figured out a teen girl was pregnant before her father did. Forbes. Available at: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#4ee09e726668. Accessed 11 Nov 2018

Hoadley MC, Xu H, Lee J, Rosson MB (2010) Privacy as information access and illusory control: the case of the facebook news feed privacy outcry. Electron Commer Res Appl (Special Issue on Social Networks and Web 2.0), 9(1):50–60

Kohavi R, Provost F (2010) Applications of data mining to electronic commerce. Data Mining and Knowledge Discovery, 5(1/2), 2001. Retrieved on September 15, 2010, from http://robotics.stanford.edu/~ronnyk/ecommerce-dm/editorial.pdf

Kohavi R, Rothleder NJ, Simoudis E (2002) Emerging trends in business analytics. Commun ACM 45(8):45–48

Kramer S (2015) The big risks of big data mining. V3B. Available at: https://v3b.com/2015/06/the-big-risks-of-big-data-mining/. Accessed 11 Nov 2018

Krishnamurthy B, Wills EC (2010) On the leakage of personally identifiable information via online social networks. SIGCOMM Comput Commun Rev 40(1):112–117

Marr B (2018) ‘how much data do we create every day? The mind-blowing stats everyone should read’. Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#2a7b7b3760ba. Accessed 10 Nov 2018

Ng K, Kakkanatt C, Benigno M, Thompson C, Jackson M, Cahan A, Zhu X, Zhang P, Huang P (2015) Curating and integrating data from multiple sources to support healthcare analytics. Stud Health Technol Inform 216:1056

Pepitone J (2010) 5 Data breaches: From embarrassing to deadly. CNN Money. Available at: https://money.cnn.com/galleries/2010/technology/1012/gallery.5_data_breaches/index.html. Accessed 11 Nov 2018

Young AL, Quan-Haase A (2009) Information revelation and internet privacy concerns on social network sites: a case study of Facebook, ACM

Zhang M, Jansen BJ, Chowdhury A (2011) Business engagement on Twitter: a path analysis. Electron Mark 21:161–175