Book chapter in: Sheryl Kruger Strydom and Moses Strydom, eds., Big Data Governance and Perspectives in Knowledge Management, 335 pages, IGI Global. Projected release date: November 2018.

Big Data Governance in Agile and Data-Driven Software Development:
A Market Entry Case in the Educational Game Industry

Lili Aunimo
Haaga-Helia University of Applied Sciences
Ari V. Alamäki
Haaga-Helia University of Applied Sciences
Harri Ketamo
Headai Ltd.

ABSTRACT
Constructing a big data governance framework is important when a company performs data-driven software development. The most important aspects of big data governance are data privacy, security, availability, usability and integrity. In this chapter, the authors present a business case where a framework for big data governance has been built. The business case is about the development and continuous improvement of a new mobile application that is targeted for consumers. In this context, big data is used in product development, in building predictive modes related to the users and for personalisation of the product. The main finding of the study a novel big data governance framework and that a proper framework for big data governance is useful when building and maintaining trustworthy and value adding big data driven predictive models in an authentic business environment.

Keywords: Big Data, Predictive Model, Data Governance, Data Privacy, Data Security, Data Availability, Data Usability, Data Integrity, Start-up Company, Market Entry, Educational Application, Personalisation

INTRODUCTION

A big data governance framework is critical when a company performs data-driven software development and business. The authors adhere to the common definitions of big data governance and of data governance, which were presented by Sarsfield (2009), Soares (2012), and DAMA (2017). To meet the requirements of an agile start-up company that needs to manage and govern big data based predictive models and the data related to them, the authors propose a framework for big data governance. This framework includes five key dimensions for big data governance: data privacy, security, availability, usability, and integrity. Each dimension is described in the big data governance framework section. These five key dimensions are important in building and managing a successful big data driven business.

In this chapter, the authors present the business case based on which the authors have derived the proposed big data governance framework. The business case concerns the development and continuous improvement of a new mobile educational application that is targeted at children and young people who wish to learn to play soccer. In this context, big data are used in three contexts: 1) product and business development, 2) building predictive modes about the learners' progress in general and 3) personalisation of the product to meet the needs of each individual learner.

Previous research and guidelines on how to govern big data governance (e.g., Soares, 2012; DAMA, 2017) have been published. However, there is a research gap concerning big data governance in agile big data driven start-up companies and especially on how to govern big data in the product development phase.

The main finding of the study is the proposed novel big data governance framework and the fact that a proper framework for data governance is necessary when developing trustworthy and value-adding big data driven software products in an authentic business environment. Without big data governance, the predictive models and other data-driven applications may not bring added value to the business because their trustworthiness is uncertain, they might violate the right to privacy of customers, or they might not be available for use when needed. In addition, without proper big data governance, they might not meet the needs of the business, or valuable data might be leaked to competitors because of the poor governance and weak security of the big data. If a company succeeds in big data governance, data can become its most valuable asset (Panian, 2010).

BACKGROUND

The field of big data governance emerged with the advent of big data. Big data are data that cannot be processed using traditional data processing software and infrastructure on a personal computer or on a dedicated server (e.g., Liebowitz, 2013). In addition, big data differ from traditional data in volume, variety, and/or velocity (Liebowitz, 2013). Typical examples of big data include web and social media data, machine-to-machine data, big transaction data, biometric data and human-generated data (Soares, 2012). The authors include Internet of Things (IoT) data and data generated in industrial processes in the broad set of transactional data.

Need for Big Data Governance in Companies

In technology companies, there is an increasing need to develop data analytics especially in launching new products and services. The diffusion of technology has increased rapidly (Downes & Nunes, 2013), increasing the need to analyse consumer behaviour in developing new products and services. Furthermore, such changes call for new models of data analytics in launching and developing new products and services in the rapidly changing digital markets. Companies that launch new mobile applications and other online services expect thousands of downloads, good user reviews in the market place, and less turnover in paid services. Thus, understanding consumer behaviour is essential for the entry of new products and services in the market, and big data analytics play a crucial role in this endeavour. To understand consumer behaviour in situations where they use the mobile application for the first time, multi-source data analytics provide richer information than a single data source does. Thus, the aim is to develop a model that combines user data and open data in the product development initiatives of new mobile services. This model could provide new data sources for personalization and predictive models about the users. In addition, it could provide open data-based opportunities for innovative ideas about product features, which then would create more value for consumers in using mobile applications.

Data Management and Data Governance

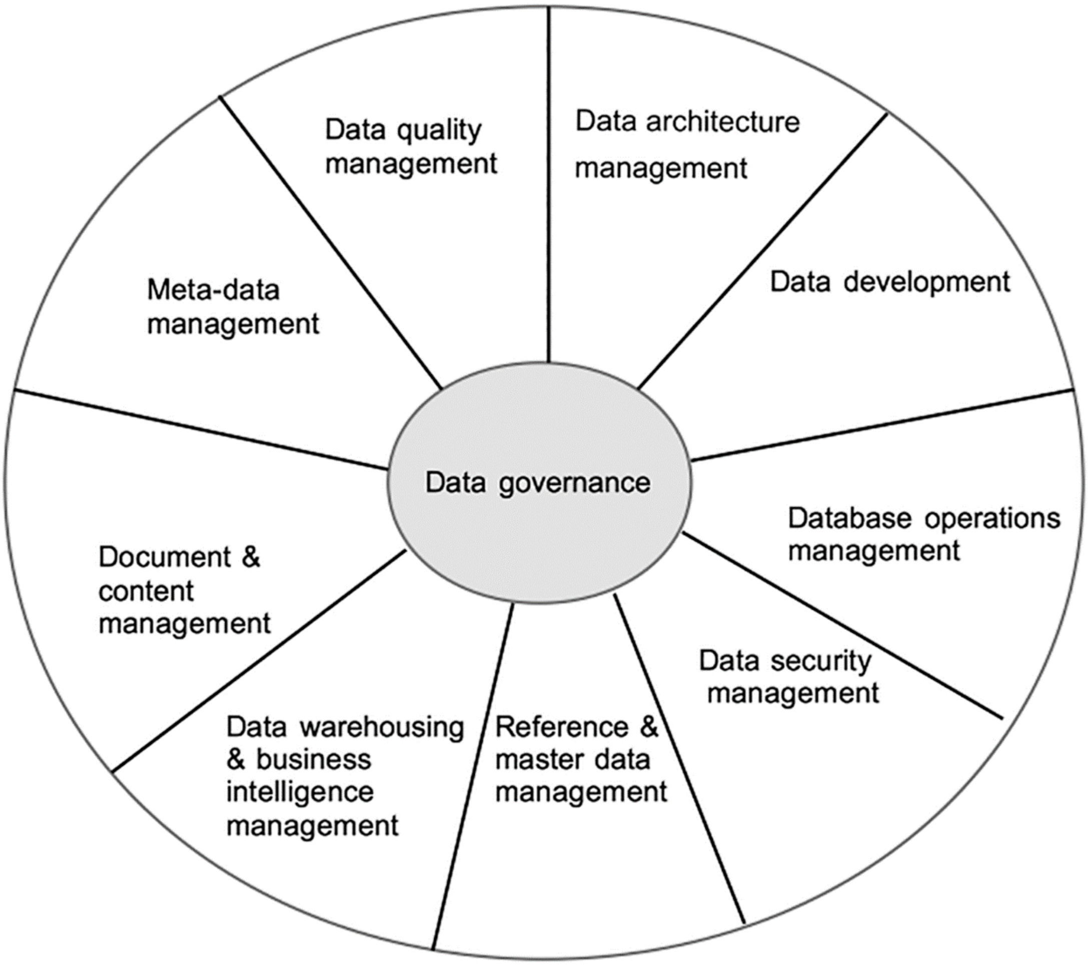
Many companies have established and mature procedures for data governance. Sarsfield (2009) defined data governance as set of processes that ensure the formal management of important data assets throughout the organisation. It guarantees that data can be trusted, and that people can be made accountable for any adverse event that happens because of poor data quality. The Data Management Body of Knowledge (DMBOK) provides the following concise definition of data governance:

The exercise of authority, control, and shared decision making (planning, monitoring and enforcement) over the management of data assets. (DAMA, 2017)

Although many companies have mature procedures for data governance, very few have any procedures for the governance of big data for two reasons: first, the field of big data is still immature; second, the existing big data applications are geared toward exploratory data analysis and discovery than toward traditional business intelligence. The latter reason has created a vicious circle: to be governed, data need to be modelled, and to be modelled, they need to be explored (du Mars, 2012). However, it is probable that big data, like any other company data, will soon be governed routinely. One indicator of this is that the Global Data Management Community (DAMA), has dedicated a subchapter to big data governance in its *International Guide to Data Management Body of Knowledge* (DAMA DMBOK, DAMA 2017, Chapter 14.6).

Data management is strongly related to data governance. DAMA International defines data governance as a part of data management. This relationship is shown in Figure 1

Figure 1. Data governance as defined by DMBOK


Data quality management
Data architecture management
Meta-data management
Data development
Data governance
Database operations management
Document & content management
Data security management
Data warehousing & business intelligence management
Reference & master data management

The nine disciplines of data management that are listed inside the sectors of the circle in Figure 1 are the following:

1. **Data architecture management**: Defining the overall process of managing all data assets in an organization.
2. **Data development**: Analysis, design, implementation, testing, deployment and maintenance of data.
3. **Database operations management:** Supporting all actions in the database lifecycle: from data acquisition to data integrity management.
4. **Data security management**: Ensuring privacy, confidentiality and appropriate access to data.
5. **Reference and master data management**: Acquiring and maintaining the relatively stable data concerning the business domain and customers.
6. **Data warehousing and business intelligence management:** Enabling reporting and analytics.
7. **Document and content management**: Managing data found outside of relational databases and data warehouses

8. **Meta-data management**: Integrating, controlling and providing descriptive information about the data assets.
9. **Data quality management**: Defining, monitoring and improving the correctness and trustworthiness of data

One could imagine that before big data governance can be established, an organisation should already have in place a framework for data governance. However, in start-up companies where big data plays a key role in their business, this is not the case. That kind of companies need to start by implementing a big data governance framework in their organisation.

Big Data in Software Development

In this study, special attention is put to the software development process from the big data perspective. The software development process is a multidimensional process where several factors affect to the success. Flyvbjerg and Budzier (2011) showed in their study how failures in the software development or implementation processes can cause significant damage for business or even fall entire companies. They showed that the overruns of development costs and schedules are not so fatal for the companies than damages that influence their business operations and customer satisfaction. Additionally, overrunning significantly budgets and schedules is not rare. Flyvbjerg and Budzier (2011) revealed that every sixth of large IT-projects overrun their budgets 200% and schedules 70 %. Thus, the successful management of mobile software projects is not only technological endeavor but it also deals with end user and business perspectives (Alamäki & Dirin, 2015). Alamäki and Dirin (2015) state that several stakeholders need to be involved to the mobile application development process. Developers and designers can evaluate the feasibility of technological features, business professionals validate features towards business needs, end users focus on the user experience and industry experts contribute to the business model of a new mobile application.

Big data provides new ways to improve product and service development and innovation (Paajanen, Valkokari & Aminoff, 2017; Tao et al. 2018). Advanced data collection and analytics help designers, developers and product managers monitor user experience and behavior of the potential users, and optimize decision making in investing to the development of new product features (Chen, Zhang & Zhao, 2017). Additionally, more effective data collection and analytics provide useful information to the decision makers of product life cycle management (Zhang, et al. 2017). Thus, big data analytics enhance design and development process in various areas of industries and service business.

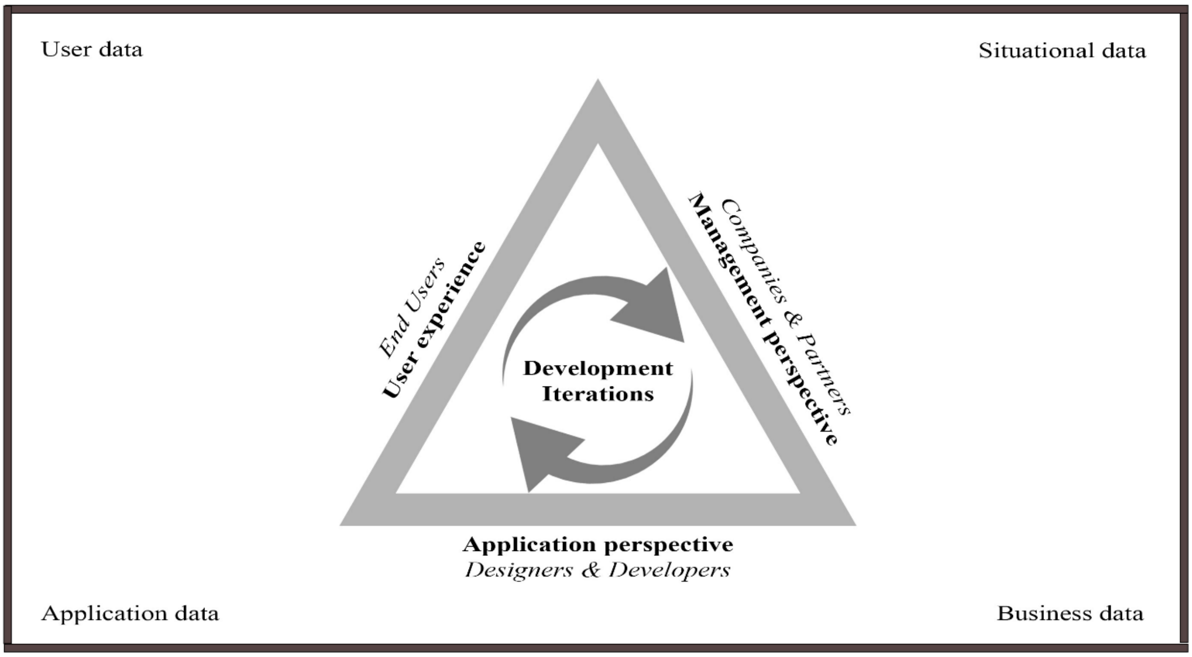


Figure 1. Data management perspectives in new application development (adapted from Alamäki & Dirin, 2015)

In figure 2, the authors summarize mobile development perspectives (Alamäki & Dirin, 2015) to the business risks of the IT-projects (Flyvbjerg & Budzier, 2011) and the principles of data-driven product and service development and management (Chen, Zhang & Zhao, 2017; Zhang, et al. 2017; Tao et al, 2018). Big data analytics provides useful information to various stakeholders of mobile application development, and each stakeholder has different perspective to the data. The figure 1 shows that mobile application development should focus on the several different data sources that the analysts can combine in searching, for example, patterns, segments and profiles. User data create information about the user experience of end users, such as usage time, user profiles, navigation paths, activities and likings. The application data generate device- and performance-related information, such as used end terminals, operating systems and IP-addresses. For example, data saved to the log files of the software systems can potentially provide both user and application data. The goals of analytics determine which category collected data belong to. The situational data show location where application is used and in which types of context. The business data assist managers to monitor customer purchase behavior, such as orders and rankings and business environment where a new product or service will compete. Each data source provides different types of information to the mobile development process

Agile big data driven software development and data governance are often seen as having completely different goals and that they cannot be applied in the same project. However, proper big data governance does enable an organization to remain agile (Panian, 2010). In an agile software development project also data governance should be done in an agile manner. In an agile software company, big data governance should be applied to the existing agile way of working and not as a completely new process (Seiner, 2014).

Privacy Issues of Personal Data

It is important to focus on users’ privacy concerns in designing and planning commercial mobile applications (Malhotra, Kim, & Agarwal, 2004). Several studies showed that privacy concerns, such as the feeling that privacy has been invaded or threatened, affect users’ online behaviour and willingness to use a mobile application (Ackerman, Cranor, & Reagle, 1999). Moreover, privacy issues have a legal dimension, which is codified in the European Union’s (EU) general data protection regulation (GDPR). The GDPR law will be applied on 25 May 2018 in EU countries, providing new rights for individuals to control how their data are collected, processed, and used by companies. The law includes strong sanctions for companies who contravene it. Furthermore, companies are required to demonstrate how they apply the principles of GDPR in their internal data management and privacy processes. In previous research on e-commerce, data privacy has been related to online trust (Bart et al., 2005). Trust between the application provider and users affects users’ willingness to provide their personal information to such providers. Trust is also essential in situations where users consider paying for the use of application or where they are eager to recommend an application to their friends (Dirin, Laine, & Alamäki, 2018). Hence, data privacy should be carefully considered in managing big data governance, as it affects to the trust of users.

The type of information that users provide when they access digital services could also affect their privacy (Acquisti & Gross, 2006). Thus, it is important to identify the type of information that companies need when they collect user data and develop policies according to such information. For example, Acquisti and Gross (2006) showed that users were more comfortable in providing general information, such as gender and age, than in revealing their home address and phone number. Information such as gender and age is easier to observe and obtain publicly than private information, such as a home address.

Security: Technologies and Processes

In addition to privacy, users consider security issues when they want to adopt new digital services and technologies (Wilkowska & Ziefle, 2012). The feeling that an application is secure and that it protects the privacy of users is closely related to the commercial success of the application. Distributed networks and data processing in cloud systems create security concerns for users. The simultaneous sharing of data on networks leads to users’ concerns about protecting their personal information (Chen & Zhao, 2012). Thus, the issue of security does not involve a single application provider, but it extends to communities that create policies regarding the standards and practices of the development and management of cloud systems (Kaufman, 2009). Thus, companies need to define how they have secured users’ personal information, not necessarily related to technology, but from the viewpoint of process management and documentation. Security issues are an essential dimension of process management in the big data governance policies of technology companies. In fact, problems in ensuring the security of users’ information have had negative consequences for the reputations of companies in which the user management servers have been hacked and users’ personal information has been published on the public Internet.

The management of big data security includes both processes and technological solutions that ensure the security of big data, which differs from the management of privacy because it includes only the data (including big data) assets of the company. In addition, this management is concerned about the company’s internal processes and the technological solutions used to ensure that each data asset has the correct level of security. The most important feature of data security is that only authorised people and software can access the data (Otto, 2011). Big data governance typically consists of both human-controlled procedures and software that automatically monitors the implementation of big data security as agreed by the organisation.

Availability of Data

Data availability means that authorised users have access to the data, software, and hardware upon request (Zissis & Lekkas, 2012). Thus, the availability of big data involves providing the right data to the right people and processes at a specific time such that when a piece of information is needed, the organisation can find it. The provision of big data has been included in defining the information architecture (Godinez et al., 2010; Morville, 2007) of a company or in designing enterprise searches (White, 2015). Information governance can be used to build metrics for assessing the availability of information in an organisation.

Usability: Value to the Business

The usability of big data means that the big data of an organisation must be in line with the organisation's business needs, the overall corporation strategy, and the corporate information strategy. Simply put, the usability of big data ensures that it can be monetized. Ensuring that data meets business needs is one of the core goals of data governance (Panian, 2010). It is not always easy to establish measures for defining the value for a business achieved by big data. For example, an organisation may gain value from big data because big data analytics helps it to build an accurate model of its customers. This model may be used to enhance the customers' experience of the company's products as well as in targeted marketing, among other uses. Although it is difficult to measure exactly the degree to which big data contributes in these endeavours, big data governance provides the appropriate tools to ensure that the big data processed by an organisation meets its business needs.

Integrity: Trustworthiness, Quality and Completeness

According to the principles of data integrity, only authorised parties can manage and modify data (Hashem et al., 2015). Thus, the aim of data integrity is to prevent the unauthorised use of data (Zissis & Lekkas, 2012) to ensure their trustworthiness and quality. This aim relates to the concept of accountability and professionalism of the management of data by companies. The term big data integrity means that the big data used by a company is trustworthy or that the organisation knows the extent to which the data may be trusted. Big data is not as trustworthy as traditional data in an organisation. This dimension of big data is called veracity (Ramesh et al., 2018). Big data often need to be cleaned to remove erroneous and unusable data. In some cases, data management must accept that the quality of the data is quite low. However, it is very important to know the level of the quality of the data. Another critical issue regarding big data integrity is the integration of big data from various sources. Unsuccessful data integration is also a potential source of erroneous data. This issue is important because many big data projects rely on the integration of data from various sources. Big data governance provides measures to ensure that the quality of the data is as agreed. It also provides accountability in the case of an error caused by low data integrity.

BUSINESS CASE DESCRIPTION

This study demonstrates through a genuine business case the development of big data governance by a company that provides mobile and web services for their customers. The business case is about the product development process of new software in an agile start-up company. The new software is an educational game.

Methodology of the Business Case Research

The method chosen to conduct this research is the case study (Eisenhardt, 1989). The aim was to develop big data governance in the new product and service development phases. The authors also adopted an abductive qualitative research approach (Dubois & Gadde, 2002) to develop a new framework for big data governance. The framework is based on observations concerning big data governance in the case company's product development project. In abductive research, the researchers simultaneously review the prior literature and theories, and they analyse data gathered through empirical research and development work (Dubois & Gadde, 2002). In this study, the adoption of this iterative research process allowed for developing a deeper understanding of the empirical data being analysed while simultaneously contributing to the theory of big data analytics in consumer research.

The Business Case

This case study is based on an empirical research on mobile application development. The case company is an agile software company that is developing a big data driven soccer training mobile application which uses an artificially intelligent bot that trains junior soccer players.

The use of big data related to the business case can be divided into three main categories:

1) Data used for product and business development,
2) data used for building predictive modes about the general progress of the players and
3) data used for personalisation of the product.

Each of these categories will be described in detail in the following. Firstly, log data originating from user actions while interacting with the software reveals all the bottlenecks or traps a user may face when using the product. This data is used in developing the product in a user-centric way. In addition to real user data, artificial data mimicking human behaviour is generated using a genetic algorithm. This data is used to simulate numerous users and let them run through the game any time during development phase (Ketamo 2008; Ketamo 2010). This kind of approach is used to mimic the payers' behaviour in large scale. Care has to be taken to make sure that all the possible steps are built and that the evolutionary algorithm constructs enough variance and unexpected cases. On the conceptual level the method is the same as letting the AplhaGO -AI play against itself to create more understanding on Go game (e.g. Silver & Hassabis 2016).

Secondly, the user data, in general, enables building predictive models on what activities or what kind of usage patterns might predict success or failure. This is very useful in terms of professional sports: An organization might be able to point out talent from extremely large population just based on big data. Typically talent in sports cannot be scouted within large populations just because of lack of resources. This might help national sports associations to build better understanding on factors that might reveal talent. Similar measures have been done in the domain of mathematics (Ketamo, Devlin & Kiili 2018). The educational application is built so that it will enables predictive models based on all user data.

Thirdly, personalisation requires understanding on human behaviour and understanding the knowledge domain of activity (e.g. Brusilovsky 2001). The domain in the educational soccer learning game was constructed as a knowledge graph when the course was built. On the other hand, user activities were measured and recorded with the same language. The user data combined to predictive modes on successful (or non-successful) behaviour enables adaptive learning and personalised content recommendations within the domain and also outside the domain. It is very useful for a talented player to get the latest clips related to his talent. On the other hand, if the user tends to follow non-successful patterns, the adaptive learning features can guide the him back to a successful track.

The biggest challenge in personalisation is privacy: How to take all the benefits on big data and at the same time ensure not revealing any information that can be identified to a specific person without his consent. In the educational game in question, all the data is collected anonymously and according to the rules and spirit of EU's GDPR (General Data Protection Regulation), meaning that the internal use for adaptation and personalisation is secured. However, when passing even small parts of information into an external service, additionally with other data from user's device, the user's consent needs to be asked explicitly for. Even data on suggesting and downloading single news like "Losing your nerves every time: try these 10 exercises" might tell something about the user. Typically this can be connected to a specific person using additional data such as data from cookies in browser.

In addition to studying the all the sources of big data that was used in the product development phase, a field study was conducted to verify how well user opinions collected by a traditional questionnaire correlated with the actual facts that could be observed in the user log data. This was important for a deeper understanding of the significance of the log data. The log data is one of the key sources of big data and important in product development, building predictive models on players' future steps and on personalization of the product.

In the field study, a traditional field study with test users was combined with the usage information of application server data and local weather information. The soccer training application was intended to be used outdoors. The application includes a training programme guided by an artificially intelligent bot that guides junior players to try selected soccer techniques, such as corner kicks, passing, and ball control. Thus, the outdoor context with situational variables such as the local weather were essential factors that affected the behaviour of the junior players who were the consumers of the application.

The authors conducted a field study in which the application was introduced to 134 junior soccer players in eight different soccer teams. Data was collected by interviewing the players after the test period to understand their user experience and the variables that affected their use of the application. The native mobile application communicated with the web service based on the cloud server architecture, which enabled us to use the server data. To gain a deep understanding of the factors that affected the user behaviour in this field study, the authors analysed also the weather data collected during the study period in the locations where the application was used by the junior soccer players. Thus, the authors formed a comprehensive understanding of the players' behaviour. The authors combined the traditional interview data with the usage log data, open data on the local weather, master data concerning the soccer teams and their training facilities as well as geographical reference data.

Book chapter in: Sheryl Kruger Strydom and Moses Strydom, eds., Big Data Governance and Perspectives in Knowledge Management, 335 pages, IGI Global. Projected release date: November 2018.

RESULTS AND RECOMMENDATIONS

This section first proposes a framework for big data governance. After that it provides recommendations concerning each of the five dimensions of the framework. Both are based on the business case study presented above.

The Proposed Big Data Governance Framework

The proposed big data governance framework is formed based on the requirements found in the business case concerning the product development project in the educational game industry. It also takes into account the widely accepted framework of DAMA International presented in the background chapter. Table 1 lists the five dimensions of the big data governance framework and gives a brief explanation of each.

Table 1. The five dimensions of the proposed big data governance framework and their brief description

Dimension	**Meaning**
Data privacy	Data containing information about a private person should be treated with special attention according to the organisation’s data privacy policy and legislation.
Data security	The processes and technologies that ensure that sensitive and confidential data about an organization are kept secure according to the organisation’s policies.
Data availability	Making data available at a given moment, including the usage of data, interface standards, metadata, and the findability of data.
Data usability	The data in an organisation can be used to meet the goals defined in the corporate strategy, including data monetisation.
Data integrity	The trustworthiness of the data, including data lifecycle management and data quality monitoring.

Figure 3 below shows the relation of the widely accepted data governance framework of DAMA to the proposed big data governance framework. As explained in the previous chapter, DAMA describes data governance as the act of authority, control and shared decision making over the nine fields of data management that are listed in the sectors of the circle. The authors will now explain in detail the dimensions of the proposed big data governance framework and how they relate to the existing data governance framework of DAMA.

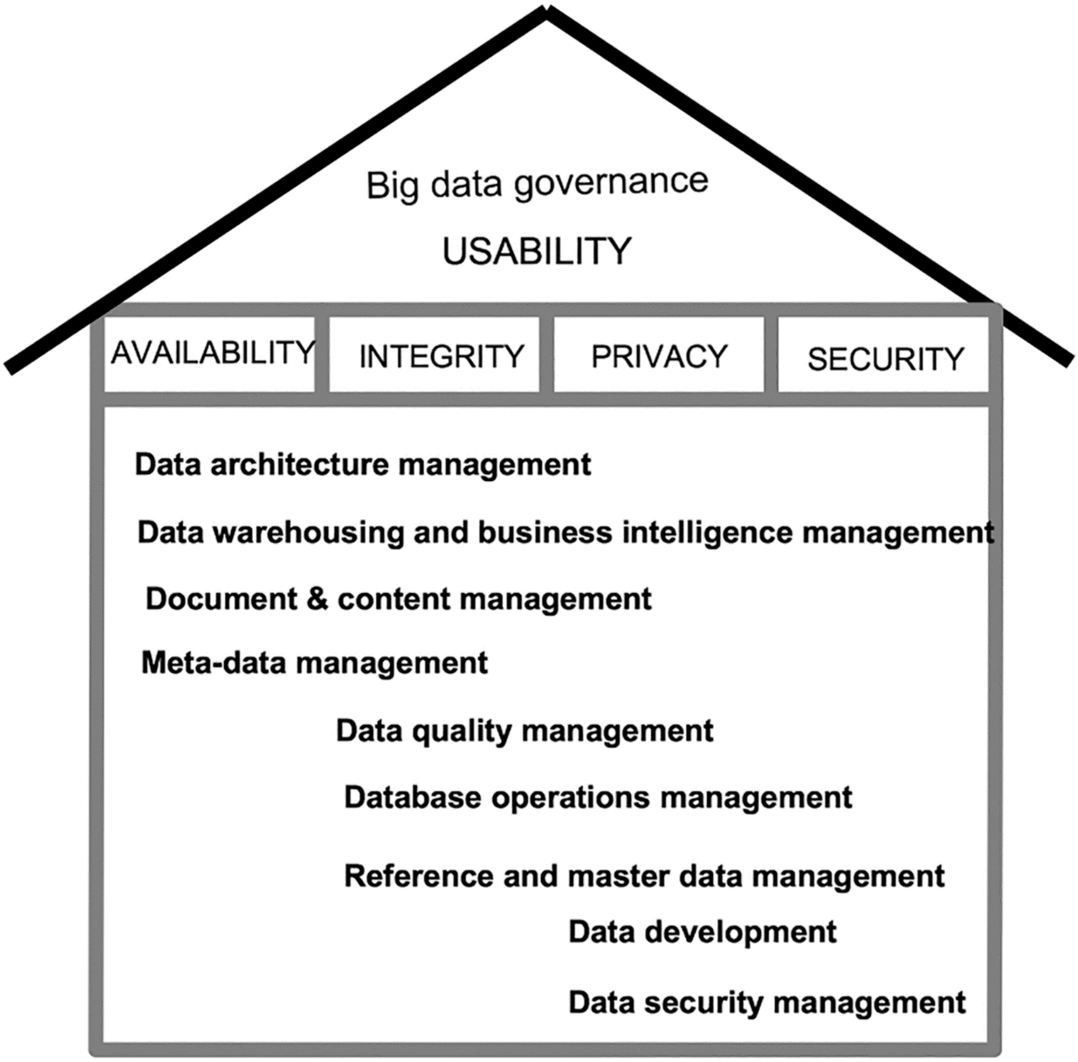

Big data governance
USABILITY
AVAILABILITY
INTEGRITY
PRIVACY
SECURITY
Data architecture management
Data warehousing and business intelligence management
Document & content management
Meta-data management
Data quality management
Database operations management
Reference and master data management
Data development
Data security management

Figure 2. Illustration of the proposed big data governance framework and its relation to the existing data governance framework presented in DMBOK. The five dimensions of the framework are written in capital letters.

The figure shows that usability is in the heart of big data governance itself. In addition to controlling all the traditional nine fields of data management, all of the nine fields should be managed with usability in mind. This is because both the amount and the technical challenges related to big data are so important that it is crucial to ensure that all the data that is being governed is also usable i.e. relevant from the business point of view. This is one reason why big data governance should not be an endeavour of the IT-department, but it should rather be responsibility of a business leader. Moreover, to ensure data usability, care should be taken to implement shared decision making about the data assets of an organisation. This shared decision making should involve all relevant parties so that all possibly valuable data assets are considered. Shared decision making is also in line with agile software development (Seiner, 201).

Data governance itself does have a more important role when dealing with big data than when dealing with traditional data. It is important that personnel of an organization is explicitly made accountable for all the data assets that are used in the organization and that are vital for the business. This ensures that all data assets are governed with authority and controlled. With the advent of big data, also those roles and departments that typically have not had any accountability at all or only a limited one for data assets will typically have a broader accountability for different data assets. For example, personnel from the marketing department is typically assigned the accountability for social media data. The accountability for sensor data (also called IoT data) is typically assigned to personnel in the operations and maintenance units of an organization. Governing big data is very much linked to the usability of the data – even more than in the framework for governing traditional data.

Now that the special requirements enforced by big data on data governance have been explained, the authors will go through the rest of the five dimensions of the framework and explain how they are related to the traditional data governance framework. In Figure 3, the four other dimensions are written in capital letters directly under the "roof" of big data governance and usability. This means that they are dimensions that need to be controlled by the framework. The nine fields of data management are written inside the box under the four big data governance dimensions. They are aligned under the dimension that most describes them. However, the fields typically can be described by several dimensions. For example, data architecture management is at the left-hand side directly under availability. This means that a good data architecture is very important for data availability, meaning that is enables the findability of the data for the right person at the right time. However, a good data architecture also helps in data integrity because it prevents a situation where the same data is stored in two places, which causes the risk that the data are not updated simultaneously.

Recommendations on Big Data Privacy

The privacy of big data is an important issue because these data are heterogeneous, and they are derived from various sources (Morabito, 2015; Soares, 2012). Many sources contain data that originate in the actions of individual customers who interact with a digital service. In the business case of the educational soccer application, the data collected on the application's usage were used to build the predictive models. The usage data revealed information about the customers, such as their location, date and time of usage, and the parts of the application they used. It is very important that these data are well protected and that only authorised people and software may access them (Otto, 2011). If need be, the proposed big data governance framework enforces data privacy related procedures for acquiring permission from users to use their personal data.

In the case study, data privacy played a crucial role. In the field test, the authors defined the privacy issues before meeting the users in person when they received a face-to-face introduction to using the application. The purpose of the study was explained to the players. Because the users were children, a letter was distributed to their parents explaining the purpose of the trial. The coaches of the junior soccer teams or research assistants delivered the information letter to the parents. Hence, the researchers obtained the parents' permission for their children to use their smartphones and access the data. The parents understood that their children would download and use the application, which was related to their hobby. The authors learned that when the persons who collected the data met the users in person, they built trust more easily than in collecting similar information through online surveys. In the field test, it was easier to motivate users to use the application, as they received a free application for participating in the study. This observation supports the findings of previous research that showed that if users received benefits, they were more willing to allow the use of their personal information (Caudill & Murphy, 2000; Chellappa & Sin, 2005). The study demonstrated that it is important to define in advance how different user groups or segments are identified and used in the analytical phase.

The authors found that by combining several data sources, the predefined research questions allowed for determining the accuracy of the user data, particularly regarding the privacy issue. For example, to obtain detailed information, the authors needed to identify small user groups in other data sources. The findings showed that some users were significantly more active than others who used the application only once or a few times. The log data collected by the server provided statistical information about the usage during the trial period, which helped the authors to validate and understand the findings of the field test. Thus, the authors were able to compare these statistics to the findings of the interviews, which increased the reliability of qualitative results. Integrating the weather information as a dependent variable was challenging because the authors could not find data on the exact location of the users when they used the application. The authors learned that weather changes rapidly within a city. For example, because rain can start and stop within minutes, it was necessary to track and analyse usage accurately in minutes rather than hours. Additionally, the exact location was also required. Thus, the authors recommend adding a geolocation feature to the application because it was not included during the field test. However, the addition of this feature could increase the privacy concerns of many users because they would reveal their location when they used the application. Although many mobile applications already have a geolocation functionality because it facilitates sending local advertisements or other information to the screens of smartphones, users could choose to disable the geolocation function. Thus, companies should provide a trade-off or a value for users who allow tracking by geolocation if it is important to know their exact location for analytical purposes.

Recommendations on Data Security

In the analyses of the companies, integrating open data did not place data security or the privacy of users at risk. The findings showed that the open weather data provided interesting information about the usage environment and the variables that affected the users’ willingness to use the application outdoors. Furthermore, the data sources were not technologically integrated, which increased the security of the data. The companies did not need to conduct analyses over networks if the data were saved to a secured server on their premises. Chen and Zhao (2012) pointed out that sharing data over networks created potential security risks and could cause privacy concerns in users. In the present study, the authors did not deal with sensitive information, such as home addresses, credit card numbers, or phone numbers. Acquisti and Gross (2006) showed that users were more willing to provide general personal information than factual private information that provides more detailed information about them.

Recommendations on Data Availability

In the case study, big data were readily available. The field study data were few and well managed. The server log data were accumulated constantly in real time. The data management principle considered that the predictive models had to be re-created from time to time to reflect changes. In the current case, the authors always used all the available log data. In a future study, the oldest data could be warehoused, and only the newest data could be used. In the current study, data warehousing procedures had to be constructed to accommodate the old models.

The external open weather data were not managed by the case company because they were sourced from the data provider. According to this experience, the authors suggest that external data should be governed such that one person or role in the organisation is accountable for them.

Recommendations on Data Usability

The term data usability means that the data should be aligned with business needs. According to the experiences of the case study, the authors suggest that businesses as well as technology stakeholders in organisations should be involved in ensuring that all data conform to the needs of the predictive models. For example, an organisational role must be accountable to ensure that the user action log data conform to the requirements of the predictive big data model generated after software updates. In this study, part of the management of data usability was the proper documentation of software updates that might affect the creation of log data.

Recommendations on Data Integrity

In the case study, data integrity concerned mainly data quality and data lifecycle management, including data warehousing. In the business case, the findings showed that in order to provide a trustworthy model based on several separate datasets, special attention must be paid to the correct integration of data.

Data integrity became a challenge when the authors combined three data sources. It was not difficult to analyse a single data set, such as clustering users in different segments. However, when there were three different data sources in different formats in the same period and location, it was difficult to identify the exact location of the anonymous soccer players or teams when the weather conditions were analysed. This finding indicates that to maintain data integrity between different data sources accurate location information and the identification of anonymous users might be required to create profiles of their usage. However, the requirement of data privacy is a limitation if users do not allow tracking their geolocation.

It is understood that a predictive model that is based on inaccurate or even erroneous data has no value (Soares, 2012). In applying the framework use in this study, the authors used big data from two sources: the company’s internal data and the company’s external data. The internal data included real-time and historical data on the company’s proprietary servers as well as master data on the soccer teams and soccer practice facilities. The external data consisted of real-time and historical data on weather and weather forecasts at the Finnish Meteorological Institute as well as through Google maps.

The predictive models were used in real time, and they were automatically updated. The findings confirmed the requirements for the automatic monitoring of data quality as well as the availability of data. Hence, special attention was paid to building not only a solid framework for ensuring data quality but also mechanisms for the automated monitoring of the most important issues in this framework. Examples of issues that this framework monitored include soccer practice facilities and teams that were not included in the master data but appeared in the customer data on the server as well as in unexpected changes to application programming interfaces in the external data sources.

Automated tests for monitoring data integration are an efficient means of ensuring data integrity. For example, if the server log data showed that a team was using a playing facility that was not in the master data, an alert would be sent to the person accountable for the master data about the playing facilities. This person would then be able to amend the master data as soon as possible.

Table 2. The summary of solutions and recommendations

Dimensions	**Solutions and recommendations based on the case study**
Data privacy	Data privacy should be taken into account very carefully in big data development. Different data types and collection methods require special attention to be paid to privacy issues. In the case study, the level of location information in the different user segments was inaccurate due to the used privacy policy. The authors recommend paying special attention to the tradeoff between privacy and value perceived by the user. This can be achieved by motivating users to share their private information that is crucial for more detailed analytics.
Data security	In many companies private and company confidential data is professionally secured and managed but sharing data over networks may create security risks. Real time data analytics on cloud applications creates more security risks than analyzing data sets offline. The authors recommend companies to partner with IT-companies who are specialized in secured infrastructure solutions and services. Additionally, the authors recommend companies to use trusted data providers in purchasing sensitive customer information, such as purchase histories or order data.
Data availability	Availability was easy to manage in the small start-up company of the case study, but in a larger corporation cross-functional communication and co-creation business insight become more challenging. It is important to design an information architecture that supports the findability of data. The development of meta-data management should also be given special attention. Data warehouses should be given special attention as they contribute to making non-transactional data available. Examples of data warehouse-related decision are: Which data to put into a data warehouse, how long to keep it and at what intervals to update it? Business intelligence is also very important in the era of big data. It is evolving very quickly: when data is no more big data, it often becomes just regular business intelligence data. For example, social media stream data used to be big data when it was too large and moving too quickly to be handled by traditional software. Document and content management is very important in a big data governance framework because the majority of big data belongs to this category. It is not nicely available in a well-structured relational database
Data usability	In the case study, the careful planning of the customer study helped to align the data collection to the business expectations. However, accuracy of data became a challenge. In the case study, location information of weather data was difficult to align to usage data due to the inaccuracy of location information. The authors found that big data analytics is an iterative process where quality of data is gradually enhanced. Especially analytics in the emerging market involves several uncertainties that are difficult to predefine exactly in advance. The authors recommend beginning big data projects as early as possible. Learning to manage a data collection and to analyze the processes related to it in multi-disciplinary teams takes time, and quality of data needs to be gradually improved by integrating various datasets.
Data integrity	To provide a trustworthy model based on several separate datasets, special attention has to be given on the correct integration of data. In

	real-time data analysis unexpected modifications in data due to changes in the software producing it have to be governed to prevent the advent of analysis based on corrupted data.

FUTURE RESEARCH DIRECTIONS

More research is needed on real-life cases of big data governance. Now that the usage of big data in creating new value for business is common, it is important that also suitable governance frameworks are developed and implemented in business. Big data governance will bring new insights also to traditional data governance and vice versa. At some point data governance will also include big data governance. Before that, big data governance is needed because it takes into account the challenges and trade-offs caused by data variety, velocity and volume. Furthermore, the frameworks and practices for governing big data are by no means ready and they need to be developed and researched in more detail and depth.

CONCLUSION

One framework of big data governance does not fit all companies. Although the key principles of privacy, security, availability, usability, and integrity should be the same in general, their implementation differs. The company’s strategy and the maturity of its product development and business model affect the implementation of a big data governance framework. The authors learned that a start-up company in an emerging market needs a flexible data governance framework that is suited for an agile software development process. In an emerging market, a company cannot predict and manage all issues in advance as it could do in a mature market. In a mature market, customer behaviour is easier to predict because the companies in it already have a long history of dealing with customer data. In emerging markets, companies are faced with a great amount of uncertainty because the products are new, and little is known about the markets and customers (Blank, 2007). Additionally, start-up companies often use lean or agile development methodologies and management, and they usually have an experimental corporate culture unlike established businesses (Ries, 2007). The goal of data analytics is to facilitate learning processes in companies. The learning cycles of start-up companies are often shorter than those of established businesses. Specifically, established businesses have longer histories of customer data, but start-ups work under conditions of uncertainty.

Despite the challenges of collecting data in the new AI-based learning technology markets, the authors learned that it is important to focus on all five key dimensions of big data governance in technology companies. The findings from the user study of the soccer learning game showed that diversified data collection methods helped us to obtain a realistic understanding of users’ thinking. The findings also showed that information about the users’ locations was important, but it required users to reveal their geolocation. This requirement calls for ways to motivate users to provide personal information, which in the literature on data privacy is called a trade-off between privacy and perceived value (e.g., Caudill & Murphy, 2000; Chellappa & Sin, 2005).

To succeed in customer-behavioural analytics, companies need to be able to manage various dimensions of data processes. In particular, service companies that launch new mobile and web services should apply a big data governance framework to manage technological data collection processes and privacy issues of customers as well as the availability, integrity, and usability of data. An adequate big data governance framework would enable data driven insights for marketers and aid in sales and product development. The findings of this study showed that users need to be motivated to provide their personal information for use by companies. The implementation of a framework for big data governance in a real-life business case increased the quality and value to the business of big data based product development and predictive model generation. The big data governance framework also facilitated the consistent and trustworthy use of customer data, which is essential in maintaining a positive company image.

Managerial Implications

From the managerial point of view, this study presents the framework of big data governance for companies that operate in emerging markets. It is essential for the success of such businesses that it learns fast from the data. Advanced multi-source data analytics provide a way to gain new knowledge for decision-making. However, maintaining data privacy when enriching the primary data with other data sources is not an obvious case. The weather data used in this study does not bring any privacy challenges. If the users of the case study would be connected to additional sources of data such as social media sources, there would be a great possibility that the privacy of the users is concerned. In fact, there are no exact rules for determining when an added data source could risk the privacy of a person. That is why all managers should be aware of the nature of big data. Learning fast through data analytics requires a big data governance framework. Thus, it is important to implement the five key dimensions into the processes of companies to manage data analytics projects.

ACKNOWLEDGMENT

This work was supported by the BIG-research program, funded by TEKES (Finnish Funding Agency for Innovation) no 2710/31/2016. The authors also thank the eight anonymous soccer teams and the students who assisted in collecting the qualitative field data.

REFERENCES

Ackerman, M. S., Cranor, L. F., & Reagle, J. (1999). Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, ACM, 1–8.

Acquisti, A., & Gross, R. (2006). Imagined communities: Awareness information sharing and privacy on Facebook. In *Proceedings of the Privacy Enhancing Technologies Symposium*, Cambridge, United Kingdom, 36–58

Alamäki, A., & Dirin, A. (2015). The stakeholders of a user-centred design process in mobile service development. *International Journal of Digital Information and Wireless Communications*, *5*(4), 270-284.

Bart, Y., Shankar, V., Sultan, F., & Urban, G. L. (2005). Are the drivers and role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study. *Journal of Marketing*, *69*(4), 133–152.

Blank, S. (2007). *The four steps to the epiphany: Successful strategies for products that win*. Quad/Graphics.

Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User- Adapted Interaction*, vol 11, p. 87-110.

Caudill, E. M., & Murphy, P. E. (2000). Consumer online privacy: Legal and ethical issues. *Journal of Public Policy & Marketing*, *19*(1), 7–19.

Chellappa, R. K., & Sin, R. G. (2005). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management*, *6*(2–3), 181–202.

Chen, D., & Zhao, H. (2012, March). Data security and privacy protection issues in cloud computing. In *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference* (Vol. 1), 647–651. IEEE.

Chen, Q., Zhang, M. & Zhao, X. (2017). Analysing customer behaviour in mobile app usage. *Industrial Management & Data Systems*, 117(2), 425-438.

DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications.

Dirin, A., Laine, T., & Alamäki, A. (2018) managing emotional requirements in a context-aware mobile application for tourists. *International Journal of Interactive Mobile Technologies* (in press).

Du Mars, R. (2012). Mission impossible? Data governance process takes on "big data." TechTarget.com. Retrieved from http://searchdatamanagement.techtarget.com/feature/Mission-impossible-Data-governance-process-takes-on-big-data

Flyvbjerg, B., & Budzier, A. (2011). Why your IT project may be riskier than you think. *Harvard Business Review*, 89(9), 23-25.

Godinez, M., Hechler, E., Koenig, K., Lockwood, S., Oberhofer, M. & Schroeck, M. (2010). *The art of enterprise information architecture: A systems-based approach for unlocking business insight*. IBM Press, Pearson Higher Ed. USA.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115.

Kaufman, L. M. (2009). Data security in the world of cloud computing. *IEEE Security & Privacy*, *7*(4).

Ketamo, H. (2008). Cost Effective Testing with Artificial Labour. In *Proceedings of 2008 Networked & Electronic Media Summit*. Saint-Malo, France, 13-15.10.2008, pp.185-190.

Ketamo, H. (2010). Balancing adaptive content with agents: Modelling and reproducing group behavior as computational system. In *Proceedings of 6th International Conference on Web Information Systems and Technologies*, WEBIST 2010, 7-10 April 2010, Valencia, Spain, vol 1, pp. 291-296.

Ketamo, H., Devlin, K. & Kiili, K. (2018). Gamifying Assessment: Extending Performance Measures with Gaming Data. In *Proceedings of American Educational Researcher Association's Annual Conference* AERA2018, New York, 13th-18th April 2018.

Ladley, J. (2012). *Data governance: How to design, deploy and sustain an effective data governance program*. Elsevier.

Liebowitz, J. (2013). *Business analytics: An introduction*. CRC Press, Taylor & Francis Group.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, *15*(4), 336–355.

Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, *45*(2), 135–155.

Morabito, V. (2015). *Big data and analytics: Strategic and organizational impacts*. Cham: Springer.

Morville, P. (2007). *Information architecture for the World Wide Web*. O'Reilly.

Otto, B. (2011). Data governance. *Business & Information Systems Engineering*, *3*(4), 1–244.

Paajanen, S., Valkokari, K. & Aminoff, A. (2017). The opportunities of big data analytics in supply market intelligence. In *Proceedings of the18th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2017,* Vicenza, Italy, September 18-20, 2017. (pp 194-205). Springer.

Panian, Z. (2010). Some practical experiences in data governance. *World Acad. Sci. Eng. Technol, 38*, pp.150-157.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (Eds.) (2010). *Recommender systems handbook*. Springer Science & Business Media.

Ries E. (2010). *The Lean Startup: How constant innovation creates radically successful businesses*. London: Penguin Books.

Sarsfield, S. (2009). *The data governance imperative*. It Governance Ltd.

Seiner, R. (2014). *Non-Invasive Data Governance. The Path of Least Resistance and Greatest Success*. Technics Publications.

Sharda, R., Delen, D., Turban, E. (2018). *Business intelligence, analytics and data science: A managerial perspective*. Pearson.

Silver, D. & Hassabis, D. (2016). *AlphaGo: Mastering the ancient game of Go with Machine Learning*. In Google Research Blog, 27.1.2016. Google.

Soares, S. (2012). *Big data governance: An emerging imperative*. MC Press, ProQuest Ebook Central.

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018). Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, *94*(9-12), 3563-3576.

White, M. S. (2015). *Enterprise search*. O'Reilly Media.

Wilkowska, W., & Ziefle, M. (2012). Privacy and data security in E-health: Requirements from the user's perspective. *Health Informatics Journal*, *18*(3), 191–201.

Zhang, Y., Ren, S., Liu, Y., Sakao, T., & Huisingh, D. (2017). A framework for Big Data driven product lifecycle management. *Journal of Cleaner Production*, *159*, 229-240.

Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, *28*(3), 583–592.

ADDITIONAL READING

Chen, H., Chiang, R. H., & Storey, V. C. (2012). *Business intelligence and analytics: from big data to big impact*. MIS quarterly, 1165-1188.

Connolly, T. & Begg, C. (2010). *Database Systems: A Practical Approach to Design, Implementation and Management*, 5th Edition, Addison-Wesley.

Dean, J. (2014). *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. SAS Institute, Inc. John Wiley & Sons, Inc., Hoboken, New Jersey.

Demirkan, H., & Delen, D. (2013). *Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud*. Decision Support Systems, *55*(1), 412-421.

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). *The 'big data' revolution in healthcare*. McKinsey Quarterly*, 2*, 3.

Gurin, J (2014). *Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation*. McGraw Hill Education.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98-115.

Book chapter in: Sheryl Kruger Strydom and Moses Strydom, eds., Big Data Governance and Perspectives in Knowledge Management, 335 pages, IGI Global. Projected release date: November 2018.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT sloan management review*, *52*(2), 21.

Verhoef, P. C., Kooge, E., & Walk, N. (2016). *Creating value with big data analytics: Making smarter marketing decisions*. Routledge.

Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. McGraw-Hill Osborne Media.

KEY TERMS AND DEFINITIONS

Big data: Data that cannot be processed using traditional data analytics software and infrastructure on a personal computer or on a data analytics server. Compared with traditional data, big data have greater volume, variety, and/or velocity than traditional data.

DAMA: The Global Data Management Community. It is a non-profit and vendor-independent association that provides a community and support for information professionals.

Data governance: The processes and technical infrastructure that an organisation has in place to ensure data privacy, security, availability, usability, and integrity.

Data privacy: Data containing information about a person should be treated with special attention according to the organisation's data privacy policy and legislation.

Data security: The processes and technologies that ensure that sensitive and confidential data about an organization are kept secure according to the organisation's policies.

Data availability: Making data available at a given moment, including the usage of data, interface standards, metadata, and the findability of data.

Data usability: The data in an organisation can be used to meet the goals defined in the corporate strategy, including data monetisation.

Data integrity: The trustworthiness of the data, including data integration, data lifecycle management, and data quality monitoring.

DMBOK: The DAMA International Guide to Data Management Body of Knowledge. A publication that is dedicated to advancing the concepts and practices of information and data management.

GDPR: The General Data Protection Regulation. It is a regulation in European Union (EU) law on data protection and privacy for all individuals within the EU and the European Economic Area.

IT governance: The processes that ensure the effective and efficient use of IT in enabling an organization to achieve its goals.

Predictive model: A data-driven model, which is used to predict a future event, in contrast to a descriptive model, which is used to explain a past event.