Journal of Positive Psychology and Wellbeing
2020, Vol. 4, Issue 2
Pp 30-44

The Future of Data Governance to Create Trusted Decision-Making Outcomes

Daniel Aristoteles Collins and Lisa de la Garza
Department of Aviation Management, Eskis¸ehir Technical University, Eskis¸ehir 26470, Turkey

A B S T R A C T
Organizations are increasingly introducing data science initiatives to support decision- making. However, the decision outcomes of data science initiatives are not always used or adopted by decision-makers, often due to uncertainty about the quality of data input. It is, therefore, not surprising that organizations are increasingly turning to data governance as a means to improve the acceptance of data science decision outcomes. In this paper, propositions will be developed to understand the role of data governance in creating trust in data science decision outcomes. Two explanatory case studies in the asset management domain are analyzed to derive boundary conditions. The first case study is a data science project designed to improve the efficiency of road management through predictive maintenance, and the second case study is a data science project designed to detect fraudulent usage of electricity in medium and low voltage electrical grids without infringing privacy regulations. The duality of technology is used as our theoretical lens to understand the interactions between the organization, decision- makers, and technology. The results show that data science decision outcomes are more likely to be accepted if the organization has an established data governance capability. Data governance is also needed to ensure that organizational conditions of data science are met, and that incurred organizational changes are managed efficiently. These results imply that a mature data governance capability is required before sufficient trust can be placed in data science decision outcomes for decision-making.

Keywords:
data lake; data governance; data quality; big data; digital transformation; data science; asset management; boundary condition

Article History:
Received: 9 September 2020;
Accepted: 30 September 2020;

1. INTRODUCTION

Over the last few years it has become more common for organizations to implement data science initiatives to support the digital transformation of their business (Provost and Fawcett2013). However, organizations continue to find it difficult to trust data science outcomes for decision-making purposes, as the data is often found to be lacking the required quality (Lin et al.2006), and it is often unclear how compliant the use of the data and the algorithms are with regards to relevant legal frameworks and societal norms and values (Nunn2009;van den Broek and van Veenstra2018). These uncertainties are a barrier to the acceptance and use of data science outcomes due to the possibility of financial risk and damage to an organization's reputation. For example, when making decisions regarding the management of physical assets, asset managers need to be able to trust the data science outcomes before they are confident enough to use these outcomes. Examples of these decisions include when and where to perform maintenance on highways or when to replace a bridge. Erring on the side of caution can be unnecessarily expensive whilst irresponsible delay of maintenance can put public safety at risk. In order for data science to be successfully adopted, it is therefore vital that organizations are able to trust the integrity of the data science outcomes (Council on Library and Information Resources 2000;Randall et al.2013). Recently, data governance has gained traction with many organizations as a means to develop this trust (Al-Ruithe et al.2019;Brous et al.2016). However, it remains unclear how data governance contributes to the development and maintenance of trust in data science for decision-making, leading to calls for more research in this area (Al-Ruithe et al.2019;Brous et al.2020).

The goal of data science is to improve decision-making. According toDhar(2013), the term data science refers to knowledge gained through systematic study and presented in the form of testable explanations and predictions. As such, data science differs from traditional science in a number of ways (Dhar2013;Provost and Fawcett2013). Traditionally, scientists study a specific subject and gather data about that subject. This data is then analyzed to gain in-depth knowledge about that subject. Data scientists tend to approach this process differently, namely by gathering a wide variety of existing data and identifying correlations within the data which provide previously unknown or unexpected practical insights. However, research has shown that favoring analytical techniques over domain knowledge can lead to risks related to incorrect interpretation of the data (Provost and Fawcett 2013).

Due to the automation of the decision-making process, it may be tempting to regard data science decision-making outcomes as being purely rational. However, as with all decision-making, the quality of the outcomes are subjected to the constraints of bounded rationality (Simon1947;Newell and Simon 1972), in that decision- making is constrained by the quality of the data available at the time. Data science models make decisions based on the information available to them at the time and also in the time given (Gama2013). According toGama(2013), bounded rationality can also appear in data science in the tradeoff between time and space required to solve a query and the accuracy of the answer. As such, it is not surprising that many organizations are implementing data governance in order to gain control over these factors (Alofaysan et al.2014;Brous et al.2020;van den Broek and van Veenstra 2018). Although recognized as being a powerful decision-making tool, data science is limited by the quality of the data inputs and the quality of the model itself.

Data governance can be defined as ―the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets‖ (DAMA International2017, p. 67), and can provide direct and indirect benefits (Ladley2012). For example,Paskaleva et al.(2017) show that adoption of data governance can change how data is created, collected, and used in organizations. Data governance can greatly improve the awareness of data science outcomes for the management of infrastructure in, for example, a smart city environment (Paskaleva et al.2017). However, information technology (IT)-driven data governance initiatives have failed in the past (Al-Ruithe et al.2019), often being affected by technical feasibility aspects carried out on system by system basis.

In this paper, a different starting point is used, and the focus is put on the investigation of data governance as a boundary condition for data science, which needs to be satisfied in order to be able to trust data science outcomes as suggested byBrous et al.(2020) andJanssen et al.(2020). In this research, boundary conditions for data science are defined as socio-technical constraints that need to be satisfied in order to be able to trust data science outcomes. These conditions refer to the ―who, where, when‖ aspects (Busse et al.2017) of data science before data science outcomes can be used. Previous research (Brous et al.2020;Janssen et al.2020) has suggested that data governance can be viewed as a boundary condition for data science. As such, our main research question asks, how is data governance a boundary condition for data science decision-making outcomes?

In order to answer this question, two explanatory data science case studies in the asset management domain were analyzed with specific regard for the role of data governance as a boundary condition for trustworthy predictive decision-making through the creation of trust in data science decision-making outcomes. The first case under study is a data science project designed to improve the efficiency of road maintenance through predictive maintenance. The project was performed under the auspices of a large European government organization using a multitude of datasets which were sourced both within the organization and externally. Open data (Zuiderwijk and Janssen2014) were also employed within this case study. The second case study is a data science project which analyzes transformer data to identify the fraudulent use of electricity within medium and low tension electrical grids without infringing privacy regulations. This project was performed under the auspices of a European distribution grid operator (DGO) which is responsible for the distribution of electricity over medium and low tension grids in a highly industrialized region of Europe.

Duality of technology theory (Orlikowski1992) is used to guide the analysis of the case studies in understanding trust in data science outcomes as a boundary value problem and specifically the role of data governance as a boundary condition for trusting data science outcomes. Duality of technology (Orlikowski1992) describes technology as assuming structural properties while being the product of human action. From a technology standpoint, data science outcomes are created by data scientists in a social context, and are socially constructed by users who attach different meanings to them and provide feedback to the data scientists. In this way, data science outcomes are the result of the ongoing interaction of human choices and organizational contexts, as suggested by duality of technology (Orlikowski1992). This approach di ffers from previous research into data science success factors, which have focused on the view that data science is either an objective, external force which has a deterministic impact on organizational properties (Madera and Laurent2016), or that trust in data science outcomes is purely a result of strategic choice and social action (Gao et al.2015). Duality of technology theory suggests that either model would be incomplete and suggests that both perspectives should be taken into account when analyzing boundary conditions of data science. The results of the case studies suggest that data science outcomes are more likely to be accepted if the organization has an established data governance capability, and we conclude that data governance is a boundary condition for data science as it enables organizational conditions and consequences of data science to be met and ensures that outcomes may be trusted.

The paper reads as follows. Section 2 presents the background of literature regarding therelationship between data governance and data science. In Section3the methodology of the research is described. Section 4 describes the findings of the case study. Section 5 discusses the findings of the case study and Section6presents the conclusions.

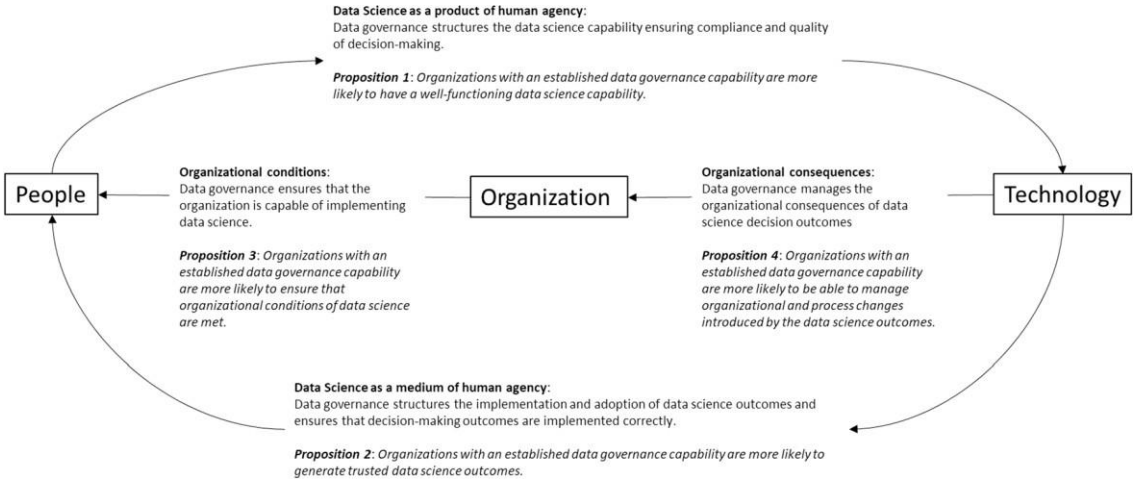
2. LITERATURE BACKGROUND

The literature review method proposed byWebster and Watson(2002) was followed to methodologically analyze and synthesize quality literature. The goal of the literature review is to gain an understanding of the current knowledge base with regards to the role of data governance for creating trust in data science decision-making outcomes. In order to understand the duality of data governance, we discuss literature which helps us understand how data governance structures organizations, taking into account research into the adoption and impact of technology on organizations as suggested by research on other disrupting technologies such as artificial intelligence (AI) and the internet of things (IoT). This paper utilizes the duality of technology theory (Orlikowski1992) as a practice lens for studying the role of data governance for the creation of trust in data science and follows the case study methodology to investigate this phenomena. The propositions that are investigated in the case studies are synthesized from the literature following the logic of duality of technology.

Based onGiddens' (1976) theory of structuration, duality of technology (Orlikowski1992) describes technology as assuming structural properties while being the product of human action.Giddens (1976) recognizes that ―human actions are both enabled and constrained by structures, yet that these structures are the results of previous actions‖ (Orlikowski1992, p. 404). In her structuration model of technology,Orlikowski(1992) identifies four main relationships, namely: (1) technology as a product of human agency, (2) technology as a medium of human agency, (3) organizational conditions of interaction with technology and, (4) organizational consequences of interaction with technology. The technology referred to in this article is data science. Data governance is about the coordination and control of the use and management of data (Janssen et al.2020;Khatri and Brown2010). The objective of this article is to understand the role of data governance as a boundary condition for data science. As such, this article looks at the role of data governance in data science using the duality of technology as a guiding logic.

Figure1below shows how the synthesized propositions and their elements are linked following the logic of the duality of technology.

Figure 1. The relationship of the propositions with duality of technology. According toOrlikowski (1992), technology is created as a result of human agency. In order for the process of technology creation to be successful, certain organizational boundary conditions need to be met. The resulting technology also has consequences for the organization, which need to be coordinated and controlled. For example, in order to develop a data science capability for an organization, it is necessary to have the required information technology (IT) infrastructure in place, available data, and sufficient data scientists with the necessary knowledge, requiring large investments (Adrian et al.2017).



2.1. The Role of Data Governance with Regards to Data Science as a Product of Human Agency

According toGao et al.(2015), data scientists develop domain expertise over time, and apply this knowledge in big data analysis to gain the best results. However, the intellectual limitations of the data scientists themselves as well as the computational limitations of the available technology (Gigerenzer and Selten2002) mean that although data scientists often seek to compensate limited resources by exploiting known regularity, bias and variance can create errors in the decision outcomes which can be exacerbated by large data sets (Brain and Webb2002). Following the logic of bounded rationality (Simon1947), data scientists develop models based on their own limited knowledge, and therefore, the models are themselves constrained by the intellectual limitations of their makers as well as the quality of the data from which they learn and the technical infrastructure in which they operate.

Big data can provide organizations with complex challenges in the management of data quality. According

toSaha and Srivastava(2014), the massive volumes, high velocity, and large variety of automatically generated data can lead to serious data quality management issues, which can be difficult to manage in a timely manner (Hazen et al.2014). For example, IoT sensors calibrated to measure the salinity of water may, over time, begin to provide incorrect values due to biofouling. Data science information products often rely on near real-time data to provide timely alerts, and, as such, problems may arise if these data quality issues are not timely detected and corrected (Gao et al.2015;Passi and Jackson2018).

Often, modern data processing systems which are required to allow large amounts of varied big data (Dwivedi et al.2017) to be ingested without compromising the data structure are generally immediately accessible, allowing users to utilize dynamic analytical applications (Miloslavskaya and Tolstoy2016;Ullah et al.2018). This immediate accessibility, as well as the retaining of data in its original format presents a number of challenges regarding the governance of the data, including data security and access control (Madera and Laurent2016), as well as in maintaining compliance with regards to privacy (Morabito2015). As such, data governance has increasingly gained popularity as a means of ensuring and maintaining compliance, andMadera and Laurent(2016) have gone so far as to posit that data governance principles should be key components of data science technologies for managing risk related to privacy and security. According toKroll(2018), a responsible data governance strategy should include strategies and programs in both information security and privacy.

Proposition 1. Organizations with an established data governance capability are more likely to have a well-functioning data science capability.

Proposition 1 considers the interaction of data science with human agency from a product perspective. In other words, data governance is believed to play an essential role in coordinating and controlling the development of data science as a capability of the organization.

2.2. The Role of Data Governance with Regards to Data Science as a Medium of Human Agency

Data science differs from traditional science in a number of ways (Dhar2013;Provost and Fawcett 2013). Traditionally, scientists study a specific subject and gather data about that subject. This data is then analyzed to gain in-depth knowledge about that subject. Data scientists tend to approach this process by gathering a wide variety of existing data and identifying correlations within the data which provide previously unknown or unexpected practical insights. Data scientists gain domain expertise and apply this knowledge in big data analysis to gain the best results (Gao et al.2015). However, the trustworthiness of data science outcomes in practice is often affected by tensions arising through ongoing forms of work (Passi and Jackson2018). According toPassi and Jackson(2018), data science isa socio-material practice in which human agency and technology are mutually intertwined. dede Medeiros et al.(2020) therefore stress the importance of developing a ―data-driven culture.‖ Data governance is important for creating value and moderating risk in data science initiatives (Foster et al.2018;Jones et al.2019), as it can help organizations make use of data as a competitive asset (Morabito2015). Data governance aims at maximizing the value of data assets in enterprises (Otto2011;Provost and Fawcett2013). For example, capturing electric and gas usage data every few minutes benefits the consumer as well as the provider of energy. With active governance of big data, isolation of faults and quick fixing of issues can prevent systemic energy grid collapse (Malik2013).

Proposition 2. Organizations with established data governance capability are more likely to generate trusted data science outcomes.

Proposition 2 looks at the interaction of data science with human agency from a medium perspective. In other words, data governance is expected to play an important role in coordinating and controlling the use of data science in organizations.

2.3. The Role of Data Governance with Regards to Organizational Conditions of Interaction with Data Science

A common challenge in data science is aligning the data science inputs and outcomes with the structure of an organization (Janssen et al.2020). This mismatch can result in unclear responsibilities and a lack of coordination mechanisms which give organizations control of the data over its entire life-cycle. This is particularly the case for data science projects which require data inputs from multiple departments. There is often a lack of established mechanisms for data governance leading to the ad hoc handling of data (Janssen et al.2020). According toWang et al.(2019) it is necessary to develop data governance mechanisms beginning with policy development to define governance goals and strategies, followed by the establishment of organizational data governance structures. Top management support (Gao et al.2015), well-defined roles and responsibilities (Saltz and Shamshurin2016), and the choice of the data governance approach (Koltay2016) are considered critical. According toJanssen et al.(2020), data governance contains mechanisms to encourage preferred behavior. Incentives such as monetary rewards or public recognition should be complemented by mechanisms such as audits. Creating sound data governance requires a balance between complete control, which does not allow for flexibility, and lack of control (Janssen et al.2020

Research has shown that favoring analytical techniques over domain knowledge can lead to risks related to the incorrect interpretation of the data (Provost and Fawcett2013).Waller and Fawcett (2013) therefore believe that a data scientist should have a good understanding of the subject matter as well as having strong analytical skills. For

example, recent years have seen a surge of interest in predictive maintenance and anomaly detection in the asset management domain (Raza and Ulansky 2017), however, when implementing data science for predictive maintenance or anomaly detection, data scientists also need to have a strong understanding of how assets deteriorate over time. Furthermore, according toKezunovic et al.(2013), much of the data may not be correlated in time and space, or not have a common data model, making it difficult to understand without in-depth knowledge of how or why the data has been generated. As the number of people with data science skills as well in-depth domain knowledge is limited (Waller and Fawcett2013), these insights suggest that data science initiatives should be governed by people with in-depth domain knowledge. According toWang et al.(2019), organizations should develop comprehensive data governance mechanisms, beginning with policy development to define governance goals and strategies, followed by the establishment of organizational data governance structures.

Proposition 3. Organizations with an established data governance capability are more likely to ensure that organizational conditions of data science are met.

Proposition 3 considers the role of data governance as being important for coordinating and controlling the organizational requirements of the data science capability.

2.4. The Role of Data Governance with Regards to the Organizational Consequences of Data Science

As well as establishing data management processes that manage data quality, data governance should also ensure that the organization's data management processes are compliant with laws, directives, policies, and procedures (Wilbanks and Lehman2012). According toCato et al.(2015), policies and principles should be aligned with business strategies in an enterprise data strategy.Panian (2010) states that establishing and enforcing policies and processes around the management of data should be the foundation of effective data governance practice as using big data for data science often raises ethical concerns. For example, automatic data collection may cause privacy infringements (Cecere et al.2015;van den Broek and van Veenstra2018), such as in the case of cameras used to track traffic on highways, which often record personally identifiable data such as number plates or faces of persons in the vehicles.

Data governance processes should ensure that personally identifiable features are removed before data is shared or used for purposes other than legally allowed (Narayanan et al.2016). Data governance should, therefore, establish what specific policies are appropriate (Khatri and Brown 2010) and applicable across the organization (Malik2013). For example,Tallon(2013) states that organizations have a social and legal responsibility to safeguard personal data, whilstPower and Trope (2006) suggest that risks and threats to data and privacy require diligent attention from organizations.

Proposition 4. Organizations with an established data governance capability are more likely to be able to manage organizational and process changes introduced by the data science outcomes.

Proposition 4 considers the role of data governance as being important for coordinating and controlling the organizational consequences of data science outcomes.

3. METHODOLOGY

This paper describes two exploratory case studies using a multi-method approach to investigate the role of data governance as a boundary condition for data science. Case study is a widely adopted method for examining contemporary phenomena, such as the adoption of data governance (Choudrie and Dwivedi2005;Eisenhardt1989). In this research, we follow the design of an explanatory case study research proposed byYin(2009), including the research question, the propositions for research, the unit of analysis, and the logic linking the data to the propositions. As suggested byEisenhardt (1989), the research was contextualized by a review of background literature.

The literature background reveals that the results of data science initiatives are often not accepted by asset management organizations (Brous et al.2017). Data science initiatives often face a number of acceptance challenges in asset management organizations due, in part, to a lack of trust in the data science outcomes (Cao et al.2016;Yoon2017). Facing these challenges has led many asset management organizations to adopt data governance as a means of coordinating and controlling the impact of data science on organizations. However, data governance remains a poorly understood concept and its contribution to the success of data science has not been widely researched. As discussed above, our main research question therefore asks, how is data governance a boundary condition for data science? FollowingKetokivi and Choi(2014), deduction type reasoning provided the basic logic for the propositions to be tested in a particular context, namely data science in an asset management domain. According toKetokivi and Choi(2014), this general logic is augmented by contextual considerations. The data analysis in this research utilizes a combination of within-case analysis (Miles and Huberman 1994) and cross-case analysis, which enabled the delineation of the combination of factors that may have contributed to the outcomes of the case (Khan and Van Wynsberghe2008). In this research, the unit of analysis was a data science project in the asset management domain

Two case studies were selected. The first case study, ―Project A‖, was a data science project for the purpose of

predictive, ―just-in-time‖ maintenance of asphalted roads. The project was conducted under the auspices of a large European public organization tasked with the maintenance of national highways. The second case study, ―Project B‖, was a data science project for the purpose of discovering fraudulent use of electricity in medium and low tension electrical grids without impacting individual privacy rights. Table1below shows the properties of the two cases according to the subject, domain, organization size, organization type, number of datasets used, and the length (in time) of the project.

Table 1. Case selection.

Property	Project A	Project B
Case subject	Asphalt life expectancy	Fraud detection in electrical grids
Subject domain	Road management	Electrical Grid Management
Organization size	±4000 staff	±4000 staff
Organization Type	Government	Semi-government
Number of data sets	40+	10
Project length	3 months	18 months

The case studies were conducted using a multi-method approach. In order to prepare the respective organizations for the case studies, both organizations were provided with information material outlining the objectives of the research. Following the suggestions ofYin(2009), the case study research followed a research protocol. The research design was multi-method, and multiple data sources were used.

Primary data sources included the use of individual interviews. The interviews were conducted by the researchers over a period of two weeks. The interviews took place six months after the completion of the projects. The interviews were limited to one hour and followed a set line of questioning, although space was given during the interviews for follow-up questions in order to clarify descriptions or subjective statements. In both cases, two data scientists (interviewee 1 and 2), one enterprise data architect (interviewee 3), and two data governance officers (interviewee 4 and 5) were interviewed.

Secondary data sources included relevant market research and policy documents as well as websites. Internal policy documents were provided to the research team by the interviewees and the researchers were also given access to the organizations' intranet and internet websites. All documents reviewed were documents that are available in the public domain.

Triangulation of factors relating to the role of data governance as a boundary condition for data science case was made by listing aspects of data governance found in internal documentation and comparing these to the aspects of data governance exposed in the interviews, and matching these with the responses of the interviewees as to the contribution of these aspects towards the success of the project. Interviewees were also requested to provide feedback with regards to possible improvements.

4. FINDINGS

The results of the case studies were analyzed using a combination of within-case and cross-case analysis. In Sections4.1and4.2the within-case analysis is reported using the theory of the duality of technology as a guiding logic. The cases have been anonymized. Section4.3reports the cross-case analysis.

4.1. Project A: Asphalt Life Expectancy

The organization under whose auspices project A is managed is a public organization in Europe tasked with the management and maintenance of public infrastructure, including the construction and maintenance of roads. The organization has a budget of approximately €200 million per annum on asphalt maintenance, with operational parameters traditionally focused on traffic safety. According to interviewee 4, ―this has led to increasing overspend due either to premature maintenance, or too expensive emergency repairs in the past.‖ Interviewee 5 stated that the prediction of asphalt lifetime based on traditional parameters has been shown to be correct ―one-third of the time.‖

According to staff members, the organization has implemented data governance for their big data in order to remain ―future-proof, agile, and to improve digital interaction with citizens and partners.‖ According to an interviewee 3, ―(the organization) wants to be careful, open, and transparent about the way in which it handles big and open data and how it organizes itself.‖

4.1.1. Data Science as Product of Human Agency

The data science model utilized more than 40 different datasets which were fed into a data lake from the various source systems using data pipelines. These datasets included data related to traditional inspections, historical data

generated during the laying of the asphalt, road attribute data, and planning data, as well as automatically generated streaming data, such as weather data, traffic data, and IoT sensor data. The current model takes about 400 parameters into consideration. According to an interviewee 2, ―this number will only grow, as the (project partners) continue to supply new data.‖ The ultimate goal of the project is a model that can accurately predict the lifespan of a highway. In the model, higher-order relationships between the datasets were discovered using machine learning techniques such as decision trees, random forests, and naïve Bayes algorithms. Neural networks were used to reduce overfitting and improve generalization error, and gradient boosting was used to efficiently minimize the selected loss function.

The organization has implemented a policy of providing knowledge, tools, and a government-wide contact network in which best practices are shared with other government organizations. These best practices refer to organization of data management, data exchange with third parties, data processing methods, and individual training. Furthermore, the organization has introduced the policy of assessing and publishing the monetary cost of data assets in order to raise awareness of the importance of data quality management. According to interviewee 2, ―managers are required to know the cost of producing their data.‖ This means that every process and every organizational unit is encouraged to be aware of its data needs and the incurred costs. The data is then considered a strategic asset and considered to be a production input.

4.1.2. Data Science as Medium of Human Agency

The goal of project A is to reduce spending by extending the lifespan of asphalt where possible while reducing the number of emergency repairs made through predictive, ―just-in-time‖ maintenance. Using available big data in a more detailed manner, such as raveling data combined with vehicle overloading data, has doubled the prediction consistency. According to interviewee 1, improving the accuracy of asphalt lifetime prediction ―has enabled better maintenance planning, which has significantly reduced premature maintenance, improving road safety and cost savings, and reducing the environmental impact due to reduced traffic congestion and a reduction in CO2 emissions.‖

4.1.3. Organizational Conditions of Data Science

The organization has translated their policy and principles into a data strategy in which the opportunities, risks, and dilemmas of their policies and ambitions are identified in advance and are made measurable and practicable. Interviewee 3 reported that the organization has also asked the data managers in the organization to appoint a sponsor or data owner. By means of the above control and design measures, the organization ensures that the data ambitions are operationalized.

The organization has invested heavily in the fields of big data, open data, business intelligence and analytics. Interviewee 5 believed that ―the return (of the investment) stands or falls with the quality of data and information.‖ As such, according to the interviewee 5, ―the underlying quality of the data and information is very important to work in an information-driven way and as much as 70% of production time has been lost in almost every department due to inadequate data quality.‖ The organization has, therefore, implemented a data quality framework to improve its control of data quality. The data quality management process follows an eight-step process, which begins by identifying: 1. the data to be produced, 2. the value of the data for the primary processes, and 3. a data owner. The data owner is the business sponsor.

4.1.4. Organizational Consequences of Data Science

Once ownership had been established, the current and desired future situations were assessed in terms of production and delivery. Interviewee 2 reported that a roadmap was then established, which was translated into concrete actions. According to the interviewee, ―the final step in the process was the actual production and delivery of data in accordance with the agreement.‖ The organization has developed their own automatic auditing tool in combination with a manual auditing tool to monitor the quality of the data as a product in order to further improve its grip on data quality. According to interviewee 3, these tools―ensure that quality measurements were mutually comparable,‖ and cause changes in the conscious use of data as a strategic asset.‖ Data quality measuring is centralized; the goal is to ensure a standardized working method. However, the organization maintains the policy that every data owner is responsible for improvements to the data management process and the data itself. The data quality framework is based on fitness for use, and data quality measurement is maintained according to 8 main dimensions and 47 subdimensions. Terms and definitions are coordinated with legal frameworks related to the environment to ensure compliance. Responsibilities relating to compliance with privacy laws are centralized, and privacy officers are assigned to this role. The CIO has the final responsibility for ensuring that privacy and security are managed and maintained, however, data owners are responsible for ensuring compliance to dataset- specific policy and regulations

4.2. Project B: Fraud Detection in Electrical Grids

Project B is a data science project designed to detect the fraudulent use of electricity within medium and low

voltage grids without infringing on personal privacy rights. The project is managed under the auspices of a large European distribution grid operator (DGO). The role of the DGO is to transport the electricity from the high voltage grid to the end-user. The project was developed to improve the discovery rates of traditional methods utilized by the expensive commercial, off-the-shelf (COTS) system, which was in place at the time.

According to interviewee 5, the organization has implemented data governance as ―an integral part of their digital transformation strategy.‖ Interviewee 4 reported that ―the data governance team, the data science team, and the data engineering teams are managed within the same department and report directly to the Chief Data Officer.‖

4.2.1. Data Science as Product of Human Interaction

Project B was one of the first data science projects undertaken in the organization. Developing the data science capability within the organization required the development of a managed data lake and data pipelines to ensure connectivity from the data sources. Initially, the data science model was developed in an external data lake. According to interviewee 1, this meant that ―no automatic data pipelines between the internal data systems and the data science model could be established, so we were forced to improvise.‖ This meant that the data science model needed to be initially fed with batch uploads of data. This situation was eventually rectified with the development of an internal data lake which allowed connectivity with the original data source systems.

According an interviewee 2, the data science model initially utilized two sets of data originating from smart grid terminals, ―but was eventually expanded to include ten data sets after we had spent quite some time on discovery and after many hours of discussion and investigation.‖ Two years of training data were made available to the data scientists. According to the data scientists involved in the case, understanding the data was exceptionally difficult in this case. For example, during the project, it was discovered that the values in certain columns had been incorrectly labeled and needed to be corrected to attain the correct value, which corresponded to the required units. The data were not supplied with metadata, and finding subject matter experts with in-depth knowledge about the data was very difficult. For example, the data scientists discovered during the project that the OBIS codes did not follow the standardized values. The OBIS code is a unique identification of the registers in the smart meter's memory, according to IEC 62056-61.

Data were supplied by a subsidiary of the organization. The subsidiary was eventually sold to a third party during the project. This led to a situation whereby data owners were not available, and no single person could be found with a definitive knowledge of how the data were collected and collated. Data were collated and managed by two data engineers assigned to the project. Interviewee 4 reported that collaboration between the data engineers and the data scientists was not optimal as code was sometimes changed without sufficient documentation or collaboration. According to a data scientist 1 ―the engineer changed quite a lot of code without checking with us (the data scientists) first.‖

4.2.2. Data Science as Medium of Human Interaction

Reducing fraudulent usage of electricity on the middle and low voltage electrical grids without infringing on personal privacy rights is of importance for a number of reasons, although few of the reasons are directly related to the DGO itself. Fraudulent usage of electricity is essentially theft, as electricity is being used without paying the provider for the service. According to interviewee 4, ―in middle and low tension grids it is especially hard to decide from whom the fraudster is stealing electricity, because there are multiple electricity providers who sell their electricity directly to the end-user but use the common grid to transport the electricity.‖ The fraudster is essentially taking electricity out of a shared service, so it is impossible to know from whom electricity is being stolen. Furthermore, fraudsters that are caught generally only have to pay the net stolen kWh, although damage is also suffered by the network operator. This amount, the so-called ―grid loss‖, is 70% lower than the price that consumers pay.

It is important to know how much energy is being used on the grid in advance in order to be able to balance the use of energy with the supply so that the grid is not overloaded. However, balancing of the entire electricity supply is generally performed by the transmission system operator (TSO), which manages the high voltage grid.

Catching fraudsters also requires collaboration with a number of parties, including the police. Moreover, European privacy laws dictate that the end-user is the owner of the data collected by the electricity meters, which means that DGOs are not able to read the values without permission from the end-users, which fraudsters are unlikely to give. According to interviewee 5 it is often difficult to coordinate a response to combating fraud, whilst the rewards for fraudulent usage remain high――we are always behind fraudsters as catching them is expensive, whilst there is almost no risk for them.‖ From a data governance perspective, this makes it especially difficult to coordinate and control the proper collection and collation of the required data

4.2.3. Organizational Conditions of Data Science

The data science projects in the organization are decided upon and prioritized by managers of the primary business processes. The data scientists work according to sprints of two weeks, according to directions suggested by the product owner. The disruption to the project caused by the sale of the subsidiary mean that a new product owner as well as new data owners need to be found within the organization. According to one of the data scientists, the data owners are necessary ―to be able to coordinate and control the proper collection, collation, and management of the data, provide input to the data scientists regarding the content of the data (metadata) and accept and control the quality of the data science outcomes.‖ This means that data governance officers and privacy officers attached to the department were required to develop a roles and responsibilities matrix for the management of the data and the use of the data, in concurrence with privacy regulations.

4.2.4. Organizational Consequences of Data Science

Despite the technological and social challenges faced during the project, the data science team reported that after an extended period of 18 months, they were able to present a workable model that greatly outperformed traditional methods of fraud detection. The model was presented to the energy management team which had been identified as the client and the main data owner. The data science team reported that the presentation was not well-received and that the model was eventually not adopted, despite the proven improvements. The data science team believed that the reason for this was that ―they didn't want to believe the results. (The organization) has spent millions on the COTS system, and they are reluctant to accept that they've made a procurement error. Their argument was that the data was unreliable, but technically it's the same data being used by the COTS system.‖ This reaction suggests that end-users as well as data owners should be an integral part of the data science project and that not only results but also intentions should be tested throughout the project.

4.3. Cross-Case Analysis

In the cross-case analysis, the results of the case studies were analyzed in comparison to the relative maturity of the data science capability as reported by the interviewees, the perceived success of the data science outcomes from the perspective of the project team, and whether or not the outcomes were accepted and adopted within the primary business processes. Table2below compares the two case studies based on data governance maturity, data science outcomes, and the adoption status of the data governance outcomes.

Table 2. Comparison of the case studies.

Project	Data Science Maturity	Data Science Outcomes	Business Acceptance of Data Science Outcomes
Project A	Established	Demonstrated improvement on traditional methods	Data science outcomes integrated within business processes in combination with traditional methods
Project B	Initial	Demonstrated improvement on traditional methods	Data science outcomes not accepted by business

Table2above shows that Project A has an established data governance capability and that the outcomes of the project were accepted by the business. In Project B, the organization does not have an established data governance capability, and the data science outcomes were not adopted by the business data science as an organizational capability is compared between the cases.

Journal of Positive Psychology and Wellbeing
2020, Vol. 4, Issue 2
Pp 30-44

4.3.1. The Role of Data Governance with Regards to Data Science as a Product of Human Agency In Table3below, the role of data governance with regards to the successful implementation of

Table 3. Comparison of the cases with regards to the role of data governance in data science as a product of human agency.

Role of Data Governance	Project A	Project B
Coordinate and control data science capabilities e.g., coordination and control of: • Data access and availability • Compliance (privacy and security) • Data quality • Data engineering, and predictive analytics skills • Algorithm management	Although much of the infrastructure and required staff capacity was available, the coordination and control provided by business leaders provided good, previously unknown insight into the available data and improved the analytical skills of the data scientists. Much of the data used in project A were open data, which were readily available. This allowed the project to progress according to schedule, even with large amounts of data.	During the project, the entire IT and data infrastructure needed to be developed. The disruption caused by the sale of the subsidiary demonstrated the importance of business leadership in coordinating and controlling the quality and trust in the data science outcomes. Data access was a major issue throughout the project, technically and ethically. The lack of data connectivity created long delays and privacy regulations needed to be checked and adhered to. This was exacerbated by the lack of business leadership to monitor and control data access.

In Table3above, we notice that although Project A required more data sets than Project B, data access was not considered an issue, and the project was able to be completed with a minimum of extra effort. In contrast, project B team members were required to set up the data infrastructure, find the data, and manage access and data quality themselves.

4.3.2. The Role of Data Governance with Regards to Data Science as a Medium of Human Agency In

Table4below, the role of data governance with regards to the acceptance, coordination, and control of data science outcomes is compared between the cases.

Table 4. Comparison of the cases with regards to the role of data governance in data science as a medium of human agency.

Role of Data Governance	Project A	Project B
Increase the business value of data assets—e.g., oversee data usage with regards to: • Acceptance of data science outcomes for risk management • Acceptance of data science outcomes for improved efficiency of primary processes • Acceptance of data science outcomes for improved effectiveness of primary processes	Project A data science outcomes showed an improved result in contrast to traditional methods. The outcomes were accepted and adopted by the business. Data owners were able to monitor the development of the data science outcomes throughout the process and had ownership of the results.	Although project B demonstrated a marked improvement in traditional methods, the business did not choose to adopt the data science outcomes. Business leaders were not a part of the data science project, and results were presented only at the end of the project.

From Table4it becomes clear that in project A, data owners were involved from the start of the project until delivery. In addition, data owners were accorded ownership of the outcomes. As a result, the outcomes were accepted by the data owners. This is in contrast to Project B, in which data owners were not available, and business owners did not accept the data science results.

4.3.3. The Role of Data Governance with Regards to Organizational Conditions of Data Science

In Table5below, the role of data governance with regards to the coordination and control of organizational conditions of data science is compared between the cases.

Table 5. Comparison of the cases with regards to the role of data governance in coordinating and controlling the organizational conditions of data science.

Role of Data Governance	Project A	Project B
Ensure that relevant coordination and control mechanisms are in place	Principles and policies have been adopted in a data strategy that follows an annual cycle of planning and control.	Project B did not benefit from a strict regime of planning and control, and the data science team were required to be self-managing.
Ensure sufficient resources (budget and staffing) are available	Budget and staffing is monitored and tested according to the data agenda.	Resources were limited to the initial project team, although there were no time pressures for delivery. As a result, the project duration was much longer than initially expected.
Ensure roles and responsibilities are sufficiently filled	Each business unit is required to appoint a data owner for every data set it manages.	Business leaders were difficult to find, and no owners were available.

From Table5we can conclude that Project A has a strict regime of coordination and control following a yearly review as well as well-defined roles and responsibilities. In contrast, team members of Project B were given little direction and no ownership was displayed by business leaders.

4.3.4. The Role of Data Governance with Regards to Organizational Consequences of Data Science In

Table6below, the role of data governance with regards to the coordination and control of organizational consequences of data science is compared between the cases.

Table 6. Comparison of the cases with regards to the role of data governance in coordinating and controlling organizational consequences of data science.

Role of Data Governance	Project A	Project B
Coordinate and control new costs related to data science: e.g., • IT infra costs • HR costs	IT infra costs and staffing costs are managed centrally according to demand.	The project team was assigned by the department manager. The project team was a part of the data office team, the costs of which were managed centrally. The acquisition of an expensive commercial, off-the-shelf (COTS) application that provided similar functionality conflicted with the data science outcomes.
Coordinate and control organizational changes as a result of data science: e.g., • new departments • new business processes • new roles and responsibilities	The management of organizational change is handled by the project manager, who oversees the data science project from initiation to adoption. Business units are required to name data owners.	The management of the new data office team fell under the control of the chief data officer who reports to the chief strategy officer. The primary business was not involved in time management. Data ownership was not defined.
Coordinate and control new risks related to data science: e.g., • non-compliance • incorrect decisions due to incorrect data • need for increased security • lack of trust in data science outcomes	Privacy and security issues are managed by means of privacy impact assessments and baseline information security testing.	Compliance was monitored by data privacy officers through privacy impact assessments, and assistance was given to the data science team by data governance officers.

From Table6it can be derived that in Project A, the implementation of the data science outcomes was managed by a dedicated project manager in conjunction with the data owners. This was in contrast to Project B, in which no data owners were involved and a rival COTS application which had previously been acquired by business leaders created an insurmountable conflict for the project team

5. DISCUSSION

Case study methodology was used in this research to identify the role that data governance plays as a success factor for data science. The choice for an in-depth case study was based on the contemporary nature of both data science and data governance. The study was conducted on the basis of two case studies in different organizations, and the results should be regarded in this light. The study was conducted in the asset management domain as asset management organizations by nature are often data-rich due to the need to monitor the state of the infrastructure assets. This may limit the applicability of the study for domains which are less data intensive, however the essence of generating value from data is likely to be the same in other domains.

5.1. Proposition 1. Organizations with an Established Data Governance Capability Are More Likely to Have a Well-Functioning Data Science Capability

With regards to Proposition 1, which proposes that organizations with an established data governance capability have better functioning data science capabilities, the results of the case studies suggest that when data governance has been actively implemented before the start of a data science project, the complexity of issues such as access to data and the understanding of the data is greatly reduced. The use of big data in data science projects often leads to serious data quality (Saha and Srivastava2014) and compliance (Narayanan et al.2016) issues which can be di fficult to manage in a timely manner (Hazen et al.2014). Data governance policies and principles (Madera and Laurent 2016) and a responsible data governance strategy (Kroll2018) should therefore be key components of data science technologies. This suggests that data governance plays an important role in ensuring the effectiveness and efficiency of the data science capability in an organization.

5.2. Proposition 2. Organizations with Established Data Governance Capability Are More Likely to Generate Trusted Data Science Outcomes

Proposition 2 suggests that organizations with an established data governance capability are better positioned to produce trusted data science decision outcomes. The results of the case studies confirm that data science projects in which data owners have a direct influence on the project from start to finish are more likely to generate trusted outcomes. Data governance is important for creating value and moderating risk in data science initiatives (Foster et al.2018), as the trustworthiness of data science outcomes in practice is often affected by tensions arising through ongoing forms of work (Passi and Jackson2018). This suggests that data governance plays an important role in creating trust in data science outcomes and positively influencing the use and acceptance of data science outcomes in the organization.

5.3. Proposition 3. Organizations with an Established Data Governance Capability Are More Likely to Ensure that Organizational Conditions of Data Science are Met

Successful data science outcomes require data governance mechanisms beginning with policy development to define governance goals and strategies (Wang et al.2019), followed by the establishment of organizational data governance structures. Top management support (Gao et al.2015), well-defined roles and responsibilities (Saltz and Shamshurin2016), and the choice of the data governance approach (Koltay2016) are considered critical. Proposition 3 proposes that organizations having an established data governance capability are more likely to be in a position to meet organizational conditions. In this regard, the case studies suggest that a regime of coordination and control of data management processes, following a regular cycle, as well as well-defined roles and responsibilities, play important roles in developing ecosystems in which data science projects are more likely to be successful.

5.4. Proposition 4. Organizations with an Established Data Governance Capability Are More Likely to Be Able to Manage Organizational and Process Changes Introduced by Data Science Outcomes

Data governance establishes data management processes which manage data quality (Passi and Jackson2018) and compliance with relevant laws, directives, and policies (Cato et al.2015). Data governance aligns policies and principles with business strategies in an enterprise data strategy (Cato et al.2015). Proposition 4 proposes that organizations with mature data governance are more likely to be able to manage changes introduced by data science decision outcomes. In this regard, the results of the case studies suggest that organizations which have a well-developed data governance capability are more likely to be able to manage new costs arising from changes in staff and technology, manage changing risks arising from changes in primary processes, and manage organizational and process changes introduced by the acceptance of data science outcomes within the business.

6. CONCLUSIONS

In this paper we analyzed two data science case studies in the asset management domain in order to understand the role of data governance as a boundary condition for creating trust in data science decision outcomes. The first case under study was a data science project which predicts the maintenance requirements of asphalt on national highways over time. The second case study was a data science project which discovers the fraudulent use of electricity in a middle- and low-level voltage grid. The results of the case studies suggest that data science decision outcomes are more likely to be accepted if the organization has an established data governance capability. Furthermore, the results suggest that organizations with an established data governance capability are more likely to have a well-functioning data science capability, are more likely to generate trusted data science outcomes, are more likely to ensure that organizational conditions of data science are met, and are more likely to be able to manage organizational and process changes introduced by the data science decision outcomes. These results confirm the propositions of the research and we conclude that data governance is a boundary condition for managing the organizational consequences of data science outcomes. Viewing the acceptance of data science decision outcomes for decision-making in organizations as a socio-material challenge in which trust plays a central role implies that the analysis and interpretation of data is tightly coupled with the governance and proper management of that data. Simply ―throwing data‖ at a problem without regard for the quality or bias of the data or the algorithm itself does not necessarily lead to acceptance of the decision outcomes. Rather, it is necessary to look at the development of trustworthy data science decision outcomes not as a purely technical problem, requiring a technical solution, but as one in which human agency and organizational forces play a significant role. This approach also has practical implications, as managers responsible for data science should ensure that the data governance capability of the organization is well established before the focus is placed on the development of the data science capability. The research was limited to two data science projects in (semi)-government organizations within the asset management domain. Further investigation with regards to data science projects with different scopes, domains, and organizations is recommended.

Conflicts of Interest:

The authors declare no conflict of interest.

References

1. Adrian, Cecilia, Rusli Abdullah, Rodziah Atan, and Yusmadi Yah Jusoh. 2017. Factors influencing to the implementation success of big data analytics: A systematic literature review. Paper presented at 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, Malaysia, July 16–17; pp. 1–6.
2. Alofaysan, Sarah, Bandar Alhaqbani, Rana Alseghayyir, and Maryam Omar. 2014. The significance of data governance in healthcare: A case study in a tertiary care hospital. In Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies-Volume 5. Eseo: SCITEPRESS-Science and Technology Publications, pp. 178–87.
3. Al-Ruithe, Majid, Elhadj Benkhelifa, and Khawar Hameed. 2019. A systematic literature review of data governance and cloud data governance. Personal and Ubiquitous Computing 23: 839–59.
4. Brain, Damien, and Geoffrey I. Webb. 2002. The Need for Low Bias Algorithms in Classification Learning from Large Data Sets. In Proceedings of the Principles of Data Mining and Knowledge Discovery. Edited by T. Elomaa,
5. H. Mannila and H. Toivonen. Berlin and Heidelberg: Springer, pp. 62–73.
6. Brous, Paul, Paulien Herder, and Marijn Janssen. 2016. Governing Asset Management Data Infrastructures.
7. Procedia Computer Science 95: 303–10.
8. Brous, Paul, Marijn Janssen, and Rutger Krans. 2020. Data Governance as Success Factor for Data Science.
9. In Responsible Design, Implementation and Use of Information and Communication Technology. Edited by Marié Hattingh, Machdel Matthee, Hanlie Smuts, Ilias Pappas, Yogesh K. Dwivedi and Matti Mäntymäki. Skukuza: Springer International Publishing, pp. 431–42.
10. Brous, Paul, Marijn Janssen, Daan Schraven, Jasper Spiegeler, and Baris Can Duzgun. 2017.
11. Factors Influencing Adoption of IoT for Data-driven Decision Making in Asset Management Organizations. In IoTBDS. Porto: Scitepress, pp. 70–79. Available online:http:
12. //www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/ 0006296300700079(accessed on 1 June 2020).
13. Busse, Christian, Andrew P. Kach, and Stephan M. Wagner. 2017. Boundary Conditions: What They Are, How to Explore Them, Why We Need Them, and When to Consider Them. Organizational Research Methods 20: 574–609.
14. Cao, Quyet H., Imran Khan, Reza Farahbakhsh, Giyyarpuram Madhusudan, Gyu Myoung Lee, and Noel Crespi. 2016. A Trust Model for Data Sharing in Smart Cities. In IEEE International Conference on Communications 2016 (ICC 2016). Kuala Lumpur: IEEE.
15. Cato, Patrick, Philipp Gölzer, and Walter Demmelhuber. 2015. An investigation into the implementation factors affecting the success of big data systems. In 2015 11th International Conference on Innovations in Information Technology (IIT). Dubai: IEEE, pp. 134–39.
16. Cecere, Grazia, Fabrice Le Guel, and Nicolas Soulié. 2015. Perceived Internet privacy concerns on social networks in Europe. Technological Forecasting and Social Change 96: 277–87.

17. Choudrie, Jyoti, and Yogesh Kumar Dwivedi. 2005. Investigating the research approaches for examining technology adoption issues. Journal of Research Practice 1: D1.
18. Council on Library and Information Resources, ed. 2000. Authenticity in A Digital Environment. Washington: Council on Library and Information Resources.
19. DAMA International. de Medeiros, Mauricius Munhoz, Norberto Hoppen, and Antonio Carlos Gastaud Maçada. 2020. Data science for business: Benefits, challenges and opportunities. The Bottom Line 33: 149–63.
20. Dhar, Vasant. 2013. Data science and prediction. Communications of the ACM 56: 64–73.
21. Dwivedi, Yogesh K., Marijn Janssen, Emma L. Slade, Nripendra P. Rana, Vishanth Weerakkody, Jeremy Millard, Jan Hidders, and Dhoya Snijders. 2017. Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modelling. Information Systems Frontiers 19: 197–212.
22. Eisenhardt, Kathleen M. 1989. Building Theories from Case Study Research. The Academy of Management Review 14: 532–50.
23. Foster, Jonathan, Julie McLeod, Jan Nolin, and Elke Greifeneder. 2018. Data work in context: Value, risks, and governance: Data Work in Context: Value, Risks, and Governance. Journal of the Association for Information Science and Technology 69: 1414– 27.
24. Gama, João. 2013. Data Stream Mining: The Bounded Rationality. Informatica 37: 6.
25. Gao, Jing, Andy Koronios, and Sven Selle. 2015. Towards A Process View on Critical Success Factors in Big Data Analytics Projects. (accessed on 10 June 2020).
26. Giddens, Anthony. 1976. New Rules of Sociological Method: A Positive Critique of Interpretative Sociology. Stanford: Stanford University Press.
27. Gigerenzer, Gerd, and Reinhard Selten. 2002. Bounded Rationality: The Adaptive Toolbox. Cambridge: MIT Press.
28. Hazen, Benjamin T., Christopher A. Boone, Jeremy D. Ezell, and L. Allison Jones-Farmer. 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics 154: 72–80.
29. Janssen, Marijn, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly 37: 101493.
30. Jones, Kerina, Elizabeth Ford, Nathan Lea, Lucy Griffiths, Sharon Heys, and Emma Squires. 2019. Developing data governance standards for using free-text data in research (TexGov). International Journal of Population Data Science 4.
31. Ketokivi, Mikko, and Thomas Choi. 2014. Renaissance of case research as a scientific method. Journal of Operations Management 32: 232–40.
32. Kezunovic, Mladen, Le Xie, and Santiago Grijalva. 2013. The role of big data in improving power system operation and protection. Paper presented at 2013 IREP Symposium Bulk
33. Power System Dynamics and Control—IX Optimization, Security and Control of the Emerging Power Grid (IREP), Rethymno, Greece, August 25–30.
34. Khan, Samia, and Robert Van Wynsberghe. 2008. Cultivating the under-mined: Cross-case analysis as knowledge mobilization. Forum qualitative Sozialforschung/Forum: Qualitative Social Research 9.
35. Khatri, Vijay, and Carol V. Brown. 2010. Designing data governance. Commun. ACM 53: 148–52. Koltay, Tibor. 2016. Data governance, data literacy and the management of data quality. IFLA Journal 42: 303–12.
36. Kroll, Joshua A. 2018. Data Science Data Governance [AI Ethics]. IEEE Security Privacy 16: 61–70. Ladley, John. 2012. Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program.
37. London: Academic Press.
38. Lin, Shien, Jing Gao, and Andy Koronios. 2006. The Need for A Data Quality Framework in Asset Management.
39. Madera, Cedrine, and Anne Laurent. 2016. The Next Information Architecture Evolution: The Data Lake Wave. In Proceedings of the 8th International Conference on Management of Digital EcoSystems. New York: Association for Computing Machinery, pp. 174–80.
40. Malik, Piyush. 2013. Governing Big Data: Principles and practices. Ibm Journal of Research and Development 57: 1.
41. Miles, Matthew B., and A. Michael Huberman. 1994. Qualitative Data Analysis: An Expanded Sourcebook. London: SAGE.
42. Miloslavskaya, Natalia, and Alexander Tolstoy. 2016. Big Data, Fast Data and Data Lake Concepts. Procedia Computer Science 88: 300–5.
43. Morabito, Vincenzo. 2015. Big Data Governance. In Big Data and Analytics. Cham: Springer, pp. 83–104. Narayanan, Arvind, Joanna Huey, and Edward W. Felten. 2016. A Precautionary Approach to Big Data Privacy. In Data Protection on the Move. Edited by Serge Gutwirth, Ronald Leenes and Paul De Hert. Cham: Springer, vol. 24, pp. 357–85.
44. Newell, Allen, and Herbert Alexander Simon. 1972. Human Problem Solving. Englewood Cliffs: Prentice-Hall, vol. 104.
45. Nunn, Sandra L. 2009. Driving Compliance through Data Governance. Journal of AHIMA 80: 50–51.
46. Orlikowski, Wanda J. 1992. The Duality of Technology: Rethinking the Concept of Technology in Organizations.
47. Organization Science 3: 398–427.

48. Otto, Boris. 2011. Data Governance. Business and Information Systems Engineering 3: 241– 44.
49. Panian, Zeljko. 2010. Some practical experiences in data governance. World Academy of Science, Engineering and Technology 38: 150–57.
50. Paskaleva, Krassimira, James Evans, Christopher Martin, Trond Linjordet, Dujuan Yang, and Andrew Karvonen. 2017. Data Governance in the Sustainable Smart City. Informatics 4: 41.
51. Passi, Samir, and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. Proceedings of the ACM on Human- Computer Interaction 2: 1–28. Power, E. Michael, and Roland L. Trope. 2006. The 2006 survey of legal developments in data management, privacy, and information security: The continuing evolution of data governance. Business Lawyer 62: 251–94.
52. Provost, Foster, and Tom Fawcett. 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data 1: 51–59. [PubMed]
53. Randall, Robert, Don Peppers, and Martha Rogers. 2013. Extreme trust: The new competitive advantage. Strategy and Leadership 41: 31–34.
54. Raza, Ahmed, and Vladimir Ulansky. 2017. Modelling of predictive maintenance for a periodically inspected system. Procedia CIRP 59: 95–101.
55. Saha, Barna, and Divesh Srivastava. 2014. Data quality: The other face of big data. In 2014 IEEE 30th International Conference on Data Engineering. Chicago: IEEE, pp. 1294–97.
56. Saltz, Jeffrey S., and Ivan Shamshurin. 2016. Big data team process methodologies: A literature review and the identification of key factors for a project's success. In 2016 IEEE International Conference on Big Data (Big Data). Washington: IEEE, pp. 2872–79.
57. Simon, Herbert A. 1947. Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization. New York: Macmillan.
58. Tallon, Paul P. 2013. Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. Computer 46: 32–38.
59. Ullah, Saeed, M. Daud Awan, and M. Sikander Hayat Khiyal. 2018. Big Data in Cloud Computing: A Resource Management Perspective. Scientific Programming.
60. van den Broek, Tijs, and Anne Fleur van Veenstra. 2018. Governance of big data collaborations: How to balance regulatory compliance and disruptive innovation. Technological Forecasting and Social Change 129: 330–38.
61. Waller, Matthew A., and Stanley E. Fawcett. 2013. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. Journal of Business Logistics 34: 77–84. Wang, Chen-Shu, Shiang-Lin Lin, Tung-Hsiang Chou, and Bo-Yi Li. 2019. An integrated data analytics process to optimize data governance of non- profit organization. Computers in Human Behavior 101: 495–505.
62. Webster, Jane, and Richard T. Watson. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly 26: 13–23.
63. Wilbanks, Debra, and Karen Lehman. 2012. Data governance for SoS. International Journal of System of Systems Engineering 3: 337–46.
64. Yin, Robert K. 2009. Case Study Research: Design and Methods. Thousand Oaks: SAGE.
65. Yoon, Ayoung. 2017. Data reusers' trust development. Journal of the Association for Information Science and Technology 68: 946–56.
66. Zuiderwijk, Anneke, and Marijn Janssen. 2014. The Negative Effects of Open Government Data—Investigating the Dark Side of Open Data. In Proceedings of the 15th Annual International Conference on Digital Government Research. New York: Association for Computing Machinery, pp. 147–52.