Data science as knowledge creation a framework for synergies between data analysts and domain professionals

van der Voort, Haiko; van Bulderen, Sabine; Cunningham, Scott; Janssen, Marijn

DOI
10.1016/j.techfore.2021.121160

Publication date
2021

Document Version
Final published version

Published in
Technological Forecasting and Social Change

Technological Forecasting & Social Change 173 (2021) 121160

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Data science as knowledge creation a framework for synergies between data analysts and domain professionals

Haiko van der Voort [a,*], Sabine van Bulderen [b], Scott Cunningham [c], Marijn Janssen [d]

[a] Delft Unversity of Technology, Faculty Technology, Policy and Management, Delft, the Netherlands
[b] IVO Rechtspraak, Utrecht, the Netherlands
[c] University of Strathclyde, Glasgow, United Kingdom
[d] Delft Unversity of Technology, Faculty of Technology, Policy and Management, Delft, the Netherlands

ARTICLE INFO

Keywords:
Data science
Knowledge
Predictive model
Value creation
Risk-based inspection
Professionalism

ABSTRACT

The road from data generation to data use is commonly approached as a data-driven, functional process in which domain expertise is integrated as an afterthought. In this contribution we complement this functional view with an institutional view, that takes data analysis and domain professionalism as complementary (yet fallible) knowledge sources. We developed a framework that identifies and amplifies synergies between data analysts and domain professionals instead of taking one of them (i.e. data analytics) at the centre of the analytical process. The framework combines the often-cited CRISP-DM framework with a knowledge creation framework. The resulting framework is used in a data science project at a Dutch inspectorate that seeks to use data for risk-based inspection. The findings show first support of our framework. They also show that whereas more complex models have a higher predictive power, simpler models are sometimes preferred as they have the potential to create more synergies between inspectors and data analyst. Another issue driven by the integrated framework is about who of the involved actors should own the predictive model: data analysts or inspectors.

1. Introduction: towards synergy between competing intelligence sources

In the past years data intelligence and analytics have sustainably proven their value to many organizations (Höchtl et al., 2016; Janssen and Kuk, 2016; Taylor and Portuva, 2019). Existing data mining frameworks, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) enable organizations to learn in concert with data (Shafique and Qaiser, 2014; Sharma, Osei-Bryson and Kasper, 2012). These processes propose a structured and iterative sequence of activities, such as problem formulation, data consultation, and analytical modelling. Most frameworks read as a functional chain from data generation towards decision-making. They show how data get generated, processed and made ready for those that have to make decisions, either on a political or on an operational level (Janssen, van der Voort and Wahyudi, 2017).

However, these models hardly consider the interests and views of the people that need to make use of the outcomes of the data science efforts. Especially professional workers may have knowledge that potentially competes with knowledge derived from data intelligence and analytics (van der Voort et al., 2019). This professional knowledge typically is derived from their personal experience in the field they work (Abbott, 2014; Freidson, 2001). For many organizations the true repository of knowledge is not a database or a warehouse, but the often tacit knowledge of daily practitioners and field workers (Polanyi, 1966). As a consequence, professionals may not see themselves as workers doing an activity at the end of a functional chain that starts with data analytics. Professionals may be able to interpret the outcomes, may disagree with them, may misinterpretate them or even neglect them.

This idea has important consequences for the way data intelligence and analytics can help to improve decision-making. The underlying premise of our research is that intelligence from data analytics and from domain knowledge are both valuable and sometimes competing. Improving decision-making processes, then, is about creating synergy between these (fallible) sources instead of taking one of them (i.e. data analytics) to the centre of the analytical process.

This article will address some critical organizational literature as a contribution to our thinking about data science processes. Our main question is:

How to amplify synergies between data analysts and domain

* Corresponding author.
E-mail addresses: h.g.vandervoort@tudelft.nl (H. van der Voort), scott.cunningham@strath.ac.uk (S. Cunningham), m.f.w.h.a.janssen@tudelft.nl (M. Janssen).

https://doi.org/10.1016/j.techfore.2021.121160
Received 13 January 2021; Received in revised form 11 August 2021; Accepted 14 August 2021
Available online 1 September 2021

professionals?

This paper is structured as follows. First – in section 2 - we will root our concern about synergy in the literature. We will distinguish two views on the process of data intelligence and analytics (from now on: 'data science') towards decision-making: a functional and an institutional view. The research approach is presented in section 3. In section 4 we will describe two prescriptive frameworks on operational (i.e. professional) decision-making that are exemplary for both views: CRISP DM and the knowledge creation process framework by Nonaka. The next step will be combining the two frameworks to respect both views. This will be discussed in section 5. In an empirical section we show the use of the framework into the Netherlands Food and Safety Authority (NVWA), using action research methods (section 6). Section 7 provides reflections and dilemmas that help the framework to be put into practice. Finally, conclusions are drawn in section 8.

2. Two views on the data science process[1]

2.1. A functional view on data science

Provost and Fawcett (2013) define data science as "a set of fundamental principles that support and guide the principled extraction of information and knowledge from data." (p. 52). Data science helps using data for all kinds of decisions. It includes a wide variety of activities, including data generation, data processing and data use (Sivarajah et al., 2017; Van der Voort et al., 2018).

A broad range of scientific publications resonate enthusiasm by listing promises of data science. Without pretending to be exhaustive, we list two key promises (Vydra and Klievink, 2019). A first promise of data science refers to its quality to produce more accurate and efficient information for decisionmaking and service provision. An increasing amount of data and – particularly – increased data processing capacity would enable analysts to serve managers and policy-makers with real-time and evidence-based information (OMalley, 2014). This makes information provision to decision-makers more accurate (Höchtl, Parycek, and Schollhammer, 2016), more efficient and more reliable (Hilbert, 2016).

A second key promise is that better information would lead to better decision-making (Höchtl et al., 2016). This claim feels intuitive since decisions need some level of intelligence. Policy decisions are assumed to be outputs of a process, for which data are the main inputs. Optimization of inputs would lead to better outputs based on inputs (Hilbert, 2016; Maciejewski, 2016).

For some organizations or in some conditions, these promises may prove enough to meet their ends. However, sometimes the promises of big data come with assumptions about organizational life that are more theoretical than practical. In fact, the promises feel functional, even a bit machine-like. If all elements work properly, it will produce better information for decision-makers. The promises rely on predefined mechanisms that potentially produce predefined results. As far as human activities are involved, humans are assumed to behave according to a common goal, being better decision-making informed by better information. Vydra and Klievink (2019) call this stream of literature "techno-optimists" that "focus on humans turning data into insight and humans making decisions in bureaucratic structures (with the help of that insight)".

2.2. An institutional view on data science

To have some more insight into the 'organizational life' it is helpful to take a further look at these human activities. The chain of activities from data generation to data use is typically inhabited by multiple persons and organizations ('actors') with different specializations and different interests. The way this process is organized is key to the quality of data use (Janssen, van der Voort and Wahyudi, 2017). An institutional view on data science takes the human actors and their behaviour to the centre of analysis. They may have their own perceptions of the goals to pursue (see for instance Arnaboldi, 2018; Van der Voort et al., 2018; Vydra and Klievink, 2019).

A key assumption behind the functional promise of data science is that data are pre-given and universal, even neutral. This refers to an ideal of machine objectivity as an objectivity of a mechanism being void of human bias, of detachment (Daston and Galison, 1992). This assumption is crucial for outcomes to be 'evidence-based'. However, from an institutional perspective this assumption can be contested. First, if we focus on actors, we see that the data science process very much relies on human interventions. Data are generated for a certain purpose found by humans, as are the systems through which the data go, humans interpret the data and transfer those interpretations to users, who are in turn make their interpretation as well. The humans involved vary, both on their knowledge and their interests. As a consequence, data science is always prone to human design, bias and error (Crawford, Miltner and Gray, 2014). The consequence is that the approaches result in false positives, relationships that do not exist and biased predictions (Janssen and Kuk, 2016). Understanding the context might be key to being able to analyze and make sense of data.

Framed positively, each actor provides his or her own added value to the process from data generation to data use, because they add their personal skills and knowledge to it. This is of crucial importance for the way we must perceive the data science process. If we assume that data is not neutral or universal, we must assume that data become personalized along the process (Daston and Galison, 1992), which would mean that the functional claim that data are universal is debatable. Data generation, processing, dashboard development and the design of all interfaces involved are done by humans, and these humans will make data – slightly and implicit – more personal. Personal knowledge and skills will leak into the data stream, finding its way to data use.

Also, the link between information and the quality of decision-making – the second assumption - can be contested from an institutional view. For instance, visualizations may prove reductions of reality in order to improve interpretation (Huff, 1993). However, here filtering and framing processes are in place (Arnaboldi, 2018), making the link between information and decision-making more subjective than assumed by the functionalists. More fundamentally, decision-makers are assumed to be a bit passive from a functional view: they are the receiving actor being served by better information. In contrast, they can also be seen as autonomous agents that may or may not be open to the information provided by data analysts (Van der Voort et al., 2018). There are plenty of possible reasons to show reluctance here, such as information overload (Feldman and March, 1981), inability to interpret (Janssen and Kuk, 2016), a wish to legitimize predefined ideas (Kogan, 1999) and accountability problems (Reddy, Cakici and Ballestero, 2019). Especially accountability issues are commonly addressed in literature. It is not clear to everyone how data are generated, how algorithms are detected, how they work out and how data are interpreted (Madsen, 2018). If data science is a black box for decision-makers, the legitimacy of their decisions is at stake (Redden, 2018; Van der Voort et al., 2018). This is even more so if the promises of data science become contested, because of possible flaws, like false positives (boyd and Crawford, 2012).

2.3. Alignment between data analysts and professionals

Table 1 serves as an illustration of the two views on data science and decision-making. Again, without claiming to be exhaustive.

Both views have their inherent logic, but also include weaknesses. As discussed, an overly functional view tends to neglect the abilities (and legitimacy) of domain experts to challenge data intelligence, as they may not be the last in line of a functional chain dominated by data

[1] We thank Deniz Özagan for doing ground work for this section

Table 1.
Comparing views on data science and decision-making.

	Functional view	Institutional view
Focus	Activities	Actors
Professionals	Receivers of information	Autonomous agents
Lead motive for data science efforts	Common goal	Common goal, but interpreted individually
Data	Neutral, objective, especially when in great numbers	Essentially prepared and treated by humans, becoming more personal along the process
Key success factor	The ability and capacity to analyse data into useful information	Alignment between supply of data and demand of information

analysts. They also possess tacit knowledge that can be taken into account when analysing data and their involvement might be key to overcomming resistance. On the other hand, an overly institutional view tends to lack overall rationality since it draws the attention towards the negotiation between actors rather than a common rationale.

If professionals - as operational decision-makers - are seen as autonomous agents rather than just receivers of better information, then the key success factor for data science may reflect both rationales. The alignment between data analysts and decision-makers (Arnaboldi, 2018; Bhimani and Willcocks, 2014) may hold sweet promises. This alignment issue can be viewed as the connection between the supply of data and the demand of information, which is in its nature an organizational rather than a technical issue (Van der Voort et al., 2018). The alignment requires to involve decision-makers in the processes of data sciences (Bhimani and Willcocks, 2014). It also refers to a problem of knowledge. Decision-makers as agents are already knowledgeable in their own right and have their own values and mental models (Bhimani and Willcocks, 2014; Boisot and Canals, 2008). The agent's knowledge does not necessarily align with knowledge as derived from data science efforts.

For drawing a prescriptive framework for the data science process - which we will do in section 5 - we will look for ways to integrate both views.

3. Research approach

Our first step was to identify frameworks representing the functional and institutional view on data intelligence and analytics. The first was CRISP DM. This is a prominent framework that resonates the functional view. The second is the knowledge creation process model of Nonaka (2000). This is a well-known example of an institutional model. Next step was to analyse the frameworks and then integrate them into a framework capturing the best of both worlds.

Finally, we put the framework into practice by conducting action research. This action research had two objectives: solving a real-world problem and contributing to science by reflecting on our concepts (Benbasat et al., 1987). We as researchers fulfilled two roles in this research: developing the risk model and reflecting upon the development process. The action research involved daily interaction with 11 employees from DSC and interviews with employees, including managers and operators. Within the case study, our research methods consisted of desk research, workshops with inspectors and data analysts, modelling and interviews.

Following the action research an evaluation took place. The first evaluation of the model has been executed in the form of model validation, using new datasets. The predictive quality of the model was researched by testing whether the model predicts the new datasets accurately. Second, four interviews have been conducted – two with inspectors and two with data analysts. The interviews were recorded and a summary was transcribed. Furthermore, the risk model and framework have been presented to inspectors and policy advisors of the Inspection agency under study and the Dutch Inspection Council.

4. Two prescriptive frameworks on the data science process

In this section we will discuss the background and principles of two frameworks that are exemplary for both views.

4.1. The data science process according to CRISP DM

Plans and programs for tabulating data drew directly from early modern science, adopting the same positivist philosophies used in the natural and physical sciences. See for example Jevons (1913) who, in his monumental work, described the basic steps for collecting evidence in pursuit of a hypothesis.

In the early days database technology largely hinged on developing a grammar which enables the collection, tabulation and reporting of data to occur at a large-scale, and without extensive human intervention (Codd, 1970). Such technologies demanded further guidance for statisticians who, now freed from routine tabulation, needed guidance on the whole modelling life-cycle.

From these needs the field of knowledge discovery in data bases (often known as data mining) was born. Furthermore, administrative process standards were developed to help data mining practitioners make the most sense out of their data. There are several standards like SEMMA and kdSS (A. I. R. L. Azevedo and M. F. Santos, 2008). The most prominent of these frameworks is known as the cross-industry standard data mining process, or CRISP-DM (Chapman et al. 2000). Figure 1 provides a graphical representation of the main steps in the modelling cycle.

A consortium of industry leaders with vested interests in databases (Teradata), statistical software (SPSS), data consulting (IBM), and transaction software (NCR) formulated a research agenda. The other participants (Daimler AG and OHRA) represented data-intensive industries, e.g., the automotive and insurance industry. Funding for the consortium was approved as part of the ESPRIT programme, with the aim of producing an open standard for data mining. Although CRISP-DM is the most commonly endorsed data mining process, it is not the only process available. The framework is nominated by many industrial participants as the most commonly used data mining process (KDnuggets, 2014). Another alternative is SEMMA, which stands for Sample, Explore, Modify, Model and Assess, developed by the SAS Corporation (SAS, 2019). SEMMA is not intended as a general purpose framework for analysis, although it has been adopted as such by some. The CRISP-DM model has even been expanded in steps and methodology by still newer authors. Azevedo and Santos (2008) offer a critical comparison of the

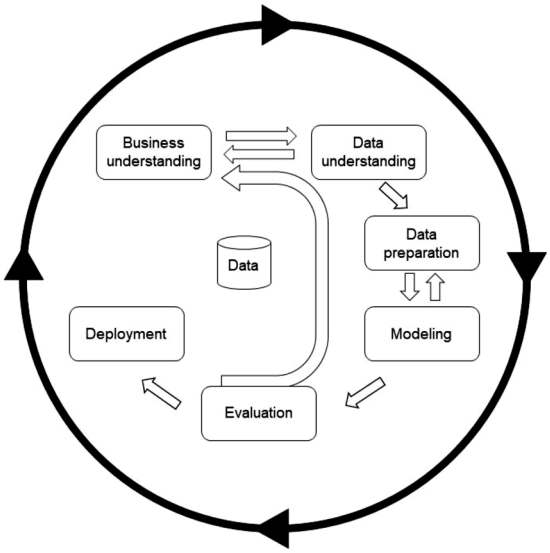

Fig. 1.. The CRISP DM framework (KDnuggets, 2014).

alternatives.

Substantial changes have occurred in business, society and technology in the nearly twenty years since the CRISP-DM framework was launched. In turn this has required proponents of the field to develop new conceptions of knowledge where the ideal data science practitioner mixes computer science, business knowledge and acumen, and statistical knowledge all in one person. Likewise the functional relationship between data mining team members and other professionals inside public and private analysis bureaus is increasingly being reconsidered. These concepts have in large part helped to define data science as a new domain of knowledge and a new kind of professional expertise. See for instance the discussion by Hicks and Irizarry of the needs of this new field of education (Hicks and Irizarry, 2017).

The CRISP-DM framework aims to improve business understanding via distinct data science activities. It clearly shows a chain of activities, with some important interactions between steps. As such, CRISP DM resonates a functional view on decision-making. Its main concern is the ability and capacity to come to useful information (see table 1). That end is assumed to be reached via data science. The framework is oriented towards activities rather than the actors pursuing those activities. It is not oriented towards decision-makers. However, a main, although implicit, assumption is that improved understanding will lead to better decisions.

Despite the success and adoption of the CRISP-DM framework, there has been relatively little critical analysis of the assumptions and the design context which underlies the framework. The sharpest criticism of CRISP-DM may well be that the framework sets up the analyst to work alone, without the necessary tools and cross-disciplinary expertise needed to achieve outcomes in a real-world data or statistical practice (Salz, Shamshurin, & Connors, 2017). What literature has been published has been of two kinds. One strand of literature cites the framework as an integral part of surveying domain knowledge (c.f. Esfandiari et al., 2014). A second strand of literature is comparative in nature; it proposes and tests alternatives in an effort to build new frameworks (c.f. Sharma, Osei-Bryson and Kasper, 2012).

4.2. The data science process as a knowledge creation process

In many instances the users of data are professional workers. As professionals, they have exclusive high-level knowledge about their jobs, exclusive abilities how to assess their work and – as a result - important discretionary freedoms to work autonomously (Abbott, 1988; Freidson, 1999). This seems at odds with the functional view on data science and decision-making. If professional workers as operational decision-makers have these discretionary freedoms, they can hardly be viewed as passive receivers of data. They will have the tendency and position to interpret data into information that fits their ideas. In other words, they will have an active role in the flow from data provision to information use. This idea aligns with the institutional view on data science. Along this line of reasoning we argue it is fruitful to take a look at knowledge creation theory. This theory rose quickly in the nineties within the organizational sciences. Theory on knowledge creation is a response to organizational theories from the fifties to the eighties that took knowledge as a pre-given. Core theorists as Ikujiro Nonaka and Georg von Krogh developed an alternative to this idea that can be broken down into the following claims:

- Knowledge is not a pre-given, but a creation by human beings (Dodgson, 1993; Nonaka, 1988; Weick and Westley, 1996), as made explicit in the previous sections.
- Knowledge involves inherently personal features, such as physical skills, experiences and perception, instead of 'universal justified true belief' (Nonaka, von Krogh and Voelpel, 2006).
- As such, there are multiple knowledge sources within one organization. Knowledge refers to senses, movement skills, physical experiences, intuition and implicit rules of thumb (Nonaka et al., 2006; Polanyi, 1966).
- As such, knowledge is both explicit and tacit (Nonaka, 1991). The latter refers to knowledge that cannot be adequately codified by either verbal means or with written documents (Polanyi, 1958). There are many possible reasons for this difficulty. For instance, tacit knowledge often contains culture-informed values, personal experiences and attitudes that are often implicit (Leonard and Sensiper, 1998). Moreover, it is informed by experience rather than explicit lessons (Lam, 2000; Nonaka, 2000; Schmidt and Hunter, 1993).
- Because knowledge is partly personal and tacit, transmission of knowledge – or knowledge conversion - between persons or institutional borders is fragile (Nonaka and Takeuchi, 1995; von Krogh, Roos and Slocum, 1994). Knowledge has to be accepted by the next actor in the chain. At the same time the transmission process is hampered by a difficulty to make knowledge explicit.
- There is an ongoing discussion about whether tacit knowledge can be made explicit, or that it is merely time-consuming (Leonard and Sensiper, 1998; Willcocks and Whitley, 2009). There is more consensus about the idea that socialisation and interaction when trying to transfer tacit knowledge would be crucial (Lam, 2000; Nonaka and van Krogh, 2009).
- Because of the complexity of this transmission process, individual knowledge often fails to benefit others in the organization and vice versa (von Krogh, 2002; von Krogh and Grand, 1999). A focus on knowledge transactions of multiple knowledge sources, acknowledge the importance of learning and knowledge conversion processes (Nonaka and von Krogh, 2009).

Nonaka (2000) developed a model that fleshes out the transmission of knowledge between actors, as this transition is found as the key to knowledge creation. He emphasises the importance of enlarging individual knowledge, finding common concepts for shared knowledge, crystallize this shared knowledge into concrete products or systems, evaluation of the product and knowledge used ('justification') and spreading the concepts through the organization ('networking'). The knowledge creation framework is depicted in figure 2.

Also Nonaka's framework emphasizes process steps and some iterations, just like CRISP DM. However, its main concern is not improving understanding via a fixed method, but by actors sharing their knowledge. Intelligence is assumed to be improved by a social process rather than a functional one. The framework then fleshes out different aspects of this social process. In this way it pays attention to the multi-actor context of a data science process, wherein per activity the knowledge will be transmitted to different persons with different mindsets. These transmission processes are its main concern, for it is assumed both fragile and critical. For these reasons we argue that the application of this framework to data science processes fits an institutional view (see table 1).

We stated that both views have their inherent logic and their inherent weaknesses. How to incorporate them into a data science process framework? In the next section we will propose such a framework.

5. Combining the frameworks for creating synergy between data scientists and domain professionals

The CRISP-DM framework is especially suited for the creation of data-mining products, such as risk models. However, this framework is made for the use of a solitary data analysts and does not incorporate the use of domain knowledge. In contrast, Nonaka's (2000) knowledge creation framework has been made to create organisational knowledge through the sharing of domain knowledge between individuals. However, Nonaka's model does not specify how a data-mining risk model can be constructed based on the organisational knowledge created.

Therefore, we hypothesize that the CRISP-DM and Nonaka's

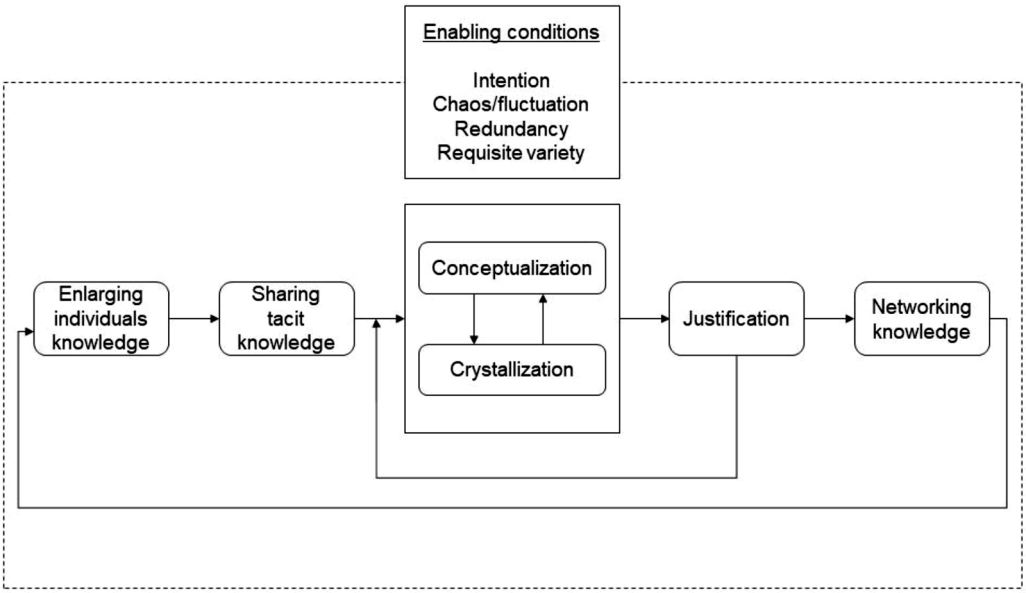

Fig. 2.. A process of generating information/knowledge in the market (Nonaka, 2000).

organisational knowledge framework complement each other. The weakness of the CRISP-DM model is the strength of Nonaka's model, namely harnessing individual tacit knowledge to create organisational knowledge. Ditto, Nonaka's model's weakness is the strength of CRISP-DM, namely the focus on the construction of data-mining models.

The new framework is shown in figure 3. The framework incorporates all the activities from both models. The activities from the knowledge creation model focus on extracting knowledge from the organisation (depicted in orange), while the CRISP-DM activities focus on transforming this knowledge into a data mining model (depicted in blue). Activities overlapping in the two models were merged, to enhance the simplicity of the model. These activities are depicted in yellow. The proposed framework contains nine activities, which are explained in detail in the next section. Of course, the activities will have iterations.

1 *Enlargement of individual knowledge*

The first activity is derived from Nonaka's organisational knowledge framework. This activity details the enlargement of individual knowledge by interaction. We assumed that different actors have different sources of knowledge and that all these sources may be valuable in the data science process. Therefore it is key to focus on learning by individual actors, both about their own – often tacit – knowledge and the knowledge of others. This activity cannot be defined as a 'step' per se. This is a continuous process that is going on prior to the other activities but also continuous after the other activities. As such, this activity can better be interpreted as a motive.

1 *Sharing tacit knowledge*

The second activity is also derived from Nonaka's organisational knowledge framework. In the first activity, individual employees have collected organisational knowledge. During the second activity, they should share their knowledge to increase the overall organisational knowledge. Different individuals with the same organisational function should share their knowledge, but also different kinds of experts can share their knowledge. This might increase general knowledge and understanding of the perspectives of different sources of knowledge within

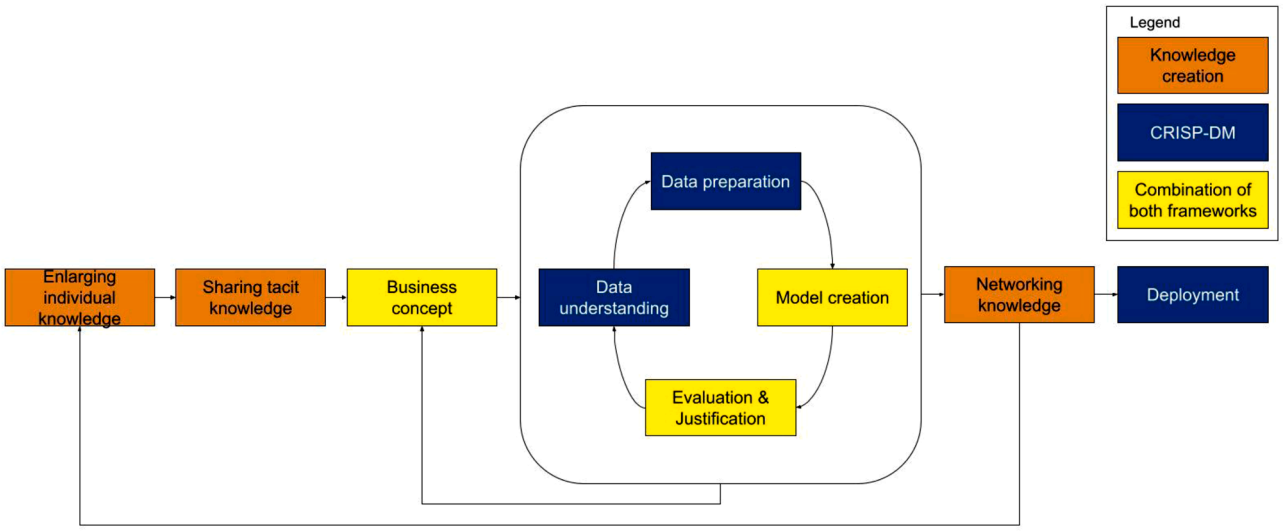

Fig. 3.. Knowledge creation model for data intellingence and analytics.

the organisation. As acknowledged in section 4 the ability of professionals to share their tacit knowledge is contested. However, within these limitations, we can assume that socialisation and interaction will allow for valuable learning processes.

Basically, the first two activities can be seen as specifications of the 'business understanding' activity from CRISP DM. However, from an institutional view these specifications are important. If we assume that domain specialists and data analysts have both unique knowledge and a unique power position, then creating a level playing field – in terms of knowledge – is a condition for business understanding. The first two steps of Nonaka's framework can be seen as efforts to create this level playing field.

1 *Business concept*

In the third activity CRISP-DM and Nonaka's framework come together. Knowledge has been enlarged and shared, The conditions have been met to develop a business concept. In line with the institutional view on data science processes this is a joint effort. The process of creating a business concept has two purposes. A first purpose is a further enlargement of business understanding, conform the first activity of CRISP DM. It is assumed here that a joint development of a business concept will add to business understanding. Secondly, the concept will be a blueprint for the model which will be created during the next activities. This obviously is the most politically sensitive step, since the business concept shows what data the stakeholders think should be added to the model, so that it represents the organisational environment correctly. Opinions and stakes about whether it does so may conflict.

1 *Data understanding*

The data understanding is a first activity of a cycle, inspired by the CRISP-DM model (see figure 3). During this activity, the business concept will be used to retrieve data which the knowledge sources think is of value to the model. Before the data can be used to create a model, the researcher should be able to understand the data. This requires the researcher to conduct some study into the source of the data: where did the data come from and how was the data collected? This ensures that the researcher can judge the reliability and value of the data.

1 *Data preparation*

The next CRISP-DM activity is data preparation. In the previous activity, the researcher has looked closely at the data, so that we know what data is available and what that data means. Now, the data can be prepared, so it can be easily used for the model. When different data sources are used, these data sources are linked together during the data preparation. Data quality issues are also looked at. For instance, values might be missing. A strategy should be devised to deal with these kinds of issues. Another important aspect is that during this activity, validation and test data is set aside, so that the model can be evaluated in later activities.

1 *Model creation*

After the data has been cleaned, the data can be transformed into a model. This activity is a combination of the modelling activity from CRISP-DM and the crystallization activity from Nonaka's framework. The framework can be used for different kind of data mining models, so this activity might change a lot per case.

1 *Evaluation & Justification*

Both the CRISP-DM and the organisational knowledge framework have a validation activity. However, they are fundamentally different. The CRISP-DM framework suggests an evaluation of the model by using test and validation data. The organisational knowledge framework suggests a justification of the model by conducting interviews with experts. This new framework suggests that a combination of both should be used, hence the activity 'Evaluation and Justification.' First, the model should be evaluated on the basis of the data. Then, the model should with justified with all the sources of knowledge. After the model has been evaluated and justified, it should be concluded whether the model is good enough to be taken into use, or if the model needs to be improved. When the model needs more improvement, the process should be restarted at activity 3: Business concept. Otherwise, the process can be continued with the next activity.

We propose this for two reasons. First, new insights about the business concept can be gained during the cycle. An updated business concept can feed the cycle again. Second, the same holds for new stakes. As noted, the development of the business concept has a large potential for conflict, especially in a multi-actor setting assumed from the institutional view. It should be avoided that actors involved – in this case domain specialists and data analysts – feel forced to build on former decisions that are no longer valid to them (i.e. feel caught in the cycle). The opportunity to reflect on the business concept will avoid this.

1 *Networking knowledge*

This is the last activity from Nonaka's framework. In this new framework, it is only the second to last activity. In the past, individual knowledge has been created and shared. During this process, more new knowledge has been created. Knowledge about the model mostly, but also knowledge about how to implement this model into the organisation. This knowledge should be broadcasted throughout the organisation, so that all relevant stakeholders know of the existence of the model. This should make them more inclined to use it, or know whom to contact to ask questions about the model.

1 *Deployment*

Nonaka's framework focuses on the creation of knowledge. The last activity of Nonaka's framework is the networking of knowledge. The CRISP-DM is a bit sharper on the way knowledge got implemented. We chose to add the deployment activity from the CRISP-DM framework. During this activity, a strategy is developed on how the model can be implemented in the organisation, so that is effectively used.

6. Putting the framework into practice

To explore the added value of the suggested framework we took the framework to daily practice. Action research has been performed for the construction of a small-scale risk model at the Dutch Food and Safety Products Authority (the NVWA), which is with about 2600 staff members the largest inspection agency of The Netherlands. The mission of the NVWA is to safeguard the safety of food and consumer products, the health of animals and plants, the well-being of animals and the enforcement of nature laws.

Like many other inspection agencies, the NVWA seeks to work risk-based and uses data analytics to define the main compliance risks. Data analysts entered the organisation of the NVWA. They are there to serve inspectors detecting non-compliance. These inspectors, however, are typically domain experts with their own source of knowledge. Within the NVWA we took an action research approach. This included daily interaction within an NVWA department dedicated to find novel ways to extract value from big data, the Data Science Cluster (DSC). The case has been executed for one food safety department of the regulation agency: the consumer craft products department (the HAP). This department is responsible for safeguarding the food safety for food products sold in bakeries, fish shops, butchers and other businesses selling edible craft products. The aim of the NVWA was to construct a risk model for HAP, that categorizes businesses in the sector based on risk factors. Then, the

model can predict which businesses have the highest risk of misconduct and – according to the principles of risk-based regulation – should be inspected more frequently. The NVWA wished the model to be based on two sources of knowledge: the inspectors and data analysts. The inspectors are asked to contribute to the risk factors that seem most important in the field. The data analysts were to construct a risk model based on these risk factors. They incorporated external data into the model to validate the risk factors suggested by the inspectors and to find new risk factors. At the end, both the inspectors and data analysts should be satisfied with the risk model.

This request from NVWA has given us the opportunity to put our framework into practice and reflect on its effects. We did so in the stepwise approach as in our framework, as described later in this section. While the NVWA was interested in the framework, our concern was basically the process towards that framework. This was to be an interactive effort between data analysts, inspectors and researchers. While working on the risk model we reflected on two issues:

From a functional view we were concerned about the way the process would lead to 'business understanding', which in this case would be an understanding of the main risks. From an institutional view we were concerned about the way the process would lead to an alignment of understanding and goals among data analysts and inspectors. What factors are critical here?

1. Enlargement of individual knowledge

The first step in the framework (figure 3) is 'enlargement of individual knowledge'. At NVWA this was already put in practice. Inspectors and data analysts were already interacting, however not in a structured way. We considered this step an ongoing activity that takes place during all days that the respondents work at the NVWA. The knowledge of each individual that works at the NVWA is enlarged during their work. It consists of their work experience and their interactions with their colleagues and other stakeholders.

2. Sharing tacit knowledge: workshops with inspectors and data analysts

This is the second step in figure 3. The framework aims to harness the knowledge of both sources of knowledge: the data analysts and inspectors. They have gained relevant knowledge through their work and study experiences. Two workshops were organised with 9 participants, one with inspectors and one with data analysts. Before the workshops, the respondents were told that the workshops would be about using data science to select which locations to inspect. They did not receive further information, because we wanted to translate their tacit knowledge from past experiences into explicit knowledge without them thinking about the subject in advance. The workshops served two functions. The first is the facilitation of knowledge sharing between the inspectors and data analysts, to increase their understanding of each other's perspective. The second function of the workshop was the transformation into a business concept. This concept serves as a guideline to what risk factors and which data sources should be incorporated into the model. Figure 4 shows the breakdown of workshop activities, the participation of the different knowledge sources and the involvement of the researcher.

As a disclaimer, we had to organize the workshops with data analysts and inspectors separately. Joint workshops would be more akin to the knowledge creation philosophy. However, transaction costs were found to be too high to merge the groups. To ensure that knowledge was shared between the two groups, the input of the inspector workshop was used as the basis for the data analyst workshop. The knowledge from the data analyst was only given back to the inspectors by an email containing only detailed results from the data analyst workshop. This knowledge feedback loop could be improved when using this model in the future by either conducting combined workshops or by having an additional workshop with the inspectors, which uses the input from the data analysts. This approach enhanced the role of us as analysts in interpreting the results of the workshops. Whether this is a price worth to pay may be subject to further research and testing.

3. Business concept: workshops with inspectors and data analysts, in iteration with DSC

This step adheres to the second goal of the workshops. During the workshops, risk factors are identified and sorted on priority. Based on these risk factors, relevant data sources are identified and categorized on feasibility and relevance. The knowledge of inspectors and data analysts has been made explicit in a business concept through the workshops. This business concept contains a list of risk factors, sorted on priority. For each risk factor, data sources have been identified which are both feasible and relevant. Based on this business concept, data could be selected to be used as input for the risk model. Before this data has been used, the data has been looked at in more detail to understand what the

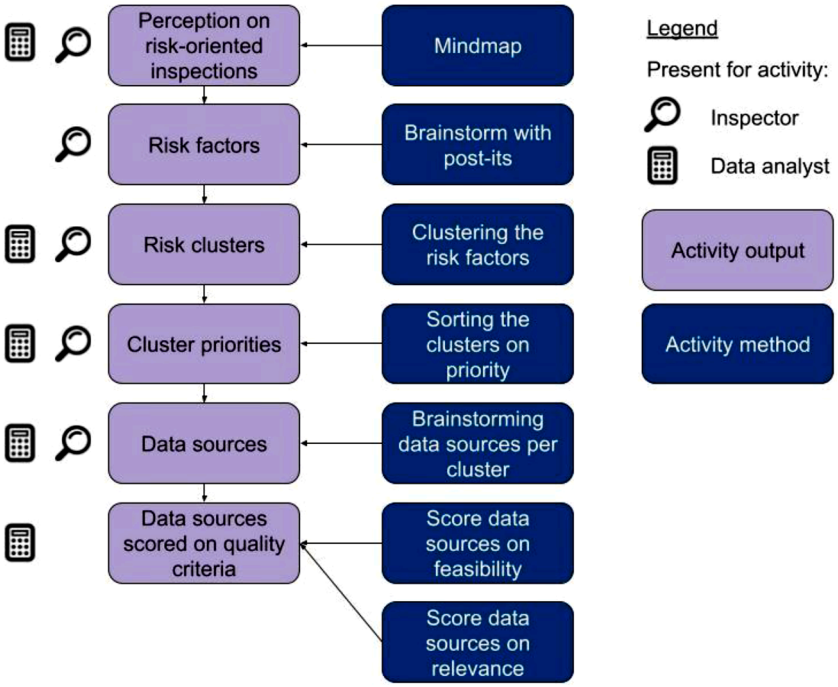

Fig. 4.. Workshop activities.

data means and what it can be used for. This has been done in iteration with DSC.

4. Crystallize knowledge into a risk model: modelling in iteration with DSC

This activity includes the cycle from data preparation to evaluation as depicted in figure 3. Data has been prepared for the risk model. This means that the relevant data has been selected and that the quality of the data has been checked and improved. Now the risk model can be constructed. As a model the decision-tree has been chosen. The tree starts with a conditional node at the top, which contains all cases. Then, the decision tree used conditional rules to split the cases into separate branches, based on the amount of risk each branch contains. The decision tree has calculated which variable is the best to split the cases on, based on the risk associated with the cases. Each of these branches can be followed by another branch. In this way the tree created a hierarchy of risks.

The choice of the model itself has been subject to interaction. The inspectors preferred a decision-tree model because of its transparency, compared to, for instance, random forest models. Both steps 3 and 4 were conducted by the researchers, however, in close interaction with DSC.

5. Networking knowledge

By using a transparent risk model the cycle from data preparation to evaluation can be done interactively with all actors involved – including data analysts and inspectors. Of course, this can be done with just a limited amount of persons. It will be a transparent process for the few. The challenge here is how to entice the larger part of the organization to understand the reasoning behind the model. This means that 'evaluation' and 'justification' will be long-lasting processes. These processes will have two functions: improving the model and spreading the word ("networking knowledge").

A first evaluation of the model has been executed in the form of model validation, using new datasets. The predictive quality of the model can be researched by testing whether the model predicts the new datasets accurately. Subsequently, the model is justified by the sources of knowledge. Four interviews have been conducted. Two with inspectors and two with data analysts. In these interviews it has been evaluated whether the prototype risk model matches their expectations. The interviewees have been asked about their experience with the workshop and whether they agreed with the implementation of the risk factors in the model, which had been based on the model input. They were also asked if they had additional ideas for model improvements or for improvements on the research approach. Furthermore, the risk model and framework have been presented to inspectors and policy advisors of the NVWA and the Dutch Inspection Council and will be further developed by the NVWA.

7. Findings and dilemmas

The main aims of our action research approach was to solve the problem of the NVWA and reflect on the process on the run. Both DSC and interviewees were satisfied with both the interaction process informed by the framework and the resulting risk model.

In this section we share some more observations from two viewpoints. From a functional viewpoint, how is business understanding served? And from an institutional viewpoint, how are understanding and goals of data analysts and inspectors aligned?

Differences between inspectors and data analysts

It was already assumed that inspectors and data analysts represent distinct knowledge sources, in this case about risks. During our workshops and interviews the differences showed in a surprising fashion. First, they had different views on what 'working risk-oriented' means. The data analysts viewed working risk-oriented almost solely as working datadriven. It appeared that inspectors viewed working risk-oriented in a broader perspective, including soft coordination and individual mental models. Second, the data analysts mostly focused on the inner working of the risk model, while the inspectors focused more on the practical use of the model and its implications for the organisation.

This observation supports the need for a framework that is aimed at a joint view. If such a joint view would not be an explicit goal of a framework, demand for intelligence and supply of data would hardly align.

The workshops led to a mutual understanding of perspectives

During the inspector workshop, it was explained to the inspectors what a risk model would look like. A first look into some methods, including the decision-tree method, was given to the inspectors. This resulted in a better understanding of the risk model methodology for the inspectors. This made them more inclined to collaborate with the data analysts, because they better understood what the data analysts wished to know of them. Furthermore, it made them accept the risk model in later stages, because the model met their expectations. Subsequently, during the data analyst workshop, the results of the inspector workshop was shown. This gave the data analysts a better understanding of the perspective of the inspectors on working risk-oriented. Also, the data analysts gained more insight in which risk factors the inspectors found relevant as input for the risk model. Thus, the workshops resulted in a common understanding of each other's perspectives.

This serves both business intelligence and alignment.

Through the workshops tacit knowledge has been shared

The second goal of the workshops was to make the tacit knowledge of the inspectors and data analysts explicit in a business concept. The inspector workshop started with a broad brainstorm on possible risk factors. At first, the inspectors came up with the most obvious risk factors. It can be hypothesized that these factors are a result of explicit knowledge, as they are commonly known within the organisation. However, after a few minutes, the inspectors began to connect their ideas with ideas written down by other inspectors. This led to new insights and less obvious risk factors began to manifest. It can be hypothesized that these risk factors are derived from tacit knowledge, because the inspectors did not previously realize that they possessed knowledge on those risk factors. Thus, during the brainstorm phase in the workshop, both tacit and explicit knowledge of inspectors became explicit. The same happened during the data analysts workshop. They did a broad brainstorm on possible data sources that could be used to add the risk factors to the model. At first, the data analysts came up with the more obvious data sources. After a while, they formed connections with other data sources and they came up with more 'out-of-the-box' ideas. Thus, during both the data analysts and inspector workshop, tacit knowledge became more explicit and has been shared.

Again, this serves both business intelligence and alignment.

The case for compromising on model sophistication

During the workshops, respondents had to write their ideas on memo's. Subsequently, they had to cluster and prioritise these memo's. This way, we created an overview of possible risk factors (clusters of factors they used to determine where to conduct inspections) which were also sorted on priority. This overview (the workshop output) was used to determine which risk factors should be added to the model and in which order. During this construction, we encountered a dilemma. Inspectors proved not to think in models. At least the models they used were implicit mental models. For this reason they had difficulties understanding the more advanced risk models proposed by data analysts. This limited understanding potentially hampers the exchange of knowledge between data analysts and inspectors in making a good risk model. The dilemma is whether or not to compromise the sophistication of the model in order to facilitate communication with the inspectors.

We have chosen to do so. For this research, the decision-tree methodology has been chosen for its transparency. A simpler model, such as a logistics regression, has also been considered. However, this model was not expected to be able to capture the complexity of the relations between different risk factors in the inspection domain. A more complex model, such as a Bayesian neural network, has also been considered. This model would be expected to have a higher model performance.

However, this model was expected to be too complex to explain to the inspectors. This would limit the acceptance and practical usefulness of the risk model. Therefore, a decision-tee model is a trade-off between model performance and complexity for this case. Another advantage of the decision tree risk model is its possibilities to develop interactively. There are multiple methods to construct a decision-tree risk model, with different degrees of interaction between the sources of knowledge. Even developing a completely interactive decision tree is an option. Interactions, while developing a model, may serve evaluation and justification purposes.

This observation shows a tension between the functional and institutional view. In fact, the potential for business understanding has been compromised here, at least from the perspective of data analysts. More sophisticated – and probably more adequate – models have been neglected for the increase of alignment.

The importance and dilemma of model ownership

The interviews revealed two important potential obstacles for the implementation of the risk model. First, the way the model has been operationalized into software is critical for its use. It is the inspectors that are going to use the risk model. The inspectors were at least hesitant to use complex software to use the risk model. A dashboard can be used to reduce complexity and provide the necessary discretionary freedoms to the inspector. For instance, a dashboard may provide the inspectors the opportunity to choose how many visits they will do risk-based and how many randomly.

Second, the ownership of the model was found to be of key importance. Who will be in charge to further optimize the model? The inspectors are in contact with the inspection environment. However, they do not have the knowledge and skills to actually change the risk model. The data analysts do, but they lack the real-world input inspectors have. A shared responsibility inevitably involves coordination problems. There are no easy answers at the moment.

For alignment, model ownership is fundamental. Who is developing the model? And on whose terms? Yet defining model ownership is hard to pinpoint in a framework. For alignment, it is a prerequisite for starting the entire operation. For business understanding, model ownership is less of a concern. But then again, this is only if business understanding is not seen as a shared concern.

8. Conclusion and discussion

Current data mining approaches provide hardly any support for incorporating the knowledge of domain experts in the approach to process data explicitly. Often these stepwise approaches assume that knowledge creation and knowledge transfer happens one after the other. An important risk of taking only a functional view is that professionals consider data science approaches as a threat to their profession and will not fully cooperate with data science initiatives. Another risk is the loss of important real-world knowledge when analyzing the data. Both risks consider the human (and subjective) influence of each contributor to the data science process.

We contrasted this 'functional view' with an 'institutional view' on data science and decision-making, that is more oriented to interaction and synergy between different and sometimes competing knowledge sources, of which data science is one and domain knowledge by professionals another. How to amplify synergies between data analysts and domain professionals?

We proposed a framework of the process from data science to data use that is based on a knowledge creation process. Our approach acknowledges that knowledge is being created in a multi-actor context, wherein per activity, the knowledge will be transmitted from person to person with different mindsets and that this transmission process is both fragile and critical. The framework developed was based on CRISP-DM – a well-known framework for the data science process – and an authoritative knowledge creation model by Nonaka (1990).

To acknowledge the multi-actor setting of data science for risk-based inspection, we added three activities to the CRISP DM model. Two prior activities are focused on respectively enlarging knowledge and sharing knowledge between professionals and data analysts. Another extra feature is a focus on valorisation after the knowledge creation phase.

The framework was used in practice by developing a risk model for the Dutch Food and Safety Authority (NVWA). Three key dilemmas were found.

First, the aim to share knowledge between different knowledge sources – data analysts and inspectors in this case – may conflict with the aim to optimize data-science-based risk models. The more sophisticated these models, the harder they serve as an interface between data analysts and inspectors, because inspectors hardly understand the more sophisticated models. In our case a simpler risk model was selected based on decision- tree modelling. This model is easy to understand for inspectors. Moreover, the transparency of this type of model eases evaluation and justification resulting in better accountability.

Second, interactions enhance the sharing of both explicit and tacit knowledge. However, the question remains how to organize this interaction. The workshops served well for the creation of common knowledge. Moreover, they helped in making the tacit knowledge of contributors more explicit, as often they were surprised about their own contributions. An issue that remains open for further development is how the practice of knowledge sharing can be embedded in the organizational structure and whether this is desirable. Obviously, sharing knowledge is a continuous activity. However, it might also be considered as overhead by the parties involved. Both data analysts and inspectors have busy jobs and will not reserve time for activities that are not explicitly part of their jobs, such as knowledge sharing with other departments. For pragmatic reasons, incidental meetings might hold promises to commit them to learn about each other's activities.

Third, ensuring ownership of the model is key, because it is the owner of the model – and the ideas behind it - that has to commit data analysts and professionals to dedicate their time to share knowledge. Main dilemma here is whether the model should be owned by either data analysts or professionals, or on a higher authority.

The findings and issues suggest that there is much potential in the framework if applied to data science for professional users, such as for risk-based inspections. Still, more research is needed to evaluate its effectiveness. This research can evaluate the impact of different forms of organizing knowledge sharing between data analysts and professionals. Another advised research question is about how framework ownership affects the continuity of knowledge creation. The framework aims to bridge the knowledge sources of data analysts and domain professionals, but who maintains the bridge? Shared ownership seems to hold most promises, however how can transaction costs of shared ownership be minimized? To establish this, we need more empirical input. With case study research we may question how business concepts are established in different framework ownership models. How would they align with the separate business cases of data analysts and domain professionals? These questions are especially important in this era, where privacy, fairness, and legality are found crucial. Furthermore, the trade-off between sophistication and transparency of predictive decision models is worth a study. For instance: how is this trade-off made within different organizations, for different purposes?

However, these questions are only part of a broader issue of legitimacy. Data science is becoming more contested because of issues of accountability, fairness, and legitimacy. On a more positive note, the contestation of data science can be seen as a maturation of the field. Obviously, any practice or method has its strengths and weaknesses, and for more mature practices, these strengths and weaknesses are better known. If we consider this, combining the strengths of data science with the wisdoms of the work floor to gain knowledge is full of promises.

References

Abbott, A., 2014. The system of professions: An essay on the division of expert labor. University of Chicago Press, Chicago.

Arnaboldi, M., 2018. The missing variable in big data for social sciences: The decision-maker. Sustainability 10, 1–18.

Azevedo, A., Santos, M.F., 2008. KDD, SEMMA and CRISP-DM: a parallel overview. Paper presented at the IADIS European Conference on Data Mining.

I. Benbasat, D.K. Goldstein, M. Mead, The case research strategy in studies of information systemsm MIS quarterly (1987) 369–386.

Bhimani, A., Willcocks, L., 2014. Digitisation, "Big Data" and the transformation of accounting information. Account. Busin. Res. 44 (4), 469–490.

in Boisot, M., Canals, A., 2008. Data information and knowledge: have we got it right? In: Boisot, M., MacMillan, I., Han, K.S. (Eds.), Explorations in Information Space. Oxford University Press, Oxford.

boyd, D., Crawford, K., 2012. Critical questions for big data: Proviactions for a cultural, technological, and scholarly phenomenon. Inform., Commun. & Soc. 15 (5), 662–679.

Chapman, P., Clinton, J., R, Kerber, Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc 16.

Codd, E.F., 1970. A relational model of data for large shared data banks. Commun. ACM 13, 377–387.

Crawford, K., Miltner, K., Gray, M., 2014. Critiquing big data: Politics, ethics, epistomology, Intern. J. Commun. 8, 1663–1672.

Daston, L., Galison, P., 1992. P. The image of objectivity. Representations 40, 81–128.

Dodgson, M., 1993. Organizational learning: A review of some literatures. Organiz. Stud. 14 (3), 375–394.

Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E., Tabar, V.K., 2014. Knowledge discovery in medicine: Current issue and future trend. Expert Syst. Appl. 41 (9), 4434–4463.

Feldman, M.S., March, J.G., 1981. Information in organizations as signal and symbol. Adm. Sci. Q. 26 (2), 171. https://doi.org/10.2307/2392467.

Freidson, E, 1999. Theory of professionalism: Method and substance. Intern. Rev. Soc. 9 (1), 117–129. https://doi.org/10.1080/03906701.1999.9971301%0D.

S.C. Hicks, R.A. Irizarry, A Guide to Teaching Data Science (2017) Retrieved from https://arxiv.org/abs/1612.07140, last accessed 9 August 2021.

Hilbert, M., 2016. Big data for development: A review of promises and challenges. Dev. Policy Rev 34 (1), 135–174. https://doi.org/10.1111/dpr.12142.

Höchtl, J., Parycek, P., Schollhammer, R., 2016. Big data in the policy cycle: Policy decision-making in the digital era. J.Org. Comput. 26 (1–2), 147.

Huff, D., 1993. How to lie with statistics. Norton, New York.

Janssen, M., Kuk, G., 2016. Big, open and linked Data (BOLD) in research, policy and practice. J. Organ. Comput. Electron. Commer. 26 (1–2), 3–13. https://doi.org/10.1080/10919392.2015.1124005.

Janssen, M., van der Voort, H., Wahyudi, A., 2017. Factors influencing big data decision-making quality. J. Bus. Res. 70 https://doi.org/10.1016/j.jbusres.2016.08.007 m.

Jevons, W.S., 1913. The Principles of Science: A Treatise on logic and scientific method. Macmillan, London.

Kogan, M., 1999. The impact of research on policy (Ed.). In: Coffield, F. (Ed.), Research and policy in lifelong learning. Policy Press, Bristol, pp. 11–18.

Lam, A., 2000. Tacit knowledge, organizational learning and societal institutions: An Integrated Framework. Org. Stud. 21 (3), 487–513.

Leonard, D., Sensiper, S., 1998. The role of tacit knowledge in group innovation. Calif. Man. Rev. 40 (3), 112–132.

Maciejewski, M., 2016. To do more, better, faster and more cheaply: Using big data in public administration. Int. Rev. Admin. Sci. 83 (1), 120–135.

Madsen, A.K., 2018. Data in the smart city: How incongruent frames challenge the transition from ideal to practice. Big Data Soc. 5 (2). Article 2053951718802321.

Nonaka, I., 1988. Toward middle-up-down management: Accelerating information creation. MIT Sloan Man. Rev. 29 (3), 9–18.

Nonaka, I., 1991. The knowledge-creating company. Harvard Bus. Rev. 96–104. November-December.

in: Nonaka, I., 2000. A dynamic theory on organizational knowledge creation (Ed.). In: Smith, D.E. (Ed.), Knowledge, groupware and the internet. Routledge, London, pp. 3–42.

Nonaka, I., Takeuchi, H., 1995. The knowledge-creating company. Oxford University Press, New York.

Nonaka, I., von Krogh, G., 2009. Tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. Org. Sci. 20 (3), 635–652.

Nonaka, I., von Krogh, G., Voelpel, S., 2006. Organizational knowledge creation theory: Evolutionary paths and future advances. Org. Stud. 27 (8), 1179–1208.

OMalley, M., 2014. Doing what works: Governing in the age of big data. Publ. Adm. Rev. 74 (5), 555.

Polanyi, M., 1966. The tacit dimension. Routledge & Kegan Paul, London.

Provost, F., Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision-making. Big Data 1 (1), 51–59.

Redden, J., 2018. Democratic governance in an age of datafication: Lessons from mapping government discourses and practices. Big Data Soc. 5 (2) article 2053951718809145.

Reddy, E., Cakici, B., Ballestero, A., 2019. Beyond mystery: Putting algorithmic accountability in context. Big Data Soc. 6 (1) article 2053951719826856.

Salz, J., Shamshurin, I., Connors, C., 2017. Predicting data sciencec sociotechnical execution challenges by categorizing data science projects. J. Assoc. Inform. Sci. Techn. 68 (12), 2720–2728.

SAS, An Overview of SAS Analytics Solutions, (2019), https://www.sas.com/en_gb/software/analytics-overview.html, last accessed 9 August 2021.

Shafique, U., Qaiser, H., 2014. H., A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). Int. J. Innov. Sci. Res. 12 (1), 217–222.

Schmidt, F.L., Hunter, J.E., 1993. Tacit knowledge, practical intelligence, general mental ability, and job knowledge. Curr. Direct. Psycholog. Sci. 2 (1), 8–9.

Sharma, S., Osei-Bryson, K.M., Kasper, G.M., 2012. Evaluation of an integrated knowledge discovery and data mining process model. Expert Syst. Applic. 39 (13), 11335–11348.

Sivarajah, U., Kamal, M., Irani, Z., Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. J. Bus. Res. 70, 263–286.

Taylor, L., Purtova, N., 2019. What is responsible and sustainable data science? Big Data Soc. 6 (2) article 2053951719858114.

van der Voort, H., Klievink, B., Arnaboldi, M., Meijer, A., 2018. Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision-making? Gov. Inform. Q. 36 (1), 27–38. https://doi.org/10.1016/j.giq.2018.10.011.

von Krogh, G., 2002. The communal resource and information systems. J. Strat. Inform. Syst. 11, 85–107.

in: von Krogh, G., Grand, S., 1999. Justification in knowledge creation: Dominant logic in management discourses (Eds.). In: von Krogh, G., Nonaka, I., Nishigusi, T. (Eds.), Knowledge creation: A source of value. Macmillan, London, pp. 13–35.

von Krogh, G., Roos, J., Slocum, K., 1994. An essay on corporate epistemology. Strat. Man. J. 15, 53–71.

Vydra, S., Klievink, B., 2019. Techno-optimism and policy-pessimism in the public sector big data debate. Gov. Inform. Q. 36 (4) https://doi.org/10.1016/j.giq.2019.05.010.

In: Weick, K.E., Westley, F., 1996. Organizational learning: Affirming an oxymoron (Eds.). In: Clegg, S., Hardy, C., Nord, W. (Eds.), Handbook of Organization Studies. SAGE, London, pp. 440–458.

Willcocks, L., Whitley, E., 2009. Devoloping the information and knowledge agenda in information systems: insights from philosophy. Inform. Soc. 25, 1–8.

Dr. Haiko van der Voort is an assistant professor Organisation and Governance at TU Delft, Faculty of Technology, Policy and Management. As a public administration scholar, he studies the (big) data use process within public organizations.

Sabine van Bulderen, Msc is a business analyst for the Dutch Administration of Justice. Her main focus is the digitalisation of the judicial process and the impact of this for the Dutch society. She used to work as a data science intern at the Dutch food and products inspectorate (NVWA).

Prof.dr. Scott Cunningham is full professor Urban Policy at the University of Strathclyde, School of Government & Public Policy at the the Faculty of Humanities & Social Sciences. He researches and teaches data science for policy analysis, mostly in an urban context.

Prof.dr.ir. Marijn Janssen is a full professor in ICT & Governance. His research is focused on the governance of ICT-architecting. He is also honorary visiting professor at Bradford University and visiting professor at Universiti Teknologi Mara. He has published over 500 refereed publications and is among the most influential digital government researchers (https://apolitical.co/lists/digital-government-world100/).