Journal of Management and Governance
https://doi.org/10.1007/s10997-022-09659-0

A fine-grained perspective on big data knowledge creation: dimensions, insights, and mechanism from a pilot study

Maryia Zaitsava[1] · Elona Marku[1] · Maria Chiara Di Guardo[1] · Azar Shahgholian[2]

Accepted: 13 October 2022

Abstract

The creation of knowledge from Big Data is increasingly drawing the attention of scholars and practitioners in management research. Valuable knowledge first requires identifying the Big Data features connected to knowledge insights creation and the mechanism beyond this creation. This paper examines Big Data dimensions and insights creations at a fine-grained level by adopting the knowledge creation lens. Specifically, what is the mechanism of creating knowledge from Big Data? How to transform raw Big Data into knowledge? We adopted a qualitative case study to explore the large-scale multinational pilot launched in three European cities. The pilot amalgamated a large amount of data feeds from different sensors and open data and created various insights to inform cities' strategies. By employing an inductive content analysis with abductive procedures and coupling it with participatory observations, we were able to ground findings on the multi-level empirical and theoretical base and build a framework that embraces all discovered complexities and fine-grained features of Big Data dimensions and guides knowledge creation from Big Data. Our research offers a more in-depth understanding of the mechanism of knowledge creation in the BD context. First, we opened up BD's black box by disentangling the knowledge creation mechanism while transforming raw BD into BD insights. Second, our study offered empirical evidence of the growth mechanism working on *Volume* and *Variety* dimensions. The uniqueness of this study lies in the fine-grained perspective of BD characteristics and the underlying mechanism of insights creation.

Keywords Big Data · Knowledge creation · Big Data insights · Big Data dimensions · Case study

✉ Maryia Zaitsava
maryia.zaitsava@unica.it

Extended author information available on the last page of the article

Published online: 14 November 2022

1 Introduction

The creation of new knowledge from enormous volumes of structured and unstructured data, namely Big Data (BD), has increasingly drawn the attention of academics and practitioners in management research (Ciampi et al., 2020; de Camargo Fiorini et al., 2018; Di Vaio et al., 2021; Lee, 2017; Sumbal et al., 2017). While scholars have pointed out that knowledge creation through data-driven insights is a critical source of competitive advantage not only in innovative and hi-tech business environments but also in traditional sectors (Akter et al., 2019; Chen et al., 2015; Ghasemaghaei, 2020; Hopkins, 2010; LaValle et al., 2011; Vidgen et al., 2017), many companies are struggling to manage this knowledge generation (Chen et al., 2015; Gandomi & Haider, 2015; Henkel & Hartmann, 2020; Johnson et al., 2017), and research on knowledge creation through BD insights is still at its infancy (Cappa et al., 2020; Ghasemaghaei, 2021; Ghasemaghaei & Calic, 2019).

Indeed, some recent investigations have shown that the key attributes of BD, *Volume*, *Velocity*, *Variety* (Cappa et al., 2020; Gani et al., 2016; George et al., 2014; Hashem et al., 2016; Johnson et al., 2017), and *Veracity* (Simsek et al., 2019; Urbinati et al., 2019) may serve as factors that affect the knowledge creation. Moreover, these studies support the idea that there is a strong relationship between BD characteristics and the firm's ability to create knowledge (Cappa et al., 2020; Ghasemaghaei, 2021; Ghasemaghaei & Calic, 2019; Johnson et al., 2017); but this line of research underestimates the complexity, the richness, the role, and the interaction between BD dimensions in knowledge creation, and there is a need to move beyond "what the big data are" to explore how to transform BD into valuable knowledge (Kar & Dwivedi, 2020).

The aim of this paper is to extend this research stream by examining BD dimensions and insights creations at a fine-grained level. Specifically, what is the mechanism of creating knowledge from BD? How to transform raw BD into BD knowledge insights?

With this aim in mind, we adopt a qualitative case study to explore the large-scale multinational pilot "Active Travel Insights" managed by three firms and launched in three European cities (Manchester, Antwerp, and Helsinki). The pilot amalgamated a large amount of data feeds from different sensors and open data and created various insights to inform cities' cycling strategies and thus create knowledge insights through BD. Employing Gioia et al. (2013) methodology with both inductive and abductive-driven content analysis coupled with participatory observations enabled us to ground findings on the multi-level empirical base and build a conceptual framework of insights creation.

In this vein, our research offers a more in-depth understanding of the mechanism of knowledge creation in the BD context. Our results showed that BD components (i.e., 4Vs) could shape and tailor the knowledge created. In particular, we identify and explore four ways to enhance knowledge creation through BD, namely via (1) numerical questions, (2) explorative questions, (3) hypotheses questions, and (4) confirmation questions. The uniqueness of this study lies in the

fine-grained perspective of BD characteristics and the underlying mechanism of insights creation. Thus, we unpacked BD into a granular phenomenon.

Finally, we provide a framework that embraces all discovered complexities and fine-grained features of BD dimensions and guides knowledge creation from BD. Specifically, the framework explains the mechanism of creating knowledge from BD and transforming BD into specific BD insights by purposefully adjusting or focusing on specific BD dimensions. This is the first to our knowledge framework that comprehensively address BD knowledge creation.

2 Theoretical background

2.1 From Big Data to Big Data insights

BD—defined as structured and unstructured data characterized by high volume, variety, speed, and veracity (Cappa et al., 2020; George et al., 2014; Ranjan & Foropon, 2021; Sumbal et al., 2017)—is a valuable resource that allows firms to gain a competitive advantage and respond promptly to fast-changing environments (Akter et al., 2019; Bergamaschi et al., 2020; Chen et al., 2015; Di Vaio et al., 2021; LaValle et al., 2011). Since the seminal work of Laney (2001) and the latest developments in the BD field, *Volume*, *Variety, Velocity*, and also *Veracity* are recognized as core features or intrinsic dimensions of BD that distinguish it from mere data (Cappa et al., 2020; Chen et al., 2012; Simsek et al., 2019; Urbinati et al., 2019).

More specifically, *Volume* is conceived as the magnitude of BD and commonly perceived as a purely quantitative dimension where "big" reflects only the size of data (Gandomi & Haider, 2015; George et al., 2014; Yoo, 2015). The recent view on *Variety* reflects the structural heterogeneity of datasets and the BD's ability to provide granular precision thanks to the variety of data sources (Günther et al., 2017; Pröllochs & Feuerriegel, 2020; Yoo, 2015). The dimension of *Velocity* captures the speed at which the firm processes and analyzes data (Akter et al., 2019; Larson & Chang, 2016; Sivarajah et al., 2017). And, BD *Veracity* identifies the credibility or trustworthiness of the output generated (de Camargo Fiorini et al., 2018; Surbakti et al., 2020), including the presence or absence of high-quality data regarding timing and relevance (Ghasemaghaei & Calic, 2019; Lee, 2017; Sukumar & Ferrell, 2013) or technical aspects (Côrte-Real et al., 2020).

Existing BD research has acknowledged that BD per se does not guarantee higher firm performance or competitive advantage (Henkel & Hartmann, 2020). These can be achieved by creating BD insights that are the outcomes of BD analytics[1] used to extract valuable meaning from BD (Chen et al., 2015; Gandomi & Haider, 2015; Ghasemaghaei, 2020; Vidgen et al., 2017; Waller & Fawcett, 2013). As highlighted in the existing literature, insights should be "actionable", "valuable", and "timely", thus, they should be related to business strategy, easy to understand and consume,

[1] BD analytics is the process of utilization of various tools and techniques with the purpose of extracting valuable insights from BD (Gandomi & Haider, 2015; Saggi & Jain, 2018; Sivarajah et al., 2017).

and created at a needed time (Chen et al., 2015; LaValle et al., 2011; Sivarajah et al., 2017; Wang & Hajli, 2017).

Scholars have primarily investigated the directionality of the relationship between the creation of insights and BD dimensions, considering 4*Vs* features an inseparable part of BD that influence insights (Chen et al., 2017; Ghasemaghaei, 2020). In particular, Ghasemaghaei and Calic (2019) found that *Velocity, Variety*, and *Veracity*, positively influenced BD insights created, while the *Volume* dimension alone had almost no effect on insights creation. Cappa and colleagues (2020) observed a linear positive effect of *Volume* and *Variety* when working together and no effect of solely *Volume* on insights creation; instead, *Veracity* directly benefits firm performance by enabling the creation of reliable insights. In this vein, the shifting focus to insights creation has become a new cornerstone for firms willing to leverage BD intrinsic dimensions to extend knowledge (Ciampi et al., 2020; de Camargo Fiorini et al., 2018; Lee et al., 2007; Sumbal et al., 2017).

Taken together, studies have provided valuable contributions about the intrinsic BD characteristics and their potential synergic forces that affect performance and insights creation (e.g., Ghasemaghaei & Calic, 2019). Although this research stream suggests synergies are embedded in the 4Vs dimensions playing a crucial role in tailoring BD insights (Ghasemaghaei & Calic, 2019), there is still a lack of examination of the mechanisms activated at the level of BD and their effect on knowledge creation. Understanding these mechanisms becomes salient for more effective and efficient use of BD in organizations (e.g., Gandomi & Haider, 2015; Vidgen et al., 2017) and moves beyond "what" BD represents to the "why" it is so (Kar & Dwivedi, 2020). To disentangle these dynamics, we draw upon the knowledge creation literature since it is particularly suitable and allows a more fine-grained look at the core dimensions of BD and investigation of the underlying processes of insights creation.

2.2 A knowledge creation lens on Big Data insights

Existing knowledge management studies have broadly highlighted the creation of knowledge as an essential process within organizations (Abbasi et al., 2016; Lapré & Wassenhove, 2001; Nonaka, 1994; Nonaka et al., 1996, 2000; Sumbal et al., 2017). Indeed, knowledge and the ability to create and use it strongly affect a firm's long-term competitive advantage (Drucker, 1993; Grant, 1996; Leonard-Barton, 1995; Nelson, 1991; Nonaka, 1990, 1994; Nonaka & Takeuchi, 1995; Nonaka & Toyama, 2015; Quinn, 1992).

Inspired and stimulated by these considerations and the conceptualization of BD and its components (4*Vs*), in this study, we are interested in understanding the mechanism of extracting knowledge from BD, and transforming BD into BD knowledge insights. In the specifics, as highlighted in the previous sub-section, BD is data characterized by *Volume*, *Variety, Velocity*, and *Veracity.* In general, data are raw facts or symbols that synthesize the properties of an object or event (Ackoff, 1989), allowing measurements and comparisons. In the case of BD, studies have shown that BD is a desirable knowledge source because of BD's distinct

nature and unique characteristics that make BD different from just large datasets (Pauleen & Wang, 2017; Ranjan & Foropon, 2021; Sumbal et al., 2017) and knowledge creation benefits from the different dimensions of BD (Cappa et al., 2020; Gani et al., 2016; George et al., 2014; Hashem et al., 2016; Simsek et al., 2019; Urbinati et al., 2019).

When BD is placed in a specific context so that it can be deduced and associated with meaning, it can create information (Ackoff, 1989; Tang et al., 2016). The aspect of the acquisition of new meaning and the importance of the knowledge creation process defines information as semantic (Ackoff, 1989; Nonaka, 1994). It is a flow of messages or meaning (Machlup & Mansfield, 1983) that enables knowledge creation, and the amount of information a signal carries determines what we can learn from it (Dretske, 1981). Indeed, several studies have shown how knowledge is captured, organized, and formalized into sources of information (Kogut & Zander, 1992; Zack, 1999).

Knowledge is commonly conceived as "justified true belief," and thus, it is embedded in the holder's commitment and beliefs (Nonaka, 1994). More specifically, knowledge has been seen to reside in many knowledge sources, from artifacts and individuals to organizations (Nonaka et al., 2000), and it can also be extracted and externalized in the form of insights (Cappa et al., 2020; Dam et al., 2019; Grover, 2020). Thus, BD insights, as extracted knowledge from BD, become a critical organizational component since they go through the process of growing and advancing the knowledge base (Balestrin et al., 2008; Ranjan & Foropon, 2021). The knowledge base can be further extended by combining different in-depth insights essential to building coherent and more comprehensive pictures for effective decision-making (Cappa et al., 2020; Ghasemaghaei & Calic, 2019, 2020; Wessel, 2016).

According to Nonaka (1994), the knowledge creation process involves two types of knowledge: explicit and tacit. Explicit knowledge refers to "know-what" and relates to the codification and formalization of knowledge in a tangible form using formulas, drawings, numbers, or words (Grant, 1996; Johnson et al., 2002; Nonaka & Konno, 1998; Nonaka & Takeuchi, 1995; Polanyi, 1962). Explicit knowledge is fully transferable, and it can be easily shared. In contrast, tacit knowledge is conceptualized as "know-how" and characterizes routines that leverage accumulated knowledge, experience, and learning (Nonaka, 1994; Polanyi, 1962). While some tacit knowledge cannot be explicated, most of the tacit knowledge can be codified and formalized (Nonaka & Takeuchi, 1995; Collins, 2010). Both tacit and explicit knowledge has been conceived as a continuum in which they constantly interact (Nonaka & Takeuchi, 1995; Nonaka & von Krogh, 2009).

Finally, as knowledge creation is the process that allows turning raw BD (and its 4V dimensions) into BD insights, BD is recognized to generate mainly explicit knowledge (since insights are codified and easily transferable). Indeed, patterns and indicators are embedded in BD and ready to be extracted in the form of insights (Kabir & Carayannis, 2013). Knowledge creation studies have recognized BD to be a form of knowledge source with a vast potential to enable both large-view and nuanced insights that can be available almost at any time (Barton & Court, 2012; Ciampi et al., 2020; Hamilton & Sodeman, 2020;

Intezari & Gressel, 2017). In this way, insights contribute to making available, amplifying, and crystallizing knowledge within an organization's knowledge system (Nonaka et al., 2000).

3 Methods

3.1 Research setting

As the present research aims to explore the knowledge creation phenomenon by focusing on BD dimensions and insights creation at a fine-grained level, we chose to adopt a qualitative approach (Lee et al., 2007; Yin, 2009). The research context is the creation of BD insights out of a large amount of raw BD flows generated. For this, the paper uses a case study method to elaborate further on existing studies on BD dimensions and derive their role in BD insights creation. We examine the "Active Travel Insights" (ATI) case, a large-scale multinational pilot project managed by three private firms operating in the Internet-of-Things (IoT) sector. The pilot amalgamated different data feeds from various sensors in three European cities: Antwerp (Belgium), Helsinki (Finland), and Manchester (United Kingdom). The ATI project aimed at creating various insights from BD to help cities gain a deeper understanding of their cycle network, including how many people use cycle routes, deep insights into the interaction between different types of road users, the impact of the road network on the environment, and where citizens are traveling to and from. The three IoT companies were responsible for the whole pilot planning and implementation, and three European cities' municipalities played the customer role.

The reasons that prompted the rationale to analyze the chosen case are twofold. First, the large-scale project with multiple stakeholders (from data scientists to managers, firstly meeting BD) and pilot's goals (unique for each city) allowed the exploration of different types of insights, from simple to complex ones. Second, the case's pilot nature made it possible to investigate the insights creation at a high-granularity level from the beginning.

The pilot used private data (exclusively produced by three types of sensors) and open data presented by each city independently. Private BD was generated with the commercial aim in real-time and remained in exclusive possession of firms after the pilot ended. Open BD was generated by state-owned sensors before the project started and was available to anyone to observe environmental conditions in a specific city. Open BD was historical and static at the moment of the pilot run; they represented one-year-old datasets. Figure 1 in Online Appendix lists three IoT technologies that generated different BD types in real-time and accumulated historical BD. Furthermore, the dashboard was tailored using the API of IoT companies and open-source data managing frameworks. The dashboard enabled raw BD accumulation and visualization either in real-time or in historical perspectives (daily, weekly, or defined time period).

3.2 Research strategy

BD dimensions exploration and insights investigation represented two main phases of the research analysis. For each phase, we used two approaches: an inductive approach with abductive procedures and a purely inductive approach. Specifically, we first employed the inductive Gioia et al. (2013) methodology with abductive procedures to analyze the vast body of formal documents and internal communication messages exchanged in Slack chat related to BD dimensions and BD insights creation. Thus, we could ground inductively derived insights on the existing BD studies. We constantly consulted existing literature throughout the coding process. A combination of abductive and inductive procedures is used especially for emerging topics (Weber et al., 2019) and allows for mitigating fast-taken or phenomenon-driven conclusions (Gioia et al., 2013, p. 26; Ramus et al., 2017).

Second, we adopted the participatory observation method (de Ven & Poole, 1995; Street & Meister, 2004; Van Bryman, 2012) based on indicative reasoning. The participatory observation method is recognized as efficient for understanding complex situations and relationships and achieving more significant and less hierarchical research practice (Clark et al., 2009). One of the authors was involved in the pilot as the Coordination Manager of one of the IoT firms and followed the insights creation process from its beginning to the end. We used participatory observations to support and enrich abductively developed findings. In this way, we enabled reliable evidence base and facts production by triangulating data (Yin, 2013) from a vast body of formal documents, communication in Slack chat, and participatory observations. To construct a detailed narrative from various contents, we adopted the "narrative strategy," one of the strategies for sensemaking of eclectic data proposed by Langley (1999).

3.3 Data collection

Data collection started in April 2019 and ended in August 2019. We collected data over five months. Internal official documents, including project milestones and deadlines, strategic plans, goals setting, monthly reports, final closing up report, and others, were collected. These documents were of special importance as they reflected the main steps, changes, rules, challenges, and plans of the pilot as the project was run mainly online. All internal documents reflected in a formal way what was happening during the project. Specifically, three obligatory monthly reports delivered to the funding body consisted of main milestones and the progress of the pilot. They were crucial to grasp summarized changes and nuances that were not possible to get from any other documents due to the large-scale nature of the project. The Final Report was a key to extracting the result, insights created, data used, and technical problems during the project. The report provided a fully summarized view of the project results and precise information on technology performance and insights created. All data sources are specified in Table 1. More than 250 pages were collected throughout the pilot.

Table 1 Summary of data sources

Method/step	Data source	Source details	Quantity
Qualitative content analysis	Internal documents:		
	1. City goals and data availability document	1. The document indicated needs, goals in launching the pilot for each city, and available open data	1. 15 pages
	2. Bid application	2. Bid application with the description of the technology to use, needs to tackle, challenges, technical specifications, expectations, possible issues	2. 15 pages
	3. Scope of the pilot for all participants	3. The document with clear goals for each participant, technical specifications, data, and privacy protection rules and regulations	3. 4 pages
	4. Hardware specification guide	4. Hardware specifications for installs, supportive infrastructure requirements, etc	4. 23 pages
	5. Re-scoping document	5. Changed Scope document according to the new GDPR requirements expressed by cities, GANTT chart	5. 13 documents (80 pages)
	6. SWOT analysis	6. SWOT analysis of the pilot	6. 2 pages
	7. Project management documents, GANNT charts	7. Managerial guidelines for the implementers of the pilot: milestones, phases, deadlines, KPIs, responsible persons	7. 2 excel files
	8. Project Risks analysis	8. Document indicating technological and managerial risks of the pilot	8. 3 reports in average 15 pages each
	9. Mid Term reports	9. Three reports from the technology companies reporting reaching milestones – for funding body	9. 29 pages
	10. Final report	10. Final report communicating the final results of the project for cities: number of sensors, insights gained, participants, KPIs reached, deadlines met/failed, challenges and issues, suggestions for the future Big Data projects	10. 14 pages
	11. Pilot showcase presentation	11. The presentation at the special event at the end of the pilot for the UK cities municipalities, road and city planners	11. 21 slides

Table 1 (continued)

Method/step	Data source	Source details	Quantity
	Internal communication chat Slack	All communication in the Slack App (internal chat) between all members of the technology firms involved in different positions (i.e., Managers, Data Scientists, Technicians). It served as the platform to communicate, plan, discuss issues, and share solutions	4 months of gathering data 85 pages transcribed according to dates and phases of the pilot
Participatory observation	Online weekly standup calls of technological companies	1. During the calls, technology companies discussed the flow, timing of the project, project, technological, data gathering and analysis challenges, dashboard design, and development, urgent needs, mutual steps to perform, plans for the next week, and results of the previous week	13 h in total of 15 calls 20 pages of the notes
	Online standup calls cities and Project Lead	1. During the calls were discussed challenges, issues, needs, changes to implement	3 calls for Helsinki 2 calls for Manchester city 10 pages of the notes
	Offline meetings: 1.Meeting with the Project Lead team during the weekly call 2. Meeting with the Funding Body team, the weekly call	2. London office of the Project Lead technology company. 1 weekly call meeting, discussion on the location choose and online visualization of the final London office of the Funding body team. Discussion on the re-scoping of the pilot goals according to the new GDPR requirements expressed by cities	1. 1.13 h 2. 40 min 10 pages of the notes
	Visualization Dashboard development	Visualized Big Data both in real-time and historical data. The research was observing 2. The design and development process of the dashboard 3. The process of testing the dashboard Changes implemented	12. 2 months of the dashboard design and development
	Raw BD, CSV files	CSV files with all Big Data gathered during the project run. The researcher was observing 1. *Volume, Variety, Velocity, Veracity* dimensions, and their characteristics and changes reflected in changes of single variables, single data sources, timing, interruption in data generation, and other aspects	Wi-Fi sensors (Helsinki, Antwerp) CCTV (Antwerp, Helsinki, Manchester)

Furthermore, we had access to internal chat messages where the main communication occurred between the IoT firms responsible for the installs and the whole pilot design. The Slack app was the place for the informal and fast exchange of messages between all participants. The city representatives were not included in the chat; thus, all project issues, challenges, and pitfalls were discussed with a high degree of openness. The chat was launched at the stage of the project preparation for the bid application, thus, before the official project started, and was closed one month after the project ended. Analysis of Slack chat was particularly helpful in following the overall flow of the project from its start to its end, understanding practical issues of BD concerning dimensions, enable the collection of more pieces of evidence on subdimensions, and other aspects, which were not or could not be explicitly reflected in formal documents. The Slack logs transcript took more than 85 pages.

Finally, a large amount of primary data based on participatory observations (de Ven & Poole, 1995; Street & Meister, 2004; Van Bryman, 2012) was collected. Specifically, the author participated in the action planning stage of the pilot and later in action taking the stage, which consisted of the sensors' location choosing, technology settings, dashboard designs, data generation, and data analysis. The author participated in the action evaluation that is reflected in the final report and attended weekly online calls (via Google Hangouts), real-life meetings with the pilot projects managers and IoT firms (London office). As the project was international, all meetings were held online, with a few offline meetings between different stakeholders. These allowed us to explore the nature and characteristics of BD dimensions within a concrete business setting.

3.4 Data analysis

Data analysis had two specific aims: to derive fine-grained aspects of BD dimensions and discover the role of fine-grained BD dimensions features in insights creation. For these, not only do we have to make sense of a vast amount of heterogeneous content and participatory observations but also base the discoveries on the existing literature to allow new insights to emerge. Therefore, the first aim was addressed by employing the qualitative methodology developed by Gioia et al. (2013). Moreover, following other comparable studies (Ramus et al., 2017; Weber et al., 2019), we combined inductive and abductive procedures to find hidden but essential aspects of predefined BD dimensions. First, two authors analyzed and identified BD-related topics across formal documentation and informal communication in Slack chat through inductive coding, supporting the inductive process with participatory observation arguments. To enable a more nuanced view of the phenomenon, we first derived the topics within informal communication (Slack chats), then performed the same coding procedure for all formal documents. Specifically, first-order topics of both datasets were related to the strategic and technical aspects of data generation, analytics, insights creation, installs, data models, etc. Second, the other two authors joined, and all four conducted the cross-datasets analysis and combined the similar first-order codes in one coding scheme. All codes of the informal coding scheme became part of the general

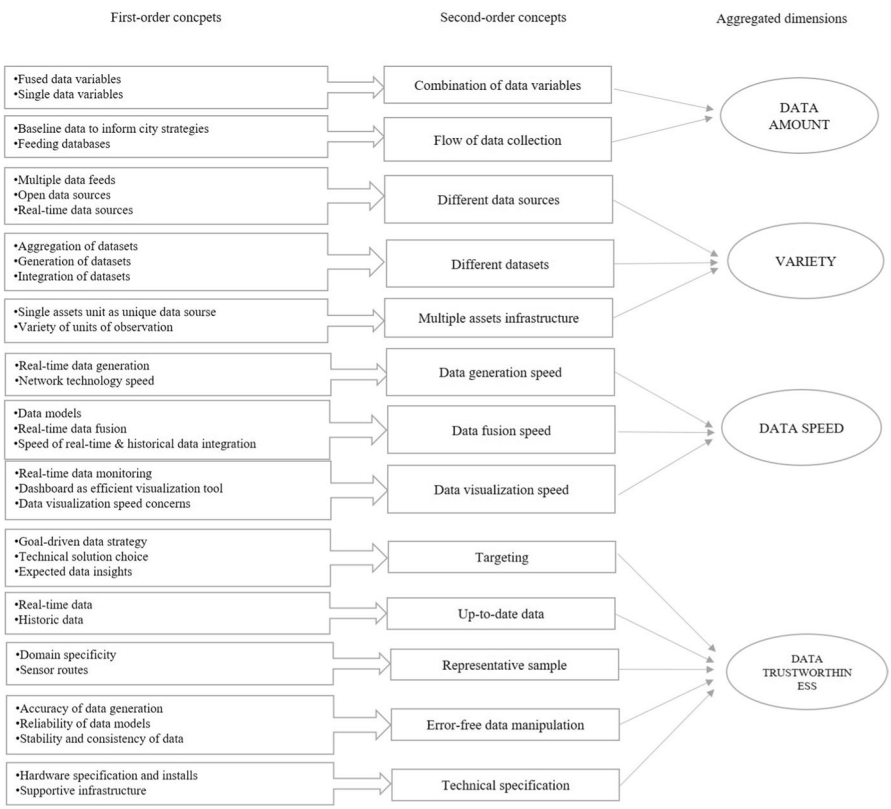

Fig. 1 Data structure

coding scheme; informal content played a great supporting role in the formal content analysis. Overall, thirty-one topics were derived. We then aggregated them into broader second-order concepts (thirteen concepts). Finally, following the abductive approach, we purposively aggregated second-order concepts into four predefined categories, namely, four BD dimensions – *Volume*, *Variety*, *Velocity*, and *Veracity* proposed by prior BD research. In this way, using both inductive and abductive reasoning, we could unveil new fine-grained aspects of established BD dimensions. The value of our distilled data structure lies in first-order and second-order concepts that disentangle predefined 4Vs (our aggregated dimensions). The final data structure of first-order, second-order, and aggregated dimensions is presented in Fig. 1 (for the more detailed data structure that also includes exemplary quotes in Online Appendix, Table 1).

Second, we analyzed insights created during the pilot. This process relied on the inductive participatory observations method (de Ven & Poole, 1995; Street & Meister, 2004; Van Bryman, 2012). Specifically, we actively used field notes created during participatory observations. Moreover, we had access to all raw BD in CSV files and access to the constant flow of insights generated and visualized

on the dashboard. We started the analysis from raw data generated in three cities (single data variables, such as geolocation, date, time, category, and the number of road users, single data sources, and an indication of data sources). Then we moved to the analysis of any types of insights created during the pilot either visualized or not (insights visualized on the dashboard, created for midterm reports, presentations, final presentation for cities or the funding body, preliminary insights, training insights, etc.).

4 Emergent findings

Our research focused on understanding the mechanism of creating knowledge from raw BD (and its 4Vs dimensions) by disentangling BD dimensions and unveiling their unique features influencing BD insights creation. Thus, when distilling the coding structure and unpacking 4Vs, a strong pattern emerged; each Big Data 4Vs has its related fine-grained characteristics and features that uniquely shape BD dimensions and BD insights created. Moreover, owing to participatory observations, we could distinguish among a large amount of BD insights created four specific types of insights. In the following paragraphs, we first expose the unique granular characteristics of each BD dimension and disentangle the 4Vs of Big Data. Second, four types of BD insights created and their role in knowledge base creation are laid out.

4.1 Disentangling Big Data dimensions

4.1.1 Making sense of Big Data volume

The coding structure unveiled that data volume was enabled by the constant flow of raw data collection and the combination of data variables. The two data flows had different functions in the project; overall, they contributed to the growing BD volume.

4.1.1.1 The flow of data collection The coding activity allowed us to see how simply generating and collecting datasets and information, such as real-time sensor data or previously collected open data, permitted the initial pull of a large amount of data. Indeed, the initial data collection was needed to start the project by feeding cities' and the funding body's databases, reassuring that cities have access to data, and making it easier for data to be incorporated into other platforms. Moreover, feeding local databases with raw data was crucial to enable the baseline data or the minimum necessary amount of Big Data to inform cities' strategies (e.g., provide initial data on how cities ran before and after the implementation of cycling interventions or enable the justification of cycling or pedestrian infrastructure investment). The baseline data amount was a collective decision in terms of the number and time sensors were operating in each city.

4.1.1.2 Combination of data variables We discovered that by increasing the number of single variables and merging them, the pilot raised the overall amount of data. Thus, in the Final Report, managers indicated the amount of data consumed and created in each Pilot city as single data variables collected (e.g., crowd flow, air quality, traffic flow). Furthermore, we observed that the growth of the data amount on servers and the platform regarded the fusion of data variables. For instance, although managers of the CCTV cameras technology recognized single data variables as a meaningful autonomous piece of data (e.g., number of specific type of vehicles), they always referred to them in combination with other variables (e.g., number of cyclists using lane at that location, and travel direction). We found that the combination with other variables enabled the creation of another unique piece of data for managers. In this way, there was a distinction for managers between a large amount of raw data that was viewed more as a large amount of not connected data variables (e.g., cities, streets, road users, time, dates) and fused data variables that were considered critical to provide a more complex insights picture (e.g., correlations between peak times of motorized vehicles and air quality).

4.1.2 Exploring the Big Data variety

We found that the Variety of BD was expressed not solely through the various data sources but also through various datasets and multiple assets infrastructure. Having them all together enabled data type "diversification," creating unique hybrid data sources and ultimately growing the BD Variety overall.

4.1.2.1 Different data sources The variety of data sources involves multiple data feeds and constant access to various data, such as open and real-time data. Moreover, various data sources enabled real-time data (e.g., Wi-Fi sensors, CCTV cameras, and air quality sensors), while open data regarded static and historical data provided by cities. Having separate data sources helped provide unique information on cities' traffic and air quality conditions. However, it was also crucial to combine them to create a sort of a hybrid data source. The pilot managers saw the fusing data feature as a unique advantage of the project, allowing a complete view of how active travel impacts the road network.

4.1.2.2 Multiple datasets Producing multiple datasets from several data sources was the focus of the pilot too. Indeed, multiple datasets enabled a richer view of movement dynamics which could be used to promote bicycle traveling. With this respect, three different aspects of data variety emerged: aggregation, generation, and integration of datasets. Specifically, not only the aggregation of both individual and combined datasets enable a rich base of various datasets, but it also was perceived as the "intelligent data collection and analysis solutions" (Final Report). Moreover, the generation of new datasets was important as it provided new and previously unavailable information. Finally, a particular interest of managers was devoted to integrating datasets into a dashboard as combining multiple datasets will help cities gain a deeper understanding of their cycle network.

4.1.2.3 Multiple assets infrastructure We discovered that asset infrastructure—which refers to any asset (sensor) or observation unit (city, street, or square)—was regarded as a valuable data holder. For instance, a single sensor as part of a specific sensors network (e.g., Wi-Fi sensor network) was considered a unique data source and even had its own "tech passport" that included unique location (city, street, and GPS coordinates with latitude and longitude), type of data and information created, and time of deployment. As we observed, it was possible to visualize unique data generated by a single device on the dashboard and compare it not only with data produced by other devices but also by devices of the same network. Finally, we found that either a single city, street, or square were considered unique units of observations and, as such unique sources of data. For example, Manchester city possessed specific data, such as air quality data, pedestrian counts, or vehicle counts, that were further compared with Helsinki city air quality data, noise level data, and weather data to derive more granular insights into how the two cities were performing. We could observe the same pattern for a single street or square.

4.1.3 Unpacking velocity in Big Data

We found that the speed of data was addressed from different angles that were each important to the overall speed topic. Thus, the further content analysis revealed intensive attention toward four aspects related to the data speed during the project: speed of data generation, data fusion, data visualization, and data use.

4.1.3.1 Data generation speed Specifically, the role of speed of data generation was essential as the project aimed at collecting real-time BD. Thus, each sensor network had an obligatory indication of the connectivity speed: "IoT Deployment Network Technology 3G/4G" (specified in all monthly reports and technical specifications). Furthermore, the quality of the network technology was crucial to enable the constant and real-time data flow. With this respect, issues of the consistent data generation speed were addressed as priority issues in the pilot.

4.1.3.2 Data fusion speed The speed of data fusion played a core role in the pilot. Specifically, data fusion speed refers to the speed of data "amalgamation" (frequently used during the pilot term) and integration of various datasets, either real-time or static historical data. In addition, as the pilot was especially interested in various datasets' interoperability to deliver insight fast, the need for fast data amalgamation and integration prompted the use of standardized data models. Thus, each sensor network specification consisted of the list of data models used to amalgamate and integrate datasets in real-time.

4.1.3.3 Data visualization speed The fast data visualization was critical, as it enabled immediate access to information for decision-makers. Specifically, pilot participants constantly referred to the visualization dashboard as a necessary and efficient tool to get fast access to data. Another aspect of data visualization, real-time monitoring, highlighted the ability of visualized data to provide immediate information about

how citizens move in real-time. Real-time monitoring of traffic intensity allowed spot changes in real-time, such as reducing traffic-related emissions based on real-time data. The importance of data visualization could not be underestimated as we found that among three main lessons learned, managers claimed the importance to "set clear expectations on the dashboard deliverables to ensure smooth delivery" (Final Report).

Finally, data use refers to the rate of data useful time or data "best before." As the pilot did not presuppose the immediate use of gained insights, we could not fully capture the data use facet. However, the Final report allowed us to conclude that the ATI solution will influence both in the short and long run if insights are used immediately.

4.1.4 Highlighting Big Data veracity

We found that the *Veracity* of data was a crucial complex aspect. *Veracity* was continuously addressed at different levels, not only at a data level but also at pilot planning activities or technological setups levels.

4.1.4.1 Targeting The pilot set up targets upfront the project launch. Targets were related to specific goals of each city, expected data insights, and technological solutions. First, goals for each city were defined very precisely via interviews and questionnaires before the project launch. Revealed cities' goals helped to crystallize which data and insights are needed. Only after this the pilot managers defined IoT technologies to employ. The final choice was made for three technologies out of five others: CCTV cameras, Wi-Fi, and environmental sensors. In this way, managers could not only enable higher trustworthiness of their generated data but also create insights.

4.1.4.2 Up-to-date data *Veracity* of data was also addressed at the data level by delivering strongly relevant, up-to-date data that would inform new road strategies. Up-to-date data for the Pilot meant real-time gathered data and historical data. In this way, by using real-time and historical data, it became possible to provide comparative overview of past events and possible future changes dictated by the need. First, analyzing road traffic even from one year ago could have caused radical bias in new cyclist strategy development, while using real-time generated data enabled the adequate level of the data newness. Secondly, using static historical data for analysis of, e.g., traffic emissions in the Helsinki metropolitan area provided a possibility to compare it with real-time air quality data and get a granular dynamic picture of how traffic emissions changed in the city.

4.1.4.3 Representative sample At the very beginning of the project, pilot managers defined good data as "accurate, stable and consistent" (Bid document). Indeed, having a limited amount of time to get necessary data, pushed managers to find a solution to access the data that are representative enough to start building a new road strategy was crucial. A representative data sample was enabled by carefully

planning sensor routes (positioning on streets and squares). Sensors routes included only those that "connect businesses to residential hubs, new residential developments, or between major transport hubs" (Scope of the Pilot). Sensor routes were identified using local authorities' transportation domain knowledge and detailed online maps (e.g., Google Earth) that enabled the latest updates on locations.

4.1.4.4 Error-free data manipulation Delivering data analysis without (un)intentional errors was another way to enable good quality data. We found that this situation is connected with the use of wrong algorithms and models, but also errors during data generation and inconsistent and not stable data. Thus, disruptions in data generation were one of the most discussed topics in Slack chat. Moreover, critical attention to an appropriate choice of data models and APIs was observed. As pre-developed open-source data models were used in the pilot, they were tested to work efficiently with traffic, the crowd flows, and environmental counts within the transportation context by the funding body validation service. Exemplarily, CCTVs' "accuracy was tested through an independent validation exercise across a total of 27,000 vehicles, achieving + 97% accuracy" (technical specification document). Moreover, the technological provider's proprietary analytics models and algorithms were set up following the Pilot context. Thus, Wi-Fi sensors were adjusted to the EU markets, following the mobile penetration rate in countries and the error rate on the number of people who might have switched off mobile devices or turned off Wi-Fi search signal for (e.g., smartphone penetration is at 87% in Europe).

4.1.4.5 Respecting technical specifications Surprisingly, we found that respecting technical specifications was another key point to enabling the gathering of reliable data. Thus, all sensors had installation requirements specified in the hardware specification guide to strictly follow by all members. The document presented exact rules to follow and clearly defined outcomes for data generation if rules are violated. For example, CCTVs "need to be installed at a height between 4 and 8 m, with a clear view of the road from the side of the carriageway" as the best position to not lose any road users. "Sensors should not be positioned in trees where foliage may grow to obstruct the sensor's view of the road" (hardware specification guide). Furthermore, attention to supportive infrastructures, such as mobile connectivity and power aspects, was observed as the enabler of constant data generation and data quality. In this case, losing power meant that sensors were not producing data, while losing mobile connectivity meant that data were not entering servers and the platform. The consortium of technology companies created the Install Risks Register plan per each city, which includes an assessment of potential issues of power and connectivity, possible solutions, and the potential impact of an issue on the project. For instance, being in control of all sensors' status in different countries and being able to spot disruptions and respond quickly was one of the preventing measures. Exemplarily, this helped to quickly react to the disruption of a sensor in Antwerp that was not generating and sending data for four hours as it was unplugged.

4.2 Creating Big Data insights in the pilot

The ATI pilot was focused on turning raw BD into BD insights (as it stands from the pilot name) to create valuable knowledge. Thus, we found that a large number of BD insights created mainly represented four types capable of answering (1) *numerical questions*, (2) *explorative questions*, (3) *hypotheses questions*, and (4) *confirmation questions*. Owed by participatory observations and findings of the previous section, we found that 4Vs characteristics played different roles for each of these types of insights. Moreover, each of the four types served distinct purposes in creating a knowledge base. Finally, newly created BD insights became truly meaningful and valuable when moved from the "know-what" (what is the number of pedestrians and vehicles or air pollution level) to the "know-how" (e.g., how to implement these insights into new cities' strategy). In the following paragraphs, we explain the four types of BD insights and how their 4Vs shaped them. We also specify the role of each type of insight in knowledge creation.

The first group of insights was able to answer simple numerical or categorical questions, such as "What are the types of vehicles using a road?", "How many bicycles are on the road?" "What are environmental conditions close to a road area?". For instance, ATI project was interested in investigating "What is the total number of vehicles per each street in Manchester per week?". In this specific case, the analysis required simple descriptive analytics using data from only one data source, namely CCTV cameras. The examination mainly included single data variables related to streets, in/out direction, date, and not defined road users. Considering the simplicity of the investigation, no fusion of any of these variables occurred. Insights were generated in real-time but analyzed and visualized semi-automatically, reducing in this way the velocity dimension. Also, insights were not highly reliable as, exemplarily, not all sensors were set up correctly yet, the target of the data gathering and analysis was defined very; generally, the dashboard was in a testing phase, and data were not visualized automatically. It happened that one CCTV camera stopped gathering data due to the electric power blackout. Therefore, insights on numerical questions group were characterized by a low degree of *Volume*, *Variety, Velocity,* and *Veracity*.

The second group consists of exploratory analysis to test hypotheses such as "Types of vehicles used on roads are diverse," "There are fewer bicycles on roads than motorized vehicles," and "Is the number of motorized vehicles higher than the number of non-motorized one in Manchester per week?" To respond to these questions, the ATI project used data gathered from two data sources (Wi-Fi sensors and CCTV cameras) and mostly single variables like the counts of motorized/non-motorized vehicles. The distinguishing feature was that these variables were fused and visualized quickly and automatically. The algorithms used were reliable, and the technology setups were respected. These insights generated proved that the types of vehicles used on roads are diverse, that there are not many bicycles on roads than motorized vehicles, and that weekdays traffic trends are clearly defined and visible to build new traffic strategies. Therefore, the hypotheses questions involved a low degree of Volume and Variety but with a high degree of *Velocity* and *Veracity.*

The third group of insights included comparative questions, such as "What is the level of air pollution with all road users or with only cyclists and pedestrians?", "What

is the cyclists' peak hour over weekends?", "What is the main destination point of pedestrians on Wednesdays?". Data collection regarded several data sources, including single CCTV cameras (each sensor is considered a single data source) located at all count lines in Manchester. Variables were fused together, accounting for specific time slots during specific dates and every 15 min. Although the *Volume* and *Variety* were highly considerable, these insights lacked speed features and reliability. In particular, the data visualization was slow because the dashboard had certain visualization functionality restrictions, such as it was not possible for cities to get cumulative and automatically visualized answers on the busiest days in the cities (weekdays trends) and other insights but only via manual data fusion and visualization. The insights generated lack *Veracity* because data from specific data sources (i.e., CCTV cameras) was missed due to technical issues.

Finally, the fourth group consists of insights that could answer questions like "Can data confirm that removing all motorized vehicles from roads will have a direct positive impact on air pollution?", or "What will be the picture of air pollution like if we remove vans from a data set?". CCTV cameras and Wi-Fi sensors were crucial data sources; data coming from each unit was fused. In addition, a wide range of time frames and a high number of variables increased the *Volume* of data necessary for the analysis. For instance, these variables included street names, country, timeslot (morning/afternoon); direction (in/out), weekend (true/false); motorized vehicles (cars, trucks, motorbikes, vans, taxis, and emergency cars) and non-motorized (cyclists and pedestrians). *Velocity* was very high using real-time data and ready data models to automatically visualize the decision trees. As long as these types of insights were created mainly at the end of the pilot, the Veracity of data was enabled not only by the well-established targets and representative samples delivery but also by respected technical specifications and correct data models employment. Finally, these insights were included in the Final Report as the most reliable.

Additionally, we observed either one type or combined types of insights contributed to building the knowledge base of the pilot. Thus, while at the beginning of the pilot, mostly numerical questions were answered, we observed that even at the end of the pilot, these questions served together with confirmation questions. Specifically, the analysis of the pilot showcase final presentation for cities revealed that all four types of insights were presented. Interestingly, either a simple road user counts during the two working days on the Deansgate road (Manchester, August 12th to 17th, 2019; with the use of only manual tools like excel for analysis and visualization; all dimensions were at a low degree) and visualization graphs and sophisticated analysis (e.g., "this graph shows the road usage of all vehicle types at Deansgate/Queen St against the Air Quality at Hardman Street. There is a correlation between the increase and decrease of Vehicle movements and Air Quality", Pilot showcase presentation) were perceived by cities as valuable insights.

5 Discussion

The present research aimed to understand knowledge creation from BD by disentangling the mechanism of turning raw BD (and its 4V dimensions) into BD knowledge insights. We qualitatively studied the case of the ATI pilot and specifically how different types of insights inform new cycling and road strategies were created. Thus, by employing Gioia et al. (2013) method and coupling it with participatory observations, we translated the mantra of 4*Vs* (*Volume, Variety, Velocity*, and *Veracity*) into related fine-grained BD characteristics. In particular, dimensions of *Volume* and *Variety* compose the core of BD and share the key ability to be split and combined in different ways. In turn, *Velocity* and *Veracity* have their fine-grained features that shape their overall traits and influence insights creation. In addition, we identified four distinct types of insights, each including a different combination of fine-grained BD characteristics. These findings prompted us to go deeper into the shared similarities and differences between BD dimensions and conceptualize them into the framework (see Fig. 2) to explain the mechanism of creating knowledge from BD and transforming BD into BD insights.

The BD core composed of *Volume* and *Variety* can grow in *breadth* and *depth*. This growth mechanism reflected in the exponential growth of BD goes through a process of data fusion boosted by splitting and combining different levels of data and data source(s). Thus, the *breadth* direction of the mechanism is based on the ability of BD to be split into autonomous portions of data that can be fused for producing new output. In its turn, the *depth* direction of the mechanism is based on the ability to combine various autonomous portions of data in a meaningful way. However, the *depth* and the *breadth* are neither considered as the ends of a continuum, nor does a higher *breadth* lead to a higher *depth*. However, they can work simultaneously.

Furthermore, the growth mechanism allows BD to produce different types of knowledge. The *breadth* enables a broader scenario and general insights by increasing the number of single variables or data sources. The *depth* offers more specific and synthetic insights through various single data variables and data sources split and combined. In the case of *Volume*, the *breadth* enabled various general insights derived from single split variables, such as a simple number of road users on roads. The *Volume depth*, in turn, enabled the transformation of a large amount of unconnected data into a more complex insights picture, such as the correlation between peak times and the number of motorized vehicles on roads. In the case of *Variety*, datasets from single data sources, such as types of motorized vehicles and the level of environmental pollution, separately provided two large pictures. In contrast, the same combined datasets represented a sort of hybrid data source that enables a deeper understanding of air pollution concerning types of vehicles. We argue that an understanding of *Volume* cannot be narrowed to a big flow of raw BD or the *breadth* of data; it is due to the *depth* produced by data fusion that higher quality of insights was created. *A Variety* of data sources does not guarantee meaningful insights, while the right combination of data sources does.

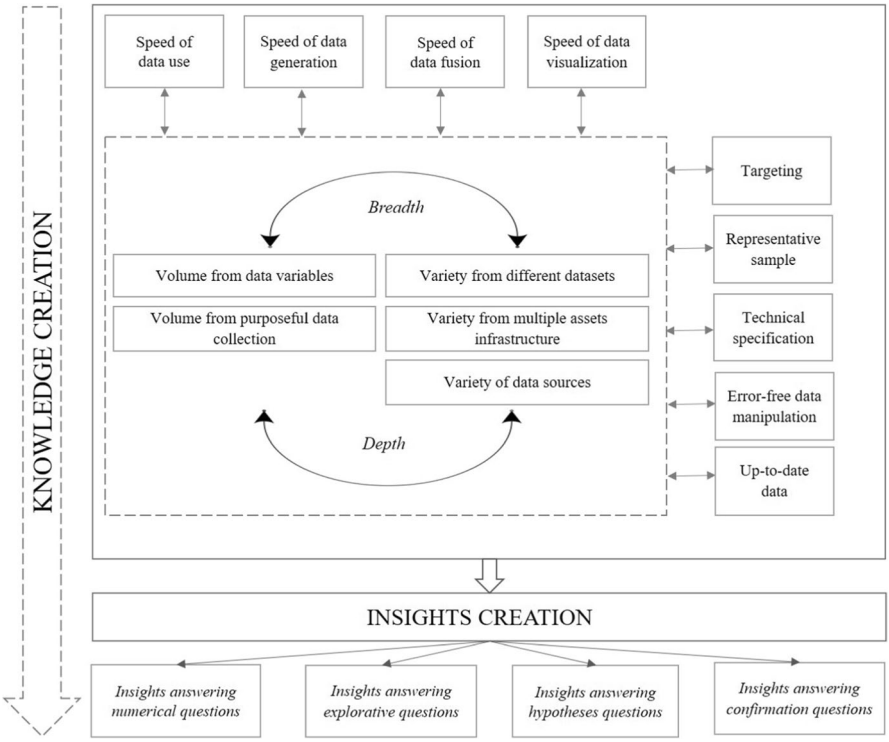

Fig. 2 A framework of Big Data and knowledge creation

Further, we found that *Veracity* and *Velocity* dimensions have their idiosyncratic features that work as sub-dimensions. Specifically, our research delineated four sub-dimensions of data *Velocity*: *data generation speed, data fusion speed*, *data visualization speed*, and *data use speed*. We also identified five sub-dimensions enabling *Veracity* of BD: *targeting, representative sample, technical specification, up-to-date data,* and *error-free data manipulation.* The specifics of sub-dimensions suggest that they are relatively independent of each other and have their unique say on BD core. Moreover, each sub-dimension separately affects either positively or negatively the ability of *Veracity* and *Velocity* dimensions to deliver fast and reliable insights. Thus, leveraging the BD mechanism and sub-dimensions, the pilot could create four different types of meaningful insights. These findings have important theoretical implications for research in BD and its role in insights creation.

5.1 Theoretical implications

Our findings suggest two theoretical contributions. First, we opened up BD's black box by disentangling the knowledge creation mechanism while transforming raw BD (and its 4*Vs* dimensions) into BD insights. Existing literature on knowledge management has pointed out that organizations are not anymore information-processing

machines, but knowledge-creating entities (Nonaka & Toyama, 2015), and knowledge from BD can be extracted and externalized in the form of BD insights (Cappa et al., 2020; Dam et al., 2019; Grover, 2020). Thus, BD insights become a critical organizational component since they go through the process of growing and advancing the knowledge base (Balestrin et al., 2008; Ranjan & Foropon, 2021). Our results showed that BD components (i.e., 4Vs) can shape and tailor knowledge created. In particular, we identify and explore four ways to enhance knowledge creation through BD, namely via (1) numerical questions, (2) explorative questions, (3) hypotheses questions, and (4) confirmation questions. In this vein, our research offers a more in-depth understanding of the granularity of the process of knowledge creation in the BD context.

Second, our study responds to the need to move beyond the "what the big data represents" to the "why it is so" (Kar & Dwivedi, 2020). Prior BD literature highlighted that BD represents a strategic source for firms, and understanding the mechanisms of creation of different insights allows firms to gain valuable knowledge to succeed (Cappa et al., 2020; Chen et al., 2012; Ghasemaghaei, 2021; Ghasemaghaei & Calic, 2019; Johnson et al., 2017). This research stream highlights the importance of exploring the relationship between BD characteristics and valuable insights (Ghasemaghaei & Calic, 2019, 2020), mainly focusing on the effects that dimensions have on insights. Our findings offered empirical evidence of the growth mechanism working on *Volume* and *Variety* dimensions. The mechanism triggers not only an exponential growth of BD dimensions but also influences insights created. While the *depth* enables broader and general insights, the *depth* offers more nuanced and synthetic insights. In this vein, our research could extend prior BD studies by unveiling the underlying mechanism influencing insights creation (Cappa et al., 2020; Ghasemaghaei & Calic, 2019, 2020).

5.2 Practical implications

The present research provides an understanding of the mechanism of producing insights needed for unique use-cases; this does not mean distinguishing between creating "good" or "bad" insights. Taking these aspects into consideration, our findings have several managerial implications. First, the growth mechanism and the notion of BD sub-dimensions can become an assessment tool for any BD project, guiding managers and assuring the necessary level of quality, depth, and speed of insights creation. Our findings suggest that, while BD as the homogeneous entity is not easy to comprehend for producing insights, BD disentangled into sub-dimensions is more manageable. This will also help in adopting a smart resource allocation; this does not mean using many computational tools if managers need simple numerical insights or adopting very advanced visualization tools to get fast insights instead of deep but slow ones. Our research showed that organizations might already intuitively implement wise resource alignment. However, doing it more consciously and strategically can lead to better value creation. Second, firms can better control their expenses, as they can increase their ability to adjust the quality of insights using the mechanism at work. They might employ more hardware to generate various data (*breadth*), which is particularly costly

or to deepen insights via algorithms (*depth*) instead. Alternatively, firms might invest in real-time data generation or buy historical data from a third party, as either source of the data can fit a project, but the latest option costs less. This can lead to better returns on investments. Finally, by having more knowledge of BD dimensions, firms and managers can enhance their planning activities and craft long-term data-driven strategy.

5.3 Limitations and future research directions

While the research question drove the choice of qualitative research, the latter generated some limitations. The explorative nature of the study can limit its potential for generalizability. Although the case study included the analysis of insights produced for three different countries, they represent a limited sample within European geography. Moreover, although we analyzed how private and open data were working together to produce insights, we explored deeper how private data were generated while not capturing details of open BD generation. The case did not work with public data. The absence of facts limits our findings related to open and public BD. Moreover, the types of BD represented in the case could not fully address the diversity of existing BD types. Additionally, the Smart Cities context and the peculiarities of IoT technologies used in the pilot represent a particular case and can restrict the findings' generalizability. Finally, the pilot nature of the case might influence the ability to fully explore the Velocity dimension and its sub-dimension, specifically, data use speed, which we believe is crucial for insights creation and value generation.

Future research might address the above-mentioned limitations by employing other methods and testing the growth mechanism by exploring the sub-dimensions role in insights creation through a multiple case study approach or by validating our framework with a quantitate approach on more massive data sets. Future research should test findings in other fields and contexts as well as on other types of data. It is of special interest that future research focuses on challenging the similarities between BD dimensions (*Volume* and *Variety*, *Velocity* and *Veracity*) that constructed the final framework by investigating their heterogeneity depending on a context, type of BD, or BD technology. An important aspect is exploring the boundaries of two directions of the mechanism, specifically, if a high degree of both *breadth* and *depth* always has a positive sign. We suggest that some other sub-dimensions of *Veracity* and *Velocity* might arise in different contexts. Lastly, BD has been seen as a meaningful technology used in the decision-making process (Aversa et al., 2018; Galliers et al., 2020; Zaitsava et al., 2022). The present work could not capture insights on decision support systems; however, future research might focus on how the quality of BD affects decision making, what is the role of cognition in this context, and how the characteristics of BD engage in an interplay with cognitive individual or collective decision-making processes.

6 Conclusion

The research offered interesting findings on BD dimensions providing a fine-grained perspective of *Volume* and *Variety* that cannot be narrowed down to purely size and variety of data sources or *breadth,* as it also has the *depth* of BD. Moreover, *Veracity* and *Velocity* cannot be narrowed to data generation speed or the reliability of data sources only as they include other sub-dimensions. A key contribution of the research was the discovery of the underlying mechanisms that enable different insights creation via the *depth* and the *breadth of Volume and Variety*. Moreover, the discovered sub-dimensions of *Velocity* and *Veracity* dimensions affect the level of usability of insights. The findings help managers to develop and manage credible knowledge from BD.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10997-022-09659-0.

Funding The authors gratefully acknowledge the financial support of Fondazione di Sardegna.

Data availability N/A.

Code availability Python Software.

Declarations

Conflict of interest The authors declared that they have no conflict of interest.

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2), 3.

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*(1), 3–9.

Akter, S., Bandara, R., Hani, U., Wamba, S. F., Foropon, C., & Papadopoulos, T. (2019). Analytics-based decision-making for service systems: A qualitative study and agenda for future research. *International Journal of Information Management, 48*, 85–95.

Aversa, P., Cabantous, L., & Haefliger, S. (2018). When decision support systems fail: Insights for strategic information systems from Formula 1. *The Journal of Strategic Information Systems, 27*(3), 221–236.

Balestrin, A., Vargas, L. M., & Fayard, P. (2008). Knowledge creation in small-firm network. *Journal of Knowledge Management, 12*(2), 94–106.

Barton, D., & Court, D. (2012). Making advanced analytics work for you. *Harvard Business revIew, 90*(10), 78–83.

Bergamaschi, M., Bettinelli, C., Lissana, E., & Picone, P. M. (2020). Past, ongoing, and future debate on the interplay between internationalization and digitalization. *Journal of Management and Governance, 25*, 983–1032.

Bryman, A. (2012). *Social Research Methods* (4th ed.). Oxford University Press.

Cappa, F., Oriani, R., Peruffo, E., & McCarthy, I. (2020). Big Data for creating and capturing value in the digitalized environment: Unpacking the effects of volume, variety and veracity on firm performance. *Journal of Product Innovation Management, 38*(1), 49–67.

Chen, D. Q., Preston, D. S., & Swink, M. (2015). How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems, 32*(4), 4–39.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Chen, H. M., Schütz, R., Kazman, R., & Matthes, F. (2017). How Lufthansa capitalized on Big Data for business model renovation. *MIS Quarterly Executive, 16*(1), Article 4.

Ciampi, F., Marzi, G., Demi, S., & Faraoni, M. (2020). The big data-business strategy interconnection: a grand challenge for knowledge management. A review and future perspectives. *Journal of Knowledge Management., 24*(5), 1157–1176.

Clark, A., Holland, C., Katz, J., & Peace, S. (2009). Learning to see: Lessons from a participatory observation research project in public spaces. *International Journal of Social Research Methodology, 12*(4), 345–360.

Collins, H. (2010). Tacit and explicit knowledge. In *Tacit and explicit knowledge*. University of Chicago Press.

Côrte-Real, N., Ruivo, P., & Oliveira, T. (2020). Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value? *Information & Management, 57*(1), 103141.

Dam, N. A. K., Le Dinh, T., & Menvielle, W. (2019). Marketing Intelligence from Data Mining perspective. *International Journal of Innovation, Management and Technology, 10*(5), 184–190.

de Camargo Fiorini, P., Seles, B. M. R. P., Jabbour, C. J. C., Mariano, E. B., & de Sousa Jabbour, A. B. L. (2018). Management theory and big data literature: From a review to a research agenda. *International Journal of Information Management, 43*, 112–129.

Di Vaio, A., Palladino, R., Pezzi, A., & Kalisz, D. E. (2021). The role of digital innovation in knowledge management systems: A systematic literature review. *Journal of Business Research, 123*, 220–231.

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.

Drucker, P. F. (1993). The rise of the knowledge society. *The Wilson Quarterly, 17*(2), 52–72.

Galliers, R. D., Leidner, D. E., & Simeonova, B. (Eds.). (2020). *Strategic information management: Theory and practice*. Routledge.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: Taxonomy and performance evaluation. *Knowledge and Information Systems, 46*(2), 241–284.

George, G., Haas, M. R., & Pentland, A. (2014). Big Data and management. *Academy of Management Journal, 57*(2), 321–326.

Ghasemaghaei, M. (2020). The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage. *International Journal of Information Management, 50*, 395–404.

Ghasemaghaei, M. (2021). Understanding the impact of big data on firm performance: The necessity of conceptually differentiating among big data characteristics. *International Journal of Information Management, 57*, 102055.

Ghasemaghaei, M., & Calic, G. (2019). Does big data enhance firm innovation competency? The mediating role of data-driven insights. *Journal of Business Research, 104*, 69–84.

Ghasemaghaei, M., & Calic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research, 108*, 147–162.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16*(1), 15–31.

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal, 17*(S2), 109–122.

Grover, V. (2020). Do we need to understand the world to know it? Knowledge in a big data world. *Journal of Global Information Technology Management, 23*(1), 1–4.

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems, 26*(3), 191–209.

Hamilton, R. H., & Sodeman, W. A. (2020). The questions we ask: Opportunities and challenges for using big data analytics to strategically manage human capital resources. *Business Horizons, 63*(1), 85–95.

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E., & Chiroma, H. (2016). The role of big data in smart city. *International Journal of Information Management, 36*(5), 748–758.

Henkel, J., & Hartmann, P. (2020). The rise of corporate science in AI: Data as a strategic resource. *Academy of Management Discoveries, 6*(3), 359–281.

Hopkins, M. S. (2010). The 4 ways IT is revolutionizing innovation. *MIT Sloan Management Review, 51*(3), 51.

Intezari, A., & Gressel, S. (2017). Information and reformation in KM systems: Big data and strategic decision-making. *Journal of Knowledge Management, 21*(1), 71–91.

Johnson, B., Lorenz, E., & Lundvall, B. Å. (2002). Why all this fuss about codified and tacit knowledge? *Industrial and Corporate Change, 11*(2), 245–262.

Johnson, J. S., Friend, S. B., & Lee, H. S. (2017). Big data facilitation, utilization, and monetiza-tion: Exploring the 3Vs in a new product development process. *Journal of Product Innovation Management, 34*(5), 640–658.

Kabir, N., & Carayannis, E. (2013, January). Big data, tacit knowledge and organizational competitiveness. In *Proceedings of the 10th international conference on intellectual capital, knowledge management and organisational learning: ICICKM* (p. 220).

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research–Moving away from the "What" towards the "Why." *International Journal of Information Management, 54*, 102205.

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science, 3*(3), 383–397.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*(70), 1.

Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management Review, 24*(4), 691–710.

Lapré, M. A., & Van Wassenhove, L. N. (2001). Creating and transferring knowledge for productivity improvement in factories. *Management Science, 47*(10), 1311–1325.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management, 36*(5), 700–710.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21–32.

Lee, B., Collier, P. M., & Cullen, J. (2007). Reflections on the use of case studies in the accounting, management and organizational disciplines. *Qualitative Research in Organizations and Management: An International Journal, 2*(3), 169–178.

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons, 60*(3), 293–303.

Leonard-Barton, D. (1995). *Building and sustaining the source of innovation*. Harvard Business School Press.

Machlup, F., & Mansfield, U. (Eds.). (1983). *The study of information: Interdisciplinary messages*. Wiley.

Nelson, R. R. (1991). Why do firms differ, and how does it matter? *Strategic Management Journal, 12*(S2), 61–74.

Nonaka, I. (1990). Redundant, overlapping organization: A Japanese approach to managing the innovation process. *California Management Review, 32*(3), 27–38.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14–37.

Nonaka, I., & Konno, N. (1998). The concept of "Ba": Building a foundation for knowledge creation. *California Management Review, 40*(3), 40–54.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.

Nonaka, I., & Toyama, R. (2015). The knowledge-creating theory revisited: Knowledge creation as a synthesizing process. In *The essentials of knowledge management* (pp. 95–110). Palgrave Macmillan.

Nonaka, I., Toyama, R., & Konno, N. (2000). SECI, Ba and leadership: A unified model of dynamic knowledge creation. *Long Range Planning, 33*(1), 5–34.

Nonaka, I., & Von Krogh, G. (2009). Perspective—Tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. *Organization Science, 20*(3), 635–652.

Nonaka, L., Takeuchi, H., & Umemoto, K. (1996). A theory of organizational knowledge creation. *International Journal of Technology Management, 11*(7–8), 833–845.

Pauleen, D. J., & Wang, W. Y. (2017). Does big data mean big knowledge? KM perspectives on big data and analytics. *Journal of Knowledge Management*. https://doi.org/10.1108/JKM-08-2016-0339

Polanyi, M. (1962). Tacit knowing: Its bearing on some problems of philosophy. *Reviews of Modern Physics, 34*(4), 601.

Pröllochs, N., & Feuerriegel, S. (2020). Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information & Management, 57*(1), 103070.

Quinn, J. B. (1992). *Intelligent enterprise: A knowledge and service based paradigm for Industry*. Simon and Schuster.
Ramus, T., Vaccaro, A., & Brusoni, S. (2017). Institutional complexity in turbulent times: Formalization, collaboration, and the emergence of blended logics. *Academy of Management Journal, 60*(4), 1253–1284.
Ranjan, J., & Foropon, C. (2021). Big data analytics in building the competitive intelligence of organizations. *International Journal of Information Management, 56*, 102231.
Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing & Management, 54*(5), 758–790.
Simsek, Z., Vaara, E., Paruchuri, S., Nadkarni, S., & Shaw, J. D. (2019). New ways of seeing Big Data. From the editors. *Academy of Management Journal, 62*(4), 971–978.
Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263–286.
Street, C. T., & Meister, D. B. (2004). Small business growth and internal transparency: The role of information systems. *MIS Quarterly, 28*(3), 473–506.
Sukumar, S. R., & Ferrell, R. K. (2013). 'Big Data' collaboration: Exploring, recording and sharing enterprise knowledge. *Information Services & Use, 33*(3–4), 257–270.
Sumbal, M. S., Tsui, E., & See-to, E. W. (2017). Interrelationship between big data and knowledge management: An exploratory study in the oil and gas sector. *Journal of Knowledge Management*. https://doi.org/10.1108/JKM-07-2016-0262
Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. *Information & Management, 57*(1), 103146.
Tang, V., Yanine, F., & Valenzuela, L. (2016). Data, information, knowledge and intelligence: The mega-nano hypothesis and its implications in innovation. *International Journal of Innovation Science, 8*(3), 199–216.
Urbinati, A., Bogers, M., Chiesa, V., & Frattini, F. (2019). Creating and capturing value from Big Data: A multiple-case study analysis of provider companies. *Technovation, 84*, 21–36.
Van de Ven, A. H., & Poole, M. S. (1995). Explaining development and change in organizations. *Academy of Management Review, 20*(3), 510–540.
Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research, 261*(2), 626–639.
Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77–84.
Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research, 70*, 287–299.
Weber, F., Lehmann, J., Graf-Vlachy, L., & König, A. (2019). Institution-infused sensemaking of discontinuous innovations: The case of the sharing economy. *Journal of Product Innovation Management, 36*(5), 632–660.
Wessel, M. (2016). You don't need Big Data—you need the right data. *Harvard Business Review*, *3*.
Yin, R. K. (2009). *Case study research: Design and methods* (Vol. 5). SAGE.
Yin, R. K. (2013). *Case study research: Design and methods*. London: SAGE.
Yoo, Y. (2015). It is not about size: A further thought on big data. *Journal of Information Technology, 30*(1), 63–65.
Zack, M. H. (1999). Developing a knowledge strategy. *California Management Review, 41*(3), 125–145.
Zaitsava, M., Marku, E., & Di Guardo, M. C. (2022). Is data-driven decision-making driven only by data? When cognition meets data. *European Management Journal*. https://doi.org/10.1016/j.emj.2022.01.003

Maryia Zaitsava is an Assistant Professor in Organization Studies at the University of Cagliari (Italy) and a visiting research fellow in Liverpool Business School (Liverpool John Moores University). Her research focus is in organizational challenges and opportunities of the early stage of digital transformation, such as data-driven decision-making first introduction, Big Data insights creation, and proof-of-concept tools and techniques.

Elona Marku is an Assistant Professor of Management at the Department of Economics and Business, University of Cagliari (Italy). She was a visiting scholar at Columbia Business School, Bayes Business School, and Warwick Business School. Her research focuses on the strategic management of technology and innovation and their implications for firm performance. Current projects also investigate the dynamics of digital transformation and the diffusion of emerging technologies.

Maria Chiara Di Guardo is Professor of Innovation Management and Organization Studies at the University of Cagliari. She is also Head of the Innovation and Entrepreneurship Centre (CREA-UniCa), and Director of the CLab-UniCa at the same university. Professor Di Guardo was visiting professor at Columbia Business School (Columbia University, NY) and Cass Business School (City, University of London), and visiting scholar at IESE Business School (Barcelona). Her research focuses on how to organize efficiently the innovation process as well as the relationship between innovation and entrepreneurship.

Azar Shahgholian is a senior lecturer in Digital Marketing at Liverpool Business School, Liverpool, UK. She holds a Ph.D. in Business and Management from The University of Manchester, UK. Her research interest includes Business Analytics, Information systems using Machine learning algorithms and big data analytics, and the role of social and knowledge networks in meeting of sustainability-related goals in project-focused organisations. She was an active member of LCR Activate; European Regional Development which provides hands-on support and funding to help Digital, Creative and Createch businesses grow using emerging technologies such as AI, Machine Learning, Virtual and Augmented Realities, Big & Open Data, High-Performance Computing and Cloud.

Authors and Affiliations

Maryia Zaitsava[1] · Elona Marku[1] · Maria Chiara Di Guardo[1] · Azar Shahgholian[2]

Elona Marku
elona.marku@unica.it

Maria Chiara Di Guardo
diguardo@unica.it

Azar Shahgholian
a.shahgholian@ljmu.ac.uk

[1] Department of Economics and Business, University of Cagliari, Via Sant' Ignazio 74, 09123 Cagliari, Italy

[2] Digital Marketing, Faculty of Business and Law, Liverpool John Moores University, 4 Rodney St, Liverpool L1 2TZ, UK