Bus Inf Syst Eng
https://doi.org/10.1007/s12599-026-00987-1

RESEARCH PAPER

Data Understanding for Data-Centric AI

Framework Development and Review of Current Methods

Joshua Holstein · Philipp Spitzer · Samuel Gensch · Marieke Hoell · Michael Vössing · Niklas Kühl

Received: 2 January 2025 / Accepted: 12 December 2025

Abstract Organizations collect growing volumes of data to extract value through analytics. However, this data growth creates challenges for effective data understanding, which forms the foundation for reliable decision-making and effective AI systems. Established analytics frameworks such as CRISP-DM and KDD acknowledge this importance but provide limited guidance to achieve this understanding, particularly for data-centric AI requiring collaboration across stakeholder groups. To address this gap, the authors conducted a systematic literature review, developing a five-dimensional framework for data understanding. They then performed a systematic mapping study analyzing how existing methods support these dimensions and accommodate different target audiences. The analysis reveals critical gaps in current methods, particularly in systematically supporting the understanding of data collection and contextualization. While most methods target data experts, the authors find a notable lack of methods supporting domain experts and decision-makers. This research advances both theoretical understanding by identifying the key dimensions that constitute data understanding and practical implementation by providing organizations with guidance on building data understanding.

Keywords Data understanding · Data analytics · Data-centric AI

Accepted after 2 revisions by Stefan Lessmann.

J. Holstein (✉) · P. Spitzer · S. Gensch · M. Hoell · M. Vössing
Karlsruhe Institute of Technology, Karlsruhe, Germany
e-mail: Joshua.Holstein@kit.edu

N. Kühl
University of Bayreuth, Bayreuth, Germany

1 Introduction

In today's digital economy, organizations continuously generate and store unprecedented volumes of data through their operations, customer interactions, and connected devices (Fassnacht et al. 2023). This growth coincides with significant advances in artificial intelligence (AI), with modern algorithms offering sophisticated ways to analyze complex data patterns and generate actionable insights (Lebovitz et al. 2021; Samtani et al. 2023). However, despite these parallel developments in data availability and algorithmic capabilities, many organizations struggle to successfully deploy AI applications and realize their promised value in practice. For example, IBM's Watson for Oncology project encountered significant challenges in providing consistent recommendations due to insufficient training data (Lohr 2021), while Amazon's recruitment system exhibited bias issues stemming from historical data (Villegas and Beachy 2021). These failures point to a fundamental challenge that lies not in the capabilities of AI models themselves but in the essential task of understanding and preparing the data that feeds these systems. As organizations work with more diverse data sources, they face increasing difficulties in developing comprehensive data understanding (Holstein et al. 2023). Yet, the growing recognition that data understanding forms the foundation for successful AI implementations has led to the emergence of data-centric AI (DCAI) – a paradigm that emphasizes systematic data engineering and understanding over algorithmic sophistication (Jakubik et al. 2024).

Traditional analytics frameworks such as CRISP-DM (Wirth and Hipp 2000) and KDD (Fayyad et al. 1996) recognize data understanding as an essential element of the analytics process. However, these frameworks typically treat it as an initial phase that precedes data preparation and

Published online: 06 February 2026

modeling, providing limited guidance on how to achieve and maintain this understanding throughout the analytics lifecycle. The DCAI paradigm reimagines this relationship by positioning data understanding as the foundation for systematic data engineering (Jakubik et al. 2024). This shift acknowledges that adequate data understanding requires the integration of multiple organizational perspectives. Data scientists must grasp technical characteristics for preparation and modeling, domain experts provide crucial business context and constraints, while decision-makers evaluate strategic relevance and implications (Lebovitz et al. 2021). This multi-stakeholder perspective suggests that systematic improvements in data quality, guided by a thorough understanding across organizational boundaries, are often more crucial for successful AI implementations than algorithmic refinements alone (Whang et al. 2023). As organizations increasingly introduce AI applications, this foundational role of integrated data understanding becomes critical for ensuring that analytics initiatives effectively support business objectives.

While DCAI's emphasis on data understanding represents a significant advancement in analytics thinking, translating these principles into practical organizational capabilities remains challenging (Whang et al. 2023). Organizations lack structured approaches for implementing DCAI's principles, particularly in environments where data understanding must be built and maintained across diverse stakeholder groups with varying technical expertise and domain knowledge (Gerhart et al. 2023; Holstein et al. 2023). Current methods and tools remain largely rooted in traditional analytics paradigms, offering fragmented support for different aspects of data understanding without providing an integrated perspective. This fragmentation becomes particularly problematic as organizations work with increasingly diverse and complex datasets, requiring a structured approach that can guide them in identifying, analyzing, and documenting relevant data characteristics (Jakubik et al. 2024). The absence of such a framework creates significant barriers where organizations struggle to establish consistent practices, stakeholders lack common ground for communication, and AI projects often fail due to poorly understood data sources. This gap between the theoretical recognition of data understanding's importance and the limited practical guidance leads to our first research question:

RQ1 *What are the dimensions of data understanding?*

The conceptualization of data understanding dimensions provides a theoretical foundation for the systematic investigation of this domain. However, the practical application of these dimensions requires an examination of existing methodological support. Prior research has introduced various methods and tools for data understanding, yet their coverage remains fragmented and unclear, particularly regarding which aspects of data understanding they cover. This leads to our second research question:

RQ2 *How do current methods and tools support different dimensions of data understanding?*

While the analysis of methodological support addresses technical aspects, effective data understanding requires integrating diverse stakeholder perspectives that each bring valuable contributions (van Giffen and Ludwig 2023; Dogan and Birant 2021; Park et al. 2021): Data scientists provide expertise on statistical properties and quality metrics, domain experts offer crucial insights by providing business rules and context, and decision-makers need to understand data's strategic implications without necessarily diving into technical details (Park et al. 2021). However, bridging these perspectives poses challenges as stakeholders use varying terminology, have different levels of technical knowledge, and focus on different aspects of the data (Gerhart et al. 2023; Lebovitz et al. 2021). This necessitates the investigation of support mechanisms across different user groups in our third research question:

RQ3 H*ow do current methods accommodate different target groups involved in data understanding?*

To investigate these research questions, we adopt a sequential research design centered on a framework-developing review (Rowe 2014). First, following the established methodology of Webster and Watson (2002), Gioia et al. (2013),and Wolfswinkel et al. (2013), we conduct a systematic literature review to develop a comprehensive framework that delineates the dimensions of data understanding. This systematic approach ensures that we capture and synthesize the currently fragmented perspectives on data understanding across different domains. Our analysis reveals five core dimensions: *Foundations*, *Collection and Selection*, *Contextualization and Integration*, *Exploration and Discovery*, and *Insights*. Building upon this synthesized foundation, we then conduct a systematic mapping study (Petersen et al. 2008) to analyze how existing methods identified in the literature cover these dimensions and accommodate different target groups. This analysis uncovers significant gaps in current methodological support, particularly in facilitating data Collection and Selection, Contextualization and Integration, with most methods focusing primarily on Exploration and Discovery. Furthermore, we find that existing methods predominantly target technical experts, while support for domain experts and decision-makers remains scarce, despite their critical role in analytics projects. Through this dual approach, we make several contributions: First, we provide a comprehensive framework that synthesizes the dimensions of data understanding, offering organizations structured guidance for improving their data analytics capabilities. Second, our systematic mapping of existing methods reveals gaps in current approaches, particularly in supporting data collection, contextualization, and integration. Finally, by identifying disparities in methodological support among different stakeholder groups, we provide direction for developing more inclusive tools that support diverse stakeholder needs, which are crucial for successful AI implementations. Together, these contributions advance our theoretical understanding and provide practical pathways for organizations to achieve appropriate data understanding for their AI initiatives.

2 Background and Related Work

Data understanding has emerged as a fundamental phase of data analytics projects, recognized as essential for extracting meaningful insights and driving successful outcomes. Comparative analyses have highlighted its integral role (Haertel et al. 2022), with data understanding appearing as a distinct phase in six of the seven major data science process models (Kutzias et al. 2023), including KDD (Fayyad et al. 1996), CRISP-DM (Wirth and Hipp 2000), and TDSP (Microsoft 2020). This widespread recognition stems from the crucial role of data understanding in enabling effective data utilization for analytical techniques and supporting reliable decision-making processes (Wirth and Hipp 2000; Janssen et al. 2017). However, traditional analytics frameworks have often treated data understanding superficially or fragmentarily (Haertel et al. 2022). The KDD process, for instance, disperses data understanding activities across multiple phases like "Creating a target dataset" and "Data preprocessing" (Fayyad et al. 1996), emphasizing technical preprocessing over comprehensive understanding. Similarly, while CRISP-DM explicitly includes a data understanding phase, its guidance remains limited to basic activities of collection, description, and quality verification (Wirth and Hipp 2000). This treatment overlooks crucial aspects like domain knowledge integration and real-world contextualization (Gerhart et al. 2023).

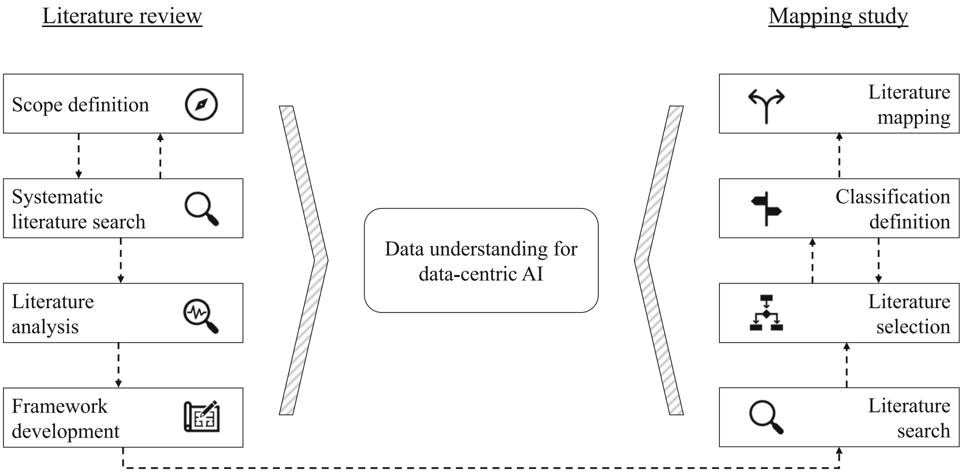
These limitations in traditional approaches to data understanding have become increasingly apparent with the emergence of DCAI, which represents a fundamental shift in how organizations approach data analytics and AI implementation (Jakubik et al. 2024). Unlike traditional model-centric approaches that focus on algorithmic refinement, DCAI emphasizes systematic design and engineering of data as the foundation for effective AI systems (Jarrahi et al. 2023). This paradigm shift promotes improving data quality and quantity while maintaining fixed model architectures, recognizing that appropriate data often drives performance improvements more effectively than model tuning (Jakubik et al. 2024). DCAI elevates the importance of comprehensive data understanding through its emphasis on domain-specific data augmentation, its recognition of data quality improvements as primary performance drivers, and its use of model performance metrics to indicate the effectiveness of data adjustments (Zhang et al. 2023).

The convergence of traditional analytics challenges and DCAI principles reveals significant gaps in current approaches to data understanding. While previous research has emphasized the need for a deeper understanding of how data represents real-world phenomena (Aaltonen et al. 2023) and the integration of domain expertise (Gerhart et al. 2023), existing frameworks provide insufficient guidance for achieving these goals. Comparative studies of analytics frameworks (Haertel et al. 2022; Fatima et al. 2020; Mariscal et al. 2010) have focused on overall framework comparison rather than an analysis of specific phases, such as data understanding. DCAI's emphasis on systematic data engineering amplifies these limitations, particularly the need for effective integration of domain knowledge when dealing with high-dimensional datasets (Jakubik et al. 2024; Jarrahi et al. 2023) and the value of exploratory analysis (Patel et al. 2023). These gaps in current research and practice motivate our development of a comprehensive framework that delineates the core dimensions of data understanding, considering both traditional business challenges and the novel requirements introduced by DCAI.

3 Research Design

To address our research questions, we employ a sequential research design combining a systematic literature review to develop a framework for data understanding (RQ1) followed by a systematic mapping study to analyze existing methods (RQ2 and RQ3). Given the distinct objectives of each phase, we employ complementary search strategies: the first phase requires a broad, exploratory approach to inductively derive theoretical dimensions, while the second phase demands a focused, systematic approach to assess methodological contributions against the established framework. This dual approach allows us to first establish a

Fig. 1 Dual research design: Integrating literature review-based framework development and systematic mapping study



theoretical foundation through rigorous literature analysis before systematically evaluating how existing methods align with this foundation (see Fig. 1).

3.1 Framework Development Through Systematic Literature Review

The first phase follows established guidelines for systematic literature reviews in information systems (Webster and Watson 2002), employing an inductive, grounded theory-inspired approach (Wolfswinkel et al. 2013) to identify the dimensions of data understanding. This inductive approach necessitates a broad search strategy to capture diverse theoretical perspectives, ensuring comprehensive coverage of the fragmented literature on data understanding. We adopt the recommendations of Gioia et al. (2013) to articulate our analysis results. We clarify the review scope using Cooper (1988)'s taxonomy, aiming to systematize research theories and methodologies while seeking to provide a neutral perspective and a literature review representative of the broad connections inherent in our topic.

Scope and Search Strategy. We began by reviewing an initial set of analytics frameworks to develop a shared understanding of data understanding and its role in analytics projects. This preliminary review helped us define our inclusion and exclusion criteria to focus our analysis on *understanding data*, which involves helping stakeholders comprehend data characteristics, provenance, quality, and real-world context, rather than *using data*, for example, by integrating external data sources for predictive modeling. Therefore, we included articles presenting frameworks for analytics (both conceptual papers and official documentation), articles offering interpretations, discussions, expansions, or comparisons of frameworks, and articles addressing challenges in data understanding within analytics. Conversely, we excluded articles focusing solely on automated technical methods rather than methodological guidance, articles not mentioning concepts related to data understanding, non-English language publications, and articles focusing solely on data preparation or cleaning without addressing understanding.

Our search covered three major databases: Web of Science, Scopus, and the AIS eLibrary. To ensure high-quality sources, we focused on premier outlets in information systems and related computer science disciplines. Our scope included all journals from the Senior Scholars' Basket of Journals, as well as selected A* and A-ranked journals according to CORE ranking that are specifically related to IS or data mining. Beyond the Senior Scholars' Basket, we specifically included IEEE Transactions on Knowledge and Data Engineering, Data Mining and Knowledge Discovery, Information Systems, ACM Transactions on Information Systems, and ACM Transactions on Database Systems. We also included major conference proceedings, including ECIS, ICIS, and HICSS, as we consider these outlets to be a representative sample for high-quality research in the discipline of data analytics frameworks in the fields of IS and computer science. Given the long history of data analytics and the early establishment of many standards, we included articles published between January 1995 and September 2023.

Search Process. Through several iterations, we developed our search string: ("data scien*" OR "data mining" OR "data analytics" OR "big data" OR "knowledge discovery" OR "data analysis") AND ("process model" OR "framework" OR "methodology") OR ("data understanding"). Our initial database search yielded 1340 articles. After removing duplicates, our sample decreased to 808 articles. Title and abstract screening following Snyder (2019) identified 40 articles for full-text review, which yielded 21 relevant papers. Through forward and backward searches (Webster and Watson 2002), we identified 17 additional publications, resulting in a final sample of 38 articles.

Analysis Process. For analysis, we followed a three-stage coding process (Wolfswinkel et al. 2013). In the first stage, three researchers independently conducted open coding on a representative subset of papers to identify concepts related to data understanding. In a collaborative workshop, the authors then synthesized their understanding of the underlying concepts. During axial coding, we established relationships between first-order concepts and developed second-order themes through collaborative workshops. One author refined the codes based on the established understanding, followed by a second workshop to finalize them. Finally, through selective coding, we aggregated themes into core dimensions, forming our framework for data understanding.

3.2 Method Analysis Through Systematic Mapping

Building on our framework, we conducted a systematic mapping study following Petersen et al. (2008) to analyze how existing methods cover the identified dimensions and support different target groups. A systematic mapping study aims to create an overview of a selected topic area by classifying identified papers into different predefined criteria. In contrast to the exploratory search strategy employed in the first phase, this phase requires a more focused and systematic approach, with search terms specifically designed to align with the conceptual dimensions established in the framework development phase.

Research Scope. The purpose of this mapping study was to provide an overview of methods that enable users to understand given datasets more holistically. We mapped the methods onto the dimensions identified in Phase 1 and analyzed their target groups to identify gaps in the current literature landscape.

Search Strategy. To identify relevant publications, we conducted a search across Web of Science, Scopus, and the AIS eLibrary, covering publications from January 2010 to May 2024. We maintained the same outlet scope as defined in our framework development phase. Our search term followed the identified dimensions from Phase 1 and included relevant synonyms: *"data understanding" OR "data exploration" OR "data integration" OR "data collection" OR "data acquisition" OR "data selection" OR "data infrastructure" OR "data overview" OR "data quality" OR "supplemental data" OR "surrogate data" OR "data visualization" OR "data summarization" OR "data provenance" OR "entity resolution" OR "schema matching" OR "schema mapping" OR ("data" AND ("human-in-the-loop" OR "domain knowledge" OR "domain expert*" OR "knowledge acquisition")).*

Selection Process. During the screening process, we included publications that propose methods facilitating the understanding of data in at least one dimension identified in the first phase of our research. We excluded publications focused solely on automated methods that do not directly contribute to user understanding, e.g., methods that automatically remove outliers. Following Snyder (2019), we identified 1714 articles after removing duplicates. Title and abstract screening reduced our sample to 167 relevant articles, of which we selected 48 during full-text screening. Through forward and backward searches, we identified nine additional publications, resulting in a final sample of 57 papers.

Classification Scheme. We developed two classification criteria: research focus and target groups. For research focus, we mapped methods to the five dimensions of data understanding identified in the first phase of our research, with each dimension consisting of three second-order themes that serve as our classification basis. Methods could address multiple dimensions simultaneously, as these dimensions are not mutually exclusive. For target groups, we employ a provisional coding procedure (Saldana 2021), which leverages predefined codes from existing research while allowing for iterative refinement. The initial coding uses a set of established roles from analytics projects (Saltz et al. 2018; Zhang et al. 2020), specifically: *method experts*, *domain experts*, and *decision-makers*. Method experts, such as data scientists or analysts, possess specialized knowledge in data analysis techniques and tools. Domain experts, while lacking formal training in data analysis, contribute deep knowledge of specific business domains, such as engineering or healthcare, which is crucial for contextualizing and validating data insights. Decision-makers, typically managers or executives, rely on data understanding to inform strategic choices but may not be directly involved in the technical analysis. Through our iterative coding procedure, we identified the necessity for an additional category: *General Users*. These are individuals who engage with data and visualizations in an informal manner, operating outside traditional project-based analytics structures, but may play a supporting role in identifying preliminary use cases or inspiring new directions for data applications.

Classification Process. One author initially classified the papers, and through collaborative workshops among the authors, we established a shared understanding. Based on this shared understanding, the author then refined the paper classifications to ensure consistency and accuracy. The dimensions were not treated as mutually exclusive; a single paper could be mapped to multiple dimensions if it addressed various aspects of data understanding. Through this systematic process, we were able to identify patterns in how existing methods support different aspects of data understanding and various user groups, revealing both the strengths and gaps in current methodological support.

4 Results

Following our sequential research design, we first present the framework for data understanding developed through our systematic literature review, followed by our analysis of existing methods based on the systematic mapping study.

4.1 Framework Development Through Systematic Literature Review

Our systematic literature review reveals five core dimensions that collectively constitute data understanding, which represent distinct but interrelated aspects of understanding data: Foundations, Collection and Selection, Contextualization and Integration, Exploration and Discovery, and Insights.

4.1.1 Foundations

The first dimension, Foundations, provides the essential groundwork for understanding data through three key themes (see Fig. 2): Infrastructure, Provenance, and Characterization and Familiarization.

Infrastructure emphasizes the importance of data warehousing models and tailored databases for data mining. Data warehousing models provide a structured environment for data storage, facilitating efficient data analysis (Dag et al. 2016). Similarly, databases designed for data mining, equipped with summary statistics, prepare data for in-depth analysis (SAS Institute Inc 2017). These infrastructure elements are crucial for organizing and setting a solid foundation for data exploration and understanding.

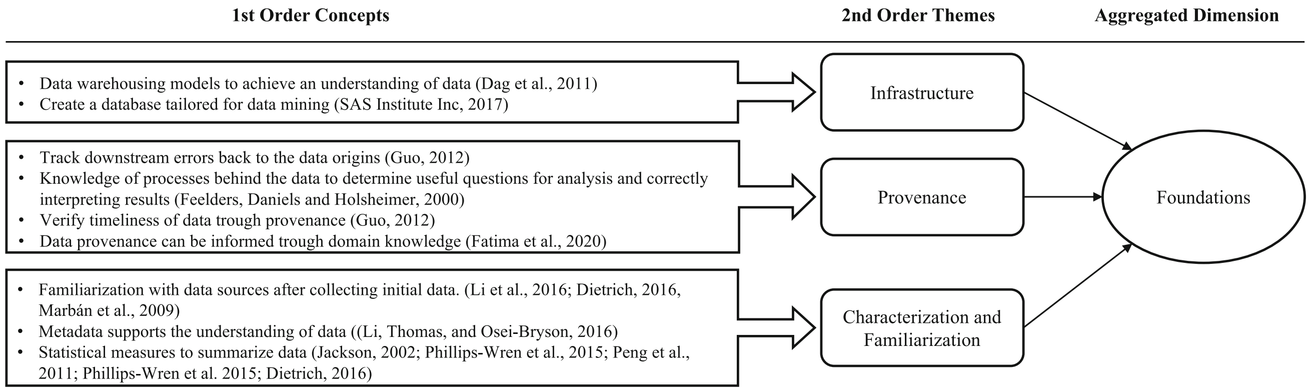
Provenance represents comprehending the data's origin, the applied transformations, and its timeliness. The ability to trace data provenance and document the transformation of data is critical to accurately interpreting results and preventing misinterpretations (Feelders et al. 2000). Tracing data provenance enables verification that the data is still up-to-date and relevant to the business problem at hand (Guo 2012). Additionally, it assures a transparent path for the data, enabling the tracing of potential downstream errors back to their origins to adjust and mitigate the root causes of these data errors. Understanding the collection processes and their respective transformations allows for formulating hypotheses that aid in later analyses and potentially creating new features.

Characterization and Familiarization, particularly through metadata, is crucial for understanding data meaning. Metadata provides detailed descriptions of data and its linkage to underlying business processes, bridging the gap between raw data and business application (Li et al. 2016). To complement metadata, statistical measures or tools can summarize data, offering insights into distribution patterns and underlying structures (Jackson 2002; Phillips-Wren et al. 2015). Examining data features and characteristics facilitates a detailed understanding of data segments and their contribution to the broader dataset (Peng et al. 2011). This involves assessing data granularity, aggregation levels, and value ranges of each source (Dietrich 2016).

4.1.2 Collection and Selection

The second dimension focuses on gathering and selecting relevant data through three themes (see Fig. 3): Data Collection, Selecting Relevant Data, and Supplemental Data.

Data Collection encompasses the acquisition of data and decisions regarding its selection and evaluation for further use. The initial data collection sets the foundation for subsequent stages of analysis (Haertel et al. 2022; Marbán et al. 2009). A profound knowledge of the data available both within and outside the organization is emphasized as crucial for effective data selection (Feelders et al. 2000). This knowledge aids in identifying gaps in the current data landscape and in making informed decisions about which



- Data warehousing models to achieve an understanding of data (Dag et al., 2011)
- Create a database tailored for data mining (SAS Institute Inc, 2017)

- Track downstream errors back to the data origins (Guo, 2012)
- Knowledge of processes behind the data to determine useful questions for analysis and correctly interpreting results (Feelders, Daniels and Holsheimer, 2000)
- Verify timeliness of data trough provenance (Guo, 2012)
- Data provenance can be informed trough domain knowledge (Fatima et al., 2020)

- Familiarization with data sources after collecting initial data. (Li et al., 2016; Dietrich, 2016, Marbán et al., 2009)
- Metadata supports the understanding of data ((Li, Thomas, and Osei-Bryson, 2016)
- Statistical measures to summarize data (Jackson, 2002; Phillips-Wren et al., 2015; Peng et al., 2011; Phillips-Wren et al. 2015; Dietrich, 2016)

Fig. 2 Dimensions, themes, and concepts of foundations

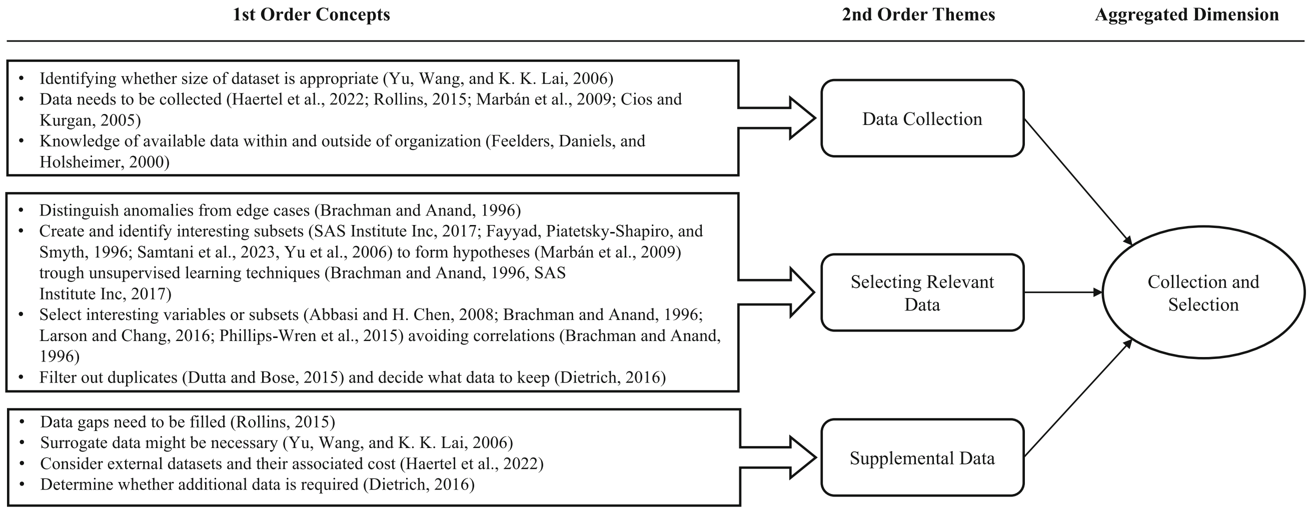

1st Order Concepts
2nd Order Themes
Aggregated Dimension
• Identifying whether size of dataset is appropriate (Yu, Wang, and K. K. Lai, 2006)
• Data needs to be collected (Haertel et al., 2022; Rollins, 2015; Marbán et al., 2009; Cios and Kurgan, 2005)
• Knowledge of available data within and outside of organization (Feelders, Daniels, and Holsheimer, 2000)
Data Collection
• Distinguish anomalies from edge cases (Brachman and Anand, 1996)
• Create and identify interesting subsets (SAS Institute Inc, 2017; Fayyad, Piatetsky-Shapiro, and Smyth, 1996; Samtani et al., 2023, Yu et al., 2006) to form hypotheses (Marbán et al., 2009) trough unsupervised learning techniques (Brachman and Anand, 1996, SAS Institute Inc, 2017)
• Select interesting variables or subsets (Abbasi and H. Chen, 2008; Brachman and Anand, 1996; Larson and Chang, 2016; Phillips-Wren et al., 2015) avoiding correlations (Brachman and Anand, 1996)
• Filter out duplicates (Dutta and Bose, 2015) and decide what data to keep (Dietrich, 2016)
Selecting Relevant Data
• Data gaps need to be filled (Rollins, 2015)
• Surrogate data might be necessary (Yu, Wang, and K. K. Lai, 2006)
• Consider external datasets and their associated cost (Haertel et al., 2022)
• Determine whether additional data is required (Dietrich, 2016)
Supplemental Data
Collection and Selection

Fig. 3 Dimensions, themes, and concepts of collection and selection

additional data sources might be beneficial for enriching the analysis.

Selecting Relevant Data becomes necessary when datasets are large and include instances or variables not relevant to the business use case. Discriminating between different subsets and identifying those worthy of more in-depth analysis is fundamental (Brachman and Anand 1996; Fayyad et al. 1996). This focused analysis enables the formulation of hypotheses based on insights obtained from critical subsets, effectively allocating resources to the most informative parts of the data. Different methods can be applied to identify potentially interesting subsets, including unsupervised learning techniques (SAS Institute Inc 2017). Data mining methods help discern relevant variables, differentiating between critical data for analysis and extraneous information. A crucial part involves distinguishing outliers from edge cases, where outliers may represent errors, while edge cases, though unusual, remain valid and relevant to the analysis.

Supplemental Data might need to be collected if not all required data is available. Often, the initial phase of data understanding reveals gaps where additional information is needed to align with project objectives. This might involve acquiring data not initially considered or delving deeper into specific areas (Rollins 2015; Dietrich 2016). In some cases, surrogate data becomes necessary (Yu et al. 2006), serving as alternative or proxy data when primary data is unavailable. Another approach involves considering external data sources and their associated costs, as understanding what external data can be leveraged and at what expense is pivotal for enriching internal datasets.

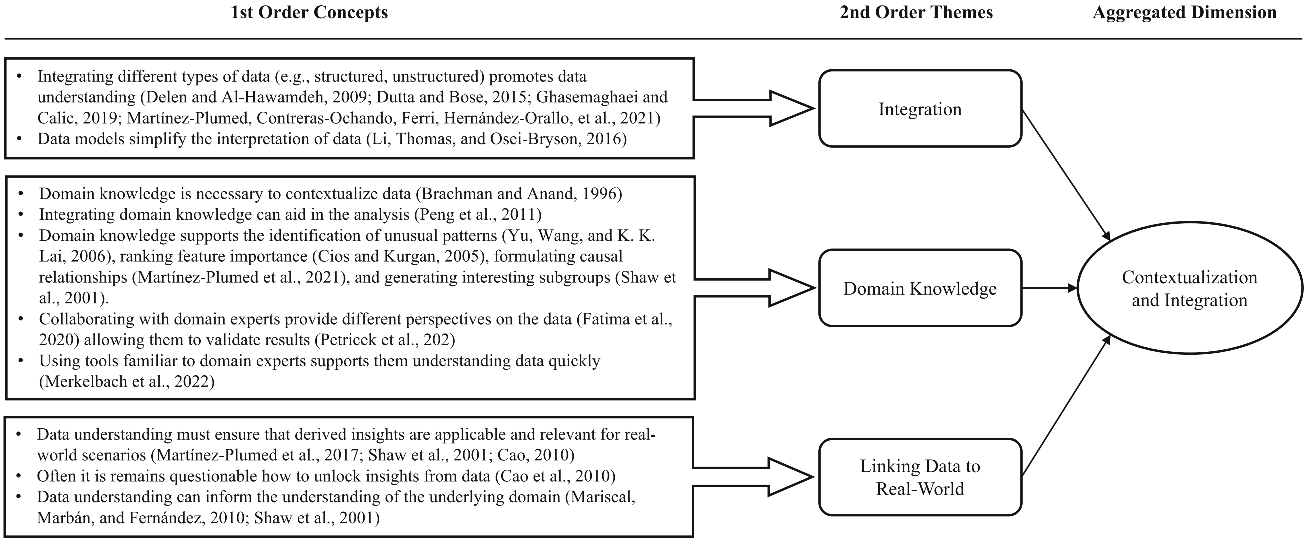
4.1.3 Contextualization and Integration

The third dimension involves contextualizing data through three themes: Integration, Domain Knowledge, and Linking Data to the Real World (see Fig. 4).

Integration of data sources allows the investigation of the interplay between various data sources and types to provide deeper insights. Incorporating different data sources, whether structured or unstructured, can provide a more comprehensive context for analysis. This process facilitates a holistic view that captures the multifaceted nature of data, leading to more informed and accurate insights (Delen and Al-Hawamdeh 2009; Martínez-Plumed et al. 2021). The utilization of preexisting data models enhances understanding by functioning as frameworks that classify and interpret diverse data types, simplifying intricate information architectures (Yu et al. 2006).

Domain Knowledge is critical to contextualize and ultimately make sense of data. Acquiring domain knowledge is crucial in comprehensively understanding data (Brachman and Anand 1996). Incorporating this knowledge can aid in analyzing data (Peng et al. 2011) by identifying uncommon patterns (Yu et al. 2006), ranking feature importance (Cios and Kurgan 2005), formulating causal relationships (Martínez-Plumed et al. 2021), and generating data subgroups (Shaw et al. 2001). Collaborative efforts beyond individual knowledge acquisition are essential, as domain experts provide different perspectives leading to a more holistic understanding (Fatima et al. 2020).

Linking Data to Real-World involves interpreting data and applying domain knowledge to address real-world complexities and challenges. This involves recognizing



- Integrating different types of data (e.g., structured, unstructured) promotes data understanding (Delen and Al-Hawamdeh, 2009; Dutta and Bose, 2015; Ghasemaghaei and Calic, 2019; Martínez-Plumed, Contreras-Ochando, Ferri, Hernández-Orallo, et al., 2021)
- Data models simplify the interpretation of data (Li, Thomas, and Osei-Bryson, 2016)

- Domain knowledge is necessary to contextualize data (Brachman and Anand, 1996)
- Integrating domain knowledge can aid in the analysis (Peng et al., 2011)
- Domain knowledge supports the identification of unusual patterns (Yu, Wang, and K. K. Lai, 2006), ranking feature importance (Cios and Kurgan, 2005), formulating causal relationships (Martínez-Plumed et al., 2021), and generating interesting subgroups (Shaw et al., 2001).
- Collaborating with domain experts provide different perspectives on the data (Fatima et al., 2020) allowing them to validate results (Petricek et al., 202)
- Using tools familiar to domain experts supports them understanding data quickly (Merkelbach et al., 2022)

- Data understanding must ensure that derived insights are applicable and relevant for real-world scenarios (Martínez-Plumed et al., 2017; Shaw et al., 2001; Cao, 2010)
- Often it is remains questionable how to unlock insights from data (Cao et al., 2010)
- Data understanding can inform the understanding of the underlying domain (Mariscal, Marbán, and Fernández, 2010; Shaw et al., 2001)

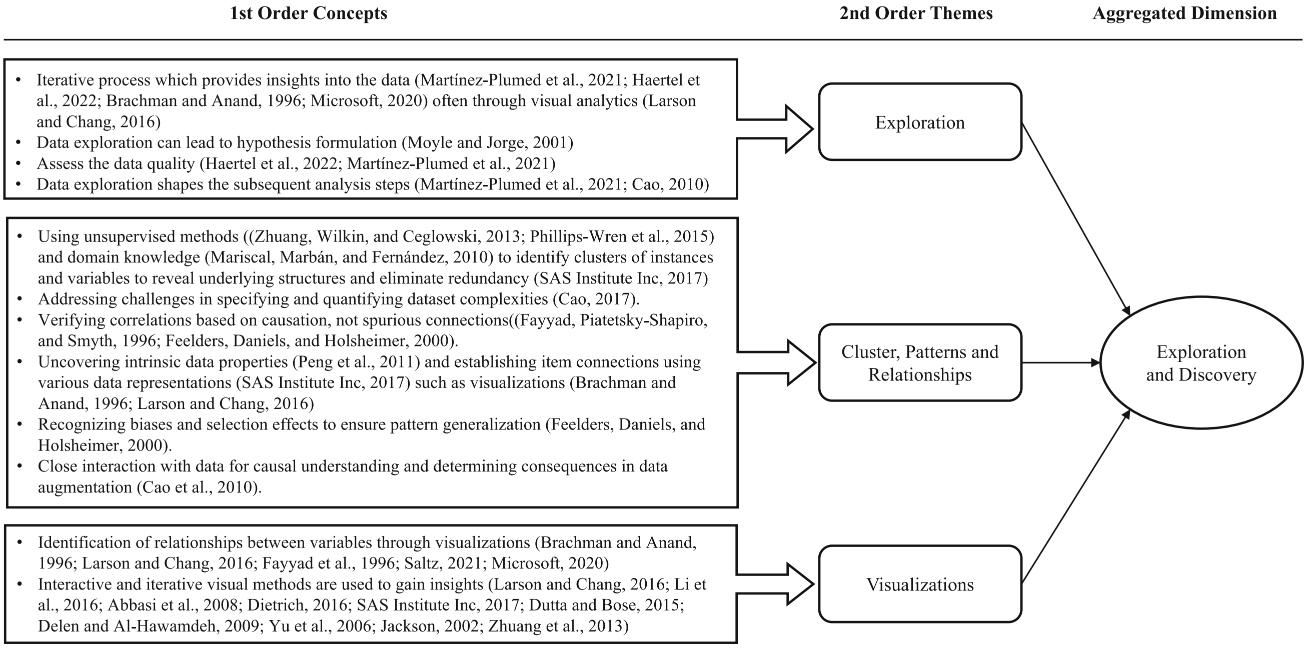
Fig. 4 Dimensions, themes, and concepts of contextualization and integration

that the complexity of real-world data necessitates thorough analysis to ensure its applicability and practical relevance (Cao et al. 2010). Through data understanding, one gains domain insights that inform and refine the analytical approach. Anticipating the context in which data will be used is critical, requiring activities to envision how insights will apply in real-world scenarios (Martínez-Plumed et al. 2017).

4.1.4 Exploration and Discovery

The fourth dimension encapsulates the critical stage of delving into data to uncover hidden patterns through three themes: Exploration, Cluster, Patterns and Relationships, and Visualizations (see Fig. 5).

Exploration of data helps understand and interpret it through iterative engagement, employing various techniques to uncover patterns, relationships, and insights that



- Iterative process which provides insights into the data (Martínez-Plumed et al., 2021; Haertel et al., 2022; Brachman and Anand, 1996; Microsoft, 2020) often through visual analytics (Larson and Chang, 2016)
- Data exploration can lead to hypothesis formulation (Moyle and Jorge, 2001)
- Assess the data quality (Haertel et al., 2022; Martínez-Plumed et al., 2021)
- Data exploration shapes the subsequent analysis steps (Martínez-Plumed et al., 2021; Cao, 2010)

- Using unsupervised methods ((Zhuang, Wilkin, and Ceglowski, 2013; Phillips-Wren et al., 2015) and domain knowledge (Mariscal, Marbán, and Fernández, 2010) to identify clusters of instances and variables to reveal underlying structures and eliminate redundancy (SAS Institute Inc, 2017)
- Addressing challenges in specifying and quantifying dataset complexities (Cao, 2017).
- Verifying correlations based on causation, not spurious connections((Fayyad, Piatetsky-Shapiro, and Smyth, 1996; Feelders, Daniels, and Holsheimer, 2000).
- Uncovering intrinsic data properties (Peng et al., 2011) and establishing item connections using various data representations (SAS Institute Inc, 2017) such as visualizations (Brachman and Anand, 1996; Larson and Chang, 2016)
- Recognizing biases and selection effects to ensure pattern generalization (Feelders, Daniels, and Holsheimer, 2000).
- Close interaction with data for causal understanding and determining consequences in data augmentation (Cao et al., 2010).

- Identification of relationships between variables through visualizations (Brachman and Anand, 1996; Larson and Chang, 2016; Fayyad et al., 1996; Saltz, 2021; Microsoft, 2020)
- Interactive and iterative visual methods are used to gain insights (Larson and Chang, 2016; Li et al., 2016; Abbasi et al., 2008; Dietrich, 2016; SAS Institute Inc, 2017; Dutta and Bose, 2015; Delen and Al-Hawamdeh, 2009; Yu et al., 2006; Jackson, 2002; Zhuang et al., 2013)

Fig. 5 Dimensions, themes, and concepts of exploration and discovery

inform the overall analysis (Larson and Chang 2016; Brachman and Anand 1996). Analysts examine data to identify trends and patterns, often suggesting new hypotheses about underlying relationships and phenomena. Data exploration also contributes to the development of data descriptions, quality reports, and understanding of how the data represents its context (Martínez-Plumed et al. 2017; Haertel et al. 2022). Further, it involves exploring data structure and encoding to inform how knowledge can be extracted from it (Cao 2017).

Clusters, Patterns, and Relationships include reducing data complexity by identifying clusters of instances and variables (SAS Institute Inc 2017), thereby revealing underlying structures and eliminating redundancy from correlated features. Analysts closely interact with the data, requiring a causal understanding to determine consequences in subsequent data augmentation (Martínez-Plumed et al. 2021) and to verify correlations, ensuring they are based on causation rather than spurious connections (Feelders et al. 2000). The process involves uncovering intrinsic data properties (Peng et al. 2011) and establishing connections between items using various data representations. Additionally, recognizing biases and selection effects is crucial to ensure the generalization of identified patterns (Feelders et al. 2000).

Visualizations are key to capturing the intricacies of data. They enable the extraction of insights and recognition of patterns. By exhibiting data points and their interrelationships, visualizations provide insights that may not be obtained through tables or summary statistics (Delen and Al-Hawamdeh 2009; SAS Institute Inc 2017). This is especially evident while exploring high-dimensional data, where coordinated visualizations reveal intricate data structures and distributions (Abbasi and Chen 2008; Dietrich 2016). Interactive and iterative visual methods are essential for exploratory data analysis, as they facilitate direct engagement with the data, enabling deeper examination of relationships between variables and identification of hidden insights (Larson and Chang 2016; Fayyad et al. 1996).

4.1.5 Insights

The final dimension focuses on the tangible outcomes obtained from analyzing the data to evaluate the Data Quality, note it in Deliverables, and inform Decision-Making (see Fig. 6).

Deliverables refer to the documentation produced during the data understanding phase. This typically includes an initial data collection report (Moyle and Jorge 2001), which outlines the specifics of the gathered data and provides a baseline for subsequent analyses. It is followed by data description and exploration reports (Haertel et al. 2022; Saltz 2021), which detail the intrinsic characteristics of the datasets and the insights captured. The data quality report (Moyle and Jorge 2001) complements this by assessing the data's reliability and appropriateness for analysis.

Data Quality describes the process of evaluating the integrity and usefulness of data. Essential tasks include verifying data quality and documenting issues (Saltz 2021), developing and implementing data quality metrics (Larson and Chang 2016), often informed by data profiling outcomes like demographics and descriptive statistics. Particularly with big data, challenges arise in maintaining accuracy and relevance due to the vastness of datasets (Martínez-Plumed et al. 2021). Enhancing data quality is crucial for modeling (Fatima et al. 2020) and involves evaluating the data's suitability for specific purposes.

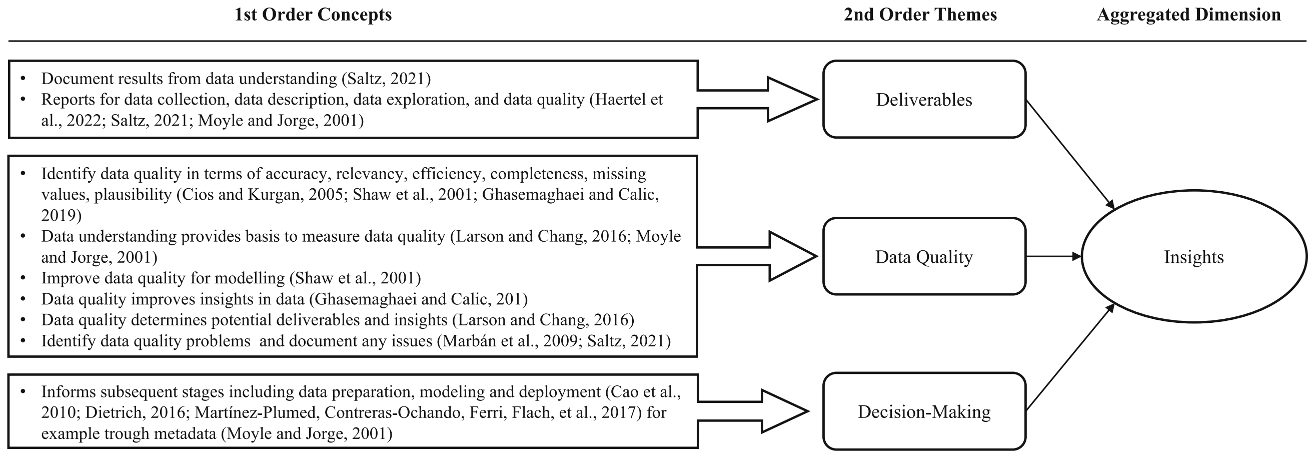
Decision-Making underscores the significance of in-depth data understanding for informed decision-making and effective application in later phases such as data preparation and modeling (Cao et al. 2010; Dietrich 2016). The phase yields critical outputs, such as metadata and data quality information (Moyle and Jorge 2001), which are integral to strategic decisions. This understanding ensures that insights derived from the data can effectively inform business decisions and strategy development.

4.1.6 Synthesis of Dimensions of Data Understanding

Data understanding serves as the critical bridge between the adjacent business understanding and data preparation phases. The five identified dimensions represent a logical order of interconnected activities, where each dimension informs and enhances the others through continuous feedback loops (see Fig. 7).

Starting with Foundations, analysts search in data infrastructures like data warehouses or data lakes for relevant data. While doing so, they check the provenance of the discovered data to ensure that the transformations applied are valid and do not hinder effective analysis concerning the underlying use cases. To get an overview of the available data sources, they familiarize themselves with and characterize the data through metadata and simple statistics.

The foundations then inform an iterative cycle that includes three central dimensions: Collection and Selection, Contextualization and Integration, and Exploration and Discovery. These interrelated elements form the core of the data understanding phase, each affecting and being affected by the others in a continuous loop of refinement and discovery. In Collection and Selection, the aim is to identify and gather relevant data sources informed by the initial foundational understanding and business requirements.. This helps to ensure that the data is comprehensive, relevant, and aligned with the analytical objectives.



- Document results from data understanding (Saltz, 2021)
- Reports for data collection, data description, data exploration, and data quality (Haertel et al., 2022; Saltz, 2021; Moyle and Jorge, 2001)

- Identify data quality in terms of accuracy, relevancy, efficiency, completeness, missing values, plausibility (Cios and Kurgan, 2005; Shaw et al., 2001; Ghasemaghaei and Calic, 2019)
- Data understanding provides basis to measure data quality (Larson and Chang, 2016; Moyle and Jorge, 2001)
- Improve data quality for modelling (Shaw et al., 2001)
- Data quality improves insights in data (Ghasemaghaei and Calic, 201)
- Data quality determines potential deliverables and insights (Larson and Chang, 2016)
- Identify data quality problems and document any issues (Marbán et al., 2009; Saltz, 2021)

- Informs subsequent stages including data preparation, modeling and deployment (Cao et al., 2010; Dietrich, 2016; Martínez-Plumed, Contreras-Ochando, Ferri, Flach, et al., 2017) for example trough metadata (Moyle and Jorge, 2001)

Fig. 6 Dimensions, themes, and concepts of insights

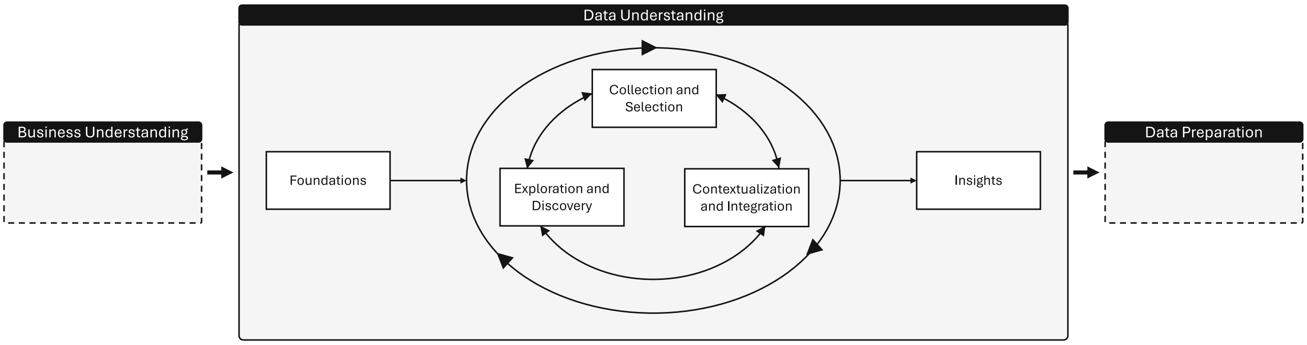


Fig. 7 Multiple interrelated dimensions together facilitate data understanding as part of the overall analytics process

Contextualization and Integration then embed these selected data elements within specific domain contexts. This is essential for creating meaningful and actionable data, utilizing domain expertise to interpret and effectively integrate the data. Subsequently, Exploration and Discovery involve thoroughly examining the data and utilizing advanced analytical techniques to reveal hidden patterns, relationships, and insights. This cycle is characterized by continuous interaction and feedback between its components. Findings obtained during exploration and discovery may reveal the need for additional data collection or indicate relationships that require domain expert interpretation. Similarly, as new data is acquired or the contextual landscape shifts, analysts must revisit exploration phases and recontextualize data to uncover additional insights. This dynamic interplay ensures that data understanding is not a linear process but rather an iterative cycle of refinement and discovery, where each phase continuously informs and enhances the others.

Ultimately, these activities result in Insights based on the acquired data understanding. Analysts must evaluate the data to determine its potential and decide whether to proceed with the project or abort it based on the quality of the data. They document their results in various reports, including data collection, description, exploration, and quality reports. If they decide to continue their project, they can use their understanding of the data to inform subsequent activities like data preparation and modeling and, ultimately, decision-making, thus leveraging the real-world value of the collected data. This final dimension serves not only as an endpoint but also as a potential trigger for revisiting earlier dimensions when new insights reveal gaps or opportunities in the understanding process.

4.2 Method Analysis Through Systematic Mapping

Our systematic mapping study reveals how existing methods cover different dimensions of data understanding and support various target audiences. We first present the mapping results organized by research focus, showing how methods address different dimensions, followed by an analysis of their target groups.

To illustrate how these methods practically support data understanding and demonstrate stakeholder collaboration, we examine an illustrative yet representative industrial scenario throughout the following analysis. Consider

GlobalTech Manufacturing, implementing a predictive quality control system for automotive component production that requires integrating IoT sensor data, quality inspection records, production schedules, maintenance logs, and supplier databases. This scenario involves data scientists (method experts), production engineers and quality inspectors (domain experts), plant managers (decision-makers), and line supervisors (general users), providing a realistic context for examining both the coverage patterns revealed in our mapping and the practical application of different methodological approaches across framework dimensions and stakeholder groups.

4.2.1 Coverage of Framework Dimensions

We map 57 research papers proposing methods for data understanding to the dimensions identified in the first phase of our research (see Fig. 8). Our analysis reveals that most methods focus on Exploration and Discovery (40), while fewer methods address Foundations (14), Collection and Selection (12), and Contextualization and Integration (12). In the Insights dimension, we identified 25 methods, though with varying emphasis across its themes. Table 1 in the Appendix provides the complete mapping of methods to dimensions and second-order themes, including specific citations for each category. Next, we describe how exemplary methods can be applied to generate an in-depth data understanding that serves as a basis for subsequent data preparation activities, and provide the number of mapped methods in brackets.

Foundations establishes the essential groundwork for understanding data by helping stakeholders comprehend how infrastructure (7), data provenance (3), and basic characteristics (4) influence data quality and interpretation. Method and domain experts must understand how existing systems shape their data, trace how data transformations affect reliability, and characterize datasets to assess their suitability for analytical purposes. Decision-makers require a foundational understanding of data infrastructure and provenance to assess whether existing data collection processes can support their strategic AI deployment objectives and identify potential systemic limitations that could affect business-critical applications.

These activities work synergistically to build comprehensive data understanding. Infrastructure analysis using platforms like the one of Scott et al. (2014) for heterogeneous data reveals how storage and processing architectures influence data availability and relationships. Provenance tracking through systems like the one of Huynh et al. (2018) allows to create provenance graphs that map data lineage to understand how collection methods and transformations affect reliability, while automated characterization tools like Voyager (Wongsuphasawat et al. 2016) enable stakeholders to explore fundamental data properties and assess completeness. Together, these approaches ensure that method and domain experts understand not just what data they have, but how system decisions, processing steps, and collection methods

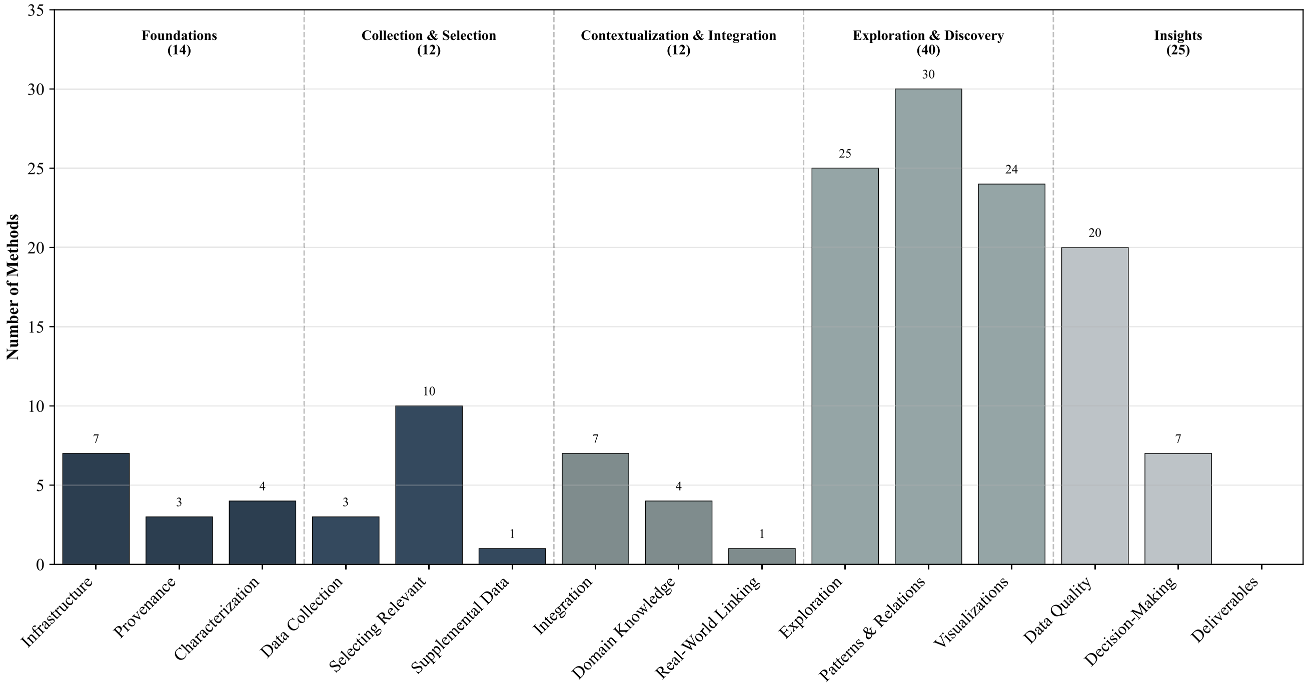


Fig. 8 Distribution of data understanding methods across framework dimensions

influence the data's meaning and reliability for their specific analytical goals.

At GlobalTech Manufacturing, building on business understanding requirements for predictive quality control established earlier, production engineers examine how the existing sensor network architecture influences data collection patterns, discovering that certain production areas have higher sensor coverage that might bias quality assessments toward detecting problems in well-monitored zones while missing those in others. Data scientists trace how raw sensor readings flow through calibration and aggregation processes, identifying that older sensors undergo additional smoothing that could mask important variation patterns needed for early failure detection. Quality inspectors use automated visualizations to understand how temperature and vibration measurements reflect actual equipment conditions, recognizing that sensor placement and calibration history create systematic gaps in coverage that could limit predictive system effectiveness in certain operational scenarios. This integrated understanding of infrastructure, provenance, and data characteristics provides the foundation for subsequent data collection and contextualization activities while establishing the groundwork for strategic decisions about where predictive AI systems can be reliably deployed.

Collection and Selection involves systematically collecting (3), selecting relevant (10), and identifying supplemental (1) data to ensure analytical efforts target the most informative and appropriate datasets while building a comprehensive understanding of data availability, relevance, and gaps. Method and domain experts must establish systematic data collection processes, identify which data subsets contain meaningful patterns for their analytical goals, and recognize where additional data sources are needed. Decision-makers require an understanding of data coverage and gaps to assess whether available data adequately represents the operational contexts where AI systems will be deployed and to make informed decisions about additional data investments needed to support strategic objectives.

These activities work synergistically to build comprehensive data coverage for analytical objectives. Continuous data collection platforms (Aydin and Anderson 2017) combine systematic data gathering with interactive monitoring that reveal data availability patterns across different systems and time periods to method experts. Interactive selection methods (Dimitriadou et al. 2016) enable domain experts to identify relevant data subsets by learning from feedback about which instances contain meaningful patterns, while goal-driven methods like the one of Liu and Yoon (2024) allow domain experts and decision-makers to specify analytical objectives in business language, automatically identifying gaps in available data and recommending supplemental sources. Together, these approaches ensure that stakeholders not only gather existing data systematically but also understand which subsets are most valuable and what additional data would enhance their analytical capabilities.

Building on the foundational understanding established previously, continuous collection platforms gather real-time sensor streams while providing production engineers with monitoring interfaces that reveal data flow patterns and coverage gaps across different production areas. Production engineers apply interactive selection methods to explore which time periods and equipment conditions produce the most informative data for quality prediction, learning from feedback to focus on sensor readings that correlate with actual defects. Quality engineers then apply goal-driven methods to specify objectives such as "predict equipment failures that impact product quality," which automatically reveals that successful implementations require vibration monitoring data, supplier quality metrics, and maintenance history records that their current dataset lacks. This systematic approach to collection, selection, and gap identification ensures comprehensive data coverage that aligns with both operational needs and analytical objectives, preparing the integrated datasets needed for later data preparation activities.

Contextualization and Integration involves understanding how heterogeneous data sources need to be integrated (7) within domain-specific contexts (4) to create comprehensive analytical datasets that are linked to real-world meaning and business objectives (1). Method experts and domain experts need to establish connections between data and physical processes, apply contextual knowledge to identify biases and constraints, and integrate heterogeneous data sources to support comprehensive analysis.

These activities work synergistically to ensure data reflects real-world complexity and business requirements. Mixed reality visualization methods (Mahfoud et al. 2018) enable domain experts to investigate data directly at physical locations where events occur, overlaying virtual data visualizations onto real environments to understand spatial relationships between data patterns and their physical sources. Bias identification methods (Cabrera et al. 2019) support domain experts in applying contextual knowledge to identify potential biases in analytical models and data, enabling targeted corrections that ensure results represent all operational conditions, while incident management frameworks (Peng et al. 2011) allow decision-makers to integrate heterogeneous data sources from multiple organizations and formats into unified datasets for decision support. Together, these approaches ensure that data understanding incorporates both technical characteristics and real-world context, bridging the gap between disparate data sources and actionable insights for critical decision-making scenarios.

Building on the systematic data collection, quality inspectors at GlobalTech Manufacturing apply, for example, mixed reality visualization to overlay sensor data directly onto production equipment, immediately identifying which specific machines generate unusual readings and understanding spatial relationships between temperature sensors, vibration monitors, and actual component quality issues. Production engineers then apply bias identification methods to examine whether their quality prediction models adequately represent all operational conditions, using their knowledge of seasonal variations and equipment aging patterns to identify periods where data might be systematically biased toward certain failure modes. Data scientists apply incident management integration methods to combine the selected sensor data with maintenance records and supplier databases into unified datasets for comprehensive quality analysis, enabling coordinated understanding when quality issues arise across multiple production systems. This progression from spatial understanding through bias correction to systematic integration ensures that the resulting datasets capture both technical measurements and operational realities, establishing contextualized data ready for subsequent exploration and pattern discovery activities.

Exploration and Discovery involves systematically exploring datasets (25) to reveal clusters, patterns, and relationships (30) that inform analytical understanding and hypothesis formation. Method experts and domain experts must iteratively explore data through visualizations (24), examine individual data instances to form mental models about underlying processes, and systematically identify clusters and relationships that reveal meaningful patterns for later modeling activities.

These activities work synergistically to enable comprehensive pattern discovery across different levels of data granularity. Collaborative visualization methods (Ben Lahmar and Herschel 2021) provide method experts with recommendation systems that suggest relevant queries and visualizations based on successful exploration patterns from multiple analysts, enabling efficient visual data discovery through content-based and collaborative filtering approaches. Instance-based exploration methods (Saghafi et al. 2022) allow domain experts to examine data as individual instances with unique properties rather than being constrained by predefined schemas, enabling them to form mental models and discover unexpected patterns without requiring deep technical knowledge of data structures. Systematic pattern identification methods (Nestorov et al. 2019) support method experts in conducting iterative preparation, visualization, and analysis stages to understand data characteristics and reveal underlying relationships between different data groups. Together, these approaches ensure that exploration progresses from collaborative visual discovery through individual investigation to systematic pattern analysis, enabling stakeholders to build a comprehensive understanding that bridges technical analysis with domain expertise.

Building on the contextualized and integrated datasets, quality inspectors apply collaborative visualization methods that recommend relevant data views based on successful quality investigations by other inspectors, enabling them to quickly identify promising visualizations for examining sensor patterns and production metrics without extensive technical expertise. Production engineers then apply instance-based exploration methods to investigate individual sensor readings and production events, allowing them to form mental models about relationships between temperature variations, vibration patterns, and equipment performance across different machine types without being constrained by predefined database structures. Data scientists utilize systematic pattern identification methods to explore discovered relationships through iterative preparation and analysis stages, identifying significant differences between high-quality and defective production patterns to reveal underlying timing relationships and attribute variations that inform predictive maintenance strategies. This progression from collaborative visual discovery through flexible exploration to systematic pattern analysis enables each stakeholder to contribute their domain expertise while building a comprehensive understanding that supports both operational insights and prepares detailed data descriptions and quality assessments that support decision-making.

Insights involves consolidating data understanding activities into actionable outcomes through data quality assessments (20), deliverables (0), and support for decision-making (7) that enable stakeholders to make informed choices about data usability and analytical strategies. Method experts must systematically evaluate data quality and reliability while creating documentation that communicates data characteristics and limitations. Decision-makers require summaries of data scope, coverage gaps, and quality limitations to make decisions about project viability, whether to proceed with current data, invest in additional data collection, or terminate projects where data limitations cannot be addressed. This enables decision-makers to assess whether available data represent the operational contexts where AI systems will be deployed and to determine boundaries for automated versus human decision-making.

These activities help transform data understanding into organizational value through evaluation and informed decision-making. Data quality assessment methods (Zhang et al. 2019) enable method experts to examine data characteristics and reveal inconsistencies, missing values, and reliability issues that affect trustworthiness. Error identification methods (Sluban et al. 2014) support method experts in ranking data points by their likelihood of being erroneous while providing evaluation interfaces that help distinguish between genuine patterns and data artifacts. Statistical reliability assessment methods (Heinrich et al. 2018) allow decision-makers to calculate the probability that datasets are free of internal contradictions, enabling informed decisions about project viability and data usability based on measures of data consistency. Together, these approaches ensure that data understanding translates into assessments of data quality, documentation of findings, and evidence-based decisions about whether to proceed with analytics projects, invest in additional data collection, or pursue alternative approaches.

Data scientists apply systematic quality assessment methods to examine sensor data and production records, identifying reliability issues such as sensor drift, calibration problems, and missing measurements while documenting these findings in data quality reports that inform strategic decisions. Quality inspectors use error identification methods to distinguish between genuine equipment problems that require maintenance attention and measurement errors that should be filtered from analysis, ensuring that operational decisions focus on actual equipment issues rather than data artifacts. Plant managers utilize statistical reliability assessments and quality documentation to make decisions on project viability, determining whether the available data foundation is sufficient to proceed, whether additional sensor installations and data collection efforts are justified, or whether alternative quality control approaches should be pursued instead. This progression from technical quality evaluation through operational error identification to strategic viability assessment enables each stakeholder group to contribute its expertise while making risk-aware decisions about project continuation, additional investments, and deployment boundaries that translate comprehensive data understanding into measurable organizational value.

4.2.2 *Support for Different Stakeholders*

Our analysis of target groups reveals that existing methods predominantly target method experts, while support for other stakeholders varies significantly. Of the 57 methods analyzed (see Fig. 9), 41 are designed for method experts, while only three target decision-makers, ten support domain experts, and ten address other stakeholders. Table 2 in the Appendix provides a comprehensive classification of methods by target audience, including the specific approaches designed for each stakeholder group.

Method Experts represent the primary target group, with 41 methods identified. These approaches encompass a broad spectrum of dimensions relevant to data understanding, offering capabilities for complex data modeling, analysis, and visualization. Despite the comprehensive coverage and advanced functionalities of these methods, as exemplified by technologies like Luo et al. (2024), Po and Sorrentino (2011), and Chung et al. (2020), they often require considerable setup and integration effort. This presents a notable barrier to immediate deployment (Liu and Li 2018; Lin et al. 2020; Schäfer and Leser 2022). Addressing these challenges through streamlined configuration processes is crucial for enabling method experts to efficiently tackle complex domain-specific challenges.

Domain Experts benefit from ten specific methods. Among these, two methods are particularly designed for the specialized task of discrimination discovery in databases (Ruggieri et al. 2010; Cabrera et al. 2019). Other methods emphasize usability and intuitiveness, particularly designed for domain experts to facilitate domain-specific applications (Cabrera et al. 2019; Saghafi et al. 2022; Lafon et al. 2013). As with most target groups, the majority of these methods significantly contribute to the dimensions of Exploration, Clusters, Patterns, and Relationships, and Visualization, enabling Domain Experts to gain deeper insights into their data (Dimitriadou et al. 2016; Vellido et al. 2013; Lafon et al. 2013).

Decision-Makers form a particularly underserved group, with only three methods designed for their needs. The first integrates data sources for analytics and decision-making (Liu and Yoon 2024) while the second supports marketing managers through sentiment analysis for brand-related decisions (Pournarakis et al. 2017). The third provides an incident response framework covering data integration, analytics, and decision-making (Peng et al. 2011).

General Users encompasses methods supporting individuals who engage with data and visualizations outside traditional analytics roles. This group is served by ten methods aimed at democratizing data understanding through intuitive interfaces, no-code and low-code systems, as well as automated support. Among these, three methods facilitate data exploration by suggesting contextually relevant search queries (Sellam and Kersten 2016; Ben Lahmar and Herschel 2021; Eirinaki et al. 2014), addressing the needs of users who may lack proficiency in formal query languages. Additional methods enhance the data interaction and exploration experience through automated visualization recommendations and generation capabilities

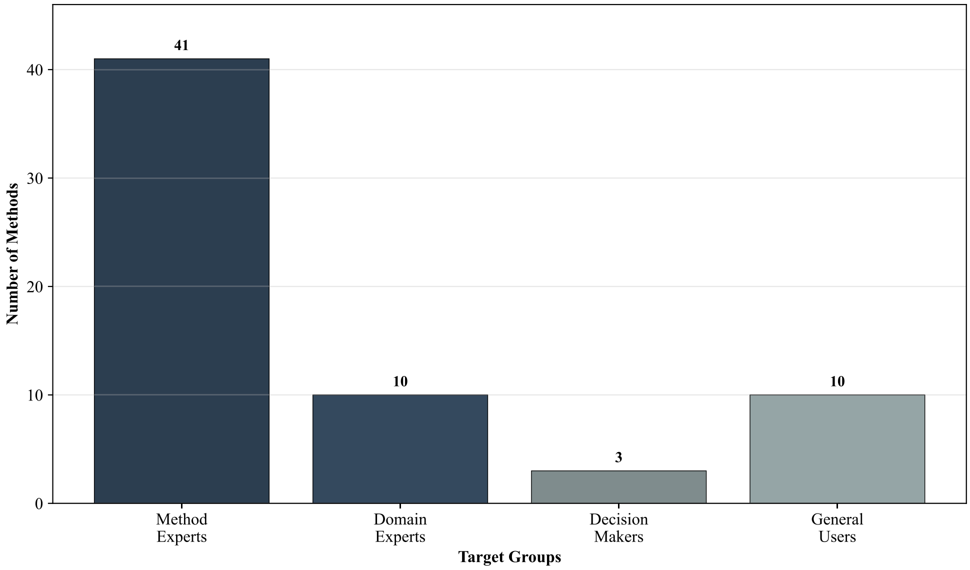


Fig. 9 Distribution of data understanding methods across target audiences

(El et al. 2020; Dibia and Demiralp 2019). These solutions lower technical barriers to data engagement, enabling general users to derive meaningful insights despite operating outside traditional project-based analytics structures. This democratization of data access and understanding supports their role in identifying preliminary use cases and inspiring novel applications of data analytics.

4.2.3 Synthesis of Method Analysis

Our mapping study of 57 methods reveals clear patterns in how existing approaches support data understanding. The majority of methods (40) focus on the Exploration and Discovery dimension, particularly addressing visualization, pattern identification, and clustering capabilities. The remaining dimensions receive less methodological support, with only 12–14 methods each addressing Foundations, Collection and Selection, and Contextualization and Integration. Within the Insights dimension, addressed by 25 methods, there is a strong emphasis on Data Quality assessment (20 methods), while no methods specifically focus on creating Deliverables. Regarding target groups, the methods show a clear focus on method experts, with 41 methods designed for this group. Domain experts and other stakeholders are each supported by ten methods, primarily focusing on making data exploration and visualization more accessible. Decision-makers represent the smallest target group with only three dedicated methods. Most methods that support non-method experts concentrate on making complex analytical tasks more accessible through intuitive interfaces and automated guidance features. This summary of our findings provides the foundation for a deeper discussion of the implications and future directions in the following section.

5 Discussion

In this study, we develop a framework for data understanding through a systematic literature review and analyze existing methodological support through a systematic mapping study. This sequential approach enables us to establish theoretical foundations for data understanding while evaluating practical support. Our findings reveal critical insights about the nature of data understanding and gaps in current approaches, particularly relevant to the paradigm of DCAI, where systematic data understanding and engineering form the foundation for successful AI implementation (Jakubik et al. 2024). Our research reveals four critical findings about the current state of data understanding in analytics projects.

Five Interconnected Dimensions for Data Understanding. The first finding centers on the identification and characterization of five core dimensions that collectively constitute data understanding: Foundations, Collection and Selection, Contextualization and Integration, Exploration and Discovery, and Insights. These dimensions extend beyond the simplified data understanding phase described in traditional frameworks (Wirth and Hipp 2000; Fayyad et al. 1996; Microsoft 2020) by operating not in isolation but as an interconnected system where each element informs and enhances the others. This interconnected nature particularly aligns with DCAI's emphasis on systematic data engineering (Jakubik et al. 2024), where comprehensive data understanding forms a foundation for data quality improvements and systematic data engineering decisions. The dynamic interplay between dimensions challenges the linear, sequential approach suggested by traditional analytics frameworks like CRISP-DM (Wirth and Hipp 2000) and KDD (Fayyad et al. 1996), instead supporting DCAI's iterative focus on data quality and understanding. The framework's iterative approach supports DCAI's focus on building a comprehensive understanding of data characteristics and quality to systematically engineer the data (Jakubik et al. 2024), which ultimately may yield improved model performance.

Methodological Gaps in Supporting DCAI. Our second key finding emerges from systematically mapping existing methods to these dimensions, revealing significant disparities in methodological support, particularly relevant to DCAI implementation. While Exploration and Discovery receives substantial attention with 40 identified methods, other dimensions crucial for DCAI, like Collection and Selection (12 methods) as well as Contextualization and Integration (12 methods), remain underserved. This imbalance is particularly problematic for DCAI initiatives, where systematic data engineering requires robust support across all dimensions (Jakubik et al. 2024; Whang et al. 2023). The limited number of methods supporting domain knowledge integration (only four methods) and real-world context linking (one method) poses significant challenges for DCAI implementation, where understanding the data's real-world implications and integrating domain expertise is essential for developing reliable AI systems. This methodological gap may explain why many organizations struggle to implement DCAI principles effectively despite recognizing their importance (Whang et al. 2023). Organizations often focus heavily on statistical data properties and sophisticated modeling techniques, achieving promising results during development. However, without proper contextualization and domain knowledge integration, i.e., establishing the relationship between the data and its real-world meaning (Aaltonen et al. 2023), these models may fail to recognize important real-world patterns or constraints. For instance, automated methods might remove certain data points as statistical outliers during data preparation when these actually represent valid edge cases crucial for the application domain. Similarly, without methods supporting real-world context linking, organizations might overlook important seasonal patterns, regulatory requirements, or business rules that affect model behavior in production environments. This disconnect between statistical optimization and real-world applicability may lead to models that perform well in controlled development settings but fail to realize their expected value when deployed in practice.

Bias Toward Method Experts. Our third key finding relates to the target groups served by existing methods, revealing a notable bias toward method experts that particularly challenges DCAI implementation. Our mapping shows that most methods (41 of 57) are designed specifically for method experts, while domain experts (10 methods) and decision-makers (3 methods) receive substantially less support. This limited methodological support creates systematic barriers to essential collaborations. With only 10 methods supporting domain experts, organizations lack tools that enable domain experts to contribute their contextual knowledge without requiring deep technical expertise (Holstein et al. 2023). For instance, domain experts can identify critical data gaps during Collection and Selection, but without specialized methods enabling them to investigate data characteristics without requiring technical expertise, this capability remains largely untapped (Park et al. 2021). The situation is even more acute for decision-makers (3 methods), who need a high-level understanding of data scope, coverage boundaries, and quality limitations to make strategic deployment decisions rather than detailed technical insights (Janssen et al. 2017). This strategic understanding enables them to assess project viability, determine appropriate investment levels in data collection, and establish where AI systems can be trusted to operate autonomously versus where human oversight remains necessary. Yet current methodological support predominantly serves only one part of this triad, potentially explaining why organizations struggle to translate DCAI principles into practice (Whang et al. 2023). The bias in current methods may reinforce silos between technical and business stakeholders, making it difficult to establish the collaborative understanding necessary for effective data-driven decision-making.

Multi-Stakeholder Collaboration as Foundation for Data Understanding. Our fourth key finding reveals that successful data understanding fundamentally depends on effective collaboration between stakeholder groups, each contributing distinct but complementary perspectives throughout the different data understanding phases (Gerhart et al. 2023; Lebovitz et al. 2021; Park et al. 2021). The dimensions of our framework reveal distinct patterns of stakeholder interaction. In Foundations, method experts lead infrastructure analysis and provenance tracking, but depend on domain experts to validate whether characterized data actually represents the business processes they claim to capture (Gerhart et al. 2023). Decision-makers require an understanding of data infrastructure and provenance to assess whether existing data collection processes can support their strategic AI deployment objectives and to identify potential systemic limitations that could affect business-critical applications. In Collection and Selection, all stakeholder groups contribute complementary expertise. Domain experts identify which data sources contain business-relevant information and recognize gaps based on their operational knowledge, while method experts contribute technical insights about sensor coverage, data availability in existing systems, and opportunities for creating derived measurements such as virtual sensors (Martin et al. 2021). Decision-makers utilize this understanding of data coverage and gaps to inform decisions about whether the data adequately represents the operational contexts where AI systems will be deployed, and to determine what additional data investments may be necessary to support strategic objectives. However, neither perspective alone suffices for comprehensive data collection. Similarly, Contextualization and Integration requires close collaboration, where domain experts provide contextual knowledge that links data to real-world processes while method experts translate this knowledge into technical integration rules (Gerhart et al. 2023). When this collaboration fails, organizations risk creating technically sound but practically ineffective datasets. In Exploration and Discovery, method experts create visualizations and apply analytical techniques, but depend on domain experts to interpret patterns and distinguish genuine business phenomena from data artifacts (Holstein et al. 2023; Lebovitz et al. 2021). Finally, Insights brings all stakeholders together, where method experts assess technical data quality, domain experts evaluate business representativeness, and decision-makers determine project viability and establish boundaries for automated versus human decision-making based on data limitations. This three-way interaction of method experts applying technical skills, domain experts providing context, and decision-makers ensuring business alignment is crucial for DCAI initiatives.

5.1 Implications for Theory and Practice

Our research advances the theoretical understanding of data understanding in the context of data-centric AI while providing practical insights for implementation. From a theoretical perspective, our framework extends current knowledge by synthesizing five interconnected dimensions of data understanding, challenging the traditional view presented in analytics frameworks like CRISP-DM (Wirth and Hipp 2000) and KDD (Fayyad et al. 1996). Rather than treating data understanding as merely an initial phase, our framework positions it as a dynamic, iterative process fundamental to successful AI implementation. This theoretical reconceptualization aligns with DCAI's emphasis on systematic data engineering and provides the structured guidance that organizations currently lack for improving their data analytics capabilities. The framework's emphasis on integration between technical exploration and domain contextualization advances our theoretical understanding of how organizations can build the comprehensive data understanding that DCAI requires (Whang et al. 2023; Polyzotis et al. 2018). Furthermore, our identification of the interconnected nature of these dimensions contributes to theory by highlighting how data understanding emerges through continuous interaction between different aspects of data work, rather than through distinct, sequential phases.

Our systematic mapping of existing methods contributes to theory by revealing specific challenges in translating DCAI principles into practice. Prior work by Gerhart et al. (2023) has highlighted how data scientists and domain experts often lack a common language for discussing data characteristics, leading to misunderstandings and inefficient collaboration in AI development. Our findings extend this insight by showing how current methodological support primarily focuses on technical data exploration while providing limited support for bridging these communication gaps for the transfer of domain knowledge relevant to the curation of data (Park et al. 2021). Some recent work has begun addressing this challenge – for instance, Holstein et al. (2023) propose methods for creating a shared understanding of feature meanings between technical and domain experts. However, our systematic mapping reveals that such approaches remain rare, with most existing methods emphasizing technical expertise over domain knowledge integration. This theoretical insight helps explain why organizations continue to struggle with implementing DCAI principles effectively, despite recognizing their importance. The gap between technical and domain perspectives is particularly problematic for DCAI initiatives, where systematic data engineering requires deep integration of domain knowledge into technical processes. Our framework provides a theoretical foundation for understanding these challenges and suggests the need for new approaches that better support knowledge integration across stakeholder groups.

The research also makes theoretical contributions to the emerging field of DCAI by providing a foundation for understanding how organizations can systematically improve data quality through enhanced understanding. Our framework suggests that data understanding needs to account for both the technical aspects of data analysis and the human factors of knowledge integration and contextualization (Gerhart et al. 2023; Lebovitz et al. 2021). This theoretical perspective challenges the predominant focus on algorithmic sophistication and suggests that successful AI implementation requires equal attention to the organizational processes that enable comprehensive data understanding. The framework's emphasis on continuous interaction between different dimensions also contributes to the theoretical understanding of how organizations can maintain and evolve their data understanding as their AI capabilities mature.

The empirical findings also yield practical contributions for organizational DCAI implementation by demonstrating the criticality of holistic data understanding approaches. The results emphasize that successful implementations require systematic engagement across all five dimensions, going beyond the prevalent focus on technical exploration alone. This might involve combining multiple tools or developing custom approaches to ensure comprehensive coverage of data understanding activities. Given the limited availability of tools supporting domain experts and decision-makers, organizations may need to develop custom interfaces or collaboration mechanisms that enable effective knowledge sharing between technical and non-technical stakeholders. Furthermore, organizations should establish systematic processes for data contextualization and integration, addressing the gaps in current methodological support. This could involve creating structured workflows for documenting data provenance, capturing domain knowledge, and linking data to real-world contexts.

5.2 Limitations and Future Research

Our development of a theoretical framework for data understanding and systematic mapping of existing methods, while comprehensive, reveals limitations that point to promising directions for future research.

First, our analysis draws primarily from published literature, which may not capture all practical approaches and informal workarounds currently used in industry. Simultaneously, our focus on literature that explicitly addresses frameworks for data understanding may overlook valuable practices embedded within domain-specific use cases, where data is applied for analytical purposes, and data understanding activities occur as secondary rather than primary research elements. Similarly, our approach may not have captured domain-specific extensions of foundational frameworks like CRISP-DM and KDD developed for scientific or other specialized contexts. This limitation opens valuable opportunities to validate and extend our findings through empirical studies of DCAI implementations across different organizational contexts. Such research could reveal how organizations overcome the methodological gaps we identified and develop effective practices for integrating diverse stakeholder perspectives, particularly focusing on how they facilitate knowledge exchange between method experts and domain specialists.

Second, a limitation stems from our treatment of AI in DCAI as relatively uniform, not fully accounting for how different AI paradigms might require distinct approaches to data understanding. For instance, generative AI models may need different types of data understanding compared to traditional supervised learning approaches, while few-shot or self-supervised learning techniques might introduce entirely new requirements for understanding training data. For example, in few-shot learning, understanding the representativeness of the few examples becomes crucial, as these samples must effectively capture the key variations within a class. This limitation suggests valuable opportunities for research examining how data understanding needs vary across different AI paradigms and data types. Future studies could investigate how organizations adapt their data understanding practices for different types of AI models, potentially leading to more nuanced frameworks that account for these variations.

Third, it is important to acknowledge that data understanding in contemporary practice also extends to broader societal and ethical dimensions. Understanding data not only requires technical and analytical clarity but also sensitivity to potential biases, fairness concerns, ethical implications, and legal compliance issues. Data may reflect or even reinforce existing social inequalities, and overlooking such risks can lead to unintended harm when insights are operationalized. Similarly, regulatory requirements such as data protection laws (European Commission 2023), impose constraints that must be considered when handling and interpreting data. Moreover, AI ethics research emphasizes the importance of addressing issues of fairness, accountability, and transparency when working with data (Chandrabose et al. 2021). While our approach does not assess these aspects, we emphasize that future research should address these aspects to incorporate them into a holistic view, ensuring responsible and trustworthy data-driven decision-making.

Finally, a limitation arises from our framework's temporal focus, which captures data understanding at a specific point in time, while the relevance of different dimensions may evolve throughout an AI project's lifecycle. For instance, the Collection and Selection dimension might be most critical during initial project phases, while Contextualization and Integration becomes increasingly important during feature engineering and model development. Similarly, the Exploration and Discovery dimension might be particularly crucial during data preparation and model debugging. Even within the Foundations dimension, the emphasis might shift from infrastructure setup to provenance tracking as projects mature. This temporal and phase-dependent variation of our dimensions suggests valuable opportunities for longitudinal research examining how data understanding practices evolve both across project lifecycles and alongside advancing AI capabilities. Future studies could investigate how organizations prioritize and balance different dimensions of data understanding across project stages, how they maintain and update their understanding as systems evolve in production, and how insights from one dimension inform and influence others throughout the project lifecycle.

6 Conclusion

As organizations increasingly adopt AI systems, our investigation of data understanding reveals both theoretical insights and practical implementation challenges. Through our sequential approach, we have established a theoretical framework delineating five core dimensions of data understanding and analyzed the current state of methodological support for these dimensions. Our analysis shows that current methods fall short in two key areas: supporting non-technical users and connecting data to real-world business contexts. These limitations are particularly significant in data-centric AI projects, which rely on a deep understanding of data for effective refinement for AI performance improvements. While existing methods provide sophisticated support for data exploration and discovery, they often fail to address the full spectrum of activities and stakeholders involved in modern analytics and AI development. The framework we present offers a foundation for future research in this area as organizations increasingly rely on data-driven decision-making and data-centric AI. Addressing the identified gaps in methodological support will be crucial for ensuring that data understanding can effectively support organizational objectives, ultimately leading to more effective data analytics practices across diverse stakeholder groups.

Funding Open Access funding enabled and organized by Projekt DEAL.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12599-026-00987-1.

References

Aaltonen A, Alaimo C, Parmiggiani E, Stelmaszak M, Jarvenpaa S, Kallinikos J, Monteiro E (2023) What is missing from research on data in information systems? Insights from the inaugural workshop on data research, Commun AIS, p 53

Abbasi A, Chen H (2008) CyberGate: a design framework and system for text analysis of computer-mediated communication. MIS Q 32:811–837

Aydin A, Anderson K (2017) Batch to real-time: Incremental data collection & analytics platform. In: Proceedings of the 50th Hawaii international conference on system sciences, pp 5911–5920

Ben Lahmar H, Herschel M (2021) Collaborative filtering over evolution provenance data for interactive visual data exploration. Inf Syst 95:101620

Brachman RJ, Anand T (1996) The process of knowledge discovery in databases, p 37–57

Cabrera AA, Epperson W, Hohman F, Kahng M, Morgenstern J, Chau DH (2019) FAIRVIS: visual analytics for discovering intersectional bias in machine learning. In: IEEE conference on visual analytics science and technology (VAST), IEEE, pp 46–56

Cao L (2017) Data science: a comprehensive overview. ACM Comput Surv 50(3):1–42

Cao L, Zhao Y, Zhang H, Luo D, Zhang C, Park E (2010) Flexible frameworks for actionable knowledge discovery. IEEE Trans Knowl Data Eng 22(9):1299–1312

Chandrabose A, Chakravarthi BR, et al. (2021) An overview of fairness in data–illuminating the bias in data pipeline. In: Proceedings of the first workshop on language technology for equality, diversity and inclusion, pp 34–45

Chung Y, Kraska T, Polyzotis N, Tae KH, Whang SE (2020) Automated data slicing for model validation: a big data - AI integration approach. IEEE Trans Knowl Data Eng 32(12):2284–2296

Cios KJ, Kurgan LA (2005) Trends in data mining and knowledge discovery. In: Advanced techniques in knowledge discovery and data mining, Springer, pp 1–26

European Commission (2023) A European approach to artificial intelligence – Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Cooper H (1988) Organizing knowledge syntheses: a taxonomy of literature reviews. Knowl Soc 1:104–126

Dag A, Topuz K, Oztekin A, Bulur S, Megahed FM (2016) A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. Decis Support Syst 86:1–12

Delen D, Al-Hawamdeh S (2009) A holistic framework for knowledge discovery and management. Commun ACM 52(6):141–145

Dibia V, Demiralp C (2019) Data2Vis: automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. IEEE Comput Graph Appl 39(5):33–46

Dietrich D (2016) Data analytics lifecycle processes. EMC Corp. US patent, No. US9262493B1

Dimitriadou K, Papaemmanouil O, Diao Y (2016) AIDE: an active learning-based approach for interactive data exploration. IEEE Trans Knowl Data Eng 28(11):2842–2856

Dogan A, Birant D (2021) Machine learning and data mining in manufacturing. Expert Syst Appl 166(114):060

Eirinaki M, Abraham S, Polyzotis N, Shaikh N (2014) QueRIE: collaborative database exploration. IEEE Trans Knowl Data Eng 26(7):1778–1790

El OB, Milo T, Somech A (2020) Towards autonomous, hands-free data exploration. In: Conference on innovative data systems research

Fassnacht MK, Benz C, Leimstoll J, Satzger G (2023) Is your organization ready to share? A framework of beneficial conditions for data sharing. In: ICIS 2023 Proceedings

Fatima F, Talib R, Hanif MK, Awais M (2020) A paradigm-shifting from domain-driven data mining frameworks to process-based domain-driven data mining-actionable knowledge discovery framework. IEEE Access 8:210763–210774

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun ACM 39(11):27–34

Feelders A, Daniels H, Holsheimer M (2000) Methodological and practical aspects of data mining. Inf Manag 37(5):271–281

Gerhart N, Torres R, Giddens L (2023) Challenges in the model development process: discussions with data scientists. Commun AIS 53(1):591–611

Gioia D, Corley K, Hamilton A (2013) Seeking qualitative rigor in inductive research. Org Res Meth 16:15–31

Guo PJ (2012) Software tools to facilitate research programming. Stanford University

Haertel C, Pohl M, Nahhas A, Staegemann D, Turowski K (2022) Toward a lifecycle for data science: a literature review of data science process models. In: PACIS 2022 Proceedings

Heinrich B, Klier M, Schiller A, Wagner G (2018) Assessing data quality – A probability-based metric for semantic consistency. Decis Support Syst 110:95–106

Holstein J, Schemmer M, Jakubik J, Vössing M, Satzger G (2023) Sanitizing data for analysis: designing systems for data understanding. Electron Markets 33(1):52

Huynh TD, Ebden M, Fischer J, Roberts S, Moreau L (2018) Provenance network analytics. Data Mining Knowl Discov 32(3):708–735

Jackson J (2002) Data mining; a conceptual overview. Commun AIS 8(1):19

Jakubik J, Vössing M, Kühl N, Walk J, Satzger G (2024) Data-centric artificial intelligence. Bus Inf Syst Eng

Janssen M, Van Der Voort H, Wahyudi A (2017) Factors influencing big data decision-making quality. J Bus Res 70:338–345

Jarrahi MH, Memariani A, Guha S (2023) The principles of data-centric AI. Commun ACM 66(8):84–92

Kutzias D, Dukino C, Kötter F, Kett H (2023) Comparative analysis of process models for data science projects. In: Proceedings of the 15th international conference on agents and artificial intelligence, Scitepress, pp 1052–1062

Lafon S, Bouali F, Guinot C, Venturini G (2013) On studying a 3D user interface for OLAP. Data Mining Knowl Discov 27(1):4–21

Larson D, Chang V (2016) A review and future direction of agile, business intelligence, analytics and data science. Int J Inf Manag 36(5):700–710

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? the dangers of training and evaluating AI tools based on experts' know-what. MIS Q 45:1501–1526

Li Y, Thomas MA, Osei-Bryson KM (2016) A snail shell process model for knowledge discovery via data analytics. Decis Support Syst 91:1–12

Lin Y, Wang H, Li J, Gao H (2020) Efficient entity resolution on heterogeneous records. IEEE Trans Knowl Data Eng 32(5):912–926

Liu X, Li XB (2018) Customer data acquisition with predictive analytics. In: ICIS 2018 proceedings

Liu D, Yoon VY (2024) Developing a goal-driven data integration framework for effective data analytics. Decis Support Syst 180:114197

Lohr S (2021) What ever happened to IBM's Watson? (published 2021) – nytimes.com. https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html. Accessed 12 Nov 2024

Luo B, Li X, Liu X, Guo J, Ren Y, Ma S, Ma J (2024) D2MTS: enabling dependable data collection with multiple crowdsourcers trust sharing in mobile crowdsensing. IEEE Trans Knowl Data Eng 36(3):927–942

Mahfoud E, Wegba K, Li Y, Han H, Lu A (2018) Immersive visualization for abnormal detection in heterogeneous data for on-site decision making. In: Proceedings of the 51st Hawaii international conference on system sciences, pp 1300–1309

Marbán O, Segovia J, Menasalvas E, Fernández-Baizán C (2009) Toward data mining engineering: a software engineering approach. Inf Syst 34:87–107

Mariscal G, Marbán O, Fernández C (2010) A survey of data mining and knowledge discovery process models and methodologies. Knowl Eng Rev 25:137–166

Martin D, Kühl N, Satzger G (2021) Virtual sensors. Bus Inf Syst Eng 63(3):315–323

Martínez-Plumed F, Contreras-Ochando L, Ferri C, Flach P, Hernández-Orallo J, Kull M, Lachiche N, Ramírez-Quintana MJ (2017) CASP-DM: context aware standard process for data mining. arXiv

Martínez-Plumed F, Contreras-Ochando L, Ferri C, Hernández-Orallo J, Kull M, Lachiche N, Ramírez-Quintana MJ, Flach P (2021) CRISP-DM twenty years later: from data mining processes to data science trajectories. IEEE Trans Knowl Data Eng 33(8):3048–3061

Microsoft (2020) Data acquisition and understanding of team data science process. https://learn.microsoft.com/azure/architecture/data-science-process/lifecycle

Moyle S, Jorge A (2001) Ramsys – a methodology for supporting rapid remote collaborative data mining projects. In: ECML/PKDD01 workshop: integrating aspects of data mining, decision support and meta-learning (IDDM), vol 64

Nestorov S, Jukić B, Jukić N, Sharma A, Rossi S (2019) Generating insights through data preparation, visualization, and analysis: framework for combining clustering and data visualization techniques for low-cardinality sequential data. Decis Support Syst 125:113119

Park S, Wang AY, Kawas B, Liao QV, Piorkowski D, Danilevsky M (2021) Facilitating knowledge sharing from domain experts to data scientists for building nlp models. In: Proceedings of the 26th international conference on intelligent user interfaces, Association for Computing Machinery, New York, NY, USA, IUI '21, p 585–596

Patel H, Guttula S, Gupta N, Hans S, Mittal RS, Lokesh N (2023) A data centric AI framework for automating exploratory data analysis and data quality tasks. J Data Inf Qual 15(4):1–26

Peng Y, Zhang Y, Tang Y, Li S (2011) An incident information management framework based on data integration, data mining, and multi-criteria decision making. Decis Support Syst 51(2):316–327

Petersen K, Feldt R, Mujtaba S, Mattsson M (2008) Systematic mapping studies in software engineering. In: International conference on evaluation and assessment in software engineering (EASE)

Phillips-Wren G, Iyer LS, Kulkarni U, Ariyachandra T (2015) Business analytics in the context of big data: a roadmap for research. Commun AIS p 23

Po L, Sorrentino S (2011) Automatic generation of probabilistic relationships for improving schema matching. Inf Syst 36(2):192–208

Polyzotis N, Roy S, Whang SE, Zinkevich M (2018) Data lifecycle challenges in production machine learning: a survey. ACM SIGMOD Rec 47(2):17–28

Pournarakis DE, Sotiropoulos DN, Giaglis GM (2017) A computational model for mining consumer perceptions in social media. Decis Support Syst 93:98–110

Rollins JB (2015) Foundational methodology for data science. Technical report, IBM

Rowe F (2014) What literature review is not: diversity, boundaries and recommendations. Europ J Inf Syst 23(3):241–255

Ruggieri S, Pedreschi D, Turini F (2010) Data mining for discrimination discovery. ACM Trans Knowl Discov Data 4(2):1–40

Saghafi A, Wand Y, Parsons J (2022) Skipping class: improving human-driven data exploration and querying through instances. Europ J Inf Syst 31(4):463–491

Saldana J (2021) The coding manual for qualitative researchers. SAGE

Saltz JS (2021) CRISP-DM for data science: Strengths, weaknesses and potential next steps. In: IEEE international conference on big data, pp 2337–2344

Saltz J, Armour F, Sharda R (2018) Data science roles and the types of data science programs. Commun Assoc Inf Syst 43(1):33

Samtani S, Zhu H, Padmanabhan B, Chai Y, Chen H, Nunamaker JF (2023) Deep learning for information systems research. J Manag Inf Syst 40(1):271–301

SAS Institute Inc (2017) https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

Schäfer P, Leser U (2022) Motiflets. Proceedings of the VLDB Endowment 16(4):725–737

Scott M, Boardman RP, Reed PA, Cox SJ (2014) Managing heterogeneous datasets. Inf Syst 44:34–53

Sellam T, Kersten M (2016) Cluster-driven navigation of the query space. IEEE Trans Knowl Data Eng 28(5):1118–1131

Shaw MJ, Subramaniam C, Tan GW, Welge ME (2001) Knowledge management and data mining for marketing. Decis Support Syst 31(1):127–137

Sluban B, Gamberger D, Lavrač N (2014) Ensemble-based noise detection: noise ranking and visual performance evaluation. Data Mining Knowl Discov 28(2):265–303

Snyder H (2019) Literature review as a research methodology: an overview and guidelines. J Bus Res 104:333–339

van Giffen B, Ludwig H (2023) How siemens democratized artificial intelligence. MIS Q Exec 22(1):3

Vellido A, García DL, Nebot A (2013) Cartogram visualization for nonlinear manifold learning models. Data Mining Knowl Discov 27(1):22–54

Villegas A, Beachy S (2021) The Amazon that customers don't see (published 2021) – nytimes.com. https://www.nytimes.com/interactive/2021/06/15/us/amazon-workers.html. Accessed 12 Npv 2024

Webster J, Watson RT (2002) Analyzing the past to prepare for the future: writing a literature review. MIS Q 26(2):xiii–xxiii

Whang SE, Roh Y, Song H, Lee JG (2023) Data collection and quality challenges in deep learning: a data-centric AI perspective. VLDB J 32(4):791–813

Wirth R, Hipp J (2000) CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining

Wolfswinkel J, Furtmueller E, Wilderom C (2013) Using grounded theory as a method for rigorously reviewing literature. Europ J Inf Syst 22(1):45–55

Wongsuphasawat K, Moritz D, Anand A, Mackinlay J, Howe B, Heer J (2016) Voyager: exploratory analysis via faceted browsing of visualization recommendations. IEEE Trans Visual Comput Graph 22(1):649–658

Yu L, Wang S, Lai KK (2006) An integrated data preparation scheme for neural network data analysis. IEEE Trans Knowl Data Eng 18(2):217–230

Zhang AX, Muller M, Wang D (2020) How do data science workers collaborate? Roles, workflows, and tools. Proc ACM on Human Comput Interact 4(CSCW1):1–23

Zhang R, Indulska M, Sadiq S (2019) Discovering data quality problems: the case of repurposed data. Bus Inf Syst Eng 61(5):575–593

Zhang T, Feng H, Chen W, Chen Z, Zheng W, Luo X, Huang W, Tung A (2023) ChartNavigator: an interactive pattern identification and annotation framework for charts. IEEE Trans Knowl Data Eng 35(2):1258–1269