



Unstructured data research in business: Toward a structured approach

Evert de Haan^{a,*}, Manjunath Padigar^b, Siham El Kihal^c, Raoul Kübler^d, Jaap E. Wieringa^a

^a University of Groningen, Nettelbosje 2, 9747 AE Groningen, the Netherlands

^b Macquarie University, 4 Eastern Road, NSW 2109, Sydney, Australia

^c WU Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

^d ESSEC Business School Paris, 3, Avenue Bernard Hirsch – 95021 Cergy-Pontoise Cedex, Paris, France

ARTICLE INFO

Keywords:

Unstructured data
Strategic framework
Organizational learning theory

ABSTRACT

Despite the unprecedented growth in both the volume of unstructured data (UD) and the associated methodological sophistication, there is a growing managerial need for a structured view of how to select data sources and methods given a specific use case or scenario. Handling UD is typically resource intensive, requires many steps, and involves high uncertainty, but UD can contain rich information not found in structured data. Recognizing the gap in clear guidelines for leveraging UD in managerial decision-making, we develop a systematic three-step approach: (1) problem identification, (2) solutions development, and (3) problem resolution. Building on organizational learning theory, we propose a solutions development framework with four conceptually distinct uses of UD based on two dimensions: organizational learning goals (exploration and exploitation) and environmental scanning scope (internal and external data sources). Finally, we discuss implications for practitioners and outline key focus areas for future research directions.

1. Introduction

The last decade has witnessed substantial growth in unstructured data (UD) – such as texts and images (Balducci and Marinova, 2018). These data are now readily available to marketing decision-makers and represent a valuable asset for advancing their understanding of the market. Managers are thus well advised to leverage these new data sources (Boegershausen et al., 2023) to improve their customers' experience. To achieve this, companies need to gather, process, and analyze the available UD to gain (consumer) insights and ultimately improve managerial decision-making.

With UD becoming increasingly abundant, the last decade also witnessed a substantial rise of new tools and methods to derive nuanced insights from UD. An associated research stream in the marketing literature has swiftly grown while trying to keep up with the pace of methodological developments, commonly originating from neighboring fields such as computer science, information systems, and computational linguistics. Tools range from traditional econometric methods to machine learning and generative AI (GenAI).

This fast-paced evolution, however, is both a boon and a bane. While research continues to develop, evaluate, and benchmark new and

improved methodologies (see e.g., Berger et al., 2020, Kübler, Colicev, and Pauwels, 2020, Hartmann et al., 2023, Humphreys and Wang 2018, or Büschken and Allenby 2016), it has become increasingly challenging for managers to cope with such rapid developments. When using UD to address a problem, managers are not only required to keep abreast with the methodological horse race but also struggle with identifying the appropriate UD research approach, depending on their goal and environmental scope. Moreover, while one may be tempted to always use state-of-the-art methods, managers must also be mindful of the return on such investments and understand how to choose the right method and data for their given purpose. Unfortunately, clear managerial guidelines on how to assess the costs and benefits of using specific data and applying UD methods is lacking.

The tension between the potential benefits and the challenges that are associated with extracting business value from UD was voiced by the five data science managers whom we asked for input for this study. These data science managers work in industries ranging from big tech to providers of shopper-focused solutions. Across the board, they indicate that roughly only 20 % of their team utilizes UD, and they acknowledge that this signifies unexplored gains and opportunities. For example, one of the managers states “For unstructured data, I think, we are not yet

* Corresponding author.

E-mail addresses: evert.de.haan@rug.nl (E. de Haan), manjunath.padigar@mq.edu.au (M. Padigar), siham.el.kihal@wu.ac.at (S. El Kihal), kubler@essec.edu (R. Kübler), j.e.wieringa@rug.nl (J.E. Wieringa).

<https://doi.org/10.1016/j.jbusres.2024.114655>

Received 31 January 2023; Received in revised form 29 March 2024; Accepted 4 April 2024

Available online 9 April 2024

0148-2963/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

there. They are really promising [...] but there are some hurdles. First of all, [...] problems are not clearly defined and then there is the technology.” The challenges outlined above are in line with the findings of IDC (2023), where the lack of experience with UD and the difficulty in quantifying the return on investment of UD projects are listed among the biggest roadblocks that prevent leveraging the potential of UD. Indeed, Davenport (2019) shows that only 18 % of organizations report being able to take advantage of UD.

Following a decade of UD research that initially focused on identifying data sources and use cases, and then transitioned to the identification, development, and assessment of methods to analyze UD, this paper aims at encouraging a new perspective. It shifts the focus towards developing processes that help ease managerial decision-making challenges related to choosing the appropriate UD sources and methods to address their problems adequately in a cost-effective manner.

The remainder of this paper is structured around three main contributions that aim at helping managers in projects involving UD to resolve the above-mentioned challenges. First, building on the organizational learning theory, we present a systematic classification of business use cases, namely - inside-out exploration, inside-out exploitation, outside-in exploration, and outside-in exploitation, based on the two dimensions of organizational learning goal (exploration vs. exploitation) and environment scanning scope (internal vs. external); for each dimension, we provide insights in how UD can be used. Second, we develop a conceptual model that provides clear guidelines for managers on how to assess and subsequently choose adequate methods depending on the project scope, aim, and context, and how UD can be integrated in this. Third, we outline the main implications and provide guidance to help further structure the growing stream of insights related to the use cases of UD and the suitability of various methods and data for business research projects.

2. UD in business practice

2.1. UD definition and trends

Balducci and Marinova (2018) conceptualize the structure of a data unit on a continuum and define UD as “a single data unit in which the information offers a relatively concurrent representation of its multifaceted nature without predefined organization on numeric values” (Balducci and Marinova, 2018; p. 557). The most common types of UD include text, images, audio, and video (Statista, 2021).

To highlight and better understand the evolution of interest in using UD in business practice, we follow common trend analyses and trace the number of Tweets mentioning UD between 2009 and 2022. Analyzing Twitter¹ data has a long tradition in business research (see e.g., Hennig-Thurau, Wiertz, and Feldhaus, 2015) and has been shown to be an adequate tool to track and understand CEOs’ and companies’ strategic orientation (see e.g. Malhotra and Malhotra, 2016). It provides us with real-time insights into rising trends and types of interests in using UD from a practitioner’s perspective. Following the approach of Yildirim and Kübler (2023, pp. 135), we extract 215,893 Tweets that mentioned UD using the Twitter Academic API. Fig. 1 shows the share of Tweets mentioning UD, corrected by the quarterly number of Twitter users, from 2019 to 2022. With the advent of social media and the availability to access consumer conversations, we see a large increase in the ratio of Tweets to active users in the third quarter of 2011. With the increase in available data and tools, we also witness a strong growth in the average share throughout late 2014 to the end of 2017, from where we find that the share of Tweets mentioning UD slowly declines again.

For deeper insights into the 215,893 Tweets, we use a structured topic model (STM). Fig. 2 displays the resulting topics and their development over time. Two main findings are noteworthy, i.e., first, from the

discussions on Twitter including “#unstructured data”, the STM model identified the following nine topics²: (1) Data Management, (2) Business Development, (3) Data Privacy and GDPR, (4) Data Access and Data Safety, (5) AI, IBM, and AI Models, (6) Social Media Listening, (7) Text Mining, (8) Value Creation, and (9) Case Studies and Tutorials. Details can be found in the Web Appendix.

Second, in addition to these various application topics, the model also identified differentiated trends over time, allowing us to get an understanding of what is discussed and how these discussions evolved over the last two decades. The results, shown in Fig. 2, reveal a steady discussion of different business applications and opportunities arising from UD (Topic 2). This finding is in line with the digital transformation model presented by Verhoef et al. (2021), as the recent increase in value-related tweets (Topic 8) highlights how companies can leverage insights from UD and generate value for the firm. Similarly, we witness how the interest in using and managing UD substantially increased and continues to increase in importance (Topic 9). Meanwhile, operational questions such as Data Storage and Management (Topic 1) and Data Accessibility and Data Safety (Topic 4) also receive increasing attention, while tools and methods such as Text Mining and Natural Language Processing (NLP) (Topic 7) and Social Media Listening (Topic 6) either gained or remained at high interest. Interest in only two subjects has declined over the past two decades: privacy issues and the impact of GDPR on data research (Topic 3) and Topic 5 mentioning initial but now outdated AI models such as IBM’s Watson AI. We can thus conclude that the interest in UD research remains at a remarkably high level, although the relative share does fluctuate per topic.

While the strong demand for methods and the supply of many tutorials for business and research applications represent a significant evolution of UD sources and methods, our content analysis also shows the lack of structure and guidance on how to compare models and methods and how to ensure that the right data with a fitting model is used to answer a business-specific research question. The importance of addressing this need becomes even more evident when we look at the body of studies in the field, which we summarize next.

2.2. Interest in different types of UD in business research

While business research has been using text data intensively and recently also started using image data, tools in the area of audio and video analytics have been developed more recently. In what follows, we highlight some examples for UD application areas for different forms of UD.

Text data: Across different business disciplines, practitioners can benefit largely from using text data to derive important insights. For example, investment firms can use text data to generate organizational insights, to assess the personalities of CEOs (Malhotra et al. 2018), the psychological traits of startup founders (Noguti et al., 2021) and measure the negative sentiment of media coverage around the announcement of CEO succession (Keil, Lavie, and Pavićević, 2022).

Marketers can use text data to understand consumers, markets, and society (Berger et al., 2020). Consumer chatter in the forms of product reviews (Chevalier and Mayzlin, 2006), social media conversations (e.g., De Haan, 2020, De Haan and Menichelli, 2020, Hewett et al., 2016, Ilhan, Kübler, and Pauwels, 2018), and search on digital platforms (Ringel and Skiera, 2016) can be used to better understand consumer behavior and map consumer perceptions. Recent advances in natural language processing can be used to design more powerful communications (Atalay, El Kihal, and Ellsaesser, 2023), create content for search engine optimization (Reisenbichler et al., 2021), and understand the linguistic drivers of engagement (Berger, Moe, and Schweidel, 2023,

² We chose the number of topics following the approach of Silge (2020), trying to maximize semantic coherence while reducing the residuals. For more details, see Web Appendix 1.

¹ Twitter was rebranded to X in July 2023.

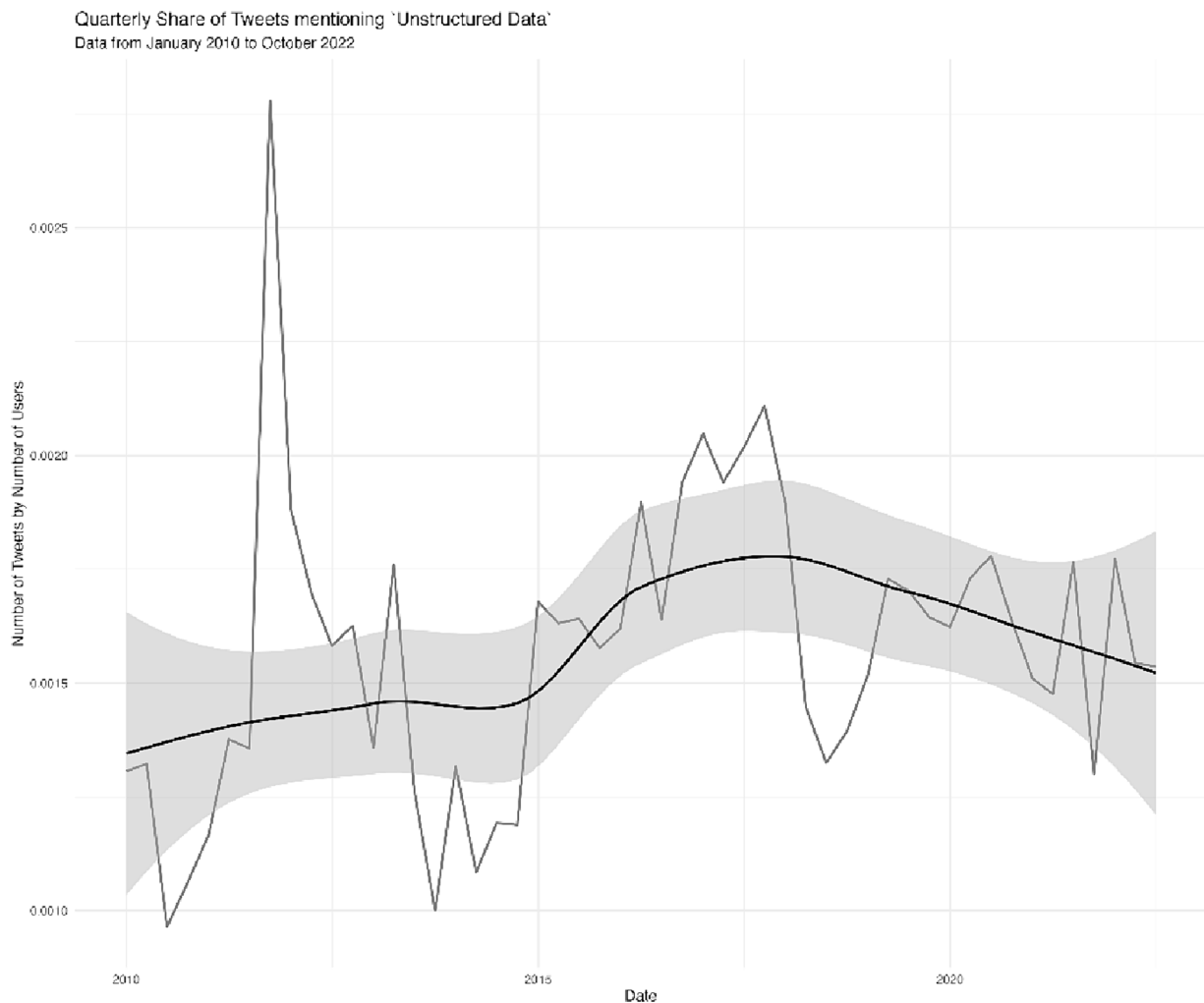


Fig. 1. Quarterly Share of Tweets mentioning 'unstructured data' from 2009 to 2022.

Kupfer et al., 2018).

Image data: In the last decade, image analytics has also garnered significant attention due to the rise in the availability of visual content (Dzyabura, El Kihal, and Peres, 2021). Practitioners can use image mining to get insights that could not be obtained otherwise, e.g., to better understand how consumers perceive their brand by analyzing images associated with the brand (Dzyabura and Peres, 2021), which could come from social media posts and could be used to (re)positioning a brand. Analyzing the images that consumers post on social media can even help in predicting businesses' future success, as Zhang and Luo (2023) have shown. Image analysis can furthermore help in forecasting product return rates (Dzyabura et al., 2023) and review helpfulness (Kübler et al., 2023), i.e., analyzing image data can help in better understanding consumers and businesses as a whole.

Video and audio data: With videos and audio podcasts gaining prominence as communication formats in the social media era, businesses can employ video and audio analytics to further understand individual preferences and new modes of communication. Voice and audio analytics can also be used to analyze quarterly earning calls, which can be valuable for investment firms, or used to measure how well a customer call with the service center has been handled, which can be valuable to analyze the efficiency and impact of customer-firm interactions (e.g., Throckmorton et al., 2015; Mayew and Venkatachalam, 2012a; Mayew and Venkatachalam, 2012b; Hobson et al., 2012).

Particularly, video content has recently gained more attention. Advances in technologies and digital platforms have made video data a

rich, relevant, and valuable source of information for businesses. Though applications using automated approaches to extract features from videos are still very limited, it shows a promising and growing area (Schwenzow et al., 2021). A recent practical example is investigating consumer reactions to influencers in sponsored and non-sponsored videos (Hwang, Liu, and Srinivasan, 2021). Video data can furthermore shed light on organizational phenomena, as Gylfe et al. (2016) have shown. Similarly, although analyzing audio data is still in an early phase, given the high volume of these data (e.g., podcasts, earning calls, customer communications), audio mining can also be useful. For example, Throckmorton et al. (2015) demonstrate how it can detect corporate financial fraud.

Next to using UD for analyses and getting valuable insights, UD can also be input for GenAI to generate new UD, e.g., for improving copywriting and creating new advertisements. As Vomberg et al. (2024) define, GenAI is "using algorithms to generate new (and meaningful) content from training data, such as text, images, or audio." For instance, Jansen et al. (2024) demonstrate that marketers can use their and competitors' ads to train GenAI to generate new ads that outperform their ads at different stages of the purchase funnel. Furthermore, Kübler (2023) provides a checklist that enables marketers to estimate the value of unstructured data in their possession with respect to GenAI training and application.

Overall, we find that the interest in using UD is growing in business research. However, as pointed out earlier, for business there is a clear lack of guidance and structure on how to ensure that data and models fit

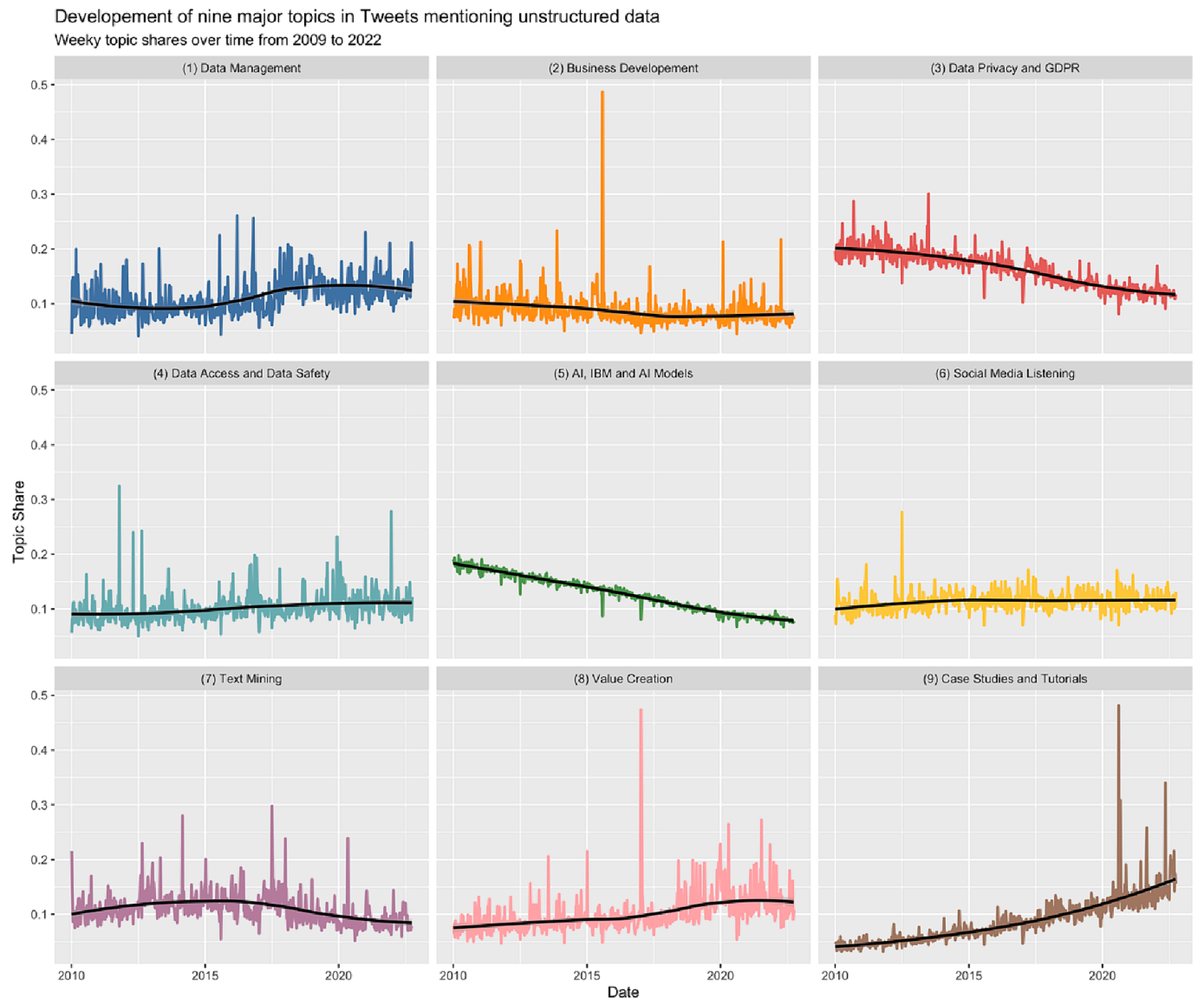


Fig. 2. Development of Topic Shares within Tweets mentioning ‘unstructured data’ from 2009 to 2022.

the problem. To understand the challenges and opportunities faced when using UD, we will next analyze the decision-making journey of UD projects.

3. Decision-making journey

We adapt the general strategic decision-making process proposed in extant literature (e.g., Agarwal and Tanniru, 1991, Harrison, 1996, Mintzberg, Raisinghani and Theoret, 1976) and propose a general project journey for UD projects consisting of three key steps: (1) problem

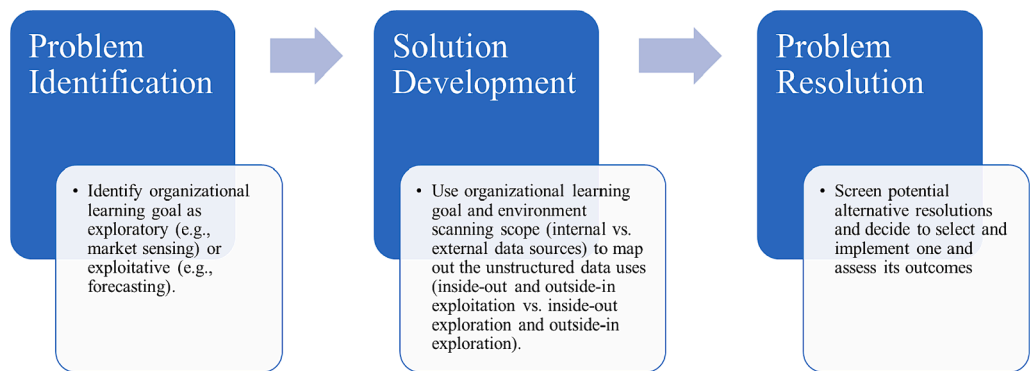


Fig. 3. General Project Decision-Making Journey.

identification, (2) solution development, and (3) problem resolution (see Fig. 3).

In the problem identification phase, managers recognize a problem that reflects an imbalance between the current position of the organization and its desired position. The manager then diagnoses the problem further by understanding the appropriate use case based on the organizational learning goal, i.e., *exploration* or *exploitation*. According to March (1991, see p. 71), exploration includes aspects captured by terms such as search, variation, risk-taking, experimentation, play, flexibility, discovery, and innovation, while exploitation includes aspects such as refinement, choice, production, efficiency, selection, implementation, and execution. In our context, explorative learning involves the acquisition of new knowledge (e.g., understanding of new markets, or shifts in competitor behavior), and exploitative learning involves leveraging the existing knowledge base (e.g., addressing specific customer segments, dynamic pricing). Table 1 provides a summary of key contrasts between the two learning goals.

Next, understanding the appropriate use case enables managers to assess the business impact of addressing the problem. Thus, the problem identification phase consists of problem recognition, learning goal assessment, and impact appraisal.

In the solutions development phase, managers either search for ready-made solutions and adapt them or design custom-made solutions. Either way, given the specific context of the previously identified problem (exploration vs. exploitation), the analyst needs to identify suitable data sources that fit the research purpose (i.e., potential internal vs. external sources of data). Thus, in this phase, one needs to answer questions such as: which data sources are suitable for the analysis and where and how can such data be accessed?

Subsequently, the type and nature of the chosen data source requires identifying suitable tools and methods for data management and analysis, to transform UD into a format that allows further analysis with traditional tools and approaches to provide managers with potential solutions. Here questions to answer include: how can data integrity be ensured (e.g., data completeness) and which methods are best suited to extract meaning (e.g., sentiment) from UD (e.g., Tweets) or which pros (e.g., time) and cons (e.g., precision) do these methods bring?

Finally, in the problem resolution phase, one needs to screen potential alternative resolutions, decide to select and implement one, and assess its outcomes. Here again managers nowadays often face a time vs. precision challenge, as new methods may arise during the resolution phase, potentially requiring the decision maker to go back to the middle stage, which again may prolong the process.

Table 1
Summary of key contrasts between explorative and exploitative learning.

	Explorative learning	Exploitative learning
Definition	Involves studying new information and knowledge beyond the scope of the organization to seek new markets, technologies, or business models.	Learning by exploiting existing knowledge and experience within the organization to improve management choices, processes, and products.
Purpose	To explore and discover novel opportunities, adapt to environmental changes, and foster innovation.	To monetize current products, technologies, or business models, and optimize existing processes for efficiency.
Context	Useful in dynamic and changing environments where adaptation is crucial for survival and growth.	Effective in stable and predictable contexts, ensuring operational excellence.
Risk levels	Higher levels of uncertainty and risk, as it seeks to bring about changes to existing norms.	Seeks to minimize risk and stabilize processes, emphasizing incremental improvements.
Time horizon	Longer-term perspective, often involving trial and error, with a willingness to invest in experimentation.	Shorter-term focus on immediate gains and operational efficiency.

3.1. Problem identification

Managers typically become aware of the organizational problems after recognizing a misfit between the actual and the desired positions. However, such misfits may be difficult to identify. A firm may, for example, desire to provide the best customer experience but may not have sufficiently positive customer feedback relative to its competition to conclude that it does. In such situations, Organizational Learning Theory (March, 1991) suggests that firms must learn and develop appropriate adaptive systems.

Organizational learning refers to the dynamic process of creating new knowledge and transferring it to where it is needed and used, resulting in the creation of new knowledge for later transfer and use (Kane and Alavi 2007). Further, firms can engage in two different learning mechanisms: exploration and exploitation. Exploration focuses on new knowledge creation with the intention to add to or replace the existing content of the organizational memory (Kane and Alavi 2007; March, 1991; Pentland 1995). Exploitation, however, emphasizes diffusion, refinement, and reuse of existing knowledge within the organization (Kane and Alavi 2007; March, 1991; Smith and Zeithaml 1996).

Building on this perspective, we propose that the problem-identification process entails developing an understanding of the organizational learning goal. Specifically, UD can be used to accomplish (1) exploratory learning - e.g., for market sensing (e.g., Netzer et al., 2012; Moe and Schweidel, 2017); or (2) exploitative learning - e.g., optimizing returns on paid search ads (Rutz, Sonnier, and Trusov, 2017) and understanding the impact of online sentiment on sales (Schweidel and Moe, 2014; Sonnier, McAlister, and Rutz, 2011).

The two learning goals can aid in addressing a problem differently. Addressing problems associated with exploitative learning typically has a short- to medium-term impact on performance by helping firms improve their current position. In contrast, explorative learning will likely have a medium- to long-term business impact by providing firms with the necessary insights to develop new or revise existing desired positions.

3.2. Solution development

3.2.1. From organizational learning goal to environmental scanning scope

Once a problem and its potential business impact is identified, managers must work with their team to find suitable solutions. One of the key choices to make is the environmental scanning scope when searching for potential sources of data. Not yet a decade ago, companies could only rely on SD that commonly originated from primary data sources, such as focus groups, customer surveys, and other forms of traditional market research. With the advent of UD, this has substantially changed, and companies benefit now not only from new - unstructured - sources but also from the fact that they can directly access these resources without having to rely on external facilitators such as market research companies and consumer panels. These opportunities, however, come with costs, as companies have to now invest more efforts in identifying UD sources to address use cases themselves. Organizations differ in their degree of environmental scanning for knowledge search, i. e., scanning and collecting information about events and trends in their external environment (e.g., Boh,Huang,and Wu, 2020; Huber, 1991; Stuart and Podolny, 1996). To give an example, companies can use call center conversations and apply voice analytics techniques to detect customer emotions (Britt, 2022) and/or mine user-generated UD (e.g., on social media) that are publicly available. Thus, companies can use “internal” UD, which can be captured within its micro-environment (e. g., call center data that are typically recorded by companies for quality and training purposes), or “external” UD, which can be collected from its macro-environment (e.g., user-generated content on social media) for similar learning goals. As it is possible that, in many cases, managers may combine both data sources for their research, we recommend that

managers rely on the extent to which internal and external data are used to make the distinction rather than assume that a problem can be addressed using only one of the two data sources.

Thus, we propose a solution development framework based on the two dimensions of organizational learning goal and environment scanning scope, leading to four key strategic use cases of UD applications (see Table 2): (1) Inside-out Exploitation, (2) Outside-in Exploitation, (3) Inside-out Exploration, and (4) Outside-in Exploration. Once potential solutions are framed within one or more of these strategic use cases, we can then proceed to understanding approaches related to data management, including data collection, preparation, and analyses. The framework and examples are mainly focused on the business research (i. e., managerial) perspective, but it can also be applicable for academic work.

While our framework can be applied to both SD and UD, it is more relevant for UD research due to the following two reasons. First, UD allows for higher levels of exploration through greater potential for flexibility and discovery due to its non-numeric and multi-faceted nature (Balducci & Marinova, 2018). De Luca et al. (2021) also note that big data sources emerge unintentionally and are mostly unstructured, providing firms with cheaper and real-time access to market information, thereby reducing the traditional barriers to exploration. Given such potential of UD, we argue that UD offers organizations the opportunity to engage in both explorative and exploitative learning, whereas SD has a greater potential to enable exploitative learning. Second, although both SD and UD can exist in internal and external data sources, UD represents the largest portion of data in both sources (Davenport, 2019; Gandomi & Haider, 2015). Therefore, our framework is more relevant from UD research as UD offers greater learning affordances and constitutes most data available for firms in both internal and external data sources.

Table 2
Framework for identifying use cases of UD.

Organizational Learning Goal	Environmental Scanning Scope	
	Internal	External
Exploitation	<p>Inside-out ExploitationGoal: optimize existing business outcomes based on internal UDExamples:</p> <ul style="list-style-type: none">● Make video-based personalized clothing recommendations (Lu, Xiao, and Ding 2016)● Use in-store video recordings to track shopper behavior (Hui et al. 2013)	<p>Outside-in ExploitationGoal: optimize existing business outcomes based on external UDExamples:</p> <ul style="list-style-type: none">● “Listen” to social media chatter to respond quickly to external events (Borah et al. 2020)● Use user-generated content to predict traditional mindset metrics like Awareness, Consideration, or Satisfaction (Kübler et al. 2020)
Exploration	<p>Inside-out ExplorationGoal: develop or revise business position based on internal UDExamples:</p> <ul style="list-style-type: none">● Use connected product ecosystems (e.g., Nike+) develop new mass-customized products and services (Kopalle, Kumar, and Subramaniam 2020)● Identify new market aspirations from customer interviews (Arunachalam et al. 2020)	<p>Outside-in ExplorationGoal: develop or revise business position based on external UDExamples:</p> <ul style="list-style-type: none">● Uncover customer needs using user-generated content on social media (Timoshenko and Hauser 2019)● Analyze patent filings to develop technology roadmaps (Jin, Jeong, and Yoon 2015; Pora et al. 2020).

3.2.1.1. Inside-out Exploitation. Inside-out exploitation allows a firm to optimize its existing business opportunities (e.g., eliminating operational bottlenecks, optimizing the customer journey, enhancing customer experience, or reducing the rate of machine failures) based on insights extracted from a company’s internal UD. Such insights can help managers develop targeted mobile promotions and/or optimize spatial configurations of their stores to maximize conversion rates. Thus, addressing the problems belonging to this quadrant can typically boost a company’s efficiency (i.e., improve profitability) and provide a competitive advantage in the short to medium run. Further, in such use cases and applications of UD, managers commonly face a situation where they start with a clear problem definition but must identify suitable data within the company’s realm to then subsequently identify suitable methods for data management.

3.2.1.2. Outside-in Exploitation. Outside-in exploitation uses of UD are use cases that allow a firm to compose an action to exploit existing business opportunities based on insights identified from UD in its external environment. Typically, firms are likely to use external UD, as illustrated by the examples in Table 2.

Answering such problems is also likely to help improve a firm’s efficiency in the short- to medium-run, but indirectly by highlighting the strategic gap between intended and desired strategic positions by analyzing environmental cues. For example, by extracting the consumers’ brand image from user-generated content on social media, a firm can boost the ROI of branding campaigns by first understanding the degree of discrepancy between the intended and perceived brand images, and then updating their campaigns. Other exemplar uses within this quadrant include trend spotting, virality hacking, quick media reactions, and firestorm predictions.

3.2.1.3. Inside-out Exploration. By inside-out exploration uses of UD, we refer to use cases that allow a firm to accomplish explorative learning goals, such as the identification of new markets, using internal UD. Table 2 provides two exemplary published cases of inside-out exploration. Addressing problems pertaining to this quadrant is more helpful for firms to revise their desired strategic position itself than minimizing the existing gap, and therefore have a more medium- to long-term business impact requiring revisiting their resource allocation strategies.

3.2.1.4. Outside-in Exploration. Table 2 provides examples of how exploration learning goals can also be accomplished by using external UD, making a case for outside-in explorative learning. Such uses help companies with organizational adaptation, which refers to intentional decision-making that leads to observable actions with the aim of reducing the distance between an organization and its economic and institutional environments (Sarta et al., 2021). Thus, by addressing such problems, a company can potentially improve not just its survival, but its business sustainability in the long run. For instance, innovation researchers hail a company’s future market orientation, which requires firms to respond to weak signals from the external environment, as an indicator of its ability to introduce radical and disruptive innovations, that can transform industries and change the rules of the game entirely (Chandy and Tellis, 1998; Govindarajan,Kopalle,and Danneels, 2011).

3.2.2. From data selection to data analysis

3.2.2.1. Data Selection. Once the managerial problem is identified (exploration vs. exploitation-related problem) and translated into a research question, researchers must put the project into motion, by a) identifying and selecting suitable data sources (internal vs. external), as well as data extraction techniques, b) ensuring sufficient data quality by controlling, preparing and merging data, and c) identifying the right tools to harvest information from the collected data, by transforming the UD into a structured format that is then ready for further analysis. Only

then, one can answer the research question in an evidence-based manner. Below, we discuss these elements.

Following from the research question, a data scientist needs to identify which data sources are needed to answer the research question. The research question typically provides some guidance but usually does not fully dictate what data sources are needed. For example, when an online retailer’s goal is to predict the return rates of products prior to launch, different data sources may be used. Besides classic SD such as price and category, one may also use UD, such as product images to improve forecasts and understand which visual features of a fashion item relate to the return rate (Dzyabura et al., 2023). While price and category data is easily available from the company’s own systems, product images have quite different data requirements and need to be either provided or crawled from the product manufacturer.

Boegershausen et al. (2022) identify three challenges faced by data scientists when selecting web scraping data sources, but these challenges apply to other types of data as well. The first concerns exploring the universe of potential data sources; the authors warn against the temptation to rely only on familiar data sources. Instead, one should explore the vast universe of potential sources to a sufficient degree.

Second, the authors advise considering all available data collection methods, including web scraping, public data platforms (e.g., Kaggle.com, Dataverse.org), and data providers (e.g., Yelp.com, IMDb). Too often, data does not get included in projects, because it seems too hard to be collected. However, even though web scraping might take more time (in terms of collection and data preparation later), it often allows analysts to access essential and valuable but often otherwise non-available information (e.g., product reviews, social media chatter, etc.). Given the fast development of scraping libraries for Python, JavaScript or R, and the possibility to access data via hidden APIs, collecting such data also became easier in recent years (Yildirim and Kübler, 2023).

Third, the authors advise to map the data context. This requires researchers to assess why and how data has been created. E.g., Twitter is more known for customer complaints, while Instagram is more known for brand appraisal (see e.g., Waterloo et al., 2018). A sentiment analysis will thus automatically find substantial differences between the general sentiment across platforms. This would become a problem, if for specific brands in a sample, only Twitter data was collected, and for others only Instagram data. Without a deep understanding of the data, any further processing steps are likely to be futile. Hence it is essential to clearly identify the data structure and the context for which it was collected. Furthermore, recent research points out that some UGC may result from fraudulent actions and thus not be representative, such as fake or bought reviews (He, Hollenback, and Prosperio, 2022). The authors thus advise to also consider the potential presence of misleading, biased, or fraudulent data, when evaluating a data source.

3.2.2.2. Data Collection and -Management. After the sources from which data needs to be collected have been identified, the actual data collection is the next step. Depending on the type of data, different data collection techniques should be employed. For internal UD that resides in relational databases, this usually requires writing SQL queries to extract the relevant data. For external structured data (SD) sources, a data scientist typically has to comply with the *modus operandi* that the external provider has designed for accessing these data. Collecting internal UD can be quite tedious, depending on how the data is stored and to what extent it is available to others than those who have generated the data. In extreme cases, it may even involve entering data points manually. External UD is to a large extent collected via APIs or web scraping. Boegershausen et al. (2022) provide an excellent discussion of the challenges that are involved when accessing this type of data.

Further, when collecting UD from external sources, managers not only have to decide on how to extract data from these sources (crawling vs. buying) but also how to identify suitable and relevant content within each media or channel (i.e., keywords, users, groups, etc.). In addition,

because of technical data management challenges, managers are also advised to spend substantial time and resources to develop an understanding of the data management approach. For instance, in data collection, there may be issues related to legal aspects, e.g., the maximum amount of data one is allowed to extract (Boegershausen et al., 2022), and completeness and censorship (see Ruths and Pfeffer (2014) for a more concrete discussion of potential biases within social media data).

Similarly, during data preparation, a key focus should lay on the predictive power as well as the validity and reliability of the processed constructs. Kübler, Colicev, and Pauwels (2020) provide, for example, an overview of different methods and show that the prediction power of simple mindset metrics systematically varies depending on which type of sentiment classifier one relies on. Managers are thus advised to invest substantial time and effort in understanding which words or constructs (e.g., sentiment, moral statements) are effective in predicting specific consumer actions (e.g., outrage, satisfaction, etc.). Depending on the application, one may focus more on minimizing true positives over false positives (e.g., in case of predicting outrage or firestorms), while other applications may focus more on accuracy (e.g., satisfaction tracking).

In Table 3, we summarize the key questions that managers should ask during data collection for the different data sources and learning goals. These questions are typical questions for each quadrant, but we do have to note that some questions can also hold for the other quadrants.

3.2.2.3. Data Preparation. Data quality issues may arise due to misspellings during data entry, missing information, or other invalid data. Verhoef et al. (2021) distinguish three dimensions of data quality: completeness, accuracy, and consistency of data. Completeness refers to whether all data are present for all entities. Incompleteness of data can be due to missing values or fragmentation. For example, fragmented online data issues may be encountered when an entity is accessing a website via multiple devices. When the data scientist does not have access to data from all devices, the full journey cannot be observed, as only fragments of it are available. Traditionally, data fusion methods

Table 3
Summary of key questions to ask during data collection.

Organizational Learning Goal	Environmental Scanning Scope	
	Internal	External
Exploitation	Inside-out Exploitation <ul style="list-style-type: none">● What data is available or can be created/ captured?● How much data is necessary to obtain robust insights?● Which formats will be most appropriate?	Outside-in Exploitation <ul style="list-style-type: none">● Which external data sources relate directly to existing business outcomes?● Which data sources more suitable for automating decisions (e.g., personalized offer based on a customer’s social media engagement) and real-time monitoring (e.g., monitoring customer sentiment)?● What are the different data extraction approaches (e.g., crawling vs. API access)?● What are the privacy and legal boundaries?● What are the limits of data completeness and censorship?
Exploration	Inside-out Exploration <ul style="list-style-type: none">● What are challenges to data storage and how can they be resolved?● How timely will the data be?	Outside-in Exploration <ul style="list-style-type: none">● To collect data in-house or use commercial solutions?● What data sources are best suited to identify potential future opportunities vs. threats?

were used to overcome fragmentation in data (e.g. Kamakura and Wedel, 1997), but recently generative adversarial networks (GANs) have shown to be very fruitful in completing data. GANs also provide a viable alternative to more traditional data imputation methods for missing values (Kubara, 2019). For an overview of more traditional imputation methods, we refer to Rubin and Little (2020).

Accuracy of data refers to whether the data values stored for an object are the correct values. To be correct, a data value must be the right value and must be represented in a consistent and unambiguous form. The threshold for desired accuracy and consequences of inaccurate data can vary significantly from one business context to another. For instance, for speech indexing – i.e., indexing speech/ audio data for an easy and effective search experience, speech must first be converted to text and then indexed. According to Tim Olson, SVP Digital Strategic Partnership at KQED Radio, this added step is trickier than it seems, particularly for news broadcasters as the bar for accuracy is “very high,” as inaccuracies that can potentially change the meaning of a topic and can cause “embarrassment for the news outlet” (Olson, 2021).

Consistency issues arise when the same value is represented in multiple ways. For example, a data set may contain inconsistent representations of New Year’s Eve such as “31-12-2022”, “12-31-2022 “, “Dec 31, 2022”, or “31 December ‘22”. Data cleaning deals with detecting and removing errors and inconsistencies from data, to improve the quality of data, and is typically a tedious task. It involves several steps, which we do not discuss here. Instead, we refer to Boehmke (2016).

Table 4 presents typical questions to address during the data preparation stage.

3.2.2.4. Data analysis. After the data is collected and prepared, insights can be generated. For UD, this step is not as straightforward as it is for SD since standard econometric techniques are not directly applicable. Fortunately, there are possibilities for UD, as discussed in Section 2 of this paper. It is important to note that for UD, feature extraction plays a crucial role. Prior to training, evaluating, and validating a model, the

manager needs to decide on the feature extraction method (Dzyabura, El Kihal, and Peres, 2021).

Depending on the managerial problem, the feature space can vary strongly. If the interpretability of the features used in the model is not necessary for deriving insights, then automatic feature extraction using deep neural networks would be the right approach. If, however, the interpretability of the extracted features is necessary for solving the managerial problem, then the type of method used to extract features would be different. For example, if a retailer is interested in forecasting return rates of fashion products in its online shop with the highest possible accuracy, to use the forecasts as input to rank-order the products in the online shop, then using deep learned features would be the appropriate way (Dzyabura et al., 2023). If, however, an understanding of which apparel features result in higher return rates, to inform designers and buyers, then using other methods to extract interpretable features (such as shape, color, or texture) would be necessary. In the example here, using color histograms or Gabor filters to extract relevant features from the image would be more appropriate.

Once a decision is made about the feature extraction method, the newly created structured variables can, in turn, be analyzed similarly as the other structured variables, and can be used in the same econometric models. The typical questions to address during data analysis are presented in Table 5.

3.3. Problem resolution

Problem resolution is the final step of our proposed strategic decision-making journey, which involves managers screening, selecting, implementing, and assessing the potential solutions developed using the framework from Table 2. Before starting a new data-driven project, one needs to determine the desired outcome, as well as the benefits of the outcome. Besides benefits, managers should also focus on the costs (e.g., money, time, or personnel) involved in reaching this. Furthermore, one should have an understanding of the resulting costs from misclassification. This may imply that researchers need to clarify whether false-positive or false-negative misclassifications lead to higher costs, to then ultimately decide for which side they wish to optimize. Based on these aspects, the ROI can be calculated. This can be done explicitly, e.g., by actually calculating the expected ROI before green-lighting the project, or implicitly, e.g., by assuming that the benefits will outweigh the costs. In some cases, the costs might be directly related to financial costs, and in others it can, for instance, be the opportunity costs (e.g., lost time which cannot be put in anything else) or costs caused by misclassification as a result of an inaccurate or inappropriate analysis (e.

Table 4
Summary of key questions to ask during data preparation.

Organizational Learning Goal	Environmental Scanning Scope	
	Internal	External
Exploitation	Inside-out Exploitation <ul style="list-style-type: none">● How can data from different internal sources be matched?● What is the data-generating process, how to interpret findings, endogeneity issues, and can we get causal insights?● Is it possible to track additional data?● What is the quality of the data, e.g., are there measurement errors, missing values, or differences in tracking method over time? How to deal with this?	Outside-in Exploitation <ul style="list-style-type: none">● How to assess data quality, specifically related to the validity and reliability of the measures and their predictive power?● What are the potential outcomes if there are measurement errors?● How can we match this to internal data, e.g., related to business outcomes?
Exploration	Inside-out Exploration <ul style="list-style-type: none">● Are variables coded consistently?● To what extent are internal sources sufficient for exploration, e.g., is it useful to find new markets? How can the data be used to uncover this sufficiently?	Outside-in Exploration <ul style="list-style-type: none">● How to match (different sources of) market data with firm data?● Are data on competitors (e.g., sentiment) measured similarly to internal data?

Table 5
Summary of key questions to ask during data analysis.

Common Concerns	For Internal Data Sources	For External Data Sources
<ul style="list-style-type: none">● Which feature extraction method is necessary?● Do we have the necessary capabilities and other resources to conduct such analyses?● How to assess output quality (e.g., causality vs accuracy)?● Is interpretability of the extracted measures necessary?● What are the potential limitations of the different options?● What is the validity and consistency of the measures (especially when new UD comes in)?	<ul style="list-style-type: none">● What additional value can be created?● What constructs can be captured with the SD and UD?	<ul style="list-style-type: none">● What level of aggregation is suitable?● How can the data be visualized and presented?● How can the external data be linked/ matched with the internal data?

g., re-acquisition costs of false-negatively classified and churning customers). Similarly, while in some cases the monetary benefits may be direct (e.g., increase in revenue or profits, or decrease in costs), in others, they may be indirect (e.g., via enhanced customer experience).

To further illustrate our point above, consider the following example. Netflix might have a managerial problem with increasing the viewership of a new series (by x%) by having a thumbnail image with a higher click-through or conversion rate. There are different ways to tackle such a problem, as also discussed by Netflix in their tech blog (Chandrashekar et al., 2017). One might be using trial and error or conducting A/B testing, without using UD. The advantage is that this is relatively easy to conduct and can provide a quick answer to the question, especially when a company like Netflix can conduct and easily implement large-scale experiments. The costs of such a research project are relatively low, the modeling procedure is straightforward, the amount of time spent, and the number of employees needed is low. However, a potential downside is that such insights are not generalizable (i.e., they are case-specific), meaning that Netflix has to conduct such large-scale experiments for every new show and movie. In addition to the scalability issue, there is also little to no insight into the specific elements of the thumbnail driving the desired action as it only provides information on which thumbnail (out of a series of thumbnails) performs best.

Another somewhat more elaborated method is to investigate all historical thumbnails and explore how users have interacted with these. Image mining could be used to classify the thumbnails into different groups based on the content in it (e.g., does it contain the leading actor or some other elements), which could subsequently be used as input for a model to predict the impact of a specific thumbnail. For such an approach, existing (out-of-the-box) tools could be used. Although this is somewhat more elaborate than the first procedure and relatively more resource-demanding, using it can result in richer insights.

Alternatively, Netflix could develop its own image mining algorithm, which can be fine-tuned for this specific task performed at a segment or even individual customer level, to find the best thumbnail for the right person at the right time (e.g., for some users a screenshot of the leading actor in a romantic setting might be most appealing, while for others an action sequence can be more effective). The “optimal” thumbnail could even be auto-generated by AI. This again will require more resources initially, but can potentially better tackle the problem of scalable, repeatable solutions. The question is if these additional benefits outweigh the additional costs. The answer to such questions highly depends on the problem at hand, the organization, resource availability, and competing problems.

Overall, in practice, many of such cost-benefit decisions must be made; what are the costs and benefits of having better data, collecting

external data, using UD, spending more time on data cleaning, incrementally improving a model, learning and estimating a more complicated model, of doing additional analysis and robustness checks, and so on. The key question is whether the incremental benefits outweigh the incremental costs. We present in Fig. 4, an illustration of the cost-benefit tradeoffs for three hypothetical problem resolution alternatives (i.e., alternative projects that can resolve the business question to varying degrees and against varying costs).

To better understand Fig. 4, let's assume that all three resolution alternatives address the same research problem. The management team must now screen and select from these three options. Resolution alternative A represents a project with existing internal data, which can be conducted very quickly, but the maximal potential return is relatively low. Such a solution is ideal for a tactical decision that must be made rather quickly (e.g., high opportunity costs) or for solving a problem where there is not much at stake. Resolution alternative B represents a project which needs more preparation time (e.g., data collection, data cleaning, studying the problem at hand), this preparation time does not directly translate to readily usable insights (i.e., the value stays for a while at zero), but as soon as the project is underway, the management can gain useful insights quickly and the benefits are significantly higher compared to resolution alternative A. Finally, resolution alternative C is extremely demanding on resources with a relatively lower incremental benefit (i.e., the line slope is less steep), but the maximum benefit over time is the highest for this project. Resolution alternative C could for instance involve a project with complicated but rich data with the potential to offer excellent insights, where with more time more insights can be obtained, but obtaining these insights does cost quite some resources.

The selection of an alternative depends on multiple aspects. The following questions can guide this selection:

- What is the organizational learning goal? On the one hand, it must be noted that an exploitative learning goal implies that the management seeks a solution with less uncertainty of returns in a relatively short time. On the other hand, explorative learning requires a more long-term view of the problem, and flexibility with experimentation may be necessary to arrive at an “optimal solution”.
- What are the opportunity costs? Firms have limited resources, and so resources allocated to pursuing one solution could have been better invested elsewhere. Thus, when conducting a cost-benefit tradeoff analysis, it is also important that managers consider opportunity costs as well.
- What is the immediacy of arriving at a solution? Depending on how well the organization is aligned with its environment, management

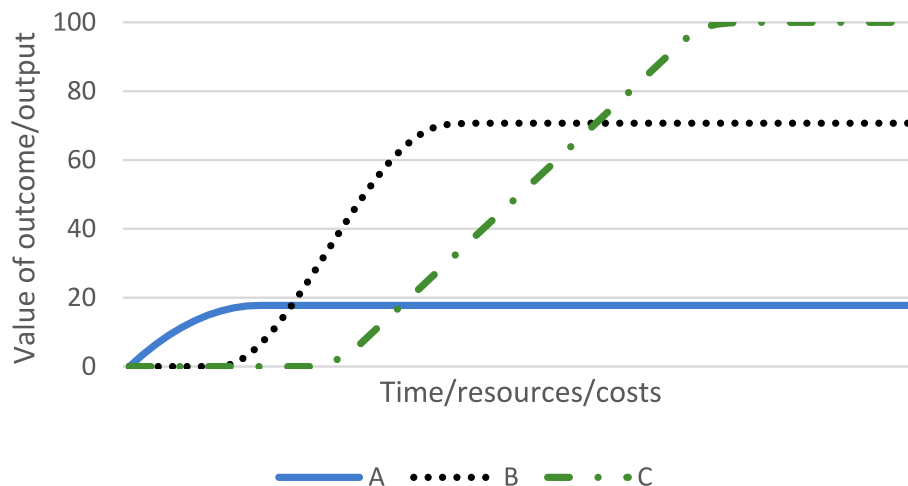


Fig. 4. Illustration of tradeoffs for three resolution alternatives.

priorities may vary. For instance, if the problem solution concerns reacting to a competitive action, the need for a solution might be more immediate.

- What is the minimum desired quality of the results/insights? E.g., how accurate does it have to be, what are the costs and implications if the estimates are off or if the implementation is sub-optimal? Are these costs driven by false positives or false negatives? They all represent a few of the critical questions the managers must answer to have clear assessment criteria to evaluate the solutions.
- What is the desired outcome? Should the outcome of the project meet or come close to this desired outcome? What are the benefits if the project exceeds the desired outcome? What are the costs if the project doesn't meet the desired outcome?

Before starting a project, investigating the various hypothetical scenarios is crucial. To achieve this, one needs to estimate the costs and benefits of having better data based on various factors – e.g., resource demands, potential impact on project delivery, and so on. Such an investigation can provide clarity on deciding (1) whether a project should be approved at all, and upon approval (2) the ideal scope of the project.

3.4. Implementation and outcome assessment

During the course of the project, expected costs and benefits must be continually monitored. Sunk costs should, of course, be ignored when deciding to cancel or change a running project; only (additional) costs and benefits that will be encountered must be taken into account for canceling or adjusting an ongoing project (Mankiw, 2009). Especially in projects related to explorative learning, given the uncertainties and experimentation involved (even more so when using external data sources), management should be prepared for potential failures, at least in terms of direct and immediate monetary benefits. Further, for every project, the costs and benefits must be quantified on the same scale (e.g., monetary value). The project can be greenlit when the gain (i.e., benefits minus the costs) is greater than zero, or above some other threshold (e.g., the project with the highest potential ROI or gains).

Coming back to Fig. 4, selecting the alternative depends on the desired outcome level, the usefulness of having an “improved” outcome, the opportunity costs (e.g., scarce resources could be used on an alternative project), and other relevant aspects. If the question is: “How will an increase in the price (by x%) to cover the rising costs impact our sales and profitability?”, and an answer is needed immediately, resolution alternative A might be best, although the answer might be less accurate. In some cases, it might even be beneficial to go for multiple alternatives, e.g., resolution alternative A to give a less accurate but quick (tactical)

answer to the question at hand and resolution alternative C to give a more precise and detailed answer to the strategic decision-making of pricing. Taking the first derivative ($f'(x)$) of Fig. 4 can help hereby since this shows us how a marginal increase in investing in scarce resources contributes to improving the outcome of the investment. See Fig. 5 for an illustration of the first derivative. As can be seen from the figure, beyond a certain threshold point, the additional investment does not pay off anymore or worse may reduce the outcome value. Thus, we can maximize the ROI by avoiding additional investments beyond the threshold point.

Of course, the examples in Figs. 4 and 5 are a simplification of reality, since in reality, there is also uncertainty about the incremental value of making additional investments. Furthermore, especially at the early stages of a project, the potential outcome might be difficult or even impossible to estimate. Hence, one must also take such uncertainties and challenges into account, for instance by checking at different stages of the project if the projected trajectory is still in line with expectations, and then adjust as necessary.

4. Discussion and implications

Firms that can “figure out” how to use insights unlocked from text, multimedia, social media, and other UD sources, which is estimated to form up to 80–90 % of all data, are likely to unlock a competitive advantage (Harbert, 2021). UD analysis is, therefore, gaining prominence in organizations (Hodgson, 2015). Yet only 18 % of organizations report being able to take advantage of UD (Davenport, 2019). Moreover, there is a lack of guidance for managers on how to create and extract value from UD with a structured approach. Our study addresses this limitation and provides a guiding framework for managers to organize their UD research. To summarize, our framework identifies four unique uses of UD based on whether the organizational learning goal is explorative or exploitative, and the degree of internal and external data sources determined in the solutions development phase.

The key implication for managers from our framework is that they should avoid having preconceived notions about the nature of a problem and the data source required to address the problem, which can limit their ability to identify appropriate uses of UD and extract value from UD. To illustrate the point, let's consider the example of customer experience management, where a firm identifies a problem when it receives a lower customer satisfaction score relative to its competition. A manager can view this as an exploitation learning opportunity to enhance customer experience by addressing key pain points in the customer journey. But customer experience can also be enhanced by developing superior products that address unmet needs of the market. Therefore, the situation can also be an opportunity for the firm to engage

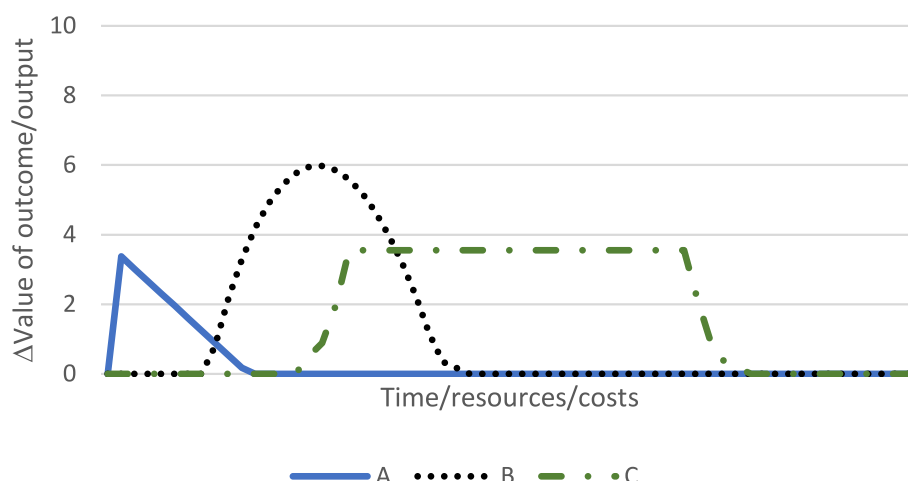


Fig. 5. The incremental contribution of the three resolution alternatives.

in explorative learning involving the identification of unmet market needs to guide superior product development for the future.

Further, both of these learning goals can be accomplished by using company internal data (e.g., clickstream, in-store video recordings, call center call logs) or external data (e.g., social media posts, and reviews on third-party websites). Depending on the firm's prioritization of the learning goal and the extent to which internal and external data are employed, four uses emerge: inside-out exploitation, outside-in exploration, inside-out exploration, and outside-in exploitation. Identifying such potential uses of UD provides managers with solution alternatives that they can evaluate and pursue instead of having a truncated view of the problem and limiting the potential sources of data. The evaluation of the potential uses should be holistic based on business impact and the tradeoff between the required and available resources to the firm.

CRedit authorship contribution statement

Evert de Haan: Conceptualization, Writing – original draft, Visualization. **Manjunath Padigar:** Conceptualization, Writing – original draft, Visualization. **Siham El Kihal:** Conceptualization, Writing – original draft, Visualization. **Raoul Kübler:** Conceptualization, Writing – original draft, Visualization. **Jaap E. Wieringa:** Conceptualization, Writing – original draft, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- Agarwal, R., & Tanniru, M. R. (1991). Knowledge extraction using content analysis. *Knowledge Acquisition*, 3(4), 421–441.
- Arunachalam, S., Bahadir, S. C., Bharadwaj, S. G., & Guesalaga, R. (2020). New product introductions for low-income consumers in emerging markets. *Journal of the Academy of Marketing Science*, 48, 914–940.
- Atalay, A. S., El Kihal, S., & Ellsaesser, F. (2023). Creating effective marketing messages through moderately surprising syntax. *Journal of Marketing*, 87(5), 755–775.
- Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, 46(4), 557–590.
- Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1), 1–25.
- Berger, J., Moe, W. W., & Schweidel, D. A. (2023). What holds attention? Linguistic drivers of engagement. *Journal of Marketing*, 87(5), 793–809.
- Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of gold: scraping web data for marketing insights. *Journal of Marketing*, 86(5), 1–20.
- Boehmke, B. (2016). *Data Wrangling with R*. Springer International Publishing.
- Boh, W. F., Huang, C. J., & Wu, A. (2020). Investor experience and innovation performance: The mediating role of external cooperation. *Strategic Management Journal*, 41(1), 124–151.
- Borah, A., Banerjee, S., Lin, Y. T., Jain, A., & Eisingerich, A. B. (2020). Improvised marketing interventions in social media. *Journal of Marketing*, 84(2), 69–91.
- Britt, R. (2022). Detecting customer emotions with CallMiner. <https://callminer.com/blog/detecting-customer-emotions-with-callminer>. (Retrieved on May 8th, 2022).
- Chandrashekar, A., Amat, F., Basilio, J., & Jebara T. (2017). Artwork Personalization at Netflix. <https://netflixtechblog.com/artwork-personalization-c589f074ad76>. (Retrieved on Dec. 21st, 2022).
- Chandy, R. K., & Tellis, G. J. (1998). Organizing for radical product innovation: The overlooked role of willingness to cannibalize. *Journal of Marketing Research*, 35(4), 474–487.
- Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345–354.
- Davenport, T. H., (2019). Analytics and AI-driven enterprises thrive in the age of with. <https://www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html>. (Retrieved on February 1st, 2024).
- De Haan, E. (2020). Satisfaction surveys or online sentiment: Which best predicts firm performance. *MSI Working Paper Series*, 20(101), 1–46.
- De Haan, E., & Menichelli, E. (2020). The incremental value of unstructured data in predicting customer churn. *MSI Working Paper Series*, 20(105), 1–49.
- Dzyabura, D., El Kihal, S., Hauser, J. R., & Ibragimov, M. (2023). Leveraging the Power of Images in Managing Product Return Rates. *Marketing Science*, 42(6), 1125–1142.
- Dzyabura, D., El Kihal, S., & Peres, R. (2021). Image Analytics in Marketing. In *Handbook of Market Research* (pp. 665–692). Cham: Springer International Publishing.
- Dzyabura, D., & Peres, R. (2021). Visual elicitation of brand perception. *Journal of Marketing*, 85(4), 44–66.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
- Govindarajan, V., Kopalle, P. K., & Danneels, E. (2011). The effects of mainstream and emerging customer orientations on radical and disruptive innovations. *Journal of Product Innovation Management*, 28(s1), 121–132.
- Gylfe, P., Franck, H., LeBaron, C., & Mantere, S. (2016). Video methods in strategy research: Focusing on embodied cognition. *Southern Medical Journal*, 37, 133–148.
- Harbert, T. (2021). Tapping the power of unstructured data. MIT Sloan. <https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data>. (Retrieved on February 1st, 2024).
- Harrison, E. F. (1996). A process perspective on strategic decision making. *Management Decision*, 34(1), 46–53.
- Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1), 75–87.
- He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science*, 41(5), 896–921.
- Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, 43, 375–394.
- Hewett, K., Rand, W., Rust, R. T., & Van Heerde, H. J. (2016). Brand buzz in the echoversion. *Journal of Marketing*, 80(3), 1–24.
- Hobson, J., Mayew, W. J., & Venkatachalam, M. (2012). Analyzing speech to detect financial misreporting. *Journal of Accounting Research*, 50(2), 349–392.
- Hodgson, K. (2015). What's the big deal about Big Data? *SDM Magazine RSS*. <https://www.sdmag.com/articles/91386-whats-the-big-deal-about-big-data>. (Retrieved on February 1st, 2024).
- Huber, G. P. (1991). Organizational learning: The contributing processes and the literatures. *Organization Science*, 2(1), 88–115.
- Hui, S. K., Huang, Y., Suher, J., & Inman, J. J. (2013). Deconstructing the “first moment of truth”: Understanding unplanned consideration and purchase conversion using in-store video tracking. *Journal of Marketing Research*, 50(4), 445–462.
- Hwang, S., Liu, X., & Srinivasan, K. (2021). Voice Analytics of Online Influencers (January 26, 2021). Available at SSRN: <https://ssrn.com/abstract=3773825>.
- IDC (2023). Untapped Value: What Every Executive Needs to Know About Unstructured Data (August 2023). IDC white paper, sponsored by Box, IDC #US51128223. Available at: <https://www.box.com/resources/unstructured-data-paper>.
- Ilhan, B. E., Kübler, R. V., & Pauwels, K. H. (2018). Battle of the brand fans: Impact of brand attack and defense on social media. *Journal of Interactive Marketing*, 43, 33–51.
- Jansen, T., Heitman, M., Reisenbichler, M., & Schweidel, D. A. (2024). Automated Alignment: Guiding Visual Generative AI for Brand Building and Customer Engagement (December 23, 2003). Available at SSRN: <https://ssrn.com/abstract=4656622>.
- Jin, G., Jeong, Y., & Yoon, B. (2015). Technology-driven roadmaps for identifying new product/market opportunities: Use of text mining and quality function deployment. *Advanced Engineering Informatics*, 29(1), 126–138.
- Kamakura, W., & Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, 34, 485–498.
- Keil, T., Lavie, D., & Pavićević, S. (2022). When do outside CEOs Underperform? From a CEO-centric to a stakeholder-centric perspective of post-succession performance. *Academy of Management Journal*, 65(5), 1424–1449.
- Kopalle, P. K., Kumar, V., & Subramaniam, M. (2020). How legacy firms can embrace the digital ecosystem via digital customer orientation. *Journal of the Academy of Marketing Science*, 48(1), 114–131.
- Kubara, K. (2019). GANs and Missing Data Imputation. <https://towardsdatascience.com/gans-and-missing-data-imputation-815a0c4e4e>. (Retrieved on Jan. 31st, 2023).
- Kübler, R. V. (2023). Will the revolution devour its children? The impact of generative and interactive AI on operative and strategic marketing. *Décisions Marketing*, 112(4), 267–288.
- Kübler, R. V., Colicev, A., & Pauwels, K. H. (2020). Social media's impact on the consumer mindset: When to use which sentiment extraction tool? *Journal of Interactive Marketing*, 50, 136–155.
- Kübler, R. V., Lobschat, L., Welke, L., & van der Meij, H. (2023). The impact of images on review helpfulness: A contingency approach. *Journal of Retailing*. forthcoming.
- Kupfer, A. K., Vor, P., der Holte, N., Kübler, R. V., & Hennig-Thurau, T. (2018). The role of the partner brand's social media power in brand alliances. *Journal of Marketing*, 82(3), 25–44.
- Lu, S., Xiao, L., & Ding, M. (2016). A video-based automated recommender (VAR) system for garments. *Marketing Science*, 35(3), 484–510.
- Malhotra, C. K., & Malhotra, A. (2016). How CEOs can leverage twitter. *MIT Sloan Management Review*, 57(2), 73.
- Mankiw, N. G. (2009). *Principles of Microeconomics* (5th ed., pp. 296–297). Mason, OH: Cengage Learning.
- March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 71–87.
- Mintzberg, H., Raisinghani, D., & Theoret, A. (1976). The structure of “unstructured” decision processes. *Administrative Science Quarterly*, 246–275.

- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521–543.
- Noguti, V., Ho, H., Padigar, M., & Zhang, S. X. (2021). Do individual ambidexterity and career experience help technological startup founders acquire funding? *IEEE Transactions on Engineering*.
- Olson, T. (2021). Using AI to explore the future of news audio. <https://blog.google/products/news/using-ai-explore-future-news-audio/>. (Retrieved Jan. 28th, 2023).
- Pora, U., Gersdri, N., Thawesaengskulthai, N., & Triukose, S. (2020). Data-driven roadmapping (DDRM): Approach and case demonstration. *IEEE Transactions on Engineering Management*, 69(1), 209–227.
- Reisenbichler, M., Reutterer, T., Schweidel, D., & Dan, S. (2021). Frontiers: Supporting content marketing with natural language generation. *Marketing Science*, 41(3), 441–452.
- Ringel, D. M., & Skiera, B. (2016). Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Science*, 35(3), 511–534.
- Rubin, D. B., & Little, R. J. A. (2020). *Statistical Analysis with Missing Data*. John Wiley & Sons.
- Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063–1064.
- Rutz, O. J., Sonnier, G. P., & Trusov, M. (2017). A new method to aid copy testing of paid search text advertisements. *Journal of Marketing Research*, 54(6), 885–900.
- Sarta, A., Durand, R., & Vergne, J. P. (2021). Organizational adaptation. *Journal of Management*, 47(1), 43–75.
- Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research*, 51(4), 387–402.
- Schwenzow, J., Hartmann, J., Schikowsky, A., & Heitmann, M. (2021). Understanding videos at scale: How to extract insights for business research. *Journal of Business Research*, 123, 367–379.
- Sonnier, G. P., McAlister, L., & Rutz, O. J. (2011). A dynamic model of the effect of online communications on firm sales. *Marketing Science*, 30(4), 702–716.
- Statista (2021). Unstructured data types in organizations in the United States and the United Kingdom (UK) in 2021. <https://www.statista.com/statistics/1262636/unstructured-data-types-organizations-us-uk/>. (Retrieved on Jan. 28th, 2023).
- Stuart, T. E., & Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17(S1), 21–38.
- Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, 74, 78–87.
- Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1), 1–20.
- Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, 122, 889–901.
- Vomberg, A., De Haan, E., Fabian, N. E., & Broekhuizen, T. (2024). Digital knowledge engineering for strategy development. *Journal of Business Research*. forthcoming.
- Waterloo, S. F., Baumgartner, S. E., Peter, J., & Valkenburg, P. M. (2018). Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. *New media & society*, 20(5), 1813–1831.
- Yildirim, G., & Kübler, R. V. (2023). *Applied Marketing Analytics with R* (1st edition). London: SAGE.
- Zhang, M., & Luo, L. (2023). Can consumer-posted photos serve as a leading indicator of restaurant survival? *Evidence from Yelp*. *Management Science*, 69(1), 25–50.
- Evert de Haan** is associate professor of marketing at the University of Groningen, the Netherlands. His research focuses on the usage of mobile devices in the online customer journey, the effectiveness of online advertising, the consequence of ad blocking, and how to use customer metrics and other marketing metrics, including data extracted from unstructured data, to improve customer service and firm performance. His research has been published in journals such as the *Journal of Marketing*, the *Journal of the Academy of Marketing Science*, the *International Journal of Research in Marketing*, and the *Journal of Interactive Marketing*.
- Manjunath Padigar** is assistant professor of marketing at the Macquarie University, Australia. His primary area of research centers around strategic value implications of a firm's innovations stemming from its responses to external changes, such as emerging technologies and evolving customer preferences. His research has been published in journals such as the *Journal of the Academy of Marketing Science* and the *Journal of Business Research*.
- Siham El Kihal** is professor of marketing at the Vienna University of Economics and Business, Austria. Her research centers on understanding customer's purchase and post-purchase decisions in E-Commerce. As an empirical researcher, she aims to study research questions that are of high relevance for consumers, practitioners, and policy makers. Her research has been published in journals such as the *Journal of Marketing*, *Marketing Science*, and the *Journal of Retailing*.
- Raoul Kübler** is associate professor of marketing at the ESSEC Business School Paris, France. His current research focuses on user generated content, digital marketing, marketing return on invest measurement and big data in general. His methodological interests are in machine learning, sentiment analysis, and time series analysis. His research has been published in journals such as the *Journal of Marketing*, the *Journal of the Academy of Marketing Science*, the *Journal of Interactive Marketing*, and the *Journal of Cultural Economics*.
- Jaap E. Wieringa** is professor of research methods in business at the University of Groningen, the Netherlands. His current research is mainly on Data Science, Machine Learning, Marketing Model Building, Statistical Quality Control, Time series analysis, Diffusion modelling, and Marketing Analytics. His research has been published in journals such as the *Journal of Marketing*, the *Journal of Marketing Research*, the *International Journal of Research in Marketing*, the *Journal of Product Innovation Management*, and the *Journal of Business Research*.