Journal of Business Research 122 (2021) 915–925

Contents lists available at ScienceDirect

Journal of Business Research

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

Data analytics in a privacy-concerned world

Jaap Wieringa[a,*], P.K. Kannan[b], Xiao Ma[c], Thomas Reutterer[d], Hans Risselada[a], Bernd Skiera[e]

[a] *Department of Marketing, University of Groningen, the Netherlands*
[b] *Department of Marketing, Robert H. Smith School of Business, University of Maryland, United States of America*
[c] *Warwick Manufacturing Group, University of Warwick, United Kingdom*
[d] *Department of Marketing, WU Vienna University of Economics and Business, Austria*
[e] *Department of Marketing, Faculty of Business and Economics, Goethe University Frankfurt, Germany*

ABSTRACT

Data is considered the new oil of the economy, but privacy concerns limit their use, leading to a widespread sense that data analytics and privacy are contradictory. Yet such a view is too narrow, because firms can implement a wide range of methods that satisfy different degrees of privacy and still enable them to leverage varied data analytics methods. Therefore, the current study specifies different functions related to data analytics and privacy (i.e., data collection, storage, verification, analytics, and dissemination of insights), compares how these functions might be performed at different levels (consumer, intermediary, and firm), outlines how well different analytics methods address consumer privacy, and draws several conclusions, along with future research directions.

1. Introduction

Digital data–rich environments provide researchers and decision makers with unique opportunities for obtaining detailed, timely, multifaceted insights into customers' behaviors and opinions. These vast data, often called "big data," primarily can be characterized by their high volume, high velocity, and high variety (3Vs; Chintagunta, Hanssens, & Hauser, 2016). The sheer volume and level of detail of these data allow for unprecedented granularity in customer analyses; their velocity provides real-time insights; and the access to varied, previously unavailable or unexplored data sources provides new insights into the needs and wants of customers. These appealing elements in turn have increased the attention devoted to data analytics, in both academia and practice (Erevelles, Fukawa, & Swayne, 2016). Yet along with these promising potential benefits, data privacy issues have come to the fore, as signaled by the passage of the General Data Protection Regulation (GDPR) in the European Union, requiring firms to adapt their data-related procedures to stricter privacy regulations. The United States does not currently have similar legislation—and at the state level, only California has passed a privacy act, set to go into effect in 2020—but increased awareness of privacy concerns has prompted self-policing by many firms (Wedel & Kannan, 2016).

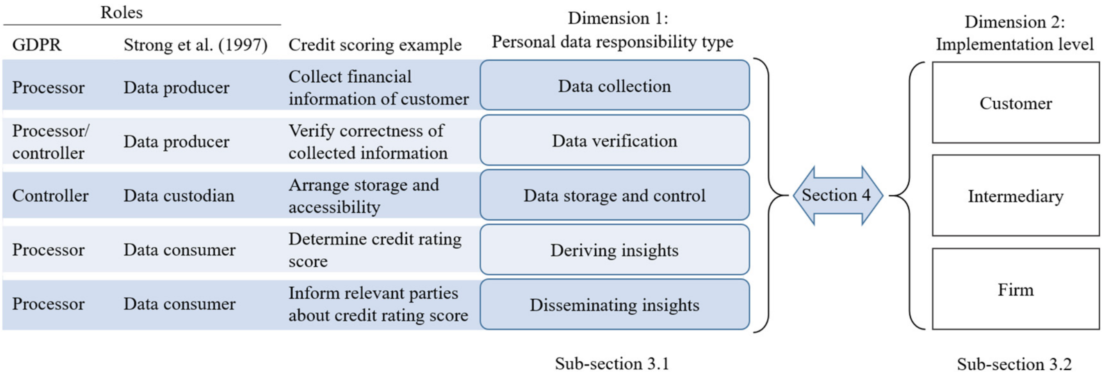
These combined trends in turn raise questions about the role and value of data analytics in a privacy-concerned society (Sivarajah, Kamal, Irani, & Weerakkody, 2017). Consumers and society could benefit from data-driven insights, but their privacy must be protected too. Although both these opposing forces have effects, the business press tends to focus on one side, whether stressing the potential of big data or warning about privacy concerns. Academic research has yet to detail the implications either. Therefore, with this study, we seek to gain insights into the best ways to conduct data analytics in a privacy-concerned world. We start by defining privacy and discuss the main privacy concerns of consumers. After that, we list and compare different functions related to data analytics and privacy (i.e., data collection, storage, verification, analytics, and dissemination of insights). Then we discuss how these functions might be performed at different levels (consumer, intermediary, firm). Finally, we outline how well different analytics methods address consumers' privacy. By combining these assessments, we draw several implications and conclusions, as well as directions for further research. In particular, we show that firms can implement various methods to collect, store, verify, and analyze big data while satisfying privacy needs and thus benefit from the information available. Even in the face of increasing privacy concerns, data analytics should be among the core capabilities that organizations pursue.

2. Privacy concerns of consumers

As Smith, Dinev, and Xu (2011) caution, no single concept of privacy exists, so we specify that for this study, privacy refers to information privacy, or access to individually identifiable personal data. This definition aligns with Westin's (1967, p. 7), in which privacy is

[*] Corresponding author at: PO Box 800, 9700 AV Groningen, the Netherlands.
E-mail addresses: J.E.Wieringa@rug.nl (J. Wieringa), PKannan@rhsmith.umd.edu (P.K. Kannan), X.Ma@warwick.ac.uk (X. Ma), Thomas.Reutterer@wu.ac.at (T. Reutterer), H.Risselada@rug.nl (H. Risselada), Skiera@wiwi.uni-frankfurt.de (B. Skiera).

https://doi.org/10.1016/j.jbusres.2019.05.005
Received 30 July 2018; Received in revised form 1 May 2019; Accepted 3 May 2019
Available online 18 May 2019



Roles			Dimension 1: Personal data responsibility type
GDPR	Strong et al. (1997)	Credit scoring example	
Processor	Data producer	Collect financial information of customer	Data collection
Processor/ controller	Data producer	Verify correctness of collected information	Data verification
Controller	Data custodian	Arrange storage and accessibility	Data storage and control
Processor	Data consumer	Determine credit rating score	Deriving insights
Processor	Data consumer	Inform relevant parties about credit rating score	Disseminating insights

Fig. 1. Central study dimensions.

"the claim of individuals … to determine for themselves when, how, and to what extent information about them is communicated to others." Information privacy protections seek to ensure personal data can be accessed only by those with the authorization to do so. The GDPR (Article 4) defines personal data as "any information relating to an identified or identifiable natural person," and further specifies that "an identifiable natural person is a person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." To prevent illegal, unauthorized uses of personal data, the GDPR requires specific efforts by firms and outlines consumers' rights over their personal data. This legislative act accordingly tries to address consumers' privacy concerns, which have emerged in response to the expanded data collection that takes place in digitalized, individualized markets. Privacy concerns reflect consumers' attitudes toward and concerns about the disclosure and processing of personal data (Malhotra, Kim, & Agarwal, 2004). They depend on the person disclosing the data, the context and setting of the data disclosure, and individual perceptions of the firm collecting the data (Bansal, Zahedi, & Gefen, 2016; Bergström, 2015). Martin and Murphy (2017) provide an excellent review of the earlier empirical findings and theoretical underpinnings on the role of privacy in the vast privacy research literature.

Beke, Eggers, and Verhoef (2018) and Bansal et al. (2016) both offer frameworks to link firms' privacy practices to consumers' privacy perceptions and concerns, which determine those consumers' behaviors and intentions to disclose. The decision to disclose personal data results from consumers' considerations of both negative and positive potential outcomes. For example, disclosing personal data could benefit consumers by increasing their access to personalized, potentially enhanced services that otherwise would be costly to obtain. Yet as Trepte and Reinecke (2011) outline, the negative consequences of disclosures include risks of unauthorized access, whether due to data breaches or unauthorized data sharing with other firms, unknown to the consumer, that could lead to identity theft or other data abuses (Martin, Borah, & Palmatier, 2016). The trade-off of these consequences implies a privacy calculus, such that consumers tend to share personal data with firms if the benefits outweigh the risks (Dinev & Hart, 2006). This privacy calculus also depends on the type of disclosed information and the ways personal data are collected, stored, and used (Beke, Eggers, Verhoef, & Wieringa, 2018).

If privacy concerns about a firm or a specific personal data disclosure episode lead consumers to refuse to share their data, getting consent to collect personal data becomes very challenging, compared with such efforts in relation to consumers with fewer or no privacy concerns. Yet GDPR demands that firms obtain consumers' consent, such that privacy concerns have direct effects on firms' ability to collect, process, and analyze personal data. Such data analytics have become critical to firms' service delivery though, especially in their attempts to optimize and personalize customer experiences by anticipating and satisfying their needs. As Beke, Eggers, and Verhoef (2018) suggest and the GDPR now requires, firms should provide consumers with transparent explanations about the data they collect and how they use them, as well as grant consumers some control over the disclosure. In turn, consumers can make more informed decisions about whether to share information, thereby affecting the amount of data disclosed, according to whether a firm provides a detailed explanation or not. Beyond this basic consideration, firms can mitigate privacy concerns and increase data disclosures by adopting different approaches to their data processing activities, as we discuss next.

3. Responsibilities for personal data and analytics

The generation and use of personal data, as is required for efficient interactions between consumers and firms, consists of several steps and processes. We identify five main steps: data collection, data verification, data storage and control, deriving insights, and disseminating insights. The responsibility for each of these steps might be assigned to or claimed by different parties in consumer–firm interactions. Accordingly, we distinguish three levels that might take responsibility for implementing each step: consumer-level, intermediate-level, and firm-level actors. Fig. 1 summarizes the two dimensions that we employ to structure this discussion, that is, the type of personal data responsibility and the level to which each responsibility is assigned.

3.1. Five types of personal data responsibilities

Our grouping of personal data responsibilities reflects the GDPR, which distinguishes controller and processor roles for dealing with personal data, as defined in its Article 4:

- A controller is "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data."
- A processor is "a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller."

Strong, Lee, and Wang (1997) propose a similar categorization but identify a separate role of data generation in data manufacturing systems. They thus specify three labels: data producers (people, groups, or other sources that generate data), data custodians (people who provide and manage computing resources to store and process data), and data consumers (people or groups that use data). Each role takes several tasks, such that data producers engage in data production processes; data custodians are linked to data storage, maintenance, and security; and data consumers adopt utilization processes, which may involve data aggregation or integration. Rather than focusing on these roles, we group the corresponding tasks and processes into five personal data responsibilities in Fig. 1.

Data collection is the first personal data responsibility; it relates clearly to a data producer role, in that it involves the generation of personal data. As the first step in the chain of data processing steps, this responsibility offers unique opportunities for implementing privacy measures early in the process, which can ensure "privacy by design" (GDPR, Article 25). According to the GDPR, data collection is a responsibility of the processor.

The second responsibility is data verification, which is the first principle of data processing in the GDPR. Article 5-1a requires personal data to be processed lawfully, fairly, and in a transparent manner relative to the data subject. This requirement also is expounded in Recital 71 of the GDPR, which mandates that inaccuracies in personal data must be corrected and that the risk of error is minimized. Data verification is strongly linked to the data producer role but also might be important for data custodians or consumers. In the GDPR, data verification is the shared responsibility of the processor and the controller.

As the third responsibility, we distinguish data storage and control, strongly tied to the data custodian (Strong et al., 1997) and controller roles (GDPR). It involves a broad range of tasks, including organization, structuring, storage, disclosure by transmission, dissemination, restriction, erasure, and destruction of personal data.

Analyzing data to obtain insights is a fourth responsibility. In many cases that require insights from personal data, the raw data likely need to be processed to generate the desired insights. Processing may involve very simple operations, such as summing or averaging, or it could encompass extensive econometric modeling efforts. This responsibility relates to the processor and data user roles.

Finally, the fifth responsibility that we identify is disseminating insights. It represents the final stage associated with the use of personal data, in which these data or the insights they provide get communicated to other stakeholders. From a privacy perspective, this responsibility is especially crucial, because it explicitly involves sharing information with different parties. In a privacy-concerned world, this fifth responsibility links closely to the first responsibility. For example, the GDPR closes the loop by allowing data collection only for specific purposes. Before collecting any data, firms must consider which insights they seek to generate and disseminate later. This responsibility therefore corresponds to the GDPR processor role or to Strong et al.'s (1997) data user role.

To illustrate these responsibilities, consider the task of determining a credit rating for a consumer who wants to apply for a new credit card. To avoid privacy issues, the five personal data responsibilities need to be arranged properly. The collection of customer-level financial data provides the input for determining the credit rating. To ensure a proper approval decision, the second responsibility is to verify the trustworthiness of these data. Because of the sensitive nature of financial data, secure storage and restricted access must be applied as the third responsibility, involving both the consumer and the credit card firm. Subsequently, the fourth responsibility is to determine the credit rating, an insight derived from the available data. Finally, the insight gets shared with (other) relevant players in the interaction, to support an informed approval decision.

3.2. Three implementation levels

These five personal data responsibilities can be delegated to either party involved in a personal data exchange. In the credit rating example, it may be the consumer's responsibility to collect and provide data such as an income statement and overview of outstanding debts (in addition to other data that will be collected by the firm), and the firm may take primary responsibility for the other four data responsibilities. This situation is relatively common, in that the majority of personal data responsibilities tend to be implemented at the firm level, but this situation is changing. In a digitally connected world, technological advances empower consumers to produce and consume information and insights (Van Bruggen, 2018) and take active control over the network that connects consumers and suppliers (Wuyts, 2010). Such empowerment benefits consumers and can improve business results (Wright, Newman, & Dennis, 2006).

Customers and firms are two obvious parties to the exchange that take on the five personal data responsibilities. However, considering only this dyad represents an overly narrow view. Multiple actors influence consumer–firm data exchanges (Henderson & Palmatier, 2010), so we identify a third level that can operate as an intermediary and handles one or more personal data responsibilities. In the credit rating example, third parties such as Equifax offer credit verification services that can facilitate all five data responsibilities. The intermediary role thus can be fulfilled by firms, as well as other trusted third parties, such as agencies or online public communities (e.g., blockchain community). These intermediaries might start out as "first-party" data processors, such that they collect and derive insights from data that they obtain directly from their customers, then transform into intermediaries when they release personal data or insights to other parties. Another type, so-called data brokers, only collect personal data from other firms, not directly from consumers. They produce insights by combining personal data across multiple sources, then disseminate those insights.

With regard to this latter group of intermediaries, consumers may benefit from data broker practices, such as if they help shoppers find products and services they prefer, but their practices have come under scrutiny, due to associated privacy concerns. The U.S. Federal Trade Commission (FTC) recently investigated nine data brokers, representing a cross-section of the industry, and concluded that they obtain and share vast amounts of consumer information, in some cases behind the scenes and largely without consumers' knowledge (Federal Trade Commission, 2014). Furthermore, the FTC report notes that personal data often pass through multiple layers of data brokers who share data for unspecified or unanticipated uses. For example, using a customer's data to identify her as a "Biker Enthusiast" could be a meaningful insight for targeted offers or discounts on motorcycles, but it also could signal a higher risk category to a potential insurance provider. Other privacy concerns arise from unnecessary and indefinite data storage; the FTC report notes the limited extent to which data brokers currently offer consumers choices about their data, most of which are invisible and incomplete.

Including intermediaries as a separate implementation level also can be justified by the size of this industry and the amount of personal data they process. An estimated 4000 data brokering companies operate worldwide (World Privacy Forum, 2013). Acxiom, one of the largest, has 23,000 servers collecting and analyzing data about 700 million consumers worldwide, with up to 5000 data points per person (Singer, 2012; Wolfie, 2017). Personal data now account for 36% of data-brokering activities globally, both legal and illegal (Transparency Market Research, 2017).

Furthermore, intermediaries have fundamentally different privacy incentives, relative to consumers or first-party data processors. That is, their primary interest is the resale value of personal data and associated consumer insights, so they have no natural motivation to restrict any collation or analysis of personal data. Instead, they are limited mainly by legislation being developed to mitigate adverse consequences for consumers. In Europe, data brokers are closely regulated by the GDPR; federal regulation attempts have thus far been less successful in the United States. For example, the Data Broker Accountability and Trust Act, which would require data brokers to establish procedures to ensure the accuracy of collected personal information, has been introduced to Congress twice but never passed (Data Accountability and Trust Act, 2011). On the state level, the varied efforts exhibit distinct levels of stringency and success (e.g., General Assembly of the State of Vermont, 2018; South Carolina General Assembly, 2018).

Table 1
Current implementation levels of personal data responsibilities.

Personal data responsibilities	Implementation level		
	Customer	Intermediary	Firm
1. Data collection	+	+	+ +
2. Data verification	+/−	+	+ + +
3. Data storage and control	−	+	+ + +
4. Deriving insights	− −	+ +	+ + +
5. Disseminating insights	−/+	+	+ +

3.3. *Current implementations of data responsibilities*

Table 1 illustrates the current implementation levels of the five responsibilities; more plus (minus) signs in a cell indicate that, in general, the level in that column takes a stronger (weaker) role in ensuring the personal data responsibility associated with that row.

This evaluation is strongly context dependent; the distribution of responsibilities in Table 1 does not hold for all countries or industries or all types of data. However, the overall conclusions that can be drawn from Table 1 offer some valuable insights; in particular, it indicates that firms currently handle most personal data responsibilities, especially after data collection. They may outsource these duties to intermediaries to some extent, but customers typically engage only in data collection and data verification. Consumers' roles for data storage and control, as well as in deriving insights, are typically minor. They might be somewhat stronger for disseminating insights though, if consumers share the raw data.

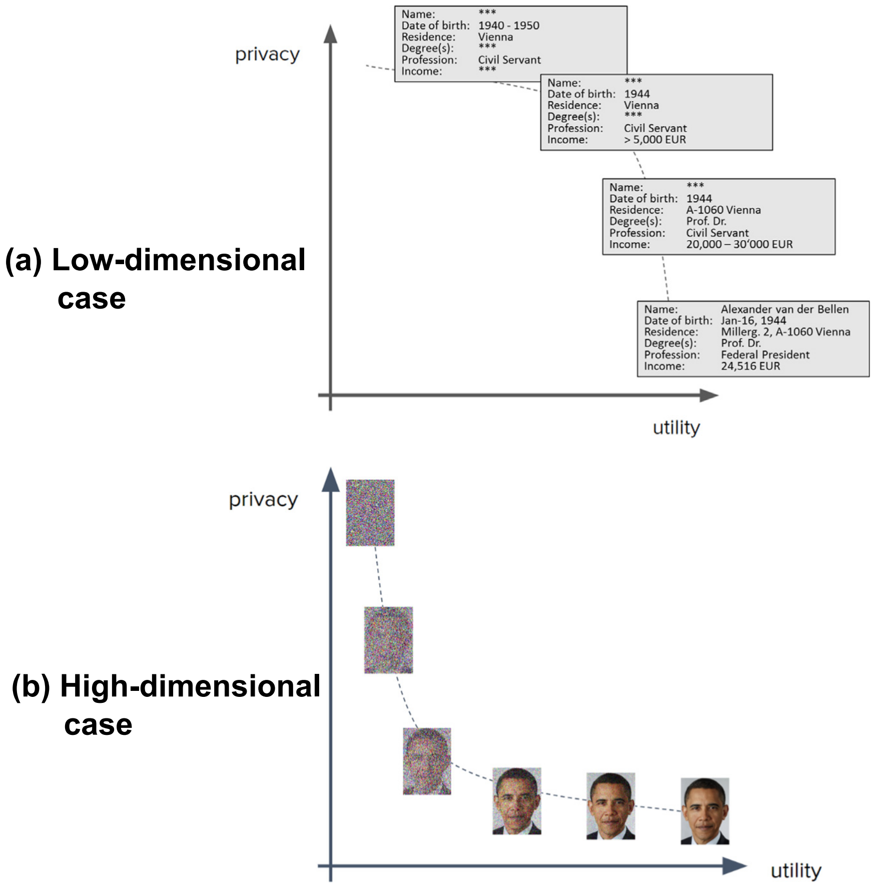
4. Comparison of implementation levels for personal data responsibilities

In this section, we discuss, for each personal data responsibility, the advantages and disadvantages of each possible implementation level. We also seek to identify key changes in the importance of their roles across responsibilities that currently or are likely to take place. Based on these predicted changes, we identify several research areas. For each responsibility, we describe solutions, open questions, and guidelines for practical applications.

4.1. *Data collection*

In a digital world, with interconnected customers and complex, multifaceted interactions with firms, data collection is not limited to a simple process of gathering potentially relevant information; it encompasses the permanent integration of multiple data sources in data warehouses and the management of their links. For example, both online and offline service providers accumulate vast customer and user data automatically, from distributed digital systems (e.g., social media, bookings, online review platforms), which can readily be combined. Thus, data collection in a digital, data-rich environment is an ongoing process that ends with data provision, requiring further storage or processing. In such an environment, the risk of damage to brand value (e.g., Facebook and Cambridge Analytica case) and customer trust are legitimate reasons to increase personal data protections.

The corresponding responsibilities traditionally accrue to the firm or data intermediary, which likely uses one of several data anonymization techniques, as we detail here. The ideal methodology to protect sensitive data would ensure that the data cannot be traced back to



Name: ***
Date of birth: 1940 - 1950
Residence: Vienna
Degree(s): ***
Profession: Civil Servant
Income: ***

Name: ***
Date of birth: 1944
Residence: Vienna
Degree(s): ***
Profession: Civil Servant
Income: > 5,000 EUR

Name: ***
Date of birth: 1944
Residence: A-1060 Vienna
Degree(s): Prof. Dr.
Profession: Civil Servant
Income: 20,000 – 30'000 EUR

Name: Alexander van der Bellen
Date of birth: Jan-16, 1944
Residence: Millerg. 2, A-1060 Vienna
Degree(s): Prof. Dr.
Profession: Federal President
Income: 24,516 EUR

Fig. 2. Trade-off between data utility and protection level for two cases.
(Source: www.mostly.ai).

individuals but still retain most of its utility or commercial value. The trade-off of risk and returns ultimately depends on the technique used for data anonymization (Schneider, Jagpal, Gupta, Li, & Yu, 2018).

The GDPR regulations do not specify processes for anonymization, but the outcome must be irreversible. Thus, pseudonymization is not an eligible technique that can comply with GDPR data protection standards. It consists of removing or hashing personally identifying information (e.g., name, email, social security number; Fig. 2) from a data set. As such, it merely reduces the direct link of a data set to the original identity of a data subject and, though it might offer a security measure, cannot be qualified as effective anonymization. Examples of pseudonymized data include social network ties (Narayanan & Shmatikov, 2009), location data points (De Montjoye, Hidalgo, Verleysen, & Blondel, 2013), or combinations of simple demographics (Sweeney, 2000) that allow for the re-identification of individual users.

Privacy concerns related to so-called first-party data, such as customer characteristics and purchase histories collected by customer relationship management systems, are relatively "controllable," from both customer and firm perspectives. Recently emerging decentralized technologies give consumers (in their roles as data owners) more control over whether and how their data may be used. For example, the transparency offered by blockchain's ledger-based technology might help mitigate consumers' concerns about how their data are being processed by marketers and advertisers (Ghose, 2018). Cloud-based services like the personal data micro-servers offered by the Hub of All Things (Section 4.3) also offer opportunities to shift control over personal data back to individual customers.

Privacy concerns gain even more relevance if the firm supplements data it collected from customers with data from another partner, such as media providers, social networks, or marketing research companies. Such secondary party data represent the proverbial 'new oil' of our increasingly digitalized economy. It enriches the firm's own customer database (and thus enhances sophisticated target marketing actions) and monetizes data collected by external providers. Against this background, the increase in the commercial value of the data, achieved by creating synergies among data-sharing parties, comes at a price of protecting at least some aspects of customers' sensitive data. Thus, the move from fully identifiable to anonymized personal data can be driven by privacy costs, which include the risk of damages to the firm's brand value or customer trust, legal penalties, and costly regulation. We distinguish non–model-based and model-based approaches for doing so (Little, 1993; Reiter, 2005; Schneider, Jagpal, Gupta, Li, & Yu, 2017).

4.1.1. Non–model-based approaches to data protection

Some simple techniques rely on generalizing (e.g., aggregating, recoding or top-coding attribute values), data swapping (i.e., changing variable values), suppression of personal identifiers, or some combination thereof. Another group of non–model-based methods employ randomization to protect micro-data by adding random noise, applying permutation techniques to alter values within a data set, or post-randomizing categorical variable labels (e.g., Gouweleeuw, Kooiman, Willenborg, & De Wolf, 1998). These widely used methods are particularly popular among governmental or statistical agencies and easily available in open-source toolkits like ARX (arx.deidentifier.org) or the R-package sdcMicro (Templ, Kowarik, & Meindl, 2015).

Some non–model-based data alteration techniques can increase data anonymity considerably. In particular, suppression and generalization techniques aim for the so-called *k*-anonymity property; it applies to a specific data release if an individual subject contained in the release cannot be distinguished from at least $k - 1$ other individual also included in the release (Samarati, 2001). This property provides some basic privacy safeguards, but it remains vulnerable to, for example, homogeneity attacks (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007), background knowledge, or intersection attacks (Francis, Eide, & Munz, 2017) when multiple, complementary data sets are released. Furthermore, *k*-anonymity typically can be warranted only for a very limited number of attributes, because the unique combinations of attribute values grow exponentially with the number of attributes (i.e., the "curse of dimensionality").

These approaches also tend to come at the price of a substantial decrease in data utility, which impairs their commercial value (Duncan, Keller-McNulty, & Stokes, 2001; Rust, Kannan, & Peng, 2002). For example, adding random noise introduces measurement error that stretches marginal distributions and attenuates regression coefficients (Yancey, Winkler, & Creecy, 2002); top-coding distorts Gini coefficient estimates (Kennickell & Lane, 2006); and swapping can destroy the correlations of swapped and non-swapped variables if used too intensively (Drechsler & Reiter, 2010).

The utility–disclosure risk trade-off also can be formalized, according to the differential privacy concept (Dwork & Roth, 2014). It quantifies the marginal impact of including an individual in a data set on the outcome of a randomized algorithm (e.g., query, summary statistic). The preceding sanitization approaches typically perform relatively poorly in terms of differential privacy, especially if they apply to high-dimensional data sets with highly intercorrelated structures (Narayanan & Shmatikov, 2008). Fig. 2 illustrates this notion for two simple cases.

Panel a in Fig. 2 represents a concave downward relationship between privacy protection and utility for low-dimensional data, such as when the personal identifiers of an individual (here, the Federal President of Austria) are characterized by just a few attributes like name, date of birth, and residence. With such low-dimensional data, the privacy gains increase notably simply by suppressing one or two attributes or generalizing some remaining attributes. In addition, the information loss is only moderate, so the data retain their usefulness for informing some query or release. This relationship between privacy gains and information loss changes completely in a case of high-dimensional, highly correlated data. Panel b in Fig. 2 illustrates such a typical case for an image of Barack Obama, represented by a high-dimensional arrangement of pixel values. The specific arrangement—or more formally, the correlational structure—of these pixels give the image meaning and makes it personally identifiable. To prevent re-identification of the individual, the image would need to be generalized (in Fig. 2, by adding noise) to such a level that the result becomes useless. As this comparison illustrates, simple, non–model-based methods might be useful for protecting data that are characterized by just a few attributes, but they need to be replaced by more sophisticated approaches when the task is to protect more complex, multidimensional marketing data structures without destroying their commercial utility.

4.1.2. Model-based approaches to data protection

More sophisticated model-based approaches for data protection typically aim to generate customer-level, "synthetic" data by mimicking an underlying data-generating process. The synthetic data-generating "engines" perform multiple imputation and bootstrap procedures to address missing data (Rubin, 1993), based on either a statistical (e.g., Bayesian) model that generate a posterior predictive distribution according to some protected, underlying probability distribution of the original data, or else some advanced machine or a deep learning approach (for an overview, see Surendra & Mohan, 2017).

Schneider et al. (2017, 2018) provide two recent marketing applications of such synthetic data generation. In one, they employ a Dirichlet-multinomial model to generate synthetic count data and thus protect histograms of market segment sizes with flexible privacy levels. They apply this model to a segmented customer base from an online ticket firm. In another application, they propose a Bayesian random effects model to estimate protected SCAN*PRO market-response functions and illustrate how data providers, such as the market research company ACNielsen, could use this model to release useful but still privacy-protected, store-level data to data users. In this model, the data provider learns which variables collected from stores might disclose store identities and thus that need to be protected through a transformation into synthetic data, before releasing the data to users. These contributions are promising starting points for identifying ways that firms, data processors, and intermediaries can protect privacy-sensitive data while still preserving their commercial value. However, both approaches tackle very specific problems, and it remains questionable if they are flexible enough to deal with more general cases, such as those characterized by high-dimensional, correlational data.

More flexibility, and thus a broader scope of applications, might result from data synthetization, which explicitly accounts for the multivariate interrelationships of high-dimensional data structures. Promising research in this direction relies on multivariate Gaussian copulas (e.g., Patki, Wedge, & Veeramachaneni, 2016); another source is the machine learning community, which benefits from significant progress in generative deep neural networks. For example, Karras, Aila, Laine, and Lehtinen (2017) train generative adversarial networks (GANs), using a set of real celebrity faces, and demonstrate that the network can generalize the structure and composition of the training data. After convergence, the network weights generate an arbitrary number of synthetic images that preserve the main characteristics of the training data but recompile them in a way that protects the original entities (i.e., in their case, real celebrities).

Such deep learning architectures also might be able to resolve complete information losses associated with efforts to protect high-dimensional and intercorrelated data (Panel b, Fig. 2). A differentially private version of a deep learning architecture with convolutional layers (Abadi et al., 2016), implemented in TensorFlow, as well as a GAN-based privacy-preserving generative deep learning approach (Beaulieu-Jones et al., 2018), represent promising attempts in this direction.

In summary, significant progress in recent years provides increasing protection of the private data collected from customers. Despite interesting, though also vague and debatable, potential opportunities offered by newly emerging technologies (e.g., blockchain, personal micro-servers), including options to grant customers complete control over their personal data, the primary responsibility for implementing privacy protections remains with firms and intermediaries. Pseudonymization and simple rule-based methods for data anonymization typically are not sufficient to protect complex, dynamic, multidimensional marketing data. Rather, research remains necessary, in particular to develop advanced model- or machine-learning based approaches that can generate synthetic, individual-level, high-dimensional data that "mimic" real-world information.

4.2. Data verification

Before the collected data can undergo further processing, it should be clear to all stakeholders that the data provided are of such quality that further processing is useful. Without sufficient data quality, users may lack confidence in the data (Martin et al., 2008), which can have financial impacts of up to 15–25% of operating profits (Olson, 2003), diminish customer confidence and satisfaction, hinder productivity, and even have serious consequences for risk and compliance (Loshin, 2011). Data quality relates closely to veracity, sometimes referred to as the "fourth V" of big data. Veracity implies the trustworthiness and the accuracy of the data (Mittal, 2013). Regardless of which level is responsible for collecting data, the other parties must be sufficiently convinced of the veracity of any data they receive. Yet there are multiple potential sources of error (Pendyala, 2018), such as:

- Incorrect observations, by humans or sensors. For example, if the value of a car is needed to determine car insurance fees, this data point may be inaccurate if humans provide the estimate, because they are exposed to subjective considerations, such as the emotional value that an owner attaches to the car or a desire to keep the fees low. Yet a human estimate is beneficial too, because it can incorporate multiple aspects to determine value. A sensor measurement, such as the number of miles the car has driven, is more objective but might be insufficient and one-dimensional. Both types of measurements thus may lead to data inaccuracies.
- Incorrect translation or extraction (e.g., automatic extraction of information from html). When a data point needs to be copied from one medium or format to another, the process can induce errors. For example, digitizing data points on paper using optical character recognition may produce some incorrect output. Or, if file formats are not compatible, conversion errors may emerge from transferring data points from one format to another.
- Incorrect entry, either manually or by sensors. Even if the data points are correctly observed or extracted, entry errors may occur, such as due to typos.

An extensive range of interrelated tools can help ensure that collected data are accurate and trustworthy. Maletic and Marcus (2009) outline a three-step data verification process: (1) define possible types of error, (2) identify error instances, and (3) correct them. We discuss several tools and techniques that can be used in each of the steps. Subsequently, we indicate whether the corresponding data verification tools should be applied at the customer, the intermediary or the firm level.

4.2.1. Metadata repositories

Metadata repositories help prevent data inaccuracies by ensuring that all data elements are named, with a clear definition (Loshin, 2011), and by accepting only those data elements that fulfill these data definitions. Thus, they limit the collection of inaccurate data and provide a means to verify erroneous data elements. The data definitions from metadata repositories provide a point of reference for the first two steps in the data verification process but are also important in the third step, in correcting erroneous data points.

4.2.2. Data profiling

Data profiling relies on analytical methods that review the data to develop a thorough understanding of their content, structure, and quality (Olson, 2003). As this definition illustrates, data profiling thus can serve multiple purposes. For example, it enables inferences of metadata and can identify anomalies, thereby contributing to the first and second steps of the data verification process, respectively. Loshin (2011) describes step-by-step, column, table, and cross-table analyses that identify data issues, and Maletic and Marcus (2009) point to clustering, pattern detection, and association rules that can recognize data errors (especially if these errors manifest as outliers).

4.2.3. Data monitoring

Even sophisticated methods for preventing or removing erroneous data points cannot completely eradicate all data issues, so some level of data inaccuracy will exist (McGilvray, 2008). Data monitoring provides a way to manage this uncertainty and identify whether the accuracy and trustworthiness of the data are sufficiently high to warrant further processing. With this ongoing error detection, data monitoring strongly reflects the second step in the data verification process.

Data monitoring tools might be transaction oriented or database oriented (Olson, 2003). The former identify issues in individual transactions, before data are stored or processed further. The latter periodically inspect stored data to find issues, often using control charts (Loshin, 2011). Berti-Équille and Borge-Holthoefer (2015) present a broad overview of methods for truth discovery, fact checking, trust computation, and detecting misinformation in networked systems. The preventive nature of transaction-oriented data monitoring might offer some advantages but processing each transaction can be too slow if it involves too much checking (Olson, 2003). Transaction-oriented data monitoring also is less effective than database monitoring, because problems might not be visible in individual transactions but could surface through assessments of counts, distributions, or aggregations.

Thus, data monitoring is most effective if it combines transaction and database monitoring.

4.2.4. Implementation levels associated with verification of accuracy and trustworthiness

Reviews of the quality of data being collected might span all three implementation levels that we identify. If customers are responsible for data collection, a firm or intermediary that takes on the further processing of those data will want to identify any anomalies or erroneous data points and correct them. If an intermediary or firm is responsible for collecting data, it should be possible for the customer to check their veracity. However, the preceding verification tools are not equally well suited for the three implementation levels. For example, most customers interact with relatively few firms, which typically require different types of data exchanges. In addition, customer-level data storage options for recording transactions with firms are not well developed (as we discuss in the next section). The basis of comparison that customers can use to verify data is smaller than the one available to intermediaries and firms, and consequently, firms and intermediaries are potentially better equipped to engage in efficient, large-scale data verification processes than customers. In contrast, customers have better options for performing detailed verifications of individual data points.

Because firms mostly dictate the types of data that need to be exchanged to complete a transaction, the format and the type of data that customers observe or produce is more heterogeneous than the data processed by the firm (or intermediary). To avoid proliferations of definitions and names for the data elements, metadata repositories should not be developed by customers. Either intermediary agencies or firms should provide the definitions and variable labels, to ensure that the collected data elements meet standardization criteria and contain appropriate information, such that they are useful for further processing.

Data profiling also may be more efficient at the intermediary and firm levels than at the customer level. Developing an understanding of various data aspects generally requires analyses of vast amounts of data, such as comparing values across many customers. More data support the use of advanced, potentially more useful types of data profiling. Customers typically conduct less sophisticated data profiling, if at all, though this assessment might change if personal data storage solutions (Section 4.3) become more commonplace.

Regarding the possible implementation levels for data monitoring, we consider transaction monitoring and database monitoring separately. Transaction monitoring is appropriate for all parties involved in a transaction, even if the focus changes for customers versus intermediaries and firms. Because customers generally are involved in relatively few transactions, they are better equipped to monitor transactions in detail. In contrast, firms and intermediaries have better options to identify problems that surface from counts, distributions, and aggregations of personal data. Database monitoring also is less well suited for customers than for intermediaries and firms; customers rarely have access to large-scale databases.

4.3. Data storage and control

As we discussed in Section 3.2, data storage and control responsibilities are strongly linked to the controller role (Article 4, GDPR). Controllers act as custodians of personal data (Diaz, Tene, & Gürses, 2013) and must be able to demonstrate compliance with the principles for processing of personal data, according to Article 5 from the GDPR: lawfulness, fairness and transparency, data minimization, accuracy, storage limitation and integrity, and confidentiality of personal data.

Baxter, Aurisicchio, and Childs (2015) identify five affordances of control that jointly affect the level of perceived control and can support the GDPR principles. First, *spatial control* is defined an ability to manipulate objects through space. For intangible, digital, personal data, this affordance relates to an ability to influence the physical location of the data servers that contain the personal data (Kamleitner & Mitchell, 2018). Second, *configuration control* pertains to the manipulation of the data collection, storage, and processing conditions, such as the ability to change access rights to data. Third, *temporal control* can be defined as the ability to use the data when desired. Fourth, *rate control* is the power to adjust the amount of personal data being used. Fifth, *transformation control* relates to the ability to alter and process personal data.

The perceived level of control, according to the customer, depends on which party is responsible—currently, it tends to be the firm. Personal data collected by firms during transactions, through enabling devices such as wearables, or from online services such as social media usually are stored on firms' hardware or software, such that they become firm assets. Firms need to exercise spatial and configuration control over the infrastructure, for maintenance purposes. Consumers typically do not possess any legal or commercial power over the infrastructure and are not allowed to exercise full spatial or configuration control. However, the firm can give a consumer some level of control over data storage, so we define the level of spatial and configuration control as medium for data stored at the firm level.

Firms also can process and generate insights from these data to improve their services. For example, supermarkets can customize vouchers according to the needs and wants of individual customers, reflecting their observed shopping behavior. In addition, firms can centralize data collected from previously separate silos and combine them with wider data sets (e.g., weather, traffic) (Ng, Scharf, Pogrebna, & Maull, 2015). Typically, consumers cannot exercise full control over these data processing steps either; they might exercise indirect control through the consent they give to the firm. Therefore, we define the level of temporal, rate, and transformation control as medium for personal data stored at the firm level.

In Section 3.2, we specified some privacy risks associated with data brokers, which store personal data gathered from many resources, often without consumers' knowledge. In these cases, there is no direct link between the consumers to whom the personal data belong and the intermediary (Boudreaux et al., 2014). Thus, data storage at the intermediary level scores low on all five control affordances, and the transparency of control remains a major concern. Yet by recognizing consumers as the owners of their personal data, the GDPR enforces consumers' right to access to (co-)created personal data (Article 15.3). It mandates that firms and intermediaries provide copies of personal data to any requesting consumer, in a commonly used digital format. This requirement creates a rather disruptive shift of power toward consumers, for two main reasons. First, consumers can function as new, potentially better aggregators of their personal data, because they may claim personal data from all firms and intermediaries and centralize previously disparate data sets across these sources. Second, consumers gain a digitally processable record of their personal data. In principle, considering the advances in personal information management systems, consumers can act more like a firm and store, control, and process their own personal data, as well as actively participating in personal data exchanges. For example, on the Hub of All Things (https://hubofallthings.com), individual users can configure their own personal data storage infrastructure. Thus, for personal data stored at the consumer level, the consumer has full control across all five control affordances.

This discussion of the storage and control of personal data leads us to conclude that, across the five control affordances, personal data storage at the consumer level provides superior control to the consumer and offers individual control by design and by default, as required by the GDPR. Provided it complies with the GDPR, personal data storage at the firm level can offer a medium level of privacy and control to consumers. The data brokerage function of intermediaries instead limits their ability to preserve privacy and control for consumers.

4.4. Deriving insights from data

Kotler and Armstrong (2014, p. 125) define customer insights as "fresh understandings of customers and the marketplace derived from marketing information that become the basis for creating customer value and relationships." Thus, an insight is the result of some analysis, based on data, that goes beyond any individual data point. Deriving insights from data requires consideration of several interrelated factors, such as the specific ways the data collection is impacted, the challenges each situation poses for the firm and its analysis, the different types of analyses that might overcome these challenges, and the extent to which insights can be derived from various methodologies. For example, related to the first factor, government oversight and regulations could restrict the collection of specific data about potential customers. As a result, data collection is affected in four ways:

(a) Some individual-level data that were collected previously may not be legal to collect, so they are no longer be available as input for any analysis.
(b) In some cases, personally identifiable information may be scrubbed, leading to anonymization of the data.
(c) Some data may be available only at the aggregate level, such as the zip-code level rather than the household level.
(d) In some cases, data may be available at the individual customer level, with permission from the customer using an opt-in mechanism.

Such scenarios imply several challenges for deriving insights using data analytics techniques (Wedel & Kannan, 2016). First, the marketing analytics techniques need to be able to use minimized (data in a compressed form or subset of original variables) and anonymized data without losing their predictive and diagnostic power. Second, intermediaries should represent customers' interests in terms of how firms use their data for targeting and marketing purposes.

Several methods currently available can address these needs and extend the four specific data collection methods. For many conditions, Bayesian methods provide possible solutions. For example, if some variables are missing, assuming models used previously to make predictions are available, together with sufficient statistics (e.g., means, variances, cross-products, posterior distributions), they could be used for Bayesian updating and analysis as new data come in, without losing any information, even in the absence of the original data (Wedel & Kannan, 2016).

A good example of an application in this genre is Holtrop, Wieringa, Gijsenberg, and Verhoef's (2017) prediction of churn at the customer level, without using past data, based on a general mixture of the Kalman filters model. Another possible methodology relies on copulas (Danaher & Smith, 2011) to deal with endogeneity correction (Park & Gupta, 2012). If joint distributions can be retained, these methods provide useful inferences about the missing dimensions. Specifically, the copula provides parameters in a distribution function, similar to a variance–covariance matrix in the multivariate normal case. With enough observations on a few variables, using the marginal distribution of the variables along with the copulas, we can construct the missing values, though not with a view to protect privacy. Bayesian estimation methods then retain information from the marginal distribution and copulas from prior data, create estimates for missing values, and update the information for new data. Copulas have been used for geostatistical interpolations of unobserved locations, as an alternative to kriging (Bárdossy & Li, 2008), and they may provide similar insights in a context of missing data.

When only aggregate data are available, it is generally the case that aggregation is performed to preserve anonymity. Wedel and Kannan (2016) describe some examples; Steenburgh, Ainslie, and Engebretson (2003) fuse data from several sources at different aggregation levels, using a hierarchical Bayesian model. Musalem, Bradlow, and Raju (2008) instead use missing data imputation methods to obtain individual-level insights from aggregate data. Such data augmentation methods can estimate consumer-level insights from aggregate data in the context of data minimization, obviating the need for individual-level data and work with anonymized aggregate data. In a related context, Jerath, Fader, and Hardie (2016) examine the possibility of estimating customer-based models using aggregated data summaries alone, namely, repeated cross-sectional summaries of the transaction data (e.g., four quarterly histograms). These hybrid models perform as well as individual-level data in deriving insights into customer behavior, but they also prevent any identification of individual customers. Another promising source of individual-level insights could be agent-based modeling techniques (Rand & Rust, 2011), which simulate individual-level behaviors to align with aggregate-level data.

If instead data are available only for customers who opt in, data imputation methods can impute values for the missing data for customers who do not opt in. Some conditions need to be satisfied for such imputation to work (see Kamakura, Wedel, de Rosa, & Mazzon, 2003). People who opt in are self-selected customers, which may create endogeneity that requires consideration, if the results serve purposes other than prediction. Continued work is needed to develop models and algorithms for obtaining insights from these data while preserving customers' privacy.

Another challenge for overcoming the data limitations imposed by privacy regulations to obtain relevant insights is the rise of institutions that might function as intermediaries between the firm and customers (see Section 3.2). Such intermediaries take various forms, such as those detailed by Rust et al. (2002). Their primary task would be to collect information from customers and provide it, in a usable form, to firms while anonymizing customers' identities. In some variations, the intermediaries retain the identities and target customers on the firm's behalf. Such activities are similar to the practices of Google, Facebook, and innumerable display advertisement intermediaries, but a key difference emerges from the fiduciary role that intermediaries may need to serve, on behalf of customers. That is, they cannot take advantage of customers or customers' data for their own profit motives by misusing their data. Such intermediaries will fall under the strict oversight of government bodies, similar to financial advisors who advise customers according to their fiduciary duties.

Such institutions are evolving, though not in the same forms. For example, the Hub of All Things (see Section 4.3) allows customers to retain control of their data and provide them to firms after they assess the benefits of doing so. Such institutions should be encouraged by governments to protect customers' privacy and harvest data for legitimate business purposes, to match products and services with customers using insights derived from data.

4.5. Disseminating insights

The dissemination of insights is crucial for ensuring the impact of the analytics, whether within the firm (internally) or across its industry (externally). For privacy, external dissemination is especially interesting. Conditional on consumers' consent to collecting and analyzing their data for a specific goal (Sections 4.1 and 4.4), sharing the relevant insights internally should not violate privacy. Instead, we focus on issues related to external dissemination of insights and thus clearly distinguish insights from data (see Section 4.4). In turn, we consider three aspects of the dissemination of insights: who initiates the dissemination (initiator), who disseminates the insight (disseminator/sender), and to whom the insight is disseminated (receiver).

4.5.1. Initiator of external dissemination of insights

According to the GDPR, the consumer must be the initiator of insight dissemination in most cases. Consumers give firms permission to collect data for specific goals only, so firms may not use these data for any other purpose. By granting consent (i.e., opting in) for data collection for a certain goal, consumers indirectly initiate dissemination of the insights related to a specific goal to third parties. For example, telecom customers might opt in to allow the firm to collect calling data, enabling it to provide relevant insights for service improvements. The resulting insights, based on the data of all customers who opt in, will be disseminated to the net operator or any third party that contributes to service improvement. Thus, by opting in, the customer initiates the insight. Exceptions exist, such that firms or intermediaries can initiate dissemination. For example, the consumer credit rating agency Bureau Kredietregistratie (BKR) is required by Dutch law to register any loan previously offered to consumers. Before providing a new loan, organizations can contact BKR and obtain a profile of the consumer applying for that loan. Although strictly speaking, the consumer initiates the dissemination by taking the loan in the first place, it is a firm that actually requests insights from the intermediary.

4.5.2. Disseminators and receivers of insights

The dissemination of insights on the customer level is not common practice, because customers do not own the insights that firms generate. For example, a customer's churn probability with a telecom service operator typically is stored and owned by the telecom firm. As we argued in Section 4.3 though, we expect that storing data at the customer level will become increasingly common in the future, such that customers might give firms access to their data for a limited time and for specific purposes. For example, a customer might share a purchase history from online retailer A with online retailer B to support analyses that lead to the development of relevant personalized offers. In exchange for the access to these data, retailer B might provide benefits, such as extended delivery options. By giving customers control over the insights generated by their data, the insights become a sort of currency that customers can decide to exchange for benefits. A key issue though is the verification of these insights, similar to the data verification issues discussed in Section 4.2.

Intermediaries are the most likely to disseminate insights externally. For example, advertising agencies collect vast amounts of data about consumers who allow tracking of their browsing behavior. Ad agencies and platforms use insights derived from these data to customize advertising and optimize ad effectiveness for clients and consumers. Strictly speaking, the intermediaries do not disseminate the insights but rather use them to serve clients. This business model is sustainable under GDPR, as long as consumers are willing to opt in to receive customization. The trade-off between protecting privacy and benefiting from customization is the responsibility of the consumer.

Finally, firms are unlikely to disseminate insights externally. Under GDPR, firms need to be transparent about what data they collect and what they intend to do with them. Most insights get used internally. As noted, consumers might give a firm permission to share the insights with other parties, but instead, they appear increasingly likely to store insights individually and share them with other parties themselves. Thus, consumers have full control over their data and insights and potentially could benefit from them.

5. Conclusions, recommendations, and research agenda

In our effort to determine how best to conduct data analytics in a privacy-concerned world, we start by identifying five responsibilities for personal data and analytics (data collection, data verification, data storage and control, deriving insights, and disseminating insights), which can be implemented at three levels (customer, intermediary, and firm). With Section 3, we reveal that most responsibilities are allocated to the firm level. For each responsibility, we also consider how the implementation might be shifted to improve consumers' privacy, which we summarize in Table 2.

Table 2
Current and preferred implementation levels of personal data responsibilities.

Personal data responsibilities	Implementation level		
	Customer	Intermediary	Firm
1. Data collection	+ ➔ + +	+ ➔ + +	+ + ➔ + +
2. Data verification	+/− ➔ + +	+ ➔ +	+ + + ➔ + +
3. Data storage and control	−➔ + + +	+ ➔ +	+ + + ➔ + +
4. Deriving insights	− −➔ + +	+ + ➔ + +	+ + + ➔ + +
5. Disseminating insights	−/+ ➔ + +	+ ➔ +	+ + ➔ + +

In the first row of Table 2 (reflecting Section 4.1), we list all levels that can take the responsibility for collecting personal data in a privacy-friendly way. From our observation in Section 4.3 that data storage and control at the customer level provide superior privacy protection, with solutions already available, we recommend that this responsibility moves to the customer level. In contrast, privacy safeguards are relatively poor at the intermediary level, because intermediaries are not directly linked to the consumers whose personal data have been collected. Yet intermediaries often constitute a large industry, such as in advertising, and are ideally positioned to combine data from various sources, such that they can generate rich, potentially novel insights. Therefore, they have an influential role in delivering on the promises of big data, and we advise maintaining their responsibility for deriving insights, despite the potential privacy issues. As long as effective privacy legislation gets implemented, firms can safely take on a sizeable portion of this responsibility too. Noting their relevant role for data storage and control, consumers also should take on more responsibilities for data verification, deriving insights, and disseminating insights (rows 2, 4, and 5 in Table 2).

Generally, the evidence in Section 4 and Table 2 indicates that increasing the role of customers relative to responsibilities can alleviate privacy concerns. The focus does not need to shift entirely to customers though. Firms and, to a lesser extent, intermediaries still should shoulder an important portion of the responsibilities. Our overview in Section 4 highlights the available solutions that can facilitate the implementation of each personal data responsibility; these techniques do not necessarily require a shift from the intermediary or firm level to the customer level to avoid privacy issues. In turn, firms and intermediaries can generate customer insights from personal data, while still respecting customers' privacy.

Together with this positive overall summary, we identify many areas that warrant further research. With respect to the data collection responsibility, we mainly identify methodological opportunities. A promising area is to develop better approaches for generating synthetic, individual-level, high-dimensional data that mimic real-world entities. Specifically, we call for the development of advanced model or machine learning approaches that are able to generate data that is applicable in a broader range of marketing applications than previously developed approaches. In turn, we encourage research that investigates the trade-off between privacy preservation and information loss, as well as the development tools that can balance this trade-off. We welcome efforts to make companies' and public institutions' uses of sensitive data more transparent (e.g., blockchain, personal data exchange services), to avoid that outcomes of newly developed approaches are neither understood, nor accepted by consumers.

Considering that most of the data verification techniques are currently available only to firms and intermediaries, techniques for large-scale, real-time data verification and protection need to be developed, ideally as part of the data collection process. When consumers take on more personal data responsibilities, they need to be better equipped to investigate data veracity, given that the type of data that they collect has fundamentally different characteristics than that of firms and intermediaries (see Section 4.2). To this end, the data verification tools that firms and intermediaries currently employ need to be adapted to suit the data verification needs of consumers. We thus call for research on data verification tools on all implementation levels to stimulate that only relevant and correct data will be stored.

Our discussion in Section 4.3 suggests that customers should be more involved in data storage and control. Consequently, intermediaries need tools to increase the level of control granted to consumers. In particular, intermediaries should find ways to strengthen their link with consumers while still protecting their privacy. Currently, intermediaries directly benefit from consumers' data, whereas consumers generally benefit only on the long run or in an aggregate sense (e.g., development of a free email service, based on their data). If consumers retained full control over their data, they might consciously provide access to only a limited set of intermediaries. In that case, consumers' personal data would function as currency, and they could make individual trade-off assessments between protecting their privacy and obtaining the benefits of sharing some data. We call for research that attempts to establish valuations of data points, aggregates, or insights. Furthermore, customer empowerment could have substantial impact on customer–firm or customer–intermediary relationships. We believe research in such areas would be fruitful.

Expanded customer roles in data storage and control also provide challenges associated with the responsibility for deriving insights. As we outline in Section 4.4, firms and intermediaries must be able to handle customer data that becomes available intermittently. In addition, customers self-select what data they share, and the endogeneity issues associated with forward looking behavior of customers in these data sharing decisions should be carefully taken into account when deriving insights from these data. We encourage further research into models that accommodate these issues.

Several open research areas also relate to the personal responsibility of dissemination. What value do consumers attach to their data? How much would consumers be willing to pay to keep or gain control over their data (e.g., in relation to content providers)? How much would they require firms to pay them to grant access to their data (e.g., online retailers)? How should firms value their access to consumers' data and insights? Which incentive schemes are most effective to motivate consumers to share data with intermediaries and firms?

The starting point for this research was the recognition of a conventional wisdom that assumes data analytics and privacy protection contradict each other. We hope to have shown that such a view is too narrow, because firms can implement a wide range of methods that satisfy different degrees of privacy, while still enabling them to address all data analytics responsibilities.

Acknowledgements

We thank Jack Pool, Kees Jan Roodbergen and Yory Wollerich for sharing their ideas with us during the Thought Leader Conference on Digital Business Models, organized by the Groningen Digital Business Centre, in April 2018. We thank the participants of the same conference for their constructive comments. We are also grateful to the special issue editor and two anonymous reviewers for their invaluable feedback during the review process.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). *Deep learning with differential privacy.* ArXiv:1607.00133 [Cs, Stat]308–318. https://doi.org/10.1145/2976749.2978318.

Bansal, G., Zahedi, F. M., & Gefen, D. (2016). Do context and personality matter? Trust and privacy concerns in disclosing private information online. *Information & Management, 53*(1), 1–21. https://doi.org/10.1016/j.im.2015.08.001.

Bárdossy, A., & Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research, 44*(7), https://doi.org/10.1029/2007WR006115.

Baxter, W. L., Aurisicchio, M., & Childs, P. R. N. (2015). A psychological ownership approach to designing object attachment. *Journal of Engineering Design, 26*(4–6), 140–156. https://doi.org/10.1080/09544828.2015.1030371.

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., & Greene, C. S. (2018). Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv,* 159756. https://doi.org/10.1101/159756.

Beke, F. T., Eggers, F., & Verhoef, P. C. (2018). Consumer informational privacy: Current knowledge and research directions. *Foundations and Trends® in Marketing, 11*(1), 1–71. https://doi.org/10.1561/1700000057.

Beke, F. T., Eggers, F., Verhoef, P. C., & Wieringa, J. E. (2018). *Consumers' privacy calculus: The PRICAL index development and validation. Working paper.* University of Groningen.

Bergström, A. (2015). Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses. *Computers in Human Behavior, 53,* 419–426. https://doi.org/10.1016/j.chb.2015.07.025.

Berti-Équille, L., & Borge-Holthoefer, J. (2015). Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics. *Synthesis Lectures on Data Management, 7*(3), 1–155. https://doi.org/10.2200/S00676ED1V01Y201509DTM042.

Boudreaux, D. J., et al. (2014). Letter to the FTC. International Center for Law and Economics. Available at: http://laweconcenter.org/images/articles/icle_ftc_nn_letter_final.pdf.

Chintagunta, P., Hanssens, D. M., & Hauser, J. R. (2016). Editorial—Marketing science and big data. *Marketing Science, 35*(3), 341–342. https://doi.org/10.1287/mksc.2016.0996.

Danaher, P. J., & Smith, M. S. (2011). Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Science, 30*(1), 4–21. https://doi.org/10.1287/mksc.1090.0491.

Data Accountability and Trust Act (2011). Available at: https://www.govtrack.us/congress/bills/112/hr1707.

De Montjoye, Y.-A.d., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports, 3,* 1376. https://doi.org/10.1038/srep01376.

Diaz, C., Tene, O., & Gürses, S. (2013). Hero or villain: The data controller in privacy law and technologies. *Ohio State Law Journal, 74*(6), 923–964.

Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research, 17*(1), 61–80. https://doi.org/10.1287/isre.1060.0080.

Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association, 105*(492), 1347–1357. https://doi.org/10.1198/jasa.2010.ap09480.

Duncan, G. T., Keller-McNulty, S. A., & Stokes, S. L. (2001). *Disclosure risk vs. data utility: The R-U confidentiality map. Technical report.* U.S. National Institute of Statistical Sciences31.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407. https://doi.org/10.1561/0400000042.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. *Journal of Business Research, 69*(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001.

Federal Trade Commission (2014). Data brokers: A call for transparency and accountability. Available at: http://purl.fdlp.gov/GPO/gpo49352.

Francis, P., Eide, S. P., & Munz, R. (2017). Diffix: High-utility database anonymization. *Privacy technologies and policy* (pp. 141–158). Cham: Springer. https://doi.org/10.1007/978-3-319-67280-9_8.

General Assembly of the State of Vermont (2018). An act relating to data brokers and consumer protection (H.764). Available at: https://legislature.vermont.gov/Documents/2018/Docs/ACTS/ACT171/ACT171%20As%20Enacted.pdf.

Ghose, A. (2018). What blockchain could mean for marketing. *Harvard Business Review,* (5), 2–5.

Gouweleeuw, J., Kooiman, P., Willenborg, L., & De Wolf, P.-P. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics, 14*(4), 463–478.

Henderson, C. M., & Palmatier, R. W. (2010). Understanding the relational ecosystem in a connected world. In S. H. K. Wuyts, M. G. Dekimpe, E. Gijsbrechts, & F. G. M. Pieters (Eds.). *The connected customer: The changing nature of consumer and business markets* (pp. 37 − 76). New York: Routledge.

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. *International Journal of Research in Marketing, 34*(1), 154–172. https://doi.org/10.1016/j.ijresmar.2016.06.001.

Jerath, K., Fader, P. S., & Hardie, B. G. S. (2016). Customer-base analysis using repeated cross-sectional summary (RCSS) data. *European Journal of Operational Research, 249*(1), 340–350. https://doi.org/10.1016/j.ejor.2015.09.002.

Kamakura, W. A., Wedel, M., de Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing, 20*(1), 45–65. https://doi.org/10.1016/S0167-8116(02)00121-0.

Kamleitner, B., & Mitchell, V.-W. (2018). Can consumers experience ownership for their personal data? From issues of scope and invisibility to agents handling our digital blueprints. In J. Peck, & S. B. Shu (Eds.). *Psychological ownership and consumer behavior* (pp. 91–118). New York: Springer. https://doi.org/10.1007/978-3-319-77158-8_6.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of GANs for improved quality, stability, and variation. ArXiv:1710.10196.

Kennickell, A., & Lane, J. (2006). Measuring the impact of data protection techniques on data utility: Evidence from the survey of consumer finances. In J. Domingo-Ferrer, & L. Franconi (Vol. Eds.), *Privacy in statistical databases. Lecture notes in computer science. vol. 4302. Privacy in statistical databases. Lecture notes in computer science* (pp. 291–303). Berlin: Springer.

Kotler, P., & Armstrong, G. (2014). *Principles of marketing.* Harlow: Pearson.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics, 9*(2), 407–426.

Loshin, D. (2011). *The Practitioner's guide to data quality improvement.* Burlington, MA: Morgan Kaufmann.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data, 1*(1), https://doi.org/10.1145/1217299.1217302 3-es.

Maletic, J. I., & Marcus, A. (2009). Data cleansing: A prelude to knowledge discovery. In

O. Maimon & L. Rokach (Red.), Data mining and knowledge discovery handbook (pp. 19–32). Boston, MA: Springer US. doi:https://doi.org/10.1007/978-0-387-09823-4_2.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research, 15*(4), 336–355.

Martin, D. L., Hoff, J. L., Gard, R. A., Gregosky, R. J., Jones, H. W., Kirkwood, C. A., ... Willott-Moore, C. L. (2008). Data collection, processing, validation, and verification. *Health Physics, 95*(1), 36–46. https://doi.org/10.1097/01.HP.0000298817.72107.48.

Martin, K. D., Borah, A., & Palmatier, R. W. (2016). *The dark side of big data's effect on firm performance. Marketing Science Institute Working Paper Series* (Report No. 16-104).

Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science, 45*(2), 135–155. https://doi.org/10.1007/s11747-016-0495-4.

McGilvray, D. (2008). *Executing data quality projects: Ten steps to quality data and trusted information.* San Francisco: Morgan Kaufmann.

Mittal, A. (2013). Trustworthiness of big data. *International Journal of Computer Applications, 80*(9), 35–40. https://doi.org/10.5120/13892-1835.

Musalem, A., Bradlow, E. T., & Raju, J. S. (2008). Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *Journal of Marketing Research, 45*(6), 715–730. https://doi.org/10.1509/jmkr.45.6.715.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *IEEE symposium on security and privacy* (pp. 111–125). . https://doi.org/10.1109/SP.2008.33.

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing social networks. *IEEE symposium on security and privacy* (pp. 173–187). . https://doi.org/10.1109/SP.2009.22.

Ng, I., Scharf, K., Pogrebna, G., & Maull, R. (2015). Contextual variety, internet-of-things and the choice of tailoring over platform: Mass customisation strategy in supply chain management. *International Journal of Production Economics, 159*, 76–87. https://doi.org/10.1016/j.ijpe.2014.09.007.

Olson, J. E. (2003). *Data quality: The accuracy dimension.* San Francisco: Morgan Kaufmann.

Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science, 31*(4), 567–586. https://doi.org/10.1287/mksc.1120.0718.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *IEEE international conference on data science and advanced analytics (DSAA)* (pp. 399–410). . https://doi.org/10.1109/DSAA.2016.49.

Pendyala, V. (2018). *Veracity of big data: Machine learning and other approaches to verifying truthfulness.* New York: Springer.

Rand, W., & Rust, R. T. (2011). Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing, 28*(3), 181–193. https://doi.org/10.1016/j.ijresmar.2011.04.002.

Reiter, J. P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association, 100*(472), 1103–1112. https://doi.org/10.1198/016214505000000619.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 461–468.

Rust, R. T., Kannan, P. K., & Peng, N. (2002). The customer economics of internet privacy. *Journal of the Academy of Marketing Science, 30*(4), 455–464. https://doi.org/10.1177/009207002236917.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering, 13*(6), 1010–1027. https://doi.org/10.1109/69.971193.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., & Yu, Y. (2017). Protecting customer privacy when marketing with second-party data. *International Journal of Research in Marketing, 34*(3), 593–603. https://doi.org/10.1016/j.ijresmar.2017.02.003.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., & Yu, Y. (2018). A flexible method for protecting marketing data: An application to point-of-sale data. *Marketing Science, 37*(1), 153–171. https://doi.org/10.1287/mksc.2017.1064.

Singer, N. (2012). Acxiom, the quiet giant of consumer database marketing. The New York times. Available at: https://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, 70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001.

Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: An interdisciplinary review. *MIS Quarterly, 35*(4), 989–1015. https://doi.org/10.2307/41409970.

South Carolina General Assembly (2018). South Carolina insurance data security act. Available at https://www.scstatehouse.gov/sess122_2017-2018/bills/4655.htm.

Steenburgh, T. J., Ainslie, A., & Engebretson, P. H. (2003). Massively categorical variables: Revealing the information in zip codes. *Marketing Science, 22*(1), 40–57. https://doi.org/10.1287/mksc.22.1.40.12847.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103–110. https://doi.org/10.1145/253769.253804.

Surendra, H., & Mohan, H. S. (2017): A review of synthetic data generation methods for privacy preserving data publishing. International Journal of Scientific & Technology Research. 6(3), 905-101.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco), 671*, 1–34.

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software, 67*(4), 1–36. https://doi.org/10.18637/jss.v067.i04.

Transparency Market Research (2017). Data broker market. Available at https://www.transparencymarketresearch.com/data-brokers-market.html.

Trepte, S., & Reinecke, L. (2011). *Privacy online: Perspectives on privacy and self-disclosure in the social web.* Berlin Heidelberg: Springer-Verlag.

Van Bruggen, G. H. (2018). Marketing and the connected customer. Brussels, Belgium: European Marketing Confederation. Available at https://mii.ie/resource/resmgr/reports/pdf/emc-february-2018.pdf.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing, 80*(6), 97–121. https://doi.org/10.1509/jm.15.0413.

Westin, A. F. (1967). *Privacy and freedom.* New York: Atheneum.

Wolfie, C. (2017). Corporate surveillance in everyday life. Vienna: Cracked Labs. Available at: https://crackedlabs.org/dl/CrackedLabs_Christl_CorporateSurveillance.pdf.

World Privacy Forum (2013). Testimony of Pam Dixon executive director, before the Senate Committee on Commerce, Science, and Transportation: What information do data brokers have on consumers, and how do they use it? Available at https://www.worldprivacyforum.org/wp-content/uploads/2013/12/WPF_PamDixon_CongressionalTestimony_DataBrokers_2013_fs.pdf.

Wright, L. T., Newman, A., & Dennis, C. (2006). Enhancing consumer empowerment. *European Journal of Marketing, 40*(9/10), 925–935. https://doi.org/10.1108/03090560610680934.

Wuyts, S. H. K. (2010). Connectivity, control, and constraint in business markets. In S. H. K. Wuyts, M. G. Dekimpe, E. Gijsbrechts, & F. G. M. Pieters (Eds.). *The connected customer: The changing nature of consumer and business markets* (pp. 77 − 103). New York: Routledge.

Yancey, W. E., Winkler, W. E., & Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer (Vol. Ed.), *Inference control in statistical databases. Lecture notes in computer science. vol. 2316. Inference control in statistical databases. Lecture notes in computer science* (pp. 135–152). Berlin: Springer.