Big Data
Volume 00, Number 00, 2019
© Mary Ann Liebert, Inc.
DOI: 10.1089/big.2018.0176

Adjusting to the GDPR: The Impact on Data Scientists and Behavioral Researchers

Travis Greene,[1] Galit Shmueli,[2,*] Soumya Ray,[2] and Jan Fell[2]

Abstract

Rapid growth in the availability of behavioral big data (BBD) has outpaced the speed of updates to ethical research codes and regulation of data privacy and human subjects' data collection, storage, and use. The introduction of the European Union's (EU's) General Data Protection Regulation (GDPR) in May 2018 will have far-reaching effects on data scientists and researchers who use BBD, not only in the EU, but around the world. Consequently, many companies are struggling to comply with the Regulation. At the same time, academics interested in research collaborations with companies are finding it more difficult to obtain data. In light of the importance of BBD in both industry and academia, data scientists and behavioral researchers would benefit from a deeper understanding of the GDPR's key concepts, definitions, and principles, especially as they apply to the data science workflow. We identify key GDPR concepts and principles and describe how they can impact the work of data scientists and researchers in this new data privacy regulation era.

Keywords: behavioral big data; data protection; GDPR; privacy and policy; information quality (InfoQ)

Introduction: The New Data Regulation Landscape

This new realm of big data has made large and rich microlevel data on individuals' behaviors, actions, and interactions accessible and usable by industry, governments, and academic researchers. Many industries, including retail, marketing, and advertising now take advantage of technologies such as GPS and facial recognition software,* originally developed by military and security agencies, to collect and process data for purposes of surveillance, anomaly detection, and prediction.[1–3] The resulting behavioral big data (BBD) include not only rich personal data but also social networks connecting individuals.[4] At the same time, this rapid technological advance has far outpaced the speed of updates to ethical research codes and regulation of human subjects' data collection, storage, and use.[5]

The ever-widening gap has motivated data science researchers to call for the creation of general ethical principles and guidelines to effectively balance the potential social and scientific benefits of BBD processing with its potential privacy costs.[6]

The European Union's (EU's) new General Data Protection Regulation (GDPR), which took effect on May 25, 2018, is poised to change the course of these developments. The GDPR is especially important because although there has been a long-standing *Directive* on the use of personal data in the EU,† a *Regulation*—which transcends national legislative processes and laws and has immediate application and enforcement in all EU Member States—has only been put in place now. The ostensible reason for updating the 1995 Directive was to keep the EU at the forefront of the modern information economy, while ensuring an "equal playing field" among the EU countries. In addition, heterogeneity in national implementations of the Directive resulted in inefficiencies in the "free

*Turow (2017) describes how retail industries use facial recognition, location tracking, biometric sensors, and other "wearables" to analyze and predict customer behavior.

†Data protection regulation, in the form of a European Union (EU)-wide directive, has applied to the processing of personal data in EU industry for over 20 years (Directive 95/46/EC).

[1]*College of Technology Management, National Tsing Hua University, Hsinchu, Taiwan.*
[2]*Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan.*

*Address correspondence to: *Galit Shmueli, Institute of Service Science, National Tsing Hua University, No. 101, Section 2, Kuang Fu Road, Hsinchu, 30013, Taiwan,* E-mail: galit.shmueli@iss.nthu.edu.tw

movement of personal data within the internal market."[7] The GDPR was designed to resolve these issues by placing limits and restrictions on the use and storage of personal data by companies and organizations operating in the EU and abroad, insofar as these organizations "monitor the behavior" of or "offer goods or services" to EU-residing data subjects (Article 24).[*] The GDPR thereby has the potential to affect any company or organization processing the personal data of EU-based data subjects, regardless of where the processing occurs.[7]

While academic research using human subjects' data in most developed countries has been strictly regulated,[†] the collection, storage, and use of personal data in industry have historically faced much less regulatory scrutiny (see, e.g., Federal Trade Commission[8]). Nevertheless, this "hands-off" approach to industry data collection and processing seems to be changing as the GDPR comes into effect and large BBD-processing corporations, such as Facebook and Google, report massive personal data breaches.[‡] Furthermore, new models of collaboration for industry/academia BBD research, such as that between Facebook and the Social Science Research Council,[9] highlight the increasing importance of academic ethical codes on industry BBD research.[§]

At the time of writing, the GDPR has just taken effect and its impact is already being felt not only by companies but also by the public, in the form of many e-mails from companies informing users of changes to the company's data privacy policies. Despite a growing number of industry-specific news articles, blog posts, marketing materials, and white papers aimed at clarifying the impact of the GDPR, developing a coherent synthesis of the complex, 261-page document is difficult. This difficulty is particularly acute for data scientists and BBD researchers, who may not be wholly familiar with the nuanced legal terminology and concepts surrounding privacy and personal data law. This article is thus a first attempt at sketching out answers to the following two questions:

1. What are the main GDPR terms and principles that a data scientist should be familiar with?
2. How should these technical and ethical principles be incorporated into data science workflows?

These questions are worth exploring because—in some cases—researchers appear to be unaware of the regulatory unification relating to the collection, access, and usage of BBD brought about by the GDPR.[10] Yet in other cases, some social scientists seem to already have incorporated key GDPR principles into their BBD research, such as the principle of data minimization and the weighing of potential benefits and harms of large-scale BBD processing. For example, Garcia et al.[11] describe the ethical considerations of their research (an apparent international collaboration) by noting:

> We consider the possible downstream consequences of our large-scale research. The resolution of the Facebook marketing API prevents the singling out of individual users, which makes all our codes useless for identifying individuals of any minority or threatened group. In addition, there is no way to identify the account of users and use our analysis for any kind of personalization or individual manipulation. From the onset, our project had the potential to reveal important relationships between social media use and gender inequalities online and off-line. These benefits greatly outweigh the minimum risks of analyzing this kind of aggregated data that are accessible to anyone with an Internet connection.

Regardless of how academic research is ultimately affected by the GDPR's principles, it would behoove data scientists and researchers to understand what is new and how the new regulations and legal environment might affect their routines, approaches, priorities, and possibilities. After all, noncompliance with the GDPR can—in the most egregious of cases—result in heavy financial penalties of up to €20 million, or 4% of the worldwide annual revenue of the prior financial year, whichever is higher.[**,††,‡‡]

Finally, the GDPR is worthy of study as it increasingly wields influence on the global discourse surrounding the debate on personal data and privacy. A recent news article concluded that the GDPR and California's new Consumer Privacy Act are "pushing the tech industry to the negotiating table," and federal legislation is currently being drafted in Congress to create

[*] The General Data Protection Regulation (GDPR) makes no provision for nationality or citizenship, but applies to all natural living persons residing within the geographic boundaries of the EU, irrespective of their immigration status (Art. 3(1)). The GDPR also applies to all natural living persons outside of the EU irrespective of their citizenship or immigration status when the data controller is an EU establishment or data processing occurs in the EU (Art. 3(2)). For example, if a company in China controls and/or processes the data of a Chinese citizen located in Germany, then GDPR applies; in contrast, an EU citizen in China whose data are controlled/processed by a non-EU entity would not be covered by the GDPR.

[†] See, e.g., compilation at www.hhs.gov/ohrp/international/compilation-human-research-standards

[‡] www.theguardian.com/technology/2018/oct/03/facebook-data-breach-latest-fine-investigation

[§] www.chronicle.com/article/Facebook-Says-It-Will-Help/243126

[**] Even corporations in the United States are not immune from the GDPR. A shareholder of Nielsen (a major data broker) recently sued the company for allegedly failing to accurately represent the degree to which GDPR would affect Nielsen's ability to collect personal data. www.jdsupra.com/legalnews/update-on-the-gdpr-six-months-in-effect-75271

[††] Firms can also take out GDPR-insurance if they are concerned about the possibility of being fined, although the legality of such insurance coverage is under question in individual EU member states.

[‡‡] www.gdpreu.org/compliance/fines-and-penalties

Table 1. The six General Data Protection Regulation principles (Article 5, Recital 39)

Principle	*Description*
Lawfulness, fairness, and transparency	Personal data must be processed lawfully, fairly, and transparently in relation to data subjects.
Purpose limitation	Personal data can only be collected for specified, explicit, and legitimate purposes (although further processing for the purposes of public interest, scientific or historical research, or statistical purposes is not considered incompatible with the initial purposes and is therefore allowed.)
Data minimization	Personal data must be adequate, relevant, and limited to what is necessary for processing.
Data accuracy	Personal data must be accurate and kept up to date.
Data storage limitation	Personal data must be kept in a form such that the data subject can be identified only as long as is necessary for processing.
Data security	Personal data must be processed in a manner that ensures security.

an "Internet Bill of Rights."[*] In addition, since 2016, there has been a wave of similar, GDPR-inspired regulations being passed—or at least seriously considered—by several other countries, including China,[†] a large group of 10 Ibero-American states (including Brazil, Mexico, Colombia, Chile, Peru, and Uruguay),[‡,§] India,[**] and Malaysia.[††] Given the economic influence of these countries and the increasing reliance on cloud technologies to collect and transfer personal data around the world, data scientists would be remiss if they did not have a basic understanding of how various governmental regulations may impact the future development of their industry as well as their day-to-day routines. Furthermore, companies hoping to benefit from collaborations with academic researchers should be aware of the major legal principles regarding personal data protection and analysis. In short, the language and legal concepts found in the GDPR have already deeply shaped the international discourse surrounding personal data collection and processing, making an understanding of the Regulation even more relevant for modern-day data scientists in the global economy.

To identify the practical impact of the GDPR on the collection, storage, and use of BBD by data scientists in companies and behavioral researchers in academia, we proceed in two main stages. First, we identify the key ideas, principles, and concepts in the GDPR that are most relevant for data scientists and organize them into four categories that are meaningful for data science researchers and practitioners: *goal*, *data*, *analysis*, and *utility*. The six key principles are summarized in Table 1, while a full list of terms and definitions is available in the Appendix. In the next step, we envision a typical workflow of a data science project or empirical study, and analyze how the GDPR can impact each step. Along the way, we try—whenever possible—to provide the reader with suggestions for alternative approaches of data collection, processing, and analysis that may be more in tune with the major GDPR principles.

The organizing framework behind our analysis and evaluation is the information quality (InfoQ) framework, which aims at "assessing and improving the potential of a dataset to achieve a particular goal using a given data analysis method and utility."[12(p.17)] The InfoQ framework can also be used to assess the value of potential, ongoing, and completed empirical studies. We therefore find it useful for analyzing the potential effects of the GDPR on data science practices and approaches.

The following sections are organized as follows. Section 2 discusses the key GDPR concepts as they relate to the four components of InfoQ: goal, data, analysis, and utility. Section 3 then examines the impact of the GDPR on data scientists by analyzing a typical data science workflow using the InfoQ framework. Finally, conclusions and future directions are given in the Conclusion section.

The Objective of the GDPR: Important Terms and Concepts for Data Scientists

A data scientist or researcher embarking on a study or project starts with either a goal and then searches for the right data set, or else starts from an available data set and identifies a useful goal to pursue with this data set. The data scientist applies data analysis methods and is continuously conscious of the metrics for measuring the study's success, such as company KPIs or successful publication in a scientific journal. In short, the four key ingredients that a data scientist

[*] www.washingtonpost.com/technology/2018/10/05/silicon-valley-congressman-unveils-an-internet-bill-rights

[†] www.iflr.com/Article/3807448/Corporate/PRIMER-Chinas-national-standards-for-personal-data-protection.html

[‡] www.jdsupra.com/legalnews/new-ibero-american-standards-to-provide-81974

[§] www.loc.gov/law/foreign-news/article/brazil-personal-data-protection-law-enacted

[**] www.prsindia.org/billtrack/draft-personal-data-protection-bill-2018-5312

[††] www.jdsupra.com/legalnews/malaysia-seeks-to-expand-personal-data-51921

works with are *goal*, *data*, *analysis methods*, and *utility measures*. These four components are also the ones that compose the concept of InfoQ, defined by Kenett and Shmueli[12] as follows:

$$\mathrm{InfoQ}(g, X, f, U) = U\{f(X|g)\} \tag{1}$$

where g is the goal, X is a data set, f is the analysis method, and U is the utility measure.

In the following we introduce the main concepts and principles of the GDPR, organized by the above components. We use *italics* to denote key terms and discuss them in nonlegal language. The formal definitions and exact GDPR article/recital are available in Table 1 (principles) and in the Appendix (terms).

Goal

Goal is the purpose for which the personal data are used. It can be a scientific question, a practical use, or any other objective that is set up by the entity using the BBD. Organizations typically have two levels of goals: a high-level "domain" goal and a more specific "analysis" goal.[12] Companies and organizations collect and use personal data for a variety of domain goals, including providing, maintaining, troubleshooting, and improving a service; developing new services; providing personalized services; and detecting fraud, abuse, and security risks. Some organizations have scientific research goals. For example, the online course provider EdX specifies in its privacy policy the goal to "support scientific research including, for example, in the areas of cognitive science and education."* There are multiple terms in the GDPR that relate to goal (Appendix). We discuss them as they relate to several guidelines.

Goals are set by *data controllers*.

According to the GDPR, the *data controller* is the entity who determines the goal of the data collection or analysis. A controller can be a company, a university, or any entity holding data on natural persons. Setting the goal is what distinguishes the data controller from the *data processor*, who works on behalf of the controller.† For example, a company using its customer data for building a customer churn model would be simultaneously considered the data controller *and* the data processor; whereas if the modeling is outsourced to a consulting firm, the controller would still be the company itself, but the processor would be the consulting firm.

Make your purpose transparent to users.

The GDPR requires companies to disclose to their subjects what personal data they are collecting on them and for which specific purposes. Only after obtaining the user's explicit consent can that data be collected and used. This requirement is based on the principle of *purpose limitation*. Indeed, the GDPR-updated privacy policies of many companies clearly contain sections detailing the data collected and their use (e.g., Facebook's "What kinds of information do we collect?" and "How do we use this information?").

Based on our personal experience with industry practitioners, however, this aspect of the GDPR tends to be mistakenly construed as meaning that the collection of any personal data without consent is not allowed. The GDPR does in fact provide legal grounds for data processing that does not require explicit consent from the data subject. Two (of several) such cases are for performing a contract signed with the data subject or on the data subject's request, as well as for the "legitimate interests pursued by the controller… except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject" (Article 6 (1))(b,f). That is, companies wishing to process personal data outside of these specific purposes must take into account the context of their business relationship with the data subject, the expectations of the data subject, and the nature of the personal data involved.

Repurposing.

The concepts of *legitimate interest*, *contractual necessity*, and *purpose limitation* are intimately connected. According to the principle of *purpose limitation* (Table 1), if a bank obtains its customers' consent to collect and process their personal data for opening and running their bank accounts (i.e., performance of a contract between the client and the business, or *contractual necessity*), then the same personal data cannot be used for purposes of direct marketing without the prior consent of the customers (Appendix). Facebook, for example, cites contractual necessity (along with consent and legitimate interests) as its most basic legal grounds for personal data processing.‡ Accordingly, only processing "defined in the contract" for providing Facebook's service is permitted.

The GDPR's stipulation of purpose limitation encompasses many common processing activities that financial institutions have only recently undertaken in the era of big data, such as using analytics to target new potential customers, improving loan decisions

*www.edx.org/edx-privacy-policy

†If two entities determine the goals of the data collection or processing (e.g., a collaboration between a company and a university), then the entities are considered joint processors.

‡www.facebook.com/business/gdpr

and fraud detection.[13] Whereas the repurposing of personal data for preventing fraud constitutes a legitimate interest of the data controller (e.g., a bank or credit card issuer), repurposing data for marketing purposes may only fall under legitimate interests when there is a relevant and appropriate relationship between a financial institution and the targeted customer. In other words, the issue of legitimate interest in marketing seems to hinge on whether a data controller's goals are client retention or targeting new clients (with whom there is no prior "relevant and appropriate relationship").

In light of this ambiguity surrounding direct marketing,* the safest course of action for data controllers would therefore be to limit processing of personal data only to those data subjects with whom some kind of documented contractual relationship currently exists. As an extra precaution, data controllers should aim for clearly-explained and defined purposes for processing and obtain explicit consent for such processing.

Goals vs. rights and freedoms. A key aspect of the personal data processing exceptions for these goals is that they must be balanced against the rights and freedoms of the data subjects. Both industry and academic data scientists should be aware of the GDPR's exceptions to these rights since many of these exceptions are concerned with the types of analyses data scientists typically perform in their work. In its discussion of exceptions to these rights, the GDPR states, "in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes [these rights can be waived]" (Article 89). The rights specifically referred to are "the right of access," "the right to rectification," "the right to restriction of processing," and "the right to object." In the particular case of "archiving purposes in the public interest," the abovementioned rights may be waived, along with the "notification obligation regarding rectification or erasure of personal data" and the "right to data portability" (Articles 19, 20). In other words, if it becomes too burdensome to conduct research due to confidentiality and security requirements, then some of the GDPR's privacy protection mechanisms (e.g., pseudonymization—see Data section) and notification requirements can be sidestepped.

Special exemption goals. The GDPR specifies four types of goals that permit special exemptions: *scientific research*, *statistical purposes*, *archiving and public interest*, and *historical purposes*. These various types of research could be carried out by the company's research and development department, by academic researchers, or other research organizations. We note that these terms are vaguely defined (if at all) in the text of the GDPR, and that many of these terms come with exceptions and additional safeguards that individual EU Member States may provide.

Scientific research. The GDPR intentionally carves out a broad swath of activities that could be construed as scientific research that includes "technological development," "demonstration," "fundamental research," "applied research," and "privately funded research." "Privately funded" research might be interpreted as applying to corporate research groups such as Microsoft Research or Facebook Research. Similarly, "technological development" may describe research by machine learning teams to improve algorithms at their companies. As an illustration, Facebook's revised Data Policy regarding "Product research and Development" reads, "We use the information we have to develop, test, and improve our products, including by conducting surveys and research, and testing and troubleshooting new products and features."†

Perhaps the only defining characteristic of scientific research as defined by the GDPR is that there should be "specific conditions… as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes." This means that if one's goal is scientific research, then special safeguards must be taken to protect personal data if the results of the research are published.

Statistical research. Two key aspects of statistical goals are the creation of "statistical surveys" and producing "statistical results." While the wording is vague as to what exactly might constitute a statistical survey, the crux of the interpretation centers on the notion that statistical research, according to the GDPR, aims to understand *aggregate*, rather than person-level results. The text then goes on to clarify that statistical results are "not used in support of measures or decisions regarding any particular natural person." Such a definition appears to bolster the idea that aggregating data

*The GDPR is opaque on the legality of processing of personal data for direct marketing. It does state, however, that "[the] processing of personal data for direct marketing purposes may be regarded as carried out for a legitimate interest," though these interests must be weighed against the fundamental rights of the data subject (Recital 47).

†www.facebook.com/about/privacy/update

and computing summary statistics—a common task in data analysis—likely fall under the scope of statistical purposes.

Public interest and archiving. While the GDPR avoids defining exactly what "archiving" or "public interest" means, it does list several of the data subject's rights that can be waived if they render "impossible or seriously impair" such research. Examples of "reasons of public interest" include "cases of international data exchange between competition authorities, tax or customs administrations, between financial supervisory authorities, between services competent for social security matters, or for public health, for example, in the case of contact tracing for contagious diseases or to reduce and/or eliminate doping in sport." An example of such processing by a company on the grounds of public interest is Facebook's Data Privacy Policy* that states it processes personal data to "Research and innovate for social good" and that they "use the information… to conduct and support research and innovation on topics of general social welfare, technological advancement, public interest, health and well-being."

Historical purposes. Genealogical research is one of the few historical research goals mentioned in the GDPR text (Recital 160).† Furthermore, regarding international transfers of personal data for scientific, statistical, and historical research purposes, the GDPR states that "the legitimate expectations of society for an increase of knowledge should be taken into consideration (Recital 113)." This research purpose then would seem to permit the exempted processing of documents such as burial certificates and birth records, which may sometimes include personal data of living relatives.‡ It should be noted that many of the details for GDPR research exemptions are currently being worked out by the individual member states and that specific details regarding which data subject rights may be overridden may differ.§

Data

A data set consists of measurements of entities. In the GDPR, the main entity of interest is the *data subject*, with a focus on measurements (variables) defined as *personal data* and *special category (sensitive) data*. Data scientists should be aware of the following key definitions (see also Appendix):

Data subject.

The GDPR's definition of *data subject* is "a [living], identifiable natural person." This differs from the definition of a *human subject* used by ethics boards in academia (mandated by the Common Rule in the United States), defined as "a living individual about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information."[14] The main difference between the academic and GDPR definitions is that the GDPR's *data subject* does not require any interaction or intervention by the data controller with the data subject.

Personal, sensitive, pseudonymized, and statistical data.

The key data measurements in GDPR are *personal data*, *special category (sensitive personal) data*, *pseudonymized data*, and *statistical data*. Due to their critical importance, we describe each as follows:

1. **Personal data,** or personally identifiable information (PII), specify a wide range of information that might identify a natural person in terms of his or her physical, physiological, genetic, mental, economic, cultural, or social identity. We note that IP addresses and cookies can be considered PII because "[they] may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them."
2. **Special category (sensitive personal) data** are categories of personal data that reveal an individual's belonging to some "special category" or group. The GDPR provides the list of categories in Article 9. Special category data are broadly similar to the concept of "sensitive personal data" under the U.K.'s 1998 Human Rights Act, except that the GDPR includes genetic data and some forms of biometric data in its definition.** By and large, processing of special category data is prohibited under the GDPR, unless users give explicit consent to such processing (Recital 51).

*www.facebook.com/about/privacy/update

†GDPR limitations only apply to living persons; deceased individuals' personal data may be freely processed.

‡www.freeukgenealogy.org.uk/news/2018/05/22/gdpr

§For an up-to-date summary of these exemptions in various EU member states, see www.twobirds.com/en/in-focus/general-data-protection-regulation/gdpr-tracker/scientific-historical-or-statistical-purposes

**www.ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data

3. **Pseudonymized data** are a subset of personal data that have had individual identifiers removed, so that it is not reasonably likely for a data processor to be able to "single out" a specific person. The GDPR states repeatedly that pseudonymizing personal data should be the foundation of a data controller's collection and storage practices.
4. **Statistical data** are synonymous with aggregated data. Statistical data are used to infer traits about groups of people, rather than specific individuals.

Publicly available data. As in the Common Rule that governs ethics boards for academic research, personal data that are publicly available are exempt from the prohibitions on processing personal data—even sensitive categories of personal data may be processed if "[they] are manifestly made public by the data subject."

Filing systems. Data are typically housed in spreadsheets or a database (including a distributed framework such as Hadoop) that must be accessed by the data scientist or data engineer. The GDPR refers to these means of data storage as the *filing system** defined as a "structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis." The GDPR requirements apply to all processing of personal data that "form part of a filing system or are *intended* to form part of a filing system" (emphasis ours). The subtle implication is that the GDPR applies even if a company merely intends to convert unstructured data to structured data, thereby creating a *filing system*.

Analysis

The GDPR uses the term *data processing* to denote a broad set of operations on personal data. *Data processing* includes not only data analysis but also operations such as data collection, recording, storage, disclosure, restriction, erasure, and destruction. The latter operations are typically handled by the system and database administrators, while data scientists primarily focus on "analysis" operations such as structuring (e.g., image or natural language processing), retrieval (e.g., sampling), consultation (e.g., exploratory analysis and visualization), adaptation, profiling (e.g., building predictive models), and automated processing (e.g., designing recommender algorithms).

Types of analyses that fall under *data processing*. Data scientists analyze personal data using a range of methods, from computing simple summaries and aggregations to sophisticated statistical models and machine learning algorithms, including text mining and network analytics. Analyses and modeling are used for training algorithms and fitting a model, as well as for deployment to new data subjects, such as providing recommendations or generating predictions for new users. Data analysis can range from manual, to semiautomated, to fully automated, as in the case of a company using off-the-shelf artificial intelligence (AI) voice or image recognition software (e.g., the ride-hailing company Uber uses an image recognition product by Microsoft to confirm the identity of drivers at the start of their shift).† This entire range of activities falls under *data processing*.

Identifying the *data processor*. The person(s) or organization(s) performing the data analysis can reside in different places: from in-house data scientists and data engineers to external consulting firms or academic researchers, as well as collaborations between these parties. In many cases, using advanced AI requires customization, as demonstrated by the growing number of consulting services offered by the providers of such software (e.g., Google's Advanced Solutions Lab that provides training in building customized systems alongside Google engineers). For this reason we consider the *data processor* related to Analysis, whereas *data controller*, the entity that sets the analysis goals, is directly related to Goal.

Utility

Utility means the objective function used by the data scientist to evaluate the performance of the analysis. It can include business objective functions such as clicks-per-view, customer churn rate, or return on investment (ROI), or more technical metrics such as precision and recall of a classifier, accuracy of predicted values, or experimental effect magnitudes.

The GDPR does not explicitly discuss metrics or performance measures. This means that companies are able to continue pursuing the same pre-GDPR objectives (e.g., optimizing ad revenues or maximizing continuous use of an app), although the means to those ends would need to change in terms of the data and algorithms used. While listing all specific applications of personal data processing and their performance

*The term *filing system* is perhaps a reference to earlier paper-based document storage systems.

†Leave it to the experts, *The Economist*, volume 426 Number 9085, March 31, 2018.

metrics would be far too onerous (and would quickly be rendered obsolete with new technology), the GDPR does lay down three important theoretical considerations for data controllers wishing to extract maximum utility from their data. These considerations may be viewed as constraints limiting the optimization of the particular objective function(s) stipulated by the data controller.

The fundamental right to privacy. The purpose in adopting the GDPR over the 1995 Directive was to "ensure a consistent and high level of protection of natural persons and to remove the obstacles to flows of personal data within the Union." This recital emphasizes the seriousness with which an individual's fundamental right to privacy must be respected under EU law. We note the sharp contrast between the EU and the US approach to data privacy: Weiss and Archick[15] aptly summarize this distinction by saying that in the United States, "collecting and processing [personal data] is allowed unless it causes harm or is expressly limited by US law," while in the EU, "processing of personal data is prohibited unless there is an explicit legal basis that allows it." Another way of generalizing the difference is that in the United States, consent to processing of personal data is implied unless data subjects opt-out ("opt-out" model); whereas in the EU, no consent is to be assumed unless data subjects explicitly opt-in ("opt-in" model). It is difficult to understate the impact this difference in underlying philosophy has had on the evolution of data processing policy in the United States and the EU. This difference has led at least one legal scholar to argue that the EU's general prohibition of automated processing (since the 1995 Directive) of personal data has "deterred entrepreneurs and investors with overbroad and rigid laws" and may help explain why many of the leading IT, social media, and cloud services originate from the United States.[16]

The principle of proportionality. This principle essentially states that the protection of personal data is not an "absolute right"; rather, the GDPR's limits on personal data processing—and thus the utility that may be extracted there from—ought to be "considered in relation to [their] function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality." The Recital asserts that the purpose of processing personal data is ultimately to "serve mankind." Such language suggests that there are cases where utilitarian arguments could be made for the processing of one's personal data against one's will or without consent, for example, in the case of a worldwide pandemic. The reference to personal data processing's "function in society" leaves open some fluidity in the interpretation of proportionality, not only because of evolving social mores but also due to future technological developments whose effects on society may, on the whole, be negative.

To understand the European perspective, recall that the Gestapo used personal information in 1930s Germany to identify Jews and various Eastern bloc secret police agencies collected vast amounts of personal information to identify potentially subversive citizens. Currently, the principle of proportionality rests on the assumption that big data processing and AI will be able to solve some of humanity's most pressing problems. Yet, if public perception of big data processing were to suddenly and drastically change, perhaps due to some malfunctioning autonomous weapon system or a massive personal data breach, the principle could be revised to reflect the fact that potential harms of personal data processing might outweigh its economic or social benefits. According to this reading, the *principle of proportionality* could thus be considered a relative of the utilitarian risk/benefit analysis for potential human subjects research first outlined in the Belmont Report and subsequently used as the basis for academic ethic boards' approval under the concept of "beneficence."

Legitimate interest. The important yet vague concept of *legitimate interest* similarly rests on the complex balance of commercial utility with respect for fundamental privacy rights. As a ground for processing, the implicit expectation is that the economic benefits of processing to the data controller (or to a third party) outweigh any potential harm done to a data subject, and thus, the controller has a "legitimate [economic] interest" in processing the data. This is the so-called balancing test. After all, according to the Charter of Fundamental Rights of the EU, data controllers have the "freedom to conduct a business," and the processing of personal data may be an inherent part of the business, as in an ad network, for example, Borgesius.[17] Yet at the same time, the same Charter bestows fundamental rights to privacy and data protection to data subjects. How these competing rights should be balanced is not obvious. Consequently, this constant tension between human rights and economic gain is a major motif in the GDPR. As the ostensible purpose

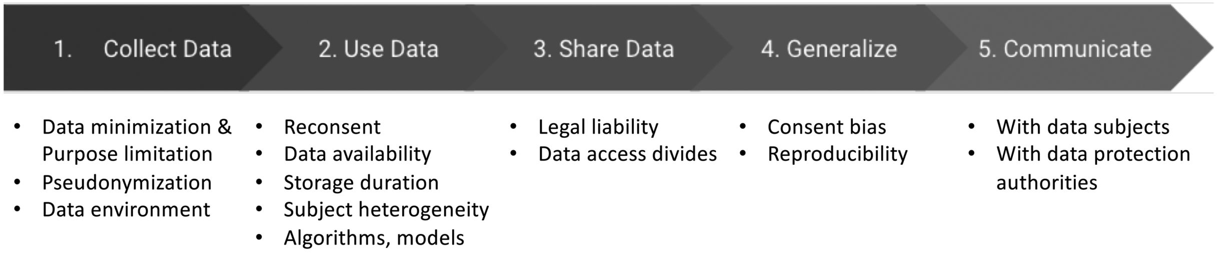

FIG. 1. An example data science workflow, tailored to behavioral big data usage.

of processing personal data is to "serve mankind," any arbitrary processing that could potentially violate a right to privacy would not pass the *proportionality* (i.e., *balancing*) *test* unless it could be shown to have significant social or business value.

The Impact of GDPR on Data Scientists: Analyzing a Typical Workflow

After our discussion of how the GDPR's principles and concepts relate to the InfoQ notions of *goal*, *data*, *analysis*, and *utility*, we now wish to analyze more concretely how GDPR will impact data scientists. Several generic data science workflows have been developed over the years, including CRISP-DM by IBM and SEMMA by SAS. However, we have tailored a workflow that highlights the specific issues encountered by many industry and academic data scientists using BBD. Figure 1 displays this workflow, from data collection to communication. We have added the steps "Sharing data" and "Generalization" to reflect the increasing growth and importance of industry/academia collaborations in the social sciences and recent academic controversies surrounding the replicability of many BBD experimental results. Nevertheless, we believe that this model workflow will be relevant to a broad swath of both researchers and practitioners in the new data regulation landscape.

To evaluate the impact of the GDPR in a methodical principled way, we again use the InfoQ framework. Since "information quality" is an abstract concept, Kenett and Shmueli[18] proposed decomposing InfoQ into eight dimensions that enable assessing the quality of information in a data set or a study. Each dimension considers an aspect of the four ingredients listed above. The dimensions are as follows: (1) data resolution, (2) data structure, (3) data integration, (4) temporal relevance, (5) chronology of data and goal, (6) operationalization (construct operationalization and action operationalization), (7) generalization, and (8) communication. We consider each of these dimensions to assess the impact of the GDPR on data scientists' routines and approaches. The eight dimensions also have the added benefit of serving as a kind of data quality "checklist" for guided analysis in each step in the workflow.

Collecting data: pre and post

Precollection: data minimization and purpose limitation. The GDPR principle of *data minimization* dictates that personal data must be adequate, relevant, and limited to what is necessary for processing (Table 1). Consequently, companies will need to carefully assess the resolution of personal data they collect and justify why it is necessary for achieving their stated goal. The GDPR also puts forth the principle of *purpose limitation*, which states that personal data can only be collected for specified, explicit, and legitimate purposes (Article 5). It is not enough for companies to say, for example, that they need to process personal data to provide a service—they must specify *how* and *why* such processing is necessary for provision of the service. Since new goals require new consent requests from users, repurposing personal data from one project to another will no longer be a viable option, even if repurposing personal data would seem to adhere to the intentions behind the data minimization principle. In other words, purpose limitation prohibits companies from collecting a small amount of personal data (the minimization principle) and then reusing it for unspecified secondary purposes.

For example, a company wishing to build a personalized pricing model must only collect and process personal data that are strictly relevant to achieving this goal. These data may include past purchase histories or website browsing times, but should not contain data relating to one's ethnicity (e.g., an "Asian-sounding name") or IP address, even if these turn out to be useful predictors of one's willingness to pay a certain price for an item.[19] Building a recommender system would proceed similarly. After choosing a suitable legal basis of processing—most likely either legitimate interest or consent in the case of a recommender system—data scientists would be limited to collecting and processing only personal data necessary for generating useful recommendations.

Not surprisingly, some companies, such as Airbnb and Uber, have successfully circumvented these limitations by building "smart pricing" algorithms that are based on anonymized (aggregated or market-driven) or nonpersonal data, usually in the form of event-based or object-related—rather than natural person-related—data. This allows, for example, Uber to make dynamic pricing decisions for individuals without needing the individual's name.[19] Such examples seem to validate the GDPR's more balanced approach to the competing interests of business and personal privacy. They suggest that it is indeed possible to develop personalized predictive models that respect users' privacy and boost company profits at the same time. Consequently, creative data scientists who can develop reasonably performing personalized predictive models from nonpersonal or aggregated data may become highly sought after in a post-GDPR world. It is also likely that at the planning stages of new analytics projects, more time will be devoted to thinking about how the goals of the project can be met without requiring the use of personal data. After all, any losses in predictive performance may be more than compensated for by savings in GDPR compliance and documentation costs.

The *data minimization* principle will likely increase the importance of statistical power calculations for A/B tests that involve personal data (e.g., in usability research or customer journey studies). For example, suppose a data scientist is tasked with finding a minimum sample size for estimating the improvement in completion rate for a user interface redesign compared with an existing design, using a 90% confidence level and 80% power. For an hypothesized 80% historical completion rate and required 20% improvement difference in completion rates between the A/B versions (a goal set by the data controller), such an analysis would require behavioral data from $\sim$49 users in each group. By reducing the power requirement from 80% to 50%, the sample size can be reduced to $\sim$14 per group.[20] To follow the principle of data minimization, data scientists will thus need to carefully consider the necessary statistical power and confidence levels needed for their particular business goals; otherwise, they risk collecting more data than needed for testing their hypotheses at their required confidence level.* And under the GDPR, companies will be required to document and justify their processing decisions, so reasoned power calculations in A/B testing may be viewed as constituting proof of compliance with the data minimization principle.

At the same time, however, data minimization efforts and privacy-preserving techniques may conflict with one another. Imagine data scientists at an online dating platform are asked to determine whether the opt-in rates for personal data processing are different for users from different countries (or even ethnic groups) at a given statistical significance level. In this situation, issues of aggregation and minimization arise: the data scientist must consider the experiment's sample size as well as potentially identity-revealing counts of data collected. If the power calculation indicates a relatively small sample is sufficient to detect a difference of a predetermined magnitude with some specified confidence level, then the probability of getting unevenly distributed counts among the different groups is increased—thereby making it much easier to single out individuals by their discordant behavior.†

This example is important because it illustrates how the GDPR requires data scientists to have a firm grasp on the interplay between the technical details of A/B testing and fundamental GDPR principles. When deciding on the particular testing goal, project stakeholders will need to ask themselves questions such as, "How big of an effect size do we expect to see?" and "How narrow must our confidence intervals be?" Companies running A/B tests with millions or billions of users will be able to detect extremely small effects with high power, but the question under the GDPR is: "Are these differences enough to justify the increased risk of reidentification?" It is precisely in these types of situations where the *principle of proportionality* arises. In essence, A/B testing under the GDPR should be done using a principle similar to Occam's razor: if sufficient power can be achieved with a smaller sample size, then the GDPR dictates that the smaller sample should be used, unless a data controller can prove the "necessity"

*For a real-life example of how Stack Overflow uses power calculations in its A/B testing, see stackoverflow.blog/2017/10/17/power-calculations-p-values-ab-testing-stack-overflow

†If Facebook's plans to create a dating app are realized, such a scenario may become commonplace, see "Facebook announces dating app focused on 'meaningful relationships,'" May 1, 2018. www.theguardian.com/technology/2018/may/01/facebook-dating-app-mark-zuckerberg-f8-conference

of such large-scale testing (see Borgesius[17] for a more nuanced discussion of the principle of proportionality, legitimate interests, and the necessity test). In the case of Facebook's controversial emotional contagion experiment,[21] where a massive online behavioral experiment conducted on more than 600,000 users' news feeds found an extremely small effect size, the controllers would need to justify the scientific contribution of such a small effect and why it outweighs the potential privacy harms inherent in large-scale data processing.

In sum, data minimization thus seems to introduce a trade-off that will need to be resolved while collecting data under GDPR: data scientists need enough data to avoid reidentification of data subjects while minimizing sample sizes to levels sufficient for detecting effects of desired magnitudes. Minimizing sample size increases the chances of reidentification of individuals through their group membership, while increasing the sample size to obscure group membership of individuals leads to collecting more data than is necessary for identifying overall group effects and may potentially lead to violations of the *principle of proportionality*.

Postcollection: pseudonymization. A key addition in the GDPR over the 1995 Directive is the introduction of the security practice of *pseudonymization*, a method for reducing the chance that any particular data value can be "attributed to a specific data subject without the use of additional information," provided that this "additional information" is kept separately and securely.* Examples of unique identifiers that might single out a data subject include names, tracking cookies, e-mail addresses, user names, or IP addresses (including dynamic IP addresses), among others. Note that *pseudonymization* is different from *anonymization*, which aims to make the process of reidentifying particular data values with specific individuals practically impossible.

Pseudonymization can affect the resolution of data available to data scientists in a few ways. First, completely removing individual identifiers from data sets impacts the ability of data scientists and researchers to make predictions at the individual level at the deployment stage. Second, data scientists will need to consider which measurements (attributes) or combinations of measurements might be used, directly or indirectly, to identify a natural person in the data set. In some cases, aggregation might be a useful approach (e.g., replacing data on a user's individual sessions with daily aggregates). The extent of this process should be based on whether analysts possess "means reasonably likely to be used [to single out individuals]" in the data (Article 26). In other words, each organization will need different privacy protocols depending on the technical means of analysis available to the data scientists, the security practices of the organization, and the intrinsic motivations for the analysis.

The lack of identifiable data may have different implications for data scientists wishing to develop personalized models (e.g., recommender systems and personalized predictions for direct marketing) that operate at the individual user level and those relying on statistical models to describe aggregated group-level behavior (e.g., A/B tests and survey research). For example, researchers studying group-level economic behavior have little to no incentive to spend the time and effort to reidentify specific individuals in a data set or single out a specific individual, since their analytical goals are not on the individual level. Machine learning researchers, however, are often interested in the predictions for individual observations, and so pseudonymization requirements may mean fewer such data sets are available for analysis.

The pseudonymization requirement is even harsher for companies with small or declining user bases. This is because identification of individuals becomes easier in aggregated data as the number of aggregated units decreases (i.e., becomes sparser), or as the number of tabular aggregations increases (i.e., more combinations of tables with different categories). As Lowthian and Ritchie[22] note, aggregated tables of group membership require large samples to keep such numbers from being used to single out individuals, due to extremely low or high values or unusual distributions of values in a frequency table. The US Census Bureau, for instance, has taken to adding statistical noise to its aggregated statistics to reduce the possibility of singling out individuals in seemingly "anonymized" summary tables.† An extra concern is when the groups by which the data are aggregated are sensitive categories of data, such as ethnic origin, religion, or sexual orientation. Thus, under the GDPR, data scientists may benefit from becoming proficient in various privacy-preserving

*[*Pseudonymization*] is the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

†www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html

techniques, such as differential privacy or k-anonymity, to reduce the likelihood of reidentification (see, e.g., Dwork and Roth[23]).

The data environment. Determining whether data are *pseudonymized* or *anonymized* requires considering the *data environment*. If one possesses a "means reasonably likely to be used" to reidentify subjects, then such data are considered *personal data* and must be pseudonymized. Consequently, data scientists and academic researchers may find themselves facing stricter controls on which data sets—public or private—might be joined to extant user data to potentially single out individual users. Mourby et al.[24] state that according to the UK Anonymisation Network, there are four main components of a researcher's data environment: *other data*, *agency*, *governance processes*, and *data infrastructure*. *Other data* refer to databases, public registers, or even social media profiles the analyst may have access to. These other data sources are important because they constitute a large portion of reidentification risk. *Agency* considers the question, "What incentives or motivations might the analyst have in reidentifying a data subject?" *Governance processes* are the formal policies and procedures that control how the data are accessed, by whom, and for how long. Finally, *data infrastructure* could be the actual hardware and software used in analyzing the data. Some data environments may have password-protected access or require encrypted flash drives, for example, to ensure the security of personal data, thereby satisfying the principle of *data security*.

The incentives of data scientists to identify individuals vary vastly and thus pseudonymization will affect them differently. Data scientists working for data brokers or marketing firms have strong economic motivation to identify specific individuals. Their ostensible goal is to map online behavior to off-line purchase behavior through first-, second-, and third-party data integration. The data broker LiveRamp (an offshoot of major data broker Acxiom), for instance, offers the product "IdentityLink" that allows advertisers a single, "omnichannel view" of the consumer. Acxiom claims to permit the identification of specific consumers across "thousands of off-line and digital channels and touchpoints," based on the individual's purchase history, web and app behavior, loyalty program history, airline and retail data, and demographic information, among many other sources.* What could be considered pseudonymized data by a data scientist working at a first-party company may not qualify as pseudonymized personal data in the case of a data scientist at a firm such as LiveRamp. Not only would a LiveRamp data scientist have a clear economic incentive for singling out individuals, he would also have access to a variety of other data sets that could be combined to increase the probability of correctly reidentifying a particular individual as well.

Data brokers and data-savvy marketers are not the only analysts who might have powerful incentives to single out individuals and thereby turn essentially anonymized data into personal data. After Facebook publicly revealed that Russian operatives had been directed to influence the 2016 US presidential election on its platform and others,† several other politically motivated operations were uncovered, one of which involved the Saudi Arabian government. The New York Times reported that a Twitter employee had been promoted to a position that allowed him access to the personal information of users, including their IP addresses and phone numbers, which could then be used to link Tweets to specific devices and single out Saudi government detractors for punishment.‡ The lesson to be learned is that privileged analysts with powerful political and financial incentives can easily circumvent the protections pseudonymization was designed to introduce. Therefore, under the GDPR, pseudonymization—or related techniques, such as k-anonymity or differential privacy—is a necessary, but not quite sufficient step for securing personal data.

The examples above illustrate that under the GDPR, the definitions of pseudonymized and anonymized data are fluid and contextual: what might appear to be anonymized data from the point of a view of an academic researcher or a data scientist at a first-party company may merely be pseudonymized from the point of view of the data broker, given the variety of methods of analysis, additional data sets, and intrinsic motivations in performing the analysis. A therefore worthwhile task for any organization processing large quantities of personal data is to assess the motivations and technical feasibility of its data scientists in singling out individuals and also to take inventory of other related sources of data (public or private) that could potentially contribute to individual reidentification.

*Meet LiveRamp IdentityLink, lp.liveramp.com/meet-liveramp-identitylink.html

†www.theguardian.com/technology/2017/oct/30/facebook-russia-fake-accounts-126-million

‡www.nytimes.com/2018/10/20/us/politics/saudi-image-campaign-twitter.html

Using data

Reconsent of pre-GDPR data. A major issue is the status of data collected pre-GDPR and its effect on later data analysis and use. In the lead-up to the GDPR, many websites and online platforms asked users to reconsent to the processing of their personal data. This is probably the simplest and safest legal route for companies to retain pre-GDPR user data, but it raises the question of what happens to personal data that the data subject does not reconsent to for processing, or chooses to have erased. Some companies have stated they will continue to use such data, although in aggregated form. For example, Kaggle notes in its revised privacy policy, "We may use aggregated, anonymized data that we derived from your personal information before you deleted it, but not in a manner that incorporates any of your personal information or would identify you personally."* Such a tactic would be legal under the GDPR because the Regulation only applies to personal data—data that could reasonably be used to identify a natural living person. Anonymized data, ipso facto, cannot be linked to a specific individual and therefore is outside the scope of the GDPR.

Furthermore, many companies are beginning to set data retention time frames based on the data collection purpose, in accordance with the principle of *data storage limitation*. For instance, data collected during an A/B test for an established, high-traffic website that already possesses a deep knowledge of its user demographics might only need to be processed for a week, and then deleted. Whereas data collected by a fledgling start-up's website may need to be processed for months while the start-up gathers basic knowledge about how users actually interact and behave with it. In this sense, the duration between data collection and use is directly considered in light of the data processor's goal.

Similarly, companies storing large amounts of customer personal data in databases will need to regularly "cleanse" them to make sure customer data are either accurate and updated or otherwise deleted, in accordance with the principle of *data accuracy*. A director at a data consulting firm remarked, "If you've got out-of-date data and you don't have a solid cleanse process, your ROI is going to be significantly impacted."† On top of this, companies will need to have protocols in place for dealing with the GDPR-granted "right to rectification" that data subjects possess regarding the accuracy of their personal data. The incentives brought on by the GDPR may therefore have positive effects on a firm's bottom line, especially for firms that struggle with data inventory, quality, and retention issues.

Data availability. Because the GDPR's reach is global (unlike the previous Directive), it will ostensibly affect firms "offering goods or services" to, or "monitoring the behavior" of, data subjects in the EU (Article 3). Due to this widened territorial scope, some data scientists and researchers, who had never previously concerned themselves with the details surrounding the use of personal data from EU data subjects, may find that variables previously available to them are no longer being collected by certain platforms, or that the ability to process them has been restricted or removed. In the wake of the GDPR, Twitter, for instance, made several changes to its popular public API, including making time zone fields private and removing background profile images of users.‡ Consequently, researchers building predictive models using these features would need to remove the features from the affected models, find creative proxies for these same predictors, or perhaps abandon the models completely, depending on the feature importance.

More generally, "special categories" of sensitive personal data may now be off-limits to some non-EU data scientists. Non-EU-based data scientists may have plausibly assumed that any European data protection laws would not or could not apply to them. A concrete example of this is found in data mining research that aims to trace public political opinions of social media users, where political opinions are considered *sensitive personal data* under the GDPR. In such situations, researchers might inadvertently process personal data of EU-residing data subjects. The authors of a highly cited data mining article using the Twitter API admit that, "Most Twitter users appear to live in the US, but we made no systematic attempt to identify user locations or even message language, though our analysis technique should largely ignore non-English messages."[25] It is probable that the authors inadvertently processed the sensitive personal data of at least some EU data subjects. In the authors' defense, however, Article 9(e) of the GDPR states that the GDPR does not apply when personal data are "manifestly made public

*www.kaggle.com/privacy

†The GDPR and its implication on the use of customer data, Royal Mail 2017. www.royalmail.com/sites/default/files/RMDS-Insight-Report-October-2017.pdf

‡www.twittercommunity.com/t/upcoming-changes-to-the-developer-platform/104603

by the data subject." However, given that use of a public API does require some technical expertise, and that it is not obvious to most Twitter users that their political statements may be mined by researchers, it could plausibly be argued that these political opinions now constitute sensitive personal data. For data scientists using political opinions as predictors in a statistical model, this could present a problem because the required predictor columns would no longer be available at the time of prediction, post-GDPR.

Data storage and duration limits. A similar issue arises for researchers interested in understanding a behavioral phenomenon using descriptive or explanatory models based on past data. Although some companies have declared they will continue to use deleted personal data, although in aggregated form, due to storage duration limits, it may not be possible to build a new model using historical data if those data were erased or if a data subject revoked consent for processing. Nevertheless, it is unclear what should happen to models and algorithms trained using these de-consented data. Should entire models be discarded or can they be kept as long as one removes the data of de-consenting users that were used to train the model?

One upshot to this dilemma, particularly in a time series forecasting context, is that using predictive models trained on pre-GDPR data might be less of a problem in rapidly evolving fields, industries, and environments. Although this may seem counterintuitive, in such cases, "disruption" is often the goal and historical patterns in data can quickly become irrelevant to future predictions. These dynamic situations call for models to be constantly updated by training on new data, which may contain new varieties of predictors as new technologies, internal policies, legal environments, and business strategies are tested. In fact, for rapidly expanding businesses, deciding which periods of data should be included in the training and testing sets can be surprisingly complex. For example, lifetime customer value models at a fledgling start-up would be expected to be constantly updated as new products and services are rolled-out, thereby affecting current and future behavior. In contrast, models used by relatively established firms with entrenched business models (e.g., for retail, Target or Walmart) may still be able to make accurate predictions using data collected further in the past, since major changes in IT and business strategy are likely to occur relatively slowly. Big established companies therefore stand to lose more useful data—in the sense that their stores of older data are more relevant to future predictions—because data collected before the GDPR must be reconsented to be used. At the same time, for fast-moving start-ups, losing pre-GDPR personal data may be considered a windfall because it forces them to use only the most relevant period of data to make predictions.

Data subject heterogeneity. A new source of uncertainty in post-GDPR data is due to the enhanced privacy settings that websites may provide under the GDPR that were not offered pre-GDPR. Given the wider variability in user privacy preferences and the ease with which they can be changed, data scientists will need to grapple with larger within-subjects and between-subjects heterogeneity. For example, changes to a user's privacy settings may result in many missing values for a single variable during a specific period of time. At the same time, different users may consent to the collection of different aspects of their online behavior. eBay's post-GDPR privacy policy, for instance, lists at least four areas where users will have a choice: marketing, communication preferences, advertising, and staying signed in.* Increased missingness of data will clearly have a negative impact on a data set's InfoQ and we may thus see an increased need for reliable imputation methods.

Choice of algorithms and models. Due to the new issues of data availability, storage duration limits, and status of pre-GDPR data, the choice of algorithms and models used by data scientists will require another layer of consideration. First, the principle of *lawfulness, fairness, and transparency* makes transparent models† such as regression models and classification and regression trees advantageous over blackbox models, such as deep neural nets, in personalized applications. Transparency requires the ability to explain why a model has produced a certain prediction or recommendation for a data subject. As a result, the GDPR's transparency requirements may shift current machine learning practices toward those more commonly used in the highly regulated financial services industry. In credit scoring, for example, Basel II regulations require that any deployed credit scoring models be highly repeatable, transparent, and auditable.[27] In this kind of

*www.ebayinc.com/our-company/privacy-center/privacy-notice

†In the book *Weapons of Math Destruction*, O'Neil[26] highlights three features needed to make an algorithm a "weapon of math destruction": Opacity, scale, and damage.

regulatory environment, predictions made by deep neural nets are much harder to explain to concerned data subjects and scrutinizing data protection authorities. As a result, simpler approaches, such as logistic regression and decision trees, are often used by practitioners.

Related to transparency is the issue of human decision makers who might use such algorithms for supporting their decisions (e.g., judges using algorithms that predict recidivism), and therefore, the transparency of the automated algorithms to the decision makers. The GDPR stipulates "data subject[s] shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." Veale et al.[28] state that recent regulatory guidance indicates that there must be "meaningful human input" undertaken by somebody with "authority and competence" who does not simply "routinely apply the outputs of the model in order to be able to avoid contestation or challenge … This serves as yet another (legal) motivating factor to create systems where human users can augment machine results." In short, models and algorithms that produce understandable outputs—especially for nontechnical audiences—are likely to be favored for communicating with the decision makers as well as the data subjects for whom decisions are made.

A second type of model selection consideration relates to data subjects' ability to reconsent to the processing of their data. In applications where the model will continue to be used for scoring data in the future (e.g., a model for direct marketing, product recommendations, customer segmentation, or anomaly detection), preferred models are those that do not require reaccessing the training data. For example, a regression model or boosted tree, once trained, no longer requires the training data to produce predictions. In contrast, for predicting an outcome for new data subjects, a k-nearest neighbor algorithm compares the new subjects to subjects in the training data.

A third consideration that also relates to the reduction in available measurements and data on subjects is the favoring of parsimonious models that require fewer measurements, as well as models that can more easily handle missing values. For example, models or algorithms that use less personal data might still be usable with post-GDPR data. Dimension-reduction methods such as unsupervised principal component analysis (or singular value decomposition) or supervised ridge regression could be less favored compared with lasso or even stepwise selection procedures. While the former require all the original measurements, the latter can lead to a subset of the original measurements. Deciding which columns of sensitive personal data can be removed while still maintaining acceptable predictive performance may become a common task for privacy-aware data scientists; it would also seem to align well with the GDPR principle of *data minimization*.

Sharing data

The GDPR is likely to significantly change the way companies share data with one another and with academic researchers. Although modern Internet companies initially collected BBD for the purpose of improving services and making better decisions, their massive stores of user-based BBD have the "potential to advance social scientific discovery, increase social good, and… ameliorate important problems afflicting human societies."[9] Whereas previously most academics had to either collect their own data or pay for data sets relevant to their research, nowadays more and more of the most socially valuable behavioral data are being collected through social media and e-commerce platforms, such as those offered by Facebook, Google, and Amazon. What is most startling, however, is that although in absolute numbers the amount of BBD has increased exponentially, the proportion of such data accessible by researchers is "[far smaller] than at any time in history."[9] Despite this trend, collaborations between academic data scientists and industry remain more important than ever, given the vast array of potential research questions that could be addressed using industry-sourced BBD.

Legal liability under the GDPR.

Before the GDPR, third-party data sharing through mutual data sharing agreements or through purchasing was extremely popular among large and small companies. However, the GDPR's stipulation that controllers, and by extension their data processors, can be held liable for damages caused by illegal processing of personal data, has made data controllers increasingly hesitant to share any data that might contain potentially PII. As a result, we are already seeing some companies eschew third-party data sharing agreements they have previously participated in, such as the recent announcement by Facebook about winding down its "Partner Categories," a feature that allowed data brokers such as Experian and Oracle to use their own reams of consumer information to target social network users. Marketing companies in particular have been deeply impacted by the way the GDPR distributes liability through the entire data collection and processing phases. One marketing blog reports that "only 20 percent of 255 brand marketers … are confident that their mar-tech vendors [will not] expose them to legal risks if [the vendors] are not GDPR compliant."* There is thus some reluctance within industry to continue relying on third-party data brokers for access to external data sets, since the personal data contained within them may not have been obtained according to the principles of the GDPR and would put them at legal risk.

Data access divides. Fear of regulatory scrutiny may also have spillover effects on BBD-focused academic research. As mentioned by King and Persily,[9] if the relative proportion of data available to academics for research continues to decrease, we may begin to see disparities in access to company data on two distinct fronts. On the corporate front, the Future of Privacy Forum[29] found the two biggest obstacles to corporate/academic data sharing were possible risks of personal reidentification and intellectual property disclosure.† Consequently, only trusted researchers from elite universities with close ties to corporations might be given access to corporate data, thereby reducing the opportunities for socially purposeful research to be done by those outside of the corporate trust network. This could impact the ability of other academics to reproduce important experimental results, for example. The other side of the data access divide is related to the types of companies that can afford to collect and process BBD after the GDPR. There is already some research showing that the GDPR has benefited large companies at the expense of smaller ones.‡ This is likely because only large companies can afford the extensive compliance costs required by the GDPR. These effects may be particularly pronounced in the ad-tracking industry because cookies are considered to be personal data. If this trend continues, then it may further exacerbate the monopoly companies such as Facebook, Google, Amazon, and Apple have on BBD. These companies could then become the de facto "gatekeepers" of academic/industry BBD-based research. Such an arrangement could hamper scientific independence, especially as it is not uncommon for companies to ask for prepublication approval or patent rights.[9]

The takeaway here is that academic researchers keen on using corporate data need to start developing, as early as possible, symbiotic relationships with corporate data providers, and they should not be surprised if more and more of their data come from the coffers of an increasingly small group of Internet companies.§ From the corporate perspective, these types of data-sharing relationships will also require greater investments in human capital in the form of compliance officers, legal counsel, and risk management teams to minimize legal exposure due to data sharing with academic researchers. For now, the GDPR's introduction of binding corporate rules, may provide a partial solution to these issues of data sharing, particularly when international transfers of personal data are required.

Generalization

The GDPR's introduction of specialized privacy and consent standards for EU data subjects may have repercussions for both the statistical and scientific generalizability of studies and research results. In large-scale behavioral experiments, such as Facebook's 2014 emotional contagion experiment,[21] EU data subjects would likely need to give explicit consent for the use of their behavioral data and they would also reserve the right to withdraw their consent for processing at any time. Such withdrawal could introduce nonsampling errors and affect model estimation and prediction. In the event of a large-scale withdrawal of EU data subjects' consent to processing, the theoretical population of interest would no longer include all EU Facebook users, but only those EU users who have agreed to their data being used (and non-EU users who do not possess rights to erasure). These users may in fact be systematically different from the users who do consent to the processing of their personal data. On top of nonsampling errors, sampling errors might also increase: the precision with which statistical effects can be estimated, for example, the width of confidence intervals for population effects might be affected by the resulting smaller sample sizes, reducing one's ability to generalize from sample to population.

*Facebook to stop allowing data brokers such as Experian to target users, The Guardian, Mar 29, 2018.

†Customer and user data can have enormous value to a firm and are often listed on a firm's balance sheet as "intangible assets."

‡www.techcrunch.com/2018/10/09/gdpr-has-cut-ad-trackers-in-europe-but-helped-google-study-suggests

§Intermediary professional organizations recognized by both industry and academia might also serve this purpose.

GDPR and consent bias. This concern about generalization, also known as *consent bias*, is further bolstered by the current debate in the scientific and statistical communities about whether requiring explicit consent from data subjects biases the results due to systematic differences in the way that data subjects are selected.[30] Under the GDPR, the problem of consent bias may be exacerbated due to differential privacy standards for EU and non-EU data subjects. Indeed, privacy-savvy users will likely be underrepresented in studies because they will not consent to their data being used for unspecified research purposes. As evidence of the possibility of such a bias, it has been reported that as of November 2018, only about one-third of US Internet users have opted in to the processing of their personal data, and as many as 17% have completely opted out.[*] The percentage of opt-outs for European users is almost guaranteed to be even higher. If this trend continues, data scientists making statistical inferences based on users' data, such as is commonly seen in A/B testing in industry or BBD-based empirical studies published in scientific journals, may end up having a significantly more accurate picture of non-EU users than EU users.

Facebook, for instance, has already publicly stated that for non-EU users in Asia, Latin America, and Africa, US privacy guidelines will apply.[†] Yet other big names in the BBD arena, such as Microsoft, have declared that they will apply GDPR protections globally.[‡] Given the extra costs of documentation and compliance of personal data processing and collection under the GDPR, it is unclear how other major BBD controllers, such as Amazon or Google, will proceed. It is telling that several months after the GDPR went into effect, there are still major media publishers such as the Los Angeles Times, Chicago Tribune, and San Diego Union-Tribune, which are blocking EU-based users from accessing content out of fear of noncompliance.[§]

If differential data processing pipelines for EU and non-EU data subjects do indeed become the norm, BBD research may then begin to resemble the ethically dubious way in which HIV vaccines were trialed in developing nations in the early 1990s. Research ethicist Iltis[31] notes that critics of such trials worried that the research benefits would only go to those in rich Western countries and that experimenters were taking advantage of "low-wage" African research subjects. Similarly, non-EU data subjects could become the new, preferred "low-privacy" BBD research subjects because of the relative ease and low cost with which their personal data could be processed. Scientific generalization would be reduced because any scientific models based on the non-EU users may not apply to EU populations for various cultural and geographical reasons. Furthermore, a key aspect of academic human subject's research, the Belmont principle of "justice," which addresses the fair distribution of benefits, risks, and costs among experimental subjects, would also be violated. Such a situation would seem to put collaborative industry/academic BBD research at odds with traditional academic human subject's research.

Concerns of scientific reproducibility. The GDPR's stringent consent standards for EU data subjects could also negatively affect the related notion of *scientific reproducibility*, which is concerned with the ability to recreate scientific conclusions and insights from previous studies.[32] Increased personal data privacy standards can hamper attempts to share or reproduce a given statistical analysis because of the legal exposure of third-party data processors and of withdrawn consent (dropouts) for processing.[29] For example, what if a data subject initially consents to his or her data being processed for statistical research purposes, but then changes his or her mind after the processing but before the analysis? It would not be possible to completely replicate the analysis if certain data subjects' personal data were removed in the time between different replication attempts.

One potential solution proposed by King and Persily[9] is a system in which all funded academic/industry research would follow a "replication standard," through which each data set used in research would come with a "universal numeric fingerprint" that would persist even if the format of the data were to change. Furthermore, all computer-code methodological details and metadata would be publicly available on the Internet, while the actual data needed for the research would be stored internally at the company and accessible to academics. Such a system could reduce the chance of an inadvertent data breach due to the sharing of personal data

[*] www.forbes.com/sites/forbesagencycouncil/2018/11/08/how-content-marketing-can-benefit-in-a-post-gdpr-world

[†] Facebook to put 1.5 billion users out of reach of new EU privacy law, Thomson Reuters, April 19, 2018, www.reuters.com/article/us-facebook-privacy-eu-exclusive/exclusive-facebook-to-put-1-5-billion-users-out-of-reach-of-new-eu-privacy-law-idUSKBN1HQ00P

[‡] Microsoft expands data privacy tools ahead of GDPR, May 24, 2018 www.theverge.com/2018/5/24/17388206/microsoft-expand-data-privacy-tools-gdpr-eu

[§] www.theguardian.com/technology/2018/may/25/gdpr-us-based-news-websites-eu-internet-users-la-times to researchers for replication purposes. Currently under the GDPR, however, if companies wish to avoid liability for potential data breaches, the simplest and most common solution is to refuse to share the data used by the original study, thereby reducing the scientific reproducibility of behavioral research.

Communication

Communication with data subjects. The GDPR lays out the major duties regarding the types of interactions between data controllers and data subjects, many of them related to using clear and simple language to explain the grounds of processing, and detailing the data subjects' right to have their information provided on request. Furthermore, data controllers must be able to clearly explain—in a nontechnical way—to data subjects, which personal data are being collected, why their data are being collected, and for what specific purposes or goal(s). For example, if there is to be communication with a child, the language used needs to be appropriate for a child (children can consent to the processing of their personal data generally starting at age 16 years),[*] and a clear "opt-out" option to the collection and processing of personal data should also be available. If data subjects do choose to opt-in, however, their "right to access" this information should not be excessively burdensome, either (Article 12 states that this information should be "easily accessible" by data subjects).[†] There are already reported cases where concerned data subjects requested information about the personal data stored about them only to receive an automated message requesting the data subject provide detailed information such as all public IP addresses, invoice IDs for purchases, credit card numbers used in purchases, dates of logins, names of user accounts, and much more.[‡] It seems reasonable to assume that for older or less technically inclined data subjects, providing this information may not be feasible. Companies may therefore want to draft multiple versions of their privacy policies with language specially crafted for children and the elderly, to conform to the GDPR principle of transparency described in Article 12. For the average data subject, however, privacy policies containing information in short clear sections such as, "What data we collect about you," "how your personal data are used," "how your information is shared," and importantly, contact information (typically including an e-mail address) for data privacy concerns may constitute adequate proof of transparency.

In addition, if companies use any type of automated means of *profiling* users, the users must be provided with a notice that algorithmic profiling is taking place, along with the "consequences of such profiling," and a choice to opt-out. As mentioned earlier in this article, organizations and their data scientists will need to ensure that decisions based on complex algorithms can be adequately understood by nontechnical users (and regulatory agencies). Data scientists will thus likely need to include communicability into their choice of algorithms and their documentation, should data subjects exercise their rights under the GDPR.

Communication with data protection authorities. Regarding documentation, the GDPR requires that data controllers and processors provide proof of compliance. Article 30, for example, stipulates that for companies with more than 250 employees, or who engage in processing "likely to result in a risk to the rights and freedoms of data subjects," detailed records must be kept that include such information as the purposes of processing, descriptions of data subjects and data categories, and expected storage duration limits for personal data. Furthermore, for companies doing large-scale data processing, regular data protection impact assessments and audits may become commonplace.

The GDPR's introduction of mandatory data breach reporting periods is highly relevant, given the recent spate of reports of massive data breaches from Facebook and Google+. According to Recital 85 of the Regulation, companies must report such a breach within 72 hours to authorities, and also to the data subjects "without undue delay." The October 2018 data breach of the access tokens of nearly 30 million Facebook users serves as a prime example of how communication with both data protection authorities and end-users will change under the GDPR.[§] Since data scientists are intimately familiar with the company's stored personal data, they will need to have a clear understanding of their role in the data controller's obligation for documentation, reporting, and communication with both authorities and users in case of such breaches. The bottom line

[*] www.twobirds.com/en/in-focus/general-data-protection-regulation/download-guide-by-chapter-topic

[†] The deluge of GDPR privacy policy emails has resulted in a new kind of phishing scheme in which scammers pretend to be data controllers requesting that the data subject re-enter personal information and credit card numbers, which are later sold on the Dark Web. www.zdnet.com/article/phishing-alert-gdpr-themed-scam-wants-you-to-hand-over-passwords-credit-card-details

[‡] www.appuals.com/epic-games-store-privacy-policy-conflicts-with-eu-gdpr-laws-sketchy-refund-policies

[§] www.theguardian.com/technology/2018/oct/03/facebook-data-breach-latest-fine-investigation is that under the GDPR, effective data scientists will need to possess strong communication skills and be comfortable interacting with diverse audiences that include data subjects, management, the data protection authorities, and other departments involved in collecting and analyzing personal data. Data scientists will also need to collaborate with other stakeholders to systematically document processing to demonstrate compliance with the GDPR principles outlined in this article. We conclude by noting that the communication skills of data scientists with less- or nontechnical audiences are therefore likely to become even more important in the future.

Conclusion

The landscape of data ethics regulation is seeing several important changes in the year 2018, with the GDPR as the most significant one. The GDPR is the first Regulation with international scope, as opposed to earlier guidelines and directives, recent attempts of companies at self-regulation (e.g., Facebook's internal ethics committee), and proposals for organizational structures (see, e.g., Polonetsky et al.[33]). As such, it is already affecting organizations around the world, which in turn affects data scientists and researchers using BBD in both industry and academia. Despite the immediate short-term impact on suppressing industry/academia collaborations, will the long run effect result in improved collaborations? We note that in 2018, another important data regulation change has taken place: an update to the Common Rule that guides academic research in the United States (the "Final Rule"). The GDPR and Final Rule updates have two notable common items: informed consent and the definition of identifiable data. While both clarify the notion of consent forms and expand the definition of identifiable data, the EU-based GDPR significantly increases the barriers and restrictions on collecting and using BBD. In contrast, the Final Rule update seems to lower the barrier for using BBD in behavioral research in academia. These US/EU and industry/academia differences may better align the ethical regulation of industry data scientists with academic researchers, thereby leading to improved collaboration.

In terms of the responsibilities of academic journals under the GDPR, two scenarios where the journal might be considered the data processor are (1) if the journal required authors to submit their data sets that contain personal data (e.g., for purposes of reproducibility) and thereafter stores or otherwise processes such data, or (2) when the published results could result in data subject reidentification.

Finally, we wish to emphasize that *research* has a special status under the GDPR. Given the confusion we have seen by data scientists and researchers, we would like to conclude by mentioning the following points regarding the nature of the data and GDPR applicability:

1. When the data are anonymous, the GDPR does not apply.
2. When the data are pseudonymous, the researcher has more lenience compared with nonresearch purposes.
3. When the researcher (or university) is the data controller, the GDPR applies equally as to nonresearch purposes.

As the GDPR just went into effect, data scientists and researchers are facing uncertainty and confusion regarding their routines and practices, as well as pressures from legal advisers and departments. We hope that this article provides a better understanding of the main GDPR concepts and principles as they pertain to the data science workflow, and can aid data scientists and behavioral researchers in this new—and often complex—global legal landscape.

Acknowledgments

The authors thank Shin-Yi Peng from NTHU and the reviewers for their valuable feedback. This work was partially funded by grant 106-2420-H-007-019 from the Ministry of Science and Technology in Taiwan.

Author Disclosure Statement

No competing financial interests exist.

References

1. de Leeuw K, Bergstra J. The history of information security: A comprehensive handbook. Amsterdam: Elsevier, 2007.
2. Turow J. The aisles have eyes: How retailers track your shopping, strip your privacy, and define your power. New Haven: Yale University Press, 2017.
3. Mansfield-Devine S. Biometrics in retail. Biometric Technology Today, 2013.
4. Shmueli G. Analyzing behavioral big data: Methodological, practical, ethical and moral issues. Qual Eng. 2017;29:57–74.
5. Zook M, Barocas S, boyd d, et al. Ten simple rules for responsible big data research. PLoS Comput Biol. 2017;13:e1005399.
6. Hand DJ. Aspects of data ethics in a changing world: Where are we now? Big Data. 2018;6:176–190.
7. Calder A. EU GDPR: A pocket guide. Cambridgeshire, UK, IT Governance Publishing, 2016.
8. Federal Trade Commission. Protecting consumer privacy in an era of rapid change. FTC report, March 2012. Technical report, 2012.
9. King G, Persily N. Working paper. A new model for industry-academic partnerships. February 4, 2019. Available online at: http://j.mp/2q1IQpH

10. Olshannikova E, Olsson T, Huhtamäki J, Kärkkäinen H. Conceptualizing big social data. J Big Data. 2017;4:3.
11. Garcia D, Kassa YM, Cuevas A, et al. Analyzing gender inequality through large-scale Facebook advertising data. Proc Natl Acad Sci U S A. 2018; 115:6958–6963.
12. Kenett RS, Shmueli G. Information quality: The potential of data and analytics to generate knowledge. Chichester, UK, John Wiley & Sons, 2016.
13. Alexander L, Das SR, Ives Z, et al. Research challenges in financial data modeling and analysis. Big Data. 2017;5:177–188.
14. Tene O, Polonetsky J. Beyond IRBs: Ethical guidelines for data research. Washington Lee Law Rev Online. 2016;72:458.
15. Weiss MA, Archick K. US-EU data privacy: From safe harbor to privacy shield. Technical report, Congressional Research Service, 2016.
16. Determann L. Adequacy of data protection in the usa: Myths and facts. Int Data Privacy Law. 2016;6:244–250.
17. Borgesius FJZ. Personal data processing for behavioural targeting: Which legal basis? Int Data Privacy Law. 2015;5:163.
18. Kenett RS, Shmueli G. On information quality. J R Stat Soc Series A. 2014; 177:3–38.
19. Steppe R. Online price discrimination and personal data: A General Data Protection Regulation perspective. Comput Law Secur Rev. 2017;33:768–785.
20. Sauro J, Lewis JR. Quantifying the user experience: Practical statistics for user research, 1st edition. Cambridge, MA, Elsevier, 2012.
21. Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci U S A. 2014;111:8788–8790.
22. Lowthian P, Ritchie F. Ensuring the confidentiality of statistical outputs from the ADRN. Technical report, Administrative Data Research Network, 2017.
23. Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci. 2014;9:211–407.
24. Mourby M, Mackey E, Elliot M, et al. Are pseudonymiseddata always personal data? Implications of the GDPR for administrative data research in the UK. Comput Law Secur Rev. 2018;34:222–233.
25. O'Connor B, Balasubramanyan R, Routledge BR, et al. From tweets to polls: Linking text sentiment to public opinion time series. ICWSM. 2010;11:1–2.
26. O'Neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Crown Publishers, 2016.
27. Saddiqi N. Intelligent credit scoring: Building and implementing better credit risk scorecards, 2nd edition. Hoboken, NJ: Wiley, 2017.
28. Veale M, Binns R, Van Kleek M. Some HCI priorities for GDPR-compliant machine learning. arXiv preprint arXiv:1803.06174, 2018.
29. Future of Privacy Forum. White paper: Understanding corporate data sharing decisions: Practices, challenges, and opportunities for sharing corporate data with researchers. Technical report, 2017.
30. Junghans C, Jones M. Consent bias in research: How to avoid it. Heart. 2007;93:1024–1025.
31. Iltis AS. Research ethics. New York, Routledge, 2006.
32. Kenett RS, Shmueli G. Clarifying the terminology that describes scientific reproducibility. Nat Methods. 2015;12:699.
33. Polonetsky J, Tene O, Jerome J. Beyond the common rule: Ethical structures for data research in non-academic settings. Colo Tech L J. 2015;13:333.

Cite this article as: Greene T, Shmueli G, Ray S, Fell J (2019) Adjusting to the GDPR: the impact on data scientists and behavioral researchers. *Big Data* 7:X, 1–23, DOI: 10.1089/big.2018.0176.

Abbreviations Used

AI	=	artificial intelligence
BBD	=	behavioral big data
EU	=	European Union
GDPR	=	General Data Protection Regulation
InfoQ	=	information quality
PII	=	personally identifiable information
ROI	=	return on investment

APPENDIX

Glossary of GDPR Terms and Their Definition

Data Controller (Article 4(7))

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, where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.

Joint Controller (Article 26)

Where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers. They shall in a transparent manner determine their respective responsibilities for compliance with the obligations under this Regulation, in particular as regards the exercising of the rights of the data subjects and their respective duties to provide the information referred to in Articles 13 and 14, by means of an arrangement between them unless and insofar as the respective responsibilities of the controllers are determined by Union or Member State law to which the controllers are subject. The arrangement may designate a contact point for data subjects.

Direct Marketing (Article 21(2))

The data subject shall have the right to object at any time to processing of personal data concerning him or her for such marketing, which includes profiling to the extent that it is related to such direct marketing.

Business Development (Recital 47)

The legitimate interests of a controller, including those of a controller to which the personal data may be disclosed, or of a third party, may provide a legal basis for processing, provided that the interests or the fundamental rights and freedoms of the data subject are not overriding, taking into consideration the reasonable expectations of data subjects based on their relationship with the controller. Such legitimate interest could exist, for example, where there is a relevant and appropriate relationship between the data subject and the controller in situations such as where the data subject is a client or in the service of the controller.

Purpose Limitation (Article 5(1)(b))

[Personal data shall be] collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (purpose limitation).

Scientific Research (Recitals 162, 159, 157)
Scientific research purposes should be interpreted in a broad manner, including technological development and demonstration, fundamental research, applied research, and privately funded research. Scientific research purposes should also include studies conducted in the public interest in the area of public health. To meet the specificities of processing personal data for scientific research purposes, specific conditions should apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes. Within social science, research on the basis of registries enables researchers to obtain essential knowledge about the long-term correlation of a number of social conditions such as unemployment and education with other life conditions. Research results obtained through registries provide solid, high-quality knowledge that can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people, and improve the efficiency of social services. To facilitate scientific research, personal data can be processed for scientific research purposes, subject to appropriate conditions and safeguards set out in Union or Member State law.

Statistical Purposes (Recital 162)
Any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.

Archiving and Public Interest (Article 89 (3))
Where personal data are processed for archiving purposes in the public interest, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18, 19, 20, and 21 subject to the conditions and safeguards referred to in paragraph 1 of this article insofar as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes.

Historical Purposes (Article 89)
Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18, and 21 subject to the conditions and safeguards referred to in paragraph 1 of this article insofar as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes.

Data Subject (Article 4(1))
An identified or identifiable natural person.

Personal Data (Article 4(1))
Any information relating to an identifiable natural 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.

Special Categories of Personal Data (Article 9(1))
Special categories of personal data that include racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health, or data concerning a natural person's sex life or sexual orientation.

Anonymized Data (Recital 26)
Information that does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.

Pseudonymized Data (Article 4(5))
The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and

organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

Filing Systems (Article 4(6))
Any structured set of personal data that are accessible according to specific criteria, whether centralized, decentralized, or dispersed on a functional or geographical basis.

Online Identifiers (Recital 30)
Identifiers provided by devices, applications, tools, and protocols, such as Internet protocol addresses, cookie identifiers, or other identifiers such as radio frequency identification tags. [These] may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them.

Statistical Data (Recital 162)
Any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.

Publicly Available Data (Article 9(2)(e))
Personal data that are "manifestly made public by the data subject."

Data Processors (Article 4(8))
Processor means a natural or legal person, public authority, agency, or other body that processes personal data on behalf of the controller.

Processing (Article 4(2))
"Processing" means any operation or set of operations performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

Profiling and Automated Processing
(Article 4(4), Recital 71)
"Profiling" [consists] of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyze or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behavior, location, or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

Principle of Proportionality (Recital 4)
The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality.

Legitimate Interest ("Balancing provision")
(Article 6(f))
[Processing is permitted only if] For the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject, which require protection of personal data, in particular where the data subject is a child.

Contractual Necessity ("Necessity Principle")
(Recital 40)
In order for processing to be lawful, personal data should be processed on the basis of the consent of the data subject concerned or some other legitimate basis, laid down by law, either in this Regulation or in other Union or Member State law as referred to in this Regulation, including the necessity for compliance with the legal obligation to which the controller is subject or the necessity for the performance of a contract to which the data subject is party or to take steps at the request of the data subject before entering into a contract.

Privacy by Design (Article 25)
Taking into account the state of the art, the cost of implementation, and the nature, scope, context, and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organizational measures, such as pseudonymization, which

are designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing to meet the requirements of this Regulation and protect the rights of data subjects.

Consent (Article 7(2,3))

If the data subject's consent is given in the context of a written declaration, which also concerns other matters, the request for consent shall be presented in a manner that is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration that constitutes an infringement of this Regulation shall not be binding. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Before giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent.