Analyzing the Impact of GDPR on Data Scientists Using the InfoQ Framework

Galit Shmueli and Travis Greene

College of Technology Management, National Tsing Hua University, Hsinchu 30013, Taiwan

ABSTRACT

The significant growth in recent years in the availability of behavioral big data to companies, governments, researchers and other organizations and individuals has led to many novel developments and technologies. This fast technological advance has far outpaced the speed of updates to ethical research codes and regulation of data privacy and human subjects data collection, storage, and use. The EU's General Data Protection Regulation (GDPR) which comes into effect in May 2018, can have a large impact on data scientists and researchers who use BBD in industry and in academia not only in the EU but around the world. We use the Information Quality (InfoQ) framework by [Kenett and Shmueli, 2014] to identify the key GDPR concepts and analyze the various potential changes that can affect data scientists in the new data privacy regulation era.

1 Introduction: The new data regulation landscape

The new realm of Big Data has made large and rich micro-level data on individuals' behaviors, actions and interactions accessible and usable by industry, governments, and academic researchers. This Behavioral Big Data (BBD) ([Shmueli, 2017a]) includes not only rich personal data but also "social graphs" connecting individuals. This fast technological advance has far outpaced the speed of updates to ethical research codes and regulation of human subjects data collection, storage, and use.

In academia, ethic boards such as Internal Review Boards (IRB) in the USA, rely on decades-old regulations for evaluating and approving human subjects research. Since then, human subjects research has grown in scope, scale, and diversity owing to data digitization and advancements in methods for data analysis and linkage. In 2015, the US Department of Health & Human Services, who sets the guidelines for IRB processes, proposed revisions "to modernize, strengthen, and make more effective the Federal Policy for the Protection of Human Subjects that was promulgated as a Common Rule in 1991"[1]. Some of the updates have created controversy ([Metcalf and Crawford, 2016]). The revised Common Rule ("Final Rule") was published in the Federal Register on January 19, 2017, to be implemented in January 2018. The Final Rule implements new steps aimed "to better protect human subjects involved in research, while facilitating valuable research and reducing burden, delay, and ambiguity for investigators."

While academic research using human subjects' data in most developed countries has been regulated, in industry the collection, storage, and use of personal data has been much less regulated. At the same time, many of today's non-military and non-security industries such as retail, marketing and advertising, make use of military technologies for data collection and processing for surveillance, anomaly detection, and prediction. This includes hardware such as sensors, fingerprint readers, and cameras, and software and especially AI such as facial and speech recognition and location prediction.

Although there have been several *directives* and *guidelines* for use of personal data in the USA and in Europe, regulation is only taking place now, with the new European Union's General Data Protection Regulation (GDPR) ([European Parliament, 2016]) that takes effect on May 25, 2018. The GDPR places limits and restrictions on the use and storage of personal data by companies and organizations operating in the EU, thereby potentially affecting every organization in the world ([Calder, 2016]). The GDPR had a multi-year journey until its approval by the European Parliament and Council, being written and re-written, with the more contentious points eradicated. Although at the time of this writing the GDPR has still not taken effect, its impact is already strongly being felt not only by companies, but also by the public, in the form of many emails from companies informing users of changes to the company's data privacy policies, and specifically mentioning GDPR as the reason.

The recent revisions to the Common Rule, and the new EU's regulation on personal data reflect the need for updating ethical and regulatory guidelines as to help protect human subjects while still supporting academic and industry research. What are the potential benefits and challenges of the new ethical codes that data scientists and researchers should consider? [Metcalf and Crawford, 2016] explain: "ethics codes also serve a number of functions beyond deterring unethical behavior, including creation of a cohesive community identity, responding to external criticism and —most importantly for our purposes—

[1] Department of Health and Human Services (2015) Notice of Proposed Rulemaking: Federal Policy for the Protection of Human Subjects. Federal Register. www.gpo.gov/fdsys/pkg/FR-2015-09-08/pdf/2015-21756.pdf

establishing the moral authority for self-regulation". These are all necessary for data scientists and behavioral researchers in today's BBD era.

While there is a growing number of news articles, blog posts, marketing materials, and white papers describing specific aspects of the potential GDPR effects, such as the impact on web developers[2], or the feasibility of using AI for advertising, it is difficult to synthesize a coherent understanding of *how data scientists and BBD researchers should approach and consider the new GDPR changes*. However, it is detrimental for data scientists and researchers to understand what is new and how the new regulations and environment might affect their routines, approaches, priorities, and possibilities. There is therefore a strong need for a methodical and clear analysis of the GDPR implications to data scientists and researchers. In this paper, we aim to identify key changes that GDPR might cause to the collection, storage, and use of BBD by companies, which are important for data scientists to consider within companies, agencies and in academia, as well as in the context of collaborations within and between these sectors. To perform this methodically, we utilize the Information Quality (InfoQ) framework ([Kenett and Shmueli, 2014, Kenett and Shmueli, 2016]) for analyzing the various aspects that relate to the use of BBD under the new GDPR. The InfoQ framework aims at assessing the quality of information contained in a dataset, for a particular goal, when analytics are applied. Specifically, the InfoQ framework is a tool for "assessing and improving the potential of a dataset to achieve a particular goal using a given data analysis method and utility." [Kenett and Shmueli, 2016, p.17]. The InfoQ framework can also be used to assess the value of potential, ongoing, and completed empirical studies. We therefore find it suitable for analyzing the potential effects of GDPR on data science practices and approaches.

The remainder of the paper is organized as follows: In Section 2 we describe the principles underlying three of the main ethics processes/regulations: The US Common Rule, ethics of biomedical research, and the EU's GDPR. These principles, as well as the differences and commonalities across them, help understand the spirit of the new regulations. Section 3 analyzes GDPR using the Information Quality framework, identifying the key GDPR definitions, and describing possible directions in which they can influence data scientists and researchers. Section 4 provides a discussion and future directions.

2 Principles of Ethical Research and Personal Data Protection

Before analyzing the details of the new regulations, we examine the principles that have guided their creation. Understanding the principles and spirit behind the regulations help create a cohesive view of the sometimes lengthy and detailed regulations, which can be daunting to those without legal background or familiarity with legal jargon, and especially to data scientists who are typically used to condensed technical forms of communication such as conference and journal papers.

Ethical guidelines for research using BBD have been formulated and implemented in academic research for nearly 40 years. Data protection regulation, in the form of a EU-wide directive, has applied to the processing of personal data in EU industry for over 20 years (Directive 95/46/EC, [European Parliament, 1995]). We therefore start by briefly describing the principles behind these data protection regulations, to highlight the *intentions* behind such regulations. This will serve as a basis when we analyze the new GDPR regulation in Section 3.

2.1 The Common Rule and Belmont Report Principles

In 1979, in an effort to codify proper ethical behavior between biomedical researchers and their human subjects, a diverse group of lawyers, doctors, and scientists was set up and funded by the US Department of Health, Education, and Welfare. Their task was to design a set of ethical guidelines for future human subjects research that balanced humanity's need for scientific and technological advancement with respect for individual human dignity. The findings of this commission became known as the Belmont Report, whose guidelines and ethical perspective would go on to shape all federally funded human subjects research in the USA under the 1974 National Research Act. The commission ultimately settled on three broad themes that would define academic human subjects research for the ensuing four decades: *respect for persons*, *beneficence*, and *justice* ([American Association of University Professors, 2006]). Later, these principles would be reformulated by the US government in 1991 as the Common Rule, which extended the ethical guidelines of the Belmont Report to 15 federally funded departments and agencies ([Iltis, 2006]). More recently, in 2012, the Department of Homeland Security convened a series of workshops that resulted in the Menlo Report, which added a fourth principle for ethical research: *respect for law and public interest.* The four principles mean the following:

Respect for humans: The principle of respect for persons is founded on the idea that research subjects willingly and voluntarily are participating in the project and have given their "informed consent" to the researcher. Informed consent is a broad and important concept in human subjects research. According to [Iltis, 2006], "consent must be documented" and investigators must also "establish a mechanism" for obtaining consent that does not "unduly influence" or "restrict" their ability to consider the study or ask questions.

[2]How GDPR Will Change The Way You Develop, Smashing Magazine, Feb 27, 2018 https://www.smashingmagazine.com/2018/02/gdpr-for-web-developers/

Beneficence: Researchers must aim to minimize harm to subjects and carefully and systematically assess both the risk and benefits of the research on the participants. The idea of "harm" here is quite broad and may cover intangible harms such as reputational harm, emotional harm, financial harm, and privacy harm, among others. On utilitarian grounds, if the harms brought on by the research outweigh the benefits to the subjects or to society, then the project will not be approved.

Justice: The principle of justice is related to the notion that the benefits of research should be fairly distributed in society and each participant deserves to be treated equally. Researchers must not arbitrarily seek out research on certain groups of subjects based on religion, health, age, race, etc. This principle is apparently aimed at stopping researchers from conducting research that only benefits certain groups in society.

Respect for law and public interest: This latest principle is less clear than the preceding three, but seems to admonish researchers to follow existing legal guidelines for obtaining data and keeping them secure. This principle requires researchers to use "due diligence" when designing studies to make sure such research respects all relevant legal contexts and information assurance procedures.

In biomedical research, the classic book *Principles of Biomedical Ethics* ([Beauchamp and Childress, 2012]), now in its 7th edition, lays out the four principles approach to biomedical research. The principles consists of four clusters of moral principles[3] that are central to, and serve as guidelines for biomedical professional ethics. These principles are related to the Belmont principles, but are more specific to the medical context ([Beauchamp and Childress, 2012, p.12]):

Respect for autonomy: A norm for respecting the decision making capacities of autonomous persons

Nonmaleficence: A norm of avoiding the causation of harm

Beneficence: A group of norms for providing benefits and balancing benefits against risks and costs

Justice: A group of norms for distributing benefits, risks, and costs fairly

While these principles are geared to clinical and biomedical research, we find them relevant also to the new realm of personal behavioral data.

2.2 Revisions to the Common Rule: The Final rule

Although the principles underlying the Common Rule are still in place, there has been a recent revision to the Common Rule aimed at keeping up with the socio-technological changes. The key changes in the Final Rule include the following ([Menikoff et al., 2017, Nichols et al., 2017]):

- Informed consent - now require that prospective participants be given the information that a "reasonable person" would want to have in order to make a decision about participating (geared mainly towards clinical and biomedical research in which long and complex informed consent forms have evolved).
- Easing of oversight for low-risk research: Annual review for low-risk projects no longer required.
- Permitting researchers to seek *broad consent*, which allows participants to agree to researchers using their identifiable private information or identifiable biospecimens, originally obtained for other purposes.
- Easing collaborations: a single-IRB review is now required for multi-institutional studies conducted in the USA.
- Pertinent to research conducted with international collaborators, the rule recognizes that there may be cultural groups or communities for which signing a consent form is unusual.
- Introducing the concept of "limited IRB review" for projects collecting sensitive, identifiable information from subjects.
- New exemption category: Research involving "benign behavioral interventions" (those that are not physically invasive, offensive, or embarrassing). This applies if the information collected is from an adult participant who prospectively agrees to the intervention and information collection and when additional criteria are met—e.g., the identity of the subject cannot readily be ascertained.
- Secondary research using identifiable private information or identifiable biospecimens is exempt without any form of review under certain circumstances (e.g., it is publicly available, the participant cannot readily be identified, or it is regulated under HIPAA for purposes of "health-care operations," "research," or "public health activities"—but not where the investigator plans to report individual research results).

[3] Principles are general norms that leave considerable room for judgment in many cases ([Beauchamp and Childress, 2012])

We note that the revision mostly eases conditions for researchers, offering a new exemption category, allowing seeking broad consent, and exempting secondary research using identifiable private information. This is in contrast to the more restrictive direction of data privacy regulation of companies in Europe.

2.3 Principles Guiding the European Union's General Data Protection Regulation (GDPR)

According to [Calder, 2016], the GDPR aims to tread the line between protecting the rights of the individual and removing barriers to the "free movement of personal data within the internal market". The regulation limits and restricts the use and storage of personal data for the purpose of keeping the EU at the forefront of the modern information economy, while ensuring an 'equal playing field' among the EU countries. Article 5 of the GDPR ([European Parliament, 2016]) outlines the six principles that should be applied to any collection or processing of personal data:

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 the public interest, scientific or historical research or statistical purposes is not considered as 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 its security.

Lastly, the GDPR assigns accountability for the above principles to the data controller (Article 5(2); see Section 3.1).

Many of the current GDPR data processing principles can be traced back to the Council of Europe's 1981 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, which were based on the Fair Information Practices (FIPS) formulated in 1973 in a report by a US government advisory committee. Around the same time, a similar set of personal data processing guidelines were outlined in the OECD's 1980 Guidelines on the Protection of Privacy and Transborder Flows of Personal Data ([Gellman, 2017]). The eight principles of the 1980 OECD document include guidelines for data collection, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability ([OECD, 1980]). According to [Gellman, 2017], Articles 5, 15, 16, and 17 of the GDPR essentially "[offer] an index to and restatement of most [of the] FIPs principles," except for the "right be forgotten," which may be viewed as either an "extension of FIPs" or "an entirely additional right." The key difference between the GDPR and the previous Data Protection Directive (95/46/EC), however, is that the GDPR is a Regulation whose articles are legally binding across the EU.

We note that the OECD's principles of data quality, purpose specification and use limitation are very much in line with the InfoQ approach, where the data, in terms of the necessary data and their accuracy, are considered through the lens of the specific goal of the analysis. In other words, data minimization does not just dictate the amount of data, but the amount relative to the goal at hand. Similarly, data must be kept accurate with respect to the specific goal.

Based on the above principles, the GDPR provides the following rights for individuals[4]:

- The right to be informed (Articles 12,13)
- The right of access (Article 15)
- The right to rectification (Articles 16,19)
- The right to erasure (Article 17)
- The right to restrict processing (Article 18)
- The right to data portability (Article 20)
- The right to object (Article 21)
- Rights in relation to automated decision making and profiling (Article 22)

[4] https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/

3 The effect of GDPR on Data Scientists: An InfoQ Analysis

The GDPR regulation ([European Parliament, 2016]) is a 261-page document containing key definitions, provisions, and recitals concerning the legal processing of personal data. We use the InfoQ framework to determine the key aspects of how the multitude of terms relate to collecting, storing, and analyzing BBD, and to identify the key concepts in GDPR that are relevant for data scientists and BBD researchers. The first step in the InfoQ framework is to identify the four InfoQ components: goal, data, analysis, and utility. For each component, we identified relevant terms in the GDPR document, and quote its exact definition. This is described in Subsection 3.1. In Subsection 3.2, we use the eight dimensions of InfoQ to elicit potential effects of GDPR on the practices of data scientists and BBD researchers.

3.1 The 4 InfoQ Components

To perform an InfoQ analysis, the first step is identifying each of the InfoQ components: goal, data, analysis, and utility. We start by briefly describing these components and linking them to GDPR.

3.1.1 Goal

Goal is the purpose for which the BBD is used. It can be a scientific question, a practical use, or any other objective that is set up by the entity using the BBD. In reality, organizations will often have two levels of goals: a high level "domain" goal and a more specific "analysis" goal [Kenett and Shmueli, 2016]. 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[5] the goal to "support scientific research including, for example, in the areas of cognitive science and education".

According to GDPR, the entity determining the goal is the *data controller* (see Table 1). In fact, setting the goal for the data collection and use is what defines the data controller and distinguishes him/her from the *data processor*, who works on behalf of the controller[6]. For example, a data controller may task a consulting firm (the data processor) to help it understand why customers are not purchasing a recently-released product. Here the high-level goal is to understand the purchase intentions and behavior of its customers. The specific analysis goal might the formulation of a set of testable hypotheses that can be plausibly confirmed or denied after analyzing the customer data held by the controller. Such a goal might be more accurately termed a "causal explanatory" goal, if it involves inferring causation, or a "descriptive" goal if it quantifies the correlation between customer purchase choices and customer (or other) data ([Shmueli, 2010]). Another common goal in a business setting is to be able to make predictions about customers' behavior. A company would like to know, for example, which customers are most likely to churn in the next three months so that preventive measures could be taken before that time. The domain goal is customer churn reduction, and the specific analysis goal might be identifying the top customers most likely to churn.

The determination of a specific analysis goal is crucial because it will normally inform project investment decisions within the company. In computing a project's net present value, for instance, determinations need to be made about a project's future cash flows so that these cash flows can be appropriately discounted and weighed against the initial start-up costs. As a more concrete illustration, a company may undertake a data processing project with the specific goal of identifying the top 10% most-likely-to-churn customers. But if it turns out that the resulting predictive model is no better than the current system for identifying churning customers, then the project will likely be abandoned as soon as it is realized that the analysis goal cannot be achieved.

Table 1 lists the main terms and definitions most likely to be relevant to a data controller's goals. A key requirement of GDPR that relates to goal is "specific purpose limitation." This is particularly relevant for businesses wishing to find a legal basis for processing that does not require explicit consent from the data subject. Article 6 details two legal grounds that are particularly relevant to the business context and the "goal" component of InfoQ. These two alternative grounds of processing are 1) "for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract"; and 2) "for 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" (Article 6 (1)(b,f)). In other words, companies wishing to process personal data outside of these specific purposes need to 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 (see Article 6 (4)(b), and recital 47 for more details).

For example, if a bank obtains its customers' consent to collect and process their personal data for opening and running their bank account (i.e., performance of a contract between the client and the business), then the same data cannot be used for purposes of direct marketing without the prior consent of the customers (see Table 1). Processing in this case would be

[5] https://www.edx.org/edx-privacy-policy

[6] 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

Term	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 subject and their respective duties to provide the information referred to in Articles 13 and 14, by means of an arrangement between them unless, and in so far 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.
Scientific Research (Recitals 162, 159, 157)	Scientific research purposes should be interpreted in a broad manner including for example 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 which 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. In order 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 & 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 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 fulfilment 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 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 fulfilment of those purposes.
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')

Table 1. Terms in GDPR and their definition, related to Goal

outside the scope of the original relationship (which was based on providing a secure place to store the client's personal funds) between the customer and the business. Nevertheless, it may be possible to argue that such direct marketing by the bank is protected under the GDPR's definition of legitimate interest. Recital 47, for example, states that "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" and finishes by noting 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. For the time being, it remains to be seen how exactly these kinds of liminal "legitimate interest" cases will be handled once the GDPR comes into effect. Until then, explicit consent for processing personal data that is based on specific purposes is probably the safest avenue for data controllers.

We note the four types of goals specified in GDPR which permit special exemptions. These are scientific research, statistical purposes, archiving & public interest, and historical purposes. These could be carried out by the company's R&D department, by academic researchers, or other researchers. We note that these terms are rather generally defined (if defined at all) in the text of the GDPR, and that many of these terms come with derogations–exceptions– and additional safeguards that individual Member States may provide.

A key aspect of the personal data processing derogations for these goals is that they must be balanced against the rights and freedoms of the data subjects. While discussing the 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 fulfilment 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 case of "archiving purposes in the public interest" the above-mentioned 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) and notification requirements can be sidestepped.

Regarding 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" (Recital 159). "Privately funded" research might be interpreted as applying to corporate research groups such as Microsoft Research or Facebook Research. Similarly, "technological development" may also apply to the work done by machine learning research teams to improve algorithms at these 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." [7]

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" (Recital 159). This should not be taken to mean that scientific research should, ipso facto, be published (or at least publishable), but rather 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. These safeguards will be determined by Union or Member State law.

Research done for "statistical purposes" is closely related to scientific research. Two key aspects of statistical purpose concern the creation of "statistical surveys" or producing "statistical results" (Recital 162). Again, the wording is quite vague when it comes to fleshing out exactly what might constitute a statistical survey. Could user segmentation and the creation of pivot tables be considered a kind of statistical survey? The crux of the interpretation centers on the notion that statistical research, according to the GDPR, aims to understand *aggregate*, rather than individual user-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" (Recital 162). Such a definition seems to bolster the idea that aggregating data and computing summary statistics—a common task in data analysis–likely falls under the scope of statistical purposes. It should also be mentioned that the GDPR relinquishes the task of defining the specific details of statistical purposes to the Union or Member States, "Union or Member State law should, within the limits of this Regulation, determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality" (Recital 162). Although not listed here (see Article 89), chief among these safeguards would be the process of pseudonymization (Sections 3.1.2, 3.2.1 and Table 2 for more on pseudonymization).

Archiving and public interest research is another goal that relies on the processing of personal data. Yet again the GDPR avoids defining exactly what "archiving" or "public interest" means, but does list several of the data subject's rights that can be waived if they render "impossible or seriously impair" such research (Article 89 (3)). Arguably the best examples of derogations due to "reasons of public interest" are found in Recital 112, which is concerned with the legal grounds for international data transfers. 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

[7] https://www.facebook.com/about/privacy/update

for social security matters, or for public health, for example in the case of contact tracing for contagious diseases or in order to reduce and/or eliminate doping in sport" (Recital 112). For an industry example of such processing on the grounds of public interest, Facebook's Data Privacy Policy states that it processes personal data in order to "Research and innovate for social good" and that they "use the information [they] have (including from research partners [they] collaborate with) to conduct and support research and innovation on topics of general social welfare, technological advancement, public interest, health and well-being. For example, [they] analyze information [they] have about migration patterns during crises to aid relief efforts."[8]

Finally, the research goal with likely the least detail is historical research purposes. The most prominent example of historical research given in the GDPR is genealogical research. It also bears worth mentioning that the scope of the GDPR only applies to natural living persons, so deceased individuals' personal data may be freely processed (Recital 160).

3.1.2 Data

In the InfoQ framework, 'data' is broadly defined to include any type of data intended to be used in the empirical analysis. Data can arise from different collection instruments: surveys, laboratory tests, field experiments, computer experiments, simulations, web searches, mobile recordings, observational studies, and more. Data can be primary, collected specifically for the purpose of the study, or secondary, collected for a different reason. Data can be univariate or multivariate, discrete, continuous, or mixed. Data can contain semantic unstructured information in the form of text, images, audio, and video. Data can have various structures, including cross-sectional data, time series, panel data, networked data, geographic data, and more. Data can include information from a single source or from multiple sources. Data can be of any size and any dimension ([Kenett and Shmueli, 2016]).

GDPR has specific definitions of data types and what they include. The focus in all cases is the 'data subject', or, in other words, "an identifiable natural person" [9]. The different data types defined in GDPR are summarized in Table 2. The most pertinent definitions for the discussion at hand are 'personal data', 'special category (sensitive personal) data', 'pseudonymized data', and 'statistical data'. Personal data is what is often called Personally Identifiable Information (PII), and specifies a wide range of information that might identify a natural person in terms of their physical, physiological, genetic, mental, economic, cultural or social identity. Recital 30 specifically states that IP addresses and cookies can be considered personally identifiable information (i.e., personal data) 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." These examples of personal data are new to the GDPR and reflect an update to the scope of personal data as previously found in the Directive 95/46/EC.

Special category (sensitive personal) data, are categories of personal data that reveal an individual's belonging to some "special category" or group. While GDPR does not include this term in the Definitions (Article 4), it does provide the list of categories in Article 9 (Processing of special categories of personal data). Special category data is broadly similar to the concept of 'sensitive personal data' under the UK's 1998 Human Rights Act, except that the GDPR includes genetic data and some forms of biometric data in its definition[10]. By and large, processing of special category data is prohibited under GDPR (Article 9).

The third data type, pseudonymized data, is a subset of personal data that have had individual identifiers removed, such that it is not reasonably likely for a data processor to be able to "single out" a specific person (see Recital 26 for more details). The GDPR states repeatedly that pseudonymizing personal data should be the foundation of a data controller's collection and storage practices.

Finally, statistical data should be thought as synonymous with aggregated data. Statistical data are used to infer traits about groups of people, rather than specific individuals.

Just as with the Common Rule, exceptions to the prohibition on processing are made if the personal data are publicly available. Recital 6 mentions the way in which technology has changed the economy and our social lives and appears to be hinting at the explosive growth of personal data posted online as part of social media profiles. In the USA, at least, there is some legal consensus that one's public social media profile constitutes the "modern public square" and thus the processing of one's public social media profile may be protected under the First Amendment. [11] Currently, however, it remains to be seen how the GDPR will deal with the issue of personal data scraped from publicly available sources, such as Facebook or LinkedIn profiles. Until then, the clearest direction on the processing of publicly posted personal data is given in Article 9(2)(e), where it is written that even sensitive categories of personal data may be processed if "[they] are manifestly made public by the data subject." If

[8] https://www.facebook.com/about/privacy/update

[9] The GDPR's definition of 'data subject' differs from the Common Rule's definition of 'human subject', which is "a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information" [Tene and Polonetsky, 2016]. GDPR's 'data subject' definition is closer to (2).

[10] https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/

[11] https://finance.yahoo.com/news/linkedin-lawsuit-determine-whether-bots-right-free-speech-192631359.html

Term	Definition
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.
Anonymised data (Recital 26)	Information which 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.
Pseudonymised 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 organisational 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 which are accessible according to specific criteria, whether centralised, decentralised 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 which are "manifestly made public by the data subject".

Table 2. Terms in GDPR and their definition, related to Data

indeed social media profiles are considered publicly-available personal data, then we may see an uptick in data science research focusing on techniques for "cloaking" user traits and behaviors, thus making it more difficult for machine learning algorithms to infer a data subjects' personal characteristics (see [Chen et al., 2017] for a recent example of such research).

The GDPR is mainly concerned with the legal grounds of processing personal data using manual or automated means. In typical data processing systems, these data will come in the form of a spreadsheet or database (including a distributed file storage and processing framework such as Hadoop) that must be accessed by the data scientist. The GDPR refers to these means of data storage as the "filing system"–a term whose generality is perhaps a reference to earlier paper-based document storage systems. The term "filing system" refers to a "structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis" (Article 4(6)). Given this definition, it seems that once personal data have been structured they constitute, by this very fact, a filing system.

A key point to note is that the Regulation applies to the processing of personal data "which form part of a filing system or are intended to form part of a filing system" (Article 2(1)). The inclusion of the phrase "intended to form part of a filing system" would seem to apply to currently unstructured data (i.e., text, video, audio, images, etc.) that a firm has plans to convert to structured data and store for processing via the use of a filing system. The implication is that firms with no intention of converting unstructured data to structured data (and subsequently storing the structured data in a filing system) are not subject to the same level of scrutiny as firms collecting and storing structured personal data. More specifically, Recital 15 states that, "Files or sets of files, as well as their cover pages, which are not structured according to specific criteria should not fall within the scope of this Regulation." Disregarding the possible reference to paper-based filing systems ("cover pages"), this recital could be interpreted as freeing firms with unstructured data from the scope of the GDPR. Nevertheless, if a firm has the intention to, at some point, convert the data such that it can be stored in a filing system, then it would be prudent to follow GDPR guidelines (i.e., use pseudonymization techniques) from the outset to minimize the risk of re-identification.

Analysis

In the InfoQ framework, 'analysis' refers to any form of analyzing the data, from computing simple summaries and aggregations to sophisticated statistical models and machine learning algorithms, including text mining and network analytics. It can pertain to a period when the data are used to estimate a model or derive knowledge (the training period), as well as to a deployment period, when, say recommendations or predictions are generated for new users. Data analysis can range from manual, to semi-automated, to fully automated, as in the case of a company using an off-the-shelf 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[12]).

Another dimension to consider is who is performing the analysis, from in-house data scientists and data engineers to external consulting firms or academic collaborators, as well as collaborations between these parties. In many cases, using advanced AI requires customization, as evident 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).

GDPR uses the term "processing" to denote a broad set of operations on personal data. Table 3 lists these operations and further related terms and their definitions. Note that we included "data processor" in the Analysis component, whereas "data controller" is listed under Goal (Table 1). We note that GDPR's term "processing" is broader than the InfoQ term 'analysis', as it also includes operations such as data collection, recording, storage, disclosure, restriction, erasure, or 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).

Term	Definition
Data processors (Article 4 (8))	'processor' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.
Processing (Article 4(2))	'Processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
Profiling/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 analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

Table 3. Terms in GDPR and their definition, related to Data Analysis

Utility

Utility means the specific objective function being used to evaluate the performance of the data 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 ([Kenett and Shmueli, 2016]).

A surprising, yet revealing issue is the stark absence of any mention of utility measures or performance metrics in GDPR. The lack of regulation or even discussion of performance measures means that companies are able to continue pursuing the exact same objectives (e.g., optimizing ad revenues, maximizing continuous use of an app) although the means to that end would need to change in terms of the data and algorithms used.

While listing all specific applications of personal data processing and their performance metrics would be far too onerous (and would quickly be rendered obsolete with new technology), the Regulation does lay down some important theoretical considerations for data controllers wishing to extract maximum utility from their data. First, the purpose in adopting the Regulation 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" (GDPR, Recital 10). This recital emphasizes the seriousness with which an individual's fundamental right to privacy must be respected under EU law (see Articles 7 & 8 of the EU Charter of Fundamental Human Rights). This is especially worth pointing out because of the sharp contrast between the EU and the USA's approach to data privacy. [Weiss and Archick, 2016] aptly summarize this distinction by saying that in the US, "collecting and

[12]Leave it to the experts, *The Economist*, volume 426 Number 9085, March 31, 2018

processing [personal data] is allowed unless it causes harm or is expressly limited by U.S. law," while in the EU, "processing of personal data is prohibited unless there is an explicit legal basis that allows it." It is difficult to understate the impact this difference in underlying philosophy has had on the evolution of data processing policy in the US and the EU.

Another key concept is the "principle of proportionality," which can be found in Recital 4 of the GDPR.The 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 therefrom– 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." In rather grandiose language the Recital asserts that, after all, the purpose of processing personal data is to "serve mankind" (Recital 4).

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. In fact, Recital 112 states that in "[cases] of contact tracing for contagious diseases" the restriction about transferring personal data to third-countries can be lifted. Returning to the notion of proportionality, the reference to personal data processing's "function in society" leaves open some interpretation of the notion of proportionality not only because of changing social mores, but also for future technological developments whose effects on society may, on the whole, be negative. To understand this European perspective, one only has to remember the ways in which the Gestapo used personal information in 1930s Germany to identify Jews or how the Stasi collected vast amounts of personal information about East German citizens in order to identify potential subversives. Currently the principle of proportionality rests on society's 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 weapons system or a terrorist group using hacked personal data to commit a crime, the principle could be revised to reflect the fact that potential harms of personal data processing could outweigh 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 IRB approval under the concept of "beneficence."

Lastly, the important yet somewhat vague concept of "legitimate interest" similarly rests on the complex balance of commercial utility with respect for fundamental privacy rights. As a grounds for processing, the implicit expectation is that the economic benefits of processing to the data controller outweigh any potential harm done to a data subject and thus the controller has a "legitimate [economic] interest" in processing the data. As detailed above, the constant interplay between human rights and economic gain is a major motif in the GDPR. As the ostensible purpose of processing personal data is to "serve mankind," any arbitrary processing which could potentially violate a right to privacy would not pass the proportionality test unless it could be shown to have significant social or business value.

Yet relying on legitimate interest for processing personal data may be a risky choice for data controllers. The legitimate interests of data controllers can be overridden if they are outside the scope of the individual's expectations for processing. Recital 47 states, for example, that "the interests and fundamental rights of the data subject could in particular override the interest of the data controller where personal data are processed in circumstances where data subjects do not reasonably expect further processing." It is therefore imperative that data controllers and processors keep these principles at the forefront of their analytics efforts in order to ensure that the utility derived from any data processing is not done at the expense of the data subject's rights. Data controllers must also communicate with data subjects to make it clear to them how their data might be processed and how that processing relates to the specific business relationship.

3.2 Potential Impact of GDPR on Data Scientists: Using the 8 InfoQ Dimensions

InfoQ is an abstract concept, and therefore [Kenett and Shmueli, 2014] proposed considering decomposing InfoQ into 8 dimensions, which can be used for assessing InfoQ of a dataset or a study. Each of the dimensions consider and affect not only the data and goal, but also the analysis method and the utility of the study. The eight dimensions are:

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
8. Communication

In the following we consider each dimension and use it to identify potential impacts of GDPR on different aspects of the data scientist's routines and approaches.

3.2.1 Data Resolution

The first InfoQ dimension of data resolution refers to the measurement scale and aggregation level of the data. In GDPR the basic unit of observation is the 'data subject' or 'natural person', with a major danger specified as the identifiability of a single data subject.

Two GDPR principles that will affect data scientists in terms of data resolution are the principles of "Data Minimization", which dictates that personal data must be adequate, relevant and limited to what is necessary for processing, and the principle of "Purpose Limitation", which states that personal data can only be collected for specified, explicit and legitimate purposes (see Section 2.3). Companies will need to carefully assess the resolution of personal data they collect and justify why it is necessary for achieving their stated goal. Since new goals may require new consent requests from users, repurposing personal data from one project to another will no longer be a viable option. Data scientists and marketing researchers building personalized recommenders or predictive models are those most likely to be affected.

A key addition in the GDPR over the DPD (95/46/EC) is the introduction of the security practice of "pseudonymization" (Recital 28). Pseudonymization is 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 (Article 4(5)). Examples of unique identifiers that could be used to single out a data subject include family name, email address, user name, or IP address, among others. It bears worth mentioning that pseudonymization is not the same as anonymization, which aims to make the process of re-identifying particular data values with specific individuals practically impossible.

Pseudonymization relates to data resolution in a few ways: First, removing individual identifiers from datasets will clearly impact the ability of data scientists and researchers to make predictions at the individual level at deployment stage. Analysts will need to consider which columns in the data (measurements) or combinations of columns might potentially 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 within "means reasonably likely to be used" (GDPR paragraph 26). Such wording implies that the degree of pseudonymization required is unique to the particular organization and will differ 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 in terms of needed data resolution. It is worth reiterating that "anonymized" data, which by definition cannot be traced back to a natural person, are not under the scope of the GDPR and may be freely processed. Lack of identifiable data may have different implications for researchers wishing to develop personalized models (e.g., data mining and machine learning researchers) that operate at the individual user level; whereas researchers who use statistical models to describe relationships among variables, or who study group-level behavior (e.g. social scientists and economists), would be less affected. For example, researchers studying group level economic behavior have little to no incentive to spend the time and effort to re-identify specific individuals in a dataset or singling out a specific individual, since their analytical goals are not on the individual level. In short, due to the GDPR's changes in the default storage format of data, i.e., pseudonymization, the regulation will likely have bigger effects on data mining and machine learning research that focuses on individual observations, while leaving social scientists using statistical models for quantifying average group effects less affected.

Second, a potential side effect of the future wide-spread use of pseudonymization of is that consumers may prefer to use products or services which purport to anonymize personal data. For example, some American users of Facebook Messenger may migrate to messenger apps such as Telegram, which are marketed as more secure. In this scenario, a sector of companies with highly-secure data will have much less rich data (fewer data columns), and at the same time the losing sector will have much fewer users (fewer data rows). Alternatively, the requirement to get consent might favor established companies, because "the cost of getting permission from users for their data [is] typically much higher for a younger company than for an established firm"[13], thereby leading to an even larger monopoly of BBD by the large data companies. In this scenario, the "data resolution divide" can lead to the large companies owning highly rich BBD, while the small and young companies suffer in terms of fewer informative columns as well as smaller data size. Such a divide will greatly disadvantage the data science capabilities of young firms to extract value from their BBD. [Martens et al., 2016] showed that when using fine-grained behavior data, there continues to be substantial value to increasing the data size across the entire range of the analyses. This suggests that larger firms may have substantially more valuable data assets than smaller firms, when using their transaction data for targeted marketing.

The pseudonymization requirement is even harsher for companies with declining user rates, thereby furthering the "data resolution divide", because identification becomes easier in aggregated data as the number of aggregated units decreases. Group

[13] How Looming Privacy Regulations May Strengthen Facebook and Google, NY Times, April 23, 2018, https://www.nytimes.com/2018/04/23/technology/privacy-regulation-facebook-google.html

membership might require large samples in order 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 ([Lowthian and Ritchie, 2017]). This issue could be especially concerning when the groups by which the data are aggregated are sensitive categories of data, such as ethnic origin, religion, or sexual orientation.

The data minimization principle can also affect data resolution: It might create challenges when attempting to aggregate users' data where there are not enough data subjects to prevent re-identification. Another issue related to the number of data subjects is choosing an appropriate sample size for group comparisons. Secondly, the GDPR's data minimization principle implies that when conducting A/B tests for usability research, statistical power calculations may become more important. For example, finding a minimum sample size while conducting a Chi-square test of successful completion rates for a user interface redesign, using a 90% confidence level and 80% power with an estimated completion rate of 80% and an expected difference in completion rates between the A/B versions of 20%, would require data from approximately 49 users in each group. By reducing our power requirement to 50% the sample size can be reduced to approximately 14 per group ([Sauro and Lewis, 2012]). Thus in order to follow the principle of data minimization data scientists will need to carefully consider the necessary power and confidence levels needed for their business goals. Otherwise they risk collecting more data than needed for testing their hypotheses at their required confidence level.

The two data minimization implications combine into a possible challenge. Consider a post-GDPR scenario, where a company with an online dating app asks its resident data scientists to determine whether the difference in opt-in rates for personal data processing is different for users from different countries (or even ethnic groups) at a given significance level. In this situation both issues arise: that of of sample size as well as potentially identity-revealing counts of data. If the sample size calculation indicates a relatively small sample, then the probability of getting unevenly distributed counts among the different groups is increased. Such scenario may be commonplace if Facebook's plans to create a dating app are realized. [14] Data minimization thus seems to introduce a tension that will need to be resolved while collecting data: researchers need enough data such that individual data subjects cannot be easily re-identified via group membership, yet power calculations are designed to set meaningful guidelines on the appropriate sample sizes needed to find an effect of the magnitude hypothesized by the researcher. Minimizing sample size increases the chances of easier identification of individuals through their group membership. And increasing the sample size to obscure group membership of individuals, leads to collecting more data than is necessary for identifying group effects.

Finally, in terms of business impact, if companies cannot ensure that user data cannot be linked to a specific individual, they may choose not to share it with third parties or academic researchers as a way of reducing any potential confidentiality breach. In their recent study of corporate data sharing, [Future of Privacy Forum, 2017] found that the two biggest obstacles to data sharing were possible risks of personal re-identification and intellectual property disclosure (customer and user data can have enormous value to a firm and are often listed on a firm's balance sheet as an "intangible asset."). Consequently, the report remarked that 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 outsiders. The takeaway here is that data processors and academic researchers keen on using corporate data need to start developing, as early as possible, symbiotic relationships with corporate data providers. 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 if they wish to minimize legal exposure due to data-sharing with researchers.

3.2.2 Data Structure

The InfoQ dimension of data structure refers to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, and network data) as well as unstructured nonnumerical data (e.g., text, text with hyperlinks, audio, video, and semantic data).

The requirements of the GDPR might favor the use of structured over unstructured data. First, large-scale data controllers and their processors will need to conduct data privacy impact assessments (DPIAs), aimed at quantifying the risks to personal privacy of holding such personal data (Article 35). Unstructured data, however, are by their nature difficult to quantify and interpret. There will thus be more incentive for companies who keep textual, video, and photo data to store them in more structured and easily-analyzable formats, as much as is technologically feasible, in order to avoid legal liability for data privacy breaches. Nevertheless, as discussed in Section 3.1.2 regarding the definition of a filing system, Recital 15 does seem to imply that unstructured data may be outside the scope of the GDPR. The crux of this distinction rests on whether the firm intends to convert these data to structured form and store them in a filing system. Given the rapid pace in the development of algorithms designed to extract meaning from unstructured data (see for example the Google Vision API [15]), it would probably be best for

[14] Facebook announces dating app focused on 'meaningful relationships', 1 May, 2018, https://www.theguardian.com/technology/2018/may/01/facebook-dating-app-mark-zuckerberg-f8-conference

[15] https://cloud.google.com/vision/

firms to ignore the ambiguity in Recital 15 and assume that any unstructured data they collect may be, at some point in time, reduced to a structured format and thereby bound by GDPR processing rules.

Social network data present another way unstructured data may potentially reveal personally identifiable information. As an example, using social media posts [Cascavilla et al., 2015] demonstrated the extent to which traditional data mining techniques and network analysis can undermine commonly-used methods of guarding against individual re-identification. Their results raise the question of whether social network data might be considered "personal data" under the GDPR and also whether it is meaningful to differentiate between personal data that is publicly posted–on a social media profile, for example–and data that is not. Again, what is considered personal data is relative to the data environment and skill of the data processor. Companies that have data scientists trained in techniques "reasonably likely to be used" to single out–identify–individuals, then such companies must then pseudonymize such data. And according to Article 71, if social network data mining techniques could be used to infer details about an individual's racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health or sexual orientation, then the processing of the above-mentioned "sensitive personal data" is expressly forbidden (though there are several legal grounds for doing so, including explicit user consent, public posting, or public interest reasons (Article 89)).

The derogation that allows publicly-posted unstructured sensitive personal data to be processed without consent will certainly raise questions about the fine line that can be crossed in predicting an individual's private characteristics, such as sexual orientation or religious beliefs, using machine learning techniques. [Wang and Kosinski, 2018], for example, showed how deep neural networks could be used to extract facial features and predict one's sexual orientation using profile pictures (unstructured data). Similarly, [Kosinski et al., 2013] showed that social media "likes" can be used to infer other sensitive personal data, such as religious beliefs, with high accuracy. As new techniques are developed to transform unstructured data into structured data, researchers in academia hoping to collaborate with companies may find it increasingly difficult to do so. As a result, academic research on novel information extraction techniques may be hampered by company preferences for keeping structured data on customers as a way to reduce legal exposure under the GDPR.

3.2.3 Data Integration

The InfoQ dimension of data integration is about integrating multiple data sources and/or data types, and linking records across databases. GDPR will likely have a large impact on integration practices, due to the process of pseudonymization, the purpose limitation principle, and by the fact that GDPR applies to EU-residing persons.

For example, the GDPR will likely have a major impact on the integration of personal data from multiple countries. To start, unless multinational firms agree to implement "GDPR level" data processing requirements on their international users, they may need to implement two separate GDPR and non-GDPR data processing pipelines. There is evidence of such action already taken by Facebook, whose CEO, Mark Zuckerberg, stated in an interview that the GDPR will affect their international data collection only "in spirit," but stopped short of creating an explicit company policy for handling non-EU user data[16].

An integration issue that might lead to changes in user behavior is the need to disclose to users how their identifiable data is collected and stored across different devices in a more transparent way. For example, the updated privacy policy of the Google-owned data mining contest platform Kaggle states to users: "When you're not signed into a Kaggle account, we store the information we collect with unique identifiers tied to the browser, application, or device you're using. When you're signed in, we collect information that we store with your Kaggle account."

Data scientists will need to consider their "data environment" and assess whether they have, or potentially have, access to other linkable data that could be used to personally identify data subjects. The concept of a data environment is useful for understanding the contextual relationship between pseudonymized and anonymized personal data. As cited in [Mourby et al., 2018], there are four main components of a researcher's data environment, according to the UK Anonymisation Network:

- Other data: the key question is what (other) data exist in a given data environment? This is what the data controller needs to know in order to assess (some of the conditions for) re-identification risk. Other data consists of co-present databases, data derived from personal knowledge and publically[sic] available sources such as public registers, social media profiles, etc.

- Agency: there is no re-identification risk without human agency – this may seem like an obvious point but it is one that is often overlooked and the least understood.The key issue her is in understanding who the key protagonists are, and how they might act and interact to increase or negate the risk of re-identification. Who are the key actors here?

- Governance processes: these processes are formally determined in policies and procedures which control how data are managed, and accessed, by whom, how and for what purposes.

[16]Facebook to put 1.5 billion users out of reach of new EU privacy law, Thomson Reuters, 19 April, 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-id

- Infrastructure: infrastructure is not dissimilar to governance, it shapes the interaction between data and environment and includes such as operational and management structures as well as hardware and software systems for maintaining security.

If data scientists possess a "means reasonably likely to be used" to re-identify subjects, then such data is considered "personal data" and must be pseudonymized. Consequently, data scientists and academic researchers may find themselves facing stricter controls on which datasets– public or private– might be joined to extant user data in order that might potentially single out individual users.

Data scientists working for data brokers or marketing firms, have a direct and strong motivation to be able to identify specific individuals so that online behavior can be mapped to offline purchase behavior through first, second, and third-party data integration. The data broker LiveRamp (an offshoot of the major data broker Acxiom), for instance, offers a product called "IdentityLink" that allows advertisers a single, "omnichannel view" of the consumer and claims to permit the identification of specific consumers across "thousands of offline 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.[17] Given the economic incentives inherent to such a service, 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 like LiveRamp. Not only would a LiveRamp data scientist have a clear economic incentive for singling out individuals, but she would also have access to a variety of other datasets that could be combined to increase the probability of correctly re-identifying a particular individual. Accordingly, the definitions of pseudonymized and anonymized data under the GDPR 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 datasets, and intrinsic motivations in performing the analysis. For a more detailed discussion on the way in which a researcher's economic incentives can impact the distinction between pseudonymized and anonymized personal data, see [Mourby et al., 2018]. Another key reduction in data integration capabilities that data scientists will be (or already are) experiencing is from cutting or drastically reducing data sharing practices with other companies. Third party data sharing through mutual data sharing agreements or through purchasing has been 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 some data controllers rightly worried (GDPR, Article 82 (2,4)). As a result, we are already seeing some companies disengage from third party data sharing that they have been engaged 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.[18] The latter is especially related to data brokers. At the same time, companies that allow users to link their accounts through well-known third party services (e.g. through Facebook or Google) should clearly express their continuing data collection and use through such an integration. For example, Kaggle's privacy policy states[19]:

> "When you link an account you may have on a third party service (such as a third party social networking site or email provider) to your Kaggle account, Kaggle collects certain information stored in connection with that account that you have configured that service to make available, including your email address, provider ID, first and last name, and profile picture."

3.2.4 Temporal Relevance

The temporal relevance dimension concerns the time line that includes the data collection, data analysis, and study deployment periods as well as the temporal gaps between these periods. These different durations and gaps can each affect InfoQ.

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 are asking users to reconsent to the processing of their personal data. This is probably the safest legal route for companies to retain pre-GDPR user data. Another issue is using deleted personal data. Some companies will continue to use such data, but in aggregate form. For example, Kaggle notes in their revised privacy policy[20]:

> "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."

A related common practice that companies are adopting is setting data retention time frames based on the reason for the data collection. Hence, the duration between data collection and use is directly considered in light of the goal.

[17] Meet LiveRamp IdentityLink, https://lp.liveramp.com/meet-liveramp-identitylink.html

[18] Facebook to stop allowing data brokers such as Experian to target users, The Guardian, Mar 29, 2018 www.theguardian.com/technology/2018/mar/29/facebook-shuts-down-partner-categories-feature-data-brokers-target-users-privacy

[19] https://www.kaggle.com/privacy Accessed May 21, 2018

[20] https://www.kaggle.com/privacy

According to GDPR, websites will need to implement a "privacy by design" approach to data collection, which, in practice, means that they should only collect and process data related to the business context of the relationship with their customers and store them for a specific duration of time (Article 25). As may soon become common practice, firms storing large amounts of customers' personal data in databases will need to regularly "cleanse" them to make sure customer data are either accurate and updated or otherwise deleted. Jim Conning, a managing director at a data consulting firm, remarks that, "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"[21].

On the one hand, asking users to reconsent to processing is one way to verify that users are still actively using the service. It also affords online service providers the chance to allow data subjects to update their information and check for its accuracy. Both of these could contribute to higher information quality in the sense that any personal data held about a data subject is more likely to be factually accurate and up-to-date. Consequently, predictive models based on these data will likely be more accurate than pre-GDPR versions. For data scientists focused on building predictive models for precision marketing, this increase in data quality may offset reduction in training dataset sizes due to data subject opt-outs.

On the other hand, major changes to the data collection policy can make longitudinal studies of user behavior–before and after the policy changes– very difficult. As [Lazer et al., 2014] point out in the case of Google Flu Trends, the way that companies modify their services over time can have a tangible impact on the results of a statistical analysis. Likewise, the exclusion of some individuals due to opting out and also to the restriction in collecting some variables of interest, pre-and-post GDPR, could make long-term studies of user behavior unfeasible.

3.2.5 Chronology of Data and Goal

The choice of variables to collect, the temporal relationship between them and their meaning in the context of the goal all critically affect InfoQ. These considerations are the essence of the chronology of data and goal dimension.

Once the GDPR goes into effect on May 25, 2018, data scientists and researchers may find that some variables previously available to them are no longer being collected, or that the ability to process them has been restricted. In particular, the previously-mentioned "special categories" of sensitive personal data may be off-limits to some data scientists and researchers, some of whom may have previously done research assuming these personal data would be indefinitely available for analysis. For data scientists looking to make predictions at a future date, this could present a problem. A similar issue arises for researchers interested in understanding some behavioral phenomenon using descriptive and explanatory models based on past data. Due to storage duration limits, it may not be possible to build a model using historical data if those data were erased or if a data subject revoked her consent for processing. However, as noted in the previous section, some companies declare that they will continue to use deleted personal data albeit in aggregated form.

The enhanced privacy settings that will likely be offered by online platforms may add another source of unpredictability in terms of which variables may be available at the time of prediction. This can cause heterogeneity in the data available for different users, resulting in many missing values for each single variable.

While GDPR's new regulation on disclosing, rectifying, and deleting personal data will affect data scientists within a company, academic researchers may be able to avail themselves of the exceptions to the GDPR that exist for the processing of personal data for statistical or scientific purposes (GDPR, Paragraph 62). It bears worth mentioning again, however, that since academic-industry collaborations are often not the original motivation for the collection and initial processing of personal data by the company, it is unclear whether secondary analysis by academics would be covered under the above-mentioned exception to the right to disclose, rectify, and delete personal data.

3.2.6 Operationalization

InfoQ describes two types of operationalization: construct operationalization and action operationalization. Construct operationalization refers to the process of utilizing measurable data to operationalize underlying abstractions of interest. For example, a researcher may operationalize the psychological concept of "openness" by counting the number of shared articles on a social media platform.

The GDPR may affect the way in which certain theoretical constructs are operationalized in a research study. Operationalization becomes an issue because GDPR Article 9 spells out the types of "sensitive" personal data that, with some specific exceptions, cannot be processed. These include data that reveal racial or ethnic origin, political opinions, religious or philosophical beliefs, sexual orientation, or trade union membership. Unless companies obtain explicit consent for the processing of such data, data scientists must turn to creative proxies for such constructs of interest, which may lower the scientific validity and generalizability of research results. Nevertheless, some research suggests that non-sensitive data, such as a Facebook user's likes, can be used to infer psychological traits with relatively high predictive accuracy [Lambiotte and Kosinski, 2014]. We reiterate that the GDPR permits several exceptions to the prohibition on processing certain sensitive types of data. Two particularly relevant exceptions are for "scientific" or "statistical research" and for personal data that are "manifestly made

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

public by the data subject" (Article 9 (4)). So although research related to these sensitive categories might be hindered by these prohibitions, the overall effect for data scientists may be mitigated by advances in machine learning techniques that allow for the recognition of extremely fine-grained patterns in data.

Action operationalization refers to deriving concrete actions from the analysis. The GDPR requires that organizations clearly state in their privacy policy "How we use personal data". In revised privacy policies we see specific examples being given, such as[22] "Cortana can use the favorite sports teams you add to the Microsoft Sports app to provide information relevant to your interests". This new type of transparency will likely make users, regulators, and data advocacy groups more aware of potential company actions based on user data. Data science projects that rely on users' ignorance of such uses can be greatly affected.

3.2.7 Generalization

[Kenett and Shmueli, 2016] discuss the difference between statistical and scientific generalization. Scientific generalization is concerned with the application of a model based on a specific target population to other populations, while statistical generalization refers to the ability to infer from a particular sample of records to the target population from which it was drawn. A common problem in statistical studies is ensuring that the sample under analysis is representative of the target population –if the sample differs from the population of interest in a systematic way, then the validity of any statistical inferences from the sample to the population cannot be guaranteed.

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 emotional contagion experiment ([Kramer et al., 2014]), EU data subjects would likely need to give explicit consent for the use of their behavioral data in such experiments and they would also reserve the right to withdraw their consent for processing at any time (Article 7). Such withdrawal can affect model estimation, if the subjects' data were previously used in the training phase. And it will affect to-be-predicted datasets in which the withdrawn subjects were previously listed. 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). In terms of statistical power, the precision with which statistical effects can be estimated, e.g., the width of confidence intervals for population effects, might be affected by the resulting smaller sample sizes. Both of these consequences reduce data scientists' and researchers' ability to generalize from sample to population.

This concern about generalization 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 ([Junghans and Jones, 2007]). The problem of "consent bias" may be exacerbated under the GDPR 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 was done in the Facebook emotional contagion experiment ([Kramer et al., 2014]). Taken to the extreme, 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. As evidence of such a possibility, Facebook has already publicly stated that for non-EU users in Asia, Latin America, and Africa, US privacy guidelines will apply[23].

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. In her discussion of the controversy surrounding biomedical research done mostly in developing regions in Africa, [Iltis, 2006] notes that critics of such trials, "were concerned that the benefits from the studies would accrue primarily to populations in the developed world, since populations in the developing world would not have access to results of the studies. They also argued that these studies would set a dangerous precedent that would lead to using research subjects as low-wage laborers for biomedical research that would benefit the developed world." There is thus precedent for this concern that non-EU data subjects might become "guinea pigs" for initial BBD research 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, the Belmont principle of "justice"–concerned with the norms of fairly distributing benefits, risks, and costs among experimental subjects–might also be violated.

The GDPR's stringent consent standards for EU data subjects could also negatively affect the related notion of *scientific reproducibility*. Scientific reproducibility is concerned with the ability to "recreate scientific conclusions and insights" from previous studies ([Kenett and Shmueli, 2015]). The Future of Privacy Forum's white paper on corporate data sharing

[22] https://privacy.microsoft.com/en-us/privacystatement Accessed May 21, 2018

[23] Facebook to put 1.5 billion users out of reach of new EU privacy law, Thomson Reuters, 19 April, 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-id

([Future of Privacy Forum, 2017]) also highlights this problem: increased personal data privacy standards can hamper attempts to share or reproduce a given statistical analysis because the legal exposure of third-party data processors and of withdrawn consent (drop-outs) for processing. If companies wish to avoid liability for potential data breaches, they may refuse to share data used by the original study, thereby reducing the scientific reproducibility of behavioral research.

Nevertheless, it is important to remember that the GDPR does grant exceptions to the right to erasure ('right to be forgotten') and the right to rectification. For example, Article 89 paragraph 1, regarding the processing of personal data for scientific or statistical purposes, states that derogations may be allowed where "such rights are likely to render impossible or seriously impair the achievement of the specific purposes." A major question that is not resolved in the text of the GDPR is whether user behavior research within a company is considered to be "scientific" or "statistical" research. Until this issue is clearly resolved, it may be safer for data scientists in companies to take a conservative approach to the processing of personal data and assume that at any time an individual user may request to withdraw consent or ask to verify the accuracy of any personal data held about him.

3.2.8 Communication

Communication is perhaps the area where GDPR will have the biggest impact on organizations that collect personal data. Regarding the types of interactions between data controllers and data subjects, Article 12 lays out the major duties, many of which are related to using clear and simple language to explain the grounds of processing. Article 12 also details the data subject's right to have information provided to her should she request it. The theme of clarity is again found in the GDPR's requirement that if personal data are processed on the grounds of consent, that such consent be explicit, with a clear ability to opt-out at any time (Article 7). Further, Recital 58 expands on Article 12, which states data subjects' right to information and spells out the ways in which data controllers must be able to clearly explain–in a non-technical way– to data subjects why their data are being collected and for what purposes. Moreover, if there is to be communication with a child, the language used needs to be appropriate for a child. Initial revisions of privacy policies by several companies provide examples of such clarity, by providing short and clear sections on "what data we collect about you", "how your personal data is used", "how your information is shared", and importantly, contact information (typically including an email address) for data privacy concerns.

Additionally, 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." (Recital 60). With these provisions, companies and their data scientists will need to spend time to make sure that the sometimes complex workings of algorithms can be adequately understood by a non-technical user. Data scientists will thus need to include communicability considerations into their choice of algorithms and solutions and their documentation.

Not only will the GDPR require better company-user communication, but there will need to be a clear communication channel for companies to communicate with their lead data protection authority (DPA). Regular data protection impact assessments and audits (for companies doing large-scale data processing) require firms to keep detailed records and justify the nature of processing (Article 5). Effective data scientists will therefore need to have strong communication skills with several audiences: management, the DPA, and other departments involved in collecting or using user data. Communication skills of data scientists with less or non-technical audiences are therefore likely to become even more important.

In the case of a data breach, companies must report such a breach within 72 hours to the authorities, and also to the data subject "without undue delay" (Recital 85). The addition of mandatory data breach notification periods may have the effect of making companies more reluctant to share user data with others such as academic collaborators. In fact, in the Future of Privacy Forum's survey of corporations that shared data with academics, privacy concerns and fear of data subject re-identification were two of the biggest reasons cited for not collaborating with outside researchers ([Future of Privacy Forum, 2017]).

4 Discussion

The landscape of data ethics regulation is seeing several important changes in year 2018. In Section 2 we described the principles underlying ethics guidelines and regulations such as the Common Rule (and the revised Final Rule) and GDPR. The Final Rule is an update to the Common Rule; GDPR is a much more intensive update of its predecessor, the European Union's Data Protection Directive (DPD) 95/46/EC in the sense that it is the first Regulation, as opposed to the earlier guidelines and directives, or even the more recent attempts of companies at self-regulation (e.g., Facebook's internal ethical committee) or proposals for organizational structures (e.g., see [Polonetsky et al., 2015]). The update to the Common Rule and the "update" of the EU DPP to GDPR have two notable common items: Informed consent and the definition of identifiable data. On consent, two updates refer to (1) the clarity of informed consent, and (2) broad consent.

On clarity of the consent form, the Final Rule includes a new requirement for a concise and understandable consent form that a "reasonable person" would want to have to make a decision about participation. In GDPR, 'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a

statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her (Article 4).

Regarding broad consent, the Final Rule includes a new option permitting researchers to seek broad consent to reuse identifiable private information for other purposes. GDPR postulates that "it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research." At the same time, "data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose" (Article 33).

On the definition of personal or identifiable data, GDPR gives a very precise definition: 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one 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" (Article 4). In contrast to this fixed list, the Final Rule leaves room for new definitions, as technology marches ahead: "the meaning of "identifiable biospecimens" and "identifiable private information" will be reexamined within 1 year and at least every 4 years thereafter in consultation with "appropriate experts"; interpretations of these terms may be altered."

Other than these commonalities, the Final Rule update seems to lower the barrier for using BBD in behavioral research in academia, while GDPR significantly increases the barriers and restrictions on collecting and using BBD by companies.

4.1 Will Industry-Academia Collaborations Become Easier?

As the new GDPR-based culture becomes adopted by industry, it has the potential of bringing ethical standards of industry research and development closer to academic ethical standards. We consider a few such examples:

1. GDPR distinguishes between the *data controllers* who typically are the 'public-facing' entities that data subjects supply their data to, and *data processors* who process the data on behalf of the controllers. In an industry-academia research collaboration that uses BBD collected by the company, such a distinction between a data controller (the company) and a data processor (the researcher) can be useful in determining the researcher's obligations and responsibilities. For example, de-identifying the data would likely be considered the company's responsibility rather than the academic researcher's.

2. Another GDPR definition that can be useful to the academic research environment is the *main establishment*, that defines the supervisory authority when the data controllers and/or processors are in more than one EU state. In an academic cross-country collaboration, defining criteria for "main researcher" can help resolve current ambiguity about ethical obligations when researchers come from countries with different ethical standards.

3. GDPR distinguishes between data *processing* (a broad terms that includes any operation on personal data including storage, organization, use, dissemination, and more) and *profiling*, which refers to any form of automated processing of personal data "in particular, to analyse or predict aspects concerning that natural person's performance at work, economic situation, healthy, personal preferences, interests, reliability, behavior, location or movements". Such a distinction in academic research would create a sharp distinction between behavioral researchers who typically use traditional statistical models and data scientists who use predictive algorithms. Hence, such a distinction might not make sense in academic research.

4. GDPR applies to organizations within the EU and any external organizations that are trading within the EU. In a cross-country collaboration, which country's ethical rules presides? Currently it is the funding source that determines ethical rules (e.g., a US-based researcher is required to obtain IRB approval). What about researchers in countries with no/very-lax ethics regulations who publish their work in major US or EU journals? This situation creates an unleveled playing field for researchers from different countries. One possibility is that journals should define their "country of applicability" to define ethics requirements.

5. As already mentioned, updates to informed consent requirements are common to both the final rule and GDPR: both require a "human-understandable" format.

4.2 Open data ethics issues

Further issues that arise in BBD research, but not addressed by GDPR or the Final Rule, also deserve careful consideration. Two such issues are data ownership and sharing. While GDPR mandates access rights to data, a different issue is data ownership. Data ownership, particularly in the area of web 2.0 and social media data, is a thorny legal issue that has yet to be conclusively settled in courts ([Kaisler et al., 2013]). For example, if companies wish to claim that they own users' social media data, then

they have an ensuing responsibility for the accuracy of such data. A related challenge is the major data brokers partnering with social media platforms (e.g., Axciom, Experian, Equifax) who aggregate online and offline data, mixing relatively verified first-party data with third-party data that has not been properly validated or verified.

Another issue raised recently in The Economist[24] is the centralization of a large proportion of BBD in the hands of a few dominant corporations (Google, Facebook, Amazon) due to network effects. These firms can use this power to extend their dominance, to eliminate competitors, and to affect how users experience the internet. A proposed solution mentioned governing data ownership and exchange, by requiring companies to make their data available for purchase as well as requiring companies, upon request from a customer, to share the customer's data with other firms.

Hacked data is another large source of BBD. Ethics governing the use of hacked datasets is another open question, which requires the use of ethics principles and assessment of risks and benefits.

The GDPR's goal of keeping the EU at the forefront of the modern information economy while ensuring an 'equal playing field' among the EU countries raises a similar need in the context of academic research both in terms of industry-academia collaborations and in terms of cross-country academic collaborations. In the former case, a large gap in ethical rules/regulations between academia and industry creates collaborative challenges, such as the one that manifested in the Facebook emotional contagion experiment that involved researchers from Cornell University. In the second case, research ethics and regulations differ across countries (e.g., in China vs. the USA or EU), thereby creating an 'unequal playing field' that gives researchers from less regulated countries an advantage, and which also creates challenges for collaborations with researchers from more heavily regulated countries. Furthermore, questions arise as to whether the ethical obligations of researchers should be determined by the country they conduct their research, the country where the journal is published, the country of coauthors in the most restrictive ethics regulation location, and more.

4.3 Opportunities for new data science developments

The availability of BBD for research creates a variety of challenges, from ethical to statistical to technical ([Shmueli, 2017b]). For example, linking records across disparate datasets poses statistical challenges (how to link), data quality issues (when some sources are less reliable), and compromises privacy and respect for persons.

By describing the ethical principles behind past and current research and industry data ethics regulation, we hope to provide data science and academic researchers with opportunities for addressing and handling data analysis needs in the BBD era. The new and updated ethical regulations on collection, storage, and usage of personal data not only affects and challenges the data scientist's work, but also provides opportunities for researchers and data scientists, in academia and in industry, for developing data analysis methods and approaches that adhere to the new restrictions and limitations. One example is privacy-preserving machine learning for data that are housed in a distributed fashion. A second example is the recent growth in approaches for "explaining a model's predictions" when blackbox predictive algorithms are used (e.g. [Chen et al., 2017]). Third, it would be useful to map different types of data analyses by their required level of data aggregation, so that individual-level user data that is collected and used for a personalized goal (e.g., in recommendation systems or personalized predictive models) might later be aggregated and reused for group-level modeling, such as through the use of sufficient statistics in generalized linear models. Moreover, perhaps the idea of sufficient statistics should be explored for new purposes such as for data masking (e.g., [Kadane et al., 2006]) and especially for prediction. Finally, there is a need to develop frameworks and processes that support reproducible research given the new limitations on data storage and sharing.

References

American Association of University Professors, 2006. American Association of University Professors (2006). Research on human subjects: Academic freedom and the institutional review board. Technical report.

Beauchamp and Childress, 2012. Beauchamp, T. L. and Childress, J. A. (2012). *Principles of Biomedical Ethics*. Oxford University Press, 7th edition.

Calder, 2016. Calder, A. (2016). *EU GDPR: A Pocket Guide*. IT Governance Publishing.

Cascavilla et al., 2015. Cascavilla, G., Conti, M., Schwartz, D. G., and Yahav, I. (2015). Revealing censored information through comments and commenters in online social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 675–680, New York, NY, USA. ACM.

Chen et al., 2017. Chen, D., Fraiberger, S. P., Moakler, R., and Provost, F. (2017). Enhancing transparency and control when drawing data-driven inferences about individuals. *Big data*, 5(3):197–212.

[24] "Taming the titans", The Economist, Jan 20, 2018, p.11-12

European Parliament, 1995. European Parliament, C. o. t. E. U. (1995). Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal L 281*, pages 31 – 50.

European Parliament, 2016. European Parliament, C. o. t. E. U. (2016). Directive (eu) 2016/680 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing council framework decision 2008/977/jha. *Official Journal of the European Union*, 119:1–88.

Future of Privacy Forum, 2017. Future of Privacy Forum (2017). White paper: Understanding corporate data sharing decisions: Practices, challenges, and opportunities for sharing corporate data with researchers. Technical report.

Gellman, 2017. Gellman, R. (2017). Fair information practices: A basic history.

Iltis, 2006. Iltis, A. S. (2006). *Research ethics*. Routledge.

Junghans and Jones, 2007. Junghans, C. and Jones, M. (2007). Consent bias in research: how to avoid it.

Kadane et al., 2006. Kadane, J. B., Krishnan, R., and Shmueli, G. (2006). A data disclosure policy for count data based on the com-poisson distribution. *Management Science*, 52(10):1610–1617.

Kaisler et al., 2013. Kaisler, S., Armour, F., Espinosa, J. A., and Money, W. (2013). Big data: Issues and challenges moving forward. In *46th Hawaii International Conference on System Sciences*, pages 995–1004.

Kenett and Shmueli, 2014. Kenett, R. S. and Shmueli, G. (2014). On information quality. *Journal of the Royal Statistical Society, Series A*, 177 (1):3–38.

Kenett and Shmueli, 2015. Kenett, R. S. and Shmueli, G. (2015). Clarifying the terminology that describes scientific reproducibility. *Nature methods*, 12(8):699.

Kenett and Shmueli, 2016. Kenett, R. S. and Shmueli, G. (2016). *Information Quality: The Potential of Data and Analytics to Generate Knowledge*. John Wiley & Sons.

Kosinski et al., 2013. Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Kramer et al., 2014. Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academies of Sciences*, 111 (24):8788–8790.

Lambiotte and Kosinski, 2014. Lambiotte, R. and Kosinski, M. (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939.

Lazer et al., 2014. Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Lowthian and Ritchie, 2017. Lowthian, P. and Ritchie, F. (2017). Ensuring the confidentiality of statistical outputs from the adrn.

Martens et al., 2016. Martens, D., Provost, F., Clark, J., and de Fortuny, E. J. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly*, 40(4).

Menikoff et al., 2017. Menikoff, J., Kaneshiro, J., and Pritchard, I. (2017). The common rule, updated. *New England Journal of Medicine*, 376(7):613–615.

Metcalf and Crawford, 2016. Metcalf, J. and Crawford, K. (2016). Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3 (1):1–14.

Mourby et al., 2018. Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., Smith, H., Aidinlis, S., and Kaye, J. (2018). Are 'pseudonymised'data always personal data? implications of the gdpr for administrative data research in the uk. *Computer Law & Security Review*, 34(2):222–233.

Nichols et al., 2017. Nichols, L., Brako, L., Rivera, S. M., Tahmassian, A., Jones, M. F., Pierce, H. H., and Bierer, B. E. (2017). What do revised us rules mean for human research? *Science*, 357(6352):650–651.

OECD, 1980. OECD (1980). Oecd guidelines on the protection of privacy and transborder flows of personal data.

Polonetsky et al., 2015. Polonetsky, J., Tene, O., and Jerome, J. (2015). Beyond the common rule: Ethical structures for data research in non-academic settings. *J. on Telecomm. & High Tech. L.*, 13:333.

Sauro and Lewis, 2012. Sauro, J. and Lewis, J. R. (2012). *Quantifying the User Experience: Practical Statistics for User Research*. Elsevier, 1st edition.

Shmueli, 2010. Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25 (3):289–310.

Shmueli, 2017a. Shmueli, G. (2017a). Analyzing behavioral big data: Methodological, practical, ethical and moral issues. *Quality Engineering*, 29(1):57–74.

Shmueli, 2017b. Shmueli, G. (2017b). Research dilemmas with behavioral big data. *Big Data*, 5(2):98–119.

Tene and Polonetsky, 2016. Tene, O. and Polonetsky, J. (2016). Beyond irbs: Ethical guidelines for data research. *Washington and Lee Law Review Online*, 72(3):458.

Wang and Kosinski, 2018. Wang, Y. and Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246.

Weiss and Archick, 2016. Weiss, M. A. and Archick, K. (2016). Us-eu data privacy: from safe harbor to privacy shield.

Acknowledgements

We thank Soumya Ray and Shin-yi Peng from National Tsing Hua University for their valuable feedback.