國 立 清 華 大 學

碩 士 論 文

Situating Data Science in the Wake of the GDPR: Practical Implications for Data Scientists, Their Managers, and Academic Researchers

International Master of Business Administration (IMBA)

Name：Travis Greene

Advisor：Prof. Galit Shmueli

Date：July 2019

Abstract

In May 2018, the European Union's (EU) General Data Protection Regulation (GDPR) went into effect. The new Regulation is global in scope and will require a shift in the way companies and researchers collect, process, and analyze personal data. To date however, little academic work has focused on how the GDPR will impact data scientists and researchers whose work relies on processing behavioral big data. Data scientists and behavioral researchers would therefore benefit from a deeper understanding of the GDPR's key concepts, definitions, and principles, especially as they apply to the data science workflow. We use the Information Quality framework by Kenett and Shmueli (2014) to identify key GDPR concepts and principles and describe how they affect typical data science work. Because of the unprecedented global reach of the GDPR, we also consider its impact on personal data transfers for international corporations and research collaborations. As many data scientists come from a STEM background, we place special emphasis on situating the GDPR within a broader social, legal, political, and economic context. In this new data privacy regulation era, data scientists and researchers must know not only their legal obligations under GDPR, but also be aware of the potential social and political implications of their work.

Key words: data science, data regulation, personal data, privacy law, behavioral big data, industry-academic collaboration

Contents

List of Figures

List of Tables

List of Abbreviations

APEC Asia-Pacific Economic Cooperation

BBD Behavioral Big Data

BCR Binding Corporate Rules

CCPA California Consumer Privacy Act (2018)

CFR Charter of Fundamental Rights of the European Union

CJEU European Court of Justice

CSL Chinese Cyber Security Law (2017)

DPA Data Protection Authority

DPD Data Protection Directive (1995)

DPIA Data Protection Impact Assessment

DPO Data Protection Officer

ECHR European Convention on Human Rights

ECSC European Coal and Steel Community

EDPB European Data Protection Board

EEA European Economic Area

FTC Federal Trade Commission

GDPR General Data Protection Regulation (2018)

HCI Human-Computer Interaction

MAR Market Abuse Regulation (2014)

OBA Online Behavioral Advertising

OECD Organisation for Economic Co-operation and Development

PII Personally Identifiable Information

SLA Service Level Agreement

Chapter 1

Introduction

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. Many industries including retail, marketing, and advertising, now take advantage of technologies such as GPS and facial recognition software,[1] originally developed by military and security agencies, to collect and process data for purposes of surveillance, anomaly detection, and prediction. (de Leeuw and Bergstra, 2007; Turow, 2017; Mansfield-Devine, 2013). The resulting Behavioral Big Data (BBD) includes not only rich personal data but also social networks connecting individuals (Shmueli, 2017). At the same time, this rapid technological advance has far outpaced the speed of updates to ethical research codes and regulation of human subjects' data collection, storage, and use (Zook et al., 2017). The ever-widening gap has motivated data science researchers to call for the creation of general ethical principles and guidelines to effectively balance the potential social and scientific benefits of BBD processing with its potential privacy costs (Hand, 2018).

The European Union's new General Data Protection Regulation (GDPR), which took effect on May 25, 2018, is poised to change the course of these developments.[2] The GDPR is especially important because although there has been a long-standing *directive* on the use of personal data in the EU,[3] *regulation* – which transcends national legislative processes and laws and has immediate application and enforcement in all EU Member States – is only taking place now. The ostensible reason for updating the 1995 Directive was to keep the EU at the forefront of the modern information economy, while ensuring an 'equal playing field' among the EU countries (Part I of this Thesis will survey these reasons in more detail.). Additionally, heterogeneity in national implementations of the Directive resulted in inefficiencies in the "free movement of personal data within the internal market" (Calder, 2016). The GDPR was designed to resolve these issues by placing limits and restrictions on the use and storage of personal data by companies and organizations operating in the EU and abroad, insofar as these organizations "monitor the behavior" of or "offer goods or services" to EU-residing data subjects (Article 24). The GDPR thereby has the potential to affect any company or organization processing the personal data of EU-based data subjects, regardless of where the processing occurs (Calder, 2016). Specific implications of the GDPR for multinational corporations will be covered in Part II of this Thesis.

While academic research using human subjects' data in most developed countries has been strictly regulated,[4] the collection, storage, and use of personal data in industry has historically faced much less regulatory scrutiny (see, e.g., Federal Trade Commission, 2012). Nevertheless, this "hands-off" approach to industry data collection and processing seems to be changing as GDPR comes into effect and large BBD-processing corporations, such as Facebook and Google, report massive personal

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

[2] The GDPR was drafted in 2016, but did not come into force until 2018. To be consistent, wherever data laws are mentioned, dates will refer to the year the law came into force.

[3] 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)

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

data breaches.[5] Further, new models of collaboration for industry-academia BBD research, such as that between Facebook and the Social Science Research Council (King and Persily, 2018), highlight the increasing importance of academic ethical codes on industry BBD research.[6]

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

1. What are the main GDPR terms and principles that a data scientist should be familiar with?

2. How should these technical and ethical principles be incorporated into data science workflows?

These questions are worth exploring because–in some cases–researchers appear to be unaware of the regulatory unification relating to the collection, access, and usage of BBD brought about by the GDPR (Olshannikova et al., 2017). Yet in other cases, some social scientists seem to already have incorporated key GDPR principles into their BBD research, such as the principle of data minimization and the weighing of potential benefits and harms of large-scale BBD processing. As just one example, Garcia et al. (2018) very clearly and prominently describe the ethical considerations of their international research collaboration. Such clear and upfront acknowledgements of the risks and benefits of large-scale personal data processing are likely to become increasingly common in published academic research using BBD.

Regardless of how academic research is ultimately affected by the GDPR's principles, it would behoove data scientists and researchers to understand what is new and how the new regulations and legal environment might affect their routines, approaches, priorities, and possibilities. After all, non-compliance with GDPR can–in the most egregious of cases–result in heavy financial penalties of up to €20 million, or 4% of the worldwide annual revenue of the prior financial year, whichever is higher. [7] [8] [9]

Finally, the GDPR is worthy of study as it increasingly wields influence on the global discourse surrounding the debate on personal data and privacy. A recent news article concluded that the GDPR and California's new Consumer Privacy Act (CCPA) are "pushing the tech industry to the negotiating table," and federal legislation is currently being drafted in Congress in order to create an "Internet Bill of Rights."[10] Additionally, since 2016, there has been a wave of similar, GDPR-inspired regulations being passed–or at least seriously considered–by several other countries including China,[11] a large group of ten Ibero-American states (including Brazil, Mexico, Colombia, Chile, Peru and Uruguay),[12] [13] India,[14] and Malaysia.[15] Given the economic influence of these countries and the increasing reliance on cloud technologies to collect and transfer personal data around the world, data scientists

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

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

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

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

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

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

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

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

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

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

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

would be remiss if they did not have a basic understanding of how various governmental regulations may impact the future development of their industry as well as their day-to-day routines. Further, companies hoping to benefit from collaborations with academic researchers should be aware of the major legal principles regarding personal data protection and analysis. In short, the language and legal concepts found in the GDPR have already deeply shaped the international discourse surrounding personal data collection and processing, making an understanding of the Regulation even more relevant for modern day data scientists in the global economy.

The present work is divided into three parts, each of which may appeal to different audiences. Chapter 2 provides a condensed summary of the scope and content of the GDPR document. Part I is of interest to the general reader and introduces the GDPR while surveying some key justifications for updating the 1995 Data Protection Directive (DPD). Following the justifications for the GDPR, several influential responses to the GDPR are discussed and situated in terms of their political, legal, economic, business, and technical implications. Part II is most relevant for data science managers and executives wishing to get a deeper understanding of how the GDPR fits into the broader legal context surrounding international personal data processing. This part will be of particular interest to firms located in both the EU and the USA. Finally, Part III makes up the bulk of this work and is aimed at practicing data scientists and researchers. Consequently, it uses a data science framework called InfoQ to organize and analyze important GDPR concepts and principles that are likely to directly impact their work. Part III can be read on its own. We believe, however, that most readers would benefit from understanding the GDPR as being situated in the various social-political, legal and economic contexts outlined in Parts I and II.

This thesis contributes to the small but growing intersection of research in data science, technology management, and the downstream social and political effects of machine learning and AI. Rahwan et al. (2019), for example, have called for a new discipline, *machine behavior*, to study these complex interactions between AI and society. Similarly, this thesis joins together several frameworks to organize the GDPR's key concepts, definitions, and principles in a meaningful way for three main audiences: the concerned global citizen, the manager of a multinational, data-centric firm, and the data scientist and modern academic researcher. Borrowed and adapted from diverse fields– the *three domain* framework comes from corporate social responsibility and the *InfoQ* framework from data science–the inclusion of these frameworks highlights the inherently interdisciplinary nature of this work. Without this organizing force, the GDPR document itself can be quite overwhelming, especially for those without any formal training in (European) privacy and data law.

It is the opinion of the author, and also of many commentators, that a society-wide debate around the social, legal, and political implications of the collection and processing of personal data is needed. In light of this need, the present work aims to properly situate the GDPR in its wider societal context and eschew its popular characterization in the news media as a mere collection of arbitrary processing rules with stiff financial penalties. If successful, data scientists and their managers will understand the GDPR as one attempt to address a crucial question: How should we regulate the ownership and processing of personal data (Harari, 2018)?

Chapter 2

A Brief Overview of the GDPR

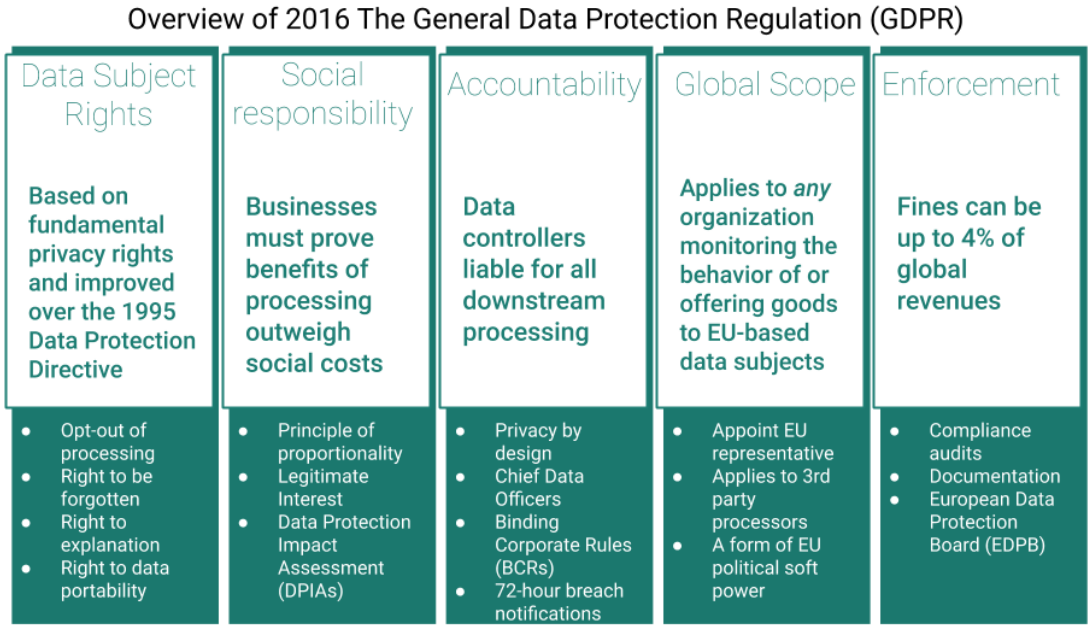

Figure 2.1: The GDPR at a Glance

Background

The 2016 General Data Protection Regulation (GDPR) received considerable media attention in the run-up to its enactment in 2018. This was partly due to the relatively short two-year implementation period and its new international scope. Nevertheless, the GDPR did not appear out of thin air: it is merely Europe's most recent attempt at addressing fundamental issues of privacy and personal data processing in the current age of Big Data.

Data protection and privacy law have a long tradition in Europe. Since the 1950 European Convention on Human Rights (ECHR) and the 1981 Council of Europe Convention regarding automatic processing of personal data, to the 1995 Data Protection Directive (1995 DPD) and finally to the 2007 Treaty of Lisbon and the 2016 GDPR, European lawmakers have for decades attempted to set down guidelines for the protection of individuals' privacy and personal data. These attempts were complicated by a variety of factors, chief among which were the gradual evolution of a purely economic association–the 1951 European Coal and Steel Community (ECSC) (Dedman, 2006)– into what we today call the European Union, and the rapid advance of digital technology in the two

decades since the 1995 DPD (Reding, 2011).[1]

It is important to note that the GDPR only regulates the processing of *personal data*, that is, data that can be directly or indirectly linked back to a natural living person.[2] Anonymized and public data are therefore outside the scope of the GDPR. However, given recent advances in de-identifying individuals by combining various datasets, the legal definition of personal data is flexible, contextual, and will likely continue to grow with time. See section 7.6.2 for a discussion of issues related to the contextual nature of personal data.

Structure of the GDPR

The GDPR is a long and complex legal document partly because it is a Regulation and not a Directive. In EU law, Directives and Regulations are legally distinct acts used to achieve EU-wide goals set forth in various EU treaties.[3] Unlike Directives, Regulations are not translated into national law by individual member states; as a result, Regulations generally see more uniform application and enforcement. The "direct horizontal effect" of the GDPR allows data subjects to litigate against both private actors and public bodies and thus requires considerable clarification relative to a Directive (Granger and Irion, 2018).

In its entirety, the GDPR contains 99 Articles, which are grouped into 11 chapters. Many of the Articles, however, deal with regulatory and bureaucratic issues irrelevant to most data scientists, managers, and academic researchers. The Articles constitute the legally binding portion of the Regulation, while the Recitals serve to give extra context and clarification of the Articles. In many cases, the Recitals will be more useful to the average reader than the corresponding Article's text. Interested readers can find many of the terms and principles related to data science in the Glossary.

As can be seen from the chapter headings, the main legal entities of interest are *data subjects*–the natural living persons about whom personal data are collected, *data controllers*–the legal entities responsible for determining how personal data are processed, and *data processors*–the legal entities who carry out the processing on behalf of the controllers. The GDPR's content consists of the following 11 chapters:

1. General provisions
2. Principles
3. Rights of the data subject
4. Controller and processor
5. Transfers of personal data to third countries or international organisations
6. Independent supervisory authorities
7. Cooperation and consistency
8. Remedies, liability and penalties
9. Provisions relating to specific processing situations
10. Delegated acts and implementing acts
11. Final provisions

[1] Interested readers should refer to Dedman (2006) to learn more about the transition from the ECSC to the modern-day EU.

[2] The GDPR expands the notion of personal data from the 1995 Data Protection Directive to now include things such as IP addresses, geolocation coordinates, and *sensitive data* such as health, genetic and biometric data.

[3] europa.eu/european-union/eu-law/legal-acts_en

Unlike the US, where personal data processing occurs on largely an "opt-out" model, the European approach is based on data subjects "opting-in" to processing. In other words, data subjects must decide to opt-in (i.e., give consent) to personal data processing when other legal bases of processing are not present. When no legal bases are present and the data subject has not consented to the processing of his personal data, then no processing can take place. The six lawful bases of personal data processing are given in Article 6:

(a) If the data subject has given consent to the processing of his or her personal data;

(b) To fulfill contractual obligations with a data subject, or for tasks at the request of a data subject who is in the process of entering into a contract;

(c) To comply with a data controller's legal obligations;

(d) To protect the vital interests of a data subject or another individual;

(e) To perform a task in the public interest or in official authority;

(f) For the legitimate interests of a data controller or a third party, unless these interests are overridden by interests of the data subject or her or his rights according to the Charter of Fundamental Rights (especially in the case of children). (See section 7.4 for a discussion of this important legal basis of processing.)

Data Subject Rights

Under the GDPR, data subjects inherit many of the same rights they had under the 1995 Directive, plus a few notable additions. Articles 12-23 spell out these rights. Data subjects now enjoy the *right to be forgotten* (data subjects can request deletion of their data) and the *right to data portability* (data subjects can request a portable, electronic copy of their data). Generally speaking, the rights of data subjects under the GDPR can be categorized as related to *transparency* (i.e., communication with data subjects should be clear and easily intelligible), *information and access* (i.e., who collected the data and for what purpose(s)?), *rectification and erasure* (i.e., how can data subjects correct false information and delete old information?), and objection to *(automated) processing* (i.e., removing consent to processing of any personal data including algorithmic decision-making).

Social Responsibility

The GDPR places emphasis on the broad social effects of personal data processing (see section 7.4 for an overview of the European historical context influencing the GDPR). Recital 4 clearly tells us that the purpose of personal data processing is to "serve mankind." Consequently, one's rights to object to processing are not absolute; they must be weighed against broader societal benefits of such processing. The process of weighing is called the *principle of proportionality.* In this sense, the GDPR can be seen as taking a utilitarian approach to personal data processing. To this end, Data Protection Impact Assessments (DPIAs) must be carried out when large-scale processing may pose privacy risks to data subjects.

Accountability

The GDPR aims to make data controllers legally accountable for the processing of personal data, even when the actual processing is done by a third-party data processor. Section 7.8 examines how this could affect the ability of companies to share data with academics. All companies need to keep detailed records of processing and compliance; in some cases, companies will also need to appoint a Chief Data Officer.

Six overarching principles of data processing form the basis of the GDPR (table 7.1 describes each principle in more detail). Many of these principles are modern versions of the OECD's 1980 *Recommendations of the Council Concerning Guidelines Governing the Protection of Privacy and*

Trans-Border Flows of Personal Dataflows.[4] Very similar principles also guided the Safe Harbor agreement that allowed US companies to process Europeans' personal data from 2000-2015 (see section 5.3 for more information). Article 5 and Recital 39 state that the GDPR is founded on the following data processing principles:

- Lawfulness, fairness and transparency
- Purpose limitation
- Data minimization
- Accuracy
- Storage limitation
- Integrity and confidentiality

Global Scope

The GDPR breaks with the Directive in a major way by applying not just to data controllers, processors, and subjects in the EU, but to *any organization* that monitors the behavior of or offers goods and services to data subjects located in the EU. In other words, a data controller with headquarters in Taiwan must respect the GDPR if processing the personal data of data subjects residing in the EU. Conversely, a German data controller must respect the GDPR when processing the personal data of a Taiwanese tourist in the EU. As far as the GDPR is concerned, location, rather than citizenship or residency status, determines whether data subjects receive must receive the rights outlined Articles 12-23. In some cases, organizations will also need to appoint an EU representative.

The global scope of the GDPR is unprecedented and has important implications for public and private international law (see 5.5.1 for a more detailed discussion). For example, European personal data can only be transferred in countries deemed to have adequate personal data protection. So far there are 13 countries that meet the adequacy standard of the European Commission. For multinational companies that rely heavily on international personal data transfers to countries without adequate protection, the GDPR has introduced Binding Corporate Rules (BCRs), which are legally binding codes of conduct approved by the Data Protection Authorities (DPAs) that allow for international personal data transfers. A treatment of the advantages of BCRs relative to more traditional standard contractual clauses can be found in section 5.3.5.

Enforcement

The GDPR again differentiates itself from the American approach to regulation and enforcement by specifying large administrative fines which may reach as high as 4% of firms' global turnover. These fines are levied by member state Data Protection Authorities (DPAs). Individual data subjects and independent organizations representing these data subjects (e.g., consumer protection and privacy interest groups) may also petition for damages when rights are violated. These damages, however, are assessed differently than the administrative fines and in general are much lower than the administrative fines. On a related note, data breaches must also be reported within 72 hours of their occurrence.

Finally, it should be noted that the GDPR provides many *derogations*, or exceptions, when the processing is done for scientific research, for law enforcement, or for private use by individuals in their homes. Many of these exemptions are summarized in section 7.1.

Figure 2.1 summarizes the above-mentioned points concerning the GDPR.

[4] www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html

Part I

The GDPR: Justifications and Responses

Overview of Part I

International transfers of personal data are essential to the global economy. Multinational corporations must keep and process the personal data of their customers, suppliers, vendors, and others in their supply chains. Human Resources departments must collect and process the personal data of thousands of employees around the world. Corporations increasingly rely on storage and processing services in the cloud, which makes tracing international movements of data complicated. Given these complexities, how can governments ensure that citizens' privacy rights are respected?

Since the 1995 Data Protection Directive (DPD), European political and legal thought regarding personal data collection and processing has continuously evolved. For example, The Charter of Fundamental Rights of the European Union (CFR), which was drafted in 2000 and went into effect in 2009 with the Treaty of Lisbon, laid down fundamental rights to privacy and protection of personal data. New digital technologies, the expansion of global data transfers, and the ability of law enforcement to access citizens' personal data prompted European policymakers to employ a new legal instrument–a Regulation–that would harmonize personal data laws in all EU Member States. No longer would Member States be relatively free to transpose the 1995 Directive into their own national laws.

The European Parliament members who drafted the GDPR cited many further reasons for the Regulation. Chief among these was to reduce bureaucracy and eliminate legal inefficiencies in international transfers of personal data. This would also have the effect of reducing regulatory uncertainty and improving legal coherence in the EU, which would benefit business. Secondly, the drafters wished to foster business innovation by creating new "privacy-friendly" technologies. It was hoped that these privacy protections would make it easier for consumers to trust new technologies. At the same time, the GDPR strengthened data subjects' privacy rights and forced data controllers to assume greater legal responsibility for violations of the GDPR. EU policy makers believed that the GDPR would serve as an example of successful governmental regulation of technology that might be transplanted around the globe and improve standards of privacy for people everywhere.

Yet the reaction to the GDPR by politicians, lawyers, academics, and other experts has not always been positive. Some have questioned whether the cultural norms of privacy in the GDPR can be exported to non-EU countries; others have stated that while the GDPR's intentions are good, it is too late as digital technology has already begun to shape everyday life. At least one commentator envisions a post-GDPR world splintered into competing government-controlled privacy regimes. If the privacy protections trump business interests, then some EU businesses will simply go elsewhere to process personal data, thus hampering technological progress.

On the technical side, critics have taken aim at the GDPR's vague provision of a right to explanation, particularly when automated profiling is involved. There have also been conceptual critiques of the GDPR that claim key aspects of it are simply incompatible with the methods of Big Data analysis. Additionally, technology experts are still undecided about how companies can best implement the GDPR's right to be forgotten.

On the economic front, some observers believe the GDPR could foster a data-backed securities market, based on data subjects' consent to processing. Similarly, the GDPR's massive fines may also help to spur a GDPR insurance market in some Member States. In terms of economic investment, economists have found reduced investment in new technology ventures just after the passage of the GDPR. Finally, some are worried that the GDPR will widen the data divide between small and medium-sized enterprises (SMEs) and large technology companies. This is because larger companies have greater access to rich BBD and also large budgets for covering GDPR compliance costs and (potential) fines.

Chapter 3

Justifications for the GDPR

Introduction

This chapter motivates the need for international data regulation and provides a deeper look into the political, social, and economic reasons for updating the 1995 Data Protection Directive (DPD) to the 2018 General Data Protection Regulation (GDPR). We note especially that some of the GDPR's updated principles were the result more of maintaining legal coherence with fundamental European privacy rights than of adjusting to new technological developments.

What follows are key arguments made by European Commission bureaucrats, legal scholars, technological commentators, and some general comments from the author regarding the validity of such arguments. Table 3.1 summarizes these ideas. In short, the major reasons behind the implementation of the GDPR are increased regulatory certainty, creation of the EU Digital Single Market, establishment of global norms and cultural influence, fostering of business innovation, and strengthening of consumer trust in new technologies.

Finally, this section will examine two criticisms of the European approach to data regulation prior to the implementation of the GDPR. In short, these are 1) national data localization requirements impede global information flows and could lead to a government-controlled "splinternet"; and 2) data regulation has historically been ineffective in the EU. The section concludes with a brief discussion of how the new provisions in the GDPR could address some of these issues.

Table 3.1: Summary of Reasons for Updating the 1995 Directive to the 2018 GDPR

Reason	Summary
Reduce bureaucracy	Under the 1995 DPD, international transfers of personal data were especially complicated and in need of streamlining
Increase regulatory certainty	For example, in some member states under the 1995 DPD, consent had to be "expressly" given (even in writing), whereas in some other states it could be implied (Reding, 2012)
Establish global norms and cultural influence	GDPR can be seen as a soft-power approach to spreading awareness of privacy rights, an alternative to the Chinese model in the "global privacy marketplace"
Foster business innovation	Promotion of existing and new privacy preserving data-mining techniques; e.g., creation of companies such as Anonos and Integris Software that allow companies to manage data privacy in novel ways
Strengthen consumer trust in new technologies	Industry lobbyists agree that better, clearer privacy laws would help consumers to trust BBD-based firms
Improve legal coherence	Some privacy-related rights in GDPR were already part of foundational EU law and therefore had to be included
Strengthen privacy rights of individuals	E.g., Clarification of consent to be "explicitly" given (no implied consent) (Article 7); addition of the right to be forgotten and the right to data portability (Reding, 2012)
Create a Digital Single Market	The EU strategy has three pillars: 1) removing barriers to cross-border online content; 2) ensuring safety and security of high speed digital networks; 3) enhancing citizens' digital skills to maximize the benefits of the digital economy
Improve regulatory enforcement	Enhanced remedies and judicial procedures for violations of data subjects' rights. For example, data protection associations can bring actions to courts on behalf of a data subject, without necessarily getting authorization from the data subject (Reding, 2012)
Increase legal responsibility for data controllers	Under GDPR data controllers follow a "principle of accountability" for adherence to the Regulation. Though it reduces bureaucratic red tape, the GDPR places more responsibility on data controllers for all downstream processing, including detailed documentation, and adherence to principles of data protection by design and data minimization, among others (Reding, 2012)

3.1 The Importance of Personal Data in the Global Economy

Today's global economy demands the unfettered and unrestricted transmission of personal data across international borders. The U.S. Chamber of Commerce reports that "nearly all businesses transfer a combination of employee, consumer, and corporate personal data across borders as part of the everyday business functions (US Chamber of Commerce, 2014)." Multinational firms in particular are reliant on the free flow of personal data[1] across international borders. For example, a multinational firm might need to transfer personal data of employees stationed abroad for performance reviews and human resources projects; similarly, the firm may need to transfer client information from one foreign subsidiary to the main headquarters in another country. A multinational with a global supply chain, such as Apple, must store and process the personal data of its thousands of employees, suppliers, and retailers around the world. Without a smooth transfer of personal data, the global economy would quickly come to a standstill.

Further, due to cost considerations, corporations are increasingly outsourcing their centralized IT systems to third party processors, often via the cloud (Moerel, 2012). Yet, cloud technologies and techniques like dynamic routing make it more difficult than ever for regulators to definitively predict where and how personal data may be stored (Moerel, 2012). What happens, for example, when a multinational's domestic privacy laws conflict with the national laws of a third-party processor? Where and how can complainants demand legal redress for privacy violations? Because of the difficulty of answering such questions, global corporations are increasingly adopting worldwide corporate privacy policies in order to deal with the enormous variation in national data protection laws. We will see in section 5.3.5 how the GDPR's Binding Corporate Rules (and other legal means) may be one solution for reducing the complexities of international data transfers among corporate entities. But more broadly, the GDPR was designed to deal with these increasingly common situations in which the personal data of Europeans is stored, processed, and analyzed in different regulatory environments. The following sections discuss in detail why European regulators felt the GDPR was needed.

3.2 Regulatory Certainty & Reduced Bureaucracy

Foremost among reasons for the passage of GDPR was improving regulatory certainty from the earlier 1995 Directive across EU member states (Voigt and dem Bussche, 2017). Regarding the long legal journey from Directive to Regulation, Calder (2016) writes:

> Across the EU, other, similar [to France and Germany's] legislation was enacted, but through a combination of time and varying national interests, no two national laws were sufficiently similar for an organization to simultaneously be compliant in its home country and across all other EU member states. That is, the free flow of information was effectively inhibited because the different regulatory environments clashed on matters of detail, requiring businesses and governments alike to arrange processes specific to an increasing array of scenarios.

By reducing legal uncertainty and bureaucracy, the GDPR affects business in at least two ways. First, in varied and uncertain legal and business environments, companies must spend more resources on legal compliance, insurance, and risk assessment. Rather than be caught out by some unforeseen event, a prudent company would rather be "safe than sorry" and prepare various contingency plans. These plans take time to develop and divert considerable amounts of resources and labor away from other important tasks in the business. Further, economic logic dictates that these costs will then be passed on to consumers and society in the form of higher prices for goods and services.[2]

[1] By "personal data" we assume the GDPR's definition of personal data as "any information relating to an identified or identifiable natural person" (GDPR Article 4 (1)).

[2] Some estimates are that Fortune 500 companies will need to spend roughly $7.8 Billion on GDPR compliance. Analysts are noticing that some of these compa-

Second, even if customers do bear the brunt of GDPR-related compliance costs, increased costs may motivate purely self-interested corporations to undertake cost-benefit analyses weighing the monetary penalties of violation and the probability of being caught against the costs of legal compliance. If the expected value of not complying is less than the price of compliance, then a profit-seeking corporation may choose not to follow the law.[3] In short, when corporations operate in uncertain business environments, we can in general expect higher costs of doing business and also increased levels of ethically and legally-dubious behavior.

One major goal of the GDPR over the 1995 Directive, however, is to make it simpler and easier for European companies to know their specific duties and obligations to data subjects and authorities. By reducing regulatory uncertainty among the various member states, the GDPR could potentially reduce corporate compliance expenses in the long-term–though the initial compliance costs will likely to be quite large for most firms.[4] To this end, the GDPR will create a new European Data Protection Board (EDPB)[5] that will force the Data Protection Authorities (DPAs) of member states to follow consistent enforcement and interpretation of the GDPR through a so-called "consistency mechanism (Albrecht, 2016)." The end result should be, according to Albrecht (2016), a dramatic "improve[ment] [in] legal certainty and coherence in the area of data protection law." This sentiment about how improved regulatory certainty benefits business and lowers compliance costs is also echoed in the United States. In the Federal Trade Commission's (FTC) call for comments on new privacy guidelines, companies such as IBM, General Electric, and AT&T all stated that the consistency between different privacy regimes promotes international competitiveness and increases compliance with privacy standards (Federal Trade Commission, 2012).

3.3 Legal Coherence and Precedent

Much of the GDPR's content borrows from European notions of fundamental human rights and freedoms. The general importance of the protection of personal data in the 1995 DPD, and later the GDPR, stems from Articles 7 and 8 of The Charter of Fundamental Rights of the European Union. These Articles, respectively, guarantee "respect for private and family life" and protection (along with access and rectification) of one's personal data under the stipulation that any personal data processing must be done "fairly" and on the basis of consent or "some other legitimate basis laid down by law (EU Charter of Fundamental Rights)." Any subsequent evolution of the 1995 Directive therefore had to include these fundamental rights.

Regarding this link between foundational European legal principles and the GDPR, Zarsky (2016) provides another salient example. He notes that the GDPR's concept of purpose limitation was "enshrined" in the bedrock of European law, (i.e., The Charter of Fundamental Rights of the European Union), and the GDPR's drafters "had no choice but to incorporate it in full within the Regulation... [because any] step short of that would have risked the invalidation of the GDPR by the European Court of Justice." This revelation is interesting because it suggests that much of the content of the GDPR was included, not because of anything related to technological advance or business concerns, but because it was needed for legal coherence with broader European ideas about human rights and privacy. If the GDPR did not contain these principles, then it would risk later invalidation when tested in the courts.

---

nies are passing on the compliance costs to their customers. iapp.org/news/a/should-vendors-be-able-to-pass-along-costs-of-gdpr-compliance/

[3]For an example of how a firms might do a cost-benefit analysis of GDPR compliance, see blogs.informatica.com/2017/02/09/deprioritising-gdpr-risk-worth-taking/#fbid=NfvnvBAqgHr

[4]Nearly 70% of multinational corporations surveyed by PricewaterhouseCoopers are expecting to spend between $1-10 Million USD on GDPR compliance. www.cio.com/article/3161920/article.html

[5]This board replaces the previous Article 29 Working Party advisory group.

3.4 GDPR as a Global Gold Standard: Advancing European Soft Power

Following the political and economic collapse of Europe in the wake of World War II, the German sociologist Jürgen Habermas wrote that Europe might be "given 'a second chance' to influence world history" through "a 'non-imperial process of reaching understanding with, and learning from, other cultures (Hettne and Soderbaum, 2005), quoting M.J. Heffernan (1998)." In other words, instead of the extractive colonialism that characterized much of Europe's foreign influence in other parts of the world prior to World War II, the ensuing post-modern era would instead be marked by the importance of ideas and cultural exchange among nations. Nearly seventy years later, "values, norms, and principles" now constitute "the backbone of the European foreign policy doctrine (Michalski, 2005)." In the post-modern period, Europe has effectively decided that the power of the idea trumps that of the sword.

Indeed, one of most outspoken framers of the GDPR, Jan Philipp Albrecht, a German European Parliament minister, boasted that the GDPR will "change not only the European data protection laws, but nothing less than the whole world as we know it" by setting a "global gold standard for every new innovation, for consumer trust in digital technologies and for an entry point to the growth opportunities of an emerging digital market (Albrecht, 2016)." Needless to say, the framers of the GDPR are endlessly optimistic about its potential to influence conceptions of human rights both at home in the EU and abroad.

Put differently, instead of acquiring foreign lands and resources by force (as in the colonial era), the modern European Union could use the GDPR as a tool to gain access to new digital markets by convincing newly-developed countries to value things like "consumer trust" and "privacy by design." It would be a win-win for both parties: Europeans would have access to new global markets and non-Europeans could enjoy increased personal data protection–assuming, of course, that the restriction of surveillance by government and big business is necessarily a good thing. We will see that in China, for example, the Chinese Communist Party leadership does not appear to be operating under this assumption.[6]

In this interpretation of Europe's broader geopolitical strategy, one could argue that the GDPR reflects the EU's "soft power" approach to international relations. The political theorist Joseph Nye defines soft power as, "the ability to get what you want by persuading and attracting others to adopt your goals (cited in Hettne and Soderbaum (2005)." Essentially, by tying implicit conceptions of human rights into the GDPR, European politicians are betting that the EU can influence policy in countries where there are no similar notions. Brussels believes that in the global marketplace of ideas, its conception of fundamental rights to privacy and data protection will become the dominant currency of global commerce in the Internet age.

3.4.1 European vs. Chinese Approaches to Data

This puts the GDPR in stark contrast to recent regulation coming out of China, a country which has come under frequent attack for its censorship of anti-government information and denial of fundamental rights to privacy.[7] The near future may therefore witness a fierce battle between these two economic, cultural, and political giants as developing nations in East Asia are forced to choose between opposing European or Chinese approaches to personal data regulation. The "European approach" could be viewed as founded on fundamental privacy rights, consent, and restriction of processing to situations in which the benefits to society outweigh any harms; while the "Chinese approach" justifies vast personal data processing by government and state-backed businesses as necessary to promote social trust and provide the backbone for an AI-based, society-wide system that seeks to label individuals as either "trust-keeping" or "trust-breaking" (Chen et al., 2018). One

[6]This however may change as China contemplates amendments to its 2017 Cybersecurity Law which gives data subjects GDPR-like opt-out rights for automated profiling and personalization algorithms: www.channelnewsasia.com/news/commentary/china-great-leap-forward-in-data-protection-11429624

[7]See for example, recent news stories about China's infamous "social credit scoring" system: www.wired.co.uk/article/china-social-credit-system-explained

major benefit of the relatively opaque data-processing laws in China is that they ease the collection of biospecimens, DNA sequences, and health data so that Chinese scientists and businesses can stay at the forefront of genetic sequencing and personalized medicine (Metzl, 2019). With a massive national database, researchers can then find interesting correlations between individuals' genetic profiles and certain behavioral and medical conditions. These findings may then provide the basis for new businesses, technologies, and scientific discoveries. In sum, the Chinese strategy rests on the wager that the fast-growing states of East Asia will value the gains in social harmony, improved allocation of business capital–through more accurate credit scoring information–and medical research over the more individualistic, restricted, and potentially business-hampering European approach.

3.4.2 The 2014 Market Abuse Regulation

The EU's 2014 Market Abuse Regulation (MAR) reflects a similar attitude towards European influence abroad, albeit this time in the realm of regulation of financial instruments. Similar to the territorial scope of the GDPR, this regulation effectively creates international jurisdiction for any insider trading of European financial instruments regardless of where the insider trading occurs (Regulation (EU) No 596/2014). The Regulation goes even further and states any trading that has an effect on financial instruments traded in European markets is also under the scope of the regulation (ibid.). With both the GDPR and the MAR, the EU could be interpreted as exerting a form of extra-territorial soft power across the world in order to gain first-mover advantage in shaping worldwide business and data processing regulations.

But the impetus for making the GDPR and the MAR applicable internationally may be based less on altruism and more on *realpolitik*: there are clearly economic advantages in setting global standards for data processing. Indeed, we see countries such as Taiwan negotiating with the EU in order to be placed on the list of countries with "adequate" personal data protection.[8] Once deemed to have adequate personal data protection, firms could freely transfer personal data across borders without the need for complex and expensive legal contracts, also known as standard contractual clauses.[9] Further, in the case of Taiwan, the acceptance of the European approach may not only be economically motivated, but may also be seen as a symbolic act of political independence from the Mainland. In any case, we should not be surprised if a Chinese version of the MAR will force neighboring countries to follow either the European or Chinese regime, with those countries that choose the "wrong" regime effectively putting themselves at risk of economic and political isolation.

3.5 GDPR as Increasing Consumer Trust

Another key justification for the 2016 GDPR was concern about the level of consumer trust in new technologies (e.g., the Internet of Things and Big Data), new industries (e.g., data brokers, social media), and new business models (e.g., Google and Facebook's increasing reliance on behavioral advertising for most of their revenues) that had emerged since the 1995 Data Protection Directive. In an effort to safeguard Europeans' rights to privacy and personal data protection, the 1995 Directive needed a makeover, especially as the Directive seemed increasingly out of touch with today's technology and applications of Big Data. In 2010, a European Parliament Commission report highlighted two key issues that would later steer the development of the GDPR.[10] First, the Commission asserted that "rapid technological developments and globalisation have profoundly changed the world around us;" and second, that "ways of collecting personal data have become increasingly elaborated and less easily detectable." The need for increased transparency of data subjects' personal data as the result of new technologies would become the *raison d'etre* of the GDPR and the later-updated e-Privacy Directive.

[8] www.taiwannews.com.tw/en/news/3655633

[9] See GDPR Article 46 for more information about international transfers of personal data.

[10] See Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Comprehensive Approach on Personal Data Protection in the European Union, at 7, COM (2010) 609 final (Nov. 4, 2010)

There is at least some incidental evidence that the GDPR may increase consumer trust. In the US, the FTC's Report on Big Data, for example, concludes that federal data privacy regulations would serve to boost consumer trust of data brokers (Federal Trade Commission, 2012). It is interesting to note that in the US, at least, some corporations are arguing in favor of federal regulation of personal data processing on the grounds that it would ultimately benefit business and make consumers more likely to give up their personal data.[11][12] And in light of the many recent data breach scandals by Facebook and Google+ (and subsequent fines by European data protection authorities), the argument that the GDPR could increase consumer trust in the "Big Four" may be getting stronger.[13] The idea is that major fines introduced in regulations such as the GDPR will make these Behavioral Big Data-based firms more accountable to the public will and less likely to make small, undetectable changes to their labyrinthine privacy policies over time.

3.6 GDPR as Slowing "Privacy lurch"

Finally, a major issue surrounding the adoption of emerging Internet technologies, such as the Internet of Things and Online Behavioral Advertising (OBA), is the creation of a new paradigm of social trust (Tene and Polonetsky, 2013). In their discussion of new and "creepy" uses of technology, Tene and Polonetsky (2013) write, "The techno-social ground is shifting, setting a complex interplay between what we can learn about each other and what we (or business or government) should know or be restricted from knowing." In other words, without a conscious formulation of a new social contract between society and technology, modern citizen-consumers are essentially the victims of a slow but steady privacy lurch, a phrase coined by the legal scholar Paul Ohm, which characterizes "an abrupt change a company makes to the way it handles data about individuals (Tene and Polonetsky, 2013)."

According to Ohm (2012), open debate of these "privacy lurches" is vital to discussions about the social implications of new technology because they "deprive their users the free choice" in deciding whether a new service is worth the privacy risk. The situation is exacerbated by users feeling "locked in" to using certain platforms because of the time and effort they have invested in building networks of friends and contacts on the platform (Ohm, 2012). Not surprisingly, Tene and Polonetsky (2013) pessimistically conclude that "The European approach—trying to establish a social norm by regulatory fiat—may not fare well in the real world." The authors instead suggest basing regulation, whether in the form of industry self-regulation or governmental legislation, on a bottom-up, inductive understanding of the social norms surrounding the use of new technologies, rather than the top-down, deductive approach of the GDPR (ibid.).[14]

The GDPR's strict processing guidelines and security standards are therefore especially important due to the potential for real-time behavioral monitoring via the Internet of Things (IoT), particularly in the healthcare and consumer goods industries. The Dutch philosopher of technology Van Den Hoven (2012), makes a similar argument when he notes that trust and privacy are key ingredients for any new technology. In the case of IoT, explicit care must be taken by designers and engineers to ensure high levels of security and personal data protection, otherwise such technology may never be trusted and used by consumers.

Extrapolating from Tene & Polonetsky's examples of creepy uses of technology, one might view the GDPR as the result of a society—in this case, the 28 member states of the European Union—drawing a line in the sand between technological innovation, business profits, and the public good. Jaroen Lanier's book *You Are Not a Gadget: A Manifesto* makes a similar argument regarding the role of technology in society. There, Lanier makes the case that "technology should be designed to serve humans and reflect their values, not the other way around (Tene and Polonetsky (2013) citing Lanier (2010))." Spurred by such claims, many citizens and politicians in the US are now pushing

[11] See Google's proposed Framework for Responsible Data Protection Regulation services.google.com/fh/files/blogs/google_framework_responsible_data_protection_regulation.pdf

[12] https://www.nytimes.com/2018/08/26/technology/tech-industry-federal-privacy-law.html

[13] In 2018, Facebook was fined £500,000 by the UK Information Commission Office (ICO) as a result of the Cambridge Analytica scandal. /www.theguardian.com/technology/2018/jul/11/facebook-fined-for-data-breaches-in-cambridge-analytica-scandal

[14] See the details of the 2018 California Consumer Privacy Act (CCPA): www.caprivacy.org/

for federal legislation similar to the GDPR. At time of this writing, however, the federal government in the US has refused to do so, even though some individual states have passed legislation similar to the GDPR.

Regardless of whether one agrees with the top-down approach of GDPR favored by Europe, one must admire the chutzpah of the drafters of the GDPR. The GDPR is the first attempt by a major political and economic entity to delineate the appropriate uses of technology in society. Even more, the Regulation proposes rough and ready legal tests to determine whether specific uses of personal data processing have benefits that outweigh potential privacy harms. For example, Recital 4 of the GDPR states very clearly, "The processing of personal data should be designed to serve mankind." Further, the GDPR mandates that the *principle of proportionality* should be used to judge when the rights to protection of personal data may properly trumped by the benefits of such processing to society (Article 4). Or course, operationalizing the GDPR's rather abstract principles through judicial rulings will be a major focus point for the next several years and will likely determine whether the GDPR will ultimately be judged as a success or failure by its proponents and critics.

3.7 General Criticisms of European Data Regulation

This section looks at two critiques of the European approach to personal data regulation just prior to the GDPR. After a brief exposition of the key points, we discuss the extent to which the GDPR could address some of these shortcomings in previous European approaches to personal data regulation. Chapter 4 will pick up where this section ends, when we review direct responses to the GDPR and.

3.7.1 Will the GDPR Help Create a Global "Splinternet?"

Kulesza (2011) predicts that (inter)national privacy regulations such as the GDPR may lead to a gloomy future of "loosely intertwined, firewall-guarded national areas in cyberspace where privacy would be secured according to varying national standards." She warns that the inclusion of adequacy decisions in the GDPR (which concern the legality of international personal data transfers to countries deemed by the European Commission to have "adequate" protections) could create a global "splinternet" where countries limit citizens' access to global networks (Kulesza, 2011). Most notably, China has its "Great Firewall," which blocks content and filters search results, not to mention the legions of government-employed "monitors" that remove social media content deemed by the Communist Party as antithetical to "social trust." Kulesza (2011) worries that without a global privacy standard, the EU and China may create their own splintered versions of the Internet based on competing political and social norms.

Discussion of Kulesza (2011)

Already in the wake of the GDPR we are seeing some major US-based news media publishers, such as the LA Times and the Chicago Tribune, block EU-based users from reaching their websites. Though Kulesza writes about government-led actions, her vision of a "splinternet" seems prescient as companies appear to be self-restricting free access to global information flows out of fear of GDPR non-compliance.

To this end, the concept of Binding Corporate Rules (BCRs)–first conceived of in a 2011 speech by EU Justice Commissioner Reding, and later implemented in the 2018 GDPR–might be the only feasible antidote against an impending splinternet. BCRs can be viewed as "sets of good business practice guidelines adopted by companies voluntarily and applied throughout their branches, regardless of where the branches are located (Reding, 2011)." We will discuss BCRs in much more detail in section 5.3.5, which discusses international data transfers under the GDPR.

3.7.2 The EU Legal-Regulatory Framework is Relatively Ineffective

Determann (2016) compares the legal and regulatory environments for data protection in the EU and the US. His assessment of the European legal and regulatory framework is not positive. He cites myriad examples of how the European approach is inferior to the US one: in the EU's 45-year history of data protection law, actual enforcement has been rare; EU regulations are too general (not sector specific, as in the US); EU data protection laws are not applied by authorities or tested in courts quickly; EU data protection is overly governed by bureaucratic agencies (the US tends to rely on civil procedure law and class actions); and finally, the EU's insistence on prohibiting automated processing and restricting international data transfers have negatively affected "freedom of information, innovation, and commerce" (Determann, 2016). Overall, Determann's views stand in sharp contrast to the glowing optimism expressed by European lawmakers and GDPR proponents.

Discussion of Determann (2016)

Determann would likely be surprised to find that many of his general criticisms of previous European data protection laws do not apply to the GDPR. Certainly, in some cases they do: the GDPR is, by its very name, a very generic law without specific sectoral guidelines, though there are exceptions for scientific research (see Recital 159, for example). He is also correct that the EU's approach is more bureaucratic than the US's, especially with the GDPR's creation of the European Data Protection Board (EDPB). The Board, consisting of Member State representatives, provides independent oversight and replaces the previous Working Party advisory council (Recital 139). Nevertheless, some of Determann (2016)'s criticisms no longer apply. For example, it is now easier for data subjects to lodge complaints for GDPR data processing violations through non-profit, public interest groups. Even more, these same non-profit organizations, designated by individual Member States, can now petition on behalf of data subjects, independently of the data subject's mandate, if they believe a data subject's rights were violated during data processing (Article 80). In short, under the GDPR, European data subjects have new forms of legal redress similar to US-style class action lawsuits, which will help keep data controllers legally accountable for any violations in processing.

Ironically, the very things Determann (2016) praises about the US data privacy regime are likely to become more European. For example, he argues that the US's sectoral approach to privacy regulation is more flexible and responsive, yet there is significant support for federal data protection regulation in Congress, and California has recently passed the California Consumer Privacy Act (CCPA). The CCPA is aimed primarily at data brokers and BBD-collecting companies, applies to any company–anywhere in the world–that processes Californians' personal data, gives data subjects more control over who collects and processes their personal data, and also requires companies to implement appropriate data security techniques.[15] There is a wave of similar state-level legislation inspired by the GDPR gradually sweeping across the US. Were the US to pass a federal data privacy law similar to the GDPR, it could significantly improve the flow of international data transfers between the EU and the US. Currently, EU-US data transfers are governed by an agreement called Privacy Shield: an agreement whose existence was based on the European Court of Justice finding that the US's protection of Europeans' personal data was inadequate. A more detailed explanation and history of this arrangement will come in Part II.

[15] www.caprivacy.org/

Chapter 4

Responses to the GDPR

Introduction

This chapter lays out several key academic reactions to and critiques of the 2018 implementation of the GDPR. These responses are categorized by their relevance to the political, legal, business, economic, and technical effects of the GDPR. For ease of presentation, I have collapsed the legal, political, and business implications into one section. It should be noted that these categories are neither mutually exclusive nor exhaustive: some responses to the GDPR cover a wide range of issues and could be justifiably placed in several categories. Tables 4.1-4.5 summarize the major currents of responses to the GDPR in the academic literature and also highlight common themes found in GDPR-related blog posts and news articles. For articles that are particularly representative of a strain of thought, or represent novel ideas, a discussion of relevant GDPR concepts and principles follows.

[1] www.theguardian.com/technology/2018/may/25/gdpr-us-based-news-websites-eu-internet-users-la-times

[2] newsroom.accenture.com/news/six-in-ten-consumers-willing-to-share-significant-personal-data-with-banks-and-insurers- in-exchange-for-lower-pricing-accenture-study-finds.htm

[3] hbr.org/2019/04/dont-acquire-a-company-until-you-evaluate-its-data-security

[4] /martechtoday.com/gdpr-introduces-new-job-position-data-protection-officer-211269

[5] iapp.org/news/a/study-gdprs-global-reach-to-require-at-least-75000-dpos-worldwide/

Table 4.1: **Economic** Implications of the GDPR by source
*denotes a source that does not directly address the GDPR

Academic literature
Improved data subject rights and the importance of consent may lead to the creation of data-backed securities markets (Allen et al., 2019)
Price discrimination using personal data will continue, though it will be curbed by the GDPR's necessity tests and data subjects' rights (Steppe, 2017)
More firms will want to hedge risk against GDPR fines and data breaches through insurance, but GDPR insurance is legal in only a few EEA member states (DLA Piper & AON, 2018)
Short-term estimated economic effects of GDPR are reduced investment in technology ventures, measured in the number of deals, total dollar amount, and dollar amount per deal, along with the loss of around 3,000-29,000 technology venture jobs (Jia et al., 2018)
Firms with greater stores of BBD can produce better predictive models, thus widening the "data divide"(Martens et al., 2016)*
Blogs and News
Rather than comply with GDPR, some US-based companies are simply withdrawing from the EU market or blocking European users[1]
According to a survey done by Accenture, post-GDPR only 40% of data subjects in the UK and Germany would be willing to exchange more personal data for personalized services and offers from financial firms; while 67% of Chinese and 50% of US data subjects would be willing to do so[2]
During the M&A process, firms must vet the personal data collection and processing practices of the target firm or else could be liable for the GDPR fines of "data lemons"[3]
Analysts forecast the GDPR may spur the creation of nearly 75,000 new Data Protection Officer jobs; nearly 9,000 in the US alone[4][5]

4.1 Economic and Business Implications

4.1.1 Creation of Data-Backed Securities Markets

Immediately after the 1995 Data Protection Directive was updated in 2018 to the General Data Protection Regulation (GDPR), many–especially smaller–firms scrambled to reevaluate their consent mechanisms to get explicit opt-in consent from their customers. This was largely due to the commonly-held misconception that consent was the only basis on which personal data could be processed under the GDPR. In light of this potential loss of personal data, Allen et al. (2019), argue that the GDPR could create a kind of "options market" for personal data that have not yet been de-consented to. The value of such a GDPR option depends on whether data subject chooses not to exercise his or her right to opt-out of processing.

Because there is always some risk that a data subject will opt-out of processing, data controllers may be indirectly incentivized to create a secondary market to hedge the risk that data subjects may exercise their "zero strike price call" data option (i.e., remove consent for processing) (Allen et al., 2019). Since holding personal data on data subjects is often a form of revenue–Google and Facebook, for example, make the vast majority of their revenues through targeted advertisements–the ability to opt-out at any moment adds a level of uncertainty into the value of these personal data, and thus data controllers will want to hedge against this uncertainty by selling other, related financial instruments. The end result could be the creation of financial derivatives ranging from simple insurance contracts that will "pay out in the event [that] personal data–and the options over said data [have] 0 or negative value," to more complex instruments such as "data backed securities" and "Collateralized Data Obligations," which divide up packages of personal data into 'tranches' of varying degrees of likelihood of losing consent (Allen et al., 2019). If these speculations prove to be accurate, then the GDPR may prove to be an unlikely boon to the financial services industry.

4.1.2 GDPR Insurance

To date there has been little to no published academic research on the implications of the GDPR on the insurance industry, although some blog posts and news reports have dealt with the issue. According to these sources, the GDPR may indirectly benefit the insurance industry through the introduction of massive fines (see, e.g., Article 83). A recent example of the regulatory impact of GDPR on business is the €50 million fine of Google, mainly due to a lack of transparency of processing and valid consent mechanisms. Because of the potentially grave economic consequences, many companies are looking to insure themselves against future violations. According to the law firm DLA Piper, in only two European Economic Area (EEA) countries, Finland and Norway, can GDPR fines legally be insured against. DLA Piper also concluded that in 20 out of the 30 countries reviewed, GDPR fines could not be insured against; and finally, in eight countries, it was unclear (DLA Piper & AON, 2018). In these latter cases, whether fines could be insured against would depend on the conduct of the specific company and whether the fine was considered "criminal" (ibid.).

Discussion

We can expect that at least in some countries, a market for GDPR insurance may arise. GDPR insurance may then become part and parcel of firms' analytics strategies in jurisdictions where such insurance can be legally obtained. At the same time, firms should remember that insurance cannot protect them against the reputational damage in the eyes of consumers and regulators if there is a data breach or misuse of personal data. GDPR insurance is also likely to make economic sense only for companies processing the personal data of a large number of data subjects (e.g., Facebook, Amazon, etc.); for most small to medium-size businesses, it is most likely unnecessary, especially given how the EU Data Protection Authorities have been fairly lenient on smaller firms.

But regardless whether a firm chooses to buy GDPR insurance, general compliance does not come cheap: it means training employees, rewriting privacy policies, overhauling IT systems, employing a Data Protection Officer (DPO), seeking legal advice, and training staff, among many other things. Unsurprisingly, firms with more cash and resources could better prepare themselves during

the two-year period from 2016-2018, when the GDPR went from draft to European law. Small and medium-sized companies struggled to comply in the face of the legal and technical uncertainties of the GDPR.[6] For example, many US-based companies were unsure whether the GDPR applied to them if they only tangentially targeted EU-based customers online.[7]

4.1.3 Deepening Data Divides among Firms

One consequence of this GDPR-driven "data-divide' is that large companies can continue to reap the benefits of BBD collection and processing for behavioral targeting and optimizing their services and products, while smaller ones face the potential loss of important customer data. Losing the ability to process customers' personal data could affect firms' bottom line, especially when behavioral targeting and analytics is a core strategy of the firm. Martens et al. (2016) illustrate that in a banking context, access to vast amounts of behavioral data can lead to considerable improvements in a model's predictive performance. The authors conclude that "large institutions have an important asset in the data they have collected, an asset from which they can get substantial competitive advantage over institutions without as much data—in this example, smaller banks" (Martens et al., 2016).

A similar power dynamic between large and small companies exists in the world of academic-industry collaboration, where the amount of personal data being collected and processed continues to skyrocket, concentrating a larger and larger proportion of these data in the hands of just a few corporations (King and Persily, 2018). Social scientists wishing to access these vast stores of personal and behavioral data will increasingly need to negotiate with companies like Facebook, Google, Twitter, Amazon, and Apple. Meanwhile, smaller companies–and society in general–lose out on the positive effects of academic-industry collaboration. If this pattern continues, then we can expect the data-divide between large and small companies under GDPR to deepen in a kind of vicious cycle, since the benefits of academic collaboration will only serve to further the ability of the large corporations to collect and process more personal data and attract more academic interest.

One solution may be to treat a company's vast BBD reserves as a public resource. According to some proponents of this view, making the data public removes legal issues surrounding processing and leads to better scientific research that relies less on personal connections to corporate networks and more on scientific merit. Of course in doing so, under GDPR the data would need to be appropriately anonymized so that the identities of the individual data subjects could not be inferred.

[6] www.bloomberg.com/opinion/articles/2018-11-14/facebook-and-google-aren-t-hurt-by-gdpr-but-smaller-firms-are

[7] www.forbes.com/sites/forbestechcouncil/2018/04/25/gdpr-and-what-it-means-for-your-business/#18cbd04c2d2f

[8] www.nytimes.com/2018/05/24/technology/europe-gdpr-privacy.htmls

[9] europa.eu/rapid/press-release_IP-19-421_en.htm

[10] qz.com/1597901/how-people-feel-about-algorithms-has-become-the-new-digital-divide/

[11] www.lawfareblog.com/road-adequacy-can-california-apply-under-gdpr

[12] www.nytimes.com/2018/05/24/technology/europe-gdpr-privacy.htmls

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

[14] /www.emarketer.com/content/how-the-gdpr-helps-consent-management-platforms

[15] www.bloomberg.com/opinion/articles/2018-11-14/facebook-and-google-aren-t-hurt-by-gdpr-but-smaller-firms-are

[16] digiday.com/media/personalization-diminished-gdpr-era-contextual-targeting-making-comeback/

[17] www.bloomberg.com/news/articles/2019-02-06/oracle-didn-t-see-the-data-reckoning-coming

[18] fortune.com/2018/06/16/gdpr-email-marketing-unsubscribe/

[19] www.theguardian.com/technology/2018/may/21/gdpr-emails-mostly-unnecessary-and-in-some-cases-illegal-say-experts

[21] discuss.okfn.org/t/gdpr-vs-part-of-open-data/6254/8

[22] www.channelnewsasia.com/news/commentary/china -great-leap-forward-in-data-protection-11429624

[23] www.theverge.com/2019/4/9/18302199/big-tech-dark-patterns-senate-bill-detour-act-facebook-google-amazon-twitter

[24] https://www.politico.com/story/2019/04/24/ireland-data-privacy-1270123

Table 4.2: **Political** implications of the GDPR by source
*denotes a source that does not directly address the GDPR

Academic literature
Brussel's Effect: GDPR "soft power" spreads privacy to world (Albrecht, 2016; Zarsky, 2016)
Member states will succeed in creating derogations to allow for innovative use of Big Data analytics under GDPR (Zarsky, 2016)
The current "datafication" of society is similar to the conflict between industrialization and environmental protection law. The GDPR may be too little, too late to stop the negative externalities (Rhoen, 2017)
The GDPR may not be easily exportable to countries in the African Union due to cultural differences surrounding privacy (Georgiadou et al., 2019)
Political party affiliation plays a major role in whether one reacts positively or negatively to online profiling and surveillance. Republicans consistently respond with "warmer emotional responses toward a variety of surveillance practices" than Democrats, especially when those surveilled are "low-income Americans," namely African Americans and Hispanics (Turow et al., 2018)*
Which humans are in the loop? We need to re-evaluate how AI is used in the classification, detection, and prediction of race and gender, particularly in facial recognition technologies and automated profiling (West et al., 2019)*
Blogs and News
Other non-EU countries begin to follow the example set by the GDPR (Brazil and South Korea, among several others) [8]
In early 2019 Japan achieved adequate personal data protection according to the European Commission, creating the "world's largest area of safe data flows"[9]
Exercising one's right to opt-out of automated profiling may become a symbol of affluence; low-income data subjects could end up on the wrong side of the "digital divide "[10]
California, in light of its Consumer Protection Act (CCPA), may try to apply for an adequacy decision from the European Commission. If so, personal data could be transferred from the EU to California without any special review or limitations[11]

Table 4.3: **Business** implications of the GDPR by source

Academic literature
GDPR spurs business innovation with new data storage and processing techniques based on dynamic de-identification (see e.g., Hintze and LaFever (2017))
There is a diversity crisis in the AI sector across gender and race, with mostly white men dominating industry & academia (West et al., 2019)
Populist pushback to EU leads to businesses and data processing operations moving to less restrictive jurisdictions (Zarsky, 2016)
Companies (i.e., data controllers) must be able to document their compliance with GDPR, in accordance with the GDPR's *Accountability principle*. This requires organizational changes, particularly in IT (Korff, 2016)
Companies should plan for GDPR by implementing 12 key changes. These include: specifying data needs and usage; considering international data processing; integrating privacy by design and default into products/services; demonstrating GDPR compliance and documentation; developing data breach procedures; planning for possible non-compliance; designating a Data Protection Officer (DPO) if necessary; providing information to data subjects; obtaining data subjects' consent for processing; and ensuring the right to be forgotten and to data portability (Tikkinen-Piri et al., 2018)
Blogs and News
Other non-EU countries begin to follow the example set by the GDPR (Brazil and South Korea, among several others) [12]
GDPR leads to decrease in advertisement trackers and third-party cookies [13]
After GDPR, more businesses are using Consent Management Platforms (CMPs) to manage data subjects' consent status and other information [14]
GDPR hurts smaller firms hurt more because they cannot afford compliance & training [15]
Contextual instead of behavioral targeting is likely to become more common in marketing and advertising[16]
Data controllers are liable for all downstream processing (i.e., Accountability Principle), leading to a reduction in third- party data due to fears of illegally-obtained data [17]
Asking customers to re-consent to processing leads to a reduction in the number of customers (particularly email lists in the USA) [18]
Smaller companies are sending (often) unnecessary emails asking to renew consent to pre-GDPR data [19]

Table 4.4: **Legal** implications of the GDPR by source
*denotes a source that does not directly address the GDPR

Academic literature
GDPR may set a global standard bringing better privacy to Americans, but the right to be forgotten may not be practical or enforceable in business settings (Safari, 2016; Buttarelli, 2016)
The GDPR endorses Data Protection by Design (DPbD). The transparency and accountability principles allow for a viable trade-off between personal control and confidentiality, yet data protection impact assessments (DPIAs) do not because a lack of publishing requirements (Veale et al., 2018a)
Pseudonymized data are considered personal data. The key determination in the definition of personal data is whether there exist "means reasonably likely to be used to identify natural persons." The definitions of anonymized and pseudonymized data are fluid and depend on the "data situation" (Mourby et al., 2018a)
The most relevant legal bases of personal data processing for behavioral targeting are: necessity for performance of a contract; necessity for the data controller's legitimate interests; the data subject's "unambiguous consent"; Companies should only rely on "unambiguous consent" as the legal basis for behavioral targeting (Borgesius, 2015)*
Data privacy concerns lead to creation of new governing bodies for industry-academic collaborations, particularly in the social sciences (King and Persily, 2018)*
Based on the seven guiding principles of the US Public Policy Council of the Association for Computing Machinery (ACM), Hosanagar (2019) proposes an algorithmic bill of rights
Blogs and News
GDPR will reopen the debate surrounding around "public" data, especially for "Open Data" initiatives [20]
The US is considering the Algorithmic Accountability Act, which is similar in intent to GDPR's Article 22 concerning Automated Profiling [21]
Public outcry in China over privacy prompts Chinese Cybersecurity Law (CSL) revisions similar to GDPR's Article 30 and also mandates opt-out of automated profiling[22]
US Congress proposed DETOUR (Deceptive Experiences To Online Users Reduction) bill to reduce "dark patterns" that reduce informed consent[23]
Ireland's Data Protection Authority (DPA) may not enforce GDPR because of the negative impact on tax receipts from major tech companies based there. Ireland could become the weakest link in GDPR enforcement in Europe, creating a "safe zone" for companies to do business in the EU, defeating the purpose of GDPR [24]

4.2 Political and Legal Implications

4.2.1 The GDPR & Environmental Protection Law

Rhoen (2017) compares the implementation of the GDPR to that of environmental protection laws. He grounds his comparison on the premise that both industrialization and "datafication" arose from technological progress and pose risks to society. In the case of environmental regulation, two social science theories influenced environmental policy: Ulrich Beck's theory of the risk society and Charles Perrow's theory of normal accidents (Rhoen, 2017). In a nutshell, Beck's theory states that technological and scientific progress create new and unpredictable risks that are borne out in different ways by various groups in society. Two important corollaries are that 1) science must have some kind of governance system that allows for control of these risks; and 2) individuals should have some say in which risks created by technological and scientific progress are acceptable (Rhoen, 2017). Meanwhile, Perrow's theory derives from a view of complex systems that categorizes them on two main axes: whether their interactions are linear or complex, and whether they are loosely or tightly coupled (Rhoen, 2017). Systems that have both complex interactions and tightly coupled components can generate erratic behavior that can quickly lead to catastrophic consequences, such as in nuclear reactor meltdowns. Most societies and governments have therefore worked to regulate these systems in order to make them more linear and less highly-coupled, so that catastrophic events are less likely to occur (Rhoen, 2017). Both theories support government regulation as playing a key role in a fair and just distribution of risk in society.

Discussion of Rhoen (2017)

Regarding the application of these social theories to the GDPR, Rhoen (2017)'s argument is essentially that the GDPR is too little, too late. The GDPR focuses too much on big data analytics' impact on the individual at the expense of its broader social impact (Rhoen, 2017). Rhoen (2017) concludes by stating that the GDPR may be an example of the Collingridge Dilemma, in which regulators realize that technology needs to be regulated only after its use has become so entrenched in society that effective regulation becomes difficult or impossible.

While Rhoen (2017) critique of the GDPR provides interesting parallels from the area of environmental protection law, it overlooks that the GDPR does in fact call attention to the social impact of big data analytics. For example, in the GDPR's requirement of data protection impact assessments during large-scale surveillance (Article 35) or its dictum that the "processing of personal data should be to serve mankind (GDPR Recital 4)." While it is true that much of the GDPR deals with the impact of analytics on the personal rights and freedoms of data subjects, the GDPR provides society-wide principles, such as the principle of proportionality, which states that the processing of personal data is not an "absolute right" and must be balanced against its social costs (GDPR Recital 4). For these reasons, I do not share Rhoen's pessimism that the EU's regulation of personal data processing is too little, too late.

4.2.2 Zarsky's Three Prognoses

Zarsky (2016) lays out perhaps the widest-ranging critique of the GDPR. He ends by listing three prognoses of varying optimism for the GDPR with political and business implications.

Best Case: The "Brussels Effect" sets a Global Standard for Businesses

In this scenario, Zarsky argues that the GDPR will "march the citizens of Europe [to new] forms of data analytics [that] will dominate the markets" and cites the so-called "Brussels Effect," by which the stringent standards of the GDPR will spur a "race to the top" in other countries who wish to access markets with European data subjects. Similarly, this effect could incentivize US firms to apply GDPR standards globally. There is some truth to this as Microsoft has claimed it will do so, however it is less clear that Facebook will apply GDPR requirements to non-EU-based data

subjects.[25]

Pragmatic View: It Depends on the Member State Derogations

In a more pragmatic middle-view, Zarsky imagines the GDPR will provide "sufficient forms of exceptions and loopholes to allow rich big data dynamics to nonetheless unfold." In this scenario, the ability of GDPR to foment technological innovation and spur business will heavily depend on the extent to which individual member states can carve out derogations and exceptions to the Regulation, particularly in the areas of automated profiling and various forms of research.

Pessimistic View: Populist Pushback from Businesses Leads to Balkanization

Given the recent trends of national separatism in the UK and the rise of populist anger towards far-away central governments, it would not be unreasonable for European bureaucrats to expect pushback from local entrepreneurs. Several commentators have also suggested that the passage of GDPR will reduce economic incentives to innovate and motivate European technology companies to take their efforts elsewhere (e.g., Determann (2016)), where they can use big data business models and analytics without problem. Second, the GDPR may not even protect privacy rights if EU firms end up outsourcing their analytics to international third countries, where GDPR enforcement is unlikely. If firms process their data in other countries, then we may end up with a competition among various national and international personal data processing frameworks. We will explore this potential situation in more detail in section 5.4.2 in Part II, where we examine the theoretical and legal implications of Binding Corporate Rules.

4.3 Technical Implications

Zarsky (2016) focuses on four key GDPR principles and forcefully asserts that each of the principles is 'incompatible' with the technical nature of big data analysis. The four GDPR principles that Zarsky believes are antithetical to big data analysis are purpose limitation, data minimization, special categories, and automated decisions. I will summarize Zarsky's criticisms of these major GDPR principles in the following paragraphs and conclude with a discussion of the key ideas and introduce related research on automated decision systems from the field of Human Computer Interaction (HCI).

4.3.1 Issues with the Purpose Limitation Principle

Zarsky claims that Big data analysis is conceptually opposed to both the idea of purpose limitation and the GDPR's requirement of "specific" processing purposes. Article 5(1) of the GDPR is centered on "compatibility tests" and that processing for "statistical purposes" would not be deemed incompatible assuming that certain "appropriate safeguards" are taken (which might include pseudonymization). Yet the GDPR in Recital 162 defines Statistical Purposes as those that, "[do not] result in decisions regarding any particular natural person." The tension, as Zarsky notes, is that Big data analysis is useful in business settings precisely because it allows companies to provide their customers with "unique and specific treatments," rather than giving them all a general "average" treatment–the type of result permitted by statistical purposes in the GDPR's sense of the phrase. Zarsky concludes that purpose specification and the disparate provenances of potentially-linkable datasets are at odds and that the current approach taken by the GDPR is "shaky, at best."

[25] www.reuters.com/article/us-facebook-privacy-eu-exclusive/exclusive-facebook-to-put-1-5-billion-users-out-of-reach-of-new-eu-privacy-law-idUSKBN1HQ00P

[26] www.linkedin.com/help/linkedin/answer/87079/linkedin-developer-resources-and-the-general-data-protection-regulation-gdpr-?lang=en

[27] techcrunch.com/2018/07/01/wtf-is-dark-pattern-design/

[28] www.developer.ibm.com/articles/s-gdpr3/

Table 4.5: **Technical** (data security and automated profiling) Implications of the GDPR by source
*denotes a source that does not directly address the GDPR

Academic literature
IoT presents many data protection and privacy risks that can be minimized using the GDPR's principle of privacy by design and default; however, the GDPR's standards need to be further clarified and integrated into future IoT designs (Wachter, 2018)
Counterfactual explanations of "black-box" automated decisions may be legally permissible explanations under GDPR (Wachter et al., 2018)
ML and NLP-based policy summarization software may be used to alert data subjects of risks to privacy in complicated privacy policies and service agreements (Tesfay et al., 2018)
The technical complexities of data subjects withdrawing consent and exercising the right to be forgotten will pose major challenges for companies; the authors survey potential solutions including: biometric authentication, digital aging systems, blockchain, Information Flow Control (IFC), Personal Data Management Architectures (PDMA), and Distributed Hash Tables (DHTs) (Politou et al., 2018)
When subjected to automated profiling, data subjects' right to receive human review of the algorithm is unclear and problematic (Veale et al., 2018b)
Incompatibility of Big Data analytics with GDPR principles of purpose limitation, data minimization, definitions of special categories, and automated decisions (Zarsky, 2016)
Blogs and News
Platform APIs restrict data access to enhance privacy and reduce possibility of data breaches and linking of datasets (e.g., Twitter removes time zones and background images; LinkedIn changes how profile picture URLs are stored using APIs)[26]
Incentives to develop 'dark pattern' websites, designed to nudge users into accepting the site's terms and conditions of personal data collection, storage, and processing[27]
Web application developers need to understand privacy preserving techniques such as anonymization, pseudonymization, hashing, masking, scrambling, and encryption[28]

4.3.2 Issues with the Data Minimization Principle

The principle of data minimization is spelled out in Article 5(1)(c) and states that data must be "limited to what is necessary in relation to the purposes for which they are processed." Also, Article 25 declares that this principle must be followed when designing IT systems. Further, the principle also refers to limited data storage duration periods and deletion after its intended use. Zarsky argues that the potential business benefits of big data analysis compels companies to collect as much personal data as possible (i.e., "data hoarding") and keep it for as long as possible. In this way, companies will be well-situated to take advantage of future advances in data science. In his view, then, the principle of data minimization and this natural business response to value of big data are fundamentally incompatible.

4.3.3 Issues with "Special Categories" of Personal Data

Zarsky details the special categories of personal data in Article 9, which need explicit consent from data subjects in order to be processed. These special categories were a carry-over from the 1995 Data Protection Directive. The GDPR essentially added genetic data and biometric data "for the purposes of identifying someone," and data "related to sexual orientation." Essentially, Zarsky's argument is that in era of big data (especially the Internet of Things) there is no clear line between special and normal categories of data because, for example, one's health status could be inferred from types of products in one's Amazon shopping history. He quotes Google's former CTO who once quipped that, "All data is credit data, we just don't know how to use it yet."

Additionally, Zarsky notes that the distinction between normal and special categories of personal data may be misguided in the age of algorithmic discrimination, because there is not necessarily "discriminator intent." In the era of Big data, discrimination often occurs without any one single individual consciously deciding to do so. Instead, it is driven by decisions in the way the data are collected and how the they were measured. He says that at the end of the day "there will be no real special treatment for these special categories as this stricter standard will be applied across the board." In other words, because machine learning algorithms can infer these private traits, the distinction between normal and special categories of personal data is so general as to be useless.

4.3.4 Issues with Automated Profiling

Article 22 of the GDPR addresses automated decision-making and profiling that have "substantial" impact on individuals. Profiling and automated decision-making should be understood essentially as synonymous with "data mining" or "predictive analytics." Specifically, Article 4(4) defines *Profiling* as:

> any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

Besides allowing data subjects to opt-out of automated processes, Article 22 has two key provisions: 1) data subjects have a right to obtain human intervention; 2) data subjects can contest the automated decision (and can also access the personal data used to make the decision) (Zarsky, 2016). Additionally, data subjects must be "informed" that such automated-processes are occurring and provided with "meaningful information" about the logic of the decisions and the possible consequences of such automated-profiling (see, e.g., Recital 71). Such "meaningful information" also includes the ability to "obtain an explanation of the decision reached (Recital 71)."

Zarsky further objects that the protections outlined in Article 22 can be easily sidestepped by inserting a human into the processing "loop," since if a human is involved in decision-making, then the process will cease to be "solely" automated. Once the decision-making process is not fully-automated, then the Article 22 provisions would not apply. Zarsky (2016) also points out that recent rulings by German courts protecting firms' trade secrets may make it difficult or impossible for data subjects to get meaningful information regarding automated decisions.

Without knowing that such profiling–even with a human "in the loop"–is occurring, and without understanding how the profiling was done, Zarsky (2016) worries that data subjects' rights to due process may be undermined. Due process is a foundational principle of modern legal systems guaranteeing that procedures of the law are fairly applied to individuals. In the case of automated profiling, due process means notification that one is being profiled and the existence of some procedure through which one can contest the results of the profiling. Without knowing how the system's decisions are made, data subjects cannot know when the system is discriminating unfairly or is based on faulty data or inferences (Zarsky, 2016).

Finally, Zarsky concludes his criticism by citing several examples of how Article 22 is incompatible with Big Data Analysis. Most fundamentally, he notes that prohibiting automated profiling effectively prohibits data mining, which is a core activity in the modern data-driven company. Secondly, he argues that interpretability requirements of algorithms could "compromise" their accuracy; thirdly, he holds that human intervention into automated processes will only slow them down and hinder innovation (Zarsky, 2016).

4.3.5 Algorithmic Explanation, Bias, and Transparency

In addition to Zarsky's critique of the GDPR's stance on automated profiling, Wachter et al. (2018) present an implicit critique of the GDPR's right to explanation. Their contribution is based on the notion of providing data subjects, data authorities, and other stakeholders with counterfactual explanations of algorithmic decisions. Besides making the classifications of "black box" neural networks more transparent, "counterfactual explanations do not attempt to clarify how decisions are made internally. Instead, they provide insight into which external facts could be different in order to arrive at a desired outcome (Wachter et al., 2018)." In other words, counterfactual explanations of algorithmic results start at the result of interest and then work backwards to see whether various combinations of pertinent facts of the data subject would change that result. So for example, if a data subject wished to know what might allow her to be accepted for a loan, data scientists would change various categories related to her (e.g., race, income level, education, etc.) and see how these counterfactual scenarios change the outcome. Changes in categories that lead to changes in the outcome are deemed important and can help data subjects and regulators to understand how the algorithm reaches its conclusions.

4.3.5.1 Bias in Automated Decisions

It is no secret that human decision-makers are fallible and subject to a number of established psychological biases (see e.g., Tversky and Kahneman (1974)). Such fallibility is especially acute when the consequences of the decisions can mean years in prison, or the loss of a job. A study by Danziger et al. (2011), for example, showed that experienced judges' became increasingly unfavorable to parolees the longer the judges went without a short meal break. On the face of it then, removing the human component from such decision-making would seem to be a positive thing. Computers do not need lunch breaks, after all. Nevertheless, the framers of the GDPR were concerned that algorithmic profiling could lead to the codification of human biases. For instance, a company might use a "biased" algorithm to screen potential job applicants. The algorithm may be biased in the sense that it unfairly bases its decisions on applicants' race or gender.[29] A similar worry is based on the thankfully-now-illegal practice of insurance and mortgage lending companies *redlining* certain locations of cities as "uninsurable.[30]"

4.3.5.2 Induction and Input Data

Two major contributors to this problem are biased input data and the lack of transparency in how these algorithms are making predictions. The input data may be biased if they do not

[29] This actually happened in the case of Amazon's recruiting tool that crawled the Web looking for promising job candidates. Eventually, Amazon stopped using the AI system when it discovered that female candidates were systematically being "downgraded" for technical roles: phys.org/news/2018-11-amazon-sexist-hiring-algorithm-human.html

[30] www.wired.com/story/ideas-joi-ito-insurance-algorithms/

contain sufficient variation or are unrepresentative of the underlying population, i.e., if a company has a history of hiring mostly male employees, the system cannot adequately learn from the few female examples. Such biases are typical of inductive learning (i.e., from particular cases to general, probabilistic rules), which is the logic undergirding most current machine learning algorithms. The philosopher Bertrand Russell explains that inductive learning rests on finding, over a great many cases, a constant connection between two or more things or properties, such as encountering hundreds of white swans and formulating the rule that therefore all swans are white (Russell, 2001). A great many of the things we hold true in life are based on this kind of inductive reasoning from observed experience.

The problem with such an inductive generalization, however, is that it does not prepare us for situations in which we might encounter a rare black swan. The strength of our inductive generalizations is a function of the number of past cases where we observe a connection between events, assuming the observed cases represent a representative sample of the phenomena in question. It turns out that, at least for swans, color is not a defining characteristic of swan-ness, though it may be a useful predictive heuristic. Similarly, in the hiring example above, it may be the case that all previous good employees were male, but it was not their male-ness which caused them to be good employees. Their male-ness was merely an accidental feature that all happened to possess. Generalizations based solely on inductive learning are limited by the quality and representativeness of the data one has. Presently, machine learning algorithms cannot judge whether their training data are adequately representative–that is the job of the human data scientist. We will resume discussion of this problem when we examine the GDPR's stipulation of a right to human intervention in cases of algorithmic profiling.

The real issue is such biases could quickly lead to discriminatory behavior. Once inductive biases become entrenched in a system's predictions–and ultimately the human decisions based on those predictions–the bias perpetuates itself in a vicious cycle as fewer and fewer female candidates become new training observations from which the algorithm can learn. In the case of redlining, once insurance companies stop lending to certain communities, those communities cannot begin to rebuild. And once urban decay has set in, insurance companies become even less likely to insure the community, thereby intensifying the decay and leading to even less investment. All this is exacerbated when the algorithm cannot give transparent explanations for its predictions, since the lack of transparency makes it more difficult for the users and maintainers of the algorithm to uncover such a bias. Although a deeper discussion of algorithmic bias is outside the scope of this work, a very readable account can be found in O'Neil (2016).

4.3.5.3 An Account of Algorithmic Bias

One of the first and most influential attempts at understanding algorithmic bias can be found in Friedman and Nissenbaum (1996). Friedman and Nissenbaum argue convincingly that bias is not exclusive to human decision-making and demonstrate three types of bias that can infiltrate computer systems (in GDPR-speak, *automated processing* or *automated profiling*): preexisting, technical, and emergent. *Preexisting bias* is reflected in the social institutions, practices and attitudes of the designers of the automated systems. *Technical bias* refers to hardware or software limitations software engineers typically encounter when building new systems, or when they converting continuous measurements into discrete categories. Finally, *emergent bias* occurs when the population of users qualitatively changes, or when social and cultural values relating to the system change (Friedman and Nissenbaum, 1996).

In their account, a "biased" computer system must satisfy two necessary and sufficient conditions. First, it must unfairly discriminate against some users. Second, it must do so systematically. In simpler words, such a system must "assign an undesirable outcome...on grounds that are unreasonable or inappropriate" and not merely as a kind of "random glitch (ibid.)." They go on to present several examples of computer decision systems that fulfill these two criteria and can lead to biased decisions, especially when paired with human decision makers. These ideas have recently become influential in the debate around algorithmic accountability (see, e.g., Shneiderman (2016)) and the GDPR's treatment of algorithmic profiling.

4.3.5.4 Issues with the Right to Human Intervention

But just because the GDPR allows data subjects to request human intervention does not mean bias on the part of the algorithms or the humans will go away. The "anchoring effect," in which humans will tend to gravitate towards decisions first made by decision-support systems, is particularly relevant to the GDPR context of automated profiling and the right to human intervention (if requested). The proviso that data subjects can request human intervention into the profiling process appears to stem from worries concerning biased training data and prejudiced "prior decision-makers," i.e., the machine learning engineers who designed the algorithms deployed on data subjects (see, for example, Barocas and Selbst (2016)). Regarding automated profiling and decision-making, Article 22 (1) of the GDPR states, "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her (italics added)." As mentioned earlier, Recital 71 gives data subjects a right to human intervention, a right to contest a system's decisions, and the right to an interpretable explanation of the system's decision. Veale et al. (2018a) suggest the drafters of the GDPR inserted the word "solely" due to fear of automation bias, whereby "humans in the loop" begin to rely on the algorithm's predictions "as a heuristic replacement for vigilant information seeking and processing ((Parasuraman and Manzey, 2010), citing Mosier & Skitka, 1996)." Zarsky (2016), in contrast, traces the idea of human intervention back to traditional legal notions of honor and respect: data subjects should have the "dignity of having a human decision-maker address his or her personal matter."

The discussion of automation bias naturally begs the question: what exactly is meaningful human input in automated decision making systems under the GDPR (Veale et al., 2018b)? Some scholars make the argument that when we expect automated profiling systems to outperform humans, they should be considered "solely automated (Veale and Edwards, 2018)." Defending this definition, however, is difficult since it implies that any human involvement in the automated decision process would *ipso facto* lead to degraded performance (Veale et al., 2018b). Yet, given the spectre of anchoring and automation bias on the part of human assistants using automated decision systems, how much difference would human intervention actually make? For now the European Commission's High-Level Expert Group on Artificial Intelligence has emphasized the need for AI to be "human-centered," driven by ethical concerns, and focused on fostering trust among its users (Koszegi, 2019). By providing data subjects with a right to human explanation and intervention, the GDPR's provisions will likely need to be court-tested and clarified as technology evolves. In any case, the issues broached by the GDPR's treatment of algorithmic profiling have spurred research on increasing the transparency of black-box algorithms.

Part II

Implications for Managers of Multinational Corporations

Overview of Part II

Data privacy regulations in the EU and the US have followed two different paths. The Safe Harbor program was therefore created to allow European personal data to be legally transferred to the US. Safe Harbor consisted of a set of seven privacy principles that enterprises voluntarily followed. Companies self-certified as following the principles. The path taken by the US's main enforcement agency, the Federal Trade Commission (FTC), favors self-regulation, varies according to industry sector, and can be described as "light touch." The European approach is "top-down," based on fundamental rights to privacy and protection of personal data. Further these rights were developed and influenced by several important Court of Justice of the European Union (CJEU) cases that eventually invalidated the Safe Harbor agreement: Schrems (2015), Digital Rights Ireland (2014), and Google Spain (2014).

The Schrems case confirmed that there was no meaningful protection in US law and thus the principles of Safe Harbor were invalidated. The Digital Rights Ireland case was important because it was the first time the CJEU invalidated an entire EU legal instrument due to its incompatibility with the EU Charter. Finally, the Google Spain case resulted in EU citizens having the right to delist links related to their personal data on search engines. The Safe Harbor agreement, which had been in effect since 2000, was suddenly struck down after Schrems and replaced by Privacy Shield in 2016.

Under Privacy Shield, self-certified companies can legally transfer personal data from the EU to the US even though the US is not considered a country with "adequate" personal data collection. There also exist other legal means of transfer: standard contractual clauses (also sometimes referred to as "model clauses") and BCRs. The former are are rigidly structured and do not adapt to data flows that grow in complexity over time. BCRs are voluntary rules created by corporations that are legally binding. They must include complaint procedures, a statement of the privacy principles, auditing and employee training mechanisms, the acceptance of liability for any rule breaches, and an explanation of how data subjects can access the rules. BCRs become legally valid once a lead Data Protection Authority (DPA) of a Member State certifies them; they are mutually recognized in over 21 Member States.

BCRs may be the safest and most efficient way for large multinationals to legally transfer personal data across borders. And if BCRs continue to grow in popularity, they may lead to an overall more consistent legal framework for international personal data transfers. Employees and customers may also appreciate the corporation's public commitment to personal data protection. However, BCRs also open up parent corporations to legal liability throughout the data processing chain. They can be expensive and time-consuming to implement relative to standard contractual clauses. There is also no guarantee that other countries will accept BCRs over public regulation or contractual tools as an instrument to regulate transborder transfers. Finally, some experts have suggested that BCRs should borrow the concepts of applicable law, jurisdiction, and enforcement from private international law (PIL) to more effectively allow for the transnational scope of the GDPR.

Chapter 5

International Personal Data Transfers under the GDPR

Introduction

This section explores the legal and historical background leading up to the GDPR, with a focus on multinational corporations with subsidiaries in Europe and the USA. First, we motivate the need for international personal data transfers. Next, we give a brief overview of personal data protection in the USA and contrast the American approach with the 1995 Data Protection Directive and its reincarnation as the GDPR. Then we examine past and current legal frameworks and means for personal data transfers. Finally, we evaluate the relative advantages and disadvantages of Binding Corporate Rules for personal data transfers under the GDPR through the *three-domain* framework advocated by Schwartz and Carroll (2003). This section may be particularly useful for data science managers and executives interested in developing a big-picture GDPR analytics strategy.

5.1 Personal Data Regulation in the US: The FTC's Role

In order to better understand how the GDPR will impact managers' analytics strategy, it will be useful to compare EU with US approaches to personal data regulation. These differences will be important later when we examine the Safe Harbor and Privacy Shield agreements, which provided legal means for European personal data to be transferred to and processed in the United States.

The de facto regulation of companies that collect and analyze data in the USA mainly falls under the purview of the Federal Trade Commission. As enforcer of the Federal Trade Commission Act of 1914, the FTC has powers to protect consumers and business against "unfair or deceptive practices in commerce.[1]" Although the FTC has no direct punitive powers, it reserves the right to seek enforcement of the FTC Act by district courts and thereby obtain civil penalties (i.e., fines) for violations that "cause or [are] likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition.[2]" The FTC will first make a determination whether it has "sufficient reason to believe" a firm is acting unfairly or deceptively towards consumers and then, with the assistance of a district court, either file a Civil Investigative Complaint (CID) or a subpoena demanding information from the company regarding the FTC's inquiry.[3]

Compared to the EU's GDPR, the FTC's regulatory influence and ability to economically disincentivize bad actors is relatively weak. In 2003, for instance, the major food companies Hershey's

[1] A Brief Overview of the Federal Trade Commission's Investigative and Law Enforcement Authority www.ftc.gov/about-ftc/what-we-do/enforcement-authority

[2] (15 U.S.C. Sec. 45(n)) www.ftc.gov/about-ftc/what-we-do/enforcement-authority

[3] (15 U.S.C. Sec. 45(n)). www.ftc.gov/about-ftc/what-we-do/enforcement-authority

and Mrs. Fields were found to violate the Children's Online Privacy Protection Act (COPPA) by obtaining children's personal information without first requiring parental consent. The companies were fined $85,000 and $100,000, respectively.[4] At that time, these penalties were the largest COPPA civil penalties ever recorded by the FTC. Considering that Hershey's 2003 sales revenues were $4.17B, this was a mere drop in the bucket.[5] More recently in 2016, the FTC targeted the Taiwanese company ASUS when it claimed that ASUS used deceptive advertising to describe its routers as safe to hacking, when in fact nearly 13,000 of them had been exploited by hackers. ASUS was forced to pay only $16,000 per violation and submit to independent security audits every two years for the next 20 years.[6]

In a nutshell, the FTC's general approach to personal data regulation is light-touch and focuses on the key ideas of legal certainty, economic incentive, and innovation. The FTC believes industry-self regulation is a laudable goal, but has largely failed in the area of internet technology (Federal Trade Commission, 2012). Instead, federal regulation is needed, but only at a base level that is "technology neutral" and leaves considerable room for individual industries and firms to best decide on how exactly to comply (Federal Trade Commission, 2012). Part of the reason in choosing to support basic regulation is that it will provide firms with the "certainty they need to understand their obligations;" additionally, the threat of punishment gives some incentive to meet those legal obligations. And finally, base regulation will give consumers confidence that businesses must respect their privacy, which will spur consumer activity (Federal Trade Commission, 2012). In short, the FTC rejects prescriptive regulation and can be viewed as a currently "default hybrid" system of governance with aspirations towards a more robust "baseline hybrid" system. Time will tell whether the FTC's exhortations will influence Congress enough to pass a baseline federal law regarding personal data protection.

5.2 EU Personal Data Regulation 1995-Present

The following section will briefly situate the GDPR as an evolution of the 1995 Data Protection Directive and examine the legal grounds of processing personal data are available for firms under the GDPR. This subsection will revisit the story of European personal data processing law through its impact on the corporation.

5.2.1 The 1995 EU Data Protection Directive

As detailed in Part I, the story of contemporary European data protection begins with the 1995 EU Data Protection Directive (95/46/EC). The Directive had two major goals: to protect natural individuals' personal data and to help establish the "free flow" of data around the European Union (DPD 95/46/EC). However, the nature of an EU Directive is such that the general guidelines are drafted and agreed upon in the European Parliament, but individual nations have some freedom in the details of implementation. The end result after nearly 20 years was a kind of patchwork implementation of the Directive where some countries' Data Protection Authorities had adopted different customs and procedures over the years, leading to sometimes contradictory procedures and customs in different countries.

For example, if a company wished to use BCRs as a means of international transfer of personal data, there exists a "mutual recognition" system in which 21 EU member countries mutually recognize the decision of the lead data protection authority as sufficient to grant national permission for the binding corporate rules.[7] If a corporation based in France, for instance, wanted to transfer its

[4] FTC Receives Largest COPPA Civil Penalties to Date in Settlements with Mrs. Fields Cookies and Hershey Foods www.ftc.gov/news-events/press-releases/2003/02/ftc-receives-largest-coppa-civil-penalties-date-settlements-mrs

[5] Hershey Foods Corporation Management's Discussion and Analysis www.annualreports.com/HostedData/AnnualReportArchive/h/NYSE_HSY_2003.pdf

[6] www.ftc.gov/news-events/press-releases/2016/02/asus-settles-ftc-charges-insecure-home-routers-cloud-services-put

[7] https://ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/binding-corporate-rules_en

personal data to a branch in Denmark or Austria– countries which are not part of the mutual recognition group–it would need to apply separately to each national data protection authority, a process which a major law firm describes as "not generally... a smooth experience" (Allen & Overy, 2016). In sum, the various national laws surrounding personal data needed to be harmonized and brought up to date with the rise of new technologies and internet-based business models, in particular machine learning and its implications for micro-targeted advertising that led to Facebook and Google disrupting the advertising industry over just a few years. With these developments changing the Internet regulation landscape, the Directive was repealed after more than twenty years in force and the GDPR went into effect in 2018.

5.2.2 The GDPR

In contrast to the light-touch, self-regulatory policy recommendations of the FTC, the GDPR will deeply impact the way data-rich companies operate. It replaces the Data Protection Directive that governed data collection and storage in the EU since 1995 and helped spur the creation of the "EU Digital Single Market," which standardized the European data-economy and provided a consistent framework for all data-related products and services in the EU.[8] All companies and institutions that process the data of EU-residing data subjects are bound by this regulation, wherever the company may be based. This includes companies in countries not located in the EU. Under the GDPR, "data controllers," the legal entities that determine how personal data are processed (i.e., Facebook and Google), will be legally liable for any infringement of GDPR law. And with hefty fines for non-compliant firms, there is a very strong economic incentive to comply: depending on the type of violation, data controllers who are found in non-compliance could be liable for up to 4% of global turnover or 20M Euros, whichever is greater.

Legal Grounds of Personal Data Processing under the GDPR

Under the GDPR, the notion of consent referenced in the Charter of Fundamental Rights has further been clarified. Essentially, if data controllers choose to use consent as a legal basis for data processing, then consent to said processing must be clear, written in unambiguous language, and not "bundled" in with other written agreements (Bird & Bird, 2017). Additionally, users must be able to withdraw their consent to processing at any time and cannot be used as a kind of "pay to play" strategy in which users must agree to processing of personal data in order to use a platform's services (ibid.).

In Article 6 (1) of the GDPR, the six grounds on which data may be processed are laid out. These are: (1) when the data subject has given consent; (2) when necessary for the performance of a contract with the subject or in order to enter into a contract; (3) when necessary in order to comply with a law; (4) when necessary to protect a "vital interest" of a data subject or another person; (5) when necessary to perform a task that is in the public interest or in the exercise of official authority granted to the controller; and finally (6) necessary for the "legitimate interests" of the controller or a third party, except where these legitimate interests are overridden or outweighed by the interests, rights, or freedoms of the data subject.

If a data controller claims "legitimate interests" as grounds for the processing of personal data, these interests may be tested using a "Legitimate Interests Assessment" or LIA.[9] The process roughly proceeds as follows. First the firm must determine whether a legitimate interest exists. Next it must then justify the processing of the data on this legitimate interest. This legitimate interest and the necessity of processing must be then balanced against the rights, freedoms, and interests of the data subject. This final step will also include factors such as the nature of the personal data (i.e., how sensitive are the data?), the expectations of the user, the potential impact of the processing, and whether appropriate data safeguards have been put in place (e.g., pseudonymization) (ibid.). Throughout these deliberations, firms that claim legitimate interest in the processing of personal data must be prepared to properly document and record their deliberations to ensure that the users' interests were not unjustifiably overridden by the data controller

[8] ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/binding-corporate-rules_en

[9] www.econsultancy.com/blog/69303-gdpr-for-marketers-five-examples-of-legitimate-interests

5.3 Safe Harbor and Key EU Court Rulings

In order to understand the creation and eventual invalidation of the Safe Harbor agreement between the EU and the US, we must review three recent European legal decisions made by the Court of Justice of the European Union (CJ/EU/Luxembourg Court) (Loidean, 2016). The first is Schrems (2015), Digital Rights Ireland (2014), and Google Spain (2014). This section will provide a short overview of the Safe Harbor regime, and then review of each of the major cases that led to its consequent demise and rebirth as Privacy Shield.

5.3.1 The Basics of Safe Harbor

The Safe Harbor agreement is founded on a set of principles closely related to the OECD's seven principles of protection of personal data, which also influenced the GDPR.[10] Since these principles closely resemble the GDPR's principles, we should not be surprised that by following them European personal data could be legally transferred to the US. In essence, by agreeing to participate in Safe Harbor, US corporations were committing to follow privacy standards nearly equivalent to those in the GDPR. In order to be Safe Harbor-certified, US corporations must promise to process Europeans' personal data according to the following principles:[11]

- **Notice** Individuals must be informed that their data is being collected and how it will be used. The organization must provide information about how individuals can contact the organization with any inquiries or complaints.
- **Choice** Individuals must have the option to opt out of the collection and forward transfer of the data to third parties.
- **Onward Transfer** Transfers of data to third parties may only occur to other organizations that follow adequate data protection principles.
- **Security** Reasonable efforts must be made to prevent loss of collected information.
- **Data Integrity** Data must be relevant and reliable for the purpose it was collected.
- **Access** Individuals must be able to access information held about them, and correct or delete it, if it is inaccurate.
- **Enforcement** There must be effective means of enforcing these rules.

But why did the European Commission feel the need to create the Safe Harbor agreement in the first place? According to Article 25 of the DPD (95/46/EC), the European Commission can come to a decision as to whether a third country, such as the US, provides adequate "protection of the privacy and fundamental rights and freedoms of individuals," which are embodied in the seven principles above (DPD Article 26(2)). In 2000, the European Commission found that the US met the minimum standards of personal data protection, provided that companies adhered to the Safe Harbor agreement, which allowed participating US companies to transfer personal data to the US for storage and processing.[12] However, only companies that were regulated by the FTC or the Department of Transportation could participate: this meant that firms in certain industries were excluded from joining. Excluded from the agreement were firms in such industries as finance, telecommunications, non-profit, and meat processing, among others.[13]

First, it is worth pointing out that the European Commission's adequacy finding did not give *carte blanche* to US companies to process EU personal data. As was discussed above in section 5 on the FTC, data protection in the USA is a mix of sectoral guidelines, state laws, and a handful of federal laws regulating credit, health data and children's privacy, for example. Second, Safe Harbor is a completely self-regulated scheme in which "self-certified organizations voluntarily" comply with

[10] marcomm.mccarthy.ca/pubs/share2.htm
[11] wikipedia.org/wiki/International_Safe_Harbor_Privacy_Principles
[12] www.privacyshield.gov/Program-Overview
[13] www.privacyshield.gov/list

the Safe Harbor privacy principles designed by the US Department of Commerce (Loidean, 2016). Judgments of non-compliance were thus left in the hands of the very corporations that had economic incentives to use personal data to improve their products and services and serve more accurate ads, in the case of companies like Google and Facebook. It was also possible for companies to pay for third-party verification services if they chose not to re-certify annually,[14] however such an option could easily lead to corporations doing "certification shopping" whereby they search for the third-party most likely to certify them. Additionally, participation in the regime was weak: after nearly four years, only 400 companies had registered with the US Department of Commerce, though by 2015 nearly 5,000 had joined.[15] These problems at the outset foreshadow the eventual demise of the agreement 15 years later.

5.3.2 The Schrems Case

The Schrems case was the culmination of several other landmark "digital rights" cases that eventually led to the invalidation of Safe Harbor. In 2013, the Austrian citizen Max Schrems made a complaint to the Irish Data Protection Commissioner in which it was claimed that Facebook-Ireland's transfer of EU citizens' personal data to the US violated EU law. At the time, these kinds of transfers were legally valid under the Safe Harbor agreement. Essentially, the argument made by Schrems was that in light of Edward Snowden's leaking of classified US government surveillance programs, "there was no meaningful protection in US law or practice" for personal data transferred to the US because US law enforcement could obtain access to personal data without a court order (Loidean, 2016). Initially, Schrems' complaint was dismissed and thrown out by the Irish DPA because Schrems could not demonstrate that his personal data were actually affected.

Yet, just one year later, the Irish High Court ruled differently and concluded that Schrems did in fact have legal standing under EU law, due primarily to the Digital Rights Ireland holding which stated that it did not matter if the complainant had been personally affected in order to show that his right to respect for private life (Articles 7 & 8 of the Charter of Fundamental Rights) had been infringed (Loidean, 2016). The court also expressed concerns about US law enforcement surveillance and the lack of personal data protections in US law. In particular, the court noted that EU data subjects had no effective means of judicial review under Safe Harbor for privacy complaints (ibid.). Ultimately, the court found that the "adequacy" of the protection given to personal data in the Safe Harbor agreement was not enough and declared it invalid in 2015.

5.3.3 The Digital Rights Ireland Case

In coming to its ruling in the Schrems case, the Irish High Court based its decision mostly on the CJEU's judgment in Digital Rights Ireland, which invalidated the Data Retention Directive (2006/24/EC). This Directive modified the 1995 Directive and allowed for the general retention and collection of communications metadata for purposes of law enforcement in the EU (ibid.). According to Loidean (2016), Digital Rights Ireland is highly significant because it "marked the first time that the CJEU has ever struck down an entire EU legal instrument due to its incompatibility with the EU Charter," which had the practical legal effect of solidifying the influence of fundamental rights on EU legal decisions. Loidean (2016) explains further that Digital Rights Ireland "established unequivocally that strict legality, necessity, and proportionality standards must underpin the safeguarding of privacy and data protection rights." It is important to note the timing of the Safe Harbor agreement, which was made in 2000 and the coming into effect of the Lisbon Treaty, which led to the EU Charter of Fundamental Rights (ECFR) becoming part of EU law in 2009. At the time the Safe Harbor principles were agreed upon, the ECFR had not yet been developed and so, looking back from today, it seems inevitable that there would be a conflict between the ECFR and the Safe Harbor principles.

[14] www.privacyshield.gov/Program-Overview

[15] www.privacyshield.gov/list

5.3.4 The Google-Spain Case

Finally, the third major case related to personal data privacy and protection came shortly after the Digital Rights Ireland decision and involved the Internet giant Google. In this case, the Grand Chamber of the Luxembourg Court established that, in brief, "EU citizens have a right to have links concerning them delisted from search engines that essentially encroach on their private lives and the protection of their personal data (Loidean, 2016)." This is the ruling that led to the now-famous "right to be forgotten" that was added to the GDPR. The Google Spain case was particularly important because it addressed two major issues of the digital age: first, that search engine sites like Google can act as a permanent store of an individual's personal data, even when those data are incorrect or removed; second, search engines like Google can have a big impact on one's online and offline identity and reputation (ibid.). Nevertheless, the Google Spain decision with hotly debated, with opponents of the decision arguing that it limited free speech and access to information (ibid.).

5.3.5 Safe Harbor is Reborn as Privacy Shield

The CJEU's sudden invalidation of Safe Harbor in 2015 was not the only reason for its demise. Several critiques of Safe Harbor were based on fears raised by the Snowden leaks. These included allegations by European Parliament members that Safe Harbor signees Microsoft and Google may have played a role in US government surveillance programs (Weiss and Archick, 2016). Other criticisms included the weakness of its government enforcement (the FTC only brought action against ten companies during the first 13 years of Safe Harbor) and the technological obsolescence of the Data Protection Directive, which had been written in the late 1990s (ibid.). All in all, the legal status of Safe Harbor was tenuous even before the Court invalidated it.

With a legal vacuum persisting for over one year, the new Privacy Shield agreement was released in February of 2016 and included revamped privacy principles found in Safe Harbor, plus additional principles such as legal recourse for European citizens (Weiss and Archick, 2016). Unlike Safe Harbor, Privacy Shield contains commitments from US national security officials that EU subject's data rights will be respected (Weiss and Archick, 2016). Probably chief among them was the inclusion of several redress possibilities for EU citizens who believe their data to be compromised by US processing and increased authority of the FTC to help monitor disputes (ibid.). Seen this way, the Privacy Shield framework may be considered something close to a "default hybrid" approach to data governance in that companies are largely left to self-regulate and self-certify adherence, but the EU Data Protection Authorities and the FTC may become involved if companies do not respond to complaints within a specified amount of time (ibid.). Further, the Department of Commerce will help to monitor compliance.[16]

Finally, in August of 2016, following the advice of the Article 29 Working Party and subsequent review by the representatives of EU member states (Article 31 Committee), the European Commission formally adopted the Privacy Shield framework to govern EU-US personal data transfers (ibid.).[17] Currently there are 3151 US companies participating in the agreement.[18] Nevertheless, there are still some residual worries about US surveillance activities and legal remedies for EU citizens, and there is even a push for Congress to pass a "Judicial Redress Act" that would allow EU citizens formal means of legal redress for privacy violations (ibid.). The argument behind this is that it would increase European confidence in that Privacy Shield framework and US data protection laws. To date, however, such an act has not been passed.

[16] urlwww.privacyshield.gov/Program-Overview

[17] When the GDPR went into effect on May 25th, 2018, the Article 29 Working Party ceased to exist and was replaced by the European Data Protection Board (EDPB)

[18] www.privacyshield.gov/list

5.4 Comparing Standard Contractual Clauses with Binding Corporate Rules for Corporate Data Transfers

Now that we have examined the history of the Safe Harbor/Privacy Shield framework for the transfer of EU-US personal data, we will now consider two other global means of transfer: standard contractual clauses (also sometimes referred to as "model clauses") and binding corporate rules (BCRs).[19] Binding corporate rules as a legal means of transfer are particularly interesting because they represent a significant new addition to the GDPR over the Directive. In fact, the former Vice-President of the European Commission, EU Justice Commissioner Viviane Reding, has publicly stated on multiple occasions that she believes BCRs are the future of multinational data transfers in the global economy. She states that BCRs represent "a way for changing legal rules [to respect] the fundamental rights of individuals whilst simultaneously promoting innovation and accelerating economic growth (Reding, 2011)." For this reason, this section will focus its analysis on this new legal instrument that seems poised to change the legal landscape for global data transfers.

5.4.1 Shortcomings of Standard Contractual Clauses

Prior to the introduction of the GDPR, contractual clauses were probably the most common way for corporations to transfer personal data outside of the EU for processing (excluding consent). They are called "standard" because these kinds of contracts follow a standard format provided by the European Commission (US Chamber of Commerce, 2014). Although popular for many businesses and perhaps useful for simple "point A to point B" transfers, these standard contractual clauses have several shortcomings. First, they are rigidly structured and do not take into account the fact that data volumes and flows can grow in complexity over time (US Chamber of Commerce, 2014). Second, *data controllers* (the legal entity determining how the data are used) need to create contracts for each and every importing group they transfer personal data to and for every type of data and for every purpose of transfer (ibid.). For large multinationals, this could easily amount to hundreds of separate contracts. Most commonly, the "importers" of the personal data would be subcontractors tasked with processing personal data on behalf of the controllers. One can quickly see how standard contractual clauses would become a bureaucratic nightmare for all but the simplest of *ad hoc* data transfers. Nevertheless, for very small firms with minimal transfer needs, standard contractual clauses might be a satisfactory option.

5.4.2 Binding Corporate Rules and the Role of Adequacy

BCRs can be summarized as a voluntary "set of binding rules that can be put in place to allow multinational groups to transfer personal data from the European Economic Area (EEA) to affiliates outside the EEA in compliance with national laws implementing the EU Directive (Allen & Overy, 2016)." Article 47 (2)(a-n) of the GDPR spells out the specific requirements for a corporation's binding corporate rules. At the very least they must include:

- complaint procedures
- a statement of the guiding privacy principles
- auditing and employee training mechanisms
- the acceptance of liability for any rule breaches
- an explanation of how data subjects can access the rules

[19] It should be noted that under the GDPR it is also possible for firms to transfer personal data internationally upon receiving consent from the data subject. I do not consider this legal means of transfer because it would be too cumbersome for a large corporation to obtain consent from thousands of employees or clients.

Once corporations have implemented these requirements, the rules become legally binding for all involved parties. Practically speaking, this means the corporation itself, its employees and data subjects–through the legal notion of third party beneficiary rights–and any independent contractors the corporation are all bound by the BCRs.[20] Note that BCRs do not cover international data transfers to firms that are not subsidiaries or contractors, unlike other data transfer agreements such as the APEC Privacy Framework. With that said, BCRs can also cover third party processors that process personal data on behalf of the data controllers, which allows them to cover the use of cloud transfers across borders (Reding, 2011). Such agreements would typically be included in the Service Level Agreement (SLA) between the controller and processor (Allen & Overy, 2016). Finally, it should be emphasized that BCRs do not allow firms *carte blanche* to collect and process personal data; the existence of BCRs merely allows a corporation to transfer personal data within the corporate network (and to its data processors), to countries in which there may not be adequate personal data protection laws.

The process through which a corporation can have its BCRs approved is roughly as follows. First the company must choose a lead data protection authority (DPA). This lead DPA will then review the draft BCRs and decide if they meet the standards set out in the Article 29 Working Party working papers, specifically WP 153.[21] Then, the lead DPA will facilitate the authorization with the other DPAs located where the firm has subsidiaries and which do not have adequate personal data protection (Allen & Overy, 2016). In order to speed up this last step, a system of mutual recognition has evolved, whereby other countries recognize the validity of the BCRs if they lead DPA judges them to be valid (ibid.). To date, 21 mostly EU countries participate in this mutual recognition group (Allen & Overy, 2016).

The same concept of "adequacy" that appeared in Safe Harbor/Privacy Shield is fundamental to understanding the impetus for the Article 29 Working party to develop and promote the idea of BCRs. According to the Court of Justice of the European Union (CJEU/Luxembourg Court), "adequacy" (in the context of personal data transfers) means that international cross border personal data transfer frameworks between an EU and non-EU country "should meet a level of protection of fundamental rights that is 'essentially equivalent' to that guaranteed by EU law," which follows the requirements of the EU Directive and the EU Charter of Fundamental Rights (Loidean, 2016). In other words, in order for a personal data transfer agreement to be "adequate" it has to at least provide a level of data protection similar to that found in an EU member country. If it can be proved that the agreement does not keep to at least this standard of protection, then personal data transfers are forbidden under the EU Directive. In fact, this is exactly what happened in the Schrems case.

At present, the Article 29 Working Party[22] has issued adequacy judgments for roughly 13 countries, most of which are extremely small sovereign states such as Jersey, the Isle of Man, the Faroe Islands, and Guernsey. Recently, however, Japan was deemed "adequate" for international personal data transfers.[23] The USA is on the approved list, but only for companies that participate in the Privacy Shield framework. There are ongoing adequacy talks with South Korea and Taiwan, among others.

---

[20] iapp.org/media/presentations/12Summit/S12_Binding_Corporate_Rules_PPT.pdf

[21] /ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/binding-corporate-rules_en

[22] The Article 29 Working Party is now the European Data Protection Board (EDPB). The EDPB is a group of independent data protection advisors from each of the member states–the Data Protection Authorities–the European Data Protection Supervisor, and the European Commission (Allen & Overy, 2016). Prior to the GDPR, the Article 29 Working Party was hugely influential for several reasons. First, the advisory group acted as a kind of independent auditing board to make sure that national laws reflected the 1995 Directive. Second, the group was tasked with issuing opinions on the codes of conduct for member states and suggesting amendments to the Directive. Third, the group was responsible for determining which countries have "adequate" personal data protections such that they can be freely transferred without special permission (EU Directive 95/46/EC). It is reasonable to assume these roles will be overtaken by the new EDPB. ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/adequacy-protection-personal-data-non-eu-countries_en

[23] http://europa.eu/rapid/press-release_IP-19-421_en.htm

5.5 Three Domain Analysis of Binding Corporate Rules

In this concluding section, we will examine binding corporate rules in more depth. Specifically, binding corporate rules will be analyzed through the *three domain* conceptual framework championed by Schwartz and Carroll (2003), which views the actions of a corporation through their legal, economic, and ethical effects. The *three domain* framework of Schwartz and Carroll (2003) has its roots in the area of corporate social responsibility (CSR) and is an update to Carroll (1979).

5.5.1 Legal Effects

Corporations that rely on binding corporate rules have several advantages over firms that do not. Firstly, the nature of BCRs is such that they can cover a broader scope of possible data transfers. Typically, if firms were to rely on consent or standard clauses, the consent or contract would only be valid for the kinds of transfer specifically designated in the contract. For example, if parent company A and subsidiary B wished to transfer personal data between themselves for the purpose of employee performance evaluations, they would need consent or a contract. Once a company wants to change the nature of the data transfers, say, for predicting employee churn, it would need to get consent again or redraft the contract to include this new purpose. Doing so could very quickly become a very time consuming and expensive endeavor. Binding corporate rules, however, allow for changes in the nature of data to be transferred (e.g., some kinds of sensitive personal data may now be included) or the structure of the corporate group (e.g., if a new subsidiary is acquired) without completely having to reapply for BCR approval (Allen & Overy, 2016).

Second, as more and more firms adopt BCRs, global standards for data privacy will become more consistent. This means that where multinational firms transfer data to countries with weakly enforced or non-existent data privacy laws, such as China, Iran, or in some sectors, even the US, the BCR standards will apply. And conversely, where local laws are of a higher standard than even the BCR, then those local laws will apply (Article 29 Working Party WP 257). If more non-EU countries recognize BCRs as providing equivalent protection to their respective national laws, then BCRs may even become the new *de facto* international standard for personal data transfers. The adoption of BCRs therefore seems like a win-win for advocates of greater personal data privacy around the world, especially for citizens of countries with a history of state surveillance of electronic communications.

Nevertheless, there are some legal drawbacks to using BCRs. Owing to the GDPR's *accountability principle*, data controllers using BCRs must assume all legal liability for any breaches of the rules by processors inside or outside of the EU. This is a considerable legal risk since fines could be up to 4% of global revenues for the most serious violations. Similarly, this liability means that firms outside of the EU will indirectly be put under EU regulatory scrutiny, something some firms may not be happy about (Allen & Overy, 2016).

Further legal drawbacks are related to the complexities of international law, cooperation, and regulation. As Wugmeister et al. (2006) point out, the inherently international nature of BCRs[24] raise questions about how they fit in with various national laws and regulatory bodies, especially when there are cross-border disputes and data breaches. In the case of the EU member state Data Protection Authorties, there are no guarantees that the lead DPA will accept a corporation's BCRs. That is, even if a firm's BCRs *are* accepted by the lead DPA, some member states are not part of the mutual recognition group that recognizes the declarations of other member states' DPAs. Making things even more complicated is that fact that some member states do not recognize the legal concept of a "unilateral declaration (Allen & Overy, 2016)." These issues will hopefully be resolved in the near future as the various member state DPAs streamline the BCR application process.

But beyond the EU, there is also no guarantee that other countries will recognize BCRs as equivalent to their national laws, even though it is highly likely most countries agree that the EU's data protection standards are high. As Moerel (2012) notes, international mutual recognition of BCRs is dependent upon the fact that "countries do recognize self-regulatory tools rather than public regulation or contractual tools as an instrument to regulate transborder transfers." In countries with

[24] Wugmeister et al. (2006) use the term "Corporate Privacy Rules" to refer to what are essentially generic versions of the GDPR's Binding Corporate Rules.

strong government presence, e.g., China or Russia, or in countries with relatively weak rule of law, e.g., Thailand, there may be less enthusiasm for the self-regulatory nature of BCRs. Despite these issues, however, it is likely that the GDPR will mitigate the uneven and sometimes contradictory national laws that made BCRs exceedingly complex under the Directive.

5.5.2 Economic Effects

Seen from an economic point of the view, BCRs are not the most attractive option for the average corporation. Due to the complicated negotiations among DPAs and the staggering amount of EU regulatory approval needed to initiate one, they do not make sense except for all but the biggest multinational corporations. Allen & Overy (2016) note that one corporate client took 11 months from start to finish, and this was unusually quick. Unless the firm has been preparing its internal data governance policies for years, it could easily take longer than one year to get approval. The initial costs for implementing the required training, auditing, and internal communications can also be substantial. The list of approved corporations with approved BCRs is thus a veritable list of global giants: Intel, BP, BMW, Citigroup, HP, and GlaxoSmithKline among others.[25] On the other hand, long run compliance costs are likely to be reduced due to the BCR, especially for firms with complex intra-group personal data transfers. Whereas with standard contractual clauses new versions would need to be continually redrafted and agreed to in every single country in which the corporation operated in, the BCRs can include a clause that allows for changes to be made in the future, should a corporate group wish to transfer to a new country. Of course, this change would still need to be approved by the lead DPA.

BCRs may also have some positive economic benefits besides lower long-run compliance costs. Firms that implement BCRs may see increased trust among employees–due to better handling of sensitive employee data–and clients and consumers. These intangible benefits felt among employees could manifest themselves in the form of increased job satisfaction and reduced employee turnover. And regarding consumer trust, there is growing evidence that firms that publicly commit to corporate codes of conduct experience increased profits (Orlitzky et al., 2003). There may thus be some market incentives for firms to "compete on privacy" in order to win the trust of consumers.

5.5.3 Ethical Effects

From an ethical perspective, BCRs may have effects analogous to corporate social responsibility codes that are growing popular among eco-conscious multinationals. BCRs force multinational corporations to take individual privacy seriously from the top down, unlike standard model clauses. As Robbins and Judge (2017) note, "Ethical top leadership influences not only direct followers, but all the way down the command structure as well, because top leaders create an ethical culture and expect lower-level leaders to behave along ethical guidelines." In other words, as a result of the required auditing, communication, and training for BCR approval by the lead DPA, along with the legally binding effects on corporate employees, an individual's right to privacy becomes a priority at all levels in the corporation. Whereas standard model clauses are created in an ad hoc fashion and function mostly as "transfer tools," BCRs represent a major undertaking that will necessarily need to be infused into the corporate culture. The result of this "infusion" of privacy principles likely leads to safer and more secure personal data in the organization, which for firms in the healthcare or insurance industries, for example, would be a decidedly good thing.

Secondly, the emphasis on "accountability" can reduce the number of cases where negligent third-party processors permit a data breach to occur in a country with sub-optimal data protection. Moerel (2012) writes that the "main purpose of 'accountability' is to use the law to hold businesses accountable for taking their responsibilities seriously by using various mechanisms to encourage or force businesses to put internal governance structures and management systems in place." As an example of this kind of thinking, the Article 29 Working Party's WP 257 stipulates that when BCRs apply to processors, "[they] have an obligation to make available to the controller all information necessary to demonstrate compliance with their obligations including through audits and inspections conducted by the Controller or an auditor mandated by the Controller (GDPR Article 28(3)(h))."

[25] ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/binding-corporate-rules_en

Accordingly, not only are controllers ultimately responsible for the conduct of processors, but processors in turn have duties to demonstrate their compliance with the rules specified in the BCRs. In third countries where the prospect of a data breach is relatively high, such accountability requirements and internal governance structures should lead to a reduction in unethical uses of personal data (i.e., data being sold to hackers or criminal organizations). An indirect benefit of this is that knowledge of stronger internal governance and management systems may be transferred to areas of the world where these are still not commonplace.

In spite of these benefits, BCRs pose challenges for the meta-norms of law. Historically data protection laws have been territorial in nature, meaning that they restrict storage and processing of data to within the country. Part of the reason for this is due to enforcement concerns. Privacy laws designed in the 1990s, for instance, were not designed with technologies like cloud computing in mind, where data may be collected in one location and then stored in another foreign location thousands of miles away. If data are processed outside the country, then what legal mechanism does a regulatory agency have for enforcing the laws?

Some international privacy frameworks, most notably APEC's Privacy Framework, have attempted to designate "accountability agents" that can be either public or private entities whose judgments of adherence to the APEC privacy principles are mutually recognized among member states.[26] There are also conditions for mutual assistance in the case of cross-border data breaches, the so-called "network approach" to cross-border data protection disputes. Nevertheless, the APEC Framework has not been a resounding success and it is likely to be soon eclipsed by the effects of the GDPR (Greenleaf, 2009). Moerel (2012) therefore suggests that the GDPR be amended to include provisions that allow BCRs to borrow the concepts of applicable law, jurisdiction, and enforcement from private international law (PIL). If this were the case, then complainants could be entitled to allege a violation of the BCRs in their native language (with translation), and then choose the applicable national law that would apply in resolving the dispute. Currently, disputes that are not resolved internally can be directed to the relevant EU DPA, but such legal recourse appears only to be available to EU citizens.

Finally, if BCRs are to operate at the global level, it stands to reason that they should reflect universal points of agreement among the majority of nations, each of which might be viewed as a rationally self-interested agent. The universality of such norms is what sets them apart from mere legal guidance and imbues them with a moral character. As Lorenzo Sacconi (as cited in (Moerel, 2012)) writes in his discussion of corporate governance:

> [social norms should be] commonly accepted on the basis of agreements and conventions, and therefore sustained by rational choices of multiple agents. Once those norms have passed the universalizability test, which is what meta-ethically distinguishes them from rules of mere prudence or from social norms not susceptible to moral meaning, they are effective by dint of their social function of solving cooperation and coordination problems.

In the Western world this may be a feasible goal: the idea of fundamental and inalienable human rights extends back hundreds of years. In the EU, for example, we saw the EU Charter of Fundamental Rights that guarantee an individual's right to data privacy. But how are we to square such rights against China's proposed "social credit" system?[27] Given Europeans' historical distrust of government surveillance programs, it would seem highly unlikely for them to agree to any kind of cooperative data privacy framework that permitted such uses of personal data.

Similarly, China's recent release of new standards for personal information protection, which took effect on May 1, 2018, appears to be competing with the GDPR (ibid.). If China does indeed become a global leader in cloud storage, AI, and data processing, then we may see a move away from GDPR-focused compliance in favor of the Chinese standards. In fact, EU-based companies may find themselves scrambling to comply with the Chinese law as the Belt and Road initiative makes intercontinental travel between the two regions cheaper and faster. If this were to happen, then multinational businesses would be forced to navigate the complex legal web of not just one, but two major personal data privacy laws. Such a scenario would place multinationals in the same place they were before the GDPR, where territoriality restrictions and reliance on simple contracts made global

[26] https://www.apec.org/Publications/2005/12/APEC-Privacy-Framework

[27] www.csis.org/analysis/new-china-data-privacy-standard-looks-more-far-reaching-gdpr

data transfers exceedingly onerous, costly, and slow. In sum, global commerce gains only when there is one standard for privacy protection. The emergence of competing privacy frameworks, such as the new Chinese standard, may force multinationals to decide whose standards they ultimately feel are "most rational" and "generalizable." Given the massive amount of infrastructure investment, it is not certain that developing countries will prefer the privacy benefits of the GDPR over the economic benefits of the Chinese approach to personal data regulation.

Part III

Implications for Data Scientists and Behavioral Researchers

Overview of Part III

Data scientists and behavioral researchers are expected to be disproportionately affected by the GDPR because their work rests on the collection, processing, and analysis of ever-increasing amounts of Behavioral Big Data (BBD). At the same time, these two groups are largely unaware of the legal complexities surrounding the processing of personal data.

Part III of this thesis lays out a data-science oriented framework, called Information Quality (InfoQ), to analyze how the GDPR might affect the modern workflows of data scientists and behavioral researchers. InfoQ decomposes a data science project into four components: goal, data, analysis, and utility measures. We can analyze the impact of the GDPR on each of these components and their relationships by examining the eight InfoQ dimensions. This section will focus on changes in data resolution, structure, integration, chronology, generalizability, and communication due to GDPR. By assessing InfoQ at the beginning of a study, data scientists and researchers can understand which of the four general components are likely to be problematic and react accordingly.

Under GDPR, goals are set by data controllers and the processing goals should be made transparent to users. Repurposing of data for other goals is generally not allowed. Further, the goals set by data controllers must be balanced against data subjects' rights to privacy. Yet, the GDPR also spells out several exemptions from these rights for the goals of statistical and scientific research. Data controllers must use security techniques, such as pseudonymization, to safeguard the personal data of data subjects. Sensitive categories of personal data must be treated with an even higher standard of care. Data processors process personal data in the name of data controllers; controllers are legally responsible for any violations of the GDPR by downstream processors. Common types of data processing include automated profiling, recommendation systems, and more generally any kind of operation on personal data. Lastly, the utility of any processing must be balanced against the risks of privacy harms to data subjects and society. In cases where privacy harms are likely, then Data Protection Impact Assessments (DPIAs) can be done, or explicit consent by the data subject is needed.

A modern data science workflow proceeds in several stages, from initial data collection and exploration, to model building, sharing, and generalization, and eventually to the communication of findings. At each step, data scientists should be aware of the relevant GDPR principles. In the first stage of collection, data scientists must adhere to the principles of data minimization and purpose limitation, while being mindful of their data environments. When building predictive models, they must deal with issues of lost consent, data availability and heterogeneity, and transparency. If corporations choose to share data with researchers, they need processing agreements; further, corporations will be liable for any violations of the GDPR by researchers. This may lead to data access divides. The GDPR could affect the generalizability of research findings due to consent bias on large BBD-generating platforms. Intellectual property laws may also hinder scientific reproducibility. Finally, the GDPR forces data scientists to document all processing for review by Data Protection Authorities (DPAs) and be able to communicate clearly and transparently to data subjects about the scope of processing. They must also be prepared for data subjects to exercise their right to be forgotten. To successfully adapt to these changes, data scientists and behavioral researchers must internalize the GDPR's principles and begin to approach their work in a new way that prioritizes the reduction of privacy harms to people and societies.

Chapter 6

Information Quality: A Framework for Analyzing the Effects of GDPR

6.1 Introduction to the InfoQ Framework

A data scientist working either in academia, a company or an organization, starts with either a goal and then searches for the right dataset, or else starts from a dataset and identifies a useful goal to pursue. The data scientist then applies data analysis methods and is continuously conscious of the relevant metrics for measuring the study's success, such as in terms of company KPIs or successful publication in a scientific journal. In short, the four key ingredients that a data scientist works with are goal, data, analysis methods, and utility measures. These four components compose the concept of Information Quality (InfoQ), defined by Kenett and Shmueli (2016) as "the potential of a particular dataset to achieve a particular goal using a given data analysis method and utility" or more formally:

$$InfoQ(g, X, f, U) = U(f(X|g)) \tag{6.1}$$

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

Nevertheless, these four InfoQ components are relatively abstract and need to be operationalized into measurable dimensions before any real-world data science studies can be carried out. Kenett and Shmueli (2016) propose deconstructing InfoQ into eight dimensions: data resolution, data structure, data integration, temporal relevance, chronology of data and goal, generalizability, operationalization, and communication. The following section will break down each of these dimensions and relate it back to the original concept of InfoQ, while at the same time discussing the dimension's relation to GDPR. The discussions that begin here will later be fleshed out in chapter 7.2, which examines the GDPR's impact on a typical data science project workflow.

6.1.1 Data Resolution

Data resolution is concerned with answering the question of whether, given the study's goal and utility measure, the data at hand are measured at the relevant scale and appropriately aggregated (Kenett and Shmueli, 2016). For example, in forecasting studies it is common for analysts to decide what temporal level the forecasts should be: are daily forecasts necessary or would weekly forecasts suffice? Typically it is easier to go from very fine-level measurements to more general measurements (i.e., from seconds to hours, or days to weeks) as the aggregation process can allow for competing sources of noise to cancel themselves out and thereby provide more stable measurements over time. Additionally, a cross-sectional study aimed at customer behavior, for example, will need decide whether predictions are most useful on the customer or item level. Another resolution aspect is the choice of aggregation function, e.g. whether a weekly average should be used or the weekly maximum. The most suitable form of aggregation will depend on the study's ultimate goal. Different aggregation functions may also have other privacy-related consequences, for example it may be easier to identify outliers if maximum instead of average values are used.

Data resolution in the GDPR

The most relevant concept of GDPR related to data resolution is pseudonymization, which will be covered in more detail in 7.6.2. There, we will see how the GDPR is likely to affect data science projects during the period before data are collected. For now, the most important thing regarding data resolution is that the GDPR's requirements may influence not only the level of predictions that can be made, but also the types of analyses that can legally be done. If individuals can be singled out–identified–then special precautions will need to be taken that include–but are not limited to– using coded (i.e., pseudonymized) data that may make individual-level predictions difficult or impossible. Similarly, during exploratory data analysis, certain visual representations of data such as scatterplots and frequency tables can make it easier for individuals to be personally identified. The GDPR calls for organizations to thus take certain steps to ensure that the probability of re-identification is sufficiently reduced (e.g., the GDPR principles of privacy by design and by default, and data minimization).

6.1.2 Data Structure

Data structure refers to the type of data collected (e.g., time series, cross-sectional, or network data) and whether they contain any corrupted or missing values (Kenett and Shmueli, 2016). For example, in the era in BBD, vast amounts of unstructured data in the form of images, text, and video are generated, however, in order to use statistical learning algorithms, these data must first be converted into a structured format (often numerical) before any models can be trained. This conversion process generally carries with it some cost in InfoQ in the form of missing or incomplete values that arise due to the extreme variability and unpredictable nature of unstructured data.

Data structure in the GDPR

A key point to note is that the GDPR only 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. Theoretically, firms without any intention to convert these unstructured data could sell them to third parties who would then presumably convert them to structured formats amenable to statistical analysis.

6.1.3 Data Integration

Data integration may be the dimension most affected by GDPR. By data integration, we mean both the addition of extra covariates (e.g., by extracting and creating new features based on textual data), and the linking of records across disparate datasets (Kenett and Shmueli, 2016). For example, predictions of the number of Kickstarter campaign backers may be improved by adding features extracted from the campaign's title. In this case, InfoQ would likely be increased by the integration of previously unstructured data into structured form. At the same time, data integration from unreliable or erroneous sources could serve to reduce InfoQ. Some commentators believe that the GDPR will have the effect of improving InfoQ because data sources will need to be vetted for GDPR compliance.

Data integration in the GDPR

As mentioned previously, the introduction of pseudonymization as standard practice under GDPR makes it much more difficult for data scientists to link records across different databases or datasets. In order to offset the potential privacy risks inherent in record-linking, the GDPR will likely have the side-effect of making various "privacy-preserving" data mining methods, such as

k-anonymity or differential privacy, more appealing. There are also new companies emerging that specialize in organizing firms' vast stores of personal data in ways that comply with data subjects' *right to be forgotten*. Because personal data may be distributed across various data warehouses in a company, actually implementing the *right to be forgotten* will be a formidable technical task for many organizations.

The GDPR's *accountability principle* will probably also reduce the number of third-party data vendors, leading firms to be much more cautious about the provenance of their personal data and any potential datasets that could combine with them to single out individuals. Facebook and Oracle, for example, are relying less on third-party data brokers who cannot prove GDPR compliance.

6.1.4 Temporal Relevance

Temporal relevance is concerned with the passage of time between data collection, analysis, and final deployment periods (Kenett and Shmueli, 2016). For example, if a project's goal was to predict whether a customer would churn in the next three months, it would likely have a higher InfoQ if the relevant datasets contained behavioral data collected in the past year, as opposed to data collected five years ago. Another temporal relevant issue is the speed at which the prediction is made: the longer it takes to generate the prediction (e.g. due to data transfer, analysis, and organizational delays), the more the relevance of the prediction diminishes.

Temporal relevance in the GDPR

Temporal relevance is directly related to GDPR through the introduction of data storage duration limits and the right to be forgotten. The GDPR stipulates that data subjects must be informed of the specific length of time that their data will need to be retained by the controller. This means that firms cannot hoard user data without having imminent plans to analyze them. Further, the specific project goal will help to determine how long personal data will need to be stored. The right to be forgotten refers to a right given to data subjects that allows them to request that data controllers delete any copies of the the data subject's data, along with any links to the original data (Article 17(2)).

6.1.5 Chronology of Data and Goal

This dimension of InfoQ is relevant to whether the data science project is looking backwards in order to explain or describe, or forwards in order to predict. In purely predictive studies, it is often the case that some variables included in a dataset cannot be used because they will not be available at the time predictions are needed. For example, if we wished to predict the number of Kickstarter campaign backers, we could not use a feature containing the amount of total funds collected, since that amount would only be known after the campaign had reached its conclusion. The InfoQ of such a dataset might thus be overstated since some features may be of no practical use for predictions. Additionally, in explanatory studies, the problem of endogeneity and reverse causation can serve to reduce InfoQ when variables that themselves are affected by the output cannot be included as explanatory variables (Kenett and Shmueli, 2016).

Chronology of data and goal in the GDPR

For the predictive case, the rights given to data subjects under GDPR can have an effect on the availability of certain behavioral data going forward. For example, with increased privacy options, some website users may choose not to consent to the collection of certain data though the initial models were trained under the assumption of indefinite collection and availability. In contrast, for explanatory and descriptive studies, the restriction of processing of certain kinds of sensitive personal data may mean that important causal factors, such as a user's race or political affiliation, cannot be used, thereby resulting in biased parameter estimates (Kenett and Shmueli, 2016).

6.1.6 Generalizability

The dimension of generalizability has both statistical and scientific aspects. Statistical generalizability refers to one's ability to generalize from a specific sample at hand to the target population; scientific generalizability (or, external validity) is concerned with applying a scientific model from one target population to another (Kenett and Shmueli, 2016). The InfoQ of a study attempting to use European users' BBD to predict US user behavior may be less than that of a similar study using data collected solely on US users. Similarly, a highly-accurate model of European customers' behavior will likely have reduced accuracy when applied to US-based users.

Generalizability in the GDPR

Because of the way in which explicit consent for processing is lionized under GDPR, consent bias may be a problem for data science projects whose goals include broader scientific generalization (as is the case in many academic-industry collaborations). If users who consent to the processing of their data are fundamentally different from users who do not consent, then the sample may only be representative of a subset of general users, and therefore not necessarily relevant to the population of interest. Further, if companies choose to run behavioral experiments on non-EU users due to the fact it is legal, cheaper and faster to collect their data, it may turn out that many of the resulting scientific models of behavior are not actually applicable to European users.

6.1.7 Operationalization

The dimension of operationalization has two major parts: construct operationalization, which is concerned with turning abstract theoretical concepts into something suitable for input into a statistical learning method; and action operationalization, which deals with the degree to which the findings of a data science project "lead to clear follow-up actions (Kenett and Shmueli, 2016)." Analyses which can more easily or accurately measure the variable(s) of interest and put the resulting predictions to use will have a higher InfoQ, all else equal. In other words, predictive models that recommend certain kinds of actions not allowed due to regulatory or legal constraints will be less useful (i.e., have a lower InfoQ) than those whose recommendations can more readily be implemented.

Operationalization in the GDPR

The notion of sensitive categories of personal data under GDPR will impact how certain theoretical and psychological (e.g., depression) constructs will be operationalized. For example, data from a survey asking about an individual's sexual orientation cannot be used except when explicit consent is given; yet, Kosinski et al. (2013) suggests sexual orientation can be inferred with nearly 88% accuracy simply by analyzing a user's Facebook Likes. The results of such studies have prompted researchers to develop techniques designed to counter or "cloak" such personal inferences (Chen et al., 2017). It remains to be seen whether the GDPR's prohibition on the processing of certain kinds of sensitive data will effectively be side-stepped by increasingly sophisticated inferential techniques.

In terms of action operationalization, the GDPR's focus on consent and on certain types of transparency and accountability rules may limit an organization's ability to meaningfully act on the results of a study. For example, data controllers, out of fear of being audited by a data protection authority, may choose not to pursue the results of some analyses if the data used in the analysis were not consented to. Likewise, transparency in the case of algorithmic profiling requires that data controllers provide data subjects with the possible consequences of such profiling. Firms whose possible actions have not been clearly related to the data subjects may not be able to actually pursue those goals.

6.1.8 Communication

The final dimension of InfoQ involves the ability to effectively communicate the results of the proposed project. The form of communication and media used will vary depending on the setting (academic or industry) and the audience (end user, manager, government agency, scientific

community, etc.). Data science projects with results that can be more clearly communicated to end users and data protection agencies will tend to have a higher InfoQ.

Communication in the GDPR

Under GDPR, data controllers have two major communicative responsibilities: the first is to the data subjects, and the second is to their respective data protection authority. Analyses relying on so-called "black-box" methods will have a harder time explaining the processes by which predictions were made to both users and government agencies. They will likely find less use in industry settings under GDPR, though they currently receive a lot of attention in the academic literature and media. Furthermore, more time will need to be spent at the experiment design and data collection phases in order to clearly spell out how the project's goals relate to the type of data collected. The GDPR places a strong emphasis on the notion of purpose limitation. Data scientists will thus need to be effective in communicating the goals and requirements of a predictive tasks to any data subjects whose personal data may be involved.

6.1.9 Assessing InfoQ

A key aspect to the operationalization of the InfoQ dimensions is that InfoQ assessments can be done. These evaluations can inform stakeholders of deficiencies in the proposed study during the initial design and collection phases, before major issues arise and become costly mistakes. Further, in the case of mutually exclusive data analysis projects, stakeholders may wish to postpone certain projects until the InfoQ assessment scores are more promising, such as when newer, cleaner data are collected, during the implementation of new data extraction pipelines, or when more unstructured data are converted into structured format.

Kenett and Shmueli (2016) present several ways in which InfoQ assessments might be carried out. The first is by using a simple ratings-based evaluation in which each of the eight dimensions are scored on a 1-5 scale and averaged using the geometric mean (Kenett and Shmueli, 2016). The advantage of such an approach is that it is easy and fast to perform. Another method is based on "scenario building," in which analysts might decide how "bad," "normal", and "good" versions of each of the eight dimensions would look and then compare them with the actual data at hand. Yet another approach to InfoQ assessment might be to conduct a pilot study in order to get an initial impression of the likely limitations and InfoQ problems a larger-scale project would encounter. Of course, exploratory data analysis–with a focus on discovering missing and suspicious values and generating summary statistics–can also augment formal InfoQ evaluation.

Finally, one particularly relevant InfoQ assessment technique is based on sensitivity analysis. Given the four InfoQ components, goal, data, analysis, and utility, change one component–the kind of analysis method used, for example–and keep the other three constant and see how InfoQ changes. This method of InfoQ assessment is especially useful under GDPR because firms may wish to know how the process of pseudonymization (or even anonymization) may impact the results of the study. It may be the case that aggregating certain kinds of categorical predictors can result in more user privacy and lead to better (or worse) predictive performance. Such an approach would also provide firms with evidence that they are complying with the GDPR's stipulation of data minimization and privacy by design.

Chapter 7

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

Now that we have introduced the InfoQ framework guiding our analysis, we will now look more closely at the main concepts and principles of the GDPR.[1] First we take a general view of the GDPR view based on the four InfoQ components of *goal*, *data*, *analysis*, and *utility*. This is visualized below in figure 7.1. Later, in section 7.5, we return to the eight dimensions to analyze more deeply the GDPR's effects on data scientists and behavioral researchers. Whenever possible, we use *italics* to denote key terms and discuss them in non-legal language. The key GDPR principles that data scientists should know about are given below in table 7.1, while a detailed list of formal definitions and exact GDPR articles and recitals are available in a GDPR glossary in Appendix D.

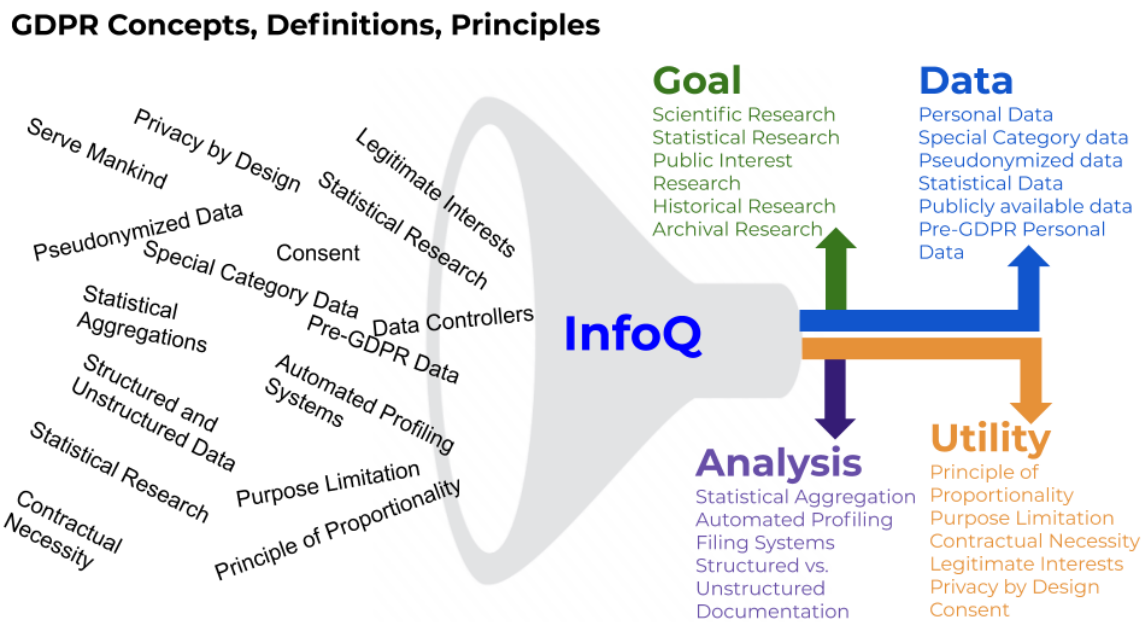

Figure 7.1: Example of key GDPR terms filtered the InfoQ data science framework

[1]Much of the material in this chapter was published in Greene et al. (2019). I would like to thank Galit Shmueli, Soumya Ray, and Jan Fell for their help in developing and presenting the ideas contained in this chapter.

Table 7.1: The Six GDPR Principles (Article 5, Recital 39)

Principle	Description
Lawfulness, fairness, and transparency	Personal data must be processed lawfully, fairly and transparently in relation to data subjects
Purpose limitation	Personal data can only be collected for specified, explicit and legitimate purposes (although further processing for the purposes of 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

7.1 Goal

Goal is the purpose for which the personal data 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. Organizations typically 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 their goal to "support scientific research including, for example, in the areas of cognitive science and education."[2] There are multiple terms in the GDPR that relate to goal (see Appendix). We discuss them as they relate to several guidelines.

Goals are set by *data controllers*

According to the GDPR, the *data controller* is the entity who determines the goal of the data collection or analysis. A controller can be a company, a university, or any entity holding data on natural persons. Setting the goal is what distinguishes the data controller from the *data processor*, who works on behalf of the controller.[3] For example, a company using its customer data for building a customer churn model would be simultaneously considered the data controller *and* the data processor; whereas if the modeling is outsourced to a consulting firm, the controller would still be the company itself but the processor would be the consulting firm.

[2] www.edx.org/edx-privacy-policy

[3] 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

Make your purpose transparent to users

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

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

Repurposing

The concepts of *legitimate interest*, *contractual necessity*, and *purpose limitation* are intimately connected. According to the principle of *purpose limitation* (see Table 7.1), if a bank obtains its customers' consent to collect and process their personal data for opening and running their bank accounts, (i.e., performance of a contract between the client and the business, or *contractual necessity*), then the same personal data cannot be used for purposes of direct marketing without the prior consent of the customers (see Appendix). Facebook, for example, cites contractual necessity (along with consent and legitimate interests), as its most basic legal grounds for personal data processing.[4] Accordingly, only processing "defined in the contract" for providing Facebook's service is permitted.

The GDPR's stipulation of purpose limitation encompasses many common processing activities that financial institutions have only recently undertaken in the era of big data, such as using analytics to target new potential customers, improving loan decisions and fraud detection (Alexander et al., 2017). Whereas the repurposing of personal data for preventing fraud constitutes a legitimate interest of the data controller (e.g., a bank or credit card issuer), repurposing data for marketing purposes may only fall under legitimate interests when there is a relevant and appropriate relationship between a financial institution and the targeted customer. In other words, the issue of legitimate interest in marketing seems to hinge on whether a data controller's goals are client retention or targeting new clients (with whom there is no prior 'relevant and appropriate relationship').

In light of this ambiguity surrounding direct marketing,[5] the safest course of action for data controllers would therefore be to limit processing of personal data only to those data subjects with whom some kind of documented contractual relationship currently exists. As an extra precaution, data controllers should aim for clearly-explained and defined purposes for processing and obtain explicit consent for such processing.

Goals vs. rights and freedoms:

A key aspect of the personal data processing exceptions for these goals is that they must be balanced against the rights and freedoms of the data subjects. Both industry and academic data scientists should be aware of the GDPR's exceptions to these rights since many of these exceptions are concerned with the types of analyses data scientists typically perform in their work. In its discussion of exceptions to these rights, the GDPR states, "in so far as such rights are likely to render impossible

[4] www.facebook.com/business/gdpr

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

or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes [these rights can be waived]" (Article 89). The rights specifically referred to are "the right of access," "the right to rectification", "the right to restriction of processing," and the "right to object." In the particular case of "archiving purposes in the public interest," the 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 - see Section 7.2) and notification requirements can be sidestepped.

Special Exemption Goals:

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

Scientific research: The GDPR intentionally carves out a broad swath of activities that could be construed as scientific research that includes "technological development," "demonstration," "fundamental research," "applied research," and "privately funded research." "Privately funded" research might be interpreted as applying to corporate research groups such as Microsoft Research or Facebook Research. Similarly, "technological development" may describe research by machine learning teams to improve algorithms at their companies. As an illustration, Facebook's revised Data Policy regarding "Product research and Development" reads, "We use the information we have to develop, test and improve our Products, including by conducting surveys and research, and testing and troubleshooting new products and features."[6]

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

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

Public interest and archiving: While the GDPR avoids defining exactly what "archiving" or "public interest" mean, it does list several of the data subject's rights that can be waived if they render "impossible or seriously impair" such research. Examples of "reasons of public interest" include "cases of international data exchange between competition authorities, tax or customs administrations, between financial supervisory authorities, between services competent for social security matters, or for public health, for example in the case of contact tracing for contagious diseases or in order to reduce and/or eliminate doping in sport." An example of such processing by a company on the grounds of public interest is Facebook's Data Privacy Policy[7] that states it processes personal data in order to "Research and innovate for social good" and that they "use the information... to conduct and support research and innovation on topics of general social welfare, technological advancement, public interest, health and well-being."

[6] www.facebook.com/about/privacy/update
[7] www.facebook.com/about/privacy/update

Historical purposes: Genealogical research is one of the few historical research goals mentioned in the GDPR text (Recital 160).[8] Further, regarding international transfers of personal data for scientific, statistical, and historical research purposes, the GDPR states that "the legitimate expectations of society for an increase of knowledge should be taken into consideration (Recital 113)." This research purpose then, would seem to permit the exempted processing of documents such as burial certificates and birth records, which may sometimes include personal data of living relatives.[9] It should be noted that many of the details for GDPR research exemptions are currently being worked out by the individual member states and that specific details regarding which data subject rights may be overridden may differ.[10]

7.2 Data

A dataset consists of measurements of entities. In the GDPR, the main entity of interest is the *data subject*, with a focus on measurements (variables) defined as *personal data* and *special category (sensitive) data*. Data scientists should be aware of the following key definitions (see also Appendix):
Data subject The GDPR's definition of *data subject* is "a [living], identifiable natural person." This differs from the definition of a *human subject* used by ethics boards in academia (mandated by the Common Rule in the US), defined as "a living individual about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information" (Tene and Polonetsky, 2016). The main difference between the academic and GDPR definitions is that the GDPR's *data subject* does not require any interaction or intervention by the data controller with the data subject.
Personal, sensitive, pseudonymized, and statistical data The key data measurements in GDPR are *personal data, special category (sensitive personal) data, pseudonymized data*, and *statistical data*. Due to their critical importance, we describe each:

1. **Personal data,** or Personally Identifiable Information (PII), specify a wide range of information that might identify a natural person in terms of his or her physical, physiological, genetic, mental, economic, cultural or social identity. We note that IP addresses and cookies can be considered PII because "[they] may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them."
2. **Special category (sensitive personal) data** are categories of personal data that reveal an individual's belonging to some "special category" or group. The GDPR provides the list of categories in Article 9. Special category data 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.[11] By and large, processing of special category data is prohibited under the GDPR, unless users give explicit consent to such processing (Recital 51).
3. **Pseudonymized data** are a subset of personal data that have had individual identifiers removed, so that it is not reasonably likely for a data processor to be able to "single out" a specific person. The GDPR states repeatedly that pseudonymizing personal data should be the foundation of a data controller's collection and storage practices.
4. **Statistical data** are synonymous with aggregated data. Statistical data are used to infer traits about groups of people, rather than specific individuals.

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

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

[9] https://www.freeukgenealogy.org.uk/news/2018/05/22/gdpr/

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

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

Filing systems Data are typically housed in spreadsheets or a database (including a distributed framework such as Hadoop) that must be accessed by the data scientist or data engineer. The GDPR refers to these means of data storage as the *filing system*[12] defined as a "structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis." The GDPR requirements apply to all processing of personal data that "form part of a filing system or are *intended* to form part of a filing system" (emphasis ours). The subtle implication is that the GDPR applies even if a company merely intends to convert unstructured data to structured data, thereby creating a *filing system*.

7.3 Analysis

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

Types of analyses that fall under *data processing*

Data scientists analyze personal data using a range of methods, from computing simple summaries and aggregations to sophisticated statistical models and machine learning algorithms, including text mining and network analytics. Analyses and modeling are used for training algorithms and fitting a models, as well as for deployment to new data subjects, such as providing recommendations or generating predictions for new users. Data analysis can range from manual, to semi-automated, to fully automated, as in the case of a company using 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).[13] This entire range of activities falls under *data processing*.

Identifying the *data processor*

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

7.4 Utility

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

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

[12]The term *filing system* is perhaps a reference to earlier paper-based document storage systems
[13]Leave it to the experts, *The Economist*, volume 426 Number 9085, March 31, 2018

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

The fundamental right to privacy

The purpose in adopting the GDPR over the 1995 Directive was to "ensure a consistent and high level of protection of natural persons and to remove the obstacles to flows of personal data within the Union." This recital emphasizes the seriousness with which an individual's fundamental right to privacy must be respected under EU law. We note the sharp contrast between the EU and the USA's approach to data privacy: Weiss and Archick (2016) aptly summarize this distinction by saying that in the USA, "collecting and 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." Another way of generalizing the difference is that in the USA, consent to processing of personal data is implied unless data subjects opt-out ("opt-out" model); whereas in the EU, no consent is to be assumed unless data subjects explicitly opt-in ("opt-in" model). It is difficult to understate the impact this difference in underlying philosophy has had on the evolution of data processing policy in the US and the EU. This difference has led at least one legal scholar to argue that the EU's general prohibition of automated processing (since the 1995 Directive) of personal data has "deterred entrepreneurs and investors with overbroad and rigid laws" and may help explain why many of the leading IT, social media, and cloud services originate from the USA (Determann, 2016).

The principle of proportionality

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

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

Legitimate interest

The important yet vague concept of *legitimate interest* similarly rests on the complex balance of commercial utility with respect for fundamental privacy rights. As a grounds for processing, the implicit expectation is that the economic benefits of processing to the data controller (or to a third party) outweigh any potential harm done to a data subject and thus the controller has a "legitimate [economic] interest" in processing the data. This is the so-called "balancing test." After all, according

to the Charter of Fundamental Rights of the European Union, data controllers have the "freedom to conduct a business," and the processing of personal data may be an inherent part of the business, as in an ad network, for example (Borgesius, 2015). Yet at the same time, the same Charter bestows fundamental rights to privacy and data protection to data subjects. How these competing rights should be balanced is not obvious. Consequently, this constant tension between human rights and economic gain is a major motif in the GDPR. As the ostensible purpose of processing personal data is to "serve mankind," any arbitrary processing which could potentially violate a right to privacy would not pass the *proportionality (i.e., balancing) test* unless it could be shown to have significant social or business value.

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

After our discussion of how the GDPR's principles and concepts relate to the InfoQ notions of *goal*, *data*, *analysis*, and *utility*, we now wish to analyze more concretely how GDPR will impact data scientists. Several generic data science workflows have been developed over the years, including CRISP-DM by IBM and SEMMA by SAS. However, we have tailored a workflow that highlights the specific issues encountered by many industry and academic data scientists using BBD. Figure 7.2 displays this workflow, from data collection to communication. We have added the steps "Sharing data" and "Generalization" to reflect the increasing growth and importance of industry-academia collaborations in the social sciences and recent academic controversies surrounding the replicability of many BBD experimental results. Nevertheless, we believe that this model workflow will be relevant to a broad swath of both researchers and practitioners in the new data regulation landscape.

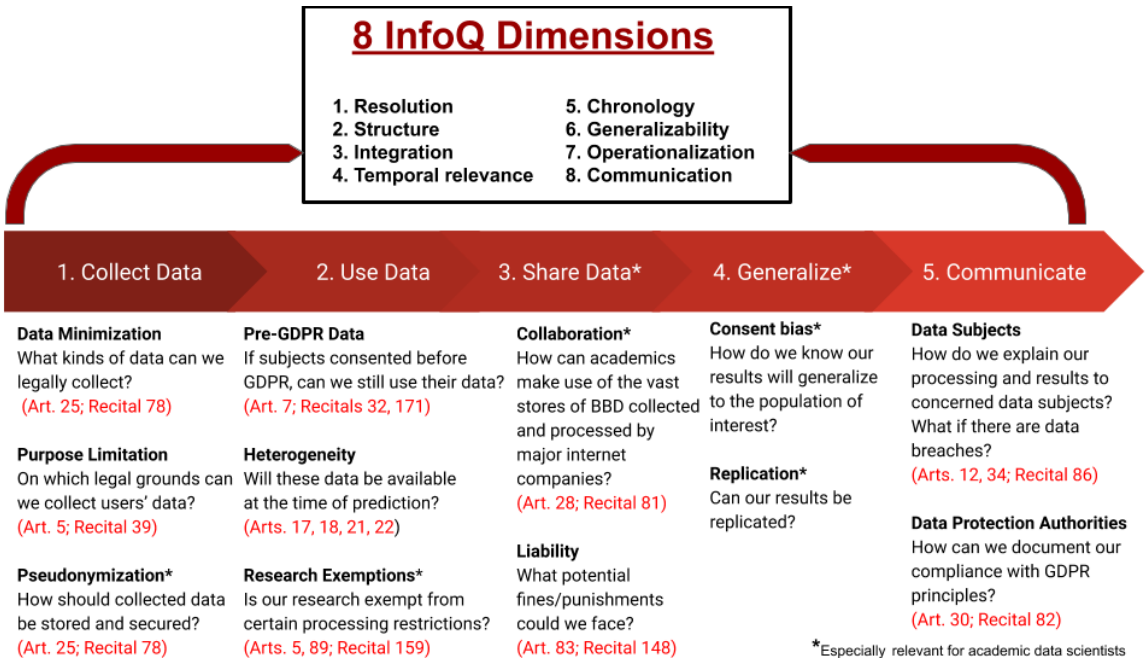

Figure 7.2: An example data science workflow under the GDPR, tailored to BBD usage

In order to evaluate the impact of the GDPR in a principled way, we again use the InfoQ framework detailed above in section 6. There, we broke down each of the eight dimensions and briefly hinted at how they relate to the principles and requirements of the GDPR. In this section, we now imagine a typical data science workflow and use these dimensions to assess the impact of the GDPR on data scientists' routines and approaches.

7.6 Collecting Data: Pre and Post

7.6.1 Pre-Collection: Data minimization and purpose limitation

The GDPR principle of *data minimization* dictates that personal data must be adequate, relevant and limited to what is necessary for processing (see Table 7.1). Consequently, companies will need to carefully assess the resolution of personal data they collect and justify why it is necessary for achieving their stated goal. The GDPR also puts forth the principle of *purpose limitation*, which states that personal data can only be collected for specified, explicit and legitimate purposes (Article 5). It is not enough for companies to say, for example, that they need to process personal data in order to provide a service–they must specify *how* and *why* such processing is necessary for provision of the service. Since new goals require new consent requests from users, repurposing personal data from one project to another will no longer be a viable option, even if repurposing personal data would seem to adhere to the intentions behind the data minimization principle. In other words, purpose limitation prohibits companies from collecting a small amount of personal data (the minimization principle) and then reusing it for unspecified secondary purposes.

For example, a company wishing to build a personalized pricing model must only collect and process personal data that are strictly relevant to achieving this goal. These data may include past purchase histories or website browsing times, but should not contain data relating to one's ethnicity (e.g., an "Asian-sounding name") or IP address, even if these turn out to be useful predictors of one's willingness to pay a certain price for an item (Steppe, 2017). Building a recommender system would proceed similarly. After choosing a suitable legal basis of processing–most likely either legitimate interest or consent in the case of a recommender system–data scientists would be limited to collecting and processing only personal data necessary for generating useful recommendations.

Not surprisingly, some companies, such as Airbnb and Uber, have successfully circumvented these limitations by building "smart pricing" algorithms that are based on anonymized (aggregated or market-driven) or non-personal data, usually in the form of event-based or object-related–rather than natural person-related–data. This allows, for example, Uber to make dynamic pricing decisions for individuals without needing the individual's name (Steppe, 2017). Such examples seem to validate the GDPR's more balanced approach to the competing interests of business and personal privacy. They suggest that it is indeed possible to develop personalized predictive models that respect users' privacy and boost company profits at the same time. Consequently, creative data scientists who can develop reasonably-performing personalized predictive models from non-personal or aggregated data may become highly sought after in a post GDPR world. It is also likely that at the planning stages of new analytics projects, more time will be devoted to thinking about how the goals of the project can be met without requiring the use of personal data. After all, any losses in predictive performance may be more than compensated for by savings in GDPR compliance and documentation costs.

The *data minimization* principle will likely increase the importance of statistical power calculations for A/B tests that involve personal data (e.g., in usability research or customer journey studies). For example, suppose a data scientist is tasked with finding a minimum sample size for estimating the improvement in completion rate for a user interface redesign compared to an existing design, using a 90% confidence level and 80% power. For an hypothesized 80% historical completion rate and required 20% improvement difference in completion rates between the A/B versions (a goal set by the data controller), such an analysis would require behavioral data from approximately 49 users in each group. By reducing the power requirement from 80% to 50%, the sample size can be reduced to approximately 14 per group (Sauro and Lewis, 2012). In order to follow the principle of data minimization, data scientists will thus need to carefully consider the necessary statistical power and confidence levels needed for their particular business goals; otherwise, they risk collecting more data than needed for testing their hypotheses at their required confidence level.[14] And under GDPR, companies will be required to document and justify their processing decisions, so reasoned power calculations in A/B testing may be viewed as constituting proof of compliance with the data minimization principle.

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

At the same time, however, data minimization efforts and privacy-preserving techniques may conflict with one another. Imagine data scientists at an online dating platform are asked to determine whether the opt-in rates for personal data processing are different for users from different countries (or even ethnic groups) at a given statistical significance level. In this situation, issues of aggregation and minimization arise: the data scientist must consider the experiment's sample size as well as potentially identity-revealing counts of data collected. If the power calculation indicates a relatively small sample is sufficient to detect a difference of a predetermined magnitude with some specified confidence level, then the probability of getting unevenly distributed counts among the different groups is increased–thereby making it much easier to single out individuals by their discordant behavior.[15]

This example is important because it illustrates how the GDPR requires data scientists to have a firm grasp on the interplay between the technical details of A/B testing and fundamental GDPR principles. When deciding upon the particular testing goal, project stakeholders will need to ask themselves questions such as,"How big of an effect size do we expect to see?" and, "How narrow must our confidence intervals be?" Companies running A/B tests with millions or billions of users will be able to detect extremely small effects with high power, but the question under GDPR is: "Are these differences enough to justify the increased risk of re-identification?" It is precisely in these types of situations where the *principle of proportionality* arises. In essence, A/B testing under the GDPR should be done using a principle similar to Occam's razor: if sufficient power can be achieved with a smaller sample size, then the GDPR dictates that the smaller sample should be used, unless a data controller can prove the "necessity" of such large-scale testing. In the case of Facebook's controversial emotional contagion experiment (Kramer et al., 2014), where a massive online behavioral experiment conducted on over 600,000 users' news feeds found an extremely small effect size, the controllers would need to justify the scientific contribution of such a small effect and why it outweighs the potential privacy harms inherent in large-scale data processing.

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

7.6.2 Post-collection: pseudonymization

A key addition in the GDPR over the 1995 Directive is the introduction of the security practice of *pseudonymization*, a method for reducing the chance that any particular data value can be "attributed to a specific data subject without the use of additional information," provided that this "additional information" is kept separately and securely.[16] Examples of unique identifiers that might single out a data subject include names, tracking cookies, email addresses, user names, or IP addresses (including dynamic IP addresses), among others. Note that *pseudonymization* is different from *anonymization*, which aims to make the process of re-identifying particular data values with specific individuals practically impossible.

Pseudonymization can affect the resolution of data available to data scientists in a few ways. First, completely removing individual identifiers from datasets impacts the ability of data scientists and researchers to make predictions at the individual level at the deployment stage. Second, data scientists will need to consider which measurements (attributes) or combinations of measurements

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

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

might be used, directly or indirectly, to identify a natural person in the data set. In some cases, aggregation might be a useful approach (e.g., replacing data on a user's individual sessions with daily aggregates). The extent of this process should be based on whether analysts possess "means reasonably likely to be used [to single out individuals]" in the data (Article 26). In other words, each organization will need different privacy protocols depending on the technical means of analysis available to the data scientists, the security practices of the organization, and the intrinsic motivations for the analysis.

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

The pseudonymization requirement is even harsher for companies with small or declining user bases. This is because identification of individuals becomes easier in aggregated data as the number of aggregated units decreases (i.e., becomes sparser), or as the number of tabular aggregations increases (i.e., more combinations of tables with different categories). As Lowthian and Ritchie (2017) note, aggregated tables of group membership 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. The US Census Bureau, for instance, has taken to adding statistical noise to its aggregated statistics to reduce the possibility of singling out individuals in seemingly "anonymized" summary tables.[17] An extra concern is when the groups by which the data are aggregated are sensitive categories of data, such as ethnic origin, religion, or sexual orientation. Thus, under the GDPR, data scientists may benefit from becoming proficient in various privacy-preserving techniques, such as differential privacy or k-anonymity, in order to reduce the likelihood of re-identification (see, e.g., Dwork and Roth, 2014).

7.6.3 The data environment

Determining whether data are *pseudonymized* or *anonymized* requires considering the *data environment*. If one possesses a "means reasonably likely to be used" to 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 to potentially single out individual users. Mourby et al. (2018b) state that according to the UK Anonymisation Network there are four main components of a researcher's data environment: *other data*, *agency*, *governance processes*, and *data infrastructure*. *Other data* refers to databases, public registers, or even social media profiles the analyst may have access to. These other data sources are important because they constitute a large portion of re-identification risk. *Agency* considers the question, "What incentives or motivations might the analyst have in re-identifying a data subject?" *Governance processes* are the formal policies and procedures that control how the data are accessed, by whom, and for how long. Finally, *data infrastructure* could be the actual hardware and software used in analyzing the data. Some data environments may have password-protected access or require encrypted flash drives, for example, in order to ensure the security of personal data, thereby satisfying the principle of *data security*.

The incentives of data scientists to identify individuals vary vastly and thus pseudonymization will affect them differently. Data scientists working for data brokers or marketing firms have strong economic motivation to identify specific individuals. Their ostensible goal is to map online behavior to offline purchase behavior through first, second, and third-party data integration. The data broker LiveRamp (an offshoot of major data broker Acxiom), for instance, offers the product "IdentityLink" that allows advertisers a single, "omnichannel view" of the consumer. Acxiom claims to permit the identification of specific consumers across "thousands of offline and digital channels

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

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.[18] 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 he 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.

Data brokers and data-savvy marketers are not the only analysts who might have powerful incentives to single out individuals and thereby turn essentially anonymized data into personal data. After Facebook publicly revealed that Russian operatives had been directed to influence the 2016 US presidential election on its platform and others, [19] several other politically motivated operations were uncovered, one of which involved the Saudi Arabian government. The NY Times reported that a Twitter employee had been promoted to a position that allowed him access to the personal information of users, including their IP addresses and phone numbers, which could then be used to link Tweets to specific devices and single out Saudi government detractors for punishment.[20] The lesson to be learned is that privileged analysts with powerful political and financial incentives can easily circumvent the protections pseudonymization was designed to introduce. Therefore under GDPR, pseudonymization–or related techniques, such as k-anonymity or differential privacy– is a necessary, but not quite sufficient step for securing personal data.

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

7.7 Using Data

7.7.1 Reconsent of pre-GDPR data

A major issue is the status of data collected pre-GDPR and its effect on later data analysis and use. It appears as though companies can continue to use personal data legally obtained prior to the GDPR, with one major caveat. Recital 171 states that "it is not necessary for the data subject to give his or her consent again if the manner in which the consent has been given is in line with the conditions of this Regulation, so as to allow the controller to continue such processing after the date of application of this Regulation." In other words, as long as the original (pre-GDPR) consent adhered to GDPR standards and that proper documentation of the consent exists, then no special actions must be taken on part of the data controller. Nevertheless, in the lead-up to the GDPR, many websites and online platforms asked users to reconsent to the processing of their personal data. Either companies did not understand Recital 171 or they never had proper consent in the first place. In any case, this was probably the simplest and safest legal route for companies to retain pre-GDPR user data, but it raises the question of what happens to personal data that the data subject does not reconsent to for processing, or chooses to have erased. Some companies have stated they will continue to use such data, albeit in aggregated form. For example, Kaggle notes in their revised privacy policy, "We may use aggregated, anonymized data that we derived from your personal information before you deleted it, but not in a manner that incorporates any of your personal information or would identify you personally.[21]" Such a tactic would be legal under GDPR because the Regulation only applies to personal data–data that could reasonably be used to identify a natural, living person.

[18] Meet LiveRamp IdentityLink, lp.liveramp.com/meet-liveramp-identitylink.html
[19] www.theguardian.com/technology/2017/oct/30/facebook-russia-fake-accounts-126-million
[20] www.nytimes.com/2018/10/20/us/politics/saudi-image-campaign-twitter.html
[21] www.kaggle.com/privacy

Anonymized data, *ipso facto*, cannot be linked to a specific individual and therefore is outside the scope of the GDPR.

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

Similarly, companies storing large amounts of customer personal data in databases will need to regularly "cleanse" them to make sure customer data are either accurate and updated or otherwise deleted, in accordance with the principle of *data accuracy*. A director at a data consulting firm remarked, "If you've got out-of-date data and you don't have a solid cleanse process, your ROI is going to be significantly impacted."[22] On top of this, companies will need to have protocols in place for dealing with the GDPR-granted "right to rectification" that data subjects possess regarding the accuracy of their personal data. The incentives brought on by GDPR may therefore have unintended positive effects on a firm's bottom line, especially for firms that struggle with data inventory, quality, and retention issues.

7.7.2 Data availability

Because the GDPR's reach is global (unlike the previous Directive), it will ostensibly affect firms "offering goods or services" to, or "monitoring the behavior" of, data subjects in the EU (Article 3). Due to this widened territorial scope, some data scientists and researchers, who had never previously concerned themselves with the details surrounding the use of personal data from EU data subjects, may find that variables previously available to them are no longer being collected by certain platforms, or that the ability to process them has been restricted or removed. In the wake of GDPR, Twitter, for instance, made several changes to its popular public API including making time zone fields private and removing background profile images of users.[23] Consequently, researchers building predictive models using these features would need to either remove the features from the affected models, find creative proxies for these same predictors, or perhaps abandon the models completely, depending on the feature importance.

More generally, "special categories" of sensitive personal data may now be off-limits to some non-EU data scientists. Non-EU-based data scientists may have plausibly assumed that any European data protection laws would not or could not apply to them. A concrete example of this is found in data mining research that aims to trace public political opinions of social media users, where political opinions are considered *sensitive personal data* under GDPR. In such situations, researchers might inadvertently process personal data of EU-residing data subjects. The authors of a highly-cited data mining article using the Twitter API admit that, "Most Twitter users appear to live in the U.S., but we made no systematic attempt to identify user locations or even message language, though our analysis technique should largely ignore non-English messages" (O'Connor et al., 2010). It is probable that the authors inadvertently processed the sensitive personal data of at least some EU data subjects. In the authors' defense, however, Article 9(e) of the GDPR states that the GDPR does not apply when personal data are "manifestly made public by the data subject." However, given that use of a public API does require some technical expertise, and that it is not obvious to most Twitter users that their political statements may be mined by researchers, it could plausibly be argued that these political opinions now constitute sensitive personal data. For data scientists using political opinions as predictors in a statistical model, this could present a problem because the required predictor columns would no longer be available at the time of prediction, post-GDPR.

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

[23] www.twittercommunity.com/t/upcoming-changes-to-the-developer-platform/104603

7.7.3 Data storage and duration limits

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

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

7.7.4 Data subject heterogeneity

A new source of uncertainty in post-GDPR data is due to the enhanced privacy settings that websites may provide under GDPR that were not offered pre-GDPR. Given the wider variability in user privacy preferences and the ease with which they can be changed, data scientists will need to grapple with larger within-subjects and between-subjects heterogeneity. For example, changes to a user's privacy settings may result in many missing values for a single variable during a specific period of time. At the same time, different users may consent to the collection of different aspects of their online behavior. Ebay's post-GDPR privacy policy, for instance, lists at least four areas where users will have a choice: marketing, communication preferences, advertising, and staying signed in.[24] Increased missingness of data will clearly have a negative impact on a dataset's information quality and we may thus see an increased need for reliable imputation methods.

7.7.5 Choice of algorithms and models

Due to the new issues of data availability, storage duration limits, and status of pre-GDPR data, the choice of algorithms and models used by data scientists will require another layer of consideration. First, the principle of *lawfulness, fairness and transparency* makes transparent models[25] such as regression models and classification and regression trees advantageous over blackbox models, such as deep neural nets, in personalized applications. Transparency requires the ability to explain why a model has produced a certain prediction or recommendation for a data subject. As a result, the GDPR's transparency requirements may shift current machine learning practices towards those more commonly used in the highly-regulated financial services industry. In credit scoring, for example, Basel II regulations require that any deployed credit scoring models be highly repeatable, transparent,

[24] www.ebayinc.com/our-company/privacy-center/privacy-notice/

[25] In the book *Weapons of Math Destruction*, O'Neil (2016) highlights three features needed to make an algorithm a "weapon of math destruction:" Opacity, scale, and damage.

and auditable (Saddiqi, 2017). In this kind of regulatory environment, predictions made by deep neural nets are much harder to explain to concerned data subjects and scrutinizing data protection authorities. As a result, simpler approaches, such as logistic regression and decision trees, are often employed by practitioners.

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

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

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

7.8 Sharing data

The GDPR is likely to significantly change the way companies share data with one another and with academic researchers. Though modern internet companies initially collected BBD for the purpose of improving services and making better decisions, their massive stores of user-based BBD have the "potential to advance social scientific discovery, increase social good, and. . . ameliorate important problems afflicting human societies (King and Persily, 2018)." Whereas previously most academics had to either collect their own data or pay for datasets relevant to their research, nowadays more and more of the most socially-valuable behavioral data is being collected through social media and e-commerce platforms, such as those offered by Facebook, Google, and Amazon. What is most startling, however, is that although in absolute numbers the amount of BBD has increased exponentially, the proportion of such data accessible by researchers is "[far smaller] than at any time in history" (King and Persily, 2018). In spite of this trend, collaborations between academic data scientists and industry remain more important than ever given the vast array of potential research questions that could be addressed using industry-sourced BBD.

7.8.1 Legal liability under GDPR

Prior to the GDPR, third-party data sharing through mutual data sharing agreements or through purchasing was extremely popular among large and small companies. However, the GDPR's stipulation that controllers, and by extension their data processors, can be held liable for damages caused by illegal processing of personal data, has made data controllers increasingly hesitant to

share any data that might contain potentially personally identifiable information. As a result, we are already seeing some companies eschew third-party data sharing agreements they have previously participated in, such as the recent announcement by Facebook about winding down its "Partner Categories," a feature that allowed data brokers such as Experian and Oracle to use their own reams of consumer information to target social network users. Marketing companies in particular have been deeply impacted by the way the GDPR distributes liability through the entire data collection and processing phases. One marketing blog reports that 'only 20 percent of 255 brand marketers ... are confident that their mar-tech vendors [will not] expose them to legal risks if [the vendors] are not GDPR compliant."[26] There is thus some reluctance within industry to continue relying on third-party data brokers for access to external datasets, since the personal data contained within them may not have been obtained according to the principles of GDPR and would put them at legal risk.

7.8.2 Data access divides

Fear of regulatory scrutiny may also have spillover effects on BBD-focused academic research. As mentioned by King and Persily (2018), if the relative proportion of data available to academics for research continues to decrease, we may begin to see disparities in access to company data on two distinct fronts. On the corporate front, the Future of Privacy Forum (2017) found the two biggest obstacles to corporate-academic data sharing were possible risks of personal re-identification and intellectual property disclosure.[27] Consequently, only trusted researchers from elite universities with close ties to corporations might be given access to corporate data, thereby reducing the opportunities for socially purposeful research to be done by those outside of the corporate trust network. This could impact the ability of other academics to reproduce important experimental results, for example. The other side of the data access divide is related to the types of companies that can afford to collect and process BBD after the GDPR. There is already some research showing that the GDPR has benefited large companies at the expense of smaller ones.[28] This is likely because only large companies can afford the extensive compliance costs required by GDPR. These effects may be particularly pronounced in the ad-tracking industry because cookies are considered to be personal data. If this trend continues, then it may further exacerbate the monopoly companies like Facebook, Google, Amazon, and Apple have on BBD. These companies could then become the de facto "gatekeepers" of academic-industry BBD-based research. Such an arrangement could hamper scientific independence, especially as it is not uncommon for companies to ask for pre-publication approval or patent rights (King and Persily, 2018).

The takeaway here is that academic researchers keen on using corporate data need to start developing, as early as possible, symbiotic relationships with corporate data providers, and they should not be surprised if more and more of their data comes from the coffers of an increasingly small group of internet companies.[29] 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 in order to minimize legal exposure due to data-sharing with academic researchers. For now, the GDPR's introduction of Binding Corporate Rules (BCRs), may provide a partial solution to these issues of data sharing, particularly when international transfers of personal data are required.

[26] Facebook to stop allowing data brokers such as Experian to target users, The Guardian, Mar 29, 2018

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

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

[29] Intermediary professional organizations recognized by both industry and academia might also serve this purpose.

7.9 Generalization

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

7.9.1 GDPR and consent bias

This concern about generalization, also known as *consent bias*, is further bolstered by the current debate in the scientific and statistical communities about whether requiring explicit consent from data subjects biases the results due to systematic differences in the way that data subjects are selected (Junghans and Jones, 2007). Under the GDPR, the problem of consent bias may be exacerbated due to differential privacy standards for EU and non-EU data subjects. Indeed, privacy-savvy users will likely be underrepresented in studies because they will not consent to their data being used for unspecified research purposes. As evidence of the possibility of such a bias, it has been reported that as of November 2018, only about one-third of US internet users have opted-in to the processing of their personal data, and as many as 17% have completely opted-out.[30] The percentage of opt-outs for European users is almost guaranteed to be even higher. If this trend continues, data scientists making statistical inferences based on users' data, such as is commonly seen in A/B testing in industry or BBD-based empirical studies published in scientific journals, may end up having a significantly more accurate picture of non-EU users than EU users.

Facebook, for instance, has already publicly stated that for non-EU users in Asia, Latin America, and Africa, US privacy guidelines will apply.[31] Yet other big names in the BBD arena, such as Microsoft, have declared that they will apply GDPR protections globally.[32] Given the extra costs of documentation and compliance of personal data processing and collection under the GDPR, it is unclear how other major BBD data controllers, such as Amazon or Google, will proceed. It is telling that several months after GDPR went into effect, there are still major media publishers such as the Los Angeles Times, Chicago Tribune, and San Diego Union-Tribune, which are blocking EU-based users from accessing content out of fear of non-compliance.[33]

If differential data processing pipelines for EU and non-EU data subjects do indeed become the norm, BBD research may then begin to resemble the ethically-dubious way in which HIV vaccines were trialed in developing nations in the early 1990s. Research ethicist Iltis (2006) notes that critics of such trials worried that the research benefits would only go to those in rich, Western countries and that experimenters were taking advantage of "low-wage" African research subjects. Similarly, non-EU data subjects could become the new, preferred "low-privacy" BBD research subjects because of the relative ease and low cost with which their personal data could be processed. Scientific generalization would be reduced because any scientific models based on the non-EU users may not apply to EU populations for various cultural and geographical reasons. Furthermore, a key aspect

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

[31] www.reuters.com/article/us-facebook-privacy-eu-exclusive/exclusive-facebook-to-put-1-5-billion-users-out-of-reach-of-new-eu-privacy-law-idUSKBN1HQ00P

[32] www.theverge.com/2018/5/24/17388206/microsoft-expand-data-privacy-tools-gdpr-eu

[33] www.theguardian.com/technology/2018/may/25/gdpr-us-based-news-websites-eu-internet-users-la-times

of academic human subjects research, the Belmont principle of "justice," which addresses the fair distribution benefits, risks, and costs among experimental subjects, would also be violated. Such a situation would seem to put collaborative industry-academic BBD research at odds with traditional academic human subjects research.

7.9.2 Concerns of scientific reproducibility

The GDPR's stringent consent standards for EU data subjects could also negatively affect the related notion of *scientific reproducibility*, which is concerned with the ability to "recreate scientific conclusions and insights" from previous studies (Kenett and Shmueli, 2015). 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 (Future of Privacy Forum, 2017). For example, what if a data subject initially consents to his data being processed for statistical research purposes, but then changes his mind after the processing but prior to the analysis? It would not be possible to completely replicate the analysis if certain data subjects' personal data were removed in the time between different replication attempts.

One potential solution proposed by King and Persily (2018) is a system in which all funded academic-industry research would follow a "replication standard," whereby each dataset used in research would come with a "universal numeric fingerprint" that would persist even if the format of the data were to change. Further, all computer code, methodological details, and metadata would be publicly available on the internet, while the actual data needed for the research would be stored internally at the company and accessible to academics. Such a system could reduce the chance of an inadvertent data breach due to the sharing of personal data to researchers for replication purposes. Currently under the GDPR, however, if companies wish to avoid liability for potential data breaches, the simplest and most common solution is to refuse to share the data used by the original study, thereby reducing the scientific reproducibility of behavioral research.

7.10 Communication

7.10.1 Communication with data subjects

The GDPR lays out the major duties regarding the types of interactions between data controllers and data subjects, many of them related to using clear and simple language to explain the grounds of processing, and detailing the data subject's right to have their information provided upon request. Further, data controllers must be able to clearly explain–in a non-technical way– to data subjects, which personal data are being collected, why their data are being collected, and for what specific purposes or goal(s). For example, if there is to be communication with a child, the language used needs to be appropriate for a child (children can consent to the processing of their personal data generally starting at age 16),[34] and a clear "opt-out" option to the collection and processing of personal data should also be available. If data subjects do choose to opt-in, however, their "right to access" this information should not be excessively burdensome, either (Article 12 states that this information should be"easily accessible" by data subjects).[35] There are already reported cases where concerned data subjects requested information about the personal data stored about them only to receive an automated message requesting the data subject provide detailed information such as all public IP addresses, invoice IDs for purchases, credit card numbers used in purchases, dates of logins, names of user accounts, and much more.[36] It seems reasonable to assume that for older or less technically-inclined data subjects, providing this information may not be feasible. Companies may

[34] https://www.twobirds.com/~/media/pdfs/gdpr-pdfs/24--guide-to-the-gdpr--children.pdf?la=en

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

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

therefore want to draft multiple versions of their privacy policies with language specially crafted for children and the elderly, in order to conform to the GDPR principle of transparency described in Article 12. For the average data subject, however, privacy policies containing information in short, clear sections such as, "What data we collect about you," "How your personal data is used," "How your information is shared," and importantly, contact information (typically including an email address) for data privacy concerns may constitute adequate proof of transparency.

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," and a choice to opt-out. As mentioned earlier in this Article, organizations and their data scientists will need to ensure that decisions based on complex algorithms can be adequately understood by non-technical users (and regulatory agencies). Data scientists will thus likely need to include communicability into their choice of algorithms and their documentation, should data subjects exercise their rights under GDPR.

7.10.2 Communication with data protection authorities

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

The GDPR's introduction of mandatory data breach reporting periods is highly relevant given the recent spate of reports of massive data breaches from Facebook and Google+. According to Recital 85 of the Regulation, companies must report such a breach within 72 hours to authorities, and also to the data subjects "without undue delay." The October 2018 data breach of the access tokens of nearly 30 million Facebook users serves as a prime example of how communication with both DPAs and end-users will change under GDPR.[37] Since data scientists are intimately familiar with the company's stored personal data, they will need to have a clear understanding of their role in the data controller's obligation for documentation, reporting, and communication with both authorities and users in case of such breaches. The bottom line is that under the GDPR effective data scientists will need to possess strong communication skills and be comfortable interacting with diverse audiences that include data subjects, management, the data protection authorities, and other departments involved in collecting and analyzing personal data. Data scientists will also need to collaborate with other stakeholders to systematically document processing in order to demonstrate compliance with the GDPR principles outlined in this Article. We conclude by noting that the communication skills of data scientists with less or non-technical audiences are therefore likely to become even more important in the future.

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

Chapter 8

Conclusion & Future Work

The GDPR is a long and complex legal document that reflects Europe's unique social, economic, political, and legal context. The present work has attempted to boil this complexity into something more easily digestible to the concerned citizen, data scientist, and data science manager and executive. In order to do this, Part I looked at the justifications for the GDPR and organized various academic responses to it based on their economic, business, political, legal, or technical implications. Later in Part II, we considered how the GDPR fit into a broader analytics strategy for a multinational firm with presence in both the US and Europe. There, we relied on the three-domain framework used in the study of corporate social responsibility to analyze the economic, legal, and ethical implications of the GDPR's Binding Corporate Rules. Finally, in Part III, we employed the data science framework InfoQ to examine the impact of the GDPR on the work of data scientists.

In section 4.2.1 we saw that there are in fact good reasons to be skeptical about the European approach to the regulation of personal data processing and ownership. Zarsky (2016) presents several conceptual, legal, and technical arguments against the success of the GDPR: in short, the nature of big data processing is incompatible with the GDPR's goals of ensuring citizens' privacy. Further, security researchers have raised concerns about how to verify data subjects' identity when respecting their *right to be forgotten* (Politou et al., 2018). Experts have also singled out data subjects' rights to explanation and human intervention as potentially problematic. Additionally, concerned citizens worry that data divides between affluent and non-affluent citizens might be made worse, since affluent citizens tend to better understand their data rights and can afford to opt-out. And in terms of business impact, there is already evidence that big corporations have benefited disproportionately from the GDPR. Smaller firms, for example, cannot afford to completely overhaul their data collection and processing pipelines. Finally, as King and Persily (2018) point out, it is still not clear how the GDPR will affect industry-academia research partnerships. More time and research will be needed to ascertain whether these issues present insurmountable regulatory hurdles.

Despite these issues, the GDPR represents the first attempt by a major political body to grapple with the broader social, political, legal consequences of digital technology. It is also the first and most notable attempt by a modern democracy to consciously direct the impact and application of new and transformative technologies. Finally, the GDPR is the first theoretically international personal data regulation (i.e., with jurisdiction beyond just the EU), as opposed to the earlier Directive, and the privacy framework promoted by the Asia Pacific Economic Cooperation area, APEC. The GDPR's top-down approach contrasts with other recent attempts of companies at self-regulation (e.g., Facebook's internal ethical committee), and proposals for organizational structures across industry and academia.

Though the title of the Regulation does not contain the word 'privacy,' it is arguably its overarching concern. As mentioned previously in section 3.2, many of the provisions in the GDPR were added in order to be legally consistent with the fundamental privacy rights outlined in the The Charter of Fundamental Rights of the European Union. Perhaps the word 'privacy' was not included in the GDPR's title because its framers felt that the GDPR was not so much about establishing new rights to privacy as it was about making sure that existing rights to privacy were adequately protected in the age of Big Data. In any case, a forthcoming Regulation, the FFD, or free-flow of data Regulation, will deal specifically with issues of non-personal data that are outside of the scope of the GDPR.

In terms of evaluating the success of the GDPR in stopping infringements of privacy, we might

look to similar international regulatory frameworks used in the banking and finance industry, such as the Basel II/III Accords. There, in an effort to prevent further international banking crises, minimum bank capital requirements were introduced (Drumond, 2009). Additionally, guidelines for data collection and modeling were announced, some of which are designed to make models' predictions more transparent to end users and regulators (Saddiqi, 2017). The success of such measures is the avoidance of future international banking crises. In judging the success of the GDPR, we might analogously look to see whether major data privacy scandals, such as Facebook-Cambridge Analytica scandal in 2018, continue to make headlines. Given that personal data has become an essential commodity in the information technology industry, we would be naive, however, to believe that the GDPR could completely deter companies and individuals from using personal data in unscrupulous ways.

It is clear that governments around the world are looking at the GDPR as a litmus test for determining whether it is possible to safely rein in the juggernaut of digital technology. On a fundamental level, the issue underlying the creation of the GDPR is: Can social and political values successfully dictate the application of new technologies, or will new technologies dictate our social and political values? The technology ethicist Van Den Hoven (2012), for example, writes that "a debate on the future values of living is necessary." And the historian Yuval Noah Harari claims that the most important political question of our time is: "How do you regulate the ownership of data?" (Harari, 2018). The GDPR represents just one participant, albeit the loudest and most outspoken, in that global debate. It attempts to answer Harari's question by positing fundamental human rights to privacy and balancing the economic benefits of personalized predictions and automated decision-making with their potential social costs and infringements on individual privacy.

As algorithms increasingly pervade daily life, technologists must begin to consider and assess the impacts of their creations on society at the outset of their work. To do this, our current system of educating engineers and computer and data scientists needs to evolve into one that properly evaluates technology's crucial role in today's digital world, especially its growing role in the political process. Spradling et al. (2008) found that in their online survey of US-based computer science undergraduates, "One hundred sixty-eight (77%) of the 218 survey program respondents that teach computer ethics answered that no training is provided." Quinn (2006) also reports that only about half of accredited undergraduate programs in computer science teach their own "computers, ethics, and society" classes. This needs to change. All new computer science and engineering graduates should receive training in human subjects research ethics on par with the training social science majors typically receive. Outdated research ethics guidelines will also need to be updated for the new GDPR and BBD-era. Specifically, how do the ethical principles of the Belmont report apply to the modern BBD context of public datasets, social networks, and personally identifiable information?

Because personal data processing, analysis, and modeling have become so embedded into modern day life, it is imperative that data scientists understand what is at stake, and why in 2018 and beyond legislators around the world feel the need to regulate their effects. Numerous commentators have weighed in on the issue. Some have called for changes to the very nature of our economic system in order to make sure the benefits of big data processing are equitably distributed (Zuboff, 2019; Scholz and Schneider, 2017). Others see the solution to Harari's question in an algorithmic bill of rights for citizens (Hosanagar, 2019). Finally, still others call for the creation of new and better regulatory bodies–modeled on those in aviation safety–to oversee the non-discriminatory use of algorithms (Shneiderman, 2016). This will likely be a fertile area of future research.

Personal data regulation is a complex issue. There is no simple solution. Each country must grapple with the problem within its unique cultural, political, legal, and social milieu. Most importantly, a society-wide debate needs to be had, otherwise we risk becoming servants to our own technological creations, proverbial frogs in the pot, unable to detect the slow changes in our social norms as we trade our privacy for convenience and a fleeting escape from boredom. We implore all data scientists and citizens to join in this debate. Ultimately, it is *their* personal data that power the algorithms that increasingly affect our daily lives.

Bibliography

Albrecht, J. (2016). How the gdpr will change the world. *European Data Protection Law Review*, 2(3):287–289.

Alexander, L., Das, S. R., Ives, Z., Jagadish, H., and Monteleoni, C. (2017). Research challenges in financial data modeling and analysis. *Big data*, 5(3):177–188.

Allen, D. W., Berg, A., Berg, C., Markey-Towler, B., and Potts, J. (2019). Some economic consequences of the gdpr. *Available at SSRN 3160404*.

Allen & Overy (2016). Binding corporate rules. (white paper).

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev*, 104.

Bird & Bird (2017). Guide to the general data protection regulation.

Borgesius, F. J. Z. (2015). Personal data processing for behavioural targeting: Which legal basis? *International Data Privacy Law*, 5:163.

Buttarelli, G. (2016). The eu gdpr as a clarion call for a new global digital gold standard.

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

Carroll, A. B. (1979). A three-dimensional conceptual model of corporate performance. *Academy of management review*, 4(4):497–505.

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.

Chen, Y.-J., Lin, C.-F., and Liu, H.-W. (2018). 'rule of trust': The power and perils of china's social credit megaproject. *Columbia Journal of Asian Law*, 32:1.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

de Leeuw, K. and Bergstra, J. (2007). *The History of Information Security: A Comprehensive Handbook*. Amsterdam: Elsevier.

Dedman, M. (2006). *The origins and development of the European Union 1945-1995: a history of European integration*. Routledge.

Determann, L. (2016). Adequacy of data protection in the usa: myths and facts. *International Data Privacy Law*, 6(3):244–250.

DLA Piper & AON (2018). The price of data security: A guide to the insurability of gdpr fines across europe.

Drumond, I. (2009). Bank capital requirements, business cycle fluctuations and the basel accords: A synthesis. *Journal of Economic Surveys*, 23(5):798–830.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9 (3-4):211–407.

Federal Trade Commission (2012). Protecting consumer privacy in an era of rapid change. ftc report. march 2012. Technical report.

Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

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

Garcia, D., Kassa, Y. M., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., and Cuevas, R. (2018). Analyzing gender inequality through large-scale facebook advertising data. *Proceedings of the National Academy of Sciences*, page 201717781.

Georgiadou, Y., de By, R. A., and Kounadi, O. (2019). Location privacy in the wake of the gdpr. *ISPRS international journal of geo-information*, 8(3):157.

Granger, M.-P. and Irion, K. (2018). The right to protection of personal data: the new posterchild of european union citizenship? In *Civil Rights and EU Citizenship*. Edward Elgar Publishing.

Greene, T., Shmueli, G., Ray, S., and Fell, J. (2019). Adjusting to the gdpr: The impact on data scientists and behavioral researchers. *Big data*.

Greenleaf, G. (2009). Five years of the apec privacy framework: Failure or promise? *Computer Law & Security Review*, 25(1):28–43.

Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now? *Big Data*, 6 (3):176–190.

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.

Hettne, B. and Soderbaum, F. (2005). Civilian power or soft imperialism-the eu as a global actor and the role of interregionalism. *Eur. Foreign Aff. Rev*, 10.

Hintze, M. and LaFever, G. (2017). Meeting upcoming gdpr requirements while maximizing the full value of data analytics. Technical report.

Hosanagar, K. (2019). *A Human's Guide to Machine Intelligence: How Algorithms Are Shaping Our Lives and How We Can Stay in Control*. Viking.

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

Jia, J., Jin, G. Z., and Wagman, L. (2018). *The short-run effects of GDPR on technology venture investment (No. w25248)*. National Bureau of Economic Research.

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

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

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

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

King, G. and Persily, N. (2018). A new model for industry-academic partnerships.

Korff, D. (2016). Practical implications of the new eu general data protection regulation for eu-and non-eu companies.

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.

Koszegi, S. T. (2019). *High-Level Expert Group on Artificial Intelligence*.

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.

Kulesza, J. (2011). Walled gardens of privacy or binding corporate rules: A critical look at international protection of online privacy. *UALR L. Rev*, 34:747.

Lanier, J. (2010). *You are not a gadget: A manifesto*. Vintage.

Loidean, N. N. (2016). The end of safe harbor: Implications for eu digital privacy and data protection law, 19 no. *J. INTERNET L.*, 8:1–12.

Lowthian, P. and Ritchie, F. (2017). Ensuring the confidentiality of statistical outputs from the ADRN. Technical report, Administrative Data Research Network.

Mansfield-Devine, S. (2013). Biometrics in retail. *Biometric Technology Today*.

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).

Metzl, J. (2019). *Hacking Darwin: Genetic Engineering and the Future of Humanity*. Sourcebooks, Inc.

Michalski, A. (2005). The eu as a soft power: the force of persuasion. In Melissen, J., editor, *The New Public Diplomacy: Studies in Diplomacy and International Relations*. Palgrave Macmillan, London.

Moerel, L. (2012). *Binding Corporate Rules: Corporate Self-Regulation of Global Data Transfers*. OUP, Oxford.

Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., and Bell, J. (2018a). 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.

Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., Smith, H., Aidinlis, S., and Kaye, J. (2018b). 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.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., Smith, N. A., et al. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm*, 11(122-129):1–2.

Ohm, P. (2012). Branding privacy. *Minn. L. Rev.*, 97.

Olshannikova, E., Olsson, T., Huhtamäki, J., and Kärkkäinen, H. (2017). Conceptualizing big social data. *Journal of Big Data*, 4(1):3.

O'Neil, C. (2016). *Weapons of Math Destruction: how big data increases inequality and threatens democracy*. Crown Publishers, New York.

Orlitzky, M., Schmidt Frank, L., and L., R. S. (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies*, 24(3):403 – 441.

Parasuraman, R. and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410.

Politou, E., Alepis, E., and Patsakis, C. (2018). Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions. *Journal of Cybersecurity*, 4:1.

Quinn, M. J. (2006). On teaching computer ethics within a computer science department. *Science and Engineering Ethics*, 12(2):335–343.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, 568(7753):477.

Reding, V. (2011). Binding corporate rules: unleashing the potential of the digital single market and cloud computing. Technical report, IAPP Europe Data Protection Congress.

Reding, V. (2012). The european data protection framework for the twenty-first century. *International Data Privacy Law*, 2(3):119–129.

Rhoen, M. (2017). Rear view mirror, crystal ball: Predictions for the future of data protection law based on the history of environmental protection law. *Computer law & security review*, 33(5):603–617.

Robbins, S. P. and Judge, T. (2017). *Organizational Behavior*. Pearson, seventeenth edition.

Russell, B. (2001). *The problems of philosophy*. OUP Oxford.

Saddiqi, N. (2017). *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*. Hoboken, New Jersey: Wiley, 2nd edition.

Safari, B. A. (2016). Intangible privacy rights: How europe's gdpr will set a new global standard for personal data protection. *Seton Hall L. Rev.*, 47:809.

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

Scholz, T. and Schneider, N. (2017). *Ours to hack and to own: The rise of platform cooperativism, a new vision for the future of work and a fairer internet*. OR books.

Schwartz, M. S. and Carroll, A. B. (2003). Corporate social responsibility: A three-domain approach. *Business ethics quarterly*, 13(4):503–530.

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

Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, 113(48):13538–13540.

Spradling, C., Soh, L.-K., and Ansorge, C. (2008). Ethics training and decision-making: do computer science programs need help? *ACM SIGCSE Bulletin*, 40(1):153–157.

Steppe, R. (2017). Online price discrimination and personal data: A general data protection regulation perspective. *Computer Law & Security Review*, 33(6):768–785.

Tene, O. and Polonetsky, J. (2013). A theory of creepy: technology, privacy and shifting social norms. *Yale JL & Tech*, 16.

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

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. (2018). Privacyguide: towards an implementation of the eu gdpr on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 15–21. ACM.

Tikkinen-Piri, C., Rohunen, A., and Markkula, J. (2018). Eu general data protection regulation: Changes and implications for personal data collecting companies. *Computer Law & Security Review*, 34(1):134–153.

Turow, J. (2017). *The Aisles Have Eyes: How Retailers Track Your Shopping, Strip Your Privacy, and Define Your Power*. New Haven : Yale University Press.

Turow, J., Hennessy, M., Draper, N., Akanbi, O., and Virgilio, D. (2018). Divided we feel: Partisan politics drive american's emotions regarding surveillance of low-income populations.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. science. 185(4157):1124–1131.

US Chamber of Commerce (2014). Business without borders.

Van Den Hoven, J. (2012). "fact sheet-ethics subgroup iot-version 4.0.". Technical report.

Veale, M., Binns, R., and Ausloos, J. (2018a). When data protection by design and data subject rights clash. *International Data Privacy Law*, 8(2):105–123.

Veale, M., Binns, R., and Van Kleek, M. (2018b). Some hci priorities for gdpr-compliant machine learning. *arXiv preprint arXiv:1803.06174*.

Veale, M. and Edwards, L. (2018). Clarity, surprises, and further questions in the article 29 working party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, 34(2):398–404.

Voigt, P. and dem Bussche, A. V. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer International Publishing, 1st edition.

Wachter, S. (2018). Normative challenges of identification in the internet of things: Privacy, profiling, discrimination, and the gdpr. *Computer law & security review*, 34(3):436–449.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):2017.

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

West, S. M., Whittaker, M., and Crawford, K. (2019). Discriminating systems: Gender, race and power in ai. Technical report, AI Now Institute.

Wugmeister, M., Retzer, K., and Rich, C. (2006). Global solution for cross-border data transfers: Making the case for corporate privacy rules. *Geo. J. Int'l L.*, 38:449.

Zarsky, T. Z. (2016). Incompatible: The gdpr in the age of big data. *Seton Hall L. Rev.*, 47:995.

Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., et al. (2017). Ten simple rules for responsible big data research.

Zuboff, S. (2019). *The age of surveillance capitalism: the fight for the future at the new frontier of power*. Profile Books.

Appendix A: Academic Research Under GDPR

Though many of the GDPR's Articles and Recitals apply to industry-focused personal data processing, there are several exemptions for certain kinds of research. Principal investigators should therefore first determine whether their research project falls under statistical/scientific/historical/archiving in public interest purposes, and is thereby excluded from honoring certain data subject rights.[1] It should be noted that what counts as scientific research under the GDPR is quite broad and includes "technological development and demonstration, fundamental research, applied research and privately funded research (Recital 159)."

Key Points to Consider

For exemptions to apply, publication must be envisaged and the publication must be in the public interest.[2] Further, all three of the following conditions must be met:

- Individuals decisions (i.e., predictions) regarding data subjects are not made
- Processing will not cause substantial damage or distress to data subjects
- Publication of results will not result in data subject re-identification[3]

If these conditions are met then researchers can also waive the "right to inform" data subjects of the processing activities if providing privacy notices would take "disproportionate effort." Further, exempted forms of research do not have to adhere to the following two GDPR principles:

- Purpose limitation
- Duration limits

GDPR Research Requirements (even if otherwise exempt)

However, researchers must still adhere to general GDPR accountability requirements that include things like data protection by design and default, third party processing contracts, keeping records of data processing, documenting and reporting of data breaches, and carrying out DPIAs for high risk processing activities. This includes using pseudonymization as the most basic form of data confidentiality and security.

[1] See Recital 159 for more guidance.

[2] These criteria are specific to the UK. Each member state may, however, create slightly different rules for exemption.

[3] This criterion is especially relevant for human subjects research involving online surveys, or any kind of secondary analysis of clinical trials.

General Research Guidelines Under GDPR

- If the research project is likely to result in privacy risks to many subjects, carry out an initial Data Processing Impact Assessment, especially if there are sensitive categories (health or ethnicity, for example)
- If collaborating with industry, typically the processor (the academic group carrying out the research) must agree that no publication will include information that could be used to identify a natural living person

In order to keep research in-line with the major GDPR principles of data minimization and data protection by design and by default, it is suggested that at the start of research the following questions are decided:

- What will be done with tables with small sample numbers in cells?
  - Potential strategies include setting a cell count threshold below which cells will be combined (collapsing categories), or replacing small count cells by reporting 0%
- What about analyses done at a level of detail below what is necessary to achieve the project's goal?
  - For example, must city or district level information be reported when only country level is needed?
- Should extreme values and outliers be published?
- Should residuals be reported?

Appendix B: Corporate Guidelines for GDPR Compliance

Informing Data Subjects of GDPR Changes and Rights[4]

- Update the company privacy policy website to reflect any post-GDPR changes in policy
- Direct employees and customers to your company privacy policy webpage
- Allow for employees and customers to update their personal information
  - Records kept must be accurate, updatable, and data subjects should be able to get copies of the data collected on them in a portable format if requested

Privacy Policies

- Check privacy policies and ensure harmony if different departments use different versions
- Communicate to data subjects what information is collected and for what purposes
  - Use clear, non-technical language appropriate for data subjects (e.g., kids and the elderly, if applicable)
  - Ensure data processing is actually necessary for the provision of services or performance of a contract with the data subject
  - Identify whether you collect any "sensitive" personal data (ethnicity/religion/etc.)
  - If using "algorithmic profiling" for individual data subjects, make sure to have an opt-out system in place
    - Your data science team should be able to explain–in a non-technical way–how the profiling is done and how decisions were made
  - Review consent mechanisms
    - Consent must be unambiguous, freely given, revocable, and demonstrable
- Allow users to opt-out of any personal data published on public websites (e.g., employee photos or personal data on the "Meet Our Team" page)

[4]These guidelines were adapted from University of Cambridge, GDPR Data Protection Working Group, Toolkit to help University Institutions Prepare for New Data Protection Legislation (GDPR). November 2017.

- Review your procedures for past job applicants, former employees, and customers Storage duration limits should be specified

Compliance and Accountability

- Hire a Data Protection Officer if a large company or doing high-risk processing
- Assign a legal representative to the EU if offering goods/services to EU data subjects
- Review your legal basis for direct marketing
  - Are you relying on consent, legitimate interest, etc?
- Consider general staff GDPR training and awareness programs in order to demonstrate compliance

IT Considerations

- Train IT such that data protection by design and default are key priorities when building a new system or project
- Review your data storage and retention procedures for the exercise of the "right to be forgotten"
  - How will IT deal with access requests by data subjects?
  - Does IT have a system in place to deal with a potential data breach?

Appendix C: GDPR Data Sharing Guidelines

The following guidelines should be used when a firm outsources (or transfers personal data) to a third party processor, such as an overseas consulting firm.[5] These same guidelines can also be used during industry-academic collaboration, where the academic(s) performing the research would be considered the data processor(s).

Definition of a Data Processor Article 4(8) defines a data processor as, "*a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.*" Examples of common third party processors may be cloud storage services, direct marketing companies, web hosting services, analytics consultants, and web-based services such as Google Analytics or Addthis social media sharing buttons on websites.

Basic Responsibilities of Processors Under GDPR Article 28(3) states that processing must include a contract that spells out the basics of processing and guarantees respect for data subjects' rights and certain organizational and security procedures. Further, controllers are legally responsible for their processors, and must only appoint those that can provide "sufficient guarantees" that the GDPR's rules will be met and data subjects' rights will be adequately protected. Controllers should therefore conduct due diligence on any potential data processors as early as possible before any contracts are signed.

Key Points to Include in a Data Processing Contract

- Basic information regarding the type of information, the duration, scope and purpose of processing
- Identification of relevant personal data and their (potentially sensitive) categories
- The Processor should only act on the written instructions of the controller
- If required, the Processor must appoint in writing a representative located in the EU[6]
- The Processor must take appropriate measures of security

[5] If the processor is outside of the European Economic Area (EEA), then the processor must either be in a country deemed 'adequate' by the European Commission (EC), or be willing to agree to specific model contractual clauses issued by the EC (though as of August 2018, the various data protection authorities had not yet made these available). Otherwise transfers of personal data cannot be made legally.

[6] The following are conditions for exemption from (Article 27 (2) (a)): "Processing which is occasional, does not include, on a large scale, processing of special categories of data as referred to in Article 9(1) or processing of personal data relating to criminal convictions and offences referred to in Article 10, and is unlikely to result in a risk to the rights and freedoms of natural persons, taking into account the nature, context, scope and purposes of the processing."

- The Processor must keep adequate documentation of processing activities, including a general description of the technical and organisational security measures taken[7]
- The person(s) responsible for processing the data should have a duty of confidence to data subjects
- The Processor cannot engage any sub-processors without the prior consent of the controller and without a written contract
- The Processor must assist the controller should data subjects wish to access their data or otherwise exercise their rights under GDPR
- The Processor must assist the Controller with its security obligations including any breach notifications or data protection impact assessments
- The Processor must be willing to submit to audits to ensure Controller and Processor are meeting their Article 28 obligations for processors
- The Processor must delete or return all personal data to the controller at the end of the contract

[7] Article 30 (2)

?? ?? ?? ?? ?? ?? ?? ?? ??
?? ?? ?? ?? ?? ?? ?? ?? ??
?? ?? ??
?? ?? ?? ?? ??

Appendix D: Glossary of GDPR terms and their definitions

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..

Direct marketing (Article 21(2)) The data subject shall have the right to object at any time to processing of personal data concerning him or her for such marketing, which includes profiling to the extent that it is related to such direct marketing.

Business development (Recital 47) The legitimate interests of a controller, including those of a controller to which the personal data may be disclosed, or of a third party, may provide a legal basis for processing, provided that the interests or the fundamental rights and freedoms of the data subject are not overriding, taking into consideration the reasonable expectations of data subjects based on their relationship with the controller. Such legitimate interest could exist for example where there is a relevant and appropriate relationship

between the data subject and the controller in situations such as where the data subject is a client or in the service of the controller..

Purpose limitation (Article 5(1)(b)) [Personal data shall be] collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation').

Scientific research (Recitals 162, 159, 157) Scientific research purposes should be interpreted in a broad manner including 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 and public interest (Article 89 (3)) Where personal data are processed for archiving purposes in the public interest, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16,

18, 19, 20 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article 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..

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"..

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 and Automated processing (Article 4(4), Recital 71) 'Profiling' [consists] of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to 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..

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

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

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

Privacy by Design (Article 25) Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

Consent (Article 7(2,3)) If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior

to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent .