Chapter 5
Legal Aspects of Applied Data Science

Michael Widmer and Stefan Hegy

Abstract Data scientists operate in a legal context and the knowledge of its rules provides great benefit to any applied data science project under consideration, in particular with view to later commercialization. Taking into account legal aspects early on may prevent larger legal issues at a subsequent project stage. In this chapter we will present some legal topics to provide data scientists with a frame of reference for their activities from a legal perspective, in particular: (1) comments on the qualification and protection of "data" from a legal perspective, including intellectual property issues; (2) data protection law; and (3) regulatory law. While the legal framework is not the same worldwide and this chapter mainly deals with Swiss law as an example, many of the topics mentioned herein also come up in other legislations.

1 Introduction and Background Information

Data science is inherently an applied science (Stadelmann et al. 2013) aiming at generating value from the data itself; thus, law is among the many disciplines to be taken into account in data scientists' activities. Considering legal aspects already at the outset of the development of data products may help address and minimize last minute legal issues, which would not only be frustrating but in many cases also costly. An example could be the development of a business solution for a financial service provider that neglects certain regulatory requirements stated by the financial supervisory authority. In this event, the regulatory barrier would impede a successful implementation, irrespective of increased efficiency standards or the potential commercial value of such solution. Implementing the regulatory requirements into an already existing product at the end may require substantial work and involves additional costs. If the legal Dos and Don'ts had been properly outlined and taken into consideration from the outset—either with or without the client being

M. Widmer · S. Hegy (✉)
ZHAW Zurich University of Applied Sciences, Winterthur, Switzerland
e-mail: wime@zhaw.ch; hegy@zhaw.ch

© Springer Nature Switzerland AG 2019

M. Braschler et al. (eds.), *Applied Data Science*,
https://doi.org/10.1007/978-3-030-11821-1_5

involved—the project would possibly have been successful without further changes. Also, understanding and taking into account legal concerns from the start may make companies more willing to conduct (big) data analytics, while currently some companies are still reluctant to do so, partly because of legal and information security concerns (KPMG and Bitkom research 2016).

While this chapter cannot provide an in-depth review of all legal aspects, it is meant to provide data scientists with a certain awareness for important issues they may encounter in the "legal minefield" of data science, if only to seek assistance from a legal department or legal advisor at the right time. Since the number of potential legal issues is extensive, this contribution will outline some of these legal topics to provide data scientists a frame of reference, in particular: (1) comments on "data" from a legal standpoint, including intellectual property issues (*see* Sect. 2 below); (2) data protection law (*see* Sect. 3); and (3) regulatory law (*see* Sect. 4). While the legal framework is not the same worldwide and this chapter mainly deals with Swiss law as an example, many of the topics mentioned also come up in other legislations. For example, the draft for a revised Swiss Data Protection Act mirrors many of the provisions of the General Data Protection Regulation of the EU. Therefore, the issues raised below will likely also come up in a similar manner in other jurisdictions. However, data scientists have to be aware that if their activities concern a multitude of jurisdictions (in particular, if they are acting cross-border) they will have to determine which laws apply and may even have to take into account various legal systems.

2 "Data" from a Legal Standpoint: Goods, Intellectual Property, and Unfair Competition Law

2.1 Introductory Comments

It is obvious that "data" is at the core of data science. Consequently, data scientists should always consider how the rights to data affect their activities, for example, whether they have obtained the necessary rights concerning the data they work with and how such rights may affect the results of their work. If they have not obtained the necessary rights, their activities may infringe third parties' rights and the exploitation, use, or even publication of the results of their activities may be adversely affected (cf. Oxenham 2016, p. 16).

2.2 Ownership of Data as Ownership of Goods?

While the legal qualification of "data" is important to the activities of data scientists, the term is not entirely clear in all respects. The question of "ownership" of data, how the commercial purposes (e.g., transfer of data) may be reached from a legal

perspective and how data should be qualified were extensively discussed in Switzerland. With respect to personal data some argue that a right similar to an intellectual property right should be introduced (Flückiger 2013, p. 864) or that data should be treated similar to goods (Eckert 2016b, pp. 247 et seq.) (Eckert 2016a). Others are against the introduction of ownership rights and propose contractual solutions of the issues arising in this context (Benhamou and Tran 2016, pp. 572 et seq.).[1] In short: how data should be qualified from a legal perspective is still somewhat disputed.[2] However, data are not "goods," and as long as the data is not protected by copyrights, other intellectual property rights, or unfair competition laws (*see* below), the only remaining solution is to solve any issues arising, for example, concerning how certain data should be used, with contracts—although this may not always be possible.

Consequently, while it is agreed that data often is a commercial asset that is widely being sold and licensed, the legal discussion in this respect is still somewhat open. Data scientists have to keep in mind that the above-mentioned issues will have an impact on their activities and have to be taken into account.

But irrespective of how the abstract "data" is qualified from a legal perspective, one has to remember that the data *carriers*, that is, the tangible media on which the data is stored, legally qualify as "goods." The ownership, etc., of such data carriers will be treated like any other goods, so they can, for example, be owned and sold.

2.3 Copyrights

In addition to the protection of the legal ownership of the "data carriers," on which the data is stored, the abstract "data" will in certain cases be protected under other legal concepts. For example, certain data may be protected by copyrights and neighboring rights.

With regard to the European Union (EU), several directives are in place to harmonize essential copyright rights throughout the member states. Reference is made in particular to the Directive on the harmonization of certain aspects of copyright and related rights in the information society (InfoSoc Directive; 2001/29/EC) and the Directive on the legal protection of computer programs (Software Directive; 2009/24/EC) (cf. European Commission n.d.). In Switzerland, Article 2 para. 2 Swiss Copyright Act (CA) provides that literary and artistic intellectual creations with an individual character constitute "works" and are protected, irrespective of their value or purpose. Computer programs also qualify as "works" in this sense (Article 2 para. 3 CA). "Collections" are protected as works in their own right insofar as they are intellectual creations with individual character with regard to their selection and arrangement (Article 4 para. 1 CA). However, this does not

[1] Further details concerning this discussion as well as alternative solutions, cf. Thouvenin (2017).

[2] With regard to the international discussion, cf., for example, Ritter and Mayer (2018).

preclude that works included in a collection may continue to be protected individually (Article 4 para. 2 CA).

Where data science is concerned, the question will arise whether certain data is copyright protected or not. While the threshold for a qualification as a work is not very high (e.g., original texts, pictures, or music are protected by copyright), mere ideas are not copyright protected and this question will have to be answered on a case-by-case basis. Depending on the data set, which is intended to form the basis of the data scientist's work, either some (or none) of the data therein may be protected as an individual work, part of the data may constitute a collection, or even the entire data set may constitute a collection.[3]

To the extent that certain data do constitute a work or a collection in the sense of the CA, the CA grants the author of such a work the *exclusive right* thereto and the *exclusive right* to decide whether, when, and how his work is used (Article 9 para. 1 and Article 10 para. 1 CA). This includes, in particular, the right to make copies of the work, such as data carriers, to offer, transfer, and otherwise distribute copies of the work as well as to decide whether (Article 10 para. 2 CA), when, and how the work may be altered or used to create a derivative work or may be included in a collection (Article 11 para. 1 CA).

In the context of data science, this means that copying data which constitutes a work or collection infringes on the exclusivity rights of the author. However, one could consider licensing or buying the copyrights to such data. Moreover, infringements of copyrights could be *justified* (e.g., by consent) or could fall within the scope of a copyright exemption, such as private use in enterprises, public administrations, institutions, commissions, and similar bodies for internal information and documentation (Article 19 CA).

However, it has to be noted that where data is available publicly on the Internet, this does not automatically mean that the potentially existing copyrights have been abandoned or implied consent to copying of such data has been given [a *different view is held by* Weber (2014, p. 22)]. Also, on the Internet it is not always clear whether the person making the work available publicly is actually the right owner. This has to be kept in mind by data scientists, in particular when creating data products, even if they are meant to mainly include or use publicly available data from the Internet.

In addition, also the tools of data scientists or the result of their activities may be protected by copyright, such as computer programs. However, individual algorithms and short computer programs that lack any complexity are not protected under Swiss copyright law (*see* Cherpillod 2012, n. 64).

Finally, there is the question, whether computer-generated works can be subject to copyright protection. Traditionally, the protection of computer-generated works by copyright was not an issue. Computers and computer programs were considered

[3]For example, the Swiss Federal Supreme Court found that the Swiss Compendium of Medicines did *not* reach the required individual character and was not protected by copyright (BGE 134 III 166).

to be merely tools that supported the creative process, very much like a brush in the hands of an artist. But, with the rapid development of machine learning software, a subset of artificial intelligence, some computer programs are no longer only supportive tools; they actually take autonomous decisions with regard to the creative process—free of human intervention (Guadamuz 2017). Therefore, it is subject to debate whether creations generated through machine learning qualify as "intellectual creations" and enjoy copyright protection. A broader concept of copyright protection in this field may help protect the results of the work of data scientists in certain cases.

2.4 Database Right Sui Generis

Furthermore, in the EU, there is another right that may serve to protect certain data. The Directive on the legal protection of Databases (Directive 96/9/EC) provides for an exclusive specific right for database producers. The holder of such database rights may prohibit the extraction and/or re-utilization of the whole or certain parts of a database. The rights are valid for 15 years and are meant to protect the investment of time, money, and effort into such database, irrespective of whether the database is in itself innovative ("non-original" databases). Thus, it applies even if the database is not protected by copyrights. (The Directive harmonized also copyright law applicable to the structure and arrangement of the contents of databases ("original" databases).)

While this sui generis database right covers databases, which would not be protected under regular copyright, one has to keep in mind that it only applies if there has been a qualitatively and/or quantitatively substantial investment in either the obtaining, verification, or presentation of the contents. Moreover, the right only prevents extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database (Article 7 para. 1 Directive 96/9/EC). If only insubstantial parts of the database are concerned, repeated and systematic extraction and/or re-utilization are still prohibited under certain circumstances (Article 7 para. 5 Directive 96/9/EC).

In many cases, the activities of data scientists will not be so extensive as to infringe these rights. However, whenever data scientists obtain databases, this so-called sui generis database right should be taken into account and it should be determined whether it applies to their specific case.

As above with copyrights, the result of the activities of data scientists (or interim results) may fall under the Directive 96/9/EC and enjoy the respective protection. This may help in the commercial exploitation of these results (e.g., as part of a licensing of such database rights to third parties).

2.5 Unfair Competition Law

In Switzerland, the legal protection of such databases is not as clear. Switzerland has not introduced a database right similar to the one of the EU mentioned above. While some databases will be covered by copyright (if the selection and arrangement of the data is an intellectual creation with individual character) even extensive data collections or databases will not constitute a copyrightable work if the respective system is simple. In any case, the copyright protection enjoyed by a collection will be limited to the structure of the database (e.g., selection and arrangement), if the respective data itself is not protected by copyright (Arpagaus 2013, n. 35). Therefore, it is important to examine how other legal concepts may be used to obtain a certain legal protection of such databases:

Article 5 let. c Unfair Competition Act (UCA) provides that it constitutes unfair competition to take the marketable results of work of another person by means of technical reproduction processes without an adequate effort of one's own and to exploit them "as such." The question is whether—and, if so, under what circumstances—this provision could be used to protect at least certain databases.

Databases, even ones that are not protected by copyright, can constitute a "marketable result of work of another person" in the sense of Article 5 let. c UCA. Downloading such a database from the Internet or otherwise copying it constitutes a "technical reproduction process." Consequently, some of the requirements for protection under Article 5 let. c UCA will be met in many cases.

The issue then becomes whether the exploitation of such a database was made "*without an adequate effort*," This will have to be considered with respect to the specific facts on a case by case basis. One will have to take into account a comparison between the effort of the first mover and the effort of the potential infringer—and also, whether the first mover already had an opportunity to amortize its efforts at the time of reproduction and exploitation.

In addition, the marketable results have to be exploited "*as such*." While this term is very vague, the Federal Supreme Court has taken this to mean that it requires not only that the reproduction of the result is made directly, but that it also must be exploited directly. *Direct reproduction* requires that the technical process directly implicates the original, while *direct exploitation* would require a commercial or professional use of the result in competition without an adequate own effort of the infringer.

Often, data scientists will not "take" the entire (or large part) of an existing database or another marketable result of a third party merely by means of a technical reproduction process, but will put more effort into their work, in particular use at least a combination of data. In addition, data scientists will often not directly exploit databases without an adequate own effort. The entire idea of data science is to apply one's own (adequate) effort to a set of data and create a new data product, which obviously goes far beyond the original set of data (Weber 2014, p. 21).

Therefore, Article 5 let. c UCA will often not be applicable in the context of data science. Nevertheless, it is important to keep this provision in mind to be able to avoid its pitfalls on a case by case basis.

2.6 Manufacturing and Trade Secrets

In addition to the provisions outlined above, certain data that data scientists want to use may be protected as manufacturing or trade secrets. While protection by Article 5 UCA as outlined in the section above does not require the data to be secret, Article 6 UCA protects *manufacturing and trade secrets* if they have become known as the result of spying or otherwise have been obtained unlawfully. In such cases, exploiting manufacturing or trade secrets or disclosing them to third parties constitutes unfair competition.

Moreover, the disclosure of manufacturing or trade secrets by a person who is under a statutory or contractual obligation not to reveal such facts, as well as exploiting manufacturing or trade secrets disclosed in such a way, is punishable under Swiss criminal law (Article 162 Swiss Criminal Code).

While data scientists in many cases will not intentionally violate manufacturing or trade secrets, they should still try to make sure that the data they use does not contain and violate such secrets.

3 Data Protection/Privacy

3.1 Background

While it may well be that only companies will survive that rigorously exploit (big) data, one should not forget that data science and data exploitation must not lead to an infringement of privacy rights (Polzer 2013, p. 6). Data protection and privacy are protected by the Swiss constitution as fundamental constitutional rights. Data protection laws are meant to specify the constitutional rights of privacy. Those data protection laws also have to be taken into account in the field of data science.[4]

Swiss data protection law is mainly set forth in the Federal Act on Data Protection of June 19, 1992 (DPA), and the Swiss Federal Ordinance to the Federal Act on Data Protection of June 14, 1993 (DPO). In the EU, the General Data Protection

[4]This contribution is not meant to be a full-blown introduction into data protection/privacy laws and the following will concentrate on a limited number of data protection law issues, which may have particular importance for data scientists.

Regulation (GDPR)[5] has entered into force and will apply as of May 25, 2018. The Swiss DPA is currently under revision and it is expected that it will be strongly influenced by the GDPR, in particular because cross-border data transfers are daily business.[6]

3.2 Personal Data

Swiss data protection laws only deal with the processing of personal data. Obviously, not all data is personal data. Under Swiss law, personal data is defined as "all information relating to an identified or identifiable person" (Article 3 let. a DPA). A person is considered to be identifiable if identification is possible without undue efforts and one has to expect that this will possibly be done (Rudin 2015, n. 10).

While this definition seems clear, there is a large spectrum between data that is clearly connected to an identifiable person and data that cannot in any way be re-identified.

De-identification of data generally is used to denominate a process of "removing or obscuring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them" (Nelson 2015, p. 12). Therefore, de-identified data may theoretically still be linked to individuals, for example, using a code, algorithm, or pseudonym.

The definition of "**pseudonymization**" in the GDPR is somewhat different: "*'pseudonymisation' means 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*" (Article 4(5) GDPR).

Anonymization on the other hand is a process of data de-identification leading to data where individual records cannot be linked back to an individual as they do not include the required translation variables to do so. Consequently, anonymized data, as it is often used in data science, is generally not subject to the DPA. De-identification may also be sufficient to exclude data from the scope of DPA, if the re-identification is not possible without undue efforts or if one does not have to expect that this will possibly be done.

However, data scientists should be aware that the process of anonymization or de-identification of data, which currently constitutes personal data, does, in itself,

[5]Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

[6]The present contribution is based on the status of legislative proceedings as of February 23, 2018.

constitute the processing of personal data and, thus, is subject to the DPA. Only the result of the anonymization (and possibly of the de-identification) is no longer relevant from a perspective of data protection laws.

Also, there is no guarantee that de-identification and/or anonymization completely precludes re-identification of the data subject. On the contrary: in particular in connection with Big Data, if there is a large amount of data, re-identification of the data subject becomes more likely and possible (Baeriswyl 2013, p. 15).[7] Once such re-identification becomes possible "without undue efforts" and one has to expect that this will possibly be done, the data becomes personal data, and the DPA applies. Consequently, if one has sufficient data to conduct such re-identification, one will have to comply with the DPA (Weber 2014, p. 20). So, while the process of re-identification itself constitutes a data processing that is relevant under the Swiss DPA, one has to be aware that the DPA becomes applicable already at an earlier stage, that is, once re-identification is "possible without undue efforts and one has to expect that this will possibly be done" (Baeriswyl 2014, pp. 50–52). If personal data is generated by accident, nevertheless, Swiss data protection laws would apply. Finally, even if the data scientist provides de-identified and/or anonymized data to third parties, data protection laws will have to be complied with, if the data scientist has to expect that re-individualization will take place (Baeriswyl 2014, p. 53).[8] This is an issue that should be further analyzed in joint research activities conducted by IT specialists and legal scholars.

Thus, the boundary between personal data and other data is somewhat vague, in particular because of the technical developments; data that cannot be re-individualized today may well become related to an identifiable person tomorrow, and, thus, become personal data (FDJP 2016, p. 43).

Consequently, even anonymization or de-individualization of the respective data does not completely exclude that data protection laws will be applicable to the activities of a data scientist. This is true irrespective of whether the data is used only internally in a data product or whether it is visible also externally and irrespective of the effect of the data product on the data subject concerned (e.g., whether you use the data for personalized pricing or to achieve better usability of a software for the data subject).

[7] With regard to de-identification, re-identification, alternative approaches, and use-cases, cf. Narayanan et al. (2016).

[8] In this context one may also point to the US Federal Trade Commission's (FTC) 2012 report *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*, in which the FTC takes the position that "data is not 'reasonably linkable' to the extent that a company: (1) takes reasonable measures to ensure that the data is de-identified; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data" (retrieved February 14, 2018, from https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers).

3.3 Privacy by Design

Sometimes, legal developments are outpaced by technological developments. Data protection laws try to address this issue by provisions concerning "privacy by design"—the GDPR as well as the draft for a revision of the DPA. Privacy by design is an approach which takes privacy into account already in the phase of designing a product or a data analysis.

The idea of including this principle into the relevant laws is that law and technology should complement each other and that technologies, which already take privacy into account, are necessary to help implement data protection laws (Legislative Message DPA 2017, p. 7029). Technology may be used to enhance data security and, at the same time, the level of protection of personal data (Kiss and Szöke 2015, p. 323).

In the draft to the revised DPA, the principle requires that technical and organizational measures have to be set up in order for the data processing to meet the data protection regulations. It has to be considered from as early as the planning stage. The purpose is to achieve that systems for data processing are engineered (from a technological and organizational perspective) from the beginning in a way that they comply with data protection principles (Legislative Message DPA 2017, p. 7029).

While this is rather vague, there are already certain reports and principles that can be used when trying to determine what "privacy by design" requires. Some guidance can be found, for example, in the following "7 foundational principles" of privacy by design (Cavoukian 2011):

1. Proactive not reactive, preventive not remedial

   The privacy by design approach aims to identify, anticipate, and prevent privacy invasive events before they arise. It does not wait for privacy risks to materialize, nor does it offer remedies in case a privacy breach occurs.

2. Privacy as the default

   The default settings deliver the maximum degree of privacy. No action is required by the individual in order to protect their privacy.

3. Privacy embedded into design

   Privacy is integral to the system, without diminishing functionality. It becomes an essential component of the core functionality being delivered.

4. Full functionality—positive-sum, not zero-sum

   Privacy by design accommodates all legitimate interests and objectives in a positive-sum "win-win" manner. It avoids the pretense of false dichotomies, such as privacy vs. security, demonstrating that it is possible to have both.

5. End-to-end-security—full lifecycle protection

   Privacy must be protected by strong security measures throughout the entire lifecycle of the data involved; from the cradle to the grave.

6. Visibility and transparency—keep it open

   The data subject is made fully aware of the personal data being collected, and of the purpose(s). Moreover, the component parts and operations remain visible and transparent.

7. Respect for user privacy—keep it user-centric

   Privacy measures are consciously designed around the interests and needs of individual users.

In addition, the European Union Agency for Network and Information Security has addressed the issue in its report "Privacy and Data Protection by Design—from policy to engineering," which tries to bridge the gap between the law and the available technologies. It can also provide further insight into this issue and is certainly a good reference for data scientists.

3.4 Privacy by Default

While "privacy by default" is listed as one of the "7 foundational principles" of privacy by design above, this principle is also explicitly mentioned in the GDPR as well as the draft for a revision of the DPA.

The respective legal provisions require that it is ensured by suitable settings that by default only such personal data are processed that are required for the respective purpose of the processing. The "default setting" is the setting that is automatically given or applied to a software application, computer program, or device, if not altered or customized by the user.

In other words, the respective data processing should—as a default—be as privacy friendly as possible, except if the data subject changes the default settings (Legislative Message DPA 2017, p. 7030), for example, to obtain additional functionalities. Such settings have to enable the data subject to make its own choices concerning privacy to a certain extent.

3.5 Automated Individual Decisions

Another provision in data protection law which could substantially affect the activities of data scientists concerns "automated individual decisions." The GDPR as well as the draft for a revision of the DPA restrict automated individual decision making under certain circumstances. The GDPR states that 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" (Article 22 para. 1 GDPR). The draft for a revision of the DPA provides that a data subject has to be informed "of a decision which is taken exclusively on the basis of an automated processing, including profiling, and

which has legal effects on the data subject or affects him significantly" (Legislative Message DPA 2017, pp. 7056 et seq.).

This may lead to substantial difficulties in data science, in particular in cases where individual decisions are taken by algorithms. However, the GDPR only covers decisions based "*solely*" on automated processing while the term used in the draft for a revised DPA is "*exclusively.*" So how should the term "solely" in the GDPR or "exclusively" in the draft to the revised DPA be interpreted?

While one could argue that it is already sufficient if a human was included at the end of the process to formally "make the decision," this would defy the purpose of the legal provisions. Rather, it should only be considered that the decision is not based *solely* or *exclusively* on automated processing, if a person "actively exercises any real influence on the outcome of a particular decision-making process" (Bygrave 2000) and actively assesses the result of the automated processing before this person takes the decision also formally.

There are, however, also exceptions to this requirement. One important exception is that the provisions will not apply if the automated process was based on the data subject's explicit consent. According to the GDPR, the data subjects must be provided with information not only of the *existence* of such automated decision making, but also of the *logic involved* and the *significance* and *envisaged consequences* of such processing for the data subject (Article 13(2)(f) GDPR), which also is a necessary foundation for a valid consent.

However, explaining (and understanding) what goes on, for example, in a neural network in terms of a generated outcome (i.e., why is this case decided that way?) is a difficult task, even for an expert (cf. Dong et al. 2017; Stock and Cisse 2017). It will be substantially more difficult to try and explain such issues (or other algorithms) to an average data subject. In particular, if one cannot easily trace the precise path of a neural network to a final answer, the description of automated decisions is open to interpretation. This difficulty may also affect the issue of validity of a data subject's consent, since such consent not only has to be freely given, specifically and unambiguously, but also has to be made on an "*informed*" basis (Article 4 (11) GDPR). And even in cases of valid consent, the data subjects will still have to be informed and the data subjects will have the (1) right to obtain human intervention; (2) right to express their point of view; (3) right to obtain an explanation of the decision reached; and (4) right to challenge the decision (Recital 71 GDPR).

Since algorithms are an important means of governing data streams, assessments of how an automated decision will affect the data subject may have to be made on a regular basis. However, this seems to be an impossible reality, should automated decisions become the norm (Naudits 2016).

3.6 Self-Regulation

In addition, the GDPR as well as the draft for a revision of the DPA to a certain extent provide that self-regulation shall have some legal effects. Self-regulation is

generally considered to be the regulation in a field, for example, an industry, by its own members often using, for example, standards or codes of conduct (such as the Code of Conduct of the German insurance industry),[9] as opposed to legislation set forth by the government.

There are various kinds of self-regulation (i.e., regulation by private, non-governmental entities). Autonomous self-regulation is solely based on the initiative of the private players, while initiated self-regulation is based on private activities initiated by governmental impulses. In some cases, government may even try to steer self-regulation and, thus, achieve a regulated autonomy (regulated self-regulation), such as in the case of data protection.

While there are some disadvantages to self-regulation (e.g., lack of transparency, democratic deficit, putting private and commercial interests over public interests), some of these disadvantages can be addressed in regulated self-regulation. Moreover, self-regulation can also have many advantages: It can avoid further governmental interventions and legal regulations; self-regulation generally is closer to actual practice and the involved parties can introduce their technical expertise. In addition, self-regulation generally is more flexible than governmental regulation and it is easier to react to (technical) changes. Finally, self-regulation can contribute to the good reputation of the field concerned (Widmer 2003, pp. 20–22).

The GDPR as well as the draft for a revision of the DPA introduce the possibility of regulated self-regulation. Article 40 GDPR provides that associations and other bodies representing categories of controllers or processors may prepare codes of conduct for the purpose of specifying the application of the GDPR in certain aspects. Such codes would then have to be submitted to the supervisory authority, which shall approve it, if it complies with the GDPR. In cases where a code of conduct concerns processing activities in several member states, the supervisory authority must, before approval, submit it to the European Data Protection Board for an opinion. If it approves, the European Commission must review the code and, if it also approves, publish it.

Such codes of conducts can be used not only to facilitate cross-border data transfers, but also help to set forth and demonstrate compliance, in particular with regard to security risks of data processing (*see*, e.g., Articles 24, 28, and 32 as well as Recitals 77 and 81 GDPR). Codes of conduct are particularly fit to address legal questions for specific industries, but also other questions of data protection, such as the requirements of privacy by design or privacy by default in specific fields (Bergt 2016, p. 671).

The revision of the DPA goes less far than the GDPR in this aspect. It provides that professional and business associations whose statutes entitle them to defend the economic interests of their members, as well as federal bodies, may submit a code of conduct to the supervisory authority. Thereupon, the authority shall comment on the submitted code and publish its opinion. However, note that the interested parties

[9] Gesamtverband der Deutschen Versicherungswirtschaft (2012). This code is currently under revision due to GDPR adaptations.

cannot derive any rights from a positive opinion or a waiver of an opinion. Nevertheless, in case of a positive opinion from the supervisory authority, it can be assumed that behavior in line with the submitted code of conduct does not entail any administrative measures (Legislative Message DPA 2017, pp. 7034–7035).

Taking this possibility into account, it may make sense, for example, for associations in the field of data science to consider initiating and/or participating in self-regulation projects concerning certain issues, which affect their activities. Not only will this give them the possibility to more closely have an effect on the regulation which concerns them and to mitigate the risks from vague legal provisions, it may also give them the possibility to more quickly influence how (technical) changes of their field are approached from a legal perspective and may even help prevent further sector-specific data protection laws.

4 Regulatory Aspects

Data science obviously does not take place in a vacuum. The application of data science to particular fields and the creation of new data products from a legal perspective will also have to take into account the context of the specific industry data science is applied to. In many industries, there are substantial regulatory requirements that have to be met, not to mention sector-specific data protection provisions to be taken into account. Creating new products without concern to such regulatory frameworks may result in commercial nonstarters or expensive rectifications before commercialization.

Space constraints hinder us from exhaustively listing and explaining such regulatory frameworks for all fields to which data science may be applied. Suffice it to say that among many others, the following fields are particularly regulated and such sector-specific laws will have to be taken into account: banking and finance, insurance, pharmaceutical sector, health care, and telecommunications.

5 Conclusion

In this chapter, we have outlined a number of legal issues that can affect the activities of data scientists.

It seems clear that data carriers should be treated as goods from a legal point of view, and that copyright protects some data or data collections. Also, in the EU databases are to a certain extent protected by a sui generis right (in Switzerland no such right exists) and, in addition, unfair competition law also prevents certain abuse of data. Data scientists should be aware that the legal discussion in this respect has not yet caught up and is still open. Moreover, data scientists should always consider how rights to data affect their activities, for example, whether they have obtained the

necessary rights concerning the data they work with and how such rights may affect the results of their work.

Data protection law is certainly a legal field that data scientists should be aware and have some knowledge of. While many activities of data scientists will not necessarily involve personal data, the risk of re-identification—and its impact on the qualification of data as "personal data"—must always be considered. Among many others, "privacy by design" and "privacy by default" are some of the provisions that have to be taken into account already early in the process of developing a data product. In addition, rules concerning "automated individual decisions" often are of concern to data science. However, it remains to be seen how they will play out in the future. Self-regulation may be one way to address some of the vagueness and uncertainties of data protection laws from the perspective of specific fields of data science and—if effective—may also help to mitigate the legal risks and preclude the potential perception of a need for further sector-specific legislation.

Finally, data scientists should always remain aware that the application of data science to specific fields may also lead to the application of certain industry-specific regulation. It is important that data scientists obtain at least a broad overview of such industry-specific laws and consider their effect on their activities and potential data products already at the beginning of a project.

References

Arpagaus, R. (2013). Commentary on Art. 5 UCA. In R. Hilty & R. Arpagaus (Eds.), *Basler Kommentar, Bundesgesetz gegen den Unlauteren Wettbewerb (UWG)*. Basel: Helbing Lichtenhahn.

Baeriswyl, B. (2013). "Big Data" ohne Datenschutz-Leitplanken. *Digma, 13*(1), 14–18.

Baeriswyl, B. (2014). Big Data zwischen Anonymisierung und Re-Individualisierung. In R. H. Weber & F. Thouvenin (Eds.), *Big data und Datenschutz – Gegenseitige Herausforderungen* (pp. 45–59). Zürich: Schulthess.

Benhamou, Y., & Tran, L. (2016). Circulation des biens numériques: De la commercialisation à la portabilité. *Sic!* (11), 571–591.

Bergt, M. (2016). Verhaltensregeln als Mittel zur Beseitigung der Rechtsunsicherheit in der Datenschutz-Grundverordnung. *Computer und Recht*, (10), 670–678.

Bygrave, L. (2000). Automated profiling, minding the machine: Article 15 of the EC data protection directive and automated profiling. *Computer Law and Security Review, 7*(4). Retrieved February 14, 2018, from http://www6.austlii.edu.au/cgi-bin/viewdoc/au/journals/PLPR/2000/40.html

Cavoukian, A. (2011). *Privacy by design, The 7 foundational principles, Implementation and mapping of fair information practices*. Retrieved February 14, 2018, from https://www.ipc.on.ca/wp-content/uploads/Resources/pbd-implement-7found-principles.pdf

Cherpillod, I. (2012). Commentary on Art. 2 CA. In B. Müller & R. Oertli (Eds.), *Stämpflis Handkommentar SHK, Urheberrechtsgesetz (URG)* (2nd ed.). Bern: Stämpfli.

Dong, Y., Su, H., Zhu, J., & Bao, F. (2017). *Towards interpretable deep neural networks by leveraging adversarial examples*. Retrieved February 14, 2018, from https://arxiv.org/abs/1708.05493

Eckert, M. (2016a). Digitale Daten als Wirtschaftsgut: Besitz und Eigentum an digitalen Daten. *SJZ, 112*(11), 265–274.

Eckert, M. (2016b). Digitale Daten als Wirtschaftsgut: Digitale Daten als Sache. *SJZ, 112*(10), 245–249.

European Commission. (n.d.). *The EU copyright legislation*. Retrieved February 14, 2018, from https://ec.europa.eu/digital-single-market/en/eu-copyright-legislation

FDJP. (2016). *Erläuternder Bericht zum Vorentwurf für das Bundesgesetz über die Totalrevision des Datenschutzgesetzes und die Änderung weiterer Erlasse zum Datenschutz*. Retrieved February 14, 2018, from https://www.bj.admin.ch/dam/data/bj/staat/gesetzgebung/datenschutzstaerkung/vn-ber-d.pdf

Flückiger, A. (2013). L'autodétérimination en matière de données personnelles: un droit (plus si) fundamental à l'ère digitale ou un nouveau droit de propriété? *AJP, 10*(6), 837–864.

Gesamtverband der Deutschen Versicherungswirtschaft. (2012). *Verhaltensregeln für den Umgang mit personenbezogenen Daten durch die deutsche Versicherungswirtschaft*. Retrieved February 14, 2018, from https://www.gdv.de/resource/blob/23938/8db1616525e9a97326e2f2303cf42bd5/download-code-of-conduct-data.pdf

Guadamuz, A. (2017). Artificial intelligence and copyright. *WIPO Magazine* (5), pp. 14–19.

Kiss, A., & Szöke, G. L. (2015). Evolution or revolution? Steps forward to a new generation of data protection regulation. In S. Gutwirth, R. Leenes, & P. de Hert (Eds.), *Reforming European data protection law* (pp. 311–331). Dordrecht: Springer.

KPMG & Bitkom Research. (2016). *Mit Daten Werte Schaffen – Study 2016*. Retrieved February 14, 2018, from https://home.kpmg.com/de/de/home/themen/2016/06/mit-daten-werte-schaffen.html

Legislative Message DPA. (2017). *Botschaft zum Bundesgesetz über die Totalrevision des Bundesgesetzes über den Datenschutz und die Änderung weiterer Erlasse zum Datenschutz* (BBl 2017, pp. 6941–7192).

Narayanan, A., Huey, J., & Felten, E. W. (2016). A precautionary approach to big data privacy. In S. Gutwirth, R. Leenes, & P. De Hert (Eds.), *Data protection on the move* (pp. 357–385). Dordrecht: Springer.

Naudits, L. (2016, August 2). *The right not to be subject to automated decision-making: The role of explicit consent*. Retrieved February 14, 2018, from https://www.law.kuleuven.be/citip/blog/the-right-not-to-be-subject-to-automated-decision-making-the-role-of-explicit-consent

Nelson, G. S. (2015). *Practical implications of sharing data: A primer on data: Privacy, anonymization, and de-identification* (Paper 1884–2015). Chapel Hill, NC: ThotWave Technologies. Retrieved February 14, 2018, from https://support.sas.com/resources/papers/proceedings15/1884-2015.pdf

Oxenham, S. (2016). Legal maze threatens to slow data science. *Nature, 536*(7614), 16–17.

Polzer, G. (2013). Big Data – Eine Einführung. *Digma, 13*(1), 6–10.

Ritter, J., & Mayer, A. (2018). Regulating data as property: A new construct for moving forward. *Duke Law & Technology Review, 16*(1), 220–277.

Rudin, B. (2015). Commentary on Art. 3 DPA. In B. Baeriswyl & K. Pärli (Eds.), *Stämpflis Handkommentar SHK, Datenschutzgesetz (DSG)*. Stämpfli: Zürich/Bern.

Stadelmann, T., Stockinger, K., Braschler, M., Cieliebak, M., Baudinot, G., Dürr, O., & Ruckstuhl, A. (2013). *Applied data science in Europe, challenges for academia in keeping up with a highly demanded topic*. Presented at the European Computer Science Summit 2013, Amsterdam, Netherlands. Retrieved February 14, 2018, from http://www.informatics-europe.org/images/ECSS/ECSS2013/slides/ECSS2013-Stadelmann-paper.pdf

Stock, P., & Cisse, M. (2017). *ConvNets and ImageNet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism*. Retrieved February 14, 2018, from https://arxiv.org/abs/1711.11443

Thouvenin, F. (2017). Wem gehören meine Daten? Zu Sinn und Nutzen einer Erweiterung des Eigentumsbegriffs. *SJZ, 113*(2), 21–32.

Weber, R. H. (2014). Big data: Rechtliche perspektive. In R. H. Weber & F. Thouvenin (Eds.), *Big Data und Datenschutz – Gegenseitige Herausforderungen* (pp. 17–29). Zürich: Schulthess.

Widmer, M. (2003). *Das Verhältnis zwischen Medienrecht und Medienethik*. PhD thesis, University of Zürich, Stämpfli, Bern.