1. Introduction

Vanessa Mak, Eric Tjong Tjin Tai and Anna Berlee

1. A NEW KIND OF SCIENCE

This book deals with one of the most important scientific developments of recent years, namely the exponential growth of data science.[1] More than a savvy term that rings of robotics, artificial intelligence and other terms that for long were regarded as part of science-fiction, data science has started to become structurally embedded in scientific research. Data, meaning personal data as well as information in the form of digital files,[2] has become available at such a large scale that it can lead to an expansion of knowledge through smart combinations and use of data facilitated by new technologies. This book examines the legal implications of this development. Do data-driven technologies require regulation, and vice versa, how does data science advance legal scholarship?

Defining the relatively new field of data science requires a working definition of the term. By data science we mean the use of data (including data processing) for scientific research. The availability of massive amounts of data as well the relatively cheap availability of storage and processing power has provided scientists with new tools that allow research projects that until recently were extremely cumbersome if not downright impossible. These factors are also often described with the term 'big data', which is characterized by three Vs: volume, velocity and variety.[3] The term

[1] The term exponential is fitting here, since the emergence of data science has followed the so-called Moore's Law, which predicted in 1965 that computer power would grow by a factor of 2 every 1.5 years. See V Mayer-Schönberger and K Cukier, *Big Data* (Mariner Books 2014) 35.

[2] The term 'personal data' has a specific meaning in European law; see Art 4(1) of Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data [2016] OJ L119/1. 'Data' can also refer to the narrower category of digital files. See eg TFE Tjong Tjin Tai, 'Data in het vermogensrecht' [2015] Weekblad voor Privaatrecht, Notariaat en Registratie (WPNR) 993.

[3] A McAfee and E Brynjolfsson, 'Big Data: the Management Revolution' [2012] *Harvard Business Review* 60–69; D Laney, *3D Data Management: Controlling Data Volume, Velocity and Variety* (META Group Inc 2001). Two other Vs that are

data science is nonetheless broader, because it can also refer to the use of data sets that are large but still limited—and therefore, unlike big data, of a manageable size for processing.

The term data science can also refer to more specific methods of working with data, such as data mining or machine learning. Furthermore, there are additional benefits to interdisciplinary cooperation, as scientists from different disciplines are beginning to share tools and analytical methods. Data science as a practice tends to transcend individual disciplines and begins to show characteristics of a discipline of its own. In that sense we can speak of data science also as a single discipline, provided we bear in mind that using data has no advantages as long as it is not aimed at results that are worthwhile for specific disciplines in the end.[4]

The widespread interest in data science methods and applications also brings to the fore new questions and risks that in the past did not concern scientists. (Informational) privacy, integrity of data, and data ownership are issues that have (only) needed serious attention since the rise of the mainframe computer and the internet.[5]

For law and legal research the rise of data science is important for two reasons. First of all, data science gives rise to many legal issues which have not yet been fully investigated. These issues are not limited to data science, on the contrary they are often shared with businesses that handle data. Secondly, data science methods can be applied to law as a discipline. What advantages and risks are involved with applying data science to law?

Given the rapid developments of recent years, there is as yet relatively little literature dedicated to data science and law. The present volume intends to fill that gap. It presents an overview of the current issues and the state of research in the field, as well as an outlook on the research

---

sometimes added to these three are veracity and value; see B Marr, 'Why Only One of the 5 Vs of Big Data Really Matters', Blog IBM Big Data & Analytics Hub, 19 March 2015, available at: <http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters>; A Lafarre, 'Recht voor big data, big data voor recht' [2016] Computerrecht 146, 147.

[4] The use of data science in law is the subject of Part II of this book. See for an introduction below, section 6.

[5] D J Solove and P M Schwartz, *Information Privacy Law* (Wolters Kluwer Law & Business 2011), 2; V Mayer-Schönberger, P E Agre and M Rotenberg, 'Generational Development of Data Protection in Europe' in P E Agre and M Rotenberg (eds), *Technology and Privacy: The New Landscape* (The MIT Press 1997) 219–41; L Floridi, *The 4th Revolution: How the Infosphere is Reshaping Human Reality* (Oxford University Press 2014) 115. A F Westin and D J Solove, *Privacy and Freedom* (ig publishing 2015) 176.

questions that are likely to inspire research in data science and law in the near future. We have been fortunate in attracting the cooperation of a group of leading experts in comparative law, regulation, technology, and issues related to the information society in which we live. In addition, our volume provides space for a number of younger, upcoming legal scholars to present their contributions. The diversity of authors in terms of area of legal expertise, and in the jurisdictions that they represent in Europe and further afield, means that we are able to present a broad palette of current issues that should paint a fairly good picture of the state of play in the field of data science and law. As a starting point for further research, the book should hopefully inspire discussions with colleagues in other jurisdictions, in particular in Asia and in the United States, which together with Europe globally are the leading breeding grounds for innovation in data-driven technologies.

Before going further into detail, we will briefly discuss some key concepts of data science, interwoven with a historical overview. Subsequently, we will discuss the kinds of issues that data science gives rise to from a legal perspective. Finally, we will discuss the application of data science to law. At this point, we will also suggest why law, in comparison to other disciplines, appears to be fairly late in using data science.

2. DATA AS DIGITIZED INFORMATION

When we consider data science, we may be inclined to gloss over the key concept of data. Traditionally and historically scientists have always used data, in the simple meaning of information. The collection of data is intrinsic to what we understand by science. The works of Aristotle[6] can already be seen as compilations of information in classical Greece. Modern science has from the beginning integrated collections of measurements (data) with the analysis and formation of theories.[7]

However, in computer science 'data' has taken on a new and more specific meaning. A computer cannot handle anything except insofar as it has become tractable for the computer, which in essence means that it has a digital representation. This means that analogue, real-world

[6] Such as his History of Animals, containing a wealth of accurate zoological observations.

[7] E.g., the collection of measurements of astronomical data by Tycho Brahe allowed Johannes Kepler to develop laws for describing planetary orbits. Similarly, scientists such as Francis Bacon (in his New Organon) stressed the importance of systematically collecting data in tables.

information needs to be digitized to allow automated processing.[8] When we talk about data science, meaning the use of computers and automated tools for processing data, we necessarily talk about digitized information.

From an information theoretical viewpoint this may appear irrelevant: information has the exact same content, regardless as to whether it is digitized or not.[9] Digitization does not add information.[10] However, for practical purposes the digitization of information makes a qualitative difference: it opens up avenues for research that were closed beforehand. For example, the widescale scanning of books by Google Books, the scanning of old newspapers,[11] or the unlocking of public records online (via open data portals): such as case law,[12] court records,[13] and historical property records,[14] have not added any information as such. But the consequent possibility of full-text searching now makes it feasible to count the number of occurrences of a specific term, thereby giving an indication as to when this term became in vogue.[15] Earlier such 'counting' would be theoretically possible as a manual task, but would not be undertaken except in rare cases due to the amount of time and work involved.[16] With the scanned

---

[8] The importance of digitization is not to be overlooked, see <https://en.wikipedia.org/wiki/Digitization>, and J S Brennen and D Kreiss, 'Digitalization' in *The International Encyclopedia of Communication Theory and Philosophy* (Wiley 2016), DOI: 10.1002/9781118766804.wbiect111.

[9] Information, following the seminal article of Claude Shannon ('A Mathematical Theory of Communication' (1948) 27 *Bell System Technical Journal* 379 and 623), may be considered as simply the actual message transmitted in communication, whereby the specific representation of the information is irrelevant, and whereby the transmitted communication may contain additional elements to enhance robustness (redundancy).

[10] See on the difference between digitization and digitalization which is as 'the way in which many domains of social life are restructured around digital communication and media infrastructures': S Brennen and D Kreiss, 'Digitalization and Digitization', Blog Culture Digitally, 8 September 2014, available at <http://culturedigitally.org/2014/09/digitalization-and-digitization/>.

[11] E.g., <http://www.britishnewspaperarchive.co.uk> or <http://www.nytimes.com/ref/membercenter/nytarchive.html?mcubz=0>.

[12] See <https://www.rechtspraak.nl/Uitspraken-en-nieuws/Uitspraken/Paginas/Open-Data.aspx> (Netherlands), and <http://www.belgielex.be/info/db/cassatie-cassation/juridat/nl/index.html> (Belgium).

[13] See <https://www.pacer.gov/>.

[14] See <http://digitalarchives.landregistry.gov.uk/1862/search>.

[15] See for example the Google Books Ngram Viewer at <https://books.google.com/ngrams>.

[16] For public records see R Gellman, 'Public Records—Access, Privacy, and Public Policy: a discussion paper' [1995] 12 *Government Information Quarterly* 393.

newspapers it is relatively easy to make a ranking of the most frequently used terms, and to see historic trends.[17] This would have been unfeasible on the basis of manual analysis.

Hence, we will understand 'data' primarily as meaning 'digital information'. Unfortunately, the necessities of the English language will also require us to sometimes speak of data in the traditional sense, as a synonym of a collection of information. We will speak of 'analogue data' when we do this, to avoid misunderstanding.

3. WHY DATA SCIENCE NOW?

One might wonder why data science has only become prominent in scientific research so recently. To understand the factors that have driven this development it is useful to present a brief sketch of technological advancement in the last half century. That overview also serves to highlight several conditions that need to be present for data science to come to fruition.

Evidently data science would not be possible without computers. But computers have been available for half a century, while the kinds of things that data scientists do were not done as a matter of course. The simple explanation is that computing power was limited. Usually commentators point to Moore's law which predicts a doubling of computing power (CPU speed) every 1.5 years.[18] However, there is more involved than the CPU itself. Driven by the need for advanced graphics for computer games several chip manufacturers such as Intel and Nvidia have created advanced graphical processing units (GPUs) which make it possible to calculate the required matrix calculations very quickly. Furthermore, the internal storage of CPUs, both cache memory and RAM, has increased exponentially. An IBM PC around 1990 had available only 640 Kb RAM. In 2000 a common Windows PC would have available some 1 Mb RAM, which is only 50 per cent more than ten years earlier. Nowadays even a cheap laptop has 4-8 Gb RAM, which is a 4000 per cent increase in 15 years.[19] Even a smartphone such as the iPhone 8 which became available in 2017 and had 2 Gb RAM!

---

[17] For case law digitization allows citation counting and network analysis, see the overview in R Whalen, 'Legal Networks: The Promises and Challenges of Legal Network Analysis' (2016) *Michigan State Law Review* 539.

[18] See n 1.

[19] See also <http://www.relativelyinteresting.com/comparing-todays-computers-to-1995s/>, comparing 1995 to 2012.

RAM and cache memory are important as they allow computers to do many operations quickly.[20] That process is similar to the way in which you can quickly combine two things you know if you can retrieve them from your own memory, while this combination takes much more time if you have to look this up from books or even from the internet.

The combination of these developments in computing power (most within the last ten years) have led to a significant speed-up of processing, which makes research feasible that was not possible before without expensive, specialized hardware. An example is a pilot project run at Tilburg University, where we process several Gb of Dutch case law data. This can be done on a several years old Dell desktop used for office work.[21] Processing the full data set with a non-optimized program took 75 minutes for a single year of data.[22] On a 2015 Apple MacBook Air this only took eight minutes for each year.[23] This makes the difference of having to plan each analysis ahead and reserve a day for obtaining results, and being able to do frequent test runs and analysis in the course of a few hours.

Another driver of data science is that programming has further matured. A wealth of new programming languages has become available, with extensive libraries, providing many functions, allowing faster and easier development.[24] One contributing factor is again processing power, whereby more user-friendly languages can still be executed rapidly.

Finally, other factors that have stimulated the rise of data science are the increased use of open source software and the availability of cheap data storage. Many crucial technologies—programming languages such as Python, operating systems such as Linux, and programs such as Apache—are available for free and low everyone to share improvements and developments. Furthermore, data science processing requires the availability of huge sets of data, which is facilitated by a significant fall in the costs of data storage disks.

At the same time, organizations are becoming aware of the value of data, and increasingly undertake projects to collect data in branches of

---

[20] A related performance gain is caused by the availability of cheap Solid State Drive (SSD) for mass storage, which are about 5 times faster than traditional Hard Disk Drives (HDD).

[21] Intel i3 Core, 4 Gb RAM, 1 Tb HDD. This was fairly up to date consumer hardware around 2010–12.

[22] Processing the available cases in 2016 would mean going through 148 603 cases.

[23] 2.2 GHz Intel Core i7, 8 Gb RAM, 512 Gb SDD.

[24] In particular the Python programming language is generally viewed as being easy to use for non-time-critical applications. E.g., K C Louden and K A Lambert, *Programming Language: Principles and Practices* (3rd edn, Cengage 2012) 38.

consumer goods and services provision, primarily with the aim of obtaining data on customer preferences and behaviour.

This combination of factors explains why data science has gained prominence in the last decade.[25] The next section explains in more detail what data science actually is.

4. THE TOOLS AND POSSIBILITIES OF DATA SCIENCE

Data science typically operates with collections of data, large data sets. This differs from traditional statistical analysis in social science or economics where typically great care was taken in data collection through field work, surveys or experiments. The data used in data science is often obtained through fairly simplistic or automatic collection methods, or a by-product from other practices. Examples are scanning of books by Google Books, or collection of communication traffic data that is obtained in the course of operating an Internet Service Provider. New data is generated by combinations of data sets from various sources, which is done with a variety of IT tools and programming languages. Even very high volumes of data can be handled with ease.

Data may be analysed by the use of fairly traditional statistical methods, using computer programs: the classic SPSS, or more modern environments like R. Less sophisticated tools may be useful as well: simple algorithms in programming languages like Python or database systems like SQL can fairly easily gather interesting data and create tables and records from which interesting research conclusions can be derived, such as the earliest use of a certain term, a list of court decisions ordered by the number of references to each decision. Other tools make network analysis easy: popular programs like Gephi or Cytoscape allow graphical analysis of networks. For full-text search the Apache Lucene framework is available, as well as further developments like ElasticSearch. Many of the tools listed above are available for free as open source software.

A further related development is the use of advanced special purpose algorithms. This is the area of Artificial Intelligence. Nowadays the term 'deep learning' has become fashionable.[26] Data may be analysed without

---

[25] An additional driver of interest are the new techniques for what is now called 'deep learning', which have been available only since 2008.

[26] This refers to a particular technique for realizing Artificial Intelligence, in which neural networks with several layers (and therefore 'deep') are trained ('learn').

requiring a set of instructions that allow the programmer to determine the outcome in advance: instead, sophisticated algorithms (neural networks) are able to learn from being exposed to a lot of data and being pointed in the right direction.[27] This kind of training provides new kinds of results.

This example shows that one possible aspect, in contradistinction to traditional IT analysis, is that it is no longer necessary to start from a fixed database structure. Instead one may operate freely with heterogeneous data sources which can easily be combined by on-the-fly transformations with a variety of tools. In this respect, data science resembles what is also called 'big data'. Indeed, one might think that data science simply is the concept of big data applied to the sciences.[28] To some extent this is true: the techniques used can be applied both in business environments as well as in scientific research. However, data science is different in that the stringency of science imposes further demands on methodology. It is not sufficient that you produce pretty graphs or interesting results, but you must be able to prove that these results are actually scientifically valid, are 'true' or correct. This imposes restrictions and additional checks, such as:

- integrity of the databases (i.e., the data must not be unknowingly contaminated or incomplete)
- correctness of processing the data (selecting, merging, combining data)
- correctness in analysis: statistical methods and network analysis must be done in an appropriate manner.

It is, however, true that many issues involved with big data may also appear in a discussion of data science. This applies in particular to legal issues, to which we turn now.

5. LEGAL ASPECTS OF DATA SCIENCE

When we talk about the legal aspects of data science, what we refer to primarily is the question: do new data-driven technologies require regulation?

---

[27] See in particular what is called 'deep learning'. Further Chapter 4 on liability for (semi)autonomous systems.

[28] To be true, data science is also used to refer simply to the application of data science techniques for business purposes.

That question is perhaps less innovative than the question whether data science in law—the subject of the next section—is developing into a self-standing discipline. Yet, even within the more traditional question about regulation, there is room for innovation. The emergence of data science can be seen as disruptive,[29] not just in relation to society or to markets, but also in relation to law.

For law, the first disruption is that data-driven technologies raise questions of justice. Algorithms or automated decision-making systems have been used for decades by public administrations as a means to efficiently streamline their tasks.[30] The tax authorities in many countries, for example, make use of automated decision-making systems for generating annual tax decisions, and executing them.[31] This practice can appear as undemocratic due to a lack of transparency—the algorithm can be a 'black box'—and the difficulty that addressees of decisions might have in challenging them.[32] Similar questions of justice arise when public tasks are transferred to private organizations who make use of algorithms,[33] such as Uber,[34] and when

---

[29] The term 'disruptive' has been used in particular in relation to the rise of the sharing economy, with firms such as Uber and Airbnb breaking into regulated markets, in these cases the market for taxi services and the market for hotel services. See e.g., A Sundararajan, *The Sharing Economy. The End of Employment and the Rise of Crowd-Based Capitalism* (MIT Press 2016). Economists have questioned whether platforms such as Uber actually fulfil the conditions for disruptive innovation in economic theory; see C M Christensen, M E Raynor and R McDonald, 'What is Disruptive Innovation?' (2015) *Harvard Business Review*, available at: <https://hbr.org/2015/12/what-is-disruptive-innovation>.

[30] See for more on this Chapter 11.

[31] See for more on this Chapter 13.

[32] J Bing, 'Code, Access and Control' in M Klang and A Murray (eds), *Human Rights in the Digital Age* (Cavendish Publishing 2005); M Bovens and S Zouridis, 'From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control' (2002) 62 *Public Administration Review* 174; F. Pasquale, *The Black Box Society* (Harvard University Press 2015), pp. 156 et seq.

[33] Or when public authorities make use of information collected by private organizations, see D J Solove 'Access and Aggregation: Public Records, Privacy and the Constitution', *Minnesota Law Review* 86/6, p. 1140 and C J Hoofnagle 'Big Brother's Little Helpers: How Choicepoint and Other Commercial Data Brokers Collect and Package Your Data for Law Enforcement', *North Carolina Journal of International Law and Commercial Regulation* 29/4, pp. 595–637. See on this also Chapter 11, section 4.

[34] E.g., Uber's algorithm sets the price of the service, which may result in low prices and hence low payment for drivers when the supply is great; see S O'Connor, 'When Your Boss is an Algorithm', *Financial Times*, 8 September 2016.

data-driven technologies are used for profiling in criminal prosecutions.[35] In private law relationships the question of justice might translate into one of (contractual) fairness.[36]

Besides justice, other concerns can arise in relation to fundamental rights such as data protection or privacy. Data protection and privacy have been, and continue to be, the subject of specialized research and this book does not therefore specifically focus on these issues. Instead, less highlighted areas are brought to the fore, such as the economic regulation of data-driven technologies through property, contract and liability law; the use of data-driven technologies in public administration and criminal prosecution; and the use of data science methods in law.

Third, looking at rules of law more specifically, data-driven technologies often have aspects that do not fit neatly into existing legal categories or concepts. Software and data in general, for example, challenged traditional notions of property law such as what can be subject to the right of ownership. Software is no longer (only) shipped using a physical medium such as a floppy disk, CD-ROM or DVD-ROM, which conveniently merged the question of ownership of a copy of software with ownership of the physical carrier. With the transfer of software being detached from a tangible or corporeal carrier, the question of whether a copy of software or a data-file can be owned,[37] becomes more difficult to answer.[38] The object, data, cannot be considered tangible property, and it is questionable whether it is intangible property akin to a claim. Data, therefore, does not neatly fall into the traditional categories of objects over which property rights can be exercised. This chiefly raises the question, whether this is problematic.[39] Or, should perhaps rights in relation to software, and data in general, be

---

[35] See in this volume Chapter 10. See also B Schneier, *Data and Goliath. The Hidden Battles to Collect Your Data and Control Your World* (WW Norton & Company 2015).

[36] For a discussion of the application of unfair terms regulation to digital consumer contracts, see N Helberger, 'Big data en het consumentenrecht' in PH Blok, Big data en het recht (Sdu 2017) 151, 159 ff. See also Chapter 2 and Chapter 5, section 4.

[37] See on software and ownership also N Härting, '"Dateneigentum" – Schutz Durch Immaterialgüterrecht?' (2016) 10 Computer und Recht 646. See also Chapter 6 of this volume.

[38] See for different approaches, Michael Dorner, 'Big Data und "Dateneigentum"' (2014) 9 Computer und Recht 617.

[39] A similar question can be asked in relation to property rights in data, see, H Zech, 'Data as a Tradeable Commodity', in: A De Franceschi (ed.), *European Contract Law and the Digital Single Market* (Intersentia 2016), 53 and H Zech, 'Information as Property' (2015) *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 192. Reinout Wibier, 'Big Data En

kept out of the scope of property law, and (remain to) be governed by intellectual property law such as copyrights and database rights,[40] or where it concerns personal data by data protection regulation?[41] If there is a role to play for property law in data,[42] a 'sui generis' right of data ownership may then need to be created.[43] The use of data has furthermore enabled companies to target product advertising to specific consumers based on the preferences that have been gathered from earlier purchases. Amazon, for example, provides personalized recommendations for other products based on the history of purchases that a customer has made earlier. One legal question that arises from this practice is whether the rules of consumer law should be applied as they have always done in the non-digital world. The notion of a 'consumer' is very general and is often defined as a natural person not acting in the course of a business or profession. But if the actual consumer is known to the trader—including perhaps his extensive expertise on the goods he buys—should the same rules apply to him

---

Goederenrecht' [2016] Weekblad voor Privaatrecht, Notariaat en Registratie 7110.

[40] See Chapter 7 in this volume. See for example criticism of the EU efforts to introduce a property right in data: P B Hugenholtz, 'Data property in the system of intellectual property law: Welcome guest or misfit', Paper presented at the conference 'Trading Data in the Digital Economy: Legal Concepts and Tools', Muenster Colloquium on EU Law and the Digital Economy, University of Muenster, 4–5 May 2017. Available at <https://www.ivir.nl/publicaties/download/Data_property_Muenster.pdf>.

[41] The question of whether personal data can perhaps (better) be protected by granting the data subject property rights over their personal data has been subject of extensive discussion already, see extensively N Purtova, Property Rights in Personal Data: A European Perspective (dissertation, Uitgeverij BOXPress 2011).

[42] For example, see J H M van Erp, 'Ownership of Digital Assets?' (2016) 5 *European Property Law Journal* 73. See further Chapter 6 in this volume. Sjef van Erp, 'Ownership of Digital Assets?' (2016) 5 European Property Law Journal 73.

[43] J H M van Erp and W Loof, 'Eigendom in het algemeen: eigendom van digitale inhoud (titel 1)' in Boek 5 BW van de toekomst. Over vernieuwingen in het zakenrecht (Sdu 2016) 23. Van Engelen suggests such a property right to 'bits & bytes' may have already been created by the CJEU in the UsedSoft-case CJEU 3 July 2012 ECLI:EU:C:2012:407, Case C-128/11 (*UsedSoft GmbH v Oracle International Corp*), see D van Engelen, 'UsedSoft v Oracle: The ECJ Quietly Reveals a New European Property Right in "Bits & Bytes"' (2012) 1 *European Property Law Journal* 317, cf. R M Wibier and J Diamant, 'Oracle gaat niet over eigendom maar over contractsvrijheid' [2012] Nederlands Juristenblad 2966.

as to other, less sophisticated consumers?[44] The emergence of data-driven technologies therefore requires a re-assessment of existing legal categories and concepts.

These three aspects—justice, fundamental rights and legal classification—give some pointers as to the issues that need to be considered when we ask whether data-driven technologies require regulation. We will see in the following chapters that other issues may arise.

Part I of this volume is structured according to legal area. The effects, possibilities and downsides of the use of data-driven technologies are looked at from twelve different areas of the law. From consumer, contract and liability law, to (intellectual) property law, corporate law, competition law, criminal law, and administrative law. Each of these areas of the law is affected by data science, and grapples in its own way with how to regulate data science. Part I is concluded with a chapter regarding data localization which exemplifies the difficulties associated with regulating something which is neither fungible nor finite and difficult to pin down to a particular geographical place.

6. DATA SCIENCE FOR LAW

Up to quite recently there were hardly any applications of data science to legal research. Most research seemed to be done from 'law-and' research, as other disciplines such as economics and sociology have for far longer used large collections of (non-digital and digital) data for research. An example is the 'legal origins' debate which arose from a law and economics research into the influence of the kind of legal system (civil law versus common law) to the development of a state.[45] The methods used for that kind of research were, however, fairly traditional economic methods that did not attract attention as such.

Similarly, network analysis has become huge after the seminal work of Barabasi.[46] In other disciplines this has been taken on enthusiastically.

---

[44] These and other issues concerning consumer and contract law are discussed in Chapter 2 of this volume.

[45] See the seminal article R La Porta, F Lopez-de-Silanes, A Shleifer and R. Vishny, 'Law and Finance' (1998) 106 *Journal of Political Economy* 1113, and criticism, for example G Tullock, 'The Case against the Common Law' in F Parisi and CK Rowley (eds), *The Origins of Law and Economics* (Edward Elgar Publishing 2005) 464–74.

[46] Albert-László Barabási, *Linked: The New Science of Networks* (Perseus 2002).

Within law such studies have started slowly, although gaining impetus recently.[47]

The lack of interest in data science among lawyers may be explained by a variety of circumstances.

First, in law there is a relative lack of publicly available data. Although legislation is in itself free, and court decisions are available for public inspection, most case law used to be only available in paper and digital forms from publishers.[48] Only since around 2000 did courts start to publish decisions on their own websites. Even then, most courts still only provide access through a search field or a list of decisions, and do not allow downloading of the whole data set.[49] Legal researchers themselves have not yet collected many data sets for sharing.

Secondly, complexity of the object of legal research (law) is an obstacle to doing data science. Law involves complicated conceptual relationships that cannot easily be derived from the available material texts. For example, analysing a case requires qualification of the facts in the light of a variety of applicable doctrines, while each doctrine can only be understood as a set of rules to be derived from a combination of statutory rules and case law rules, all of which are laid down in natural language with its inherent fuzziness. This may go some way to explain the lack of practical success over decades of research into AI and law.

Thirdly, a peculiarity of law is that law is bound to a jurisdiction. Hence legal researchers are mostly concerned with data for their own jurisdiction only, and have to grapple with numerous idiosyncratic details for their own system. This makes it hard to learn from data analysis by foreign researchers, as each system requires a new approach.

---

[47] The perceptive overview article of R Whalen, 'Legal Networks: The Promises and Challenges of Legal Network Analysis' (2016) *Michigan State Law Review* 539 states that the first studies appear starting from 2007. See further N Peterson and E V Towfigh, 'Network Analysis and Legal Scholarship' (2017) 18 *German Law Journal* 695, 696, also J Frankenreiter, 'Network Analysis and the Use of Precedent in the Case Law of the CJEU – A Reply to Derlén and Lindholm' (2017) 18 *German Law Journal* 687. Mark A Hall, Ronald F Wright, 'Systematic Content Analysis of Judicial Opinions' (2008) *California Law Review* 96/1, p. 63–122. See further on the possible relevance of network analysis for law: T F E Tjong Tjin Tai, De grenzen van het privaatrecht (Tilburg University 2011).

[48] M van Eechoud and L Guibault, 'International copyright reform in support of open legal information', draft 2016, Open Research Symposium Madrid 2016, available at <http://www.ivir.nl/publicaties/download/OpendataCopyrightReform_ODRSdraft-WP_sep16.pdf>.

[49] An exception is the Dutch judiciary, offering all public court decisions on rechtspraak.nl in a large ZIP-file. See n 12 above.

Fourthly, lawyers are by training used to applying hermeneutic, qualitative techniques for analysis of texts. They, may therefore be disinclined and untrained for turning to quantitative research. Also, it is uncommon for lawyers to have extensive training or experience in coding.

There has been considerable progress in the area of automated searching of case law (and literature).[50] But these developments were based on proprietary software and algorithms, hence have not furthered legal research using data science. It has only made traditional legal research easier by facilitating finding and accessing legal materials. The analysis of the materials themselves, however, still required human labour consisting of reading the texts by a legally trained researcher.

However, the tide seems to have turned. The rising interest in big data, as well as the fear of being 'disrupted' by Legal Tech[51] may have given greater urgency to lawyers to pick up data science. The question that arises is: how to do this?[52] What are the risks and opportunities involved? Can we learn from data science in other disciplines? Are there idiosyncrasies in law that require particular care? In Part II of this volume several authors address these and other issues in a series of contributions. Chapter 15 by *Custers* serves as an excellent introduction to the different methods of data science that may be useful for law and its practitioners. However, employing data is not without its legal effects, which may touch upon fundamental rights of those about whom decisions are made based on the use of data, as Chapter 16 explores, and may extend beyond the rights to privacy and data protection. Chapter 17 in turn explores the possibility of the vast amount of data

---

[50] See among others for the US: R C Berring, 'Collapse of the Structure of the Legal Research Universe: The Imperative of Digital Information' (1994) 69 *Washington Law Review* 9. As well as in the area of visualization of networks in case-law, such as Ravel law (now part of LexisNexis).

[51] We will not discuss Legal Tech as such. Legal Tech is a movement driven mostly from legal practice. In essence, it involves the application of new technology, in particular data technology or 'data science' broadly speaking, to the advancement of legal practice. This can consist simply of applying data science analysis methods to legal problems. It might also involve fairly mundane applications (from a scientific point of view) of technology, such as a tool to facilitate citing case law in a legal document. Think of the way EndNote or Zotero is a research tool: it does not lead to research conclusions, but it does help researchers in their work.

[52] Two noteworthy descriptions of this kind of research are Frank Fagan, 'Big Data Legal Scholarship: Toward a Research Program and Practitioner's Guide' (2016) 20 *Virginia Journal of Law & Technology* 1 and Kevin D Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age* (Oxford University Press 2017). Furthermore see M. Truyens, P. Van Eecke, 'Legal Aspects of Text Mining' (2014) 30 *Computer Law & Security Review* 153.

that can be collected and utilized to develop 'granular legal norms' that are tailored to the individual addressees. Also tailored to an individual is the ruling of a judge in a particular case. The penultimate chapter, Chapter 18, discusses the particular challenges and transformation that data science brings to the judiciary. In it, the authors develop a framework which may assist in the decision of whether a particular technology should be applied by the judiciary.

Finally, in Chapter 19 we conclude by providing a brief overview of some recurring themes that emerged from the different chapters and some of the answers and questions that they have generated. As such this chapter might conclude the book but this will not be the last word you hear of this topic.