Introduction to the research handbook in data science and law (second edition)

Mak, V.; Tjong Tjin Tai, T.F.E.; Berlee, A.

1. Introduction to the *Research Handbook in Data Science and Law (Second Edition)*

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. The first edition of this book examined the legal implications of this development. Do data-driven technologies require regulation, and vice versa, how does data science advance legal scholarship?

In the Introduction of the first edition of this *Research Handbook* we stated that the field of data science was a relatively new field. Six years later, the novelty has faded and the practice of using data (including data processing) for scientific research, has become more common. Also, regulation, judicial lawmaking and legal scholarship have developed new perspectives on data, in some fields more than in others.

For science, the availability of massive amounts of data as well the relatively cheap availability of storage and processing power has provided scientists with the tools that allow research projects that until a few years ago were extremely cumbersome if not downright impossible. In the first edition of this *Research Handbook* in 2018 we

---

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 e.g. TFE Tjong Tjin Tai, 'Data in het vermogensrecht' [2015] Weekblad voor Privaatrecht, Notariaat en Registratie (WPNR) 993.

discussed these developments in terms of ‘big data’, which is characterised by three Vs: volume, velocity and variety.[3]

Nevertheless, the term data science is nonetheless broader than ‘big data’ because it can also refer to the use of data sets that are large but still limited and therefore also manageable without resorting to the use of Large Language Models (LLMs).

As already made clear above, 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 sharing tools and analytical methods. Data science as a practice tends to transcend individual disciplines and is showing 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. Data protection, 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 still relatively little literature dedicated to data science and law. The present updated 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 questions that are likely to inspire research in

---

[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 sometimes added to these three are veracity and value; see Emerging India Analytics, ‘Understanding the 5 V’s of Big Data: Unveiling its Characteristics’ (*Medium*, 17 August 2023) available at: <https://medium.com/@analyticsemergingindia/understanding-the-5-vs-of-big-data-unveiling-its-characteristics-27ec4a0f35a6> accessed 4 April 2024; 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 volume. See for an introduction below, section 6.

[5] DJ Solove and PM Schwartz, *Information Privacy Law* (Wolters Kluwer Law & Business 2024) Ch. 1, para A; V Mayer-Schönberger, PE Agre and M Rotenberg, ‘Generational development of data protection in europe’ in PE Agre and M Rotenberg (eds), *Technology and Privacy: The New Landscape* (The MIT Press 1997) 219–241; L Floridi, *The 4th Revolution: How the Infosphere is Reshaping Human Reality* (OUP 2014) 115. AF Westin and DJ Solove, *Privacy and Freedom* (Ig publishing 2015) 176.

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 DIGITISED 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 information needs to be digitised 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 digitised information.

---

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.

8 The importance of digitisation is not to be overlooked, see <https://en.wikipedia.org/wiki/Digitization> accessed 4 April 2024, and JS Brennen and D Kreiss, 'Digitalization' in *The International Encyclopedia of Communication Theory and Philosophy* (Wiley

From an information theoretical viewpoint this may appear irrelevant: information has the exact same content, regardless as to whether it is digitised or not.[9] Digitisation does not add information.[10] However, for practical purposes the digitisation 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 scanned newspapers it is relatively easy to make a ranking of the most frequently used terms, and to see historic trends.[17] Moreover, digitisation allows—barring any legal impediments—the data to be used for training machine learning models, to perform more complex pattern recognition. All of which would have been unfeasible on the basis of manual analysis.

2016) <https://onlinelibrary.wiley.com/doi/10.1002/9781118766804.wbiect111> accessed 4 April 2024.

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 digitisation and digitalisation 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: <https://culturedigitally.org/2014/09/digitalization-and-digitization/> accessed 4 April 2024.

11 See <www.britishnewspaperarchive.co.uk/> or <https://help.nytimes.com/hc/en-us/articles/115014772767-New-York-Times-Archived-Articles-and-TimesMachine?mcubz=0> accessed 4 April 2024.

12 See <www.rechtspraak.nl/Uitspraken/Paginas/Open-Data.aspx> (Netherlands), and <https://belgielex.be/en/jurisprudence> (Belgium), accessed 4 April 2024.

13 See <https://pacer.uscourts.gov/> accessed 4 April 2024.

14 See <https://digitalarchives.landregistry.gov.uk/start/overview> accessed 4 April 2024.

15 See for example the Google Books Ngram Viewer at <https://books.google.com/ngrams/> accessed 4 April 2024.

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

17 For case law digitisation 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.

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 in the last decade. 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 perform the required matrix calculations very quickly. This calculating power has in recent years generated a run on GPUs for the purposes of training AI models (and mining bitcoin).[19]

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% more than 10 years earlier. Nowadays even a cheap laptop has 4–8 Gb RAM, which is a 4000% increase in 15 years.[20] Even a smartphone such as the iPhone 6 which became available in 2014 already had 1 Gb RAM and the latest model iPhone 15 available in 2023 starts with 6 Gb RAM.

RAM and cache memory are important as they allow computers to do many operations quickly.[21] 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.

---

18 See (n 1).

19 See also <www.nytimes.com/2023/08/16/technology/ai-gpu-chips-shortage.htmlwww> accessed 4 April 2024, discussing the shortage of GPUs this had generated.

20 See also <www.relativelyinteresting.com/comparing-todays-computers-to-1995s/> accessed 4 April 2024, comparing 1995 to 2012.

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

The combination of these developments in computing power (most within the last 10 years) have led to a significant speed-up of processing, which makes research feasible that was not possible before without expensive, specialised hardware. An example is a pilot project run at Tilburg University, where we process several Gb of Dutch case law data.[22] This can be done on a decade-old Dell desktop used for office work.[23] Processing the full data set with a non-optimised program took 75 minutes for a single year of data.[24] On a 2015 Apple MacBook Air this only took 8 minutes for each year.[25] On a 2020 Apple MacBook Air M1 the same process took a mere 1 minute and 19 seconds for a single year.[26] Performance improvements like these make the difference between 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.[27] 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 and R, operating systems such as Linux, and programs such as Apache—are available for free and allow 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 for data storage disk. Not only can one purchase physical storage at a lower price per unit, but there's also the flexible option of leveraging a cloud storage solution whereby disk space is rented.

At the same time, organisations are becoming aware of the value of data, and increasingly undertake projects to collect ever increasing amounts of data in branches

---

22 The figures are for showing the relative change in speed for similar cost. Professional use and research would, of course, employ more costly hardware that can perform several orders of magnitude faster. A popular choice nowadays is to rent a server with a cloud provider, and running analyses in the cloud: this does not require an investment in hardware and only incurs costs for each individual run.

23 Intel i3 Core, 4 Gb RAM, 1 Tb HDD. This was fairly up to date consumer hardware around 2010–2012.

24 Processing the available cases in 2016 would mean going through 148,603 cases.

25 2,2 GHz Intel Core i7, 8 Gb RAM, 512 Gb SDD.

26 M1 processor, 16 Gb RAM, using an external SSD (Samsung T5) over USB 3.1. The comparison is not direct as this result uses Python 3 (the earlier results were in Python 2) and therefore required a few minor changes in the code.

27 In particular the Python programming language is generally viewed as being easy to use for non-time critical applications. See KC Louden and KA Lambert, *Programming Languages: Principles and Practice* (3rd edn, Cengage 2012) 38.

of 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.[28] 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 as a by-product from other practices. Examples are scanning of books by Google Books or the 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 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. Programs and libraries like Gephi, Cytoscape or EasyGraph facilitate graphical analysis of networks. For general full-text search the Apache Lucene framework, while a database like ElasticSearch makes sophisticated search facilities easily available. Many of the tools listed above are available for free as open-source software.

A related development is the use of advanced special purpose algorithms. This is the area of Artificial Intelligence (AI). Since the first edition of this *Research Handbook* the developments in the area of AI appear to have skyrocketed. In particular the general application of specific models has taken flight. So much so that the term 'big data' has even gone out of fashion and has been replaced with a new fascination for all things 'generative AI'. In common parlance 'AI' seems to have become associated mainly with one type of machine learning that can also be used for data science purposes; the large language model (LLM), more specifically generative AI such as the generative pre-trained transformer architecture model

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

(GPT)[29] and Google's Pathways Language Model (PaLM). Examples of generative AI are ChatGPT, GPT-x,[30] Bard, Llama, and Claude. All of these models provide the user with the ability to generate (high) quality text based on natural language input. These models thus allow for the creation of text, based on text. This is the difference between asking a search engine a question and getting a resulting page full of links where one may find the answer, and typing in a question and getting an answer directly.[31]

Around the same time that these text-to-text based generative AI garnered broader application, the models which produce images from text such as DALL-E, Stable Diffusion, Midjourney and Google Imagen, also started gaining traction with the general population. These models are examples of text-to-image models. Combined with advances to the text-to-speech models popularised by for example ElevenLabs which allows the creation of audio and speech based on text and the video-to-video models and text-to-video such as GEN-2 or Imagen Video that allow for the creation of video based on natural language text input, it seems that our own creativity has become the only limitation to what may be created.

While these advances in text-to-media models are important to note, the same holds true for the reverse; media-to-text. To take but one example, the developments in Computer Vision also merit our attention. These are algorithms which are designed to enable machines to interpret and understand images, whether they are static or in motion, such as image recognition.

The combination of all (or some of) the models mentioned above, enables 'multimodal deep learning', which is another subfield of machine learning that involves relating information from multiple sources, such as text, audio, images, and video,[32] allowing the processing and linking of information across different types of data.[33] This allows for the creation of (actually) funny memes, but also recognition of hateful ones, identification of a location based on an image, but also in identifying possible dangerous traffic situations in autonomous vehicles, allowing for healthcare analysis at a deeper level whereby medical images are combined with health records, and also for sentiment analysis not restricted to text only but including visual cues, etc.

---

29 Which in essence is a type of artificial intelligence model designed to generate text by predicting the next word in a sequence, the prediction is based on the large dataset of text it was trained on.

30 As of November 2023 Open AI released GPT-4 Turbo; previous models were GPT-4, GPT-3.5, GPT-3, GPT-2, and GPT.

31 Google's search engine has also started providing snippets at the top of the search results page with the most likely answer, however this is a direct quote from the specific website and therefore differs from the answer ChatGPT for example would give you, as the latter provides an (in essence *new*) generated answer.

32 J Ngiam and others, 'Multimodal Deep Learning' in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (2011) 689–696.

33 Google for example has released its Gemini models which are created specifically with multimodality in mind.

Perhaps this also accounts for the term ‘big data’ being in decline, as it presumes the data to be so large that it is no longer of a manageable size for processing. The volume, velocity, and variety of the data no longer seems to be problematic, but is rather a feature when looked at from the perspective of for example LLMs.

In particular the training of LLMs thrives on voluminous (high quality) data sets, in which the variety of the data formats is no longer an issue. LLMs can digest heterogeneous documents without specific preprocessing or formatting, and have the potential to differentiate and categorise on their own, given sufficient data.

Application of these models has found widespread application in both business environments[34] and data science.[35] However, data science is different in that the stringency of science imposes further demands on methodology. It is not sufficient to produce pretty graphs or interesting results, but one must be able to prove that these results are actually scientifically valid, are ‘true’ or correct. This imposes restrictions and additional checks,[36] 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? 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

---

[34] OpenAI, the company behind ChatGPT and GPT-4, has stated that more than 92% of Fortune 500 are using OpenAI, see also <www.washingtonpost.com/technology/2023/11/06/openai-app-store-chat-gptstore/> accessed 4 April 2024.

[35] OpenAI specifically targets data science with its ‘code interpreter’ for which it identifies ‘[d]oing data analysis and visualization’ as one of its use cases. See <https://openai.com/blog/chatgpt-plugins#code-interpreter> accessed 4 April 2024.

[36] Although to some extent, the stringent quality requirements that the high-risk AI in the currently-being-negotiated AI Act come very close to, if they do not overlap entirely with, these requirements. See for example the proposed art 10 of the AI Act.

emergence of data science can be seen as disruptive,[37] 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. The tax authorities in many countries, for example, make use of automated decision-making systems for generating annual tax decisions, and executing them.[38] 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.[39] Similar questions of justice arise when public tasks are transferred to private organisations who make use of algorithms,[40] such as Uber,[41] and when data-driven technologies are used for profiling in criminal prosecutions.[42]

---

[37] 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* (The MIT Press 2016). Economists have questioned whether platforms such as Uber actually fulfil the conditions disruptive innovation in economic theory; see CM Christensen, ME Raynor and R McDonald, 'What is Disruptive Innovation?' (2015) *Harvard Business Review*, available at: <https://hbr.org/2015/12/what-is-disruptive-innovation> accessed 4 April 2024.

[38] See for more on this Chapter 14.

[39] 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) 156 et seq.

[40] Or when public authorities make use of information collected by private organisations, see DJ Solove, 'Access and Aggregation: Public Records, Privacy and the Constitution' (2002) 86 *Minnesota Law Review* 1140 and CJ Hoofnagle 'Big Brother's Little Helpers: How Choicepoint and Other Commercial Data Brokers Collect and Package Your Data for Law Enforcement' (2003) 29 *NCJ Int'l L & Com Reg* 595–637.

[41] For example, 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). In the Netherlands, this was also subject to litigation specifically challenging the use of algorithms for '*batched matching*' system and algorithms used to determine whether a driver's app should be deactivated due to (suspected) fraudulent behaviour. See cases Hof Amsterdam 4 April 2023, ECLI:NL:GHAMS:2023:796 and ECLI:NL:GHAMS:2023:793.

[42] See in this volume Chapter 17. See also B Schneier, *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World* (WW Norton & Company 2015, reprinted 2016).

In private law relationships the question of justice might translate into one of (contractual) fairness.[43]

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 specialised 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,[44] becomes more difficult to answer.[45] 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, of whether this is problematic.[46] Or, should perhaps rights in relation to software, and data in general, be kept out of the scope of property law, and (remain to) be governed by intellectual property law such as copyright and database rights,[47]

---

[43] For a discussion of the application of unfair terms regulation to digital consumer contracts, see N Helberger and others, 'Big data en het consumentenrecht' in PH Blok (ed), *Big data en het recht* (Sdu 2017) 151, 159 ff. See also Chapters 2 and 3 in this volume.

[44] See on software and ownership also N Härting, '"Dateneigentum" – Schutz Durch Immaterialgüterrecht?' (2016) 10 *Computer und Recht* 646. See also Chapters 8 and 9 in this volume.

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

[46] A similar question can be asked in relation to property rights in data, see, H Zech, 'Data as a Tradeabe 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.

[47] See Chapters 8 and 9 in this volume. See for example critical on the EU efforts to introduce a property right in data: PB Hugenholtz, 'Data Property in the System of Intellectual Property Law: Welcome Guest or Misfit'. Paper presented at the OOD: '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

or where it concerns personal data by data protection regulation?[48] If there is a role to play for property law in data,[49] a 'sui generis' right of data ownership may then need to be created.[50] 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 personalised 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.[51] 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 extensive expertise on the goods they buy—should the same rules apply to them as to other, less sophisticated consumers?[52] 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 thirteen different perspectives ranging from consumer, contract and liability law to (intellectual) property, corporate, competition, criminal, administrative and taxation 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.

---

at: <www.ivir.nl/publicaties/download/Data_property_Muenster.pdf> accessed 4 April 2024.

48 The question of whether personal data can perhaps (better) be protected by granting the data subject rights 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).

49 For example, see JHM van Erp, 'Ownership of Digital Assets?' (2016) 5 *European Property Law Journal* 73. See further Chapter 8 in this volume.

50 JHM van Erp and W Loof, 'Eigendom in het algemeen: eigendom van digitale inhoud (titel 1)' in LCA Verstappen (ed), *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, Case C-128/11 *UsedSoft GmbH v Oracle International Corp* ECLI:EU:C:2012:407, see D van Engelen, 'UsedSoft v Oracle: The ECJ Quietly Reveals a New European Property Right in "Bits & Bytes"' (2012) 1 EPLJ 317, cf. RM Wibier and J Diamant, 'Oracle gaat niet over Eigendom maar over contractsvrijheid' (2012) *Nederlands Juristenblad* 2966.

51 See further Chapter 4 in this volume.

52 These and other issues concerning consumer and contract law are discussed in Chapters 2, 3 and 4 in this volume.

6. DATA SCIENCE FOR LAW

Up to quite recently there was hardly any application 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 (digital and non-digital) data for research. An example is the 'legal origins' debate, which arose from law and economics research on the influence of the kind of legal system (civil law versus common law) on the development of a state.[53] 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.[54] In other disciplines this has been taken on enthusiastically. Within law such studies have started slowly, although gaining impetus recently.[55]

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

Firstly, 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 formats from publishers.[56] Courts only started to publish decisions on their own websites around the year 2000. Even then, most courts still only provide access through a search field or a list of decisions, and do not allow the downloading of the whole data set either via an API or in bulk.[57] Legal researchers themselves have not yet collected many data sets for sharing.

---

[53] See the seminal article, R La Porta and others, 'Law and Finance' (1998) 106 *Journal of Political Economy* 1113. Critique of this thesis can be found in 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) 464–474.

[54] A-L Barabási and J Frangos, *Linked: The New Science of Networks* (Perseus Books 2002).

[55] The perceptive overview article of Whalen (n 17) states that the first studies appear starting from 2007. See further N Peterson and EV 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. MA Hall and RF Wright, 'Systematic Content Analysis of Judicial Opinions' (2008) 96 *California Law Review* 63–122. See further on the possible relevance of network analysis for law: TFE Tjong Tjin Tai, *De grenzen van het privaatrecht* (Tilburg University 2011).

[56] M van Eechoud and L Guibault, 'International Copyright Reform in Support of Open Legal Information' (2016) Working Paper draft, Open Research Symposium Madrid 2016, available at: <www.ivir.nl/publicaties/download/OpendataCopyrightReform_ODRSdraft-WP_sep16.pdf> accessed 4 April 2024.

[57] An exception is the Dutch judiciary, offering all public court decisions on *rechtspraak.nl* up until recently in a large ZIP-file, and still via a REST API. See (n 12) above.

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 and case law rules, all of which are laid down in natural language with its inherent fuzzyness. This may go some way to explaining the lack of practical succes over decades of research into AI and law, notwithstanding exciting and innovative attempts at new forms of analysis (such as network analysis of legal materials[58]).

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.

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 still 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).[59] 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 requires human labour consisting of reading the texts by a legally trained researcher. Although one of the benchmarks used for the quality of the LLM models is to see whether they could pass the standardised tests for admission to law school in the United States (LSAT)[60] and GPT-4 even passed the Uniform Bar Exam (UBE),[61] there are still serious doubts as to whether such models can actually accomplish the tasks of coping with new developments and overcoming historical bias that are expected from any professional lawyer.

---

58 For further (critical) discussion and references see Whalen (n 17); C Coupette, *Juristische Netzwerkforschung* (Mohr Siebeck 2019).

59 See among others for the US: RC 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 visualisation of networks in case-law, such as Ravel Law (now part of LexisNexis).

60 See <www.semafor.com/article/03/15/2023/how-gpt-4-performed-in-academic-exams> accessed 4 April 2024.

61 DM Katz and others, 'GPT-4 Passes the Bar Exam' (15 March 2023) available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389233> accessed 4 April 2024.

However, the tide seems to have (nevertheless) turned. The rising interest in generative AI, as well as the fear of being 'disrupted' by Legal Tech[62] may have given greater urgency to lawyers to pick up data science. The question that arises is: how to do this?[63] 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 serves as an excellent introduction to the different methods of data science that may be useful for law and its practitioners. Chapter 16 in turn explores the possibilty of the vast amount of data that can be collected and utilised 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. Chapter 17 delves into the interplay between data science, data crime and legal frameworks, discussing the evolution from information to data ethics and the impact of big data on cybercrime, including the emergence of data crime modelling and the exacerbating effects of ransomware and AI on data-related criminal activities. It also examines how data science and law can reciprocally evolve to better detect, understand, and combat data crimes. The penultimate Chapter 18 discusses the particular challenges and transformation that data science and in particular new advances in AI bring to the judiciary. In it, the authors continue on their framework developed in the first edition of the volume which may assist in the decision of whether a particular technology should be applied by the judiciary and offer a wide array of examples of the current uses and limitations of the technologies that are or have been used.

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, although Chapter 19 might conclude the book, this will not be the last you hear of this topic.

---

[62] We will not discuss Legal Tech as such, although Chapter 10 in this volume does touch upon the topic. 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.

[63] Two noteworthy descriptions of this kind of research are F Fagan, 'Big Data Legal Scholarship: Toward a Research Program and Practitioner's Guide' (2016) 20 *Virginia Journal of Law & Technology* 1 and KD Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age* (OUP 2017). See also the contributions in MA Livermore and DN Rockmore (eds), *Law as Data: Computation, Text and the Future of Legal Analysis* (Santa Fe Institute Press 2019). Furthermore, see M Truyens and P Van Eecke, 'Legal Aspects of Text Mining' (2014) 30 *Computer Law & Security Review* 153.