Chapter 14
Methods of Data Research for Law

Bart Custers

Abstract

Methods of data research are becoming increasingly important in the legal domain. After explaining the concept of *legal big data*, to show that law is an area in which a lot of big data is available, this chapter discusses and illustrates several existing and potential applications of data research methods for lawyers and legal researchers. Particular opportunities exist with regard to (1) predictions, (2) searching, structuring and selecting, and (3) decision-making and empirical legal research. These methods constitute an important contribution to legal practice and legal scholarship as they may provide novel unexpected insights and considerably increase efficiency (less resources, more results) and effectiveness (more accurate and reliable results) of legal research, both in legal practice and legal scholarship. This may, among other things, result in improved legal services, new business models, new knowledge and a more solid basis for evidence-based policies and legislation. However, there are also several limits to and drawbacks of the use of these data research methods for law. From a methodological perspective, these include the lack of human intuition, an abundance of results that are not always relevant, limited insights in underlying causality, issues with repurposing, self-confirmation, self-fulfilling prophecies and reliability issues. It is concluded that, given the opportunities these developments provide for new business models for legal services and for legal research (both in legal practice and in legal scholarship), it is likely that these methods will be used on a larger scale in the near future and that new and additional methods will be developed. This will change to some extent the way legal work looks like and the job market for lawyers.

Keywords: legal big data, legal predictions, legal research, legal decision-making, empirical legal research, legal methodology, predictive policing, artificial intelligence, machine learning, computer forensics

14.1 Introduction

Data science and big data offer many opportunities for researchers, not only in the domain of data science and related sciences, but also for researchers in many other disciplines. Typical examples of such other disciplines are medicine (for instance, the 'Google flu trends' in which search engine query data was used to predict the epidemiological development of influenza viruses),[1] health (for instance, smartphone data that was used to detect physical activity patterns of people and patterns in obesitas),[2] sociology (for instance, the GDELT database,[3]

[1] J Ginsberg, MH Mohebbi, RS Patel, L Brammer, MS Smolinski, and L Brilliant, "Detecting influenza epidemics using search engine query data", 457 *Nature* 7232, p. 1012–1014, 2009.
[2] T Althoff, R Sosič, JL Hicks, AC King, SL Delp and J Leskovec, "Large-scale physical activity data reveal worldwide activity inequality", *Nature*, 2017, See DOI: 10.1038/nature23018.

an open source database with news media and social media data that was used to try to predict the Arab Spring),[4] psychology (for instance, the use of smartphones for research in psychology),[5] politics (for instance, the alleged use of big data in the US presidential elections to influence voting behavior)[6] and even oenology (for instance, to predict the quality of Bordeaux wines).[7]

The use of data science and big data in all these disciplines may add new opportunities to existing research methods. Typically, the use of big data involves a data driven approach, comparable to explorative data analyses and in contrast with more traditional explanatory research approaches that are usually hypothesis driven or theory driven or both. The latter approach starts with formulating a hypotheses based on theory (sometimes phrased as a research question) that is then tested against available data (that often have to be collected first via empirical research). After this testing, the hypothesis can be either be accepted or rejected and the existing theory can be expanded or amended with the research results. The use of hypotheses requires a certain amount of intuition, prior knowledge and/or theory regarding what a researcher is looking for, mainly because not all hypotheses can be tested due to limited resources.

Data science, particularly tools like data mining and machine learning,[8] offers opportunities to go through larger amounts of data in automated ways, not only to find particular data, but also to discover patterns in data, even when the data is unstructured.[9] Search algorithms can combine *all* available attributes, in order to see whether they are correlated. This approach may yield all kinds of unexpected statistical relationships that may not always be explainable or causally related (at least not at first sight). Statistical relationships may be useful for decision-making, even when underlying causality is missing or unclear. Furthermore, statistical relations may be a first step in discovering underlying mechanisms and causal relationships,[10] which may also lead to revision of theories.

The fact that data science and big data are playing an increasingly important role in so many research areas raises the question whether this also applies to the legal domain. Do data science and big data also offer methods of data research for law? As will be shown in this chapter, the answer to this question is positive: yes, there are many methods and applications that may be also useful for the legal domain.[11] This answer will be provided by discussing

---

[3] P Schrodt, "Automated Production of High-Volume, Near-Real-Time Political Event Data". Paper presented at the New Methodologies and Their Applications in Comparative Politics and International Relations Workshop. Princeton University, 4-5 February 2011.

[4] K Leetaru, "Did the Arab Spring Really Spark a Wave of Global Protests? The world may look like it's roiling now, but the 1980s were far worse". *Foreign Policy*, 30 May 2014.

[5] G Miller, "The Smartphone Psychology Manifesto", *Perspectives on Psychological Science*, Vol. 7, Issue 3, p. 221-237, 2012.

[6] H Grassegger and M Krogerys, "The Data That Turned the World Upside Down", *Motherboard*, 28 January 2017.

[7] I Ayres, "*Supercrunchers; How Anything Can Be Predicted"*, London: John Murray, 2007. The equation based on rainfall and temperature is the following: Wine Quality = 12,145 + 0,0017 * Winter Rainfall + 0,0614 * Average Growth Season Temperature – 0.00386 * Harvest Season Rainfall.

[8] For more details on how these technologies work, see, for instance, KS Candan and ML Sapino, "*Data management for multimedia retrieval"*, Cambridge: Cambridge University Press (2010); T Calders and BHM Custers, "What is data mining and how does it work?", in: BHM Custers et al. (eds.), *Discrimination and privacy in the information society*, Heidelberg: Springer, p. 27-42, 2013.

[9] Estimations are that approximately 95 % of big data is unstructured, see A Gandomi and M Haider, "Beyond the hype: Big Data concepts, methods and analytics", *International Journal of Information Management,* Vol 35, nr. 2, p. 137-144, 2015.

[10] V Mayer-Schönberger and K Cukier, "*Big Data: A revolution that will transform how we live, work and think"*, New York: Houghton, Mifflin, Harcourt Publishing Company, 2013.

[11] H Surden, "Machine Learning and Law", *Washington Law Review*, Vol. 89, No. 1, 2014.

these methods of data research for law in this chapter. As such, this chapter provides an overview of these methods, but also serves as an introduction to the second part of this book, which focuses on developing a new discipline.[12]
This chapter is structured as follows. In Section 14.2 the concept of legal big data is introduced and explained. In Section 14.3 several methods of data research for law are discussed. These methods are illustrated with several existing and potential applications for lawyers and legal researchers. The methods discussed are: (1) predictions, (2) searching, structuring and selecting, and (3) decision-making and empirical legal research. In Section 14.4 the limits to and drawbacks of the use of data research methods for law are discussed. The focus is on methodological limits and drawbacks, as ethical and legal issues will be discussed in the next chapters of this book. In Section 14.5 future developments regarding the way legal work and the legal job market will look like are discussed. In Section 14.6 conclusions are provided.

14.2 Legal big data

When asked, lawyers and legal experts are not usually inclined to consider themselves data analysts. They usually consider working with data an activity that occupies people with other professions. Results of a research project in the Netherlands in which legal researchers of all law schools were surveyed show that legal researchers hardly associate their work with sciences, although there is, to some extent, a connection with social sciences.[13]
Neither do lawyers and legal experts generally consider themselves to be professionals working with data. However, from the perspective taken in this chapter, this is not entirely true. They actually work with big data, since the data they work with consists of large volumes of data (sizable texts) that are, at least from a technological perspective, unstructured (different formats). Hence, at least two major aspects of big data (i.e., volume and variety) are met.[14] Legal documents, including legislation, case law, policy documents and academic journal publications, cannot only be considered data, but also as big data. This is referred to as "legal big data".[15]
Legal documents have existed for centuries. Whereas in the past these documents were sometimes hard to access (for instance, because of language issues and limited numbers of copies), nowadays many legal documents are being digitalised or digitally created. Digitalising legal documents (for instance, when old documents are scanned) results in better access to these documents. Digitally creating (new) legal documents also results in significantly enhanced searchability of these documents, which can be done in automated ways rather than manually. As a result of these developments, the accessibility of legal documents has considerably improved in the last decades. A lot of legal documentation that was in the past only available on paper and through a limited number of copies, is now available online for everyone who is interested, often including experts from non-legal

[12] Note that some parts of this chapter were also published in a Dutch journal, see BHM Custers and F Leeuw , "Legal big data: Toepassingen voor de rechtspraktijk en juridisch onderzoek", *Nederlands Juristenblad,* 34, p. 2449-2456, 2017.
[13] WH van Boom and RAJ van Gestel, "Rechtswetenschappelijk onderzoek – een samenvatting van de uitkomsten van een landelijke enquête", *Nederlands Juristenblad*, Vol. 20, p. 1336-1347, 2015.
[14] In general, data are considered big data when they meet several of the so-called 3Vs (Volume, Variety and Velocity). See D Laney, "*3D Data Management: Controlling Data Volume, Velocity and Variety"*. Gartner. Stamford CT: META Group Inc., 2001.
[15] FL Leeuw and H Schmeets, "*Empirical Legal Research, A Guidance Book for Lawyers, Legislators and Regulators"*, Cheltenham: Edward Elgar Publishing, Inc. p. 8, 2016.

disciplines and the general public. Also the searchability, i.e., the number of methods to search through the documents, has considerably improved. With the use of big data technologies it is increasingly possible for computers (although sometimes it may still be difficult) to recognize the text in scanned documents (which are actually images rather than texts). Sometimes also crowdsourcing is used for this purpose, in which large numbers of internet users (sometimes in return for access to online services) indicate which texts are in the images (so-called captcha tests).[16] When automated processing is required to distil texts from images, this is often done with the use of so-called Optical Character Recognition (OCR) software, which can recognise which parts of an image are text and, subsequently, which words and characters are in these text areas.[17] When a legal document is digitally created, such a 'translation' is not necessary; the documents can directly be analysed with the appropriate software. The difference can easily be seen in pdf-files: in digitalized (scanned) documents the search function does not work, whereas in digitally created (printed) documents the search function can be used.

Obviously big data is not merely about finding words. Using text mining, i.e., software that can (on a large scale) recognise patterns in texts, legal documents can be automatically analysed.[18] Such pattern recognition can be used to analyse social media data, for instance, with sentiment analyses, see the example of Coosto in the next section.

For developing a new discipline in the domain of data science and law, the topic of the second part of this book, this offers several important new perspectives. The use of big data offers possibilities for new methods of data research for law. Apart from legal research that is guided by hypotheses or theories, it also enables legal research that is driven by the data that is available. This involves automated pattern searching in legal documents and may reveal, for instance, which factors and circumstances determine the amount of financial compensations for data breaches[19] or risks levels of individual criminals for violating their probation or parole.[20] The search for such patterns may yield novel, unexpected results. Furthermore, such automated analyses can be performed with increasing efficiency and efficacy. New methods and their applications are discussed in the next section.

14.3 Methods

In this section, several methods of data research for law are discussed. These methods are illustrated with several existing and potential applications for lawyers and legal researchers both in legal practice and in legal scholarship. In the following subsections, (1) predictions,

[16] Captcha stands for "completely automated public Turing test to tell computers and humans apart". A Turing test is intended to test intelligent machines, to see whether they show intelligent behavior that is indistinguishable from human behavior. In 2017 researchers were able to create an artificial intelligence system that was able to break captcha challenges, see A Sulleyman "Bot 'breaks' captcha, making the most annoying thing on the internet pointless", *The Independent*, 31st October 2017.

[17] For a more detailed explanation of this technology, see: TD Duan, TLH Du, TV Phuoc and NV Hoang, "Building an Automatic Vehicle license-Plate Recognition System", Intl. Conf. in Computer Science, RIVF05, 21-24 February 2005, Can Tho, Vietnam, 2005.

[18] KB Cohen and L Hunter, "Getting Started in Text Mining". *PLoS Computational Biology*. Vol. 4 nr. 1, p. 20, 2008.

[19] DJ Solove and DK Citron, "Risk and Anxiety: A Theory of Data Breach Harms" (December 14, 2016). GWU Law School Public Law Research Paper No. 2017-2; GWU Legal Studies Research Paper No. 2017-2, 2017. See Chapter 3 and 4.

[20] BE Harcourt, "*Against prediction; profiling, policing and punishing in an actuarial age*", Chicago: University of Chicago Press, 2006. See also Chapter 10.

(2) searching, structuring and selecting, and (3) decision-making and empirical legal research are examined respectively.

14.3.1 Predictions

With the help of legal big data it is possible to predict the outcomes of court cases. In the United States, justices for the Supreme Court are nominated and appointed by the president. As a result of this, Supreme Court justices may favour the political views of the president that appointed them. When important rulings are expected, there is a lot of speculation in the media and by experts to predict whether the behaviour of the justices will be in line with their expected political views or, perhaps, their rulings will be surprising. The cases cover many different legal questions, including tax law, environmental law, equal treatment law, patent law, freedom of expression, right to life or criminal law. The Supreme Court of the United States consists of nine justices and the ruling in a specific case is determined by the majority vote of the justices.[21]

In 2004, US law professor Theodore Ruger organised a competition to find out who would best predict the rulings of the Supreme Court of the United States: a computer or a team of experts.[22] The computer (or rather a model) was provided with input on the outcomes of all cases that were processed by the Supreme Court the previous year. On the bases of over 600 cases, a model was created to predict the outcome of new rulings. The team of experts consisted of 83 reputable law professors and seasoned law practitioners. Both the computer and the experts were to predict the voting behaviour of each individual Supreme Court justice and the outcome of the majority vote.

The results were astonishing. When predicting the votes of individual justices, both the computer and the experts performed equally well (68 % and 67 % correct predictions respectively). However, when predicting the majority vote, the computer was able to easily beat the experts. The computer model predicted 75 % of the outcomes of the cases correctly, whereas the experts were able to only predict 59 % of the outcomes of the cases correctly, which is hardly any better than throwing a coin. Predicting how nine justices together behave turned out to be so difficult for the experts that they were hardly able to correctly predict the outcome, whereas the computer could more easily make correct predictions. For justices with strong ideological or political views the experts easily recognised the pattern and made correct predictions on the justice's behaviour. However, for the more moderate justices the experts were unable to predict the behaviour, whereas the computer still was able to discover some patterns.

In 2014, US law professor Daniel Katz and his team published a significantly improved model, for which almost 8000 cases of the past 60 years were used as input.[23] The predictions of this model were correct for 70 % of the cases and for 71 % of the individual justices. This model tried to incorporate a variety of dynamics that may influence court outcomes, such as public opinion,[24] changing membership and shifting views of justices[25] and changing judicial

[21] Sometimes the Supreme Court divides evenly on a case, for instance, because of recusals or vacancies. In such cases, resulting in confirmation of the lower court's decision. However, such a ruling of the Supreme Court does not establish binding precedent.

[22] TW Ruger, PT Kim, AD Martin and KM Quinn, "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-making", Columbia Law Review, Vol. 104, p. 1150, 2004.

[23] D Katz, M Bommarito, and J Blackman, "Predicting the Behavior of the Supreme Court of the United States: A General Approach", *PLoS ONE,* Vol. 12, nr. 4., e0174698, 2014.

[24] JA Segal, "Separation-of-powers games in the positive theory of congress and courts". *American Political Science Review*, 91(1), p. 28-44, 1997.

norms and procedures.[26] In 2016, British and US researchers also presented an accurate prediction model (79 % correct predictions) for the European Court of Human Rights.[27] A more detailed discussion on predicting judicial decision-making can be found in Chapter 19 of this book.

Predicting the outcomes of court cases can be very useful for legal practitioners, as it may help assessing whether presenting their case to a court is a good idea at all. When the likelihood of success is low, a lawyer or legal counsellor could perhaps better advise a settlement for his client. Judicial decision-making prediction models can also be a useful instrument for legal researchers to establish what existing positive law is and how it should be interpreted in specific cases. In the sociology of law domain, prediction models may be interesting to reveal which non-legal factors may play a role in sentencing (such as in: "justice is what the judge had for breakfast").[28] When looking at prediction models for areas in which little or no legal or criminological theories are available, legal big data may disclose patterns that contribute to (further) developing such legal theories. For instance, correlations between crime levels and weather may be further investigated and yield detailed predictions for further modelling particular types of criminal behaviour.[29] Another correlation worth investigating is that between crime levels and the release of new computer games (assuming such games will keep some potential offenders at home for some time) or providing further facts for solving heated controversies on theories on whether violent computer games are a predictive indicator for violent behaviour. Issues like underlying causality may be a concern in such cases – this will be discussed below, in Section 14.4.

Also other types of predictions can be very useful in legal practice.[30] For instance, in policing it may be very useful to analyse criminal data to make predictions of who will commit crimes, where crime will take place and which persons, buildings and objects may be at risk as a crime target. This is usually referred to as *predictive policing*.[31] Obviously it may be very helpful for law enforcement agencies to know who is likely to commit crimes and where it is likely that crimes will take place. This knowledge will enable them to focus their limited resources towards specific people and areas, maximising the effectiveness of their resources and successes.[32] This, in turn, may increase police legitimacy.[33] At the same time, it should be

---

[25] L Epstein, AD Martin, KM Quinn, and JA Segal, "Ideological drift among Supreme Court justices: Who, when, and how important". *Northwestern University Law Review*, 101(4), p. 1483-1542, 2007; AD Martin and KM Quinn, "Assessing preference change on the US Supreme Court". *Journal of Law, Economics, and Organization*, 23(2), p. 365-385, 2007.

[26] GA Calderia and C Zorn, "Of time and consensual norms in the Supreme Court". *American Journal of Political Science*, 42(3), p. 874-902, 1998.

[27] N Aletras, D Tsarapatsanis, D Preoţiuc-Pietro and V Lampos, "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective". *PeerJ Computer Science* 2:e93, 2016. See https://doi.org/10.7717/peerj-cs.93.

[28] S Danziger, J Levav and L Avnaim-Pesso, "Extraneous factors in judicial decisions", *Proceedings of the National Academy of Sciences*, Vol. 108, nr. 17, p. 6889-6892, 2011.

[29] For preliminary results indicating such correlations, see: R Murataya, and DR Gutierrez, "Effects of Weather on Crime", *International Journal of Humanities and Social Science*, Vol. 3, No. 10, p. 71-75, 2013.
Other are also experimenting with this on the basis of open access data. See, for instance: http://crime.static-eric.com/

[30] Note that when applying predictions to individuals, it means ascribing personal data to them. In the EU, processing such personal data is regulated by the General Data Protection Directive (GDPR), which aims to protect the informational privacy of people. See also Chapter 7.

[31] WL Perry, B McInnis, CC Price, SC Smith, and JS Hollywood, "*Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations"*. Santa Monica: RAND Corporation, 2013. See Chapter 10.

[32] It should be noted, though, that the exact effectiveness of these methods is hard to determine, see BHM Custers and SJ Vergouw, "Promising policing technologies: Experiences, obstacles and police needs regarding

noted that when the police profile frequent offenders or high risk neighbourhoods, this may be a self-fulfilling prophecy, as it may aggravate over time the perception of a correlation between them and crime (see the next section).[34]

Similarly, predictions and profiling may be useful not only in combating crime, but also in the fight against terrorism. Because terrorism is much less prevalent than high volume crime, it requires a different approach, that is usually based less on general police surveillance activities in specific neighbourhoods and focused more on specific individuals. Due to the high impact of terrorist attacks, prevention is even more important. For this reason it may also be valuable for law enforcement to have risk assessments of people, buildings and objects that may need increased observation and protection. Due to the low prevalence and incidence of terrorist attacks and the changing modi operandi, it is very difficult to make predictions, though.[35]

Another legal domain in which predictions may be very useful is that of probation and parole. In most countries, criminal courts heavily base their sentencing on (1) whether someone is a first offender or a repeat offender and (2) risk assessments of how likely recidivism is. These two factors are strongly related to each other, however. In many models for risk assessment used for sentencing, probation and parole decisions, prior convictions play an important role, resulting in the paradigm that "if you offend once, you are likely to offend again; if you offend twice, you will definitely reoffend again and again".[36] Although these relations may be statistically correct, on a group level they may prevent any other conclusions for those individuals who actually are willing and managing to improve their behaviour. This type of use of big data and profiling may then aggravate the difficulties that profiled persons already have obtaining work, education and a better life. Limiting these legitimate options, some convicted criminals may feel inclined or forced to fall back to their previous criminal behaviour after serving their time in prison.

14.3.2 Searching, structuring and selecting

Predicting outcomes of court cases is a major topic in the United States, but much less so in continental Europe. This might be related to cultural aspects (such as claim cultures and the costs of litigation)[37] and/or to the differences between common law systems and civil law systems.[38] Former US Supreme Court justice Oliver Wendell Holmes once stated that "prophesies of what the courts will do in fact, and nothing more pretentious, are what I mean

law enforcement technologies", *Computer law & security report,* 31, p. 518-526, 2015; BHM Custers, "Technology in Policing: Experiences, Obstacles and Police Needs", *Computer law & security report,* 1, p. 62-68, 2012.

[33] TR Tyler, "Enhancing Police Legitimacy", *The Annals of the American Academy of Political and Social Science*, 593, 84e99, 2004.

[34] BE Harcourt, "*Against prediction; profiling, policing and punishing in an actuarial age"*, Chicago: University of Chicago Press, 2006.

[35] J Rae, "Will it Ever be Possible to Profile the Terrorist", *Journal of Terrorism Research*, Vol. 3, Nr. 2, Autumn 2012. J Jonas and J Harper, "Effective counterterrorism and the limited role of predictive data mining", *CATO Institute Policy Analysis*, 584, p. 1-12, 2006. Note that this limit the reliability of such profiles, see BHM Custers, "Effects of Unreliable Group Profiling by Means of Data Mining". In: Grieser G, Tanaka Y, Yamamoto A (eds.) Lecture Notes in Artificial Intelligence. Heidelberg, New York: Springer Verlag, p. 290-295, 2003.

[36] BE Harcourt, "*Against prediction; profiling, policing and punishing in an actuarial age"*, Chicago: University of Chicago Press (2006). See also K O'Neill, "*Weapons of Math Destruction"*, New York: Crown, 2016.

[37] B Levin, "Addicted to welfare". *The Times*. 17 December 1993, London. p. 20, 1993.

[38] One (oversimplified and scientifically unsubstantiated) explanation would then be that common law systems focus more on (large amounts of) case law (yielding less predictable outcomes), whereas civil law systems focus more on legislation and legal theory (yielding more legal certainty).

by the law".[39] In civil law systems, however, other methods of data research for law may be much more important.

A completely different domain in which legal big data may contribute to legal practice and academic legal research is by facilitating research. The term legal research is used in two meanings in this chapter. On the one hand, legal research may refer to preparatory research for a specific legal case. This is the process of identifying and retrieving information necessary to support legal decision-making,[40] including finding primary sources of law in a given jurisdiction, previous and related case law, searching secondary sources like academic law papers and legal dictionaries and searching non-legal sources. This is referred to in this chapter as legal research in legal practice. On the other hand, legal research may refer to academic legal research. This type of legal research focuses less on individual case law and more on, among other things, finding patterns across case law, interpreting legislation, developing legal theory, investigating sociology of law and developing (suggestions for) new legislation. This is referred to in this chapter as legal research in legal scholarship. Legal big data may enhance both types of legal research.

Technology company IBM developed a computer system called Watson, that can interpret questions in natural language and answer those questions after consulting a collection of (digital) encyclopedias, books, journals, scientific publications and websites. In 2011, Watson competed in the TV quiz show *Jeopardy*, in which participants are presented with answers to which they must phrase their responses in the form of questions. Watson was playing against the two best players in the history of the TV show. One of them managed to play even against Watson in the first round, but all other rounds were won by the computer. Watson is very much like the onboard computer of starship *USS Enterprise* in the science fiction series *Star Trek*: when crew members ask a question to the onboard computer, the computer can understand the question and answer it. This once was science fiction, but now more or less exists.[41]

Watson is a typical example of artificial intelligence. The computer is fed with large amounts of data and equipped with software that is able to recognize patterns. Currently a spin-off of Watson is developed by IBM, called ROSS, that is specifically aimed at answering legal questions. Some people expect that legislation needs further specification in axioms and definitions before computers can interpret the legislation, but that is no longer necessary in the era of big data, in which large amounts of data are available. The large amounts of data allow the computer to distinguish definitions and interpretations on the basis of the context in which they appear. It is important to note that for judges and academic legal researchers this is no different – they also continuously keep further interpreting many legal concepts and legal provisions.

Another (again US based) example of facilitating legal research via legal big data is Ravellaw, an innovative company offering access to legal big data for legal research.[42] Ravellaw can be accessed online. A typical functionality it offers is looking for case law via search strings. The results are not shown in a list, like for instance, search results on Google, but in a way that underlying patterns in case law are shown. In Figure 14.1 a screenshot is shown in which

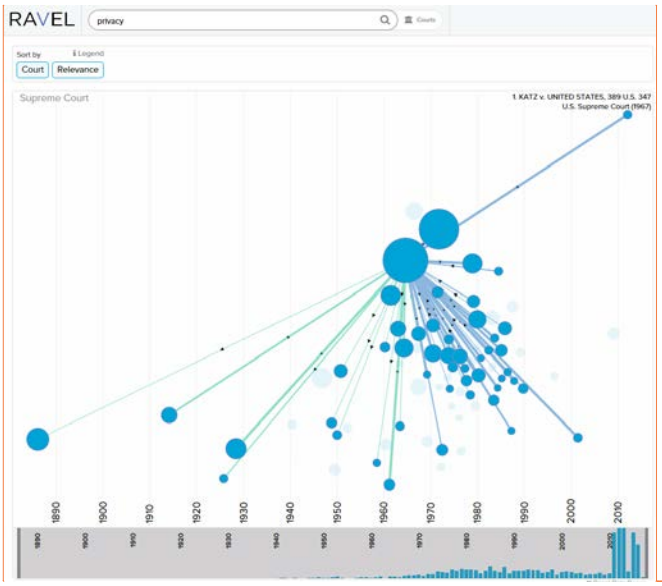
[39] Oliver Wendell Holmes, Jr., The Path of the Law, 10 Harv. L. Rev. 457, 461, 1897. See also M Radin, "The Theory of Judicial Decision: Or How Judges Think", *American Bar Association Journal*, Vol. 11, p. 357-362, 1925.

[40] JM Jacobstein and RM Mersky, "*Fundamentals of Legal Research"*, Santa Barbara: Foundation Press, p. 1, 2002.

[41] Natural language user interfaces are also available for consumers, in intelligent personal assistants and related applications like Siri, Amazon Echo and OK Google, but these have limited performance compared to artificial intelligence like Watson.

[42] E Eckholm, "Harvard Law Library Readies Trove of Decisions for Digital Age". *New York Times*, 28 Oktober 2015.

Ravellaw is used to search for case law on privacy. Ravellaw shows all case law (in this case of the US Supreme Court) in which privacy plays a role. The relevance of each case is presented with the size of smaller and bigger dots – the landmark cases are *Katz vs United States* en *Roe vs Wade*, the two biggest dots in the middle of the figure.[43] At the lower side of Figure 14.1 also the prevalence of case law is indicated for each year. A remarkable (and unexplained) result is a significant increase in privacy cases in 2010.
A typical advantage of the use of legal big data for legal research is that the likelihood of missing important information (such as a highly relevant case) considerably decreases. Because large amounts of case law and other legal documents can be processed (which would take a person days or weeks to go through), the accuracy and reliability of legal research can be considerably improved. Furthermore, novel, unexpected patterns may be discovered as shown above. This application of legal big data can be a tool to determine the relevant existing positive law and to interpret it in particular cases. Also, this application of legal big data may be used to discover underlying socio-legal patterns, further develop legal theory and (as will be discussed in the next subsection) improve or develop new legislation, regulations and policies.



Commented [A1]: See Contributor Guidelines from Elgar for the use of images and delivery of these to the typesetter.

Figure 14.1: Visual representation of US Supreme Court cases on privacy in Ravellaw.

Searching, structuring and selecting information may not only be useful in legal research, but also in other legal domains. A typical example may be forensics. For instance, in computer forensics sometimes an abundance of information is available, in which it can be hard to find

[43] Note that visualizations such as provided by Ravellaw may strongly influence the ways in which humans understand and process such information. When the data used as input is not complete or correct or the data analyses are flawed, this may raise several concerns, for instance, regarding reliability and accuracy. See also the discussion in Chapter 1.

the useful pieces of information that help to establish what happened (truth finding) and evidence that can be used in courts (evidence finding). Such pieces of information have to be found in large volumes of data, for instance, on seized computer systems, in online e-mailboxes, darkweb forums or ledgers of virtual currencies.[44] Search tools can be useful for these purposes, but also tools that may disclose patterns and relations in data may be useful.[45] The question is to which extent such disclosed patterns and relations may actually be used as evidence.[46] This may become particularly complicated when data science and big data are used to predict missing pieces of information (see the previous section). A more detailed discussion on data analysis in legal practice can be found in Chapter 20 of this book.

14.3.3 Decision-making and empirical legal research

A third area in which data science and big data may contribute to the domain of law is that of improving law and regulations. Part of much legal work (for instance, of judges, legal scholars, policymakers and legislators) is developing laws and regulations. Legal big data may be of added value in this area, for instance, when legal big data is combined with big data from social media. Social media data can also be regarded as big data, because it concerns large amounts (millions of users) of unstructured (text, pictures, videos), fast (real-time messages) data. These data can be used to investigate how large amounts of people view particular legal topics and their behaviour related to this. This is the domain of empirical legal research.[47]

A typical example is Coosto, a software company from the Netherlands, that performs so-called sentiment analyses on social media data. For this purpose messages on social media like Twitter and Facebook are analysed with text mining tools. Based on these analyses it is determined whether particular messages are positive or negative towards a particular topic. In Figure 14.2 a screenshot is shown in which Coosto is searched for Leiden University. It can be seen that in the previous period there were 1194 messages on social media on Leiden University, of which 14 % was positive and 7 % was negative.[48] These types of sentiment analyses may be useful for those who are developing laws and regulations to determine which proposed policies, regulations and legislation can count on public support.[49] Public support may increase and decrease over time, so sentiment analyses may also be used to choose the appropriate timing for launching new ideas, such as legislative proposals and new policies.[50]

---

[44] For a more detailed account, see JJ Oerlemans, "*Investigating Cybercrime"*, Amsterdam: Amsterdam University Press, (2017). See also RLD Pool and BHM Custers, "The Police Hack Back: Legitimacy, Necessity and Privacy Implications of The Next Step in Fighting Cybercrime", *European journal of crime, criminal law and criminal justice,* 2017(25), p. 123-144, 2017.

[45] P Adriaans and D Zantinge, "Data mining", Harlow, England: Addison Wesley Longman, 1996.

[46] T Zarsky, "Data mining as Search: Theoretical Insights and Policy Responses". in: B.H.M. Custers et al. (eds.), *Discrimination and privacy in the information society*, Heidelberg: Springer, p. 325-338, 2013.

[47] FL Leeuw and H Schmeets, "*Empirical Legal Research, A Guidance Book for Lawyers, Legislators and Regulators"*, Cheltenham: Edward Elgar Publishing, Inc., 2016.

[48] The other message were considered to be neither positive nor negative and, thus, neutral.

[49] Note that such applications of sentiment analyses would only work well when large amounts of data are used and even then, it has to be verified whether the samples are sufficiently representing public opinion. For instance, social media users typically do not represent the entire population. See, for instance, BHM Custers S van der Hof and BW Schermer, "Privacy Expectations of Social Media Users: The Role of Informed Consent in Privacy Policies", *Policy and Internet* 6(3): 268-295, 2014.

[50] For opportunistic use of big data during the US presidential elections in 2016, see: H Grassegger and M Krogerys, "The Data That Turned the World Upside Down", *Motherboard*, 28 January 2017.

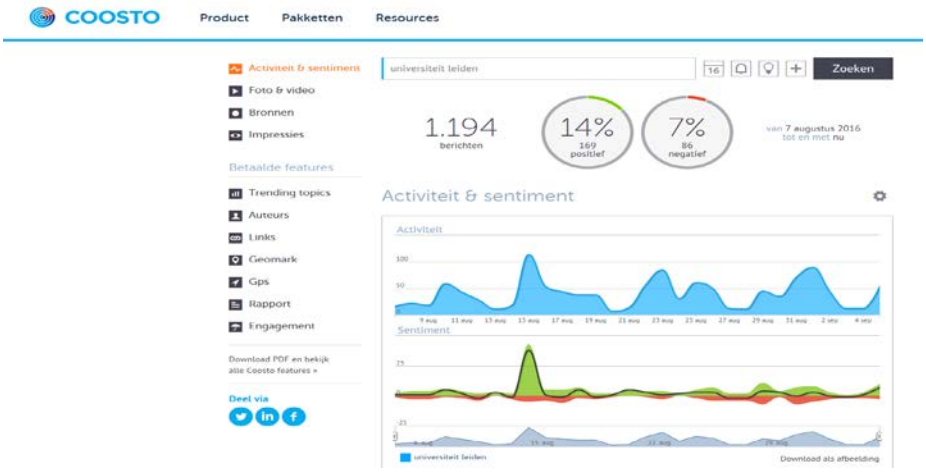


Figure 14.2: Visual representation of a sentiment analyses on social media by Coosto.

Legal big data may not only be useful in assessing public acceptance, but also in improving the contents of laws and regulation. By combining legal data with behavioral data, it becomes possible to assess which rules (or which types of rules) are best complied with and/or may be easiest to enforce. A typical example is walking on lawns instead of the pavement in public parks.[51] When the pavements constitute too much of a detour, people will be inclined to take a shortcut and walk via the lawns. After some time the grass will disappear on these often-used routes, inviting even more people to use these new routes. The newly created paths may show a completely new network compared to the initial pavements. The new network may supplement and to some extent replace the old network and will be better 'complied with'.
When attempting to create rules that are better complied with, it may be helpful to use the concept of 'nudging'.[52] Nudging is the offering of incentives, such as positive reinforcement or indirect suggestions, in order to try to make desirable behavior attractive, without forcing people into this behavior or limiting their liberties. The aim is to (slightly) redirect behavior via a choice architecture. Typical examples are the use of fake plastic houseflies in men's public toilet urinals and the use of attractive garbage bins. The UK government has set up special nudging teams (called Behavioural Insight Teams) to adjust behavior of the public in desirable directions. One of their most successful projects is on reducing fraud. By introducing new reminder letters that informed recipients that most of their neighbors had already paid, they were able to increase tax receipts. An adjusted message stating that "not paying tax means we all lose out on vital public services like the [national health care system], roads and schools" worked even better.[53]

Social media data can reveal plenty of behavioral patterns of people, including their travels, purchases, food preferences and physical exercise. Combined with legal big data, this may

[51] P Ball, "*Critical Mass; How One Thing Leads to Another*", New York: Farrar, Straus and Giroux, 2004.
[52] RH Thaler and CR Sunstein, "*Nudge: Improving Decisions About Health, Wealth, and Happiness*", New York: Penguin Books, 2009.
[53] T Rutter, "The rise of nudge – the unit helping politicians to fathom human behavior". *The Guardian*, 23 July 2015.

yield insights on how people will behaviorally react to proposed laws and regulations. These data can also be used to evaluate policies and legislation afterwards.[54] A typical example is the evaluation of judicial measures, such as camera surveillance in inner cities or community service for shoplifters.[55] Ideally, the effectiveness of such measures is evaluated in a so-called *randomized controlled trial*, similar to the evaluation of medical treatments. This approach involves the comparison of a test group and a control group that are completely similar, except for the factor that is investigated. Such an approach may in many situations not be realistic, because of the high costs involved or because there may be ethical concerns or practical limitations. Big data may then offer an alternative in the form of a quasi-experimental design, in which control groups are assembled afterwards. Within very large datasets, it is more likely that there will be 'twins' (i.e., individuals or groups that are similar in every aspect except the attribute under investigation). Also by introducing new measures in stages (for instance, region by region) groups may be created that are comparable. This may result in stronger evaluations and this, in turn, may contribute to evidence-based policies and legislation. Furthermore, the increasing amounts of available data may reveal more fine-grained insights. For instance, it may be discovered that convicting shoplifters to community service is only effective (in terms of preventing recidivism) for youth whose parents are not divorced.[56]

The use of legal big data for developing and improving laws and regulations may contribute to the work of socio-legal research, developing legal theories and evidence based legislation. The use of data science and big data may not only play an important role in decision-making regarding new laws and legislations, but also in decisions in specific cases. In Subsection 14.3.1 we already discussed predicting court decisions and court decisions based on predictions. However, decisions can also be made based on big data (not only on predictions, but also on the raw data or disclosed patterns) or even by algorithms developed in data science - which is usually referred to as automated decision-making or algorithmic decision-making.

Decision-making based on big data may be useful, for instance, to increase consistency in sentencing. A sound democratic legal systems is based on equality: similar misbehaviour should be sentenced in similar ways. Large differences in sentencing for similar crimes and misdemeanours within one jurisdiction would not be just. Although some judges or courts may be known for milder or harsher sentencing within particular jurisdictions, for a suspect and for the objectivity and fairness of the legal system it should not matter (too much) who is deciding. The criminal acts of the suspect and the circumstances in which a crime took place should be the most relevant factors in sentencing, rather than the personality of a judge. For this reason, in many jurisdictions guidelines for sentencing exist, often based on data on previous court decisions. For instance, all cases on shoplifting of the past five years are collected and the average sentence is calculated. This average may serve as a guideline for a judge. According to specific circumstances (first offender/second offender, use of violence during the crime, etc.) a judge may add to or subtract from this starting point.

Automated decision-making may also be (increasingly) important in an online environment. There exists a large number of smaller disputes not worth going to court for, mainly due to the costs and time involved. That is why alternative dispute resolution (ADR) and online dispute resolution (ODR) may be important alternatives. It is beyond the scope of this chapter to discuss these topics in further detail, but these approaches obviously offer solutions to

[54] F Willemsen and F Leeuw, "Big Data, real world events and evaluations", in: G. Petersson e.a. (eds.), *Big Data and evaluation*, Piscataway NJ: Transaction Publishers, 2016.

[55] For more examples, see F Willemsen and F Leeuw, "Big Data, real world events and evaluations", in: G. Petersson e.a. (eds.), *Big Data and evaluation*, Piscataway NJ: Transaction Publishers, 2016.

[56] Note this is a hypothetical example.

handling large numbers of cases. Furthermore, automated decision-making is often highly consistent. A typical weakness, however, may be that what is perceived as just decisions may change over time, whereas automated decision-making is based on (patterns in) historical data. As such, gradual changes may be included in automated decision-making, but more disruptive changes in society may be much harder to take into account. A combination with crowd-sourced online dispute resolution may be helpful in this respect.[57]

14.4 Limits and drawbacks

As was explained in the previous sections, the use of methods of data research for law and legal big data may have several advantages that may be a valuable contribution to legal practice and legal scholarship. It may yield novel, unexpected insights and it may considerably increase efficiency (less resources, more results) and effectiveness (more accurate and reliable results) of legal research, both in legal practice and legal scholarship. This may, among other things, result in improved legal services, new business models, new knowledge and a more solid basis for evidence-based policies and legislation.

However, the use of methods of data research for law also has some limits and drawbacks, which are discussed in this section. A distinction can be made between methodological issues and legal/ethical issues, but this section only focuses on the methodological limits and drawbacks. Legal and ethical issues, for instance, issues regarding privacy and discrimination are discussed in the next chapters, particularly in Chapter 15 on law and ethics and in Chapter 17 on data and fundamental rights.

Perhaps the most important methodological drawback is that the use of data research methods does not allow for the (direct) use of human intuition. The volumes of legal big data are usually too large to easily obtain useful overviews and insights. That is why instruments for automated data analyses were developed as described in the previous sections. The *lack of human intuition* makes it difficult to interpret data, to determine which algorithms to use and to interpret patterns and relationships that are discovered.

Another drawback is that in many situations data science methods may yield an *abundance* of patterns and relationships, many of which may not be novel or useful. For instance, correlations may be discovered that show increased risk of driving under influence for adults or people with driving licences, but this is not very remarkable. When there are many of these trivial relations, it may be hard to distinguish the really novel results.

Furthermore, these methods only yield statistical results. In many situations these may be sufficient knowledge to use as a basis for decision-making. For instance, if a relation is discovered between the buying oranges and macaroni, this may be useful for marketing purposes, without knowing underlying reasons of this relationship. However, in many situations, also in the legal domain, it may be useful to know underlying causal mechanisms. Discovering or even proving *underlying causality* may be much more difficult and often requires further research.

Another issue is that, although large amounts of legal big data may be available, these data may have been collected in the past for other purposes. Apart from legal restrictions on data reuse,[58] this *repurposing* may cause problems, because when the data are used for new purposes they may no longer entirely match these purposes. As a result, many of the discovered patterns may be based on indicators by proxy, rather than the actual factors determining outcomes. This may affect the reliability of research results. Contrary to other

[57] D Dimov, "*Crowdsourced Online Dispute Resolution"*, PhD Thesis. Leiden: Leiden University, 2017.

[58] H Ursic and BHM Custers, "Legal Barriers and Enablers to Big Data Reuse - A critical Assessment of the Challenges for the EU Law", *European Data Protection Law Review* 2(2): 209-221, 2016.

research methods, the use of data research methods does not allow for adjustment, simply because this approach is based on existing rather than new data.[59]
Closely related to this is the problem of *self-confirmation*.[60] Since these data research methods are based on available historical data, the research results will mostly look into the past rather than into the future. As explained in the previous section, gradual changes may be discovered and used for making predictions about the future, but more disruptive changes may be much harder to take into account.
Self-confirmation is a bias type of problem. Another example in this category of bias problems is that of *self-fulfilling prophecies*. A typical example is when surveillance of law enforcement agencies focuses on neighbourhoods with ethnic minorities. The probable result of such a policy would be that law enforcement databases get filled with people from these ethnic minorities. This is a form of selective sampling. When the law enforcement databases are subsequently used to find patterns on which people are more prone to fall into criminal behaviour, it may not be surprising to discover that people from these ethnic minorities may be profiled as showing increased levels of criminal behaviour. However, since the data was biased, this is a mere self-fulfilling prophecy.
A final methodological limit to discuss here is that of *reliability*. As explained above, the use of legal big data allows for taking into account much larger volumes of data (in fact, *all* data) to find patterns and relations. This may actually increase the reliability of the findings. However, at the same time there may be reliability issues because the findings are statistical relations. As such, they describe probabilities that may have limited use for decision-making, particularly in a legal context. For instance, when the data shows that the probability that a suspect committed a murder is 85 %, this obviously is not sufficient for a conviction. In such cases it is obvious that the algorithm should not replace the judge in the decision-making processes.[61]

14.5 Future developments

The developments and opportunities described in the previous sections will reshape, at least to some extent, the way legal work and the legal job market looks like. When developing a new discipline on the cross section of law and data science, it is important to closely examine these developments. This section will provide a brief overview of some of these developments and the ways in which they influence the characteristics of legal jobs and the legal job market.
In the United States, the use of legal big data seems to have permeated much further in legal practice and legal scholarship. The results of these developments are already becoming clearly visible. The employment opportunities for law graduates have significantly decreased over the past years, in part due to significant changes in business models for legal services.[62] Both increased technological opportunities and the rise of new innovative business models are putting traditional legal services under pressure. There are three important changes in the

[59] For a taxonomy on data reuse, see BHM Custers and H Ursic, "Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection", *International Data Privacy Law* 6(1), p. 4-15, 2016.
[60] K O'Neill, "*Weapons of Math Destruction*", New York: Crown, 2016.
[61] For more on reliability, see BHM Custers, "Effects of Unreliable Group Profiling by Means of Data Mining". In: Grieser G, Tanaka Y, Yamamoto A (Red.) Lecture Notes in Artificial Intelligence. Heidelberg, New York: Springer Verlag. p. 290-295, 2003.
[62] J Gershman, "Law School Applicant Pool Still Shrinking", *The Wall Street Journal*, 23 April 2015.

legal services market, of which at least the first two are enabled by legal big data and technological developments:[63]

1. More affordable, standardized and commoditized services are offered, sometimes even online.
2. With the help of new technologies, traditional lawyers can boost their productivity and perform the same amount of work with fewer lawyers.
3. The traditional rationale for granting lawyers a monopoly on the practice of law is breaking.

The job market for lawyers is thus becoming less attractive in two ways. First, there are less jobs and, as such, less employment opportunities. Second, the jobs pay less because employers can offer lesser payment conditions in a job market with an abundance of job seekers and a scarcity of jobs. In the US the number of law students that graduate each year is roughly twice the number of annual job openings.[64] As a result, employers can select the best candidates for each position without having to offer highly competitive wages. Another result is that it has become less and less attractive for prospective students to choose for law schools. The number of applications for law schools has drastically decreased in the period 2005-2015 with 40 %.[65]

In other countries, the job market for lawyers is still fine. For instance, in the Netherlands, the number of legal jobs is still increasing.[66] Nevertheless, it seems a good idea for lawyers (and the people and institutions educating them) to closely follow these developments. The new ways of working will increasingly be characterised by lower labour costs, mass customization, recyclable legal knowledge and pervasive use of IT.[67] As a result, at least part of the legal work may be outsourced or performed in other ways by people in other disciplines. Legal practice will (also in the long term) not be superfluous or replaceable, but data science and legal big data will definitely change its characteristics.

14.6 Conclusions

In this chapter, methods of data research for law were explained. First, the concept of legal big data was introduced, to show that law is an area in which a lot of big data is available. Even though many lawyers do not seem to regard it as big data, many legal documents, including legislation, case law, policy documents and academic journal publications, meet many of the characteristics that are typical of big data. Next, several methods of data research for law were discussed and illustrated with several existing and potential applications for lawyers and legal researchers. Particular opportunities exist with regard to (1) predictions, (2) searching, structuring and selecting, and (3) decision-making and empirical legal research. However, there are also several limits and drawbacks of the use of data research methods for law. From a methodological perspective, these include the lack of human intuition, an abundance of results (that are not always relevant), limited insights in underlying causality, issues with repurposing, self-confirmation, self-fulfilling prophecies and reliability issues.

[63] MR Pistone and MB Horn, "*Disrupting Law School: How disruptive innovation will revolutionize the legal world*". San Francisco: Christensen Institute, 2016.
[64] B Tamanaha, "*Failing Law Schools*", Chicago: University of Chicago Press, 2012.
[65] J Gershman, "Law School Applicant Pool Still Shrinking", *The Wall Street Journal*, 23 April 2015.
[66] Yacht, "*Trends en ontwikkelingen op de Legal arbeidsmarkt vierde kwartaal 2016*". Amsterdam: Yacht, (2017).
[67] R Susskind, "*Tomorrow's Lawyers*", Oxford: Oxford University Press, 2013.

Finally implications of these developments were discussed for the way legal jobs and the legal job market may look like in the future.

From this general overview, it can be concluded that methods of data research are becoming increasingly important in the legal domain. Many of the methods described in this chapter are already in use, although sometimes only on a small scale.[68] Given the opportunities these developments provide for new business models for legal services and for legal research (both in legal practice and in legal scholarship) it is likely that these methods will be used on a larger scale in the near future and that new methods will be developed.

This will change to some extent the way legal work looks like and the job market for lawyers. Those[69] who argue that the legal domain is unique and (therefore) incompatible with developments in data science and big data and those[70] who argue that legal practice is highly complex and require advanced cognitive abilities that data science is unable to provide, may not have a complete overview of the current developments. Data science and big data have found their applications in many other disciplines, from medicine to archaeology, and it is very unlikely that these developments will bypass the legal domain. Even though data science may not have the technological ability to match human-level reasoning, it can still have an impact on law.[71] At the same time, however, given the limits and drawbacks of the methods described in this chapter (not to mention the ethical and legal issues discussed in the next chapters), it is unlikely that these developments will entirely replace or invalidate legal work, like legal scholarship or legal decision-making. Lawyers will be needed in the near future and in the long term, to some extent for their expert knowledge, but probably mostly for the human factor they contribute to legal work. For instance, given the current lack of ethicality of the decision-algorithms used, it is unlikely that they will replace judges in decision-making. Law is a normative discipline and is (therefore) likely to benefit from human experts.

However, that does not mean that law cannot be combined with and benefit from data research methods and big data. These methods constitute an important contribution to legal practice and legal scholarship as they may provide novel unexpected insights and considerably increase efficiency (less resources, more results) and effectiveness (more accurate and reliable results) of legal research, both in legal practice and legal scholarship. This may, among other things, result in improved legal services, new business models, new knowledge and a more solid basis for evidence-based policies and legislation – all opportunities not to opt out of.

The research leading to the presented results has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 731873.

[68] See also DM Katz, "Quantitative Legal Prediction", 62 *Emory Law Journal*, 909, 936, 2013.

[69] See, for instance, K Okamoto, "Teaching Transactional Lawyering", 1 *Drexel Law Review*, Vol. 69, No. 83, 2009.

[70] See, for instance, Chicago Law School, "Symposium Legal Reasoning and Artificial Intelligence: How Computers 'Think' Like Lawyers", 8 *University of Chicago Law School Roundtable* 1, 19, 2001.

[71] H Surden, "Machine Learning and Law", *Washington Law Review*, Vol. 89, No. 1, 2014.