Data as Facts

Frederike Zufall

A. Legal data sets

As digitalisation increases, the amount of legal resources in the form of data is growing exponentially. Not just legal scientific publications,[1] but also court decisions have been published for many years,[2] with the databases being constantly expanded. Nowadays, statutes, parlamentary documentation, regulatory guidelines or administrative proceedings are widely available as open source. Regulatory approaches towards increasing open data by the EU[3] have likewise accelerated this process, particularly given the economic value of data as training data for machine learning.[4]

At the same time, the availability of data in and on the legal domain, has always been a prerequisite for empirical legal research.[5] The existence and accessibility of digital representations of court decisions not only facilitates research on legal sources beyond the single document, but also

1 Compare databases for legal research such as, e.g., LexisNexis, HeinOnline, Westlaw or beck-online in Germany.

2 Compare the European Court of Human Rights' database of court decisions, available at https://hudoc.echr.coe.int, or of the European Court of Justice, at https://curia.europa.eu or the curated U.S. Supreme Court database by H. J. Spaeth, L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, S. C. Benesh, and M. J. Nelson, 'Supreme Court Database (SCDB)'.

3 Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, *OJ L 172,* 26.6.2019, pp. 56–83; Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), *OJ L 152*, 3.6.2022, pp. 1–44 .

4 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European Strategy for Data, COM/2020/66 final, p. 2–3.

5 M. Heise, 'The Past, Present, and Future of Empirical Legal Scholarship: Judicial Decision Making and the New Empiricism' (2002) 2002 *University of Illinois Law Review* 819 at 829–30; compare M. A. Livermore and D. N. Rockmore (eds), *Law as Data: Computation, Text, & the Future of Legal Analysis* (Santa Fe Institute Press, 2019).

enables quantitative analysis on a whole data set. Investigating, how law works (or not), and affects reality (in the way it aims and claims), may have a broader base and a much greater impact if it can be demonstrated through quantitative methods.

However, empirical legal research did not stop at analysing metadata about judges[6] and their background,[7] or other circumstances, but has, with vector presentation from natural language processing allowed to move this analysis to the text[8] – the actual content of legal decisions, and the way how they apply law to the real world.

Besides this empirical perspective, the rennaissance of the field AI and law in recent years has yielded a number of contributions to the task of 'judgement prediction'.[9] Using huge corpora of court decisions, a machine learning classifier can be trained on the overall outcome decision, indicicating whether or not a new unknown case would be won, or on specific outcome parameters such as the content of a sentencing in criminal procedure. The underlying idea is that similar judgment texts will likely lead to a similar court decision. However, one problem with these approaches is that the training data may contain both the facts and the reasons of the judgments. Knowing already the normative verdict, and reasoning of the text, the prediction may not only be easier,[10] but it also

6 C. Engel, 'Lucky You: Your Case is Heard by a Seasoned Panel – Panel Effects in the German Constitutional Court' (2022) 19 *Journal of Empirical Legal Studies*.

7 Y. Lim, 'An Empirical Analysis of Supreme Court Justices' Decision Making' (2000) 29 *The Journal of Legal Studies* 721–52.

8 E. Ash and D. L. Chen, 'Case Vectors: Spatial Representations of the Law Using Document Embeddings' in M. A. Livermore, D. N. Rockmore (eds), Law as Data, (SFI Press, 2019), pp. 313–37.

9 Compare 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' (2016) 2 *PeerJ Computer Science* e93; I. Chalkidis, I. Androutsopoulos, and N. Aletras, 'Neural Legal Judgment Prediction in English' Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (Florence, Italy: Association for Computational Linguistics, 2019), pp. 4317–23; L. Yuan, J. Wang, S. Fan, Y. Bian, B. Yang, Y. Wang, and X. Wang, 'Automatic Legal Judgment Prediction via Large Amounts of Criminal Cases' 2019 IEEE 5th International Conference on Computer and Communications (ICCC), (2019), pp. 2087–91; see for an overview: Y. Feng, C. Li, and V. Ng, 'Legal Judgment Prediction: A Survey of the State of the Art' (2022), pp. 5461–69; and J. Cui, X. Shen, and S. Wen, 'A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges' (2023) 11 *IEEE Access* 102050–71.

10 M. Medvedeva, M. Wieling, and M. Vols, 'Rethinking the field of automatic prediction of court decisions' (2022) *Artificial Intelligence and Law* at 3.2.

confuses the actual process of subsumtion and reasoning that a human lawyer would perform.

B. Law and the real world

Judgment prediction tasks and the way in which they harness existing judgment corpora invite us to consider the extent to which these tasks have a parallel in judicial decision-making or more general in the process of legal decision-making.

The way in which judgments are written, with a clear distinction between the 'facts' and the 'reasons' of the case, is based on the legal theoretical distinction between normativity and reality, between 'is' and 'ought'.[11] This has also been reflected in general procedural principles and codifications of procedure law prescribing different sections for facts and reasons.[12] Distinguishing between the facts found and the legal, normative decision based on these facts corresponds to the different questions and dimensions that court proceedings need to address beforehand: what has actually happened? What can be proven? These questions of fact are subject to specific procedural rules and are addressed through the testimony of witnesses, judicial inspection or documentary evidence. By contrast, questions of law, and the application of law to these facts, are then situated in the normative realm.[13]

These theoretical grounds imply that predicting the outcome from judgment corpora containing the verdict or judgment texts containing the reasons does not equate to the judicial reasoning process. But neither does using the facts section of the texts, because this section was written *after* the parties had been heard and the evidence assessed in order to establish a procedural truth.[14] Moreover, the facts do not only comprise what has

11 D. Hume, *A Treatise of Human Nature* (1739) bk 3, part 1; H. Kelsen, *Reine Rechtslehre: Einleitung in die rechtswissenschaftliche Problematik* (F. Deuticke, 1934) pp. 37–38; F. Schauer, *Thinking like a lawyer: a new introduction to legal reasoning* (Harvard University Press, 2009) pp. 203–6.

12 Compare, for instance, § 313(1) of the German Civil Procedure Code and § 117(2) of the German Administrative Procedure Code. The Seventh Amendment (Amendment VII) of the U.S. Constitution also reiterates this distinction by preserving fact-finding to a jury.

13 Schauer, *Thinking like a lawyer: a new introduction to legal reasoning*, p. 204.

14 J. Bentham, *Rationale of Judicial Evidence: Specially Applied to English Practice : in Five Volumes* (Hunt and Clarke, 1827) bk 1, ch. 1.

been deemed 'true', but only those facts that correspond to the specific underlying legal issue. The facts section mirrors the factual side of what is prescribed by the normativity of the law. Hence, it can only be written once the final decision and the assessment that led to it have been made. This is different from perceiving reality without the normative lens.

Still, the idea and endeavour of the judgment prediction task and the use of case law corpora for machine learning leads to another, more general underlying question: to what extent can data represent facts? Or to what extent can we treat existing data the same as we treat facts for a legal decision?

C. Reality as data

Beyond the facts that we find written down in existing judgement corpora, the amount of data about the real world is much greater. Think about the data collected by sensors in the context of autonomous vehicles, or CCTV settings. The potential sources here are not only much broader and bigger than the facts in case law data sets. In comparison to the facts section in judgment texts, they are also a rather direct representation of the real world –without the normative lens. It is more closely comparable to what humans perceive as reality. This is also similar in that it corresponds to the factual realm *before* it has reached the courtroom and been subject to evidence, procedure, and filtering to determine its relevance to the legal question at hand. In this light, we can say that the real task of judgment prediction still lies ahead of us: predicting a normative outcome from the actual reality as it presents itself to us.

When we, as humans, speak of reality – and I think this is important when referring to the theoretical legal distinction between is and ought – this is also very much tied to the cognitive dimension. The idea of the *is* has always been tied to what humans may perceive and cognitively process. This is also one reason why the Seventh Amendment to the US Constitution entrusts fact-finding to a jury of non-lawyers, detaching this dimension from the assessment by judges.

If legal decision-making is no longer performed by humans, the question not only arises of how normativity and law can be applied by machines. There is also the question of the associated reality, the questions of fact and how they are processed by machines. This is particularly pertinent given that the range of perception (if we may still call it that way)

through hardware is potentially broader than what humans may perceive through their senses. Ultimately, the range of what we consider to be the starting point of our 'is' becomes broader. Or in other words, as our reality becomes more digital, we could argue that data is then not only a potential representation of reality, but extends it. And this has not only consequences on the questions of facts, and how we establish what is true, but it equally extends the potential realm and scope of the *ought*.

D. Data as facts

In fact, we have already witnessed an extension of the *ought* with increasing regulation on data or digital phenomena over recent years. Data as a reality is already subject to various forms of regulation that address (partly or wholly) digital phenomena, such as data protection law, the regulation of online platform providers or illegal content on the internet.[15] Some of this regulation was designed for the offline world but has since been adapted for similar online situations, some has genuinely been adopted to regulate digital phenomena.

But beyond data being subject to data regulation or to digital regulatory frameworks, to what extent can data serve as facts in legal decision-making? This would most likely be the case in domains where the reality is already digital; not because the subject of regulation is data, but because the data is more readily available and more readily susceptible to processing. In criminal law, for instance, there are an increasing number of offences in the digital realm that manifest as data, such as various forms of illegal online content. Other use cases range from data processed by IoT devices to autonomous vehicles.

However, there are still and will always be factual scenarios that only appear non-digitally. For instance, many criminal offenses are committed in the real world and involve physical human activity. In these cases, data

15 Compare, for instance, the EU General Data Protection Regulation, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of such Data, and repealing Directive 95/46/EC; or the Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act).

has a different role. It can serve as potential evidence, for instance, if the real-world scenario has been recorded by a CCTV. The facts of the case are then not the CCTV recording, but the criminal conduct itself.

The existence of data may blur the lines between what is subject to law as facts and what is merely a data representation of it. When regulating data, the subject of normativity is data; but when hardware collects data on reality, does it only qualify as potential evidence? When we rely on our own human senses to assess what is real, we do not question its existence. However, we do so when we hear testimony from third parties, such as witnesses in court. Existing procedure and evidence law has been built on the trust in human cognition and perception. This does not mean, however, that heuristics and biases do not come into play when finding the facts. [16] But the question we ask when hearing witnesses, is whether we trust what they say, and trust in how they have perceived reality.

In contrast, if we treat data as facts, we would ask whether and under what conditions we believe and trust the data, and what it appears to stand for. As then the facts are not based on human perception, procedural rules for fact-finding would also need to adapt accordingly. And this would open up legal procedure to quantitative empirical methods like never before.

Ultimately, the increasing availability of data challenges our legal and dogmatic framework, not only as we try to adapt to digitalisation, and not only as its use as training data for machine learning increases. But it does so at its theoretical core, when we consider what it is that law addresses, and whether it does so at a procedural – or at the substantive level.

References

Aletras, N., D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, 'Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective' (2016) 2 *PeerJ Computer Science* e93.

Ash, E. and D. L. Chen, 'Case Vectors: Spatial Representations of the Law Using Document Embeddings' in M. A. Livermore, D. N. Rockmore (eds), Law as Data, (SFI Press, 2019), pp. 313–37.

Bentham, J., *Rationale of Judicial Evidence: Specially Applied to English Practice : in Five Volumes* (Hunt and Clarke, 1827).

16 C. Engel, 'Judicial decision-making a survey of the experimental evidence' (2022) p. 7.

Chalkidis, I., I. Androutsopoulos, and N. Aletras, 'Neural Legal Judgment Prediction in English' Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (Florence, Italy: Association for Computational Linguistics, 2019), pp. 4317–23.

Cui, J., X. Shen, and S. Wen, 'A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges' (2023) 11 *IEEE Access* 102050–71.

Engel, C., 'Lucky You: Your Case is Heard by a Seasoned Panel – Panel Effects in the German Constitutional Court' (2022) 19 *Journal of Empirical Legal Studies*.

Engel, C., 'Judicial decision-making a survey of the experimental evidence' (2022).

Feng, Y., C. Li, and V. Ng, 'Legal Judgment Prediction: A Survey of the State of the Art' (2022), pp. 5461–69.

Heise, M., 'The Past, Present, and Future of Empirical Legal Scholarship: Judicial Decision Making and the New Empiricism' (2002) 2002 *University of Illinois Law Review* 819.

Hume, D., *A Treatise of Human Nature* (1739).

Kelsen, H., *Reine Rechtslehre: Einleitung in die rechtswissenschaftliche Problematik* (F. Deuticke, 1934).

Lim, Y., 'An Empirical Analysis of Supreme Court Justices' Decision Making' (2000) 29 *The Journal of Legal Studies* 721–52.

Livermore, M. A. and D. N. Rockmore (eds), *Law as Data: Computation, Text, & the Future of Legal Analysis* (Santa Fe Institute Press, 2019).

Medvedeva, M., M. Wieling, and M. Vols, 'Rethinking the field of automatic prediction of court decisions' (2022) *Artificial Intelligence and Law*.

Schauer, F., *Thinking like a lawyer: a new introduction to legal reasoning* (Harvard University Press, 2009).

Spaeth, H. J., L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, S. C. Benesh, and M. J. Nelson, 'Supreme Court Database (SCDB)'.

Yuan, L., J. Wang, S. Fan, Y. Bian, B. Yang, Y. Wang, and X. Wang, 'Automatic Legal Judgment Prediction via Large Amounts of Criminal Cases' 2019 IEEE 5th International Conference on Computer and Communications (ICCC), (2019), pp. 2087–91.