Law as a Lamppost

Janet Freilich[*]

ABSTRACT: Law produces all manner of public information: court documents, securities filings, patents, property records, and much more. This information is used in a multitude of ways—it teaches readers about individual cases, transactions, or entities, and is also aggregated to inform policymaking, set priorities, and drive predictive analytics and artificial intelligence. But choices about the information produced (or hidden) by law are often unintentional. Doctrines and institutions that appear facially unrelated to information production—like subject matter jurisdiction—nonetheless affect the shape and quantity of data produced. And even doctrines focused on information—like property recordation—create data used for purposes never envisioned by the law. We can only count what we can see, so law is inadvertently deciding which transactions, cases, and people are influential and which are invisible. Choices about how law produces information are directly responsible for selection bias—and thus for incorrect decisions—in areas as varied as how we calculate risk of child abuse, classify causes of death, create contract drafting software, and automate adjudicative processes. After demonstrating the prevalence of law's accidental information spillovers and their effects—which are unaccounted for by existing theories of law and information—this Article provides an updated framework for incorporating information into the theory and structure of law and highlights new roles for legal doctrines and institutions. The Article concludes with concrete ways that this new understanding of information can affect policy, including how it can be factored into institutional and doctrinal decisions, how to update the cost-benefit analysis of legal information to account for new uses and audiences, and how law might address harmful biases in available legal information.

[*] Professor, Boston University School of Law. I thank Atinuke Adediran, Pamela Bookman, Kevin Brown, Jack Conrad, Courtney Cox, Nestor Davidson, Debby Denno, Charles Duan, Howard Erichson, Martin Gelter, Caroline Gentile, Patrick Goold, Bernice Grant, Abner Greene, Tracy Higgins, Joseph Kupferman, Ron Lazebnik, Todd Melnick, Christopher Morten, Tejas Narechania, Aileen Neilsen, Ngozi Okidegbe, Sepehr Shahshahani, David Simon, Olivier Sylvain, Zephyr Teachout, Salome Viljoen, Ari Waldman, Maggie Wittlin, Joy Xiang, and Ben Zipursky.

Introduction

There is a well-known story about a drunk man looking under a lamppost for his keys. A passerby stops to help him look and asks, "Are you sure that you lost your keys under this lamppost?" "No," answers the drunk man, "but this is where the light is." Social scientists tell this story to explain that we focus

our attention where we can see best.[1] The questions we ask, answers we find, the policies we make—and, in the information age, the tasks that can be automated with artificial intelligence—are driven by where we have information and data: under the light of the lamppost.

Law is a lamppost. Law produces enormous amounts of public information and so influences the path of data-driven work both inside and outside the legal system. But much of these effects are accidental and the light—or shadow—from legal information can have serious inadvertent effects and create surprising biases. Securities filings, for example, are the largest public source of contracts and are used to train contract drafting and analytics software, but these contracts are not representative, and thus the software may not translate well to other types of contracts.[2] Property sales are public records, but leases are usually not, so there is comprehensive data on foreclosures but poor data on evictions, making it difficult to study or enact policy on the latter.[3] Workers' compensation data on employee injuries is used to allocate funding and determine priorities for safety improvements, directing finite resources away from independent contractors, whose injury data are not systematically collected.[4] In each of these examples, legal doctrines are responsible for the selection bias; they impact publicly available data which in turn is outcome determinative for the information-based question, policy, or AI application.[5]

Outside of the legal system, legal information is used to train artificial intelligence and, in doing so, dictates the model's output and is "a foundation for how the system will perceive observable reality."[6] For example, text from patents is the largest component of Google's C_4 dataset, which is used to train many large language models.[7] Government legislation is a major part of the Brown Corpus, a dataset of words and phrases used by linguists to study the English language and by computer scientists to build software that can parse text.[8] Court opinions make up a significant portion of The Pile, a dataset used

1. Noam Chomsky wrote that: "Science is a bit like the joke about the drunk who is looking under a lamppost . . . because that's where the light is. It has no other choice." ROBERT F. BARSKY, NOAM CHOMSKY: A LIFE OF DISSENT 95 (1998).

2. *See infra* Part II.

3. *See infra* Section II.C.

4. *See infra* Section II.B.

5. A related point—that appellate legal cases are not representative of the legal system or social behavior more generally—was made by George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1, 1–2 (1984).

6. KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 96 (2021).

7. Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, WASH. POST (Apr. 19, 2023, 6:00 AM), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning (on file with the *Iowa Law Review*).

8. W. Nelson Francis & Henry Kučera, *Brown Corpus Manual*, ICAME CORPUS MANUALS (rev. ed. 1979), http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM [https://perma.cc/85RT-E6F5]; *Text Type Analysis*, SKETCH ENGINE, https://app.sketchengine.eu/#text-type-analysi

for language modeling.[9] Mugshots formed an early basis for image-recognition software.[10] No legal policy was designed with these uses in mind; rather, legal data was created and made public for other reasons and used by computer scientists because it was available, free, extensive, and comprehensive. Data makes the world go 'round, and law makes data.

This Article's core argument is that every aspect of the legal system affects, often unintentionally, the creation and dissemination of information that can be aggregated for use in research, policy, and data analytics. The descriptive argument is expansive—legal information includes statutes, court documents, private information regulated by law, disclosures required by law, communication between private parties in the shadow of law, and more, and the pathways by which law can intervene in information production and availability are similarly varied.[11] Beyond the descriptive, this Article makes two additional claims. First, traditional conceptions about the relationship between law and information—although well-developed—are outdated and do not account for the modern era's hunger for big data and data-driven policymaking and law enforcement.[12] Second, law's data spillovers inadvertently create both substantial harms and noteworthy benefits. Careful attention to how legal doctrines shape the production and dissemination of information can mitigate these harms and enhance the beneficial applications of legal information.

The relationship between law and information is central to the legal system and the subject of a venerable and voluminous scholarship; however, existing scholarship overlooks several key points.[13] First, prior scholarship

---

s?corpname=preloaded%2Fbrown_1&tab=basic&filter=containing&onecolumn=1&wlattr=brown doc.genre&wlminfreq=1&include_nonwords=1&itemsPerPage=50&showresults=1&cols=%5B%22frq%22%5D&showtimelines=0&diaattr=&showtimelineabs=0&timelinesthreshold=5&wlsort=frq [https://perma.cc/PK2H-WK2R]; *Brown Corpus*, KAGGLE (2018), https://www.kaggle.com/datasets/nltkdata/brown-corpus/data (on file with the *Iowa Law Review*).

9. Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, ARXIV 1, 4 (Dec. 31, 2020), https://arxiv.org/pdf/2101.00027 [https://perma.cc/6USB-YA92].

10. CRAWFORD, *supra* note 6, at 89–96.

11. For a fuller definition of "legal information" and additional description of the ways in which law can affect information, see *infra* Part I.

12. *E.g.*, Daniel E. Ho & Larry Kramer, *Introduction: The Empirical Revolution in Law*, 65 STAN. L. REV. 1195, 1202 (2013) (noting the rise in "scholars, courts, and decisionmakers, [who] are grappling with data"); Kristina McElheran & Erik Brynjolfsson, *The Rise of Data-Driven Decision Making Is Real but Uneven*, HARV. BUS. REV. (Feb. 3, 2016), https://hbr.org/2016/02/the-rise-of-data-driven-decision-making-is-real-but-uneven (on file with the *Iowa Law Review*).

13. For a very small sample of this literature, see, e.g., James Boyle, *A Theory of Law and Information: Copyright, Spleens, Blackmail, and Insider Trading*, 80 CALIF. L. REV. 1413, 1416 n.1 (1992) (noting that few legal scholars "have seemed interested in dealing holistically with information, rather than with the doctrinal categories into which law has traditionally divided it"); Zachary D. Clopton & Aziz Z. Huq, *The Necessary and Proper Stewardship of Judicial Data*, 76 STAN. L. REV. 893, 949 (2024) (characterizing judicial data as a public asset). There has been particular emphasis on law that either directly affects information collection and dissemination, like copyright (e.g., Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 745

often relates to laws that deliberately consider information, like property records,[14] disclosure mandates,[15] or efforts to make court and agency documents more easily available.[16] But much of the informational function of law is unintentional. *Every* choice about design of legal doctrines and institutions across both public and private law affects the information that law produces. Doctrines and institutional structures that facially have nothing to do with information, and whose creators did not consider information effects, nonetheless impact informational output and consequently influence how data can be used. Second, the informational function of law has historically been viewed as communicating information about a transaction, entity, or case to a party who is directly interested in that piece of information—like a potential purchaser of property, a party who might be influenced by the outcome of an earlier case, or an inventor who might build on information contained in a patent.[17] This Article emphasizes that, in the information era, much legal information is aggregated and used by parties who are uninterested

(2021); and Pamela Samuelson, *Generative AI Meets Copyright*, 381 SCIENCE 158, 159 (2023)), agency disclosure regimes (e.g., Daniel E. Ho, *Fudging the Nudge: Information Disclosure and Restaurant Grading*, 122 YALE L.J. 574, 582 (2012); and Cynthia A. Williams, *The Securities and Exchange Commission and Corporate Social Transparency*, 112 HARV. L. REV. 1197, 1209–10 (1999)), and areas where law affects information transmission between private parties (e.g., Richard Craswell, *Taking Information Seriously: Misrepresentation and Nondisclosure in Contract Law and Elsewhere*, 92 VA. L. REV. 565, 567, 595 (2006)). There are also prominent scholars of law and information whose work studies the regulation of data, information platforms, and privacy. *E.g.*, Danielle Keats Citron, *Reservoirs of Danger: The Evolution of Public and Private Law at the Dawn of the Information Age*, 80 S. CAL. L. REV. 241, 246 (2007); Julie E. Cohen, *The Regulatory State in the Information Age*, 17 THEORETICAL INQUIRIES L. 369, 370 (2016); Woodrow Hartzog & Daniel J. Solove, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230, 2265 (2015); Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 119 (2004); Salomé Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573, 577 (2021).

14. *E.g.*, Maureen E. Brady, *The Forgotten History of Metes and Bounds*, 128 YALE L.J. 872, 875–76 (2019); Henry E. Smith, *The Language of Property: Form, Context, and Audience*, 55 STAN. L. REV. 1105, 1167 (2003).

15. *E.g.*, Cass R. Sunstein, *Informational Regulation and Informational Standing:* Akins *and Beyond*, 147 U. PA. L. REV. 613, 616 (1999); William M. Sage, *Regulating Through Information: Disclosure Laws and American Health Care*, 99 COLUM. L. REV. 1701, 1707 (1999).

16. *E.g.*, Merritt E. McAlister, *Missing Decisions*, 169 U. PA. L. REV. 1101, 1107–08 (2021); David E. Pozen, *Freedom of Information Beyond the Freedom of Information Act*, 165 U. PA. L. REV. 1097, 1098–1102 (2017); Amy J. Schmitz, *Measuring "Access to Justice" in the Rush to Digitize*, 88 FORDHAM L. REV. 2381, 2393 (2020).

17. *See, e.g.*, Thomas W. Merrill & Henry E. Smith, *Optimal Standardization in the Law of Property: The* Numerus Clausus *Principle*, 110 YALE L.J. 1, 3 (2000) (explaining that property law communicates information to interested parties); Jeffrey L. Furman, Markus Nagler & Martin Watzinger, *Disclosure and Subsequent Innovation: Evidence from the Patent Depository Library Program*, 13 AM. ECON. J. 239, 240 (2021) (discussing how patents provide information to scientists in related fields); Arthur R. Miller, *Confidentiality, Protective Orders, and Public Access to the Courts*, 105 HARV. L. REV. 427, 490 (1991) (exploring how protective orders affect third parties interested in evidence presented in litigation).

in any individual transaction but care instead about patterns seen across groups of transactions.[18]

Law's influence on information availability can be a powerful tool if wielded deliberately and with good intention to shape socially beneficial policies and AI applications.[19] But data is presently merely an underappreciated accident of the legal system, and so the uses it enables are arbitrary and sometimes injurious: Some groups are more likely to litigate, be arrested, disclose contracts, buy houses, or be otherwise represented in publicly available legal data, meaning that AI applications or policies based on legal data will overrepresent those groups and underrepresent others (for better or worse).[20] When the resultant information is used without careful consideration of these biases, the effects can be harmful. For example, an algorithm used in child welfare determinations incorporated variables including whether parents were receiving mental health treatments.[21] Because the government only had access to that information for individuals whose treatments were funded by the state, not those who could pay privately, the algorithm was in part making decisions based on income levels.[22] It is also more difficult to study problems and make policy where law does not provide data.[23] Situations or groups with less data may be invisible.[24] And the development of legal technology will occur faster in data-rich areas and may be focused primarily on groups overrepresented in datasets, meaning the technology may work poorly for others.[25]

Law's data spillovers can also create a circular, self-reinforcing effect. Legal policy impacts data production and availability, which in turn impacts the shape of legal policy. Sometimes this cycle is explicit and intended: The 1996 Dickey Amendment banned federal funding for research on gun violence, decreasing the amount of data on gun violence which in turn made it difficult to enact policies to prevent gun violence.[26] Because data availability impacts

18. *See infra* Part II (discussing how legal data is aggregated for applications including predictive analytics and automating legal decision-making and to inform policy choices and allocate funding).

19. Janet Freilich & W. Nicholson Price II, *Data as Policy*, 66 B.C. L. REV. (forthcoming 2025) (manuscript at 25) (on file with the *Iowa Law Review*).

20. *See, e.g.*, Mary Madden, Michele Gilman, Karen Levy & Alice Marwick, *Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans*, 95 WASH. U. L. REV. 53, 57 (2017) (explaining, in the context of privately collected data, that "low-status" individuals experience "heightened vulnerability to online surveillance and associated adverse outcomes").

21. Virginia Eubanks, *A Child Abuse Prediction Model Fails Poor Families*, WIRED (Jan. 15, 2018, 8:00 AM), https://www.wired.com/story/excerpt-from-automating-inequality (on file with the *Iowa Law Review*).

22. *Id.*

23. *See infra* Section III.A.3.

24. *See infra* Section III.B.

25. *Id.*

26. Representative Jay Dickey, the amendment's sponsor, later expressed regret for this effect, noting that "[a]s a consequence [of the amendment], U.S. scientists cannot answer the most basic question: What works to prevent firearm injuries?" Jay Dickey & Mark Rosenberg, Opinion,

what is studied and consequently the policies that can be proposed, law's unintended effects on data also (inadvertently) impact policy. For instance, because federal court records are more easily available than state court records, doctrines like subject matter jurisdiction (which is about such things as protecting out of state litigants from bias and managing court caseloads—not information dissemination) unintentionally dictate whether a case will be part of a data analysis to inform policy.[27]

In emphasizing law's impact on information, this Article has implications not only for how data is used but also for legal scholarship that relates to law and information.[28] Scholarship on privacy law addresses law's influence on dissemination and collection of large-scale data.[29] However, this important line of literature can be narrowly focused on privacy harms from aggregate legal information.[30] While acknowledging the severity of the problem, this Article emphasizes the other side of the privacy coin: Harms can flow from information that is not public and is therefore invisible to data aggregators.[31] If some information is not publicly available, its subjects will have additional privacy, but their policy interests may also be ignored. Thus, while privacy is an important consideration in determining how legal doctrines should incentivize or suppress information flows, it is not the only consideration.

This Article suggests reforms to both theory and practice to better account for law's data spillovers. First, this Article lays out an updated view of law and information, emphasizing that technology has created new uses and expanded

---

We Won't Know the Cause of Gun Violence Until We Look for It, WASH. POST (July 27, 2012), https://www.washingtonpost.com/opinions/we-wont-know-the-cause-of-gun-violence-until-we-look-for-it/2012/07/27/gJQAPfenEX_story.html (on file with the *Iowa Law Review*). *See generally* Janet Freilich, W. Nicholson Price II & Aaron Kesselheim, *Disappearing Data at the Federal Government*, NEW ENG. J. MED. 1 (March 26, 2025), https://www.nejm.org/doi/10.1056/NEJMp2502567 [https://perma.cc/V2X8-5J5X] (quantifying and discussing removal of CDC datasets).

27. In part because of data availability, much legal scholarship focuses on federal courts and judicial interpretation of federal laws, not state laws (this is also because federal laws have a country-wide impact). *See, e.g.*, Michael Risch, *From Patents to Trade Secrets*, *in* RESEARCH HANDBOOK ON EMPIRICAL STUDIES IN INTELLECTUAL PROPERTY LAW 103–04 (Estelle Derclaye ed., 2023) (noting that trade secret litigation was difficult to study prior to the creation of a federal law that brought trade secret cases into federal court).

28. Some of this extensive body of scholarship is summarized *infra* Section I.A.

29. *E.g.*, Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137, 1139–40 (2002). *See generally* ANDREW GUTHRIE FERGUSON, THE RISE OF BIG DATA POLICING: SURVEILLANCE, RACE, AND THE FUTURE OF LAW ENFORCEMENT 107 (2017); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 103 (2014); Ari Ezra Waldman, *Privacy, Notice, and Design*, 21 STAN. TECH. L. REV. 74, 91 (2018).

30. *See* Mary D. Fan, *The Right to Benefit from Big Data as a Public Resource*, 96 N.Y.U. L. REV. 1438, 1446 (2021) (criticizing the "myopic" privacy literature for ignoring "the right of the public to benefit from the collection of our data by private entities").

31. This is also recognized by privacy scholars. *E.g.*, Ari Ezra Waldman, *Gender Data in the Automated Administrative State*, 123 COLUM. L. REV. 2249, 2252–53 (2023).

audiences for legal information.[32] This leads to new roles for law and legal institutions. Law has traditionally had the important function of determining whether information is *correct*. As law's effect on information expands, law increasingly takes on the role of deciding which information is *visible*.[33] Legal institutions—as channels for most legal information—also have increased ability (and responsibility) to shape information flow.[34]

With respect to policy, this Article advocates for several complementary approaches. First, that information production should be a factor in setting policies. Scholars and policymakers consider many different priorities—justice, administrability, political feasibility, and others—informational effects must be part of the discussion.[35] Second, that new uses and audiences for legal information in the modern era change the costs and benefits of producing (or suppressing) legal information. Because most doctrinal and institutional choices about information production rely on some form of balancing those costs and benefits, existing rules about information production may need updating.[36] Third, if there are harmful gaps or biases in the information produced by legal doctrines, policymakers have a variety of tools to help. These include additional funding for research, government-run national surveys, development of statistical correction techniques, standardization across institutions, and more.[37] Fourth, data users and intermediaries should be both aware of and transparent about the makeup and shortcomings of the data that they use. Some of the unintentional harms of legal information can be mitigated with additional disclosure and better data literacy.[38]

The Article proceeds as follows. Part I explores existing theories linking law and information and then turns to how legal information is presently used. Part II consists of a series of case studies on how law affects information, highlighting the breadth of ways that legal doctrines can unintentionally affect information production and consequently shape information-intensive applications. Part III provides a more general look at the impact of legal information, including the informational biases and harms law creates, the role of legal information in shaping the development of legal technology, and the problem of invisible law. Parts IV and V turn to reforms, both theoretical and practical.

32. *See infra* Section IV.A.

33. This has always occurred, but the role of law in disseminating information is more important as that information is used in a broader array of functions. *See infra* Section III.A.

34. *See infra* Section IV.B.

35. *See infra* Section V.A.

36. *See infra* Section V.A.

37. *See infra* Section V.B.

38. *See infra* Section V.C.

I. THE RELATIONSHIP BETWEEN LAW AND INFORMATION

Law produces a great diversity of information.[39] It includes the output of legal processes like statutes and court opinions, private information publicized by legal institutions like inventor-written patents and databases of property owners and home prices, research conducted by agencies like the Environmental Protection Agency, disclosures incentivized by laws like voluntary reports of product safety risks, communication between private parties that is shaped by background legal doctrines like disclosures in contracts, and a great deal more. Because this Article is focused on how law affects information, it uses an expansive definition of information which includes any information linked, even tangentially, to law.

The universe of information that legal doctrines and processes might affect is vast, and the pathways by which law can shape the volume, contents, and availability are also numerous. Law can directly regulate information by requiring or forbidding disclosure of information. Law can incentivize or disincentivize disclosure by providing carrots or sticks for information release or by creating settings where certain types of information are encouraged or expected.[40] Legal institutions can create platforms to disseminate information and can change those platforms to increase or decrease dissemination, or to change the audience for dissemination. Laws can make it easier or harder for private intermediaries to capture legal information, which in turn affects how those intermediaries can process and distribute the information.[41] This is not a complete list but illustrates the wide array of possible connections between law and information. When this Article refers to the relationship between law and information, it encompasses the breadth of potential ways in which law can affect information.

Given the multiplicity of types of legal information and ways in which law and information connect, it is unsurprising that legal scholars have focused a tremendous amount of attention on these topics. Section I.A summarizes this literature, emphasizing its historical orientation toward how information about an individual piece of legal information is communicated to parties specifically interested in that material—a focus that reflects how legal information has traditionally been used. But with the rise of big data, legal information is used differently. Section I.B turns to how legal information is used in the aggregate, explaining that the value of legal information now often lies not in the individual

39. Legal information is not only varied, it is often also important and impactful. *E.g.*, Jessica Silbey, *A Matter of Facts: The Evolution of the Copyright Fact-Exclusion and Its Implications for Disinformation and Democracy*, 71 J. COPYRIGHT SOC'Y (forthcoming 2025) (manuscript at 372) (on file with the *Iowa Law Review)* (calling law an industry "that produce[s] predominantly fact-based works central to the socio-political institutions at the heart of U.S. democracy").

40. *E.g.*, Yonathan A. Arbel & Murat Mungan, *The Case Against Expanding Defamation Law*, 71 ALA. L. REV. 453, 463 (2019) (illustrating the point in the context of defamation law, which affects incentives for speech by private parties).

41. *See infra* Part IV.

transaction but in the patterns, sums, and trends seen across a population or field of law.

A. EXISTING THEORIES OF LAW AND INFORMATION

Law can affect information directly—for instance, trade secret law dictates conditions under which law deters public release of private information,[42] copyright law relates to control of information,[43] and privacy law governs what information can be collected[44] and how it must be protected.[45] But the relationship between law and information is far more complex. The very structure of law reflects deep thinking about what information is produced by legal systems, how law can incentivize or dissuade the creation or dissemination of information, and how legal information is used by both the public and the legal system itself. This Section discusses existing theories of how law affects information from private law, public law, and legal institutions.

Private law is, roughly, the law of rights and duties between private individuals.[46] The informational structure of private law is concerned with how information is best communicated between these private individuals.[47] Consider property law. Property conveyances must take one of a small number of recognized forms (unlike contracts, which are infinitely customizable), which facilitates communication of information about property by making it simpler for third parties to ascertain the boundaries and attributes of the property right.[48] Similarly, property sales are recorded in a public register so that third parties are aware of the transaction to prevent, for instance, a property owner from selling land twice to different people.[49]

Contract law also deals with communication between individuals. It raises questions about how the structure of contracts incentivizes exchange of

---

42. *E.g.*, Courtney M. Cox, *Legitimizing Lies*, 90 GEO. WASH. L. REV. 297, 318 (2022).

43. *E.g.*, Lemley & Casey, *supra* note 13, at 744. Although copyright law is often thought of in the context of creative works, it also has significant impact on the dissemination of factual information. *See* Silbey, *supra* note 39, at 385 (compiling a list of cases disputing copyright over factual information and discussing the implications of the doctrine).

44. Jessica Litman, *Information Privacy/Information Property*, 52 STAN. L. REV. 1283, 1290–91 (2000).

45. Olivier Sylvain, *The Market for User Data*, 28 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 1087, 1094 (2019).

46. John C.P. Goldberg, *Introduction: Pragmatism and Private Law*, 125 HARV. L. REV. 1640, 1640 (2012) (noting the concept of private law "eludes precise definitions").

47. For examples from property law, see, e.g., Brady, *supra* note 14, at 875; Nestor M. Davidson, *Property and Relative Status*, 107 MICH. L. REV. 757, 760 (2009) (exploring how property communicates information about status); Carol M. Rose, *Introduction: Property and Language, or, the Ghost of the Fifth Panel*, 18 YALE J.L. & HUMANS. 1, 5–6 (2006) (discussing the relationship between property and expression); and Smith, *supra* note 14, at 1139–40.

48. Merrill & Smith, *supra* note 17, at 3; Nestor M. Davidson, *Standardization and Pluralism in Property Law*, 61 VAND. L. REV. 1597, 1599 (2008).

49. Richard A. Epstein, *Notice and Freedom of Contract in the Law of Servitudes*, 55 S. CAL. L. REV. 1353, 1354–56 (1982).

information between parties,[50] the impact of information asymmetries between contracting parties,[51] the extent to which one party to a contract must disclose information,[52] and when contracts disclose so much information that they overload the reader and fail their communication goal.[53] Tort law similarly concerns itself with information. Tort law may, for example, create incentives for disclosure of information about known hazards by penalizing failure to do so[54] or encourage parties to gather information about possible dangers of their actions by defining foreseeability capaciously.[55]

These are only examples—private law and information has been discussed in great length and from many angles in legal scholarship. But note that each example above concerns transmittal or gathering of information to or by a directly interested party. Property law is interested in information that a potential buyer, neighbor, trespasser, or creditor has about the property. Contract law focuses on information exchange between parties to the contract. Tort law is concerned with information flows from a party taking an action to a party who may be injured.

In the realm of public law—roughly defined as law governing the rights and duties of individuals with respect to governments or government entities with respect to each other[56]—policy and scholarship is also deeply concerned with information.[57] Here, some government agencies take an explicitly aggregation-oriented view toward information. The Centers for Disease Control and Prevention ("CDC"), for instance, collects and publicizes certain health metrics, not because the audience is interested in the individual patient, but so that researchers can search for patterns that affect public health.[58] The Department

50. Lucian Arye Bebchuk & Steven Shavell, *Information and the Scope of Liability for Breach of Contract: The Rule of* Hadley v. Baxendale, 7 J.L. ECON. & ORG. 284, 285–86 (1991); Craswell, *supra* note 13, at 567.

51. Ian Ayres & Robert Gertner, *Strategic Contractual Inefficiency and the Optimal Choice of Legal Rules*, 101 YALE L.J. 729, 759–62 (1992).

52. *E.g.*, Anthony T. Kronman, *Mistake, Disclosure, Information, and the Law of Contracts*, 7 J. LEGAL STUD. 1, 1–2 (1978) (asking "if one party to a contract knows or has reason to know that the other party is mistaken about a particular fact, does the knowledgeable party have a duty to speak up or may he remain silent and capitalize on the other party's error?").

53. *E.g.*, Omri Ben-Shahar & Carl E. Schneider, *The Failure of Mandated Disclosure*, 159 U. PA. L. REV. 647, 687–89 (2011).

54. Benjamin C. Zipursky, *Rights, Wrongs, and Recourse in the Law of Torts*, 51 VAND. L. REV. 1, 20–21 n.73 (1998).

55. *Id.* at 47; *see also* Oren Bracha & Patrick R. Goold, *Copyright Accidents*, 96 B.U. L. REV. 1025, 1038 (2016) (discussing torts and information in the context of copyright).

56. Goldberg, *supra* note 46, at 1640.

57. *E.g.*, Abner S. Greene, *The Concept of the Speech Platform:* Walker v. Texas Division, 68 ALA. L. REV. 337, 344–45 (2016); Janet Freilich, *Government Misinformation Platforms*, 172 U. PA. L. REV. 1537, 1541–43 (2024); Beth Simone Noveck, Essay, *Rights-Based and Tech-Driven: Open Data, Freedom of Information, and the Future of Government Transparency*, 19 YALE HUM. RTS. & DEV. L.J. 1, 3–5 (2017).

58. *About the Vaccine Adverse Event Reporting System (VAERS)*, CTRS. FOR DISEASE CONTROL & PREVENTION, https://wonder.cdc.gov/vaers.html [https://perma.cc/8Q3X-TSZF].

of Housing and Urban Development collects and shares data on housing because, in the aggregate, the information can help evaluate housing programs and policy.[59] But these instances of focus on data aggregates are siloed in the sense that they primarily arise in the context of one subject area or one law and there is little comprehensive scholarship on the relationship between law and aggregate information.[60]

Rather, field-spanning scholarship on information in the public law sphere often relates, like its private law counterparts, to how information about individual events or entities is transmitted to parties interested in the specific piece of knowledge. This can be seen in the literature on information regulation and the use of information to achieve regulatory goals.[61] Here, the government produces information to aid consumers in making informed decisions (warning labels on cigarettes) or nudge companies into desired behavior by ensuring publicity for the companies' decisions (public disclosure of pollution emissions).[62] Another example is transparency-focused legislation such as the Freedom of Information Act ("FOIA"),[63] which envisions disclosure of information to an interested party who serves as a watchdog for corruption, like a journalist.[64] The focus on specific pieces of information, rather than aggregates, also characterizes literature on the design of legal institutions. For instance, scholarship on information produced by the court system tackles questions such as how procedural rules affect communication between parties—such as in pleadings and discovery[65]—and from parties to the interested

59. Nestor M. Davidson, *Affordable Housing Law and Policy in an Era of Big Data*, 44 FORDHAM URB. L.J. 277, 288–90 (2017).

60. One exception is the burgeoning movement towards open government, which encourages public entities to make public as much information as possible so that it can be used for a variety of purposes, including aggregation in data analytics technologies. Noveck, *supra* note 57, at 4.

61. ANTHONY I. OGUS, REGULATION: LEGAL FORM AND ECONOMIC THEORY 121 (1994).

62. *E.g.*, David M. Grether, Alan Schwartz & Louis L. Wilde, *The Irrelevance of Information Overload: An Analysis of Search and Disclosure*, 59 S. CAL. L. REV. 277, 301 (1986); Christine Jolls, Cass R. Sunstein & Richard Thaler, *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471, 1533–34 (1998).

63. 5 U.S.C. § 552 (2018).

64. For example, the Freedom of Information Act is designed to let the public know "what the government is up to." U.S. Dep't of Just. v. Reps. Comm. for Freedom of the Press, 489 U.S. 749, 773 (1989) (quoting EPA v. Mink, 410 U.S. 73, 105 (1973) (Douglas, J., dissenting)). Scholars noted that private companies aggregate information from FOIA requests. *See also* Margaret B. Kwoka, *FOIA, Inc.*, 65 DUKE L.J. 1361, 1365 (2016) (documenting how "commercial requesters . . . [have] dominat[ed] the [FOIA] landscape at some agencies").

65. *E.g.*, Scott Dodson, *New Pleading, New Discovery*, 109 MICH. L. REV. 53, 73 (2010); Alex Reinert, *Pleading as Information-Forcing*, 75 LAW & CONTEMP. PROBS. 1, 1 (2012).

public,[66] including when documents should be deemed confidential[67] and when settlement prevents the creation of information that would be useful to third parties.[68]

B. How Legal Information Is Used Today

Much of the literature just described focuses on communication of individual pieces of information. But, with the rise of big data analytics, legal information is increasingly aggregated and used in large-scale analyses.[69] This Section tracks the increased influence of collections of legal information, the many uses of aggregate legal information, and explores why legal information is a popular target for data aggregators. Throughout, this Section's emphasis is on how changing technology is changing how legal information is used.

First, there has been an empirical revolution in legal scholarship and policymaking.[70] Across fields of law, policies are increasingly driven by empirical evidence, courts are swayed by data, and legal decision-making has harnessed—or has been taken over by—predictive analytics, which depend on data to make decisions. This has been documented in fields as diverse as corporate law,[71] family law,[72] international law,[73] criminal law,[74] intellectual property law,[75] and more.[76] The impact of data on policies, decision-making, and enforcement

66. *E.g.*, Andrew D. Bradt & D. Theodore Rave, *The Information-Forcing Role of the Judge in Multidistrict Litigation*, 105 CALIF. L. REV. 1259–65 (2017); Samuel Issacharoff & Geoffrey Miller, *An Information-Forcing Approach to the Motion to Dismiss*, 5 J. LEGAL ANALYSIS 437, 438 (2013); Kishanthi Parella, *Reputational Regulation*, 67 DUKE L.J. 907, 912–13 (2018).

67. Howard M. Erichson, *Court-Ordered Confidentiality in Discovery*, 81 CHI.-KENT L. REV. 357, 357 (2006); Miller, *supra* note 17, at 436.

68. *E.g.*, Xinyu Hua & Kathryn E. Spier, *Information and Externalities in Sequential Litigation*, 161 J. INST. & THEORETICAL ECON. 215, 216 (2005); Michael J. Meurer, *The Settlement of Patent Litigation*, 20 RAND J. ECON. 77, 78 (1989).

69. *See generally* Michael A. Livermore & Daniel N. Rockmore, LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS (2019) (exploring advancements that allow for large-scale studies based on the text of court documents, opinions, and other data).

70. Ho & Kramer, *supra* note 12, at 1195–96.

71. Randall Thomas, *The Increasing Role of Empirical Research in Corporate Law Scholarship*, 92 GEO. L.J. 981, 982–84 (2004) (reviewing MARK J. ROE, POLITICAL DETERMINANTS OF CORPORATE GOVERNANCE: POLITICAL CONTEXT, CORPORATE IMPACT (2003)).

72. Clare Huntington, Essay, *The Empirical Turn in Family Law*, 118 COLUM. L. REV. 227, 267–71 (2018).

73. Gregory Shaffer & Tom Ginsburg, *The Empirical Turn in International Legal Scholarship*, 106 AM. J. INT'L L. 1, 7–8 (2012).

74. Tracey L. Meares, *Three Objections to the Use of Empiricism in Criminal Law and Procedure—and Three Answers*, 2002 U. ILL. L. REV. 851, 852–53.

75. Jeremy de Beer, *Evidence-Based Intellectual Property Policymaking: An Integrated Review of Methods and Conclusions*, 19 J. WORLD INTELL. PROP. 150, 150–51 (2016).

76. *E.g.*, William M. Landes, *The Empirical Side of Law & Economics*, 70 U. CHI. L. REV. 167, 167–69 (2003); Holger Spamann, *Empirical Comparative Law*, 11 ANN. REV. L. & SOC. SCI. 131, 132–34 (2015); Emanuel V. Towfigh, *Empirical Arguments in Public Law Doctrine: Should Empirical Legal Studies Make a "Doctrinal Turn"?*, 12 INT'L J. CONST. L. 670, 672–73 (2014).

will only grow as artificial intelligence is incorporated into these areas. As data increasingly drives how policies are crafted, research is conducted, judges' decisions are made, and laws enforced, underlying decisions that control what information is available for data analysis are increasingly salient.

Legal information is one of the primary sources of data both for legal applications and uses entirely unrelated to law.[77] Within law, legal information is an integral part of efforts to automate certain legal functions. For instance, patent prosecution software is trained on patent documents[78] and policing algorithms use past arrests and crime-related data.[79] Legal information drives a variety of predictive analytics tools, like software that predicts case outcomes.[80] And legal information is aggregated for research purposes, for example, studying the impact of laws relating to redlining[81] and to inform policymaking.[82]

But legal information is also aggregated for applications entirely unrelated to law. Data scientists, whether they are studying a research question or training an algorithm, gravitate toward legal information because it has many desirable characteristics. Legal datasets are often free, easy to access, well-formatted, contain a large number of data points, and sometimes have labeled characteristics.[83] Because legal data is available and high quality, it gets used frequently.[84] Property records are used to target advertising to people who bought houses in a particular price range.[85] Court records are used to train algorithms that can detect whether text contains sensitive information (which

77. Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 424–25 (2017).

78. Tabrez Y. Ebrahim, *Automation & Predictive Analytics in Patent Prosecution: USPTO Implications & Policy*, 35 GA. ST. U. L. REV. 1185, 1198–99 (2019); Sean Tu, Amy Cyphert & Sam Perl, *Limits of Using Artificial Intelligence and GPT-3 in Patent Prosecution*, 54 TEX. TECH L. REV. 255, 259 (2022).

79. Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 132–33 (2017) (discussing sources of bias and problems with this training data).

80. *E.g.*, Kevin D. Ashley, *A Brief History of the Changing Roles of Case Prediction in AI and Law*, 36 LAW CONTEXT 93, 94–96 (2019); Tammy W. Cowart, Roger Lirely & Sherry Avery, *Two Methodologies for Predicting Patent Litigation Outcomes: Logistic Regression Versus Classification Trees*, 51 AM. BUS. L.J. 843, 844 (2014).

81. *E.g.*, Price Fishback, Jonathan Rose, Kenneth A. Snowden & Thomas Storrs, *New Evidence on Redlining by Federal Housing Programs in the 1930s*, 141 J. URB. ECON. 1, 6–8 (2024) (assembling loan data from county land records).

82. *E.g.*, Ryan P. Kelly, Phillip S. Levin & Kai N. Lee, *Science, Policy, and Data-Driven Decisions in a Data Vacuum*, 44 ECOLOGY L.Q. 7, 9–10 (2017) (discussing data analysis by the National Marine Fisheries Service under the Endangered Species Act).

83. For instance, inventions described in patents are labeled with industry/technology classifications. Labeling is useful to train supervised machine learning systems to categorize items. *E.g.*, Kate Crawford & Trevor Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets*, 36 AI & SOC'Y 1105, 1107 (2021).

84. Freilich & Price, *supra* note 19, at 11.

85. *E.g.*, Xinyu Chen & Filip Biljecki, *Mining Real Estate Ads and Property Transactions for Building and Amenity Data Acquisition*, URB. INFORMATICS, 2022, at 1, 1.

has both legal and non-legal applications).[86] Information obtained during discovery in litigation proceedings is used to teach computers how to classify emails into folders.[87] Patents are used to train large language models.[88] Government legislation and judges' opinions are used to study language.[89] Mug shots are used to train facial recognition software.[90] The utility of legal information extends well beyond the traditional borders of law, serving as a vital input into applications in data science, technology, and more.

Some legal institutions are quite sensitive to the data aggregation uses of the information they produce. For example, the Office of the Chief Economist at the Patent and Trademark Office makes bulk data sets available "[t]o advance research on matters relevant to intellectual property, entrepreneurship, and innovation."[91] The National Institute of Standards and Technology has created a data portal to "provide[] a user-friendly discovery and exploration tool for publicly available datasets."[92] Some courts have implemented efforts to make records more accessible for bulk download.[93] Scholars who aggregate legal information for research are well aware of the uses and challenges surrounding the datasets available in their fields.[94] But discussions that focus on how information is *used* generally do not deeply explore doctrines affecting what information is *created*.

---

86. *E.g.*, Jan Neerbek, Morten Eskildsen, Peter Dolog & Ira Assent, *A Real-World Data Resource of Complex Sensitive Sentences Based on Documents from the Monsanto Trial*, 2020 PROC. 12TH CONF. LANGUAGE RES. & EVALUATION 1258, 1258.

87. Bryan Klimt & Yiming Yang, *The Enron Corpus: A New Dataset for Email Classification Research*, *in* MACHINE LEARNING: ECML 2004, at 217, 219 (Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti & Dino Pedreschi eds., 2004).

88. Schaul et al., *supra* note 7, at 1.

89. *See supra* notes 8–9 and accompanying text.

90. CRAWFORD, *supra* note 6, at 96.

91. *Research Datasets*, U.S. PAT. & TRADEMARK OFF., https://www.uspto.gov/ip-policy/economic-research/research-datasets [https://perma.cc/RTA5-WBEL] ("To advance research . . . the Office of the Chief Economist (OCE) releases datasets . . . .").

92. *About NIST Data*, NAT'L INST. OF STANDARDS & TECH., https://data.nist.gov/sdp/#/about [https://perma.cc/KM9B-QYCU].

93. For example, Massachusetts provides some datasets on criminal trials. Dep't of Rsch. & Plan., Mass. Trial Ct., *Criminal Court Reports and Dashboards*, MASS.GOV (June 2, 2023), https://www.mass.gov/info-details/criminal-court-reports-and-dashboards [https://perma.cc/R6BB-CQKL]. In general, court data is difficult to access in bulk.

94. *E.g.*, Jens Frankenreiter & Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANN. REV. L. & SOC. SCI. 39, 40 (2020) ("Techniques from the fields of artificial intelligence, natural language processing, text mining, network analysis, and machine learning are now routinely taken up by legal practitioners and law scholars. . . . This law-as-data approach uses computer-based tools to extract useful information from high-dimensional legal data sets, and in particular from collections of legal documents. This information can be analyzed to gain traction on long-standing research questions within law scholarship . . . .").

II. CASE STUDIES: THE NEW LEGAL INFORMATION

This Part provides examples of how legal doctrines (1) have unintended and unanticipated impacts on information production, dissemination, and use; and (2) interact in unexpected ways with modern uses and audiences for legal information, leading both to harmful consequences unforeseen by the law's designers and to happy surprises. Because law affects information through an uncountable number of pathways, this Part highlights the broad relationship between law and information by selecting examples that span very different areas of law, including public and private law, and different institutional structures. Further, because even laws designed to create information often produce results used for purposes beyond their original intent, this Part includes examples of doctrines deliberately directed toward information creation and those that are not.

This Part is divided roughly by whether information relates to litigation (Section II.A), regulation (Section II.B), or transactions (Section II.C). Note, however, that these categories have substantial overlap and the distinctions here are intended to be organizational, not definitional.

A. LITIGATION

The courtroom has always been a vital source of public information.[95] This is all the more true in the present day when written court documents, particularly judges' opinions, have taken on new importance in big data analytics of all sorts,[96] are a significant part of training data for large language models,[97] and are a critical source of input for efforts to automate adjudication.[98] But not all court information can be used in large-scale data analytics—these applications require *recorded*, *accessible*, and *aggregable* information.

1. Jurisdiction and Jury Trials

This Subsection begins with three procedural doctrines that (along with many other doctrines) determine whether information from a case can be incorporated into a data intensive application: the right to a jury, subject matter jurisdiction, and personal jurisdiction. None of these three doctrines were intended to be about information production, nor are they currently conceptualized that way, yet they mediate information production in ways that are likely to significantly bias efforts to automate many aspects of litigation.

---

95. *E.g.*, Richmond Newspapers, Inc. v. Virginia, 448 U.S. 555, 567–69 (1980) (discussing the long history of the open courtroom in English and American law and calling it "an indispensable attribute of an Anglo-American trial").

96. *See, e.g.*, Stephen J. Schultze, *The Price of Ignorance: The Constitutional Cost of Fees for Access to Electronic Public Court Records*, 106 GEO. L.J. 1197, 1216–17 (2018); Solove, *supra* note 29, at 1139–40.

97. Gao et al., *supra* note 9, at 1–4.

98. Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1168 (2019).

Note that the discussion here provides a stylized description of case proceedings—there are many additional rules and practices that affect information production from courts (the availability of arbitration,[99] standing rules,[100] settlement,[101] filing requirements and seals,[102] and many others[103]).

i. Jury Trials

When cases go to trial, they can be decided by either a judge or a jury. In a bench (judge) trial, the judge issues an opinion on the record—often a written opinion that includes both findings of fact and legal conclusions, as well as the judge's reasoning and response to the parties' arguments.[104] These written statements are publicly available, with some limitations.[105] By contrast, in jury trials, the jury renders a decision as to liability but there is no public record of its factual findings or reasoning.[106] Judge and jury trials therefore produce significantly different quantities of public information.[107] Some aspects of cases that go to juries are, from a big data perspective, invisible.[108]

However, the rules that determine whether a case is decided by a jury do not contemplate the informational function of the choice of the decision-maker and certainly do not consider the possibility of big data analytics. In the federal system, the Seventh Amendment gives parties the right to request a jury trial in civil cases in "[s]uits at common law."[109] The meaning of "suits at common law" (in contradistinction to suits "at equity") is complex, but it can be simplistically summarized for purposes of this Article as a determination based on remedies: Remedies at law involve damages whereas remedies at equity

99. Pamela K. Bookman, *Arbitral Courts*, 61 VA. J. INT'L L. 161, 167 (2021).

100. F. Andrew Hessick, *Standing, Injury in Fact, and Private Rights*, 93 CORNELL L. REV. 275, 289–90 (2008).

101. Owen M. Fiss, *Against Settlement*, 93 YALE L.J. 1073, 1078 (1984).

102. Stephen Wm. Smith, *Gagged, Sealed & Delivered: Reforming ECPA's Secret Docket*, 6 HARV. L. & POL'Y REV. 313, 313 (2012).

103. For a comprehensive overview, see generally Clopton & Huq, *supra* note 13. A related line of literature concerns selection of cases for litigation and how that influences the creation of legal doctrine. *E.g.*, Sepehr Shahshahani, *Hard Cases Make Bad Law? A Theoretical Investigation*, 51 J. LEGAL STUD. 133, 135–36 (2022).

104. For federal courts, see FED. R. CIV. P. 42(a)(1).

105. Although many opinions are "unpublished," they may still be available on court websites and through commercial services like Lexis and Westlaw. However, not all written opinions are easily publicly accessible. *E.g.*, McAlister, *supra* note 16, at 1103–05.

106. *E.g.*, Mason v. Ford Motor Co., 307 F.3d 1271, 1274 (11th Cir. 2002) ("[W]hen a typical general verdict is employed, the jury is asked to articulate no factual findings other than the ultimate finding of which party wins.").

107. Note that jury trials may still produce some written opinions by judges. For instance, judges may provide a written opinion on a motion for a directed verdict or on appeal, FED. R. CIV. P. 50, and at early stages of the case judges may publish decisions about motions to dismiss, FED. R. CIV. P. 12(b)(6), or motions for summary judgment, FED. R. CIV. P. 56.

108. Some information is still available, such as the complaint, which is filed in every case regardless of whether it goes to a judge or jury. FED. R. CIV. P. 3, 8.

109. U.S. CONST. amend. VII.

are nonmonetary, such as injunctions, specific performance, and accounting for profits.[110]

Thus, plaintiffs seeking damages may have a jury trial while plaintiffs seeking an injunction will not. This distinction—and the ability of litigants to waive jury trials if both sides so agree[111]—creates systematic differences across areas of law and types of litigants in jury versus bench trials. For example, business litigants were more likely than individual litigants to have cases heard by judges.[112] 98.7% of medical malpractice and 92.1% of motor vehicle trials were heard by a jury whereas only 3.5% of mortgage foreclosure trials and 15% of eminent domain cases.[113] Because certain types of cases are more likely to go to a jury, less information is available about those areas of law.

ii. Jurisdiction

Jury rules are not the only procedural doctrines that appear unrelated to information production but have big practical impacts. Jurisdictional rules have the same characteristics and consequences. Constitutional and statutory bounds on federal courts' subject matter jurisdiction limit the types of cases that can be heard in federal court to, most commonly, cases arising under federal law[114] and cases between citizens of different states.[115] Because federal court opinions are more likely to be indexed by private sources like Westlaw, Lexis, or Google, this has implications for what information is available and how it can be used.[116] For example, Google Scholar indexes cases for all federal district courts but does not index lower court cases for all states.[117] But subject matter jurisdiction is not a doctrine motivated by information production—rather its rationales include ensuring uniformity of federal law, protecting out-of-state parties from bias or hostility in state courts, funneling federal legal questions to more expert federal judges, and managing the caseload of federal courts.[118]

---

110. Samuel L. Bray, *The System of Equitable Remedies*, 63 UCLA L. REV. 530, 533–35 (2016).

111. FED. R. CIV. P. 38(d).

112. LYNN LANGTON & THOMAS H. COHEN, BUREAU OF JUST. STAT.: U.S. DEP'T OF JUST., CIVIL BENCH AND JURY TRIALS IN STATE COURTS, 2005, at 2 (2009), https://bjs.ojp.gov/content/pub/pdf/cbjtsc05.pdf [https://perma.cc/8D7H-FAED].

113. *Id.* (percents are of cases that went to trial, not cases commenced).

114. 28 U.S.C. § 1331.

115. *Id.* § 1332(a)(1).

116. *See* Christina L. Boyd, Pauline T. Kim & Margo Schlanger, *Mapping the Iceberg: The Impact of Data Sources on the Study of District Courts*, 17 J. EMPIRICAL LEGAL STUD. 466, 466–68 (2020).

117. *Select Courts*, GOOGLE SCHOLAR, *https://scholar.google.com/scholar_courts* [https://perma.cc/F52V-W5RA].

118. James W. Moore & Donald T. Weckstein, *Corporations and Diversity of Citizenship Jurisdiction: A Supreme Court Fiction Revisited*, 77 HARV. L. REV. 1426, 1431–32 (1964) (discussing the caseload management rationale for diversity jurisdiction); John F. Pries, *Reassessing the Purposes of Federal Question Jurisdiction*, 42 WAKE FOREST L. REV. 247, 248–49 (2007) (describing the first three rationales for federal question jurisdiction).

Rules about personal jurisdiction, which have the effect of determining which state courts can hear cases, also inadvertently impact information production.[119] Different states have different systems for making court dockets publicly available, with some more accessible and amenable to bulk download than others.[120] Private legal research databases provide comprehensive coverage of some states but not others.[121] Thus, the state in which a case is litigated affects whether information from that case will be easy to find and incorporate into a data-intensive analysis.

As with subject matter jurisdiction, the rationale for personal jurisdiction is entirely unrelated to information production. Rather, personal jurisdiction rules arise from considerations of due process and limits on state powers.[122] Many other procedural rules also affect the likelihood that information about a case will be public.[123] The overarching point is that rules assigning cases to particular courts and affecting the likelihood that a case will produce a written opinion influence the production and dissemination of information through court dockets—yet the rationales undergirding these rules are quite unrelated to information. And like rules about juries, rules about jurisdiction severely skew the pool of cases that are aggregable for large-scale analysis.

2. Discovery

Judicial opinions are not the only source of large-scale information produced by litigation. Procedural rules governing discovery allow collection and dissemination of otherwise inaccessible data sources. This Subsection traces how information obtained in litigation discovery was instrumental for computer scientists working on email sorting and classification technology.

Countless AI applications have been trained by what "remains the largest public domain database of real e-mails in the world—by far."[124] This email dataset, called the Enron Corpus, came from an investigation by the Federal

119. The personal jurisdiction inquiry is concerned with whether the court has the power to render a judgement against the defendant. Int'l Shoe Co. v. Washington, 326 U.S. 310, 316 (1945).

120. JASON TASHEA, FED. OF AM. SCIENTISTS, DAY ONE PROJECT: DIGITIZING STATE COURTS, EXPANDING ACCESS TO JUSTICE 2–3 (2021), https://fas.org/wp-content/uploads/2021/01/digitizing-state-courts.pdf [https://perma.cc/9HRF-VE5X].

121. For instance, Lex Machina, a division of LexisNexis, recently added partial coverage for Oregon state courts. Gloria Huang, *Lex Machina Launches Enhanced Legal Analytics for Oregon Court Modules*, LEX MACHINA: BLOG (July 19, 2023), https://web.archive.org/web/20240422220544/https://lexmachina.com/blog/lex-machina-launches-enhanced-legal-analytics-for-oregon-courts-modules (on file with the *Iowa Law Review*).

122. World-Wide Volkswagen Corp. v. Woodson, 444 U.S. 286, 291–92 (1980).

123. Most notably, those affecting the likelihood that a case will be brought in court in the first place and the likelihood that the case will settle.

124. Jessica Leber, *The Immortal Life of the Enron E-mails*, MIT TECH. REV. (July 2, 2013), https://www.technologyreview.com/2013/07/02/177506/the-immortal-life-of-the-enron-e-mails [https://perma.cc/W4TU-W3EJ].

Energy Regulatory Commission ("FERC").[125] In 2001, fraudulent accounting practices caused energy company Enron to go bankrupt, resulting in the FERC investigation. Because information discovered in these types of proceedings can be made public in the absence of a protective order,[126] and because of strong public interest in the case, FERC publicly released the hundreds of thousands of documents—mostly emails—it obtained in discovery from Enron.[127]

Emails are generally private, and it is difficult to acquire large-scale email databases, so the Enron Corpus directly influenced the development software on fraud detection, counterterrorism, email foldering, detecting sensitive text and studies of email etiquette, use of expressions, and patterns of speech.[128] From the perspective of the legal system, perhaps the most impactful use of the Enron Corpus was in developing one of the earliest widely-accepted uses of machine learning in law: predictive coding in ediscovery.[129]

Although improvements in computer science and advancements in ediscovery software were unintended spillover effects of discovery doctrines, these doctrines were certainly made with information exchange and publicity in mind. As a default, documents obtained through discovery can be publicly released by either side in the absence of a confidentiality agreement or protective order.[130] Courts assessing proposed protective orders seriously consider the merits of public disclosure of information.[131] But case law on discovery disclosure

---

125. John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N.Y. TIMES (Mar. 4, 2011), https://www.nytimes.com/2011/03/05/science/05legal.html (on file with the *Iowa Law Review*) ("Such [ediscovery tools] owe a debt to an unlikely, though appropriate source: the electronic mail database known as the Enron Corpus.").

126. Rules of Discovery for Trial-Type Proceedings, 52 Fed. Reg. 6957, 6963–64 (Mar. 2, 1987); Harvey L. Reiter, *The FERC's New Rules of Discovery: A Welcomed Approach*, 8 ENERGY L.J. 35, 55–56 (1987).

127. Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 610–11 (2018) (discussing the Enron emails and their use in computer science research).

128. *Id.* at 611; *see also* Nathan Heller, *What the Enron E-Mails Say About Us*, NEW YORKER (July 17, 2017), https://www.newyorker.com/magazine/2017/07/24/what-the-enron-e-mails-say-about-us (on file with the *Iowa Law Review*) (describing various studies from universities and scholars).

129. Machine learning has been part of the discovery process in litigation for over a decade. *E.g.*, Da Silva Moore v. Publicis Groupe & MSL, 287 F.R.D. 182, 197 (S.D.N.Y. 2012); Order Approving the Use of Predictive Coding for Discovery at 1, Glob. Aerospace Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012); Kleen Prods. LLC v. Packaging Corp. of Am., No. 10 C 5711, 2012 WL 4498465, at *5 (N.D. Ill. Sept. 28, 2012).

130. Seattle Times Co. v. Rhinehart, 467 U.S. 20, 36–37 (1984); FED. R. CIV. P. 26(c)(1). Such agreements and orders are quite common, and district court judges have "wide latitude in designing protective orders." Poliquin v. Garden Way, Inc., 989 F.2d 527, 532 (1st Cir. 1993).

131. For examples of courts finding disclosure is in the public interest, see, e.g., *Poliquin*, 989 F.2d at 535 (holding disclosure in the public interest to "avoid wasteful duplication of discovery in other cases"); Shingara v. Skiles, 420 F.3d 301, 308 (3d Cir. 2005) (finding that the public had an interest in the evidence supporting the claims and defenses of the case); Glenmede Trust Co. v. Thompson, 56 F.3d 476, 485 (3d Cir. 1995) ("[S]haring of information among current and potential litigants is furthered by open proceedings."). Courts have hidden information from

generally does not contemplate use of aggregate data or use of data for purposes entirely unrelated to the parties or issues in the case.[132] Discovery rules were certainly not intended to produce large public datasets of emails and other documents for purposes of developing discovery software and other research on texts and language. The release of the Enron Corpus was not, therefore, a deliberate strategy to incentivize the development of computer science and legal technology but instead a happy accident (from the perspective of the twelve-billion-dollar ediscovery industry).[133]

B. REGULATION

This Section turns from litigation procedure to regulatory frameworks. Regulatory agencies are often conduits for public data dissemination and thus play an outsized role in mediating the relationship between law and information. The regulator's view of information may, however, be too narrowly focused on the specific area under regulation and not account for broader informational effects. This is illustrated below with the example of workers' compensation systems. The example additionally explores how new uses and audiences for information on employee injuries have substantially changed the informational consequences of some tort doctrines and regulatory structures, leading to inadvertent harm.

Workers can be either employees or independent contractors.[134] The distinction between the categories was originally devised to determine when employers could be held liable for employees' actions.[135] The test has evolved, but information production is not and has never been a consideration in creating or applying the doctrines.[136] Yet whether a worker is classified as an employee or an independent contractor has an enormous bearing on the information publicly available about that worker. There are public databases that compile all injuries suffered by employees, but there are no equivalent

public disclosure after finding that discovery proceeds more efficiently when confidentiality is guaranteed or for avoidance of embarrassment, oppression, or undue burden or expense. *Id.*; FED. R. CIV. P. 26(c)(1).

132. When courts discuss using information obtained in discovery for purposes beyond the case in which it was obtained, these purposes are generally linked to broader public interest in the parties to or subject of the case. *E.g.*, Int'l Union v. Garner, 102 F.R.D. 108, 113 (M.D. Tenn. 1984) ("There has been no showing that plaintiffs have used discovery to develop information unrelated to the case before the Court. If information developed at trial and made part of the public record affects public opinion or fuels unfair labor practice charges, that is not a reason to seal all records in this case.").

133. *EDiscovery Market Insights*, SKYQUEST (Oct. 2024), https://www.skyquestt.com/report/e-discovery-market [https://perma.cc/J8VK-JSER].

134. *E.g.*, NLRB v. United Ins. Co. of America, 390 U.S. 254, 258 (1968) ("There are innumerable situations which arise in the common law where it is difficult to say whether a particular individual is an employee or an independent contractor . . . .").

135. Stephen F. Befort, *Revisiting the Black Hole of Workplace Regulations: A Historical and Comparative Perspective of Contingent Work*, 24 BERKELEY J. EMP. & LAB. L. 153, 168 (2003).

136. For a discussion of the modern multifactorial test, see *United Ins. Co.*, 390 US at 259.

public databases for independent contractor injuries. This difference arises from the workers' compensation system, which benefits employees who are injured at work.[137] Like the employee/independent contractor distinction, production of information was not an underlying rationale or consideration in workers' compensation systems. Rather, the primary goal of workers' compensation is to protect and compensate workers.[138] The specifics vary by state, but generally employers must file a detailed report of the injury with the state government.[139] Many states release those reports—without identifying information—for public use.[140]

Workers' compensation programs do not apply to independent contractors.[141] The rationale is that, because they operate mostly independent of the employer's control, the employer cannot take measures to prevent injuries to independent contractors.[142] While this explanation may be reasonable with respect to an employer's ability to control risk, it means that injuries to independent contractors are invisible.[143] There is no reporting system for injuries to independent contractors. It is very difficult to track injuries to, for instance, workers in gig economy jobs like drivers for Uber and Lyft.[144]

1. Consequences of Changing Uses and Audiences for Information

The main goal of workers' compensation programs is compensation. But the programs also have major informational effects. Workers' compensation reports, because they are an easily accessible and richly detailed source of data, are aggregated and used in both government and third-party applications. For instance, Massachusetts uses workers' compensation data to set policies to

137. 58 AM. JUR. *Workmen's Compensation* § 1 (1975).

138. 82 AM. JUR. 2D *Workers' Compensation* § 10 (2013) ("The primary goal . . . is to aid the injured employee and protect him or her against the special risks of employment with comprehensive coverage for his or her injuries." (footnote omitted)).

139. *E.g.*, DEP'T OF INDUS. ACCIDENTS, COMMONWEALTH OF MASS., EMPLOYER'S GUIDE TO THE MASSACHUSETTS WORKERS' COMPENSATION SYSTEM 4 (2019), https://www.mass.gov/doc/employers-guide-to-workers-compensation-english-0/download [https://perma.cc/5W9C-HEGT].

140. For example, Oregon provides record-level information for claims including data on the date of the injury, the age and gender of the injured party, a description of their occupation and the injury, the source of the injury, details about the employer, and more. *E.g.*, *Oregon Workers' Compensation Record Level Claims*, OREGON.GOV OPEN DATA PORTAL (Nov. 29, 2023), https://data.oregon.gov/Business/Oregon-Workers-Compensation-Record-Level-Claims/t9t7-8a2y [https://perma.cc/UH4Y-5M9G].

141. 99 C.J.S. *Workers' Compensation* §§ 205, 207 (2023). There are some exceptions to the rule that workers' compensation does not cover independent contractors. These vary by state. For instance, Louisiana workers' compensation covers independent contractors who do manual labor. Gaspard v. Travelers Ins. Co., 284 So. 2d 104, 108 (La. Ct. App. 1973).

142. *E.g.*, Potter v. Hawaii Newspaper Agency, 974 P.2d 51, 59 (Haw. 1999).

143. Other types of work are also invisible, for instance injuries to those laboring in unpaid domestic work. Miriam A. Cherry, *People Analytics and Invisible Labor*, 61 ST. LOUIS U. L.J. 1, 2 (2016).

144. Molly Tran & Rosemary K. Sokas, *The Gig Economy and Contingent Work: An Occupational Health Assessment*, 59 J. OCCUPATIONAL & ENV'T MED. e63, e64 (2017) (noting that "there is little information regarding the health aspects of gig work").

prevent occupational injuries.[145] At the federal level, the National Institute for Occupational Health and Safety ("NIOSH") uses workers' compensation records to determine "where hazards exist and what interventions are effective"[146] and to direct its research funding.[147] The private sector uses workers' compensation data to train AI programs that give advice on preventing injuries,[148] prioritize allocation of safety resources,[149] and compare injuries among one's own employees to those of competitors.[150]

Thus, information produced by workers' compensation laws is important and impactful. This information has substantial benefits in preventing employee injuries.[151] However, the data spotlight on employee injuries also creates an unintentional harm: funneling funding and focusing policymaking *away* from independent contractor injuries.

The readily available data for employee injuries contrasted with lack of data for independent contractors means that employee-based data drives policy and sets priorities.[152] For instance, when government agencies use workers' compensation data to decide which projects should be funded, that suits the needs of employees but may not fit the needs of non-employees. And if a company uses predictive analytics software trained on workers' compensation claims to prioritize funding safety projects, it will focus on the needs of

145. MASS. DEP'T OF INDUS. ACCIDENTS, MASS. DEP'T OF PUB. HEALTH & MASS. DEP'T OF LAB. STANDARDS, USING MASSACHUSETTS WORKERS' COMPENSATION DATA TO IDENTIFY PRIORITIES FOR PREVENTING OCCUPATIONAL INJURIES AND ILLNESSES AMONG PRIVATE SECTOR WORKERS (2019), https://www.mass.gov/doc/dph-dia-and-dls-release-new-study-on-utilization-of-workers-compensation-data/download [https://perma.cc/X7AH-QWE9].

146. DAVID F. UTTERBACK, ALYSHA R. MEYERS & STEVEN J. WURZELBACHER, WORKERS' COMPENSATION INSURANCE: A PRIMER FOR PUBLIC HEALTH iii (2014), https://stacks.cdc.gov/view/cdc/21466 [https://perma.cc/CK39-7C6Z].

147. Alysha R. Meyers, *AI and Workers' Comp*, CTRS. FOR DISEASE CONTROL & PREVENTION: NIOSH SCI. BLOG (May 1, 2019), https://blogs.cdc.gov/niosh-science-blog/2019/05/01/ai-workers-comp [https://perma.cc/BF6E-G8US].

148. Griffin Schultz, *Using Advanced Analytics to Predict and Prevent Workplace Injuries*, 88 OCCUP. HEALTH & SAFETY 90, 90 (2012).

149. Lisa Romeu, *7 Ways Data Analytics Can Improve Your Workers' Compensation Program*, PMA COS. (Dec. 20, 2021), https://www.pmacompanies.com/blog/7-ways-data-analytics-can-improve-your-workers-compensation-program [https://perma.cc/KZB3-UYWM].

150. *Workers' Compensation Benchmarking Model*, RISKONNECT, https://riskonnect.com/solutions/claims-administration-software/workers-compensation-benchmarking-model [https://perma.cc/J53W-C7C7].

151. *E.g.*, NAT'L INST. FOR OCCUP. SAFETY & HEALTH, *Center for Workers' Compensation Studies* (Oct. 17, 2024), https://www.cdc.gov/niosh/centers/workers-comp.html [https://perma.cc/PBW5-34CM] ("Workers' compensation employer data can help develop 'leading indicators' that identify workplace hazards and controls that most impact future injuries/ illnesses. Prediction leads to prevention.").

152. More generally, it is difficult to set policy for and uncover problems affecting workers who are not employees. *E.g.*, Orly Lobel, *The Gig Economy and the Future of Employment and Labor Law*, 51 U.S.F. L. REV. 51, 64 (2017) (noting that data gathering can be sparse for workers in the gig economy and that discrimination can be combatted if "regulatory agencies more actively enforce data mining and reporting").

employees, not independent contractors.[153] Though non-employee worker injuries are difficult to study, the Bureau of Labor Statistics notes that the safety risks for independent workers differ from those of employees, and thus the needs of independent contractors may not be captured in employee data.[154]

The negative effects of invisibility described here are at least in part a direct consequence of the rise of big data analytics. For applications that do not require data aggregation, the availability of information about workplace injuries to both employees and independent contractors is roughly equivalent. The injured party and their employer would both know about the injury, as would any party asked to pay for the injury. But publicizing and aggregating workers' compensation records, coupled with an increasing reliance on data analytics to make decisions and set priorities, means that employee interests are prioritized over independent contractors simply because there is more data available about the former.

C. TRANSACTIONS

This Section turns to examples of transactional information: contracts and property records. These examples both have substantial regulatory components but are included in this Section because the primary underlying source of legal information comes from transactions between private entities. Legal rules govern the extent to which private transactions become public information. Legal rules often do this deliberately and with significant consideration to information but changing uses of and audiences for transactional information nonetheless produce unexpected effects.

1. Contracts

Contract drafting software is an example of how legal rules about disclosing transaction leads to unintentional consequences. Contracts between private parties are generally not public documents and it is difficult to find large datasets of contracts.[155] For applications that require large datasets for

153. Perhaps it is suitable for companies to focus primarily on employees, but the line between employees and non-employees is increasingly blurred, and company decisions can greatly affect independent contractors as well as employees. *See generally* Miriam A. Cherry & Antonio Aloisi, *"Dependent Contractors" in the Gig Economy: A Comparative Approach*, 66 AM. U. L. REV. 635 (2017).

154. Stephen Pegular & Matt Gunter, *Fatal Occupational Injuries to Independent Workers*, *in* 8 WORKPLACE INJURIES No. 10 (U.S. Bureau of Lab. Stat., Beyond the Numbers, 2019), https://www.bls.gov/opub/btn/volume-8/fatal-occupational-injuries-to-independent-workers.htm#_edn3 [https://perma.cc/ZG8Z-P9RS].

155. *E.g.*, Michael Curtotti & Eric C. McCreath, *A Corpus of Australian Contract Language*, 2011 INT'L CONF. ON A.I. & L. 199, 200 (noting that there are few publicly available contracts corpora and that "[a] likely reason for the slower development of this field is that until recently it would have been extremely difficult to obtain contract texts").

training or evaluation—like data analytics and contract drafting AI[156]—a major source of bulk contract data is the Securities and Exchange Commission's ("SEC") EDGAR system, an online repository of filings.[157] Contracts from EDGAR are used both to train AI applications and as a source of data for researchers to study contracts.[158]

EDGAR, and its parent institution, the SEC, are focused on information disclosure. The SEC requires companies to disclose information because "[t]he system is designed to provide investors with material information, foster investor confidence, contribute to . . . fair and orderly markets . . . and inhibit fraud . . . ."[159] When an electronic database of SEC filings was first proposed in the 1980s, it was with full recognition of the potential for data aggregation and its value for market research and analysis.[160] Doctrines governing SEC disclosure requirements affect whether information will be posted and aggregable on EDGAR, and information-provision is the major policy goal underlying those doctrines.[161]

---

156. Beverly Rich, *How AI Is Changing Contracts*, HARV. BUS. REV. (Feb. 12, 2018), https://hbr.org/2018/02/how-ai-is-changing-contracts [https://perma.cc/LUY4-3735]. For examples of AI applications trained on contract data, see Dipankar Chakrabarti et al., *Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support*, 2018 INST. ELEC. & ELEC. ENG'G REGION 10 CONF. 683, 684.

157. For example, LexPredict and Bloomberg's contract analysis software trained on contracts from EDGAR. Kathryn D. Betts & Kyle R. Jaep, *The Dawn of Fully Automated Contract Drafting: Machine Learning Breathes New Life into a Decades-Old Promise*, 15 DUKE L. & TECH. REV. 216, 226 (2017). Other sources of data are a company's own contracts, contacts disclosed as exhibits in cases, and contracts that are placed online and can be collected from web searches. *See, e.g.*, Andrew Antos & Nischal Nadhamuni, *Practical Guide to Artificial Intelligence and Contract Review*, in RESEARCH HANDBOOK ON BIG DATA AND LAW 467, 472 (Roland Vogl ed., 2021) (discussing use of a company's own contracts as training data); Michael Curtotti & Eric C. McCreath, *A Corpus of Australian Contract Language*, 2011 PROC. 13TH INT'L CONF. ON A.I. & L. 198, 199 (providing an example of contracts collected online).

158. *E.g.*, Yuanyuan Chen & Anandhi Bharadwaj, *An Empirical Analysis of Contract Structures in IT Outsourcing*, 20 INFO. SYS. RSCH. 484, 488 (2009); Gabriel Rauterberg & Eric Talley, *Contracting Out of the Fiduciary Duty of Loyalty: An Empirical Analysis of Corporate Opportunity Waivers*, 117 COLUM. L. REV. 1075, 1121 (2017); Julian Nyarko, *We'll See You in . . . Court! The Lack of Arbitration Clauses in International Commercial Contracts*, 58 INT'L REV. L. & ECON. 6, 9–10 (2019); Sarath Sanga, *Choice of Law: An Empirical Analysis*, 11 J. EMPIRICAL LEGAL STUD. 894, 903 (2014).

159. U.S. SEC. & EXCH. COMM'N, FIFTY-FIFTH ANNUAL REPORT 16 (1989).

160. James Packard Love, *The Ownership and Control of the U.S. Securities and Exchange Commission's EDGAR System*, 20 GOV'T PUBL'N REV. 61, 63 (1993) ("From the very beginning, the SEC proposed regulating the prices and services the contractor would charge for bulk or wholesale access to EDGAR filings . . . .").

161. *E.g.*, TSC Indus. Inc. v. Northway, Inc., 426 U.S. 438, 448 (1976) (describing the purpose of disclosures as "to enable the shareholders to make an informed choice"); Basic Inc. v. Levinson, 485 U.S. 224, 231 (1988) ("Acknowledging that certain information concerning corporate developments could well be of 'dubious significance,' . . . the Court was careful not to set too low a standard of materiality; it was concerned that a minimal standard might bring an overabundance of information . . . 'bury[ing] the shareholders in an avalanche of trivial information . . . .'" (citation omitted)); Matrixx Initiatives, Inc. v. Siracusano, 563 U.S. 27, 28 (2011) (defining the

However, policies underlying EDGAR and doctrines governing disclosure requirements are about securities markets; they were not created with contract drafting software in mind. The ability to generate large datasets of contracts and subsequently use them to them for contracts-related AI and research studies about contracts is an unexpected side effect of the SEC rules. While this has enabled many useful applications and studies, it also leads to bias because contracts from EDGAR are not representative of contracts in general. SEC rules require registered companies to file "contract[s] not made in the ordinary course of business that is material to the registrant."[162] These are generally loan contracts, stock purchasing agreements, and employment agreements for CEOs and other important employees.[163] Additionally, only companies with "total assets exceeding $10,000,000 and a class of equity security . . . held . . . [by five hundred]" or more persons need to file contracts with the SEC.[164] Because most contracts are not filed with the SEC, and those that are are not a representative sample, both the AI applications and the research conclusions may not generalize.[165]

2. Property Leases

AI, although a major topic of discussion at present, is not the only way in which legal information has unforeseen effects. As a general matter, law and policymaking have turned toward the empirical, with increased emphasis on data and numerical evidence.[166] Legal information can impact policymaking in ways that were unintended by the underlying doctrines.

This can be seen in the context of property. Property law is deeply concerned with the communication of information and there is abundant public information about property transactions including (depending on the

materiality requirement as "satisfied when there is a substantial likelihood that the disclosure of the omitted fact would have been viewed by the reasonable investor as having significantly altered the total mix of information made available" (internal quotation marks omitted) (quoting *Basic Inc.*, 485 U.S. at 231–32)).

162. 17 C.F.R. § 229.601(b)(10)(i)(A) (2024).

163. Nyarko, *supra* note 158, at 10.

164. Securities Exchange Act of 1934 § 12(g), 15 U.S.C. § 78l.

165. SEC documents are not the only public source of contracts, but a similar selected sample problem applies to other contract data. For instance, court records are a public source of contracts, but certainly not a representative source. Litigation in general is not representative of all underlying disputes. Priest & Klein, *supra* note 5, at 1–2. And there will be systematic biases in the types of contracts that can be found in court cases—some types of contracts are more likely to contain arbitration clauses, and therefore disputes are resolved in (private) arbitration, not (public) court. *See* Pamela K. Bookman, *The Adjudication Business*, 45 YALE J. INT'L L. 227, 279–80 (2020).

166. *See generally* Richard Lempert, *Empirical Research for Public Policy: With Examples from Family Law*, 5 J. EMPIRICAL LEGAL STUD. 907 (2008); Jennifer Arlen, *The Essential Role of Empirical Analysis in Developing Law and Economics Theory*, 38 YALE J. ON REGUL. 480 (2021) (discussing the growing importance of empirical legal research).

jurisdiction), the parties' names, sale price, taxes paid, and more.[167] The origins of property law's concern with information lie in the need to provide information about a property to others who might interact with the property—for instance, public recordation of property's boundaries and limitations on the forms property transactions can take.[168] Property's focus on communicating individual-level information creates difficulties for data collection, and therefore policymaking, in certain areas.

One example is evictions. Because leases are not a matter of public record, involuntary terminations of a lease for failure to pay—evictions—are not public and are difficult to collect and study.[169] Eviction rules are governed by state law and vary, but many states do not require public disclosure of eviction unless the landlord needs a court's help to complete the eviction process.[170]

This means that a large subset of evictions is invisible. That can benefit tenants—a foreclosure affects credit scores and may show up in background checks, whereas a voluntary eviction may not—but can also hurt tenants because it is difficult to study evictions on a large scale. Although evictions are undeniably a major policy issue,[171] there are "no comprehensive" local or federal statistics on evictions.[172] It is therefore challenging to spot, for instance, patterns of discrimination.[173] Current empirical approaches rely on court records, which are not always available (one study found that 3.6 million eviction cases were filed annually, and estimated that this undercounted eviction cases by about one million per year) and are biased in the sense that not all evictions produce court records.[174] Because eviction is often invisible, it is also hard to understand its causes and consequences. Scholars note that this sort of

167. Christopher L. Peterson, *Foreclosure, Subprime Mortgage Lending, and the Mortgage Electronic Registration System*, 78 U. CIN. L. REV. 1359, 1365–66 (2010).

168. Merrill & Smith, *supra* note 17, at 9; *see also* Henry Hansmann & Reinier Kraakman, *Property, Contract, and Verification: The* Numerus Clausus *Problem and the Divisibility of Rights*, 31 J. LEGAL STUD. S373, S402 (2002); Molly Shaffer Van Houweling, *The New Servitudes*, 96 GEO. L.J. 885, 899 (2008).

169. *E.g.*, James F. O'Rourke, *Some Considerations to Be Observed in the Recording of Leases*, 12 REAL PROP. PROB. & TRUST J. 256, 256 (1977) ("Few leases of real property ever find their way into the land records.").

170. *E.g.*, MISS. CODE ANN. § 89-7-35 (2024) (issuance of warrant for removal).

171. Ashley Gromis et al., *Estimating Eviction Prevalence Across the United States*, PROCS. NAT'L ACAD. SCIS. 1, 1 (2022), https://www.pnas.org/doi/full/10.1073/pnas.2116169119 [https://perma.cc/88HM-DTKT] ("Court-ordered eviction and displacement due to eviction are primary causes of homelessness and have long-term effects on material hardship and health." (footnotes omitted)).

172. *Id.*

173. *See, e.g.*, Deena Greenberg, Carl Gershenson & Matthew Desmond, *Discrimination in Evictions: Empirical Evidence and Legal Challenges*, 51 HARV. C.R.-C.L. L. REV. 115, 118 (2016) ("[I]t is imperative that legal scholars design methods to detect and prevent discrimination in eviction decisions to ensure that protected groups are not disproportionately subjected to the negative consequences of involuntary displacement.").

174. Gromis et al., *supra* note 171, at 1.

understanding requires longitudinally and geographically comprehensive or at least representative data that is not available for evictions.[175] Further, although policymakers are interested in interventions to reduce evictions and consequent displacement and homelessness, "[u]nderstanding the scope and geography of the problem and evaluating the effectiveness of different policy interventions requires having access to accurate eviction data . . . ."[176] With respect to evictions, lack of data hampers policy efforts.

The case studies above are only a sample of the many areas where law produces data and casts light on the shadows that limited data leave behind. They illustrate the wide variety of doctrines that impact information production, the breadth of uses for legal information, and many unintentional consequences of that information.

III. UNINTENDED CONSEQUENCES OF LEGAL INFORMATION

The Part above provided several examples of how legal doctrines and institutions can inadvertently impact information availability in ways that in turn shape the development of policy, research, and technology. This Part turns to a more general discussion of the relationship between law and information. Section III.A expands on the benefits of legal information and the harms that arise from the current system where law produces information in often unplanned and arbitrary ways. Sections III.B and III.C address two specific areas where legal information has an outsized effect: privacy and the development of legal technology.

A. BENEFITS AND BIASES

Legal information has a host of benefits. It informs the public about legal transactions, aiding transparency and accountability.[177] It permits the growth and evolution of law and legal arguments. It improves understanding of law's impact and footprint on society. It allows for empirically-informed policymaking.[178] Beyond law, it facilitates the development of linguistic technologies and other data analytics applications.[179] Many of these benefits are either made possible or enhanced by new uses and audiences for legal information. With the rise of big data analytics and more widespread distribution of legal information, it is easier to see large-scale patterns in legal

175. *Id.* at 2 (noting that even when court records are available, "[d]ata coverage varied significantly across these states and counties over time . . . preventing comprehensive collection of records across years [and states]").

176. Adam Porton, Ashley Gromis & Matthew Desmond, *Inaccuracies in Eviction Records: Implications for Renters and Researchers*, 31 HOUS. POL'Y DEBATE 377, 378 (2021).

177. *E.g.*, Gannett Co. v. DePasquale, 443 U.S. 368, 376 (1979) (discussing the "public's vital interest in open judicial proceedings").

178. *See supra* Part II (providing examples).

179. *See supra* Part II (providing examples).

transactions, spot problems, conduct panoramic studies, and incorporate legal information into non-legal technologies such as large language models.

These applications of legal information are useful and beneficial. But the way that law affects information also poses concerns. When legal rules allow for aggregation of information about some types of transactions, cases, organizations, and people but not others, it creates the potential for biases in informational analysis.[180] This Section surveys how legal doctrine can create such biases and highlights the problems caused by the legal system's current failure to consider how legal rules affect information availability. Although many of these problems existed before the advent of large-scale data analysis, the increased reliance on predictive analytics and artificial intelligence, coupled with the empirical turn in many areas of law, means that choices about what data to collect and where empirical analysis will be possibly take on new importance.

1. Generalizability and Transferability

Algorithms trained on data from one setting do not always transfer well to different settings, even settings that are relatively similar.[181] In order for algorithms to be generalizable—to be accurate beyond the training data—training data should be a representative sample.[182] If an algorithm is trained with data that overrepresents some types of law or some types of people, it may not produce optimal results when applied to others.[183] The notion that a conclusion from one group may not transfer to another is not new,[184] but the more that we rely on data-driven algorithms, the more that transferability problems matter.

Take, for example, the workers' compensation injury database described above, which contains data on injuries to employees.[185] The database is not representative of all workers, because it does not contain data on independent contractors and other non-employee workers. If workers' compensation data is used to train an algorithm intended to identify risk and reduce injuries, it may work quite well for employees, because the dataset is representative of employees and the injuries they suffer. But it may not work well for independent contractors. Most obviously, the algorithm might not have any

180. An example discussed above is the workers' compensation. *See supra* Section II.B.1.

181. *E.g.*, W. Nicholson Price II, *Distributed Governance of Medical AI*, 25 SMU SCI. & TECH. L. REV. 3, 12 (2022) (discussing transferability of AI systems in the medical context).

182. David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 713 (2017); Selbst, *supra* note 79, at 134 ("Training data must also be a representative sample of the whole population. The ultimate goal of data mining is pattern-matching and generalization, and without a representative sample, generalizing introduces sampling bias." (footnote omitted)).

183. Dominik Stammbach, Boya Zhang & Elliott Ash, *The Choice of Textual Knowledge Base in Automated Claim Checking*, 15 J. DATA & INFO. QUALITY 3:1, 3:1 (2023).

184. *E.g.*, W. Nicholson Price II, *Medical AI and Contextual Bias*, 33 HARV. J.L. & TECH. 65, 91 (2019) (discussing the problem in the context of clinical trials).

185. *See supra* Section II.B.1.

input data related to independent contractor injuries. But even if data on independent contractor injuries is available—for instance, Uber gathers data on its drivers[186]—software that is designed using training data from workers' compensation may misinterpret and produce inaccurate results if independent contractor data is inputted.

A similar lack of transferability may occur with litigation data. Federal court data is more accessible than state court data, so algorithms that use court documents as input data might reasonably choose to use data from federal courts rather than state courts.[187] Because federal courts hear some state law cases,[188] the creators of such an algorithm would be able to use state law cases as input data. But would an algorithm so trained provide accurate predictions when applied to state law questions in general, rather than the subset that appear in federal court?

Perhaps not. Federal courts can hear state law questions under diversity-of-citizenship jurisdiction. This limits state law questions to those where the parties are citizens of different states and where the amount-in-controversy is greater than $75,000.[189] State lawsuits that are generally not heard in federal court may be systematically different than those in the training data—for example, they may be more likely to involve local controversies or disputes between neighbors (thus there would be no diversity of citizenship) and state law cases with small amounts in controversy would certainly be underrepresented in federal court data.[190] Thus an algorithm trained on state law cases that appear in federal court might not transfer to state law cases that are heard in state courts. This bias could affect not only attempts to predict outcomes of

186. Yupeng Fu & Chinmay Soman, *Real-Time Data Infrastructure at Uber*, PROC. 2021 INT'L CONF. ON MGMT. DATA 2503, 2503.

187. In fact, federal court data availability, particularly for bulk analysis, is shockingly bad. For a catalogue of problems with court data access, see generally Charlotte S. Alexander & Mohammad Javad Feizollahi, *On Dragons, Caves, Teeth, and Claws: Legal Analytics and the Problem of Court Data Access*, *in* COMPUTATIONAL LEGAL STUD. (Ryan Whalen ed., 2020).

188. Under diversity jurisdiction, for example. David L. Shapiro, *Federal Diversity Jurisdiction: A Survey and a Proposal*, 91 HARV. L. REV. 317, 317 (1977).

189. There are other doctrines that allow federal courts to hear state law cases, for instance, supplemental jurisdiction, but diversity jurisdiction is a major source of state law cases in federal courts. 28 U.S.C. § 1367.

190. Scholars of law and data are increasingly studying biases in legal datasets. *E.g.*, Keith Carlson, Michael A. Livermore & Daniel N. Rockmore, *The Problem of Data Bias in the Pool of Published U.S. Appellate Court Opinions*, 17 J. EMPIRICAL LEGAL STUD. 224, 224–25 (2020); Nina Varsava, *Opinion Authorship and Precedential Status*, 101 WASH U. L. REV. 1593, 1596 (2024).

cases but also many other uses of legal texts, for instance, sentiment analysis[191] or semantic analysis to train software to tag concepts or categories in legal texts.[192]

2. Policy

Policy setting and legal decision-making are increasingly data-driven.[193] But data-driven decisions can happen only where data is available. Take the example of short-term rental platforms. It is easy to stumble across anecdotal complaints about these platforms,[194] but policymakers have struggled with whether and how best to regulate short-term rentals.[195] One challenge is that short-term rentals are invisible transactions. Unlike sales—and unlike hotels, perhaps a closer analogy[196]—short-term rental agreements need not be disclosed. Of course, platforms are, by their nature, public listings of information, but the public data from platforms is not complete. Airbnb, for example, does not provide the precise geographic location of their rentals nor is it easy to tell how many listings are actually occupied.[197] Further, while the platforms themselves share a select amount of information, they are presumably motivated

191. *E.g.*, Yi-Hung Liu & Yen-Liang Chen, *A Two-Phase Sentiment Analysis Approach for Judgement Prediction*, 44 J. INFO. SCI. 594, 594 (2017); Douglas R. Rice & Christopher Zorn, *Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies*, 9 POL. SCI. RSCH. & METHODS 20, 21 (2021).

192. *E.g.*, Kaiz Merchant & Yash Pande, *NLP Based Latent Semantic Analysis for Legal Text Summarization*, 2018 INT'L CONF. ON ADVANCES COMPUTING, COMMC'N & INFORMATICS 1803, 1803; Fusheng Wei, Han Qin, Shi Ye & Haozhen Zhao, *Empirical Study of Deep Learning for Text Classification in Legal Document Review*, 2018 IEEE INT'L CONF. ON BIG DATA 3317, 3317.

193. *E.g.*, Huntington, *supra* note 72.

194. *E.g.*, Andrew Williams, *As Housing Crunch Intensifies Across the Country, Data Gives a Peek at Airbnb Impact*, 10 BOS. (May 31, 2023, 12:01 PM), https://www.nbcboston.com/news/national-international/as-housing-crunch-intensifies-across-the-country-data-gives-a-peek-at-airbnb-impact/2821373 [https://perma.cc/67ZH-LUE2] ("With U.S. cities facing clear housing shortages, homes being used as short-term rentals on Airbnb and other platforms are drawing renewed scrutiny . . . ."); Tracy Jan, *Faced with Complaints of Discrimination, Airbnb Partners with NAACP to Recruit Black Hosts*, WASH. POST (July 26, 2017, 11:34 AM), https://www.washingtonpost.com/news/wonk/wp/2017/07/26/faced-with-complaints-of-discrimination-airbnb-partners-with-naacp-to-recruit-black-hosts (on file with the *Iowa Law Review*) (reporting that short-term rentals are "beleaguered by discrimination complaints").

195. Nicole Gelinas, *Airbnb Is a Problem for Cities Like New York and San Francisco*, N.Y. TIMES (June 16, 2015, 6:51 AM), https://www.nytimes.com/roomfordebate/2015/06/16/san-francisco-and-new-york-weigh-airbnbs-effect-on-rent (on file with the *Iowa Law Review*).

196. Many state and local governments maintain registries of lodging operators which serve as a public database for information about hotels. *E.g.*, *Public Registry of Lodging Operators*, MASS.GOV, https://www.mass.gov/info-details/public-registry-of-lodging-operators#search-the-registry [https://perma.cc/E9MY-U74V].

197. Jennifer Combs, Danielle Kerrigan & David Wachsmuth, *Short-Term Rentals in Canada*, CANADIAN J. URB. RSCH. 123 (2020), https://web-p-ebscohost-com.proxy.lib.uiowa.edu/ehost/detail/detail?vid=0&sid=f15347ac-546c-40be-a570-334b11cd7626%40redis&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSx1aWQsdXJs#db=afh&AN=146921920 (on file with the *Iowa Law Review*).

to only share information that paints them in a positive light.[198] Because short-term rentals are invisible, it is hard for local governments to make data-based policy about short-term rental platforms.

Data also drives legal decision-making on an individual level such as in parole[199] and sentencing decisions.[200] These can, naturally, only incorporate existing data.[201] Data availability therefore affects how decisions are made in these contexts. For instance, the Department of Human Services in Allegheny County, PA, requested proposals to "better use data already available to us to improve decision-making through predictive-risk modeling."[202] Researchers used the county's existing data to create the Allegheny Family Screening Tool, one of the most-used predictive analytics tools in family law,[203] which scores a child's risk for certain future events such as abuse or an out-of-home placement.[204] If the algorithm score is above a certain threshold, the state must investigate the allegations.[205] The tool is praised as useful, but is also much criticized.[206] Data used in making the predictions include variables such as whether a family is receiving mental health treatment, accessing supplemental nutrition assistance program benefits or welfare benefits.[207] This information was readily available from the county and does have some correlation with a child's risk of abuse but "many of the variables in the algorithm that are used

198. Yang Wang, Mark Livingston, David P. McArthur & Nick Bailey, *The Challenges of Measuring the Shorter-Term Rental Market*, HOUSING STUD. 2261 (2023), https://www.tandfonline.com/doi/full/10.1080/02673037.2023.2176829 [https://perma.cc/66XA-QR6F] ("The challenge with any analysis is the lack of official data to assess [short-term rental] activity. . . . Airbnb does not make the data available that would allow a proper evaluation.").

199. Rashida Richardson, Jason M. Schultz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 15, 23 (2019).

200. Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 804 (2014); Ngozi Okidegbe, *Beyond More Accurate Algorithms: Takeaways from* McCleskey *Revisited*, 121 MICH. L. REV. 1109, 1111 (2023).

201. There are significant problems with existing data; some scholars investigate better data sources, such as community data. Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. 2007, 2008 (2022).

202. Emnet Almedom, Nandita Sampath & Joanne Ma, *Algorithms and Child Welfare: The Disparate Impact of Family Surveillance in Risk Assessment Technologies*, BERKELEY PUB. POL'Y J., Fall 2020, at 14, 21 (2020).

203. For an overview of how predictive analytics are used in family law, see Huntington, *supra* note 72, at 267.

204. *Allegheny Family Screening Tool*, ALLEGHENY CNTY., https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx [https://perma.cc/2EPZ-3EST].

205. *Id.*

206. Virginia Eubanks, *A Child Abuse Prediction Model Fails Poor Families*, WIRED (Jan. 15, 2018, 8:00 AM), https://www.wired.com/story/excerpt-from-automating-inequality [https://perma.cc/K9WZ-TY9G]; Dan Hurley, *Can an Algorithm Tell When Kids Are in Danger?*, N.Y. TIMES (Jan. 2, 2018), https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html [https://perma.cc/TB9K-GMAW].

207. Almedom et al., *supra* note 202, at 22.

to predict neglect and abuse are about whether a family has taken advantage of public services . . . . This means that poorer families are penalized more harshly."[208] Here, existing data created for other purposes shaped child welfare policies and decision-making.

3. Data Use

The structure of legal systems and institutions affects who owns and can use data. Public data is widely available and useable by many. When a legal doctrine does not release information publicly or in easily aggregable format, it restricts who can use that information. This has privacy benefits and may satisfy certain instincts about who should own data.[209] But nonpublic information is also an implicit choice to limit the entities who can use legal data, access certain aspects of the legal system, and create applications based upon legal data. Though this is true outside the big data context, it is expanded with the advent of big data analytics.

As an example, take the rules governing litigation procedure. Some documents produced by the parties during litigation are filed with the court and become part of the public record.[210] Other documents are served on opposing counsel, meaning that they are disclosed to the other parties in the case, but not to the judge and are not part of the public record.[211] This means that law firms frequently involved in litigation likely have large databases of documents that are not filed with the court, both their own and those received from opposing counsel. These firms can use these documents as resources to improve their own drafting or to look for patterns about what strategies work best in different situations. To be sure, a low-tech version of this practice has always occurred, with experienced lawyers using documents drafted for prior cases as models for later cases and their own extensive experience to predict which strategies will be successful. But data analytics may provide a bigger advantage to the possessors of large databases of documents.

4. Questions and Answers

Researchers prefer to study questions that can be answered, so they gravitate toward areas where data is available.[212] If legal doctrines hide information

208. *Id.*; *accord* Stephanie K. Glaberson, *Coding Over the Cracks: Predictive Analytics and Child Protection*, 46 FORDHAM URB. L.J. 307, 338 (2019).

209. *E.g.*, Jorge L. Contreras, *The False Promise of Health Data Ownership*, 94 N.Y.U. L. REV. 624, 626–31 (2019) (discussing data ownership in the context of health information).

210. Complaints, for example. *E.g.*, N.J. LOC. CIV. R. 5.1 (specifying that complaints are filed with the court).

211. Expert reports, for example. *E.g.*, N.J. LOC. CIV. R. 26.1(c)(1) (specifying that initial and expert disclosure materials must be served on other parties but "shall not be filed until used in a proceeding or upon order of the Court").

212. For an example in the medical context, see Daniel M. Doolan, Jennifer Winters & Sahar Nouredini, *Answering Research Questions Using an Existing Data Set*, MED. RSCH. ARCHIVES, Sept. 2017, at 1, 2 (2017) ("Increasingly, research studies are being conducted using existing data sets.").

about certain groups or topics, those groups or topics may be understudied or misunderstood. For instance, as explored above, although evictions are an enormously important policy issue, they are very difficult to study because there are no comprehensive legal records. Another type of legal record, death certificates, undercount Native Americans, who are concerned that they have become "invisible tribes."[213] Data-driven limitations on what questions can be asked are not restricted to researchers. Home buyers looking to make an offer can use public records to gather information on comparable past sales; renters must rely on incomplete data from private platforms for similar comparisons.[214] Companies seeking to launch a new product often want to understand the patent landscape and their risk for patent litigation—this is in part a function of whether competitors own an interest in a particular patent, which is public information as to sales[215] but not as to licenses.[216]

Data limitations may also mean that researchers get answers wrong or arrive at misleading conclusions. For example, law review articles often measure the significance or prevalence of a term, case, or statute based on the frequency of a proxy for its prevalence or importance, but this strategy will get the answer wrong if the question involves cases decided by juries, which do not produce written opinions.

Legal rules therefore affect what questions we can ask and how accurate our answers will be. To be sure, legal doctrines cannot be designed to provide data on all topics or enable answers to all questions—and such an effort is neither possible nor desirable. However, given that the set of questions we can ask and answer is presently an unintentional effect of historical legal rules, it may be preferable to deliberately consider which questions are particularly important and how legal rules affect data gathering for those inquiries.

Although, as emphasized above, the notion that data availability affects the questions asked and answers obtained is not a new one, it has become increasingly salient. As our technological capability for big data analysis improves, more questions can be asked using data, meaning that the implicit choices law makes about what to include in that data are more impactful. As legal

---

213. URB. INDIAN HEALTH COMM'N, INVISIBLE TRIBES: URBAN INDIANS AND THEIR HEALTH IN A CHANGING WORLD 5 (2007).

214. Robert P. Berrens & Michael McKee, *What Price Nondisclosure? The Effects of Nondisclosure of Real Estate Sales Prices*, 85 SOC. SCI. Q. 509, 509 (2004).

215. Albeit imperfectly. Alan C. Marco, Amanda Myers, Stuart J.H. Graham, Paul D'Agostino & Kirsten Apple, *The USPTO Patent Assignment Dataset: Descriptions and Analysis* 6 (U.S. Pat. & Trademark Off., Working Paper No. 2015-2, 2015).

216. It can also be relevant for private actors who want to understand the patent landscape or their litigation risk. FED. TRADE COMM'N, THE EVOLVING IP MARKETPLACE: ALIGNING PATENT NOTICE AND REMEDIES WITH COMPETITION 130 n.333 (2011), https://www.ftc.gov/sites/default/files/documents/reports/evolving-ip-marketplace-aligning-patent-notice-and-remedies-competition-report-federal-trade/110307patentreport.pdf [https://perma.cc/33VP-9EPY] ("[L]itigation risk is 'a function . . . [of] underlying business considerations' that depend on knowing: 'Who's holding the patent?'").

information becomes accessible to wider audiences, more people can ask questions using legal information, again, rendering choices about the contents of that information more important. And as the way that society makes decisions both inside and outside the legal system becomes more data-driven and data-dependent, the nature of the data available becomes more significant.

B. PRIVACY AND INVISIBLE LAW

Some groups have more privacy than others.[217] Not all legal information has personally identifying details, but some does—for example, state and local governments may have publicly accessible databases with personal information including birth dates, marriage details, party affiliation, and property ownership.[218] Public court records may include details on immigration status, arrests, employment, medical conditions, and more.[219] Privacy scholars have long complained that policies about publicizing legal information are outdated because they evolved before legal records were digitized. What previously languished in practical obscurity[220] because it was time-intensive to access physical records is now easily available on electronic databases.[221]

Without comprehensive thinking about the biases created by publicizing legal data, some groups are at heightened risk for privacy-based harms from their interactions with the government and legal systems. As described in the example above about child welfare models,[222] individuals and families who interact with the government in ways unrelated to child welfare—such as receiving supplemental nutrition assistance program benefits—can then have that information used against them in the child welfare setting. By contrast, individuals who do not need governmental assistance to buy food do not have this information trail. Scholars have documented how low-income communities are subject to higher rates of both government and private monitoring and data collection.[223]

---

217. KHIARA M. BRIDGES, POVERTY OF PRIVACY RIGHTS 6 (2017) (describing how poverty leads to private invasions by government actors).

218. Solove, *supra* note 29, at 1139.

219. *Id.*

220. In the context of surveillance, the Supreme Court noted that "[i]n the precomputer age, the greatest protections of privacy were neither constitutional nor statutory, but practical. Traditional surveillance for any extended period of time was difficult and costly and therefore rarely undertaken." United States v. Jones, 565 U.S. 400, 429 (2012).

221. Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity*, 101 CALIF. L. REV. 1, 20–23 (2013); Aniket Kesari, *The Privacy-Fairness-Accuracy Frontier: A Computational Law & Economics Toolkit for Making Algorithmic Tradeoffs*, PROC. 2022 SYMP. ON COMPUT. SCI. & L. 77, 77–78; Nissenbaum, *supra* note 13, at 126; DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 20 (2004).

222. *Supra* notes 202–08 and accompanying text.

223. *E.g.*, BRIDGES, *supra* note 217, at 20; Michele Estrin Gilman, *The Class Differential in Privacy Law*, 77 BROOK. L. REV. 1389, 1389–90 (2012); Torin Monahan, *Questioning Surveillance and Security*, *in* SURVEILLANCE AND SECURITY: TECHNOLOGICAL POLITICS AND POWER IN EVERYDAY LIFE 1, 15–16 (Torin Monahan ed., 2006); Madden et al., *supra* note 20, at 55.

Bias also flows in the other direction: Too much privacy—being invisible to data aggregators—can also harm.[224] For example, as discussed above, policymaking can be driven by information. If some information is not publicly available, its subjects will have additional privacy, but their policy interests may also be ignored. When legal transactions and interactions are invisible, it is also more difficult to build data-based applications that will use the data for beneficial ends. An AI application that overlooks or does not transfer well to a particular group might mean that that group does not get the benefit of the AI application. Property owners, whose purchases are matters of public record, might understandably wish that marketing software did not know the price of their homes.[225] But independent contractors might prefer that their injuries could be used as inputs for software that assesses how best to improve safety at work.

Thus, while privacy is an important consideration in determining how legal doctrines should incentivize or suppress information flows, it is not the only consideration. To be sure, concerns about privacy are not necessarily irreconcilable with concerns about being invisible to data aggregators. In some circumstances, this needle can be threaded by making anonymous data available. But in other circumstances, anonymized data may be impractical, unworkable, or at risk of hacking, and so concerns about privacy and invisibility in aggregate data may be at odds. This Article emphasizes the harms of invisibility in a data-based world. The harms of invasion of privacy must be weighed against the ills of one's data not being counted in applications that use aggregate information.

C. AUTOMATING LAW

The field of legal technology is one area where the availability of aggregated information is particularly important. For a computer to learn how to be a lawyer or a judge, it needs data: examples of lawyers' or judges' work. Where datasets are not available, automation will be more difficult or not possible. Thus, the availability of legal information drives development of legal technology. It impacts both whether a particular field of law or legal task can be automated, the speed at which automation can occur, and the accuracy with which a computer can perform an automated legal task.

Although legal information affects the development of technology more generally, the effect is likely strongest in the field of *legal* technology because there are few other sources of training data beyond those provided by legal doctrine. For instance, while a repository of patent documents can be used to train large language models that have general applications, other sources of information can and are also used (newspapers, social media posts, Wikipedia,

224. Waldman, *supra* note 31, at 2252–53.

225. Or alternatively, they may appreciate being able to track the changing value of their homes on various platforms.

journal articles, and many others).[226] But there are no good alternative sources to train a model that can adjudicate patentability decisions or decide if trademark applications should be granted.[227]

Where there are biases in data produced by legal doctrines, legal technology trained on those datasets may have corresponding biases. For instance, computer scientists are using artificial intelligence to build automated judges.[228] AI judges will have many written opinions from the field of admiralty law, but fewer about negligence law.[229] They will provide well-reasoned opinions about medical malpractice and motor vehicle cases but miss the mark on mortgage foreclosure cases and title disputes.[230] They will know a great deal about New York law, but less about North Carolina law.[231] AI judges will be capable of evaluating arguments in briefs but not those in expert reports.[232] They will be more accurate in cases involving business litigants as compared to those involving individual litigants.

A similar problem arises in the development of automated contract drafting and analysis. If SEC data provides the source of training data for such applications, it will be skewed toward the specific types of contracts disclosed under SEC regulations and tailored toward the sorts of entities who are parties to those contracts—generally large businesses and top executives.[233] Such software may perform less well for contracts, transactions, and entities who are unrepresented or underrepresented in large contracts datasets.

And data availability will also affect *who* can develop legal technology and who can review the algorithm, which has critical implications for transparency

---

226. Schaul et al., *supra* note 7, at 1.

227. *E.g.*, Sonia K. Katyal & Aniket Kesari, *Trademark Search, Artificial Intelligence, and the Role of the Private Sector*, 35 BERKELEY TECH. L.J. 501, 553 (2020) (explaining that trademarks and decisions regarding the marks can be part of training data).

228. For a discussion of current developments and arguments regarding the desirability of automated judges, see, e.g., Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 615–17 (2020); Ian Kerr, *Prediction, Pre-Emption, Presumption*, *in* PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN 35 (2013); Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1137–43 (2019) (arguing in favor of AI judges); and Tim Wu, Essay, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001, 2001–04 (2019). A true automated judge in the United States does not exist as of the time this Article was written, but artificial intelligence is an increasing part of adjudication in both the judicial and administrative branches and other countries are experimenting with automated adjudication. Cary Coglianese & Lavi M. Ben Dor, *AI in Adjudication and Administration*, 86 BROOK. L. REV. 791, 795 (2020).

229. The former are always federal cases, the latter often in state court. LANGTON & COHEN, *supra* note 112, at 4.

230. The first two are often decided by judges, the last two by juries. *Id.* at 2.

231. New York has digitized many court documents, but North Carolina remains in the process of digitizing its court records. *North Carolina Courts, Cases, and Rules*, UNC LIBR. GUIDE (Oct. 6, 2023, 5:33 PM), https://guides.lib.unc.edu/c.php?g=914799&p=6590910 [https://perma.cc/WQ9S-PYQT].

232. Briefs are part of the public court record; expert reports are often not. FED. R. CIV. P. 26(a)(2)(B) (requiring disclosure of written expert testimony).

233. Nyarko, *supra* note 158, at 7, 18.

and democracy. For an area where legal transactions are invisible, like private contracts, private data aggregators may have an advantage in creating software and the public cannot readily review the input data. For areas where the government possesses a database of information, but it is not public or the public information is not readily amassed, like court records that can be accessed only by individual searches, the government can (and does) create AI applications[234] but the public cannot, nor can the public check, the government's work.[235]

Legal rules about information disclosure that were not devised with data aggregates or legal technology in mind have had and will have great impact on the process of legal automation. There is a substantial debate about the merits, ethics, direction, and pace of legal automation[236]—but whatever one's view on that debate, surely accidental development based on outdated informational rules is not optimal. Further, there are trenchant criticisms about both lack of transparency and over-transparency as to input and output data and the algorithms themselves.[237] Some of this is driven by data availability, because whether the underlying data is public dictates who can create software and whether the input and training data are visible. Again, the current system is partially a consequence of old rules about legal information not created with big data in mind.

IV. INTENTIONAL INFORMATION IN THE THEORY AND STRUCTURE OF LAW

Scholars and policymakers extensively discuss the relationship between law and information.[238] The existing literature and thinking on law and information does not reflect the complex modern information ecosystem, the scope of unintentional informational effects, nor the way that use of legal information has changed and is continuing to change. This Part recommends new

234. DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 21 (2020), https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf [https://perma.cc/YAH9-LE6H].

235. Lack of transparency in algorithmic decision-making in law is a major challenge. *E.g.*, Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1356 (2018).

236. *E.g.*, Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797, 799–802 (2021); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 7–8 (2014); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1152–54 (2017); Woodrow Hartzog, Gregory Conti, John Nelson & Lisa A. Shay, *Inefficiently Automated Law Enforcement*, 2015 MICH. ST. L. REV. 1763, 1764–68; Milan Markovic, *Rise of the Robot Lawyers?*, 61 ARIZ. L. REV. 325, 331–35 (2019); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 2–4 (2019); Ari Ezra Waldman, *Power, Process, and Automated Decision-Making*, 88 FORDHAM L. REV. 613, 613–16 (2019).

237. *E.g.*, Wexler, *supra* note 235, at 1356.

238. *See supra* Part I.

approaches to integrating informational considerations into legal theory, discourse, and policy.

Section A synthesizes how information use differs from its traditional descriptions in the legal literature. Section B explores how this conception of information suggests new roles for law and legal institutions.

A. *New Paradigms of Information Use*

Legal scholarship on law and information has historically been characterized by an emphasis on laws that explicitly and deliberately address information and a focus on communication of individual pieces of information to directly interested parties. Those features, while still important, no longer describe how law affects information or how legal information is used.

First, *every* choice about legal doctrine affects information, no matter how apparently unrelated it may appear. Legal doctrines that say nothing about information have made an implicit choice not to require or forbid public disclosure of information relating to the doctrine, thereby making the status quo the default state for that information. Such doctrines might also affect information availability more directly, even if the doctrine does not explicitly affect information. For instance, a case interpreting the definition of independent contractor or a statute classifying certain workers as employees are not, on their face, rules about information. But because more information is gathered about employees through the workers' compensation system as compared to independent contracts, these rules affect what information will be collected and publicly released. Each doctrine therefore has some power to influence the extent to which information is produced, how datasets are shaped, what policy questions can be investigated, and how data-driven analytics will develop.

This universality means that the impact of many legal doctrines on information can be quite unintentional. Particularly when a doctrine does not outright mention information or when the path between the specific legal transaction or entity governed by the doctrine and the ultimate release of information is lengthy and involves multiple steps, discussion of the underlying doctrine may not account for its informational effects. Further, because the way in which information is disseminated changes over time, a doctrine's effect on information is not static. Changes to interrelated doctrines, modification of institutional policies and—in recent years very significant—changes to how data is made available, accessed, and analyzed, all affect the relationship between law and information. A doctrine's impact on information may not be as originally intended when the doctrine was first created.

A further important characteristic of the relationship between law and information is that information is often used in the aggregate—with users collating legal data to make large-scale observations about the world. The ability to aggregate legal data at scale has greatly expanded in recent years,

enabled by modern technology.[239] Further, the uses for aggregate legal data have also expanded and now encompass both learning about the legal transaction or issue as well as broader uses both within and beyond the legal system. Old rules about when and why information should be made public may operate under outdated assumptions about how information is used.

Some traditional justifications for how law governs information fit poorly with this model of information use. For instance, property law's rationale for not requiring public disclosure of leases relates to individual-level information and is not always sensible in the context of data aggregates. Leases were historically not recorded because a tenant's occupancy of land is often easily noticeable upon visual inspection.[240] Recordation systems are thus less necessary.[241] Further, if a buyer of property is unaware of a previously signed short-term lease on the property, the short duration of the lease reduces the scope of the problem.[242] "[M]uch of the material in a lease agreement has little importance in providing notice to third persons."[243]

These rationales clearly do not contemplate uses of information that require aggregating data across properties. Visual inspection of property is not practical when collecting a dataset of thousands or millions of properties. Similarly, material in a lease agreement or the presence of a short-term lease agreement may not be essential information to the buyer of an individual piece of property, but to a policymaker interested in understanding rental patterns or housing practices, such information is critical.

Uses for legal information have changed; audiences for legal information have also changed. Historically, legal information was relatively difficult to access. Someone interested in learning about a case, property record, or patent would have to visit the local courthouse, county record office, agency archive, or a specialized library. There, the reader could review only a small number of cases. Accordingly, the general public's view of legal information was less relevant because few members of the public would ever see most legal information.

---

239. Examples are discussed in Section I.B.

240. ANDREW R. BERMAN, FRIEDMAN ON LEASES § 31.1 (6th ed. 2017). In general, "possession requires a reasonably diligent inquiry into the nature of the right asserted by the one in possession and charges the purchaser with knowledge of whatever facts such an inquiry would reveal." Nat. Res., Inc. v. Wineberg, 349 F.2d 685, 690 (9th Cir. 1965).

241. There are certain circumstances where leases must be recorded, and these are driven by the need for notice. *E.g.*, Paul B. Zion, Note, *Mandatory Recording of Personal Property Leases in South Carolina: An Examination of the South Carolina Bailment Statute as Affected by U.C.C. Article Nine*, 30 S.C. L. REV. 557, 560 (1979) ("The underlying purpose [of South Carolina's lease recordation requirement] is to prevent the apparent ownership of the person in possession of personal property from misleading his bona fide creditors or purchasers.").

242. BERMAN, *supra* note 240, § 31.1 (noting that "[n]o authority has been found discussing whether such unrecorded [short-term] leases are valid against bona fide purchasers under the recording acts, perhaps because . . . the lease is short term and rarely would justify taking any dispute to litigation").

243. Brunson v. Centennial Am. Props., LLC, No. CV208-059, 2010 WL 11613662, at *4 (S.D. Ga. May 27, 2010).

Modern constituencies for legal information are different. Because technology has made legal information more accessible and easier to analyze, more people can see legal information. When lawyers write legal documents, their primary intended audience is often either the other parties to the case or transaction or an adjudicator. But there are many other constituencies for legal information—the casual reader who searches the internet for a person's name and comes across a court case, a computer scientist who downloads SEC filings, an organization that analyzes patent filings to study innovation patterns, and many more. New constituencies for legal information may add to the unintended effects of that information and thereby necessitate new thinking about how legal information is communicated. For instance, legal documents that use specialized terminology and writing conventions may be clear to lawyers and judges, but impenetrable and misunderstood by broader audiences.[244] Specific formats or pieces of information may also be particularly useful to new audiences for legal information.[245] Both the harm and benefits of legal information change as the audiences for that information expand—a change that legal doctrines presently often do not take into account.[246]

Further, there has been a growth in *negative* constituencies for legal information. These are audiences who deliberately use legal information for deleterious purposes. Stalkers who track their victims through legal filings.[247] Extortionists who gather information from legal sources and threaten to publicize it unless paid (for instance, websites posting mugshots and offering to remove them for a fee).[248] Conspiracy theorists who search for apparently-relevant legal documents to bolster their claims (for example, those who point to patents on "chemtrails" as evidence that the government has validated the theory's existence[249]). Purveyors of misleading claims who point to the claims'

244. One example comes from patents, which often contain experimental data supporting the patented invention. The Patent and Trademark Office permits these experiments to be fictional, meaning that the numerical results presented in patents are often fabricated. This is well-known by some audiences for information in patents—namely the patent examiner and other lawyers—but is confusing and misleading for casual readers without a deep understanding of patent law. Indeed, prior work has shown that ninety-nine percent of scientists who cite fictional data from patents do so as if it were factual. Janet Freilich, *Prophetic Patents*, 53 U.C. DAVIS L. REV. 663, 699 (2019). Lawyers who draft patents acknowledge that patents can be hard for lay readers to understand but maintain that their primary responsibility is not about information communication, but rather to obtain the strongest legal right possible for their client. Janet Freilich & Lisa Larrimore Ouellette, *Science Fiction: Fictitious Experiments in Patents*, 364 SCIENCE 1036, 1037 (2019).

245. Standardization, for example, is particularly useful for data aggregators, but less important for readers of individual documents.

246. Freilich, *supra* note 57, at 1585.

247. *See* Solove, *supra* note 29, at 1173.

248. Michael Polatsek, *Extortion Through the Public Record: Has the Internet Made Florida's Sunshine Laws Too Bright?*, 66 FLA. L. REV. 913, 917 (2014).

249. Jason Daley, *Science Officially Debunks Chemtrails, but the Conspiracy Will Likely Live On*, SMITHSONIAN MAG. (Aug. 22, 2016), https://www.smithsonianmag.com/smart-news/science-officially-debunks-chemtrails-conspiracy-live-180960139 [https://perma.cc/AP3K-G2XG].

presence in legal information as a way to legitimize and sanitize their assertions (vaccine skeptics who note that government agencies collect information about vaccine side effects[250]). As the audience for legal information evolves, the ways in which legal information can be used to harm also grows—but legal doctrine is not keeping up.

B. NEW ROLES FOR LAW AND LEGAL INSTITUTIONS

The nature of legal information, its uses, and its audiences are shifting. This suggests a need for rethinking the role of law and legal institutions. To begin, consider law's information-sorting function. One classic way in which legal processes sort information is by hearing two sides of a dispute and deciding which set of facts is correct (or, more precisely, meets a particular legal standard).[251] With the growth of data analytics and expanded audience for legal information, legal processes and institutions now have additional information sorting functions. Legal doctrines make choices about which pieces of information to render more visible, to make more authoritative, and to standardize. In this way, law sorts information by elevating certain pieces of information—and, by extension, rendering other information invisible. This new role is not entirely a function of the information era—legal processes have always highlighted certain pieces of information. For instance, prominent cases have long received more media coverage, and facts with significant legal implications drawn more attention because of the law's role.[252] However, as uses of legal information shift, law's information-sorting role also shifts from a primary emphasis on deciding which information is *correct* to an increasing emphasis on deciding which information is *visible*.

Legal institutions also play an important role in the new paradigm of information described in this Article. As a general matter, it is hard to systematically gather or publicize information without the intervention of some form of central institution. For example, areas of private law often rely on a regulatory or institutional overlay to effectuate their informational outcomes. Property transactions are recorded in public registers run by local governments.[253] Information on torts is publicized by court records.[254] Tort alternatives like workers' compensation report injuries to state governments

250. Freilich, *supra* note 57, at 1580.

251. Because of burdens of proof and other procedural complexities, the adjudication process is better thought of as determining whether facts are sufficient to meet a particular legal standard, not whether they are true.

252. *E.g.*, David Margolick, *Not Guilty: The Overview; Jury Clears Simpson in Double Murder; Spellbound Nation Divides on Verdict*, N.Y. TIMES (Oct. 4, 1995), https://www.nytimes.com/1995/10/04/us/not-guilty-overview-jury-clears-simpson-double-murder-spellbound-nation-divides.html (on file with the *Iowa Law Review*).

253. Brady, *supra* note 14, at 875–76, 888–89.

254. *E.g.*, Michael J. Saks, *Do We Really Know Anything About the Behavior of the Tort Litigation System—And Why Not?*, 140 U. PA. L. REV. 1147, 1183–86 (1992).

which collect the information and host it online.[255] And most government agencies gather and publicize substantial amounts of information through their own channels—patents from the Patent Office,[256] environmental data from the EPA,[257] drug use and safety information from the FDA,[258] and many others.

The relationship between legal information and institutions has several consequences. First, it provides an opportunity for central governance of data and suggests avenues for policy interventions. As discussed in Section IV.B, below, institutions are well positioned to survey information use and enact policy changes. Second, it indicates an important role for private institutions. Private institutions aggregate, disseminate, or suppress legal information, and this shapes the influence of legal information.[259] For instance, many readers access court data from private data collection systems like Google's case law database or Lexis and Westlaw. These systems do not index all court records from all systems, and their decisions about what to make available dictates which cases are read and aggregated.[260]

Private institutions are not only channels for legal information; some private institutions work affirmatively to reduce the biases created by existing legal information. For example, the U.S. Government's system to allow public access to court records, PACER, has historically charged ten cents per page for access to case information,[261] making the aggregation of thousands or millions of records cost prohibitive for many.[262] Additionally, the format of PACER does not facilitate large-scale analysis of cases.[263] In response, several organizations have sought to make court records available for free bulk download in order to expand access to the records and increase the public's ability to

255. *See supra* Section II.B.

256. *Research Datasets*, *supra* note 91.

257. *TRI Toxics Tracker*, U.S. ENV'T PROT. AGENCY, https://edap.epa.gov/public/extensions/TRIToxicsTracker/TRIToxicsTracker.html [https://perma.cc/JB3E-5SQJ].

258. *Drug Safety and Availability*, U.S. FOOD & DRUG ADMIN. (Sept. 27, 2024), https://www.fda.gov/drugs/drug-safety-and-availability [https://perma.cc/4B6R-Z5SN].

259. *See* Olivier Sylvain, *Intermediary Design Duties*, 50 CONN. L. REV. 203, 218 (2018) (making a related point in the context of platforms that disseminate private information, like social media companies).

260. This is not an issue unique to data aggregation—these biases apply to any use of cases. For a particularly stark example, see Neil C. Thompson et al., *Trial by Internet: A Randomized Field Experiment on Wikipedia's Influence on Judges' Legal Reasoning*, *in* CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURISPRUDENCE (Kevin Tobia ed., forthcoming May 2025) (on file with the *Iowa Law Review)* (testing Wikipedia's influence on judges' use of cases).

261. The Judicial Conference recently approved a plan to remove PACER fees. ADMIN. OFF. OF THE U.S. CTS., REPORT OF THE PROCEEDINGS OF THE JUDICIAL CONFERENCE OF THE UNITED STATES 12 (Mar. 15, 2022), https://fingfx.thomsonreuters.com/gfx/legaldocs/egvbkwemjpq/jcus_mar_22_proceedings_o-1.pdf [https://perma.cc/L2G6-7TAX].

262. Dru Stevenson & Nicholas J. Wagoner, *Bargaining in the Shadow of Big Data*, 67 FLA. L. REV. 1337, 1360 (2015).

263. *Id.* at 1359.

conduct data analytics using court records.[264] Although these are not perfect solutions, any efforts to account for information in designing legal rules must understand how that information is disseminated through private institutions.

Further, legal doctrines sometimes make choices about whether to channel information into public or private institutions. This choice has significant consequences for public control over and access to information. And the choice is often unconscious, in the sense that policymakers are focused on aspects of doctrine other than the informational and consequently do not consider when information is privatized. Workers' compensation systems are an example—when instituted, they move information about employee injuries from public court dockets to private insurance companies.[265] This increases the quantity of information available (to insurance companies) about employee injuries because all injuries are reported whereas without a workers' compensation system only some injuries would become court cases. However, it requires a regulatory overlay—an agency to create a public database of claims—in order to bring that information back into public view.

Finally, understanding the changing relationship between law and information affects how legal scholars conceptualize the relationship between public and private law. Although public and private law are often described as distinct spheres, this distinction blurs in the context of information use.[266] At first blush, it may be unintuitive that laws concerning pieces of information about individuals might legitimately be concerned with aggregation of that information across a population. Much of the information discussed above is produced through private law mechanisms—property, torts, contracts—which are traditionally focused on interactions between private parties and may be at odds with the public purpose of information aggregation.[267]

But because information generated through private law channels has aggregate effects beyond the individuals involved in the underlying dispute, private law may need to be more concerned with population-level informational effects. Certainly, there are many existing examples of settings where information generated by private law is shared publicly and can be aggregated. Property records and court records, for example, share information publicly because policymakers decided that the need for public scrutiny of the information

264. *Id.* at 1360–62.

265. Some states also have public insurance options for workers' compensation.

266. The structure of private law has always been interested in both individual interactions and broader public policy goals, although how to prioritize the two "is a perennial concern of legal theory." Hanoch Dagan, *The Limited Autonomy of Private Law*, 56 AM. J. COMPAR. L. 809, 809 (2008).

267. Aditi Bagchi, *Private Law and Public Discourse*, 65 ARIZ. L. REV. 541, 557 (2023) ("[P]rivate law quite literally holds itself out as private, concerning itself with bilateral relationships rather than society at large. Traditional accounts of private law are reluctant to link the principles of justice at stake in private law with the principles of justice that govern major political choices.").

outweighed any individual desire to keep the information private.[268] Although the public policy purposes underlying these releases of information are not about aggregate collection of data (but rather about the importance of publicizing individual pieces of information), the principle that private information might be released for public purposes also applies to data aggregation.

Further, there are existing instances where information is collected for shared individual and aggregate purposes. One example is death certificates. When you die, your death—and some attendant information on your demographics and cause of death—will be recorded and collected by local, state, and federal government agencies.[269] Death certificates are generally public records.[270] Death certificates have a clear private function—they are used to settle estates.[271] But from their earliest inception, governments have intended death certificates to provide aggregable public information for public health and epidemiological purposes.[272] In short, death records are designed and used both for private purposes and for data aggregation and public informational purposes.[273] Individual and aggregate uses for information can co-exist and need not be in opposition.

V. INFORMATION POLICY

Accounting for information in legal theory and understanding the complex and multifaceted effects of doctrinal change on legal information can lead

268. Although, there are instances in both contexts where information is kept secret. *E.g.*, Does I Thru XXIII v. Advanced Textile Corp., 214 F.3d 1058, 1063 (9th Cir. 2000) (permitting plaintiffs to conceal their identities because they had "an objectively reasonable fear of extraordinarily severe retaliation"); Dale A. Whitman, *Secrecy and Real Property*, 27 AM. U. L. REV. 251, 252 (1978) (discussing policy considerations in allowing property owners to keep their identities secret).

269. Jeffrey R. Boles, *Documenting Death: Public Access to Government Death Records and Attendant Privacy Concerns*, 22 CORNELL J.L. & PUB. POL'Y 237, 254 (2012).

270. *Id.* at 260. Some states limit access to death records. *E.g.*, N.H. REV. STAT. ANN. § 5-C:9 (LexisNexis 2021) (restricting release of death certificates to those with a "direct and tangible interest").

271. Erin G. Brooks & Kurt D. Reed, *Principles and Pitfalls: A Guide to Death Certification*, 13 CLINICAL MED. & RSCH. 74, 74 (2015).

272. Kathryn Schulz, *Final Forms*, NEW YORKER (Mar. 31, 2014), https://www.newyorker.com/magazine/2014/04/07/final-forms (on file with the *Iowa Law Review*) (tracking the origin of death certificates to Bills of Mortality in 14th Century England, which tabulated plague deaths, and whose "intended purpose seems to have been to help the healthy steer clear of the most infectious parts of town"). The New York Times has called death certificates "the most widely used statistical tools to monitor serious diseases" and argued that "no document has as much impact on the health of a population as does the death certificate." Lawrence K. Altman, *The Doctor's World: New Certificate May Ease Criticism of Death Data*, N.Y. TIMES (Oct. 18, 1988), https://www.nytimes.com/1988/10/18/science/the-doctor-s-world-new-certificate-may-ease-criticism-of-death-data.html?searchResultPosition=1 (on file with the *Iowa Law Review*).

273. Although there are challenges in aggregating information from death records, particularly that the information therein can be "notoriously inaccurate." Frances Stead Sellers, *Rise in Infant Deaths Hits Black Families Hardest, Study Finds*, WASH. POST (March 13, 2023, 12:05 AM), https://www.washingtonpost.com/health/2023/03/13/sids-sudden-infant-death (on file with the *Iowa Law Review*).

to concrete policy. Because every legal doctrine affects information use, a comprehensive list of policies is beyond the scope of this Article. Rather, this Part attempts to provide illustrative examples of how better understanding informational effects can improve policy. This Part begins with a general discussion of how to think more broadly about law's unintentional informational effects (Section V.A), then turns to specific suggestions for rethinking when public disclosure of information is appropriate (Section V.B), solutions that are alternatives to changing levels of disclosure (Section V.C), and recommendations for information users (Section V.D).

A. INTEGRATING INFORMATION INTO LEGAL DOCTRINE

Having shown that the relationship between law and information has shifted, this Section suggests how legal scholars and policymakers might incorporate this expanded notion of law and information into legal theory and the design of legal systems and doctrine. This Section discusses overarching, general approaches to thinking about law and information; the following Section turns to specific policy proposals.

First, policymakers should deliberately consider the impact of law on information. The field of privacy law has been deeply influenced by the notion of "privacy by design," that privacy must be a proactive, essential consideration when designing technology and throughout its life cycle.[274] So too should information access and aggregation be a core consideration for every legal doctrine and legal institution—disclosure by design, as it were.[275] This does not mean that questions of information use should weigh more heavily than other central goals of lawmaking, or that release of more information will necessarily be desirable, but that the law's potential influence on information production and dissemination should be considered and weighed as either a positive or negative aspect of the new doctrine. This is very much in keeping with how legal rules are traditionally discussed—the difference is that this Article advocates for updated attention to new uses of and audiences for information to achieve more deliberate informational effects.

Importantly, the relationship between a legal doctrine and the information it produces changes over time. As technology develops and new audiences are able to access information, the impact of that information may shift. Thus, even after a law comes into effect, the information it produces and uses of that information should be periodically reviewed to ensure that they conform to the original goals. Laws from before the advent of large-scale data analysis

274. Ann Cavoukian, *Privacy by Design: The Definitive Workshop*, 3 IDENTITY INFO. SOC'Y. 247, 247–48 (2010). This principle has been influential in scholarship and policy recommendations. *E.g.*, FED. TRADE COMM'N., PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS, at vii (2012) (articulating the baseline principle of privacy by design that "[c]ompanies should promote consumer privacy throughout their organizations and at every stage of the development of their products and services").

275. Solove, *supra* note 29, at 1139.

may also need to be audited to determine where information goes and how it is used.

Further, policymakers should also focus on the blank spaces and silences—the areas where law does not produce information. These silences may be appropriate, protecting privacy, for instance, or the result of other important considerations, like juries' silence on their reasoning process in reaching a verdict in litigation.[276] Or there may be areas where more production of or access to information would be beneficial, like better public access to legislation and court records, in which case legal design may be able to facilitate those goals.[277]

Finally, when considering information effects, a broader, more integrated view of legal information and its uses would be helpful. It is relatively easy to understand linear relationships or siloed effects of law on information, but a systems thinking approach to legal information can reveal both problems with how information is used and potential solutions.[278]

For example, return to the workers' compensation scheme discussed above. If this program is observed in isolation, it appears to function well with regard to information. Workers' compensation laws incentivize workers to report injuries which are then aggregated in a public government database. This creates a large-scale, mostly complete, and relatively unbiased dataset for researchers, insurance companies, and data scientists to draw on in efforts to manage and reduce worker injuries—a result that corresponds nicely to the goal of the original workers' compensation laws: worker protection. The problems with information produced by the workers' compensation system are visible only if one looks beyond employees. Non-employee injuries are not tracked and, because research is conducted and funding is sometimes allocated based on workers' compensation data, the information from the scheme may skew policy against non-employee workers.

B. RETHINKING DISCLOSURE

Better understanding the relationship between legal doctrines and information production and dissemination may push in favor of changing doctrines to increase or decrease the availability of public information. Most

276. *E.g.*, United States v. Thomas, 116 F.3d 606, 608 (2d Cir. 1997) (discussing "the importance of safeguarding the secrecy of the jury deliberation room").

277. The Supreme Court recently held that annotations to Georgia's official code could not be copyrighted, relying in part on strong public interest arguments favoring widespread access to this information. Georgia v. Public.Resource.Org, Inc., No. 18-1150, slip op. at 17 (2020) ("Imagine a Georgia citizen interested in learning his legal rights and duties. If he reads the economy-class version of the Georgia Code available online, he will see laws requiring political candidates to pay hefty qualification fees (with no indigency exception), criminalizing broad categories of consensual sexual conduct, and exempting certain key evidence in criminal trials from standard evidentiary limitations—with no hint that important aspects of those laws have been held unconstitutional by the Georgia Supreme Court.").

278. VIRGINIA ANDERSON & LAUREN JOHNSON, SYSTEMS THINKING BASICS: FROM CONCEPTS TO CAUSAL LOOPS 17 (1997).

explicit choices about how law makes information available are balancing tests, weighing the costs of public information against the benefits.[279] As this Article has emphasized, the impact of legal information is broader than previously recognized in the literature and has changed (and is continuing to change) significantly as law takes an empirical turn and society increasingly relies on artificial intelligence. New uses and audiences for legal information can change both the costs and benefits of providing that information, potentially leading to different answers about whether information should be available.

To illustrate, consider patents. One important purpose of patents is to disclose information about cutting edge inventions so that others can build on the technology.[280] By statute, patents are publicly available eighteen months after they are filed.[281] The cost of the eighteen-month lag is that it delays the public's ability to use the knowledge in the patent; the benefit is that patent applicants are given a longer period of secrecy to continue developing their invention.[282] AI is increasingly using information in patents to aid in scientific discovery and suggest research and development strategy.[283] This means that the information in patents is more useful, increasing the benefit of publication. It may also mean that secrecy to inventors is more valuable, increasing the cost of publication, or perhaps that the faster pace of technological development renders secrecy less useful after a shorter period of time. Although weighing the costs and benefits is an empirical question, this Author's hunch is that the benefits of earlier publication outweigh the costs, meaning that new uses for information in patents push for earlier public disclosure of that information.

This example is only illustrative—technological changes may require rebalancing the costs and benefits of legal information in a host of fields. In property law, for example, courts took the position that "material in a lease agreement has little importance" for public notice.[284] But there may well be a public interest in aggregating information in rental agreements to better craft policy. Revisiting questions of when information is public is increasingly important as the uses of and audiences for legal information evolve.

279. FOIA, for example. *See, e.g.*, NARA v. Favis, 541 U.S. 157, 171 (2004) (requiring "balanc[ing] the family's privacy interests against the public interest in disclosure").

280. Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 480–81 (1974).

281. 35 U.S.C. § 122(b)(1)(A).

282. David Popp, Tedd Juhl & Daniel K.N. Johnson, *Time in Purgatory: Examining the Grant Lag for U.S. Patent Applications*, TOPICS ECON. ANALYSIS & POL'Y 5 (Jan. 2004), https://www.proquest.com/docview/21720942 64/fulltextPDF/17EF8DAF19144DBEPQ/1?accountid=14663&sourcetype=Scholarly%20Journals [https://perma.cc/CG2U-K9TQ].

283. Janet Freilich, *Patents' New Salience*, 109 VA. L. REV. 595, 623–24 (2023).

284. Brunson v. Centennial Am. Props., LLC, No. cv-208-509, 2010 WL 11613662, at *4 (S.D. Ga. May 27, 2010).

C. ALTERNATIVES TO DISCLOSURE

Sometimes, the best approach is to change laws to gather or publicize more (or less) information. But, although much of this Article highlights the harms of insufficient or overabundant public data, the most effective policy solution is not always to directly alter the legal doctrines producing the information. Legal doctrines have many priorities—justice, efficiency, administrability, political feasibility, and so on—and information production is only one (and not always the most important) of those considerations.

Even where recognition of the informational effects of law does not militate in favor of directly changing a doctrine, it can suggest more oblique approaches. Take, for instance, the problems created by insufficient information about independent contractor injuries, in contrast to abundant information about employee injuries from workers' compensation systems. One solution is to expand the workers' compensation system—and its data collection apparatus—to independent contractors. But while scholars have advocated for this move on other grounds,[285] it seems too extreme a solution for the narrower problem of data insufficiency. Instead, state and federal labor agencies could fill the data gap by conducting extensive surveys to gather data on injuries to non-employee workers to supplement the existing comprehensive data on employee injuries.[286]

Put more formally, there are several different intervention points for legal doctrines. Law can affect information creation, for instance by requiring disclosure. Law can affect information dissemination by creating platforms for or barriers to information sharing. And law can also affect incentives for third parties to use information—changing whether and how private parties choose to incorporate information into their own applications or highlight legal information in other venues.

Thus, policy changes need not target only the quantum of public information, but also its nature and availability. This is happening, for instance, with court records. Many states are currently engaged in efforts to digitize their court documents to make those documents more accessible,[287] and the federal court system continues its longstanding discussion on making data

285. *See, e.g.*, Michael Babcock & Michael Oldfather, *The Role of the Federal Employers' Liability Act in Railroad Safety*, 15 WORKERS' COMP. L. REV. 531, 531 (1999) (assessing work safety rules for railroad workers); John A. Pearce II & Jonathan P. Silva, *The Future of Independent Contractors and Their Status as Non-Employees: Moving on from a Common Law Standard*, 14 HASTINGS BUS. L.J. 1, 2–3 (2018) (discussing health and safety considerations for independent contractors); Tran & Sokas, *supra* note 144, at e63 (discussing lack of occupational health protections for gig workers).

286. The U.S. Bureau of Labor Statistics already does this to some extent. *E.g.*, Pegular & Gunter, *supra* note 154.

287. Schmitz, *supra* note 16, at 2389–92.

from federal courts more easily available.[288] Jurisdictions can also coordinate information policy in order to reduce the bias that arises from uneven accessibility of data. For instance, some state and local courts (and the federal court system) redact social security numbers in court records while others do not.[289] To avoid making disclosure of one's social security number inadvertently dependent on doctrines such as personal jurisdiction and venue, all court systems could redact social security numbers.[290]

Relatedly, standardization across jurisdictions, industries, or areas of law is an area where policymakers can intervene to reduce problems created by legal information.[291] Lack of data standardization across collecting agencies is a barrier to data aggregation. Federal agencies could create standard templates for data collection that may be adopted by local governments, as was done for the standard death certificate.[292] Institutions can also disseminate best practices for data collection to ensure standardization. For instance, death certificates were a main source of data on levels and severity of COVID-19 in different geographic areas,[293] but it was not always clear how best to categorize COVID-19 with respect to other co-morbidities on death certificates, nor was it clear how to read death certificates to distinguish between deaths caused by COVID-19 and those where the patient died of something else while infected.[294] In an attempt to resolve this confusion, the CDC released best practices guidelines to standardize reporting of COVID-19-related deaths.[295]

---

288. *See* JUD. CONF. OF THE U.S., REPORT OF THE PROCEEDINGS OF THE JUDICIAL CONFERENCE OF THE UNITED STATES 12 (2022), https://www.uscourts.gov/sites/default/files/jcus_mar_22_proceedings.pdf [https://perma.cc/EAH8-7VRQ].

289. U.S. GOV'T ACCOUNTABILITY OFF., GAO-05-59, GOVERNMENTS COULD DO MORE TO REDUCE DISPLAY IN PUBLIC RECORDS AND ON IDENTITY CARDS 3 (2004), https://www.gao.gov/assets/gao-05-59.pdf [https://perma.cc/SD2J-B6QT].

290. *E.g.*, Anita Ramasastry, *Can States Legally Put Residents' Social Security Numbers and Other Identifying Data Online? The Troubling Answer Is That They Can, and Do*, FINDLAW (Apr. 17, 2006), https://supreme.findlaw.com/legal-commentary/can-states-legally-put-residents-social-security-numbers-and-other-identifying-data-online-the-troubling-answer-is-that-they-can-and-do.html [https://perma.cc/DEN9-Y2XV] (discussing the practice in Ohio to publicly list UCC lien filings online, sometimes with social security numbers).

291. For a call for standardization of information reporting requirements in the environmental law context, see Leehi Yona, *Emissions Omissions: Greenhouse Gas Accounting Gaps*, 49 HARV. ENV'T L. REV. (forthcoming 2025) (manuscript at 64) (on file with the *Iowa Law Review*).

292. *U.S. Standard Certificate of Death*, CDC (Nov. 2003), https://www.cdc.gov/nchs/data/dvs/death11-03final-acc.pdf [https://perma.cc/7KUJ-U5MK].

293. Debra Houry, *We Are Not Overcounting Covid Deaths in the United States*, WASH. POST (Jan. 30, 2023, 1:58 PM) (on file with the *Iowa Law Review*) (stating, as the Chief Medical Officer of the CDC, that "[t]he most reliable way CDC gathers data on covid deaths is through provisional covid death counts based solely on death certificates").

294. Stephanie Pappas & LiveScience, *How Covid-19 Deaths Are Counted*, SCI. AM. (May 19, 2020), https://www.scientificamerican.com/article/how-covid-19-deaths-are-counted1 [https://perma.cc/MUM6-79AK].

295. *Reporting and Coding Deaths due to COVID-19*, NAT'L CTR. HEALTH STAT. (May 20, 2020), https://www.cdc.gov/nchs/covid19/coding-and-reporting.htm [https://perma.cc/9XAK-UXDA].

Another policy approach is to target the uses to which information can be put. These types of interventions can be nimble responses to evolving technologies and new applications of information. For instance, law enforcement agencies have long made mugshots and arrest information part of the public record.[296]

With the advent of the internet, enterprising third parties gathered these mugshots, posted them online so that they were readily findable in internet searches and offered to take them down if the subject of the mugshot so requested—and paid a fee.[297] Policymakers recognized that this sort of extortion was not in the public interest and several states passed laws preventing this use of mugshots.[298] In another example, courts often grant protective orders to bar public dissemination of information obtained from the opposing party during discovery.[299] But in some circumstances, the public interest in this information can override confidentiality concerns, and some states have statutes governing the scope of protective orders when cases involve safety hazards or questions of public health.[300]

Yet another policy device is to provide additional funding for research in areas with less data. Legal mechanisms can lead to skewed data in ways that hamper research and policymaking, for instance, by providing comprehensive records on foreclosures but not on eviction. The state could stimulate research and informed policymaking by giving supplementary grants for research in important areas with little data. This approach recognizes that the legal system can subsidize and encourage research by making public data available—but can sometimes achieve the same ends by making funding available in the absence of data. If evictions are an important policy problem that cannot easily be studied because of data deficiencies, the government can provide additional funding to fill that gap.

A further method to fix data distortions is to develop statistical correction factors. Sometimes recognizing ways in which legal mechanisms bias data allows statisticians and computer scientists to mathematically or programmatically "correct" the data.[301] To illustrate, death certificates are systematically inaccurate as to the race of the deceased because the person filling out the certificate must guess at the deceased's race based on visual observation, since they cannot, of course, ask the deceased. Inaccuracies are particularly high for Native Americans—a study from the CDC found that almost half of self-

296. Polatsek, *supra* note 248, at 916.

297. *Id.* at 917–18.

298. *Id.* at 916, 954.

299. Richard L. Marcus, *Myth and Reality in Protective Order Litigation*, 69 CORNELL L. REV. 1, 1–5 (1983).

300. *E.g.*, FLA. STAT. § 68.081 (2023).

301. *E.g.*, Adam Zewe, *Can Machine-Learning Models Overcome Biased Datasets?*, MIT NEWS (Feb. 21, 2022), https://news.mit.edu/2022/machine-learning-biased-data-0221 [https://perma.cc/LEE3-KKEG].

identified Native Americans were later misclassified as white on their death certificate.[302] The CDC has worked to develop correction factors to quantify this under-counting through links to other data sources that include self-reported racial identification like birth records, tribal registries, and some electronic medical records.[303] The CDC then estimates how public health data should be adjusted.[304]

Legal institutions can also reduce the consequences of biased legal data by being transparent about how the information is created and what it does and does not include. For instance, when the SEC mandates disclosure of certain types of information, the agency might indicate, in lay language, the parameters of the information that the rule is likely to make public. The agency could specify the thresholds for disclosure, the areas to which it relates, what is not included, and possibly summary descriptive statistics about certain information. Although this information is already sometimes accessible from reading the regulations, clearer documentation would make it easier for non-lawyers to understand the information and account for biases when using the information.

D. USER SELF-HELP

Users of legal information also have a role to play in preventing harms from biased data. If users understand the source of information, including what information is present and what is missing, they can sometimes avoid harmful misuse of information. Although better understanding will not solve all problems surrounding how legal information is used, it is an important component of improving legal information.

First, actors who aggregate pieces of legal information should take efforts to understand the information's shortcomings. Often, this is already done. For instance, many researchers who use legal information in empirical studies have deeply sophisticated understandings of the data's strengths and weaknesses.[305] They are often careful to specify that conclusions apply only to the groups whose data is included in the study, rather than more generally.

---

302. Elizabeth Arias, Melonie Heron & Jahn K. Hakes, *The Validity of Race and Hispanic-Origin Reporting on Death Certificates in the United States: An Update*, VITAL & HEALTH STAT., Aug. 2016, at 1, 6.

303. Robert N. Anderson, Glenn Copeland & John Mosely Hayes, Commentary, *Linkages to Improve Mortality Data for American Indians and Alaska Natives: A New Model for Death Reporting?*, AM. J. PUB. HEALTH S258, S258 (Supp. 2014), https://pmc.ncbi.nlm.nih.gov/articles/PMC4035882/pdf/AJPH.2013.301647.pdf [https://perma.cc/9ZDA-NUHP]. *See generally* CTRS. FOR DISEASE CONTROL & PREVENTION, IMPROVING ADOPTION OF EHR-BASED ELECTRONIC DEATH REPORTING (2017) (summarizing findings from California Department of Public Health initiative to enable the electronic health system to capture data from death certificate medical sections), https://www.cdc.gov/nchs/data/nvss/evital/Adoption_EHR_EDRS_Final_Observations_July_2017.pdf [https://perma.cc/G4UX-FN2V].

304. *See generally* Anderson, Copeland & Hayes, *supra* note 303.

305. *See supra* Section I.B.

But there is still much progress to be made in ensuring that the shortcomings of legal data are accounted for in data analysis. One area for improvement is in media reporting. Despite widespread recognition that death certificates are highly inaccurate with respect to race—and the existence of techniques for adjusting death certificate data to account for these inaccuracies—popular news outlets often report unadjusted racial data from death records.[306]

With respect to artificial intelligence, it is useful to know what data was used to train the model in order to evaluate the accuracy of the model's output and its applicability to different contexts. That knowledge can help mitigate the problems that are created when an AI model trained on one type of data (say, one type of court case) is applied in a different setting where it may perform more poorly. However, not all AI models disclose their training data.[307] This Article joins the chorus of calls for AI systems to disclose where their data comes from.[308]

Finally, improved public knowledge of data is important. The legal system is often viewed as authoritative and the information it produces assumed to be produced or reviewed by experts.[309] This may mean that readers are predisposed to trust legal data sources without thinking about their biases. For instance, death records can drive public opinion because they are trusted.[310] That can be a problem—death records are often used as a source of information about police killings, but several research studies have found that more than half of deaths due to police violence were not recorded in death certificates.[311] The reasons for undercounting are varied and may include lack of space on the form to report police involvement, the certifier's lack of training in how to fill out the certificate, incorrect translation of deaths due to police violence into classification codes used for standardization, and coroner complicity with the

306. *E.g.*, Paul Overberg & Jon Kamp, *Covid-19 Deaths Skew Younger Among Minorities; Coronavirus Infections Take a Heavy Toll on Latino People in Their Prime Working Years*, WALL ST. J. (Aug. 17, 2020, 3:54 AM) (on file with the *Iowa Law Review*); Sari Horwitz, Debbie Cenziper & Steven Rich, *As Opioids Flooded Tribal Lands Across the U.S., Overdose Deaths Skyrocketed*, WASH. POST (June 29, 2020) (on file with the *Iowa Law Review*).

307. Jack Hardinges, Elena Simperl & Nigel Shadbolt, *We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models*, HARV. DATA SCI. REV., Dec. 2023, at 2, 2–3.

308. *E.g.*, Mahima Pushkarna, *Documentation for Responsible AI*, 2022 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 1776, 1777 (2022) (exploring the use of data cards as a disclosure device); R. Stuart Geiger, et al., *Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?*, 2020 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 325, 326 (calling for additional disclosure relating to human-labeled training data); David A. Simon, Barbara J. Evans, Carmel Shachar & I. Glenn Cohen, *Should Alexa Diagnose Alzheimer's?: Legal and Ethical Issues With At-Home Consumer Devices*, 3 CELL REPS. MED. 1, 4 (2022) (discussing disclosure in the context of medical AI).

309. Freilich, *supra* note 57, at 1539–40.

310. Tim Arango & Shaila Dewan, *More Than Half of Police Killings are Mislabeled, New Study Says*, N.Y. TIMES (Sept. 30, 2021) (on file with the *Iowa Law Review*).

311. *Id.*

police.[312] This undercounting in official records contributes to public apathy and official inaction against police violence.[313] Better public understanding of the underlying legal information (here, death certificates) and its strengths and weaknesses might help.

CONCLUSION

Information is power. In choosing how to allocate access to information, legal doctrines and institutions implicitly make choices about who has access to information-driven power and whether that power will be held privately or distributed publicly. Keeping aggregate information about certain practices, transactions, or groups private gives power to those select entities who are able to access and aggregate nonpublic information.[314] And when automated systems rely on aggregate information, discrimination proliferates where members of the public and regulators do not have access to the aggregate information driving the system. In the information age, law is an ever-brighter lamppost, shining its light to facilitate information use in some areas and for some audiences, while keeping others in the shadows, with accidental harms and benefits. More deliberate consideration of law's information spillovers and their effects would help. For law, it is "better to illuminate than solely to shine."[315]

312. Fablina Sharara et al., *Fatal Police Violence by Race and State in the USA, 1980–2019: A Network Meta-Regression*, 398 LANCET 1239, 1250 (2019); Justin M. Feldman, Sofia Gruskin, Brent A. Coull & Nancy Krieger, *Quantifying Underreporting of Law-Enforcement-Related Deaths in United States Vital Statistics and News Media Based Data Sources: A Capture-Recapture Analysis*, PLOS MED., Oct. 2017, at 2, 2; Colin Loftin, David McDowall & Min Xie, *Underreporting of Homicides by Police in the United States, 1976-2013*, 21 HOMICIDE STUD. 159, 160 (2017).

313. Sharara et al., *supra* note 312, at 1250.

314. One quantifiable example is financial firms that aggregate and leverage information in high-frequency trading. *See* Olivier Sylvain, *Network Equality*, 67 HASTINGS L.J. 443, 471 (2016).

315. Quote by Thomas Aquinas, translated in JEAN-PIERRE TORRELL, SAINT THOMAS AQUINAS 166 (1996).