LAW SEARCH IN THE AGE OF THE ALGORITHM

Michael A. Livermore, Peter Beling, Keith Carlson, Faraz Dadgostari, Mauricio Guim & Daniel N. Rockmore[*]

2020 MICH. ST. L. REV. 1183

ABSTRACT

The process of searching for relevant legal materials is fundamental to legal reasoning. However, despite its enormous practical and theoretical importance, law search has not been given significant attention by scholars. In this Article, we define the problem of law search and examine the consequences of new technologies capable of automating this core lawyerly task. We introduce a theory of law search in which legal relevance is a sociological phenomenon that leads to convergence over a shared set of legal materials and explore the normative stakes of law search. We examine ways in which law scholars can understand empirically the phenomenon of law search, argue that computational modeling is a valuable epistemic tool in this domain, and report the results from a multi-year, interdisciplinary effort to develop an advanced law search algorithm based on human-generated data. Finally, we explore how policymakers can manage the challenges posed by new machine learning-based search technologies.

---

* This project has been carried out in extensive collaboration with a large group of colleagues, including Allen Riddell, Gregory Leibon, and Reed Harder. Our thanks for helpful comments from participants at the Online Workshop on the Computational Analysis of Law (especially Eugenie Dugoua), a faculty workshop at the University of Virginia School of Law, and the 2016 Conference on Empirical Legal Studies in Europe. We thank the Presidential Fellowship program at the University of Virginia Data Science Institute (now the School of Data Science) and MITRE for providing financial support for this project. Some of the work and ideas reported here first appeared in: Faraz Dadgostari et al., *Modeling Law Search as Prediction*, 29 A.I. & L. 3 (2021) and Greg Leibon et al., *Bending the Law: Geometric Tools for Quantifying Influence in the Multinetwork of Legal Opinions*, 26 A.I. & L. 145 (2018). We thank our coauthors and the anonymous peer reviewers for those papers for their helpful feedback. We also thank Andrew Teal for his research assistance and the editors at the *Michigan State Law Review* for their excellent work. Replication files, including data and code, can be found at: https://math.dartmouth.edu/~mslrlawsearchalgorithm/.

TABLE OF CONTENTS

INTRODUCTION

At its most fundamental, legal decision-making involves the interpretation of legal materials. Some scholars have argued that the use of legal materials is *the* distinctive feature of all legal reasoning.[1]

1. *See* Frederick Schauer & Virginia J. Wise, *Legal Positivism as Legal Information*, 82 CORNELL L. REV. 1080, 1082 (1997) [hereinafter *Legal Positivism as Legal Information*] (stating that "the distinctive character of legal reasoning . . . is the information set on which legal argumentation and legal decision[-]making relies"); *see also* FREDERICK SCHAUER, THINKING LIKE A LAWYER: A NEW INTRODUCTION TO LEGAL REASONING 5–7 (2009) (stating that for Schauer, the "special oddness" of legal reasoning—compared to other forms of rationality deployed "throughout our decision-making lives"—lies in that fact that it provides "route[s] towards reaching a

Others might add something more, such as analogical approaches to deciding cases or concerns with natural law, morality, well-being, or fairness.[2] But positive law plays at least some role in almost all accounts of legal reasoning.[3] Whether it is a judge issuing an opinion, a lawyer advising a client, or a business conforming to regulatory requirements, legal decision-making necessarily involves identifying, interpreting, and applying the law.

Although scholarship related to the interpretation and application of the law fills the shelves of libraries, the process of identifying relevant law—what we refer to in this Article as *law search*—is relatively little studied. The practical importance of finding relevant law is widely recognized and is reflected both in the law school curriculum (which has included courses on legal research for several decades) and in countless billable hours.[4] But scholarly accounts of law search are largely missing.[5]

This lacuna is especially problematic given the rapid and consequential technology-driven changes in law search in recent years, which have profound transformational potential. Recent advances in fields such as natural language processing, predictive analytics, machine learning, and computational text analysis create many new possibilities for understanding legal phenomena, and there

---

decision *other than* the best all-things-considered decision for the matter at hand"). *See generally* Frederick Schauer & Virginia J. Wise, *Nonlegal Information and the Delegalization of Law*, 29 J. LEGAL STUD. 495 (2000) [hereinafter *Nonlegal Information and the Delegalization of Law*].

2. *See generally* RONALD DWORKIN, LAW'S EMPIRE (1986); ROBERT P. GEORGE, IN DEFENSE OF NATURAL LAW (1999); RICHARD A. POSNER, HOW JUDGES THINK (2008).

3. *See* JOSEPH RAZ, THE AUTHORITY OF LAW: ESSAYS ON LAW AND MORALITY 38–40 (1979). Positive law means law as a matter of social fact, the identification of which does not require moral (or other) considerations. It is the law itself that provides a test for the identification of the content and determination of the existence of the sources authorized to produce binding legal rules.

4. *See, e.g.*, *First-Year Legal Research and Writing Program*, HARV. L. SCH., http://hls.harvard.edu/dept/lrw/ [https://perma.cc/CZV4-SJ57] (last visited Feb. 15, 2021) (reviewing the year-long legal research and writing course at Harvard Law School and describing its content as "series of sequenced, interrelated exercises introducing students to the way lawyers conduct legal research, analyze and frame legal positions, and present their work in writing and in oral argument"). *See generally* David Houlihan, *How Research Efficiency Impacts Law Firm Profitability*, LAW360 (Sept. 11, 2014, 2:13 PM), https://www.law360.com/articles/575667/how-research-efficiency-impacts-law-firm-profitability (reviewing how inefficiency in legal research impacts law firm profitability based on its impact on the utilization and capacity of associate attorney resources and research "write-offs").

5. *See infra* Section I.C for exceptions.

is an important trend in legal scholarship toward "law as data" research.[6] Law search is one area where these tools have begun to be applied more broadly within the profession, with both startup companies and the legacy commercial legal databases experimenting with the use of new technologies to enhance their law search offerings.[7]

This Article describes some initial steps in what we hope becomes a sustained interdisciplinary research program. We begin by defining the problem of law search and examining the consequences of new technologies capable of automating this core lawyerly task. We also describe the concepts of legal relevance and convergence and explain some of the related normative issues. We then turn to the study of law search. There are a number of different empirical methods that can be used to study this phenomenon, but we argue that computational modeling can provide a unique epistemic lens in this area. To illustrate the value of this approach, we report the results from a multi-year, interdisciplinary effort to develop an advanced law search algorithm. Finally, we explore some policy challenges created by new law search technologies and discuss some possible responses that can facilitate open and nonbiased access to the law.

Law search is a catchall phrase used to encompass a general process of finding case law, statutes, or other materials relevant to a legal question or argument. It is through search that judges, administrative hearing officers, lawyers, and laypeople identify the subset of authoritative legal texts that apply to a legal matter of interest. Often, the term "legal research" is used to capture a similar idea.[8] We use the term "law search" rather than "legal research" to clarify that the process we are discussing is that of identifying extant relevant information, rather than creating new knowledge. Our usage better accords with how these terms are deployed in other fields; for example, economics or medical "research" builds the existing stock of knowledge, while a "search" might be carried out in those fields as part of a literature review.

It is worth emphasizing that legal reasoning occurs both prior and subsequent to law search. Before relevant law can be identified, personal, social, and business conduct must first be translated into

6. *See* Michael A. Livermore & Daniel N. Rockmore, *Distant Reading the Law*, *in* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS 3 (Michael A. Livermore & Daniel N. Rockmore eds., 2019).

7. *See infra* Section I.B for examples.

8. *See generally* AMY E. SLOAN, BASIC LEGAL RESEARCH: TOOLS AND STRATEGIES (4th ed. 2009).

legal questions arising from relationships, obligations, and liabilities. After the relevant law is identified, it must be applied to the question at hand. But law search serves as a linchpin in this process, acting as the focal point of both the initial process of translation and subsequent processes of interpretation and application.

From a practical perspective, legal professionals are hired, at least in part, for their ability to find the relevant law and bring it to bear on their clients' concerns. Major private companies (such as Westlaw, LexisNexis, and Bloomberg Law) offer search-facilitating services on a large-scale commercial basis.[9] Attorneys now rely on these services, and for many, they are the primary means of accessing the law. Technological advances associated with the digitization of legal texts and computerized search engines are only the most recent in a long tradition that seeks to lower the cost of identifying relevant legal materials. As far back as the sixth century, the Byzantine Emperor Justinian I attempted to organize then-extant legal traditions into a comprehensible whole in the *Corpus Juris Civilis* that he commissioned.[10] Common law traditions presented particular problems and led to texts such as Blackstone's Commentaries in eighteenth-century England and the West American Digest System in the early twentieth-century United States.[11]

There is a small body of research that examines how efforts like these to organize the body of law both reflect and influence the practice of legal reasoning.[12] What has been missing from this prior work is an effort to systematize the study of law search. This lacuna exists despite the extensive body of research on search more generally, in fields such as economics, computer science, and psychology.[13] The normative values that are implicated by law search remain murky, and no general conceptual and methodological apparatus has been

9. *See Lexis Advance*®, LEXIS NEXIS, https://www.lexisnexis.com/en-us/products/lexis-advance.page [https://perma.cc/334R-XDB3] (last visited Feb. 15, 2021); *Westlaw Edge*, THOMSON REUTERS, http://legalsolutions.thomsonreuters.com/law-products/westlaw-legal-research/ [https://perma.cc/5ZJ6-LNTM] (last visited Feb. 15, 2021); BLOOMBERG LAW, https://www.bna.com/bloomberglaw/ [https://perma.cc/UMQ7-JBNL] (last visited Feb. 15, 2021).

10. For an English translation, with original text, of the Codex of Justinian see generally THE CODEX OF JUSTINIAN (Bruce W. Frier ed., 2016).

11. *See* WILLIAM BLACKSTONE, COMMENTARIES ON THE LAWS OF ENGLAND (David Lemmings ed., 2016) (1893); *see, e.g.*, 4 THE AMERICAN DIGEST ANNOTATED (1909).

12. *See infra* Section I.C.

13. *See infra* Section III.A for a brief discussion of this literature.

developed to facilitate consistent, cumulative research into law search as an empirical phenomenon.

There are several distinctive features of law search that separate it from other forms of search and make it a worthwhile object of study in its own right. For example, there are normative issues raised by law search that are not implicated by other forms of search. From a consequentialist perspective, law search can be understood as an optimization problem minimizing search costs while achieving a socially desirable level of legal accuracy or cooperation. Social decision-makers should consider whether private incentives alone will generate optimal search behavior, or whether government intervention is needed to correct for market failures. From a deontological perspective, law search may implicate social justice or one's obligations as a law-abiding subject, a lawyer advising clients, or a judge issuing decisions. The interaction of law search and these deontological considerations is an area ripe for continued inquiry.

Law search is also distinct from other kinds of information search in that often, the goal is to *converge* on a set of shared materials that can be used to analyze a legal question. The related concept of *legal relevance* is subtle and nuanced, involving subjective judgments along several potential dimensions. One approach to understanding legal relevance is sociological and takes the relevance of legal documents to a particular legal question to be a matter of social fact, determined by the judgments made by members of the legal community in question. Convergence occurs when competent members of a legal community faced with the same legal question identify the same sources of relevant legal authority. Convergence can be (at least theoretically) complete, imperfect, or lacking altogether. One normative question we will explore in more detail is whether complete convergence is a desirable feature of a legal order. At first blush, complete convergence may seem to be an unobjectionable good that should be promoted. We will complicate that intuition by arguing that the values of flexibility and dynamism in a legal order can, at least plausibly, make desirable some departure from complete convergence.

From an empirical perspective, there are several different lenses through which law search can be seen and understood. In the behavioral and social sciences, there are many strands of research that examine information-seeking behavior as an economic, sociological, or psychological phenomenon using traditional disciplinary tools.[14]

14. *See generally* THEORIES OF INFORMATION BEHAVIOR (Karen E. Fisher, Sanda Erdelez & Lynne (E. F.) McKechnie eds., Am. Soc'y Info. Sci. & Tech. 2005)

An alternative approach to understanding search has arisen in the fields of computer science and software engineering, which poses the question in terms of user assistance with the "[i]nformation retrieval" problem.[15]

Drawing on both the information behavior and information retrieval approaches, we engaged in a multi-year interdisciplinary collaboration to construct an advanced computational model of law search, which we refer to as *LexQuery*. This model performs well against two useful benchmarks: the ability to predict the citations in judicial opinions, and conformity with human judgments concerning legal similarity. By defining these benchmarks and releasing the code and data that embody *LexQuery*, we hope to spur other researchers to take on the problem of law search and improve on this early-stage effort.

Given the importance of law search to the legal order, there is also a place for public policy to promote social values that new technologies might threaten. We identify several ways in which the private market for law search technologies might not align with overall social welfare or other important values. However, we also argue that direct regulation of this market carries substantial downsides. We propose instead that the government support a robust open access

---

(containing an edited volume collecting multiple perspectives) [hereinafter THEORIES OF INFORMATION BEHAVIOR]; Dale T. Mortensen & Christopher A. Pissarides, *Job Creation and Job Destruction in the Theory of Unemployment*, 61 REV. ECON. STUD. 397, 397 (1994) (modeling "a job-specific shock process in the matching model of unemployment with non-cooperative wage" behavior); Robert S. Ledley & Lee B. Lusted, *Reasoning Foundations of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason*, 130 SCIENCE 9, 9 (1959) (analyzing the "complicated reasoning processes" involved in medical diagnoses). *See generally* Meir G. Kohn & Steven Shavell, *The Theory of Search*, 9 J. ECON. THEORY 93 (1974) (developing basic results on the problem of economic behavior search); Peter A. Diamond, *Aggregate Demand Management in Search Equilibrium*, 90 J. POL. ECON. 881 (1982) (analyzing search equilibrium as a simple barter model with identical risk-neutral agents where their interactions are coordinated by a stochastic matching process); Gabriel Ramos-Fernández et al., *Lévy Walk Patterns in the Foraging Movements of Spider Monkeys (Ateles Geoffroyi)*, 55 BEHAV. ECOLOGY & SOCIOBIOLOGY 223 (2004) (describing the foraging patterns of spider monkeys); STEVE ALPERN & SHMUEL GAL, THE THEORY OF SEARCH GAMES AND RENDEZVOUS (2003) (describing the theory of search games and rendezvous search).

15. *See* Calvin N. Mooers, *The Theory of Digital Handling of Non-Numerical Information and Its Implications to Machine Economics*, 48 ZATOR TECH. BULL., 1, 3 (1950) (describing the information retrieval problem as a nonnumerical problem). *See generally* CHRISTOPHER D. MANNING, PRABHAKAR RAGHAVAN & HINRICH SCHÜTZE, AN INTRODUCTION TO INFORMATION RETRIEVAL (2009) (describing the basics of information retrieval).

research environment, which can provide search tools that are widely available to the public. Such tools can augment the private market for law search in ways that help address its limitations.

The remainder of this Article proceeds as follows. Part I defines law search and discusses its importance and coevolution with information technology. Part II delves more deeply into the normative stakes of law search, discussing the concept of legal relevance and exploring the social function of law search. Part III focuses on efforts to study law search empirically, articulates the difference between the study of information seeking behavior and computational models of information retrieval, and reports the results of our multi-year effort to develop a computational law search model, which culminated in the *LexQuery* model. Part IV describes the public policy implications of law search, describing potential market failures and the ways that policy can promote social goals in this domain.

I. FINDING THE LAW

Law search has played a role in the practice of law since the dawn of the profession, with technologies of the day—from books, to library organizing systems, to artificial intelligence—strongly influencing how this aspect of the legal profession is practiced. Despite its importance, law search has not been subject to a commensurate level of study from a scholarly perspective. In this Part, we briefly discuss the history of law search and its coevolution with information technology and provide an overview of prior law scholarship on search.

A. Corpus Juris to LexisNexis

Information systems are tools that allow users to search and find information of interest.[16] Every information system has two parts. The first is the body of underlying documents, data, or other primary media that holds the information. The second is an organizing principle that allows the user to locate the information. For small collections of information, an organizing principle can be as simple as chronological or alphabetical order. As collections grow in size, more elaborate organizing principles are required to facilitate information retrieval.

16. *See* Robert C. Berring, *Collapse of the Structure of the Legal Research Universe: The Imperative of Digital Information*, 69 WASH. L. REV. 9, 16 (1994).

The law provides an example of a large collection of information that, as it grows, requires increasingly sophisticated information systems to navigate. The written law itself can be understood as a system for collecting widely diffuse information—in the form of behavioral norms and conventions—and publishing them for ease of access.[17] The process of translating diffuse community-held information about norms into written form necessarily leads to information loss because the medium of written text is incapable of capturing all the information embedded in community-level practices and individual experiences and understandings. In addition, after translation to the written word, there may be both intentional and unintentional alteration of the content of the norms. Indeed, an advantage of written norms is that they are amenable to relatively rapid explicit change through recognized procedures.[18]

In the early sixth century, Byzantine Emperor Justinian I commissioned his *Corpus Juris Civilis*, which at the time was an extremely resource-intensive and sophisticated effort to collect the extant law, summarize it in written form, and organize it for publication and diffusion.[19] The *Corpus Juris* included four multivolume books: a collection of imperial ordinances; a collection of the writings of jurists; an elementary legal textbook for use in education; and a collection of new ordinances issued by Justinian.[20] All law not included in the *Corpus Juris* was declared invalid.[21] As an information technology, the *Corpus Juris* compressed (by excluding many texts), summarized (in explanatory notes), organized (according to subject matter), and disseminated (via publication) a vast body of information that was included in the law. This was an incredible advance and served as the foundation for many hundreds of years of legal development.

---

17. *See generally* GILLIAN K. HADFIELD, RULES FOR A FLAT WORLD: WHY HUMANS INVENTED LAW AND HOW TO REINVENT IT FOR A COMPLEX GLOBAL ECONOMY (2017).

18. *Cf.* H.L.A. HART, THE CONCEPT OF LAW 317 (1961) (describing "rule of recognition" as a defining feature of legal systems).

19. *See* JOHN HENRY MERRYMAN & ROGELIO PÉREZ-PERDOMO, THE CIVIL LAW TRADITION: AN INTRODUCTION TO THE LEGAL SYSTEMS OF EUROPE AND LATIN AMERICA 7 (2007).

20. *See generally* Wolfgang Kaiser, *Justinian and the Corpus Iuris Civilis*, *in* THE CAMBRIDGE COMPANION TO ROMAN LAW (David Johnston ed., 2015) (providing historical context for the *Corpus Juris Civilis* and describing its preservation into the contemporary period).

21. *See* ANDREW M. RIGGSBY, ROMAN LAW AND THE LEGAL WORLD OF THE ROMANS 39–40 (2010).

One advantage of code-based legal regimes is that by deemphasizing judicially created precedent, they substantially constrain the number of legally relevant documents. The tradeoff for that limitation comes in terms of comprehensiveness, as code-based systems leave underspecified areas of law that judicial doctrine could help fill out. Common law systems, with their emphasis on judicial precedent, have the inverse problem. The law may be more comprehensive, in the sense of specifying the outcome of more legal questions, but the profusion of documents makes it vastly more difficult to identify the relevant law.[22]

One approach to dealing with this case law is via summaries, such as *Blackstone's Commentaries on the Laws of England* from the late eighteenth century.[23] These commentaries are not meant to supplant the law, but rather to compress the information held in a large set of diffuse documents into a more readily accessible form. An alternative approach is codification, which aims to replace judicially created law with centralized documents.[24] In the United States, the American Law Institute issues Restatements concerning various areas of common law that are intended to distill the body of judicial decisions into a set of more general principles and rules that can be more readily identified and applied.[25]

More sophisticated information systems can promote the retrieval of case law, providing direct access to primary legal sources (unlike summaries) and reducing pressure for codification. In the United States, a major leap forward in developing an organizing system suited to the growing body of law came with John B. West and The West Publishing Company's American Digest System, first published in 1889.[26] West had the most comprehensive collection of

22. A larger number of plausibly relevant prior cases may also allow judges to "cherry pick precedents" in ways that expand judicial discretion. For a statistical examination of this theory, see generally Anthony Niblett, *Do Judges Cherry Pick Precedents to Justify Extra-Legal Decisions?: A Statistical Examination*, 70 MD. L. REV. 234 (2010).

23. *See* Robert C. Berring, *The Ultimate Oldie but Goodie: William Blackstone's* Commentaries on the Law of England, 4 J.L. 189, 190 (2014) (tracing Blackstone's *Commentaries on the Law of England* to a series of lectures that Blackstone gave to students at Oxford).

24. *See* Nuno Garoupa & Andrew P. Morriss, *The Fable of the Codes: The Efficiency of the Common Law, Legal Origins, and Codification Movements*, 5 U. ILL. L. REV. 1443, 1443, 1445 (2012).

25. *See generally* RESTATEMENT (SECOND) OF CONTS. (AM. L. INST. 1981).

26. *See generally* 1 THE AMERICAN DIGEST ANNOTATED (West Publ'g Co. 1888).

documents of the day in the National Reporter publications and he made them available along with an organizing system that could allow a user to find and retrieve the cases.[27]

The American Digest System is a subject matter-based index of legal documents. It creates a taxonomy by assigning every reported case to one of seven broad categories: Persons, Property, Contracts, Torts, Crimes, Remedies, and Government.[28] Inside these seven categories are 430 topics.[29] These 430 topics are further divided into subtopics, called key numbers; these key numbers contain headnotes, which are small abstracts of each point of law contained in a decision.[30]

The categories were meant to encompass the entire universe of the law so that a lawyer with good judgment would know the appropriate category for every legal question. The act of classifying cases had implications beyond mere organization. Robert Berring has argued that the act of classifying cases affects how law is understood and develops.[31] When a West editor assigns a headnote to one topic and not another, that person makes at least an implicit statement on the scope of a legal rule. Some have argued that, as the National Reporter and the American Digest System became the primary method of case retrieval, it influenced thinking about the law.[32] It is worth noting that categories need not be conceptual to affect the development of the law. Even the regional classification system developed by West, which groups together the state law reporters for several states, appears to have affected the diffusion of ideas within the judiciary.[33]

After West, the next major innovation in legal information systems came during the broader digital information technology

27. *See* JOSEPH L. GERKEN, THE INVENTION OF LEGAL RESEARCH 204 (2016).

28. *See* Robert C. Berring, *Full-Text Databases and Legal Research: Backing into the Future*, 1 HIGH TECH. L.J. 27, 31 (1986).

29. *See id.*

30. *See id.* at 31–32.

31. *See id.* at 32.

32. *See, e.g.*, *id.* at 33 ("The Key Number System provided a paradigm for thinking about the law itself."). Legal classifications may also interact with broader social norms and understandings to jointly affect how legal disputes are perceived. In recent work, researchers show that the language of criminal trial transcripts in the London Central Criminal Court changed substantially over the course of the nineteenth century as the role of violence in society shifted during that period. *See* Sara Klingenstein, Tim Hitchcock & Simon DeDeo, *The Civilizing Process in London's Old Bailey*, 111 PROC. NAT'L ACAD. SCIS. 9419, 9419 (2014).

33. *See* Gregory A. Caldeira, *The Transmission of Legal Precedent: A Study of State Supreme Courts*, 79 AM. POL. SCI. REV. 178, 181 (1985).

revolution. In 1970, Lexis launched its first computer-based database, which allowed for the use of keyword searches to retrieve documents.[34] The new electronic databases allowed researchers to find cases based directly on the words they contained. This new system had several advantages over book-based resources. For example, full-text electronic databases allowed lawyers and judges to search for particular words, like the name of a party, statute, judge, geographic location, product, or trademark.[35] This search approach was not possible using the legal categories in the American Digest System index. Keyword searches also allowed direct access to legal documents unmitigated by professionally organized subject matter categories. Just as the West system tended to herd and partition legal issues into distinct categories, keyword searches, with their tendency to cut across traditional doctrinal areas, opened up the potential for thinking about the law in new ways. For Berring, the new search technology ultimately undermined "judicial and professional conformity and conservatism."[36]

Although seemingly straightforward, the technology behind keyword searches created the platform for today's more sophisticated natural language processing and machine learning approaches. For example, one early approach to representing quantitatively the content of documents is the "tf-idf" metric, which stands for term-frequency/inverse-document frequency. This metric (which depends on a given document and term) captures how frequently a word appears in a document compared to how common it is over the relevant corpus. A rare word—from the point of view of the corpus—that appears many times in a given document has a high tf-idf. It turns out that tf-idf is a useful way to mathematically represent documents to inform the generation of results from keyword searches.[37] These types of metrics served as the basis for the more sophisticated machine learning and natural language processing algorithms to come.

34. *See* Berring, *supra* note 28, at 38.

35. *See id.* at 42.

36. *Id.*

37. *Cf.* H. P. Luhn, *A Statistical Approach to Mechanized Encoding and Searching of Literary Information*, 1 IBM J. RSCH. & DEV. 309, 309 (1957) (proposing a novel statistical approach to keyword searches); Karen Spärck Jones, *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, 28 J. DOCUMENTATION 11, 12 (1972) (discussing the application of a term weighting system on information retrieval).

B. New Technologies

In recent years, the major commercial search engines have expanded their capabilities by offering users access to increasingly sophisticated natural language search queries, some of which are informed by machine learning algorithms.[38] This algorithmic approach to electronic law search holds the promise of reducing the cost of search and increasing the accuracy and value of results, but is also considerably less transparent to users, raising a host of important practical and normative questions.

The shift to more sophisticated algorithmic law search encompasses a range of technologies. At one end of this range are fairly straightforward approaches, such as knowledge representation systems that generate search inquiries from a string of natural language text based on linguistic structure.[39] These types of knowledge representation systems, which rely on relatively simple, deterministic logical operations hand-generated by human programmers based on substantive expertise, are sometimes referred to as "good old fashion artificial intelligence" (GOFAI) because they were the foundation of the first wave of artificial intelligence research in the 1970s and 1980s.[40] Systems for translating natural language into search queries reduce start-up costs because searchers do not need to learn a syntax (such as Boolean terms and connectors) to successfully conduct searches.

GOFAI sought to translate human expertise into executable code by representing knowledge directly as logical operations. The more recent generation of artificial intelligence takes data—typically large volumes of data—and then manipulates that data with algorithms to extract patterns that are useful for purpose of prediction and description. An important hallmark of these approaches is that they are generally quite naïve and nonparametric, are not based on in-depth

38. *See* Jack G. Conrad & Qiang Lu, *Next Generation Legal Search—It's Already Here*, LEGAL INFO. INST. (Mar. 28, 2013), https://blog.law.cornell.edu/voxpop/2013/03/28/next-generation-legal-search-its-already-here/ [https://perma.cc/Q6A4-UBU6].

39. For an overview of knowledge representation, see generally BUILDING EXPERT SYSTEMS (Frederick Hayes-Roth, Donald A. Waterman & Douglas B. Lenat eds., 1983).

40. *See* JOHN HAUGELAND, ARTIFICIAL INTELLIGENCE: THE VERY IDEA 112–16 (1985); AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE 32 (2018) (drawing contrast between traditional knowledge representation and machine learning models).

substantive knowledge, and run with relatively little human intervention.

Machine (natural language) translation provides a useful illustration of the contrast between GOFAI and new data-based artificial intelligence. A knowledge representation translation system would seek to encode the linguistic knowledge of human experts into a set of rules, which could then be deployed on natural language to engage in translation tasks. Such knowledge representation systems performed notoriously poorly, even when substantial resources were devoted to them.[41] A data-based approach uses translated texts that appear in multiple languages (for example, the Bible or various United Nations documents) as inputs, and then deploys algorithms such as neural networks (the foundation of so-called deep learning methods), on that data to construct models for how best to "predict" the English version of a document from the French version (for example). In practice, the data-based approaches substantially outperform those based on human expertise.[42]

Translation is most commonly approached as a problem in "supervised learning" in which data is labeled with appropriate metadata (in this case, texts are matched with each other and labeled by their language). But data-based artificial intelligence also encompasses unsupervised analysis as well. The goal of unsupervised learning approaches is to extract patterns even from unstructured data that lacks labels. One particularly important computational text-analysis tool with implications for search is *topic modeling*, a family of natural language analytic tools.[43] Topic models extract latent "topic" variables, represented as distributions over the vocabulary of the corpus, which match intuitive subject matter categories.[44] Topic models essentially act as a way to automate subject matter labeling in a manner akin to the headnote categorization used in the American Digest System.[45] Their potential value for search has been recognized,

41. *See id.*

42. *See id.*

43. *See* David M. Blei, *Probabilistic Topic Models*, 55 COMMC'NS ACM 77, 77 (2012).

44. A "distribution" over a set of words assigns weights (nonnegative numbers) to each of the words; these weights sum to one. The most highly weighted words in the distribution give a sense of what the topic is about. *See id.* at 77–78.

45. *See* Michael A. Livermore, Allen B. Riddell & Daniel N. Rockmore, *The Supreme Court and the Judicial Genre*, 59 ARIZ. L. REV. 837, 842 (2017).

including in the context of legal documents.[46] Topic models have also seen broader application by scholars interested in articulating patterns in corpora of documents in the social sciences and the humanities.[47] A related technology (and mathematical representation) generate word and document embeddings, which provide another approach for extracting high-level semantic content from large corpora of data.[48]

Existing commercial legal databases have considerable amounts of data that can be used by AI-based systems. Expert-generated annotations, such as Westlaw's headnotes, contain extremely high-quality information that can be used to train a supervised learner or validate unsupervised analysis such as a topic model. Citation networks are also an important source of information, especially for judicial documents. Finally, and perhaps most important, commercial databases have access to large volumes of user information, including searches, click-through rates, and level of engagement with different documents. All of this information can be used by AI-based systems.

Leading legal search engines have begun to use their data advantages to leverage artificial intelligence to inform search results.[49] New entrants are also moving quickly to gain a foothold in AI-informed law search. One example is the ROSS Intelligence AI-supported legal research platform.[50] The ROSS platform is powered by the company's own artificial intelligence system, which combines natural language processing with machine learning capabilities to

---

46. *See* Talia Schwartz, Michael Berger & Juan Hernandez, A Legal Citation Recommendation Engine Using Topic Modeling and Semantic Similarity (2015) (unpublished manuscript) (on file with authors).

47. *See, e.g.*, Kevin M. Quinn et al., *How to Analyze Political Attention with Minimal Assumptions and Costs*, 54 AM. J. POL. SCI. 209, 210 (2010); Allen Beye Riddell, *How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models*, *in* DISTANT READINGS: TOPOLOGIES OF GERMAN CULTURE IN THE LONG NINETEENTH CENTURY 91 (Matt Erlin & Lynne Tatlock eds., 2014) (applying topic models in humanities context).

48. *See, e.g.*, Quoc Le & Tomas Mikolov, *Distributed Representations of Sentences and Documents*, 32 PROC. 31ST INT'L CONF. ON MACH. LEARNING 1188 (Eric P. Xing & Tony Jebara eds., 2014) (describing embeddings approach).

49. *See generally* Qiang Lu & Jack G. Conrad, *Bringing Order to Legal Documents: An Issue-Based Recommendation System Via Cluster Association*, 1 PROC. INT'L CONF. ON KNOWLEDGE ENG'G & ONTOLOGY DEV. 76 (Joaquim Filipe & Jan Dietz eds., 2012) (describing work on WestlawNext search algorithm).

50. *See generally* Ron Friedmann, *Exploring the ROSS and Fastcase Partnership*, LAC GRP. (Dec. 11, 2019), https://lac-group.com/blog/exploring-ramifications-ross-fastcase-partnership/ [https://perma.cc/X86J-CPDY] (interviewing ROSS CEO Andrew Arruda regarding the aligning of two legal research companies, ROSS and Fastcase, to create a new legal research platform).

identify legal authorities that are relevant to particular questions. Users formulate a legal question in plain natural language, and ROSS's artificial intelligence system returns answers that are responsive to that question.[51] Casetext's CARA is another AI-based law search engine that uses whole documents as inputs and then uses natural language processing tools to generate a set of recommended relevant legal authorities.[52] "Vincent," a similar artificial intelligence feature offered by vLex Justis, applies natural language processing and machine learning to whole document inputs to suggest related legal authorities across international jurisdictions.[53] Casetext has also developed a separate platform, "Compose," which uses machine learning to draft legal arguments by recommending conceptually similar legal authorities that fit a user's fact pattern.[54] Startups like LawGeex and ThoughtRiver have taken natural language processing beyond litigation, using artificial intelligence to analyze contracts and recommend relevant clauses present in similar legal documents.[55] Some legal startups have even published benchmarking data with the goal of spurring research that can deliver "qualitative improvements" akin to the "dramatic improvements in computer vision and deep learning" that was achieved on the basis of large repositories of

51. According to a recent study based on user feedback, ROSS retrieved 42.9% more relevant authorities than natural language and Boolean searches. *See generally* DAVID HOULIHAN, ROSS INTELLIGENCE AND ARTIFICIAL INTELLIGENCE IN LEGAL RESEARCH (2017).

52. *See* CASETEXT, http://www.casetext.com/cara-ai/ [https://perma.cc/AUA2-Q8GW] (last visited Feb. 15, 2021) (providing brief background on CARA and its process); *cf.* Beth Hoover, *Introducing Clerk: Win More Motions with Intelligent Brief Analysis*, JUDICATA (Oct. 5, 2017), https://blog.judicata.com/introducing-clerk-848abbed8fd3 [https://perma.cc/4TBK-7KVP] (discussing a similar product, Clerk, which is limited to California state law but seeking to expand to more states).

53. *See Quick Start Guide*, VLEX JUSTIS, http://justis.com/vlexjustis-user-guide.pdf [https://perma.cc/Q76E-T4XX] (last visited Feb. 15, 2021) (describing Vincent's ability to search for relevant legal authorities in both English and Spanish across nine international jurisdictions).

54. *See* COMPOSE, https://compose.law [https://perma.cc/98HD-5WMY] (last visited Feb. 15, 2021) (explaining Compose's process).

55. *See* LAWGEEX, https://www.lawgeex.com [https://perma.cc/9QA8-WA8Q] (last visited Feb. 15, 2021) (discussing LaxGeex's process in analyzing contracts); *see also* THOUGHTRIVER, https://www.thoughtriver.com [https://perma.cc/Z6J7-7WSC] (last visited Feb. 15, 2021) (discussing how ThoughtRiver's prescreening and artificial intelligence technology helps lawyers in contract negotiation).

labeled images that was made available to the artificial intelligence research community.[56]

The push to incorporate ever more sophisticated computational tools into law search is driven by users' desire for high quality, fast, and usable search results at low cost. Competitive market pressures provide ample incentive for both legacy actors such as LexisNexis and Westlaw and startups to chase technological advances. To the extent that the costs and benefits of law search are fully internalized to private actors, and consumers of these services can be expected to act rationally, market innovation can generally be expected to produce net social benefits. Although law search has characteristics of private market behavior, the *legal* nature of law search also implicates public values in ways that may create distance between the market equilibrium and the social optimum. Especially as the technologies that undergird law search become more complex and more opaque, it is worth inquiring into whether market and technological forces are working in ways that are consistent with social values.

C. Prior Work

Although the amount of scholarly attention devoted to law search has generally not matched its conceptual and practical significance, its importance has been recognized by some, especially as the digital revolution called attention to how changes to law search can affect how law is practiced.[57]

For example, some scholars have focused on the ways that digitized law search makes the boundary between the legal and non-legal worlds more permeable. M. Ethan Katsh argues that the reduced cost of accessing legal materials will make the law more accessible to nonlawyers, essentially democratizing the process of legal development.[58] Frederick Schauer and Virginia J. Wise point to an opposite effect as lawyers more easily access nonlegal information.[59] These authors raise the potential for a "[*d*]*elegalization of* [*l*]*aw*" as

---

56. *See* Itai Gurari, *Legal Search: Sharing Judicata's Data to Drive Progress*, JUDICATA (Aug. 2, 2017), https://blog.judicata.com/legal-search-sharing-judicatas-data-to-drive-progress-811eed64f04b [https://perma.cc/K6UR-5RQ9].

57. *See* Stefan H. Krieger & Katrina Fischer Kuh, *Accessing Law: An Empirical Study Exploring the Influence of Legal Research Medium*, 16 VAND. J. ENT. & TECH. L. 757, 759–60 n.6 (2014) (citing to and collecting sources).

58. *See* M. ETHAN KATSH, LAW IN A DIGITAL WORLD 57–59 (1995).

59. *See Nonlegal Information and the Delegalization of Law*, *supra* note 1, at 495; *see also Legal Positivism as Legal Information*, *supra* note 1, at 1091.

the distinction breaks down between traditional legal materials and other types of documents—for example, work in the social sciences.[60] Some judicial commentators, most notably Judge Richard Posner, have argued in favor of increasing reliance of legal institutions on the many nontraditional sources that are enabled by widespread digitization and public availability of information.[61]

Scholars have also discussed the interaction between law search, information systems, and legal thinking. As discussed earlier, Berring has written on the importance of the West American Digest System in structuring legal thinking according to categories and on the liberating potential of keyword-based searches.[62] Berring has also argued that practice-based searches—such as those focused on individual judges or legal adversaries—will foster a more "realistic" practice of law and potentially lower transaction costs in the legal system.[63] Along similar lines, Richard Delgado and Jean Stefancic argue that innovative jurisprudence involves challenging old categories, a process facilitated by the less rigidly structured keyword-based information systems.[64] Others have expressed more skepticism, arguing, for example, that electronic search will make it more likely that lawyers will "tilt[] at windmills" by advancing "marginal cases, theories, and arguments"[65] or will encourage lawyers to "neglect broader issues and legal concepts" in favor of fact-oriented searches.[66] Still, others argue that search technology may have less influence than suspected, in part

60. Schauer and Wise show that, in fact, there is an increasing use of nonlegal information by lawyers and judges in the period after the major legal information companies merged with nonlegal information provision firms. *See* Frederick Schauer & Virginia Wise, *Bundling, Boundary Setting, and the Privatization of Legal Information*, *in* MARKET-BASED GOVERNANCE: SUPPLY SIDE, DEMAND SIDE, UPSIDE, AND DOWNSIDE 134, 141 n.18 (John D. Donahue & Joseph S. Nye Jr. eds., 2002) (emphasis added).

61. *See* RICHARD A. POSNER, REFLECTIONS ON JUDGING 37 (2013).

62. *See* Berring, *supra* note 28, at 42–43.

63. *See id.* at 42.

64. *See* Richard Delgado & Jean Stefancic, *Why Do We Tell the Same Stories?: Law Reform, Critical Librarianship, and the Triple Helix Dilemma*, 42 STAN. L. REV. 207, 209 (1989).

65. Katrina Fischer Kuh, *Electronically Manufactured Law*, 22 HARV. J.L. & TECH. 223, 226 (2008).

66. Carol M. Bast & Ransford C. Pyle, *Legal Research in the Computer Age: A Paradigm Shift?*, 93 L. LIBR. J. 285, 298 (2001); *see also* F. Allan Hanson, *From Key Words to Key Numbers: How Automation Has Transformed the Law*, 94 L. LIBR. J. 563, 582 (2002) (arguing the electronic search has facilitated an "image of the law as a relatively unorganized assortment of facts and doctrines" rather than "a hierarchy governed by general principles").

because of the conservative influence of the system of informal apprenticeship that remains central to lawyers' professional development.[67] The limited amount of empirical work that has been done on changes in prevailing search technologies has generally found important, if not overwhelming, effects on certain attorney behaviors such as citation choices and research approaches.[68] More generally, scholars have speculated on the potential for new technologies, and especially automation and artificial intelligence to affect how law is enacted and practiced. For example, scholars have argued that new technologies will render old distinctions—such as between rules and standards—obsolete as algorithmically defined rules allow for both the flexibility of standards and the predictability of rules.[69] Several commentators have focused on the potential for predictive algorithms to replace some of the work currently carried out by lawyers.[70] Advanced natural language search tools and sophisticated machine learning techniques of search and summarization are part of this broader trend toward automation of some legal tasks.

Outside the legal domain, a range of social issues relating to search have emerged. Given the importance of search in managing the flow of information in society, some scholars have argued that major players such as Google should be subject to antitrust regulation or at least face "publicly funded alternatives" that limit the ability of private

---

67. *See* Judith Lihosit, *Research in the Wild: CALR and the Role of Informal Apprenticeship in Attorney Training*, 101 L. LIB. J. 157, 175–76 (2009); Paul Hellyer, *Assessing the Influence of Computer-Assisted Legal Research: A Study of California Supreme Court Opinions*, 97 L. LIBR. J. 285, 298 (2005).

68. *See* Krieger & Kuh, *supra* note 57, at 777; Casey R. Fronk, *The Cost of Judicial Citation: An Empirical Investigation of Citation Practices in the Federal Appellate Courts*, 2010 U. ILL. J.L., TECH. & POL'Y 51, 51 (2010).

69. *See* Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 IND. L.J. 1401, 1401 (2017); *see also* John O. McGinnis & Steven Wasick, *Law's Algorithm*, 66 FLA. L. REV. 991, 991 (2014).

70. *See id.* at 1024–25; *see also* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1147–48 (2017) (analyzing administrative rule-making and enforcement by machine learning algorithms); Daniel Martin Katz, *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909, 910 (2013) (arguing that prediction is a core component of a lawyer's services and suggesting that quantitative analysis of big data set can outperform the lawyers subjective predictions). *See generally* RICHARD SUSSKIND & DANIEL SUSSKIND, THE FUTURE OF THE PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS (2015) (analyzing how new technologies from telepresence to artificial intelligence will fundamentally change today's professions, particularly doctors, teachers, accountants, architects, consultants and lawyers).

actors to "coloniz[e] the web."[71] Safiya Umoja Noble, among others, has argued that search results on popular search engines reflect and perpetuate racists and sexists stereotypes.[72] Beyond simply providing access to problematic content, search results prioritize as well, drawing users attention to some content providers at the expense of others. Even seemingly neutral algorithms, when coupled with certain types of data, can generate search results at odds with broader social values of nondiscrimination.

More specifically, there has been a spate of recent scholarship focusing on the potential for machine learning to reinforce bias within the legal system.[73] The crux of the argument is that sophisticated machine learning algorithms operate as a "black box" that defies straightforward interpretation, making it difficult to understand how prediction is actually achieved in the model. This becomes a problem if some of these factors should not bear on a legal decision. For example, if machine learning approaches are used to predict criminal behavior for purposes of sentencing, then even if the underlying data is stripped of impermissible factors (such as race), the algorithm may nevertheless identify variables or combinations of variables (e.g., place residence, employment history, family status) that end up—intentionally or not—as proxies for the impermissible factor.[74] In at least some machine learning approaches, it may be extremely difficult to tease out whether this has occurred. There may be related concerns as law search gravitates toward natural language searchers that rely on opaque—indeed proprietary—algorithms that translate user queries into returns. These natural language searches may end up replicating some of the issues associated with the "categorization" of the law under West headnotes, without the transparency of the explicit American Digest System.

Law search technologies at least potentially bear on a wide range of social values, from how lawyers ought to engage in legal reasoning, to the automation of the legal profession, to discrimination and bias.

---

71. *See* Frank Pasquale, *Dominant Search Engines: An Essential Cultural & Political Facility*, *in* THE NEXT DIGITAL DECADE: ESSAYS ON THE FUTURE OF THE INTERNET 401, 402 (Berin Szoka & Adam Marcus eds., 2010).

72. *See generally* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018).

73. *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 671–72 (2016); *see also* Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 900 (2016).

74. *See* Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1267 (2020).

For some of these issues, there are already well fleshed out normative theories that can provide some guidance when evaluating the consequences of new technologies.[75] In other areas, such as the automation of legal services, there is considerable debate about basic normative questions—such as the relative importance of efficiency versus economic stability—but law search does not itself pose any distinctive normative challenges. But on certain matters, the normative stakes of law search technologies remain murky: Even assuming that electronic law search tends to drive attorneys toward more fact-intensive styles of legal argument that rely less on background principles, it is not clear whether that is a good or bad development. To help gain purchase on these questions, a theory is needed about the nature of law search and how it intersects with normatively important values, such as the rule of law or social welfare. In the following Part, we turn to the task of sketching out the contours of such a theory and mapping intersections with normative values.

II. SEARCH MATTERS

Although law search is central to legal reasoning and has been addressed by a limited number of scholars, it remains undertheorized.[76] In particular, the normative stakes of law search have not been properly articulated, which makes it difficult to evaluate seemingly important trends, such as the digitization of legal material or the move away from the American Digest System. In this Part, we begin by providing a more fleshed out definition of law search that is grounded in a descriptive understanding of legal relevance. Working with this definition, we go on to describe the relationship of law search technologies to convergence, which is an emergent property of a legal system that is defined by the degree of agreement over relevant legal materials. We then go on to discuss how the normative stakes of law search can be understood for social well-being and rule-of-law values.

75. *See generally* DEBORAH HELLMAN, WHEN IS DISCRIMINATION WRONG? (2008) (providing a general theory of the underlying justification for wrongfulness judgments concerning discrimination).

76. The lack of a sophisticated theoretical discourse of law search is particularly glaring when contrasted against the very robust scholarly conversation on some of the perennial questions in jurisprudence.

A. Relevance and Convergence

When someone sits down to search through the law, the starting point is typically some legal question, and the person engaged in law search is attempting to gain information that is relevant to that legal question. For example, a chemical manufacturer may be contemplating expanding production at a facility, which potentially opens a range of legal issues related to environmental permitting, workplace safety requirements, employment contracts or union bargaining agreements, and local land use limitations. The task for the "law searcher" is to identify the statutes, case law, regulations, and contracts that bear on the many legal questions that are raised by the contemplated expansion.

How the law searcher goes about this task will depend on the technologies of the day. Prior to the advent of digital commercial databases, law reporters and indices would have been the tools of choice for many lawyers. With more recent technologies, law searchers can take advantage of many different resources, perhaps toggling between keyword or natural language searches in a commercial database, querying free online search engines, clicking back through a commercial annotation system, examining secondary sources, following sequentially through a statutory or regulatory text, and examining citations and cross-references.

The purpose of search is to identify *relevant* legal information. Generally speaking, there are two approaches to defining legal relevance: a normative approach and a descriptive approach.[77] A normative account of relevance would attempt to define the types of authorities that a legal actor ought to consider when determining the content of the law. In Sections II.B and II.C, we will discuss some of the factors that might inform a normative account of relevance.

The normative *a priori* assumption of legitimate authority and reliance on that authority is in distinction to a more crowd-sourced, sociological view of relevance, which is prevalent in other areas. For example, in the subfield of computer science focused on search and information retrieval, human agreement on the relevance of information is used to benchmark the success of different systems.[78]

77. The distinction we draw between normative and descriptive relevance tracks similar distinctions in other contexts. For example, the term "legitimacy" can be understood in a normative or a descriptive manner. *See* Richard H. Fallon, Jr., *Legitimacy and the Constitution*, 118 HARV. L. REV. 1787, 1851 (2005).

78. *See* Jean Carletta, *Assessing Agreement on Classification Tasks: The Kappa Statistic*, 22 COMPUTATIONAL LINGUISTICS 249, 249 (1996). Researchers in

The information science scholar Tefko Saracevic argues for a multifaceted understanding of relevance that includes several types of subjective judgments that people make.[79]

Borrowing from these approaches, we define a *descriptive* notion of legal relevance as follows: *A document is (descriptively) legally relevant to a legal question when it is understood by the dominant legal community as containing information that bears on that legal question*. This definition is descriptive and is based entirely on judgments made by a legal community. Under this definition, those judgments make up the ground truth, and can be neither correct nor incorrect.[80] This definition is close to Saracevic's notion of "situational relevance," which has to do with the "[u]sefulness" of information for a particular purpose.[81] A document is legally relevant to a question when it can be used as the basis for legal argumentation and analysis. Relevance is determined functionally with respect to

---

information retrieval have developed a number of metrics for assessing the performance of systems based on subjective human judgments, such as precision (relevant retrieved document over all retrieved documents), recall (relevant retrieved documents over relevant documents), and f-score (a combined metric of precision and recall). *See* David C. Blair & M. E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMC'NS ACM 289, 289–90 (1985) (comparing full-text electronic databases against manually indexed databases based on criteria of precision and recall).

79. *See generally* Tefco Saracevic, *Relevance Reconsidered*, *in* INFORMATION SCIENCE: INTEGRATION IN PERSPECTIVES, PROCEEDINGS OF THE SECOND CONFERENCE ON CONCEPTIONS OF LIBRARY AND INFORMATION SCIENCE 201 (1996) (providing account of relevance). The different types of relevance identified by Saracevic include topical relevance (aboutness), cognitive relevance (informativeness and novelty), situational relevance (usefulness), and motivational relevance (satisfaction). Building on the work of Saracevic, Marc van Opijnen and Cristiana Santos have described a similar multi-dimensional relevance model specific to law. *See* Marc van Opijnen & Cristiana Santos, *On the Concept of Relevance in Legal Information Retrieval*, 25 A.I. & L. 65, 65 (2017).

80. Note that a search algorithm like the original Pagerank algorithm that undergirded the original Google platform conferred "authority" on those webpages (digital resources) that were effectively most likely to be landed upon by a searcher moving randomly (according to the link structure) through the space of pages relevant to a query. Jon M. Kleinberg's related "HITS" algorithm distinguished the quantitatively scored properties of "authority" and "hub" in a metric evaluation of each node. *See* Jon M. Kleinberg, *Authoritative Sources in a Hyperlinked Environment*, 46 J. ASS'N COMPUTING MACH. 604, 604 (1999).

81. *See* Saracevic, *supra* note 79, at 12.

norms and practices concerning legal reasoning and argumentation within a legal community.[82]

The descriptive understanding of legal relevance offered here is related to Oliver Wendell Holmes's predictive theory of the law, summed up by Holmes's declaration that "[t]he prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law."[83] *Mutatis mutandis*, we take "prophecies of [the materials] courts will [examine] in fact, and nothing more" to be the definition of legal relevance.

These "prophecies"—i.e., predictions—are both substantive and procedural. They are substantive in that the searcher predicts the judgements of others concerning whether a document bears in a substantive way on the legal question at hand. The predictions are procedural in the sense that there are norms, conventions, and practices concerning how search is conducted. A document that is difficult to locate using the law search methods of the dominant legal culture is less relevant precisely because it is unlikely to be found, and as a consequence is unlikely to be deemed relevant by other legal searchers. And so, the process of search as actually practiced has consequences for what is or is not relevant law.

The predictive theory has been criticized as falling short as a normative theory of the content of the law.[84] In particular, the predictive theory does not appear to provide a way to criticize the relevance judgments made by members of a legal community. This is a fair critique and we turn to normative questions below. But describing relevance as a sociological phenomenon can also be useful and is perhaps best understood as arising from H.L.A. Hart's "external" perspective of the social scientist or historian interested in understanding legal phenomena, rather than the "internal" perspective of bona fide legal actors.[85]

The cornerstone of the descriptive account of relevance is that legal actors make judgments about the types of legal materials that

82. *See* Stuart A. Sutton, *The Role of Attorney Mental Models of Law in Case Relevance Determinations: An Exploratory Analysis*, 45 J. AM. SOC'Y INFO. SCI. 186, 187 (1994).

83. *See* O. W. Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 461 (1897).

84. One concern is that the predictive theory also does not appear to give guidance to judges. A related claim is that it is incoherent to think of unearthing the law's content absent an exercise in normative reasoning.

85. *See* HART, *supra* note 18, at 89. For a critique of the distinction, see Charles L. Barzun, *Inside-Out: Beyond the Internal/External Distinction in Legal Scholarship*, 101 VA. L. REV. 1203, 1212 (2015).

others will find useful for answering legal questions. From this decentralized behavior, a collective judgment about legal relevance emerges through the behavior and beliefs of individual legal actors in the community. These relevance judgments can be understood as roughly analogous to the way that individual views about the definition of a word generate some collective sociological fact about definitions.

Although this process is mostly decentralized, there is some role for centralized coordination. Entities like Westlaw or LexisNexis and texts such as the American Law Institute's Restatements play a moderate coordinating role by standardizing search practices and providing focal points. But, at least for more sophisticated questions, legal searches are carried out in a decentralized fashion. To the extent that there are system-wide properties associated with law search those properties emerge from uncoordinated behavior of individuals.

One important macro-feature of a legal system that emerges from individual search behavior is the degree of agreement in legal relevance judgments, what we call *convergence*. When there is agreement about the relevant law for a given legal question or dispute, the parties will converge on some set of sources. When, in general, parties tend to agree on the law that is relevant in their disputes, then the level of convergence in that legal system is high.

The level of convergence in a legal system results from several interacting features. The law itself can be one. Legal systems that have a larger number of documents can expect, other things being equal, less convergence. The information systems used to organize the law and facilitate document retrieval are another factor. Some information systems may facilitate convergence by channeling searchers in particular ways, while others may facilitate idiosyncratic search patterns. The community of legal searchers, and the norms, practices, and conventions that structure their behavior, is another feature that might influence convergences. If informal practices are widely shared or well known, that may enhance convergence.

The institutional setting of courts and the role that texts play in legal argumentation also affects convergence. In an adversarial context, lawyers will seek out the authority that best supports their position, but they will also seek out the authority that best supports their opponent's case, so that they can prepare and develop counterarguments. Both sides will seek any authority that they believe a judge would find binding or persuasive. The judge, in carrying out (or having a clerk carry out) additional research, will search for an authority that a reviewing panel will find binding or persuasive. Even

a final reviewing court searches for authority in this predictive fashion, both to project influence down the judicial hierarchy (if lower courts better conform to precedent that they respect) and forward in time (assuming that appropriately defended decisions will be more resistant to change). Highest courts must also be mindful of their institutional legitimacy, which they maintain in part conforming to the expectations of a broader legal community about appropriate sources of legal authority.

To the extent that there are agreed-upon rules or conventions concerning what sources of authority are binding, they can have a coordinating effect for lawyers and judges who are involved in adversarial proceedings.[86] The duty to follow relevant authoritative sources creates a focal point for search, which will facilitate convergence. Strategic-predictive considerations also mitigate effects of individual biases if the adversarial context or judicial hierarchies punish lawyers or judges who fail to anticipate and respond to arguments offered by opponents.[87] Lawyers whose search is influenced by some individual bias will fail to find relevant legal authority and will, accordingly, weaken their chances of success.

From a normative perspective, convergence may appear to be a good to be maximized by a legal system. But this need not be the case.

---

86. *See* RICHARD H. MCADAMS, THE EXPRESSIVE POWERS OF LAW: THEORIES AND LIMITS 57 (2015) (suggesting that sustained cooperation depends on the existence of stable focal points that coordinate behavior among people with different moral and empirical views).

87. One source of bias might be the difficulty for the law searcher of separating the question of what the law is from what the law should be. Psychologists have discussed how normativity bias and confirmation bias can affect cognition and perception in many contexts. *See generally* HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT (Thomas Gilovich, Dale Griffin & Daniel Kahneman eds., 2002); Martin Jones & Robert Sugden, *Positive Confirmation Bias in the Acquisition of Information*, 50 THEORY & DECISION 59 (2001) (presenting strong evidence of confirmation bias in information acquisition and information use, and suggesting that the bias results from a pattern of reasoning which, although producing sub-optimal decisions, is internally coherent and self-reinforcing); Geoffrey L. Cohen, *Identity, Belief and Bias*, *in* IDEOLOGY, PSYCHOLOGY, AND LAW 385, 389 (Jon Hanson ed., 2012) (explaining that beliefs are tied to long-held identities that resist change and bias the processing of new legal information). Given that the law carries direct normative weight, it may be that these biases are particularly powerful in the law search context. Normative bias is distinct from the stronger view that it is incoherent to think of unearthing the law's content absent an exercise in normative reasoning. *See, e.g.*, RONALD DWORKIN, LAW'S EMPIRE 90–96 (1986); Lon L. Fuller, *Positivism and Fidelity to Law—A Reply to Professor Hart*, 71 HARV. L. REV. 630, 646 (1958); JOHN FINNIS, NATURAL LAW & NATURAL RIGHTS 18–19 (2d ed. 2011).

As will be discussed in the next two Sections, there may be tradeoffs posed by convergence so that some level of disagreement in legal relevance determinations may be a desirable feature of the legal system.

B. Efficiency

There is a variety of normative frameworks that could be used to develop a notion of legal relevance and evaluate law search practices and the level of convergence in a legal system. In this Section, we begin by taking the perspective of a social decision-maker concerned with overall well-being. This classic law and economics framework provides one very general approach for normative legal analysis. In the following Section, we broaden the scope to include rule-of-law values that may have independent moral force, beyond their interaction with individual well-being.

As a matter of efficiency, law search can be thought of as a transaction cost that, other things being equal, would be good to minimize. Technologies or practices that reduce search costs would allow valuable social resources to be allocated to other uses. The costs of law search can also propagate through the market system by adding costs to business transactions, which might result in deadweight loss if otherwise efficient transactions do not occur.

The costs of search can be compared to its benefits, which include convergence. Where law plays a coordinating function, social welfare is promoted when there is a high degree of convergence. If there are many different ideas about how contracts are properly formed or, for that matter, which side of the road to drive on, it will result in coordination costs as parties bicker over contract form or drivers get into accidents. A high degree of convergence will mean that most of the time, parties will at least agree on the relevant legal authorities that bear on some legal question. Where coordination is important, increasing that base level of agreement (i.e., increasing convergence) is useful.

Another value implicated by law search is accuracy. Imagine a society that has adopted an efficient set of rules, but there is some uncertainty about what rules apply to any given case. That uncertainty may result in the application of the wrong rule to a case, leading to undesirable outcomes. With accuracy in mind, it is possible to think of better and worse ways of engaging in law search. Practices or information systems that tend to favor retrieval and use of legal documents that result in accurate legal judgements are welfare

enhancing, while practices or systems that promote use of legal documents that lead to inaccurate legal judgments are welfare reducing.

This notion of accuracy provides one way to think of legal relevance in normative terms. One could define a document as relevant inasmuch as its use in legal analysis and argumentation tends to result in accurate legal judgements. If the legal rules in a system are themselves efficient, the use of relevant documents would be normatively desirable from an efficiency perspective. If the rules are not efficient, accurate judgments might not be welfare maximizing, and, therefore, the use of relevant documents might result in bad outcomes.

The costs and benefits of search generate tradeoffs. In a given system, more law search may reduce the risk of inaccuracy, but on a declining marginal basis. In such a system, the socially optimal amount of search will balance the benefits of accuracy against the costs of search, leading to some level of investment in search, and some level of residual inaccuracy. Thus, efficient search may leave potentially relevant documents undiscovered, which would lead to an irreducible overhang of inefficient rule-application to all law-dependent transactions. The same tradeoff can be stated in terms of search costs and convergence: Even if search tends to lead to convergence, and parties prefer more convergence to less convergence, they will still stop short of perfect convergence because search is costly.

There is also a potential for tradeoffs between accuracy and convergence. Imagine a typical legal system with a large number of laws, regulations, and cases, but where a universally adopted information system returned a single document in response to all queries. In such a system, there would be perfect convergence, because all legal actors would work from the same legal materials. But those materials (presumably) would be highly inaccurate, in that they would not generally aid in the formation of correct legal judgments.

For example, imagine all U.S. law searchers used the "DumbTech" legal search service, and only that service. For all queries, DumbTech returns a single statutory provision, say 16 U.S.C. § 167, a provision of the conservation law having to do with the sale of timber on government lands. For some legal matters, that document would aid legal decision-making, but for most, it would not. Nevertheless, if that same document was returned for all queries, convergence would be complete. Less extreme versions of the tradeoff

are also imaginable, whereby an information system or practice tended to channel searchers, but in the wrong direction.

Other features of a legal system can also interact with search, creating their own tradeoffs. For example, it may be welfare enhancing for a legal system to be relatively comprehensive, in the sense of stating a large number of explicit rules that cover many specific questions. But comprehensiveness interacts with search costs to reduce the amount of convergence. A simple legal system consisting of a single four-word text (say, "maximize aggregate social welfare") would lead to high levels of convergence, but would give little guidance and would lead to radical uncertainty and disagreement in the application of the (agreed-upon) legal text to individual cases.

There are also dynamic effects to take into consideration. Legal change over time occurs in part through the process of rule identification, explication, and application. Convergence over legal sources may result in a relatively information-poor environment because there is less contestation about foundational questions of the best law to apply to a given matter. When convergence is low, there is a broader range of potential legal rules that can be brought to bear on a given legal question. This larger field provides decision-makers with greater latitude to push the law in a desirable direction. On the other hand, if convergence is so low that decision-makers find themselves drowning in irrelevant information, that can serve as an alternative drag on their ability to enact desirable legal change.

Even within the relatively narrow scope of the law and economics framework, law search has a wide range of normative consequences. Search itself is costly and utilizes resources that could be devoted to other pursuits. But it also generates benefits as private parties are better able to coordinate their behavior, and legal decision-making is rendered more accurate as relevant legal authorities are considered. A simple legal system may facilitate search but may lack comprehensiveness. Too much agreement on the law may also make the law less flexible and less able to adapt to new circumstances. Collectively, these dynamics create important normative stakes for law search behavior and tradeoffs with substantial social consequences.

Given the importance of law search to legal reasoning, there are other potential normative frames that could be brought to bear, and in particular rule-of-law values, which create additional evaluative dimensions, which we turn to in the next Section.

C. Rule of Law

Concern for the rule of law suggests that the law should be clear, determinate, and predictable, and that the decisions of legal actors (and most importantly judges) be made in a law-like manner, rather than on the basis of personal whim or preference.[88] These rule-of-law values provide an alternative, non-welfarist basis to engage in normative reasoning about law search, legal relevance, and convergence.

In a legal system in which judges are bound to follow legal precedent, those judges would presumably have an associated obligation to seek out and find relevant law. A judge who does not find and follow relevant precedent that should have been found has (arguably) failed in a duty to be faithful to the law. Assuming that judges are under a duty to follow the relevant law, the natural question arises of how much effort judges are required to expend in order to identify that law. Perhaps it is sufficient to rely on the parties' briefs. But, perhaps there is an independent requirement to search out binding authority.

Presumably, the obligation to seek out and find relevant law is not absolute and must be balanced against other demands, such as the timely deciding of cases. Judges must decide how to trade off their practical limitations—including search costs—against the need to engage in law search. The obligation of judges also filters down to lawyers engaged in advocacy and (to a lesser extent) counseling. Clients and lawyers are free to bargain over the time that should be spent on search, but there may be minimal requirements concerning obligations to the court (in the advocacy context) or legal competence (in the counseling context).

Obligation to obey the law is also related to a normative understanding of legal relevance—a document can be considered relevant inasmuch as it provides useful information for legal actors toward understanding rights and duties. But, in some very broad sense, all legal documents could be understood as providing at least some useful information. Ronald Dworkin's famous thought experiment posits a "lawyer of superhuman skill, learning, patience and acumen" (Hercules) who is able to read and understand every extant legal

88. *See* TOM BINGHAM, THE RULE OF LAW 48 (2010); RAZ, *supra* note 3, at 210–33; BRIAN Z. TAMANAHA, ON THE RULE OF LAW: HISTORY, POLITICS, THEORY 119 (2004) ("The rule of law [in the sense of formal legality] entails public, prospective law, with the qualities of generality, equality of application, and certainty."); Frederick Schauer, *Formalism*, 97 YALE L.J. 509, 528 (1988).

document.[89] Hercules does not need to economize on search time or direct his energies to the most relevant law—every text can be considered and weighted in direct proportion to its bearing on a legal matter. But for legal actors with non-Herculean human-level skills, learning, patience, and acumen, only a relatively small number of legal texts can be considered. The limited nature of human attention raises the question of which legal documents a lawyer or judge is bound to consider (i.e., the legally relevant documents), and which can be left unexamined (i.e., the less relevant or irrelevant documents).

Rule-of-law values are also implicated beyond the context of individual decision-makers. At the system level, higher levels of convergence imply, *ceteris paribus*, greater predictability and determinacy. Comprehensiveness also has a rule-of-law cast, however, by constraining official discretion and reducing uncertainty.[90] But, as discussed in the prior Section, comprehensiveness is in tension with convergence. The more distinct types of subject matter that are regulated and the higher the specificity of those regulations, the harder it will be to achieve convergence on the same legal materials.

Debates over the relative merits of rules and standards can also be usefully informed by attention to convergence and law search.[91] For example, greater notice and predictability concerning the law's content may be theoretically facilitated by specific rules. But if those specific rules are unlikely to be found, they do little practical good. If so, a vague but findable standard may ultimately prove more desirable from the perspective of notice.

There is more that can be said about the relationship between rule-of-law values and law search, legal relevance, and convergence. This area is generally under-explored in jurisprudential theory. The preceding discussion provides only a cursory summary of some of the most obvious issues, and future work could develop these topics in considerably more detail. In Part IV, we return to these normative issues in the context of a focused discussion on the value of policymaking in the area of law search.

---

89. *See* RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 105 (1978).

90. *See generally* GERALD J. POSTEMA, BENTHAM AND THE COMMON LAW TRADITION (1986).

91. *See generally* Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557 (1992); Kathleen M. Sullivan, *Foreward: The Justices of Rules and Standards*, 106 HARV. L. REV. 22 (1992); Pierre Schlag, *Rules and Standards*, 33 UCLA L. REV. 379 (1985).

III. STUDYING LAW SEARCH

The disciplinary richness of contemporary social sciences and humanities has many advantages, but that richness can also create a challenge for organizing a coherent research program.[92] Although it may be easy enough to define law search as an object of study, different conceptual and disciplinary frameworks will generate different questions answered using differing methods.[93] An economist approaching the empirical phenomenon of law search might be concerned with the effect of incentives on the development of new search technologies; a psychologist, by contrast, may be interested in how the framing of legal questions affects search results. Scholars of both of these disciplines will tend to ask different questions and use tools different from those used by historians, sociologists, or anthropologists studying the same general phenomenon.

The empirical approach that we have pursued in most detail involves constructing a computational model of law search, which we believe provides one useful lens on law search as an empirical phenomenon. In this Part, we begin by situating our approach through a discussion of the data that can be gathered on law search and the broad categories of research questions that have arisen in the study of search. Next, we argue that computational modeling can provide useful insights into the phenomenon of law search and we discuss trends in law scholarship that relate to the work we describe. Finally, we describe *LexQuery*, a computational algorithm for how law searchers navigate through a corpus of legal documents.

92. A sustained and cumulative scholarly project requires what Joan Fujimura has referred to as a "standardized package." *See generally* Joan H. Fujimura, *Crafting Science: Standardized Packages, Boundary Objects, and "Translation*," *in* SCIENCE AS PRACTICE AND CULTURE 168 (Andrew Pickering ed., 1992); Joan H. Fujimura, *The Molecular Biological Bandwagon in Cancer Research: Where Social Worlds Meet*, 35 SOC. PROBS. 261 (1988). This package includes a theoretical framework capable of generating well defined research questions as well as methods that can provide answers to those questions within known and accepted epistemic parameters. Developing a fully articulated standardized package for the study of law search is beyond the scope of this Article and is likely to involve some trial and error as well as an ongoing interdisciplinary discourse.

93. Search, as a general phenomenon, has been analyzed by many disciplines. *See supra* note 14 and accompanying text.

A. Empirical Study of Search

Data is a useful starting place to begin a discussion of empirical research. In the domain of law search, the best types of data are likely to derive from direct observation of search behavior. Regardless of the kind of question asked, data that traces how individuals engage in law search will often be capable of producing useful insights.

Prior to the introduction of digital search databases, search activity was extremely difficult to observe. Now, however, the major commercial legal databases collect massive amounts of data about how their users go about the business of search. These firms can use this information, but it is not generally available to researchers. Although unfortunate from a scholarly perspective, this practice is understandable given the information's proprietary nature and high value.

Even without the ability to directly observe search behavior, there are other sources of data that can provide insight into how search is carried out. One particularly important source is the law itself. Legal documents are the object of search (i.e., what is sought out) and are useful to study for that reason alone. In addition, their content and structure also influence how search is carried out. One searches through a collection of statutes (which are organized hierarchically by subject matter) differently from how one searches through judicial opinions (which are organized chronologically but have dense citations to other subject-related documents).[94] And search, in turn, affects the documents in the corpus, as authors make choices of content and citation based in part on the results of prior law searches, carried out by themselves or others. In this way, legal documents provide a source of data both on the effects of law search (in terms of citation) as well as corpus features that affect how search is carried out.

Once search-related data is collected, there are many different conceptual frameworks that could be brought to bear on its analysis. One convenient way to categorize these different approaches is the contrast between research into information retrieval and research on information seeking behavior.

Computer scientist Calvin Mooers coined the phrase "information retrieval" as "[t]he problem of directing a user to stored

94. Daniel N. Rockmore, Keith Carlson, Faraz Dadgostari & Michael A. Livermore, *A Multinetwork and Machine Learning Examination of Structure and Content in the United States Code*, 8 FRONTIERS PHYSICS 625241 (2021).

information, some of which may be unknown to him [or her]."[95] Since that time, a substantial research program in computer science has developed to analyze and respond to the information retrieval problem, especially in the context of unstructured corpora of text documents.[96] The advent of the World Wide Web and other new collections of digitized documents boosted the field with both research and commercial opportunities. Information retrieval intersects with the field of natural language processing and computational text analysis more generally because statistical representations of language are fundamental to effective text-oriented information retrieval systems.[97] Statistical modeling of documents, such as tf-idf, helped demonstrate the usefulness of mathematical representations of documents for information retrieval purposes.[98]

Researchers in the field of information behavior take a user-centric approach and focus on how people approach the problem of information search across media.[99] Early work emphasized the personal and social factors that give rise to the need for information.[100] Subsequently, scholars have focused on the iterative nature of search (where early results influence subsequent searches), the complexity of the roles and tasks that motivate search, and the importance of environmental factors that can facilitate or inhibit successful search.[101] Researchers have also focused on the cognitive and psychological aspects of search.[102] In recent years, search behavior has been

95. *See* Mooers, *supra* note 15, at 3; Calvin Mooers, *Information Retrieval Viewed as Temporal Signalling*, 1 PROCS. INT'L CONG. MATHEMATICIANS 572, 572 (1950).

96. *See generally* MANNING, RAGHAVAN & SCHÜTZE, *supra* note 15.

97. *See generally* Alan F. Smeaton, *Using NLP or NLP Resources for Information Retrieval Tasks*, *in* NATURAL LANGUAGE INFORMATION RETRIEVAL 99 (Tomek Strzalkowski ed., 1999).

98. *See generally* Luhn, *supra* note 37; Jones, *supra* note 37.

99. *See* THEORIES OF INFORMATION BEHAVIOR, *supra* note 14.

100. *See generally* T. D. Wilson, *On User Studies and Information Needs*, 37 J. DOCUMENTATION 3 (1981).

101. *See generally* Gloria J. Leckie, Karen E. Pettigrew & Christian Sylvain, *Modeling the Information Seeking of Professionals: A General Model Derived from Research on Engineers, Health Care Professionals, and Lawyers*, 66 LIBR. Q. 161 (1996).

102. *See generally* Lokman I. Meho & Helen R. Tibbo, *Modeling the Information-Seeking Behavior of Social Scientists: Ellis's Study Revisited*, 54 J. AM. SOC'Y INFO. SCI. & TECH. 570 (2003).

examined across a range of user categories, including social and natural scientists,[103] doctors,[104] engineers,[105] and lawyers.[106]

To summarize the distinction between information retrieval and information behavior: The former can be understood as a technical research program with the goal of developing optimized methods for search. The field most closely associated with this research program is computer science, and sub-questions involve natural language processing and extracting latent structure from unorganized collections of documents. Research into information behavior, by contrast, focuses on how people actually engage in search tasks, their mental processes and goals, and how they make relevance determinations. Disciplines such as economics or psychology can take information seeking behavior as an object of study, and the field of information studies is in part focused on these questions. These disciplines will typically engage in research by constructing either formal or informal behavioral models and then using data generated naturally or through experimentation to test and refine those models. Humanities disciplines that engage in empirical work—such as historians or anthropologists—can also engage in research on information seeking behavior, using tools such as surveys, interviews, or archival research.

The empirical study of law search can be undertaken through both information retrieval and information behavior perspectives. Search for the law has much in common with other search contexts, but there are also important differences. One important category of differences is normative, and in particular, the importance of rule-of-law values for defining legal relevance and evaluating law search practices and related systemic macro-phenomena, such as convergence. There are empirical differences as well, including the institutional setting, the incentives and stakes for the parties, the costs, the role of legal education in shaping practice, and many others. Understanding and mapping the similarities and differences between

---

103. *See generally* Bradley M. Hemminger et al., *Information Seeking Behavior of Academic Scientists*, 58 J. AM. SOC'Y INFO. SCI. & TECH. 2205 (2007).

104. *See* Karen Davies, *The Information-Seeking Behaviour of Doctors: A Review of the Evidence*, 24 HEALTH INFO. & LIBRS. J. 78, 78 (2007).

105. *See* Mark. A. Robinson, *An Empirical Analysis of Engineers' Information Behaviors*, 61 J. AM. SOC'Y INFO. SCI. & TECH. 640, 640 (2010).

106. *See generally* C.C. Kuhlthau & S.L. Tama, *Information Search Process of Lawyers: A Call for 'Just for Me' Information Services*, 57 J. DOCUMENTATION 25 (2001); Margaret Ann Wilkinson, *Information Sources Used by Lawyers in Problem-Solving: An Empirical Exploration*, 23 LIBR. & INFO. SCI. RSCH. 257 (2001).

law search and other forms of search is an important task for future research.

B. Modeling Law Search

The approach to studying law search that we will describe in the following Section draws from the fields of both information retrieval and information behavior and can best be described as falling within a new wave of scholarship referred to as computational social science. In recent years, several technological developments—including an explosion in data collection and computational processing power—and new mathematical and statistical techniques to take advantage of these developments, have led to the emergence of new lines of research in many social science disciplines.[107] This research is grounded in traditional social science methods that have been scaled up and adapted to take advantage of new data or technologies.[108] Examples include social network modeling and computer simulation. The shared methodological affinities in these new approaches have led some commentators to group them together as a new field.[109] Although somewhat nebulous, what holds computational social science together as a category is a focus on algorithmic (i.e., computational) tools that are used to either analyze massive data sets or to engage in large scale simulation experiments related to social, human-centered, phenomena.

A computational social science approach to the study of law search sits at the intersection of research on information behavior and information retrieval. The approach described below begins with insights into information behavior, which is then used to construct formal agent-based models of law search. These agent-based models are essentially stylized representations of how people engage in the activity of law search. The "agents" are effectively data points with attributes that interact according to a specified and possibly evolving rule set. These behavioral models can then be combined with methods from the field of information retrieval, such as metrics to evaluate the performance of recommendation systems. Retrieval systems that are

107. *See generally* David Lazer et al., *Computational Social Science*, 323 SCIENCE 721 (2009).

108. *See id.* at 722; *see also* JOHN H. MILLER & SCOTT E. PAGE, COMPLEX ADAPTIVE SYSTEMS: AN INTRODUCTION TO COMPUTATIONAL MODELS OF SOCIAL LIFE 4 (Simon A. Levin & Steven H. Strogatz eds., 2007).

109. *See generally* COMPUTATIONAL SOCIAL SCIENCE: DISCOVERY AND PREDICTION (R. Michael Alvarez et al. eds., 2016).

constructed based on different agent-based models can be compared against each other according to these metrics to determine which better conform to real world or experimental data. As these formal models improve, they will ultimately come to better represent the underlying behavioral phenomenon of law search.

The computational social science framing of law search also resonates with two other threads of law scholarship: a longstanding research program on "artificial intelligence and law," and the more recent growth of "law as data" scholarship. The field of artificial intelligence and law started during the 1980s with a group of researchers interested in using computational tools to represent the law as a set of executable code.[110] Much of this work can be described as GOFAI-style knowledge representation. In the context of law, knowledge representation could be used to translate statutes or case law rules into a machine-native language.

An alternative approach to applying computational methods to the law creates a framework of "law as data."[111] This approach draws from related advances in several fields that use computational tools to analyze text, including the move to "text as data" in the social sciences, and "distant reading" in the humanities.[112] Law as data research builds on multiple traditions in legal scholarship including quantitative empirical legal studies as well as the qualitative close reading and interpretation of legal texts.[113] Researchers in the field of law as data have tackled a wide range of questions, from the role of corporate opportunity waivers in influencing firm performance, to the growing divergence between opinions issued by the U.S. Supreme Court and U.S. appellate courts.[114]

---

110. *See generally* Trevor Bench-Capon et al., *A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law*, 20 A.I. & L. 215 (2012); *see also* Edwina L. Rissland, Kevin D. Ashley & R.P. Loui, *AI and Law: A Fruitful Synergy*, 150 A.I. 1, 7 (2003).

111. *See generally* LAW AS DATA: COMPUTATION, TEXT, AND THE FUTURE OF LEGAL ANALYSIS (Michael A. Livermore & Daniel N. Rockmore eds., 2019).

112. *See* Justin Grimmer & Brandon M. Stewart, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, 21 POL. ANALYSIS 267, 272–73 (2013); FRANCO MORETTI, DISTANT READING 48–49 (Verso ed., 2013).

113. *See* Jens Frankenreiter & Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANN. REV. L. & SOC. SCI. 39, 41 (2020).

114. *See* Gabriel Rauterberg & Eric Talley, *Contracting Out of the Fiduciary Duty of Loyalty: An Empirical Analysis of Corporate Opportunity Waivers*, 117 COLUM. L. REV. 1075, 1076–77 (2017); Livermore, Riddell & Rockmore, *supra* note 45, at 862.

The *LexQuery* model discussed in the next Section spans the divide between the fields of artificial intelligence and law and law as data. The computational model of law search contributes to the broader project in artificial intelligence and law of "understand[ing] and model[ing] legal argument," given the centrality of law search to legal argumentation.[115] Legal texts are also used as data in the construction and validation of the model, and so law as data is also central to the project.

C. LexQuery

The *LexQuery* model described in this Section provides a parsimonious description of law search that, along with data generated by a corpus of legal opinions, can be used for the purpose of simulation and to make predictions about human search behavior. For some purposes, accurate prediction is enough: For example, if an algorithm could cheaply and quickly generate results that are equivalent to those of a trained professional, highly skilled human capital could be reallocated to other tasks. A well-calibrated model can also provide insights into structural features of the law, such as the amount of search-informing information that is encoded in judicial citations.

Computational modeling also has significant limitations that should be recognized. In particular, it is important to keep in mind the distinction between prediction and causal inference, which bears on the ability to draw conclusions concerning interventions and counterfactuals. As just one example, predictions based on semantic similarity (as captured in the model) may, in fact, be a consequence of how human searchers rely on the ALI's Restatements. If changes to a Restatement (which would not be observed in our model) *cause* human searchers to change their behavior, there would be a loss of predictive accuracy of the model. Because our approach is not meant

115. Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 YALE L.J. 1957, 1957 (1990). In a classic formulation of the research agenda for the field, Edwina L. Rissland laid out a "unifying theme" of "understand[ing] and model[ing] legal argument." Rissland described three steps toward achieving this goal: first, "represent[ing] several types of knowledge, such as cases, rules, and arguments;" second, understanding "how to reason with [this legal knowledge], such as to manipulate precedents, to apply and make inferences with rules, and to tailor arguments to facts;" and third, understanding "how to use [this legal knowledge] ultimately in a computer program that can perform tasks in legal reasoning and argumentation, such as analogizing favorable cases and distinguishing contrary ones, anticipating parries in adversarial argument, and creating artful hypotheticals." *Id*.

to capture the full universe of causal variables and relationships, the accuracy of out-of-sample predictions is contingent on the stability of the data-generating process, a fact which may be difficult to confirm.

LexQuery has two general components: a search space and search strategies.[116] The search space is a mathematical representation of the corpus of documents, which we refer to as the *legal landscape*. Law searchers traverse this landscape by navigating between documents, relying on one or another formalized search strategy. The following discussion provides more detail on these concepts and reports the results of several validation tests performed on the *LexQuery* model on different law corpora.

1. *The Legal Landscape*

Broadly, *LexQuery* codifies the act of navigation through a corpus of documents. Navigation focuses on the portion of search in which a searcher actively moves from one document to another within a corpus. In its current form, *LexQuery* does not include any modeling of the initial query that leads a searcher to a starting place within the corpus. Rather, it takes a starting place—what we call a *source document*—as a given. Navigation focuses on the steps of the search process that occur after a source document is identified.

The space of documents through which the searcher is navigating (i.e., the search space or legal landscape) is represented as a *network* in which documents are *nodes* with *edges* between them.[117] Other terminology common in the field refers to *vertices* rather than nodes, and *links* rather than edges.

There are two types of edges between documents. One type is based on cross-reference information. For judicial opinions, that means that there are edges between two documents whenever one of them cites to the others. This structure is grounded in the qualitative observation discussed above that searchers often use citations as one way to identify documents of interest. A second set of edges between

116. More detail on the *LexQuery* model can be found in two technical papers. *See generally* Faraz Dadgostari et al., *Modeling Law Search as Prediction*, 29 A.I. & L. 3 (2020); Greg Leibon et al., *Bending the Law: Geometric Tools for Quantifying Influence in the Multinetwork of Legal Opinions*, 26 A.I. & L. 145 (2018). We draw liberally from those papers in the following discussion. This Section also reports results from Rockmore, Carlson, Dadgostari & Livermore *supra* note 94. Readers interested in greater detail on these models are referred to these papers.

117. The search space is more accurately categorized as a "multinetwork" because it is based on two network structures: the one derived from citation and the network structured based on textual similarity.

documents is constructed based on their semantic content (i.e., the words contained in those documents). Semantic content can be understood as a proxy for several different actual search mechanisms used by law searchers, including keyword searches, curated categorizations (i.e., Westlaw headnotes), and other sources (such as treatises). The operating assumption is that documents with similar words will show up together in keyword searches, be grouped under similar headnotes, and appear in the same treatises.

There are many ways to represent semantic content, and there is a balance that needs to be struck between completeness, level of coarse-graining, and computational costs. We opt for a *topic model* representation. The use of topic models has become widespread in a broad range of academic disciplines interested in texts (e.g., political science and digital humanities).[118] Topic models are based on term frequency vector representations of documents (i.e., "bag-of-words" representations), which are effectively lists of word frequencies or proportions, indexed by a vocabulary of words.

The highest possible dimensionality would effectively assign a dimension to the word type and position of each word in the document, leading to a large explosion of dimensions that would prove computationally intractable. By ignoring word order, bag-of-words representations achieve substantial dimension reduction from this baseline.

Topic models reduce dimension even further from the vocabulary size to a relative handful of dimensions equal to the number of topics. Very roughly, topic models use word co-occurrence to construct subject matter categories represented as distributions over the vocabulary (these are the "topics"). Documents are then represented as distributions over topics. Prior research has shown that topic model representations of judicial opinions retain a considerable amount of the original data found in a full-term frequency vector.[119] Thus, topic models achieve some level of coarse-graining, which can reduce the influence of idiosyncratic language.

2. *Search Strategies*

The search space is a collection of documents connected as some weighted combination of citation connectivity (reflecting whether one document cites the other or vice versa) as well as degree of similarity

118. *See, e.g.*, *supra* note 47 and accompanying text.
119. *See generally* Livermore, Riddell & Rockmore, *supra* note 45.

between their topic composition (measured with respect to the probability distributions representing their respective content). These two forms of connectivity combine to produce a distance measure on the document space.[120] Searchers navigate between documents based on those connections according to different abstract search strategies that are meant to capture the behavioral and cognitive nature of law search as a sequential decision process. In addition to searching for relevant legal authority, searchers may take other information into account, such as whether a judicial opinion was written by a well-respected jurist or is widely recognized as persuasive. A general notion of *quality* can capture those non-content-based attributes of a document. In *LexQuery*, the two proxies for quality are an impact factor measure based on citation history and a temporal factor in which more recent decisions are favored. The search strategies that are described below operate in common same search spaces, but users navigate through it differently, resulting in different search outcomes.

LexQuery is tested using three different search strategies. By way of analogy, imagine a robot that is programmed to navigate through some physical space, say for purposes of cleaning the floors. The space is the floor plan and the strategies determine how the robot navigates it. The robot could move about using different strategies—one might be to cover all of the ground that is close to its starting position and move outward; another might just randomly set off in a given direction and periodically make ninety degree turns; another might clean in a single location for a while and then periodically move to a distant portion of the space.

In the case of *LexQuery*, we have an abstract space of documents and search strategies. The first strategy is the *proximity strategy*. Beginning with a source document as a starting place, the proximity algorithm simply picks up all of the documents that are closest to the source document within the space. An important feature of the proximity algorithm is that its results are very closely related to the search space itself, as the results that are generated using this strategy directly reflect proximity as represented in the search space.

The second search strategy is the *covering strategy*. The covering algorithm is meant to capture the fact that there are often multiple legal issues within a single document. The covering strategy begins with a source document and then identifies the most proximate document. Then, based on some fixed parameters reflecting the

120. *See supra* note 116 and accompanying text.

various legal issues relevant to the opinion, it determines whether to continue navigating along that line of documents or to return to the source document and begin the search again along a different line. The idea is that there can be multiple legal issues present in a document, and once a searcher is satisfied with the results on one issue, it may make sense to explore a second or third legal issue, rather than continuing to collect documents on the first. Embedded within the covering algorithm is an assumption about how to make the tradeoff between depth (i.e., exploring one issue in more detail) and breadth (i.e., exploring a larger number of issues). This tradeoff is expressed in terms of a set of parameter values concerning the number of issues to explore and how deeply to explore them.[121]

The final search strategy that we introduce is the *adaptive strategy*. This strategy is akin to the covering strategy, but rather than using defined values for the breadth–depth tradeoff, the parameters are learned from the corpus, using a *reinforcement learning* program. The reinforcement learner uses the documents in the corpus and the citations included in those documents as data for a training procedure in which parameter values that correctly predict citation are reinforced. One can think of the adaptive algorithm as akin to a law student who learns the types of cases to identify (and cite) by studying the cases that have been identified (and cited) by the experts who produced the existing stock of documents in the corpus.

Each of these strategies takes different approaches to capturing features of how law search is carried out. Because they must be formalized and converted into an executable program, as models they are by nature simplified representations of the complex, idiosyncratic, and stochastic human search process processes. Nevertheless, they capture many important features of law search: the relevance of semantic content, guidance via citations, characteristics of document quality, and the tradeoff between depth and breadth. In practice, human legal researchers rely on a variety of sources of information not explicitly represented in the models, including their background understanding of the relevant law or secondary sources, such as treatises or the ALI's Restatements. But, a good deal of this "out-of-model" information may be proxied in features that can be extracted from the documents and therefore can be, even if loosely, captured by the models.

121. We set these parameters based on initial data exploration to improve the performance of the model.

3. *Results*

We interrogate the *LexQuery* model using three different legal corpora and two different procedures. The corpora are: (1) opinions issued by the U.S. Supreme Court, (2) Supreme Court opinions combined with opinions issued by U.S. appellate courts, and (3) the statutes of the U.S. Code.

The first measure is a *citation prediction task*. For this task, we remove the citation information from a document and then use the semantic information only (as represented via the topic model) to predict the citations in that document. We use two measures to evaluate performance. The first is *recall*, which is defined as the fraction of citations correctly predicted by the strategy. The second measure is *precision*, which is the fraction of predicted citations that are correct. Recall and precision are standard measures used in evaluating predictive algorithms.[122]

Table 1: *LexQuery* Citation Prediction

	Precision@10	Precision@20	Recall@10	Recall@20
SCOTUS				
Proximity	13%	10%	3%	7%
Covering	16%	18%	5%	13%
Adaptive	19%	18%	7%	15%
Combined				
Proximity	35%	19%	11%	12%
Covering	38%	20%	12%	13%
Adaptive	48%	31%	14%	18%
U.S.C.				
Proximity	7.5%	3.8%	3%	3%
Covering	16.5%	11.2%	6.2%	8.5%

122. Both of these measures can be estimated against a given number of predicted citations, denoted "@N" where N is the number of predictions. "Recall@10" reports the recall estimate based on the top ten predictions.

Table 1 reports the performance of *LexQuery* on the citation prediction task for the various corpora and strategies.[123] For purposes of comparison, the precision and recall estimates based on ten results are illustrative. For the U.S. Supreme Court opinions, in a group of ten results, the expected number of matches from the proximity algorithm to actual citation is roughly 1.3, the covering algorithm will generate 1.6, and the adaptive algorithm will generate roughly 2. This performance is an order of magnitude better than random choice. The performance measures are somewhat better for the combined corpus of U.S. Supreme Court opinions and appellate court opinions, in part because appellate court opinions have fewer citations to predict. The U.S.C. presented a more difficult prediction challenge because citation (represented as cross-references between sections) is much more sparse. The data for the U.S.C. was insufficient to train the learning algorithm for the adaptive strategy, and so only the proximity and covering results are presented.

The second validation measure, which is only carried out for the U.S. Supreme Court corpus, compares the outputs of the *LexQuery* model to *human relatedness prediction*. A group of research assistants was given a list of ten randomly selected Supreme Court opinions and asked to return ten "similar" opinions. The lists generated by the research assistants were compared to the opinions generated by the *LexQuery* model using the prompt opinions as the source documents. The validation measure compares the degree of overlap between *LexQuery* and the research assistants with the degree of overlap among the research assistants.

123. These reported estimates are averages over several clusters in the data. For more detail, see generally Dadgostari et al., *supra* note 116. The expected number of matches in a pool of ten predicted citations is the precision rate times ten.

Figure 1: *LexQuery* Relatedness Prediction

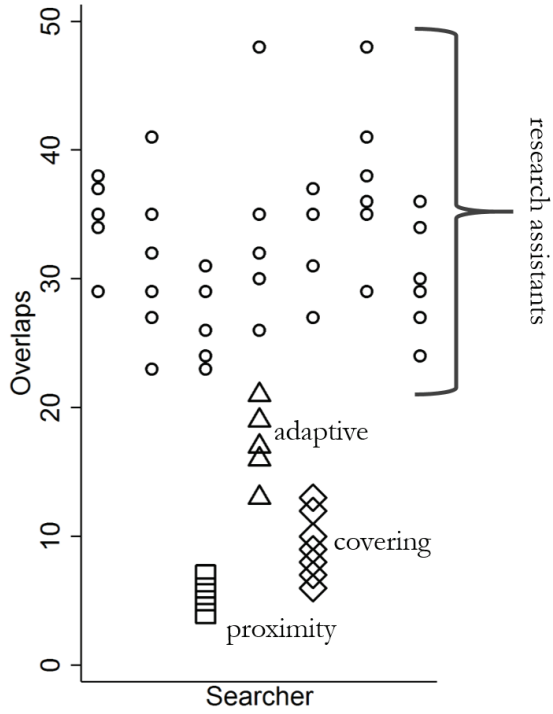

Figure 1 reports the performance of *LexQuery* on the relatedness prediction task. The search results for each research assistant were compared to each other to determine the degree of overlaps. Each two research assistant pairs generated a number of overlaps, generally between twenty and fifty (There was a total of one hundred possible overlaps.). In Figure 1, the cross-research assistant overlaps are represented by circles. The search results of each of the search strategies were also compared to the research assistants, with an overlap generated for each research assistant-model pair. The adaptive strategy performed the best, with an average of seventeen overlaps. The covering strategy averaged nine overlaps and the proximity strategy averaged six.

4. *Future Directions*

The *LexQuery* model provides one way to represent computationally the law search task, and its performance against the benchmarking tasks of citation prediction and relatedness prediction

indicate that the model captures important features of how law search is carried out. From an information behavior perspective, the *LexQuery* model can provide insights into the nature of law search, for example by estimating how law searchers made tradeoffs between the breadth and depth of search inquires. The performance of *LexQuery* could also be used to estimate the ease of search for different corpora—for example, one way to interpret the lower performance of *LexQuery* for the U.S.C. is that search is more difficult for statutes than an equivalent corpus of judicial opinions.

The learning approach used in the adaptive strategy is also ripe for improvement in future iterations. The citation prediction benchmark creates an easy measure of performance, and recent years have seen the performance of machine learning algorithms grow exponentially for similar tasks, such as machine translation and text and voice recognition. Contemporary flexible machine learning algorithms, which include deep neural networks, perform best when there is a very large number of observations that can be used to train these models. Law search is a context where such "big data" exists, in the form of large numbers of citation-bearing legal texts (including opinions and statutes, as well as briefs and similar secondary documents).

In the final Section of Part IV, we will discuss how tools like *LexQuery* could form the backbone of a public policy initiative that is designed to facilitate public access to legal knowledge. Before discussing that proposal in greater depth, we turn to the question of whether and how public policy intervention in the area of law search might be justified.

IV. SEARCH POLICY

As discussed in Part II, law search has important normative consequences, both from the perspective of social welfare and for rule-of-law-values. The natural operation of the market for legal services may not always result in an efficient tradeoff between competing values, and individual decisions may not always well-reflect society's overall interest in protecting the rule of law. Given these realities, there may be a role for policymakers to improve outcomes. This Part begins by exploring the ways that private markets might fail to protect some of the values implicated by law search. We then explore some of the potential downsides associated with direct regulation of the market for law search. Finally, we argue in favor of government support for the production of public access law search tools—a "public

option" for law search. This intervention avoids some of the downsides associated with direct regulation while nonetheless promoting values that are inadequately protected by private markets.

A. Limitations of Private Markets

There is a robust private market for law search, most recently evidenced by the significant amount of technological innovation that has occurred in this sector in the past several years. Nevertheless, there are reasons to think that private markets alone will not arrive at an efficient tradeoff between the various costs and benefits of law search and may well ignore or under-protect rule-of-law values. This Section briefly describes some of the potential limitations of the marketplace in the context of law search.

1. *Externalities*

Externalities are perhaps the archetypical example of market failure. Externalities occur when transactions in the marketplace have effects on third parties that are not accounted for by the transacting parties.[124] Externalities can be positive or negative, in that the effects on third parties can be harmful (as is the case of pollution) or could provide a benefit (such as private provision of habitat for beneficial species).

Law search may create positive externalities because a large number of social interests can be affected by legal decision-making—not only the parties with the most direct stakes. To take a stylized example, imagine that there is a social interest in avoiding evictions due to declines in neighboring home values, disruptions to children's schooling, and similar costs. Accordingly, there are various mandatory contractual terms involving matters such as notice and grace periods that are intended to favor renter continuity. Assume that these terms are efficient. Landlords and renters both have some incentive to understand these terms so that they can structure their behavior prior to or during a dispute. But these private incentives will not necessarily align with the overall public interest in avoiding evictions—even a party who wishes to fight an eviction order will not adequately invest

124. *See generally* Michael A. Livermore & Richard L. Revesz, *Environmental Law and Economics*, *in* THE OXFORD HANDBOOK OF LAW AND ECONOMICS 509, 511 (Francesco Parisi ed., 2017). Specifically, this definition is for "real" externalities, which lead to market failures.

in understanding his or her legal rights if there are social values at stake that are not internalized by that party.

A related pathway for externalities associated with law search involves legal regimes that attempt to internalize other externalities. Pollution control laws are an example. The possibility of insolvency or inadequate penalties could lead to under-investment in law search by parties that are charged with legal duties under the pollution control regime. The cost of legal compliance can be thought of as having two components: *primary costs*, which include investments or behavioral changes to conform to legal requirements, and *information costs*, which include law search and analysis of the relevant law. We can assume that the risk of a penalty declines in both primary and information costs (i.e., as parties understand the law better and do more to comply). Parties will continue to spend on both compliance and understanding up to the point where the marginal costs of doing so equals the marginal benefits (in terms of reduced exposure to the risk of penalty). If the combination of the probability of detection and penalties accurately reflect the social costs of noncompliance, then private parties will incur the optimal costs. But parties can be partially shielded from the full negative consequences of their action due to the possibility of insolvency, and society can under-penalize harms or invest too little in detecting law violations. In such cases, parties will spend less than they should to understand and comply with their legal obligations.

A final type of prevalent externality in the law search context is in the development of the law. The public resolution of disputes in the court system generates a broad social benefit through the articulation of legal norms. Private parties have incentives to invest in law search in ways that promote their cases, but they do not take account of the social interest in legal interpretation and articulation when making that investment.

Generally, the positive externalities associated with law search imply that, from the perspective of efficiency, too little is invested in this activity. The classic response to a positive externality is a subsidy, which helps "internalize" the positive social benefits of the underlying activity. In the context of legal development, one existing subsidy comes in the form of resources for judges and judicial staff to carry out their own law search, above and beyond the description of the law in parties' briefs. This subsidy is one way to help correct for the positive externality generated by private investments in law search in the course of litigation.

2. *Network Effects*

A second type of limitation of the private market for large search arises out of "network effects" or "network externalities."[125] Network effects exist when additional users increase the value of a shared resource.[126] Telecommunications systems provide classic examples network effects: As each additional person connects to the telephone network or Internet, the value of those systems for other users increases. Systems with network effects have a natural tendency toward monopoly.[127]

The importance of search to contemporary commerce has led to some focus on the potential for network effects to contribute to monopoly rents for dominant search providers. One source of such network effects is through a feedback loop in which first, the best performing search engine attracts the most users; second, the large user base attracts advertisers; third, advertising revenue is used to improve search technology, which attracts even more users. A second network effect that is relevant for search providers is data. As a provider collects data on search patterns, search results, and user preferences, it can use this data to improve performance. The performance boost from data is especially important as new machine learning algorithms, predictive analytics, and artificial intelligence tools have become available.[128]

---

125. *See* Michael L. Katz & Carl Shapiro, *Systems Competition and Network Effects*, 8 J. ECON. PERSPS. 93, 94 (1994) (internal quotation marks omitted).

126. *See* Mark A. Lemley & David McGowan, *Legal Implications of Network Economic Effects*, 86 CALIF. L. REV. 479, 495 (1998) ("Network effects are demand-side effects."). Network effects are akin to club goods, but with negative marginal costs per user. Club goods are nonrival, in the sense that adding an additional user imposes no marginal costs on existing users. An example is an uncrowded subway, where an additional user causes no additional cost. Network effects occur when there is value to adding an additional user. *See* RICHARD CORNES & TODD SANDLER, THE THEORY OF EXTERNALITIES, PUBLIC GOODS, AND CLUB GOODS 11 (2d ed. 1996).

127. *See* Michael L. Katz & Carl Shapiro, *Network Externalities, Competition, and Compatibility*, 75 AM. ECON. REV. 424, 425 (1985) ("[C]onsumption externalities [i.e., network effects] give rise to demand-side economies of scale, which will vary with consumer expectations. . . . [I]f consumers expect a seller to be dominant, then . . . it will, in fact, be dominant.").

128. *See* Robert Wayne Gregory et al., *The Role of Artificial Intelligence and Data Network Effect for Creating User Value*, ACAD. MGMT. REV. (forthcoming Mar. 2020) (manuscript at 4) (on file with authors) (providing, as examples, Google's web search algorithm and Tesla's self-driving car algorithm, both of which improve with data from their users).

Network effects are relevant for law search. Well-performing law search tools that attract users who provide revenue (typically through subscription fees), which can be used to fund innovation to improve those tools, set off a feedback loop of expanded user base. Law search engines can also collect data on their users, search inquiries, and the performance of their search tools. These data can be used to improve performance, again with this effect potentially boosted through the use of new algorithmic tools. In addition to these general search-related bases for network effects, law searchers may be particularly risk-averse about failing to find authority that could be used by potential opponents in negotiations or litigation. Accordingly, if a legal searcher anticipated that a counterparty is likely to use a particular information system—for example the American Digest System, or one of the large commercial databases—then there will be value in using that information system as well, at the very least as a supplement to whatever other tools would otherwise be used. The larger the number of legal agents who use an information system, the more valuable it will be to existing users. This is a classic pathway for a network effect.

3. *Biases*

In the current market for law search, there are substantial differences in the results returned by different services. One study comparing six legal databases (Casetext, Fastcase, Google Scholar, Lexis Advance, Ravel, and Westlaw) found that identical search terms entered in the search box generated little overlap between each database's top ten search results.[129] Differences in law search engine algorithms can lead to dramatically varied results, even if each algorithm's goal is to return the most relevant law.[130] Per a 2018 study, Lexis and Westlaw return the most relevant search results by a significant margin, but the field continues to evolve with the arrival of newer companies like Casetext, Fastcase, and Ravel.[131]

If we assume that for each user and search query there is a most useful (or best) set of result returns, the differences between companies could indicate there the is some *error* in search returns (i.e., search algorithms), or alternatively that the returns (algorithms) are

129. *See* Susan Nevelow Mart, *Results May Vary*, 104 A.B.A. J. 48, 49 (2018).

130. *See* Mart, *supra* note 129, at 50. *See generally* LAWRENCE LESSIG, CODE: AND OTHER LAWS OF CYBERSPACE (1999).

131. *See* Mart, *supra* note 129, at 53.

systematically biased. Here, we use the term error to indicate uncorrelated divergence from the best returns, and systematic bias to indicate there are underlying model features that generate departures from the best returns.

In general, error in search results is relatively harmless, in that they simply reflect an imperfect technology. Systematic bias is more troubling for at least two reasons. The first somewhat less problematic issue is that systematic bias can lead to legal blind spots, where appropriate and valid legal arguments are not made because law searchers do not come across the documents that might prompt those arguments. Such legal blind spots might stunt the development of the law and result in a less-than-optimal set of legal rules and legal outcomes.

More problematic consequences of systematic bias can arise if that bias maps onto important social categories.[132] The problem of gender and racial bias in the context of search results, predictive analytics, and algorithmic decision-making has been subject of substantial scholarly and public attention.[133] In one study of targeted advertising campaigns, algorithmic systems were found to have generated biased advertisements for particular individuals based on their name alone, explained in part by the race associated with each name.[134] Gender-related bias has also been revealed in various artificial intelligence contexts.[135]

Without further study, it is difficult to know whether systematic biases in law search results have the added problematic feature of corresponding to existing social categories. One could imagine that search results might privilege opinions written by some judges over others, or that results contain law that is more favorable to parties in

---

132. *See* OSONDE OSOBA & WILLIAM WELSER IV, AN INTELLIGENCE IN OUR IMAGE: THE RISKS OF BIAS AND ERRORS IN ARTIFICIAL INTELLIGENCE 17 (RAND Corp. ed., 2017).

133. *See supra* Section I.C.

134. *See* Latanya Sweeney, *Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising*, ACM QUEUE, Mar. 2013, at 10–13 (detailing a study in which the author manually searched thousands of black-identifying and white-identifying names on Google.com and Reuters.com to record resulting advertisements).

135. *See* Susan Leavy, *Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning*, *in* 2018 ACM/IEEE 1ST INT'L WORKSHOP ON GENDER EQUAL. SOFTWARE ENG'G 14–15 (2018) (explaining that women are often associated with domestic roles like "mother," and descriptions of occupations women hold are often qualified like "female lawyer" or "woman judge," while occupations men hold are not similarly described).

some social categories than others. Additional analysis is required to investigate these possibilities.

It is not clear that private markets will necessarily work to eliminate undesirable bias in law search results. Generally, law searchers will seek out results that inform their decision-making, and this tendency will create a natural pressure on firms to provide results with low error and bias. However, there are social interests in reducing certain kinds of systematic bias (such as bias that harms racial minorities), which private law searchers will not take account of when choosing between search engines—this effect generates a kind of externality. In addition, the preferences or interest of law searchers helps determine search results and ultimately the shape of the law. If this is the case, groups with the resources to fund a larger number of law searchers will have the greatest influence—there is no *prima facie* reason to believe that this influence will push the law in a welfare maximizing direction.

4. *Access to the Law*

One rule-of-law value that may not be adequately protected by private markets is access to the law. In *Georgia v. Public.Resource.Org*, the U.S. Supreme Court recently affirmed the longstanding principle that "no one can own the law."[136] As the Court noted, quoting an 1886 opinion of the Massachusetts Supreme Court, "Every citizen is presumed to know the law . . . and it needs no argument to show that . . . all should have free access to [its contents]."[137] That case concerned whether the State of Georgia could hold copyright to the official annotated version of the state statutes. The Court held that it could not, and that any word created by a "judge or legislator . . . in the course of his [or her] judicial or legislative duties is not copyrightable."[138]

But theoretical access to an enormous body of law is different from the practical ability to identify the relevant law for a legal question of interest. In *Georgia v. Public.Resource.Org*, the Court raised the following dystopian possibility:

> If everything short of statutes and opinions were copyrightable, then States would be free to offer a whole range of premium legal works for those who can afford the extra benefit. A State could monetize its entire suite of

136. Georgia v. Public.Resource.Org, Inc., 140 S. Ct. 1498, 1507 (2020).

137. *Id.* at 1516 (internal punctuation omitted) (quoting Nash v. Lathrop, 6 N.E. 559, 561 (Mass. 1886)).

138. *Id.* at 1513.

> legislative history. With today's digital tools, States might even launch a subscription or pay-per-law service.[139]

In reality though, law search is dominated by commercial databases—many documents that are ostensibly publicly available, are in practice difficult to acquire outside of those channels. Although states cannot "monetize [the] entire suite of legislative history" or offer a "pay-per-law service," commercial databases do exactly that, and without an alternative mechanism to find the law, the formal "access" provided by the state is illusory.

B. The Downsides of Regulation

The limitations of private markets are a primary justification for government intervention. For example, because firms operating in markets will not account for the harms caused by their pollution, governments impose requirements on emitters to adopt pollution-control technologies. In the context of law search, the most obvious existing government intervention to address a market limitation is the direct provision of law search by government officials—judges and judicial staff—during the course of litigation. There is a social interest in identifying the relevant law to address a dispute, and private parties cannot be counted on to invest adequate resources in identifying those authorities. This social investment appears to have paid off, at least to some extent: Judicial opinions frequently cite to cases that are not identified in either of counsels' briefs.[140]

Opportunities for directly regulating law search, however, appear limited. It is possible to imagine a regulatory regime that attempts to internalize the relevant positive externalities through a system of subsidies for law search, provided either to individual parties or to search providers. Network effects could justify breaking up the largest search providers or taking other steps to avoid too much concentration of market power. Anti-discrimination laws could be used to penalize search providers when their results exhibit socially undesirable systematic biases. Vouchers could be provided on a means-tested basis to provide access to commercial law search engines for those who lack sufficient financial resources.

---

139. *Id.* at 1512–13.

140. *See* Kevin Bennardo & Alexa Z. Chew, *Citation Stickiness*, 20 J. APP. PRAC. & PROCESS 61, 84 (2019) (finding that of 7552 cases cited across 325 federal court of appeals cases, 51% were "endogenous" and did not appear in either counsels' brief).

Although some of these regulatory interventions are imaginable, they would come with a host of complications and difficulties. Any system of subsidies to correct for the social interest in legal understanding would be difficult to design and potentially subject to gaming. Breaking up the large commercial providers might do more harm than good for the provision of useful, low-cost law search services. Anti-discrimination law has had difficulty being implemented even in its most traditional domains. It is far from clear how anti-discrimination law could be used to address the problem of bias in law search. The entire process of regulation would be subject to public choice failures and the threat of interest group capture.

Skepticism is warranted of many imagined regulatory interventions in the area of law search, notwithstanding the limitations of private markets. Nevertheless, that skepticism does not mean that public institutions need be entirely paralyzed. In the following Section, we discuss how government policy can support an open-access regime for public law search that could encourage a vibrant community of volunteers and researchers who could develop law search tools that could be disseminated for free on an open platform. Such a "public option" could help augment the private market for law search without the many limitations and shortfalls of direct regulation.

C. A Public Option

With respect to general search engines, it has been argued that the market dominance by a handful of firms has a number of pernicious effects and has proposed "publicly funded alternatives" as one way to address the problem of market concentration.[141] Although we do not take up the broader question of whether such an intervention is justified for search engines generally, we think that there is a particularly powerful argument in favor of government policy to support broad public access to well performing law search tools. Although there are many important details that would need to be worked out, in this Section we sketch the overall outline for our proposal.

Free access to the law is fairly limited. As mentioned above in Section IV.B, the law cannot be copyrighted, and so private entities are free to republish and distribute the law as they see fit. The two main commercial databases, Westlaw and LexisNexis, embed this noncopyrightable information behind a paywall along with a great

141. *See* Pasquale, *supra* note 71, at 402.

deal of proprietary information (such as various forms of mark-up), and then provide users with access to this material via various search tools. Several not-for-profit entities now provide access to digitized version of certain legal materials: these include the Free Law Project, Public.Resource.Org, the Legal Information Institute at Cornell University, and Harvard University's Caselaw Access Project. All of these initiatives play an important role in providing free access to legal materials. However, especially when compared to the commercial providers, the search and navigating capabilities of these free alternatives are fairly limited, which reduces their ability to provide usable access for the typical person.

Rather than supporting the community of not-for-profits and volunteer developers interested in providing public access to the law, many government entities in the United States have actively thwarted their efforts. The case *Georgia v. Public.Resource.Org* discussed above is an example. That litigation was brought by a state government against a not-for-profit that had released public versions of the state's official annotated statute. Another well-known impediment to public access to the law are the fees charged by the U.S. Courts for access to its PACER docket management system. This fully digitized resource includes all documents filed with the federal courts. Any document in the PACER system is ostensibly available to the public, but at a steep fee that makes bulk access impracticable. The U.S. Courts have jealously guarded this source of revenue.

There is a place for a more productive role for government entities in facilitating access to the law. The first and most obvious would be to avoid efforts to monetize the law. The PACER system could be made free to access, and efforts, like Georgia's, to copyright the law or engage in exclusive contracts with commercial providers can end. Such steps would mark a significant improvement from the status quo.

There also are more proactive steps that governments can take to facilitate access. A single unified source for legal documents, including statutes, court documents, legislative history, and the like could be maintained. States could be encouraged (and funded) to contribute their legal materials to this clearinghouse. Such a "USLaw.gov" site could include a user-friendly Application Programming Interface (API) to allow developers to easily explore and extract data. An easily used API is one way that developers could be encouraged to create tools that help users navigate these resources. The clearinghouse could also hold regular competitions between different search tools, with winning developers recognized for their

innovation on the site. Funding through the National Science Foundation or other sources could also support research in this area. A system of anonymized data collection on search patterns could be periodically released to researchers with the goal of improving search tools.

Given the widely recognized value of access to the law, as recently articulated and affirmed in *Georgia v. Public.Resource.Org*, surprisingly little has been done by states or the federal government to affirmatively promote genuine access. The first step in such a program is to halt efforts to affirmatively impede access to the law—eliminating the fees charged to users of the PACER system would be one important move along these lines. Creating a well-structured, easily accessed central clearinghouse for legal documents with an easy-to-use API could cultivate a community of researchers and developers interested in creating search tools that could be made publicly available. Funding for academic projects related to developing law search tools could also help expand the pool of knowledge that would be available to this developer community. Any of these steps would be an improvement on the status quo, and collectively, they could make substantial progress toward fulfilling the promise of providing functional access to the law for the public.

CONCLUSION

Law search is an integral component of legal reasoning, but it has not been given the attention that it deserves. From a theoretical perspective, law search implicates a wide range of important normative concerns ranging from welfare considerations to the rule-of-law values at the heart of any legal system. These normative concerns have not been well-articulated, and this Article takes a few small steps to define and explore their contours. We provide an empirical, descriptive account of law search and the related concept of legal relevance, and also describe how legal relevance could be understood in normative terms. On the empirical front, we outline some of the ways that a research program into law search would fit within existing information retrieval and information behavior paradigms. We go on to argue that tools from computational social sciences can be productively applied to the question of law search, and we discuss the *LexQuery* model, which draws its motivation from observations related to search behavior and implements a well-performing information retrieval platform. Finally, we discuss some of the policy implications of law search and propose government

policy for a law search "public option" generated by a developer community committed to creating open access law search tools.

Michigan State Law Review