March 5, 2014

Bargaining in the Shadow of Big Data

Dru Stevenson, *South Texas College of Law*
Nicholas J. Wagoner

Available at: https://works.bepress.com/dru_stevenson/18/
BARGAINING IN THE SHADOW OF BIG DATA

Dru Stevenson* & Nicholas J. Wagoner**

Attorney bargaining has traditionally taken place in the shadow of trial, as litigants alter their pretrial behavior—including their willingness to negotiate a settlement—based on their forecast of the outcome at trial and associated costs. Lawyers bargaining in the shadow of trial have traditionally relied on their knowledge of precedent, intuition, and previous interactions with the presiding judge and opposing counsel to forecast trial outcomes and litigation costs. Today, however, technology for leveraging legal data is moving the practice of law into the shadow of the trends and patterns observable in aggregated litigation data. In this Article, we describe the tools that are facilitating this paradigm shift, and examine how lawyers are using them to forecast litigation outcomes and reduce Coasean bargaining costs, in both litigation and transactional scenarios. We also explore some of the risks associated with bargaining in the shadow of big data and offer guidance to lawyers for leveraging these tools to improve their practice.

Our discussion pushes beyond the cartoonish image of big data as a mechanical fortuneteller that predicts who will win or lose a case, supposedly eliminating research or deliberation. We also debunk the alarmist clichés about newfangled technologies eliminating jobs. Demand for lawyers capable of effectively bargaining in the shadow of big data will continue to increase, as the legal profession catches up to the data-centric approach found in other industries. Ultimately, this Article paints a portrait of what big data really means for practicing attorneys, and provides a framework for exploring the theoretical implications of lawyering in the era of information analytics.

I. INTRODUCTION ................................................................. XX
II. BACKGROUND .................................................................. XX
   A. Intuition is Overrated .................................................... XX
   B. Innovation in the Legal Industry ................................. XX
   C. Law as Data ................................................................. XX
      1. Big Data in Disparate Domains .............................. XX
      2. Viewing the Law from a New Perspective .............. XX
      3. Public Access to Legal Data ................................. XX
I. INTRODUCTION

For the rational study of the law the black-letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics.

—Oliver Wendell Holmes, Jr., The Path of the Law, 10 HARV. L. REV. 457, 469 (1897).

The practice of law includes prediction. Clients expect their lawyers to answer questions like, “What are the odds of winning this case, and what’s it going to cost me?” Justice Oliver Wendall Holmes observed this fact

* Hutchins Research Professor, South Texas College of Law; Baker Institute Scholar, James A. Baker III Institute at Rice University. Many thanks to my institution for providing support for this research, and to Josh Blackman, Matt Festa, and Val Ricks for helpful comments and suggestions.

**I would like to thank Daniel Lewis of Ravel Law and Owen Byrd of Lex Machina for sharing their experiences with me regarding their work at startups focused on developing software that will facilitate lawyering in the shadow of data; as well as Keith Lee and Josh Blackman for providing thoughtful commentary on earlier drafts of this Article.

1 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. REV. 909, 912 (2013) (“Do I have a case? What is our likely exposure? How much is this going to cost? What will happen if we leave this particular provision out of this contract? How can we best staff
more than a century ago when he wrote, “the prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law.”

Today, practitioners still resemble the Holmes’ “black-letter men” of yesteryear more than the number-crunching masters of economics that Holmes had envisioned. The vast majority of attorneys today still rely solely on their own experience, knowledge of case precedents, and intuition to predict what the courts will actually do. But as mere mortals, attorneys—even the exceptional ones—are inherently limited in their capacity to retain and process the information necessary to make well-informed judgments. Likewise, attorneys are limited in their range of personal experiences. Computers, on the other hand, while lacking the ability to frame interesting questions or draw conclusions, are far better at storing, processing, and summarizing large volumes of information. Thus, by leveraging the quantitative strength of computers, lawyers can more accurately forecast how events will play out in litigation. This foresight, in turn, will allow lawyers and their clients to avoid costly mistakes, better appreciate the strengths and weaknesses of their case, and increase their odds of securing a favorable outcome.

Others in the legal academy have already begun to discuss the value of big data decisionmaking, though not necessarily in the context of lawyering. Professor Ian Ayers recently popularized the term “super crunching,” referring to the statistical analysis of big data to guide real-world decisions. Super crunching improves decisionmaking by leveraging the quantitative strength of computers so that human users may spend more time doing what they do best—identifying relevant data to crunch, formulating questions to ask of their data, and then acting on the results.

---

2 Oliver Wendell Holmes, Jr., The Path of Law, 10 HARV. L. REV. 457, 461 (1897).

3 See CLARENCE MORRIS, HOW LAWYERS THINK 11–19 (1937).

4 IAN AYRES, SUPER CRUNCHERS 21 (2007).

5 See id. at 10.
In popular culture and semiotics, on the other hand, *Moneyball* has become the vernacular shorthand for data-driven decisions. In 2002, after making a fortune in finance by analyzing data to uncover and exploit inefficiencies in the financial markets, John Henry led a group of investors in the purchase of the Boston Red Sox baseball franchise. Using the data-driven principles that Professor Ayers would later espouse in *Super Crunchers*, Henry built a team that went on to win the 2004 World Series and break the “Curse of the Bambino.” In a letter to a friend, he explained, “People in both fields operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage.”

Henry’s insight applies equally to the practice of law. Lawyers, just like decision makers in every other field, operate with beliefs and biases that can cloud judgment. Yet more than a century ago, Justice Holmes pointed to a better path to predicting the law than that traveled by today’s “black-letter man”—one that leads to a future where “the rational study of the law” requires a “man of statistics and master of economics.” Like John Henry’s Boston Red Sox, lawyers who embrace data-driven decisionmaking will gain a clear advantage over their counterparts who still cling to their Dictaphones, feather-quilled pens, and thumb-in-the-air predictions when handling their clients’ legal affairs.

---

6 Dan Shaughnessy, *Reversing the Curse* 60 (2005).


BARGAINING IN THE SHADOW OF BIG DATA

Why then, has most of the legal profession yet to find the path envisioned by Justice Holmes? The profession lags behind most other industries in crunching data to improve its marketing and delivery of services. A significant bottleneck to progress has been the historical inaccessibility of litigation data, and the inadequacy of the tools with which lawyers presently conduct legal research.

This is not to say that legal practitioners have not embraced modern technology. To the contrary, law firms were early adopters of office computers and word processors, scheduling and billing software, in-house networks, conference phones, and online databases (Lexis and Westlaw). Before smartphones became ubiquitous, lawyers commonly had PDA’s to keep track of appointments, hourly billing, and client contacts.

“Big data” is different. Previous technological advances mostly made firms more efficient at tasks that they were already doing—scheduling meetings, drafting documents, sharing ideas, and looking up cases. A few technologies considerably changed how lawyers approached a task. The most obvious example of this was the profession’s massive shift toward precision-timed billing (in minutes or fractions of hours) in the early 1980s, rather than ball-parked or “scheduled” fees, once computer software made such time-tracking more feasible.\(^{12}\) For the most part, however, the underlying nature of the work remained largely the same—researching, writing, meeting with clients and opposing counsel—and the technology merely made these tasks more convenient, or allowed lawyers to handle more cases. Even the research that attorneys now perform through Lexis and Westlaw is analytically analogous to the old approach in law libraries—finding cases in bound digests and reports, using intricate indexing systems like West’s Keys or Lexis’s Headnotes.

Big data, by contrast, invites lawyers to make a fundamental change in

\(^{12}\) See Ed Wesemann, Alternative Pricing, 70 OR. ST. B. BULL. 38 (2010) (“The popularity of billing by the hour was supported by technology, first through ‘one-write’ record books, then electronic accounting machines, and eventually, computers until it became the pricing standard for the legal profession.”); H. Edward Wesemann, The Power of Price, 23 LEGAL MGMT. 22, 32 (2004) (“The billing-by-the-hour trend didn’t earnestly begin until the 1960s, and it took computers to bring firms to complete dependence on hourly billing.”); MORTIMER SCWARTZ ET AL., PROBLEMS IN LEGAL ETHICS 138–39 (10th ed. 2012). Ironically, even as computers facilitated the entrenchment of hourly billing, the advent of computers prompted leaders in the legal profession to predict that computers would also force the demise of time-based fees, as the new efficiency would make legal tasks too brief to generate sufficient time-based revenue. See, e.g., Sandra L. Yost, Alternative Billing Strategies—New Wave or Passing Fad?, 13 LEGAL MGMT. 18, 20 (1994) (discussing comments by state bar leaders at panel discussions about the impact of computers).
their approach to the law itself by looking to statistical patterns, predictors, and correlations, in addition to the legal rules that purportedly control outcomes—case law, statutory law, procedural rules, and administrative regulations. Traditional lawyering required knowledge of the pertinent legal rules and the ability to apply them to a given set of facts, whether in litigation or in transactional work. This application of law to facts would yield a type of estimate about probabilities; that is, a prediction of the likelihood that a given rule would govern a given scenario. The question was whether a feature of the client’s current situation would trigger a rule and its mandatory result. Analogies, comparisons, and normative judgments all figured into this assessment. Lawyer’s fees reflected, in theory, the time and resources required to determine the relevant law and analyze the likely outcome.

Big data turns this approach on its head. Rather than assuming that rules dictate outcomes as the basis for making specific predictions, big data looks for patterns and correlations. For example, historical litigation data, in the aggregate might reveal a judge’s tendency to grant or deny certain types of pretrial motions, an opponent’s historical avoidance of expert witnesses, or a party’s typical timing for settlements may all be more relevant for a client or lawyer than the published court opinions in prior cases that ran the full course of litigation.

Coasean bargaining between attorneys traditionally took place in the shadow of the trial, which simply refers to the fact that litigants alter their pretrial behavior, including willingness to negotiate a settlement, based on their perception of the likely outcome at trial and the anticipated cost to get there.13 Traditionally, lawyers have formed their perception of the relative strength of their clients’ cases based on case precedents as well as their prior interactions with the presiding judge, opposing counsel, and experience handling similar cases. Today, however, the availability of tools for leveraging legal data are increasingly moving bargaining into the shadow of the trends and patterns observable in aggregated litigation data.

What exactly does this mean for the profession? A traditional lawyer with many years of experience would acquire awareness of some relevant patterns, tendencies, and habits—for example, the fact that certain judges were always severe; certain lawyers were more affable than others, even as opposing counsel; certain parties were unusually litigious; certain courthouses painfully slow. Such experiential knowledge supplemented the

lawyer’s knowledge of the legal rules, and these personal factors could sometimes matter more than the legal rules themselves. Seasoned lawyers practiced not just in the shadow of the trial, but also in the shadow of the traits and tendencies of everyone involved. These intangible aspects of traditional lawyering, ironically, are the most analogous to lawyering in the shadow of data. The new technology, however, makes available in minutes insights that once might have taken thirty years of practice experience to acquire. The latest legal-data services becoming available for practitioners yield accurate statistical information about the preferences, persistence, and pace of the judges, parties, and lawyers involved in a matter. Big (legal) data not only compresses the time component of such knowledge from decades to minutes, it dramatically expands the potential breadth of observations—that is, a lawyer can now “profile” complete strangers from other jurisdictions for factors the lawyer has never before encountered in his or her limited circle of acquaintances.

At the outset, we should dispense with the cartoonish image of big data as a mechanical arbitrator or adjudicator that will simply tell lawyers which party will win or lose a case, sparing them the need to do any research or give the matter much thought. To the contrary, big data potentially makes the lawyer’s job much more complicated and nuanced. Working in the shadow of data poses more ambiguity, both normative and positive, than working in the shadow of trials, which, in turn, occur in the shadow of rules. In fact, big data poses a new set of hazards or perils for its users. The line dividing successful from unsuccessful attorneys will shift from those skilled in “reading” the other parties to a lawsuit in the traditional sense of the word toward those familiar with the capabilities and limitations of lawyering with the benefit of big data. Specifically, the Holmesian lawyers of tomorrow will know when patterns in litigation data make predictions, when predictions run the risk of becoming self-negating or self-fulfilling, and when to reduce their exposure to the systemic risks that will inevitably result from more attorneys beginning to practice law in the shadow of data.

We also must move beyond the tired clichés about newfangled technologies eliminating jobs. It is a truism that automation replaces the workers who once manually performed the newly automated task.  

\[14\] Some have argued that the legal profession historically has been slow to embrace any new technologies because reduced hours (through automation) would reduce billable revenue. William G. Ross, *The Ethics of Hourly Billing by Attorneys*, 44 Rutgers L. Rev. 1, 30 (1991) (“A more troubling reason for the widespread failure of attorneys to use word processors may be that managers of law firms fear that the time saved by use of computers would result in a net reduction of billable hours.”).
believe, like many economists, that automation eventually shifts the workforce elsewhere, to other nonautomated tasks, rather than demobilizing any section of the workforce permanently. Some legal tasks, like document review, are becoming increasingly automated, meaning associate attorneys are in less demand for such tasks; but we believe that demand for associates who can handle data-driven lawyering will continue to increase as the legal profession catches up to the data-centric approach to information found in other industries.

The thrust of this Article is that lawyering in the shadow of data is qualitatively different than the technology-enhanced lawyering of previous decades. We attempt to take a more theoretical approach than previous literature in this area in order to paint a picture of what big data really means for lawyers and judges—and we hope to debunk some common misperceptions along the way. This includes a description of the newest legal-technology startups, and some predictions about anticipated market entrants in the near future. In addition, we define—for the first time—one of the hazards inherent in basing legal decisions on historical data that other commentators have largely overlooked. Some of these concerns are normative, but most of them are descriptive and theoretical. Our point, however, is not to discourage the use of big data, but rather to help the legal profession understand the advantages and pitfalls of lawyering in the shadow of the data.

The ensuing discussion proceeds along the standard roadmap: Part II will familiarize readers with the very recent technological upheaval in the American legal profession and will discuss the emergence of the “law-as-data” movement that has resulted. This includes a brief overview of PACER and the commercial alternatives. Part III then delves into the nature of legal “predictions” that big data will enable, and will attempt to define the value and limitations of such data-driven predictions. This section contains the crux of our argument, that big data significantly enhances Coasean bargaining by lowering information costs, which traditionally formed the most easily identifiable transaction costs for purposes of Coasean analysis of legal transactions and disputes. Part IV initiates a long-

---

overdue discussion about the theoretical problems with predictions, forecasts, and lingering uncertainty in the legal arena. An important subsection of Part IV addresses the special problem of “Black Swans”—unforeseeable systemic risks that creep in as information-based decisionmaking allows tighter integration and coordination of independent players in the system. Part V provides a concise recap of the main points and suggests areas for further research.

II. BACKGROUND

A. Intuition is Overrated

Lawyers, like accountants, doctors, and any other professionals, are human. Human nature has a tendency to cloud judgment by introducing cognitive biases and emotion into the decision-making process. Lawyers are often unaware of the subtle saboteurs of rational thought that have infiltrated their subconscious. Practitioners are also inherently limited by their range of experiences and finite capacity to retain information. Their ability to make sound judgments takes years of practice to develop and a lifetime to master.

Data, on the other hand, is indifferent. It is impartial. It can be comprehensive in scope where lawyers are limited in experience. Data does not care whether it is consistent with someone’s preconceived notions or personal experiences, nor is it particularly concerned with intuition. Rather, data is simply a historical reflection of reality. It seems rather obvious then, that aggregated data about relevant, past litigation should play a key role in informing the inherently subjective judgment of lawyers; a “reality check,”


18 See Stephanos Bibas, Plea Bargaining Outside the Shadow of Trial, 117 HARV. L. REV. 2464, 2496-527 (2004). The same is true, of course, for judges. See Richard A. Posner, The Rise and Fall of Judicial Self-Restraint, 100 CAL. L. REV. 519, 553–54 (2012) (“Justices operate on limited information; because there are no sensible algorithmic methods of deciding difficult cases, most constitutional decisions have only weak claims to objective validity; the parts of the Constitution that generate litigation at the Supreme Court level are too old and general to be directive; the issues presented in constitutional cases tend to be both emotional and momentous and the decisions resolving them inescapably reflect the Justices’ personal values, psychology, background, peer pressures, political anxieties, professional experiences, ideological inclinations, and other non-legalistic factors, often operating unconsciously. . . .”).
so to speak.

Yet lawyers continue to rely mostly on their experience, intuition, and instinct. This approach to legal decisionmaking is problematic for several reasons. First, as noted above, intuition takes years to develop. To make matters worse, studies have shown that lawyers are generally overconfident when assessing their likelihood of success, and that the

---


20 In 1993, the American Bar Association published the results of a study in which roughly 3,000 lawyers were asked to take a Myers-Briggs Type Indicator test to measure the personality types prevalent in the profession. Larry Richard, The Lawyer Types, ABA J. 74, 74 (July 1993). The results showed that, while a mere 30% of the general population in the United States indicated that they relied primarily on their intuition rather than concrete facts, 57% of lawyers were inclined to trust their intuition over concrete facts when making decisions. Id. at 76. This may be due in large part to the fact that lawyers are often called upon to make judgment calls based on incomplete facts, and perhaps are more accustomed to wrestling with ambiguity than the general population.


23 Some have suggested that this is due in part to the fact that overconfidence is a necessary trait to attract and retain clients. Loftus & Wagenaar, 1988, p. 450; Simon, 1988; Brehm 1956.
accuracy with which lawyers can predict outcomes does not improve with experience.\textsuperscript{24}

Considering what is at stake, it seems imprudent to rely on experience, intuition, or instinct alone in predicting the path of the law. As the next section will explain, however, lawyers are not entirely to blame. Only recently have economic, technological, and political forces come together to form the perfect storm for introducing the legal profession to a new, data-driven approach to practicing law.

B. Innovation in the Legal Industry

This Article is set against the backdrop of the economic turmoil that has rocked the legal profession in recent years, the disruptive technology that is reshaping the decision-making processes in other industries, and the growing body of free and easily accessible legal data that now exists. For nearly a century the practice of law remained partly sheltered from the winds of technological change that have swept through other industries in recent years. That all changed in 2008. Some of the nation’s largest white-shoe law firms collapsed under the crushing weight of overhead costs that clients could no longer afford to subsidize.\textsuperscript{25} Many more law firms have experienced dramatic downsizing over the last five years.\textsuperscript{26} With more

\textsuperscript{24}Jane Goodman-Delahunty et al., Insightful or Wishful: Lawyers’ Ability to Predict Case Outcomes, 16 PSYCH., PUB. POL’Y, & L 133, 133 (2010).

\textsuperscript{25}Larry Ribstein, The Death of Big Law, 2010 WISC. L. REV. 749 (2010); James B. Stewart, Dewey’s Fall Underscores Law Firms’ New Reality, N.Y. TIMES (May 4, 2012); Steven Pearlstein, Why Howrey Law Firm Could Not Hold It Together, WASH. POST. (Mar. 12, 2011), http://articles.washingtonpost.com/2011-03-19/business/35207602_1_law-firm-howrey-simon-robert-ruyak (observing that that rising overhead costs and fee cuts demanded by clients precipitated the collapse of the Howrey law firm in 2011, which, just two years earlier, “appeared to the outside world to be a thriving global enterprise with more than 700 lawyers”).

empty chairs at the office and tighter lines of credit at the bank, law firms searched desperately for a cheaper alternative to rehiring pricey junior associates when legal grunt work—namely, document review—picked back up.\(^{27}\)

Many law firms were quick to embrace e-discovery software, which automated the process of reviewing documents for relevancy with greater accuracy than human reviewers, at a fraction of the cost.\(^ {28}\) As a result, the e-discovery services industry sprang up virtually overnight, raking in revenue of approximately $2.8 billion in 2009 alone.\(^ {29}\) It was not long before the New York Times picked up on the significance of this shift toward automated legal services, heralding a day that many in the profession thought would never come with a headline that read, “Armies of Expensive Lawyers, Replaced by Cheaper Software.”\(^ {30}\)

C. Law as Data

1. Big Data in Disparate Domains

While technological innovation has crept into certain discrete facets of the legal profession, it has dramatically improved other sectors of the economy. Technology giants like Google, Amazon, and Facebook have demonstrated just how valuable making the world’s information more

---

\(^{27}\) See generally Farhad Manjoo, Will Robots Steal Your Job?, Slate (Sept. 29, 2011 2:42 A.M.), http://www.slate.com/articles/technology/robot_invasion/2011/09/will_robots_steal_your_job_5.single.html (“E-discovery software has been similarly revolutionary. These systems can mine huge volumes of material (like all the email correspondence in a civil suit) for damning evidence. The simplest software looks for specific keywords, but more sophisticated systems can detect patterns of behavior that might interest lawyers. This was the sort of work that once consumed the lives of first-year associates; now computers do it faster, at lower cost, and with about as much success as humans.”).

\(^{28}\) Cara H. Drinan, Getting Real about Gideon: The Next Fifty Years of Enforcing the Right to Counsel, 70 WASH. & LEE L. REV. 1309, 1341 n.173 (2013); Smith, supra note 15 at 141 (“Now, thanks to advances in artificial intelligence, “e-discovery” software can analyze documents in a fraction of the time for a fraction of the cost.”); Stewart Macaulay, Notes on the Margins of Lawyering, in Three and a Half Minutes, 40 Hofstra L. Rev. 25, 33 n.43 (2011); Ed Poll, Technology Investments and Two-Handed Decision Making, 30 LEGAL MGMT. 52, 54 (2011)(“E-discovery software can analyze documents required for litigation discovery in a fraction of the time for a fraction of the cost when compared to using lawyers for the task.”).


useful can be. Even traditional brick-and-mortar industry leaders like Wal-
Mart and General Electric now heavily rely on sophisticated computer
networks and software to collect, store, distill, and act on the deluge of data
that pours into their servers daily. The ability to crunch big data in this
manner largely stems from two recent developments: the abundance of data
that now exists and the emergence of software that makes it more accessible
and useful.

So what exactly is “big data”? For some, it is a tired cliché; nothing
more than “[a] meme and a marketing term.” For others, it is gold, or the
new oil, “valuable, but if unrefined it cannot really be used.” NPR likened
it to dust: using a cell phone, browsing the web, buying groceries,
connecting with friends on Facebook, driving with a toll tag—each day
“[w]e kick up clouds of it wherever we go.” Mckinsey Global Institute
defined the term more formally as “large pools of data that can now be
captured, communicated, aggregated, stored, and analyzed.” However one
describes or defines big data, there is no denying that its time has come.

---

31 Data Deluge, The Economist.
32 AYERS, supra note 4, at 145–65; Katz, supra note 1, at 913–19.
are Mining the Information Mother Lode for Pricing, Practice Tips and Predictions, ABA J. (May
2013).
34 Id.
35 Michael Palmer, Data is the New Oil, ANA Marketing Maestros (Nov. 3, 2006 5:43 A.M.),
36 Geoff Nunberg, Forget YOLO: Why 'Big Data' Should be the Word of the Year, NPR (Dec. 20,
2012 10:58 A.M.), http://www.npr.org/2012/12/20/167702665/geoff-nunbergs-word-of-the-year-
big-data.
37 James Manyika et al., Big Data: The Next Frontier for Innovation, Competition, and
38 Lohr, supra note 33, at SR1.
40 See Katz, supra note 1, at 916.
The abundance of data and computers’ increasing capacity to store it are only half of the equation. The other half consists of computers that are now capable of scouring big data for patterns that might help human analysts solve problems and answer questions. Baseball teams crunch obscure statistics to spot undervalued players. President Obama’s 2012 reelection campaign identified likely donors and undecided voters by combing through reams of polling data. Commercial industries as diverse as healthcare, national defense, fraud detection, travel, online dating, telecommunications, oil and gas, and banking have enjoyed

41 Lewis, supra at ___.


That data-driven decisionmaking played a huge role in creating a second term for the 44th President and will be one of the more closely studied elements of the 2012 cycle. It’s another sign that the role of the campaign pros in Washington who make decisions on hunches and experience is rapidly dwindling, being replaced by the work of quants and computer coders who can crack massive data sets for insight. As one official put it, the time of “guys sitting in a back room smoking cigars, saying ‘We always buy 60 Minutes’” is over. In politics, the era of big data has arrived.

Id.


45 The Data Deluge, THE ECONOMIST (FEB. 25, 2010), http://www.economist.com/node/15579717 (“Credit-card companies monitor every purchase and can identify fraudulent ones with a high degree of accuracy, using rules derived by crunching through billions of transactions. Stolen credit cards are more likely to be used to buy hard liquor than wine, for example, because it is easier to fence. Insurance firms are also good at combining clues to spot suspicious claims: fraudulent claims are more likely to be made on a Monday than a Tuesday, since policyholders who stage accidents tend to assemble friends as false witnesses over the weekend. By combining many such rules, it is possible to work out which cards are likeliest to have been stolen, and which claims are dodgy.”).

46 Data, Data Everywhere, supra note 39 (“Microsoft’s search engine Bing, can advise customers whether to buy an airline ticket now or wait for the price to come down by examining 225 billion flight and price records.”).

47 Lohr, supra note 33 at SR1 (“Online dating services, like Match.com, constantly sift through their Web listings of personal characteristics, reactions and communications to improve the
similar benefits by figuring out how to capture and glean insight from the deluge of domain-specific data that now exists.

The legal profession has been noticeably absent from the parade of progress that has lead to valuable breakthroughs in other industries. It seems anomalous that “a singularly information-dependent profession”\textsuperscript{51} like the legal profession would arrive late to the age of big data. Some commentators have suggested that lawyers have resisted technological progress for fear that efficiency gains might decrease the number of billable hours for which they may charge their clients.\textsuperscript{52} While some lawyers may indeed harbor such unsavory motives for resisting innovation, we believe that the best attorneys welcome new solutions that let them perform at an even higher level, for a greater number of clients. The more plausible explanation for the lack of big-data solutions in the legal-research space is the historical inaccessibility of legal data, as discussed below.

2. Viewing the Law from a New Perspective

There are many lenses through which to view the law. Justice Holmes viewed the law through the eyes of a “bad man.”\textsuperscript{53} “For many observers of the American legal system, law is what judges write in appellate opinions.”\textsuperscript{54} Having entered the age of big data, a growing number of scholars are now urging scholars and practitioners alike to view the law as

\begin{quote}
\textsuperscript{49} Data Deluge, \textit{The Economist} (Feb. 27, 2010) (“Mobile-phone operators, meanwhile, analyse subscribers’ calling patterns to determine, for example, whether most of their frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, he or she can then be offered an incentive to stay.”).

\textsuperscript{50} See \textit{id}. (“The oil industry uses supercomputers to trawl seismic data before drilling wells.”).

\textsuperscript{51} Data, Data Everywhere, \textit{supra} 39 (“Personal-finance websites and banks are aggregating their customer data to show up macroeconomic trends, which may develop into ancillary businesses in their own right.”).


\textsuperscript{53} Ross, \textit{supra} note 14, at 30.

\textsuperscript{54} Holmes, \textit{supra} note 2, at 459.

data. For instance, professors Daniel Katz and Josh Blackman routinely write and speak about the value of aggregating bulk legal data and then analyzing it using modern computing. This “law-as-data” movement stems from the legal academy’s growing interest in legal empiricism, which, until recently, remained largely relegated to the realm of political science.

To date, the vast majority of empirical studies of the law, or more precisely, litigation, have focused on the Supreme Court. This is understandable given the high court’s position at the top of the Judicial Branch; greater visibility than lower courts; and smaller, more manageable pool of cases. A handful of scholars have, however, stressed the importance of empirically analyzing litigation data generated at the district-court level.

For purposes of this Article, “litigation data” means (1) the basic information about a given case, such as the names of the parties, attorneys, nature of suit, and other characteristics about the case (otherwise referred to

55 See generally Katz, supra note 1.


57 See generally Gerald N. Rosenberg, Across the Great Divide (Between Laws Political Science), 3 GREEN BAG 2d 267 (2000).


as “case metadata”); (2) information reflected on the electronic docket sheet for a given case; and (3) the electronic court records associated with a given case. When combined, these sources paint a detailed portrait of litigation on the front lines of the federal Judiciary. When aggregated on a large scale and then filtered using software, patterns and trends begin to surface that shed light on how parties, attorneys, and judges typically behave during the litigation process.

Case metadata becomes even more valuable when combined with information scraped from case docket sheets and then classified using a machine-learning algorithm. Docket sheets are ideally suited for studying the ebb and flow of litigation because they consist of a sequentially ordered timeline of every event that occurs over the life of a lawsuit. When a lawyer electronically files something or a judge communicates in some fashion with the parties, the court clerk will create a new docket entry that consists of a brief description of the event, the date of that event, and a hyperlink to any documents associated with that event.

The text in docket entries lend themselves to large-scale analysis because they are far easier to automatically parse and classify with machine-learning algorithms than court records (a key component of of software enabled, trend-based lawyering), as they are typically short and formulaic. Docket sheets are also an important source of information about district courts because the vast majority of litigation events do not produce written opinions. As one commentator has pointed out, “a district judge

---

60 This Article limits its analysis to litigation data collected from district courts at the federal level, although many of the principles discussed herein apply equally to litigation data at the state level.


62 For purposes of this Article, the authors collaborated with an expert in scraping and machine learning to develop a proprietary algorithm capable of automatically reading case docket sheets, identifying summary judgment events, and classifying the outcome, movants, nonmovants, and other metadata that the authors may discuss at greater length in a subsequent article.

63 See Kim et al., supra note 59, at 107–08.

may make many decisions of varying types at different points in time in a single case. Most are not accompanied by written reasons and therefore are recorded on the docket sheet or in a brief order as simply a decision to ‘grant’ or ‘deny’ a particular motion.\textsuperscript{65} Decisions that do not produce written opinions, however, can sometimes dramatically alter the cost and ultimate outcome of litigation.\textsuperscript{66}

Readers are probably most familiar with the third and final category of litigation data discussed in this Article—court records. Court records tell the story of a given case. To initiate a lawsuit in district court, a plaintiff must file a complaint that sets forth a valid legal basis for relief and factual allegations that would, if true, justify such relief. A defendant will then respond to the complaint by filing an answer,\textsuperscript{67} motion to dismiss,\textsuperscript{68} or any number of other legal instruments. After reaching a final decision in a given case, a judge may issue an order, ruling, or judgment accompanied by a written opinion in which the judge will articulate his or her findings of fact and conclusions of law in support of their decision.

3. Public Access to Legal Data

The previous section of this Article answered the “What” question; that is, “What type of data may lead to a better understanding of litigation?” Before moving on to the “Where” question (as in, “Where can this data be found?”), it is important to first take a step back and ask, “Why should this data be accessible for empirical analysis?” The answer to this question requires a brief journey back in time.

Since the age of the Pharaohs, men and women have conditioned their willingness to submit to the judgment of another on the guarantee that such decisions would be carried out in public. Inscribed inside the tomb of an Egyptian official who reported to King Tuthmosis III were the instructions,

\begin{quote}
Lo, whenever an administrator hears cases let there be publicity and let water and air report all that he may do. Lo,
\end{quote}

\textsuperscript{65} See Kim et al., supra note 59, at 98 (discussing “opinion bias” in empirical studies of district courts).

\textsuperscript{66} Id. (describing the ways in which discovery decisions and nondispositive rulings can influence case settlements).

\textsuperscript{67} FED. R. CIV. P. 12(a).

\textsuperscript{68} FED. R. CIV. P. 12(b).
then his conduct is by no means unperceived. If he does anything (unseemly) and he is to blame he is not to be reinstalled on the authority of an acting official but men shall learn of it on the authority of his (proper) judge. . . .

The strong preference for transparency in American courts traces its roots back to Medieval Europe. In the mid-thirteenth century, lawyers began keeping records of their appearances before the English court of common pleas. At the end of their career, a lawyer would hand their records down to their apprentice, who would then carry on the tradition. Lawyers quickly realized the value of keeping detailed records of proceedings, often citing them to show that their proffered position was consistent with the judge’s previous rulings, or that the position advanced by their adversary was inconsistent with previous rulings. This practice developed into “the concept of precedent as a means towards achieving uniformity and consistency in pleading and in judicial decision-making.”

Lawyers and judges were not the only ones to benefit from accessibility to public records. English peasants used court records to prove their entitlement to develop land under the feudal system. Before long, court records were considered to be a “public resource.” The importance of transparency was not lost on the colonists who established the first courts in America either. For example, the 1677 Concessions and Agreements of West New Jersey formally recognized that “inhabitants of the said Providence may freely come into, and attend the said courts, and hear and be present, at all or any such tryals as shall be there had or passed, that justice may not be done in a corner nor in any covert manner.”

---

69 From the tomb of Rekhmire, Vizer to Tuthmosis III, King of Egypt, XVIII Dynasty (1580–1321 B.C.).


71 Id.

72 Id.

73 Id. at 103.

74 Id.

75 Id.

colonial government of Pennsylvania similarly declared in 1682 that "all courts shall be open." Early settlers "sensed from experience and observation that . . . the means used to achieve justice must have the support derived from public acceptance of both the process and its results."

Today, federal courts carry on the nation’s tradition of transparency by preserving the public’s broad, common-law “right to inspect and copy public records and documents, including judicial records and documents.” This right of public access to court records applies to both civil and criminal, pretrial and trial proceedings. Some courts have gone a step further by recognizing the right to access public records in certain cases as one that is protected by the First Amendment.


79 Nixon v. Warner Comms., Inc., 435 US 589, 597 (1978); see also Ex parte Drawbaugh, 2 App. D.C. 404, 407 (1894) (“Such claims of right, and contests over them, are not the ordinary incidents of judicial proceeding; and any attempt to maintain secrecy, as to the records of the court, would seem to be inconsistent with the common understanding of what belongs to a public court of record, to which all persons have the right of access, and to its records, according to long established usage and practice.”).

80 In re Continental Illinois Securities Litigation, 732 F.2d 1302, 1308-9 (7th Cir. 1984); United States v. Edwards, 672 F.2d 1289, 1294 (7th Cir. 1982); FTC v. Standard Fin. Mgmt. Corp., 830 F.2d 404, 409 (1st Cir. 1987) (“We rule that relevant documents which are submitted to, and accepted by, a court of competent jurisdiction in the course of adjudicatory proceedings, become documents to which the presumption of public access applies.”); Smith v. U.S. Dist. Ct. for the S. Dist., 956 F.2d 647, 649 (7th Cir. 1992) (“The federal common law right of access to judicial records and documents is well recognized.”); Foltz v. State Farm Mut. Auto. Ins. Co., 331 F.3d 1122, 1134 (9th Cir. 2003) (noting that the common law right to public access extends “to pretrial documents filed in civil cases, including materials submitted in connection with motions for summary judgment”); Hartford Courant Co. v. Pellegrino, 380 F.3d 83, 93 (2d Cir. 2004) (protecting the public’s right to inspect and copy docket sheets).

81 See United States v. Ochoa-Vasquez, 428 F.3d 1015, 1029 (11th Cir. 2005); In re Providence Journal Co., 293 F.3d 1, 10–13 (1st Cir. 2002); Hartford Courant Co. v. Pellegrino, 380 F.3d 83, 93 (2d Cir. 2004); Phoenix Newspapers, Inc. v. United States Dist. Court, 156 F.3d 940, 948 (9th Cir. 1998) (concluding that the First Amendment requires the “release of transcripts [of closed criminal proceedings] when the competing interests precipitating hearing closure are no longer viable”); United States v. Antar, 38 F.3d 1348, 1359–60 (3d Cir. 1994) (stating that “the right of access to voir dire examinations encompasses equally the live proceedings and the transcripts which document those proceedings” and observing that “[i]t is access to the content of the proceeding—whether in person, or via some form of documentation—that matters”); Grove Fresh Distrib., Inc. v. Everfresh Juice Co., 24 F.3d 893, 897 (7th Cir.1994) (“The First Amendment presumes that there is a right of access to proceedings and documents which have historically been open to the public and where the disclosure of which would serve a significant role in the functioning of the process in question.” (internal quotation marks omitted)); Globe Newspaper Co. v. Pokaski, 868 F.2d 497, 505 (1st Cir. 1989) (finding that, “after Richmond Newspapers, a
Courts and scholars alike have identified a variety of benefits that flow from public access to court records, including (1) courts that are more accountable to the people,82 (2) courts that are more accountable to policymakers,83 (3) increased public respect for and confidence in the court system,84 (4) greater consistency in the law’s application,85 (5) increased public interest in how laws are made and applied,86 and (6) the “community therapeutic value” of openness that squelches the public’s urge to take justice into its own hands.87

82 See Nixon, 435 U.S. at 598 (1978) (noting that public’s right to access court records arises from “the citizen’s desire to keep a watchful eye on the workings of public agencies”); Smith v. U.S. Dist. Ct. for So. Dist., 956 F.2d 647, 650 (7th Cir. 1992) (quoting FTC v. Standard Fin. Mgmt. Corp., 830 F.2d 404, 410 (1st Cir. 1987)); see also CESARE BECCARIA, ON CRIMES AND PUNISHMENTS 22 (Henry Paolucci trans., Bobbs-Merrill Co. 1963) (“Let the verdicts and proofs of guilt be made public, so that opinion, which is, perhaps, the sole cement of society, may serve to restrain power and passions; so that the people may say, we are not slaves, and we are protected - a sentiment which inspires courage and which is the equivalent of a tribute to a sovereign who knows his own true interests.”); Peter A. Winn, Symposium, Technology, Values, and the Justice System: Online Court Records: Balancing Judicial Accountability and Privacy in an Age of Electronic Information, 79 WASH. L. REV. 307, 307 (2004).


84 Press-Enterprise Co. v. Superior Court of Cal., 464 U.S. 501, 508 (1984) (“Openness thus enhances both the basic fairness of the criminal trial and the appearance of fairness so essential to public confidence in the system.”).

85 LoPucki, supra note 83, at 508.

86 Richmond Newspapers, 448 U.S. at 573 (Brennan, J., concurring).

87 Richmond Newspapers, Inc., 448 U.S., at 570; see also Brown & Williamson Tobacco, 710 F.2d at 1178 (“The crucial prophylactic aspects of the administration of justice cannot function in the dark; no community catharsis can occur if justice is ‘done in a corner [or] in any covert manner.’”)(quoting Richmond Newspapers, 448 U.S. at 571); Peter A. Winn, Judicial Information Management in an Electronic Age: Old Standards, New Challenges, 3 FED CT. L. REV. 135, 137 (2009).
4. Sources of Litigation Data

a. PACER

Federal litigation data is available electronically to the public through the Public Access to Court Electronic Records, or “PACER,” an online repository for court records, docket sheets, and case metadata maintained by the Administrative Office of the United States Courts. Before there was PACER, there was paper. Litigation data was “practically obscure” in that it took far more effort and expense to locate, copy, and aggregate case files when they were stashed away in filing cabinets buried in dimly lit basements of federal courthouses across the country. Not surprisingly, empirical studies of district courts were relatively modest in size, few, and far between during the infancy of the Internet. Litigants had little fear that their personal affairs documented in court records would leak into the public domain, notwithstanding the fact that these documents were technically part of the public record. And district courts were free to manage their dockets with minimal scrutiny from the public.

That all changed with the creation of PACER. The Judiciary’s switch from paper to electronic case filings made litigation data available to


89 Justice Stevens appears to have coined the term “practical obscurity,” which he used to refer to the inaccessibility of paper documents due to the inconvenience associated with acquiring access to them. U.S. Dep’t of Justice v. Reporters Comm. for Freedom of the Press, 489 U.S. 749, 763–64 (1989); see generally Peter A. Winn, Judicial Information Management in the Electronic Age: Old Standards, New Challenges, 2009 FED. CTs. L. REV. 135, 152 (2009) (describing the practical barriers to access associated with paper-based practical obscurity).


anyone with a computer, Internet connection, and credit card. Practical obscurity had suddenly become a relic of the past. As discussed below, however, the Judiciary has managed to recreate practical obscurity to a degree by erecting PACER’s pay wall, which makes practice-focused empirical analysis of bulk litigation data prohibitively costly.

“Those tracing the history of PACER date its birth in 1990, when an appropriations act authorized the federal judiciary to build a system furnishing remote access to court records, to be supported by funds generated by access fees.” The federal Judiciary’s shift from keeping paper records to electronic records resulted “an explosion of use [of PACER]; . . . . There were 20,028 user accounts in 1995, 39,408 in 1999 and 270,000 in 2003.”

Initially, observers heralded PACER as “a huge improvement over the existing system of paper records,” “tremendously ahead of its time,” and “one of the great success stories of the federal Judiciary.” The enthusiasm with which users initially greeted PACER is not surprising considering the old method of inspecting court records, which entailed traveling to the courthouse in which the desired documents were located, thumbing through boxes of dusty files, and photocopying them for a fee of fifty cents per page. Despite the Judiciary’s quantum leap from paper to pixels roughly two decades ago, PACER is beginning to show signs of aging. Critics now complain that the service imposes arbitrarily high access fees; provides

---

93 Peter W. Martin, Symposium, The New “Public Court”: Online Access to Court Records—From Documents to Data, Particulars to Patterns, 53 VILL. L. REV. 855, 860 (2008).

94 Id.


restrictive, inadequate search capabilities, and has an outdated user interface that renders the data largely inaccessible.

The most common complaint about PACER appears to be directed toward the service’s access fees. Rather than subsidizing PACER’s creation and maintenance with taxpayer dollars, Congress authorized the Administrative Office of the United States Courts to charge reasonable user fees for access to electronic litigation data. To access court records through PACER, members of the public must therefore register an account with the site and provide credit-card information.

The service currently charges users ten cents per page to view and download PDFs of public court records. The service also charges ten cents per search, regardless of whether the search results turn up relevant cases. While ten cents per page may seem like a trivial fee to most users, it makes compiling a sizeable dataset for empirical analysis prohibitively expensive. And ten cents per page seems far less reasonable when

---


105 Id.

106 To PACER’s credit, users are provided with free access to PACER data so long as their account does not accrue charges of more than $15.00 in a quarterly billing cycle. Id. The service also provides free access to judicial opinions, and courts may waive PACER fees for, among other user categories, individual researchers affiliated with a particular education institution. Id. That said, the steep PACER fees remain prohibitively costly to commercial ventures that might otherwise develop new tools around the data that improve the public’s access to litigation information and understanding of the courts.
compared with the cost per printed page. One article pointed out that, under PACER’s pricing scheme, “a weekday copy of the New York Times would go for $7 or $8, and Walter Isaacson’s biography of Steve Jobs would cost $65.60.”

One of the most vocal critics to speak out against PACER’s pay wall has been Carl Malamud, who has earned the reputation as being an “open access gadfly” thanks to his aggressive efforts to make digital copies of government records available to the public for free. According to Malamud, “putting the nation’s legal system behind a wall of cash and kludge separates the people from . . . the ‘operating system for democracy.’” So, when the Administrative Office and the Government Printing Office announced that they would be providing users with free access to PACER records at seventeen libraries across the country as part of a pilot program, Malamud saw the opportunity that he had been waiting for, urging his supporters to download as many free court records through the PACER pilot program as possible, and then send them to him so that he could make them available to the public for free on his website.

Steve Schultze, then-Princeton University’s Associate Director of Information Technology Policy, felt compelled to join Malamud’s “PACER liberation front,” as he would later call it, after becoming frustrated with the system’s inaccessibility. Schultze soon realized, however, that PACER houses hundreds of millions of court records, and that the number of records that he could manually download would amount to nothing more than a drop in the ocean. So, instead, Schultze created a computer

---

107 Beato, supra note 99.

108 Id.


111 Id.

112 Schultze, supra note 96.

113 Id.
program that could download bulk records from PACER and automatically save them to a thumb drive.\textsuperscript{114} Schultze gave the program to a fellow hacker named Aaron Swartz, who then proceeded to download around 2.7 million public court records from PACER for free from an account created by one of the libraries participating in the pilot program.\textsuperscript{115}

The federal government was not happy when it learned that Swartz had downloaded metadata, docket sheets, and court records for roughly 20\% of the information stored in PACER.\textsuperscript{116} Government officials promptly suspended PACER’s free-access pilot program, “pending an evaluation.”\textsuperscript{117} According to a \textit{New York Times} report on the incident, “A couple of weeks later, a Government Printing Office official . . . told librarians that ‘the security of the Pacer service was compromised,’” and that “‘[t]he F.B.I. [was] conducting an investigation’” into the matter.\textsuperscript{118} Nevertheless, the data that Swartz had downloaded from PACER found its way to Carl Malamud’s website where it remains free to download.\textsuperscript{119} The FBI eventually dropped its investigation into Swartz’s mass download,\textsuperscript{120} perhaps realizing that Swartz had broken no laws by downloading public records from PACER and reposting them to the web.

The recent efforts of Malamud, Schultz, Swartz, and countless other open-access advocates have reignited the debate over whether PACER should be accessible to the public free of charge. Sensing that this would not happen anytime soon, Schultz and a team of undergraduates from Princeton created a clever web application called “RECAP” (whose slogan is “turning PACER around”) that users can download and run from within

\textsuperscript{114} Id.

\textsuperscript{115} Id.

\textsuperscript{116} John Schwartz, \textit{An Effort to Upgrade a Court Archive System to Free and Easy}, \textit{N.Y. TIMES}, Feb. 13, 2009, at A16 (reporting that Aaron Swartz, “a 22-year-old Stanford dropout and entrepreneur . . . managed to download an estimated 20 percent of the entire database: 19,856,16 pages of text,” prompting the government’s suspension of its pilot program providing free access to PACER).

\textsuperscript{117} Id.

\textsuperscript{118} Id.

\textsuperscript{119} \textsc{BULK.RESOURCE.ORG}, https://bulk.resource.org/courts.gov/ (last visited June 12, 2013).

\textsuperscript{120} Aaron Swartz, \textit{Wanted by the FBI}, AARON SWARTZ BLOG (Oct. 5, 2009), http://www.aaronsw.com/weblog/fbi/article.
their browser when they log into PACER. For every court record that a user downloads from PACER, the program saves a duplicate copy to its own database, which it then makes available to the public for free. In return, the program alerts users when they come across court records in PACER that are available for free in RECAP’s database, thereby saving the user the expense of having to download the document from PACER. Since creating the program, Schultz estimates that RECAP has freed millions of records previously trapped behind PACER’s pay wall. As a result, startups are now able to develop and tinker with new web applications that let lawyers leverage PACER’s litigation data on a large scale at little or no cost.

b. PACER Alternatives

No single company has had a more profound impact on disseminating the rule of law than West Publishing. Established in 1879 by two brothers, John and Horatio West, the company now dominates the roughly $18 billion market for legal information. Over the last few decades, the publisher’s model for distributing legal information has changed dramatically due to technological innovations. In 1973, LexisNexis shook up West’s comfortable monopoly in the legal-publishing space by introducing the world’s first database of electronic legal texts. Two years later, West introduced its own electronic case database, and so began the race to digitize the law.

122 Id.
123 Id.
125 John W. Heckel, Questions and Answers, 75 L. LIBR. J. 299, 305 (1982).
128 Id.
The next major paradigm shift in legal publishing occurred with the widespread adoption of web browsers that provided lawyers with instant, remote access to a wealth of electronic legal information that previously required a trip to the courthouse. Westlaw and Lexis quickly moved their respective digital content to the web and began charging law firms considerable subscription fees for access. Rather than developing sophisticated natural-language search algorithms to help lawyers more effectively identify relevant sources, Westlaw simply applied its print-based topical index, called “KeyCites,” to its web-based legal content, which has since grown into a hyperlinked matrix of more than 100,000 topics and subtopics. Even today, it is clear that Westlaw still views itself as a traditional publishing company rather than a technology company. Rather than developing and refining machine-learning algorithms capable of automatically indexing new cases with a high degree of accuracy as they are added to its databases, Westlaw still employs an army of pricey legal experts to manually sift through, summarize, and classify each source before making it available online.\footnote{129}

Having fattened on the fruits of their duopoly over the last couple of decades, Westlaw and Lexis’s failure to innovate is hardly surprising.\footnote{130} Rather than investing in new features and competing on the basis of technology, the two companies have, in the past, simply used their deep cash reserves to acquire promising startups that might one day pose a competitive threat.\footnote{131}

\footnote{129}Susan Nevelow Mart, *The Case for Curation: The Relevance of Digest and Citator Results in Westlaw and Citator Results in Westlaw and Lexis*, 32 LEGAL REFERENCE SERVICES QUARTERLY 13 (2013).

\footnote{130}David Hall, *Google, Westlaw, LexisNexis and Open Access: How the Demand for Free Legal Research Will Change the Legal Profession*, 2012 SYRACUSE SCI. & TECH. L. REP. 119, 122 (2012) (“West and Lexis have struggled to keep up with the times and have been slow to react to customers’ evolving expectations with electronic research. A large part of their reluctance to change was due to their position as a firmly entrenched duopoly.”).

\footnote{131}See generally KENDALL F. SVENGALIS, LEGAL INFORMATION BUYER’S GUIDE AND REFERENCE MANUAL 8-15 (2005) (“In the last twenty years of the 20th century, the world of legal publishing has been dramatically altered as a consequence of corporate acquisitions and mergers. While small legal publishers continue to enter the fray, many leading legal publishing houses have been acquired by major international conglomerates.”). Between 1980 and 2005, Thomson Corporation had acquired over a dozen competitors in the legal-information space:

But a new breed of legal-technology startups are emerging that offer a promising glimpse into the future of legal research. A future where web applications allow attorneys to not only search for and sift through legal authority, but also to learn about the tactics of their opposing counsel, the tendencies of their judge, and other behavioral aspects of litigation that tend to influence the outcome and cost of litigation ahead of time. In fact, lawyers in certain practice areas have already begun to make important strategic decisions based on trends and patterns in litigation data revealed through their use of these services.

Leading the way in the emerging renaissance in legal research is a web-startup named Lex Machina, which is Latin for “law machine.” Lex Machina was founded in 2006 as an interdisciplinary project between the Stanford Law School and Stanford University’s computer science department to bring transparency to intellectual-property litigation. Today, attorneys handling bet-the-company patent litigation use Lex Machina’s web-based analytics service to surface trends and patterns in historical patent litigation to more accurately forecast costs and more effectively evaluate various case strategies.

The service scrapes PACER for new cases involving intellectual property on a nightly basis. Whereas Westlaw manually assigns KeyCites to the legal opinions stored within its databases, Lex Machina does so through a combination of human reviewers and a proprietary algorithm that automatically parses and classifies the outcome for each case that the service extracts from PACER. After extracting, processing, and scrubbing the data, Lex Machina visualizes aggregated case data for a particular judge, party, attorney, or law firm with analytics that allow users to quickly discern trends and patterns in the data that may affect the cost or outcome of their case.

A number of other recent startups have similarly pledged to build

---

Findlaw, one the Internet's leading legal index and information sites, and the Harrison Company. West itself had already acquired Banks-Baldwin to go with its earlier purchases of Foundation Press, the American Law Book Company, the Edward Thompson Company, and Boston Law Book, among others.

Id.

132 Formerly the “Stanford Intellectual Property Litigation Clearinghouse.”
applications that provide attorneys with “a new view on legal research.” Ravel Law, for example, is a search engine for legal opinions that displays the citation network for a given legal concept. Whereas Westlaw and Lexis stack search results in a column, often burying important cases pages back in search results, Ravel Law visually represents the most important case on a particular topic as a hub, with numerous spokes pointing to subsequent cases that have cited it. The size of the hub reflects the relative number of cases that cite it. The frequency with which courts cite a particular case often signals that case’s influence over a given area of law or relevance to a particular legal concept. Until Ravel Law’s launch in 2013, lawyers using Westlaw could not glean such insights into the Judiciary’s citation networks without manually plotting each case citation by hand. Whereas manual plotting would have taken days and perhaps even weeks, Ravel Law condenses the process into a matter of seconds.

Another noteworthy startup is Judicata, which is attempting to develop a Google-like search engine for finding relevant legal authority. The ambitious startup will attempt to compete directly with Lexis and Westlaw by offering a more refined search experience that “uses highly specialized case law parsing and algorithmically assisted human review to turn unstructured court opinions into structured data.” The company’s goal is to then “leverage that data to build legal research and analytics tools that are an order of magnitude better than existing offerings.”

The sudden explosion of technology startups focused on utilizing data to improve the delivery and consumption of legal services is remarkable for several reasons. First, for decades, the legal profession experienced very little technological disruption. While new technologies like fax, computers, electronic databases, and mobile devices enabled lawyers to become more efficient, they did not help them to become more effective. Moreover, these developments went virtually unnoticed by the profession, as they were incremental, intermittent, and only made lawyers more efficient rather than more effective. As one reporter observed, Christina Farr of Venture Beat similarly noted that, “For all the talk of innovation in Silicon Valley,” law has been the “one industry that has remained largely untouched by

133 RAVEL.LAW.COM.


135 Id.
BARGAINING IN THE SHADOW OF BIG DATA

But things are starting to change as data continues to play an increasingly important role in the decision-making process for lawyers and their clients. After returning from a legal-technology conference held in early 2014, Robert Ambrogi—a Massachusetts lawyer who “has been writing about legal technology for more than two decades” now—wrote, “I am convinced that I have never before seen a time of such intense and creative innovation in legal technology.”

The proliferation of legal-technology startups today is also remarkable because of the sudden spike in investment activity that has made it possible. For years, investors perceived the legal-technology startup market as one of the most unattractive markets in which to invest. In fact, in 2012, TechCrunch categorized startups based on their target market, and ranked them based on the amount of capital that each category of startups had raised in recent years. TechCrunch reported that legal startups were “in last place” in terms of new-venture fundraising, raising a mere $104 million in startup funding. In stark contrast to the dismal investment figures published by TechCrunch, Thomson Reuters more recently reported that an “estimated $458 million . . . has been invested in legal startups in recent years . . . with an average valuation of $4.3 million[ per startup].”

And the list of investors who are now pouring money into legal-technology startups reads like a Who’s Who of Sand Hill Road. Lex Machina, for example, has already raised a reported $6.8 million in funding so far from a pool of investors that includes, among others,

---


139 Id.


• Jerry Yang, founder of Yahoo!
• Naval Ravikant, cofounder of AngelList;
• Jeff Hammerbacher, founder and chief scientist at Cloudera;
• Joe Lonsdale, founder of Palantir Technologies and Addepar;
• Radhika Shah, founder and CEO of SocialVilla;
• Eugene Zhang, founder and president of Innospring;
• Michael (Xuecheng) Jin, cofounder of Simple Silicon and Magnision Technology;
• Costanoa Venture Capital;
• Cue Ball Capital;
• Stanford University;
• Weiyiing Ding, partner at TEEC Angel Fund;
• Miriam Rivera and Clint Korver, partners at Ulu Ventures;
• Michael Borrus, Robert Siegel, and Alan Chiu, venture capitalists at XSeed Capital;
• Nick Adams, the managing director of AMECLOUDVC;
• Lenny Mendonca, the director of McKinsey & Company;
• Bill Tobin, a partner at Strayer Consulting Group;
• Chris Byrne, vice president of intellectual property strategy at Samsung Innovation;
• David Chao, cofounder and general partner at DCM;
• Richard Chen, trustee of the U.C. Berkeley Foundation;
• Dan Cooperman, a partner at Bingham McCutchen;
• Eric Goldman, a law professor at Santa Clara University School of Law;
2014] BARGAINING IN THE SHADOW OF BIG DATA 33

• Mark Haynes, a partner at Haynes Beffel & Wolfeld;
• Dan Cooperman, an attorney and board advisor;
• Shane Johnson, managing director of PaintBrush Capital Partners;
• Brad Jones, a partner at Redpoint Ventures; and
• Jean Kovacs, president of the Northern California Harvard Business School Alumni Angels.142

Since 2012, Judicata has raised nearly $8 million in venture funding from a host of big-name investors, including:

• Peter Thiel, cofounder of PayPal, billionaire investor at Founders Fund and Clarion Capital, and former attorney;
• Keith Rabois, partner at Khosla Ventures and former attorney;
• David Lee, managing member at SV Angel;
• David Perla, managing director at 1991 Ventures; and
• Micah Siegel, an angel investor and professor of engineering at Stanford University.143

III. BARGAINING IN THE SHADOW OF BIG DATA

A. Understanding the Lawyer–Machine Symbiosis

To appreciate the power and limitations of software as a tool for gleaning insights from aggregated legal data, it is helpful to consider the application of data-crunching software to another domain: chess.144 For

142 Meet Our Team, LEX MACHINA, https://lexmachina.com/about/our-team/#prettyPhoto (last visited Feb. 13, 2014); Lex Machina, ANGEL LIST, https://angel.co/lex-machina (last visited Feb. 13, 2014). Note that some of the individuals identified in this list invested in Lex Machina on behalf of a principal, while others invested in their personal capacities. This Article includes the institutional affiliations of some of the investors who fall into the latter category solely as a point of reference.


144 A portion of this section expands upon the ideas presented in a lecture given by the cofounder of Palantir Technologies at Stanford University, during which he traced the roots of modern-day predictive analytics to early computer programs capable of beating humans at chess. See Stephen
centuries, chess was viewed as a game of human intellect; a yard-stick for measuring the human mind. People used words like “genius”\textsuperscript{145} and “prodigy”\textsuperscript{146} to describe those who had demonstrated mastery of the game. As computers became increasingly powerful, chess players began to ask whether a computer could out maneuver a human opponent in a chess match. In 1997, computer scientists at I.B.M. accepted the challenge by “teaching” a computer, dubbed “Deep Blue,”\textsuperscript{147} to play chess. The scientists trained Deep Blue by feeding it loads of data from previous matches played by chess grandmasters.\textsuperscript{148} When Deep Blue faced off against Garry Kasparov later that year, the computer seemed to predict the grandmaster’s every move,\textsuperscript{149} beating him in what marked the first time in history that the world’s best chess player had lost to a computer under tournament conditions.\textsuperscript{150}

As computers became increasingly powerful over the years that followed, the question of whether computers could outperform humans in processing information had become rather mundane. Instead, people began to wonder whether the combination of humans and computers working together could outperform a super computer working alone. In 2005, the online chess website playchess.com tested this theory by hosting a “freestyle” chess tournament in which players of varying levels of skill formed teams that could use any resource at their disposal, including computers running chess software.\textsuperscript{151} To the shock and amazement of the chess world, two amateurs from New Hampshire using off-the-shelf

\begin{itemize}
  \item Referring to Kasparov’s 1997 match against Deep Blue, IBM’s website boasts, “Kasparov isn’t playing a computer, he’s playing the ghosts of grandmasters past.”\textsuperscript{148} FREQUENTLY ASKED QUESTIONS: DEEP BLUE, IBM, https://www.research.ibm.com/deepblue/meet/html/d.3.3a.shtml (last visited Apr. 27, 2013).
  \item Kasparov, \textit{supra} note 147.
  \item Id.
  \item Dark Horse ZackS Wins Freestyle Chess Tournament, CHESS NEWS (June 19, 2005), http://en.chessbase.com/home/TabId/211/PostId/4002461.
\end{itemize}
computers and chess software took home the $20,000 first prize, handily defeating some of the world’s best chess players, equipped with the world’s most powerful chess software.\textsuperscript{152} So what set them apart from the best players in the world equipped with the most powerful chess software in world? As Garry Kasparov would later explain, “Their skill at manipulating and ‘coaching’ their computers to look very deeply into positions effectively counteracted the superior chess understanding of their grandmaster opponents and the greater computational power of other participants.”\textsuperscript{153} In other words, the average players were able to leverage their computers to a much greater degree than their higher-ranking opponents because they were keenly aware of the limitations of relying on their own personal experiences and intuition, yet also recognized the human component that is so vital to leveraging the computational strength of computers when making decisions.\textsuperscript{154}

While this insight may have seemed novel and profound to the world of chess at the time, decades earlier, professor J. C. R. Licklider described the working relationship between humans and computers that explains the chess amateurs’ success,

> In the anticipated symbiotic partnership, men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions in technical and scientific thinking. Preliminary analyses indicate that the symbiotic partnership will perform intellectual operations much more effectively than man alone can perform them.\textsuperscript{155}

What is remarkable about this article is that, despite the dramatic changes in technology since its publication, the administrative tasks that eat up valuable time and Licklider’s vision for how computers might eliminate such waste remain just as prescient today. Licklider offered the following

\textsuperscript{152} Id.

\textsuperscript{153} Kasparov, \textit{supra} note 147.


\textsuperscript{155} J. C. R. Licklider, Man-Computer Symbiosis, HFE-1 IRE TRANSACTIONS ON HUMAN FACTORS IN ELECTRONICS 4, 4 (1960).
description of how he went about a typical work-day in trying to solve a complex problem:

About 85 per cent of my “thinking” time was spent getting into a position to think, to make a decision, to learn something I needed to know. Much more time went into finding or obtaining information than into digesting it. Hours went into the plotting of graphs, and other hours into instructing an assistant how to plot. When the graphs were finished, the relations were obvious at once, but the plotting had to be done in order to make them so.\(^{156}\)

Most lawyers have probably experienced similar frustrations when trying to discern winning arguments from losing arguments by reading dozens of opinions over the course of several days or even weeks at the client’s expense. Even more lawyers likely do not even bother, opting to go with their gut instead. Consider, for example, a hypothetical attorney who is considering whether filing a motion to dismiss early in litigation would be a prudent decision. Before an informed prediction about the likely cost or result of such a motion can be made, the attorney must perform a number of mechanical tasks better left to a computer. Such tasks might include searching Westlaw for similar cases, noting the arguments made in each case, and then plotting the outcomes of each case to determine the likelihood that the presiding judge will grant their motion to dismiss, which could take hours or even days to accomplish. Only then will the attorney be in a position to make a well-informed prediction about the likelihood of success, forecast the fees and costs associated with preparing such a motion, and communicate these considerations to their client. At this point the attorney will have already exhausted a considerable amount of resources, raising the stakes of the decision.

Unfortunately, as the above hypothetical illustrates, it currently takes far too long to search for and compile the information needed to make an informed prediction about the best course of action. Services like Lex Machina present a solution to this problem by relying on lawyers, judges, and scholars to do what they do best—coming up with hypotheses (for example, the United States District Court for the Eastern District of Virginia is the most “Plaintiff-friendly” forum for disability discrimination lawsuits)—while letting computers do what they do best—testing these hypotheses against a large collection of historical data (for example, “of the

\(^{156}\) Id.
+20,000 disability discrimination cases decided in federal courts over the last twenty years, the Eastern District of Virginia has decided the highest percentage in favor of plaintiffs”). Note that, in the lawyer–computer symbiosis illustrated in the example above, it is the attorney, not the computer, who must apply his or her domain expertise to determine which questions are important to their client’s case, which data might contain answers to such questions, and which outcomes should be considered “favorable.”

B. How Big (Legal) Data Changes the Game

As mentioned in the Introduction, we believe that big data differs qualitatively from the technological changes of the last few decades. Word processors, photocopiers, billing software, email, and online case databases mostly streamlined tasks that lawyers were already doing. Lawyers were sending notes and letters before email, making carbon copies before Xerox, and drafting legal documents long before WordPerfect. A few tasks gave way to complete automation, but most merely became more efficient, more replicable, and used less space.

Despite this electronic streamlining, the approach to lawyering was much the same. Legal rules dominated the approach to practice. Legal advice would center on the rules and their mandated outcomes—case law, statutes, and regulations. Legal analysis meant applying the law to the facts, making comparisons, distinctions, analogies, and normative arguments. Legal predictions were assessments of the probability that a client’s situation—set of facts—would trigger a particular legal rule, thereby compelling a given outcome. Legal advice instructed clients how to work within or around the rules. Where the rules conferred broad discretion on judges or regulators, lawyers had to cope with uncertainty.

Most cases settle before trial, and most transactions do not result in litigation. Even so, negotiations and bargaining, whether over contracts, plea bargains, or lawsuit settlements, have traditionally taken place in the shadow of the trial, as most cases are resolved in lead up to trial based on the parties and lawyers’ expected outcome at trial, discounted by the money that they expect to save by not having to proceed to trial.\textsuperscript{157} The

technologies that lawyers embraced still functioned within this paradigm. Online case research, conference calls, drafting documents on laptops, scheduling with smartphones—still takes place in the shadow of the trial.

Big data moves lawyers away from the shadow of the trial and in to the shadow of data. Software designed to surface trends and patterns in legal data as described above provide attorneys with powerful tools for gleaning relevant, practical insights into dynamics that affect litigation beyond the black-letter law. A judge’s tendency to grant or deny certain motions may tell the lawyer more about her next decision than the statutory verbiage governing the circumstances. Similarly, knowing which judges or jurisdictions, if any, move dockets along quickly, micromanage settlements or discovery, love or eschew special masters—in the last year or last decade—are trends that lawyers can now obtain and consider. Big data yields clear behavioral profiles, charted across time, of opposing parties, opposing counsel, judges in an alternate forum, and so on. Discretionary rulings shed their uncertainty and become a matter of pattern matching and predictions. Big data changes the approach to lawyering, because it adds to decision making a large dimension besides the legal rules. Of course, the computer will not tell the lawyer what to do—but it completely changes the context for strategic decision-making.

When lawyers refer to relying on their “instinct” or “experience,” we assume that they are not talking about the legal rules. Lawyers are not claiming that they trust hunches about statutes or case precedents that they could simply look up, but rather their hunches about the leanings of a particular judge or jury, the resolve of the opposing party or lawyer, or the salesmanship of a witness. Such intuitions are experience-based inferences about the tendencies of certain individuals to behave in certain ways, drawn from the lawyer’s personal observations or discussions with his or her colleagues. In other words, in addition to looking up legal rules to ascertain the mandated disposition of an issue, lawyers also engage in unscientific profiling of the other characters involved in a matter. Therein lies the main value, both to the client and to prospective lateral-hire employers, of legal experience—the rather intangible knowledge about how things work and how key players in a legal matter behave.

Up to now, technological advances in law practice had almost nothing to do with this component of lawyering, that is, the lawyer’s experience-based inferences, hunches, or predictions. Instead, rolodexes moved to databases, bookkeeping moved to spreadsheets, meetings moved to conference calls, and so on. Certain manual tasks gave way to automation. The software enabling lawyers to make decisions based on data differs from previous technological changes in that it directly relates to experience-based inferences, except that a lawyer can glean insights in minutes that previously would have taken years of observations and personal interactions to acquire.

C. Coasean Bargaining in the Shadow of Big Data

A feature of big data completely ignored in the legal and social science literature up to now is its potential to affect bargaining and negotiations by lowering information costs and information asymmetries. Nearly everything written about big (legal) data has been in either marketing material158 or policy literature discussing privacy concerns.159

From a Coasean160 standpoint, lower transaction costs make bargains more likely—and make legal rules in the surrounding environment less important for the eventual outcome.161 Information costs and asymmetries

---


160 See R.H. Coase, The Problem of Social Cost, 3 J.L. & ECON. 1 (1960). Coase’s article helped launch the law and economics movement in the legal academy; citations to it have become ubiquitous in the academic literature. For an excellent overview of the subsequent literature and Coase’s reaction to the impact of his article, see Daniel Farber, Parody Lost/Pragmatism Regained: The Ironic History of the Coase Theorem, 83 VA. L. REV. 397 (1997).

161 The Coase Theorem posits roughly that legal rules or rights matter least where parties have the most opportunity to negotiate; conversely, rules and rights matter more when parties have less opportunity to bargain around the laws. This is my paraphrasing; Coase himself did not provide a one-sentence version of his argument in his original article, and he attributed the moniker “Coase Theorem” to the economist Joseph Stigler. See R.H. COASE, THE FIRM, THE MARKET, AND THE LAW 157 (Univ. Chi. Press, 1988).
are primary transaction costs for all forms of lawyering outside the courtroom\textsuperscript{162}—contract negotiation and drafting, settlement talks, plea bargaining, and the full range of transactional work.\textsuperscript{163} Lower transaction costs make it easier to negotiate under, or around, the legal rules.\textsuperscript{164} The Coase Theorem suggests\textsuperscript{165} that bargains become more likely to occur as transaction costs decrease,\textsuperscript{166} and numerous scholars have identified information costs as a primary category of transaction costs overall.\textsuperscript{167} Big data lowers information costs and information asymmetries between

\textsuperscript{162} See Eugene Kontorovich, The Constitution in Two Dimensions: A Transaction Cost Analysis of Constitutional Remedies, 91 VA. L. REV. 1135, 1150 n.29 (2005) (“Transaction costs, broadly defined, sometimes purchase real benefits. Some types of transaction costs-- such as the costs of negotiating contracts, which involve obtaining information about the value of the thing being contracted for--allow prospective purchasers to evaluate whether the transaction should proceed.”).

\textsuperscript{163} See George M. Cohen, The Fault Lines in Contract Damages, 80 VA. L. REV. 1225, 1242 (1994) (“Transaction costs involved in making contracts include costs of collecting information, negotiating, and drafting.”).

\textsuperscript{164} As Coase observed, “It is always possible to modify by transactions on the market the initial legal delimitation of rights. And, of course, if such market transactions are costless, such a rearrangement of rights will always take place if it would lead to an increase in the value of production.” Coase, supra note 160, at 15.


\textsuperscript{166} See also Johnathan R. Macey, Transactional Costs and the Normative Elements of the Public Choice Model: An Application to Constitutional Theory, 74 VA. L. R. 471, 478–79 (1988) (pointing out that the transaction costs don’t actually have to be that low—a purely rational actor will invest $99.99 to get $100 from the government, while a more “rational” rational actor will still spend a good deal of money). Macey also points out that this cost is highly inefficient—it’s money spent to transfer wealth, rather than money spent to generate wealth, and represents a “deadweight cost”. Id.

parties; bargains are therefore more likely to occur. Bargains become easier to initiate and consummate.

For example, in the context of litigation and settlement negotiations, parties are far more likely to reach an agreement to settle before trial – thereby saving both sides the costs and anxiety of litigation – if each side has better information about their opponent.\(^\text{168}\) This would include information about the opposing party and their litigation history, as well as opposing counsel and their track record and patterns of decisionmaking, all of which are now available through big data. Particularly important would be information about previous settlement agreements—whether involving the opposing party, their counsel, or a related nexus or *casus belli* of litigation, and this information is unavailable through the conventional Westlaw and Lexus services. Professor Ben Depoorter has recently argued that information about previous settlements is supremely important for those trying to settle a new case before trial,\(^\text{169}\) and the data analytic firms provide invaluable information in this regard. Even post-verdict settlement negotiations, which seem to occur more frequently than previously thought,\(^\text{170}\) depend on information and are, in turn, a valuable source of information for future litigants.\(^\text{171}\) Settlement negotiations are a classic case of Coasean bargaining, and big data facilitates such negotiations by providing both parties with better information. Not only do the parties have less uncertainty to interfere with agreements, their positions are likely to move closer together as each side has a more realistic picture of the case and its likely outcome, meaning there is less disagreement to keep them apart.\(^\text{172}\) Lawyering with big data means that cases are more likely to settle overall, and that parties should reach agreements more quickly.\(^\text{173}\)


\(^{171}\) See id.

\(^{172}\) See Steven Shavell, *Any Frequency of Plaintiff Victory at Trial Is Possible*, 25 J. LEGAL STUD. 493, 493–94 (1996) (explaining that cases are less likely to settle and more likely to go to trial when the parties disagree on the likely outcome).

Of course, an increase in settlement rates due to Coasean bargaining can present downsides as well. To the extent that settlement lowers the expected total cost of litigation, marginal plaintiffs (especially those with smaller claims) are more likely to bring suit—settlements greatly decrease the number of trials but may also increase the number of claims brought. Thus, the Coasean bargaining benefits of big data, in the context of litigation settlements, will have a partial offset as marginal plaintiffs are more likely than before to initiate new lawsuits.

Litigation is not the only arena for Coasean bargaining. Property transactions also depend on low bargaining costs, and information costs have historically been one of the most prevalent transaction costs for agreements to transfer real estate or intellectual property. Coasean bargaining occurs regularly between city planners and developers, and the information now available through big data can significantly facilitate such bargaining—again, by reducing information costs for both sides. Energy trading is a growing area of transactional work, and information costs have traditionally been barrier to Coasean bargaining in this area as well; big data should lower these costs and streamline such transactions. Any transaction complex enough to require lawyers to consummate the deal becomes easier in the shadow of big data—valuations become more

Litigation as a Curative Tool, 4 DREXEL L. REV. 41, 77 (2011) (“Regulators can act more quickly and effectively with full information on the number of lawsuits, the settlements, and information revealed in pretrial discovery.”); Charles Silver, “We’re Scared To Death”: Class Certification and Blackmail, 78 N.Y.U. L. REV. 1357, 1378 (2003) (“Settlement values can also change quickly in response to new information.”).


176 See David Schleicher, City Unplanning, 122 YALE L.J. 1670, 1682 (2013) (discussing Coasean bargaining between city planners and developers).

177 Troy A. Rule, Property Rights and Modern Energy, 20 GEO. MASON L. REV. 803, 833 (2012)(discussing Coasean bargaining and property rights in the energy arena, and mentioning information costs as a traditional obstacle.).
accurate and harmonized between parties, and prospective buyers and sellers can retain the benefits of arms-length transactions while gaining the assurance of information that previously came only through familiarity.

A more subtle effect of big data on lawyering, from a Coasean perspective, is that legal rules or procedures governing the situation have less import—a corollary of the Coase Theorem.178 As big data facilitates bargaining by lowering information costs and asymmetries, a side effect is the dilution of various legal rules surrounding the subject of negotiation. Transaction costs—the things that prevent people from agreeing automatically—are what animate the legal rules, or give them their verve.179 The Coase Theorem implies that ex ante legal rules and entitlements matter less when parties can easily transact around them.180

In the real world, Coase emphasized,181 transaction costs vary greatly but are never zero. The import of the relevant legal rules or rights would therefore correlate with the transaction costs present in a given situation.182 Where transaction costs are low, bargaining becomes easy, and the legal rules exert less control over the eventual allocation of resources. As big data provides parties with significantly more information and reduces uncertainty about other parties, trends in the courts, and even external

178 See Dru Stevenson, Jury Selection and the Coase Theorem, 97 IOWA L. REV. 1645, 1654–58 (2012); Russell Korobkin, The Status Quo Bias and Contract Default Rules, 83 CORNELL L. REV. 608, 623 (1998) (“Economic analysis thus suggests that default rules matter in only a subset of the total number of contractual relationships. If transaction costs are low and information is distributed symmetrically between the parties, they will negotiate efficient contracts regardless of the content of default rules.”).

179 See Thomas W. Merrill & Henry E. Smith, Making Coasean Property More Coasean, 54 J.L. & ECON. 77, 78 (2011) (coining the phrase “Coase corollary” for the notion that “in a world of zero transaction costs, wealth maximization will occur regardless of the nature and scope of property rights”); see also Coase, supra note 160, at 15–19; R.H. COASE, THE FIRM, THE MARKET, AND THE LAW 178 (1988) (“The same approach which, with zero transaction costs, demonstrates that the allocation of resources remains the same whatever the legal position, also shows that, with positive transaction costs, the law plays a crucial role in determining how resources are used.”).


182 See Merrill & Smith, supra note 179, at 92–99; Stevenson, supra note 178, at 1654–58.
transaction costs (such as the length of time typically involved in enforcing an agreement in the courts), a primary hurdle to agreements disappears.\footnote{\textit{\textsuperscript{183}} See Joseph Blocher, \textit{Institutions in the Marketplace of Ideas}, 57 DUKE L.J. 821, 839 (2008) ("Coase . . . divided transaction costs into four categories: search, information, negotiation, and enforcement. Contrary to the neoclassical model's assumptions of perfect information and costless exchange, Coase recognized that these costs distort the market, making the mathematical predictions of neoclassical models largely inapplicable in the real world. Every transaction cost, he realized, is a small market failure.".).}

In sum, an important implication of lawyering in the shadow of data is that the newly available information should facilitate more bargaining, agreements, settlements, and other legal transactions. As bargaining or consummating deals is an important part of what lawyers do, big data should prove incredibly useful for practitioners who strike the right balance between intuition, personal experience, and behavioral insights gleaned from aggregated legal data. At the same time, a side effect of the reduced information costs is that it is also easier to bargain around certain legal rules and assignments of rights, which could have important policy implications beyond the scope of this Article.\footnote{\textit{\textsuperscript{184}} See, e.g., Barry E. Adler & Ian Ayres, \textit{A Dilution Mechanism for Valuing Corporations in Bankruptcy}, 111 YALE L.J. 83, 148 (2001) ("Where markets are thick, information full, and transaction costs low, either markets or structured negotiation backed by judicial determination will allow a financially distressed firm to reorganize easily in a manner consistent with absolute priority."). For a similar application of transaction cost analysis to policymaking, outside the arena of big data, see Guy Halfteck, \textit{Legislative Threats}, 61 STAN. L. REV. 629, 701 (2008) ("Group organization increases the likelihood of group-wide compliance and renders legislative threats more effective as catalysts of reform. Moreover, the tendency towards organization reduces the transaction costs of regulatory bargaining, consequently enabling legislators and groups to share information, negotiate, and design superior regulatory measures.").} Surprisingly, the Coasean implications of big data have so far received zero attention in the legal academic literature, the business academic literature, and the social science literature.

D. Ambulance Chasing and Big Data

A recent article demonstrated how mapping data from social networking sites precisely traced the impact of an east coast hurricane on individual neighborhoods.\footnote{Taylor Shelton, Ate Poorthuis, Mark Graham, and Matthew Zook, \textit{Mapping the Data Shadows of Hurricane Sandy: Uncovering the Sociospatial Dimensions of 'Big Data,'} GEOFORUM (forthcoming) (January 16, 2014), available at SSRN: \url{http://ssrn.com/abstract=2379966} (last visited march 3, 2014).} The Center for Disease Control can monitor Google search trends to track the spread of seasonal flu outbreaks.\footnote{Mohammad Tufail Malik, Abba Gumel, Laura H. Thompson, Trevor Strome and Salaheddin M. Mahmud, \textit{"Google Flu Trends" and Emergency Department Triage Data Predicted the 2009}
are well-known instances where researchers knew what they were looking for. Big data, however, can also show analysts what to look for in the first place, as patterns and correlations emerge that nobody would have guessed. As an example of the more conventional use of big data, plaintiffs’ lawyers who have heard about a major chemical spill or product recall announcement can use data culled from social networking sites or search engines to find as many of the victims (potential clients or witnesses) as possible. This mimics what lawyers in previous generations would have done with newspaper notices and short television ads.

A more novel approach, however, would be to use big data to identify mass torts (or even serial tortfeasors, such as a negligent contractor or architect, that are yet unknown. For example, one shipment of products to a particular store may have been a defective batch, resulting in localized or clustered injuries that are unlikely to attract media attention, or a particular scammer may have a spate of victims within a short period. In other words, big data could help identify tortfeasors as well as victims. Lawyers themselves could be the subject of such searches – big data will eventually help state disciplinary authorities uncover unauthorized practice of law by

---


188 Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock and Duncan J. Watts, Predicting Consumer Behavior With Web Search, 107 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 17486 (2010).
non-lawyers or foreign lawyers (which often goes undetected now), legal malpractice, and other forms of misconduct. The availability of rich data can yield metrics to measure lawyer competence more objectively that we do at present.

On the defense side, it is probably unnecessary to suggest that lawyers could utilize big data to identify tortfeasors or victims, as the insurers and manufacturers that retain defense counsel are generally years ahead of the legal profession in the use of big data and analytics. Presumably, many insurers and large companies are already using big data to predict the location and scope of upcoming liabilities, just as police use it to predict the neighborhoods that need to most patrols. Lawyers could add to this analysis by using data analytics to predict what incidents will turn into class actions rather than individual lawsuits, and how much time defendants probably have before they face actual liability.

The use of big data by potential defendants (insurers and tortfeasors) leads to another likely development – the impact of data availability on liability itself. Companies now have more information than before about their past, current, and future customer, making misuses of their product, as well as injuries from defects, much more foreseeable. This enhanced ability to foresee consumer injuries directly affects the duty of care, the duty to warn, the duty to recall products, and so on. It is also easier to trace lines of causation, both before and after an accident. In addition, online purchasing directly from manufacturers already eliminates many of the jointly liable co-defendants of the past, that is, retailers and distributors.

IV. NORMATIVE CONSIDERATIONS

So far, this Article has explored how summarizing trends and patterns across historical data allows individuals to make better predictions about various measurable aspects of litigation. The ability to predict litigation outcomes in particular rests on the assumption that such outcomes are not entirely random. It is hardly controversial to suggest that litigation outcomes are not entirely random given the important role that precedent plays in guiding judicial decisionmaking. Historical litigation data lends itself to predictive modeling because judges, particularly district court judges who cannot “make” new law, are generally bound by the decisions of their appellate-level predecessors. This principle is often referred to as “stare decisis,” which is Latin for, “To stand by things decided.”

A. The Predictive Value of Precedent

The Judiciary’s preference for predictable decisions serves several salutary interests that are vital to the preservation of a healthy legal system. First, basic fairness dictates that judges reach similar results in similar cases. Predictable decisions “foster[] reliance on judicial decisions, and contribute[] to the actual and perceived integrity of the judicial process.” Moreover, a judge who delivers rulings that are consistent with his or her previous rulings reduces legal uncertainty, which, in turn, allows individuals and organizations to plan their affairs with little fear of unanticipated legal consequences. The absence of predictable outcomes, on the other hand, stifles personal growth and economic progress. For these reasons and many more, courts have a vested interest in deciding issues in a manner that is consistent with precedent. This, in turn, enhances the predictive value of historical trends and patterns gleaned from litigation data analytics.

The fact that laws change over time does not strip historical litigation data of its predictive value. While some laws do indeed change, the change tends to be incremental, thereby reducing the number of outlier decisions that defy precedent.

B. Self-Fulfilling Prophecies and Self-Defeating Predictions

Even forward-looking forecasts based on highly predictive historical data, however, can, and do, fail for a variety of reasons worth exploring. Predictions can fail due to people confusing signals and noise in the data


191 Payne v. Tennessee, 501 US 808, 828 (1991); see also Burnet v. Coronado Oil & Gas Co., 285 U. S. 393, 406 (1932) (Brandeis, J., dissenting) (arguing that, “in most cases,” it is more important “that the applicable rule of law be settled than it be settled right”).


194 Andrew Stranieri & John Zeleznikow, Knowledge Discovery from Legal Databases—Using Neural Networks and Data Mining to Build Legal Decision Support Systems, in INFORMATION TECHNOLOGY AND LAWYERS 81, 82 (2006).
set; others fail because of misinterpreted lines of causation, still others because of unique, unpredictable “black swan events”; and some predictions become self-cancelling or self-fulfilling as richer information becomes available to a wider range of actors. Noise-signal problems are not unique to law, and are an ever-present hazard for any superficial or amateur attempt at data analysis.

Particularly problematic for law, however, are the self-fulfilling and self-nullifying tendencies of certain predictions. For example, there are instances where statistical predictions related to law or enforcement become self-fulfilling or reinforcing (as when police use statistics to identify the high-crime neighborhoods, shift resources to that neighborhood, and then find that the arrest numbers, and hence the “crime rate,” increase or get worse). This is a type of feedback effect—an informational input for decision makers skews the decisions toward more of the same information.

In contrast, other cases illustrate how legal data becomes self-negating, as when a court with a record of favoring plaintiffs thereby attracts plaintiffs with increasingly meritless cases, so eventually numbers revert to the mean. This is a type of rebound effect, where a decision maker’s informational input affects the decisions in such a way as to make the information an inaccurate predictor. In other words, a disparity with strategic significance for parties, once widely known, will tend to attract marginal parties for the side favored, and repel marginal parties for the disadvantaged side, so that outcomes eventually even out over time. Statistician-turned-columnist Nate Silver points to drivers’ use of GPS traffic information to find the least congested route across Manhattan as one example of self-defeating predictions; when too many drivers follow the data about the least-congested thoroughfare, that route quickly becomes the most congested.

---


198 The “judicial-hellhole” phenomenon—where a streak of plaintiff win creates a type of litigation gold rush—seems to reflect a bandwagon benefit, at least in the short term, but becomes self-nullifying over time due to the self-selection or screening effects of marginal plaintiffs and defendants.

199 SILVER, supra note 195, at 219.
Assuming that predictions become easier and more accurate with the advent of accessible data, a crucial question for the legal system is, which predictions carry the risk of becoming self-fulfilling, and which could become self-defeating. The answer seems to depend primarily on two variables: information asymmetries and binary strategic incentives. Returning to Nate Silver’s work, apparently the primary factor determining which way a prediction will tip is whether the individual decision maker seeks “bandwagon benefits” or, in contrast, hopes to gain an insider advantage. Politicians hope to gain bandwagon benefits in the form of additional votes by courting public endorsements from influential celebrities. One study showed that, in 2008, President Barack Obama gained as many as 1 million additional votes thanks to the “Oprah Effect” alone.\footnote{See generally Craig Garthwaite & Tim Moore, Can Celebrity Endorsements Affect Political Outcomes? Evidence from the 2008 US Democratic Presidential Primary, 29 J.L., Eco., & Org. 355 (2012).} Savvy politicians know that elections are won and lost based on crowds lurching in their direction; the most effective politicians know how to garner support by jumping on a bandwagon that has momentum.

Infected patients (Silver’s second example) are better off when there is a higher chance of a doctor’s visit yielding a quick diagnosis and readily available treatment, which is more likely when the disease is a case that the physician is seeing frequently, as in a publicized outbreak or minor epidemic. Drivers, in contrast, have the exact opposite utility function—they want the road less traveled, or at least the one less congested at that moment. For litigants, the same dichotomy is also supremely important in determining which way legal predictions will tip. These examples highlight how strategic incentives—especially in instances with binary choices or two-sided competitions—can turn informational richness into a self-cancelling or self-fulfilling phenomenon.

Yet information asymmetries also play a critical role in deciding the reliability of predictions, especially in determining whether a prediction will trend at all toward self-fulfillment or self-defeat, or whether the original prediction will prove accurate. Bandwagon benefits tend to occur where information is widely distributed among decision makers in the same “ecosystem” (for example, traffic patterns and celebrity endorsements). Private information, on the other hand, leads to information asymmetry, which changes the dynamic by eliminating crowd responses and allowing one party to exploit the prediction to the disadvantage of another. With crowd-response examples, public information is desirable for those seeking bandwagon benefits, as bandwagon effects thrive on information feedback.
loops. When incentives are more competitive, rather than collaborative, public information eliminates many chances for opportunism by the parties, as everyone is on equal footing, can glean from the same predictions, and can predict the reactions of others to the original predictions.

Litigation is inherently competitive and access to PACER data on the scale required for super crunching, however, is prohibitively expensive for many litigants. Predictions based on big (legal) data can become self-fulfilling or self-nullifying depending on the degree of information asymmetry present in a given case and the extent to which the parties’ incentives are a zero sum game. If all the parties know, for instance, that a particular judge or jurisdiction is more favorable to plaintiffs, the tendency will be for marginal defendants to avoid it if at all possible and marginal plaintiffs to seek it. This, in turn, will eventually lead to fewer plaintiff-friendly outcomes as the number of frivolous cases rises. Thus, widely available information (low informational asymmetry) coupled with an incentive to beat the crowds will normally yield self-defeating predictions as rich data becomes increasingly accessible to the masses.

Litigation is inherently competitive, while transactions are inherently collaborative—though both can have elements of exploitation/opportunism or mutual benefit. Data-driven legal predictions can become self-fulfilling or self-nullifying depending on the information asymmetries and the extent to which the parties’ incentives are a zero sum game. If all the parties know, for instance, that a particular judge or jurisdiction is more favorable to plaintiffs, the tendency will be for marginal defendants to avoid it and marginal plaintiffs to seek it, eventually pushing a reversion to the mean. Thus, widely available information (low asymmetries) combined with beat-the-crowd strategic incentives may sometimes yield self-negating predictions as rich data becomes available.

On the other hand, a type of bandwagon effect for lawyers can occur where a jurisdiction has a large body of precedents on an issue, as this benefits all parties by reducing uncertainty and fostering earlier settlements. In that case, the information becomes self-reinforcing. The widely available information in this scenario, combined with bandwagon benefits, yields self-fulfilling prophecies.

In litigation, as in military conflict, business, politics, and just about any other competitive facet of life, success often depends on access to information. Better intelligence leads to more accurate predictions and smarter decisions. In the context of litigation, this might include predictions about the strategic advantages and disadvantages of filing a lawsuit in a
particular forum, which arguments will appeal to a particular judge, the type of tactics to anticipate from the opposing counsel, the cost of litigating a case to trial, and the list goes on. Recognizing that relevant information is a valuable resource in the adversarial context of a lawsuit, law firms and their clients are willing to pay sizeable sums for access to the vast trove of information stored in Westlaw and Lexis. Some have gone so far as to accuse wealthier litigants who are able to afford access to unpublished opinions compiled by Westlaw or Lexis as gaining an unfair advantage over their adversaries without such access. 201 Thus, the predictive value of legal data in the context of litigation will sometimes depend on information asymmetries. The degree of asymmetry will, in some cases, determine whether the information remains advantageous for its users, or becomes self-defeating over time.

Information asymmetries will likely remain prevalent given the high costs required to access, aggregate, process, and analyze large pools of litigation data. 202 The last two decades have already seen a widening gap in the financial status of firms and parties, and this gap will translate into wide gaps in how much information each side is willing or able to purchase. To the extent that asymmetries prevail, the observations and predictions from big (legal) data will generally avoid becoming self-defeating or self-fulfilling, and will instead merely work to the advantage of the side with better intelligence.

From a public policy standpoint, self-negation of information can produce a social good. For example, it can produce more consistency or uniformity between judges, which serves the goal of having similar litigants obtain equal results in every court. A judge may feel concerned about her reputation if the data reveals a pattern of bias, delays, or haste, and the judge could accelerate the pace of self-correction of the prediction by modifying her own behavior. The information would self-correct eventually anyway as parties will become aware of the bias and settle before trial, or become aware of the delays or haste and seek other venues. Self-negation

---


202 See supra Part II.
could also lead to earlier settlements where defendants determine, from an actuarial standpoint, that they stand to lose more by trying their luck at trial rather than settling the matter at the outset of litigation.

C. Black Swans

Surprise events can upset predictions, especially those based on the frequency of prior occurrences. Natural and man-made disasters, if they strike by surprise, can interrupt court processes, abruptly alter public opinion (thus impacting jurors, parties, and judges), and can disrupt markets (including currency values), which can change the stakes of a given case. Even personal events that come as a surprise—a judge or lawyer having to step aside due to health problems—can nullify even a well-researched, well-analyzed prediction. Such “shocks” are ironically a familiar part of life, of course. The other side of the equation, however, is the tendency for predictions to foster systemic fragility. This is the core insight of bestselling author Nassim Nicholas Taleb, in his trilogy on uncertainty—Fooled by Randomness, The Black Swan, and Antifragile—that information abundance and the aggregate of predictive decisions can create an entirely new set of looming disasters, apart from what nature serves up.

Predictions and information present hazards in three ways: (1) decision makers tend to be overconfident when relying on forecasts, so they take more risks; (2) forecasters and decision makers tend to undervalue low-probability, high-stakes risks; and (3) over time, entire systems (institutional strategies and processes) grow up around the information, so that sudden variations can cause exponentially greater catastrophes due to the increased interdependency within the entity. Systemic efficiency—the elimination of redundancies, usually by centralized planning—makes systems more fragile, more like a house of cards. The insidious flipside of the “too big to fail” idea is that the same entities have become too big not to fail.203 Taleb contrasts “fragile” institutions and individuals with those that are “robust”—independent enough to endure systemic shocks and unforeseen events. Nature offers many examples of robustness—the ability of vegetation to regrow after defoliation, animal populations rebounding after an epidemic, and so forth. A third category—much rarer, and to Taleb, much more desirable—is what he dubs “antifragile,” that is, individuals and entities that benefit from shocks of all types. In the abstract, the legal profession is quite robust, as evidenced by its antiquity. In the American context, the legal profession has even been rather antifragile, as Taleb

would define the term, as sudden events causing social upheaval (catastrophic storms, wars, plane crashes, mass torts, crime sprees, market spikes or collapses, etc.) often create a gold rush for lawyers.

Yet the legal profession is changing in the age of big data. The availability of tools for harvesting insights from big (legal) data enables tighter coordination between parties, lawyers, and judges, as each of these players has access to far more information about the others, and can more accurately predict their responses to various courses of action. Reduced information costs, as a subset of transaction costs, fosters more bargaining, yielding earlier and more frequent agreements. Tighter coordination means greater efficiency in that the reduction in information costs, as a subset of transaction costs, encourages bargaining in a greater number of cases, at an earlier stage in the pretrial litigation process. Thus, practitioners are beginning to step outside of the shadow of trial and are instead beginning to make important value judgments about their case in the shadow of data.

The downside is that tighter coordination raises the stakes for systemic errors or systemic effects from a breakdown in any part of the arrangement. This is not a reason to eschew big data, but rather an inherent feature of it that users should keep in mind. Taleb’s main insight is not that unpredictable events happen, but that modern economic and political systems are more vulnerable to systemic shocks or disruptions because their tight coordination centers around precise, statistical predictions, allowing no room for error, play, or disruption by a low-probability event.

Taleb’s “black swan” scenarios, fragile systems, and so on are largely a response to the advent of big-data-based decisionmaking in the financial sectors (but also in the health case and educational systems, which have become increasingly complex, but also increasingly interdependent). Simply put, data—especially statistical data—fosters interdependency in complex systems, and interdependency carried the inherent potential for system-wide disruptions or catastrophes. As the legal system begins to embrace the amazing potential of big data, we expect more interdependency and tight coordination as statistical predictions furnish the basis for more decisions (by lawyers, parties, and judges). Such increased interdependency, of course, increases the vulnerability to systemic disruptions or shocks. Again, this is not a reason to avoid big data, but it is something that sophisticated users of the data should recognize, and for which they may want to compensate or hedge in some way.

D. Ethical Considerations
1. Unauthorized Practice of Law and Fee Sharing with Nonlawyer Owners

Does big data present new ethical quandaries or pitfalls for lawyers? Josh Blackman has raised several points that he believes present looming ethical problems for companies such as Lex Machina that provide data analytics to lawyers. First, Blackman predicts that legal data analytic firms, which generally have lawyers and nonlawyers among their owners, will soon provide services that constitute the practice of law, and will violate ABA Rule 5.4, which bars fee sharing between lawyers and nonlawyers. He writes, “It is only a matter of time before several players in the field who are coy about the manner in which they are structured are going to have to deal with this [Rule].”

Blackman’s worries assume that data analytic firms will 1) engage in the practice of law, as defined by state bars, and 2) will have nonlawyer owners when they do, and 3) will base their operations outside Washington, D.C., which permits fee-sharing with nonlawyers. At present, we are not aware of any data analytic firms providing anything more than a sophisticated search engine for an online archive of legal information – no more the practice of law, say, than the familiar Westlaw or Lexis database services. Westlaw and Lexis both provide limited customer service or telephone assistance with searches, which to date no one has seriously challenged as the practice of law. Blackman seems to be suggesting that (in response to inevitable market demand) some data analytic firms will go a step further and provide personal consultant or data concierge services, where a nonlawyer entrepreneur can obtain specific advice about where to file a lawsuit, whether to respond to a demand letter, and so on. We are skeptical that the data analytic firms will wade into such murky waters – their specialty, and the value-add of their service, lies entirely in their database and search features, not with operating a clumsy legal help call center where lawyers provide anonymous representation from a cubicle to each caller. Nevertheless, concerns about variations on this idea are bound to arise, so it seems appropriate to work through the issues here for the sake of argument.

Blackman later raises the unauthorized practice of law (UPL) as a separate point: “Reading graphs to offer advice on how a case should settle, or where it should transfer to, is at its heart the practice of law. That an

204 See Josh Blackman, http://joshblackman.com/blog/2014/02/18/the-looming-ethical-issues-for-legal-analytics/
205 See id.
algorithm spit it out doesn’t really matter.” As mentioned above, defining the practice of law for disciplinary purposes is a necessary antecedent for his first point about sharing legal fees with nonlawyers, so it seem more appropriate to address these issues together. Applying the UPL issue to data analytics, Blackman asserts, “Non-lawyers, or even lawyers not working for a law firm, are unable to give this type of advice. Data analytics firms should tread carefully about handing out this type of personalized advice outside the context of the attorney client relationship.”

The first step in responding to these concerns, therefore, is to address the issue of whether the activities of a data analytic firm will fall within the definition of practicing law. Unfortunately, there is no clear, much less uniform, definition for this phrase. Few, if any, state bars have a clear definition of what constitutes “the practice of law,” but certain actions have clearly fallen under this rubric, such as preparing and filing legal forms or pleadings, appearing as a client’s representative before a tribunal, and so forth.

The American Bar Association concedes there is no clear definition of the “practice of law,” but suggests functional proxies: “Although there is no single, specific definition of the practice of law, functionally the practice of law ‘relates to the rendition of services for others that call for the professional judgment of a lawyer.’” Merely providing descriptive information or analysis to lawyers or other customers, which is the current service provided by the data analytic firms, is no more the practice of law than the algorithmic Google Trends or WestlawNext, albeit with richer data and interesting graphs and charts. If a data analytic firm took the unusual step of drafting forms or filing documents on behalf of customers, outside the patent, trademark, Social Security, or immigration arenas, this would probably constitute representation and the provision of legal services.

Less clear, however, is the provision of consulting services. Jury consultants, which provide empirical-based advice about the likely biases or leanings of prospective or actual jurors, are pervasive in the legal system, though most are nonlawyer sociologists, psychologists, and so forth. Before Lex Machina or similar analytic firms begin offering services that

---

206 Id.

207 Id.

208 See Cleveland Bar Assn. v. Boyd, 859 N.E.2d 930 (Ohio 2006)(“ With few exceptions... the unauthorized practice of law occurs when a layperson prepares legal pleadings and other papers for filing in court on another's behalf without the supervision of a licensed attorney.”).

verge on the practice of law, jury consultant firms are likely to use these resources to provide additional services to lawyers, and it seems unlikely that they would thereby trigger accusations of unauthorized practice of law. The ABA has issued no opinion letters about jury consultants, and there are no published opinions indicating that jury consultants have ever faced prosecution for unauthorized practice of law.

In a published formal opinion in 2008, the ABA approved of legal outsourcing by law firms to lawyers and nonlawyers, the ABA refused to define what might constitute the unauthorized practice of law by an outsourced nonlawyer:

Finally, the outsourcing lawyer must be mindful of the admonition of Rule 5.5(a) to avoid assisting others to “practice law in a jurisdiction in violation of the regulation of the legal profession in that jurisdiction . . . .” This Committee lacks the authority to express an opinion as to whether the provision of legal services by any particular lawyer, nonlawyer, or intermediary constitutes the unauthorized practice of law. Ordinarily, an individual who is not admitted to practice law in a particular jurisdiction may work for a lawyer who is so admitted, provided that the lawyer remains responsible for the work being performed and that the individual is not held out as being a duly admitted lawyer. We note only that if the activities of a lawyer, nonlawyer, or intermediary employed in an outsourcing capacity are held to be the unauthorized practice of law, and the outsourcing lawyer facilitated that violation of law by action or inaction, the outsourcing lawyer will have violated Rule 5.5(a).²¹⁰

A survey of earlier ABA Opinions is more illuminating. The historical trend has been to find improper fee sharing with lawyers – overlapped with the unauthorized practice of law by the nonlawyers – almost exclusively in situations where the nonlawyer received contingent fees tied to the outcome of a case. For example, in a 1986 Formal Opinion about lawyers hiring outside researchers, the ABA forbid the payment of contingent fees to research corporations, but did not forbid hiring them on an hourly basis:

A corporation (the Corporation) is engaged in the business of providing legal research and analysis services to lawyers. In rendering these services the Corporation contracts only with lawyers. Its professional staff consists of law graduates, most of whom are licensed to practice in at least one American jurisdiction. The Corporation is not a professional law corporation and a substantial portion of its stock is owned by persons who are not lawyers. The Corporation has always billed on an hourly basis for its services to lawyers. It is now considering a contingent fee arrangement for situations where lawyers contemplate using the Corporation’s services on a more extensive or continuing basis in the course of more complex matters, including complex litigation.\footnote{ABA Informal Op. 86-1519 (1986)}

The ABA reached a similar conclusion in 1980 regarding lawyers who work for research firms that provide “sophisticated economic and economic-related analysis to enable lawyers to deal with problems for which such analysis is useful.”\footnote{ABA Informal Op. 1445 (1980).} The owners of the research firm were lawyers and a nonlawyer economist.\footnote{See id.} The opinion described the services in terms that resemble those that Blackman contemplates the data analytic firms will soon offer:

This organization provides to lawyers advice on the most promising lines of economic argument, names of possible economic witnesses, consultation with witnesses, conducts economic studies to assist witnesses in testimony, and prepares comprehensive memoranda analyzing the facts as they bear on the economic issues in cases. [.. .] This consulting corporation, as described, does not preempt the attorney-client relationship. The final responsibility and supervisory powers remain with the attorney who hires this corporation. Your letter requesting a Committee opinion does not reveal in what manner the corporation is paid for its services to attorneys. In order to avoid ethical violations, the consulting corporation’s compensation would have to be computed without regard to the legal fees collected by the attorneys requesting this service. Since employees of the corporation, or its agents, might be used as witnesses at trial (although this matter is unclear in your
letter) it would be a violation of DR 7-109(C)(1)(3) to arrange a contingent fee agreement. Even if the consulting corporation had no witness role in a trial, a contingent fee agreement between attorney and his non-lawyer employees is forbidden (exceptions are noted in DR 3-102(A)(1)(3)). Formal Opinion 297 states that an accountant employee of a law firm must be ‘paid a regular salary computed without regard to fees collected for legal services rendered to particular clients.’ If the corporation is not practicing law, the monies paid to it by the employing attorney are not legal fees. Thus income can be ethically shared among lawyers and non-lawyers alike.

The foregoing illustrates how the definition of “practicing law” appears to blur with the analysis of whether fee sharing with nonlawyers occurred. The dispositive question thus appears to be whether the consulting firm bills its clients on a contingent fee basis – hourly or flat-fee billing appears to avoid any violations of the Model Rules.214

---

214 On the other hand, ABA Informal Op. 1264 (1973) dealt with a corporate-owned Institute that provided legal research for other attorneys:

[The Institute] advertises that it offers 'a full research organization for attorneys' promising a work product which will be presented in a 'creative, influential and convincing manner.' This service is stated by the advertisements to be available only to attorneys and to be performed by attorneys employed by The Institute. The Institute enters into contractual arrangements based upon an estimated fee with its customers at the outset of its service and bills them from time to time with the anticipation that it will make a profit. It, of course, compensates the attorneys working for The Institute in accordance with separate arrangements made with them. The Institute states it plans no contact with clients of individual attorneys and plans to limit its advertising to journals such as The American Bar Association Journal distributed primarily, but not exclusively, to members of the Bar.

The ABA Ethics Committee concluded that this was clearly unauthorized practice of law:

At the outset it must be pointed out that the corporation in question is not a professional association or corporation of lawyers, permissible under DR 2-102(B). Rather it is nothing more or less than a typical corporation for profit, utilizing lawyers for the performance of legal services for other lawyers. Ordinarily the definition of practice of law is a matter for which this Committee declines responsibility. However, we have little hesitancy in this instance in concluding that this is a classic case of an unauthorized practice of law and that lawyer participation is thus forbidden by DR 3-101(A).

Id.
Similarly, the ABA in 1987 took a dim view of “medical-legal consultants” primarily because the fee arrangement included contingent fees:

The Consultant provides an initial ‘work product consultant report,’ usually through its ‘Medical Directors,’ consultation of its Medical Directors and, when a case so warrants, assistance to the lawyer with depositions of expert witnesses and opposing parties and at trial. It makes available expert witnesses from its ‘independent consulting staff.’ The Consultant offers services on a direct fee or contingent fee basis. The direct fee contract requires the client to pay agreed fees for certain services (including an ‘expert report charge’ and a ‘testimony service charge’ for each expert provided by the Consultant), reimbursable expenses, and all charges by the expert witnesses provided by Consultant (which are paid to the expert witnesses). The Consultant offers three contingent fee options: (1) a modified contingency fee of 20% of the total recovery under which the client pays a reduced report fee, reimbursable expenses and agreed reduced fees to expert witnesses; (2) a straight contingency fee of 30% of the total recovery under which the client pays only reimbursable expenses and no fees for services of the Consultant or the expert witnesses provided by the Consultant; and (3) a fee to ‘maximize recovery’ after a settlement offer which works like the straight contingency fee arrangement except that the percentage is negotiated and applies only to the amount of the recovery in excess of the settlement offer.\(^\text{215}\)

At present, data analytic firms will probably shield themselves from nonlawyer fee sharing or unauthorized practice of law accusations by marketing themselves to lawyers and firms (including in-house counsel at corporations), refraining from offering case-specific strategic advice or predictions, and including disclaimers about their information not constituting legal advice. Firms that want to branch out into case-specific consulting should probably package themselves along the lines of jury consultants, like DecisionQuest, Malekpour & Ball, or High Impact, which tout themselves as “full-service trial consultants.” Moreover, if a new data analytic startup were to base its headquarters in the District of Columbia (regardless of its server locations), law firms could share fees with them.

They need to avoid charging contingent fees, as that has been the historical tripwire for the ABA to find outsourcing improper.

Moreover, as mentioned above, the District of Columbia allows nonlawyer ownership of law firms, within certain parameters, as recognized by the a recent ABA Opinion:

In contrast to the Model Rule, District of Columbia Rule 5.4(b) permits “an individual nonlawyer who performs professional services which assist the organization in providing legal services to clients” to hold an ownership interest in a law firm; and District of Columbia Rule 5.4(a) permits the sharing of legal fees with such persons . . . A fee-sharing issue may arise when a lawyer undertakes the representation of a client in a matter that involves the services of another lawyer or law firm governed by different rules. For example, a lawyer in a Model Rules jurisdiction may reasonably conclude that the client requires the assistance of a specific lawyer in a District of Columbia firm, in which a nonlawyer happens to hold an ownership interest, on a matter involving federal government contracts because that lawyer is uniquely qualified in such matters. With informed client consent, the two lawyers may work together on the matter. If the requirements of Model Rule 1.5(e) are met, a typical fee arrangement in such matters is for the client to receive a single billing for the work of both lawyers. In this situation, there may be a question whether the lawyer from the Model Rules jurisdiction, by participating in this common inter-firm fee arrangement, shares a legal fee in violation of Model Rule 5.4(a) because the District of Columbia firm’s portion of the fee will presumably become part of that firm’s overall revenues, revenues from which distributions may ultimately be made to the nonlawyer who holds an ownership interest.  

If a data analytic firm with nonlawyer owners were to base their operations in Washington, D.C., law firms in other jurisdictions could share fees with the lawyers employed by the data analytic firm. Given that

\[^{216}\text{ABA Formal Op. 13-464.}\]

\[^{217}\text{See id.}\]
their services are entirely online (with some additional phone service), there
is no reason that they could not incorporate and register their business in
D.C. as opposed to Silicon Valley.

It is also worth mentioning that there is wide variation in how states
enforce the rules against unauthorized practice of law – for example, a
number of states have an Unauthorized Practice of Law Commission that
functions somewhat separately from the state bar disciplinary authority, and
may even have a different definition of "unauthorized practice of law" than
the state entity that promulgates ethics opinions or the entity that
reprimands or disbars lawyers.

2. Conflicts of Interest

Professor Blackman next identified conflicts of interest as a likely
problem for big data. “They will need to put safeguards into place to ensure
that they are not offering advice to adverse parties, or more likely, clients
who may have interests adverse to another client.” To the extent that the
data analytic firms have outside funders or venture capitalists, this would
further complicate the range of potential conflicts of interest. Moreover,
Blackman expects that data analytic firms will want to harvest data from
their clients, as well as providing information, in order to enrich their
database and enhance their algorithms. Blackman anticipates that users of
the services (customers) would find it objectionable that their search trails
would be visible – even partly or indirectly – to litigation opponents.

Presumably, user history will be invisible to other users, or
invisibility/anonymous searching would at least be an option for users, as is

Lawyers subject to the Model Rules may work with other lawyers or law
firms practicing in jurisdictions [such as the District of Columbia] with rules
that permit sharing legal fees with nonlawyers. Where there is a single billing
to a client in such situations, a lawyer subject to the Model Rules may divide
a legal fee with a lawyer or law firm in the other jurisdiction, even if the
other lawyer or law firm might eventually distribute some portion of the fee
to a nonlawyer, provided that there is no interference with the lawyer's
independent professional judgment.

Id.

218 Blackman, supra note 204.

219 See id.

220 See id.
the case with online sites like YouTube and LinkedIn. Even so, user-informed algorithms produce results from which sophisticated users could, in theory, infer information about other users, especially when one’s litigation opponent is the only other user likely to have searched for information about one’s firm, one’s client, the judge assigned to the case, and so on. Most cases even with seasoned litigators have unique combinations of parties, counsel, experts, and judges. Just as Google’s autocomplete feature in its search box reveals information about the most common searches around the world beginning with the letters typed so far, if a data analytic site were to have a similar autocomplete feature, it could reveal search queries by an opposing party.

In response to these concerns, we first offer the observation that this concern raised by Blackman runs counter to his concerns about sharing fees with nonlawyers and unauthorized practice of law, because the conflicts of interest rules will not apply unless the data analytic firms are or include lawyers engaged in the practice of law. Conversely, if a big data firm turns out to be lawyers providing legal (but mostly statistics-based) services for clients, rules about conflicts would apply but concerns about unauthorized practice would not (and sharing legal fees with nonlawyers would also be less likely).

Moving on to the rules about conflicts of interest, the ABA Model Rules of Professional Conduct cover conflicts of interest starting with the basics at Rule 1.7, then addresses special types of conflicts of interest in Rule 1.8, including the use of “information relating to representation of a client to the disadvantage of the client.”221 Both 1.7 and 1.8, however, have consent-waiver provisions with low thresholds. Data analytic firms providing services to lawyers will presumably obtain informed consent form users as part of their boilerplate terms and conditions of use. The services as they function now present no more conflict of interest than would Westlaw or Lexis providing search results to litigation opponents who use their services at the same time. Even if these firms eventually offer concierge or individualized consulting services, the potential conflicts here are any greater than a jury consultant who represents different firms who at times have been on different sides of litigation or transactions regarding certain parties. Rule 1.8(b)’s prohibition applies only to use of information that disadvantages a client, so non-injurious use of data harvested by the entity from its clients would not necessarily trigger a

221 ABA Model Rules 1.8(b).
3. Legal Malpractice

Blackman’s third ethical concern pertains to legal malpractice insurance. He offers a hypothetical of a big data firm that advises a user to file suit in a particular forum or venue, based on the favorable patterns there, but fails to account for a recent change in the law in that jurisdiction, which ensures an unfavorable outcome from now on. Blackman sees this as a potential malpractice suit, and he expresses concern that data analytic firms will not carry malpractice insurance. “In most jurisdictions, active lawyers are required to maintain some level of malpractice insurance to protect against this,” he writes.

Actually, only one state (Oregon) requires lawyers to carry legal malpractice insurance. A few states, including Texas, require lawyers to provide notice to clients that the lawyer is uninsured at the beginning of representation—a mere formality, fine print in the retainer agreement. And a few states require all LLCs and LLPs—not only lawyers, but certainly including them—to carry a minimum amount of liability insurance, usually

---

222 See Blackman, supra note _.

223 See id.

224 Jack A. Guttenberg, Practicing Law in the Twenty-First Century in a Twentieth (Nineteenth) Century Straightjacket: Something Has to Give, 2012 MICH. ST. L. REV. 415, 472 n. 339 (2012) (“Oregon is the only American jurisdiction to require malpractice insurance of all lawyers who are admitted to practice in the state. OR. REV. STAT. § 9.080(2) (2011). As of August 9, 2011, seven jurisdictions require disclosure directly to the client of the lawyer’s failure to maintain malpractice insurance. Another seventeen jurisdictions require lawyers to report their lack of malpractice insurance on their annual registration statements.”); Leslie C. Levin, Bad Apples, Bad Lawyers Or Bad Decisionmaking: Lessons from Psychology and from Lawyers in the Dock, 22 GEO. J. LEGAL ETHICS 1549, 1588 (2009) (“. . .legal malpractice insurance is not required in the United States in order to practice law.”); Carole Silver, Regulating International Lawyers: The Legal Consultant Rules, 27 HOUS. J. INT’L L. 527, 535 (2005) (noting that 17 states require foreign legal consultants to have proof of insurance to be licensed, but only Oregon requires lawyers to carry malpractice insurance); Martin C. McWilliams, Jr., Who Bears the Costs of Lawyers’ Mistakes? -- Against Limited Liability, 36 ARIZ. ST. L.J. 885, 948 (2004) (“As of this writing only one state requires lawyers to carry insurance, and in other states movements to require legal malpractice insurance have failed.”); Christel Walther, LLC and Lawyers: A Good Combination?, 50 LOY. L. REV. 359, 392-93 (2004) (“Some LLP and LLC statutes provide that the liability shield only applies if the LLP or LLC carries some requisite amount of insurance coverage. Other statutes do not condition liability protection on maintaining some level of insurance or providing evidence of financial responsibility.”); Susan Saab Fortney, Seeking Shelter in the Minefield of Unintended Consequences—The Traps of Limited Liability Law Firms, 54 WASH. & LEE L. REV. 717, 729–30 (1997).
$100,00. Thus, firms using the corporate form of LLP or LLC in those states would have to have liability insurance, but not necessarily legal malpractice insurance in particular. In general, therefore, lawyers in forty-nine states do not have to carry professional malpractice insurance.

According to a 2013 survey of the seven largest legal malpractice insurers in the country, conflicts of interest generate the lion's share of malpractice claims, with "procedural errors" and "failure to know/properly apply law" tied as distant seconds (only 1/4 as common as conflicts as the cause). The category of "litigation errors" (it is unclear how they distinguish this from "procedural errors") is the least common cause of malpractice claims that they actually report.

Let us return to Blackman’s hypothetical, in which a nonlawyer consultant gives misinformed advice about where to file a suit. Again, this depends heavily on the firms meeting some definition of the practice of law, and there is no clear definition – and legal malpractice cases against nonlawyers are extraordinarily rare. This scenario is analogous to legal malpractice suits against jury consultants – again, rare or nonexistent. Moreover, a legal malpractice action typically requires the plaintiff to prove damages and causation – that the case could have prevailed but for the defendant’s negligence. It is not clear under Josh’s example that the party could not re-file the case in another jurisdiction, which could undermine a malpractice claim.

E. Privacy

Recent academic commentary about big data has focused disproportionately on privacy concerns, and we hope to turn the discussion in another direction. We do not offer here a full-blown rebuttal to the alarmist, Orwellian predictions of others (as that would turn this article into yet another piece focused on an overdone subject), it seems appropriate to provide a brief justification for our call to move beyond the excessive hand-wringing about big data and personal privacy.225

We believe that academic fretting about privacy began midcentury as a legitimate reaction against McCarthyism, but became overblown in the 1980’s and 1990’s as an artifact of the narcissism and hyper-individualism of the culture in that era. The endless dire predictions that we were on the

---

225 For a more thorough rebuttal to scholars such as Julie Cohen, we refer the reader to Jane Bambauer, Is Data Speech?, 66 STAN. L. REV. 57, 96 (2014), whose arguments align with our own.
A similar strain of alarmist literature centers on a phobia of societal taboos being easier to enforce if people cannot hide their vices. Yet this is not as much of a problem with privacy as it is with the fact that our society’s taboos are changing, so there are divergent sets of values or taboos that groups seek to impose on each other. We have emerging taboos in our society against racism, misogyny, and other forms of bigotry, and greater transparency would help new norms of tolerance to emerge.

Our second reason for moving beyond the tired privacy discussion flows from this last point, about the new values in our society. We believe that our society cannot pull together its balkanized factions, and newly arrived subcultures, without greater transparency—anonymity and community are inherently in tension.

V. Conclusion

A seismic shift is underway in the legal profession, whereby lawyers are increasingly supplementing their intuition and practice experience with insights gleaned from big (legal) data to inform their judgment. The increasing accessibility of legal data and continued refinement of analytics for mining it will alter the skillset and approach of lawyers who embrace these tools. The analytics platforms that are facilitating this shift will not, however, supplant practicing attorneys, rather, these platforms will complement their skills, knowledge, and experience. While algorithms may be able to process and summarize trends and patterns in historical litigation data far more accurately and efficiently than their human counterparts, it is ultimately the lawyer’s job to draw on their experience and intuition to determine which trends and patterns are relevant to a given situation, and how to act on such insight.
Even so, our point here has not been to promote technophilia. Tempering our exuberance for new predictive tools should be a more reflective, nuanced approach to predictions generally—a recognition that predictions can backfire, become self-fulfilling, or can even create new systemic risks or institutional fragility. Knowing the limitations of a method or technology ultimately makes it more useful. Risks and recklessness are not the necessarily the same.

Our discussion has therefore been primarily descriptive, but also modestly normative. The descriptive sections explained the new platforms that are available for legal data analysis, and the normative sections explored the pros and cons of using these tools in practice. Most advances in technology merely made lawyers more efficient at doing the same tasks they were already doing. Big data changes the nature of making legal predictions—and thereby promises to change the role of lawyers overall.