2017

Is the Future of Law a Driverless Car? Assessing How the Data Analytics Revolution Will Transform Legal Practice

Eric L. Talley
Columbia Law School, etalley@law.columbia.edu

Follow this and additional works at: https://scholarship.law.columbia.edu/faculty_scholarship

Part of the [Law and Society Commons](#), [Legal Profession Commons](#), and the [Science and Technology Law Commons](#)

Recommended Citation

Eric L. Talley, *Is the Future of Law a Driverless Car? Assessing How the Data Analytics Revolution Will Transform Legal Practice*, JOURNAL OF INSTITUTIONAL & THEORETICAL ECONOMICS, VOL. 174, P. 183, 2018 (2017).

Available at: https://scholarship.law.columbia.edu/faculty_scholarship/2061

This Working Paper is brought to you for free and open access by the Faculty Publications at Scholarship Archive. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Scholarship Archive. For more information, please contact scholarshiparchive@law.columbia.edu.

Is the Future of Law a Driverless Car? Assessing How the Data Analytics Revolution Will Transform Legal Practice

Eric L. Talley¹

November 2017
(First Draft: May 2017)

Abstract: Machine learning and artificial intelligence technologies (“data analytics”) are quickly transforming research and practice in law, raising questions of whether the law can survive as a vibrant profession for natural persons to enter. In this article, I argue that data analytics approaches are overwhelmingly likely to continue to penetrate law, even in domains that have heretofore been dominated by human decision makers. As a vehicle for demonstrating this claim, I describe an extended example of using machine learning to identify and categorize fiduciary duty waiver provisions in publicly disclosed corporate documents. Notwithstanding the power of machine learning techniques, however, I remain doubtful that data analytics will categorically displace lawyers from the practice of law, for two reasons. First, many of the most powerful approaches in data analytics as applied to law are likely to continue to require human practitioner inputs to train, calibrate and supervise machine classifiers. And second, the underlying evolutionary process that characterizes legal doctrine and precedent is irreducibly dynamic and complex – traits that are poorly adapted to pure algorithmic decision-making. Consequently, aspiring legal researchers and practitioners should not fear entering the field, but in doing so they would be well advised to invest in skill sets that are complementary to data analytics techniques.

¹ Isidor & Seville Sulzbacher Professor of Law, Columbia Law School. Portions of this paper came from developing my thoughts for the Fifty-Ninth Annual John R. Coen Lecture, at the University of Colorado School of Law (2016) and the Keynote Address for the CELS East Asia conference in Taipei (2017). Many thanks to Michael Linneman and Hannah Song, who provided invaluable research assistance, participants in the JITE 2017 conference in Siracusa Italy for helpful discussions and to Elliott Ash and Michael A. Livermore for their substantial suggestions and insights. All errors are mine.

1. Introduction

The technologies of machine learning and artificial intelligence are quickly colonizing vast swaths of academic research and professional practice across several domains, and law is proving to be an accommodating target. Already a staple of the discovery process and citation analysis, data analytics technologies are overcoming years of motivated resistance to have significant penetration in transactional law as well. In this article, I analyze some of the ways that this transition is unfolding, I explore some the techniques that are appear to be particularly promising, and I speculate on what it portends for the study and practice of law over the longer term.

As a metaphorical “vehicle” for this exercise, I analogize the data-analytics revolution in law to the advent of the driverless care. As is now well understood, the technology that enables driverless cars² is developing rapidly, geared generally around using *data analytics* (defined more rigorously below) to assess, predict and respond to physical hazards and risks around an automobile.³ Within the next decade or so, it is widely predicted that vehicles with such abilities will be commonplace on the streets and highways of the US and the world.⁴ Its infusion into our everyday life will without a doubt carry great convenience to many. But at the same time, the driverless car portends significant dislocation to others, such as chauffeurs, truck drivers, bus drivers, taxi drivers, and even the ascendant workforce of Uber and Lyft. Society’s widespread gains will be their collective loss.

² Sometimes called the “autonomous car,” the “robotic car” or the “self-driving car,” the driverless car has been under development at Google and other companies for years. See, e.g., <https://www.google.com/selfdrivingcar/>.

³ See *Id.* at 8-13.

⁴ See, e.g., Leslie Hook and Tim Bradshaw, “Driverless cars inspire a new gold rush in California,” *Financial Times* (May 24, 2017).

Is the evolution of autonomous vehicles and their impacts—both positive and negative—an appropriate metaphor for the evolution of legal practice and scholarship? On first blush, the metaphor seems apt: much of what goes into effective lawyering consists of assessing, predicting and responding to *legal* hazards and risks that will plausibly befall one's client. Over the course of centuries, law schools have developed tools to train professionals to develop such skills as a concomitant of their professional competence and judgment. And, legal researchers have developed tools to understand and evaluate the ecosystem and production of legal actors. But can this last? Might the very same core technologies that sit behind autonomous automobiles have a similar displacing force in the legal profession? As these technologies continue to develop, are there any corollary benefits to this shift, in the same way that consumers would benefit from the plight of drivers? More specifically, the astounding advances in data analytics (and in the related but distinct subfields of machine learning, natural language processing, big data and deep learning) over the last two decades have virtually upended several brick-and-mortar industries. Is data-driven automation destined to do the same for the practice of law?

Legal scholarship and research are similarly at risk of being consumed by data analytics—both as means and end. Although quantitative analysis of law (also called empirical legal studies) is nothing new, textual analysis methods have become significantly more powerful over the last half decade. Are seasoned law professors, whose claim to authority emanates from their deep and rich knowledge of legal institutions, similarly at risk of being pushed off-stage by algorithms that are likely to do a more complete and coherent job of summarizing and classifying the law? More to the

point, if the very practice of law is destined to be colonized by deterministic algorithms, is there anything interesting left for researchers and scholars to study? And thus, the central question taken up below concerns whether (and how) both the practice and study of law might also be upended by data analytics methods, and what the implications as well as the opportunities are for those who currently occupy those fields.

The conclusion I offer to this set of questions is less alarmist than the motivating question might suggest. Although data analytics techniques will no doubt change (and will in many ways overtake) many of the key functions that lawyers and legal researchers now perform, I argue that the longer term effect is unlikely to eliminate the demand for lawyers *per se* as much as it will change the nature of what they do (and do not) perform. Savvy aspiring legal professions would thus be wise to steer their ships in the honing skills that are complements to (rather than substitutes for) data analytics tools. More to the point, I argue that one of the most unyielding (and indeed unique) aspects of law is its irreducible complexity – a complexity that (in my view) will necessarily implicate significant human input over the longer term.

Before proceeding, it is important to define more clearly what I mean by terms such as “data analytics”, “big data”, and “machine learning.”⁵ By and large, these terms are susceptible to considerable gimmickry, and they warrant somewhat greater precision than they are usually accorded if one is to discuss their implications meaningfully. For instance, what is the difference between “big data” and data? Is there a real difference?

⁵ Other very popular but equally abstruse terms that refer to more or less the same umbrella of data practices and methods include “data science”, “artificial intelligence”, and “informatics.” See, e.g., Bernard Marr, “What Is The Difference Between Artificial Intelligence And Machine Learning?”, *Forbes* (December 6, 2016).

Although the nuances of the following definitions vary depending on whom one asks, I take *data analytics* to represent a broad umbrella term under which big data and machine learning reside. Functionally, the term differentiates prior eras of empirical inquiry to this one, where the amount of data available (and consequently, the area of opportunity for data analysts) is not only growing but becoming increasingly rich and more detailed, demanding more sophisticated approaches. Machine learning (ML) is an artificial intelligence term referring to those algorithms and methods that allow computers to learn without being explicitly programmed.⁶ ML methods are often conceptualized through the paradigm of analogous conscious and unconscious human learning processes. For the purposes of this paper, then, I perceive six traits of data analysis, the combination of which is unique to data analysis but several of which are familiar:

- 1) Using quantitative data to analyze problems (here legal ones);
- 2) Calibrating and validating that analysis within statistical models;
- 3) Using such models for prediction (as opposed to testing a deductive “theory”);
- 4) Expansive / imaginative view about the sources of data;
- 5) Marshaling emergent machine learning techniques for analysis; and
- 6) Utilizing enhanced computing power to perform the aforementioned tasks.

To be sure, several of these characteristics have long been familiar to both legal practitioners and legal scholars. The first two, for example, have been around for decades, and are a hallmark of the empirical legal scholarship movement. The third is familiar as well; though in some respects, those schooled in quantitative social sciences tend habitually to describe the enterprise of pure prediction enterprise with the pejorative “data mining”—a term that used to reflect an instinctual disdain for those using empirical methods to predict, but without understanding the causal drivers behind prediction. (Only

⁶ See, e.g., Andrew Ng, “Lecture 3 - What is Machine Learning?”, Coursera.

recently has the pejorative connotation apparently flipped to a badge of honor.) The last three factors, however, are relatively new, and are likely the key characteristics that distinguish what is (arguably) new about modern data analytics, at least as I use the term in this paper.

2.1: Data in the Practice of Law: eDiscovery

With these working definitions in mind, there is much in lawyering and legal scholarship that lends itself to modern data analysis. Consider an area of legal practice that has already been invaded by it: the e-Discovery industry, which has clearly established the biggest beachhead for data analytics methods in the law. (See Figure 1).

Figure 1: Law Firm Leverage (Source: National Law Journal)

e-Discovery has driven the drastic change in leverage ratios inside litigation firms, i.e., the reduced need for associates within litigation-oriented firms. Early in my own career, I had exposure to this process in a patent litigation firm. Associates played a key role in processing discovery—with as many as a half dozen (or more) effectively acting as high-paid research assistants for each partner, combing through banker's boxes of produced

discovery materials. Flash forward to today, where, with the assistance of outsourced e-discovery services, this process can be executed with one. A partial result of this trend, as depicted in Figure 1, is a recalibration of law firm “leverage” ratio of associates to partners in litigation-oriented law firms, which notoriously crossed below the focal 1:1 benchmark in 2010.

At the same time, the business model of data analytics service proves to lawyers has proven fertile. Already a \$6-\$7 billion industry, by 2019 the global size of the e-Discovery industry is projected to rise to almost \$11 billion (see Figure 2).

Figure 2: e-Discovery Market Size (Source: ComplexDiscovery.com)

Two-thirds of that growth, moreover, is projected to be from inside the U.S. Note one other aspect of Figure 2: some of this anticipated growth will come from the sale of software, an area where lawyers enjoy little comparative advantage. But much of it will come through the provision of services—something that we are good at as lawyers. (I shall return to this point below.)

To many, it is already clear that e-Discovery is but the tip of the iceberg when it comes to using data analytics techniques for lawyering. Indeed, on the front lines of the data analytics revolution, companies and organizations are increasingly collecting and leveraging their own quantitative data. When these entities are involved in litigation, it is often necessary to evaluate these data as part of crafting and developing a legal argument. It is also not difficult to find areas of law that are natural foci of legal scholarship and which are susceptible to data analytics approaches. In the next section, I walk through an extended example of one such area to demonstrate the logic of how data analytics can be used in an area far outside the litigation environment: transactional law.

2.2: Data in Legal Scholarship: Identifying Corporate Opportunity Waivers

How does one harness data analytics technology in practice, and how might it affect legal scholarship? There are many different algorithms and approaches one might use; but for current purposes, it suffices to concentrate on a simple example. I emphasize that this is but one example using one possible algorithm of many⁷, but one whose general approach may be particularly useful both to lawyers and researchers interested in empirical work.⁸

The example I develop below comes from jointly authored work with Gabriel Rauterberg (Rauterberg & Talley 2017), using machine learning techniques to develop a first-of-its-kind data set of “corporate opportunity waivers” (COWs) in public disclosures. (Because space and audience considerations did not permit a detailed

⁷ For an expanded discussion on what algorithms are available using machine learning methods, refer to Elliott Ash’s and Michael A. Livermore’s responses to this paper.

⁸ An early version of this approach was developed in Talley & O’Kane (2012).

exposition of the techniques used in Rauterberg & Talley (2017), this paper also provides a helpful methodological companion piece.)

Before getting to the specifics of the machine learning details, a little background of the problem studied in Rauterberg & Talley (2017) is in order. A critical doctrinal and conceptual anchor of business law is the concept of “fiduciary duties” owed by managers, officers, directors and significant block equity holders to the business entity, the violation of which is the touchstone for nearly every type of significant form of corporate liability exposure. Fiduciary duties generally fall into one of two categories: the duty of care and the duty of loyalty. Of these branches, most would agree that the most weighty is the duty of loyalty – which governs actions involving financial conflicts of interest between the fiduciary and the company – has long been said to be immutable (non-waivable) by the parties. Indeed, most textbooks and treatises spend significant time covering the duty of loyalty, often noting that the duty is so sacred that it is immutable—not susceptible to any attempted end runs around it (contractual or not).

That traditional characterization, as it turns out, is not strictly accurate anymore: Beginning in 2000, a handful of states (starting with Delaware) began to amend their statutes to permit waivers of a specific manifestation of the duty of loyalty: *the corporate opportunities doctrine*, which prohibits corporate fiduciaries from pursuing (in their personal capacities) new business opportunities in the corporation’s line of business, without first permitting the corporation a right of first refusal. While the doctrine is traditionally not waivable (just like other forms of loyalty), the new statutory reforms created an important carve-out that permitted business entities to promulgate “corporate opportunity waivers” (“COWs”). The statutory reforms, however, were also notoriously

inexact as to how a company should go about waiving corporate opportunities, giving notoriously imprecise instructions as to *where* a COW was to be memorialized (e.g., in a corporate charter, bylaw, resolution of the board, or some other contractual agreement). Indeed, the enabling statutes in the nine states that have promulgated them give firms substantial freedom to embed waivers in virtually any of these ways.⁹

This lack of statutory precision makes empirical analysis of waivers (and thus their effects) challenging to study using traditional empirical methods. Indeed, prior to Rauterberg & Talley (2017), there had been no systematic empirical analysis of COWs, no doubt in part because of the difficulty of data collection. Under conventional approaches to data collection, it was all but impossible to manually collect these data in a cost-efficient manner, akin to looking for needles inside hay stacks embedded within larger hay stacks.

The key to unlocking the data set came from a machine learning (ML) classifier. Recall that machine learning was defined above as an approach allowing computers to learn without being explicitly programmed. Where a traditional approach would require using trained research assistants to comb through the entirety of a corpus, classifying each public disclosure one at a time, a supervised ML approach instead endeavours to train a machine learning algorithm to automatically classify documents. I describe the approach below.

Because the ML approach may be unfamiliar to some readers, and in the light of its great potential across other areas of law and finance research, this note explains the basic components of the ML approach using a simple example, and demonstrates strategies for calibrating and evaluating the classifier. (The description below also largely

⁹ For a theoretical model of optimal contracting over corporate opportunities, see Talley (1998).

tracks the procedure developed originally by Talley & O’Kane 2012.) There are six principal steps:

1. Text Extraction from Raw Data
2. Text Cleaning
3. N-Gram Parsing
4. TF-IDF Transformation
5. Dimensionality Reduction
6. Supervised Calibration, Out-of-Sample Extrapolation, and Evaluation

Text Extraction from Raw Data

The starting point of the analysis is a set of raw textual documents, assumed here to be in ASCII format (but could be HTML, PDF, UTF-8, etc. Each comes with its own challenges and prescriptions). This set of documents constitutes the set of raw unstructured “inputs” the researcher is interested in (e.g., merger agreements, constitutions, corporate charters, contracts, legal opinions, etc.) In Rauterberg & Talley (2017), the raw documentary inputs were the full text of all public securities filings for all companies between 1995 and March 2016. From these raw inputs, Rauterberg & Talley extracted text “snippets” that were identified as *candidate* COWs. Candidate COWs were flagged through a deliberately general and over-inclusive Boolean key-word search of SEC/Edgar filings. Specifically, document snippets consisted of the flagged key words, plus a 150-word margin of text proceeding and succeeding them in the document. Where the key-word search flagged multiple sections of a single document, the resulting snippet included the union of the collected individual snippets and 150-word margins. While the Boolean key-word query deliberately flagged many “false positive” documents, it also helped both to pare down both the number of documents requiring classification, and to narrow the text to be analyzed within each document. (See Rauterberg & Talley 2017 for details.)

Text Cleaning

From the set of candidate snippets, Rauterberg & Talley (2017) undertook measures to clean and parse the extracted data. These measures typically consist of several steps:

- **Typesetting Code:** If the document was in a non-ASCII format, any typesetting codes (e.g., HTML tags) are stripped out.¹⁰
- **Punctuation:** All punctuation is stripped out from the text (such as periods, commas, exclamation points, question marks, ellipses, colons, semi-colons, etc.).
- **Stems:** Word inflections are also typically stripped to their word stem (or base form). For example, the terms “walking”, “walked”, “walks”, and “walkable” would all be reduced to “walk”, and treated identically thereafter. A stemming library is necessary for this step (available in many Python modules).
- **Stop Words:** Finally, although not implemented in Rauterberg & Talley (2017), it is often thought desirable to drop “common” words that contribute little to the semantic content of the document. For example, words like “the” or “an” or “are” or “is” are frequently dropped in some applications. Stripping out stop words similarly requires using a library utility.

Note that while these are considered typical data cleaning steps, it is conceivable, for example, that a researcher trying to classify the sentiment of a document (another common machine learning task) might find that it is useful for his purposes to include punctuation in their data set. Different research problems require different approaches, and researchers often try many combinations of data cleaning steps.

N-Gram Parsing

From the cleaned and parsed documents, the next step is to extract a numerical matrix of *N-grams* tabulating raw numerical counts of each unique permutation of “N” consecutive words across the entire set of relevant documents. For instance, the most common n-gram representation is a matrix of “1-grams”, “unigrams” or “bag-of-words”, constituting the raw frequency counts of single stemmed terms. Denote this matrix by

¹⁰ In the current project, strings are converted into bytes and back to strings to clean out non-ASCII characters.

$\mathbf{N}$, where representative element n_{ij} represents “count frequency” -- the number of times term j appears in document i . Matrix $\mathbf{N}$ is a foundational intermediate result.

To elucidate this step, consider a simple example inspired by a familiar literary canon.¹¹ Suppose one were interested in a set of five “documents” (labelled D1 through D5, respectively), whose contents are as follows:

- D1: *Hickory, dickory, dock;*
- D2: *The mouse ran up the clock;*
- D3: *The clock struck one;*
- D4: *The mouse ran down;*
- D5: *Hickory, dickory, dock.*

Parsing these documents into 1-grams yields the following matrix $\mathbf{N}$:

	hickory	dickory	dock	the	mouse	ran	up	clock	struck	one	down
D1	1	1	1	0	0	0	0	0	0	0	0
D2	0	0	0	2	1	1	1	1	0	0	0
D3	0	0	0	1	0	0	0	1	1	1	0
D4	0	0	0	1	1	1	0	0	0	0	1
D5	1	1	1	0	0	0	0	0	0	0	0

Table 1: Raw 1-gram Matrix $\mathbf{N}$

The extraction of N-grams (in this case at the 1-gram level¹²) is a typical step in summarizing the “latent” semantic content of a document. The rows of Table 1 effectively summarize the content of each document as a vector in 11-dimensional space, 11 being the number of distinct words in all five documents. It is possible using standard

¹¹ See Goose, M. (c. 1697).

¹² It is possible also to parse the documents at the 2-gram (or 3-gram level, etc.), so that the columns of matrix $\mathbf{N}$ would consist of the universe of consecutive pairs (or triples, etc.) of terms. Surprisingly, in many cases, a 1-gram parsing has significant explanatory power. Rauterberg & Talley (2017), for example, extract 1-grams from their raw data set, and on that basis alone calibrate a machine-learning classifier with an accuracy rate in the mid 90% range. See *infra*.

measures of distance to assess the similarity of any two of these vectors, and thus, documents.

One common measure of similarity is a vector cosine measure, in which the distance between two vectors X and Y is given by the cosine of the angle formed between the two vectors, $\frac{X \cdot Y}{\|X\| \|Y\|}$, where $\|X\|$ denotes the Euclidean norm of X . The grid of similarity scores in the running example from above is pictured in Table 2, bounded between zero (no similarity) and one (identical content).¹³

	D1	D2	D3	D4	D5
D1	1				
D2	0	1			
D3	0	0.53033	1		
D4	0	0.707107	0.25	1	
D5	1	0	0	0	1

Table 2: Cosine Similarity of Documents D1 through D5

Many e-Discovery and plagiarism detection utilities make use of proximity scores such as those in Table 2, assessing the similarity between documents in the corpus and “flagged” specimens (such as hot discovery documents or known student essays circulating on the internet). For example, if document D2 were tagged in litigation as being incriminating, a search of the remaining documents in the corpus would likely flag D4, and possibly also D3; but neither D1 nor D2 appear to bear any syntactical similarity.

Other more sophisticated approaches can refine the document summaries even further to accentuate unique attributes of each document. For instance, note from Table 1 that the common article “*the*” appears several times. In many applications, a stop-word

¹³ Clearly, cosine is one of many alternative distance measures one might employ in practice. It has the advantage, however, of being scale independent, and always falling between 0 and 1 (assuming non-negative word counts)

dictionary typically strips this type of term out, and the use of such a dictionary is subject to the preferences of the researcher. (In some cases, including the example above, removing stop words visits a loss on predictive power, and thus some researchers choose to leave them in).

Having extracted matrix $\mathbf{N}$, a common next step is to transform the raw frequency counts into “term frequency – inverse document frequency” (or TF–IDF) measures. The purpose and effect of this transformation is to accord greater proportional weight to the counts of terms that appear frequently in a *particular* document and yet are relatively *uncommon overall*. The resulting transformed matrix, $\mathbf{T}$, contains representative element t_{ij} for document i and term j , defined by the expression:

$$t_{ij} = \left(\frac{n_{ij}}{\sum_m n_{mj}} \right) \times \ln \left[\frac{|\{j : n_{ij} > 0\}|}{M} \right]^{-1}, \quad (1)$$

where $m \in \{1, \dots, M\}$ indexes the universe of documents analyzed. The first bracketed element of (1) represents the raw count of a given term in document i relative to its total across all documents. The second term consists of the log of the inverse frequency with which term j appears (at least once) across the universe (with cardinality M) of documents analyzed. By “rewarding” the frequent intra-document use of terms that are rare on the whole, the TF–IDF transformation tends to be better able to differentiate unique documents (Salton and Buckley, 1988).

In the example from above, transforming the raw counts of matrix $\mathbf{N}$ into a TF–IDF matrix $\mathbf{T}$ is a relatively straightforward computational task. Consider again the term “*the*”, which appears twice in D2. This term appears a total of four times across all documents. And thus, in D2, the relative frequency of “the” is given by:

$$\left(\frac{n_{ij}}{\sum_m n_{mj}} \right) = \frac{2}{4} = 0.5 \quad (2)$$

Additionally, the term “the” makes at least one appearance in 3 out of a total of 5 documents, and thus its document frequency is 3/5. The natural log of the inverse of the document frequency is therefore:

$$\ln \left[\frac{|\{j : n_{ij} > 0\}|}{M} \right]^{-1} = \ln \left(\frac{5}{3} \right) = 0.223, \quad (3)$$

and thus the TF-IDF value for the word “the” in D2 is:

$$t_{2,4} = \left(\frac{n_{2,4}}{\sum_m n_{m,4}} \right) \times \ln \left[\frac{|\{j : n_{2,4} > 0\}|}{5} \right]^{-1} = \left(\frac{2}{4} \right) \times \ln \left(\frac{5}{3} \right) = 0.112. \quad (4)$$

Applying the identical transformation to the other elements of the raw unigram matrix is given in Table 2 below. Notice from the Table that D1 and D5 have identical components, indicating that they are substantially similar (indeed identical) to one another. In addition, D2 and D4 share many similar components, but are far from identical. D3 is the most unlike the others.

	hickory	dickory	dock	the	mouse	ran	up	clock	struck	one	down
D1	0.458	0.458	0.458	0	0	0	0	0	0	0	0
D2	0	0	0	0.112	0.458	0.458	1.609	0.458	0	0	0
D3	0	0	0	0.056	0	0	0	0.458	1.609	1.609	0
D4	0	0	0	0.056	0.458	0.458	0	0	0	0	1.609
D5	0.458	0.458	0.458	0	0	0	0	0	0	0	0

Table 3: TF-IDF matrix $\mathbf{T}$

Note that like the raw count matrix $\mathbf{N}$, the TF-IDF transformed matrix $\mathbf{T}$ has a fixed point at $n_{ij} = 0$, and thus any term that appeared with a frequency of zero in the raw matrix would remain at zero after transformation. Consequently, in most “real world”

applications $\mathbf{T}$ usually remains both extremely large (many columns) and sparse (many cells with value 0). The next step, then, is to employ a technique known as *Singular Value Decomposition (SVD)* that allows one to reduce the dimensionality of the data set with minimal information loss.

The concept of SVD is a straightforward matrix operation. Consider our transformed TF-IDF word count matrix $\mathbf{T}$ (with elements t_{ij}). Singular value decomposition involves using the algebraic structure of $\mathbf{T}$ to produce synthetic variables that are designed to explain internal variation within $\mathbf{T}$. The key algebraic relationship for accomplishing this task is given by the decomposition:

$$\mathbf{T} = \mathbf{U} \times \mathbf{\Sigma} \times \mathbf{V}^T, \quad (5)$$

where $\mathbf{U}$ is a column orthonormal basis; $\mathbf{V}$ is a row orthonormal basis; and $\mathbf{\Sigma}$ is the diagonal matrix of eigenvalues, where the component eigenvalues σ_{ii} are ordered from largest to smallest as one proceeds from down the diagonal, from upper left to lower right.¹⁴

A machine learning classifier typically makes use of the first k components of $\mathbf{U}$ to estimate the parameters of a predictive model (e.g., logit, probit), where k is chosen by the researcher. See, e.g., Bishop (2006); Jolliffe (2002). These eigenvectors are decreasing in strength, in that the first eigenvector explains the largest fraction of the variability in the columns of matrix $\mathbf{T}$, the second eigenvector explains the second largest, and so on. Usually, the researcher will need to select a criterion for determining the value of k . One common rule of thumb is to pick the number k that retains some specified percentage γ of the “energy” of the data, as measured by the sum of the squares

¹⁴ The SVD is done in Python using Scipy’s SVD function.

of the principal eigenvalues in the diagonal matrix Σ . That is, the value of k is the smallest value for which:

$$Energy_k = \frac{\sum_{i=1}^k (\sigma_{ii})^2}{\sum_i (\sigma_{ii})^2} \geq \gamma \quad (6)$$

Leskovec et al. (2014). The subset of the columns of $\mathbf{U}$ selected becomes the variables of interest for calibrating a predictive model.¹⁵

In the running example involving five simple documents, the singular value decomposition of matrix $\mathbf{T}$ yields the following set of non-zero eigenvalues: {2.327, 1.883, 1.636, 1.122}. Note that there are only four non-zero singular values, even though the example data set has five observations. This observation reflects the fact that the data set is not of full rank, since D1 and D5 have identical content. The first two eigenvalues manifest aggregate energy of approximately 63%, indicating that two factors can explain roughly three fifths of the variation in the data. Adding a third factor would increase the energy to 84%, and adding a fourth would increase it further to 100%. For illustrative purposes, we limit our attention below to the first two eigenvalues, which have associated eigenvectors denoted as F1 and F2. The numerical values of these factors as follows:

	F1	F2
D1	0	0
D2	0.10223301	0.77838211
D3	0.99457099	-0.0921333
D4	0.01941521	0.62099336
D5	0	0

Table 4: 2-Factors

Although some information is clearly lost in reducing the dimensionality of the parsed data set from 11 columns to 2, the retained factors still manage to capture a

¹⁵ Elliott Ash discusses additional dimension reduction techniques as well as benefits and potential downsides of each in his response to this paper.

significant amount of the variation in the documents. D1 and D5, for example, still share identical coordinates, and D2 and D4 still appear similar (albeit not identical) in both dimensions. D3 remains the most unlike the others. A scatter plot of the documents in factor space helps illustrate this point, as pictured in Figure X.

Figure 3: D1-D5 as Depicted in 2-Dimensional Factor Space

Even reduced dimensionality space, the documents effectively cluster in three areas, and are amenable to analysis as such.

Rauterberg & Talley (2017) apply a virtually identical approach to the approximately 10,600 snippets of candidate corporate opportunity waivers flagged from public SEC filings. The eigenvalues that emerge from that analysis are pictured in Figure 4, along with their cumulative energy.

As illustrated Figure 4, substantial variation in the textual corpus can be captured with a relatively small number of factors. The number of factors used for model calibration must be selected with some care. Select too few, and the classifier will be insufficiently nuanced to be accurate. Select too many, and the researcher is sure to overfit the model, making its extrapolation to out-of-sample data unreliable. Rauterberg & Talley (2017) ultimately select the first 14 principal eigenvectors, corresponding to a

90% energy threshold. Based on calibration tests (see below), this choice appears to work extremely well in the corporate opportunity waiver context.

Figure 4: Principal Eigenvalues for Candidate COWs and Energy

Supervised Calibration, Out-of-Sample Extrapolation, and Evaluation

The processes described above – i.e., summarizing latent textual data as a matrix and reducing its dimensionality – are common steps in a variety of distinct machine learning applications to law. What happens from there turns on the nature of the project. Broadly speaking, machine learning techniques applied to textual data tend to divide into one of two approaches: unsupervised learning and supervised learning. *Unsupervised learning* explores the structure of the data itself, with the goal of uncovering textual patterns, similarities, sequential regularities, and syntactical “clusters” internal to the text itself (i.e., without the researcher’s input). The process of identifying such patterns, or “topics” associated with the data often (though not always) involves extracting principal eigenvalues, as described above.¹⁶

¹⁶ Formally, a topic model estimates a probability distribution across words in a fixed vocabulary. See Blei et al. (2003).

While unsupervised learning techniques can be incredibly useful in uncovering hard-to-detect patterns in textual data, evaluating the significance of such patterns is sometimes challenging unless one has an independent means for assessing importance and/or gravity of various types of document. This limitation is particularly salient in legal applications of text analysis, which tend to turn on whether the language used in a contract / regulation / judicial opinion imposes a net legal burden or benefit on populations of interest, and the magnitudes of such effects. Consequently, it will often be desirable – in addition to summarizing textual content of the corpus – to involve “real world” human classifiers to assess the language and import of some subset of the textual data being analysed, as one does in *supervised learning* techniques. In turn, this human-coded data can be used to “train” a predictive model to classify the universe of documents, including those that human coders have never previously assessed.

Rauterberg & Talley (2017) pursue the latter course of *supervised learning*. From a data set of 10,682 snippets that included candidate waivers, the authors and a team of research assistants manually coded 1,000 randomly selected snippets along over 40 dichotomous dimensions, including not only the presence/absence of a waiver disclosure (our primary topic of interest), but also the scope, reach, and location of such disclosures when they occurred. The most significant variable, of course, is whether the candidate disclosure was a *bona fide* disclosure of a waiver (in whatever form). They find approximately 62% of the hand-classified sample fits this description. The hand-coding of this sample, in turn, creates a “training” data set for a machine learning *perceptron* – an algorithm for predicting the presence or absence of a genuine disclosure in a given document.

Stripped down to its essentials, calibrating a perceptron reduces to estimating a qualitative regression such as:

$$\Pr\{y_i = 1\} = f(X_i, \varepsilon_i | \beta), \quad (7)$$

where y_i denotes the dependent dichotomous variable (here, the presence or absence of a waiver disclosure), f is some (potentially non-linear) function of data attributes X_i and an error term ε_i , and β is a vector of estimated coefficients. In what follows, we employ logistic regression, but a similar approach would apply to probit, linear probability, SVM estimation, and others.

Once the elements of β are estimated, it is possible to assess how well the model “fits” the patterns manifest within the sample data set. Unlike more conventional regression analysis approaches, however, where the elements of X_i are easily interpreted variables of interest, here these elements consist of the principal eigenvectors extracted from the decomposition of the text data matrix. As such, their interpretational content is recondite, and thus the interpretation of the estimated coefficients is of little interest.

ML Classifier	Hand Coded			Fit Measure (50% Classifier)		
	Present	Absent	Total			
	Present	595	35	630	Sensitivity (True Positive Rate)	95.20%
	Absent	30	340	370	Specificity (True Negative Rate)	90.67%
	Total	625	375	1000	Positive predictive value	94.44%
Logit Estimation				Negative predictive value	91.89%	
				False Positive Rate (1 - Specificity)	9.33%	
				False Negative Rate (1 - Sensitivity)	4.80%	
	Pseudo					
	R ²	0.7194		Converse False Positive	5.56%	
	Log lik.	-185.65		Converse False Negative	8.11%	
	$\chi^2(14)$	951.83***		Correctly classified	93.50%	

Table 5: Within-Sample Classification Accuracy of COW Disclosures

What *is* usually of interest, however, is whether the model does a good job in practice of classifying the textual data. Several diagnostics are helpful in measuring predictive performance, such as the rate of false positive and false negative classifications. Table 5 provides several metrics along these lines, employing a classification “assignment rule” that assigns the value of 1 whenever the estimated model probability of a COW is at least 50%. Using this assignment rule, we find a correct classification rate of 93.5% across the entire sample, and reasonably good rates of both false positive classifications (9.33%) and false negative classifications (4.8%). The underlying prediction model – built on a logistic regression – delivered strong predictive power at all conventional levels of statistical significance.

Figure 5: ROC Mapping (Within-Sample)

Another approach for assessing classification accuracy makes use of the (so-called) “Receiver Operating Characteristic” (or ROC) function. This function represents parametric curve that emerges as one varies the assignment rule from 0% to 100%, plotting the false positive rate (horizontal axis) against the true positive rate (vertical axis) along the way. Good classifiers will tend to “bow” towards the northwest corner of

the unit square, a point that coincides with perfect classification. Consequently, the area under the curve of the ROC (sometimes called the ROC-AUC) is a common proxy for classifier performance: The closer the ROC-AUC area is to 1, the better the classifier. Figure 5 illustrates the ROC function for our within-sample estimations. As can be seen from the Figure, the quality of the machine classifier appears exceptionally good, with a ROC-AUC measure of 0.9764.

Although in-sample calibrations (here, the hand-coded 1,000 documents) are instructive, we expect that they will necessarily degrade somewhat when the predictive model is taken outside of the training sample to the rest of the data set (i.e., the full set of 10,600 documents less the 1,000 hand-coded documents). Nevertheless, it is out-of-sample prediction where the ML approach can be useful in economizing the time and expense of hand coding. We cannot check the accuracy of the model's classification of the 9,600 documents without hand-coding them as well, so we utilize Monte Carlo simulations on the hand-coded 1,000 documents for which we can definitively check the model's classification accuracy against the hand-coded classifications. The Monte Carlo simulation approach was employed in Talley & O'Kane (2012), and previously proposed by Breiman (1996) and Friedman et al. (2000; 2003). Within each iteration of the simulation, the sample data set of 1,000 hand-coded documents were randomly segregated into two groups: A provisional "training" data set, consisting of roughly 75% of our sample data set, and a provisional "testing" data set, consisting of the remaining 25% of the sample data set. We then fit equation (2) to the training data, marshalling the resulting coefficient estimates to generate predictions of the presence/absence of the waiver in the testing data, and generating predictive metrics similar to those discussed

above. We repeated the Monte Carlo simulation for 1,000 iterations, which produced empirical distributions for each of these metrics.

Table 6 presents classification performance metrics from the simulations. The results are surprisingly good. Overall, while the percentage of correct classifications declined (as expected), they fell by an extremely modest one percent (from 93.5% to 92.5%), a pattern generally replicated symmetrically in both positive and negative classifications. Similarly, the area under the ROC curve declined, but again only trivially—from 0.9764 in the full sample to a mean value of 0.9727 across the Monte Carlo simulations. Moreover, the standard deviations associated with each of the Monte Carlo metrics also appear notably modest – indicating precise estimation.

	MC Iter.	Mean	St. Dev.	Median	Min	Max
% Correctly Classified	1,000	92.4980	1.5123	92.4000	86.8000	96.8000
Sensitivity (50% classifier)	1,000	94.3740	1.8423	94.4828	87.3418	98.6928
Specificity (50% classifier)	1,000	89.3847	3.1513	89.5294	78.4946	98.9362
Area under ROC	1,000	0.9727	0.0086	0.9728	0.9363	0.9945

Table 6: Monte Carlo Classification Metrics, Simulated Out-of-Sample Data

The respectable correct classification rates might also be put into perspective by comparing it to the rate of *human* error. Interestingly, it is possible to deploy the ML classifier to back-test and audit the accuracy of the human coders whose classifications were used to train the very algorithm testing them. A hand-audit of “disagreements” between the ML classifier and human classifications in the original sample, human classifiers were found to be accurate approximately 97.8% of the time. Thus, the

difference between human and machine classification is even smaller than reflected in Table 5.¹⁷

A final diagnostic measure of how well the ML classifier performs is to extrapolate the calibrated within-sample model onto the full data set, thereby generating estimated probabilities of the presence/absence of a waiver disclosure over the entire universe of snippets. The frequency distribution of estimated probabilities can provide some sense of how definitively the ML classifier discriminates between positives and negatives.

Figure 6

Figure 6 illustrates the histogram of estimated waiver probabilities extrapolated over the full data set ($N=10,620$). The figure does *not* correct any of the known misclassifications in the hand-coded set, and thus it accurately illustrates the noise that the ML classifier introduces. Note the strongly bimodal distribution in the empirical frequency, indicating that the ML classifier not only discriminates between likely waivers and non-waivers, but

¹⁷ This auditing process also permitted correction and re-estimation of the ML classifier using corrected hand-coded data. Tables 4 and 5 reflect this corrected model.

does so with some degree of statistical definitiveness.¹⁸ Thus we are confident that the model trained by the 1,000 observation sample data set can be satisfactorily applied to classify the rest of the out-of-sample data set. I pause here to point out the magnitude of labour saved – whereas the absence of ML methods would have demanded an army of research assistants hand-coding 10,600 documents, ML methods allow for automated classification of 90% of the entire data set that is as accurate as hand-coding.

3. Implications for Legal Researchers and Practitioners

The above example is admittedly particularized, but many parts of the basic approach generalize broadly to legal practice and research. As noted above, there have now been for many years several applications of data analytics in the field of law, particularly in litigation domains. Machine learning oriented citation of an opinion analysis has been around for over a decade, and can be extremely valuable to discern patterns and clusters of precedent citations as well as parallel lines of precedent (Katz and Stafford 2010). In addition, legal document comparison is a well-established machine learning approach, with e-Discovery serving as a prime example. Indeed, as noted above, the efficiency of data analytic approaches in discovery helps explain the de-leveraging trends in law firms illustrated in Figure 1. Rather than having a large team of associates looking for inculpatory documents in produced records, it has now become commonplace

¹⁸ Fortunately, the Monte Carlo simulations yielded a ML classifier that performed extremely well by conventional measures. We thus deemed it unnecessary to iterate our calibrations further. That said, in many applications, the researcher must often continue to iterate on the calibration, for example by selecting a different preceptor model, adjusting the number of retained factors to optimize the out-of-sample performance, extracting 2-grams or 3-grams rather than 1-grams, or even going back to the original data set to code additional training documents by hand.

to staff just one or two associates on such a task, attempting to find a smaller number of “hot” documents that can serve as a comparison template for algorithmic search.

Increasingly, transactional attorneys (and researchers) are coming to utilize machine learning tools to great effect. For example, mergers and acquisitions practitioners are obsessed with “what is market?” – a shorthand for understanding the trends in terms that other parties tend to include in definitive merger agreements. Talley & O’Kane (2012) utilized a methodology similar to that which is described above to classify “material adverse effect” (or Act of God) clauses in M&A agreements. (I and a group of research assistants are currently working with the American Bar Association to produce an automated version of their annual “Dealpoints” study that hand-codes merger provisions in large deals on an annual basis.)

The far more tempting (though still largely undeveloped) frontier is to use data analytics and machine learning to predict judicial outcomes in ongoing cases. By conceptualizing the set of complaints, briefs, expert reports, existing case law (and so forth) as the corpus, can one predict a plaintiff’s chances of prevailing should a case go to trial? More generally, could insurers, lenders and other contractual entities predict ex ante litigation risk exposure from specific practices (e.g., marketing a new pharmaceutical or consumer product)? Can one predict regulatory or political risk associated with merger activity or pricing? Each of these approaches seem readily amenable to data analytics and machine learning methodologies.

Somewhat tellingly, however, the above tasks are also predominantly those carried out by most lawyers and legal researchers. If an algorithm can exercise legal judgment, make predictions, and assess risks more efficiently (and possibly more

accurately) than a human, we come full circle to the question posted by the title of this essay: Are lawyers and legal researchers soon to be displaced by data analytics methods in both the practice and research of law?

To get some traction on this question, it may help to advert to an academic discussion more familiar within law and philosophy about whether legal argumentation and rhetoric is something akin to a complex game (such as chess), with attorneys playing the role of players and judges the role of judge. H.L.A. Hart, Stanley Fish, Andrei Marmor, and several others have developed this topic at some length. And if it is correct that law and chess are roughly analogous, then we indeed have something to worry about. Consider:

- IBM's Deep Blue, knocking off chess grandmaster Garry Kasparov,
- Google's AlphaGO defeating the top five "Go" players in the world, or
- IBM's Watson pulverizing several former champions in "Jeopardy."

If the practice and study of law analogize to these types of complex games of skill, things indeed appear bleak for most all legal professionals. This concern has no doubt given rise to some degree of angst among people in the legal profession. The number of LSAT takers and applicants to law schools have fallen along with the ratio of associates to partners. Law schools are closing. Large firms are still laying off attorneys, well after the financial crisis.

Is this an existential moment for law? Are we approaching the "singularity event" that is so often talked about by technologists, where machine knowledge overtakes human knowledge, never to return? Are we about to turn into the out-of-work drivers for what will become the law's driverless car?

Perhaps; but I am skeptical for a variety of reasons. And it is not because I believe data analytics and ML methods to be unpromising. As noted above, they are proving to be extremely powerful (and promising). And yet, at least three considerations suggest that the apocalyptic story (at least in its most extreme forms) will not come to pass.

First, the most powerful and promising sorts of data analytic methods – those that hinge on *supervised learning* approaches – necessarily implicate human judgment as an input in ways that more technocratic tasks (such as engineering a bridge) may not. The corporate opportunities project detailed above, for example, required human classifiers to train the algorithm, those classifiers, in turn, had to draw on their own legal training to identify and understand the implications of what is (and is not) in the text of a contract, email, warranty, disclosure, etc. Practicing lawyers must identify “hot” documents in order to train an e-Discovery algorithm to have any predictive power. In nearly every sub-field of law, moreover, the same legal terms can have entirely different connotations. Thus, even if viewed as a perfectly consistent and unchanging institution (more on this below), the practice and study of law is sufficiently complex that it would take decades to distill into a set of reliable algorithms. Throughout this process, human knowledge and machine learning would necessarily serve as complements – not substitutes – for one another in relevant legal domains. Humans plus machines, at least in the context of law, are going to be able to do far more than either of them independently.

Second, the law simply is not a static system, and it never has been. There are some elements of how legal reasoning evolves that are unavoidably and irreducibly complex. Consider, for example, the cases that we spend our time developing in first-year law school classes: *Marbury v. Madison*, *Buick v. MacPherson*, *Brown v. Board*, *Walker*

Thomas Furniture, Citizens United, Obergefell. These seminal cases are interesting for at least two reasons: First, they were hard, legitimate “close calls” going in; but second, they quite literally changed the course of law: Each marked an important conceptual “shock” in the application of time-honed legal principals to factional situations. Each of these watershed moments changed the way we conceive the law not just in a positive / descriptive way, but also in *normative / prescriptive* dimensions. Such evolutionary moments are difficult for algorithmic approaches to predict.

The free parameter played by the normative commitments underlying law arguably motivated another well-known legal philosopher, Ronald Dworkin, in his consideration of the analogy of law to chess. He famously rejected the analogy as follows:

*“In adjudication, unlike chess, the argument for a particular rule may be more important than the argument from that rule to the particular case; and while the chess referee who decides a case by appeal to a rule no one has ever heard of before is likely to be dismissed or certified, the judge who does so is likely to be celebrated in law school lectures.”*¹⁹

Dworkin’s observation nests closely to the idea that law, particularly common law systems, gain much of their vitality because they are constantly in flux, influenced by a persistent background normative dialogue (even if one that plays out through seemingly technocratic doctrinal distinctions). The forces of such a dialogue can (and do) affect the application of law to facts, and can do so in ways that may cause legal outcomes to exhibit (to the unschooled algorithmic eye) a “careening” characteristic, seemingly manifesting stark and dramatic shifts from past practices and norms. Whenever the “black letter law” seems settled, it can still be destabilized by public policy, theory, normative commitments, and value trade-offs. These commitments—and how they

¹⁹ Dworkin (1978), at 112.

evolve over time—are perhaps the most unpredictable of all forces that shape law and legal institutions, and among the least amenable to algorithmic prediction.

A final reason that law as an institution is likely to be particularly resistant to algorithmic takeovers is that law must continue to play a central role in mediating *other* areas where we are *really* worried about technological singularity events. Such domains include information privacy, intellectual property, securities market trading, and many other areas are now trying to cope with the fact that in fact this may be a real danger in other venues. Almost by definition, law must play a central role in determining the appropriate role (and legal bounds) on automation. It is difficult to imagine that such a regulatory role itself could be easily co-opted by an algorithm, charged with creating regulations to stem bad consequences of algorithmic approaches elsewhere.

4. Conclusion

Data analytics will continue to affect and shape law for decades (if not centuries) to come. As it does so, however, I would wager that lawyers and data analytics are not likely to appear to be substitutes for one another as they will prove to be complements. To the extent this prediction is true, then it is probably good news for human aspiring lawyers who wish to embrace it (and eventually we must all embrace it). New lawyers entering the field would be well advised to spend some time brushing up on data analytics, coding, and statistics skills—and many of the brightest students, up-and-coming practitioners, and promising researchers have similarly begun to stake this area out as prime real estate. It is these individuals—and not their data algorithms *per se*—who will represent the greatest risks to established, traditional interests. Put differently,

the biggest existential threat (and opportunity) posed by the data analytics revolution to law is almost certainly ourselves, by being unaware of how quickly these applications are growing, how they are likely to transform both practice and research far into the future, and how quickly they will come to dominate the landscape. If law is to become a driverless car in some capacity, lawyers and legal researchers cannot and should not be left standing on the sidewalk.

References

1. Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer Science+ Business Media LLC, New York.
2. Blei, D., Andrew Ng and Michael Jordan (2003), *Latent Dirichlet Allocation*, *Journal of Machine Learning Research* 3: 993-1022.
3. Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140.
4. Dworkin, R., *Taking Rights Seriously* (Harvard Press 1978).
5. Friedman, J., T. Hastie, and R. Tibshirani (2000). "Additive logistic regression: a statistical view of boosting." *Annals of Statistics*. 28: 337-407.
6. Friedman, J., T. Hastie, & R. Tibshirani, (2003). *Elements of Statistical Learning*.
7. Goose, M., (c. 1697). "Hickory Dickory Dock," *Stories or Tales from Times Past: Tales of Mother Goose* (Charles Perrault, ed.).
8. Jolliffe, I.T. (2002), *Principal Component Analysis* (Springer-Verlag 2d ed).
9. Katz, D. & D. Stafford, *Hustle and Flow: A Social Network Analysis of the American Federal Judiciary* (2010), *Ohio State Law Journal* 71:457.
10. Leskovec et al. (2014), "Mining of Massive Data Sets" (Stanford CA).
11. Rauterberg, G. & E. Talley (2017), "Contracting Out of the Fiduciary Duty of Loyalty: An Empirical Analysis of Corporate Opportunity Waivers." Forthcoming, *Columbia Law Review*, <http://ssrn.com/abstract=2822248>.
12. Salton, Gerard and C. Buckley, (1988), "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management* 24 (5): 513–523.
13. Surden, H. & M. Williams, *Self-Driving Cars, Predictability, and Law* (March 14, 2016). Available at SSRN: <http://ssrn.com/abstract=2747491>
14. Talley, E. (1998), *Turning Servile Opportunities to Gold: A Strategic Analysis of the Corporate Opportunities Doctrine*, *Yale Law Journal* 108:277-398.
15. Talley, E. & D. O’Kane (2012). "The Measure of a MAC: A Machine-Learning Protocol for Tokenizing Force Majeure Clauses in M&A Agreements, 168 *J. Inst. & Theor. Econ.* 181.