CONTRIBUTORS

Charlotte S. Alexander, Georgia State University
Daniel Argyle, FiscalNote
Elliott Ash, ETH Zurich
Jon Ashley, University of Virginia
Adam B. Badawi, University of California, Berkeley
Keith Carlson, Dartmouth College
Daniel L. Chen, University of Toulouse
Ryan Copus, Harvard University
Giuseppe Dari-Mattiacci, University of Amsterdam
Marion Dumas, London School of Economics and Political Science
Vlad Eidelman, FiscalNote
Mohammad Javad Feizollahi, Georgia State University
Adam Feldman, SCOTUSBlog
Jens Frankenreiter, Max Planck Institute for Research on Collective Goods
Brian Grom, FiscalNote
Yosh Halberstom, University of Toronto
Ryan Hübert, University of California, Davis
Khalifeh al Jadda, Georgia State University
Anastassia Kornilova, FiscalNote
Manoj Kumar, NYU Center for Data Science
Hannah Laqueur, University of California, Davis
Michael A. Livermore, University of Virginia
Allen Riddell, Indiana University Bloomington
Daniel N. Rockmore, Dartmouth College
Anne M. Tucker, Georgia State University
Anna Venancio, Primer AI
Alan Yu, University of Chicago
CONTENTS

Introduction: From Analogue to Digital Legal Scholarship

The Editors ....................................................... xiii ~ ix ~

Frameworks

1: Distant Reading the Law
   Michael A. Livermore and Daniel N. Rockmore ....................... 3

2: Big Data, Machine Learning, and the Credibility Revolution in Empirical Legal Studies
   Ryan Copus, Ryan Hübert, and Hannah Laqueur ....................... 21

3: Text as Observational Data
   Marion Dumas and Jens Frankenreiter ................................. 59

4: Prediction Before Inference
   Allen Riddell ..................................................... 73

Investigation

5: Style and Substance on the US Supreme Court
   Keith Carlson, Daniel Rockmore, Allen Riddell, Jon Ashley, and Michael A. Livermore ................................. 83

6: Predicting Legislative Floor Action
   Vlad Eidelman, Anastasija Kornilova, and Daniel Argyle ...... 117
7: Writing Style and Legal Traditions
   Jens Frankenreiter .......................................................... 153

8: A Computational Analysis of California Parole Suitability Hearings
   Hannah Laqueur and Anna Venancio ........................................ 193

9: Analyzing Public Comments
   Vlad Eidelman, Brian Grom, and Michael A. Livermore .......... 235

~ x ~

Exploration

10: Using Text Analytics to Predict Litigation Outcomes
   Charlotte S. Alexander; Khalifeh al Jadda, Mohammad Javad Feizollahi, and Anne M. Tucker ......................... 275

   Elliott Ash and Daniel L. Chen ............................................. 313

12: Reference Networks and Civil Codes
   Adam B. Badawi and Giuseppe Dari-Mattiacci .......................... 339

13: Attorney Voice and the US Supreme Court
   Daniel L. Chen, Yosh Halberstom, Manoj Kumar, and Alan Yu 367

14: Detecting Ideology in Judicial Language
   Marion Dumas ................................................................. 383

15: Opinion Clarity in State and Federal Trial Courts
   Adam Feldman ............................................................... 407
Law is an artifact of language. In this chapter, we ask what can be gained by applying to the law new techniques from natural language processing that translate words and documents into vectors within a space. These vector representations of words and documents are information dense—in the sense of retaining information about semantic content and meaning—while also being computationally tractable. This combination of information density and computational tractability opens up a wide potential realm of mathematical tools that can be used to generate quantitative and empirically testable hypotheses about the law.

This new approach to legal studies addresses the shortcomings of existing methods for studying legal language. At a theoretical level, even the best formal models of legal decision-making require strong simplifying assumptions that treat the law metaphorically. The case-space literature, for example, assumes the language of law to be a function over an idealized geometric space, where the law separates the fact space into “liable” and “not liable” or “guilty” and “not guilty.” Case-space models give us some intuition about the legal reasoning process, but

---

1 Cameron and Kornhauser (2017) provide a recent review of this literature.
they have been somewhat limited because it has been unfeasible to empirically realize the legal case-space model in any formal mathematical way.

Likewise, because law consists of text, the standard empirical research methods are somewhat limited in the questions that can be asked. Traditionally, text-based empirical legal studies research has relied on small-scale datasets, where legal variables are manually coded (e.g., Songer and Haire 1992). Hand-coding legal documents is labor-intensive and requires subjective and simplifying decisions.

However, new computational approaches to the study of text are enabling new kinds of large-scale quantitative text-based projects across a wide range of disciplines. For example, recent work in computational linguistics has made breakthroughs in vector representations of language (Jurafsky and Martin 2000). Topic models such as latent Dirichlet allocation serve to automate the coding of texts by generating topics as sets of words that tend to cooccur (Blei, Ng, and Jordan 2003; Blei 2012). These algorithms have provided a window to the relations between documents at scale.

An active literature in computational legal studies has begun to apply these methods to legal documents. Livermore, Riddell, and Rockmore (2017) use a topic model to understand agenda formation on the US Supreme Court (see also Carlson, Livermore, and Rockmore 2016). Leibon et al. (2018) construct a geometric and network-based model to study any kind of citation-linked document corpus and use it to represent the geometric relations between US Supreme Court cases and even predict domains of heightened or diminished legal activity. Ganglmair and Wardlaw (2017) apply a topic model to debt contracts, while Ash, MacLeod, and Naidu (2018) apply one to labor union contracts.
But legal scholars are still mostly unfamiliar with a new counterpart to topic models: embedding models. The success of word embedding models, such as Google’s Word2Vec algorithm, is that they “learn” the conceptual relations between words; a trained model can produce synonyms, antonyms, and even analogies for any given word in a large text corpus (Mikolov, Sutskever, et al. 2013; Levy, Goldberg, and Dagan 2015). The derived word vectors serve well as features in downstream prediction tasks by encoding a good deal of information in relatively rare word features. More recently, analogously constructed “document embeddings” have built upon the success of word embeddings to represent words and documents in a joint geometric space (Le and Mikolov 2014). Like word embeddings, these document embeddings have advantages in terms of interpretability and serve well in prediction and classification tasks.

This chapter serves to introduce document embeddings into the legal literature and illustrates the method using a corpus of US appellate court cases. Our data include the universe of US Supreme Court and US Circuit Court cases for the years 1887–2013. We construct document embeddings for each opinion in the corpus. We then construct judge vectors by taking the average of the document embeddings for the cases authored by the judge. These case vectors are used to analyze the geometry of federal appellate case law.

We ask whether the information recovered by our model provides a meaningful signal about the legal content in cases. We find that spatial clustering in these embeddings encodes differences between cases in different courts, between cases in different years, and between cases in different legal topics. The vectors can also discriminate judges based on birth cohorts but do not do well in encoding the partisan affiliation of judges or
law school attended. We also demonstrate that the vectors can be used to produce a measure of similarity between the legal writings of any two judges.

In the final section, we outline a range of potential future applications for the use of embedding models in computational analysis of law. These include examining analogies and associations, experimenting with structured and categorical embeddings, and constructing embeddings for citation networks.

Vector Space Embedding Models and the Law

A first-order problem in empirical analysis of text data is the a priori high dimensionality of text. There is an arbitrary number of approaches for representing plain text as data. One must trade off informativeness, interpretability, and computational tractability (Ash 2018). For example, one could represent a document as a frequency distribution over words, but with a large vocabulary, say, 20,000 words, a document is still a high-dimensional vector and interpreting such a representation is difficult.

Word embeddings came about as a dimension reduction approach in deep learning models for prediction tasks in computational linguistics (Mikolov, Sutskever, et al. 2013). Such tasks include, for example, predicting the next word in a sequence given a set of words in a sentence. To that end, the model represents a word as a small and dense vector (say 100 dimensions). Initially, words are randomly distributed across the vector space. But the word locations then become features in a learning model: they move around during training to improve performance on a prediction task. In natural language settings, this process typically leads to words clustering near similar words.
Document embeddings, such as Le and Mikolov’s paragraph vectors, use a separate embedding layer for both the word and the document to solve the prediction task (Le and Mikolov 2014). These models locate documents in a vector space, where documents that contain similar language tend to be located near to each other in the space. Embedding models are different from topic models (e.g., Blei 2012) because the dimensions have a spatial interpretation rather than a topic-share interpretation.

Embedding models have become popular because the spatial relations between the vectors encode useful and meaningful information (Levy, Goldberg, and Dagan 2015). To illustrate, a word embedding can identify similar words in the vocabulary. For example, the embedding of “judge” might be close to that of “jury” but far away from “flowerpot.” Similarly, the proximity of document embeddings can identify similar cases in a corpus of decisions based on use of similar language. For example, in our data *Engel v. Vitale* (1962) is spatially close to *Everson v. Board of Education* (1947), presumably because they are both early US Supreme Court decisions that deal with religious freedoms in the states. Finally, judge embeddings constructed from these documents could be used to identify similar judges in the legal system. For example, we find (see below) that Antonin Scalia appears to be close to Clarence Thomas (perhaps since they are both conservative judges who tend to use originalist arguments).

**Application to Federal Appellate Courts**

This section illustrates the use of document embeddings in the federal appellate courts. We begin by discussing the data and how the document vectors are constructed. We then explore the visual relations between the cases. Finally, we explore similarity relations between judges.
LAW AS DATA

DATA AND DOCUMENTS

The analysis utilizes a corpus of all US Supreme Court cases and all US Circuit Court cases for the years 1887–2013. We have detailed metadata for each majority opinion; we mainly use the court, date, case topic, and authoring judge. The circuit court data does not include unpublished opinions. Per curiam opinions and discretionary opinions (concurrences/dissents) are excluded from the analysis.

For case topic, we use the seven-category “general issue” designation coded for Donald Songer’s Courts of Appeals Database (e.g., Haire, Songer, and Lindquist 2003). To make these categories, Songer’s research team classified cases to a single major topic such as crime, civil rights, first amendment, due process, privacy, labor relations, economics/regulation, and others.

The cases are linked to biographical information on the judges obtained from the Federal Judicial Center. This includes a plethora of demographic and career details by judge. In the illustrative analysis, we use the birth date and political affiliation of the appointing president.

Finally, the dataset includes the full text of the authored judicial opinions. We remove HTML markup and citations. We then have each case as a list of tokens. These tokens provide the inputs for the embedding model.

CONSTRUCTION OF DOCUMENT VECTORS

The next step is to construct document vectors for each case. The model we use is Doc2Vec (Le and Mikolov 2014), implemented in the Python package gensim. The objective function solved by this model is to iterate over the corpus and try to predict a given word using its context (a window of neighboring words), as well as a bag-of-words representation of
the whole document. The model uses an embedding layer for the context features and the document features. Therefore the spatial location of documents encodes predictive information for the context-specific frequencies of words in the document.

We feed the case documents in random order into Doc2Vec using standard parameter choices (Dai, Olah, and Le 2015). We used the distributed bag-of-words model over the distributed memory model, with 200 dimensions per document vector. Other parameter choices include a context window of size 10, capping the vocabulary at 100,000 words (based on overall document frequency) and excluding documents shorter than 40 words in length. The algorithm reads through the corpus documents in random order five times. As this chapter is an exploration and illustration, we did not explore the parameter space deeply or broadly.

VECTOR CENTERING AND AGGREGATION

We now have a set of vectors \( \vec{t}_i \) for each case \( i \). We normalized each vector to length 1. Each case has an authoring judge \( j \) working in court \( c \) at year \( t \). Besides author and time, the other metadata feature is the case topic \( k \). We use these categories for descriptive statistics, as well as to “control” for these features for more targeted analysis.

For visualization and other analysis we center and aggregate the document vectors in several ways. Let \( I_j \) be the set of cases authored by \( j \). Let \( I_{jt} \) be the set of cases authored by \( j \) at year \( t \). We construct a vector representation for a judge using

\[
\vec{\bar{j}} = \frac{1}{|I_j|} \sum_{i \in I_j} \vec{t}_i
\]
where $| \cdot |$ gives the count of the set. Similarly, the vector for judge $j$ at year $t$ is given by

$$\vec{j}_t = \frac{1}{|I_{jt}|} \sum_{i \in I_{jt}} \vec{i}_t$$

and the vector for all cases on topic $k$ in court $c$ during year $t$ is given by

$$\vec{c_{kt}} = \frac{1}{|I_{ckt}|} \sum_{i \in I_{ckt}} \vec{i}_t$$

Meanwhile, the same notation and corresponding aggregation formula is used to construct a vector for a year $\vec{t}$, for a court $\vec{c}$, for a topic $\vec{k}$, or for the cases in court $\vec{c}$ during a particular year $\vec{t}$.

We are interested in recovering the ideological component of the judge vectors. Therefore we explore the following steps to center the document vectors before aggregating. We represent the year-centered vector for case $i$ as $\vec{i}_t = \vec{i} - \vec{t}_i$, where $\vec{t}_i$ corresponds to the average vector for all cases in the same year as $i$. Similarly, let a subscripted judge vector $\vec{j}_t$ be defined as

$$\vec{j}_t = \frac{1}{|I_{j}|} \sum_{t \in I_{j}} \vec{i}_t$$

with the average for judge $j$ of the year-centered vectors $\vec{t}_i$.

The preferred centering specification depends on the context of the analysis. We center by interacted groups, in particular. In the results below, we variously center by topic-year $\vec{kt}$, by court-year $\vec{ct}$, and by court-topic-year $\vec{ckt}$. Only after this centering step do we aggregate by judge and perform analysis of the spatial relations between vectors. The hope is that the remaining spatial variation is purged of court-specific, topic-specific, and year-specific differences in language. The remaining variation will provide a cleaner summary of the ideological differences between judges.
Chapter 11: Case Vectors: Spatial Representations of the Law

Here we have used the unweighted average of the case vectors, where each case is weighted equally. Future work might explore the use of other weighting schemes. A sensible alternative would be to weight the cases by their length (in words or sentences), for example. In addition, it would be reasonable to weight the cases by the number of citations they later received (as a proxy for importance). Finally, one might normalize the vectors after centering and/or aggregating.

VISUAL STRUCTURE OF CASE VECTORS AND JUDGE VECTORS

In this section we present a variety of visualizations to understand better the spatial relationships encoded by our case vectors and judge vectors. Our visualization is a t-SNE (t-distributed stochastic neighbor) embedding plot (van der Maaten and Hinton 2008), which projects the vectors down to two dimensions for visualization purposes. We use t-SNE plots because the dimension reduction algorithm is designed to preserve local distances between points and therefore recover informative clusters (Lee and Verleysen 2007). In tests, we generated better visualizations using t-SNE than other manifold learning algorithms such as principal components, multidimensional scaling, or isomap.

We begin by exploring the institutional, temporal, and judge-level features encoded in the vectors. For figure 11.1, we centered the case vectors by topic interacted with year, as described above. We then averaged by judge and plotted the judge vectors. The vectors are labeled by court. One can see that, conditional on topic and year, the document vectors separate the courts quite well. This is consistent with systematic differences in legal language across courts, conditional on topic and year, being captured by the embedding.
For figure 11.2, we centered on court interacted with topic. We then averaged by court-year and plotted the court-year-level averaged vectors. The dots are labeled and colored by the decade the case was published. One can see a steady linear development of case law across the geometric space. This shows that, controlling for court and topic factors, the embedding captures systematic differences in language across time.

For figure 11.3, the cases were centered on judge interacted with year; this controls for any judge-level time-varying components of language. We then averaged and plotted by topic-year.
The labels and colors distinguish the seven-digit general issue topic. We can see that the document embeddings discriminate topics, effectively capturing differences in language across recognized issue areas.

Next we look at whether the vectorized language in the case vectors encodes information about judge characteristics. For figure 11.4, we centered on an interacted grouping for court, topic, and year. This centering controls for any time-varying topic and court-level language variation. We then averaged by judge and plotted the judge vectors. The labels and colors are by political party—Democrat or Republican. These are randomly distributed across the graph. It appears that the language features encoded by the document embeddings are not
Figure 11.3. t-SNE plot centered by judge-year, averaged by topic-year, labeled by topic.

informative about political party.

Figure 11.5 considers another judicial biographical feature: birth cohort. As before, we centered on court-topic-year and averaged/plotted by judge. In this case, the labels and colors are by birth cohort decade (1910s through 1950s). In stark contrast to political party, there is clear segmentation across the geometric space across cohorts. Remember that this is conditioned on court-topic-year, so is not driven by time trends over the sample. The vectorized language recovers differences in the legal language used by judges from different generations.

Finally, for figure 11.6, we consider law school attended as a final source of linguistic differences across judges. Conditional
on court, topic, and year, we see apparent random distributions across the space in terms of law school. As with political party, it seems like language or ideological differences by school do not show up in the vectors.

**Analysis of Relations Between Judges**

This section uses our vector representation of judges to produce a similarity metric between courts and judges. We adopt a measure of vector similarity that is used often for document classification—*cosine similarity* between two vectors,

$$s(v, w) = \frac{\bar{v} \cdot \bar{w}}{||\bar{v}|| \cdot ||\bar{w}||}$$

computed from the cosine of the angle between the vectors. In the case of word embeddings, high similarity means that the words are often used in similar language contexts.
In the case of judges, we can say that similarities approaching 1 mean that the judges tend to use similar language in their opinions. Similarities approaching -1 mean the judges rarely use the same language. In between, we have a continuous metric to rank which judges are relatively more or less similar to each other.

First we look at similarity between court vectors to complement the spatial representation in figure 11.1. We centered the vectors by topic and year and then aggregated by court. We then computed the pairwise similarities between the court vectors. These are reported in table 11.1.

The colors provide a gradient for similarity, with white and light gray indicating that the courts are relatively similar and dark gray meaning they are relatively dissimilar. The table has
some interesting features. First, the DC circuit is most similar to the Supreme Court of the United States, which is intuitive since they both have a large portion of their document made up of cases that involve issues of federal government functioning,
such as separation of powers. Second, the 11th Circuit is similar to the 5th Circuit, which is intuitive since the 11th Circuit used to be a part of the 5th Circuit and they share many legal precedents.

Unlike in the case of courts, the current model does not appear to capture similarity between judges. Starting with the Supreme Court, we center the document vectors on topic and year. Then we take the average of these centered vectors by judge as our representation of judge writing, reasoning, and beliefs. Unfortunately, this analysis yields few immediate insights—for example, although Justice Scalia is close in the space to Justice Thomas, he is even closer to Souter, Stevens, and O’Connor, a result that it is at odds with more conventional measures of ideological space based on their voting patterns (Epstein et al. 2007). Understanding these similarities is an area for future work. It could be that there are not enough decisions to be predictive of partisanship. Or perhaps these vectors are capturing something new about judge reasoning that is not captured by party.

Although the model does not perform well on Supreme Court justices, it is worth investigating whether appellate court judges might be better represented. For this we focus on one notable circuit court judge, Richard A. Posner, who holds well-known judicial views and has published over 3,300 opinions during his tenure. The document vectors are de-meaned by court, year, and topic and aggregated by judge. Based on this information, we rank all circuit court judges by the similarity of their vector to Posner’s vector.

These are reported in table 11.2. Interestingly, the most similar judge is Frank Easterbrook, who, like Posner, is known for the use of economic analysis in opinions. Stephen Breyer, of all recent justices, is most closely associated with the law and
economic movement (for example, he has a published article in *The Economic Journal* on “economic reasoning and judicial review” (Breyer 2009)). In addition to his law and economics orientation, Posner has a conservative reputation, and we see other conservative judges such as Neil Gorsuch and Antonin Scalia. Henry Friendly makes an appearance—he is a well-known pragmatist, as is Posner. Finally, Michael McConnell has coauthored academic articles with Posner (McConnell and Posner 1989). These document vectors are more intuitive for connections between circuit court judges but may raise some new interesting questions about connections between Supreme Court judges.

**Discussion and Future Work**

To recap, we applied the document vectorization algorithm Doc2Vec to 12 decades of opinion texts from US appellate courts. While previous work has applied LDA to large corpora of legal opinions, this is the first to introduce document embeddings to the area of empirical legal studies. Our analysis of the resulting vectors serves to validate their informativeness in terms of legal jurisdiction (distinguishing courts), time (distinguishing decades), and topics (distinguishing broad legal areas). These vectors would therefore be useful for downstream prediction or inference tasks on these categories. The method has the advantage of requiring few subjective decisions by the researcher, and the resulting features are easier to work with than high-dimensional sparse representations such as N-grams.

In terms of distinguishing judges, the results are more mixed. The vectors capture judge birth cohort and provide some intuitive rankings for the similarity of circuit judges to Richard Posner. But the vectors do not show a clear signal for political party or judge law school. Similarities between
Table 11.2. Most similar circuit court judges to Richard A. Posner.

<table>
<thead>
<tr>
<th>Circuit judge name</th>
<th>Similarity</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSNER, RICHARD A.</td>
<td>1.000</td>
<td>1</td>
</tr>
<tr>
<td>EASTERBROOK, FRANK H.</td>
<td>0.663</td>
<td>2</td>
</tr>
<tr>
<td>SUTTON, JEFFREY S.</td>
<td>0.620</td>
<td>3</td>
</tr>
<tr>
<td>NOONAN, JOHN T.</td>
<td>0.596</td>
<td>4</td>
</tr>
<tr>
<td>NELSON, DAVID A.</td>
<td>0.592</td>
<td>5</td>
</tr>
<tr>
<td>CARNES, EDWARD E.</td>
<td>0.567</td>
<td>6</td>
</tr>
<tr>
<td>FRIENDLY, HENRY</td>
<td>0.566</td>
<td>7</td>
</tr>
<tr>
<td>KOZINSKI, ALEX</td>
<td>0.563</td>
<td>8</td>
</tr>
<tr>
<td>GORSUCH, NEIL M.</td>
<td>0.559</td>
<td>9</td>
</tr>
<tr>
<td>CHAMBERS, RICHARD H.</td>
<td>0.546</td>
<td>10</td>
</tr>
<tr>
<td>FERNANDEZ, FERDINAND F.</td>
<td>0.503</td>
<td>11</td>
</tr>
<tr>
<td>EDMONDSON, JAMES L.</td>
<td>0.501</td>
<td>12</td>
</tr>
<tr>
<td>KLEINFELD, ANDREW J.</td>
<td>0.491</td>
<td>13</td>
</tr>
<tr>
<td>WILLIAMS, STEPHEN F.</td>
<td>0.481</td>
<td>14</td>
</tr>
<tr>
<td>KETHLEDGE, RAYMOND M.</td>
<td>0.459</td>
<td>15</td>
</tr>
<tr>
<td>TONE, PHILIP W.</td>
<td>0.459</td>
<td>16</td>
</tr>
<tr>
<td>SIBLEY, SAMUEL</td>
<td>0.459</td>
<td>17</td>
</tr>
<tr>
<td>SCALIA, ANTONIN</td>
<td>0.456</td>
<td>18</td>
</tr>
<tr>
<td>COLLOTON, STEVEN M.</td>
<td>0.445</td>
<td>19</td>
</tr>
<tr>
<td>DUNIWAY, BENJAMIN</td>
<td>0.438</td>
<td>20</td>
</tr>
<tr>
<td>GIBBONS, JOHN J.</td>
<td>0.422</td>
<td>21</td>
</tr>
<tr>
<td>BOGGs, DANNY J.</td>
<td>0.420</td>
<td>22</td>
</tr>
<tr>
<td>BREYER, STEPHEN G.</td>
<td>0.414</td>
<td>23</td>
</tr>
<tr>
<td>GOODRICH, HERBERT</td>
<td>0.412</td>
<td>24</td>
</tr>
<tr>
<td>LOKEN, JAMES B.</td>
<td>0.410</td>
<td>25</td>
</tr>
<tr>
<td>WEIS, JOSEPH F.</td>
<td>0.408</td>
<td>26</td>
</tr>
<tr>
<td>SCALIA, ANTONIN (SCOTUS)</td>
<td>0.406</td>
<td>27</td>
</tr>
<tr>
<td>BOUDIN, MICHAEL</td>
<td>0.403</td>
<td>28</td>
</tr>
<tr>
<td>RANDOLPH, A. RAYMOND</td>
<td>0.397</td>
<td>29</td>
</tr>
<tr>
<td>MCCONNELL, MICHAEL W.</td>
<td>0.390</td>
<td>30</td>
</tr>
</tbody>
</table>
Supreme Court judges and the similarities of circuit judges to Supreme Court judges are not intuitive or informative.

One interpretation of these results is that judicial language is not very politicized along partisan lines. This would be consistent with the finding in Ash, Chen, and Liu (2017) that judicial language is less polarized than congressional language. In contrast, differences in policy approach, such as use of economic analysis, might be more salient for distinguishing judges. This would be consistent with the finding in Ash, Chen, and Naidu (2017) on the importance of economics (language and training) in judicial decision-making.

But there is the alternative possibility that Doc2vec representations of language are not rich enough to encode some dimensions of judicial ideology. Richer representations, such as those constructed from grammatical relations between words (Levy and Goldberg 2014), may be needed. Another possibility is that document embeddings may require a large number of documents to form coherent ideological dimensions. Understanding the limitations of Doc2Vec and related models is important for future research.

In the rest of this section, we outline some parallel and future work in using embeddings models for empirical analysis of law. This research may serve to address the limitations identified during the research for this chapter.

**ANALOGIES AND ASSOCIATIONS**

The exploratory analysis taken in this chapter focused on distances and similarities. But the metric nature of these embedding models allows for richer analysis using matrix algebra. In particular, an intriguing use of word embeddings is to encode analogies. A well-known example is that word embeddings “know” that “man” is to “woman” as “king” is
to “queen,” through the vector algebra king - man + woman = queen (Mikolov, Sutskever, et al. 2013). In a legal milieu, Ash (2016) shows that “personal income tax” – “person” + “corporation” = “corporate income tax.”

Dai, Olah, and Le (2015) show that document embeddings also encode analogical relations between documents, with an application to Wikipedia articles. The document vector for the “Christina Aguilera” article, minus the vector for the “America” article, plus the vector for the “Japan” article, results in the vector for the Japanese pop star “Ayumi Hamasaki.” In the case of the law, a document embedding could say something like “Everson v. Board of Education is to Engel v. Vitale as Griswold v. Connecticut is to Roe v. Wade.” These cases share an analogical relation, in that the latter case is a related application of the constitutional principle articulated in the former case. In the vector math, that would be represented as Everson – Engel + Griswold = Roe. Finally, a judge embedding could say something like “Scalia is to Thomas as Ginsburg is to Breyer,” in the sense that Scalia – Thomas + Breyer = Ginsburg.

This discussion of analogies is exemplary of the feature that directions in the embedding space encode semantic meaning, for example, those related to singular-vs.-plural, verb tense, etc. Bolukbasi et al. (2016) show how to isolate a vector direction for a semantic concept such as gender in the embedding space. Construct a list of word pairs that share the gendered analogical relation (man–woman, men–women, boy–girl, boys–girls, father–mother, etc.) and then take the average of the vectors defined by the pairwise differences. This “gender” vector defines a semantic concept rather than any particular word or pair of words. It can then be used to identify and analyze the use of gendered language.

In the law, we would be interested in isolating other types
of language dimensions—notably, legal and political concepts and distinctions. For example, there might be a direction for liberal vs. conservative or procedural vs. substantive. Similarly, it would be worth investigation whether there are directions or clusters for originalists or pragmatists, or economic analysis vs. more traditional doctrinal methods.

Some of the recent work on embeddings has used these associational features to analyze partisan language (Iyyer et al. 2014) or to analyze cultural biases such as sexism and racism. Caliskan, Bryson, and Narayanan (2017) show that the results of implicit association tests are reproduced in aggregate language associations. Garg et al. (2018) and Kozlowski, Taddy, and Evans (2018) use long-run historical corpora to analyze trends in biased language over the last century.

The issue of unconscious bias is of particular significance in the legal system (e.g., Fagan and Ash 2017). Rachlinski et al. (2009) show that trial judges demonstrate the same implicit biases as the broader population on a standard psychological test, but that test was confidential, and it was not matched with the judges’ actual decisions. The construction of a language-based measure of bias, available for all judges from their written opinions, would be quite useful for understanding the importance of prejudice in the judicial system.

Toward this end, Ash, Chen, and Ornaghi (2018) analyze implicit associations in judicial language. The broad descriptive results from Caliskan, Bryson, and Narayanan (2017) are replicated in the judiciary, and relations between “innocent” and “guilty” are also analyzed. Male names tend to have a stronger connotation with “guilty” (relative to “innocent”) than female names. In addition, African American- and Hispanic-associated names are more closely related to guilty than Caucasian-associated names.
Future work should analyze whether and how these biases in language are associated to biases in decisions. One could ask, for example, whether judges with a racially based lexical bias also tend to reject discrimination complaints or to give longer criminal sentences to certain defendants. Similarly, having more traditional gender views, as detected in one’s implicit gender bias, might be reflected in more conservative judicial decisions related to gender discrimination cases. We could also look for peer effects by testing whether sitting with a biased judge has an impact on a peer judge’s subsequent decisions.

STRUCTURED EMBEDDINGS AND CATEGORICAL EMBEDDINGS

The document embeddings developed in the previous section were trained on the whole corpus. The embedding model did not explicitly model a time component, a court component, or other metadata categories. Differences across courts, time, and judges were encoded only through aggregating by different categories. Future work might explicitly account for differences between these categories in how the embeddings are constructed.

Along these lines, recent work in embedding models seeks to include these relations more flexibly and elegantly as a part of the data-generating process. Rudolph and Blei (2017) provide a model for learning dynamic embeddings and examine how language has changed over time in the US Congress over the last century. Rudolph et al. (2017) provide a model for structured group embeddings and allow word and document vectors to have a group component and an individual component.

In parallel work, we found difficulties in initial applications of structured embeddings to judge groups (Ash, Chen, and Ornaghi 2018). Word similarities across groups seem to be
sensitive to model parameters. Systematic differences in word similarities between Republican and Democrat judges can flip based on the embedding dimension and vocabulary size, for example. While structured embeddings do not work off the shelf, we expect that there is still potential in this research area.

As discussed, the Doc2Vec embedding was not able to discriminate between judges based on assumed ideology. This may be because the language style of written decisions may not encode ideology. This information may be mostly contained in the direction of the decision (e.g., for or against plaintiffs) or in some interaction between the decision and the language. Embedding layers in the deep-learning literature provide an alternative approach for identifying spatial relations between judges in prediction of decisions.

As described, Word2Vec and Doc2Vec work by colocating words that are most similarly predictive for a deep-learning task. In that case, a word is the embedded categorical variable but embedding layers can be used for any sort of categorical variable. In future work, the judge identity could be represented with an embedding lookup layer to a relatively low-dimensional dense vector space. The location of the judge vectors, initialized randomly, would be endogenous to the model. As the model goes through further training, the locations of these vectors would be pushed around to improve predictiveness. As a by-product of the model, the judges that locate together in the vector space would be predicted to behave similarly in court, holding other factors constant. This type of model may work to analyze ideological dimensions of judging.

EMBEDDING OF CITATION NETWORKS

In this chapter, the focus has been on the language of opinions as representing legal ideas. But in a common law system, the
cases cited in an opinion are another potential means to express ideological content. Ash, Chen, and Liu (2017) show that citations are more predictive of the political party of a judge than the writing style. Therefore, in the context of the geometry of law, citations could be included as features in the document embedding. This might reveal more differences, such as those between political parties.

Another approach to embedding citations is based on Rudolph et al. (2016) and Ruiz, Athey, and Blei (2017). In that paper, the model predicts occurrence of a product in a grocery shopping cart based on the cooccurrence of other products. In the legal analogue, cases could be treated as a bundle of citations to precedents in the same way that Rudolph et al. (2017) treat grocery baskets as a bundle of products. The citation embedding model would predict the presence of a particular citation using the list of cooccurring citations. As with word embeddings, cases that tend to be cited together would locate near each other in the embedding space. The model would thereby construct a “precedent space” as opposed to a language space.

An intriguing feature of the grocery cart model is that the learned parameters encode complementarity or substitutability of items. In the context of Rudolph et al. (2017), that means coffee being substitutable with tea but complementary with milk, for example. In the context of the law, we would learn which precedents are complementary (tending to be cited together) and which are substitutable (tending to appear in similar contexts but not together). By pairing substitutability metrics with ideological valence (liberal vs. conservative), we can analyze the parallel histories of liberal and conservative jurisprudence in the United States.
Conclusion

One of the fundamental challenges to using text for purposes of data analysis is the very high dimensional nature of textual artifacts. Various tools have been developed to reduce this dimensionality while preserving the kind of information that is useful to researchers, and some—such as topic modeling—have become popular as a tools for legal scholarship. Word and document embeddings have important advantages as a dimension reduction tool that researchers in law may find useful. This chapter describes a word embeddings model for a legal corpus and discusses some of the legal information, such as jurisdiction and time, that the embeddings model seems to capture well. Based on these early results, there is every reason to believe that there is substantial opportunity for future work to use these models to address significant questions of interest to scholars in empirical legal studies.

Acknowledgments

We thank Brenton Arnaboldi, David Cai, Matthew Willian, and Lihan Yao for helpful research assistance. We thank Michael Livermore, Daniel Rockmore, and participants at the Santa Fe Institute Law as Data Workshop for helpful feedback on this research.