

# Using Text Analytics to Predict Litigation Outcomes: A Preliminary Assessment

Charlotte S. Alexander,<sup>\*</sup> Khalifeh al Jadda,<sup>†</sup>  
Mohammad Javad Feizollahi,<sup>‡</sup> Anne M. Tucker<sup>§</sup>

<sup>\*</sup>To whom correspondence should be addressed.

## *Introduction*

This chapter describes the goals, methodologies, and preliminary results of an ongoing litigation outcomes prediction project conducted by the Legal Analytics Lab (the Lab) at Georgia State University. The Lab was

---

<sup>\*</sup>Georgia State University, J. Mack Robinson College of Business, 35 Broad Street, Room 1142, Atlanta GA 30303; calexander@gsu.edu. Director, GSU Legal Analytics Lab. We thank the student members of the sprint team, Harry Alex, Ayushri Bhargava, Colt Burnett, Vivian Chew, Chris Cirelli, Pearson Cunningham, Brad Czerwonky, Nathan Dahlberg, Fei Drouyor, Hayden Hillyer, Ziyang Huang, Amanda Iduate, John Lesko, Xiaotong Li, Siddhant Maharana, Ojasvi Maleyvar, Vahab Najari, Babak Panahi, Lucas Perdue, Kyle Price, TJ Sizemore, Pallavi Srinivas, Renate Walker, Zhe Wang, Chad Williams, and Caroline Xu; and the sprint sponsor, Amanda Farahany of Barrett and Farahany.

<sup>†</sup>Georgia State University, J. Mack Robinson College of Business, 35 Broad Street, Atlanta GA 30303; khalifeh@southernndatascience.com

<sup>‡</sup>Georgia State University, J. Mack Robinson College of Business, 35 Broad Street, Atlanta, GA 30303; mfeizollahi@gsu.edu

<sup>§</sup>Georgia State University, College of Law, 85 Park Place, Atlanta, GA 30303

established in 2017 as a site for collaboration between the university's data science and analytics faculty housed within the business school, and the faculty of law. Faculty collaborators pursue a variety of grant-funded research projects under the umbrella of the Lab. In addition, teams of faculty, graduate analytics students, and law students conduct what are known as "sprints," or semester-long, short-term projects commissioned by outside sponsors (law firms or companies) that focus on discrete real-world, law-related data problems. The project described here began as a sprint, conducted during the spring 2018 semester on behalf of a plaintiffs' side employment law firm in Atlanta, and is continuing as a stand-alone research project within the Lab.

The initial sprint project took as its subject all employment law cases filed and closed in the U.S. District Court for the Northern District of Georgia in the period 2010 – 2017. This included 5,111 cases, for which we had approximately 8,600 court documents in PDF form (complaints, magistrates' reports and recommendations on summary judgment, and district court judges' summary judgment decisions), and all docket sheets in a .csv file, containing about 200,000 text entries.

The law firm that sponsored the sprint was interested in answering a set of descriptive and predictive questions.

Descriptive:

- What was the frequency and distribution of case-ending events: settlement pre-discovery, settlement after discovery had begun, dismissal, granted motion for summary judgment for plaintiff, granted motion for summary judgment for defendant, and trial?
- In summary judgment decisions, what were the legal doctrines used and cases cited most frequently by judges?
- Could we identify defense lawyer and judge "playbooks"?

Predictive:

- What features of a lawsuit, observable at different points in the intake and litigation processes, predicted its case-ending event?

The law firm envisioned three uses for the deliverables that the sprint would generate. The first was to improve the firm's intake process by identifying characteristics of successful cases, defined as those that settle pre-discovery or withstand a defendant's motion for summary judgment. The second was to gain intelligence on the litigation strategies used by opponents and favored by judges. The third was to develop an empirical portrait of judges' summary judgment behavior on the Northern District of Georgia, given other researchers' findings of relatively high summary judgment grant rates in that district (Eisenberg and Lanvers 2008).

The research team, in turn, had its own set of goals: to develop code that would classify all entries on a docket sheet into stages in a "life cycle" model of litigation; to write code that would extract all case law citations from a document and count their frequency (including short and long form citation formats); to test the application of techniques such as topic modeling to fact-heavy documents such as summary judgment decisions; and, more generally, to experiment with whether litigation outcomes are susceptible to predictive modeling.

The remainder of this chapter describes the data assembly process, methodologies employed to extract features from our text and build a predictive model, preliminary results, and areas of continuing work. Along the way, the chapter offers observations about the challenges inherent in applying data science techniques to legal text. This chapter is thus primarily a methodological description, rather than a discussion of substantive results, though preliminary results are reported briefly throughout.

## ***I. Data Assembly***

The project began with a set of docket sheets and court documents that we received from the law firm sponsor. The firm had originally assem-

**Table 1.** *NOS Code Distribution.*

<b>NOS Code</b>	<b>Frequency</b>	<b>Percent</b>
442 Civil Rights – employment	2,596	51
710 Fair Labor Standards Act	1,934	38
445 Americans with Disabilities Act – employment	429	8
751 Family and Medical Leave Act	152	3
Total	5,111	100

bled the materials from the U.S. Courts’ Public Access to Court Electronic Documents (PACER) system and paid the associated download fees. Together, this material was associated with 5,111 lawsuits that had opened and closed in the U.S. District Court for the Northern District of Georgia within the study period and bore one of PACER’s four employment law-related Nature of Suit (NOS) codes: 442 Civil Rights- employment; 445 Americans with Disabilities Act- employment (ADA); 710 Fair Labor Standards Act (FLSA); and 751 Family and Medical Leave Act (FMLA). In rough terms, these codes cover lawsuits concerning employment discrimination of all types (442 and 445), wage and hour violations (710), and disputes around employees’ family and medical leave from work (751).

The NOS code is assigned by the plaintiff or his/her lawyer at the time a case is filed, chosen from a list on a required document known as a civil cover sheet (“U.S. Courts, Services and Forms, Civil Cover Sheet” n.d.). Plaintiffs may choose only one NOS code; they are instructed to “select the most definitive” if more than one could apply. Table 1 shows the distribution of NOS codes across cases in our data set.

Notably, relying on the self-reported NOS code as a threshold filtering device likely produced results that were both under- and over-inclusive, sweeping in cases that did not, in fact, contain claims made under that statute, and excluding cases that may have contained those statutory allegations, but others were deemed more “definitive” by the plaintiff. There-

fore, this project did not actually, as stated in the Introduction, “take as its subject all employment law cases filed and closed in the U.S. District Court for the Northern District of Georgia in the period 2010 – 2017,” but instead all cases in which the plaintiffs or their lawyers deemed an employment law claim to be “most definitive.” Christina Boyd and David Hoffman have recently explored this issue and proposed a series of smart reforms to the NOS code assignment process that would greatly improve the quality of case-filing data for researchers and policy-makers (Boyd and Hoffman, forthcoming 2018).

In addition to the NOS code, the plaintiff or plaintiff’s attorney also assigns each case a “cause of action” classification. The civil cover sheet instructs, “Cite the U.S. Civil Statute under which you are filing (Do not cite jurisdictional statutes unless diversity),” and directs the plaintiff to provide a “brief description of cause” (“U.S. Courts, Services and Forms, Civil Cover Sheet” n.d.). Within PACER’s data, the values of these fields appear as something like “29:216(b) Labor - Minimum wage or overtime compensation.” The cause classification thus provides a greater level of detail on the statutory violations and claim types alleged within any given lawsuit.

However, attempting to align NOS codes with their associated cause classifications reveals some strange pairings. Of the 2,596 cases with “442 Civil Rights- employment” as their NOS code, four percent had cause classifications that were seemingly unrelated to civil rights violations in an employment setting (e.g. “Qui tam False Claims Act,” which allows private plaintiffs to sue federal contractors for defrauding the government, or “Libel, assault, slander”) or causes that had their own, separate NOS code (e.g. “29:2601 Family and Medical Leave Act” or “42:12101 et seq. Americans with Disabilities Act of 1990”). The other NOS code categories fared better: one percent or fewer of cases with FLSA and ADA NOS codes had mismatching cause classifications, and no cases with the FMLA NOS code were mismatches.

There could be multiple explanations for NOS code-clause classification mismatches, and for the higher rate of mismatches within the 442 NOS code category specifically. Instead of treating the cause classification as providing greater detail about the same claim captured by the more general NOS code, the plaintiffs or their attorneys may have treated the cause classification as a way to record additional claims present in the case, beyond the single claim captured by the NOS code. Or the plaintiffs or their attorneys may have been mistaken or confused.

Further, with respect to the higher mismatch rate within the 442 Civil Rights- employment NOS code, cases with employment discrimination claims may be particularly likely to include additional, non-discrimination employment law claims. As the author has explored in previous joint work with Camille Gear Rich and Zev Eigen, plaintiffs' lawyers report adding non-discrimination employment law claims to their employment discrimination lawsuits, to increase those cases' viability in the face of perceived hostility toward discrimination claims by the federal courts (Alexander, Eigen, and Rich 2016). The cause classification mismatches within the 442 NOS code may therefore be capturing these add-on claims, reflecting the plaintiffs' lawyers' belt-and-suspenders approach.

Thus, while using PACER's NOS code or claim classification fields to construct a corpus of docket sheets and court documents is problematic, it can be the best approach out of a relatively bad set of data assembly options. An alternative strategy would require downloading all docket sheets and complaints, for all NOS codes, from PACER for a given court within a given time period, and then parsing the complaint text to classify a case according to its statutory allegation and claims made. Because PACER charges ten cents per page downloaded, such a cast-a-wide-net strategy would be cost-prohibitive.<sup>1</sup>

---

1. The charge applies to docket sheets and party-filed documents. There is no charge for judges' opinions accessed via PACER's Written Opinions Report (WOR), which, according to the U.S. Judicial Conference, is supposed to contain "any document issued by a judge or

Other sources for docket sheets and court documents such as Westlaw, LexisNexis, or Bloomberg Law allow keyword searching to identify relevant lawsuits, but do not make their search algorithms public, and produce sometimes dramatically different results sets. As an example, in a separate project, the author ran an identical search across all three vendors, which produced 9,712 results in Westlaw, 7,261 results in LexisNexis, and 5,644 results in Bloomberg Law.

In the end, as explained further below, the sprint team took the NOS code-filtered data set provided by the sponsor, and attempted to identify the full set of statutory allegations and claim types for each lawsuit from the text of the complaints themselves, when that text was machine-readable. Part II.b below explores the extent to which the plaintiff-provided NOS codes and cause classifications derived from PACER mapped onto the equivalent statute and claim type classifications generated by our code.

---

judges of the court sitting in that capacity, that sets forth a reasoned explanation for a court's decision." (PACER Service Center 2005). However, the author's own work has revealed that the coverage of the WOR is woefully inconsistent across judges and districts. For example, in a separate project examining judges' summary judgment decisions in employment law cases, two districts – Wyoming and the Southern District of Iowa – had zero WOR entries in the period 2008-2016, despite having hundreds of summary judgment decisions in employment law cases available on Westlaw. In addition, the number of WOR entries per civil case filed in a district (obtained from Federal Judicial Center data) – a normalized measure of WOR activity – ranged from a high of 0.55 for the Northern District of California to a low of 0.0007 in the District of South Dakota. Courts' ranking by WOR ratio roughly tracks the underlying number of civil cases filed, suggesting that higher-volume courts issued more opinions, or perhaps designated more of their opinions as free via the WOR. However, the Northern District of Georgia, which was the eleventh highest-volume court during the time period, was thirty-ninth with respect to its WOR ratio.

## ***II. Features Extraction***

Each of the 5,111 cases in the data set had an associated docket sheet and a complaint, and some had a magistrate's report and recommendation on summary judgment and/or a district court judge's ultimate decision on summary judgment. The sections below describe the methodologies employed to extract from each document type the features of the plaintiffs, defendants, lawyers, judges, claims, and litigation. These sections also provide preliminary summary results, addressing the three descriptive questions set out in the Introduction. The Part that follows then describes the predictive model into which we fed the extracted features as independent variables, in an attempt to predict any given lawsuit's case-ending event: settlement pre-discovery, settlement after discovery had begun, dismissal, granted motion for summary judgment for plaintiff, granted motion for summary judgment for defendant, or trial.

### ***a. Docket Sheets***

A docket sheet is a chronological index of all activity in a case, listing all documents filed by the plaintiff and defendant and all court actions in the case. Figure 1 provides a snippet of a docket sheet, downloaded from PACER's website and then converted to .csv format, for purposes of illustration.

In the sprint, the text of the docket sheet entries (the rightmost column shown in Figure 1) provided information about the "players" in each case: the district court judge's and magistrate judge's names, the number and names of the plaintiffs' and defendants' lawyers, and whether the plaintiff filed the case pro se or moved to proceed in forma pauperis, i.e. to be relieved of the requirement to pay a filing fee. In each instance, the students wrote relatively simple code to extract the key words or names associated with each feature from the relevant section of the docket sheet.



hard to read this -- graphic designer will work with authors to improve format  
probably should convert to native LaTeX table

case_number	activity_date	acti	docket_text	row	Rel	period
1:15-cv-02410-TWT	2015-07-06	1	COMPLAINT with Jury Demand filed by Marvin A. Guzman. ( Filing fee \$ 400.00 receipt number 113E-5915)	0	1	1
1:15-cv-02410-TWT	2015-07-06	2	Electronic Summons Issued as to Souto Foods, LLC. (eop) (Entered: 07/08/2015)	1	0	
1:15-cv-02410-TWT	2015-07-13	3	Return of Service Executed by Marvin A. Guzman. Souto Foods, LLC served on 7/9/2015, answer due 7/30/2	0	0	
1:15-cv-02410-TWT	2015-07-13	4	NOTICE Of Filing Reissued Summons by Marvin A. Guzman re 2 Electronic Summons Issued (Attachments: )	3	0	
1:15-cv-02410-TWT	2015-07-14	5	Electronic Summons Issued as to Souto Foods, LLC. (dr) (Entered: 07/14/2015)	4	0	
1:15-cv-02410-TWT	2015-07-28	6	ANSWER to 1 COMPLAINT with Jury Demand by Souto Foods, LLC. Discovery ends on 12/28/2015.(Bolet, A	5	1	2
1:15-cv-02410-TWT	2015-07-28	7	Certificate of Interested Persons by Souto Foods, LLC. (Bolet, Albert) (Entered: 07/28/2015)	6	0	
1:15-cv-02410-TWT	2015-07-29	NA	Clerks Notation re 7 Certificate of Interested Persons, OK, per TWT (ss) (Entered: 07/29/2015)	7	0	
1:15-cv-02410-TWT	2015-08-13	8	Application for Leave of Absence for the following date(s): 9/18/2015 - 09/24/2015, by Peter Andrew Lam	8	0	
1:15-cv-02410-TWT	2015-08-14	NA	Clerks Notation re 8 Leave of Absence 9/18/2015 - 09/24/2015, by Peter Andrew Lampros. Judge Thrash	9	0	
1:15-cv-02410-TWT	2015-08-26	9	JOINT PRELIMINARY REPORT AND DISCOVERY PLAN filed by Marvin A. Guzman. (Lampros, Peter) (Entered: 10	1	5.1	
1:15-cv-02410-TWT	2015-08-27	10	CERTIFICATE OF SERVICE of Discovery by Marvin A. Guzman.(Lampros, Peter) (Entered: 08/27/2015)	11	1	5.1
1:15-cv-02410-TWT	2015-08-27	11	SCHEDULING ORDER: re: 9 Joint Preliminary Report and Discovery Plan. Discovery ends on 12/28/2015. Su	12	1	5.1
1:15-cv-02410-TWT	2015-08-28	12	Initial Disclosures by Souto Foods, LLC.(Bolet, Albert) (Entered: 08/28/2015)	13	1	5.1
1:15-cv-02410-TWT	2015-11-02	13	Joint MOTION to Approve Settlement by Marvin A. Guzman. (Attachments: # 1 Exhibit Settlement Agree	14	1	8.2
1:15-cv-02410-TWT	2015-11-12	14	ORDER GRANTING 13 Joint Motion to Approve the Settlement Agreement. Signed by Judge Thomas W. Thi	15	1	8.2
1:15-cv-02410-TWT	2015-12-07	15	STIPULATION of Dismissal by Marvin A. Guzman. (Lampros, Peter) (Entered: 12/07/2015)	16	1	8.4
1:15-cv-02410-TWT	2015-12-08	NA	Clerk's Entry of Dismissal APPROVING 15 Stipulation of Dismissal with prejudice pursuant to Fed.R.Civ.P.41	17	1	8.4
1:15-cv-02410-TWT	2015-12-08	NA	Civil Case Terminated. (adg) (Entered: 12/08/2015)	18	1	11

Figure 1. Docket Sheet Extract.


The docket sheet text also provided a wealth of information about the litigation itself, requiring both simple and more sophisticated methodological approaches. On the simple end of the spectrum, the students calculated the number of days each case was open by extracting the first and last date on each docket sheet for each case, and counted the number of docket sheet entries – both could be used as rough proxies for a case’s level of activity and/or complexity. Students also identified whether the case had been removed to federal court from state court by identifying the presence the trigram, “Notice of Removal” in the first docket sheet entry, and counted the number of depositions taken by each party by identifying “Notice of Deposition” in proximity to the name of a lawyer for the plaintiff or defendant. The deposition count feature was used as a rough proxy for each party’s investment in the case, as well as the development of the factual record and an indication of factual complexity.

More sophisticated work – which remains ongoing – was required to capture the stages of litigation through which each lawsuit progressed, before concluding with one of six case-ending events: settlement pre-discovery, settlement after discovery had begun, dismissal, granted motion for summary judgment for plaintiff, granted motion for summary judgment for defendant, or trial.

replace with em-dash

might be interesting to hear some more about this analysis -- what decision tree algorithm? did you try other methods (e.g random forest)? How was the precision/recall?

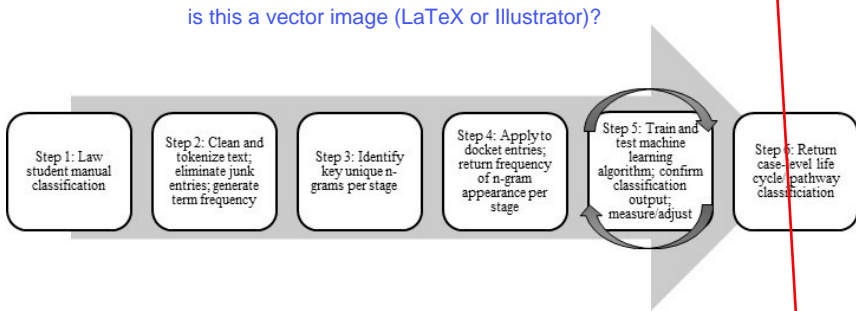
**Table 2.** *Litigation Life Cycle Stages.*



Stage Number	Stage Description
1	Complaint
2	Answer
3	Motion to Dismiss
4	Motion to Dismiss decision
5	Discovery
6	Motion for Summary Judgment
7	Motion for Summary Judgment decision
8	Settlement
9	Trial

Here, the students and faculty created a set of stages and substages through which a lawsuit might progress, shown in Table 2, and then developed a machine learning algorithm, using decision trees, that assigned each docket entry to a single stage, after first excluding some docket entries as junk. (Note that though “Settlement” appears as stage 8 in Table 2, it could occur at any point in litigation; the remainder occurred in sequence, though not every lawsuit reached every stage.) Law students manually classified approximately 1,000 docket entries into the life cycle stages as part of this process, and the predictive model then identified the key n-grams that were unique to each stage and persisted across all docket entries that were manually assigned to that stage. Using those n-grams, the model iterated across all docket entries. The docket entry-level classifications were then rolled up to a rougher set of case-level classifications, generating a picture of each case’s pathway through litigation. Figure 2 illustrates this process.

During the spring semester sprint, we were unable to identify whether a case-ending motion for summary judgment was granted in favor of the plaintiff or defendant. We were also unable to identify trials successfully, as they were exceedingly low-occurring events within our data set, and



**Figure 2.** *Docket Entry Classification Procedure.*

in civil litigation more generally. Therefore, our initial results grouped case-ending summary judgment decisions into a single event, and omitted trials. Work continues on these tasks.

Part IV below describes the ongoing refinement of our techniques, including distinguishing between granted motions for summary judgment for plaintiffs and for defendants, adding more granular litigation stages (e.g. default judgment and sua sponte dismissals), and using a **neural network** trained on a much larger set of docket sheet entries to achieve better classification performance.

Nevertheless, by the close of the spring semester, the team was able to classify eighty-four percent of the cases in our data set into one of nine pathways, ending in settlement pre-discovery, settlement after discovery had begun, dismissal, or a granted motion for summary judgment (for either party). Table 3 shows the pathways and their frequency; Figure 3 provides a visualization, where bar width represents number of cases that followed that pathway; Table 4 shows the frequency of each case-ending event for which we were able to generate data.

It is tempting to compare these results to the work of other researchers who have studied employment law case outcomes, with a particular focus on the way that employment discrimination cases end. In the Eisenberg and Lanvers study mentioned above, the authors found that the judges on the Northern District of Georgia granted summary judgment at

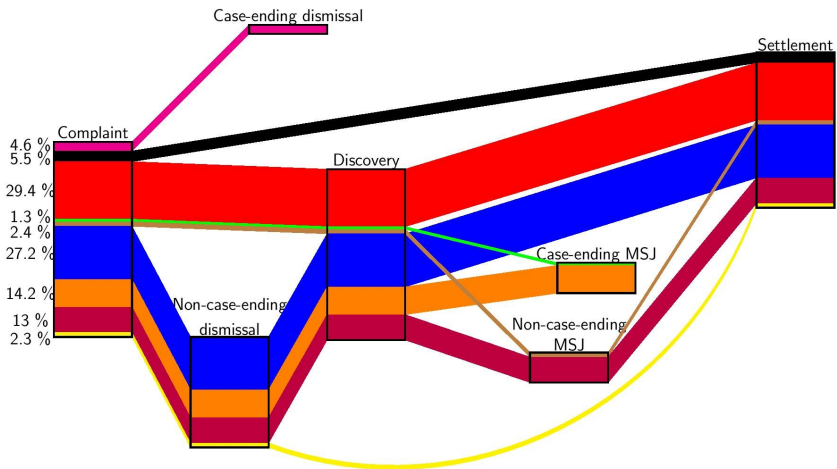
**Table 3.** *Pathways Through Litigation.*

<b>Pathway</b>	<b>Frequency</b>	<b>Percent</b>
complaint   case-ending dismissal	197	4
complaint   discovery   case-ending motion for summary judgment	57	1
complaint   discovery   non-case-ending motion for summary judgment   settlement	100	2
complaint   discovery   settlement	1,248	24
complaint   non-case-ending dismissal   discovery   case-ending motion for summary judgment	603	12
complaint   non-case-ending dismissal   discovery   non-case-ending motion for summary judgment   settlement	618	12
complaint   non-case-ending dismissal   discovery   settlement	1,152	23
complaint   non-case-ending dismissal   settlement	97	2
complaint   settlement	233	5
unknown	806	16

**Table 4.** *Case-Ending Events.*

<b>Event</b>	<b>Frequency</b>	<b>Percent</b>
Settlement	3,448	67
Pre-Discovery	552	16
Post-Discovery	2,896	84
Case-ending motion for summary judgment	660	13
Case-ending dismissal	197	4
Unknown	806	16
Total	5,111	100

this is a touch busy, and will need to be in gray-scale. Can work with SFI graphics editor.



**Figure 3.** Pathways Through Litigation Stages - Visualization.

a higher rate than judges from other districts. Further, as they state, “The most striking effect was the approximate doubling—to almost 25%—of the NDGA summary judgment rate in employment discrimination cases and a substantial increase in the NDGA summary judgment rate in other civil rights cases.” (Eisenberg and Lanvers 2008). In another study focused exclusively on the outcome of employment discrimination lawsuits, Nielsen, Nelson, and Lancaster found that forty percent of plaintiffs lost their Title VII claims on dispositive motions or at trial, fifty-eight percent of cases settled, and two percent of plaintiffs ultimately won at trial (Nielsen, Nelson, and Lancaster 2010). Finally, in a study of employment discrimination cases filed in federal courts between 1979 and 2006, Clermont and Schwab report an overall plaintiff win rate – regardless of stage – of fifteen percent (Clermont and Schwab 2009).

All of this research provides a useful baseline or base rate against which to measure the sprint project’s outcomes, and particularly our findings with respect to employment discrimination lawsuits’ case-ending events. Because of the high number of “unknown” pathways and un-

specified case-ending events, however, as well as the underperformance of our classification methodology, our results are not yet reliable at this stage. The other features extracted from the docket sheets – e.g. number of days open, number of docket entries, and number of depositions taken by plaintiffs and defendants – are all reliable and able to be included in the final predictive model described later in the chapter.

Moreover, this initial work on litigation pathways and case-ending events has created a foundation on which to build as we continue to refine our docket sheet classification code. This work will also allow us to answer the first and last of the descriptive questions listed above: the frequency and distribution of case-ending events, and lawyer/ judge playbooks. By overlaying judge and lawyer identifiers atop the lawsuits' litigation pathways, in addition to other filters such as claim type, we will be able to determine whether certain lawyers' or judges' cases follow certain patterns through the stages of litigation.

## ***b. Complaints***

In addition to docket sheets, another focus of the sprint project was the complaint filed in each case. Here, our goal was to extract from the complaints' text the allegations of statutory violations, claims nested within those allegations, defendant industry, and plaintiff characteristics including occupation and – for discrimination claims – race and/or national origin.

It soon became apparent, however, that the complaints would not be the rich source of information that we had hoped. Of the 5,111 complaint PDFs, only 3,263, or about sixty-four percent, were machine-readable. The unreadable complaints were either hand-written or scanned images, and were largely filed by pro se plaintiffs.

Using the readable 3,263 set, the first task was to extract all references to statutes from the text. The students and faculty developed a lexicon of all relevant statutes' full names and abbreviations, and variations thereof. This list was not limited to the options available as NOS codes or cause

classification choices on the civil cover sheet, but resulted from law students' initial reading of a sample of complaints to get a sense of the range of statutory violations alleged. Students then wrote code to generate a frequency count of those key terms' appearance within each document, and decided on an acceptable frequency threshold for classifying a complaint as containing an allegation under one or more statutes.

For the complaints that could not be classified using these methods, the students created a citation-finder that extracts citations to the U.S. Code and the Code of Federal Regulations from the text. Law faculty and students reviewed the output to classify the remaining complaints according to the statutes and regulations cited therein.

Table 5 reports the total number of allegations made under Title VII of the Civil Rights Act of 1964 (Title VII), the Age Discrimination in Employment Act (ADEA), the ADA, the FMLA, and the FLSA. These statistics capture whether a complaint alleged a violation of any of those five statutes. They therefore represent allegation counts, not lawsuit counts, so the total number exceeds the 3,263 machine-readable complaints. For the same reason, these totals are not directly comparable to the NOS code totals shown in Table 1 above, which capture only one NOS code per case. Nor do the statutes align perfectly with the NOS codes. However, for purposes of rough comparison, Title VII and ADEA together would fall under 442 Civil Rights- employment, and the remaining statutes would align with their own, stand-alone NOS codes (445 ADA, 710 FLSA, and 751 FMLA).

The methodology described above had varied success in replicating the statutes identified in either the NOS code or the cause classification derived from PACER for a given case. The sprint team's text analytics identified the presence of an ADA, Title VII, or FLSA violation allegation in a complaint in seventy-seven percent of the cases with either the relevant NOS code or that statute listed in the "cause" field. The figure for FMLA cases was ninety-seven percent, and eighty-two percent for ADEA cases. Without reading every complaint, it is hard to determine whether

**Table 5.** *Allegations of Statutory Violations Extracted from Complaint Text.*

<b>Statute</b>	<b>Frequency</b>	<b>Percent</b>
FLSA	1,600	45
Title VII	1,089	30
FMLA	377	11
ADA	258	7
ADEA	258	7
Total	3,582	100

the lower match rates are driven by problems with our text-based classification system, or problems with plaintiffs' assignment of NOS codes and cause classifications on the civil cover sheet. Additional work is ongoing to confirm the output and improve accuracy.

Moreover, this protocol was highly supervised. Another, less manual approach might have used machine learning to train an algorithm to recognize the statutory allegations and claim types within a complaint, but the time constraints of the sprint during the spring semester did not permit experimentation with unsupervised techniques on the text of the complaints. Lab faculty and research assistants are continuing to explore these possibilities in ongoing work.

In addition to classifying each case by the statute(s) mentioned in the complaints, the sprint team created an additional, more granular classification: claim type. Here again, the law student team members read a sample of complaints and constructed a lexicon of keywords and phrases that identified the particular claim being made under the statute that was alleged to have been violated. For most statutes, this required one step down in granularity. For ADA allegations, for example, the team identified keywords that were associated with the following claims: discriminatory hiring, promotion, discipline, transfer, demotion, termination, and constructive discharge; hostile work environment; retaliation, and reasonable accommodation. As in the statutory allegation classification



process, the team also established an acceptable keyword appearance frequency threshold.

For Title VII allegations, there were two steps down in granularity. First, the team identified a set of claim-type keywords, similar to the ADA list above, e.g. hiring, promotion, transfer, demotion. Next, the team identified keywords that were associated with the particular protected class that was at issue: race, sex, national origin, religion, and color. Finally, the team constructed a set of rules that identified all combinations of claim-type keywords and protected class keywords in close proximity to one another within the text, and established relevant frequency thresholds. This protocol allowed any complaint alleging a Title VII employment discrimination violation to be classified further as alleging hiring discrimination by religion, for example.

Applying this methodology to complaints that contained allegations under only one statute was relatively simple: the students simply ran the statute classification code and then the claim classification code, which produced both levels of classification. However, complaints that contained allegations of violations of more than one statute presented a trickier challenge, as some claims could be nested within more than one statute. For example, a single complaint might make both Title VII and ADA violations, and discrimination in the form of harassment and demotion. Because harassment and discriminatory demotion are actionable under both statutes, the task is to sort the claim types properly into their statutory buckets.

To solve this problem, the student teams wrote code that first identified the statutes, as described above, and then created a roughly two-sentence window before and after the statute identifiers as they appeared in the text. The code then searched for the claim-type keywords within that window. Here, the code searched only for the keywords that were relevant to the particular statute at hand – searching for overtime-related keywords only within FLSA windows, for example, and not in the windows around any of the discrimination-related statutes. Once the code

identified keywords within any given window, it moved to the next instance of a statute in the text and drew a new window, but did not permit the windows to overlap. Through this process, and again applying acceptability thresholds, the students were able to assign to each of the 3,263 machine-readable complaints one or more statute-claim classification pairs.

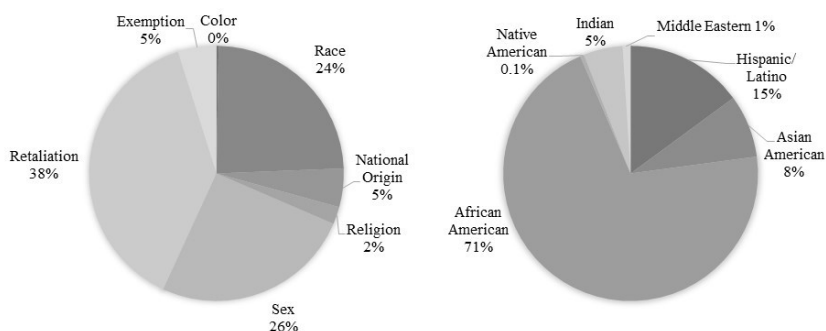
It is more difficult to compare the output of this code to the PACER-provided fields, because granular claim classifications may or may not appear in PACER's cause classifications. For example, all 152 cases with an FMLA NOS code also had "29:2601 Family and Medical Leave Act" as their cause classification, providing no additional detail. In contrast, the sprint team was able to classify FMLA cases into those that made "interference" and/or "notice" claims – two types of FMLA violations – introducing a greater level of granularity than what is available via PACER.

The clause classification field for cases with the 442 Civil Rights-employment NOS code did provide some detail as to Title VII cases, indicating the relevant protected class that we could then map onto our claim classification output. Here, the best match rate was for religious discrimination at seventy-seven percent, meaning that our code identified religious discrimination claims in the text of seventy-seven percent of the complaints which carried a PACER cause classification that indicated religious discrimination.

As with the statute classification process described earlier in this section, Lab researchers are working to improve the claim classification process in a variety of ways, including experimenting with less supervised methods such as topic modeling. The challenge there, however, is the large amount of case-specific factual detail that appears in complaints, which can result in topics that are not illuminating as to the types of allegations being made. Further, we are now focusing less on improving match rates of the kind reported here, and more on validating the output of the code by performing manual statute and claim type classifications. This is because the PACER-provided NOS code and claim classification

data is flawed in all of the ways described above, making it an unreliable metric against which to measure code performance.

Separate from the statute and claim classifications just described, the sprint team also constructed an additional set of classifications using the text of the available 3,263 complaints – work that is also still ongoing. First, for the cases that alleged race or national origin discrimination, the students wrote code that attempted to identify those plaintiffs’ specific race or national origin by extracting keywords from windows of text. This exercise produced the results shown in Figure 4, which also illustrates the Title VII claim types extracted from the text. The chart on the right can be read as providing more detail for the “Race” and “National Origin” slices of the chart on the left.



**Figure 4.** *Title VII Claim Types and Plaintiffs’ Race or National Origin Extracted from Complaint Text.*

Setting aside whether these text-derived distributions map onto the distributions of the PACER NOS codes and cause classifications, it is interesting to note that the results shown in Figure 4 are consistent with other research on the frequency of different types of Title VII claims. The chart on the left shows retaliation as the largest slice, followed by race discrimination claims; this squares with research that shows that the largest category of discrimination charge filed with the Equal Employment Opportunity Commission (a predecessor step to filing a Title VII

lawsuit in court) has been retaliation, followed by race discrimination, since 2010 (“U.S. Equal Employment Opportunity Commission, Charge Statistics, FY 1997 Through FY 2017” n.d.).

Finally, apart from race and national origin characteristics, the sprint team began work on identifying the occupation of the plaintiff and industry of the defendant, using both a keyword-centric approach that draws on the U.S. Bureau of Labor Statistics’ industry and occupation lexicons, and a less supervised natural language processing approach involving tools such as Part-of-Speech tagging and other semantic web methodologies.

In sum, the team’s work on the complaints is incomplete but promising. Yet it is important to remember throughout that the results presented here were derived from only sixty-four percent of the complaints associated with the full set of 5,111 cases in the study set. Further, the non-machine readable complaints were not randomly distributed across case and plaintiff types: large numbers were filed by pro se plaintiffs. This means that the preliminary results presented here are incomplete in another important way, as they do not capture all plaintiff types. Manual work may be needed to convert those complaints into a readable format; options include transcribing them in full by hand, or manually extracting the relevant features.

Stepping back from the particulars of this sprint project, this issue points to a much larger challenge that many legal analytics projects must face: that of data quality and access. If many law-related insights are locked up in unreadable PDFs, then the conclusions that can be drawn from only their readable counterparts are necessarily, and potentially troublingly, limited. Research will exclude populations with limited resources for representation, whose complaints are hand-written or so idiosyncratic that they are not susceptible to analysis. More work needs to be done in the areas of image processing and handwriting recognition to address this problem.

## *Summary Judgment Documents*

The final set of documents from which the sprint team extracted features of the lawsuits were the Reports and Recommendations (R&Rs) issued by magistrate judges on the parties' motions for summary judgment and the district court judges' final summary judgment opinions and orders (SJs).

This portion of the research was, in a sense, its own separate mini-project within the larger sprint, intended to answer the second descriptive question set out in the Introduction: In summary judgment decisions, what were the legal doctrines used and cases cited most frequently by judges? To this initial question, we added a question about the dynamics of judging: To what extent did district court judges adopt the magistrates' R&Rs in ruling on motions for summary judgment? In other words, where was the actual decision-making power exercised on summary judgment?

These questions swept in both case-ending summary judgment decisions, which had already been captured by the litigation life cycle work described in Part II.a above, and non-case-ending summary judgment decisions. Thus, the unit of analysis here became the summary judgment decision, not the lawsuit, which was the unit of analysis in the docket sheet and complaint portions of the research.

The team used a combination of strategies to answer these questions. First, we tried to use topic modeling on the corpus of R&Rs and SJs to get at the legal doctrines upon which judges were relying. Even after substantial cleaning steps, including the removal of all proper nouns, however, the topics that were produced were too specific to the individual facts of the cases to provide much insight into the legal rules that judges were deploying.

We then shifted to an approach that combined supervised and unsupervised techniques, in which the law students first identified keywords that were associated with doctrines upon which a judge might rely in deciding a summary judgment motion in any case, as well as in the particular types of employment law cases at issue in this research. Students

generated a simple frequency count for each key word, and selected a threshold for inclusion. Next, students used a Word2vec model to pull additional important terms from the context in which the keywords were used in the set of R&Rs and SJs. Specifically, students used continuous skip-gram architecture, which can predict a window of context words in which a given word appears. Faculty reviewed the context output to identify any additional terms that should be added to the keywords list, and the students then re-ran the frequency table. From the table, the students created a word cloud in Tableau, with filters by judge, and additional filters to be added for statute and claim types, year, and other case features.

Students also wrote code that extracted all citations to case law, in both long form and short form, from each R&R and SJ, and displayed citation frequency across the data set in a similar Tableau word cloud dashboard. Ongoing refinements to the citation dashboard include displaying citations on a per-opinion basis and creating greater filtering ability, as well as perhaps combining the keyword and citation dashboards into a single interface. Figure 5 provides a screenshot of each dashboard in its preliminary form.

Finally, the team turned to the additional question about the balance of summary judgment decision-making between magistrates and U.S. district court judges. The students began by writing code that would extract the magistrates' recommendations from the text of the R&Rs (grant, deny, partial), on the one hand, and the district court judges' action on the R&Rs (adopt, reject, partial) from the text of the SJs, on the other. However, this approach was stymied by the wide variation in the language that judges, and particularly the magistrates, used. Students briefly explored a less supervised approach, but lacked the number of documents and time necessary to train a machine learning model to classify R&Rs and SJs accurately by outcome.

The team then turned to the text of the docket sheet entries associated with R&Rs and SJs, realizing that the docket text was written in a more

standardized format. After first writing code that paired each R&R with its relevant SJ order, the students were then able to identify the frequency with which a district court judge adopted the magistrate’s recommendation. The students also determined which party filed the initial motion for summary judgment and whether the SJ order ended the case, as an alternative way to get at the same summary judgment-related case ending events explored in Part II.a above. Figure 6 gives two examples of the data structure produced by these protocols.

This code is still a work in progress, but preliminary results are shown in Figure 7.

In eighty-five percent of cases with both an R&R and an SJ order, the district court judge adopted the magistrate’s recommendation wholesale. This finding sheds light on where and how summary judgment decision-making was actually happening during our study period.

Work continues on this summary judgment-focused mini-project, including revisiting the task of extracting outcomes from the R&R and SJ text, and exploring how best to exploit the text of judges’ opinions to identify patterns.

### ***III. Predictive Model***

Taken together, the processes described above generated a set of features that we could then feed into a random forest model, as explanatory or independent variables, to attempt to predict cases’ termination in one of four case-ending events: dismissal, motion for summary judgment, pre-discovery settlement, and post-discovery settlement. As our techniques for classifying cases according to their case-ending events improve, we will add to this list to capture which party won on summary judgment and will add more events, including trial, default judgments, and sua sponte dismissals (as distinguished from rulings on a motion to dismiss). We will then re-run our predictive models.

Using our preliminary data, we constructed four models, using only the information that would be available to a plaintiffs’ attorney at four

points in time: (1) the pre-filing intake stage, (2) early in litigation just after a case is filed, (3) at the close of discovery, and (4) in a state of “omniscience,” our term for an all-in model involving the full set of known case features. We assigned features as independent variables only to the version of the model in which that information could be known. As an example, the judge assigned to a case is a feature that would not be known at pre-filing intake; therefore, that variable was included only in models 2-4 and omitted from 1.

The model was the richest at the omniscience stage (4), incorporating twenty-eight case features; when it simulated intake (1), only ten features were available. Using each relevant set of available features, our four models attempted to bucket cases into the four case-ending events.

The preliminary results of our modeling are shown in Figure 8, along with the four most important features in each. Accuracy is listed; this refers to the model’s success in bucketing cases by outcome, as described above.

Though this output is extremely preliminary, one interesting result is the predominance of the attorneys’ own prior case handling patterns as important predictors of case outcomes. Notably, this echoes other researchers’ findings on the importance of attorney-related explanatory variables in predicting litigation outcomes (Ashley 2017).

This result is not useful as a business matter: the law firm sprint sponsor is not helped at intake by its own historical case outcome rate. Nevertheless, these attorney-centric results may be acting as proxies for some filtering or case selection process performed by plaintiffs’ attorneys, indicating that attorneys’ own non-quantitative predictive modeling may be quite effective. Further work is needed to unpack these results, and to attempt to extract further case features from the docket sheets and court documents that might themselves be the subject of the attorneys’ own filters.

Here, the gap between a purely predictive approach to an analytics problem and an inference-based, causation-focused, data modeling ap-



proach becomes clear. If a given variable, such as attorneys' records, predicts with high accuracy how a case will end, that may be enough to satisfy some business goals, and the "why" question need not be answered. However, if the "why" is not self-evident from the nature of the variable itself, or if the "why" is important in shaping future actions or interventions, then the utility of the sort of predictive modeling employed here decreases.

In his insightful commentary on what he calls "the two cultures" within statistical modeling, Leo Breiman puts it this way, "Approaching problems by looking for [an inference-based, causation-focused] data model imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems." (Breiman 2001). On the other hand, techniques such as the random forest model used here "are A+ predictors. But their mechanism for producing a prediction is difficult to understand." *Id.*

The sprint research team is continuing its modeling work, considering different regression model specifications (consistent with the former, data modeling approach), and continuing to refine its decision tree modeling (consistent with the latter). By jumping between the two cultures, and deploying each one's tools, we hope to come closer to answering the predictive question posed above: What features of a lawsuit, observable at different points in the intake and litigation processes, predicted its case-ending event?

#### ***IV. Continuing Work***

As the foregoing Parts have noted, Lab researchers are still engaged with the tasks that remain unfinished or unrefined after the end of the spring semester sprint project. In general terms, researchers are testing approaches that are less supervised, and less reliant on expert-generated keywords. These endeavors require more data than our 5,111 subject lawsuits can provide, however.

As a solution, the team is accessing the RECAP archive of 2.2 million docket sheets and millions more court documents from all ninety-four U.S. district courts available via Court Listener, a non-profit, free legal search engine run by Free Law Project. The law firm sprint sponsor will also likely provide a complete set of docket sheets, R&Rs, and SJs from three additional U.S. district courts.

This larger data set will enable the use of a neural network or other deep learning techniques to process the docket sheet entries and more accurately classify them into a more complete set of litigation stages. The team continues to experiment with various machine learning approaches to classifying complaints by statutory allegations and claim types, and R&Rs and SJs by outcome.

Taken together, these tasks are akin to bench science in this growing field called, variously, legal analytics, legal informatics, or computational law. The Legal Analytics Lab team is continuing to experiment, to explore the potential of the tools of data science to unlock text, detect patterns, predict outcomes, and derive insight.

## ***Citations & Bibliography***

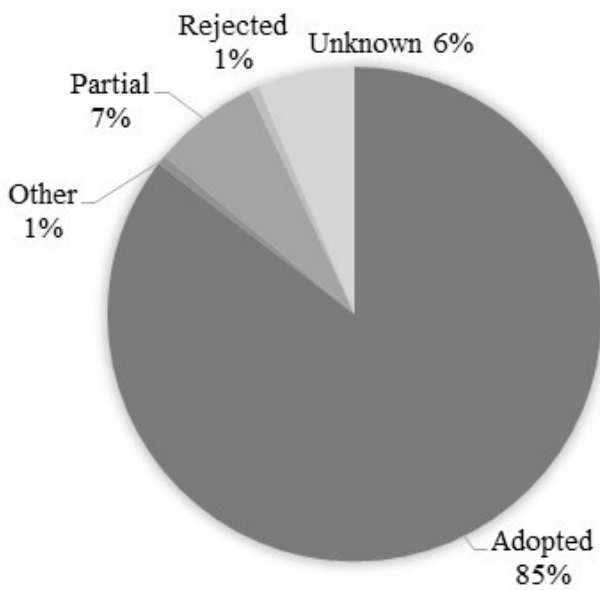
## References

- Alexander, Charlotte S., Zev J. Eigen, and Camille Gear Rich. 2016. "Post-Racial Hydraulics: The Hidden Dangers of the Universal Turn." *New York University Law Review* 91:1–58.
- Ashley, Kevin D. 2017. *Artificial Intelligence and Legal Analytics*. Cambridge: Cambridge University Press.
- Boyd, Christina L., and David A. Hoffman. forthcoming 2018. "The Use and Reliability of Federal Nature of Suit Codes." *Michigan State Law Review* xx:xx–xx.
- Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16:199–215.
- Clermont, Kevin M., and Stewart J. Schwab. 2009. "Employment Discrimination Plaintiffs in Federal Court: From Bad to Worse?" *Harvard Law and Policy Review* 3:103–132.
- Eisenberg, Theodore, and Charlotte Lanvers. 2008. "Summary Judgment Rates Over Time, Across Case Categories, and Across Districts: An Empirical Study of Three Large Federal Districts." *Cornell Law School working paper No. 08-22*.
- Nielsen, Laura Beth, Robert L. Nelson, and Ryon Lancaster. 2010. "Individual Justice or Collective Legal Mobilization? Employment Discrimination Litigation in the Post Civil Rights United States." *Journal of Empirical Legal Studies* 7:175–201.
- "U.S. Courts, Services and Forms, Civil Cover Sheet." n.d. PACER Service Center. <http://www.uscourts.gov/forms/civil-forms/civil-cover-sheet>.
- "U.S. Equal Employment Opportunity Commission, Charge Statistics, FY 1997 Through FY 2017." n.d. PACER Service Center. <https://www.eeoc.gov/eeoc/statistics/enforcement/charges.cfm>.



case_number	docket_text	paired_order_text	filer	order_adopted	case_ending
1:10-cv-00007-JEC	FINAL REPORT AND RECOMMENDATION recommending GRANTING 22 MOTION for Summary Judgment as to all of Plaintiff's claims. Signed by Magistrate Judge Gerrilyn G. Brill on 06/14/11. (Attachments: # 1 Order for Service)(fap) (Entered: 06/15/2011)	ORDER regarding Magistrate Judge's Final Report and Recommendation. IT IS HEREBY ORDERED that the Court ADOPTS the Magistrate Judge's Final Report and Recommendation 28 GRANTING defendants' Motion for Summary Judgment as to all of plaintiff's claim 22 . The Clerk is directed to close this action. Signed by Judge Julie E. Carnes on 7/5/11. (cem) (Entered: 07/06/2011)	D	Adopt	TRUE
1:10-cv-00383-ODE	REPORT AND RECOMMENDATION that Defendant's 42 MOTION for Summary Judgment be GRANTED IN PART AND DENIED IN PART. Specifically, the undersigned RECOMMENDS that the motion for summary judgment be DENIED with respect to Plaintiffs' Title VII claims, DENIED with respect to Plaintiffs' S. 1981 claims, and GRANTED with respect to Plaintiff Lewis, Power, and Johnson's claims under the Equal Protection Clause. Signed by Magistrate Judge Alan J. Baverman on 1/31/2012. (rej) (Entered: 02/01/2012)	ORDER that City of Kennesaw's Objections are 56 SUSTAINED and the 55 Report and Recommendation is ADOPTED IN PART AND REJECTED IN PART. The City of Kennesaw's 42 Motion for Summary Judgment is GRANTED in accordance with Federal Rule of Civil Procedure 56(a) because Plaintiffs have failed to establish a prima facie case of discrimination. Signed by Judge Orinda D. Evans on 3/30/2012. (anc) (Entered: 03/30/2012)	D	Partial	FALSE

**Figure 6.** *Docket Sheet Analysis Output Extract.*



**Figure 7.** *District Court Judges' Actions on Magistrate Judges' Report and Recommendations on Motions for Summary Judgment.*

Model 1: Pre-Filing Intake Accuracy 67%				Top 4 Predictors	
Actual \ Predicted	Case ending dismissal	Case ending SJ	Settlement	Plaintiffs' attorneys' settlement rate	Plaintiffs' attorneys' total previous cases
Case ending dismissal	1105	9	222	Plaintiffs' attorneys' dismissal rate	
Case ending SJ	53	14	22	Year filed	
Settlement	351	6	371		
Model 2: Early Litigation Accuracy 80%				Top 4 Predictors	
Actual \ Predicted	Case ending dismissal	Case ending SJ	Settlement	Plaintiffs' attorneys' settlement rate	Defendants' attorneys' dismissal rate
Case ending dismissal	781	0	4	Plaintiffs' attorneys' dismissal rate	
Case ending SJ	57	0	1	Defendants' attorneys' settlement rate	
Settlement	339	0	110		
Model 3: Close of Discovery Accuracy 92%				Top 4 Predictors	
Actual \ Predicted	Case ending dismissal	Case ending SJ	Settlement	Non-case ending dismissal	Defendants' attorneys' settlement rate
Case ending dismissal	1309	0	27	Plaintiffs' attorneys' settlement rate	
Case ending SJ	40	27	22	Defendants' attorneys' dismissal rate	
Settlement	66	1	661		
Model 4: Omniscience Accuracy 94%				Top 4 Predictors	
Actual \ Predicted	Case ending dismissal	Case ending SJ	Settlement	Non-case ending dismissal	Defendants' attorneys' dismissal rate
Case ending dismissal	773	0	25	Plaintiffs' attorneys' settlement rate	
Case ending SJ	15	30	7	Defendants' attorneys' settlement rate	
Settlement	20	0	422		

Figure 2. Prediction Model Predictions Results (Predicted vs. Actual)