Scenario Analytics

Analyzing Jury Verdicts to Evaluate Legal Case Outcomes

Jack G. Conrad
Thomson Reuters
Research & Development
610 Opperman Drive
Saint Paul, Minnesota 55123, USA
jack.g.conrad@thomsonreuters.com

Khalid Al-Kofahi
Thomson Reuters
Center for Cognitive Computing
120 Bremner Blvd.
Toronto, Ontario M5J 0A8, Canada
khalid.al-kofahi@thomsonreuters.com

ABSTRACT

Scenario Analytics is a type of analysis that focuses on the evaluation of different scenarios, their merits and their consequences. In the context of the legal domain, this could be in the form of analyzing large databases of legal cases, their facts and their claims, to answer questions such as: Do the current facts warrant litigation?, Is the litigation best pursued before a judge or a jury?, How long is it likely to take?, and What are the best strategies to use for achieving the most favorable outcome for the client? In this work, we report on research directed at answering such questions. We use one of a set of jury verdicts databases totaling nearly a half-million records. At the same time, we conduct a series of experiments that answer key questions and build, sequentially, a powerful data-driven legal decision support system, one that can assist an attorney to differentiate more effective from less effective legal principles and strategies. Ultimately, it represents a productivity tool that can help a litigation attorney make the most prudent decisions for his or her client.

CCS CONCEPTS

•**Information systems** → **Data analytics;** *Environment-specific retrieval;* Clustering;

KEYWORDS

Data Mining, Data Analysis, Decision Support Systems, Legal Applications, Evaluation

ACM Reference format:
Jack G. Conrad and Khalid Al-Kofahi. 2017. Scenario Analytics. In *Proceedings of ICAIL '17, London, United Kingdom, June 12-16, 2017,* 10 pages.
DOI: http://dx.doi.org/10.1145/3086512.3086516

1 INTRODUCTION

Conceptually, a significant proportion of what knowledge workers do, including attorneys, can be described as falling under one of three tasks: *find* information, *analyze* information found and *decide* based on such analysis. Certainly attorneys do other things, including representing and negotiating on behalf of their clients, but the categorization above is nonetheless a useful generalization. It provides us with a directional guide as to the kind of capabilities we need to develop to support legal professionals.

We observe that, from a content perspective, legal information providers like Thomson Reuters and LexisNexis understood the importance of supporting each of these knowledge tasks as evident by the types of content they publish. For example, indices and citator databases are intended to help the *Find* task. Case summaries, the key number system and analytical material are designed to help legal practitioners with the *Analyze* and *Find* tasks, while practice guides are designed to help attorneys with the *Decide* task. But from a technology perspective, with a few exceptions, most applications focus on the Finding task. Certainly this is the case for applications targeted at legal researchers.

This work is an attempt to inject technology (data mining, natural language processing and machine learning) into the last two tasks: Analyzing and Deciding. Our central hypothesis is that within frequent disputes certain patterns repeat and that practitioners would benefit from seeing such patterns comprised of facts, claims, counter-claims, legal principles applied, analysis and decisions. But statistical discovery of patterns is only possible in large datasets. Otherwise one could not distinguish between incidental occurrences and actual patterns. Theoretically, one could discover strong correlations between certain patterns and outcomes. But given the large number of dimensions (variables) in the data, this requires an even larger dataset.

Scenario Analytics permits attorneys to interact with the data, identify patterns and formulate and validate hypotheses. This data-driven approach helps parties formalize more effective legal strategies. It also provides a better communication tool between attorneys and their clients because attorneys are able to use data to explain decisions, as opposed to solely basing decisions on experience. This is certainly not the first work that attempts to support a data-driven decision support process (e.g., Lex Machina [5]), but our work is different in that it does not focus on providing summary statistics on how judges ruled on certain motions, but focuses on unearthing deeper patterns in the data.

The remainder of the paper is organized as follows. Section 2 provides some of the underlying motivations behind this work. In Section 3, we review related research. Section 4 describes the data set that we harness for our research. We report on our experimentation in Section 5. In Section 6, we discuss an important jurisdictional-based component associated with a state's treatment of negligence for cases. We summarize our results in Section 7 and draw additional conclusions in Section 8. Finally, in Section 9, we outline future work.

ICAIL '17, London, United Kingdom
 978-1-4503-4891-1/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3086512.3086516

2 MOTIVATIONS

The motivation for this work is to introduce data-driven decision-making processes into the practice of law. This does not only make the system more efficient, but also increases the equitable application of the law. We chose jury verdicts and settlements because they cover a diverse set of litigation categories in all 50 U.S. states. That they consist of shorter paragraphs of free text, created using editorial guidelines, made them a prime target for the kind of analysis that we had in mind. Applying such analysis to, e.g., case law documents would have been a much harder task in our opinion.

Given such a repository, it should be possible to identify, organize and analyze the underlying fact patterns and legal strategies used for similar cases and, moreover, to determine which strategies have been more effective and which less effective. For such scenarios, effectiveness can be defined in terms of key parameters such as award levels or trial lengths. Such an approach is not intended to replace an attorney; rather, it can provide a means for legal practitioners themselves to weigh their strategic options and to select the most promising among them. Beyond this, it would be possible to harness this data in such a way as to build predictive models for parameters such as award level and trial length, for instance, to forecast the award and duration of a trial based on the chosen strategy. Of course such models would need to be created and refined based on a specific jurisdiction and litigation type (e.g., premises liability and injury from slipping on uncleared pavement) along with additional features associated with properties like plaintiff profile (gender, age, health, personal limitations, etc.). We have taken a series of investigative steps to validate some of these propositions in the act of researching and developing foundational tools for realizing such a decision support application and the litigation strategy assistance it can provide.

3 PRIOR WORK

The origin of fact-based information retrieval arguably stems from the field of legal case-based reasoning [7, 11]. Case-based reasoning is the act of devising solutions to unsolved problems based on pre-existing solutions for problems of a similar nature. Case-based reasoning developed out of the more general field of artificial intelligence, which, unlike general computing applications, tries to solve more general purpose problems, and at least in theory, strives to replicate the general functioning of human intelligence. Expert Systems or Knowledge-Based Systems (KBS) are subsets of case-based reasoning. And legal knowledge-based systems are yet a domain-specific application of KBS [4].

Late in the first decade of the 2000's Morales and Moens conducted research in the field that would come to be known as "Argumentation Mining" or "Argument Mining" [10]. This work was extended in subsequent workshops of the same name [12] and in subsequent works at ICAIL [2] and the AI and Law Journal [15]. Argumentation mining aims to automatically extract structured arguments from unstructured textual documents. Recent advances in machine learning methods promise to enable breakthrough applications for such expert assistance via information technology: something that a few years ago was not deemed feasible. In this survey article, the authors introduce argumentation models and methods, review existing systems and applications, and discuss challenges and perspectives of this relatively new research area [9].

More recently, legal applications in the field of IP litigation have focused on, for example, early case assessment tools that can give practitioners a notion of their chance of success in pursuing a case, based on the success rate of cases with the same features [5].

4 DATA

Thomson Reuters possesses several collections of jury verdicts and settlements records. Largest among these is the LRP-JVS database. It contains approximately 400,000 records spanning all 50 states and covering a relatively broad range of topics such as premises liability, medical malpractice and employment discrimination.[1] These records contain a wide variety of informative fields such as the following:

- date of activity (accident, filing, trial or settlement)
- event-type (rear-end collision, sexual harrassment, ...)
- docket no.
- jurisdiction (county, state, court)
- case-type (liability, discrimination, malpractice ...)
- description (general and specific)
- injury type (primary, secondary ...)
- award (award category, award range, exact award)
- damage summary (plaintiff profile)
- unstructured textual description, including
  - fact paragraph
  - plaintiff claims
  - defendant claims

In total, there are over 25 fields of case-related information in each JVS record. An example of the unstructured textual description is shown in Figure 1. It includes a section containing the seminal facts of the event (green), the plaintiffs' claims (blue) and the defendants' claims (red). These textual summaries are authored by our company's employees who are trained to use a standard, semi-closed vocabulary in describing the facts and claims of a case.

These unstructured textual summaries accompany a rich set of metadata that is partially itemized above. The cardinality of the major metadata elements is presented in Table 1.

No.	Information Type	No. Uniq. Entries
1	State Jurisdiction	52
2	Court	3,278
3	Gen. Description of Event/Accident (NP)	530
4	Spec. Description of Event/Accident (NP)	5,483
5	Primary Injury (NP)	2,674
6	Secondary Injury (NP)	2,673
7	Case Type [Liability / Other (e.g., Discrim.)]	376,921 / 13,481

Table 1: Key Metadata Fields Associated with LRP-JVS Documents

Roughly three-quarters of the cases in the LRP-JVS collection receive an award granted, while the remaining one-quarter receive zero award (Table 2 and bullets). One can also determine award percentages per jurisdiction, for example,

- State with the highest award percentage:
  - West Virginia (2668/2874) 92.83%

[1] In these experiments, we use the largest of our JVS databases, LRP-JVS. In total, our JVS databases comprise over 500K records.

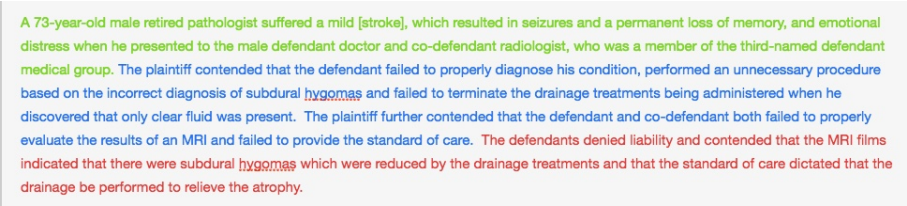

Figure 1: Segmented Unstructured Text Summary in LRP Jury Verdicts Record

Totsl Cases w/ Awards	286,785
Total Cases w/ Zero Awards	103,617
Total Cases	390,402
Percent Cases w/ Awards	73.5%

Table 2: Cases with/out Awards

- State with the lowest award percentage:
  - Massachusetts (5450/9451) 57.67%

Some of the most frequently occurring event types recorded in the LRP-JVS database are show in Table 3.

1. General Description	Award Count	No Award	Total	% Award
2. Rear-End Collision	44,894	11,338	56,232	79.84%
3. Premises Liability	31,562	14,300	45,862	68.82%
4. Doctor Malpractice	10,658	11,272	21,930	48.60%
5. Products Liability	10,955	4,269	15,224	71.96%
6. Pedestrian Accident	10,484	3,682	14,166	74.01%

Table 3: Most Prevalent Event-types in LRP-JVS Documents

In addition to categories and descriptions of jury verdict events, the LRP database also has a series of fields containing information about the award associated with the cases (Table 4).

Value Type	Value Range	Count	Percent of Total
VAL0	$0	104,617	26.79%
VAL1	$1 - $49,000	167,003	42.77%
VAL2	$50,000 - $99,999	26,760	6.85%
VAL3	$100,000 - $199,999	23,119	5.92%
VAL4	$200,000 - $499,999	25,576	6.54%
VAL5	$500,000 - $999,999	15,807	4.04%
VAL6	$1,000,000 - $1,999,999	12,244	3.13%
VAL7	$2,000,000 - $4,999,999	9,273	2.37%
VAL8	$5,000,000 - Up	6,431	1.63%
TOTAL	$0 - Up	390,402	100.00%

Table 4: Distribution of Awards by LRP Defined Categories

There are different ways in which to represent the distribution of awards in JVS cases. From a data science perspective, a series of equally sized bins, e.g., $25K each, may be most typical. The problem with such bins is that with award categories in the tens of millions of dollars, we would need to create hundreds of bins. For that reason, we use the increasingly larger bins shown in Table 4 for reporting purposes.

5 EXPERIMENTS

We have harnessed our LRP jury verdicts and settlements collection to conduct a series of experiments that explore the prospects of creating an analytical and predictive tool for litigating attorneys. We have designed a series of NLP technology-supported tasks that would be instrumental to the development of the kind of legal decision support system application described above. These tasks or research functions are investigated in the sub-sections that follow.

5.1 Automatic Topic Classification

In any operational environment where previously unseen jury verdicts are processed, it is essential to automate the classification process along topical lines. The Key Number System (KNS) represents a legal taxonomy consisting of approximately 100,000 leaf nodes and 200,000 total nodes. The depth of the taxonomic tree ranges from 3 to 11, with the average depth being about 6. The system is maintained by Thomson Reuters. There also exists a key number assigner classification tool that has been trained on O(10M) editorially produced and KNS-classified points of law (a.k.a. headnotes). We conducted an experiment wherein we selected jury verdict documents from three distinct litigation areas, applied the key number assigner to the unstructured text portion of these documents' factual descriptions of the case and subsequently had a separate set of legal domain experts judge the key numbers assigned to the documents. One of the motivations behind this effort was to potentially leverage the KNS to classify the unstructured textual descriptions of the facts and plaintiffs' claims, group like classes of these summaries, and to further analyze similar classes being litigated by the plaintiffs' attorneys.

To grade the key number assignments, we relied on a 5-point Likert scale where 5 represented "on point" classifications, 4 highly relevant, 3 correct, 2 close to topic and 1 poor classification. Cohen's Weighted Kappa Score is used to measure the degree of disagreement between the editors.[2] The results can be seen in Table 5.

What these results illustrate is that, in general and with a substantial degree of editorial agreement, the automatic key number assignments are reliable for the task of capturing essential features of the facts and plaintiff-side arguments for representing the case. This is especially true since we used an out-of-the-box version of the key number assigner and not one specially trained on jury verdict summaries. As such, this performance represents a type of lower bound. With more focused training data representing the LRP summaries themselves, these assignments could potentially become still more accurate.

5.2 Clustering Plaintiff Claims

Given the ability to automatically organize incoming jury verdict summaries by reasonably fine-grained topic, the next logical step in

[2]To address not only the agreement between editors, but also the *degree* of disagreement, the Weighted Kappa Score is used. The greater the disagreement between the editors, the greater the weight that comes into play. http://www.real-statistics.com/reliability/weighted-cohens-kappa/

No.	Categories	No. Reviewers	No. Cases	Mean Score	Percent	Weighted Kappa
1	Premises Liability	2	50	4.4/5.0	88%	0.925
2	Medical Malpractice	2	50	4.1/5.0	82%	*
3	Racial Discrimination	2	50	3.9/5.0	79%	0.553

Table 5: Human Assessment of Automatic Key Number Assignments to Jury Verdict Claims

an analysis process is to differentiate one set of cases from another within these classes based on the underlying legal principles or strategies they invoke. For this task, we segment the jury verdict summaries into four sections: background facts, plaintiff claims, defendant claims, and remaining details. We then apply a k-means clustering algorithm over the plaintiff claims for values of k in the low single digits. Other clustering algorithms were also explored, including agglomerative, partitional (e.g., repeated bisection) and graphical. They did not produce results that were superior to k-means. We used the NLTK 3.0 toolkit to conduct the clustering [3].

In order to distinguish one set of clustered plaintiff claims from another in terms of utility, we use a metric that helps differentiate those claims that have been more effective from those that have been less effective. The metric is based on the award behavior for a given cluster and is called the 'award_quotient.' We define award_quotient as the ratio of a cluster's non-zero awards to its zero awards:

$$award\_quotient = \frac{\text{(cases w/ non-zero award)}}{\text{(cases w/ zero award)}} \quad (1)$$

The idea behind the award_quotient is that the metric is augmented when cases result in a positive award and diminished when cases result in a zero award. It provides a means of quickly identifying when a cluster has a high degree of awards. In addition to award_quotient, we have examined a number of other characteristic features that can help contrast one cluster's properties from another. One feature that has occasionally factored into differentiating one cluster from another is average length in tokens of the plaintiff claims. Based on our empirical study, we used a stopword list of about 100 terms that was expanded with additional terms that should not be permitted to impact the clustering, e.g., 'plaintiff' or variations thereof, alternatives for 'claimed' (contended, asserted, alleged, argued, maintained, ...), and age modifiers (e.g., year-old). We conducted trials where 'male' and 'female' and their variants were both included and excluded from our stopwords, except for certain discrimination cases, where they were not stopped.

We conducted a series of clustering experiments as above using three distinct litigation areas: Premises Liability, Medical Malpractice and Racial Discrimination, the latter set being roughly one magnitude smaller than the others due to the less frequent nature of discrimination cases. Upon forming different clustered claims in this manner, we examined distinct language patterns associated with each. In addition, we look for clusters with higher than average award_quotients (e.g., $AQ >= 2.5$) along with other distinguishing properties, including case details. In cluster C_3 under Premises Liability in Table 6 below, for example, we see a higher award_quotient (in bold), albeit with an unremarkable average token length relative to the others.

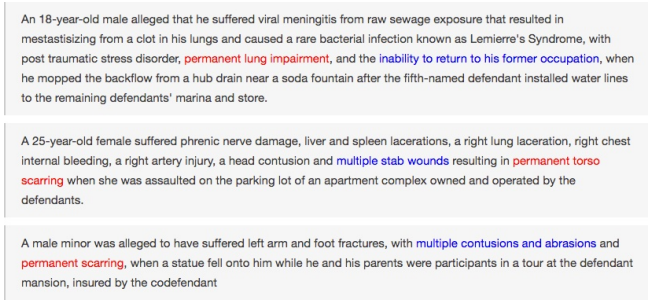

Figure 2: Language Patterns in Clustered Plaintiff Claims

When we examine the language used in clusters like this, we notice the plaintiffs' attorneys emphasizing certain details in the case. In Figure 2, we see examples from these cases in which the attorneys are underscoring the permanent nature of the injuries (scarring or lung impairment), and emphasizing the multiple injuries that the victim received (multiple stab wounds, contusions, abrasions), and in at least one case, where the plaintiff is unable to return to his former occupation. Although this study appears anecdotal, our repeated empirical examination of language patterns in each of the clusters has corroborated these kind of distinct per-cluster patterns.

In subsequent experiments, using our complete set of over 37K premises liability cases, we found that clusters consisting of plaintiff claims with longer average length tended to carry more substantive and meaningful evidence of consistent language patterns. This can be seen quantitatively in Table 7 where we filter out records with relatively short plaintiff's claims and only use records with claims consisting of 55 or more tokens. This number was determined empirically.

Among the complete set of premises liability cases (row 1, Table 7), one can see that cluster 2 has the highest award_quotient of the group, approaching 2.5. A more informative view of these clusters and their award distributions can be seen in Figure 3. One can note clearly that only for cluster 2 is the first zero-award bar (in blue) significantly surpassed by the first non-zero-award bar (in red).

The two clusters with the highest award quotients in Table 7, among the longer plaintiff claims, are clusters 0 and 3, both with quotients well above 3.0. Some of the representative language associated with these clusters is shown in Figure 4, including, "known dangerous condition." "failed to provide ..." , "failed to remedy or warn ..." , "failed to ensure ...", "failed to properly maintain ..." In cluster 0, there is also the notion of repetition and a compounding of the evidence, including emphasis upon "known dangerous conditions." By contrast, in cluster 3, there is the notion of "failed to inform ...", "failed to correct ...", "failed to adequately ..." as well as "negligently leave ..." This is what would distinguish these clusters from the properties of the others. We have found that, in

No.	Categories	Cases	C_0 Mean Lgth	C_0 Award Ratio	C_1 Mean Lgth	C_1 Award Ratio	C_2 Mean Lgth	C_2 Award Ratio	C_3 Mean Lgth	C_3 Award Ratio	C_4 Mean Lgth	C_4 Award Ratio
1	Premises Liability	5,460	63.1	(664/666) = 0.997	43.1	(239/187) = 1.278	39.0	(472/479) = 0.985	**41.2**	**(640/241) = 2.656**	42.5	(890/816) = 1.091
2	Medical Malpractice	3,709	38.3	(122/374) = 0.326	39.3	(79/266) = 0.297	71.1	(221/995) = 0.222	37.8	(186/473) = 0.393	42.6	(227/585) = 0.338
3	Racial Discrimination	213	30.8	(5/10) = 0.500	39.2	(35/21) = 1.667	27.5	(21/17) = 1.235	46.3	(22/19) = 1.157	27.4	(22/31) 0.710

Table 6: Properties by Topic, Award Patterns

Categories 1	Cases	C_0 Mean Lgth	C_0 Award Ratio	C_1 Mean Lgth	C_1 Award Ratio	C_2 Mean Lgth	C_2 Award Ratio	C_3 Mean Lgth	C_3 Award Ratio	C_4 Mean Lgth	C_4 Award Ratio
Premises Liability	37,048	36.9	(3001/1564) = 1.92	36.9	(2010/1139) = 1.77	**27.0**	**(5146/2136) = 2.41**	30.7	(6778/4043) = 1.68	37.3	(4966/2952) = 1.68
Premises Liability (Lgth > 55)	2,513	**75.5**	**(177/51) = 3.47**	65.0	(435/244) = 1.78	70.2	(267/120) = 2.23	**72.5**	**(435/136) = 3.20**	74.0	(463/185) = 2.50

Table 7: Properties by Topic, Award Pattern (Premises Liability specific)

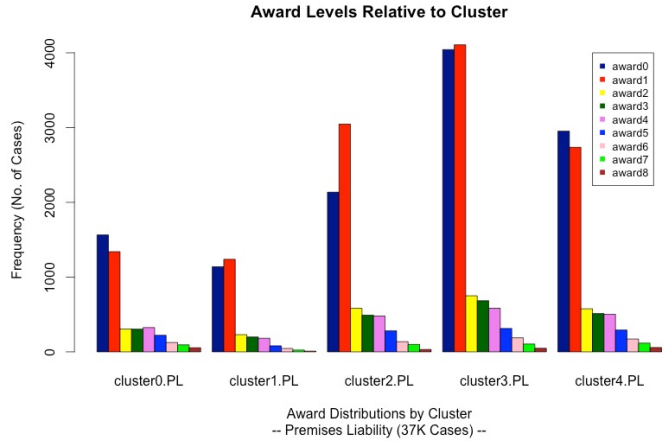

Figure 3: Award Distributions by Cluster

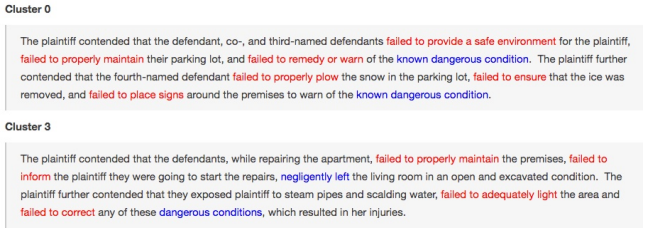

Figure 4: Language Patterns: Longer Clust. Plaintiff Claims

general, values of $3 \leq k \leq 6$ produce the most effective results. Clearly optimal values of k can vary depending on litigation type and underlying fact pattern.

The instances presented above are intended to serve as informative illustrations of utility rather than any form of definitive findings. They present some of the properties of the most effective evidence and language, treated here as a proxy for arguments. We also note that such evidence is envisioned to be used as exemplar material in a legal decision support application rather than definitive results from a fully automated system. Ultimately, they would serve to inform the attorney-client of what some of the most effective techniques thus far marshalled have been.

5.3 Associating Language Patterns with Award Distributions

Given the ability to systematically cluster cases on the same topic and to differentiate these cases from one another based on their award_quotient and other features such as language patterns, a useful next step in the analysis process would consist of presenting the actual award *distributions* associated with each of these cluster sets. An example of such a presentation is shown in Figures 10 and 11 located in the Appendix. Here topically clustered cases assigned the same key number are shown in terms of the cluster's award distribution. The topics in this example come from Employment Discrimination cases.

For finer-grained analysis purposes, we divide the distributions shown into three distinct categories, (a) those that follow a standard Zipfian curve shape (red rectangle), (b) those whose zero-award bar is overshadowed by the award bars immediately to the right (blue rectangles), and (c) those whose zero-award category is negligible (green rectangles). Below each distribution in Figure 11 is the key number associated with it. Figure 10 provides the textual labels for each of the key numbers and their distributions appearing in the red rectangle while Figure 11 does the same for the key numbers appearing in the blue and green rectangles. One can observe that there are no dominant or concentrated themes among the KN labels in Figure 10, and the Zipfian curves they are associated with. They run the gamut of discrimination and harrassment-related categories under EMPLOYMENT PRACTICES, and even a couple outside the area of EMPLOYMENT PRACTICES, involving FEDERAL EMPLOYMENT.

By contrast, the KN labels shown in Figure 11 tied to the blue rectangles and their diminuative zero-award bars and the green rectangles and their negligible zero-award bars, are more homogeneous

and coherent; they are aligned with the topics of "sex discrimination", "sexual harrassment" and "retaliation for exercise of rights." This evidence suggests that "Retaliation for exercise of rights" is usually rewarded by juries, which are instances where both race and gender can factor in. And again in cases where "affirmative action" and "remedial action" factor in, non-zero awards are also forth coming. Given a data-driven tool built atop jury verdicts like these, one that would permit a litigation attorney to identify attractive award distributions, and subsequently to explore the language used to produce them, the attorney could then determine what legal principles could be harnessed to wager similar arguments and to produce similarly favorable outcomes for the client. As such, the tool would be an effective means of testing and formulating various legal strategies to consider for the trial.

Whereas these samples are not extensive, the patterns identified are nonetheless valid, and whether zero or negligible 'No Awards' are indicated, these types of categories may still be worth tracking in larger samples or other JVS data sets. Another observation is that as compensatory as these types of cases are, we recognize that attorneys are not at liberty to 'manufacture' such facts for the benefit of their current client. Figures 10 and 11 are nonetheless shown as illustrations of the kinds of differences that exist among these distributions, and how attorneys can sift through them to distinguish between less remarkable and more remarkable case distributions.

5.4 Analyzing Relations between Trial Length and Award Level

The two properties of a case that a litigation attorney would be most attentive to are trial length and award level. For this reason, in a separate investigation we examined relationships between the length of a trial and the level of award in our LRP jury verdicts collection. We investigated competing hypotheses regarding this relationship, described below.

5.4.1 *Short Case Hypotheses.*

A No Award: The case had virtually no merit and was quickly dismissed.

B Award: The case and culpability of the defendant was so absolutely clear and had few mitigating circumstances (e.g., certain rear-end collisions), that the case and its award were promptly determined.

5.4.2 *Long Case Hypotheses.*

A No Significant Award: The case was complex and difficult to assign blame because of the complicated nature of the issues. As a result, no large or requested amount was awarded.

B Significant Award: The case was complex and took time to sort through and analyze all of the issues, but once that was done, and blame was sufficiently determined, a significant award was granted.

The LRP jury verdicts record contains up to three court dates. These include (1) incident date; (2) filing date; (3) trial or settlement date. We calculated the trial length by taking the difference between (2) and (3). We examine this variable generally and at the state level, especially for all of the states examinined in Table 8. These states were chosen for the role they play in Section 6 on Negligence. They nevertheless represent six sizable states whose combined populations cover roughly one-third of the total U.S. population.

The mean trial length ranged from 15.5 months (North Carolina) to 29.4 months (Illinois).[3]

No.	State	Population	Mean Trial Length (mos.)	Total LRP Recs
1	Maryland	5,976,407	19.2	6,500
2	North Carolina	9,943,964	15.5	5,504
3	California	38,802,500	20.5	31,609
4	Florida	19,893,297	26.6	18,499
5	Georgia	10,097,343	25.4	7,581
6	Florida	12,880,580	29.4	16,048
Cum.	Six States	96,594,091		85,741

Table 8: Mean Trial Length for Six Select States

In addition, we created a stacked bar graph representing the distribution of awards across the award categories we presented in Table 4. These are shown in Figure 5.

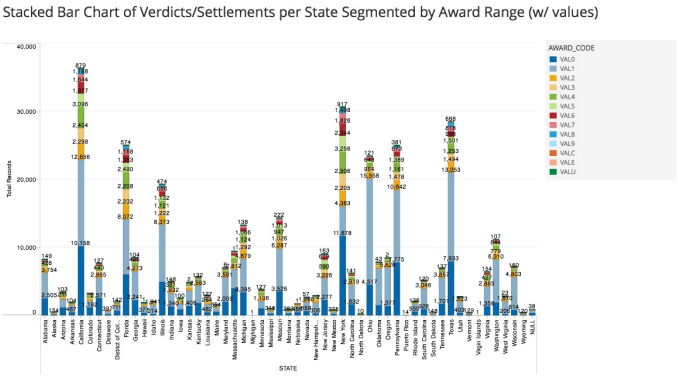

Figure 5: Distribution of Award Categories by State

As is clear from the subsections above, trial length and award level statistics may be informative to some extent, but are of limited utility of if examined separately. For this reason, we combined the two in a single plot, Trial Length vs. Award Level for non-zero award levels (Figure 7) and Trial Length vs. Award Level for zero award levels (Figure 6). We selected these data points from premises liability cases from our largest source of state data collected: California. The Pearson Correlation Coefficient for the combined set of 360 recent cases shown in Figures 6 and 7 is 0.123 while the same coefficient for the set with zero awards removed is 0.170 (p-value < 0.05 for each) [13].[4] The removal of zero awards improves the correlation, yet both coefficients remain relatively small along the 0 to 1.0 scale. This suggests that zero awards are associated with neither short cases nor long cases, but can be found along the entire trial length axis (Figure 6). Overall, the best correlation results from using the entire data set without zero award cases. We can nevertheless observe that short cases tend to be associated

[3]The backlog of cases in Cook County, Illinois, the jurisdiction where Chicago is located and by far the state's largest county, is well know and has been reported on in publications like the [Chicago Tribune].

[4]Because not all JVS records possess both a filing date and a trial date, when we restrict our analysis to recent cases within a single state, we may limit the overall pool of available cases to $<$ O(1K).

with lower awards (Figure 7, left-hand side, awards < $200K). By contrast, longer cases can be associated with no awards, modest awards, or large awards (Figure 7, right-hand side).[5]

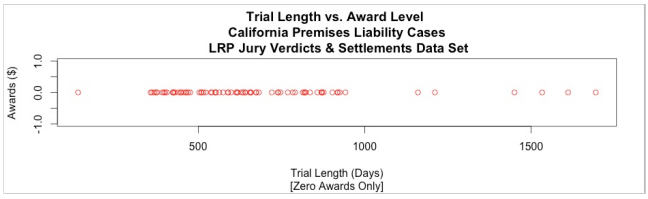

Figure 6: California JVS: Length of Trial for Zero-Awards

What this study indicates is that there is no simple or reliable relationship between trial length and award level. In order to probe this relationship futher, one would need to consider other factors or incorporate other features in order to develop a predictive model. This topic is discussed further in Section 9.

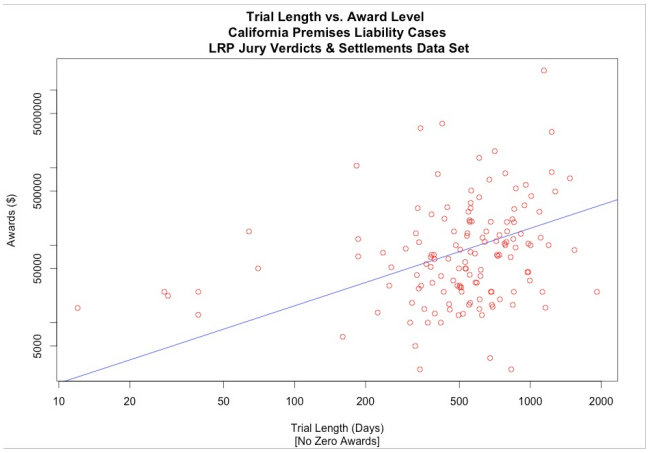

Figure 7: California JVS: Length of Trial vs. Level of Award

6 NEGLIGENCE MODELS

A key element in analyzing collections of legal cases and jury verdicts is the jurisdiction underlying the cases. One significant reason why jurisdiction plays a pivotal role in one's analysis is because of the varying 'treatments' of negligence that different jurisdictions apply. In short, different states treat negligence with assorted degrees of severity. Whereas in some states the plaintiff receives no award if found negligent in any degree, in other states the plaintiff needs to be found over 50% negligent in order for no award to be given. The three primary models include the following:

- **Pure Contributory Negligence System** – where the plaintiff receives no award if found negligent even in part.
- **Pure Comparative Fault System** – where the award to the plaintiff is reduced by percentage the plaintiff is found negligent in the case, e.g., the plaintiff was found to be 25% negligent in the accident, so the award is reduced by this amount.
- **Modified Comparative Fault System** – where the plaintiff receives no award if plaintiff negligence is usually found to be 50% or greater.
  - Slight/Gross Negligence Comparative Fault System (variant of Modified CFS) – plaintiff barred from any recovery for anything more than slight negligence (South Dakota only).

These primary models can witness small but significant variations, for example, in a few states the Modified Comparative Fault System model grants no award to the plaintiff if the plaintiff negligence is found to be 51% or greater. To measure the role that the treatment of negligence can play in award distribututions, in our research, we have examined three different sets of state jurisdictions, a pair of states with sizable population representing each of the primary negligence model treatments. Collectively, these six states cover roughly one-third of the U.S. population. The states included in this study are shown along with their negligence-type and populations in Table 9.

The different negligence models are significant because they can influence the degree to which a defendant's attorney plays what has been called "the negligence card," that is, accuses the plaintiff of being negligent, in whole or in part, in the accident that may have transpired. For example, a defendant's attorney might claim that the plaintiff was negligent in a product liability case for not having read the instructions before (mis)using a tool. Our hypothesis was that the more severe the treatment of negligence is within a state, the more likely the defendant's attorney will be to try to accuse the plaintiff of being responsible, in whole or in part, for the mishap. We have explored the role of jurisdiction and neglgence models extensively in our later investigations.

We will examine our findings based on jurisdiction in order to determine if there is any pattern that is related to how the jurisdictions treat negligence (contributory vs. comparative). For the six selected states, we have calculated the award_quotient for four distinct litigation categories. These include premises liability, product liability, pedestrian accident, and rear-end collisions. With the exception of rear-end collision (pure contributory and pure comparative) and pedestrial accident (pure comparative), we see no significant spikes in Table 9 in the award_quotients for those states with treatments of negligence where one would most expect it (pure contributory and pure comparative).

In Table 10, using the data from these same states, we differentiate between two types of properties of cases, those where the defendant accuses the plaintiff of being negligent, in whole or in part, versus those cases where the defendant does not. And again, with the exception of the rear-end collision category, we witness no noticable reduction in the award_quotient when the negligence claim is made. The suggestion here is that other factors may be playing a role in these outcomes and that one needs to examine the cases under a finer microscope, one that takes other variables into consideration. With the exception of rear-end collision, these observations are counter to our expectations. We have conducted an additional, finer-grained study into some of these variables, and have found no definitive explanations, except that the age of the plaintiff, especially in a state like Florida, may play a role in the sentiment of a jury. The details of that study are beyond the scope of this current report.

[5]Returning to our original hypotheses, in more formal terms, these findings may disprove the Null Hypothesis (no correlation) in favor of an alternative hypothesis (non-negligible correlation), but it is clearly not a strong finding.

No.	Negligence System	State	Population	Premises Liability	Product Liability	Pedestrian Accident	Rear-end Collision*	All Cases*
1	Pure Contributory	Maryland	5,976,407	0.78	0.10	0.83	3.67	2.51
2	Pure Contributory	North Carolina	9,943,964	1.86	2.33	1.33	2.64	2.82
3	Pure Comparative Fault	California	38,802,500	2.01	1.30	4.85	3.03	2.58
4	Pure Comparative Fault	Florida	19,893,297	1.10	1.36	2.97	2.45	3.15
5	Mod. Comparative Fault	Georgia (50%)	10,097,343	1.83	1.08	1.08	3.62	2.80
6	Mod. Comparative Fault	Florida (51%)	12,880,580	0.67	1.00	1.77	3.86	2.89
Cum.	—	Six States	96,594,091	1.54	1.27	3.07	3.08	

Table 9: Award Quotients per Category for Three Negligence Systems and Six States Examined

Role of Negligence	(Y/N)	Premises Liability	Product Liability	Pedestrian Accident	Rear-end Collision	All Cases
Defendent Did Not Accuse Plaintiff of Negligence	N	(12,034/8,123) = 1.48	(440/327) = 1.35	(4,502/2,059) = 2.19	(5,536/2,046) = 3.49	(23,192/14,575) = 1.57
Defendent Did Accuse Plaintiff of Negligence	Y	(12,158/4,733) = 2.57	(147/113) = 1.30	(3,668/1,331) = 2.76	(1,225/681) = 1.80	(17,393/7,512) = 2.32
Corresponds to Expectations	(Y < N)	[No]	[No]	[No]	[Yes]	[No]

Table 10: Differences in Award Quotients Relative to Defendent Negligence Claims

7 RESULTS

The work presented here provides an approach for data-driven legal strategies. By building up a set of capabilities, from automatically classifying previously unseen case summaries to clustering topically similar cases and identifying distinct language patterns used in these cases, we have shown that it would be possible to develop a legal decision support assistant around such data and techniques.

Moreover, this capability previsions legal professionals harnessing such a tool in an iterative, exploratory, and ultimately insightful manner. Before taking on a new case, an attorney could enter the plaintiff's facts into the application and then proceed to enter the anticipated claims (arguments) worth considering (Appendix, Figure 8). In response, the application would return a set of result summaries based on case outcomes that shared the same properties as the instant case (litigation type, event type, fact pattern, claim type ...) (Appendix, Figure 9). The application could also suggest additional claims available to the attorney. From these result summaries and award distributions, the attorney could also decide whether the case was a prospective litigation worth pursuing.

The experiments show that certain facts patterns are correlated with certain claims and result in bigger award ratios. But these relationships, between facts and awards, vary by topic and are not uniformly strong. This may point to the need for more data, but it is also indicative of the nuances associated with legal reasoning. It can also be indicative that the variability in the system is considerably large – both in terms of the claims resulting from a set of facts as well as the award level. These observations thus provide additional motivations for the need for this type of work.

8 CONCLUSIONS

In this work, we have explored the breadth and depth of a large jury verdicts and settlements collection in order to develop an approach and an application to assist legal professionals to examine and analyze prospective legal strategies for conducting their litigation. In the past, much of the focus of legal information providers has been on delivering accurate, comprehensive and timely information to their customers. More recently, certain enterprises have started to build applications that integrate data and experiences around specific tasks. In general, however, these efforts have come up short of helping customers make decisions – or allowing them to interact with the data, to see patterns, and to formulate and validate hypotheses.

By contrast, scenario analytics delivers the ability to examine the underlying facts in a case destined for litigation in order to determine how these facts match fact patterns and legal principles associated with prior cases and ultimately the legal strategies used by attorneys in those cases. The goal of this research has been to see how such patterns correlate with certain outcomes, for example, award levels or trial lengths. A practical application of scenario analytics permits a lawyer to explore the use of various legal strategies in order to differentiate the most favorable from the least favorable results.

Moreover, as a capability, this form of analytics will allow legal professionals to determine the most promising avenues for litigation. A resulting tool would thus save practitioners time, effort and the prospect of having to consider numerous less productive litigation paths. As researchers, our long-term objective is to make the administration of the law more effective and equitably applied. This is not only good for business, but it is also good for society. When most people cannot afford effective legal representation, this is an access to justice issue. When one sees significant variability in outcomes for cases with similar facts, this is an equitable application of the law issue. The solution is not to take away the discretion of judges; rather, it is to make them aware of the data, to ensure their decisions are as informed as they can be.

9 FUTURE WORK

There have recently been a pair of well publicized computer science journal articles reporting on experiments in modeling and subsequent predictions of high court decisions [1, 8]. There have

also been academic publications that have addressed the subject of how attorneys predict civil jury verdicts while also seeking second opinions [6]. There exist online databases to facilitate some of this research, some, claiming to have over 180,000 verdicts.[6] In addition, there have been less formal articles on how the state of trial court predictions can be improved [14]. To our knowledge, the current work is the first of its kind to formally explore the prospects of analyzing jury verdicts from a data science perspective, one that leverages a classification engine trained on O(10M) human assignments using a taxonomy consisting of 100K leaf nodes. In the following section, we describe how we have begun to pursue the next logical direction in this research: providing predictions for cases based on the facts of the case and the prospective claims that can be used.

9.1 Prediction

In the next phase of this research, we are assembling sets of representative features that will support predictive models directed at case duration, case award levels, and, ultimately, case outcomes. The features derive from both case data, for example, the unstructured text of a jury verdict summary as illustrated in Figure 1, as well as the metadata, for instance, the type of litigation, the categories of injury, the characteristics of the plaintiff, the patterns of past settlements for the state and county involved (Section 4, bullets). Given such features, we can begin to harness machine learning techniques to train models on some of our variables of interest and measure their performance, that is, their predictive capabilities, on previously unseen data sets. Our preliminary models are relying on feature sets of lower cardinality, e.g., O(10), in order to keep them initially simple.

ACKNOWLEDGMENTS

The authors would like to thank Arun Vachher for executing the Key Number Assigner operations supporting our experiments. We are also grateful to Connie Hall and Megan Putler from the New Product Development team for their grading assessments.

REFERENCES

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preeotiuc-Pietro, and Vasileios Lampos. 2016. Predicting Judicial Decisions of the European Court of Human Rights: a Natural Language Processing Perspective. *PeerJ Computer Science [Online Journal]* (24 October 2016). 2:e93.

[2] Kevin Ashley and Vern Walker. 2013. Toward Constructing Evidence-Based Legal Arguments Using Legal Decision Documents and Machine Learning. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and the Law*. ACM, New York, NY, 176–180.

[3] Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics (ACL), 69–72.

[4] Maria Jean J. Hall and John Zeleznikow. 2001. Acknowledging Insufficiency in Evaluation of Legal Knowledge-based Systems: Strategies Towards a Broad-based Evaluation Model. In *Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAL 2001) (St. Louis, Missouri)*. IAAIL, ACM Press, 146–156.

[5] Karl Harris. 2016. Using Data-Driven Insights to Set Litigation Strategy. In *Legal Text, Document and Corpus Analytics Workshop*. Center for Computation, Mathematics and Law (CCML), Univ. of San Diego School of Law, San Diego, CA.

[6] Jonas Jacobson, Jasmine Dobbs-Marsh, Varda Liberman, and Julia A. Minson. 2011. Predicting Civil Jury Verdicts: How Attorneys Use (and Misuse) a Second Opinion. *Journal of Empirical Legal Studies* 8, S1 (December 2011), 99–119.

[7] E. Jukes. 1995. Case Based Reasoning in Legal Information Retrieval. In *IEEE Colloquium on Case Based Reasoning: Prospects for Applications*. IEEE Press, 10/1–10/4.

[8] Daniel Martin Katz, Michael James Bommarito II, and Josh Blackman. 2017. A General Approach for Predicting the Behavioral of the Supreme Court of the United States. *SSRN* (17 January 2017), 18 pgs.

[9] Marco Lippi and Paolo Torroni. 2016. Argument Mining: State of the Art and Emerging Trends. In *ACM Transactions on Internet Technology*. Association for Computing Machinery, ACM, New York, NY, 1–25.

[10] Rachel Mochales-Palau and Marie-Francine Moens. 2011. Argumentation Mining. *Artificial Intelligence and Law* 19, 1 (April 2011), 1–22.

[11] Marie-Francine Moens. 2002. What Information Retrieval Can Learn from Case Based Reasoning. In *Proceedings of the 15th International Conference on Legal Knowledge and Information Systems*. IOS Press, Amsterdam, NL, 83–91.

[12] Chris Reed, Kevin Ashley, Claire Cardie, Nancy Green, Iryna Gurevych, Diane Litman, Georgios Petasis, Noam Slonim, and Vern Walker. 2016. 3rd International Workshop on Argument Mining. In *Proceedings of the Conference of the Association of Computational Linguistics*. 1–171.

[13] Sidney Siegel and N. John Castellan Jr. 1988. Chapter 9: Measures of Association and Their Tests of Significance. In *Nonparametric Statistics for the Behavior Sciences* (2nd ed.). McGraw-Hill, Boston, 235–238.

[14] Christina A. Studebaker. 2014. Can Case Outcome Predictions Be Improved? *American Psychology-Law Society Newsletter* (March 2014), 2 pgs.

[15] Vern R. Walker, Nathaniel Carie, Courtnev C. Dewitt, and Eric Lesh. 2011. A framework for the extraction and modeling of fact-finding reasoning from legal decisions: lessons from the Vaccine/Injury Project Corpus. *Artificial Intelligence and Law* 19, 4 (Nov. 2011), 291–331.

A APPENDIX

Shown in Figures 8 and 9 are the UIs for the prototype litigation decision support application.

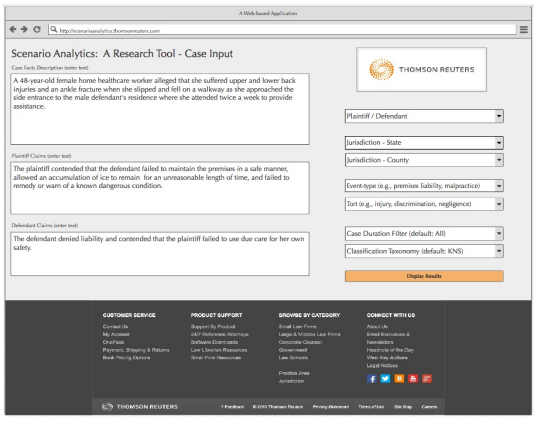

Figure 8: Prototype Fact-based Case Input GUI

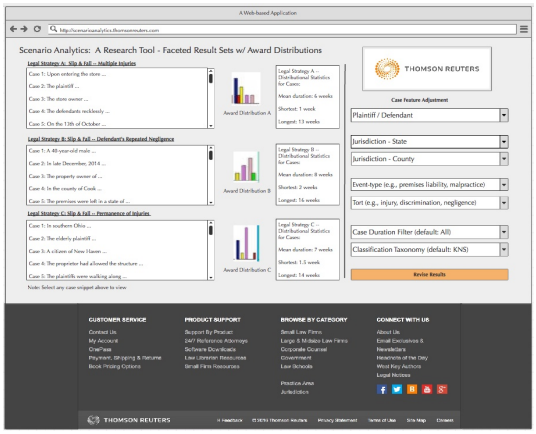

Figure 9: Prototype Application Analysis Output GUI

[6] http://www.verdictsearch.com

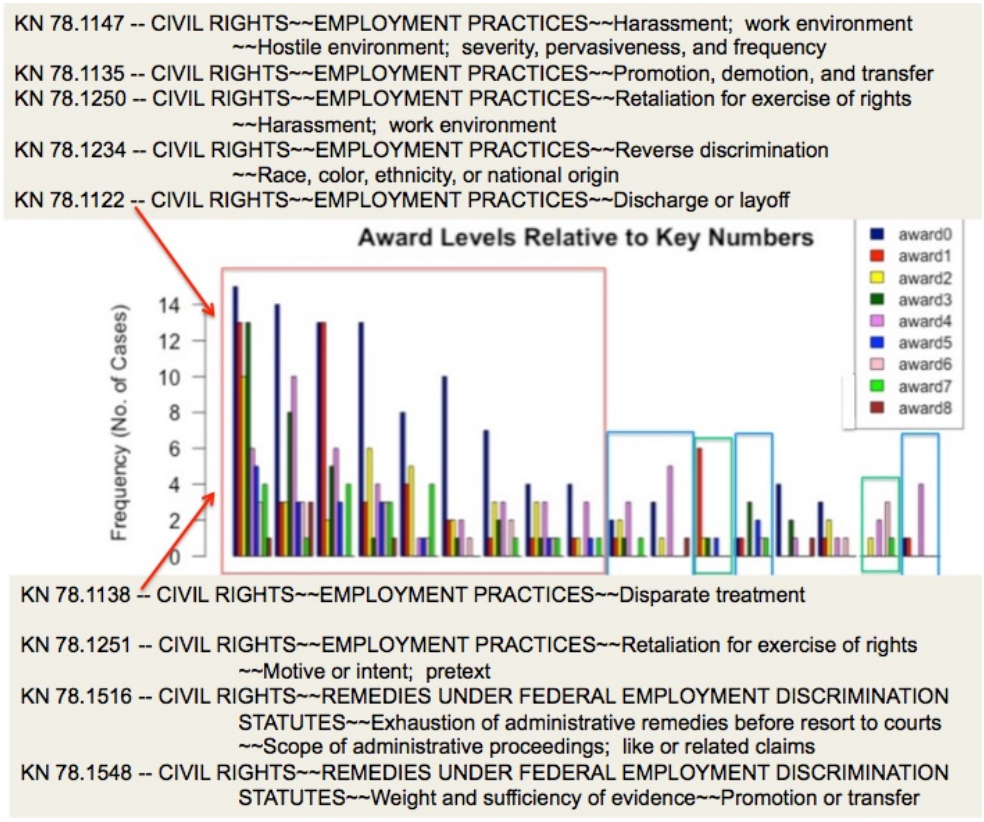

Figure 10: Award Levels Relative to Key Numbers - initial distributions

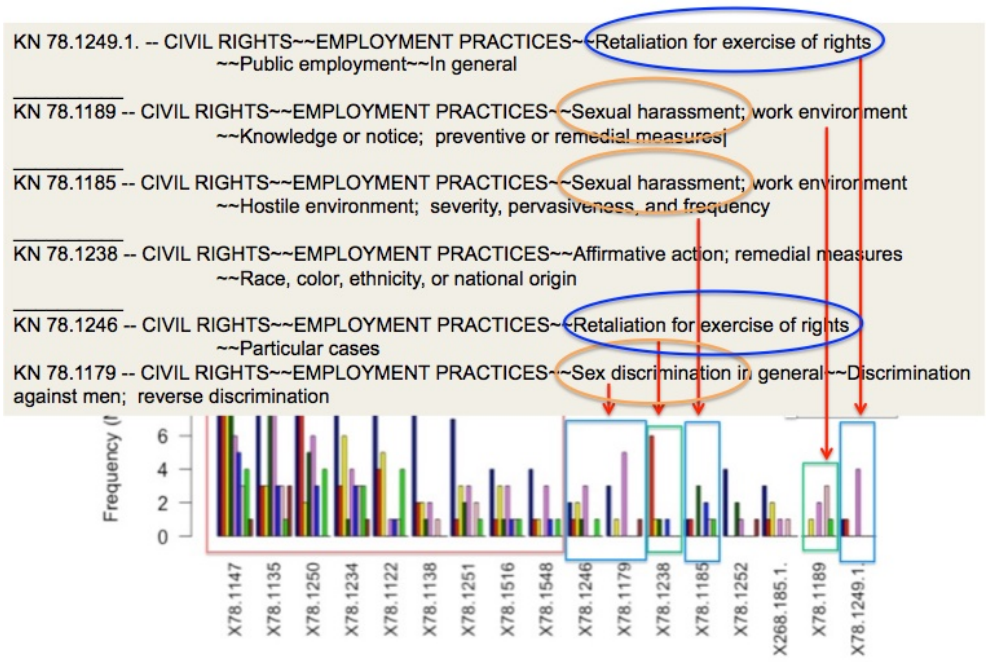

Figure 11: Award Levels Relative to Key Numbers - other distributions