Predicting Bad Patents

William Ho
Vladimir Stojanovic, Ed.

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-63
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-63.html

May 11, 2017

Predicting Bad Patents
Final Capstone Report

William Ho
University of California, Berkeley
May 2017

In collaboration with:
Joong Hwa Lee, Dany Srage, David Winer, Tzuo Shuin Yew

Faculty Advisor: Lee Fleming
Faculty Committee Member: Vladamir Stojanovic

Table of Contents

Executive Summary

Legal disputes over patent quality have increased rapidly in frequency and cost, especially in recent years. As a response, the 2012 Leahy-Smith America Invents Act established a faster, lower-cost method for the public to challenge the validity of patents before the Patent Trial and Appeals Board (PTAB). Despite this advance, patent challenges remain expensive in terms of time, money, and energy. Our capstone team aims to help reduce the frequency of such disputes by developing an automated predictor of patent quality. These predictions about the probability of post-grant invalidation before the PTAB can help inventors write original, high-quality patents and avoid legal challenges down the road.

In this report, I will discuss my technical contributions to the team's efforts of using machine learning to analyze thousands of PTAB cases for building an automated predictor of dispute outcomes. My work towards this goal include evaluating data sources, designing the workflow for acquiring and transforming data into a useful format for our machine learning algorithms, working in depth with one of our data sources, and performing exploratory analysis of high-level descriptive statistics. I will then discuss the engineering leadership aspects of our project, which include industry and legal contexts in which our project will operate, competitors and how they inform our marketing strategy, trends in machine learning that make our project possible, and potential ethics issues and our steps to mitigate them. The result of our project is a predictor that performs slightly better than the background probability of PTAB invalidations and denials, and as the PTAB dispute process becomes more well understood with more case data, we hope to improve the quality of our patent analytics to achieve our goal of reducing patent disputes.

Chapter 1: Technical Contributions

1. Project and team overview

The United States Patent and Trademark Office (USPTO) reviewed over 620,000 patent applications in 2015, a time- and labor-intensive endeavor for applicants and patent examiners alike (Patent Technology Monitoring Team, 2016). In addition, the number and cost of recent high-profile legal disputes about patent quality speak volumes about the potential impact of any good patent quality predictor, with the intellectual property licensing industry sized at $40.5 billion (Rivera, 2016, p. 7). Our project aims to provide such a predictor by using machine learning techniques to find clues that can identify bad patents ahead of time, as well as providing summary statistics to help inform inventors and patent lawyers during the application process. Our goal with this type of predictor is to help screen out invalid or weak patents before they are granted, thus saving time and money for all parties that would otherwise be involved in disputes.

My team has divided up the work of building a predictor based on our individual exposure to relevant technologies. In this chapter, I will discuss my technical contributions to the project, including evaluating new data sources, designing the pipeline for transforming acquired data into a useful format at a shared location, working with one of our data sources, and contributing to high-level descriptive statistics. My teammates' contributions are as follows: Joong Hwa Lee extracted data from one of our data sources and tried out advanced machine learning techniques designed for textual data; Dany Srage handled a different data source and computed high-level descriptive statistics on various patent features; David Winer owned the effort of setting up, tuning, and analyzing the effectiveness of various machine learning models; and Tzuo Shuin Yew set up a shared database for the whole team and designed the graphical user interface intended for end users to interact with our predictor without resorting to running

computer code themselves. The work breakdown structure below in Figure 1, created by teammate David, provides an overview of how we organized our work.

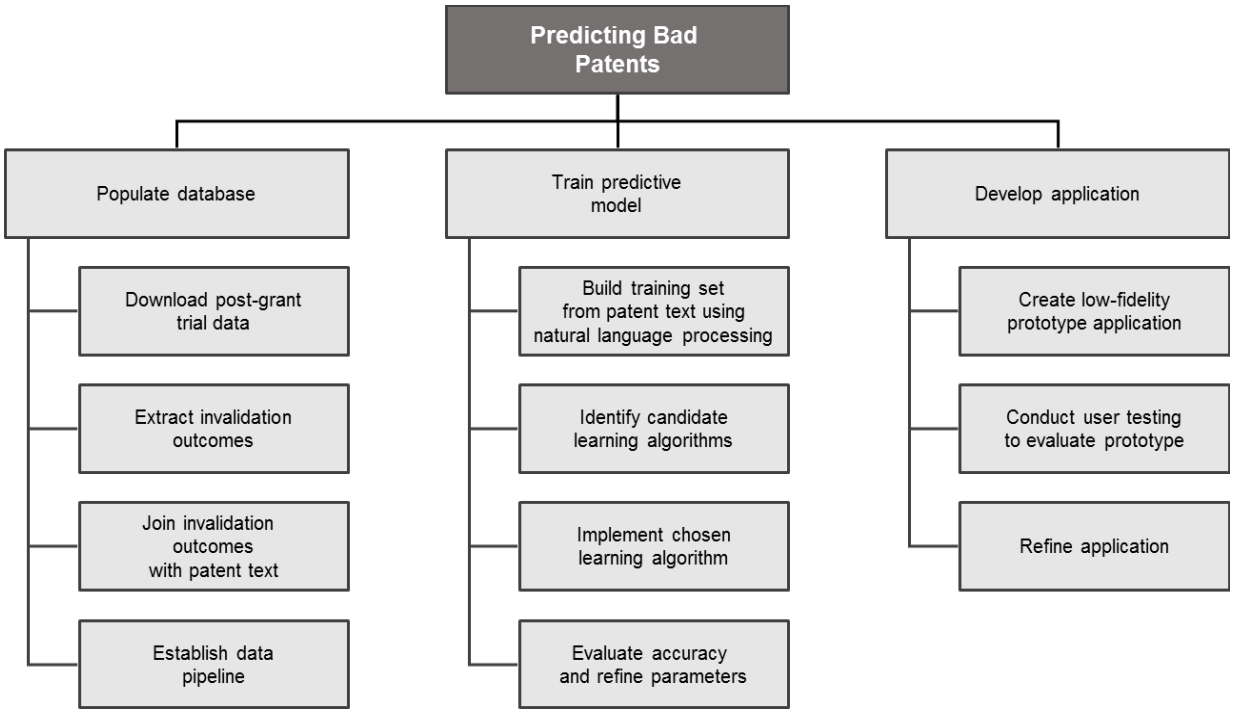

Figure 1: Project work breakdown structure

2. Overview of data sources

Since clean data is essential to any machine learning project, we created a set of criteria in mind for evaluating which external data sets to use. The criteria are as follows: the data must be reasonably clean and available for the patents and Patent Trial and Appeals Board (PTAB) cases that we want to examine; the data must be reasonably available to download; and the data set must be large enough to be useful, but small enough to store locally.

We decided to use three data sets from the United States Patent and Trademark Office, each with varying scores for the above criteria. The first one is the Electronic Freedom of Information Act (e-FOIA) data set, which has shortcomings in its consistency and availability

with respect to the cases of interest. Its largest weakness from a data acquisition perspective is its design for human users. As such, results are paginated with a capped number of results per page, and the site rate-limits downloads by serving error pages if requests are made in rapid succession. Despite these flaws, this data set still contains enough useful information to justify the effort of working with it. My teammate Joong Hwa discusses the process of downloading from this data set in his technical contributions paper.

Our second data source is the Patent Application Information Retrieval (PAIR) data set discovered by teammate Dany, who discusses his work of acquiring data from it in his paper. This data set provides a machine-friendly application programming interface (API) to facilitate automated usage of data for patents submitted to the Patent Office, including those disputed before the PTAB (United States Patent and Trademark Office, n.d.). It also contains cleaner data than the previous data set and has unique fields such as art units and patent examiners. I will discuss my work on optimizing Dany's data acquisition script later in this chapter.

The third data source we used comes from the PTAB, a division of the patent office formed by the America Invents Act in 2012 and responsible for hearing cases brought forth by the public wishing to challenge a patent's validity, among other cases (Love, 2014, Background). This data set is unique in that it was made available by the patent office via an API about one month into our project, prompting us to re-evaluate our road map to accommodate the work of acquiring data from it (United States Patent and Trademark Office, 2016). For each case brought before the PTAB, this source includes the petitioner who brought the case, the defendant, the patent at stake, and links to documents presented during the trial. While its data is not perfectly clean, such as our decision parser failing to find a decision in 4% of documents labeled as having one, it has enough data relative to noise to be a valuable addition to our project. I was responsible for downloading and extracting data from this source, as discussed later.

3. Designing the data acquisition and transformation pipeline

Even though we have identified the data sets to use in our machine learning model, we must first gather, unify, and transform the data into a standardized format to form the training data for the model. The reason for this is twofold: first, those responsible for creating the model should focus on machine learning instead of data parsing; and second, transforming text documents and other types of difficult-to-parse data into more amenable features such as categories can provide notable improvements in prediction accuracy. Taking the time to design the data acquisition and transformation pipeline, lay out its goals and restrictions, and adapt nonconforming scripts has resulted in a set of programs that each perform their job well, can be updated independently, and work together to create the foundational data set on which our machine learning model will train. Figure 2 below summarizes the process by which we acquire and preprocess data for our machine learning algorithms.

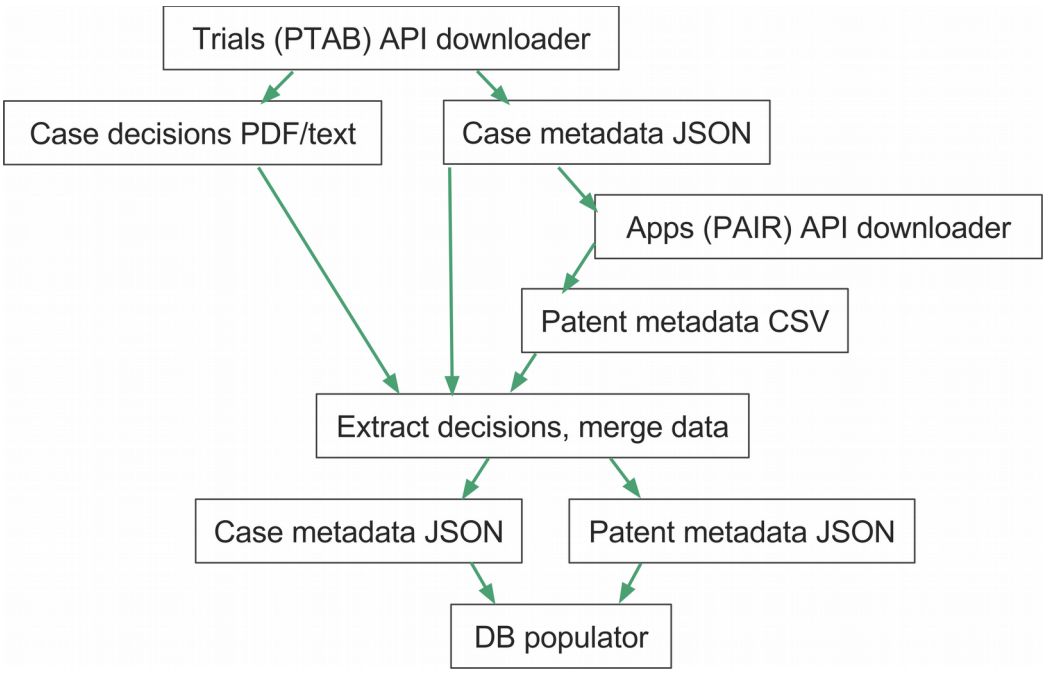

Figure 2: Our data acquisition pipeline is built for modularity, extensibility, and performance.

A well-designed pipeline for acquiring, transforming, and exporting data for later use has several features, the first of which is modularity. The pipeline should not be written as a monolithic program because such a design, with all pieces of functionality interdependent on each other, will be unnecessarily difficult to extend or maintain later. Therefore, each program should be responsible for exactly one task, and we realize this design by implementing each data downloader and parser in a separate script.

With modularity comes extensibility, or the ability to add features at a later date with minimal changes to existing code. Passing data between scripts is a prime concern here, as the choice of format can determine what changes must be performed to accommodate new functionality. We decided on JavaScript object notation (JSON) as a data exchange format because it natively supports arrays, where key-value pairs are stored and later retrieved (Crockford, 2016). Our scripts look up keys relevant to them to retrieve their associated values, while extraneous keys have no effect on correctness and require no changes. Another concern for extensibility is to define clear dependencies between scripts. In the interest of keeping the design simple, we decided to disallow cyclic dependencies, thus ensuring that each script only needs to be run once in a specified order to produce the data set for our machine learning model.

The third feature of a well-designed data pipeline is performance, which is facilitated by the acyclic graph of dependencies between our scripts. Since we want to minimize the amount of time spent on downloading data, which is much slower than extracting features locally, we would like to maximize the number of data sets from which we can download concurrently. At the same time, one data set may provide clues that let us limit the amount of data downloaded from another source. Such is the case where our downloads from the PAIR data set can be limited to the challenged patents listed in the PTAB data set. These two means are in tension with each other, but the end goal is the same, and we ultimately chose to have the PAIR downloader depend

on the output of the PTAB downloader because we can download data for less than 4000 challenged patents rather than the full set of millions of patents (Taylor, 2015). By doing so, we saved ourselves millions of network requests and possibly days or weeks of run time.

4. Challenges and solutions to acquiring data

With the data acquisition and transformation pipeline designed, the next step is to implement the scripts to fetch the actual data from the three data sets, each of which presented unique challenges. As mentioned earlier, the website hosting the e-FOIA data set was not designed for automated downloads and therefore had rate-limiting policies in place preventing full-speed acquision. My teammate Joong Hwa wrote the initial version of the downloader for this data set and ran into this rate-limiting issue. Our solution involved adding random delays of a few seconds between consecutive requests to lower our average request rate below the remote server's limit, which had to be discovered through trial and error.

In contrast, the PAIR data set did not present rate-limiting problems, but the data set's size posed a problem since millions of patents exist, yet only a few thousand have been challenged before the PTAB. As discussed previously, our solution to reduce the number of patents was to read the patent numbers of the disputed patents from the PTAB downloader's output, remove any duplicate values that come from an individual patent being challenged multiple times, and then fetch data for only these patents.

The PTAB data set, like the PAIR data set above, can be reached through a machine-friendly API, but unlike the PAIR data set, some of the data of interest is stored in human-readable decision documents stored as portable document format (PDF) files. The presence of many such large files means that transient errors can significantly affect our ability to download

the data we needed. Computers and networks do not always perform flawlessly, and transient failures crop up for various reasons outside of our control. Therefore, it was deemed necessary to have the ability to continue downloading after an interruption. This was implemented by first downloading all the metadata, which is much smaller and can be downloaded in batches of 100 cases at a time to minimize the probability and cost of interruptions. Then, for each case, the associated case decision file is downloaded only if a local copy does not already exist. Downloaded files are named according to their PTAB case number to facilitate this checking. This same resuming feature was implemented in the e-FOIA downloader and is arguably even more important there because rate-limiting restrictions on the remote server, coupled with the sheer number of documents, make retrying downloads even more expensive. With these optimizations, we successfully acquired data from all three sources with minimal waste of work, and can re-run these downloaders to acquire new data efficiently as it becomes available.

5. Transforming the data into useful formats

For our data to be useful, they need to be transformed into formats that machine learning models generally expect, such as continuous data like those from measurements, or categorical data like names. David and Joong Hwa worked on transforming full-length patent application texts into machine learning-friendly vectors and attempting to predict invalidation and denial outcomes based on these vectors. Meanwhile, our data sets provide other information in human-readable formats that need to be transformed in other ways.

To extract case decisions encoded within the PDF files provided by the PTAB and e-FOIA data sets, a Unix program called `pdftotext` was used to extract the plain text from the files (Pdftotext, n.d.). Regular expressions, a language for pattern matching, were then used on

the extracted text to identify key phrases that indicate whether the case was denied a hearing or whether one or more claims of the challenged patent was found to be invalid. Regular expressions allow searching for patterns of text while accounting for variations such as upper- or lower-case words, misplaced punctuation, and extraneous spacing added between words during the PDF conversion process. The expressions used in our project are shown below in List 1.

```
1. ORDERED.{0,240}?(unpatentable|UNPATENTABLE|anticipated|ANTICIPATED|cancell?ed|
   CANCELL?ED|obvious|OBVIOUS)
2. ORDERED.{0,120}?[aA]dverse\s+[jJ]udgment.{0,120}?(granted|GRANTED)
3. [jJ]udgment\s+(is\s+)?entered
```

List 1: Our regular expressions for extracting decisions from case documents maintain a balance between expressiveness and simplicity.

A balance between regular expression complexity and understandability had to be made, since more complex patterns can deal with more formatting cases and thus extract more data, but are less readable and maintainable by humans. We compromised on the PTAB data set by using a set of three regular expressions, shown above, to cover a representative set of cases while keeping each expression simple. This compromise is made possible by only parsing documents specified by the PTAB API to contain decisions, compared to earlier efforts at parsing documents from the e-FOIA data set with no prior knowledge of their contents. This latter case required the use of many complex regular expressions and resulted in high misclassification rates. The result of this parsing is a pair of categorical, yes-or-no answers to case denial and patent invalidation which can be easily used in a machine-learning algorithm.

6. Exporting the data for concurrent use

After all the data from our three data sets has been downloaded and transformed, the last step in the data pipeline is to export the transformed data into a shared location so that all team members can work with a single copy of the data. We decided on storing our data in a MySQL relational database since MySQL natively supports concurrent access to data, allows for a rich set of queries that machine learning algorithms can use to fetch data for further processing, and runs on Linux, the operating system on the shared machines available to us. My teammate Tzuo Shuin set up the database and defined the table schemas, which are a set of rules describing the types of each field of data to be stored.

The database populator script, which reads data transformation script's output, was rewritten in a modular fashion to fit the goals of maintainability and extensibility. The initial version of the populator implemented nearly everything in one big function, making changes and extensions difficult and error-prone. That monolithic function was refactored into individual functions and objects that abstract away the details of connecting to the database and ensuring all pending data is safely stored. This change reduced the number of lines of code by 10%, even as two database tables had to be updated instead of just one from the addition of the PTAB database after the start of our project. Adding a third table is straightforward should we need to do so in the future. Furthermore, the script was redesigned during the refactoring operation such that only one line needs to be changed to export to a different database on a different machine. This greatly facilitates the creation of redundant databases in case one machine becomes unavailable and allows team members to stay productive even in the face of database technical problems.

7. Time interval statistics for patents; early prediction results

The first step to building a machine learning model after the data is in place is to create summary statistics to understand the data at a high level, since this understanding can help inform the choice of model and provide context for its performance. My teammate David set up various machine learning models and evaluated their effectiveness, including true positive and false positive rates, to analyze how different ways of vectorizing text and using features affect classification accuracy. To support this effort, three date-based statistics were computed for patents challenged before the PTAB: application-to-grant time, grant-to-challenge time, and challenge-to-decision time. These three intervals were chosen for their ease of computation from the data available to us. In addition, they are easily understood both in terms of what they are and how they could explain prediction results should they prove useful. For the purpose of this analysis, denied cases were counted as not invalidating the patent in question. The means and standard deviations of these dates are shown below in Table 1. In addition, histograms of these dates were visually inspected for any immediate patterns.

	Application-to-grant	Grant-to-challenge	Challenge-to-decision
Denied	1173.3 (793.2)	2413.5 (1893.4)	160.3 (41.6)
Not denied	1088.5 (704.4)	2411.9 (1821.7)	177.6 (27.7)
Invalidated	1073.4 (703.3)	2452.2 (1893.4)	177.3 (28.2)
Not invalidated	1173.9 (786.3)	2387.9 (1878.4)	162.0 (40.8)

Table 1: Means and standard deviations of various time intervals for patents in the PTAB database show no significant separation between classes. All numbers are in days.

There exists a strong similarity between the statistics for patents whose challenges were denied and those that were not invalidated by the PTAB because denied cases cannot possibly end up invalidating a patent. In addition, the vast majority of patents whose cases were accepted ended up being invalidated by the PTAB. This correlation is shown below in the number of cases fitting each pair of outcomes in Table 2.

	Invalidated	Not invalidated	Total
Denied	0	2052	2052
Not denied	1406	187	1593
Total	1406	2239	3645

Table 2: There exists a perfect correlation between case denial and lack of invalidation, and most accepted cases invalidate one or more claims of the disputed patent.

In addition to the above statistics, scatter plots of pairs of intervals, shown below in Figures 3a, 3b, and 3c, were generated in an attempt to find correlations that could help predict denial of challenges or invalidation of patent claims. Because of the above correlations between denial and invalidation, the scatter plots for the invalidation classes are very similar and thus are omitted.

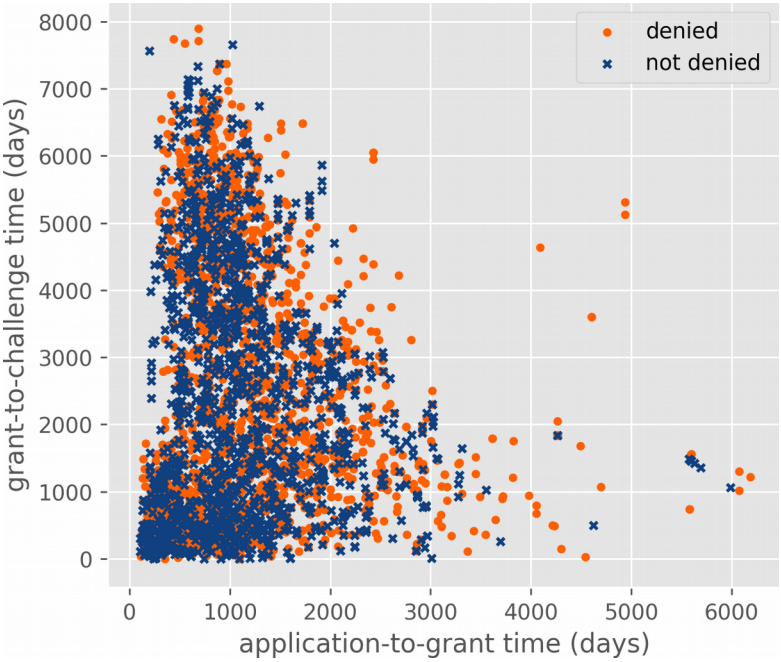

Figure 3a: There is no significant visual separation for patents whose challenges were denied vs. accepted when plotting application-to-grant time against grant-to-challenge time.

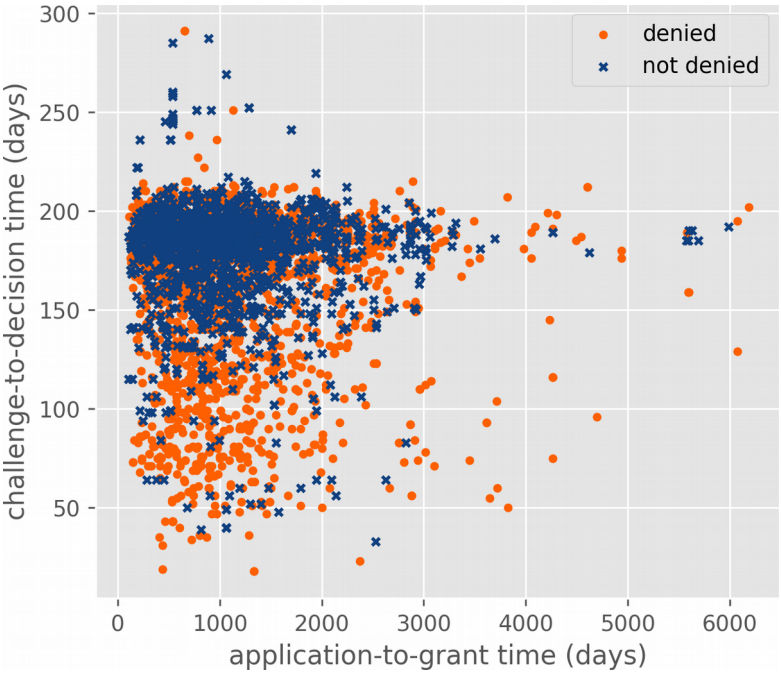

Figure 3b: Patents whose cases were denied appear to have both shorter application-to-grant and challenge-to-decision times than patents whose cases were accepted by the PTAB.

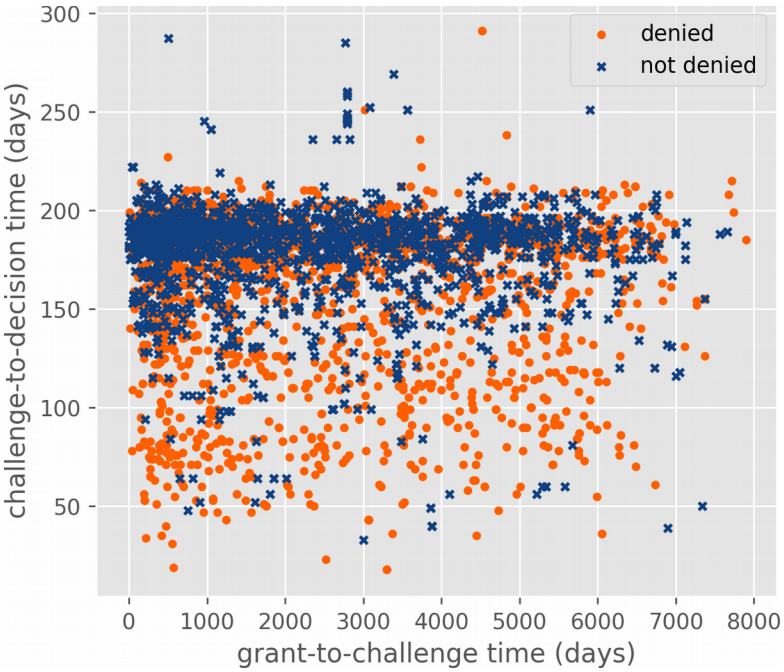

Figure 3c: There is no significant visual separation for patents whose challenges were denied vs. accepted when plotting grant-to-challenge time against challenge-to-decision time.

Unfortunately, only the application-to-grant vs. challenge-to-decision scatter plot appeared to have a significant visual difference, but since it is infeasible to plot all combinations of features that go into our machine learning model, there is always a possibility of the model uncovering patterns that the human eye has missed. Despite the mostly negative results, this type of exploratory data analysis is expected in any machine learning project before running any machine learning model.

Nevertheless, these time intervals, along with other features from our data sets, proved sufficient to give better-than-chance accuracy. As my teammate David shows, the best results come from our support linear classification (SVC) model, which trades expressiveness for simplicity by attempting to find a high-dimensional plane that separates classes of data (Boser, Guyon, and Vapnik, 1992). This model obtains an accuracy of 72% on test data on which it has

not been trained, a result 11 percentage points higher than the baseline rate of 61%. These results encourage further exploration with more features and more powerful models such as neural networks.

8. Conclusion and future work

We have seen some encouraging early results with respect to our goal of building a classifier that can predict patent invalidation at the Patent Trial and Appeals Board. These results are made possible by constructing a solid data acquisition and preparation process by which training data is made available for the machine learning model. We have successfully defined and implemented programs for gathering data from three data sources from the United States Patent and Trademark Office, and the careful design of our data pipeline ensures that other data sets can be integrated easily. There remains much room for exploration and research, both with different data sets and more expressive machine learning models, such as neural networks for classifying large collections of text documents (Lai, Xu, Liu, and Zhao, 2015). Revisiting the problem of predicting patent invalidation at a later date may also prove helpful as more cases proceed through the PTAB and provide a more detailed picture of how challenges are denied and how patents are invalidated.

Chapter 2: Engineering Leadership

1. Introduction and overview of technology

In this section, my teammates and I examine the industry context and business considerations associated with building our post-grant review prediction algorithms. First, we will discuss the current patent landscape and how it informs the marketing strategy for potential customers. Second, we will analyze the different competitors in the legal services space and define how our tool differs from existing offerings. Third, we will discuss the current state and trends in the machine learning field today and how they can be applied to our tool. Fourth and finally, we will close with an ethics section that will examine the ethical issues we considered in designing and deploying the algorithm in the form of a website.

2. Marketing strategy in an expanding patent analytics market

It is becoming increasingly challenging for research-oriented firms and their attorneys to navigate the intellectual property landscape in the United States. In addition to the increase in the sheer number of patents, recent changes in US law have made it significantly easier for members of the public to challenge existing patents. In 2012, the US federal government enacted the Leahy-Smith America Invents Act (AIA). This legislation substantially expanded the role of the US Patent and Trademark Office in post-grant patent reexamination (Love, 2014, Background). The AIA opened the gates of post-grant patent opposition to members of the public by providing a much less costly and more streamlined avenue for post-grant opposition through the Patent Office’s Patent Trial and Appeals Board (PTAB). Any member of the public could challenge an existing patent for only a few thousand dollars—relatively inexpensive compared to litigation (Marco, 2016).

Accordingly, the patent application process is under two types of strain: it is resource constrained—since there are more and more patents being filed every year—and it is coming under more scrutiny due to the America Invents Act. There are two main sets of stakeholders that have an interest in improving the current application process: (1) the USPTO and (2) patents filers and their lawyers.

First, because "IP-intensive industries accounted for about […] 34.8 percent of U.S. gross domestic product […] in 2010," reducing the time it takes to effectively examine a patent—perhaps through assistance from a computerized algorithm—is a critical priority for the USPTO (Economics and Statistics Administration and United States Patent and Trademark Office, March 2012, p. vii). Indeed, helping patent examiners reduce the time they spend on each patent while still maintaining the quality of examinations would mean reducing the cost and time associated with filing patents, proving economically accretive and reflecting well on the US Patent and Trademark Office. In fact, the USPTO has expressed interest in a predictive service in the past and has conducted its own research into predicting invalidation (United States Patent and Trademark Office, 2015, p. 38).

Secondly, when applying for patents, patent filers and their attorneys have a strong interest in preempting potential litigation through effective framing and wording of their patents. Patent litigation is becoming more and more common as evidenced by an IBISWorld report: "Demand for litigation services has boomed over the past five years" (Carter, 2015, p. 4). Therefore, a tool that could help patent filers prevent litigation would be valuable during the application process. One industry that may be especially interested in this sort of tool is Business Analytics and Enterprise Software. In the past several years, the costs associated with protecting "a portfolio of patents" have disproportionately increased in this industry (Blau, 2016, p. 22).

3. Competition in the patent analytics space

Patent validity is a major concern for the $40.5 billion intellectual property licensing industry, whose players often must decide whether to license a patent or challenge its validity (Rivera, 2016, p. 7). These decisions are currently made through manual analysis conducted by highly-paid lawyers (Morea, 2016, p. 7). Because of the cost of these searches, data analytics firms such as Juristat, Lex Machina, and Innography have created services to help lawyers perform analyses more effectively.

One common service is semantic search for prior art and similar patents, where queries take the form of natural language instead of mere keywords. Other services include statistics about law firms, judges, and the patent-granting process. These services build their own high-quality databases by crawling court records and other public data sources, correcting or removing incomplete records, and adding custom attributes to enable such search patterns and reports. Their prevalence reflects the trend towards data analysis as a service, since law firms are not in the data-analysis business (Diment, 2016).

The above services lie outside the scope of predicting invalidation from patent text and metadata but become relevant when discussing commercialization because high-quality data improves model accuracy and enables techniques like concept analysis that are difficult or impossible with raw unlabeled datasets. As such, partnering with existing firms that provide clean datasets or otherwise cross-licensing our technologies may be advantageous.

While these services help lawyers make manual decisions with historical statistics, we have found no service that attempts to predict invalidation for individual patents. Juristat is the only major service we found that performs predictions on user-provided patent applications. Specifically, Juristat predicts how the patent office will classify a given application and

highlights key words contributing to that classification, with the goal of helping inventors avoid technology centers in the patent office with high rejection rates (Juristat, n.d.).

Our project, if successful, can become a Juristat-like service for predicting post-grant invalidation. While we cannot speculate on existing firms' development efforts, the lack of similar services on the market suggests a business opportunity. Whereas existing services target law firms and in-house IP lawyers, our project aims to help the USPTO evaluate post-grant petitions, which are brought forth by parties attempting to invalidate a patent.

4. Machine learning technology trends

This work has been enabled by many recent advances in the application of machine learning to large data problems. Even though machine learning has been around for several decades, it took off within the past decade as a popular way of handling computer vision, speech recognition, robotic control, and other applications. By mining large datasets using machine learning, one can "improve services and productivity," for example by using historical traffic data in the design of congestion control systems, or by using historical medical data to predict the future of one's health (Jordan & Mitchell, 2015 p. 255-256). For this project, we had access to a large dataset of historical patent filings since 1976, for which recently developed machine learning techniques proved especially useful.

Machine learning algorithms generally fall into one of two categories: supervised and unsupervised (Jordan & Mitchell, 2015 p. 256). Supervised learning algorithms need to be run on training data sets where the correct output is already known. Once the algorithm is able to generate the correct output, it can then be used for regression or clustering. In contrast, unsupervised learning algorithms use data sets without any advance knowledge of the output,

and perform clustering to try to find relationships between variables which a human eye might not notice. Recent trends indicate that supervised learning algorithms are far more widely used (Jordan & Mitchell, 2015 p. 257-260). Our historical data set indicated whether or not patents were invalidated/denied during past disputes, which made a supervised learning algorithm the appropriate choice.

5. Ethical considerations

As with all engineering projects, we anticipated the possibility of running into potential ethical conflicts. We used the Code of Ethics, written by the NSPE (National Society of Professional Engineers), as a guideline for our planning (NSPE, 2017). We identified two components of the Code of Ethics, which our project could potentially violate if left unchecked.

The first is Rule of Practice 2: “Engineers shall perform services only in the areas of their competence” (NSPE, 2017). One of our potential target customers is the United States Patent and Trademark Office, who would ideally use our project to aid with their patent approval decisions. If our project were seen to be an automated replacement, rather than a complement, for trained patent examiners or attorneys, that may be considered an attempt to perform services outside of our “areas of competence.” While we cannot dictate how our customers ultimately utilize our product, we can mitigate the issue through thorough written recommendations in our documentation to hopefully encourage responsible usage.

The second ethical consideration is Professional Obligation 3: “Engineers shall avoid all conduct or practice that deceives the public” (NSPE, 2017). While we fully intend our project to be used in service to the public, we recognize the possibility of bias in our supervised machine learning algorithm (Reese, 2016), with the resulting output capable of unfairly swinging the

outcome of a patent decision. Unlike the prior ethical issue, we have more control in this situation, since we do not have a viable product without a sufficiently trained algorithm. By verifying our datasets to ensure equal representation and objective input, we can avoid inserting biases and thus maintain ethical integrity.

6. Conclusion

Collectively, recent economic and regulatory trends have made now an exciting but uncertain time for inventors, attorneys, and the US Patent and Trademark Office. Thoughtful applications of machine learning and statistics can make sense of these recent changes and assist stakeholders in truly understanding what drives patent invalidation. As we pursue this technology, our understanding of the industry landscape of potential customers/competitors, leading trends in machine learning research, and the ethical considerations associated with our technology will drive our research. Ultimately, we hope that our technology contributes to the continued development of a patent ecosystem that enables inventors to do what they do best: developing novel and socially valuable inventions.

References

Blau, G. (2016). *IBISWorld Industry Report 51121c Business Analytics & Enterprise Software Publishing in the US*. IBISWorld. Retrieved October 16, 2016

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (COLT '92). ACM, New York, NY, USA, 144-152. doi:10.1145/130385.130401

Carter, B. (2015). *IBISWorld Industry Report OD4809 Trademark & Patent Lawyers & Attorneys in the US*. IBISWorld. Retrieved October 15, 2016

Crockford, D. (2006, July). The application/json Media Type for JavaScript Object Notation (JSON). Retrieved March 03, 2017, from https://www.ietf.org/rfc/rfc4627.txt

Diment, Dmitry (2016, March). *IBISWorld Industry Report 51821. Data Processing & Hosting Services in the US*. IBISWorld. Retrieved October 11, 2016

Economics and Statistics Administration and United States Patent and Trademark Office. (March 2012). I*ntellectual Property And The U.S. Economy: Industries in Focus.* U.S. Department of Commerce. Retrieved October 15, 2016

Jordan, M. I., & Mitchell, T. M. (2015, 07). Machine learning: Trends, perspectives, and prospects. *Science, 349* (6245), 255-260. doi:10.1126/science.aaa8415

Juristat - Patent Analytics. (n.d.). Retrieved October 13, 2016, from https://juristat.com/#etro-1

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. *Association for the Advancement of Artificial Intelligence*. Retrieved March 7, 2017 from arXiv.

Love, B.J., & Ambwani, S. (2014). Inter partes review: an early look at the numbers. *University of Chicago Law Review, 81(93)*. Retrieved October 14, 2016 from: https://lawreview.uchicago.edu/page/inter-partes-review-early-look-numbers

Marco, A. (2016, October 5). Phone interview with USPTO Chief Economist.

Morea, S. (2016). *IBISWorld Industry Report 54111 Law Firms in the US.* IBISWorld. Retrieved October 16, 2016

NSPE. (n.d.). Code of Ethics. Retrieved February 03, 2017, from https://www.nspe.org/resources/ethics/code-ethics

Patent Technology Monitoring Team. (2016, October 16). *U.S. Patent Statistics Chart*. Retrieved from USPTO: https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm

Pdftotext(1) - Linux man page. (n.d.). Retrieved March 18, 2017, from https://linux.die.net/man/1/pdftotext

Reese, H. (2016, November 18). Bias in machine learning, and how to stop it. Retrieved February 04, 2017, from http://www.techrepublic.com/article/bias-in-machine-learning-and-how-to-stop-it/

Rivera, Edward (2016, April). *IBISWorld Industry Report 53311. Intellectual Property Licensing in the US*. IBISWorld. Retrieved October 10, 2016

Taylor, K. (2015, January 16). Patent Trial and Appeal Board - Multi-Petition Challenges of a Patent. Retrieved March 5, 2017, from https://www.knobbe.com/sites/default/files/PTAB%20Multi%20Petition%20Challenges%20of%20a

United States Patent and Trademark Office. (n.d.). PAIR Bulk Data. Retrieved from https://pairbulkdata.uspto.gov/

United States Patent and Trademark Office (2015, January). Patent Litigation and USPTO Trials: Implications for Patent Examination Quality. Retrieved October 9, 2016 from https://www.uspto.gov/sites/default/files/documents/Patent%20litigation%20and%20USPTO%20trials%2020150130.pdf

United States Patent and Trademark Office. (2016). *Patent and Trial Appeal Board (PTAB) Application Programing Interface (API) User Manual*. Retrieved February 21, 2017, from https://developer.uspto.gov/sites/default/files/dh-ptab_ug.pdf