# Confidential Informants Project

CS506 Spring 2020

# Outline Of Report

I. Project overview
II. Data Collection
III. Data Cleaning
IV. Data Analysis
V. Caveats About the Interpretability of the Data
VI. Appendix: structure of the data

# I. Project Overview

The goal of the project was to analyze all criminal cases within the state of Massachusetts in order to find instances and patterns of when the state had used informants improperly in criminal cases.

Informants are witnesses who may at times testify anonymously. For the purposes of our project, the term "informant" is synonymous with "confidential informant". Confidential informants are sometimes criminals who are freed from jail or offered lighter sentences in exchange for their collaboration with law enforcement. The state of Massachusetts usually argues that confidential informants must have their identity protected in order to keep them safe from reprisals, but this obviously leads to potential for abuse. A motivation for this project was several high profile cases uncovered by WGBH and other news outlets in which the state had abused the confidential informant system to bribe inmates into lying to obtain false convictions in cases where there was little other evidence.

This project was requested by Paul Singer, head of investigative reporting from WGBH.

# II. Data Collection

## A. *Data We Started With*

Per direction from the client, we began our analysis with data from project teams that worked on a similar project in 2018 and 2019. This data, contained in the .json files cases.json and mass_appeals.json, represented the decisions of all *published* cases from the Massachusetts Supreme Judicial Court (SJC) and Massachusetts State Appeals Court respectively in the 2008-2018 period.

For an example of the type of dense legal language used in the decisions, see the following, which is the beginning of a case:

*"'The present case is the most recent in a series of cases concerning the egregious misconduct of Annie Dookhan, a chemist who was employed in the forensic drug laboratory of the William A. Hinton State Laboratory Institute (Hinton drug lab) from 2003 until 2012. On January 23, 2007, the defendant, Admilson Resende, pleaded guilty on indictments charging distribution of a class B controlled substance (cocaine), G. L. c. 94C, § 32A (c) (five counts); violation of the controlled substances laws in proximity to a school or park, G. L. c. 94C, § 32J (three counts); and possession of a class B controlled substance (cocaine) with intent to distribute, G. L. c. 94C, § 32A (c) (one count).He completedPage 3service of his sentences.On October 2, 2012, the defendant filed in the Superior Court a motion to withdraw his guilty pleas pursuant to Mass. R. Crim. P. 30, as appearing in 435 Mass. 1501 (2001), based on Dookhan\'s malfeasance.Prior to the issuance of a ruling on the defendant\'s motion, this court decided Commonwealth v. Scott,(2014), in which we articulated, in reliance on Ferrara v. United States, 456 F.3d 278, 290-297 (1st Cir. 2006), a two-prong framework for analyzing a defendant\'s motion to withdraw a guilty plea under rule 30 (b) in a case involving the misconduct of Dookhan at the Hinton drug lab. Scott, supra at 346-358. Under the first prong of the analysis, a defendant must show egregious misconduct by the government that preceded the entry of the defendant\'s guilty plea and that occurred in the defendant\'s case."*

## B. *How we Scraped*

Within the initial data, it became clear that there were only about 96 cases that contained informants (see the analysis section). Thus, on the direction of Mr. Singer, we decided to scrape about 10 more years of cases, from 2000-2008. We ran into problems when we realized that the site that all three previous teams had used to scrape court decisions was no longer operational. We eventually found the somewhat equivalent site *masscases.com*--data there is not labelled as thoroughly as data from the site that was previously used, and in addition to having to completely rewrite the code used for scraping because of the different site, the lack of tagging made scraping more difficult. We decided to use

selenium webdriver to accomplish the scraping, and once the code was written, it took only a few hours to execute.

# III. Data Cleaning

## A. *How we Combined the Data*

The new data was scraped to have a similar format to the old data, so merging the two data sets would be easier. We settled on pandas DataFrame for the final structure, as the efficiency and implementation of it compared to python lists is vastly superior. With this change we had to reformat the inner fields of the dictionaries to strings instead of lists. This makes it easier for pandas to search for and work with the data. This change led to a necessity for new cleaning functions and new search functions that implemented pandas and worked with our new data structure. Now, a search for cases that contain informants, or cases that satisfy other criteria, is easy returns the full cases in a new pandas DataFrame, which is in turn searchable and has all the usability of pandas. This is a significant improvement in terms of searching the intersections and finding more complex feature groups in the data.

## B. *How we Cleaned the Data*

### 1. *Dissents / footnotes*

We noticed that the decision of many cases also includes a dissent, or contradictory, implemented opinion of the minority of judges, and the dissent or a footnote in our cases can alter the results of our analysis. This is due to the fact that they may contain keywords that are not relevant to the outcome of the case. While scraping the data we were careful with the footnotes and tried not to include any, regrettably, we think we still have some left. This is due to the fact that we searched for the keyword “Footnote” and found 500 examples of it in our data. However, distinguishing these footnotes from citations is difficult.

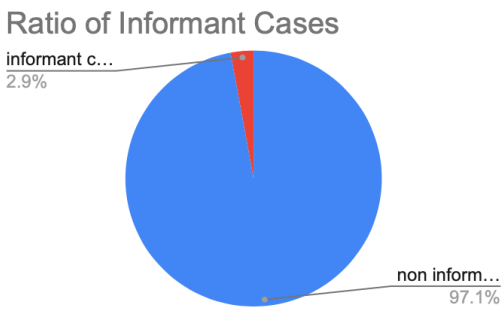
### 2. *The Issue of Punctuation*

We found that the data was formatted in such a way that there were no spaces between words and punctuation so when the data was put into word2Vec the punctuation would throw off the algorithm. For example, “informant” and “infromant.The” were considered two different words. We

fixed this issue by editing all of the text to have spaces on each side of any punctuation or special characters.

## IV. Data Analysis

### A. *How we located cases that contain informants*



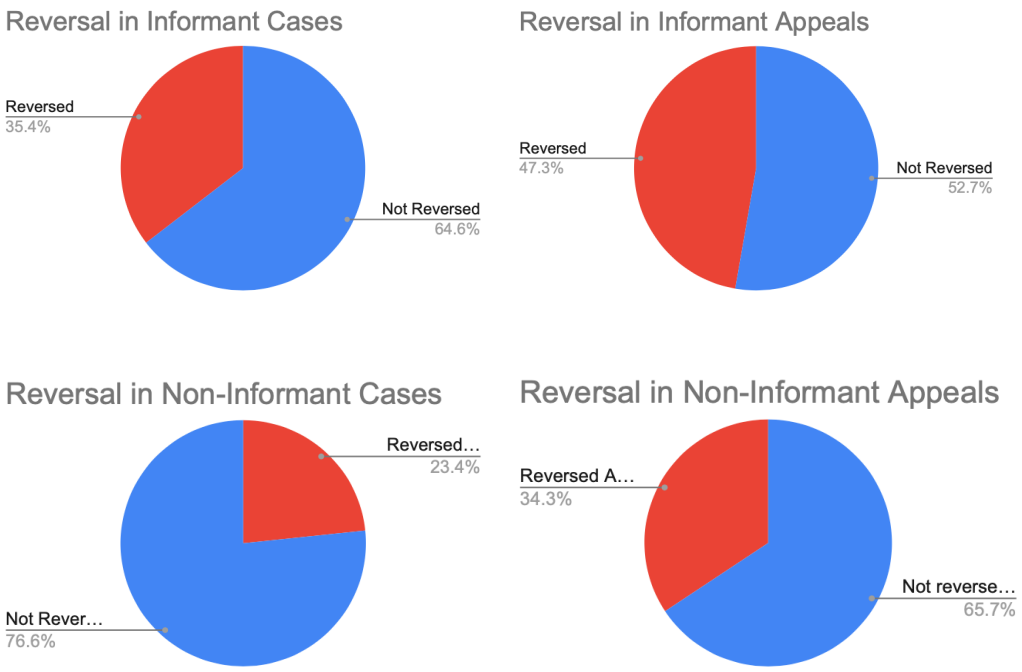

Initially we were looking at cases that had informants by simply searching through the data for the keyword “informant”, we realized we needed to expand our search and found a list of keywords that were related to informants., including "confidential informant", " ci ", "snitch", "informant", and "gang informant".

This expanded our search, but the amount was still low. Our search returned 79 cases in our SJC cases, and 146 cases in our appeals court cases. The graph shows how these cases make up only 3% of our dataset.

A step that we intend on taking is using our word2vec algorithm and using it to find words that are similar to “informant” and “snitch” in order to broaden our search with more related terms.

### B. *How much more often are informant cases overturned?*

During our analysis we focused on the verdict of cases in order to give us insights on how informant cases play out in court. We focused specifically on the “reversed” keyword as we found that the cases mentioned it to produce their verdict. Here are our results.




The top two charts are reversals in informant cases, while the bottom two focus on the cases that had no connection to informants (note that cases here is shorthand for SJC cases and appeals here is shorthand for Statewide Appeals Court appeals). We see that the relative reversal of informant cases is higher (35% and 47%) than the cases that had no informant (23% and 34%). This suggests that there is a higher likelihood of a case being overturned if an informant is involved. We also notice that in both categories, the appeals court has a higher proportion of overturned cases. There are multiple possible explanations to why cases with informants are overturned more often, up to this point we don't have the answer, but we have narrowed the search from 7000 cases, to 100 that are overturned and have informants, to search for newsworthy material for Mr. Singer.

C. *The state of our overturned analysis*

We started by searching through cases to understand the language that they used in order to determine if the case is to be overturned or not. This preliminary analysis showed that the verdict was not in the header, but in the text. It came as "reversed", "reversal", "we reverse" or "reversing". With this in mind we searched the dataset and found a large number of cases. With our move to pandas, we were able to create a DataFrame that has all instances of cases with these keywords. There are two ways we plan on going, one is expanding our search keywords with word2vec in order to refine our search and eliminate errors, and the second is a more sophisticated method that we shall discuss in part VI.

### D. *The state of our word2vec analysis*

After finishing the scraping of the new data, we looked at different ways we could vectorize the data. After doing research on other projects similar to our own online, emailing with Gowtham, and speaking with professor Galletti, we chose that Word2Vec was the best way to vectorize our data. We took in the 4 JSONs of cases and created a tokenization of all the text from all the cases, keeping them separated between SJC and Statewide Appeals Court cases. We ran into a slight issue regarding punctuation once we put this data into Word2Vec. The fashion in which we fixed this is described above. We fixed this issue by editing all of the text to have spaces on each side of any punctuation or special characters. Once this cleaning was finished, we found that word2Vec provided a great system to discover more about our data. We have discovered multiple words that are used in a similar fashion as “affirm’ and “reverse” to now use in future scraping schemes. Going forward, we are going to use this word2Vec embedding to do more advanced analysis on our data.

## V. Caveats About the Interpretability of the Data

### A. *Note on Massachusetts court system*

The Massachusetts Court system is made up of several levels, with the SJC as the highest court of final appeal and the Statewide Appeals Courts as the court system immediately below that in importance and in line of appeal. While in some cases the SJC may exercise what is known as its "right of superintendency" and chose to hear cases from much lower courts and skip over the Statewide Appeals Courts, most of the cases in the SJC are further appeals from the Statewide Appeals Courts and are thus duplicates in some sense. We have set aside the problem of figuring out which cases are duplicates in this sense because there is not sufficient metadata in the pages scraped to do this unambiguously without a complex algorithm that would involve cross referencing data and the names of the parties with an additional website containing records of what cases were tried on which date at which courthouse.

For this reason, we generally analyzed cases from the SJC and from the Statewide Appeals Courts separately.

### B. *Caveats About Unpublished Cases*

Note that as well, we only had access to *published* cases. Some fraction of cases

are selected as not being important enough, or not modifying precedent enough, to have their decisions recorded in official legal books or to have their decisions published online. It is unknown what the ratio of published to unpublished cases are, though from anecdotal evidence in talking to the law student Betsy Byra and searching specialized expensive databases which contain unpublished cases (but which cannot be scraped without breaking Terms of Service), unpublished cases seem somewhere between a quarter and a third.

## VI. Challenges still to be addressed

### A. *More advanced overturned analysis*

The next stage of analysis of overturned analysis is identifying more overturned cases through the similar words identified by word2vec. This will allow for more cases to be taken account for that have not been by prior analysis.

### B. *More advanced word2vec*

Going forward, we would like to use our word2vec scheme to be used as an embedding system for more complex analysis. We will be using word2Vec to compare known suspicious cases to other informat cases to see if we can make predictions about the validity of their ruling.

### C. *Logistic Regression*

Now that we have access to the word2vec embedding and vectorization scheme, we plan to combine this with our analysis of which cases are overturned to figure out which keywords are most important in determining whether a case is overturned, based on their coefficients in the logistic regression.