

# County of Santa Clara

Office of the District Attorney

County Government Center, West Wing  
70 West Hedding Street  
San Jose, California 95110  
(408) 299-7400  
www.santaclara-da.org

(ENDORSED)  
**FILED**  
AUG 15 2018

Clerk of the Court  
Superior Court of CA County of Santa Clara  
BY Janice Jones DEPUTY



Jeffrey F. Rosen  
District Attorney

August 9, 2018

Santa Clara County Civil Grand Jury  
191 North First Street  
San Jose, CA 95113

Dear Civil Grand Jurors:

Pursuant to Penal Code section 933(c), here is the comment of the District Attorney's Office in response to the Civil Grand Jury Report released on June 21, 2018, entitled *2016-2017 Civil Grand Jury Report Follow-Ups: Justice Still Delayed*.

**Finding 1a: The true extent of the slower-than-average felony-case resolution in the County is masked because:**

- A. The Santa Clara County District Attorney's Office (DA) and the Santa Clara County Superior Court (Court) use a different benchmark than the State to report felony- case resolutions.
- B. The Santa Clara County District Attorney's Office and the Santa Clara County Superior Court's figures disagree with one another even when using the same benchmark.

**Finding 1b: Efforts to improve 12-month felony-case disposition rates are weakened by the DA and Court presenting higher figures than the State Judicial Council, perpetuating the "culture of complacency" cited in "Justice Delayed".**

**Recommendation 1: The Santa Clara County District Attorney should publish in its next annual report a reconciliation of its 12-month felony disposition figures with those of the Court and explain how and why its figures differ from those of the Judicial Council.**

**Response: Somewhat Agree.**

The District Attorney's Office worked with New York University's Center for Urban Science and Progress to carefully evaluate case-aging and delay in the criminal process. The

methodology used by NYU and the DA to analyze case delay are outlined with precision in the attached report.<sup>1</sup> The DA shares the Grand Jury's concern regarding the discrepancy between the Court's data and the DA data. We will continue to work with the Court to understand the discrepancy but cannot speak to the Court's methodology. Further, the DA pledges to continue evaluating case delay in Santa Clara County and reporting our findings annually to the public. In our annual report we will explain our Felony Disposition figures and explain our methodology.

The DA agreed with the finding in the initial Grand Jury Report released on June 15, 2017 that there was a delay in the resolution of felony cases in the county due to a culture of complacency within the criminal justice administration. Our office took the Grand Jury's recommendations to heart and some improvement has been made. There has been an increase in attention to felony in custody cases that have aged significantly. These cases are often the most complicated cases to reach resolution. The Grand Jury report spurred greater cooperation between the Court, Defense Bar, DA and County Executive in alleviating this backlog. This cooperation resulted in an unprecedented number of homicide cases proceeding to trial in 2017. We have continued to prioritize these cases in the first six months of 2018.

Case Management meetings have continued to occur between the criminal justice parties in order to discuss relevant topics including delay in each of the departments.

**Finding 2: It is difficult for the public to judge the performance of the DA and the Santa Clara County Public Defender Office (PDO) in improving the speed of felony-case resolutions because they:**

- **Have not publicly provided details about how their respective offices are educating and training their staffs about the ramifications of slow felony-case dispositions.**
- **Do not detail how they are holding their staffs accountable.**

**Recommendation 2a: The DA should issue a report explaining how it holds staff accountable by December 31, 2018.**

**Response: Do not agree.**

We do not feel it is necessary nor legal to publish a report on this matter. Personnel matters are confidential under the County's civil service merit system rules. Media coverage has shown, however, that there is a high level of accountability under the current

---

<sup>1</sup> The attached report from NYU depicting our methodology was completed a few months after our initial response to the Civil Grand Jury report last year. In their final analysis, 81% of our felony cases were resolved in a year rather than 47% (state report) and rather than 73% (NYU's initial calculation which we used in our response last summer). This final number is more aligned with the number reached by the PDO.

administration as to the rules and regulations set forth within the Santa Clara County District Attorney's Office. Contained within the Policy & Procedural Manual (PPM) of the DA there are several sections that require prosecutors in the office to diligently seek the expeditious completion of criminal cases. Section 2.02 of the PPM states:

It is the duty of the prosecutor to eliminate delays in the criminal justice system. A primary goal of this office is to have all cases resolved expeditiously. From the moment a case is brought to us from review through sentencing or final ruling, it is the responsibility of every attorney to keep it moving. Cases should not languish on desks, or in file cabinets, for any reason. Monthly, each supervisor will do a timeliness review of the team's cases and report those results to his or her assistant district attorney. Attorneys should make their supervisor aware of dated cases. Alternatives to support the timely resolution of cases should be considered, including but not limited to, Penal Code 872(b) preliminary hearings, grand jury proceedings and opposing unjustified continuances of trials.

Further, there are additional PPM sections that address a prosecutor's obligation to seek the timely resolution of criminal matters by adhering to Penal Code section 1048, seeking prompt arraignment and pleas, holding preliminary hearings as soon as possible, bringing cases to trial as soon as possible, and not agreeing to continuances of preliminary hearings or trials when there is not good cause. (PPM sections 5.02 (b)(xiv)(1)(a), 5.02(b)(iv), 5.02(b)(viii), 5.02(b)(xi), and 5.02(b)(xiv)(1) respectively.)

In addition to these clear set guidelines that each Deputy District Attorney in the office is required to read, understand and know, there are further measures to monitor this endeavor. Each team is led by a Supervising Deputy District Attorney who is responsible for monitoring the aging of cases on their team. This topic is often discussed at supervisory and team meetings. In addition, the Chief Trial Deputy oversees all matters that are on the trial calendar and meets with attorneys to insure they are managing their trial matters in a timely manner. Each attorney hired by Mr. Rosen attends initial office training, part of which stresses the importance of managing their caseload and their responsibility to seek a timely resolution of their criminal cases.

Pursuant to our responses to the last CGJ report we have instituted several positive changes to resolve cases earlier. We have begun to utilize the criminal grand jury even more in order to expedite the setting of a trial date. We have been proactive in reassigning vertically assigned cases set for preliminary hearing and/or trial when necessary to avoid a continuance even though this causes a new attorney to have to prepare in a shorter amount of time and often cause some angst to victims of crime. We have continued to implement our new discovery procedures on additional teams to provide as much discovery as early as possible on a majority of our cases. We have created a new system in which attorneys are required to input into our case management system when and why a case set for preliminary hearing or trial was continued. This new procedure makes the aging of cases easier to track by Supervisors, Assistants and the Chief Trial Deputy.

**Finding 3a:** It is difficult for the public to judge the performance of the DA and PDO in improving the speed of felony-case resolutions because neither office makes public its felony-case tracking data.

**Finding 3b:** It is difficult to improve 12-month felony-case resolutions when the DA is tracking cases for special attention only at the 12-month mark and the PDO at nine months.

**Finding 3c:** There is potential for more disparity in case-resolution statistics, since the County, PDO and DA are implementing a data management system that differs from the new system being implemented by the Court.

**Recommendation 3a:** The DA and PDO should use identical benchmarks when publishing felony cases statistics.

**Response: Agree**

As indicated in our response to Recommendation 1 we have been transparent about how we reach our data regarding felony disposition and we encourage the PDO to join us in this endeavor.

**Recommendation 3b:** The DA and PDO should start tracking cases for special attention when they have been in the process for six months, starting December 31, 2018.

**Response: Partially Agree**

A six-month resolution period is an appropriate expectation for some but not all criminal cases. Very few serious sexual assault, multi-defendant gang cases or homicide cases can resolve within six months. However, these are not the majority of cases filed in our county and we concur with the Civil Grand Jury that most cases should be resolving in our county more quickly. Many less serious and less complicated cases sent to the Case Management departments are either resolving or being set for preliminary hearing within six months. The cases that do not resolve proceed to preliminary hearing and trial thereafter if they cannot resolve. Our office will continue to work toward a more timely resolution of complicated cases as well.

Efforts by the DA have already been implemented in this area in the form of more significant supervisory oversight of each team's cases and the age of each case. Two Superior Court Judges have begun to set post-preliminary hearing trial setting dates on complicated sex and gang cases in an effort to expedite their readiness for trial or resolution. We look forward to the Case Management departments continuing to stress the expeditious resolution of matters or the setting of preliminary hearings within 90 days of the case appearing on their calendar.

**Recommendation 3c: The DA should publish an annual report on the number of felony cases that remain unresolved after six months and include estimates on how many of those cases could be resolved within 12 months, starting December 31, 2018.**

**Response: Partially Agree**

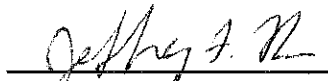
We will include in our annual report the number and type of cases that we believe were resolved in the previous year within six months or a year and differentiate these from the type of cases that were not able to resolve in that timeframe. As to making future predictions we decline to do so. We want to provide the public with accurate information but not predictors based on guess work. As pointed out in our Grand Jury Response dated August 4, 2017, it is the defendant and his or her attorney that control the clock in a criminal case.

Other Findings and Recommendations related to Other Entities and Departments:

The Civil Grand Jury Report has made additional findings that address the other participating entities of the criminal justice system and County government.

The District Attorney's Office looks forward to continuing to implement progressive solutions to this issue and to working with our colleagues in the defense bar and the bench to improve the fair and expeditious resolution of criminal matters.

Sincerely,



Jeffrey F. Rosen  
District Attorney  
Santa Clara County

cc: Dave Cortese, Board President, Board of Supervisors  
Cindy Chavez, Member of the Board of Supervisors  
Joe Simitian, Member of the Board of Supervisors  
Mike Wasserman, Member of the Board of Supervisors  
Ken Yeager, Member of the Board of Supervisors  
Dr. Jeff Smith, County Executive  
Miguel Marquez, County Chief Operating Officer  
The Honorable Patricia M. Lucas  
The Honorable Deborah Ryan  
The Honorable Vanessa Zecher  
Molly O'Neal, Public Defender



## A Data-Driven Evaluation of Delays in Criminal Prosecution

Hrafnkell Hjorleifsson

Michelle Manting Ho

Christopher Prince

Achilles Edwin Alfred Saxby

Mentors: Federica B. Bianco (NYU Center for Urban Science and Progress)

Sponsors: NYU BetaGov/Litmus and the Santa Clara District Attorney's Office.

### Abstract

The District Attorney's office of Santa Clara County, California has observed long durations for their prosecution processes. It is interested in assessing the drivers of prosecutorial delays and determining whether there is evidence of disparate treatment of accused individuals in pre-trial detention and criminal charging practices. A recent report from the county's civil grand jury found that only 47% of cases from 2013 were resolved in less than year, far less than the stat-wide average of 88%. We describe a visualization tool and analytical models to identify factors affecting delays in the prosecutorial process and any characteristics that are associated with disparate treatment of defendants. Using prosecutorial data from January through June of 2014, we find that the time to close the initial phase of prosecution (the entering of a plea), the initial plea entered, the type of court in which a defendant is tried, and the main charged offense are important predictors of whether a case will extend beyond one year. Durations for prosecution are found not significantly different for different racial and ethnic populations, and do not appear as important features in our modeling to predict case durations longer than one year. Further, we find that, in this data, 81% of felony cases were resolved in less than one year, far greater than the value reported by the civil grand jury.

### 1 INTRODUCTION

In the United States, criminal cases are settled through an adversarial system between the prosecuting attorney who represents the public and the defense attorney who represents the accused. The responsibility of the District Attorney (DA) who prosecutes the case is to bring charges against the accused defendant and prove guilt beyond a reasonable doubt. DA performance is frequently measured by the rate of convictions, plea bargains, or diversions (Nugent-Barakove, Budzilowicz, and Rainville 2007). This capstone project focuses on how long it takes for felony cases to be resolved by a District Attorney's office.

This time metric is important to consider because delays in felony case resolutions, or dispositions, places a burden on government resources, leaves defendants uncertain about their futures, and prolongs the victims' wait for closure. Making criminal justice more efficient while maintaining fairness and due process is beneficial for all parties involved. (Association 2006)

The New York University Center for Urban Science and Progress (CUSP) is working with the Santa Clara County (SCC) in California and *BetaGov* at the New York University's Marron Institute to investigate the duration and outcome of SCC's felony cases.

In a recent report issued by the SCC Civil Grand Jury, it was found that Santa Clara was the slowest of all California counties in resolving its felony cases (Santa Clara County Civil Grand Jury 2017). While the rest of the state is able to process 88% of felonies within a year, Santa Clara falls far short. Only 47% of SCC cases in 2013 were resolved within a year. The SCC Civil Grand Jury cited figures from the 2015 Court Statistics report issued by the California Judicial Council (Judicial Council of California 2016)

According to this report, while the rest of the state is able to process 88% of felonies within a year, Santa Clara falls far short: only 47% of their cases are resolved within a year.

The SCC Civil Grand Jury report found through interviews with officials that a "culture of complacency" that tolerates delays in the county and the DA's approach to charging were contributing factors to case delays, among other reasons. The "culture of complacency" refers to a supposed belief among public officials in the criminal justice system that everyone in the criminal justice system is already doing their best to move the process forward and that the state Judicial Council's standard of one year for felony case dispositions is unreasonable for complex cases such as gang crimes.

While SCC had independently observed delays in their case resolutions, this report further motivates SCC to assess its

efficiency in resolving felony cases. The SCC District Attorney's office would like to know how delays and disparities might be explained by case characteristics such as prior convictions, charge enhancements, and defendant characteristics. However, the results of the SCC Civil Grand Jury report are not reproducible, as it is unclear (to us at least) which data sources were used to make these assessments and whether those data sources are publicly accessible. We are tasked with independently quantifying, evaluating, and identifying the delays in case resolution, understanding the drivers of these delays from a data-driven perspective, and relating them to case characteristics.

The deliverables of this capstone project are two-fold. The first is to provide District Attorneys with an interactive dashboard for exploring and visualizing case progression based on key variables such as the number of charges, number of defendants, race and age of defendant, and other case characteristics. The second is to provide an in-depth statistical analysis into what variables change the outcome and lengthen the timeline of cases.

## 2 PRIOR WORK ON EVALUATING PROSECUTORIAL EFFICIENCY

### 2.1 Existing measurements of prosecution performance

Previous studies have examined case processing time as a standardized measurement allowing comparison across jurisdictions (Klemm 1986). In order to use case processing time, researchers first must subdivide case timelines into appropriate time frames and reduce the scope to time under the control of the court system (Neubauer 1983). Early studies have also shown that case complexities such as prior convictions, mandatory minimums, and the number of defendants in specific jurisdictions may contribute to the length of a case (Luskin and Luskin 1986; Walsh et al. 2015). These findings align with the expectations of prosecutors at the SCC District Attorney's Office and form the basis for our capstone project.

In recent years, there have been other data-driven efforts to evaluate and compare court system performance. One such effort is (Measures for Justice 2017), an initiative to aggregate and compare the performance of criminal justice systems from arrest to post-conviction for the entire country via an interactive public dashboard. One of the largest challenges is that criminal justice data is neither recorded uniformly across local jurisdictions nor is it publicly available. The solution from Measures for Justice is to reach out individually to jurisdictions to obtain data and then create standardized core measurements for evaluating performance.

In addition to parsing and understanding case timelines, another motivation of this capstone is to determine whether the addition of defendant characteristics can explain delays in resolution, which would indicate the presence of disparities. It is widely perceived that racial and ethnic disparities

exist in the criminal justice system, and much research has been conducted on biases at the point of arrest and police interaction (Ross 2015). However, no previous work has found the presence of racial disparities in criminal case processing times.

### 2.2 Previous analytical techniques

Machine learning models can be helpful in decision making in the presence of a large amount of data. To be adopted by policy makers, though, they must be easily interpretable and cost-effective. Previous studies on the topic of time to disposition have been dominated by linear regression and basic exploratory analysis. The use of machine learning techniques in the field of criminology is just beginning to emerge. The use of tree-based classifiers to model the outcomes of cases (Katz, Li, and Blackman 2017) and advanced techniques in modeling cost-effective treatment regimes to optimize bail decisions (Lakkaraju and Rudin 2016) focus on accuracy of prediction and optimization. The employment of advanced models on case processing time could help inform prosecutors in making decisions that both minimize case length and prioritize fair outcomes.

## 3 DATA

### 3.1 Data sources

The data used in this project were obtained from the DA's office of SCC, which stores its case information in a case management database called CIBERlaw. We received data for all felony cases charged by the SCC DA's office between January 1st and June 30th 2014 for adult defendants. This allows us to follow case evolution for a time span of three years, which is expected to be sufficient for the vast majority, if not all cases, to have been resolved. The data arrived as four separate datasets:

- **Case Information.** Contains the basic information for each felony case: case ID numbers, defendant ID numbers, and other book-keeping information (e.g. file number). This file contains the time-stamps of case opening (logging) and issuing. It also contains demographic information for each defendant: race/ethnicity, gender, age, and zip code of residence. This file contains 4,794 observations, with the same number of unique defendant ID's and 4,405 unique case ID's.
- **Defendant Charges.** Contains information on the charges faced by each defendant as relayed in penal codes. It contains 34,421 observations with 15,668 unique case IDs, for both felony and misdemeanor cases issued by SCC in the same time frame.
- **Charge Enhancements.** Contains limited information on defendants' prior convictions as well as enhancements on current charges. When a charge is enhanced it mandates harsher sentencing. An example of this is: if a driver is

charged for driving under the influence of alcohol, should the driver refuse to have their blood tested, or their blood alcohol level tests above a certain threshold, the original charge would be enhanced. This dataset contains 10,831 observations with 3,640 unique case IDs, for both felony and misdemeanor cases.

- **Case Events.** Contains information on all events relative to the prosecutorial process other than the case-issuing and case-opening date (which are recorded in Case Information) through 2016. Each event is time-stamped, and falls into one of 155 distinct case event *types* and one of 469 distinct case event *results*. The dataset contains 481,614 observations with 14,983 unique case IDs.

Our unit of observation is a unique case-defendant pair. In order to protect the privacy of the defendants, both defendant and case ID's are anonymized from their entries in SCC's database. These ID's are then used to merge the four datasets into a single set containing all the relevant information on each felony case-defendant pair (each case-defendant pair contained in the Case Information file). We first merge the Case Information dataset with Defendant Charges, and in the process 6 observations are lost (6 cases-defendant pairs missing in the Defendant Charges file), taking the total number of observations from 4,794 to 4,788. Merging Charge Enhancements with the resulting dataset from the previous merge has no affect on the number of observations although only some cases will carry enhancements. Finally, we merge the resulting file with the Case Events file. In this process, 278 observations are lost, taking the number of observations from 4,794 to 4,510.

Misdemeanor data were included in the Defendant Charges and Case Events tables which explains why we find much higher numbers of case ID's in those sets of data than in Case Information. All of the observations with only misdemeanor charges are discarded in the merge process. The reason we lose felony observations in the merging process stems from the fact that some case and defendant ID pairs found in Case Information are missing in Defendant Charges and Case Events. Without direct access to the CIBERlaw system, we do not know what caused these discrepancies, but we only lose 5% of the initial observations, and we do not anticipate this will affect our analysis.

Information on prior convictions of defendants turned out to be incomplete: specifically, *strike priors*, which in the California penal system significantly modify the charges and following procedure, were missing, and it was not possible to know from the data we obtained whether or not a defendant had a strike prior.

### 3.2 Construction of timelines

To understand what causes delays in the prosecutorial process, one must first understand the timeline of a case. To construct a simplified timeline of a case, with guidance from the SCC

DA's office and in accordance with prior work (Neubauer 1983), we identified four key events, which need to be recognized and extracted from our data: *arraignment*, *plea*, *case disposition*, and a case's last event. In some cases an event is explicitly stated in the categories originally listed in the Case Event file, in others it must be inferred. Identifying these events is key to gaining insights about prosecution durations.

From the point of view of a prosecutor, a case generally ends at disposition, or resolution. A disposition usually takes the form of a dismissal, guilty plea, guilty verdict, or acquittal. In the CIBERlaw system, there is no single event that explicitly logs the disposition of a case. Instead there is a number of case *event type* and *results* combinations that can represent disposition.

**Disposition** is defined as the first occurring instance of one of the following event *results*:

- *formal probation granted*,
- *credit time served*,
- *summary probation granted*,
- *sentenced*,
- *prison sentenced imposed*,
- *defendant deceased*,
- *found guilty*,
- *found not guilty*,
- *defendant released by court*,
- *defendant discharged*,
- *deferred entry of judgment PC1000*,
- *cases consolidated*,
- *charges suspended per civil compromise*,
- *motion to dismiss interest justice granted*, or
- *motion to dismiss case granted*.

By these case results we identified the disposition event for 90% of our observations; the remaining 10% are missing clear disposition dates in our data: disposition may not have been reached yet, or may have been logged differently. Disposition events are also logged, probably more straightforwardly, in a separate database by the SCC courts. Access to this database may provide a more solid determination of disposition date and decrease uncertainty in our results.

We are also interested in looking at two other key events for each case: *arraignment* and *plea*. **Arraignment** is defined as the first event for a case-defendant pair of *event type* "Arraignment". **Plea** is defined as the first event where *event result* is one of:

- *Plead guilty*,
- *Plead not guilty*,
- *Not guilty plea entered by court*, or
- *Plead nolo contendere*.



A plea of *nolo contendere*, or no contest, is a plea where the defendant neither admits nor disputes charges. While it is not technically a guilty plea it has the same immediate effect. 3% of cases have no identifiable arraignment event and 7% of cases have no identifiable plea event.

Finally, we extract the time stamp of the very **last event** registered to a case-defendant pair. Notice that we cannot tell from our data if further events related to this case-defendant pair will occur past the finite time span of our data (three years). Thus the *last event* should not be interpreted as *final* event for a case, but rather as *latest* event. The case-load of a court may be significantly increased if court dates are scheduled after the dismissal of a case, for example to monitor probation, thus it is interesting to assess how long a case remains alive in the SCC court system even after dismissal.

Thus defined, these four phases can take negative values if the case is re-issued and the original issue date in the Case Information file is over-written. One specific example of this is a case where disposition happens in September 2014 and the last event registered to the case is in December 2015. However, the case is then re-issued in April 2016 following the approval of Proposition 47 which was passed in a referendum on November 4th, 2014, and which reclassified a number of drug-related offenses, including the most common felony charge in our data (violation of Health and Safety Code 11377(a), possession of methamphetamine). Other reasons for re-issuing cases include motions to re-open the case, for example in the light of new evidence, or consolidating and splitting of cases with multiple defendants. Out of the 4,510 observations in the merged dataset we find that days-to-arraignment has a negative value for 76 observations, days-to-plea is negative for 69 observations, days-to-disposition is negative for 71 observation and days-to-last is negative for 29 observations. All in all, we have 79 cases-defendant pairs with negative values for at least one phase of the process. In obtaining the results that are described below, these 79 observations have been dropped from the dataset, taking the number of observations from 4,510 to 4,431. However, our main results (e.g. the median, 25th, and 75th percentile of the distribution of time-to-disposition) have also been evaluated redefining the time-line for these observations as days passed since case-opening, as in most cases the time elapsed between case-opening and case-issuing is only a few days. None of our results change significantly by choosing to drop or to redefine the time-line for these 79 cases.

### 3.3 Engineered Features

From the attributes of the original sets of data new features were engineered to retain all relevant information we are interested in examining and encode it in a format that enables visualization and modeling. The variables are encoded as either integers (e.g. number of charges for a defendant/case pair), binary (e.g. whether there was a preliminary hearing or not), categorical (e.g. pleas guilty, not guilty, *nolo contendere*), or continuous interval variables (e.g. defendant's

age). The full set of features used in our analysis is listed, in alphabetical order, in Table 1.

## 4 VISUAL ANALYSIS TOOLS

### 4.1 Visual tool to enable data exploration

While we will perform a statistical analysis on the data, this work will be generated from a typical data science approach: finding, comprehending, merging and sorting data, and applying statistical tools and other filters to identify trends in the data. These are not tasks that are suited for a DA's office, which generally has little training for this purpose, and has many other important legal tasks to perform. Therefore, it is desirable to automate much of this process and provide a way for the prosecutors to interactively engage with their data so that they can identify trends without advanced data skills.

Even before the final SCC dataset was in our hands, we generated concepts for the visualization using synthetic data sets. These data sets were constructed with a small set of features of various types that we expected would be of interest to the attorneys. This includes the durations of four phases of prosecution, variables for race and gender, and a value for the age of the accused. Although these are only some of the important variables to consider in our visualization and modeling activities, we chose these for development purposes so that we could determine the best ways to handle arbitrary variables we may want to display. In particular, we have been able to prototype the ability to filter our data based on binary, categorical, and continuous variables. All of the engineered features are enabled in the final version of the dashboard, and new variables can be easily added on an as-needed basis.

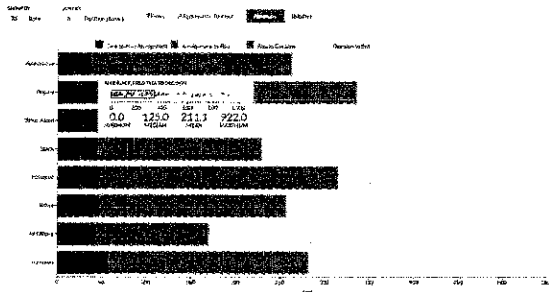
The simplest form of this visualization is a stacked horizontal bar plot (Figure 1). Each bar represents a category of comparison that is selected by the user, (e.g. ethnicity, age group, or court category referring to the court where disposition happened). Visual comparisons are made via three information channels for each bar: its location on the  $x$ -axis, its width, and its color.

The location of the bar encodes the time for a given phase to commence relative to the start of some other chosen phase. Location attributes are most easily compared by a user when they are placed on the same scale (Munzner 2014; Wilkinson 2005). Therefore, we provide the ability to choose which phase to compare against and align the  $x$ -axis (time) such that the phase begins at time  $t = 0$ , and earlier phases are displayed on the negative portion of the scale. For overall case duration comparison, we align to the beginning of the first phase, where the start of each case is displayed at  $t = 0$ .

The width of the bar encodes the duration of each phase. These values are calculated as the difference of the times from the beginning of each case to the ends of two consecutive phases. We enable an absolute time scale visualization (in days) and a relative time scale visualization where each aggregated observation (each bar) is measured and visualized as

Type	Feature	Description	Possible values/range	Missing values
Categorical	Courtroom	in what courtroom did the disposition take place	various	460
Integer	Courtroom Count	through how many different courtrooms did the case go	1-17	123
Categorical	Courtroom Type	what is the type of courtroom where disposition happened	case management court, domestic violence court, drug court, general felony court, north county court, south county court, unknown	123
Binary	Custody	was the defendant in custody or not at the start of the case	0,1	325
Categorical	Defender type	type of defender at disposition	public, private, independent, alternate, unknown	123
Binary	Gang enhancements	are gang enhancements present	0,1	0
Categorical	Initial Plea	did the defendant plea guilty, not guilty or no contest	guilty, not guilty, no contest	325
Integer	Ncharges	number of charges a defendant is facing	1-47	0
Integer	NcourtDates*	the number of court dates for a case	1-76	0
Integer	Ndefendants	number of defendants per case	1-7, 20	0
Integer	Nenhancements	number of charge enhancements	1-26	0
Integer	Nfelories	number of felony charges for a case	1-44	0
Integer	Nfa	number of times a defendant failed to appear	1-12	0
Integer	NHS	number of charges due to violation of the Health & Safety code	1-24	0
Integer	NPC	number of charges due to violation of the general Penal code	1-32	0
Integer	NpleaDates*	the number of plea dates in a case	1-30	402
Integer	NVC	number of charges due to violation of the Vehicle code	1-16	0
Binary	PCI2022	are there other critical enhancements connected to the case (the use of a weapon or presence of injury)	0,1	0
Binary	PCI368	was the defendant deemed incompetent to stand trial at any point	0,1	0
Binary	Prelim	was there a preliminary hearing or not	0,1	0
Categorical	Possible Outcome	what was the inferred sentence outcome	prison, probation/jail, unknown	0
Binary	Public Defender	was the defendant represented by a public defender at any point	0,1	123
Integer	Time to Plea*	the number of days between when case got created until the defendant's initial plea	0-1222	325
Binary	Time waived	was there time waived at any point	0,1	0
Binary	Trial	did the case go to trial or not	0,1	0
Binary	More than a year	is the time to disposition less than or greater than one year	0,1	0
Integer	(Time to Arraignment*)	the number of days between when case got created until the defendant was arraigned	0-1093	120
Integer	(Time to Disposition*)	the number of days between when case got created until disposition	0-1181	460
Integer	(Time to Last*)	the number of days between when case got created until the last event registered to a case	0-1232	0

**Table 1.** Engineered features. *Time to disposition*  $\geq 1$  year is the feature on which the classification, is based. See Section 6. Features marked with a \* are timeline related features, meaning that they intrinsically convey information about the duration of the case resolution, and will be considered differently in our analysis (see subsection 6.4). Features indicated in parenthesis are visualized through our dashboard section 4 and used in the exploratory analysis section 5, but are not used as input features for the models



**Figure 1.** Screenshot of our visualization tool designed to enable exploration of SCC prosecutorial data running in the Chrome web browser. The visualization tool breaks down the prosecutorial process into four phases: case issue-to-arraignment, arraignment-to-plea, plea-to-disposition, disposition-to-last logged event, and enables aggregation, filtering, and sorting on other axes: demographic, court related categories, etc. Here the visualization is using synthetic data, binned and aggregated on age ranges and sorted by the duration of the second phase (“arraignment to plea”). Note that the x-axis (days) is aligned such that the second phase starts at  $t = 0$  and the first phase is shown extending in to the negative portion of the domain. Also shown is an example of the distribution information that is displayed when the user hovers over a bar using a pointing device: minimum, maximum, and a box plot showing the entire distribution for that prosecutorial phase (arraignment-to-plea) and the category belonging to that bar (defendant between 21 and 25 years of age).

a fraction of the longest aggregated observation. Since these times are determined by our own categorization scheme for the case events, the phase durations will be subject to some error depending on how well we can identify the demarcations between the phases in the data and how well the data is entered into the DA’s case management system.

The color of the bars encode which of the four phases is being represented. We use four colors drawn widely and uniformly from the *viridis* color palette (van der Walt and Smith 2015). The colormap was developed for the Matplotlib python graphics package. Viridis has two desirable properties: it is perceptually uniform (meaning that the scale is uniformly smooth and does not induce a perception of structure) and robust to common forms of colorblindness.

An additional channel of information is available when hovering the mouse pointer over any aggregated bar, showing a one-dimensional horizontal scatter plot of the underlying data along a time axis. Also displayed is a box plot of the distribution, as well as the elementary statistics of minimum, maximum and median.

A prototype of this dashboard using synthetic data is available at <http://bit.ly/2hbPqRL>.

## 5 EXPLORATORY DATA ANALYSIS

### 5.1 Measuring the duration of SCC Cases

Having extracted the time of arraignment, plea, disposition and the last event of a case, timelines for each case can now be constructed. Statistics of the phases of the prosecutorial

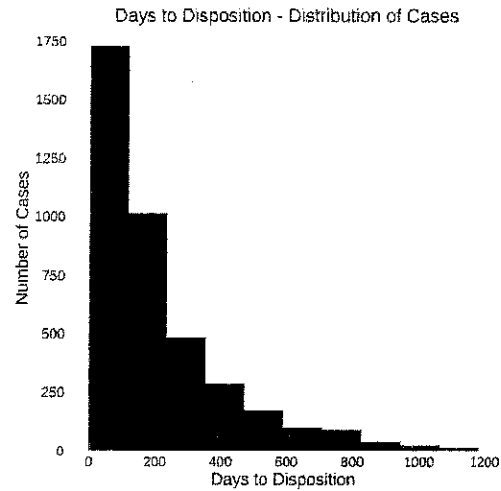
days to	min	25%	median	75%	max	mean
Arrestment	0	1	5	31	1093	30.5
First plea	0	36	90	180	1222	137.4
Disposition	0	63	141.5	281.8	1181	210.8
Last Court Event	0	178	378	684.5	1232*	455

**Table 2.** Statistics on the duration of the prosecutorial process in four phases from the day the case was issued for the 4,431 cases issued between January and June of 2014 by SCC with *complete* information (i.e. missing data were removed by row). (\*) The last event is the latest event logged, but we have no information to indicate whether future court events are possible or expected.

process for the 4,431 cases issued in January through June of 2014 can be seen in Table 2.

The first, important result of our work is already available from this analysis: the median time to disposition in our data is 141.5 days (i.e. 50% of the cases are resolved within 141.5 days). This directly contradicts the findings of the report issued by the SCC Civil Grand Jury which, as discussed in (section 1) states that only 47% of cases in SCC are resolved within a year. Furthermore, according to our findings, 81.5% of cases in SCC were resolved within a year. Figure 2 shows us the distribution of case duration and further emphasizes the point that most cases are resolved early in the process. Because the SCC Civil Grand Jury report is not reproducible, we cannot reconcile or explain this discrepancy. We note that the SCC Civil Grand Jury data was not the specific dataset on which our analysis is based (cases issued in January through June 2014), and that the details of their methodology, including the definition of “duration”, may likely differ. Nonetheless, the difference is very significant, with the upper quartile of our distribution (281.8 days) being below their median, and it would be extremely surprising if the six months of data we have in hand were so severely uncharacteristic to explain the difference. Similarly, the uncertainties we identified in the extraction of the time to disposition (the fraction of cases with missing disposition, the removal of observations with missing data) are not sufficient to explain this discrepancy. The choice of marking the beginning of the timeline for the four prosecutorial phases with the *issuing* date of the case, a time stamp which may be overwritten in our data by a more recent re-issuing date, may bias our statistics to estimate systematically shorter durations. However, even choosing to mark the beginning of these timelines with the *case-opening* date, which conversely biases our analysis towards systematically longer durations, we still estimate the median duration of the time-to-dismissal to  $\sim 150$  days.

Even though the picture we get is not as grim as the one depicted in the Grand Jury report, a rate of 81.5% case closure within a year would still be below the state average of 88% quoted in the report (which however is also obtained through a different analysis).

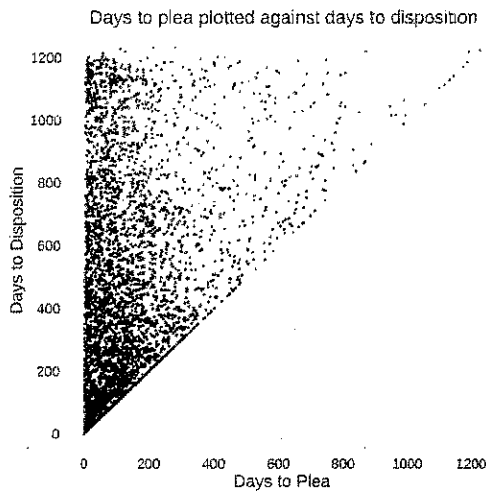


**Figure 2.** Histogram of the days it takes a case to reach disposition starting from the day the case is issued. 71 cases have negative time to disposition due to the case being re-issued in the course of the prosecutorial process. These cases have been dropped.

## 5.2 Exploratory analysis of prosecutorial data

The engineered features allow us to easily examine specific factors that might be believed to cause delays in the prosecutorial process. We begin by assessing if any difference in the statistical distribution of durations can be identified by slicing the dataset along each of the engineered features and extracting the relevant statistics. Below we explore some of the features that we expect, from domain knowledge, could lead to different durations.

A variable of particular interest to the SCC DA is time to plea: is it possible that most prosecutorial delays are determined by a late plea? This is particularly important for SCC since in the SCC judicial system, unlike for most other counties, including the rest of California, the defendant is not bound to take a plea at first arraignment, but instead can defer the plea indefinitely (and then, at the time of entering the first plea, the defendant has the right to a prompt trial, or right to waive time). Figure 3 shows time to disposition plotted against time to plea. Disposition cannot take place before a plea is entered, so if the time to plea is long, time to disposition is also long (these are the data points in the top right quadrant of the plot). However, the opposite is not necessarily true: when time to plea is short (the left portion of the  $x$ -axis), time to disposition can still be long, as indicated by the relatively high density of points in the upper left quadrant of the plot: time to plea cannot be the only variable determining prosecutorial delays.



**Figure 3.** Days to disposition plotted against days to plea for felony cases issued by SCC in January-June 2014. No disposition can happen before the plea, hence the bottom right portion of the plot is empty. The SCC DA indicated the long duration of the prosecutorial process up to plea, which is uncharacteristically long due to peculiarities of the laws that in SCC do not require a defendant to enter a plea early in the case, would drive the long duration of the prosecutorial process to disposition. However, in this plot we see a large fraction of defendant-case pairs at the top left of the plot, with short time to plea, and yet long time-to-disposition, indicating that delays in entering a plea are only partially responsible for delays in the prosecutorial process up to disposition.

To examine what other case factors might be the key drivers of delay we look at case duration for cases with specific characteristics independently. In Figure 4 we look at the distribution of case duration (days-to-disposition) through multiple violin plots. When the data can be split in a binary fashion, a violin plot allows an intuitive comparison of the two distributions. The different colors (blue and green) represent case duration distributions for two different subsets of the dataset. The distributions are normalized and smoothed via kernel density estimate with a Gaussian kernel. The minimum and maximum values of each distribution reflect the shortest and longest case in the dataset (and they need not be equal for the two subsets). We visualize the distributions of days-to-disposition in this fashion for the following binary split of the data:

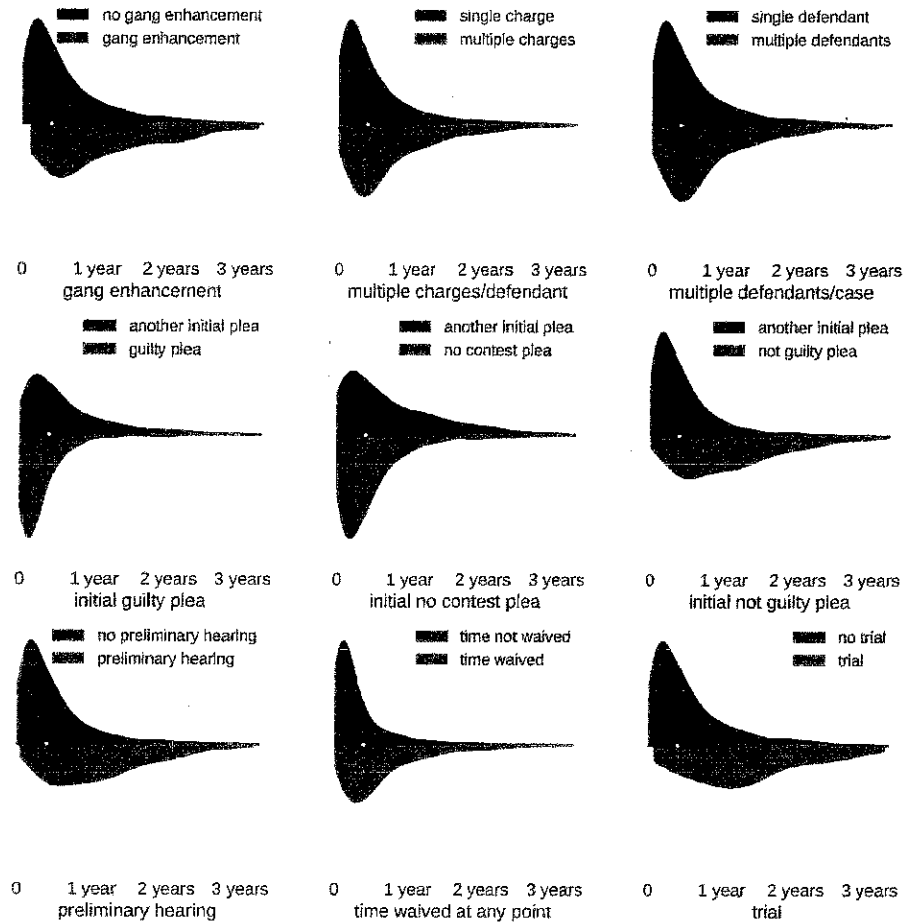
- gang enhancement vs no gang enhancement on the charges,
- single vs multiple charges on the case,
- single vs multiple defendant,

- guilty plea vs any other plea,
- no contest plea vs any other plea,
- not guilty vs any other plea,
- there was a preliminary hearing vs no preliminary hearing on the case,
- time waived vs time not waived,
- trial vs no trial.

More information of these features can be found in Table 1. We see that cases where the defendant initially pleads guilty or no contest to charges are generally resolved early in the process, while cases where the defendant pleads not guilty have a flatter distribution, indicating more variability in the prosecutorial process duration. Cases where the defendant pleads not guilty are more likely to go to trial, so this is consistent with what we see in the distribution of durations for cases that have preliminary hearing and/or a trial. The presence of enhancements and number of defendants are of specific interest as they had been clearly identified by the SCC DA as possible key contributors to delays in the prosecutorial process, but, while the first shows more power in the tail, the presence of more than one defendant on a case or more than one charge against a defendant do not, somewhat surprisingly, show significant differences in case duration. However, we emphasize that the number of cases with multiple defendants and multiple charges is small, so this difference may not be statistically robust.

We can extend this examination to include the other key milestones of a case (subsection 3.2). In Figure 5, we show a comparison of all four prosecutorial phases, time to arraignment, plea, disposition, and the last event for defendants, for defendants initially in custody against defendants initially not in custody. We see that both arraignment and plea most commonly happens very early in the process for those defendants initially in custody. Based on data from January through June of 2014 the median time to disposition for defendants in custody was 89 days. For those out of custody it was 192.5 days.

In Figure 6 we see the same breakdown for defendants that have at some point during a case been found to be not competent to stand trial plotted against all other defendants. While this is a very rare occurrence, it drives the most significant difference in the distribution of durations. The most common time of arraignment, plea and disposition for these defendants doesn't show significant differences, as the motion to evaluate competence to stand trial would occur later in the process. However, the median duration to disposition extends past a year, and most commonly the last event of these cases happens very late in the process, after 1000 days. Based on data from January through June of 2014 the median time to disposition for defendants that are at some point not competent to stand trial was 353 days. For other defendants (excluding the aforementioned group) it was 135 days.



**Figure 4.** Distributions of duration of the full prosecutorial process, from case issuing to disposition, for felony cases issued between January and June 2014 by the SCC DA. In each plot a distribution is shown as a histogram smoothed with a kernel density estimate for two samples (blue and green) split along the vertical axis for comparison: a so-called *violin* plot. Each violin plot shows the time-to-plea distribution for two subsets of our data. The horizontal bar indicates the inter-quartile range (thick bar), full statistical distribution without outliers (thin bar) and median (white dot) for the *entire* distribution. We compare time-to-disposition for defendants (from the top left) going vs not going to trial, charged of crimes with vs without a gang enhancement, which plead guilty vs not guilty or *nolo contendere*, *nolo contendere* vs guilty or not guilty, charged with one vs more than one charge, who waived vs did not waive time (Table 1), who had vs did not have a preliminary hearing, charged as a single defendant vs with others (often occurring in gang related charges), and that pleas guilty vs not guilty or *nolo contendere*

Even though it takes more than twice as long to reach disposition for defendants that have at some point been found to

be not competent to stand trial, this or any of the other engineered features will not explain delays in the prosecutorial

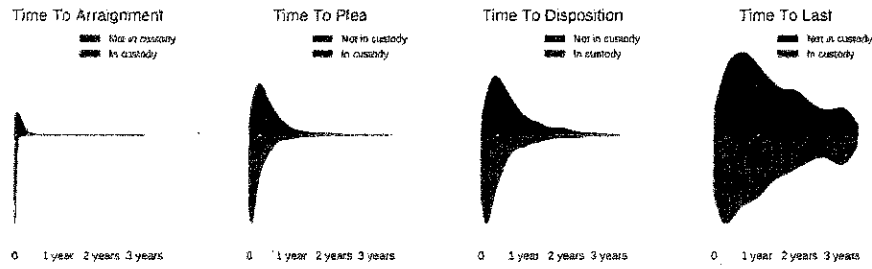


Figure 5. Duration of the prosecutorial process to disposition of cases for defendants initially in custody (blue) compared to defendants initially not in custody (green). No significant differences are observed. Details of the graphics are as in Figure 4.

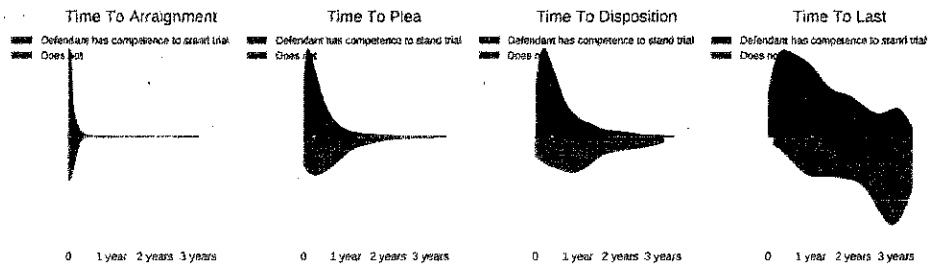


Figure 6. Duration of the prosecutorial process to disposition of cases for defendants initially in custody (blue) compared to defendants initially not in custody (green). Significant differences are observed, especially in the time-to-sentence, the distribution of which peaks later and has more power in the tail, and in the post-sentence duration, with an accumulation of defendant continuing to have court dates scheduled years after the beginning of the case. While these events occur after sentence and do not affect the primary metric we are testing (time-to-disposition and particularly when time-to-disposition extends past a year) it may affect the efficiency of the courts and cause delays in other cases. Details of the graphics are as in Figure 4.

process on their own. The case of being incompetent to stand trial is an exception, applicable to 2% of the defendants in the dataset. In the last section of this paper, we construct models to identify the most prominent drivers of prosecutorial delay.

### 5.3 Demographic data

Having information on age and race/ethnicity allows us to explore the demographics of the data. In Figure 7 we see how the defendants' race/ethnicity breakdown compares to that of the population of SCC. There are some disparities between the two with some ethnicities over- or under- represented in the data. The defendants' age decreases steadily (Figure 8) and the majority of defendants are male (Figure 9). In Figure 10 a choropleth indicates the number of defendants by zip code of residence. The largest number of defendants come from the zip codes around San Jose, as well as zip codes 95037 and 95020 to the south of San Jose

Using the demographics of the data and the timeline of cases gives us an access to case duration for different sections of the population. In Figure 11 we see case duration

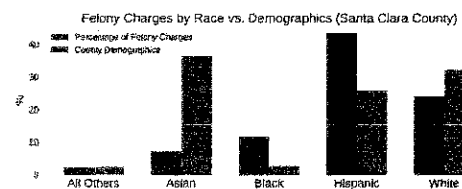


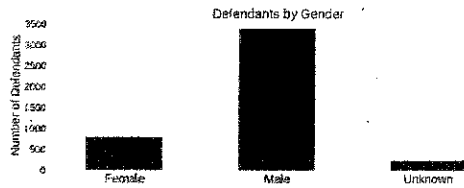
Figure 7. Ethnic breakdown of defendants in SCC felony cases issued between January and June of 2014 (blue) compared to the ethnic breakdown of the population of the county of Santa Clara (green).

by race/ethnicity. Case duration corresponds to “days-to-disposition”, and it is measured in days from when a case gets issued until it is resolved through sentencing or dismissal, as described in subsection 3.2. This analysis does not reveal statistically significant differences in the duration of the process for different races/ethnicities in our case dataset.

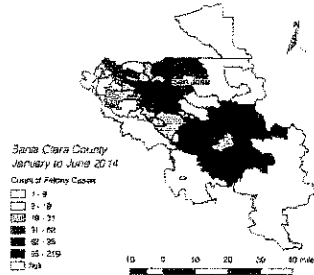
In Figure 12 we have isolated the most commonly found charge in the data, violation of Health and Safety Code



**Figure 8.** Age of defendants at crime commission in SCC felony cases issued between January and June of 2014 by 5 year age bins. Notice that the data only include defendants tried as adults.

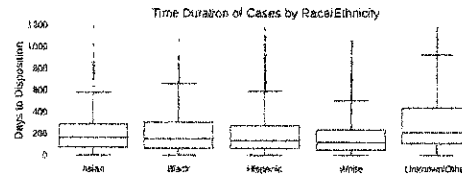


**Figure 9.** Gender of defendants in SCC felony cases issued between January and June of 2014. 77% of the defendants identified as males, 18% as females, and the gender is unknown for 5%.

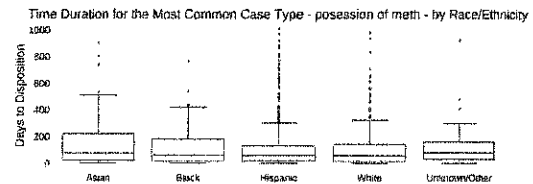


**Figure 10.** Number of defendants by zip code of residence. Most commonly defendants live in the areas in and around San Jose. Zip codes 95037 and 95020 to the south of San Jose are also prominently featured.

11377(a) which is the possession of methamphetamine, in order to control for compounding factors as some felonies may be committed preferentially by certain ethnic groups, leading to systematically different durations in relation to the specific charges. Again, we conclude that no statistically significant difference in the duration of the process for different races/ethnicities appears from this exploratory analysis.



**Figure 11.** Case duration, measured as days between case issue and disposition, by different race/ethnicity. For each ethnic group, the horizontal line within the box represents the median case duration. The box represents the inter-quartile range (IQR), the "whiskers" represent the full distribution, excluding statistical outliers, which are shown as individual data points. No statistically robust differences appear, as all the medians fall in the 25–75 percentiles of all other groups. Curiously, the distribution for Unknown/Other (missing and uncommon ethnic groups) is only marginally consistent with most of the other distributions. We speculate this may be due to cases issued against defendants that are not in custody and not reachable/fleeing from custody, and wish to test this in the future



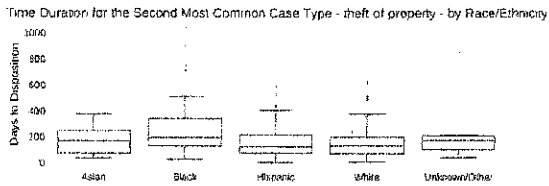
**Figure 12.** As in Figure 11: prosecutorial process duration by ethnicity/race for the most common charge issued by the SCC's DA in January through June of 2014, HS 11377(a) – possession of methamphetamine – to correct for compounding biases in crime by race (e.g. frequency of crime by ethnic group). As for the full charges sample, the distribution of prosecutorial duration is consistent for all ethnic groups.

Lastly, in Figure 13 we have isolated the second most common charge found in the dataset, theft of property –PC 459-460(b)– since Proposition 47 reclassified HS 11377(a) and this would no longer be a felony in charges issued after 2014. There are only 312 observations of this kind, but again, no differences appear.

## 6 ANALYTICAL MODELS

### 6.1 Decision Tree Models

We chose to use decision tree-based models to build a classifier that predicts whether a given case would be disposed within a year or not. Decision trees are considered one of the most versatile machine learning methods (James et al. 2013). Importantly, the model *predicts* the duration but our interest is not in the prediction itself, i.e. we do not intend to build a method that allows the DA to know in advance whether a case



**Figure 13.** As in Figure 11: prosecutorial process duration by ethnicity/race for the second most common charge issued by the SCC’s DA in January through June of 2014, PC 459-460(b) – theft of property. As for the full charges sample, the distribution of prosecutorial duration is consistent for all ethnic groups.

is likely to be processed quickly or not, but rather in identifying the importance of specific input features in making this prediction, and tree based methods allow the importance of a feature to be extracted. Once we have in hand a model with good predictive power, we can identify what input features were used to classify a case-defendant pair as one that will take more or less than a year to reach disposition: those are drivers of prosecutorial delays.

Binary classifiers take the values of the input features  $x_i$  and output  $y_i \in \{0, 1\}$  (the simple binary output we set for our models is time-to-disposition longer, or shorter than one year). Decision trees do this by partitioning the feature space into subspaces, such that the divisions give rise to final regions with learned classifications. The subspaces divided at each step of the construction of the tree represent branches of the tree, and the final set of subspaces after the desired number of partitions are created are the tree’s terminal leaves. Each node of a tree makes an independent binary decision to split the input data: for example, the tree considers age, and splits the sample at a critical value that maximized the so-called “purity” of the outcome with respect to the goal classification, which in our case is whether the case took more or less than one year to reach disposition. Purity can be measured by various metrics, such as variance, Gini impurity, or entropy (subsection 6.2).

These partitions can be complex, with many splits (the so-called “depth” of the tree), leading to many branches with high accuracy (the purity of the leaves). These trees are “strong” learners (performing well on the training data), but generally exhibit poor performance on unseen data in high dimensions since they overfit the training data. Conversely, the partitions can be simple, with few splits (possibly even only one) having nodes with lower purity. These trees are called “weak” learners, but have the advantage of being simple and not overfitting the training data. A decision tree model based on our data, with a depth of 6 resulting in 10 leaves, is shown in Figure 14.

Trees, importantly, enable the use of diverse feature types in input. Since each node generates a binary split based on a single input feature to optimize the purity of the two resulting branches, numerical (continuous and ordinal) variable can be

used along side of binary variables. One must however pay attention to categorical variables that cannot be interpreted in an ordered sense, as for example race/ethnicity, or court type, since a binary decision may not enable a meaningful partition. Mathematically, the correct way to input these variables is by generating  $n$  binary variables for a feature with  $n$  categories (dubbed *one-hot* encoding). However, this tends to wash out the importance of these features and to complicate interpretation. Mostly, if  $n$  is low ( $\lesssim 5$ ) categorical encoding by converting the categories to numerical values is considered not problematic, although technically incorrect. For our data, this is a concern when looking at the courtroom of disposition (which can take 53 possible values) and courtroom type (7 possible values). We will discuss this further in subsection 6.3, and assess the robustness of our results to categorical variable encoding choices by implementing different encoding schemes.

Robust against outliers and data transformations, decision trees are fast and their results are interpretable. In isolation, decision trees can perform well and have low bias, but they tend to exhibit high variance as errors in the first node quickly propagate through the children nodes of the tree when applied to data unseen by the model (James et al. 2013). In order to reduce this variance, ensemble methods are frequently employed. To improve the performance of our models we use two such techniques: Random Forest (RF) and Gradient Boosted Decision Trees (GBDT). Both models are implemented in Python using the `scikit-learn` RF implementation (Pedregosa et al. 2011) and the `XGboost` implementation of the GBDT (Chen et al. 2016).

## 6.2 Leaf purity and features importance

The Gini impurity is calculated as the sum of the products of the population ratio and the classification error rate over each of  $N$  classes,

$$I_{Gini} = \sum_{i=1}^N p_i e_i,$$

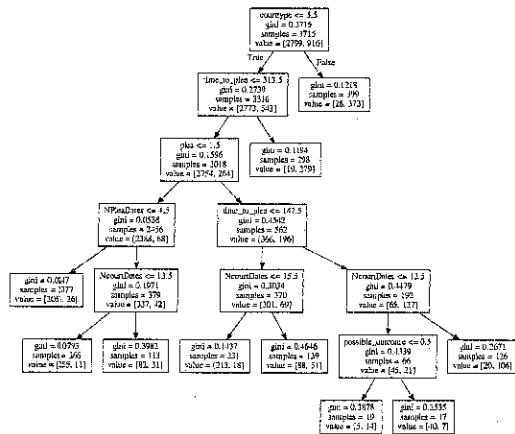
with  $p_i$  being the population ratio and  $e_i$  being the misclassification rate, both for class  $i$ . In the case for  $N = 2$ , this can be simplified: for a leaf having  $a$  members correctly classified and  $b$  members misclassified, the impurity can be calculated as

$$I_{Gini} = \frac{2ab}{(a+b)^2}$$

For example, in the tree in Figure 14, for the rightmost leaf on the bottom level that contains 40 correctly classified and 7 incorrectly classified, the coefficient is calculated as  $\frac{2 \times 40 \times 7}{(40+7)^2} = 0.2535$ .

Leaf impurity measures can be used to determine which features of a decision tree model have the most importance to





**Figure 14.** A single decision tree using the training data set, separating the data into the two classes "time-to-disposition less than one year" and "time-to-disposition greater than one year". At each decision node of the graph, the tree splits that subsample into two subsamples (branches) on the variable indicated. In each node, the graph indicates the boolean test on which the split is performed, the Gini coefficient (representing the purity of the node with respect to the final classification scheme), the number of samples on which the test is performed, and the size of each of the two true classifications. The data is split with the data for which the boolean test is True going to the left child node, and the data for which the boolean test is False going to the right child node. The performance of this tree would be evaluated by measuring how well it classifies a labeled test data set, using the final classifications in the terminal nodes (the "leaves" of the tree), which are assigned to the class having the larger number of observations in the population.

determining the final classifications. The Gini variable importance measure for a variable  $X_m$  in a RF of  $N$  trees is given by Louppe et al. (2013)

$$Imp_{Gini}(X_m) = \frac{1}{N} \sum_T \sum_{t \in T: v(s_t) = X_m} p(t) \cdot (I_{Gini}(t) - p_L I_{Gini}(t_L) - p_R I_{Gini}(t_R))$$

where the summations are over all nodes  $t$  in trees  $T$  having  $X_m$  as the splitting variable,  $p(t)$  is the proportion of observations in the forest that are evaluated at node  $t$ , and  $p_L$  and  $p_R$  are the proportions of the population split to the left and right children nodes  $t_L$  and  $t_R$ , respectively. We use this measure for variable importance throughout the rest of this paper.

In our analysis, the actual performance of the classification is less important than determining the variables that influence the classification. We evaluate the receiver operator characteristic (ROC) plots to validate that the models have reasonable predictive power, but once that is established, the Gini variable

importance measures are our primary interest. Weaknesses of this measure include a bias towards (higher reported values for) continuous variables and away (lower reported values for) variables with a small number of categories. It is also possible for a combination of lower importance variables to be jointly predictive, which would not be detected in a simple evaluation of importance rankings (Epifanio 2017).

### 6.3 Treatment of categorical variables

Categorical variables cannot be split at a tree node in a natural way, as a numerical or boolean variable can be. Two techniques are commonly used to transform categorical variables into other types: "one-hot encoding", which produces multiple boolean variables, one for each category; and a simple numerical mapping that assigns integers  $i \in \{0, 1, \dots, n-1\}$  to each of  $n$  classes, such that each category gets a distinct integer label.

There are several weaknesses introduced with either of these methods. For one-hot encoding, the observations within a single category become sparse which might undermine that category's importance. Also, one-hot encoded features of the original feature are dependent on each other, and this covariance is lost in the model. For the second method, by casting categories into integers we are imposing an order relationship to features that may not possess a natural sense of "greater than" or "less than". To assess if this is a concern, the classification scheme can be permuted to determine if the order affects the outcomes.

To test how the choice of encoding scheme affects the resulting classification, we run RF models using both methods and compare the resulting top feature importance. Here, one-hot encoding extends the feature space from 26 to 248 covariates. The second method performs the numerical cast on each of the categorical variables as described above, keeping the same number of covariates before and after the transformation. The feature importance is similar between the two runs of the model, after observing that the features are themselves split in the one-hot encoded method. Because both methods (one-hot encoding and casting categories into integers) give similar results, we take this to be an indicator of robustness with respect to encoding choice, and in the remaining modeling we use only numerical encoding.

### 6.4 Random Forests

RFs are an ensemble learning method based on decision trees. The prediction of the RF classifier is determined by majority voting across multiple trees fit on subsamples of the data and subsamples of the features.

We ran the RF model using four different sets of input variables. For each we optimize the hyperparameters using a grid search routine from the `scikit-learn` python module (Pedregosa et al. 2011)

In our first iteration, we use all of the engineered features. Using hyperparameter grid-searching, we fit 50 trees with a

minimum of 10 samples at each leaf node, each tree having a minimum of five features, using a Gini impurity criterion (subsection 6.2) to measure leaf purity.

In order to detect disparities, we then recalibrate and run the RF model with the demographic features removed. If the predictive power increased with the inclusion of demographic variables (beyond the expected increase due to a larger feature space), that would indicate that these variables are influential on the model, and suggest the presence of disparities in defendants' treatment based on demographic information. We fit a RF classifier of 50 trees with a minimum of 2 samples at leaf node, again with the Gini criteria. Each of these trees considered a maximum of 20% of the feature space.

Our third iteration of RFs is a classifier without timeline related features (see Table 1). Having timeline related variables, such as time-to-plea, but also number of court dates, and number of plea dates, as input features in the trees may be problematic because of their correlation with the target variable. Moreover, we hope to predict the length of cases with information exogenous to the case proceedings; keeping timeline related variables in the classifier is helpful for pointing out where delays may be happening during a case progression, but we would also like to identify which features of our classifier become important when run without this retrospective information.

The fourth model removes both the timeline related variables and the demographic variables, again comparing the performance of each to identify any impact the demographic variables have on the resulting classification.

### 6.5 Gradient Boosted Decision Trees

Whereas the RF is an ensemble learning method that operates on many decision trees in parallel, the technique known as GBDT is an ensemble method that operates on trees in a serial, recursive fashion.

Boosted models are constructed by adding many weak learners into a single model,

$$f_M(x) = \sum_{i=1}^M T(x; \Theta_i)$$

where  $M$  is the number of learners. Generally,  $T$  could be any type of learner, but in a GBDT  $T(x; \Theta_i)$  is the  $i$ -th tree of the model defined on the input variables  $x$  and whose parameters  $\Theta_i$  define the structure of the tree. The  $i+1$ -th weak learner is generated iteratively by fitting the tree on the residual errors from the model of the first  $i$  summed trees. In practice, this is a difficult problem to solve analytically, so numerical methods are substituted to estimate the next optimal tree. In the GBDT technique, gradient descent is used to find the local minimum of the loss function with respect to the current model. As with other gradient descent learning models, the rate of descent is an additional hyperparameter to tune. The number of trees

$M$  may be chosen *a priori* or be allowed to increase until a desired performance is achieved.

Similarly to the RF models, we run the models four times using the same variable sets as identified above, using the same grid search algorithm to optimize the hyperparameters of the model.

## 7 RESULTS

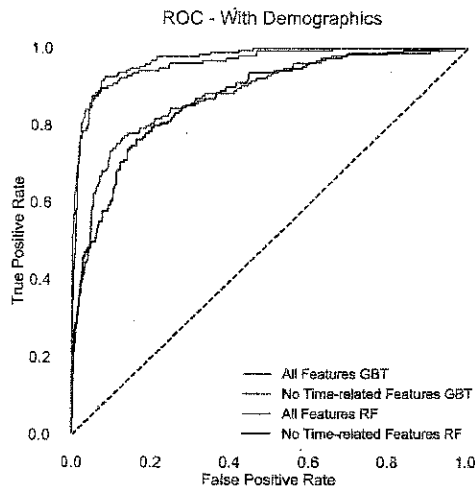
In summary, we generated eight models, without demographic features and without timeline-related features, with one but not the other, with both, for both RF and GBDT.

Because models determine a binary class, their performance can be evaluated using a Receiver Operating Characteristic (ROC) curve, which is based on the probability estimates of the positive class at different false positive tolerance thresholds. In Figure 15 and Figure 16, we show ROCs for all our models, for RF and GBDT respectively. In these figures, the dashed diagonal line represents randomly guessing the class, with contamination (false positives) growing along the  $x$ -axis, and completeness increasing along the  $y$ -axis. The farther left and above the model curve falls from the diagonal, the more accurate the model is at predicting the positive class, with a perfect classifier generating a single point at  $x=0$  (no false positives) and  $y=1$  (no false negatives). Our RF and GBDT models perform similarly for each set of input variables. Removing the timeline-related data degrades the performance of the model significantly, pulling the curves closer to the 45 degrees line. From this we infer, as expected, that the timeline-related variables are strong predictors of long-duration dispositions. By the same analysis, the fact that there is little to distinguish the performance of the models when demographic variables are removed from consideration in either set of models indicates that demographic do not affect the prosecutorial duration. This is consistent with the exploratory analysis performed above: we find no evidence of disparities in the treatment of felony cases *as measured by the duration of the prosecutorial process alone* for cases issued by SCC in January-June 2014.

The feature importance for each model is shown in Figure 17, where the top 10 features of each models are shown in a horizontal bar chart with the relative importance indicated by the length of the bar.

We note that the relative importance has a significant drop after the very first feature for both RF and GBDT models when time-related features are omitted, indicating that the following features have significantly less predictive power than the first and, more importantly, that they are all roughly of the same importance, and we also see a significant drop after the first two features in GBDT when time-related features are included, but the decline in importance is smooth for the first several features in the RF models with time-related variables included.

We found that **time to plea** is the most important feature for both RF and GBDT models, regardless of the inclusion of

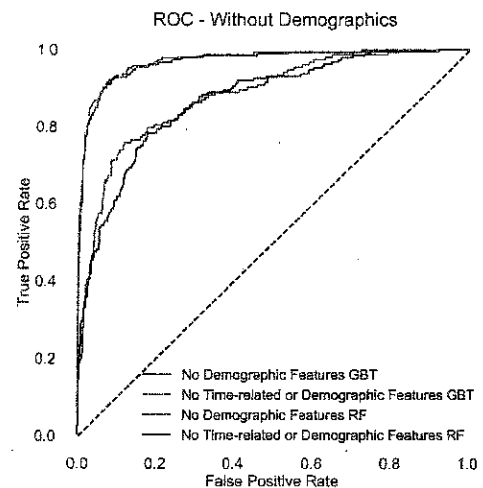


**Figure 15.** ROC curves for the RF and GBDT models, including demographic data in input. The ROC plots show True Positive (completeness) vs False Positive rate (purity) of a classifier. Curves close to the outer edge of the upper right quadrant of the plot indicate good predictors. The ROCs change depending on the subset of variables in the model, but for a given set of input variables, the two ensemble methods exhibit similar performance. However, the inclusion of the time-related features in the input set significantly improve the accuracy of the predictions.

demographic variables (see subplots (a), (b), (c), and (d) in Figure 17). This corresponds with our expectation based on Figure 3, because a case can not be concluded until a plea has taken place. Splitting the cases on that feature at 365 days will yield a perfectly pure node in the tree.

In addition, the **number of court dates** and **type of initial plea** appear in the top five of both models, with and without demographic variables. The importance of the time-related features (time to plea and number of court dates) is not surprising since they are correlated with time to disposition. The **number of plea dates** is an important feature, but only showing in the RF. The relatively high power of prediction for this features also aligns with our expectations, but its correlation with time to plea may weaken the importance, and may explain the discrepancy in the RF and GBDT models.

With the addition of demographic variables, we found no change in the predictive power of our models with time-related variables in input. Although **age at offense** and **zip code** appear in the top 10 important features of the GBDT, the models are still dominated by the first few time-related features.



**Figure 16.** ROC curves for the RF and GBDT models, without demographic data in input. A change in the shape of the ROC curves compared to Figure 15 (with demographic variables in input) would indicate a change in the predictive power. An significant improvement in predictive power would point to a disparate treatment of accused individuals. However we do not observe that here. This suggests that there is no evidence of significant disparities in the duration of the prosecutorial process based on defendant's demographic information (age, gender, race/ethnicity, and residence area).

After removing time-related features, we find agreement between RF and GBDT that **possible sentence outcome**<sup>1</sup> is among the most important features. The **main charge** and **type of initial plea** are the primary importance features for both GBDT and RF models, respectively (panels (e) through (h)), however they only appear as at best the fifth-most important feature in the complementary model. It is not clear why the models would predict such different results, although likely the explanation resides in the existence of covariance between features: for example, covariance exists between main charge and courtroom type. Main charge is the first feature by importance in GBDT, and sixth in RF, however courtroom type is only fifth for GBDT and third for RF. It is harder to explain why initial plea appears only in the seventh place by importance for the GBDT model, while it is the first and foremost feature in the RF models. There are likely hidden covariances with initial plea and many other variables, court type, number of felony charges, and certainly possible sentence outcome.

<sup>1</sup> This feature is indicates as possible sentence outcome because, like disposition, the outcome of disposition is not clearly logged in CIBERlaw, and parsing through the event *results* and *types* gave us a insight into the outcome, categorized as probation/jail, prison, or other, but with some uncertainty.

**Courtroom count** (the number of courtrooms that had scheduled events for this case) now appears as an important feature in both models, but should be noted that this feature could also be interpreted as time-related.

The feature with the third highest importance in the RF models without time-related features is the **courtroom type**, type of court where case disposition took place. Courtroom types are associated with specific types of criminal cases such as drug or domestic violence crimes, or they are specific to a different geographic area of the county. Court types are associated with specific types of criminal cases such as drug or domestic violence crimes, or they are specific to a different geographic area of the county. Thus, they may be related to the main charge, which is an important feature in the GBDT model, as indicated above.

Surprisingly, whether a case went to trial or had a preliminary hearing was not considered important by our models in determining case duration, even though they appear to affect the distribution of durations (Figure 4). However, only 100 cases went to trial and only 848 held a preliminary hearing, therefore their measured importance may be suppressed in the classifier. Furthermore, preliminary hearing and trial may correlate with other features, such as court type and initial plea. This correlation may also weaken their importance.

Enhancements on a case were very weak features in determining case disposition within a year.

Full details of the feature importance for all eight runs of our models are given in Table 3.

## 8 CONCLUSIONS AND FUTURE WORK

Understanding why some criminal cases take a long time to resolve is a complex but important task. We attempted to shed light on this issue by creating a dashboard for exploration of case timelines and by modeling assess which features are important in determining a case's duration.

We received data describing all felony cases issued by the SCC DA office between January and June 2014, including all case events through 2016, allowing us to follow a case evolution through 2016, for 2.5 to 3 years (section 3). Several tasks of parsing and preprocessing case events and characteristics were necessary, which required close collaboration with domain experts, our colleagues at the SCC DA's office.

We described the design and construction of a case timeline visualization tool (section 4). This tool eases exploratory analysis by representing timelines as horizontal bar charts that can be grouped, filtered, and sorted according to a user's choices.

We explored the data's statistical properties (section 5) and found that the time to disposition in our data appears rather shorter than expected, with 50% of cases completing within 141.5 days. This directly contradicts the findings of a recent report issued by the SCC Civil Grand Jury (Santa Clara County Civil Grand Jury 2017) that identifies SCC as the slowest

county in California in resolving felony cases, and which states that only 47% of cases in SCC are resolved within a year. Because the methodology in the SCC Civil Grand Jury report is undisclosed, we cannot explain these discrepancies, but we document our methodology and could perform a closer comparison if documentation on their methodology became available.

We constructed decision tree models (section 6), Random Forest (RF) and Gradient Boosted Decision Trees (GBDT) models, to isolate important features in determining whether a case is resolved within a year. We found general agreement in the feature importance between RF and GBDT models, with some notable discrepancies, which highlight the perils of using decision trees to assess feature importance in the presence of covariant features. The features' covariance can obfuscate the importance of a feature, since the information content of the two covariance features is the same.

We found that the time to the first plea event is one of the most important features in determining case length (section 7). This statement is somewhat misleading, however, because time to plea is strongly correlated with time to disposition, our target variable, since disposition cannot happen before plea. There are however several cases where plea occurs early and yet disposition is not reached until much later (Figure 3). We also found that the type of initial plea and the type of court room for disposition were important features. Because the type of initial plea is highly important in determining duration, the SCC DA could investigate their plea bargaining strategies and change initial plea offers to influence overall case durations. In addition the SCC DA may want to investigate which courtroom types and which court rooms host the slowest cases.

Lastly, we found no evidence of disparate treatment of accused in the duration of the prosecution for different races/ethnicities or different age ranges, neither in our exploratory data analysis, by extracting the statistical profiles of the data subsets as split on demographic variables, nor by modeling the prosecutorial duration and comparing models that use and do not use demographic information in input, which have similar explanatory power.

There is much future potential work that could be done with this dataset. Augmenting this data with external datasets such as sentencing outcomes, bail amounts, concurrent individual court room, attorney case loads, and arrest and incarceration rates in SCC may yield more robust models. A comparison with case data from similar jurisdictions in California would also be valuable.

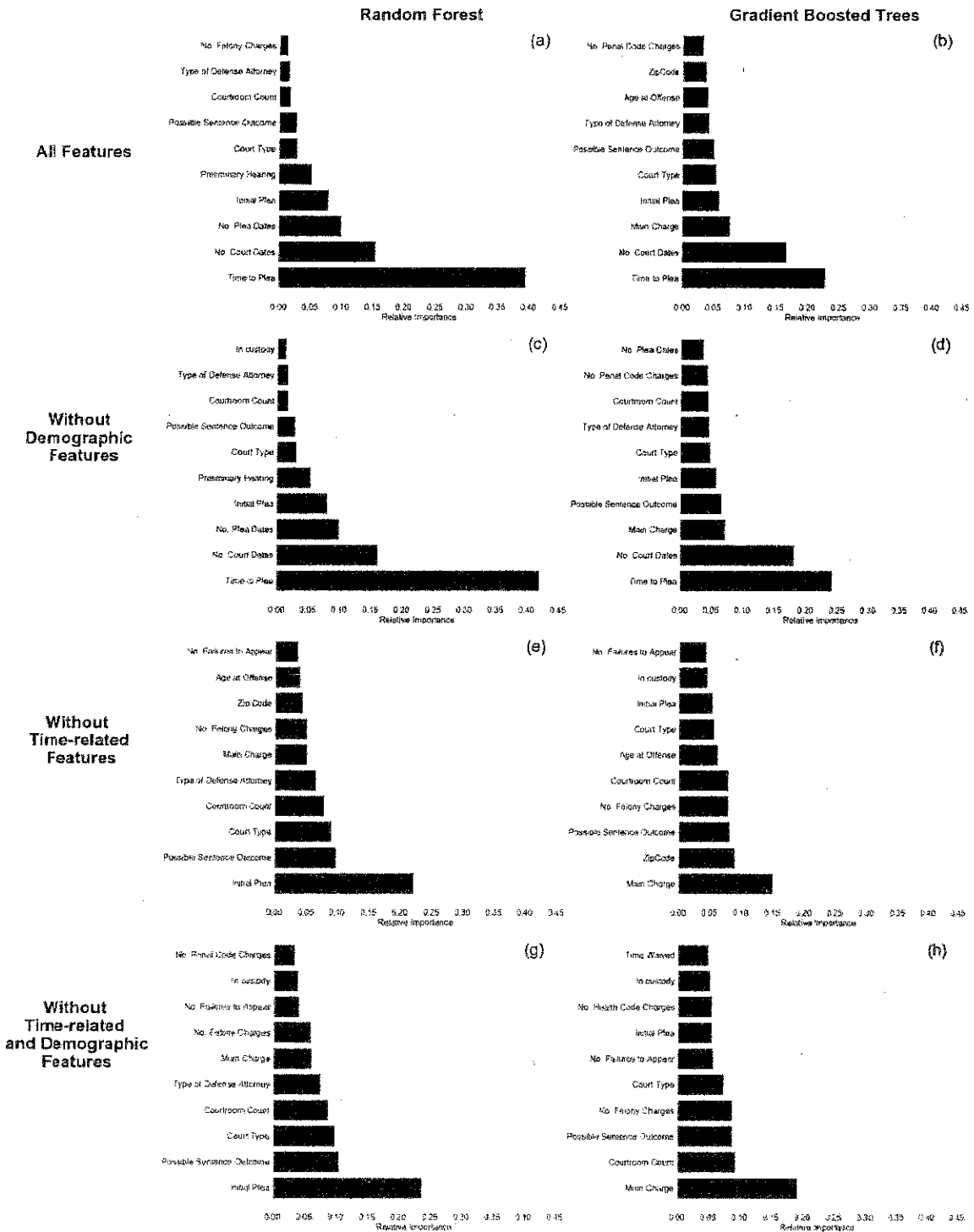


Figure 17. Top ten feature importance for each of the models run for RF (panels on left, (a), (c), (e) and (g)) and GBDTs (panels on right, (b), (d), (f) and (h)).

	Random Forest				Gradient Boosted Tree			
	With time-dependant vars		W/o time-dependant vars		With time-dependant vars		W/o time-dependant vars	
	w/ Demogr.	w/o Demogr.	w/ Demogr.	w/o Demogr.	w/ Demogr.	w/o Demogr.	w/ Demogr.	w/o Demogr.
Age at Offense	0.011	—	0.04	—	0.042	—	0.062	—
Court Type	0.03	0.031	0.09	0.098	0.054	0.048	0.056	0.073
Courtroom Count	0.018	0.018	0.078	0.087	0.032	0.045	0.079	0.092
Enhancement PC 12022	0.003	0.003	0.012	0.015	0	0	0.003	0
Enhancement PC 1368	0.001	0	0.004	0.004	0.002	0.002	0.028	0.032
Gang Enhancement	0	0	0.002	0.002	0.002	0.002	0.002	0
Gender	0.002	—	0.006	—	0	—	0.008	—
In custody	0.013	0.013	0.036	0.039	0.018	0.021	0.045	0.052
Initial Plea	0.08	0.08	0.222	0.236	0.059	0.057	0.054	0.055
Main Charge	0.013	0.013	0.051	0.06	0.077	0.072	0.151	0.191
No. Charges	0.006	0.006	0.021	0.026	0.024	0.032	0.011	0.019
No. Court Dates	0.155	0.161	—	—	0.167	0.182	—	—
No. Defendants	0.003	0.003	0.011	0.014	0.005	0.006	0.005	0.006
No. Enhancements	0.007	0.007	0.028	0.032	0.005	0.016	0.012	0.021
No. Failures to Appear	0.006	0.006	0.036	0.04	0.019	0.014	0.044	0.057
No. Felony Charges	0.014	0.012	0.051	0.059	0.022	0.021	0.079	0.087
No. Health Code Charges	0.006	0.006	0.017	0.022	0.006	0.011	0.036	0.055
No. Penal Code Charges	0.008	0.008	0.029	0.034	0.034	0.043	0.031	0.042
No. Plea Dates	0.1	0.1	—	—	0.024	0.037	—	—
No. Vehicle Code Charges	0.001	0.001	0.004	0.005	0.003	0.005	0.011	0.021
Possible Sentence Outcome	0.029	0.029	0.097	0.105	0.051	0.065	0.081	0.087
Preliminary Hearing	0.052	0.053	—	—	0.006	0.011	—	—
Public Defender	0.002	0.002	0.005	0.006	0.014	0.013	0.006	0.011
Race/Ethnicity	0.005	—	0.017	—	0.014	—	0.023	—
Time Waived	0.006	0.006	0.021	0.025	0.008	0.01	0.039	0.049
Time to Plea	0.396	0.42	—	—	0.229	0.242	—	—
Trial	0.004	0.004	0.013	0.016	0	0	0.002	0.002
Type of Defense Attorney	0.017	0.017	0.066	0.074	0.043	0.046	0.042	0.047
Zip Code	0.013	—	0.043	—	0.038	—	0.09	—

Table 3. Feature importances found for each of the eight runs of models. “—” indicates the variable was not used in the model.

## REFERENCES

- Nugent-Barakove, M. Elaine, Lisa M. Budzilowicz, and Gerard Rainville. 2007. "Performance Measures for Prosecutors: Findings from the Application of Performance Measures in Two Prosecutors' Offices". National District Attorneys Association, American Prosecutors Research Institute. [http://www.ndaa.org/pdf/performance\\_measures\\_findings\\_07.pdf](http://www.ndaa.org/pdf/performance_measures_findings_07.pdf).
- Association, American Bar. 2006. *ABA Standards for Criminal Justice*. Washington, D.C.
- Santa Clara County Civil Grand Jury. 2017. "Justice Delayed: Why Does It Take So Long To Resolve Felonies In Santa Clara County?". County of Santa Clara Superior Court of California. [http://www.sccscourt.org/court\\_divisions/civil/cgj/2017/Why\\_Does\\_It\\_Take\\_So\\_Long.pdf](http://www.sccscourt.org/court_divisions/civil/cgj/2017/Why_Does_It_Take_So_Long.pdf).
- Judicial Council of California. 2016. "Statewide Caseload Trends, 2005-2006 through 2014-2015". <http://www.courts.ca.gov/documents/2016-Court-Statistics-Report.pdf>.
- Measures for Justice. 2017. "Measures for Justice". <https://www.measuresforjustice.org/>.
- Klemm, Margaret F. 1986. "A Look at Case Processing Time in Five Cities". *Journal of Criminal Justice* 14 (1). Elsevier BV: 9–23. doi:10.1016/0047-2352(86)90023-1.
- Neubauer, David W. 1983. "Improving the Analysis and Presentation of Data on Case Processing Time". *The Journal of Criminal Law and Criminology* 74 (4). <http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=6410&context=jclc>.
- Luskin, Mary L., and Robert C. Luskin. 1986. "Why So Fast, Why So Slow: Explaining Case Processing Time". *Journal of Criminal Law and Criminology* 77 (1). <http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=6512&context=jclc>.
- Walsh, Wendy A., Tonya Lippert, Meredyth Goldberg Edelson, and Lisa M. Jones. 2015. "Length of Time to Resolve Criminal Charges of Child Sexual Abuse: A Three-County Case Study". *Behavioral Sciences & the Law* 33 (4). Wiley-Blackwell: 528–45. doi:10.1002/bsl.2187.
- Ross, Cody T. 2015. "A Multi-Level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States 2011-2014". Edited by Peter James Hills. *PLOS ONE* 10 (11). Public Library of Science (PLoS): e0141854. doi:10.1371/journal.pone.0141854.
- Katz, Daniel Martín, Michael J. Bommarito II, and Josh Blackman. 2017. "A General Approach for Predicting the Behavior of the Supreme Court of the United States". *PLOS ONE* 12 (4): e0174698. doi:10.1371/journal.pone.0174698.
- Lakkaraju, Himabindu, and Cynthia Rudin. 2016. "Learning Cost-Effective and Interpretable Regimes for Treatment Recommendation". In *ArXiv:1611.07663 [Stat]*. Barcelona, Spain. <http://arxiv.org/abs/1611.07663>.
- Munzner, Tamara. 2014. "Visualization Analysis and Design". CRC press.
- Wilkinson, Leland. 2005. *The Grammar of Graphics*. Springer-Verlag. doi:10.1007/0-387-28695-0.
- van der Walt, Stéfan, and Nathaniel Smith. 2015. "Matplotlib Colormaps". <http://bids.github.io/colormap/>.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer New York. doi:10.1007/978-1-4614-7138-7.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python". *Journal of Machine Learning Research* 12: 2825–30.
- Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press. doi:10.1145/2939672.2939785.
- Louppe, Gilles, Louis Wehenkel, Antonio Suter, and Pierre Geurts. 2013. "Understanding Variable Importances in Forests of Randomized Trees". In *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, 431–39. Curran Associates, Inc. <http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests.pdf>.
- Epifanio, Irene. 2017. "Intervention in Prediction Measure: a New Approach to Assessing Variable Importance for Random Forests". *BMC Bioinformatics* 18 (1). Springer Nature. doi:10.1186/s12859-017-1650-8.