Analysis of Transparency within the Utah Criminal Justice System

Jess Campbell
University of Utah

UUCS-20-014

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

3 December 2020

Abstract

The state of Utah has numerous laws and regulations governing the process that begins when a person is arrested and booked into jail. These include requirements such as when the probable cause affidavit or charges should be filed, or when pretrial status should be determined. Considering the deleterious effects even a few additional days of incarceration can have on a person's life it is important that there is sufficient data to determine if these statutes are being followed. Using custom web scrapers, I analyzed whether it is possible to make this determination from publicly available booking information. This analysis also includes recommendations to the state of Utah to improve the current system of accountability.

ANALYSIS OF TRANSPARENCY WITHIN THE UTAH CRIMINAL JUSTICE SYSTEM

by

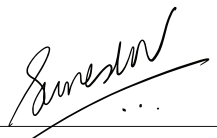
Jess Campbell

A Senior Thesis Submitted to the Faculty of
The University of Utah
In Partial Fulfillment of the Requirements for the Degree

Bachelor of Computer Science

School of Computing
The University of Utah
November 2020

Approved:



/ 12/02/2020

Suresh Venkatasubramanian, PhD
Thesis Faculty Supervisor

/

H. James de St. Germain
Director of Undergraduate Studies
School of Computing

Acknowledgements

There are a great number of people without whose support I would not have been able to complete this thesis. Firstly, many thanks go to my advisor, Suresh, who was willing to take me on and present me with this topic, even after I besieged him in his office.

Secondly, to Jason Groth, Brittany Urness, and the Smart Justice Team at the ACLU of Utah: my deepest thanks for allowing me to use my skills to shed more light on this vital topic, as well as being willing to assist with the legal aspects of my research.

And finally, to my amazing husband, Jon, and our three children: I could not have accomplished so much without your continued love and support.

Abstract

The state of Utah has numerous laws and regulations governing the process that begins when a person is arrested and booked into jail. These include requirements such as when the probable cause affidavit or charges should be filed, or when pretrial status should be determined. Considering the deleterious effects even a few additional days of incarceration can have on a person's life, it is important that there is sufficient data to determine if these statutes are being followed. Using custom web scrapers, I analyzed whether it is possible to make this determination from publicly available booking information. This analysis also includes recommendations to the state of Utah to improve the current system of accountability.

Contents

1 Introduction **6**

2 Background **8**
- 2.1 The U.S. Criminal Justice System . . . 8
  - 2.1.1 Racial and Socioeconomic Disparities . . . 9
  - 2.1.2 The Negative Effects of Pretrial Detention . . . 10
- 2.2 Criminal Justice Algorithms . . . 11
  - 2.2.1 Algorithmic Transparency . . . 11
- 2.3 Why We Need Transparency . . . 14

3 Utah's Current Transparency Laws **16**
- 3.1 HB 288 . . . 16
  - 3.1.1 County Jails . . . 16
  - 3.1.2 Prosecutorial Agencies . . . 17
  - 3.1.3 Administrative Office of the Courts . . . 18
  - 3.1.4 Commission on Criminal and Juvenile Justice . . . 19
- 3.2 Analysis . . . 19

4 Project Description **21**
- 4.1 Goals . . . 21
- 4.2 Design . . . 22
- 4.3 Limitations . . . 23

5 Implementation **24**
- 5.1 Web Scraping . . . 24
  - 5.1.1 Master Scraper . . . 25
  - 5.1.2 Scrapers . . . 25
  - 5.1.3 Database . . . 26
- 5.2 Initial Analysis . . . 27
  - 5.2.1 Salt Lake County . . . 29
  - 5.2.2 Sanpete . . . 30
  - 5.2.3 Washington . . . 31
  - 5.2.4 Morgan . . . 32
  - 5.2.5 Cache . . . 33
- 5.3 The Scraped Websites . . . 34
  - 5.3.1 Beaver . . . 35
  - 5.3.2 Davis . . . 36
  - 5.3.3 Tooele . . . 37
  - 5.3.4 Utah . . . 38
  - 5.3.5 Weber . . . 39
- 5.4 Beaver Court Data . . . 39

6 Conclusion **41**
6.1 Transparency Analysis 41
6.2 Recommendations 42

List of Figures

1 Detention Flowchart. 21
2 Scraper and Analysis Sequence Diagram. 26
3 Database Schema. 27

List of Tables

1 Analysis Overview of Selected Websites 29
2 Analysis Overview of Scraped Websites 34

1 Introduction

As Bryan Stevenson stated in his book, *Just Mercy*, "[t]he true measure of our character is how we treat the poor, the disfavored, the accused, the incarcerated, and the condemned" [40]. This is true of our entire criminal justice system. Yet how can we measure our character when we cannot measure how we are treating these groups? The lack of comprehensive and standardized data surrounding the criminal justice system renders it nearly impossible to fully understand how disenfranchised groups in America are being affected.

It is no secret that the criminal justice system in the United States is broken. With over 10 million arrests every year and 2.3 million people incarcerated at any given time, the United States holds 25 percent of the world's prison population, while only accounting for five percent of the total population [25]. In addition to this, numerous sources point out the racial and social inequity prevalent in our system: people of color are more likely to be stopped and questioned, whether in a vehicle or no [9, 18]; people of color are more likely to be searched, ticketed, or arrested following such a stop [27]; people of color and those within a lower socioeconomic status are more likely to be denied bail pretrial, or granted bail they cannot afford [3]. In a 2018 report to the United Nations, the Sentencing Project stated that the "United States in effect operates two distinct criminal justice systems: one for wealthy people and another for poor people and people of color" [42]. In light of these systemic problems, many states and local authorities are turning to machine learning algorithms, arguing that the objectivity of a computer can help overcome these biases. Yet, an algorithm can only learn based on the data it is given and in researching this topic, time and time again in various forms I have encountered the disclaimer: "where data is available." Thus the question becomes, how much do we know right now about the criminal justice system within the state of Utah? In particular, for those who are arrested, can we determine if the laws governing the process from arrest to sentencing are being adhered to? Data, in the end, is the most important thing. Without the data, how can we keep law enforcement, the jails, and the courts accountable? Without the data, how can we ensure that algorithms that are being implemented are having the effect they claim, instead of perpetuating the status quo? As part of this thesis, I give an overview of the current status of the criminal justice system within the United States, as well as briefly examining some of the causes that have lead to the problem of mass arrests and incarceration. I focus specifically on the state of Utah as well, in ad-

dition to the laws currently in place regarding data collection and transparency. Furthermore, I also examine the use of machine learning algorithms within the criminal justice system and argue that the lack of transparency - both in the data used to train the algorithms and in how they work - undermines the basic tenets of scientific theory and due process. I then propose a method of analyzing the level of transparency within Utah and whether the information provided publicly by the local jails and courts is sufficient for a third party to determine if state statutes are being followed. This was accomplished through a combination of web scraping and requests through the Administrative Office of the Courts (AOC). I also have conducted a thorough analysis of H.B. 288, the data collection law passed earlier this year, and address how it will potentially affect transparency when it goes into effect. I also address how it could potentially change the outcome of this research, as well as its shortcomings. In particular, I argue that it does not go far enough to answer the questions I have posited, or to address the problems of transparency and the fragmented and disparate entities that make up Utah's criminal justice system. Without the benefit of a centralized database or a trusted entity to manage it, the new law falls far short of its intended goals. Lastly, based on the results of my research, I have presented a recommended course of action for the state of Utah, including a centralized criminal justice database, accessibility for researchers and legislators, and privacy and security for those who are processed through the system.

2 Background

2.1 The U.S. Criminal Justice System

As stated previously, there is estimated to be more than 10 million arrests every year, with approximately 2.3 million people currently incarcerated. Between 600,000 and 700,000 of those are held in local jails alone, and a significant number - approximately 460,000 - have not been convicted of any crime [37, 28]. This has not always been the case. The incarceration rate remained fairly steady from the 1920's through the 1970's. However, in the past four decades, it has quadrupled and remained steady ever since [44]. The National Research Council states that most of this growth can be attributed to "the likelihood of imprisonment and in lengths of prison sentences", where likelihood of imprisonment refers to the fact that far more crimes are punished with imprisonment than before the act. Along with this, some states have legislated "truth-in-sentencing" acts, that work to abolish parole. This paper also references the federal Violent Crime Control and Law Enforcement Act of 1994, which mandated that states applying for a federal grant for prison construction showed that it:

> (A) has increased the percentage of convicted violent offenders sentenced to prison; (B) has increased the average prison time which will be served in prison by convicted violent offenders sentenced to prison; (C) has increased the percentage of sentence which will be served in prison by violent offenders sentenced to prison.

This act also created 50 new federal offenses and 60 new death penalty offenses, eliminated eligibility for the Pell grant for incarcerated people (putting higher education out of reach for lower-income people), and increased funding for additional law enforcement officers [45]. This act was passed shortly after the overall crime rates for the United States started decreasing. Some, including former President Bill Clinton, claim this crime bill and the additional police officers it put on the street was largely responsible for the drop in crime, but researchers have estimated that, at best, it contributes less than five percent of a double-digit decrease [16]. All this to say that while crime has gone down over the past three decades, it has not been due to the tough-on-crime policies of the nineties and early 2000s.

2.1.1 Racial and Socioeconomic Disparities

In conjunction with the troubling rise in arrests and incarceration, there has been an equally troubling issue with how the criminal justice system affects people of color and those who are live at or near the poverty level. The report to the United Nations highlights the severe racial disparities present at every step of the criminal justice system: Black people are more likely to be arrested and convicted than white people, and once convicted, receive harsher sentences [42]. In this same report, the statistics are put into startling terms: "[a]s of 2001, one of every three black boys born in that year could expect to go to prison in his lifetime ... compared to one of every seventeen white boys." Likewise, the majority of the country still requires bail to be paid to secure one's release pretrial. This leaves those who are able to afford it free to live their lives, while those who cannot are left in prison to await their trial. In fact, the median income of those who cannot afford their bail is $16,000, compared to $33,000 for those who can [28]. To compound this issue, even though the number of annual arrests has decreased since the 1990s, the percentage of incarcerated people who have not been convicted of any crime has increased, as has the average amount of time spent in pretrial detention. We have also seen a significant decrease in the number of defendants released on their own recognizance (meaning no bail is required or paid), or who are denied bail altogether [28]. This indicates that the greatest increase in people who are held pretrial are from those who are unable to afford bail. Matt McLoughlin, co-founder of an organization called the Chicago Community Bond Fund (CCBF) - which raises money for people unable to afford their bail - states "[o]f the 88 people who we paid bond for whose cases have completely resolved, 20 were not convicted of anything. These 20 people spent a combined 2,946 days in the jail before CCBF posted their bond — an average of 147 days per person" [33].

In addition to this, those who do secure their release may still be required to pay recurring fees that cover electronic monitoring. Here in Utah, one such advertised service - Onpoint Court Services - charges $100 setup fees along with a $10 to $12 daily service fee for most types of monitoring [32]. For a defendant potentially waiting months for their trial, this can add up to thousands of dollars in fees. And none of this takes into account the effects that incarceration and pretrial detention has on individuals, families, and even the economy.

2.1.2 The Negative Effects of Pretrial Detention

As stated previously, an estimated 460,000 people are incarcerated on any given day that have not been convicted of a crime. While some of these have been denied bail, an estimated 90 percent are unable to afford bail [28]. While the effects listed above specifically refer to those serving sentences following a conviction, numerous studies have shown that even a short pretrial detention can have severe consequences as well. These can range from losing a job, or facing loss of wages at work, losing homes or vehicles, or even custody of children. It has been stated that consequences such as these are simply that: the consequences for breaking the law. But, according to United Nations Office on Drugs and Crime, "family and employment circumstances are key factors in accounting for desistance" where desistance refers to offenders "maintain[ing] crime-free lives." [12]. If, as a society, we are truly interested in reducing crime rates, we should be taking steps to ensure we do not strip people of the tools needed to be a productive member of society.

According to Alexander and Kristi Holsinger, nearly 60% of those detained for three days or more suffered a job loss or change. This number jumps to 76% when including those who kept their job, but faced consequences from their employer [24]. Financial and residential stability were also negatively affected as incarceration time increased. These statistics show just how important it is to adhere to the timelines as every unnecessary day in jail can have a significant negative affect on the lives of these people.

Finally, Worden and Clark state that "[r]ural counties' geographic, demographic, socioeconomic, and political characteristics have the potential to shape the way their courts function" [55]. We can see this in the way Utah's own rural counties function. For example, the Justice Court in Carbon County, which has a population of just over 20,000, only meets weekly on Tuesdays, with check-in required by 9:00 am [11]. This means that someone who is arrested and booked that same day may not go before a judge until a week has passed, and these hearings are often where information or charges are filed and bail determined.

The inability to afford bail in conjunction with the negative effects of pretrial detention described above leads to a significant rate of people taking plea deals instead of waiting for their trial [21]. These plea deals often have the desired effect of getting the individual out of prison, but leaves a permanent conviction on their record and also leads to higher rates of recidivism. And, considering that less than five percent of arrests are for violent offenses, we thus end up with

hundreds of thousands of indigent and minority people suffering life-altering events for low-level offenses [52].

2.2 Criminal Justice Algorithms

To address these biases, a number of people and institutions have turned to machine learning algorithms. Currently, there are two main categories of criminal justice algorithms in use. The first is predictive policing, which goes by names such as PredPol, currently in use in parts of Utah, or HunchLab. These are used to determine where to send police officers on patrol, identifying "hot spots" and tracking officers and vehicles in real time to ensure they hit these areas. The other set of algorithms are referred to as Risk-Assessment Tools which are utilized in multiple stages of the criminal justice system. A well-known implementation of this is called COMPAS, while the lesser-known PSA, or Public Safety Assessment, has been implemented throughout Utah. Used to some degree in nearly every state, these algorithms purport to predict recidivism, flight risk, and even the optimal bail to set for a person awaiting trial.

The idea behind utilizing these algorithms is that computers are unbiased and objective arbiters of truth, but the reality is that they learn based on the data we give them. Jennifer Lynch of the Electronic Freedom Foundation further explains, "... the data used by predictive policing algorithms is colored by years of biased police practices ... [which] means it will continue to predict crime that looks like the crime we already know about" [26]. And therein lies the problem with any machine learning algorithm. If we train it with data that shows police have always more heavily policed poorer and minority neighborhoods, which leads to more arrests in those areas, predictive policing algorithms will send more patrols there. Likewise, if judges regularly grant far higher bail to people of color, a risk-assessment tool trained on that data will "learn" that people of color are higher risk. In addition to these fears, there is also the fact that many of these algorithms are completely black-box, their inner working known only to those who sell them to law enforcement and the courts.

2.2.1 Algorithmic Transparency

Both PredPol and HunchLab utilize proprietary algorithms. PredPol in particular has stated that its algorithm is "complicated for normal mortal humans" [22]. However, in a 2016 article, researchers determined the algorithm could be boiled down to a sliding window, or moving average [29]. By further testing

this algorithm on real crime data, the authors find that contrary to the company's claims, this simplistic model reinforces the biases already present. Their analysis shows that "black people would be targeted by predictive policing at roughly twice the rate of whites" and this bias repeats for low-income groups. Frustratingly, we do have an idea of what data is being used to train this algorithm and it is messy and incomplete. Andrew G. Ferguson emphasizes this point in his 2017 article, stating that:

> "Crime data is notoriously incomplete. Certain crimes like murder, burglary, and auto theft tend to be consistently reported to authorities, while other crimes like sexual assault, domestic violence, and fraud tend to be underreported. Some communities ... simply decline to report crimes. The Department of Justice has reported that half of crimes with victims go unreported" [17].

In addition to this, Ferguson also points out that crime statistics and police reports in multiple jurisdictions have "been shown to be inaccurate, misleading, and occasionally fraudulent." Furthermore, a significant majority of jurisdictions have fewer than 24 officers, meaning the datasets generated by their interactions with the community will be quite small. And given small datasets, this means that the algorithm trained on it will likely be unable to generalize to real data and the real harms caused by current practices will continue to be perpetuated.

COMPAS is another well-known criminal justice algorithm. A risk-assessment tool, it was widely derided by a ProPublica article in 2016 which claimed the model was racially biased [2]. In response to these claims, a number of researchers performed their own assessments. In a 2018 article, researchers at Berkley, Harvard, and Duke, determined that the accuracy of COMPAS could be matched with a simple if-else block, based only on age, gender, and prior convictions, compared to the 137 features COMPAS purports to use [1]. In a separate study, researchers opted to crowdsource the likelihood of recidivism and found that random, untrained strangers on the internet were better at prediction than COMPAS [15].

As stated above, PSA is the risk-assessment tool currently in use across the state of Utah. According to the Utah courts website, the PSA report is determined when a person is booked into jail [49]. This report is made available via Xchange (a database of court information accessible by subscription) for the life of the case, and is accessible to both prosecution and defense. Unlike COMPAS,

the creators of PSA have made the risk factors and associated weights publicly available [41]. However, similar to COMPAS, the accuracy of the algorithm leaves much to be desired. While data is currently being collected within Utah and research conducted to assess the validity and usefulness of the algorithm, results from Cook County, Illinois indicate it may be overstating the risk of many defendants: "[b]etween October 2017 and December 2018, 99 percent of people flagged as high risk for violence who were released before trial were not charged with any new violent crimes during the release, a percentage virtually identical to the one for those deemed low to moderate risk" [13]. In a New York Times op-ed, researchers point out that the best accuracy would come from predicting "that every person is unlikely to commit a violent crime while on pretrial release" [4]. Until the data being collected is made publicly available, it is possible that similar rates of unnecessary pretrial detention are occurring here as well as in Illinois.

Similar to the desire for the data behind and because of these algorithms, it should also be transparent where they are being used. In a 2019 Vice article, it was revealed that PredPol was being secretly tested in localities across the country, including in South Jordan, Utah [22]. As Andrew G. Ferguson states in his article "Policing Predictive Policing,"

> "[T]he criminal justice system has eagerly embraced a data-driven future without significant political oversight or public discussion. Worse, the temptations of new technology have at times overwhelmed considerations of utility or effectiveness and ignored considerations of fairness or justice" [17].

The use of these algorithms in any setting should be cause for concern and further study. We should not discard fairness and justice for convenience, but should actively work towards fully transparent data collection that allows us to see the systemic issues and work proactively to effect change.

Bilel Benbouzid, in his attempts to recreate the PredPol algorithm (which is based on earthquake prediction algorithms), points out that "[c]ontrary to the seismologist, police officers cannot experience 'failed' predictions ... the seismologist conceives prediction in terms of its practical consequences, and the developers conceive it in terms of an absolute duty to act" [7]. In my own favorite misquote of a popular adage, "don't just do something, sit there," acting may seem like the correct response, but without taking the time to thoroughly examine the problem with all of the data available, any action taken is likely to

exacerbate, rather than correct the issue.

2.3 Why We Need Transparency

So far, the majority of the statistics cited in this proposal are couched in careful terms - "approximately," "estimated" - and this is not by accident. As Alice Speri states in her 2019 Intercept article, "we know remarkably little about who is arrested, where, and why" [39]. This is largely because the 18,000 law enforcement agencies within the United States voluntarily self-report their data to the FBI and Bureau of Justice Statistics (BJS), and what they do report is not consistent from one agency to the next. Similarly, in a Politifact article, House Representative Ted Lieu (D-CA) was fact-checked when he stated that more than 450,000 people are held because they can't afford bail. This was Politifact's analysis of his claim:

> "Ultimately, we found that the number in Lieu's tweet is an overestimate of the number of Americans who are in jail because they can't afford to pay bail. In addition, the statistic Lieu is referring to is not one that is kept nationally. In other words, no accurate count exists of the amount of people in jail in America who are too poor to pay bail, so there's no way to fully verify Lieu's claim" [20].

Similarly, the Vera Institute of Justice, which aggregates statistics from across the country, estimates arrests for 2016 to be 10.6 million, yet only 8.9 million were actually reported [52]. Building on this is the fact that not all interactions with law enforcement are recorded. Stop and frisk tactics are still used in many parts of the country, but, as Dean Knox, a professor of politics at Princeton states, "[t]he vast majority — 99.9 percent of the data — we never get to see ... [w]e just don't see all the times when police officers are encountering civilians on the street. And that's a huge problem, because among the data that you do get to see — the stops, and the arrests, and the use of force that officers record — those are already contaminated, because officers have discretion in who they choose to engage" [9]. In an article co-authored by Knox titled "Administrative Records Mask Racially Biased Policing," the authors state:

> "We show that when there is any racial discrimination in the decision to detain civilians — a decision that determines which encounters appear in police administrative data at all — then estimates of the

effect of civilian race on subsequent police behavior are biased absent additional data and/or strong and untestable assumptions" [27].

This lack of data prevents accountability. We cannot identify if police officers are disproportionately pulling over people of color, or judges are consistently setting bail higher for indigent defendants. We cannot know what biases we may be training into criminal justice algorithms, and we cannot know if harms against minorities are being exacerbated. It is imperative that data be collected and made public so the people can hold those in power, as well as the algorithms used by them, accountable.

3 Utah's Current Transparency Laws

Initial research into the laws that currently govern transparency within Utah turned up very little. The primary transparency law is referred to as GRAMA (Government Records Access and Management Act) and is the state's own version of the Freedom of Information Act (FOIA). From GRAMA's official website, "[g]overnment records belong to the citizens of the state, who have a legal right to open and fair access" [47]. The primary concern with this statement is that the records and data the government is collecting are too few or not easily accessible. As emphasized in the previous section, without complete, comprehensive data about all aspects of the state's criminal justice system, there is no way to accurately determine how bias affects it, or if due process is being followed. Even HunchLab, in an unreleased white paper titled "Using Data to Reduce Policing Harms," argues that we cannot effectively minimize harms unless we have the data to show what harms are being committed [46].

3.1 HB 288

More recently, in March of 2020, House Bill 288 was passed which mandates that the local jails, prosecutorial agencies, and the AOC collect specified data and submit it quarterly to the Commission on Criminal and Juvenile Justice (CCJJ) on a quarterly basis [23]. The requirements laid out in the bill are to go into effect beginning January 1, 2021, which is after the research for this thesis will have concluded. It is important, however, to note how the state is approaching data collection and whether these changes will affect the outcome of future research in this area.

The four agencies represented in this bill all have separate reporting requirements and responsibilities.

3.1.1 County Jails

Each county jail is required to collect the following information for all bookings:

- full name
- offense tracking number
- gender
- date of birth

- race
- ethnicity
- zip code

The offense tracking number refers to a single offense "that requires a mandatory court appearance and for which an individual is booked into a jail facility." This means that an individual, booked into jail under multiple offenses, will have multiple offense tracking numbers. It should also be noted that the booking date and time - critical for determining if probable cause and information were filed in a timely manner - are missing from this list.

3.1.2 Prosecutorial Agencies

The prosecutorial agencies are required to collect and report the following information for each case they oversee:

- full name
- offense tracking number
- date of birth
- zip code
- referring agency
- whether the prosecutorial agency filed charges, declined charges, initiated a pre-filing diversion, or asked the referring agency for additional information
- if charges were filed, the case number and the court in which the charges were filed
- all charges brought against the defendant
- whether bail was requested and the amount
- date of initial discovery disclosure
- whether post-filing diversion was offered and, if so, whether it was entered
- if post-filing diversion or other plea agreement was accepted, the date entered by the court

- the date of conviction, acquittal, plea agreement, dismissal, or other disposition of the case

From the law, a pre-filing diversion refers to "an agreement between a prosecutor and an individual prior to being charged with a crime, before an information or indictment is filed, in which the individual is diverted from the traditional criminal justice system into a program of supervision and supportive services in the community." Post-filing diversions are much the same, though they take place before conviction and after charges have been filed.
Here we note that while each offense has its own tracking number, there is nothing aside from full name and date of birth to identify an individual, and nothing to identify an individual booking or grouping of charges. Zip code is mentioned, but it is not specified whether this is the zip code of the individual's place of residence or of where the offense was committed. These agencies are also required to publish online specific policies, including screening and filing criminal charges, sentencing recommendations, and discovery practices.

3.1.3 Administrative Office of the Courts

For every criminal case filed with the court, the AOC shall collect and report the following:

- case number
- full name
- offense tracking number
- date of birth
- charges filed
- initial appearance date
- bail amount, if any
- represented by public defender, private counsel, or pro se
- final disposition of the charges

3.1.4 Commission on Criminal and Juvenile Justice

The above agencies all submit quarterly reports with their respective data to the CCJJ. The CCJJ is responsible for using this data to publish annual reports. The bill lists a total of 23 responsibilities for this agency. A primary focus is to "promote research and program evaluation as an integral part of the criminal and juvenile justice system." The intent of the submitted information is that it will be used to find ways to reduce recidivism and to see where programs have been effective in accomplishing these goals. They are also charged with promoting communication and coordination of all the criminal justice agencies and developing information systems with common standards that will make sharing this data easier.

The CCJJ is also responsible for studying other jurisdictions where there have been successes in these areas and make recommendations for adopting them as appropriate. It is also charged with performing annual audits of the criminal history information they collect for completeness and accuracy.

3.2 Analysis

Several concerns were addressed above, but the primary concern comes back to what the data is being used for. The bill, as written, appears to address who owns the data, but not necessarily its use. Having a tracking number for each offense is reasonable, but tracking numbers for each individual and booking would be able to answer questions such recidivism for an individual, or how many charges are laid per booking based on age, race, or gender. Individual tracking numbers would also be ideal to prevent misidentification. In a white paper, three authors ran the mathematical probability of collisions (matching names and dates of birth) within a population [5]. In their conclusion, they point out that only 8.3% of the population is at risk of misidentification using these identifiers, which may seem a reasonable rate of error, but as mentioned in the previous section, the effects of even a few days of pretrial detention can have severe and long-lasting effects on an individual. This means 8.3% is an unacceptable risk and should be mitigated with alternative forms of identification. In addition to these potential issues, the CCJJ's responsibility to promote communication and collaboration as well as developing common data standards among the agencies naturally lends itself to the idea of a state-wide database. As will be discussed in later sections, more rural counties often lack the resources and funding to create or maintain criminal justice databases. In fact, many of

the local counties we surveyed as part of this project appear to be using manual entry for many of their records, which is prone to errors. A central database would help to promote the collaboration necessary to maintain accurate and up-to-date records and mitigate the overhead required for the individual agencies to generate these reports.

4 Project Description

4.1 Goals

The primary goal for this research project is to effectively measure transparency within the state of Utah. This is accomplished by scraping booking data from county jail websites - where available - and from this gathered information, attempt to determine if due process is being followed. Specifically, we attempted to identify counties that are not filing probable cause affidavits or charges within the required time frame of 24 hours and four business days, respectively.

If either of these do not occur within the requisite time frame, the person must be released on their own recognizance, meaning no bail is to be paid. Instead, the person only has to promise, in writing, that they will return for their court date. Several additional points of interest in the process have been identified and are highlighted in red in Figure 1, provided to us by Brittany Urness, a Legal Volunteer for the Smart Justice Campaign at the ACLU of Utah. These statutes can be found in the *Utah Rules of Criminal Procedures* [50], with our particular focus being Rule 9.

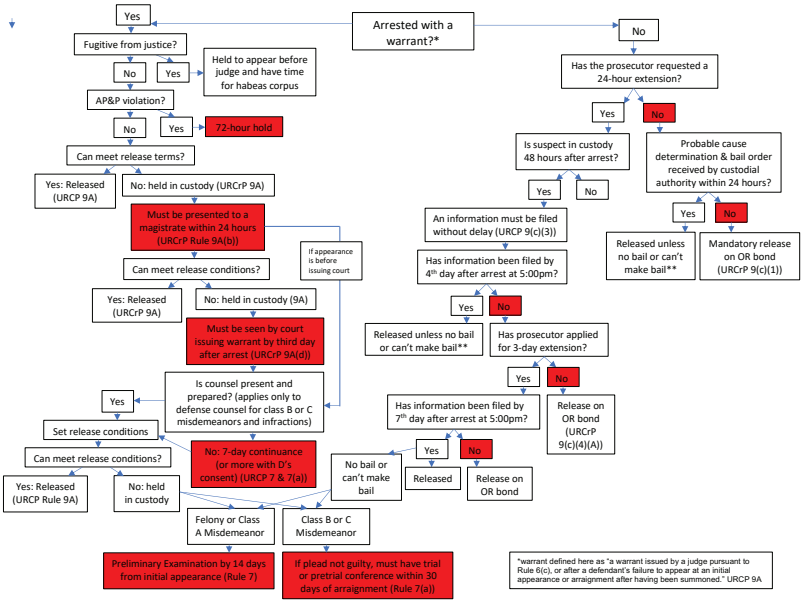


Figure 1: Detention Flowchart.

To answer these questions, we first ask the following:

1. What data is needed to answer the question?

2. How accessible is this data?

To answer the first question, we can use the information above. If probable cause has to be filed within 24 hours of booking, we need to know the exact date and time of the booking as well as the probable cause filing. As information - or charges - have to be filed by 5:00pm of the fourth business day following a booking, we also need the date and time this was filed, as well as the extension, if any. Along with this, we need some sort of unique identifier that can join these together, as booking data is maintained by the jails and filing information is handled by the AOC. In addition to this, there are other data that may be interesting to researchers and legislators, such as race, gender, age, zip code or city, the booking charges and bail or bond for each charge, and some way to identify people who have been booked multiple times. The second question is less easily answered. For a computer scientist, creating a web scraper that can port all information into a database is a relatively minor task, though still constrained by the availability of both website, data, and time. Likewise, for a lawyer or other members of the criminal justice system, accessing the appropriate records through the courts is only a matter of time and potentially money. To that end, for each county in the state of Utah, we have examines how easy it is to collect booking and court data. For the counties that do have websites, we analyzed the amount of information available, as well as its accuracy and ease of access (paywall, anti-scraping tools, page design, etc.). From this point, we determined how easy it is to collect further information regarding an individual's progress through the court system. Where these datasets are disjoint, we determined how difficult it is to combine the booking and court information. In addition, we continually monitor the websites we are scraping to see if the information displayed changes and in what way. The loss or addition of publicly available information will also factor into the transparency evaluation.

4.2 Design

In order to perform this analysis, we began by scraping the booking websites. Because each website is different, every county has its own scraper. This scraper parses the website daily and deposits the gathered information into the appropriate database tables. A separate analysis script for each county is run, usually

on a weekly basis, to detect inmates who have been in for longer than four days and who are not serving a sentence. We decided to use this very loose metric, as the websites often list the arrest charges, but these are not the same as the charges that need to be filed within the four-day timespan. This also helps keep the cost of additional information low, since GRAMA requests are time-consuming and the AOC charges a fee to access records. This is particularly important for counties such as Salt Lake and Weber which have thousands of inmates on any given day. When an inmate is flagged by this process, the script generates an email with each flagged inmates information which is then sent to the Smart Justice Coordinator at the ACLU of Utah.

Using this list of flagged inmates, the ACLU of Utah submits a request for further information from the AOC. This dataset contains information from both the District and Justice Courts for that county. The District Court oversees felonies and class A misdemeanors, while the Justice Court handles all other charges. Parsing this information, we can determine when charges and the probable cause affidavit was filed for each person and pair it up with the flagged people listed in our database.

Like the scraper and the database, this parsing is county-specific as well. At this time, we have only recently received our first batch of data for Beaver County, which contains multiple ways these events are noted, making it difficult to easily identify where issues may lie. For example, probable cause is listed as "affidavit/declaration," "affidavit/declaration of probable cause," and "probable cause affidavit." For other counties, it is possible this has even more variations.

As it takes some time to get the data necessary to do a full analysis from the AOC, this is an historical analysis. It is our hope that the results of this thesis will highlight the need for transparency, as it should not take three to six months just to determine a person has been held beyond what the law dictates.

4.3 Limitations

The main limitations of this research project included the time and money required to collect the data that we need, as well as in the actual measurement of transparency and how to present this measure in a way that is easy to understand.

5 Implementation

Prior to writing any code, we went through each county in Utah and attempted to find a booking website, or county jail inmate roster. Currently, out of 29 counties, there are eight without booking websites, though Daggett county does not have a jail to build a website for. Out of the remaining 21 that do, we identified five that looked promising and wrote scrapers for them. These include Beaver, Davis, Tooele, Utah, and Weber. We also chose five others to perform a thorough assessment based on the above discussion as well as more detailed parameters, which are listed below. We begin this discussion with the details of our implementation, followed by the analysis of the various websites. For the ones that were scraped, we discuss the issues and roadblocks we encountered, particularly in the context of transparency.

5.1 Web Scraping

As stated in the previous section, due to the varying nature of each county's website, the implementation of this project covers only the high-level design, which has changed as the project has progressed. The initial goal for the project was to collect the booking data, flag any inmates that were incarcerated for more than four days, and forward these names to the ACLU of Utah who would then determine if information was filed within the requisite timeframe. To this end, we wrote individual scrapers for each county, including the code to collect and forward the list of names and information for flagged individuals to the ACLU of Utah on a weekly basis. Since each scraper was separate, there was also a separate Cron job that would run each one. This data was collected and stored in XML format.

One of the soft goals of this project is to make the project relatively easy to shift to additional counties and states. Because of this, the separate scrapers and storage in XML wasn't viable for long term. To that end, we designed an overarching paradigm, built largely on the work we had already completed, which focused on database storage and a single master file running all of the scrapers together. This current design is described in detail, beginning with the master file.

5.1.1 Master Scraper

A single master file runs each of the individual scripts. This file utilizes a config file that contains the name of each county with a current script, the filenames of the various scripts and possibly ReadMe files (created to document shifts in data collection due to website changes), and the email addresses of the Smart Justice Coordinators as well as the code developers. The master script is run as part of a daily Cron job, and utilizes try-except blocks when running each county-level script. If a county script fails to run, the master file notes the failed script and creates a separate Cron job for it that runs every 15 minutes until it is successful. This is to account for temporary website outages, or to catch when the scraper fails due to a change.

Once all the county-level scripts have run, the master script emails the developer with the output, including any failed scripts. This is due to using the school's computers. If the computer being used is turned off for any reason and the master script not run, the developer is alerted by the lack of an email. While it would be preferable to only receive emails when something fails, we have had numerous instances in the past where the computer running the scrapers has been turned off and our only indication is the lack of a success email. For future expansion and robustness, a more secure and reliable server would need to be engaged.

5.1.2 Scrapers

For the scraper and analysis code, we are using Python 3.7. For the scraping in particular, we are using the BeautifulSoup package which allows for relatively easy parsing of html. On a daily basis, each scraper collects the data for any new inmates or bookings, and updates the information for existing inmates. In particular, this includes whether they are still incarcerated. Anyone who reaches the threshold of more than four days incarcerated is flagged. This is done by either adding the names to a text file as they are flagged, or doing a database search for flagged inmates. The second method requires an additional database column that checks whether the individual's information has already been forwarded. Other considerations for the scrapers include what is being displayed on the website. As we will see below, the various websites display different information. For instance, Beaver County does not have an inmate roster, but a complete list of bookings for the past 30 days. This means that someone who is in custody for longer than this will drop off the website, even

though they are still incarcerated. The Weber County website uses an inmate roster which is updated frequently. Unlike Beaver, where the people are marked as in custody, a person on Weber's site is automatically in custody. It is only when they drop off the website that we can assume they've been released. Thus for Beaver, we only update incarceration time when we see an inmate that is already in our database. For Weber, we have an additional column for "last seen" which is what is used to determine the incarceration time. In tandem with this, while Beaver has a "status" column, that shows whether or not they are released, Weber only has the "last seen" column, which can be interpreted as the day of their release. Lastly, for our purposes, we collect all information displayed on the page with the exception of the booking photo. This is largely due to the storage costs, but also because it is not necessary to answer any of the project's questions.

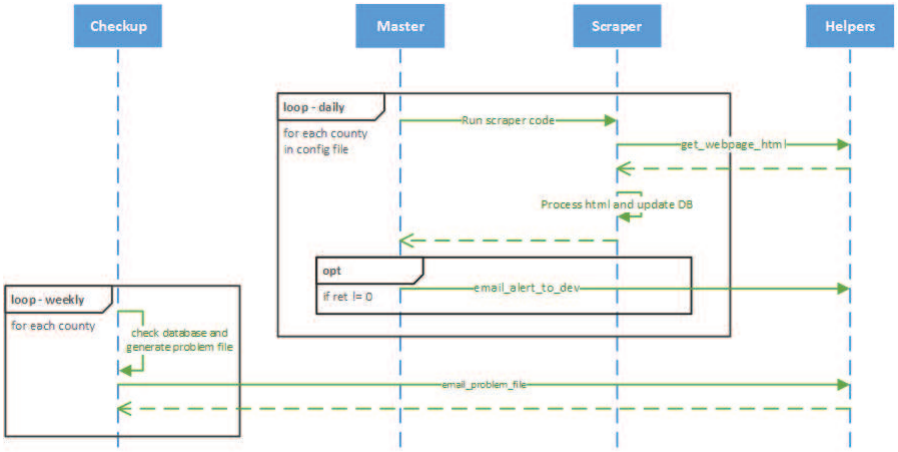


Figure 2: Scraper and Analysis Sequence Diagram.

5.1.3 Database

When the project first began, instead of using a database to store our scraped information, we opted to use XML. This proved to be quite cumbersome, in particular due to having to check all previously stored inmates and update their information as appropriate. For Beaver County, this was not an issue, as they average one booking a day in any given 30-day period and to date have fewer than 300 people in the database - not accounting for duplicates. Weber County, on the other hand, hosts approximately 1,000 inmates on any given

day. Currently, the Weber County scraper takes nearly 20 minutes to run. Along with parsing and storing the data in the database, we also have to create unique scripts to port the XML data to the database.

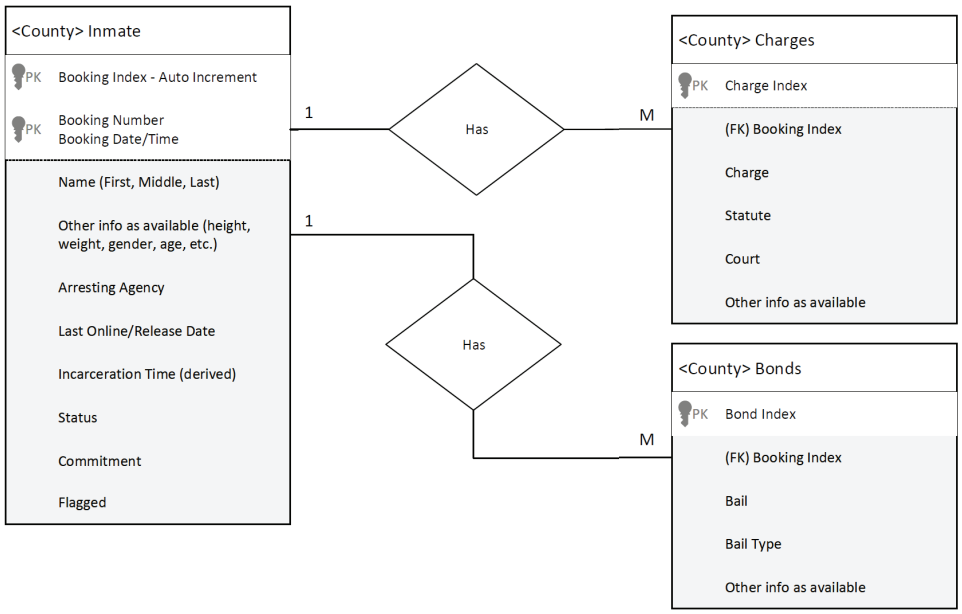


Figure 3: Database Schema.

5.2 Initial Analysis

To analyze the websites, we examine at the following:

1. Is the website easily accessible?
    - Specifically, can it be located with a search engine? Is a log-in required for access, or is it protected by a paywall? Is there a way to see a complete list of inmates or bookings?
2. What information is displayed?
    - Is booking date and time displayed? Is there a way to identify people uniquely, preferably with an individual identifier, or with a booking number and date and time? Is demographic information, such as gender, age, and race, included?

3. Is the information accurate?

    - Are there discrepancies in spelling for entries such as locations and charges? This indicates that manual entry is likely being used, leading to errors in the displayed information. It also indicates that other fields that aren't able to be verified may also be inaccurate, such as name and booking dates and times.

4. Is the data on the website easily scraped?

    - Are there features of the website that make it difficult to scrape the data? Particularly, does it require Captcha entries, or use JavaScript to display information? Or is the information displayed in a format that is difficult to parse, such as PDF or variable text? Scraping is simplified with the use of class and id names for html elements, as well as with clear links when more information is listed on a separate page.

5. How does the website address privacy?

    - This is nearly always the antithesis to the previous question, as anti-scraping measures help ensure the privacy of the people listed. In recent years, a number of predatory sites have been created and used the ability to scrape booking data and collect mugshots to populate their websites [38]. For those who want their mugshots removed, payment is usually required. In one such case, the website required $399 for each charge listed in order to remove the mugshot for a man who had 90 charges, all of which were dismissed. Thus, while preventing scraping can negatively affect transparency and accessibility, it does help protect the privacy of booked individuals.
    - In addition to this, there exists a law in the Utah Code, 17-22-30, which restricts the distribution of booking photos to these predatory sites, known as "publish-for-pay" websites [51]. Likewise, there are requirements for removing booking photos from these sites following a request from the individual, for free within seven days if the person was not convicted, among other situations, or for no more than a fee of $50 and within 30 days otherwise. This law does not address if the publish-to-pay site collects the photos from a publicly accessible website.

6. Are there other features or supplemental data provided by the county that affect the available data, transparency, or privacy?

County	Accessibility	Demographic Information	Accuracy	Scrapable	Privacy
Salt Lake	✓	✓	✓	x	✓
Sanpete	✓	x	✓	x	✓
Washington	✓	x	✓	✓	x
Morgan	x	x	✓	x	✓
Cache	✓	x	✓	x	✓

Table 1: Analysis Overview of Selected Websites

5.2.1 Salt Lake County

1. Salt Lake County uses an inmate roster that is easily found through a search engine [35]. The main page has the option to search by name, booking number, permanent number, or by state ID. All of these return the full list of inmates, in order by the lookup method, and in groups of 30.

2. Clicking on a name brings up a new tab with detailed information, far more information than any other website displays. It includes gender, age, weight, race, hair and eye color, as well as the arrest charges and bail amounts. Booking date is included, however, booking time is not. This may not be as problematic as other counties, as each charge has the case number and offense date combined with the charge details. This may not seem pertinent as we have not yet discussed the data from the AOC, but this data will ensure when linking the two data sets that we have the correct join.

3. The information displayed on the website appears to be accurate. There are no obvious spelling or typographic errors, though a test inmate has been left in the database and comes up at the top of the list when performing an empty search on booking number or permanent ID. While it is relatively obvious this is a test entry in the database, and not an actual inmate, other counties have not made their test entries as clear.

4. The website relies heavily on JavaScript both to execute the search queries and access the individual profile pages. Once on the detailed page, it is

possible to navigate through a series of tables with the information. While lacking class names or IDs for the elements, both label and data are table data entries in a consistent order. This is not ideal, but would make scraping the individual page (once accessed) relatively easy.

5. While Salt Lake County does have enough information on its inmate roster that individuals could potentially be identified, the limitations to scraping presented by the use of JavaScript help keep the information secure from scrapers. There are solutions to this, but most are large-scale industrial solutions, such as robotic process automation (RPA), that are far more costly than the average person could afford [8]. Also, it does not display booking photos, preventing their use or display on predatory websites.

6. Just last year, Salt Lake County released a public dashboard that has multiple pages of visualizations of the inmate population [34]. This includes breakdowns of the population by race, gender, booking agency, as well as the total population over time. This anonymized data is readily accessible with interactive charts and is updated daily. This approach showcases how relevant information can be made transparent without sacrificing the privacy of the inmates. Overall, as a candidate for web scraping, Salt Lake County ranks very low. But, even though scraping is difficult, this is largely negated by the dashboard which gives statistics about a number of questions researchers might have. It still fails to answer the due process posited by this research, but does far more than any other county in terms of transparency.

5.2.2 Sanpete

1. Sanpete County booking website appears in search engine results and posts bookings on a weekly basis, with data going back a full year [36]. However, these postings are PDF files, with no way to search or see who is still incarcerated. As these are linked files, it is also possible to go back further than a year, as a standard naming convention for the PDF files is generally used.

2. The displayed information includes a booking photo, name, city of residence, arrest date (which may not be the same as the booking date), along with charges and bail. The arrest location and agency are also included. Starkly absent is any demographic information or the booking date and

time. No unique identifiers are provided for either the booking or the individual, which will make this information difficult to join with AOC data.

3. It is not immediately apparent if the information is accurate. No apparent spelling or typographic errors were detected and the consistent generation of the PDF files indicates that manual entry is unlikely to be used.

4. While accessing the booking data is not difficult, particularly as each file is listed as an href element with all of the names beginning with "files," parsing data from a PDF file is notoriously difficult. There are solutions, similar to the RPA mentioned above as well as others that are strictly dependent on the consistency of the documents. For going back further than the dates listed on the main page, there is some variance in the file names of the booking reports, with the majority named with the date, such as 9-30-2019.pdf, and others given names such as Dec_2_booking_report.pdf, or Dec_4_2019.pdf.

5. Intentional privacy considerations do not appear to have been made, particularly with the ability to access files that predate the displayed weekly bookings. However, due to the difficulty in scraping the booking data, as well as the booking photos, the website has decent privacy protections.

5.2.3 Washington

1. Washington County also uses a booking website which can be found with a search engine [53]. It is publicly accessible and lists all of the bookings for the past five days on a single page.

2. The information listed for each person includes a booking photo, their name, city and state, PCF number, arrest date and time, and the arresting agency. To reiterate, arrest date and time is not the same as booking date and time, which is what the time constraints for filing probable cause and information are dependent on. Also included is the list of charges and whether they are still incarcerated. Similar to Sanpete, there is no demographic information included.

3. There is a significant level of consistency in the charge descriptions, and no obvious spelling or typographic errors.

4. The design of the page is fairly straightforward, with each booking given its own table nested within a div element with a class name. Each of the table data elements also have class names, such as name, picture (for the booking photo), or charge. Each charge can be clicked to give more information, such as the statute, classification, and bail. Unlike previous sites, clicking this simply toggles visibility of another element which is also given the class name "chargedetails" and can be scraped without having to click. All of these features make the website fairly straightforward to scrape. The only mitigating factor is the short timespan the bookings cover, making daily scrapings of the website necessary.

5. While the website is simple to scrape, the limited time frame and lack of demographic details provides a small degree of privacy to the people displayed. However, the easily scraped website, combined with name, picture, and home city provide little in the way of privacy.

5.2.4 Morgan

1. The Morgan County website was difficult to find, and did not come up in search engine results [31]. We were able to locate it by following a link through a blog post on WordPress [30]. There is a search, but it requires a minimum of two letters for the last name and one letter for the first name before executing a query. In addition, this does not appear to be a traditional booking or inmate roster site, as it includes people both on probation and parole.

2. Once a successful query is executed, a table with all of the results is displayed. Each row contains the offender number, name, sex, and date of birth. A row can then be clicked which brings up a statistics tab with more details. This includes the same information as on the search results tab, as well as height, weight, location (such as parole or probation), housing facility, parole date, and a list of aliases. There are no photos and booking date and time are notably absent.

3. The displayed information appears to be accurate, with no apparent spelling or typographic errors.

4. The website is designed as a front-end for a database. Everything, including the initial search, is executed with a jQuery script, making scraping

nearly impossible without industrial-scale tools. This is in addition to the limitations of the search, which requires a minimum number of letters for both the first and last name to execute the query. At a maximum, this would mean more than 17,000 queries. Even minimizing based on valid beginnings (avoiding combinations such as "qz," for example) would still leave a considerable amount of searches. In addition to this, the search executes not only on the name, but also all aliases, so there is a strong possibility for overlap in the search results. Once the information is displayed, there are elements labeled with IDs, but getting to this point would take a significant amount of work.

5. Though the website provides a significant amount of information about each person, it is not only difficult to find, but also incredibly difficult to scrape due to the abundance of jQuery scripts. The limitations with the search query also provide additional privacy.

5.2.5 Cache

1. Cache County has both booking data for the past thirty days as well as an inmate roster that can be found via search engine [10]. The data is publicly available with no log-in or other requirements.

2. Both the booking website and inmate roster are near identical in their initial display. A single table spanning multiple pages, with 15 rows to a page, is shown with a date and time and first and last name. On the inmate roster, middle name, gender, and age are also displayed. If a name is selected, further information is shown. On the booking page, arrest date and time, name number (likely a unique identifier for each person), age at arrest, agency, related incidents, bail, and offenses are displayed. On the inmate roster, booking date and time (which corresponds to the date and time displayed in both tables), age, gender, height, weight, hair and eye color, bail, and the offenses are shown. These offenses may differ between the two sites. Booking photos are not shown on either site, but there is an option to download them. Clicking on this brings up a statement that requires the person downloading it abides by the Utah code mentioned earlier regarding publish-for-pay websites. To get the download, the person must enter their email, signature, and complete a Captcha.

3. There are no obvious spelling or typographic errors, aside from the inmate

roster listing the booking date and time as "Bookied Date/Time." There are some discrepancies between listed offenses on the different websites, but these include listing a warrant as the charge on the bookings website, and the actual charges on the inmate roster.

4. Both websites are similarly designed, using JavaScript to display the additional information for each person. Likewise, JavaScript is used to move through the pages in the table, making it difficult to get more than the first page. This may not be problematic as for the purposes of this research, only booking date and time and unique identifying information is necessary. With the consistency of information between the two websites it is possible to grab the requisite information, though it may require multiple scrapings a day to ensure no bookings are missed. Then, in conjunction with the inmate roster, we can determine length of incarceration.

5. As explained above, it is possible to gather the information needed for this research, as well as age and gender, but all other information is relatively protected with JavaScript. The requirement to submit a name and email to get booking photos, along with displayed links pointing to the the "publish-for-pay" law above, maintains a balance between publicly available data and privacy for the people who are booked.

5.3 The Scraped Websites

Scrapers were created for each of the counties below. As above, we will answer the previously stated questions, with the added discussion of problems or stoppages that we encountered. It should be noted that, unlike the websites listed above, the ones for which we created scrapers were chosen because they were easily scraped and had a reasonable amount of information available for data collection.

County	Accessibility	Demographic Information	Accuracy	Scrapable	Privacy
Beaver	✓	x	x	✓	x
Davis	✓	✓	✓	x	✓
Tooele	x	✓	x	✓	✓
Utah	✓	✓	x	x	✓
Weber	✓	✓	✓	✓	x

Table 2: Analysis Overview of Scraped Websites

5.3.1 Beaver

1. The Beaver County booking website is easily located with a search engine [6]. All of the bookings for the last thirty days are displayed on a single page.

2. Prior to March 18th, 2020, Beaver County displayed the full name, gender, age, home town and state, booking date and time, booking number, and BCCF number. Charges were also displayed with the classification, court, and statute. Following updates to the website - which included changing the web address - age and gender were no longer displayed, and neither was the booking number. The displayed middle names were replaced with a middle initial, and the BCCF was also relabeled as "Number" instead of specifying if it was the booking or BCCF number. Arresting agency was also added. Both iterations displayed whether the person was currently incarcerated, however, the original site also specified if they were released. The updated site simply removes the "IN CUSTODY" label when someone is released.

3. There are no apparent spelling mistakes on the website, but the accuracy of the displayed information was called into question when the changeover happened. Initially, we were using the booking date and time, booking number, and BCCF number as the key to identify a specific booking. The removal of the booking number forced us to omit it from the key. This was not as problematic as the appearance of duplicates in the database caused by discrepancies in the booking time. For example, according to the original site, one inmate was booked on March 4th, 2020 at 20:25:45. The updated website changed the time for this individual to 21:39:26 and thus our scraper recorded them as two separate bookings. All told, there were eight such instances. Six of them had differences of around two minutes, while the most egregious had a difference of nearly eight and a half hours. This may not seem extreme, but with a 24-hour requirement to file the probable cause affidavit, accurate timestamps are essential. It likewise calls into question the accuracy of all booking times and makes determining if due process was followed far more complicated.

4. The information on the web page is easily scraped. The information is displayed as a series of tables with one table for each booking and the HTML elements have class names. The charge details are hidden until

the charge is clicked, but this is done through an "onclick" event and visible in the HTML.

5. Though Beaver only displays thirty days of bookings, the information is easy to find and scrape. Booking photos are also displayed and can also be pulled from the website and there are no safeguards or privacy protections.

5.3.2 Davis

1. The Davis County inmate roster is easily located with a search engine [14]. The page also has a search by name option. Like Beaver, this website has changed since the beginning of this project, both in design and web address, but the displayed information has remained largely the same.

2. On the original web page, first, middle, and last name were displayed, along with age and gender. Booking number, date and time were also displayed, as well as arresting agency, housing unit, and charges. The charges held additional information, including the statute description, court, fine, and type. The updated website no longer displays the middle name or court, but does display the state statute for each charge. Booking photos are not displayed and individuals drop off the page when they are released.

3. The information displayed appears to be accurate, with no obvious spelling or other typographical errors.

4. The original web page was easily scraped, with table displays similar to Beaver above. The new website displays a table with each row containing the booking date, first and last name, gender, age, and a link for inmate details. This link, when clicked, shows an overlay on the page with the remaining information. Within the HTML of the page there is an element with an ID that holds this information, but the data within it is populated via jQuery request and is thus not able to be scraped with traditional means.

5. In addition to the difficulty posed by the design of the page, on the inmate details overlay is a "Terms of Use/Booking Photo Information" button. Clicking this drops down additional information, emphasizing that there are no guarantees on the quality of the displayed data. It also includes a link the Davis County Records Request page where a GRAMA request can be submitted for the complete records and booking photo. So while

the original design did very little to address privacy concerns, the new website balances the desire for public accountability and the privacy of booked individuals.

5.3.3 Tooele

1. The Tooele County inmate roster cannot be found via search engine [43]. It shares the main part of the web address with the Tooele County Sheriff's page, which can be found via search engine, but it is not directly linked on the page.

2. Full names are displayed, along with age, gender, height, weight, and hair and eye color. There are spaces to display charges and bonds as well, though more often than not, neither are populated. The booking date is displayed, but the only time that can be found is on the booking photo itself. The booking photo also displays a number, but this is different from the ID for each inmate within the HTML.

3. There are no apparent spelling or other typographical errors. The times on the booking photos, though they could potentially be collected, may not be the actual booking times. Likewise, the discrepancy between the number on the booking photo and the ID number in the HTML with neither being clearly identified can potentially be a cause of further discrepancies.

4. The Tooele website is quite similar to Davis, with a table on the main page that displays limited information, and a placeholder HTML element that populates with a jQuery request when a magnifying glass symbol is clicked. This is also an overlay and displays all of the above information. On the main page, nothing is displayed until the search button is clicked (and the name fields can be empty) and then the inmates are displayed in groups of ten. Only first, middle, and last name are displayed in these tables. As with Davis, this design makes scraping incredibly difficult, though we were able to identify a separate link that utilized a database query using the ID in the HTML. This returns an unformatted page with the individual and all the information displayed on the details overlay.

5. With the difficulty this website presents, both in finding it and with scraping it, Tooele County is helping to protect the privacy of the inmates. This does affect transparency, but as there are no blocks on the website (aside from knowing the address), it can still be regarded as publicly accessible.

5.3.4 Utah

1. The Utah County inmate roster can be found easily with a search engine [48]. Options are displayed to use an inmate search, see all of the current inmates, or to view inmate statistics.

2. Aside from Salt Lake County, Utah County appears to have the most complete information displayed. Included in the individual display are: full name, arrest date and time, arresting agency, booking date and time, booking number, release date (though this is generally blank), status (such as active or electronic monitor), height, weight, eye and hair color, gender, year of birth, and birth country. The charges are listed below, individually. Each one displays the court, case number, whether they are being held for that particular charge, bail, whether it is bondable, and a description of the charge. Booking photos are displayed, but only for thirty days following booking. Though both the web address and web design for this page have changed, the information displayed has been consistent.

3. In previous iterations of the website there have been issues with the accuracy of the displayed information. In particular, very few inmates have had release dates displayed, even though their status may have changed to something other than active. Additionally, Yogi Bear was once an inmate of Utah County Jail and while this was likely a test entry within the database, it does call into question the accuracy of other entries.

4. Previous iterations of the website were easily scraped, with the ability to collect booking numbers and append to a standard web address. This allowed scraping of each individual's information by simply accessing their web page. The new website, though publicly accessible, is virtually unscrapable. Clicking on the link for current inmates brings up a table with ten inmates per page. Name, year of birth, booking date, ID, and status are all displayed. Clicking on the name executes a JavaScript command that brings up a Captcha. The box stating "I'm not a robot" must be clicked before selecting continue. Once this is done, the individual's page is brought up. The web address for each individual is identical, indicating that the page is a place holder and populated by JavaScript. These make scraping the new site a near impossibility without commercial-grade solutions.

5. Utah County has done a commendable job protecting the information of its inmates. While it does mean we have not been able to restore the scraper for this site, and likely will not be able to, the privacy protections are top notch.

5.3.5 Weber

1. Easily located via search engine, the Weber County inmate roster is publicly accessible and shows all current inmates in a table on one page which can be filtered on first and last name [54].

2. On the main page the booking date, full name, gender, and age of each inmate is shown. It is ordered by last name and each entry has a link to the individual's details page. The details page contains the same info as the main page, as well as a booking photo, booking number, name, case number (though this is generally blank), height, weight, hair and eye color, and a list of charges and associated bonds.

3. There are no apparent spelling or typographic errors.

4. The data is easily scraped as the displayed rows on the main page are table rows and the web address for each individual's details is visible within the HTML element. The scraper can use the main page to access each of these web pages in turn and collect the data. The only major impediment is that the only visible time that could be collected is part of the booking photo.

5. Weber County appears to take no steps to protect the privacy of its inmates. When someone is released their information drops off, but as we have shown, it is incredibly easy to collect all pertinent information.

5.4 Beaver Court Data

In addition to the data collected from the websites, we worked with the ACLU of Utah to gather data from the AOC. The AOC allows for creating custom documentation, essentially filtering for the inmates we're looking for as well as all pertinent information. The ACLU of Utah was able to request court records from the Beaver County Justice Court spanning July 2019 to January 2020. It took three months to receive this data and cost $384. Each offense was listed with an ID and case number, but the ID did not match anything we

had for the booking data. Additionally, neither arrest nor booking date was included in the data, only the offense date, which may or may not correspond to the booking date. In order to join this data set with our booking data, we had to manually compare dates and names and essentially make our best guess. This was especially difficult with people who were booked multiple times in our time frame, which is unfortunately common. Also, some of the people listed in the court data were never incarcerated for their offenses but had to appear in court. Based on our initial analysis of this data, none of the cases violated the four-day filing requirement for information. However, a total of 18 did not meet the deadline to file probable cause. The majority of these cases could not be found in the collected booking data, either because the original booking predated the beginning of the project or the individual may not have been booked. Unfortunately, the difficulty in joining the data sets, as well as the questionable accuracy of the booking site itself, implies that this analysis is nowhere near thorough enough to determine if probable cause and information are being filed within the time constraints set by the state of Utah.

6 Conclusion

The number of issues we encountered while working this project implies that this method of transparency analysis is likely not viable for the long term. Three of the five scrapers were broken at least once (with one breaking three times) due to design and address changes. The majority of these changes also made scraping the websites more difficult as well. Accuracy is the other concern with scraping. Several of the websites have disclaimers stating the data is "AS IS" and that no guarantees are made to the accuracy. As we have seen, in particular when Beaver County updated their booking website, the booking times (if available) may not be exact. This makes determining if probable cause was filed within 24 hours impossible just from the booking data. The lack of booking times on other sites further compounds the issue. Lastly, collecting the pertinent data from the AOC is time-consuming and costly. The court data also does not include relevant data such as booking date and time. To further confuse the issue, there is no way to cleanly join the booking data and court data which can lead to other issues, including missing the very people we are trying to identify.

6.1 Transparency Analysis

Even though we were able to scrape a good deal of information from the various sites, the data we collected did not enable us to answer the questions posed at the beginning of the project. Measuring the transparency depends on acquiring data on a single person from arrest to court ruling or plea deal, with systems in place to ensure we have the same person from start to finish. The lack of unique identifiers and inconsistency in the data between the various sites leads us to state that the current level of transparency is not sufficient for the public to ensure due process is being followed. The new data collection law passed earlier this year does present an opportunity to increase transparency and ensure adherence to the law [23]. Specifically, transparency does not require that anyone have access to this data. This bill already requires submission of data to the CCJJ on a quarterly basis, in conjunction with unique identifiers for each offense that will allow for simple joining of the county jail, prosecutorial, and court data. While this is an admirable first step, the data collection required by the bill still would not be able to answer the questions posed at the beginning of this process.

6.2 Recommendations

As stated in a previous section, while the data collection law is a good first step, it does not go far enough. Primarily, the data submitted by the county jails should include the booking date and time. Without this particular data point, it is not possible to determine if probable cause and information were filed within a timely manner. Similarly, the offense tracking number should not be the only unique identifier. Utilizing a unique identifier for both individuals and bookings could also provide a wealth of information about recidivism and how people are charged when arrested. The responsibilities of the CCJJ as outlined in the bill appear to focus on diversion and reducing recidivism, meaning this analysis may lay outside its scope. It is possible, however, to designate some organization as a trusted entity, someone who has access to the collected data and can analyze and report on the statistics. This is already being done with the PSA algorithm and researchers at Harvard University [19].

Our second recommendation is for minimum requirements for the county jail websites as far as the displayed information and privacy protections. Again, while there are laws that help to protect individuals from predatory "publish-for-pay" websites, having protections on the booking websites that prevent scraping would prevent it from being published in the first place.

Lastly, and perhaps most loftily, we recommend a state-wide criminal justice database. Currently, every county appears to collect different information and uses different identifiers. In order to have a unique identifier that the jails, prosecutorial agencies, and the courts use for a single offense, an open form of communication is a necessity. With a state-wide database, all information could be entered and collected from the same place. Entries could be created by the jails when someone commits an offense, with additional information collected as the individual moves through the system. Then, again with a designated trusted entity, regular analyses could be performed to see not only improvements in recidivism, but also to ensure due process is being followed by all parties.

Transparency in our criminal justice system is critical. H.B. 288 is a step in the right direction, but we need to ensure that not only the people in the system are monitored, but also the system itself. With transparency and accountability, the detrimental and often unnecessary effects of pretrial detention can be mitigated and trust in the system will increase.

References

[1] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research 18* (2017), 234:1–78.

[2] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Arnold, D., Dobbie, W., and Yang, C. S. Racial bias in bail decisions. *The Quarterly Journal of Economics 133*, 4 (2018), 1885–1932.

[4] Barabas, C., Dinakar, K., and Doyle, C. The problems with risk assessment tools. https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html.

[5] Barr, J. R., Coggeshall, S., and Zhao, W. The trouble with names/dates of birth combinations as identifiers. https://www.idanalytics.com/media/The_Trouble_With-Names_White_Paper_FINAL.pdf.

[6] Beaver County. Beaver county booking website, 2020. https://www.beaverutahsheriff.com/468/Recent-Bookings.

[7] Benbouzid, B. Values and consequences in predictive machine evaluation. a sociology of predictive policing. *Science & Technology Studies 31* (2018).

[8] Boulton, C. What is rpa? a revolution in business process automation. https://www.cio.com/article/3236451/what-is-rpa-robotic-process-automation-explained.html.

[9] Bronner, L. Why statistics don't capture the full extent of the systemic bias in policing. https://fivethirtyeight.com/features/why-statistics-dont-capture-the-full-extent-of-the-systemic-bias-in-policing/?utm_source=pocket&utm_medium=email&utm_campaign=pockethits.

[10] Cache County. Cache county arrests and bookings, 2020. https://www.cachesheriff.org/news/inmate-roster.html.

[11] Carbon. Carbon County Justice Court, 2020. https://www.carbon.utah.gov/Administration/Judicial/Justice-Court.

[12] Chin, V. Introductory handbook on the prevention of recidivism and the social reintegration of offenders, 2018. https://www.un.org/ruleoflaw/blog/document/introductory-handbook-on-the-prevention-of-recidivism-and-the-social-reintegration-of-offenders/.

[13] Corey, E. How a tool to help judges may be leading them astray. https://theappeal.org/how-a-tool-to-help-judges-may-be-leading-them-astray/.

[14] Davis County. Davis county inmate roster, 2020. https://www.daviscountyutah.gov/sheriff/inmate-roster#.

[15] Dressel, J., and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*, 1 (2018), eaao5580.

[16] Evans, W. N., and Owens, E. G. Cops and crime. *Journal of Public Economics 91*, 1–2 (2007), 181–201.

[17] Ferguson, A. G. Policing predictive policing. *Washington University Law Review 94*, 5 (2017).

[18] Fletcher, M. A. For black motorists, a never-ending fear of being stopped. https://www.nationalgeographic.com/magazine/2018/04/the-stop-race-police-traffic/.

[19] Friedman, G. Poor people are trapped behind bars. how utah is using an algorithm to get some of them out. https://www.deseret.com/2018/6/17/20647170/poor-people-are-trapped-behind-bars-how-utah-is-using-an-algorithm-to-get-some-of-them-out.

[20] Geng, L. Politifact - do more than 450,000 americans sit in jail because they are too poor to pay bail?, Jun 2018. https://www.politifact.com/factchecks/2018/jun/29/ted-lieu/do-more-450000-americans-sit-jail-because-they-are/.

[21] Gupta, A., Hansman, C., and Frenchman, E. The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies 45*, 2 (2016), 471–505.

[22] Haskins, C. Dozens of cities have secretly experimented with predictive policing software. https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software.

[23] H.B. 288, Prosecutor Data Collection Amendments, 2020 General Session, Utah 2020. https://le.utah.gov/~2020/bills/static/HB0288.html.

[24] Holsinger, A. M., and Holsinger, K. Analyzing bond supervision survey data: The effects of pretrial detention on self-reported outcomes. *Federal Probation 82*, 2 (Sep 2018), 39–45. https://www.uscourts.gov/federal-probation-journal/2018/09/analyzing-bond-supervision-survey-data-effects-pretrial-detention.

[25] Johnson, D. Criminal justice fact sheet, 2020. https://www.naacp.org/criminal-justice-fact-sheet/.

[26] Kirkpatrick, K. It's not the algorithm, it's the data. *Communications of the ACM 60*, 2 (2017), 21–23.

[27] Knox, D., Lowe, W., and Mummolo, J. Administrative records mask racially biased policing. *American Political Science Review* (2020), 1–19.

[28] Liu, P., Nunn, R., and Shambaugh, J. The economics of bail and pretrial detention. https://www.hamiltonproject.org/assets/files/BailFineReform_EA_121818_6PM.pdf.

[29] Lum, K., and Isaac, W. To predict and serve? *Significance 13*, 5 (2016), 14–19.

[30] Monroe County Jail. Morgan county jail, ut inmate search, mugshots, prison roster, Aug 2020. https://monroecountyjail.net/prisons/utah/county-jail/morgan-county-jail/.

[31] Morgan County. Morgan county jail website, 2020. https://corrections.utah.gov/index.php/2014-10-30-20-13-59.

[32] Onpoint Court Services, 2019. https://onpointmonitoring.com/pricing/.

[33] Rockett, D. Poor people often can't afford to pay bail - even when they're innocent. an app developed in chicago offers help. https://www.chicagotribune.com/lifestyles/ct-life-appolition-making-bail-20190124-story.html.

[34] Salt Lake County. Salt lake county jail inmate lookup, 2007. http://iml.slsheriff.org/IML.

[35] Salt Lake County. Salt lake county jail dashboard, 2019. https://slsheriff.org/page_jail_dashboard.php.

[36] Sanpete County. Sanpete county booking website, 2010. https://www.sanpetesheriff.org/Booking.html.

[37] Sawyer, W., and Wagner, P. Mass incarceration: The whole pie 2020, Mar 2020. https://www.prisonpolicy.org/reports/pie2020.html.

[38] Solon, O. Haunted by a mugshot: how predatory websites exploit the shame of arrest. https://www.theguardian.com/technology/2018/jun/12/mugshot-exploitation-websites-arrests-shame.

[39] Speri, A. Police make more than 10 million arrests a year, but that doesn't mean they're solving crimes. https://theintercept.com/2019/01/31/arrests-policing-vera-institute-of-justice/.

[40] Stevenson, B. *Just Mercy: A Story of Justice and Redemption*. New York: Spiegel & Grau, 2014.

[41] The Arnold Foundation. Risk-assessment fact sheet public safety assessment (psa), May 2019. https://www-cdn.law.stanford.edu/wp-content/uploads/2019/05/PSA-Sheet-CC-Final-5.10-CC-Upload.pdf.

[42] The Sentencing Project. Report to the united nations on racial disparities in the u.s. criminal justice system. https://www.sentencingproject.org/publications/un-report-on-racial-disparities/.

[43] Tooele County. Tooele county inmate roster, 2019. http://inmate.tooelecountysheriff.org/Roster/Search.

[44] Travis, J., Western, B., and Redburn, S. *The Growth of Incarceration in the United States: Exploring Causes and Consequences*. The National Academies Press, 2014. https://www.nap.edu/read/18613/chapter/1.

[45] United States. Congress. House. Committee of Conference. Violent crime control and law enforcement act of 1994 : conference report to accompany h.r. 3355, 1994. [Washington, D.C.?] :[U.S. G.P.O.].

[46] Using Data to Reduce Policing Harms, document in progress not yet publicly available.

[47] Utah Code. Government records access and management act, Utah 2020. https://le.utah.gov/xcode/Title63G/Chapter2/63G-2.html.

[48] Utah County. Utah county inmate roster, 2020. https://api.utahcounty.gov/sheriff/corrections/inmateSearch.

[49] Utah Courts. Utah public safety assessment frequently asked questions, Jun 2018. https://www.utcourts.gov/resources/reports/psa/faq.html.

[50] Utah Courts. Utah rules of criminal procedure, Feb 2019. https://www.utcourts.gov/resources/rules/urcrp/.

[51] Utah General Assembly. 17 utah code 17-22-30, 2019. https://le.utah.gov/xcode/Title17/Chapter22/17-22-S30.html.

[52] Vera Institute of Justice. Arrests: How many arrests are made annually, and for what?, Jan 2019. https://arresttrends.vera.org/arrests.

[53] Washington County. Washington county booking website, 2007. https://news.washeriff.net/divisions/corrections-division/bookings/.

[54] Weber County. Weber county inmate roster, 2020. https://www.webercountyutah.gov/sheriff/roster/index.php.

[55] Worden, A., and Clark, A. *Misdemeanor Justice in Rural Courts*. May 2019, pp. 55–65.