A Computational Decision-Tree Approach to Inform Post-Conviction Intake Decisions

Kalina Kostyszyn, Carl J. Wiedemann, Rosa Bermejo, Amie Paige, Kristen W. Kalb-DellaRatta, & Susan E. Brennan

Stony Brook University

June 29, 2023

Abstract

How might data analytic tools support intake decisions? When faced with a request for post-conviction assistance, innocence organizations' intake staff must determine (1) whether the applicant can be shown to be factually innocent, and (2) whether the organization has the resources to help. These difficult categorization decisions are often made with incomplete information (Weintraub, 2022). We explore data from the National Registry of Exonerations (NRE; 4/26/2023, N = 3,284 exonerations) to inform such decisions, using patterns of features associated with successful prior cases. We first reproduce Berube et al. (2023)'s latent class analysis, identifying four underlying categories across cases. We then apply a second technique to increase transparency, decision tree analysis (WEKA, Frank et al., 2013). Decision trees can decompose complex patterns of data into ordered flows of variables, with the potential to guide intermediate steps that could be tailored to the particular organization's limitations, areas of expertise, and resources.

Table of Contents

This material is based upon work supported by the National Science Foundation under Award No. 2125295 (NRT-HDR: *Detecting and Addressing Bias in Data, Humans, and Institutions*). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

1. Introduction

A. The Promise and Pitfalls of Data-Intensive Methods

Data and data-intensive methods are increasingly promoted—and indeed, sometimes mandated—as solutions within domains that call for people to make difficult decisions about pressing human problems. The fair and ethical use of such methods requires transparency, especially when the stakes are high. However, AI/machine learning tools can be notoriously opaque to human users. Opacity in algorithms can result in biased decisions that, once made, are difficult to challenge. The damage done can be life-changing and difficult to reverse (e.g. firing good teachers for the wrong reasons; Turque, 2012; O'Neil, 2016). When data-intensive algorithms are "black boxes," it's difficult to understand the reasoning behind the outcomes. Therefore, it is necessary to advocate for transparency in two key ways: first, the variables included in the training data for algorithms should be justifiable, and second, it should be clear how these variables are evaluated or weighted in outcomes or predictions. This is particularly important for algorithm-aided decisions in the domain of criminal justice, which can have profound impacts on vulnerable individuals.

Bias can creep into algorithms in different ways. When machine learning models are trained on datasets that are missing relevant information, the models produce outcomes that are unreliable for those cases that are underrepresented in the datasets; this may result in reproducing the biases of the past, or in otherwise unreliable outcomes. For example, "state-of-the-art" facial recognition algorithms learned to detect White faces better than Black faces and male faces better than female faces (Buolamwini & Gebru, 2018), making errors when classifying new cases that were not well represented in the training data. And when intersectional identity is sparsely represented and unevenly distributed in training data, "fairness gerrymandering" may

result (Kearns & Roth, 2020), as it did when the faces of Black women were recognized least accurately of all (Buolamwini & Gebru, 2018). Transparency in the kind of data used to develop, train, and test algorithms is necessary to understand and ultimately prevent the potential misclassification of underrepresented individuals down the line.

Bias can also stem from the variables chosen for inclusion in training data. During training, models learn to represent underlying patterns among variables in the data in ways that are hidden even from their developers, and that may perpetuate undesirable stereotypes. This can occur even when key variables such as race or gender are removed from a dataset (and thus are considered to be "protected"). It may seem sufficient that a protected variable isn't included in training a model, yet it can nevertheless still influence the outcome when other (proxy) variables that are correlated with the protected variable are included (see, e.g., O'Neil, 2016 & Angwin et al., 2016 for discussion of risks associated with proxies). For example, Amazon discontinued using an algorithm trained to identify successful job candidates after it was discovered that, despite removing gender as a variable, the algorithm still recommended men over women (Dastin, 2018); the resumés in the training data still included information such as extracurricular activities strongly correlated with gender. As another example, some states require that inmates fill out questionnaires that are used to support automated predictions about recidivism. Although asking about race is illegal in some jurisdictions and therefore avoided, questions about family members' unemployment or welfare status, or about the age at which an individual first interacted with the police, can differentiate privileged from underprivileged individuals (and may divide them along race-based lines; Angwin et al., 2016). In this way, variables associated with privilege or lack thereof can serve as a proxy for race while ignoring that one's first contact with the police may be a result of biased policing practices (O'Neil, 2016; see also Harcourt, 2015).

Another source of bias can arise when algorithms are deployed as decision aids without taking into account whether the distribution of errors is equitable and fair, or whether there are disparate impacts on individuals. For example, Northpointe's COMPAS (Correctional Offender Management Profiling for Alternative Sanctions; T. Brennan & Dieterich, 2017) algorithm derives individuals' recidivism risk scores from questionnaires given to them when they are incarcerated. COMPAS does not include an overt race variable, and its developers claimed that it was unbiased because its error rate in predicting recidivism for Black parolees was the same as for White parolees (39% for both). In an adversarial audit, the public interest group ProPublica obtained a dataset through a FOIA (Freedom of Information Act) request of more than 7,000 parolees in Broward County, FL over a 2-year period; all had been given the COMPAS algorithm's lengthy input questionnaire when incarcerated (Angwin et al., 2016). The ProPublica team painstakingly unearthed the ground truth about whether these individuals actually recidivated over the next several years and merged this information with COMPAS's predictions about them (as recounted in Christian, 2020). They found that the *types* of errors were dramatically different for Black and White parolees: approximately two-thirds of errors for White parolees were false negatives (where COMPAS had recommended release, but the parolee had recidivated), whereas two thirds of errors for Black parolees were false positives (where COMPAS had recommended denying parole, but the individual was paroled and there was no recidivism) (Angwin et al., 2016). This distribution of errors privileges one group while being grossly unfair to another. Yet data-intensive decision-making aids such as COMPAS are mandated in many jurisdictions around the U.S. (and with limited or no oversight; see Christian, 2020).

Biases can be further compounded when a decision-support algorithm is deployed blindly by those who should ultimately be the ones accountable for a decision, but who don't understand the limitations of the algorithm within their context of use. The COMPAS algorithm was designed to assist with judges' decisions about eligibility for parole or treatment programs (Angwin et al., 2016). Yet it has been applied to decisions about bail, pre-trial detention, and sentencing (uses that even the developers deem to be inappropriate; Angwin et al., 2016 & Christian, 2020).

Finally, although decision-support algorithms such as COMPAS are often used to assess risks posed by individuals accused or convicted of crimes (for the benefit and protection of society), these data-intensive methods can also be used to determine when and how to provide benefit and support to the accused or convicted individuals themselves. Whenever algorithms are used to recommend life-changing decisions, transparency is essential, not only to prevent unintended harms, but also to undo the harms that may have resulted from multiple sources of bias. Here, we explore the use of data-intensive methods in the domain of *wrongful convictions.*

B. Wrongful Convictions and the Intake Process

Wrongful convictions, by their very nature, are not readily observable. Accordingly, the true rate of wrongful convictions is a *dark figure*, that is to say, a figure that is typically recognized as unknown or even unknowable, but at the very least, extremely difficult to ascertain (Bedau & Radelet, 1987; Gross et al., 2014). One estimate based on a thoroughly-vetted survey of state prisoners (with non-parametric tests used to account for possible false innocence claims) suggests that 6% of incarcerations are based on wrongful convictions (Loeffler et al., 2019). Among capital cases, a conservative estimate of the rate of wrongful convictions is about 4%

(Gross et al., 2014). It can therefore safely be said that many people are actively serving prison sentences for crimes that they did not commit, or that did not even occur in the first place.

Through exoneration, the official alleviation of legal culpability for a crime that a person was originally found to be guilty of, victims of wrongful conviction may find an avenue to justice. Innocence organizations are a group of legal representatives and advocates for the wrongfully convicted. With more than 900 contributions to exonerations, innocence organizations play a vital role in exonerating the wrongfully convicted. As of 2023, there are 72 member organizations in the Innocence Network spread across the United States. The Innocence Network serves as a community that provides various forms of support for newly exonerated individuals in addition to providing resources for legal organizations that join its mission in exonerating the innocent. Whether an innocence organization accepts an application can depend on the availability of resources like the number of staff and budget. Innocence organizations can receive anywhere from 20 to 2,400 requests for assistance a year, and moving a case to exoneration is an extremely time-consuming process that intake staff estimate to take around seven years (Weintraub, 2022). The investigative processing of a case alone can take more than a year to complete (Krieger, 2011). Therefore, innocence organizations and staff must strategically allocate resources to cases they determine are most likely to be successful.

The inner workings of individual innocence organizations impact the types of cases they can investigate and litigate. A qualitative study of 19 innocence organizations by Weintraub (2022) found that intake procedures vary among organizations. Such variations include 1) length of application, 2) whether the application is reviewed by either intake staff or directors, attorneys, or law students, 3) whether an organization conducts a screening interview with the applicant, and 4) intake criteria. Common intake criteria of most innocence organizations include

factual innocence and geographic restrictions within a certain state or region, but organizations vary on acceptance or consideration of cases involving child sexual abuse and sustained abuse, whether an applicant was involved in the criminal action, cases with DNA evidence, arson, shaken baby syndrome, guilty pleas, new evidence of innocence at intake, indigent status, and sentence length (Weintraub, 2022).

To support intake staff as they categorize and evaluate post-conviction requests, data-intensive decision support tools should empower them to effectively interpret and communicate about the results of multi-step data-driven analyses.

C. The NRE and the Six Canonical Factors

To understand the myriad of factors that contribute to wrongful convictions, data from successful exoneration cases can be illuminating. To this end, the Innocence Project actively maintains, updates, and consults a national dataset containing information on DNA-based exonerations (Innocence Project, *Cases*, 2023). Through examination of this dataset, the Innocence Project has identified such "contributing causes" of wrongful convictions exposed via DNA evidence as: eyewitness misidentification, misapplication of forensic science, false confession or incriminating statement, incentivized informants' statements, misconduct by government actors, and inadequate defense counsel (West & Meterko, 2016). These factors are particularly relevant to the Innocence Project's internal investigations and goals that focused originally on DNA evidence, but are not generalizable to the larger set of wrongful conviction cases that include non-DNA cases as well (Acker & Redlich, 2019).

The much broader National Registry of Exonerations (NRE) database aims to include all exonerations; for this reason, we focus here on the NRE. Founded in 2012, this database is maintained by a dedicated group of scholars, lawyers, and journalists who have cataloged data on

successful exonerations both for DNA- and non-DNA-based cases that have occurred since 1989. As of April 26th, 2023 when we did our analyses, the database contained information on 3,284 cases in total, making it the most comprehensive and most-frequently cited (Gross, 2008) source of raw information on known wrongful convictions to date.

Each case in the NRE database includes at least one of six "canonical" factors that have been identified as common contributors to wrongful convictions: False Confession (FC), Mistaken Witness Identification (MWID), False/Misleading Forensic Evidence (F/MFE), Perjury/False Accusation (P/FA), Official Misconduct (OM), and Inadequate Legal Defense (ILD) (Acker & Redlich, 2019). Given the greater diversity of cases in this dataset, the six so-called canonical factors are presumably more appropriate for analyses seeking to shed light on wrongful convictions in general, compared to the causal factors related to DNA-based exonerations (ibid).

Due to their dichotomous nature, the six canonical factors can be used as indicator variables for the technique known as latent class analysis (LCA). LCAs are informative for datasets such as the NRE, as they identify latent (i.e., not directly observable) subgroups within populations (McCutcheon, 2002). This method can be considered analogous to factor analysis albeit for categorical data: Both analyses demonstrate the interrelatedness of indicator variables whose associations are explained by unobserved factors, rather than direct causal relationships (McCutcheon, 2002). Many cases in the NRE include more than one of the six canonical factors, as they frequently co-occur. A benefit of applying LCA to a dataset such as the NRE is that classes extracted from the analysis would account for co-occurrences of the relevant subsumed canonical factors.

Our present project is inspired by the results of an LCA analysis of the NRE database, reported by Berube et al. (2023). In their paper, Berube and colleagues sought to identify patterns that broadly underlie wrongful convictions. Through applying LCA to the NRE, they found that a four-class model best fit the data and named the four extracted classes as follows: *Intentional Errors*, *Witness Mistakes*, *Investigative Corruption*, and *Failures to Investigate*. They then performed correlations with other NRE variables, such as exoneree demographics, measures of case severity, and process/evidence-related variables to examine how trends within the six canonical factors, as represented by the latent classes, related to other case factors.

D. The Current Analysis

Although LCA offers a useful method for extrapolating underlying patterns within a dataset such as the NRE, there are a number of important limitations that should be considered alongside its implementation (Weller et al., 2020). According to current best practices for LCA, as described by Weller et al. (2020), proper class assignment and percentages of representation within a particular class are not always guaranteed because LCAs rely on probability estimates to assign members of a dataset to a particular latent class (Muthén & Muthén, 2000). Weller et al. (2020) also warn of the heightened potential for "naming fallacies" to occur when researchers attempt to create labels for the extrapolated classes. Such labels may fail to appropriately capture the complexities of the determining factors in class memberships. Therefore, we first aimed to reproduce Berube et al.'s (2023) analysis, both to demonstrate the stability of their findings with a larger data set, and to simultaneously allow for a possible re-assessment of the labels originally conferred upon the latent classes. Second, to account for the inherent limitations of LCAs, we aimed to use the extracted classes as targets for a predictive analysis that would be more transparent and interpretable.

Models of complex data often use regression-based analyses to make predictions. However, these models can be nonoptimal for guiding human decision-making because of difficulties in interpreting and applying data to ambiguous, novel, and idiosyncratic cases. We introduce what we propose may be a more transparent framework using *decision trees* (Flach, 2012; Duda et al., 2001). Decision trees identify and lay out the impacts of variables one by one in a graphical representation similar to a flow chart, in a form that can be scrutinized by a human decision-maker. It may be possible to use decision trees to identify combinations of features relevant at different stages of evaluating a post-conviction case, to help with prioritizing new cases, and to direct attention to the most promising path to pursue next. Once the algorithm segments the dataset based on a particular feature, subsequent branches (or steps) can be interpreted more easily than the outcomes of classic regression analyses. Furthermore, integrating the grouping variables/classes identified by an LCA can improve model fit for decision trees (Gañan-Cardenas et al., 2022), making the pairing of these two approaches promising. A decision tree approach may uncover previously undetected trends in the data and increase the interpretability of results derived via LCA.

Ultimately, the framework we propose in this paper uses successful exonerations to evaluate and identify potential pathways that may be used during the intake process for new applicant cases. This might conserve work hours, identify specialized resources needed for a particular applicant, and transparently support efforts to expedite and communicate about decisions within an innocence organization. Of course, this framework will need to be tested within the context of use. To the extent that this framework may reveal previously unknown biases introduced prior to conviction, it may also allow for more effective communication with

law enforcement, legislators, and other policy-making entities in an effort to reduce future wrongful convictions.

2. Method

A. Sample

The analyses presented here were based on data from the National Registry of Exonerations, downloaded on April 26th, 2023. There were a total of 3,284 exonerees in the database at the time of download. Notably, Berube et al.'s (2023) latent class analysis was conducted on data from the same source, but at the time of their download the database included a total of 2,880 exonerees.

B. Variables

In accordance with our goal of assessing the reproducibility of Berube et al.'s (2023) analysis, we based our analyses on the same variables and the same coding scheme to the greatest extent possible. We therefore relied on variables included in the NRE dataset, such as the six canonical factors, exoneree demographic information, case severity measures, and process/evidence-related variables.

Covariates

Exoneree Demographic Information. Several NRE database variables concern demographic information about the exoneree and geographic information (jurisdiction) about the case; while we initially remove state information, we add it back later while experimenting with manipulations of the tree structure. We excluded exoneree names from analyses, as well as

counties. [1] In addition, in the absence of intuition about how these variables are distributed or interact with other variables, we removed any continuous variables from our initial study—ages and dates, for example. We do, however, differentiate juveniles at time of conviction from adults, without further differentiating within those classifications. We retain information about race and sex (in fact, 'female exoneree' is a separate variable listed with the process information).

Case Severity Measures. Another set of variables deals with case severity. The 'worst crime display' variable contains values for the single most severe crime associated with each case; there are additional binary variables specifying whether attributes (such as homicide, sexual assault, etc.) were part of a case, distributing this information in a way that makes comparison simpler. Berube and colleagues (2023) also use the sentence length as a measure of case severity; just as we removed ages and dates, we remove this information. The sentence length may interact, in ways we currently are unable to discern, with conviction date, length of incarceration, and details associated with post-conviction actions.

Process/Evidence Related Variables. The bulk of the variables in the data set are divided by the NRE as either 'tags', which include information about the crime or the exoneration, or 'official misconduct tags', which contain more specific information about misconduct that led to wrongful conviction; these are coded as binary true/false values. A small number of them reiterate information in other variables - female exonerees and juvenile

[1] We discovered that six exonerees were in the database twice, from multiple exonerations pertaining to the same case. For these cases, we used data from the chronologically later exoneration, which generally has more specific characteristics (or "tags" in NRE database parlance). For two of these exonerees, the LCA analysis assigned each entry a different class because of significant differences in tags, despite it being the same individual.

defendants have been mentioned, but there is in addition a variable that marks whether a case was held at the federal level and that is repeated in the state information.

Handling of Exceptional Covariates. As we examine the trees, a number of these covariates will be manipulated due to somewhat exceptional status. First, because the NRE dataset represents a snapshot in time (with the outcomes of cases that may have taken decades to adjudicate), we consider a class of variables that are determined only at the end of the exoneration case. We use the 'no crime' and 'DNA used in exoneration' variables as examples of information that did not contribute to the original conviction but was a basis for overturning it. Second, we code based on whether an innocence organization and/or conviction integrity unit participated in the exoneration process, though we recognize that these variables may not be useful for all intended analyses. Manipulating these variables, however, is important to demonstrate the flexibility of these decision trees and how they are able to make similar generalizations even when provided with different input data.

C. Data Analysis Plan

Latent Class Analysis (LCA)

We began by reproducing Berube et al.'s (2023) latent class analysis (LCA), which identified four underlying classes in the NRE dataset, such that each case could be categorized based on its highest probability of belonging to one of the four classes. Following Berube (ibid), we used the *Six Canonical Factors* that contribute to wrongful convictions (Acker & Redlich, 2019) as latent class indicators, coded dichotomously. These are: Mistaken Witness Identification (MWID), False Confession (FC), Perjury or False Accusation (P/FA), False or Misleading Forensic Evidence (F/MFE), Official Misconduct (OM), and Inadequate Legal Defense (ILD). Goodness of fit was assessed using multiple criteria, including the Bayesian

information criterion (BIC; Schwarz, 1978) and Akaike's information criterion (AIC; Akaike, 1987).

Decision Tree Analysis

Using decision trees (WEKA, Frank et al., 2016), we decomposed and reanalyzed the four classes modeled in the LCA. First, we used decision trees to predict classification from the LCA approach, using only the Six Canonical Factors, to assess the validity in combining these approaches (*six-factor model*) and determine what ordered combinations of features could predict LCA-based classification. Second, we explored other trends within the four latent classes by examining covariates other than the canonical factors (*extended model*). The decision trees determined other associated features that predict the classification of a case. Case tags - such as *withheld exculpatory evidence* or *juvenile defendant* - were recoded as binary features where possible, then ordered by the decision tree to see how accurately combinations of these features could predict LCA-based classification.

3. Summary of Results

A. Latent Class Analysis Reproduction

All statistical analyses pertaining to the LCA reproduction were conducted in R (R Core Team, 2020), with associated figures produced via the ggplot2 package (Wickham, 2016). In accordance with recommended best practices for LCAs (Weller et al., 2020), we successively fit a series of models, starting with a one-class model. A four-class model provided the best overall fit according to statistical criteria, thus supporting Berube et al. (2023) while expanding their analysis to a larger dataset. The optimal BIC value was associated with a four-class model (21052.42), as compared to a three-class model (21180.39) and a five-class model (21054.58). Similar to results reported by Berube et al. (2023), fit improvement, as indicated by AIC, from

the four-class model to the five-class model (a difference of 40.52) was much smaller than fit improvement from the three-class model to the four-class model (a difference of 170.64). So, we maintain, in agreement with Berube et al. (2023), that a four-class model seems to best fit the data, despite a more favorable AIC value being associated with the five-class model. We also note that one of the classes in the five-class model included a membership of only 7% of cases. Such a low representation of the dataset could lead to issues both with generalizability and interpretability. Again, deferring to the four-class model appears to be the optimal solution.

Table 1

Results of Model Fit Comparisons

Model	AIC	BIC	BLRT	*P*-value	χ^2	*P*-value	Mixing Proportions
1-Class	22452.86	22489.44	1724.51	-	1849.51	-	-
2-Class	21528.89	21294.40	472.80	<.001	484.89	<.001	.37/.63
3-Class	21058.45	21180.39	302.10	<.001	315.37	<.001	.19/.64/.17
4-Class	20887.81	21052.42	117.46	<.001	115.32	<.001	.18/.21/.30/.32
5-Class	20847.29	21054.58	62.94	<.001	63.00	<.001	.16/.30/.16/.07/.31

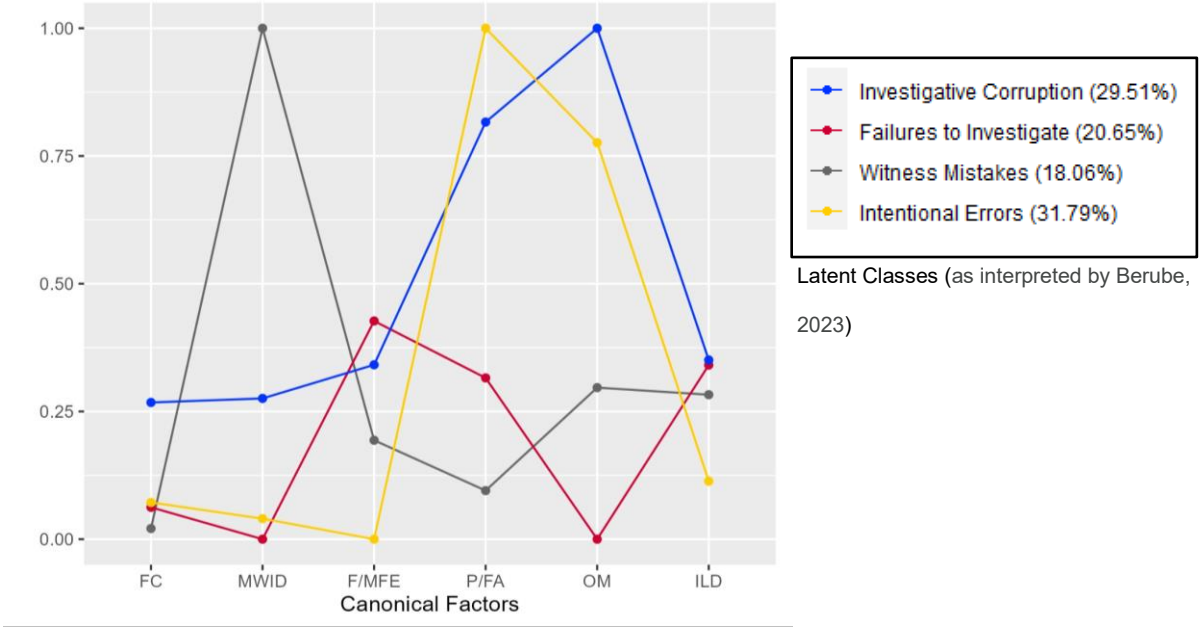
Note. AIC = Akaike's information criterion; BIC = Bayesian information criterion; BLRT = bootstrap likelihood ratio test; χ^2 = chi-square. *P*-values were reported for BLRT and χ^2. Mixing proportions were based on the most likely latent class membership.

Distributions of predicted class memberships and the profiles of their respective representations of the Six Canonical Factors were highly similar to those reported by Berube et al. (2023). For example, 100% of the cases estimated to be members of the class originally labeled "Witness Mistakes" were associated with MWID. The canonical factor OM (Official

Misconduct) was also highly indicative of cases assigned to the “Witness Mistakes" class. The “Investigative Corruption” class was most strongly characterized by OM (100% of assigned cases), P/FA (Perjury/False Accusation), and ILD (Inadequate Legal Defense), while the “Intentional Errors" class was most strongly characterized by P/FA (100% of assigned cases) and OM. The “Failures to Investigate" class was most strongly characterized by F/MFE (False/Misleading Forensic Evidence), but did not have as high of an association with this canonical factor as was observed for the factors that most strongly characterized the other latent classes. These patterns are largely in alignment with Berube et al.’s (2023) results. We therefore tentatively retain the labels reflecting Berube et al.’s (2023) interpretations, but a critical evaluation of these labels and their interpretability follow in the next paragraph and in subsequent sections of this paper. Percentages of predicted class membership were as follows: Intentional Errors, 31.79%; Witness Mistakes, 18.06%; Investigative Corruption, 29.51%; and Failures to Investigate, 20.65% (see *Figure 1*).

Figure 1

Latent Class Analysis of NRE Data as of April 26th, 2023 (Reproduction of Berube et al., 2023)



Note. FC = False Confession; MWID = Mistaken Witness Identification; F/FME = False/Misleading Forensic Evidence; P/FA = Perjury/False Accusation; OM = Official Misconduct; ILD = Inadequate Legal Defense.

Because a goal of the present work is to use these extracted classes as targets of prediction via decision trees, a critical evaluation of the extent to which classes are mutually exclusive is warranted. To that end, it is worth noting that both the Intentional Errors class and the Investigative Corruption class were characterized by high degrees of P/FA and OM. Visually, their patterns of representation were quite similar. The similarities in these patterns suggest that our subsequent decision tree analysis may be more likely to misidentify cases assigned

membership to the Intentional Errors class as Investigative Corruption, or vice versa. Examining the LCA's posterior probabilities can help us predict the directionality of errors that a decision tree might make in predicting membership within these two classes. In LCA, posterior probabilities represent the probability of a given case to have otherwise been assigned to one of the alternative classes in the model. We thus ran a Welch's two sample *t*-test, comparing the mean posterior probability of members in the Intentional Errors class to have been categorized as Investigative Corruption ($M = 0.18$, $SD = 0.11$), to the mean posterior probability of members in the Investigative Corruption class to have been categorized as Intentional Errors ($M = 0.15$, $SD = 0.18$). Results of the *t*-test indicated that there was a statistically significant difference in these mean posterior probabilities, $t(1566.5) = 4.66$, $p < .001$. In other words, cases that were assigned membership to Intentional Errors had a higher mean posterior probability of being assigned to the Investigative Corruption class than the mean posterior probability of cases classified as Investigative Corruption to have been assigned to the Intentional Errors class. It should therefore be expected that, when using decision trees, classification disagreements will manifest such that cases originally assigned as Intentional Errors will be more often classified as Investigative Corruption, as opposed to the converse.

B. Decision Trees

We use WEKA's J48 package (WEKA, Frank et al., 2016) to build our decision trees. J48 is an implementation of the C4.5 algorithm (Quinlan, 1993), which is a Classification and Regression Tree (or CART model; Breiman et al., 1984) that, given the input database, will make partitions within the data based on how well a partition is able to generalize for classification. Variables used for these partitions have a high *information gain* at that point in the algorithm; the higher information gain a variable has, the more evenly its values subdivide the

space, which minimizes the number of additional variables needed to classify a data point. A variable having low information gain does *not* mean that a given value is not representative of a class, but rather that the other values are not sufficiently discriminatory for other classes. At the point these low-gain variables are found in the decision tree, competing branches have been eliminated, making categorization based on the variable's value more likely.

To take advantage of the decision trees' learning ability, we ran several models, training two instances of each with either a 75% train-25% split or with 10-fold cross validation. In the first training regimen, 75% of the dataset was used to train the tree a single time, and evaluation was done over the held-out 25% of the dataset. In the second, we use 10-fold cross-validation (Stone, 1976), where we first partition the dataset into 10 equal sets, train a model over 9 of those ten, and rotate which model we test on the remaining 10th set. All other standard settings are untouched; in particular, we did not 'prune' the tree, or remove low-occurrence branches, as we wanted to examine the breadth of generalizations.

For evaluation, we primarily use precision, recall, and f-score measures. Precision is the percentage of selected items that belong to the target group versus selected items that were not targets; recall is the percentage of target items selected versus target items the model did not select. The f-score is the harmonic mean of these two. Additionally, we will list confusion matrices, which will display how many items in each group were correctly classified and, if not, which other category they were classified into. These values will elucidate the error rates and patterns of classification disagreements (here, disagreements in categorization between the LCA analysis and the decision tree), which we will analyze below.

Six-Factor Model

We begin with the six-factor model, which trains itself on the canonical factors associated with each case and predicts which latent class is attributed to each. This demonstrates how well the decision trees are able to interpret the underlying data given to the latent class models. Hearteningly, these models perform near-perfectly, easily using the canonical factors to categorize cases.

Extended Model

Next, we created an 'extended' model, in which we train on the set of covariates rather than the six canonical factors, while still predicting the latent class for each case. High performance here will demonstrate that the decision tree is finding underlying patterns in the covariates that align with the latent classes.

Table 2

Evaluative accuracy scores for the 6-factor and baseline extended models

Model	*Precision*	*Recall*	*F-Score*
6-factor, cross validated	1.000	1.000	1.000
6-factor, 75-25	0.999	0.999	0.999
Extended, cross validated	0.722	0.721	0.720
Extended, 75-25	0.737	0.737	0.737

Given high performance of the extended model, we can now manipulate the tree by excluding covariates that offer little predictive power regarding new cases under consideration, or including covariates which were initially set aside.

Table 3

Confusion matrices for baseline extended models (75/25 split)

	Classed FtI	Classed IE	Classed IC	Classed WM
True FtI	131	16	0	12
True IE	19	173	56	11
True IC	0	53	176	12
True WM	10	11	15	124

Removing 'No-Crime' and 'DNA' Cases

In order to more closely model the incomplete information that may be available to intake staff, in this section we remove variables that refer to the outcome of the exoneration. In the NRE, the variable 'DNA' refers specifically to new DNA evidence introduced in post-conviction that directly led to exoneration. In 'no-crime cases,' the exoneree was initially convicted of a crime that did not happen. This could be a crime that was entirely fabricated, or an incident that was mistaken for a crime, such as an accident or a suicide. Because these variables may be unknown at intake, we present a model here that makes predictions without them. These perform comparably to the baseline extended model, which used these variables for partition, suggesting that it is able to use other information in the data set to make similar generalizations.

Table 4

Confusion matrices for no DNA/no-crime extended models (75/25 split)

	Classed FtI	Classed IE	Classed IC	Classed WM
True FtI	114	25	0	20
True IE	7	172	57	23
True IC	0	35	200	6
True WM	18	12	26	104

Removing Innocence Organization and Conviction Integrity Unit Information

In this section, we remove variables that refer to the involvement of an innocence organization (IO) or conviction integrity unit (CIU), as they are characteristics of those who are vetting the case post-conviction rather than variables in place at the time of the crime, investigation, or prosecution. However, considering that our analyses may be used in the future to inform whether to accept a case, such information may be of value. It is worth noting that, while these two variables were included in the baseline model, that baseline was outperformed by the 75-25 split of this manipulation. This suggests that the high information gain of these variables may be preventing the model from revealing other, more informative partitions downstream. Because of these differences, it may be useful to observe a model with and without these variables to see where their predictions differ. This is not to say that the claims made by one model are inadequate, but rather that when the data is restructured by removing carefully selected variables, the model will compensate by taking advantage of new generalizations it previously could ignore.

Table 5

Confusion matrices for no IO/CIU extended models (75/25 split)

	Classed FtI	Classed IE	Classed IC	Classed WM
True FtI	136	11	0	12
True IE	20	172	54	13
True IC	0	45	185	11
True WM	11	5	23	121

Adding State-Wise Information

In this section, rather than removing variables, we add the state where the case occurred as a variable. Because this includes over 50 discrete values - every state or U.S. territory, plus federal and military cases, which are listed separately - this was removed from the initial tree for interpretability, but reintroduced to assess the predictive power of these variables. These models are comparable to the baseline, though state information, where it appears, partitions trees close to the leaves.

Table 6

Confusion matrices for extended models with state information (75/25 split)

	Classed FtI	Classed IE	Classed IC	Classed WM
True FtI	127	17	0	15
True IE	18	173	58	10
True IC	0	46	185	10
True WM	17	8	22	113

Table 7

Evaluative accuracy scores for modified extended models

Model	*Precision*	*Recall*	*F-Score*
No DNA, no-crime, cross validated	0.714	0.711	0.711
No DNA, no-crime, 75-25	0.723	0.720	0.719
No IO, CIU, cross validated	0.718	0.717	0.716
No IO, CIU, 75-25	0.750	0.750	0.749
Added state, cross validated	0.722	0.719	0.719

Added state, 75-25	0.731	0.730	0.730

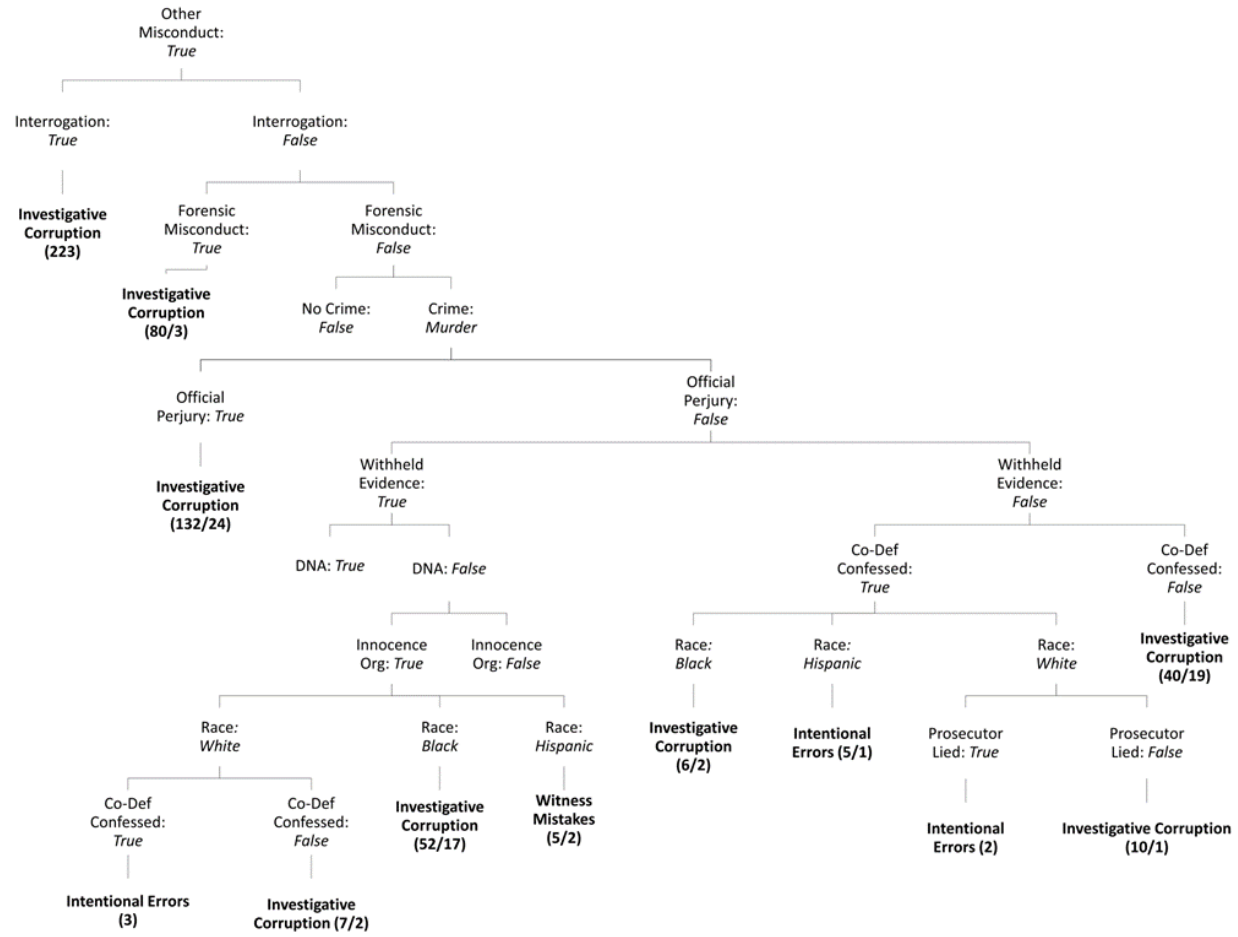
4. Discussion

Once these models are generated, a user can easily traverse them in the order displayed for the variables associated with a case, to examine similar cases and differences in classification. In *Figure 2*, we present an image of certain branches in our extended baseline model, with some sub-branches removed for ease of reading (for the entire tree, see Appendix A). Some of the case classifications are straightforward; if a case in the database has 'true' values for the variables 'Other Misconduct' (shortened from the NRE variable 'Misconduct That Is Not Withholding Evidence') and 'Interrogation' (or 'Misconduct in the Interrogation of the Exoneree'), it is immediately sorted into the Investigative Corruption class, with no differences in classification between the LCA and the decision tree.

There is more nuance deeper within the trees, where more variables must be known for classification. As an example, consider the presence of *race*, referring to the race of the exoneree at the end of many of the branches in *Figure 2*. The generalizations for racial groups depends on the value of 'Withheld Evidence' (referring to exculpatory evidence withheld in the original trial); when this variable is true, the variable 'Co-Defendant Confessed' determines classification, but when that same variable is false, 'DNA' (whether DNA evidence was a deciding factor in exoneration) and 'Innocence Org' (whether innocence organizations were involved in exoneration) are prioritized. Even at that point, *race* is considered before *co-defendants*.

Figure 2

Example Branch of Extended Model



Note. A zoomed-in view of a branch of the extended model (predicted latent classes are in bold). The first number in parentheses reflects cases that follow this branch; after the slash is the number of classification disagreements with the LCA classes, or cases whose assigned class doesn't match the decision tree's generalization.

A. Classification Disagreements

At the ends of these branches, the numbers in parentheses refer to the number of cases whose values follow that branch; when there is a second number after a slash, that number quantifies instances where the decision tree disagrees with the LCA for how to classify individual cases. Take, for example, the 'Withheld Evidence: True' branch that terminates with 'Investigative Corruption' (52/17). This means that, given cases with a Black exoneree, where an Innocence organization took part in exoneration and DNA was a substantial factor in exoneration, etc., there are 52 individual cases whose facts match this branch. However, only 35 of these cases are agreed to be members of the Investigative Corruption class; of the 17 items classified differently between the LCA and decision tree, the LCA categorized 15 cases as members of the Intentional Errors class, and 2 as Witness Mistakes. The fact that this branch does not continue to subdivide means that the decision tree algorithm has determined there is no further generalization that can distinguish between the classes. There may be, for example, some variable whose values are distributed evenly within the next branching, but because it does not add further diagnosticity, that branching is not included in the model's output.

B. Next Steps

Coping with Continuous Variables

In future models, we plan to consider how best to approach the inclusion of continuous variables. This may be informed by Gañan-Cardenas et al. (2022), given the similarity of their methods to our own. In their work, they identify coefficients used to measure dissimilarity in continuous variables. Our goal would be to reintegrate age and chronological dates using these measures to find commonalities between exonerees whose timelines are comparable.

Re-Assessing Latent Class Labels

A secondary goal of this project is to critically evaluate the latent class labels conferred by Berube et al. (2023) for their informativeness to human users (such as innocence organization intake staff members) in understanding the factorial patterns by which exonerations can be characterized. Toward this end, we might first consider whether the canonical factors listed in the NRE database provide the most useful basis for our analyses. An empirical investigation of the differences between wrongful conviction cases and cases in which the person charged with a crime escaped conviction was carried out by Gould et al. (2014). They argued that factors commonly associated with wrongful convictions such as police misconduct, false confessions, eyewitness misidentification, and reliance on jailhouse informants, should perhaps not be considered as "causal" factors, but rather as contributors. Their results pointed instead to age and criminal history of the person charged with a crime, punitiveness of the state, Brady violations, forensic error, weak defense, weak prosecution case, family defense witness, non-intentional misidentification, and lying by a non-eyewitness, as better candidates for "causal" factors (Gould et al., 2014; see also Acker & Redlich, 2019, pp. 20-21). However, an important distinction can be made in that the causal factors pointed to by Gould et al. (2014) are specifically relevant to processes that result in a wrongful conviction. Although informative for identifying wrongful convictions, the presence of any one of these factors may not be specifically predictive of whether a case will result in exoneration. Because the NRE contains only cases that resulted in successful exonerations, the six canonical factors in the NRE inherently lend themselves to more confident inferences about exonerations, as opposed to wrongful convictions in general.

As discussed in the *Introduction*, there are notable differences between the set of six “contributing causes” identified by the Innocence Project as of Aug 1, 2018 and the NRE’s six “canonical” contributing factors from 2,253 cases as of that date (Acker & Redlich, 2019, pp. 15-

16), with quite different percentages across the two distributions. Recall that the Innocence Project cases all involve DNA evidence (often from sexual assault), whereas the NRE set is much larger and more representative of wrongful convictions in general (ibid, p. 17). A benefit of our decision tree approach, constrained on the latent classes extracted from the six canonical factors in the NRE database, is that it can help determine which combination(s) of factors present in a wrongful conviction are likely to result in an exoneration, while at the same time prioritizing transparency.

The benefit offered by predicting latent class membership from covariates within the NRE database via decision trees hinges on the extent to which the latent classes themselves are easily distinguished, and thereby interpretable. For the sake of distinguishability, the so-called "Witness Mistakes" and "Failures to Investigate" classes appear to have suitably unique patterns. However, as noted in our results, the so-called "Investigative Corruption" and "Intentional Errors" classes display markedly similar patterns of underlying canonical factors. Accordingly, our post-hoc analysis of the LCA's posterior probabilities suggested that cases originally assigned to the Intentional Errors class in our LCA would be more often classified as Investigative Corruption, as opposed to the converse, in the subsequent decision trees. Indeed, this pattern was borne out by each of the models produced. It appears, therefore, that Intentional Errors cases may be harder to distinguish than Investigative Corruption cases.

This inference is consistent with Berube et al.'s (2023) assessment of differences between the Investigative Corruption and Intentional Errors classes. Based on correlational analyses of the covariates, they identified fewer discriminating factors for Intentional Errors than for Investigative Corruption. More specifically, Berube et al. (2023) suggested that federal and no-crime cases should be particularly indicative of Intentional Errors. While both of these variables

are present in the output trees, there are differences in their patterns of distribution. In our models, no-crime cases appear in the decision tree before federal cases, suggesting that the no-crime label has a higher discriminatory power. In our dataset, 640 of the 1039 cases categorized as Intentional Error are labeled as no-crime, for a rate of about 62%. However, while there are more individual cases with this label in the Intentional Error category, they make up a higher percentage of the Failure to Investigate category - 474 out of 678, for 70% (the percentages for Investigative Corruption and Witness Mistakes, respectively, are 20% and 0.2%). Conversely, the Intentional Errors category does have the highest raw count and percentage of federal cases, with 61 out of 1039 (about 6%). While this means that the label will more likely indicate an Intentional Error case, these are very small proportions of the dataset. This is reflected in the decision tree with no-crime cases being represented at an early point where this information can easily eliminate a case from a category, while federal cases are much later in the hierarchy, providing discriminatory power only when additional information has already been considered.

Berube et al. (2023) also pointed out that the second-highest rate of F/MFE was observed in the Investigative Corruption class, so F/MFE might be a helpful distinguishing factor. Our decision tree results supported this; the models in which forensic misconduct was included to predict Investigative Corruption showed high information gain. In Berube et al.'s (2023) LCA, , about 42% of juvenile defendants were assigned to the Investigative Corruption class, and in our reproduction, this was true for about 44% of juvenile defendants. Indeed, Intentional Errors are less easy to distinguish from Investigative Corruption than vice versa; the confusion matrices in Tables 3 through 5 highlight this, showing that the raw number of cases identified by the LCA as Intentional Errors but classified by the decision tree as Investigative Corruption is always higher than vice versa (although the degree of difference varies).

This underscores a benefit of the concurrent use of decision trees and LCA methods for understanding patterns in the NRE, particularly when the ultimate goal is to inform real-world intake decisions. Decision trees allow for a more nuanced window into how covariates may predict latent class membership, as opposed to inferences made through correlations alone.

Potential Uses

Our eventual aim is to understand how data-intensive methods could support post-conviction intake decisions. A qualitative study of 22 innocence organizations in 2011 found that organizations on average reviewed more than a thousand requests for every one successful exoneration (Krieger, 2011). Innocence organizations are often the last resort for wrongfully convicted applicants, and the organizations carefully consider each application with this in mind. When asked about their work practices, innocence organizations estimated that initial reviews of applications took around 21% of their time, and investigations, 50% (ibid). It can be difficult to decide whether to conduct an investigation without advance knowledge of the eventual outcome. Organizations burdened by the need to conduct excessive numbers of investigations actually achieved a lower rate of successful exonerations (ibid). Accordingly, Krieger (2011) recommended that future studies should focus on identifying patterns and trends of characteristics among cases that required serious investigation, in order to assist innocence organizations with reviewing new cases:

> "A future study should attempt to analyze all the cases seriously investigated (within a particular project or from many projects) to determine if particular characteristics or trends can be found that will help projects improve their selection of cases for serious investigation or review" (p. 378, footnote # 240).

Some innocence organizations are already using patterns of data from previous exonerations to help them identify cases with high likelihoods of success (Weintraub, 2022). However, in their raw form, these patterns of data are not easy to observe, nor easy to infer, through rote case-by-case examination. The use of a decision tree algorithm in this context extracts patterns that exist within the available data and presents such patterns in a fashion that is easily-readable, interpretable, nuanced, and transparent. Therefore, innocence organizations may refer to decision trees to inform and perhaps deepen their understanding of these patterns, such that they might be better equipped to identify cases with high likelihoods of success.

Innocence organizations may also use this framework as a training tool for law students undergoing internships/practicums, newly admitted lawyers working in post-conviction litigation, or newly-hired intake staff. For example, law students are an invaluable resource for innocence organizations in providing support to their applicants and their cases (Ricciardelli et al., 2012). These students spend the majority of their time screening applicant cases, which provides an increased educational benefit (Stiglitz et al., 2002). By having a better understanding of the pathways that lead to wrongful convictions, students may be better equipped to assess and apply their knowledge to these cases. Yet, certain critical case elements may be missing, or overlooked, in the initial legal proceedings of a criminal investigation (Findley & Scott, 2006). Accordingly, law students working at innocence projects through internships/practicums are often tasked with finding and collecting this information (Ricciardelli et al., 2012). A potential benefit of the decision tree framework is that it might make this process more efficient; it focuses on the factors most relevant to a particular applicant's case. An arduous information search could therefore be bolstered by efficient and communicable data-intensive methods.

C. Policy Implications

The analytical approach we outline here facilitates adding and removing variables in data-intensive models, which in theory could allow both users and policy-makers to better understand the models on which they rely for life-altering decisions. It could make it easier for policy-makers to audit and monitor algorithms for biases, especially for those that negatively impact vulnerable individuals (see Kalluri, 2020). It could shape the policies underlying decision-making by innocence organizations to be more efficient, as well as empower them to audit their own practices for bias if they wish (e.g., intake staff members are well aware that deciding to take on a client who has submitted a complete questionnaire is much easier than one for whom key information is missing, ambiguous, or incoherent). It could allow policy-makers to communicate about AI and data-intensive models with lawmakers (as well as with the general public) by demonstrating that removing a sensitive variable (such as race) from consideration by a model does not ameliorate the bias that can be created by proxies. It could allow stakeholders who may be injured or disadvantaged by the outcome of a particular algorithm to discover, document, and contest that decision. This sort of transparency is not present in black-box approaches such as deep learning and complex regressions (which cannot be explained even by their developers).

Bias is not simply a characteristic that exists "in" the algorithms and their training data; the emergence of biases (as well as their unintended consequences) depends on the context of use. Recall the Amazon algorithm that recommended qualified men but not qualified women for hiring (Dastin, 2018); if the context of use had simply been to find men to hire (reflecting the data patterns of the past), that algorithm would have been deemed successful. The point is that decision-support algorithms should be monitored and evaluated regularly, as the impacts can

change over time. In the domain of exonerations, such changes may include changes in the law, new developments in forensic techniques and caveats about reliability (e.g., Fabricant, 2022), evolving precedents about the use of predictive technologies, and public literacy about such technologies.

Moreover, policy should provide regulators with the tools and "teeth" to establish transparency baselines and standards in the use of AI/machine learning, even (or especially) from tech corporations and other powerful institutions who commonly claim that their data-intensive methods and algorithms are proprietary trade secrets. Any black-box methods should be linked to laws requiring accountability from those in power, as well as clear and available policies for stakeholders who wish to contest the decisions recommended by an algorithm. In another high-impact domain for the use of AI, healthcare systems, frameworks for ethical use have identified algorithm monitoring and deimplementation as a final phase in mitigating bias in an algorithm's life cycle (Chin et al., 2023). While agencies such as Health and Human Services (HHS) have begun to take steps towards regulatory monitoring of AI (HealthIT, 2023), similar transparency and regulatory power should be established to monitor and de-implement potentially harmful tools used in criminal justice decision making.

For the post-conviction context of use, the methods proposed here might be useful not only to innocence organizations, but also to CIUs (conviction integrity units) and others who review potentially wrongful convictions. This will require developing easy-to-use tools that users can understand when they explore large datasets—an area for further research.

D. Conclusion

Although data-intensive methods make promises of efficiency and accuracy, resulting decisions may be biased when an algorithm's inner workings are neither transparent nor

interpretable. Our approach uses LCA coupled with decision tree analysis on successful exoneration data; this reverse-engineering approach to intake data relies on patterns already present in the data to clarify trends within the latent class categorization and find further similarities between successful cases. These commonalities may be useful to determine what information would be needed for future post-conviction cases, while also directing resources for policy reform or educating staff in the use of data-driven frameworks.

5. References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317–332. https://doi.org/10.1007/BF02294359

Acker, J. R. & Redlich, A. D. (2019). *Wrongful Conviction: Law, Science, and Policy* (2nd ed.). Durham, North Carolina: Carolina Academic Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bedau, H. A., & Radelet, M. L. (1987). Miscarriages of justice in potentially capital cases. *Stanford Law Review*, *40*(1), 21–179. https://doi.org/10.2307/1228828

Berube, R., Wilford, M. M., Redlich A. D., & Wang, Y. (2023). Identifying patterns across the six canonical factors underlying wrongful convictions. *The Wrongful Conviction Law Review*, *3*(3), 166-195. https://doi.org/10.29173/wclawr82

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Int. Group, 37(15), 237-251

Brennan, T., & Dieterich, W. (2017). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). *Handbook of Recidivism Risk/Needs Assessment Tools*, 49–75. https://doi.org/10.1002/9781119184256.ch3

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparity in commercial gender classification. In Conference on Fairness, Accountability, and Transparency (pp. 77-91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., Jain, A., Jordan, W. B., Konya, S., Moore, R. H., Moore, T. T., Rodriguez, R., Shaheen, G., Snyder, L. P., Srinivasan, M., Umscheid, C. A., … Ohno-Machado, L. (2023). Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Network Open*, *6*(12). https://doi.org/10.1001/jamanetworkopen.2023.45050

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Dastin, J. (2018, October 10). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). John Wiley & Sons.

Fabricant, M. C. (2022). *Junk Science and the American Criminal Justice System.* Akashic Books.

Findley, K., Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review*, *2*, 291–397. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=911240#

Flach, Peter. 2012. *Machine Learning: the art and science of algorithms that make sense of data* (1st ed.). Cambridge University Press.

Frank, Eibe, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (4th ed.). Morgan Kaufman.

Gañan-Cardenas, E., Pemberthy, J. I., Rivera, J. C., & Mendoza-Arango, M. C. (2022). Operating Room Time Prediction: An Application of Latent Class Analysis and Machine Learning. *Ingeniería y Universidad*, *26*, 1-23. https://doi.org/10.11144/Javeriana.iued26.ortp

Gould, J. B., Carrano, J., Leo, R. A., & Hail-Jares, K. (2014). Predicting erroneous convictions. *Iowa Law Review*, *99*(2), 471-522.

Gross, S. R. (2008). Convicting the innocent. *Annual Review of Law and Social Science*, *4*(1), 173-192. https://doi.org/10.1146/annurev.lawsocsci.4.110707.172300

Gross, S. R., O'Brien, B., Hu, C., & Kennedy, E. H. (2014). Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20), 7230–7235. https://doi.org/10.1073/pnas.1306417111

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter, 26*(4), 237-243.

HealthIT.gov. (2023, December 20). *Health Data, technology, and interoperability: Certification Program updates, algorithm transparency, and information sharing (HTI-1) final rule*. https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program?et_rid=941383285&et_cid=5033060

Innocence Project. (2023). *Cases*. https://innocenceproject.org/all-cases/#exonerated-by-dna

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature, 583, 169. *doi: https://doi.org/10.1038/d41586-020-02003-2*

Kearns, M. & Roth, A. (2020). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Krieger, S. A. (2011). Why our justice system convicts innocent people, and the challenges faced by Innocence Projects trying to exonerate them. *New Criminal Law Review*, *14*(3), 333–402. https://doi.org/10.1525/nclr.2011.14.3.333

Loeffler, C. E., Hyatt, J. & Ridgeway, G. (2019). Measuring self-reported wrongful convictions among prisoners. *J Quant Criminol*, 35, 259–286. https://doi.org/10.1007/s10940-018-9381-1

McCutcheon, A. (2002). Basic Concepts and Procedures in Single- and Multiple-Group Latent Class Analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 56-86). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.003

Muthén B. O., Muthén L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical & Experimental Research*, 24(6), 882-891. https://doi.org/10.1111/j.1530-0277.2000.tb02070.x

National Registry of Exonerations. (n.d.) Retrieved April 26th, 2023 from https://www.law.umich.edu/special/exoneration/Pages/detaillist.aspx

O'Neil, C. (2016). *Weapons of Math Destruction*. New York: Broadway Books.

Quinlan, R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Ricciardelli, R., Bell, J. G., & Clow, K. A. (2012). "Now I see it for what it really is": the impact of participation in an innocence project practicum on criminology students. *Albany Law Review*, *75*(3), 1439–1466.

Schwarz, Gideon (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Stiglitz, J., Brooks, J., & Shulman, T. (2002). The hurricane meets the paper chase: Innocence projects new emerging role in clinical legal education. *California Western Law Review*, *38*:*2*(5), 413–431. https://scholarlycommons.law.cwsl.edu/cwlr/vol38/iss2/5

Stone, M. (1976). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society (Series B - Methodological)*, *36*, 111-133.

Turque, B. (2012, March 6). 'Creative... motivating' and fired. *The Washington Post*. https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/gIQAwzZpvR_story.html

Weintraub, Jennifer N. (2022). The Dark Figure of Wrongful Convictions: How Intake Decisions Impact Exonerations. Unpublished doctoral dissertation, the University at Albany, State University of New York, Albany, NY.

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *46*(4), 287-311. https://doi.org/10.1177/0095798420930932

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer New York.

West, E., & Meterko, V. (2016). Innocence project: DNA exonerations, 1989-2014; review of data and findings from the first 25 years. *Albany Law Review*, *79*(3), 717-795, 2015/2016. https://ssrn.com/abstract=2986970

6. Appendix A.

Expanded Decision Tree, Figure 2.

```
OTHER_MISCONDUCT = True
|   INTERROGATION = True: Investigative_Corruption (223.0)
|   INTERROGATION = False
|   |   FORENSIC_MISCONDUCT = True: Investigative_Corruption (80.0/3.0)
|   |   FORENSIC_MISCONDUCT = False
|   |   |   NO_CRIME = False
|   |   |   |   CRIME = Murder
|   |   |   |   |   OFFICIAL_PERJURY = True: Investigative_Corruption (132.0/24.0)
|   |   |   |   |   OFFICIAL_PERJURY = False
|   |   |   |   |   |   WITHELD_EV = True
|   |   |   |   |   |   |   DNA = True
|   |   |   |   |   |   |   |   PROSECUTOR_MISCONDUCT = True: Investigative_Corruption (28.0/2.0)
|   |   |   |   |   |   |   |   PROSECUTOR_MISCONDUCT = False
|   |   |   |   |   |   |   |   |   WINTESS_TEMP = True: Intentional_Errors (14.0/6.0)
|   |   |   |   |   |   |   |   |   WINTESS_TEMP = False: Investigative_Corruption (2.0)
|   |   |   |   |   |   |   DNA = False
|   |   |   |   |   |   |   |   INNOCENCE_ORG = True
|   |   |   |   |   |   |   |   |   RACE = White
|   |   |   |   |   |   |   |   |   |   CO_DEF_CONFESSED = True: Intentional_Errors (3.0)
|   |   |   |   |   |   |   |   |   |   CO_DEF_CONFESSED = False: Investigative_Corruption (7.0/2.0)
|   |   |   |   |   |   |   |   |   RACE = Black: Investigative_Corruption (52.0/17.0)
|   |   |   |   |   |   |   |   |   RACE = Hispanic: Witness_Mistakes (5.0/2.0)
|   |   |   |   |   |   |   |   INNOCENCE_ORG = False
|   |   |   |   |   |   |   |   |   PERMITTING_PERJURY = True
|   |   |   |   |   |   |   |   |   |   INTEGRITY_UNIT = True: Investigative_Corruption (12.0/4.0)
|   |   |   |   |   |   |   |   |   |   INTEGRITY_UNIT = False
|   |   |   |   |   |   |   |   |   |   |   JAIL_INFORMANT = True
|   |   |   |   |   |   |   |   |   |   |   |   CO_DEF_CONFESSED = True: Intentional_Errors (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   CO_DEF_CONFESSED = False: Investigative_Corruption (14.0/5.0)
|   |   |   |   |   |   |   |   |   |   |   JAIL_INFORMANT = False: Intentional_Errors (51.0/14.0)
|   |   |   |   |   |   |   |   |   PERMITTING_PERJURY = False
|   |   |   |   |   |   |   |   |   |   CHILD_VICTIM = True: Investigative_Corruption (12.0/3.0)
|   |   |   |   |   |   |   |   |   |   CHILD_VICTIM = False
|   |   |   |   |   |   |   |   |   |   |   PROSECUTOR_MISCONDUCT = True
|   |   |   |   |   |   |   |   |   |   |   |   JUVENILE_DEF = True: Intentional_Errors (7.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   JUVENILE_DEF = False: Investigative_Corruption (43.0/18.0)
|   |   |   |   |   |   |   |   |   |   |   PROSECUTOR_MISCONDUCT = False: Intentional_Errors (37.0/11.0)
|   |   |   |   |   |   WITHELD_EV = False
|   |   |   |   |   |   |   CO_DEF_CONFESSED = True
|   |   |   |   |   |   |   |   RACE = White
|   |   |   |   |   |   |   |   |   PROSECUTOR_LIED = True: Intentional_Errors (2.0)
|   |   |   |   |   |   |   |   |   PROSECUTOR_LIED = False: Investigative_Corruption (10.0/1.0)
|   |   |   |   |   |   |   |   RACE = Black: Investigative_Corruption (6.0/2.0)
|   |   |   |   |   |   |   |   RACE = Hispanic: Intentional_Errors (5.0/1.0)
|   |   |   |   |   |   |   CO_DEF_CONFESSED = False: Investigative_Corruption (40.0/19.0)
```