NeuLaw's Criminal Record Database
Datasheet

I. Motivation

I-A For what purpose was the dataset created?

According to the creators of the dataset, the dataset was created "To allow large-scale, cross-jurisdictional analyses of criminal arrests" and "enhance many types of research – for example, identification of high-frequency offenders, measurement of changes in policing strategies, and quantification of legislative efficacy – giving policy makers the best data upon which to base law enforcement decisions" [1].

I-B Who created the dataset?

Is it an official law enforcement or government body? An academic research team? Other?

The codebook lists Gabe Haarsma, Sasha Davenport, Pablo A. Ormachea David M. Eagleman as authors [2].

I-C Was there a specific task in mind, or gap that needed to be filled?

The dataset was created to improve on the information available from the UCR SRS program. Specifically, according to the creators, the advantages of this novel dataset include: (1) individual identifiers allow for recidivism analysis—albeit only for repeated bookings within the same jurisdiction (2) the presence of all the charges allows for deeper understanding of all crime, not just a subset, (3) more and different offender-specific variables than the UCR, (4) the data represent a comprehensive and growing picture of information available to judges and prosecutors, and (5) more and different disposition-specific variables, enabling assessment of small variations in punishment [1].

I-D Any other comments?

There maybe an updated version of this dataset that is not freely available, from here: hrefhttp://scilaw.org/risk-assessment/ http://scilaw.org/risk-assessment/.

II. Composition

II-A What do the instances that comprise the dataset represent?

For example: crimes, offenders, court cases, police officers

Records of criminal charges. The specific variables varies depending on the jurisdiction as described below.

II-B Are there multiple types of instances?

For example: offenders, victims, and the relationship between them.

Not within each jurisdiction.

II-C How many instances are there in total?

Of each type, if appropriate.

Harris County, TX: 3.1 million records, spanning from 1977 to April, 2012.
New York City, NY: 9.8 million records spanning from 1977 to 2013.

Miami-Dade County, FL: 5.7 million records spanning from 1971 to 2012.

II-D Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

For example, if it is traffic stops from a territory, is it all traffic stops conducted within that territory within a specific time? If not, is it a representative sample of all stops? Describe how representativeness was validated/verified. If it is not representative, please describe why.

The dataset includes all records from each jurisdiction, within the stated time frame. Some data instances were removed in pre-processing. In addition:

(1) The database contains no juvenile records, as those are not included in basic Freedom of Information Act requests. We note that juvenile is defined differently in each locale, so 17 year olds are included in Harris County records whereas only 18 year olds appear in New York City and Miami-Dade County records.

(2) The database does not include sealed or expunged records, as those are typically removed from the underlying county databases. It is likely that this disproportionately affects certain crime types (e.g., traffic offenses).

II-E What data does each instance consist of?

If there is a large number of variables, please provide a broad description of what is included.

In the **Harris County dataset**, each instance contains Information regarding the:
1. Offense: date, code, name, degree, bond amount at the time of arrest, category, broad category.
2. Defendant: unique ID, race, gender, DOB (mm/yyyy), height, weight, citizenship status.
3. Case: unique case ID, date filed, offense degree, case bond, case status.
4. Attorney: hired or assigned. 5. Grand jury: date, defendant present, and jury action code.
6. Disposition: date, plea, disposition (e.g., dismissed).

In the **New York City dataset**, each instance contains Information regarding the:
1. Offense: month, year.
2. Arrest: county, month, year, charge, crime category, broad crime category.
3. Defendant: race, gender, age at arrest.
3. Disposition: county, month, year, charge, disposition.

In the **Miami-Dade County dataset**, each instance contains Information regarding the:
1. Arrest: date, code, crime category, broad crime category.
2. Case: date filed, date closed, offense degree, trial type (Bench / Jury), case code, case status.
3. Defendant: race, gender, DOB (mm/yyyy).
4. Disposition: code, plea, disposition.

II-F Is there a target label or associated with each instance?
Please include labels that are likely to be used as target labels, e.g. recidivism.

There is not a pre-specified target label. However, disposition is most suitable to be used as a target label.

II-G Are there recommended data splits (e.g., training, development/validation, testing)?
If so, please provide a description of these splits, explaining the rationale behind them.

No.

II-H Does the dataset contain data on race and ethnicity?
If so, is it based on the individual's self-description, or based on officer's impression? Was it collected or derived in post-processing? For example, by name analysis.

Yes. For race, this originates from the raw data and it is not clear whether it is based on the individual's self-description.

The jurisdictions within the datasets do not identify offenders of Hispanic descent. To obtain a better understanding of the demographics, the creators have estimated the Hispanic population by last name [1].

II-I Are there any known errors, sources of noise, bias or missing data, or variables collected for only part of the datasets?
If so, please provide a description.

All the records in the database were originally entered by humans. The creators attempted to fix typographical errors. However, a larger problem is missing data. For example, some fields have become more populated with time. Birth date was not as commonly entered in some of the earlier records from the 1970s and 1980s, but becomes more rigorously entered with time 6 [1].

The dataset does not contain corrections records, as most states do not consider those public. Therefore, while we know each offender's sentence at the end of trial or plea bargaining, we cannot know how long an offender actually served [1].

II-J Does the dataset contain data on criminal history or other data that might be considered confidential or sensitive in any way?
For example: sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; If so, please provide a description.

The dataset contains partial information on criminal offending, as well as demographic information. The partial criminal offending can be constructed as the dataset contains unique identification numbers that can be linked across multiple offenses in an area. For example, in Harris County, Texas, 44% of the 1.2M uniquely identified offenders have multiple offenses – and therefore a partial record of offense (see Figure 1).

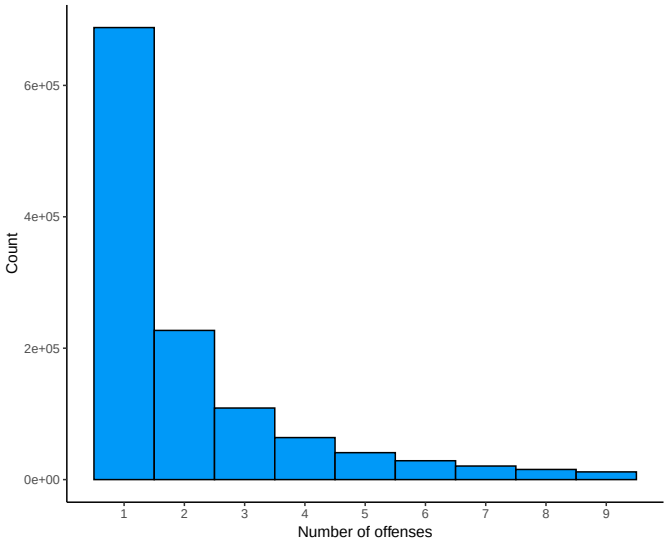

Fig. 1. A histogram of the number of offenses per offender in Harris County, Texas. The visualization has been limited to offenders that have less than 10 offenses for conciseness. Individuals with more than 10 offenses represent less than 3% of the dataset.

II-K Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
If so, please describe how.

Possibly, if comparing to other sources such as news articles. Only relevant for cases that attracted media attention.

III. Uses

III-A What type of tasks, if any, has the dataset been used for?
If so, please provide examples and include citations.

Examples of papers that have used this dataset are [3], [4], [5], [6].

III-B Is there a repository that links to any or all papers or systems that use the dataset?

No.

III-C What (other) tasks could the dataset be used for?
For example: testing predictive policing systems, predicting recidivism.

The dataset can be used for research questions around case disposition and sentencing. A partial criminal record can be constructed from the Harris County dataset.

III-D Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that might impact future uses?
For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The dataset only contains arrest data and not incident-based data, thus providing a picture of crime at the courthouse level. This means that previous stages in the law enforcement process (e.g., 911 calls, house calls, etc.) could skew the arrests that make it into courthouse databases [1].

The recidivism analysis allowed by this only applies for repeated bookings within the same jurisdiction. This approach will systematically undercount the true recidivism rate due to relocation [1].

The dataset does not have victim data, precluding the analysis of, for example, whether ethnicity or age of victim affects sentencing [1].

Some jurisdictions have more limited data than the rest. For example, New York City's records only list the most serious offense per arrest and do not yet include an identifier [1].

While our Broad categorization allows for comparisons across jurisdictions, the detailed categorization does not. The subcategories become populated only if the jurisdictions' labels or code citations provided enough detail [1].

IV. Collection Process

IV-A How was the data associated with each instance acquired?
e.g. the data collected survey, the raw data is routinely collected by the courts.

To acquire the underlying data, the dataset creators "contacted New York City (New York), Harris County (Houston), and MiamiDade County (Miami), to obtain copies of their criminal records from their justice information management systems. As public records, the data were obtained via Freedom of Information Act requests" [1].

IV-B Was the information self-reported?
If the data was self-reported, was the data validated/verified? If so, please describe how.

No. The data was derived from a dataset of criminal records used by respective local authorities. It was not collected for research purposes.

IV-C Who was involved in the data collection process?
Was this done as part of their other duties? If not, were they compensated?

The data was entered into the courts data systems by employs of the courts.

IV-D Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?
If not, please describe the timeframe in which the data associated with the instances was created. If the collection was not continuous within the timeframe, please specify the intervals, for example, annually, every 4 years, irregularly.

Harris County, TX – 1977 to April, 2012.
New York City, NY – 1977 to 2013.
Miami-Dade County, FL – 1971 to 2012.

IV-E Were any ethical review processes conducted (e.g., by an institutional review board)?
If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown. The dataset creators do state that "The Institutional Review Board at Baylor College of Medicine exempted this release of an anonymized dataset from human subject research oversight because they consist of publicly available records" [1].

IV-F Were the individuals in question notified about the data collection? Did they give their consent?
If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

It is likely the individuals know of their criminal charges. It is unlikely they knew or gave consent for it to be used as part of a research dataset.

IV-G Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown.

V. Pre-processing, cleaning, labeling

V-A Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, removal of instances, processing of missing values)?
If so, please provide a description and reference to the documentation. If not, you may skip the remaining questions in this section.

Yes. Data processing is described is detail in [1] and in the codebook [2]. Broadly, the data was cleaned and standardized, and duplicated entries were removed. Entries have

been de-identified by removing names, addresses, etc. DOB was replaced with the month and year only. In the Harris County dataset, defendants and cases were given a unique identifiers. The creators added seven calculated variables for all the datasets: 1. Broad crime category (32 categories), 2. Detailed crime category ($\sim 150 - 175$ categories) 3. Standardized disposition[1] 4. Gender, using given name to determine gender when missing or unknown. 5. Race, using surname to add Hispanic ethnicity. 6. The defendant age at the time of case filed or the arrest date. 7. The year the case is filed. 8. Aggregated case numbers to combine multiple offenses into single case (Harris County only).

V-B Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?
If so, please provide a link or other access point to the "raw" data.

Yes. The calculated age, race and gender variables are added to the dataset alongside the raw variables.

V-C Is the software that was used to preprocess/clean/label the data available?
If so, please provide a link or other access point.

No.

VI. Distribution

VI-A Is the data publicly available? How and where can it be accessed (e.g., website, GitHub)?
Does the dataset have a digital object identifier (DOI)?

Yes. The dataset can be found here [2].

VI-B Is the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

VII. Maintenance

VII-A Is the dataset maintained? Who is supporting/hosting/maintaining the dataset?

The dataset is not maintained. There maybe an updated version of this dataset that is not freely available, from here: hrefhttp://scilaw.org/risk-assessment/ http://scilaw.org/risk-assessment/.

VII-B How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Unknown.

[1]standardized disposition has 7 possible dispositions: No Action, Dismissal, Transfer, Acquittal, Guilty, Guilty by Plea, Conditional Dismissal, and Unknown/No Final Disposition

VII-C Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

No.

VII-D Are older versions of the dataset continue to be supported/hosted/maintained?

N/A.

VII-E If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
If so, please provide a description.

No.

References

[1] P. A. Ormachea, G. Haarsma, S. Davenport, and D. M. Eagleman, "A New Criminal Records Database for Large-scale Analysis of Policy and Behavior," *Journal of Science and Law — jscilaw.org*, vol. 1, no. 1, p. 2, 2015.

[2] D. M. Eagleman, "Neulaw Criminal Record Database."

[3] P. A. Ormachea, S. Davenport, G. Haarsma, A. Jarman, H. Henderson, and D. M. Eagleman, "Enabling Individualized Criminal Sentencing while Reducing Subjectivity: a Tablet-based Assessment of Recidivism Risk," *AMA Journal of Ethics*, vol. 18, no. 3, pp. 243–251, 2016.

[4] J. A. Bouffard and L. N. Askew, "Time-series Analyses of the Impact of Sex Offender Registration and Notification Law Implementation and Subsequent Modifications on Rates of Sexual Offenses," *Crime & Delinquency*, vol. 65, no. 11, pp. 1483–1512, 2019.

[5] M. N. Cooper, "What a Difference a Year Makes: An Examination of Prosecutorial Decision-making for Persons Under the Age of 18 in the Harris County, Texas Adult Criminal Justice System. Patterns and Predictors." Ph.D. dissertation, 2018.

[6] R. Pfeffer, P. Ormachea, and D. Eagleman, "Gendered Outcomes in Prostitution Arrests in Houston, Texas," *Crime & Delinquency*, vol. 64, no. 12, pp. 1538–1567, 2018.