MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] The original raw data files were sought by J. Kalven,
a journalist in the City of Chicago, as part of his investigation into police abuse. After the original FOIA requests and legal case, the non-profit Invisible Institute (https://invisible.institute) began to collaborate with Kalven and the University of Chicago's Mandel Legal Aid Clinic to follow up on earlier FOIA requests and to file new ones. The data disclosed in response to these earlier and now ongoing FOIA requests were made available online as part of the Citizens Police Data Project.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] The Chicago Police Department (CPD),
Civilian Office of Police Accountability (COPA), and the City of Chicago produced the raw data files in response to FOIA requests. The raw data were curated and released publicly by the Invisible Institute and its collaborators. The cleaned and linked data were produced as part of research by the authors of this document.

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3] The acquisition of the original raw data was funded by
the Invisible Institute.

[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] There are multiple types of instance
in this data.
- Officer: information about an individual police officer
- Unit assignment: a single unit assignment for an officer
- Complaint: a complaint filed against a police officer, either internally or by a civilian
- Tactical Response Report: a form that an officer is required to fill out after their response requires use of force
- Award request: a request to grant an award to an officer
- Salary: a record of an officer's salary, pay grade, and position across multiple years

[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] There are roughly 35,000 unique officers
in the cleaned roster appearing in roughly 130,000 profiles throughout the data, 730,000 award request records, 194,000 salary records, 108,000 unit assignment records, 109,000 complaints, and 10,500 tactical response reports.

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then

what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] This data contains information regarding all sworn officers in the Chicago Police Department / City of Chicago databases for the stated date ranges (which differ for each source of raw data).

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Officer: officer unique ID, race, gender, age, appointment date, resignation date, badge number(s), position title(s)

- Unit assignment: officer unique ID, start date, end date, unit number
- Complaint: complaint ID, involved officer IDs, allegation, result of the investigation, resulting sanction (where available)
- Tactical Response Report: report ID, event location, date, and time, environmental conditions, who was notified, weapons discharged, weapon information, subject demographic information
- Award Request: awardee unique ID, requester, request date, award reference number, award type, request tracking number, incident dates, ceremony date
- Salary: officer unique ID, salary, position title, pay grade, year

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] Not explicitly. However, labels could be constructed from the data that exists. For example, one could aggregate complaints to produce an integer “number of complaints” for each officer in the data, and use that as the response variable in a prediction task.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] In the original raw data files, missing data (of all fields) is quite common (see Appendix D). In the cleaned and linked data files, we are able to aggregate multiple profiles of a single officer appearing throughout the data to “fill in the gaps,” although this process is not perfect and there are still missing entries.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] In the raw data, no. In the cleaned data, we provide a unique officer identification that enables linking the activities and records regarding individual officers across datasets. There is no relational data (i.e., network edges) explicitly contained in the data. However, it is possible to use the data to construct a network, e.g., by linking officers co-listed on complaints.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] No, although the officer database is likely to be incomplete

prior to roughly 1980.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13]

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14]

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[answer15]

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16]

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17]

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18]

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, 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; criminal history)? If so, please provide a description.

[answer19]

[question20] Any other comments?
[answer20] No.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
[answer21]
[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
[answer22]
[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?
[answer23]
[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?
[answer24]
[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
[answer25]
[question26] Were any ethical review processes conducted (for example, 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.
[answer26]
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipB] NO
[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
[answer27] The
raw data was acquired from public links provided by the Invisible Institute (https://invisible.institute). The Invisible Institute acquired the data through FOIA requests made to the CPD and

the City of Chicago.

[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

[answer28] It is unknown whether the individual officers were notified by the CPD when the raw data was released.

[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

[answer29] Not explicitly.
The Chicago Police Department was compelled by law to produce these records per FOIA requests

[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

[answer30] Not applicable

[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.

[answer31] Not known.

[question32] Any other comments?

[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

[answer33] Yes; the main section of this documentation provides details the cleaning and linking of the raw data resulting from FOIA requests made to the City of Chicago.

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

[answer34] Yes; the raw data is available in the raw/ folder in the repository.

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

[answer35] Yes; the source for cleaning and linking is provided in the src/ folder in the repository.

[question36] Any other comments?

[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] Not the newly cleaned and linked version. The raw data itself has been used previously; see Section 5 for details.

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

[answer38] Not that the authors of this work are aware of.

[question39] What (other) tasks could the dataset be used for?

[answer39] This data set has a rich variety of possible uses; for example, network analysis (and in particular, analysis of dynamic events occurring on networks) and predictive regression/classification. See Section 5 for more details

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/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 (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40] Yes; the data are less reliable in earlier years (e.g., pre-1980). See Section 4 for more details.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] This data should not be used to single out, study, or identify individual officers.

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] Yes, the data is publicly available

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] It is available on GitHub at https://github.com/chicago-police-violence/data. Release versions will be marked using the "release" feature on GitHub.

[question45] When will the dataset be distributed?

[answer45] It is currently publicly accessible

[question46] Will 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.

[answer46] Yes; the source code is released under the MIT license, and the data output by the cleaning code is released under the Creative Commons 4.0 BY-NC-SA license.

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47] No.

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48] No.

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The repository will be hosted on GitHub.
As of August 2021, the repository owners are Thibaut Horel, Trevor Campbell, and Lorenzo Masoero, but ownership may change over time.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] Issue threads on GitHub are the primary channel of contact for the repository maintainers.

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] Not as of yet. For each major release version, notes will be included and hosted in the repository that will detail cleaning/linking errors that have been fixed.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] The original raw source data from FOIA requests will not be modified. More raw data files may be added over time corresponding to new FOIA requests. The data cleaning and linking code will be edited over time to fix errors; release versions will be clearly marked on GitHub. There is no set schedule for updates

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

[answer54] No; this data was released per FOIA requests and is in the public domain.

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

[answer55] Yes; a full version-controlled history of the project exists on GitHub.

[question56] 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. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56] Yes; the repository for the dataset is hosted on GitHub, where pull requests are a usual channel for external contribution.

[question57] Any other comments?

[answer57] No.