Penn
University of Pennsylvania

Landscape of Criminal Justice Data Sources

Greg Ridgeway
Rebecca W. Bushnell Professor
Department of Criminology
Department of Statistics and Data Science

ASA March 2025

Criminal Justice Data Sources

- Federal sources
  - FBI/CJIS
  - Bureau of Justice Statistics
  - US Sentencing Commission
- City open data portals
- Special data collections
  - Pennsylvania Sentencing Commission
  - North Carolina convictions
  - Florida court data
  - New Jersey use-of-force data
  - Criminal Justice Administrative Records System (CJARS)

Federal sources

FBI/CJIS, BJS, USSC

Uniform Crime Report

https://cde.ucr.cjis.gov

- The nation's crime data collection for almost 100 years
  - Summary Reporting System (SRS) – monthly counts of crimes reported to law enforcement
  - National Incident-Based Reporting System (NIBRS) – detailed incident level information
  - Law Enforcement Officers Killed and Assaulted (LEOKA) – Officers killed or assaulted while on duty
  - Supplementary Homicide Report – details on homicides starting in 1961

Participation is voluntary

Summary Reporting System reports counts

- Crime counts are broken down by
  - Law enforcement agency (ORI)
    - Several ORIs can cover the same geography
  - Month
  - Crime type
    - Part I crimes: murder, rape, robbery, aggravated assault, burglary, motor vehicle theft, larceny/theft, arson
    - Part II crimes: simple assault, fraud, weapons, drug violations, drunk driving, various vices, vandalism, and others
- Notes for working with UCR (NACJD files)

  https://github.com/gregridgeway/R4crim

Transition to NIBRS

- After 90 years, SRS ended in 2020
- All reporting required to go through NIBRS, insisting on incident-based reporting
- Substantial gap in national data between 2020 and 2024
  - 2012
    - 32 states participated (no CA, FL, NY, PA)
    - 28% of all crime reported through NIBRS
  - 2024
    - All 50 states participate
    - 82% of population covered
  - Some states have a long history
    - CO, DE, ID, KY, MI, MT, ND, RI, SC, TN, VT, VA

State	% population covered
CA	64%
FL	42%
PA	43%
NY	70%

Working with NIBRS

https://github.com/gregridgeway/R4crim

- Six primary tables, all linkable
  - Administrative
  - Offenses
  - Property
  - Victims
  - Offender
  - Arrestee

NIBRS includes wide range of "Group A" offenses

- Arson
- Assaults
- Bribery
- Burglary/B&E
- Counterfeiting/Forgery
- Destruction/Damage/Vandalism of Property
- Drug/Narcotic Offenses
- Embezzlement
- Extortion/Blackmail
- Fraud Offenses
- Gambling Offenses
- Homicide Offenses (including justifiable)
- Human Trafficking
- Kidnapping/Abduction
- Larceny/Theft Offenses (including retail theft, theft of car parts)
- Motor Vehicle Theft
- Pornography/Obscene Material
- Prostitution Offenses
- Robbery
- Sex Offenses (Forcible)
- Sex Offenses (Non-Forcible)
- Stolen Property Offenses
- Weapon Law Violations

“Group B” offenses reported only when there is an arrest

- **Bad Checks**
- **Curfew/Loitering/Vagrancy Violations**
- **Disorderly Conduct**
- **Driving Under the Influence**
- **Drunkenness**
- **Family Offenses, Non-Violent**
- **Liquor Law Violations**
- **Peeping Tom**
- **Trespassing**

Important notes

- Raw NIBRS data are in fixed-width format (500Mb) and all tables are interleaved

```
0150AK0010200CT0B-A39728N20221130 06010010101    N                          N
0250AK0010200CT0B-A39728N2022113023CCN  24          88
0350AK0010200CT0B-A39728N20221130519000001900 20221130
0350AK0010200CT0B-A39728N202211307190000019 00
0450AK0010200CT0B-A39728N2022113000123C                  B
```

- Follow up data (e.g. property recovered, arrest) censored at the end of March of the following year
  - January average time to arrest = 11.8 days
  - December average time to arrest = 2.9 days
- Codebook has many important details
  - Value of property stolen = 1 means “unknown”
  - Same property can appear in multiple entries (stolen, recovered)

National Crime Victimization Survey (NCVS)

https://github.com/gregridgeway/R4crim

- Nationally representative survey
  - Conducted by the U.S. Census Bureau
  - Asks about the number and characteristics of crime victimizations they experienced during the prior 6 months
- In 2022 reached 185,616 persons in about 152,794 households
  - 67% response rate for households
  - 82% response rate for individuals
- Households remain in the sample for 3½ years
  - Eligible respondents interviewed every 6 months, in person or by phone, for a total of seven interviews
- Identifiable/linkable data available at the Federal Statistical Research Data Centers

Data at NACJD back to 1992

https://www.icpsr.umich.edu/web/NACJD/series/95

- Since survey window is prior 6 months, need two "collection year" datasets to cover a calendar year
  - Example: to obtain 2024 estimates need to wait until 2025 data posted since respondents continue to report 2024 victimizations through June 2025
- Three files are of primary interest
  - Household-level file
    - Details on location, type of residence, family structure
  - Person-level file
    - Age, race, sex, marital status, education
  - Incident-level file
    - Where, when, who, what
    - Includes data on self-defense

Important NCVS notes

- Variables may change from year to year
- Spelling and punctuation of categorical levels can change from year to year
- If conducting “calendar year” analysis, need to modify sampling weight to match BJS reports

```
dataInc <- dataInc |>
  # drop non-US incidents
  filter(V4022!="(1) Outside U.S.") |>
  # cap series incidents at 10
  mutate(V4016 = case_when(
    V4019=="(2) No (is series)" & V4016>=11 & V4016<=996 ~ 10,
    V4016 >= 997 ~ NA,
    .default=V4016)) |>
  # scale collection year weight by series count
  mutate(WGTVICDY=case_when(
    V4019=="(2) No (is series)" ~ WGTVICCY*V4016,
    .default = WGTVICCY))
```

- Then make sure to use weights in all calculations

US Sentencing Commission posts case-level data

https://www.ussc.gov/research/datafiles/commission-datafiles

- Includes data on individuals (and organizations) convicted and sentenced in federal court
- Does not include data on defendants found not guilty of all charges (or who had all charges dropped)
- Case-level data on individuals include
  - Offender demographics
  - Criminal history
  - Primary offenses
  - Sentencing decisions
  - Enhancements and departures from guidelines
- Easily accessible data are from 2002-2023

Impact evaluation of the 2010 Fair Sentencing Act on crack convictions

- Fair Sentencing Act (FSA) (2010)
  - Reduced crack-to-powder sentencing disparity from 100:1 to 18:1
  - Eliminated 5-year mandatory minimum for simple possession
- USSC's 2015 report to Congress concluded
  - Number of crack cocaine convictions decreased by half
  - Average sentence for dropped from 106 months to 71 months, while no change in average sentence for powder cocaine
    - FSA reduced 15,000 person-years of incarceration
  - Little change in the race distribution of trafficking convictions
  - Little change in the role convicted dealers had in crack distribution

Local open data portals

Los Angeles

Philadelphia

City Open Data Portals

- Los Angeles
  - Crime data (2010-present)
  - Arrest data (2010-present)
  - Traffic collisions (2010-present)
- Useful for studying local effects
  - Gang injunctions in Los Angeles reduced crime by 5% in the short term and 18% in the long term, mostly due to reducing assaults

    Ridgeway, Grogger, Moyer, & MacDonald (2019). "Effect of gang injunctions on crime: A study of Los Angeles from 1988–2014," *Journal of Quantitative Criminology*, 35(3), 517–541
  - Rail transit expansion in Los Angeles had no effect on crime

    Ridgeway & MacDonald (2017). "Effect of rail transit on crime: A study of Los Angeles from 1988–2014," *Journal of Quantitative Criminology* 33(2), 277–291
- Quarterly aggregated crime reports back to 1988
  - https://github.com/gregridgeway/LAPDcrimedata

Local datasets can have anomalies

Crime counts by year and LAPD Area

District	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1	7146	7175	8084	7598	8407	10071	11240	11947	13224	13606
2	8707	8444	8626	8148	8557	9004	9531	9662	9689	9567
3	13659	12934	13126	12738	13014	13390	14699	14324	14395	13549
4	7406	6542	7085	6712	6916	7771	9545	9078	8829	8384
…										
18	11080	11060	10612	10197	10574	10824	11854	11888	12012	11764
19	10566	10553	10605	10301	9977	7981	22026	10891	10208	9199
20	8764	7988	8544	8305	9164	3	21202	11173	11038	10282
21	9919	9113	8938	9110	8799	0	22065	10841	10601	9709

Philadelphia Open Data portal

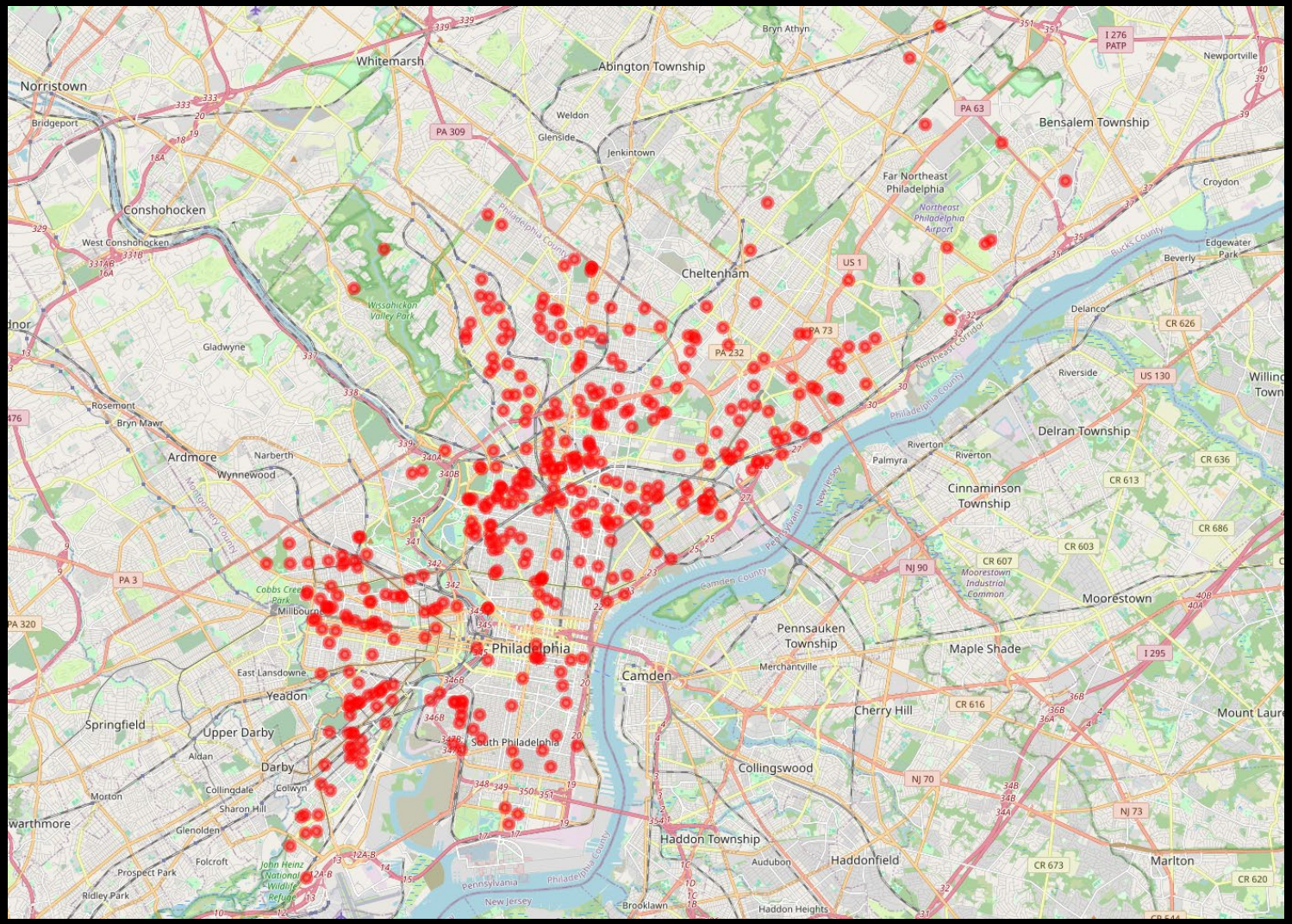
- Crimes reported to police
- Complaints against police
- Car crashes
- License & Inspection violations
- Data on shooting victims (including police shootings), age, race, sex, date, time, injuries, outcome
- Terry stop data (“stop and frisk”)
- DA data (daily counts of bail, charges, etc)

API access, frequent updates, and incident-level data are benefits

Philadelphia Example

```
library(jsonlite)
library(sf)
library(leaflet)

base_url <- "https://phl.carto.com/api/v2/sql?q="

sql_query <-
"SELECT dc_key, dc_dist, psa, dispatch_date, dispatch_time, location_block,
        ucr_general, text_general_code, point_x, point_y
 FROM incidents_part1_part2
 WHERE dispatch_date LIKE '2025-%' AND
       text_general_code='Aggravated Assault Firearm'"

full_url <- base_url |>
   paste0(sql_query) |>
   URLencode()
df <- fromJSON(full_url)$rows |>
   filter(point_x!=0 & point_y!=0) |>
   st_as_sf(coords = c("point_x", "point_y"), crs = 4326)

leaflet(df) |>
   addTiles() |>
   addCircleMarkers(radius = 5, color = "red", fillOpacity = 0.5,
   popup = ~paste("Incident Key:", dc_key, "<br>",
                  "District:", dc_dist, "<br>",
                  "PSA:", psa, "<br>",
                  "Date:", dispatch_date, "<br>",
                  "Time:", dispatch_time, "<br>",
                  "Location:", location_block, "<br>",
                  "UCR:", ucr_general, "<br>",
                  "Description:", text_general_code))
```

API access, frequent updates, and incident-level data are benefits



Other unique sources

Pennsylvania Sentencing Commission

North Carolina convictions

Florida court data

New Jersey use-of-force

Pennsylvania Sentencing Commission

https://pcs.la.psu.edu/research-data/request-data-and-reports/

- Complete data on sentences for all convictions since 2001
  - Offense charges (lots of details)
  - Plea bargain/trial
  - Judge identifier
  - Sentence terms (fine, community service, probation, jail, prison)
- Detailed codebook at website

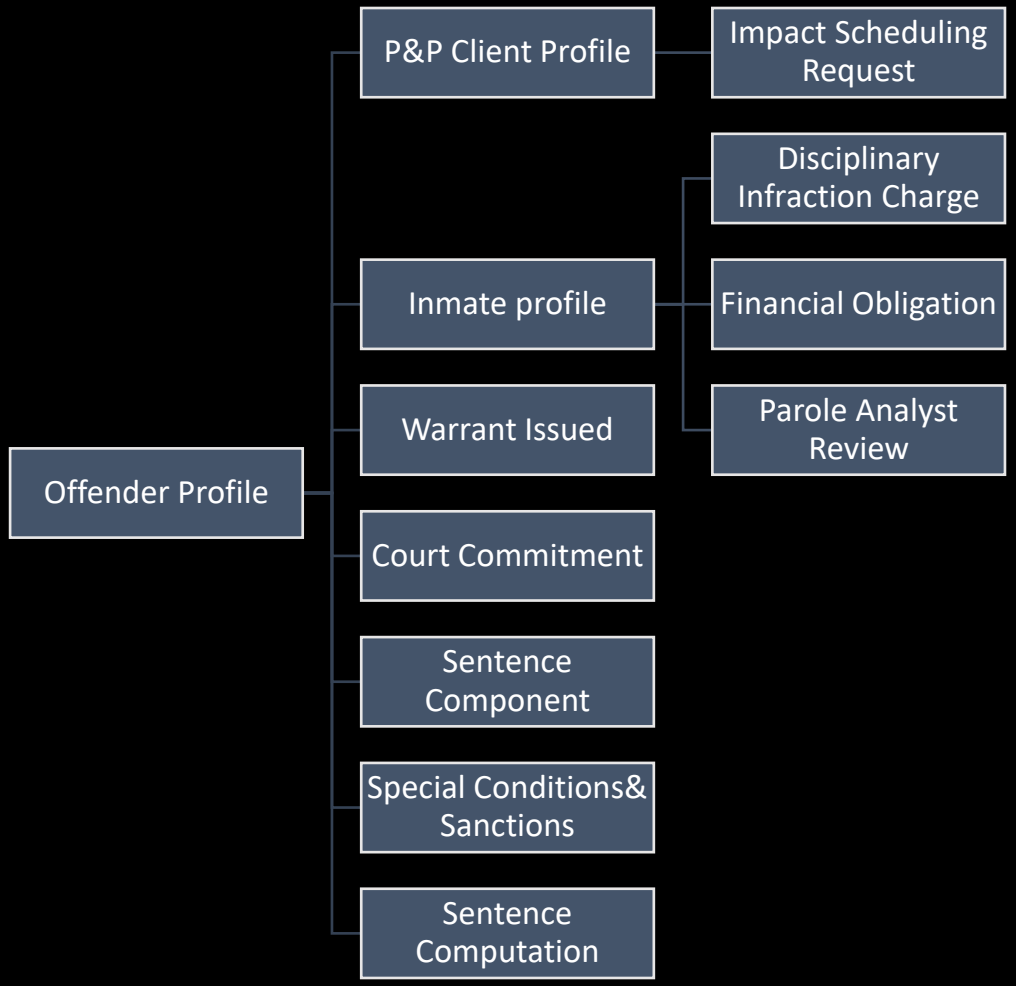
We benchmarked sentencing habits of PA judges

V. Nguyen & G. Ridgeway (2023). "Judges on the Benchmark: Developing a Sentencing Feedback System," Justice Quarterly 41(1):1–37

- Judges receive feedback when up for retention vote every 10 years
- Used detailed facts about cases (1,003 unique offense charges, 83 mandatory minimum codes) to customize a sentencing benchmark for 424 PA judges
- Judges with a high propensity to incarcerate
  - imposed a custodial sentence a rate 22-percentage points higher than their benchmarks
  - average sentence length that is 118% longer
  - thousands of individuals arbitrarily receiving longer and harsher sentences

North Carolina Corrections Data

https://webapps.doc.state.nc.us/opi/downloads.do?method=view



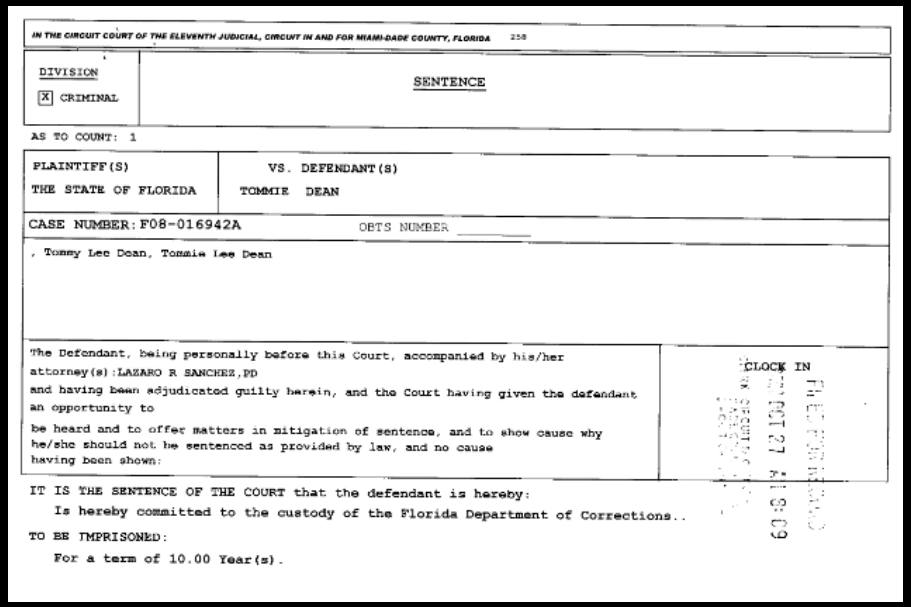
- NC posts data on all adult convicted offenders since 1972
- Identifiers, demographics, citizenship, conviction offense, sentence, and much more

Florida Court Data

https://acis.flcourts.gov/portal/home

- Each county's Clerk of the Court has data access

https://www.stateofflorida.com/clerks-of-court/



IN THE CIRCUIT COURT OF THE ELEVENTH JUDICIAL CIRCUIT IN AND FOR MIAMI-DADE COUNTY, FLORIDA 258

DIVISION
[X] CRIMINAL

SENTENCE

AS TO COUNT: 1

PLAINTIFF(S)
THE STATE OF FLORIDA

VS. DEFENDANT(S)
TOMMIE DEAN

CASE NUMBER: F08-016942A OBTS NUMBER

, Tommy Lee Dean, Tommie Lee Dean

The Defendant, being personally before this Court, accompanied by his/her attorney(s): LAZARO R SANCHEZ,PD
and having been adjudicated guilty herein, and the Court having given the defendant an opportunity to
be heard and to offer matters in mitigation of sentence, and to show cause why he/she should not be sentenced as provided by law, and no cause having been shown:

CLOCK IN

IT IS THE SENTENCE OF THE COURT that the defendant is hereby:
Is hereby committed to the custody of the Florida Department of Corrections..

TO BE IMPRISONED:
For a term of 10.00 Year(s).

- Data on DOC inmates

https://pubapps.fdc.myflorida.com/OffenderSearch/Search.aspx

DC Number:	B06320
Name:	DEAN, TOMMIE L
Race:	BLACK
Sex:	MALE
Birth Date:	01/07/1988
Custody:	MEDIUM
Release Date:	07/12/2023

Date In-Custody	Date Out-Custody
10/31/2008	03/08/2017
05/09/2018	07/12/2023

O. Mitchell, D. Mora, T. Sticco & L. Boggess (2022). "Are progressive chief prosecutors effective in reducing prison use and cumulative racial/ethnic disadvantage? Evidence from Florida," *Criminology & Public Policy* 21:535–565

New Jersey Office of the Attorney General Use-of-Force data

https://njoag.app.box.com/s/upgf6yyi9g0fyjg6ednhqmr69dg9crfh

- Incident level data on all use-of-force incidents in New Jersey between 10/2020 and 01/2025
- Officer name, age, race, sex, department, date, video available (Y/N), incident type, type of force, subject age, race, sex, injuries to officers and subjects
- Percentage of police shootings that are fatal: 44%
- Race distribution of subjects by force

Race	OC spray	Taser
Black	50%	36%
Hispanic	20%	19%
White	23%	39%

Criminal Justice Administrative Records System (CJARS)

https://cjars.org

- A University of Michigan and Census Bureau collaboration
  - collects longitudinal records from criminal justice agencies
  - harmonizes these records to track a criminal episode across all stages of the system
  - linkable at the person-level to social, demographic, and economic information from national survey and administrative records
  - Tends to have data at the prison/parole end of the justice system, less data at the arrest end
- Requires a proposal and access to a Federal Statistical Research Data Center

Summary

- A lot of publicly available data on policing and sentencing
- Police are generally the most open with data
  - If data are not publicly available, contact the police… they might just give it
  - Some prefer FOIA requests
- Courts are generally the most difficult
  - Judges least likely to make data available even though court proceedings are open to the public
  - Some state laws/policies make access difficult (PA CHRIA), cumbersome (NY DCJS), or incredibly easy (FL Sunshine laws)

Penn
University of Pennsylvania

Landscape of Criminal Justice Data Sources

Greg Ridgeway
Rebecca W. Bushnell Professor
Department of Criminology
Department of Statistics and Data Science

ASA March 2025