



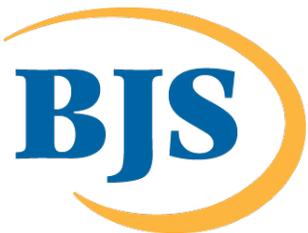
U.S. Department of Justice  
Office of Justice Programs

*Presented by*  
**Ryan Kling, Abt Associates**  
**Christopher Cutler, Abt Associates**  
**Mark Motivans, Ph.D., BJS**

# Record Linkage of Bureau of Justice Statistics (BJS) Federal Criminal Case Processing Data

FCSM Research and Policy Conference

March 9, 2018



**BUREAU OF JUSTICE STATISTICS**

---

# Outline

- Brief overview of the Federal Justice Statistics Program (FJSP)
- Data linking methodology
  - Examples
  - Diagnostics
- Conclusion

# Federal Justice Statistics Program (FJSP)

- Created by BJS in 1982 to serve as national clearinghouse of federal criminal case processing data
- The *National Justice Improvement Act of 1979* authorized BJS to: “collect, analyze and disseminate comprehensive federal justice transaction statistics ... and to provide technical assistance to and work jointly with other federal agencies to improve the availability and quality of federal justice data.” (42 U.S.C. § 3732 (15))

# FJSP data sources

- Administrative data are provided to BJS by six federal justice agencies spanning the stages of federal criminal justice (arrest to imprisonment)

## STAGE/DATA SOURCE

### Arrest

- **U.S. Marshals Service:** Contains data on suspects arrested for Federal offenses and booked by U.S. Marshals Service.
- **Drug Enforcement Administration:** Contains data on suspects arrested by DEA agents

### Prosecution

- **Executive Office for U.S. Attorneys:** Contains information on the investigation and prosecution of suspects in criminal matters received and concluded and criminal cases filed and terminated.

### Pretrial Release

- **Administrative Office of the U.S. Courts:** Contains data on defendants interviewed, investigated, or supervised by pretrial services.

### Adjudication

- **Administrative Office of the U.S. Courts:** Contains information about defendants whose cases were filed in U.S. district courts (felony defendants and Class A misdemeanants).
- **U.S. Sentencing Commission:** Contains information on criminal defendants sentenced pursuant to the provisions of the Sentencing Reform Act of 1984.

### Appeals

- **Administrative Office of the U.S. Courts:** Contains information on criminal appeals filed and terminated in U.S. Courts of Appeals.

### Corrections

- **Administrative Office of the U.S. Courts:** Contains information about defendants under Federal supervision (probation, parole, and supervised release).
- **Federal Bureau of Prisons:** Contains data on offenders admitted, released and present in the Federal Bureau of Prisons at fiscal year-end.

# Standard Analysis Files (SAFs)

- The goal of the FJSP is to provide uniform case processing statistics across all stages of the federal criminal justice system, including arrest, prosecution, pretrial release, adjudication, sentencing, incarceration, and supervision
- Data files are received in a variety of formats and are converted to Standard Analysis File (SAFs)
  - Standard unit of count (person-case)
  - Common reporting period (October 1 – September 30)
  - Common offense classification
  - Standardized disposition and sentence outcomes
- These steps “standardize” the data to improve comparability

# Data standardization

- SAFs are the “building blocks” of the FJSP
  - Annual cross-sectional data files are created at each stage for each cohort (in, out, and stock)
  - SAFs are used for statistical reports, a web query tool, and data are archived for researcher access at the University of Michigan’s National Archive of Criminal Justice Data (NACJD)
  - Data series begins with 1994 data

# Data linking methodology

- SAFs were first used to link person-case records across stages (arrest to imprisonment) in 2002
- The first generation of the link system used the one set of identifiers that were available across participating agencies:
  - Federal judicial district (there are 94)
  - Court docket number
  - Name of defendant
- A key was created for each record in each cohort based on this combined information and all cohort keys across all years were matched in a pair-wise fashion (all-to-all)

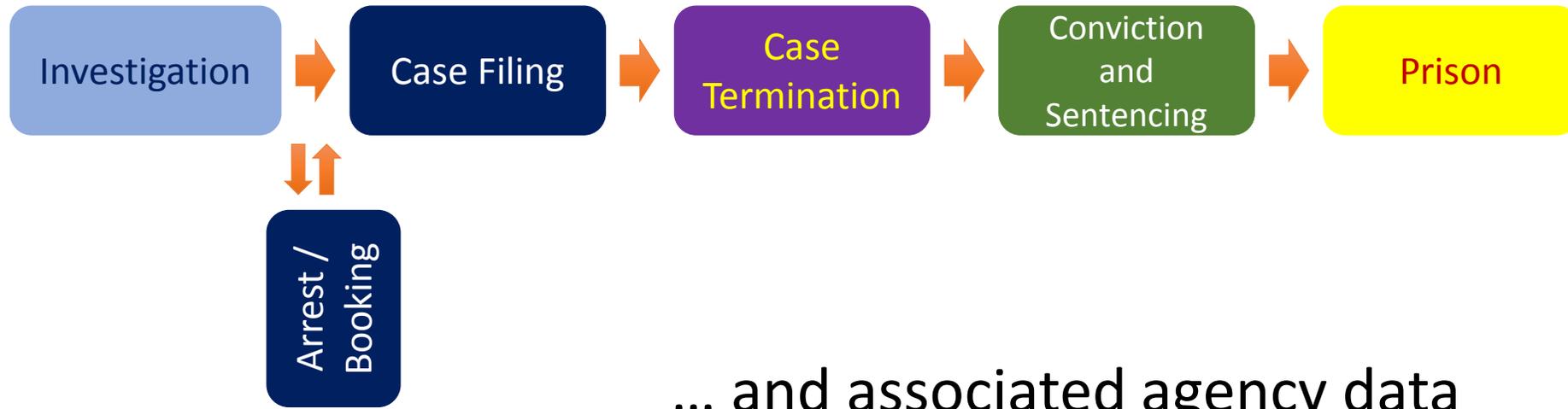
# Shortcomings of the first generation data linking method

- Shortcomings:
  - Non-directional
    - A match could be made with a record irrespective of the stage in which it appeared (e.g., out of case processing order)
  - No screening rules were used and this made it difficult to gauge link quality. It was difficult for users to distinguish false negatives from true negatives (matches that should have been found but weren't from matches that should never have been made)
  - Link file was cumbersome to use
    - All links across all stages and all years were stored in one very large index file
  - The link approach did not maximize use of all available identifiers

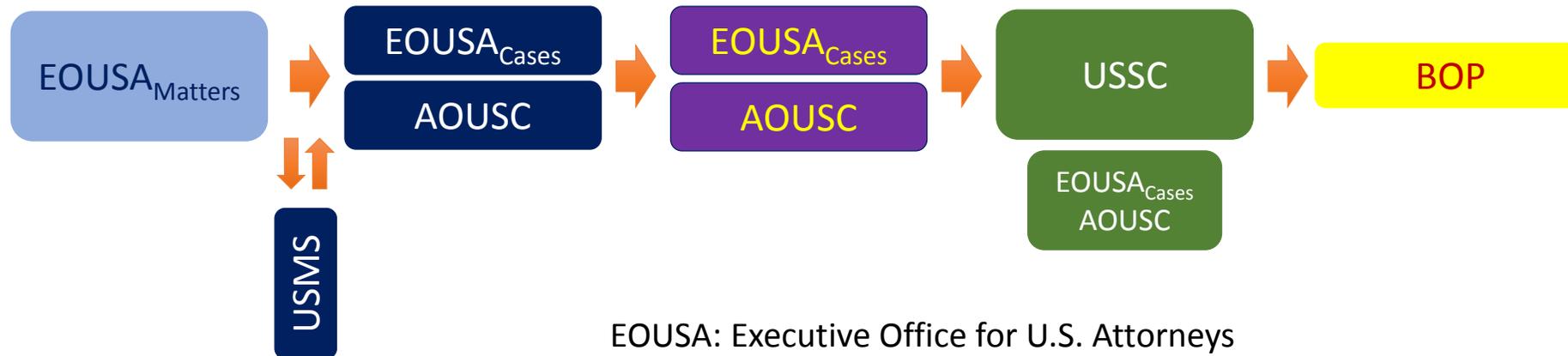
## Second generation data linking method

- The second generation link system was released in 2010 and sought to improve upon short-comings. This led to improved link rates
  - Designed as a dyad-based system -- that is, links are established between two agency SAFs at a time (from adjacent stages or dyads)
  - More criterion variables are used for each agency-to-agency link
    - Maximizes identifiers available in both agencies' data
  - Links based on a conditional subset of records
    - Records that will not link to next stage are dropped (e.g., cases dismissed won't match with a sentencing record and are dropped)

# Connecting the federal criminal case process



... and associated agency data



EOUSA: Executive Office for U.S. Attorneys

USMS: U.S. Marshals Service

AOUSC: Administrative Office of the U.S. Courts

USSC: U.S. Sentencing Commission

BOP: Federal Bureau of Prisons

# PII availability in the FJSP data

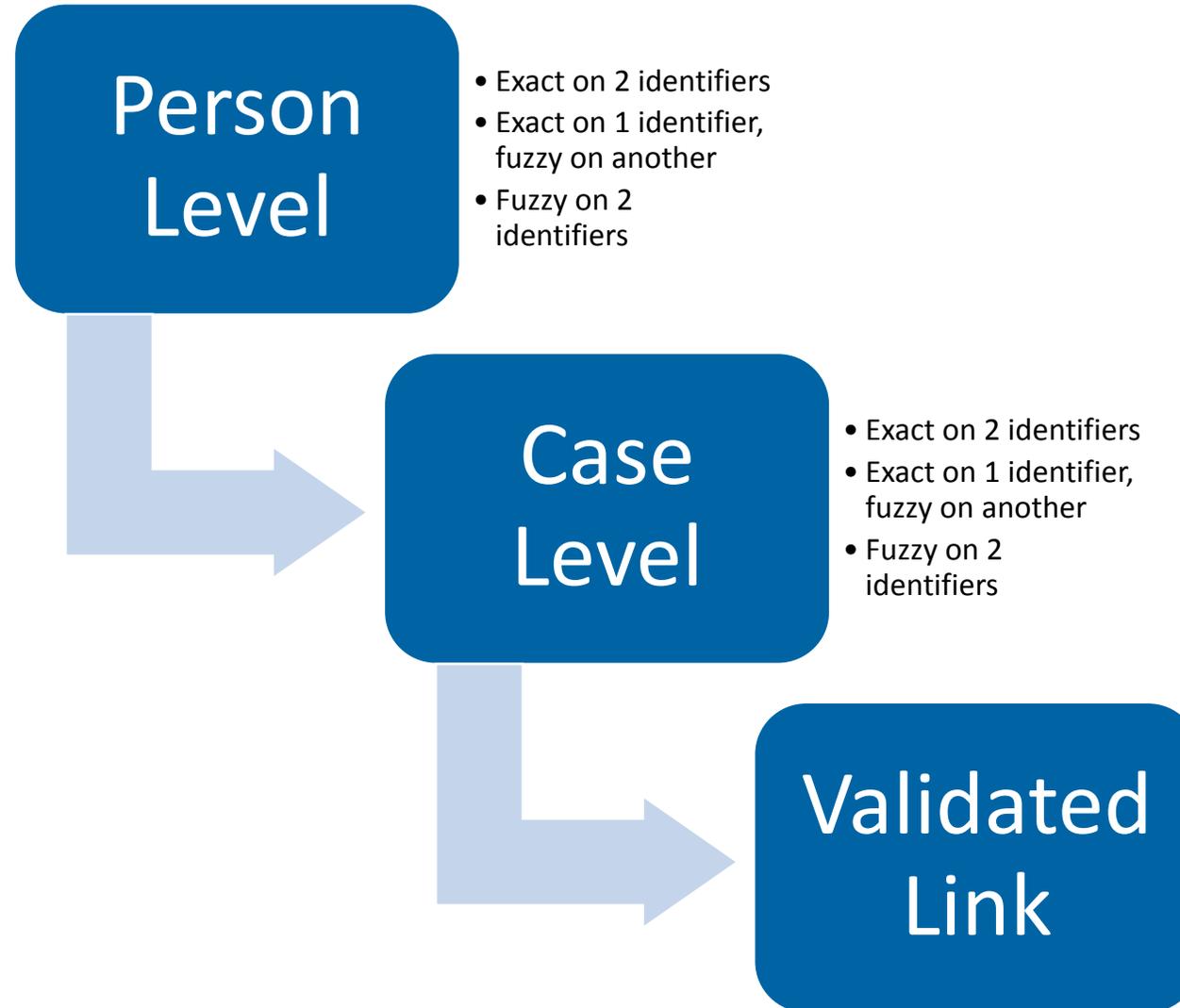
- Person-level identifiers in the FJSP available for linking (%), 2014

Stage	SSN	FBI Number	Register Number	Name	Date of Birth	DHS A-Number	Case Docketing Number
Arrest/Booking	42%	93%	100%	100%	100%	51%	84%
Investigation	0	24	15	100	45	0	83
Case Filing	0	35	78	100	76	0	99
Case Termination	0	0	0	96	0	0	100
Convictions and Sentencing	65	95	92	100	100	32	100
Prison Admission	67	100	100	100	100	39	95

# Case-level linking variables

- Federal judicial district and filing year
- Important dates
  - Arrest/booking
  - Investigation start and end
  - Case filing
  - Case termination
  - Conviction
  - Sentencing
  - Prison admission
- Other shared variables (e.g., judge; offense charge)

# Link strategy



# Challenges to data linking

- **Challenges: Validating identifiers**
  - **SSN:** 9 digits; can have leading a 0; remove “-”
  - **FBI number:** Up to 9 characters, two alphabetic in the 3<sup>rd</sup>- and 2<sup>nd</sup>-to last position, numbers elsewhere
  - **Register number:** 8 digits, can have 0-2 leading zeroes
  - **A-number:** 9 digits, leading A – frequently no A, leading zeroes often missing
  - **Docket number:** Circuit/district/filing year/docket/defendant – Not standardized within or across agencies



# Challenges (cont.)

- Challenges: Standardizing names

Type of Name	Middle Names	Other challenging name formats	
Colloquial	Anthony Michael Hall	Miguel Cabrera	Jackie Chan
Given	Anthony Michael Hall	Jose Miguel Cabrera Torres	Chan Kong-sang
Some Possible Variants	Anthony Michael Hall Hall, Anthony Hall, Anthony Michael Michael Hall, Anthony	Jose Miguel Cabrera Torres Miguel Cabrera-Torres Miguel Cabrera Torres Cabrera Torres, Miguel Cabrera Torres, Jose Miguel Cabrera, Miguel Torres Cabrera, Jose Miguel Torres	Kong sang Chan Kong-sang Chan Kong Chan

# Data linking example #1

- Cases termination <-> Conviction and sentencing example
  - Merge case terminations to conviction and sentencing
  - Linking hierarchy
    - Exact district/office/year/docket; exact sentencing date
      - Exact name; judge; defendant number
      - Exact name
      - No other qualifiers (for sealed cases)
    - Exact district/office/year/docket; defendant number; fuzzy match on either first or last name
    - Exact district/office/year/docket; fuzzier match on either first or last name
    - Block on filing year and last 3 letters of last name; fuzzy match on full name
      - Results in multiple matches
      - Sort matches by preferring same district, sentencing date, judge



## Data linking example #2

- Conviction and sentencing <-> Prison admission link example
  - Merge conviction and sentencing to prison admissions (new court commitments)
  - Linking hierarchy
    - Always require both agency sentence dates within ½ year, sentence date within 2 years of prison admission
      - Exact match on any two identifiers: FBI number; A-number; SSN; Register number; Name/DOB; Docket year and number
      - Exact match on one, fuzzy match on another
      - Fuzzy match on any two



# Data linking rates for example #1 and #2

- Inter-agency linking rates for 2014

Stage	Direction of Link	Match in 2014 (%)
Case termination (a) <->	<i>a -&gt; b</i>	91%
Conviction and sentencing (b)	<i>b -&gt; a</i>	94%
Conviction and sentencing (b)	<i>b -&gt; c</i>	64%
<-> Prison admission (c)	<i>c -&gt; b</i>	87%

# Assessing the quality of links: false positives

- Confirmatory links mitigate the number of false positive links
- Use logical sequences to assure reasonable links (e.g., a sentence occurs after a case filing)
- Synthesize data to check for false positives generated by linking procedure
  - Take a set of known good links
  - Modify the data to make them look “worse” (e.g., randomly generate messy or missing identifiers)
  - Apply linking algorithms
  - Compare linking rate on degraded data to the real rate
  - Make modifications to linking algorithm

# Assessing the quality of links: false negatives

- Quantify the true false negative rate for analysts
  - Quantify a baseline rate from known good links
  - Compare to a candidate, poorer linking rate
- Using simulation, similar to the procedure for false positives
  - Take a set known good links
  - Modify the data to make them look “worse” (e.g., randomly generate messy or missing identifiers)
  - Apply linking algorithms
  - Compare linking rate on degraded data to the real rate
  - Make modifications to linking algorithm

# Future ideas

- Complete assessment of linking rates and implement lessons learned into matching procedures
- Redesign documentation
  - Demonstrate linking rates
  - Create curriculum for users using sample research questions
  - Solicit and incorporate feedback from external users
- For each intra-agency link file *where identifiers support it*, add a within-agency unique serial number for each person to facilitate recidivism studies
- Assess and create additional link files as necessary (e.g., AO Appeals)

# Contact Information

Mark Motivans, Ph.D.  
Federal Justice Statistics Program  
Bureau of Justice Statistics  
202-514-4272  
[Mark.Motivans@USDOJ.GOV](mailto:Mark.Motivans@USDOJ.GOV)

Ryan Kling  
Senior Associate  
Abt Associates  
617-349-2460  
[Ryan\\_Kling@abtassoc.com](mailto:Ryan_Kling@abtassoc.com)

Christopher Cutler  
Senior Analyst  
Abt Associates  
617-520-2441  
[Christopher\\_Cutler@abtassoc.com](mailto:Christopher_Cutler@abtassoc.com)