Managing, Securing and Maintaining Case-Level Data

Presented by the CUNY Institute for State and Local Governance (ISLG)

September 30, 2022

SAFETY+JUSTICE CHALLENGE

Supported by the John D. and Catherine T. MacArthur Foundation

Learning Objectives

By the end of the workshop, participants will be able to:

- Gain knowledge of various technical tools and processes to work with criminal justice data;
- Understand the various use-case applications in working with criminal justice data across system points; and
- Learn about general practices and approaches in managing criminal justice data.

Agenda

- Introduction
- Overview of SJC Data Repository
  - Data Collection Process
  - De-Identification of Case-Level Data
  - Data Inventory Management and Deliverables
- Introduction of External Criminal Justice Data Sources
- Open Discussion around Technical Limitations and Challenges

Introduction

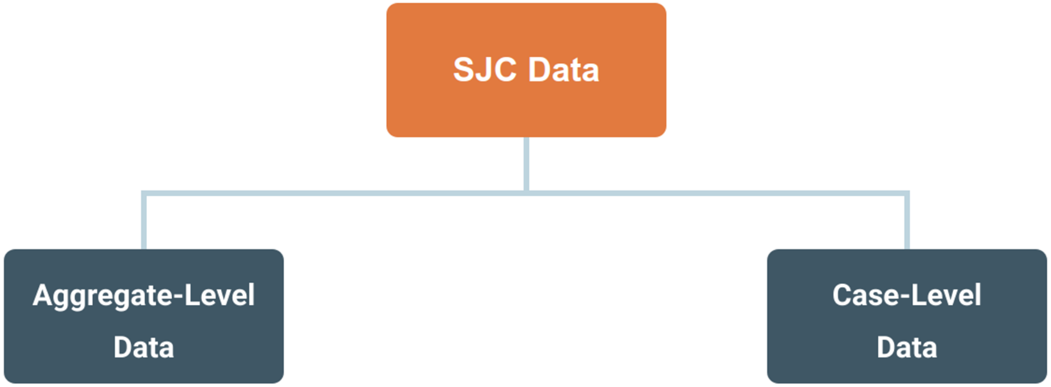
CUNY’s Institute for State and Local Governance (ISLG)

- ISLG serves as the national intermediary and primary data and analytic partner for the Safety & Justice Challenge (SJC)
- Specifically, ISLG is tasked with:
  - Collecting comprehensive, system-wide criminal justice data from sites
  - Creating and tracking performance metrics, and conducting in-depth analysis of jail population and other criminal justice trends
  - Providing analytic and data-capacity building assistance
  - Providing SJC initiative partners and approved external researchers with de-identified site data for research and technical assistance purposes
- ISLG’s Data Operation Team (DOT) is responsible for data management activities, including maintaining SJC data and fulfilling external data requests

ISLG receives two distinct types of data from SJC sites

SJC Data

ISLG receives two distinct types of data from SJC sites



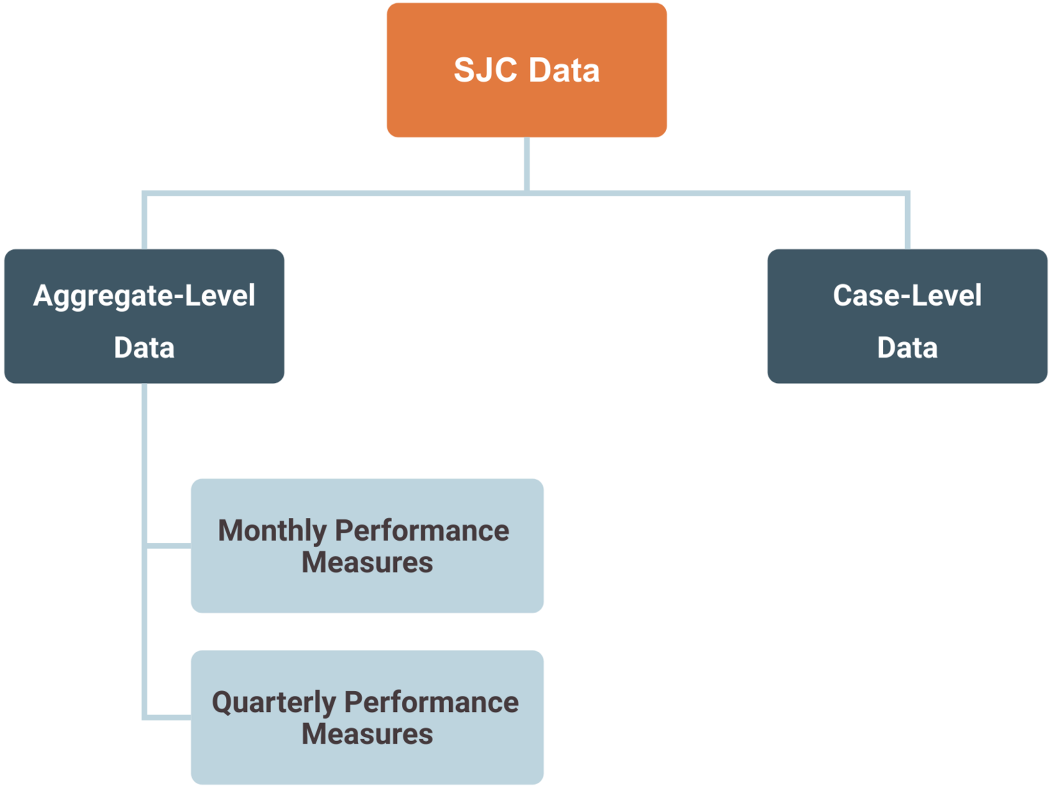
SJC Data

Aggregate-Level Data

Case-Level Data

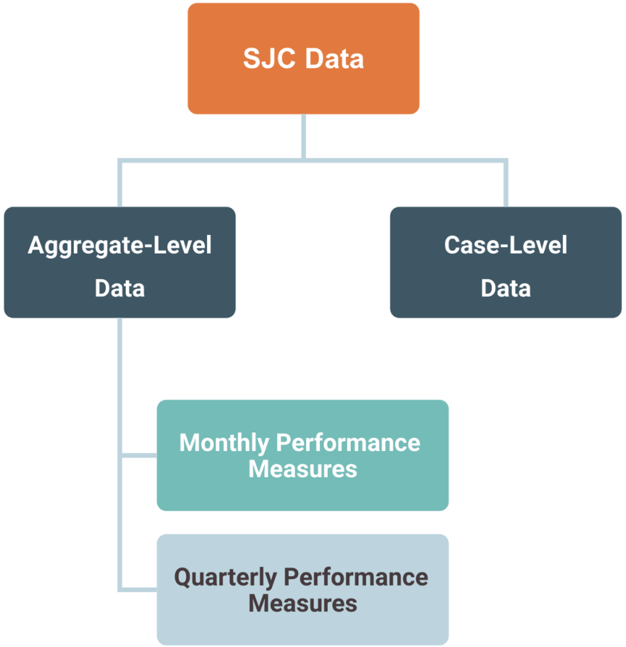
- Aggregate-Level data are collected monthly, in contrast with Case-Level data, which allows us to see much more detail, but is only collected annually
- Both types of data are useful, but there are trade-offs between detail and recency

ISLG receives two distinct types of data from SJC sites



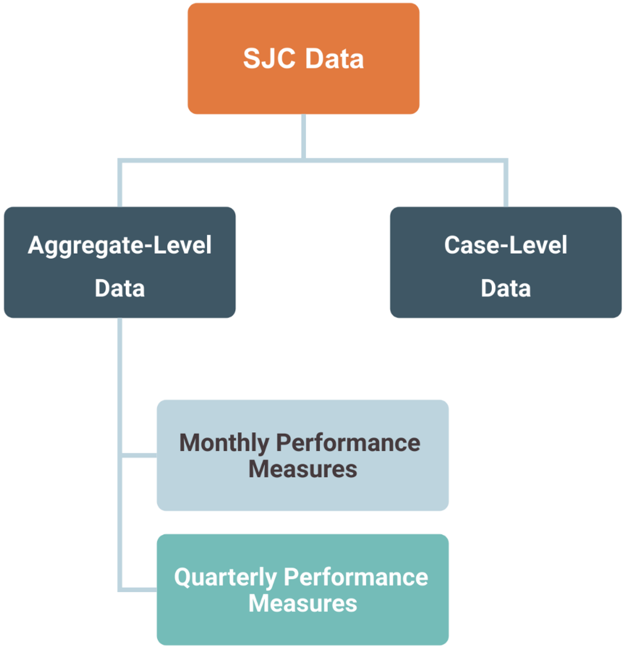
Aggregate-Level Data

- **Aggregate Monthly Performance Measure Data** is submitted by all 26 SJC implementation sites on a monthly basis
- It includes measures such as average daily population (ADP), bookings and releases, and average length of stay (ALOS)



Aggregate-Level Data

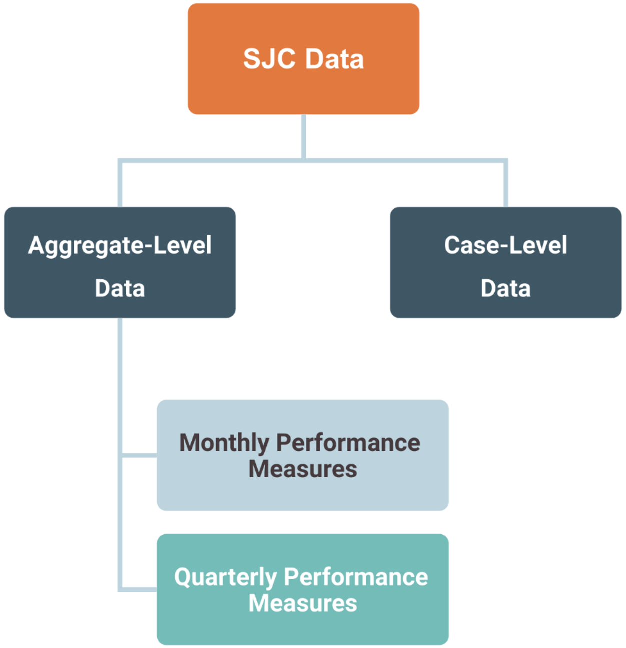
- **Aggregate Monthly Performance Measure Data** is submitted by all 26 SJC implementation sites on a monthly basis
- It includes measures such as average daily population (ADP), bookings and releases, and average length of stay (ALOS)
- **Quarterly Monthly Performance Measure Data** is provided to ISLG by five (5) SJC implementation sites on a quarterly basis
- It is provided in a standardized template that includes jail performance metrics calculated by the site



Supported by the John D. and Catherine T. MacArthur Foundation

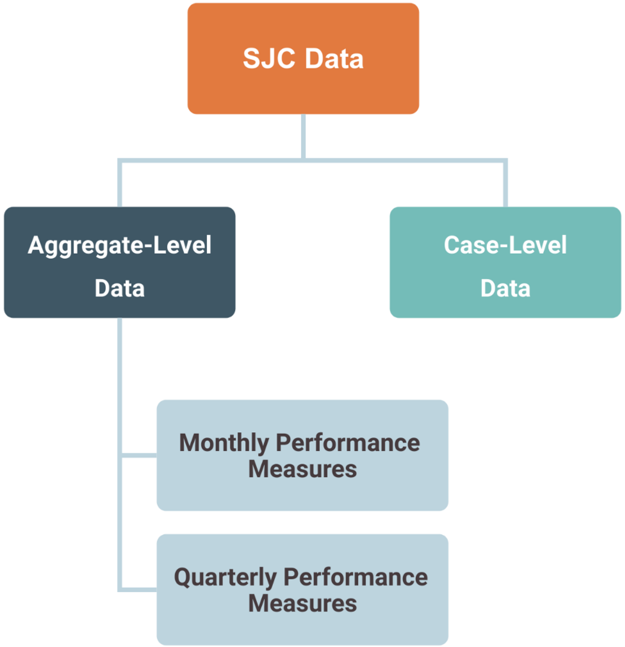
Aggregate-Level Data

- **Aggregate Monthly Performance Measure Data** is submitted by all 26 SJC implementation sites on a monthly basis
- It includes measures such as average daily population (ADP), bookings and releases, and average length of stay (ALOS)
- **Quarterly Monthly Performance Measure Data** is provided to ISLG by five (5) SJC implementation sites on a quarterly basis
- It is provided in a standardized template that includes jail performance metrics
- <u>This data does NOT include any identifiable information</u>



Case-Level Data

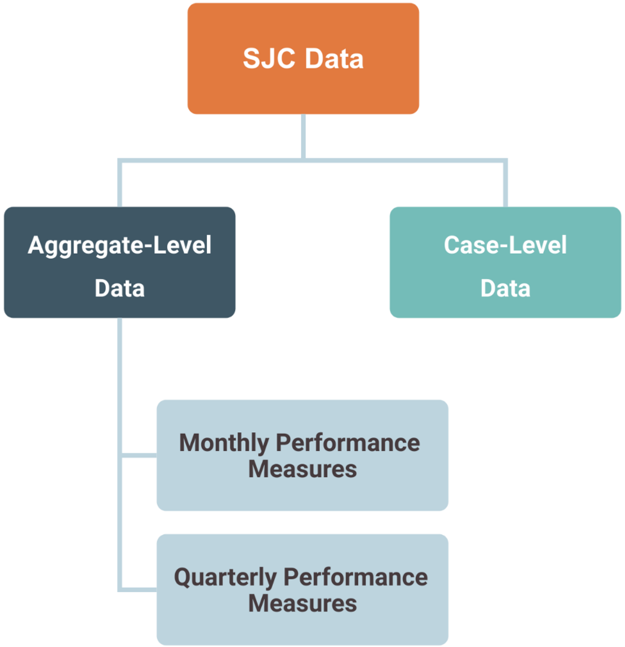
- **Case-level data** is provided by 18 SJC sites on an annual basis at the end of each SJC year for key decision points in the criminal justice process
- This data generally contains:
  - Unique person and case identifiers
  - Available demographic information
  - Important dates of key events and decisions
  - Descriptions of key events and decisions (e.g., disposition or sentence type)
  - Charge information
- As such, case-level data includes confidential data and Personal-Identifiable Information (PII), which are very sensitive in nature



Considerations for working with Case-Level data

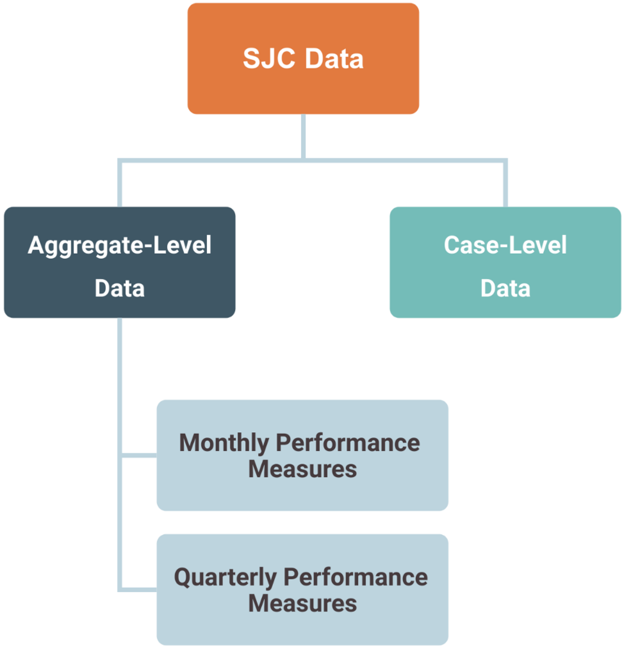
Case-Level Data

- PII and confidential data is protected against unauthorized disclosure or modification, both when the data is in use and when the data is stored or transmitted electronically
- When sharing case-level data with external stakeholders and research partners, ISLG ensures that all PII is removed and other identifiers are de-identified



Case-Level Data

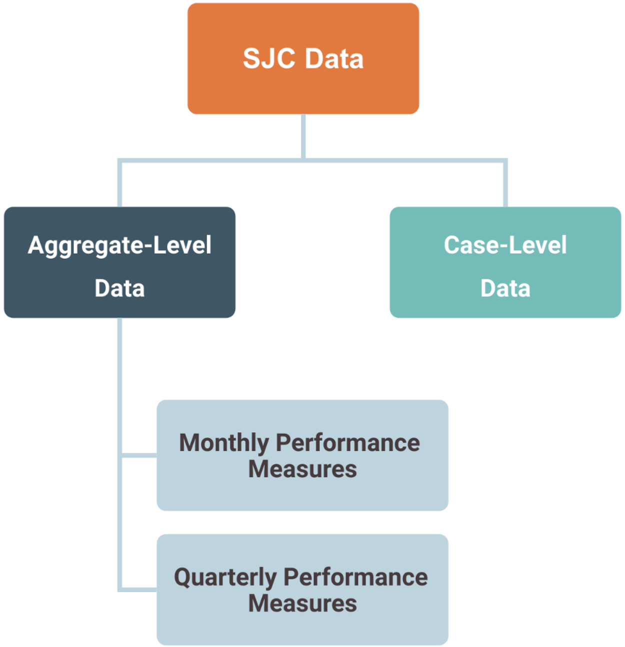
- De-identifying case-level data has several components, including:
  - Remove any PII elements (including any relevant person-level administrative (agency) identifiers)
  - Scramble Person- and Event-level identifiers
  - Extract the year of birth from all Date-of-Birth variables
  - Truncate 9-digit Zip codes to only the first 5 digits



Supported by the John D. and Catherine T. MacArthur Foundation

Case-Level Data

- Identifiable data elements include person-level identifiers and data points, such as:
    - Full Name (First, Last)
    - Full Home Address (Unit Number, Street, City, State)
    - Social Security Number
    - City or State Identification Number
    - Full Date of Birth
    - Zip Code (of home residence)
    - **Any information which can be used to distinguish or trace an individual's identity directly through linkages with other information**



The need for a centralized database

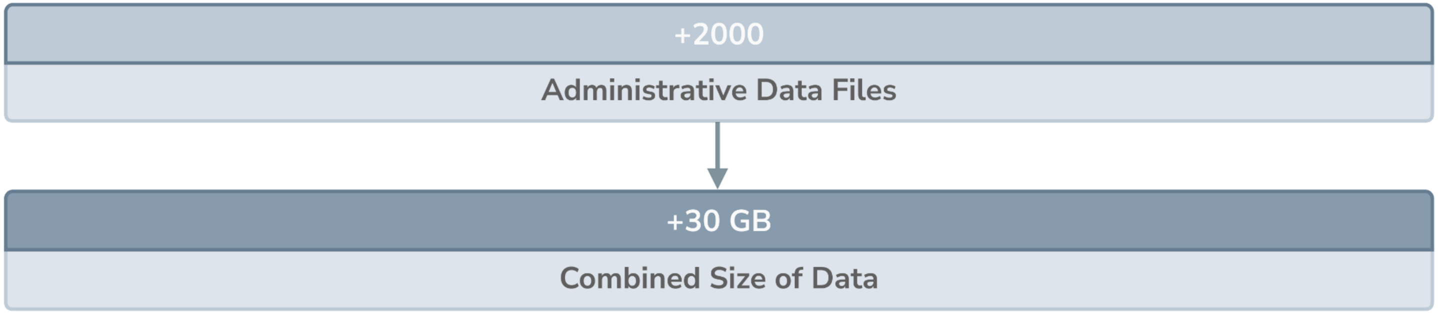
Supported by the John D. and Catherine T. MacArthur Foundation

The need for a centralized database

+2000

Administrative Data Files

The need for a centralized database



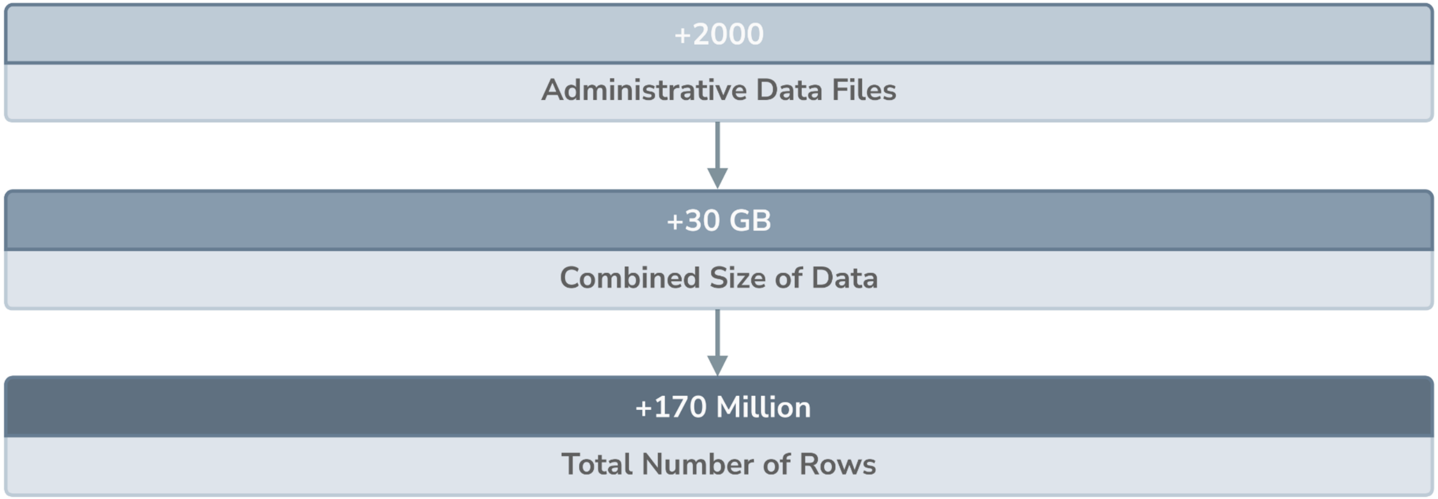
+2000

Administrative Data Files

+30 GB

Combined Size of Data

The need for a centralized database



+2000

Administrative Data Files

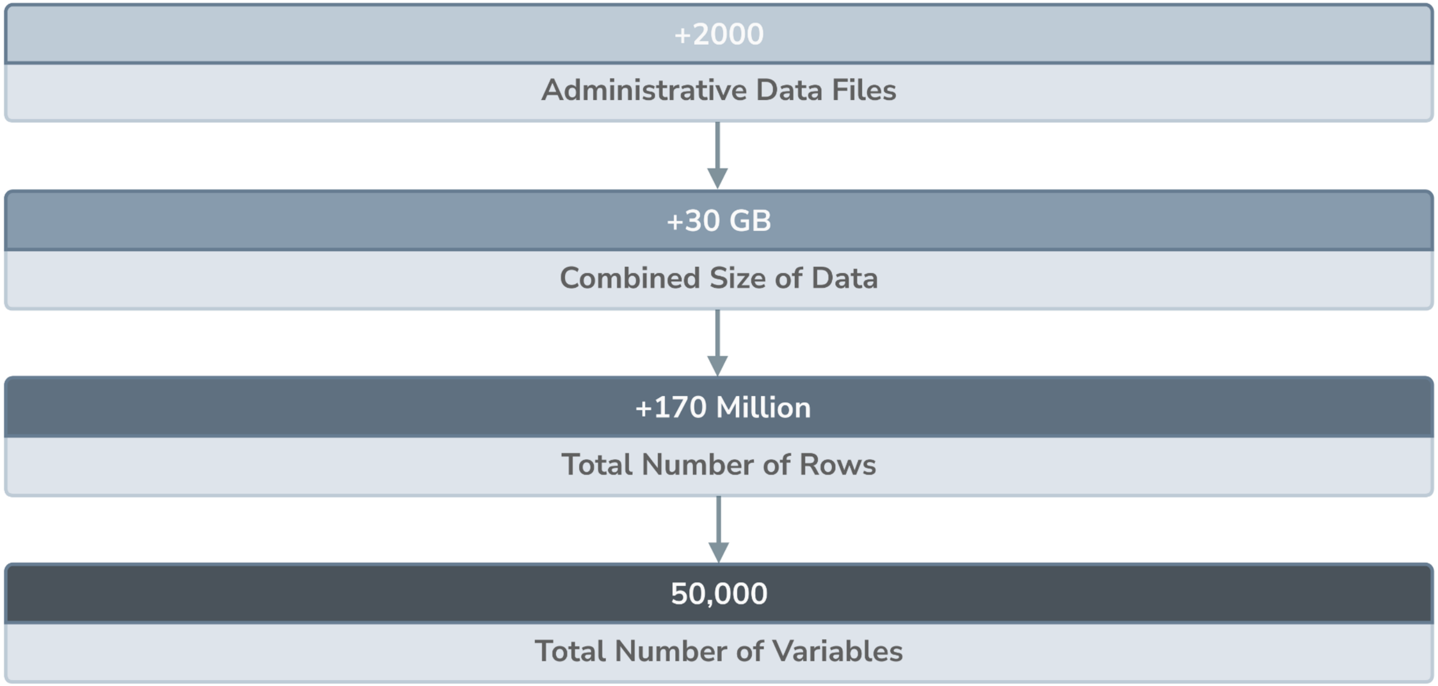
+30 GB

Combined Size of Data

+170 Million

Total Number of Rows

The need for a centralized database



+2000

Administrative Data Files

+30 GB

Combined Size of Data

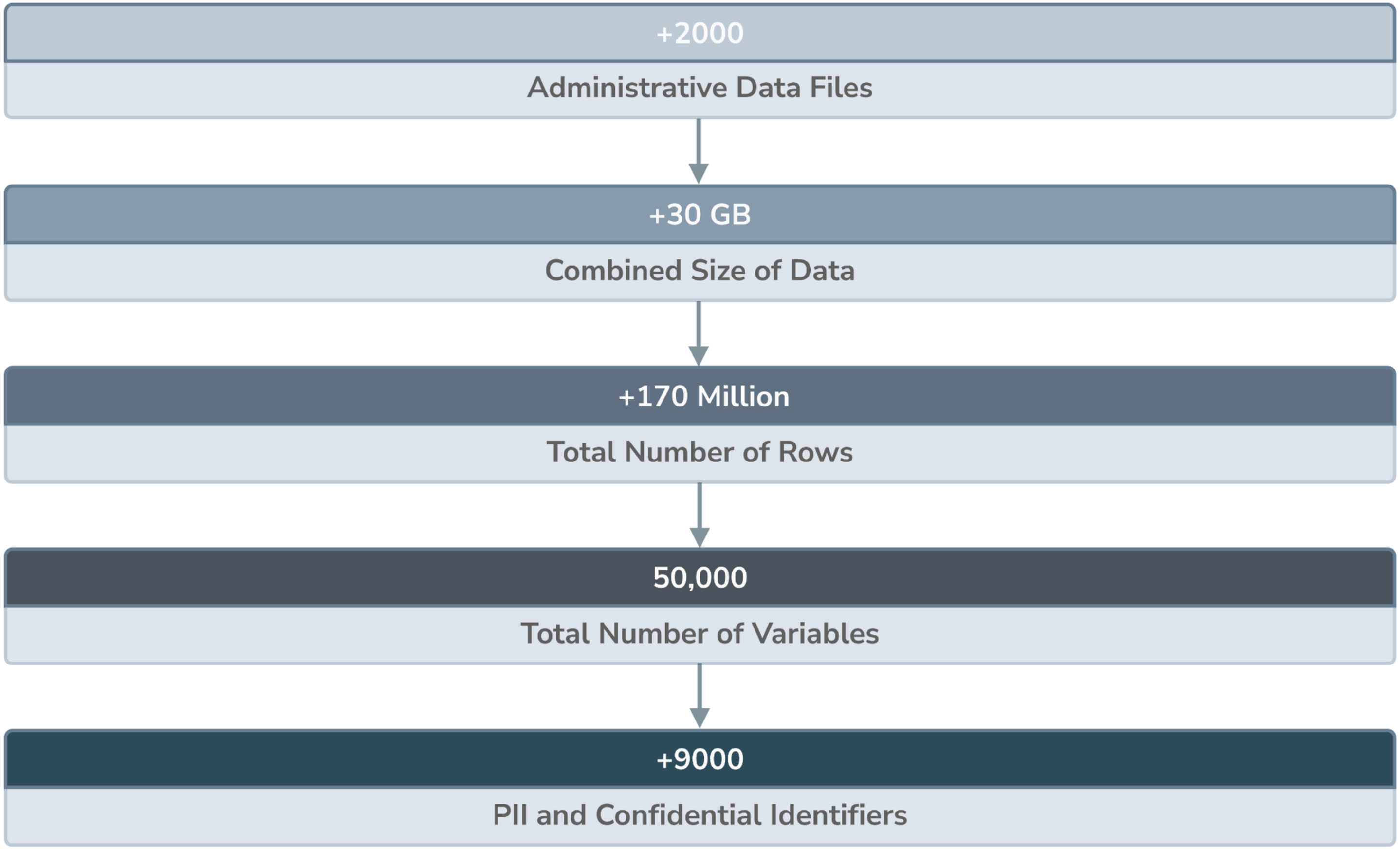
+170 Million

Total Number of Rows

50,000

Total Number of Variables

The need for a centralized database



+2000

Administrative Data Files

+30 GB

Combined Size of Data

+170 Million

Total Number of Rows

50,000

Total Number of Variables

+9000

PII and Confidential Identifiers

SJC Data Repository

Data Repository

- ISLG has developed an internal database of de-identified data, called the SJC Data Repository
- ISLG built the Repository to capture full populations of cases across years, from each key decision point in the adult criminal legal process
- This multi-year, system-wide, case-level data collection effort is unique both in its scope and in its focus on supporting many different data uses
- The Repository allows ISLG to manage its entire SJC data holdings, including:
    - De-identified case-level data that has been
    - Additional working files prepared by research staff
    - Documentation

Data Repository

- The Repository serves as the **primary source of data** needed for performance measurement and analysis work, and other research
- All the data stored in the Repository is lightly cleaned, formatted and de-identified by ISLG
- The Repository is maintained by a suite of internally developed software, that is managed by a team of ISLG data scientists and researchers
- Additionally, the Repository also contains several ancillary products as well, including:
    - Data Diagnostic (Quality Assurance) Reports
    - Supplementary Data Codebooks
    - Data Archives
    - Meta Data Files
    - Log Records

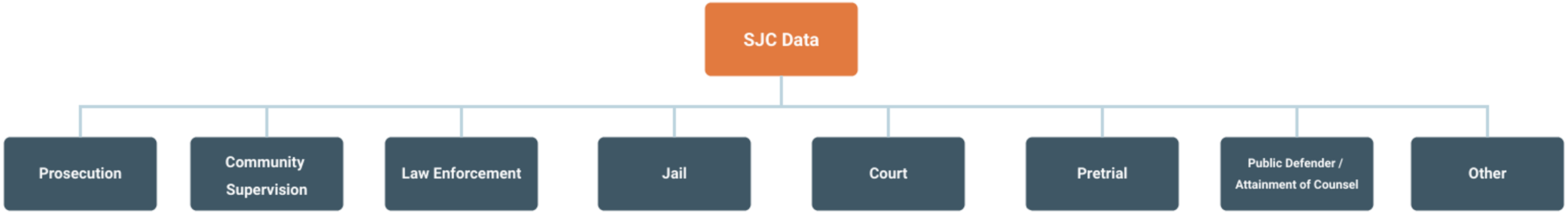
Data Repository

Data Classification for the Repository

- ISLG has developed a classification system for categorizing Criminal Justice data, allowing it to efficiently manage and organize SJC data
- This classification system has three levels of categorization:
    - System Point (Agency)
    - Sub-System Point (File Type within Agency)
    - Inclusion Period

Data Repository

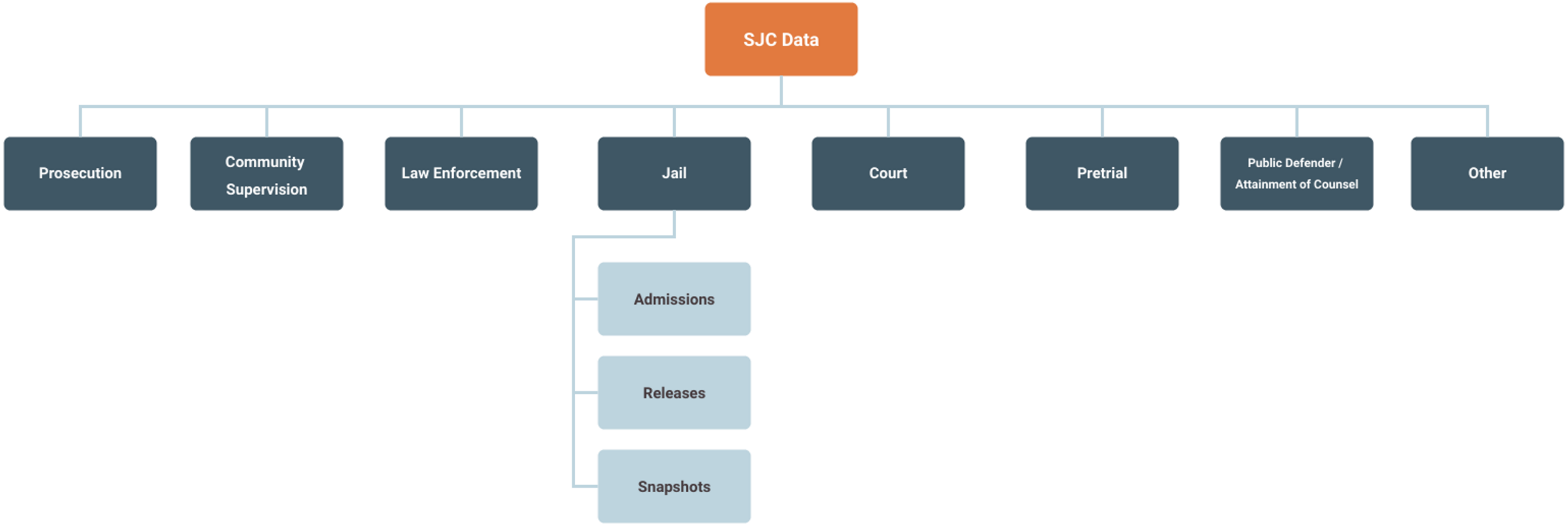
- System Points correspond to key decision points in the criminal justice process, and are classified according to eight standardized categories (specific system point/agency names vary by SJC site so ISLG created standardized categories)
- Sub-System Points are associated with each System Point, but are not standardized



Supported by the John D. and Catherine T. MacArthur Foundation

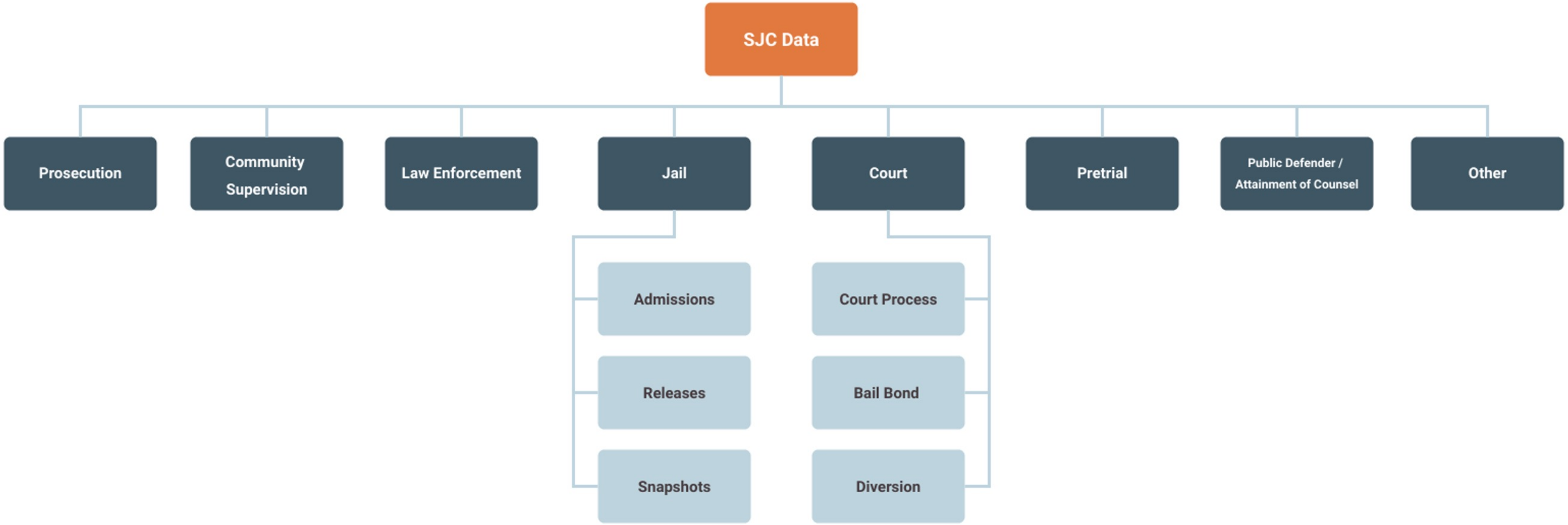
Data Repository

- System Points correspond to key decision points in the criminal justice process, and are classified according to eight standardized categories (specific system point/agency names vary by SJC site so ISLG created standardized categories)
- Sub-System Points are associated with each System Point, but are not standardized



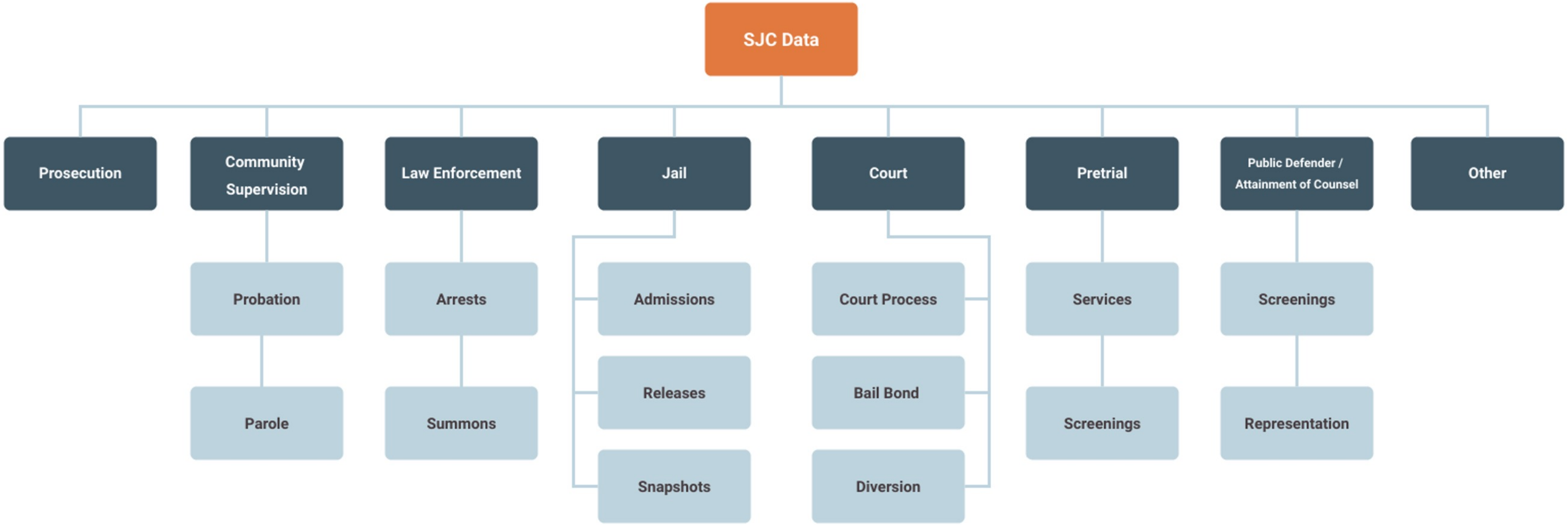
Data Repository

- System Points correspond to key decision points in the criminal justice process, and are classified according to eight standardized categories (specific system point/agency names vary by SJC site so ISLG created standardized categories)
- Sub-System Points are associated with each System Point, but are not standardized



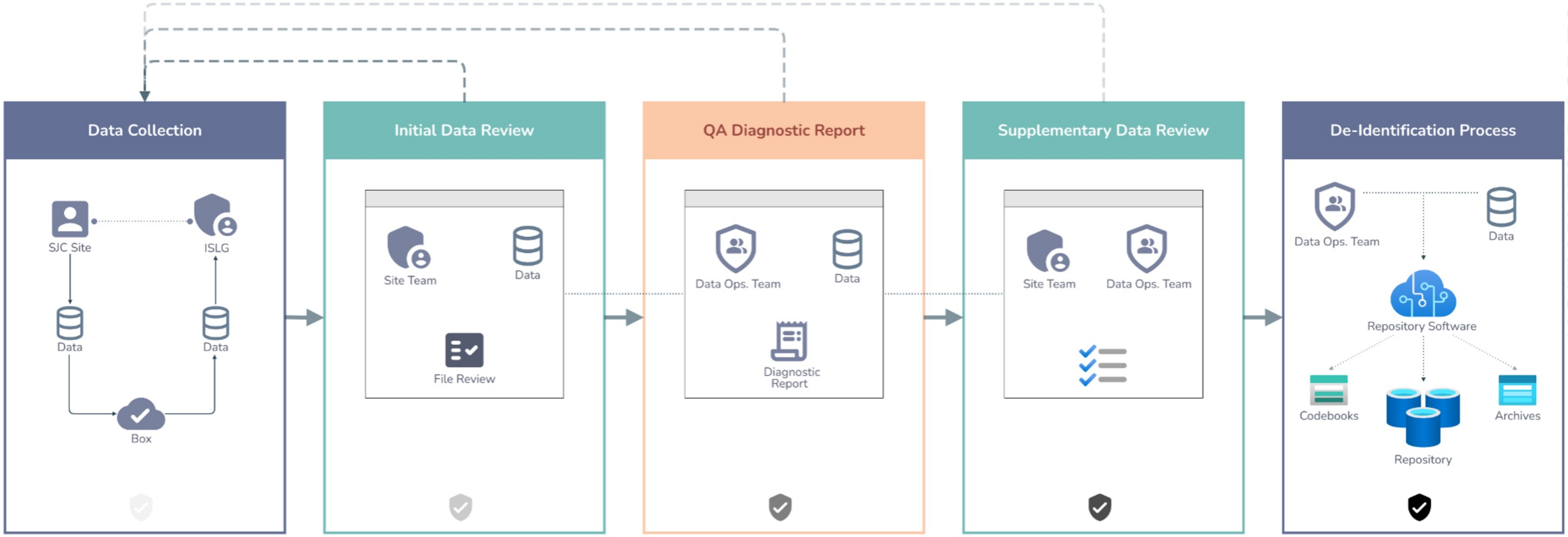
Data Repository

- System Points correspond to key decision points in the criminal justice process, and are classified according to eight standardized categories (specific system point/agency names vary by SJC site so ISLG created standardized categories)
- Sub-System Points are associated with each System Point, but are not standardized



Managing the Repository

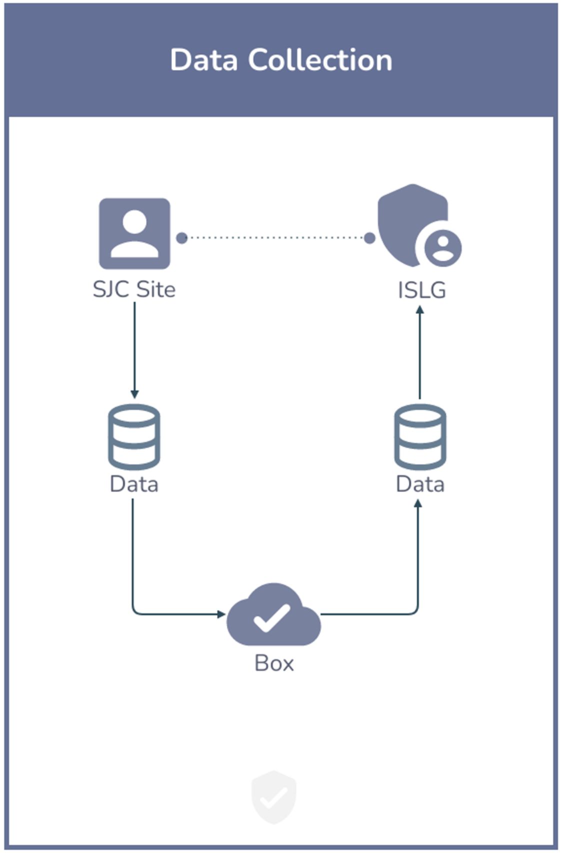
- From data collection to data de-identification, managing the Repository is a multi-stage, collaborative effort, involving multiple partners, including data liaisons from SJC sites, ISLG staff researchers and data scientists



Supported by the John D. and Catherine T. MacArthur Foundation

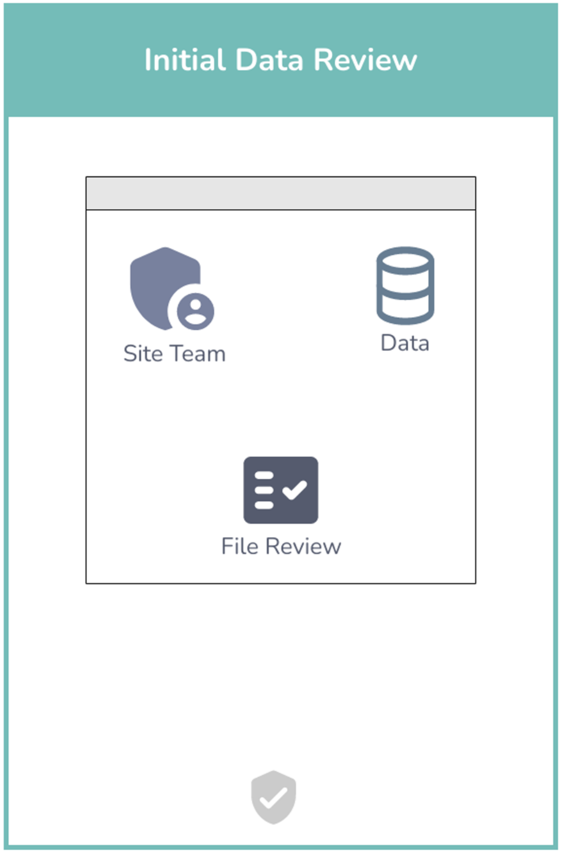
Data Collection

- The first step of the process is securely collecting data from agencies within SJC sites
- Agency staff submit data to ISLG on to Box, a secure file transfer platform
- ISLG downloads the data to ISLG's internal encrypted shared drive, before removing it from the cloud platform



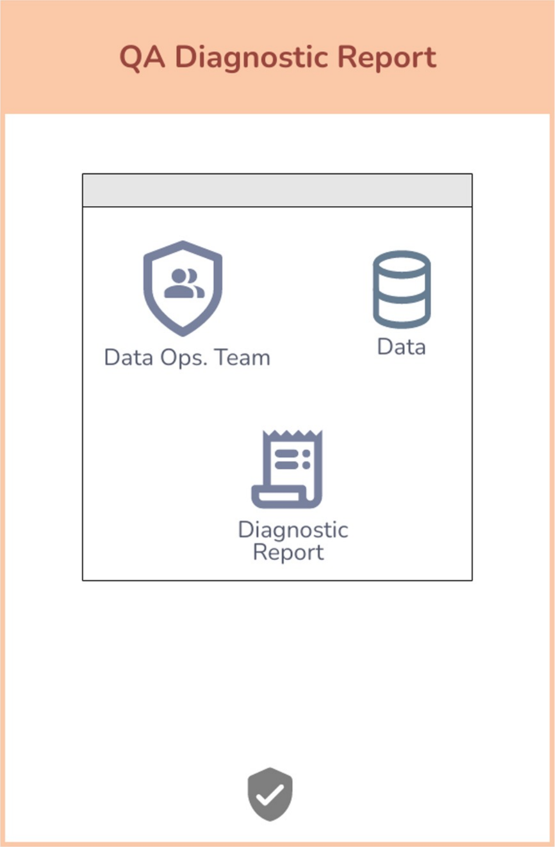
Initial Data Review

- SJC data goes through multiple stages of Quality Assurance (QA) and reviews before it can be processed into the Repository
- The first stage in the QA process is an Initial Data Review performed by ISLG site teams
- ISLG conducts a high level check of the data, and based on their knowledge and expertise of the data, attempt to identify potential issues or points of concern
- ISLG works with SJC sites when issues are identified

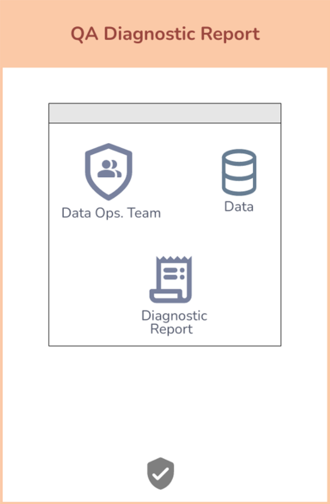


QA Diagnostic Report

- The DOT generates a QA Diagnostic Report developed to identify potential data quality issues and errors
- The Diagnostic Report provides summarized variable-level statistics to review the quality of the data
- The report seeks to detect potential anomalies in the data, serving as an important QA layer before adding any data to the Repository



QA Diagnostic Report



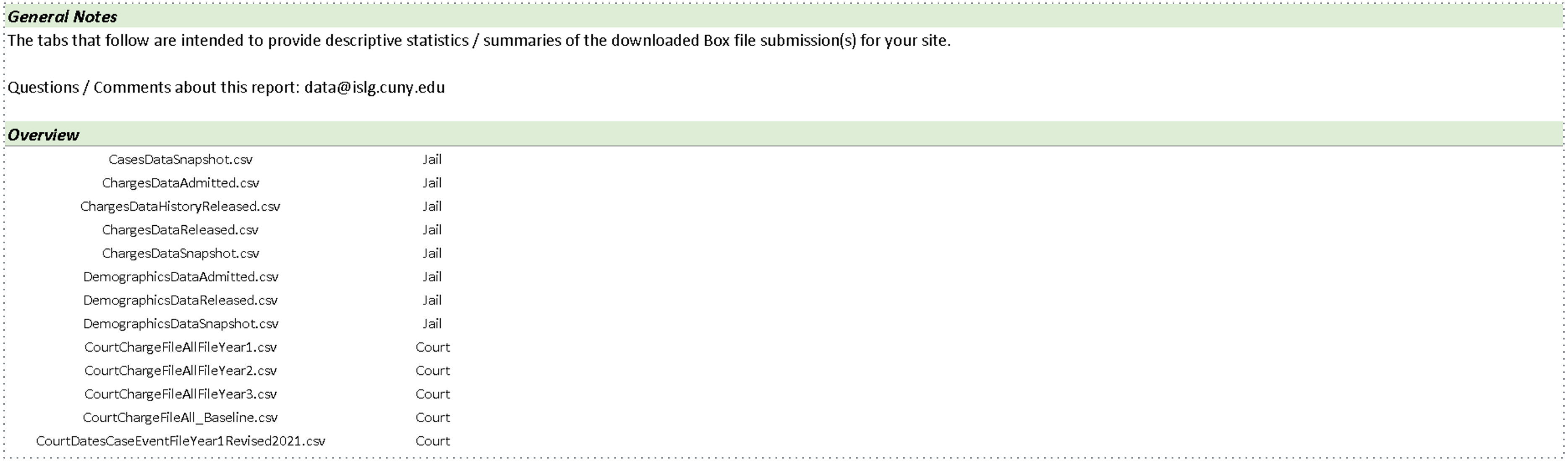
QA Diagnostic Report

Data Ops. Team

Data

Diagnostic Report

- The diagnostic report includes multiple components, each analysing the the data through a different lens, or form a different aspect of data
- These include: **Summary**



General Notes

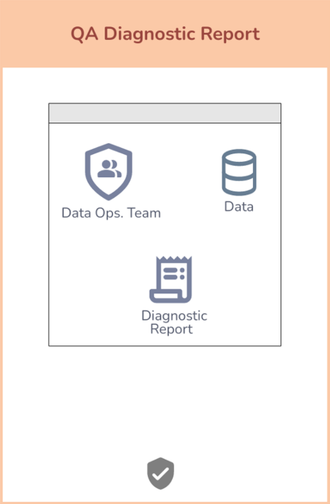
The tabs that follow are intended to provide descriptive statistics / summaries of the downloaded Box file submission(s) for your site.

Questions / Comments about this report: data@islg.cuny.edu

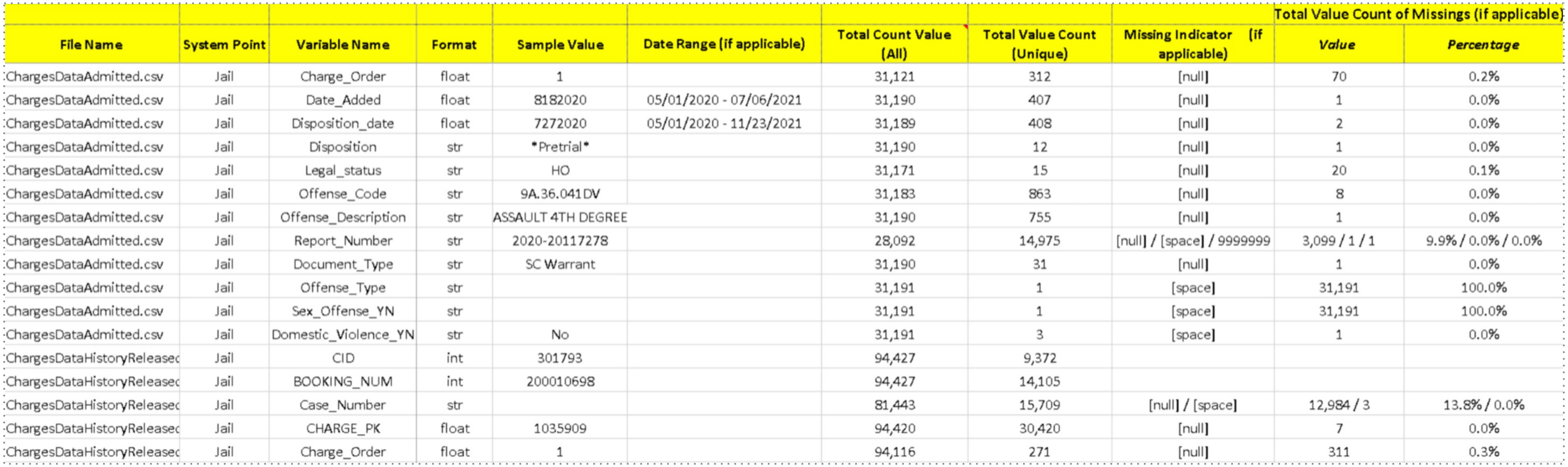
Overview

CasesDataSnapshot.csv	Jail
ChargesDataAdmitted.csv	Jail
ChargesDataHistoryReleased.csv	Jail
ChargesDataReleased.csv	Jail
ChargesDataSnapshot.csv	Jail
DemographicsDataAdmitted.csv	Jail
DemographicsDataReleased.csv	Jail
DemographicsDataSnapshot.csv	Jail
CourtChargeFileAllFileYear1.csv	Court
CourtChargeFileAllFileYear2.csv	Court
CourtChargeFileAllFileYear3.csv	Court
CourtChargeFileAll_Baseline.csv	Court
CourtDatesCaseEventFileYear1Revised2021.csv	Court

QA Diagnostic Report

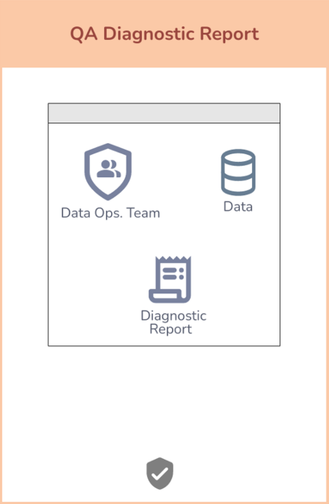


- The diagnostic report includes multiple components, each analysing the the data through a different lens, or form a different aspect of data
- These include: **System Point Analysis**

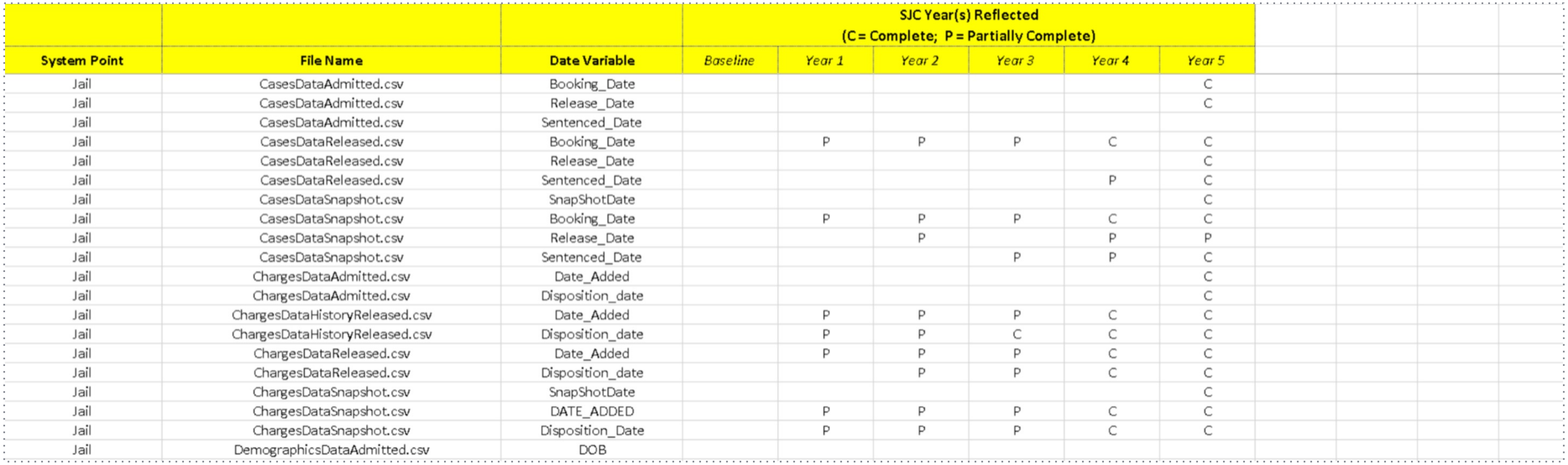


									Total Value Count of Missings (if applicable)	
File Name	**System Point**	**Variable Name**	**Format**	**Sample Value**	**Date Range (if applicable)**	**Total Count Value (All)**	**Total Value Count (Unique)**	**Missing Indicator (if applicable)**	***Value***	***Percentage***
ChargesDataAdmitted.csv	Jail	Charge_Order	float	1		31,121	312	[null]	70	0.2%
ChargesDataAdmitted.csv	Jail	Date_Added	float	8182020	05/01/2020 - 07/06/2021	31,190	407	[null]	1	0.0%
ChargesDataAdmitted.csv	Jail	Disposition_date	float	7272020	05/01/2020 - 11/23/2021	31,189	408	[null]	2	0.0%
ChargesDataAdmitted.csv	Jail	Disposition	str	*Pretrial*		31,190	12	[null]	1	0.0%
ChargesDataAdmitted.csv	Jail	Legal_status	str	HO		31,171	15	[null]	20	0.1%
ChargesDataAdmitted.csv	Jail	Offense_Code	str	9A.36.041DV		31,183	863	[null]	8	0.0%
ChargesDataAdmitted.csv	Jail	Offense_Description	str	ASSAULT 4TH DEGREE		31,190	755	[null]	1	0.0%
ChargesDataAdmitted.csv	Jail	Report_Number	str	2020-20117278		28,092	14,975	[null] / [space] / 9999999	3,099 / 1 / 1	9.9% / 0.0% / 0.0%
ChargesDataAdmitted.csv	Jail	Document_Type	str	SC Warrant		31,190	31	[null]	1	0.0%
ChargesDataAdmitted.csv	Jail	Offense_Type	str			31,191	1	[space]	31,191	100.0%
ChargesDataAdmitted.csv	Jail	Sex_Offense_YN	str			31,191	1	[space]	31,191	100.0%
ChargesDataAdmitted.csv	Jail	Domestic_Violence_YN	str	No		31,191	3	[space]	1	0.0%
ChargesDataHistoryReleasec	Jail	CID	int	301793		94,427	9,372			
ChargesDataHistoryReleasec	Jail	BOOKING_NUM	int	200010698		94,427	14,105			
ChargesDataHistoryReleasec	Jail	Case_Number	str			81,443	15,709	[null] / [space]	12,984 / 3	13.8% / 0.0%
ChargesDataHistoryReleasec	Jail	CHARGE_PK	float	1035909		94,420	30,420	[null]	7	0.0%
ChargesDataHistoryReleasec	Jail	Charge_Order	float	1		94,116	271	[null]	311	0.3%

QA Diagnostic Report


QA Diagnostic Report

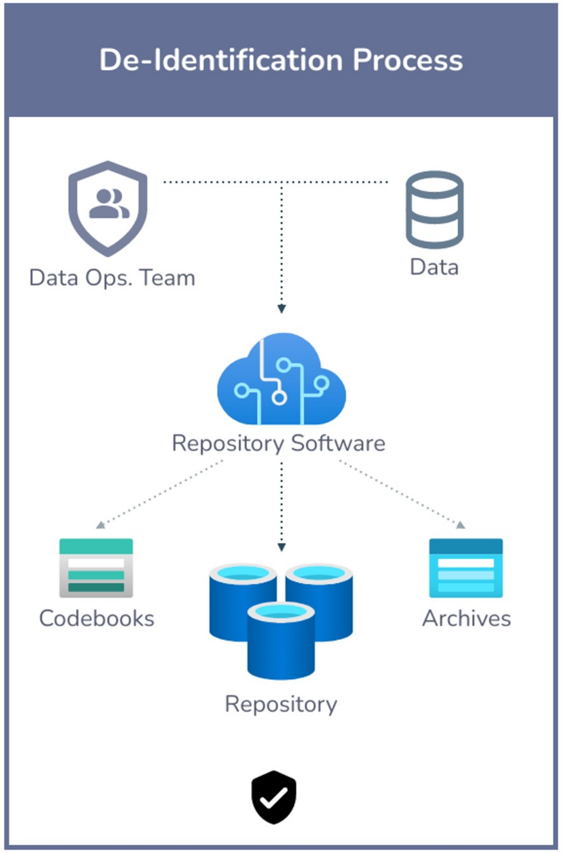
- The diagnostic report includes multiple components, each analysing the the data through a different lens, or form a different aspect of data
- These include: **Date Variable Overview**



			SJC Year(s) Reflected (C= Complete; P = Partially Complete)					
System Point	**File Name**	**Date Variable**	*Baseline*	*Year 1*	*Year 2*	*Year 3*	*Year 4*	*Year 5*
Jail	CasesDataAdmitted.csv	Booking_Date						C
Jail	CasesDataAdmitted.csv	Release_Date						C
Jail	CasesDataAdmitted.csv	Sentenced_Date						
Jail	CasesDataReleased.csv	Booking_Date		P	P	P	C	C
Jail	CasesDataReleased.csv	Release_Date						C
Jail	CasesDataReleased.csv	Sentenced_Date					P	C
Jail	CasesDataSnapshot.csv	SnapShotDate						C
Jail	CasesDataSnapshot.csv	Booking_Date		P	P	P	C	C
Jail	CasesDataSnapshot.csv	Release_Date			P		P	P
Jail	CasesDataSnapshot.csv	Sentenced_Date				P	P	C
Jail	ChargesDataAdmitted.csv	Date_Added						C
Jail	ChargesDataAdmitted.csv	Disposition_date						C
Jail	ChargesDataHistoryReleased.csv	Date_Added		P	P	P	C	C
Jail	ChargesDataHistoryReleased.csv	Disposition_date		P	P	C	C	C
Jail	ChargesDataReleased.csv	Date_Added		P	P	P	C	C
Jail	ChargesDataReleased.csv	Disposition_date			P	P	C	C
Jail	ChargesDataSnapshot.csv	SnapShotDate						C
Jail	ChargesDataSnapshot.csv	DATE_ADDED		P	P	P	C	C
Jail	ChargesDataSnapshot.csv	Disposition_Date		P	P	P	C	C
Jail	DemographicsDataAdmitted.csv	DOB						

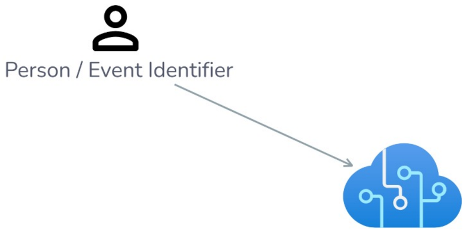
De-Identification Process

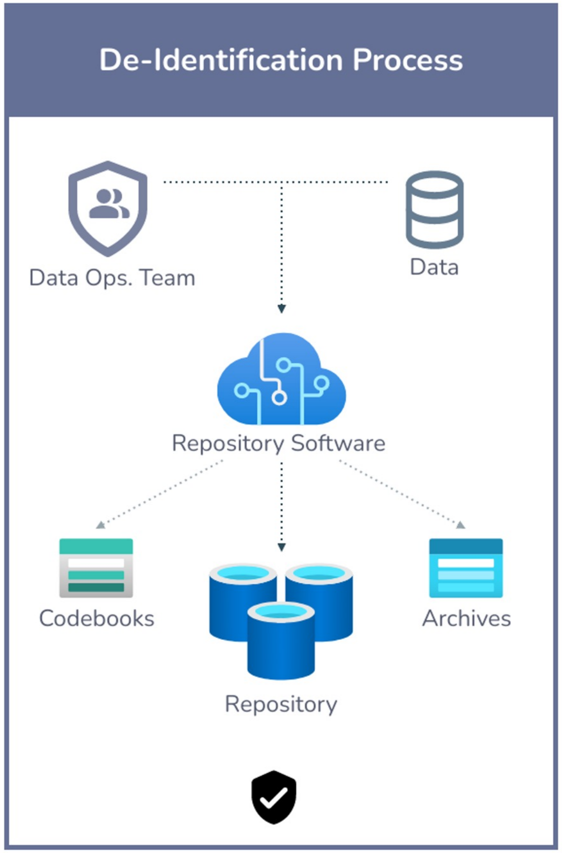
- The De-Identification Process for adding files to Repository is also performed programmatically by internally developed software, managed by ISLG's DOT
- The software processes each file into the Repository, and simultaneously generates supplementary data products
- The primary component of this process is de-identifying the data



De-Identification Process

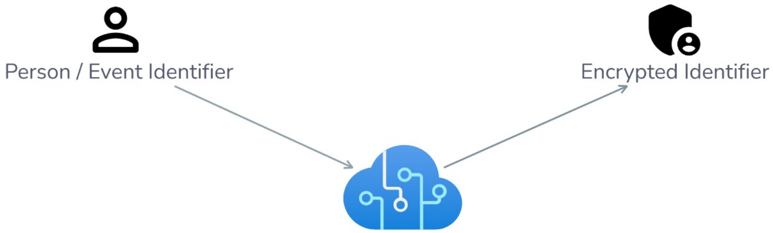
- Based on the PII and confidential flagged during the review process, data de-identification includes:
  - Scrambling Person- and Event-level Identifiers

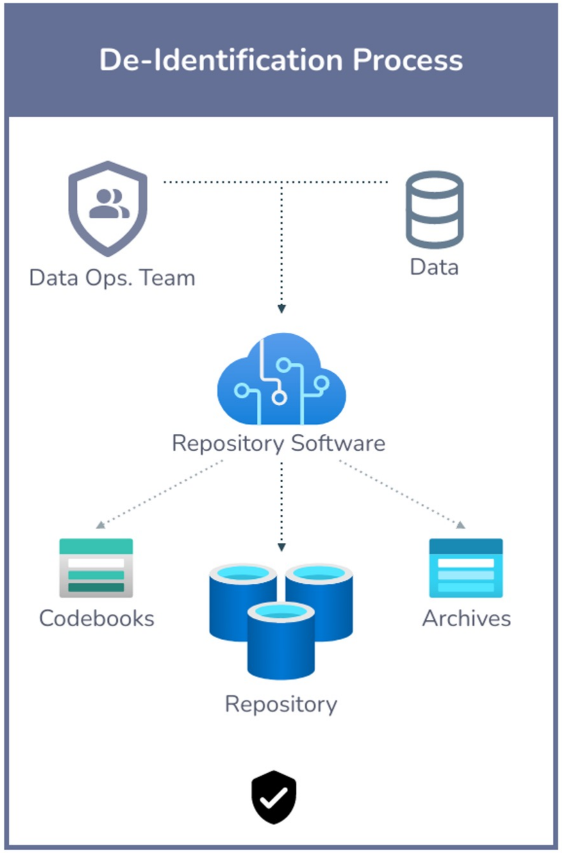




De-Identification Process

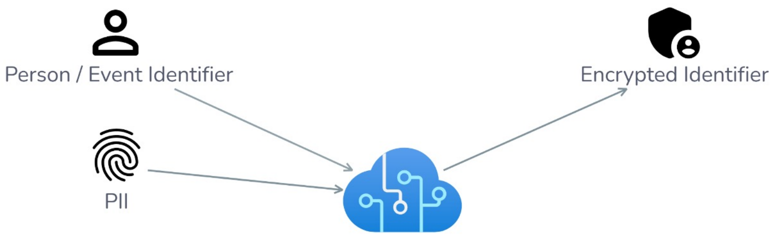
- Based on the PII and confidential flagged during the review process, data de-identification includes:
  - Scrambling Person- and Event-level Identifiers

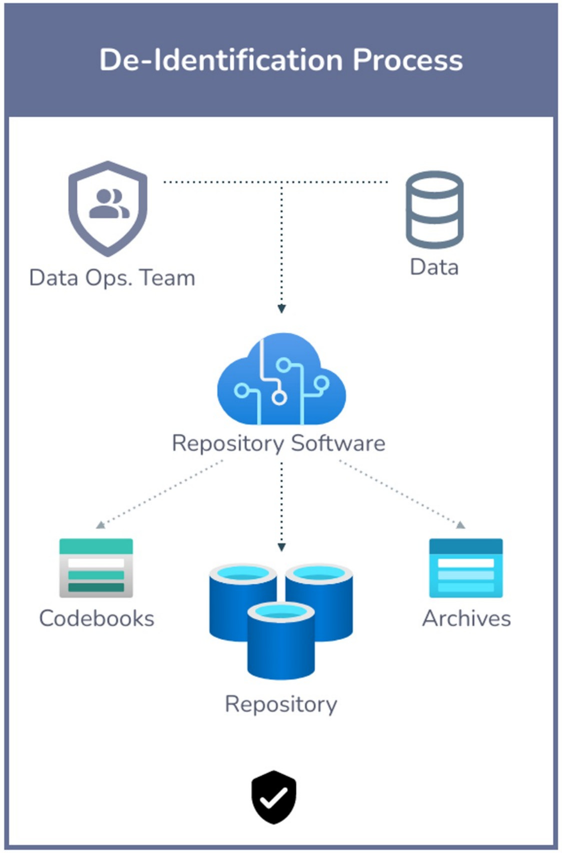




De-Identification Process

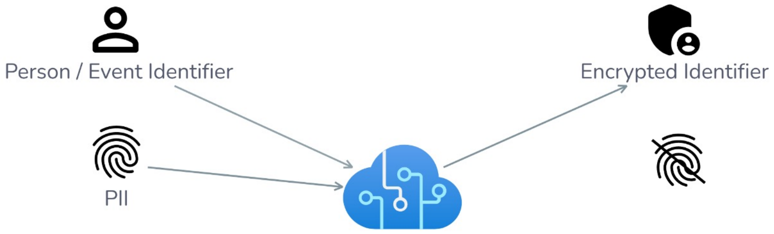
- Based on the PII and confidential flagged during the review process, data de-identification includes:
    - Scrambling Person- and Event-level Identifiers
    - Removing Person-level identifiable elements

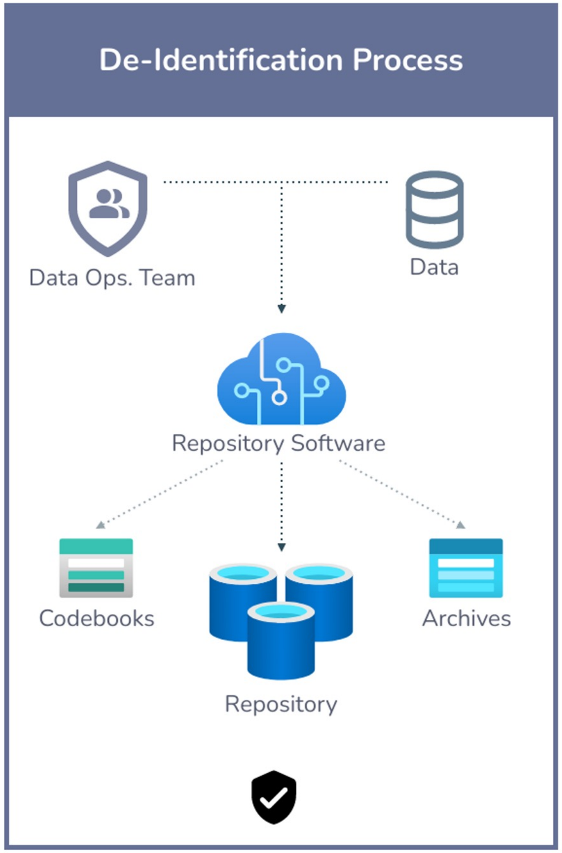




De-Identification Process

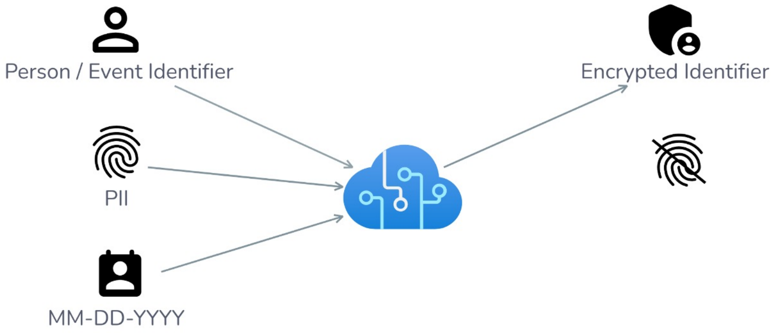
- Based on the PII and confidential flagged during the review process, data de-identification includes:
  - Scrambling Person- and Event-level Identifiers
  - Removing Person-level identifiable elements

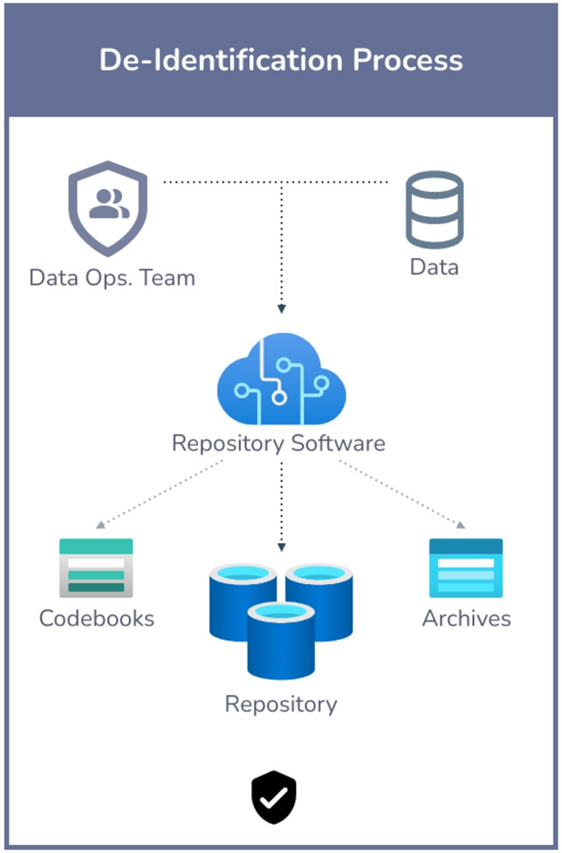




De-Identification Process

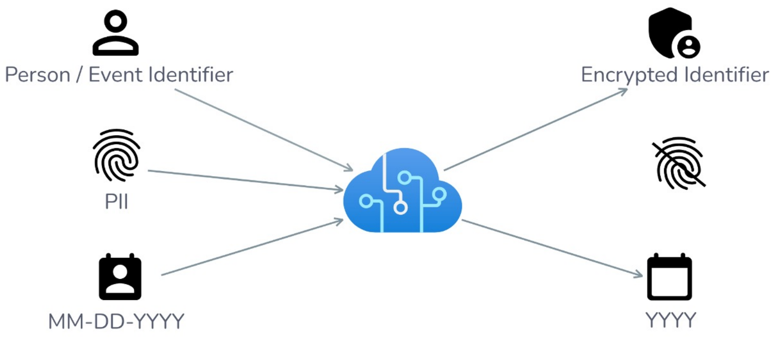
- Based on the PII and confidential flagged during the review process, data de-identification includes:
  - Scrambling Person- and Event-level Identifiers
  - Removing Person-level identifiable elements
  - Extracting Year-of-birth from Date-of-Births

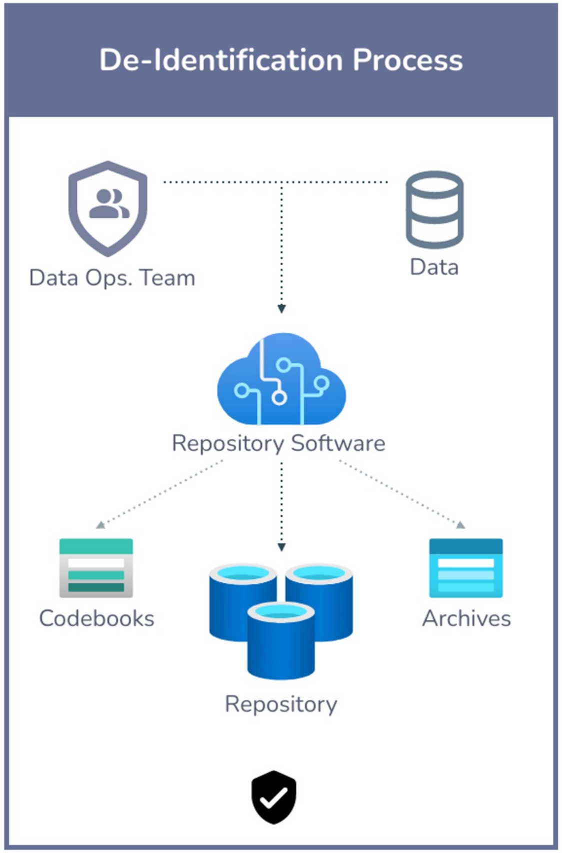




De-Identification Process

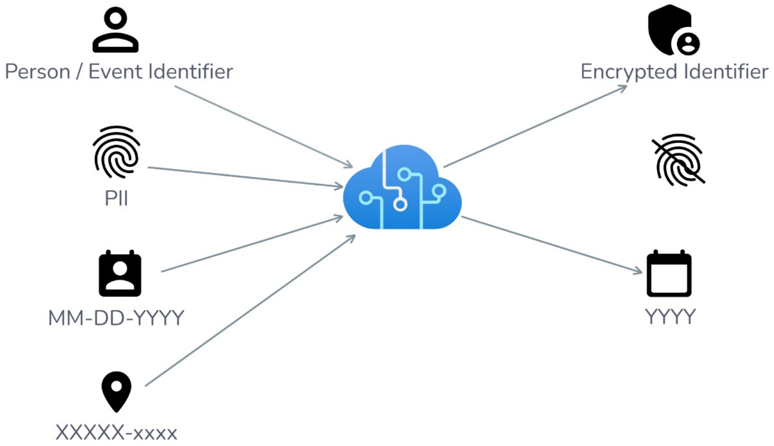
- Based on the PII and confidential flagged during the review process, data de-identification includes:
    - Scrambling Person- and Event-level Identifiers
    - Removing Person-level identifiable elements
    - Extracting Year-of-birth from Date-of-Births

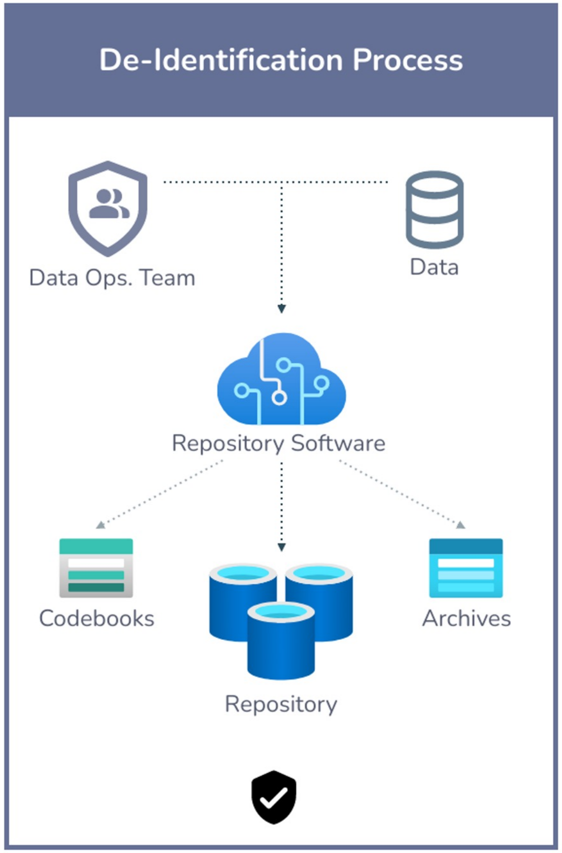




De-Identification Process

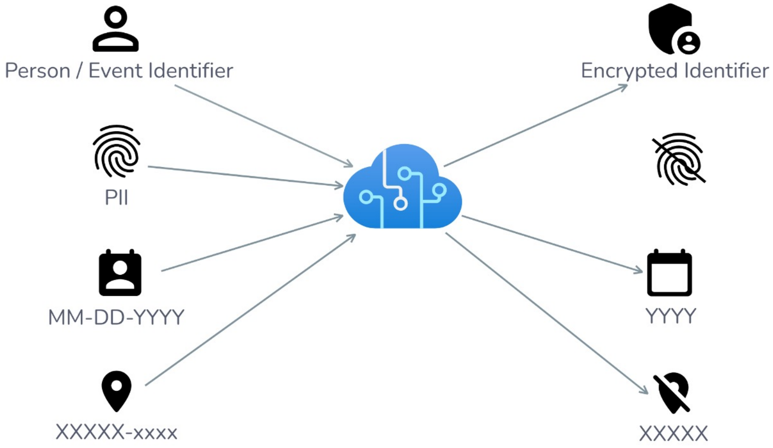
- Based on the PII and confidential flagged during the review process, data de-identification includes:
  - Scrambling Person- and Event-level Identifiers
  - Removing Person-level identifiable elements
  - Extracting Year-of-birth from Date-of-Births
  - Truncating Zip Code variables to first 5 Zip code

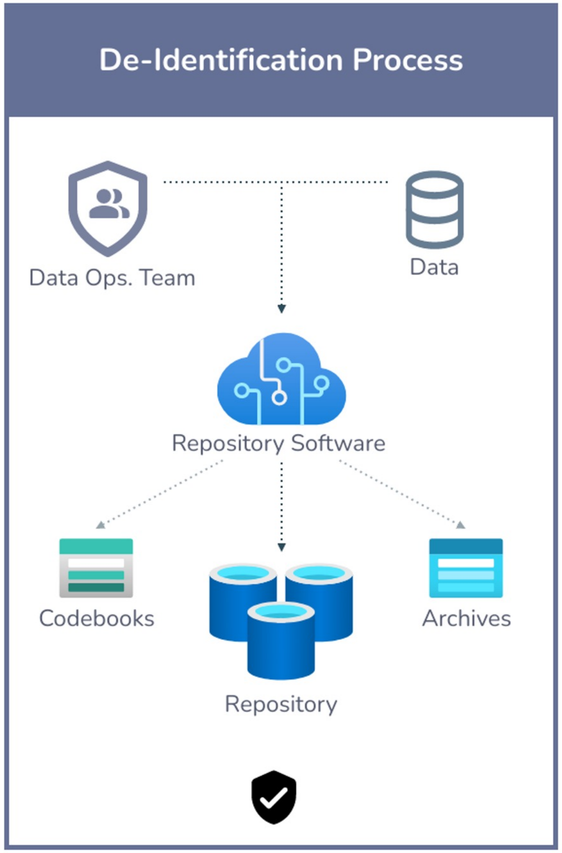




De-Identification Process

- Based on the PII and confidential flagged during the review process, data de-identification includes:
  - Scrambling Person- and Event-level Identifiers
  - Removing Person-level identifiable elements
  - Extracting Year-of-birth from Date-of-Births
  - Truncating Zip Code variables to first 5 Zip code

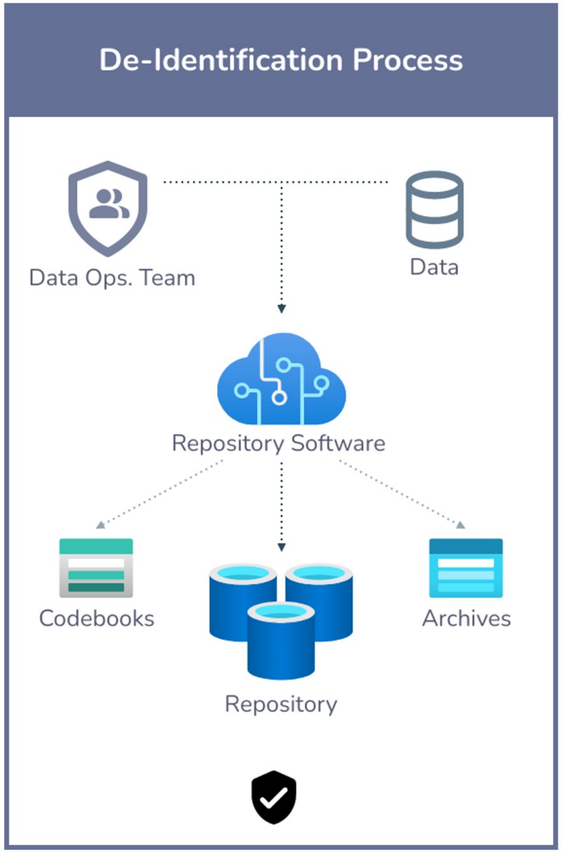




De-Identification Process

Data Hashing

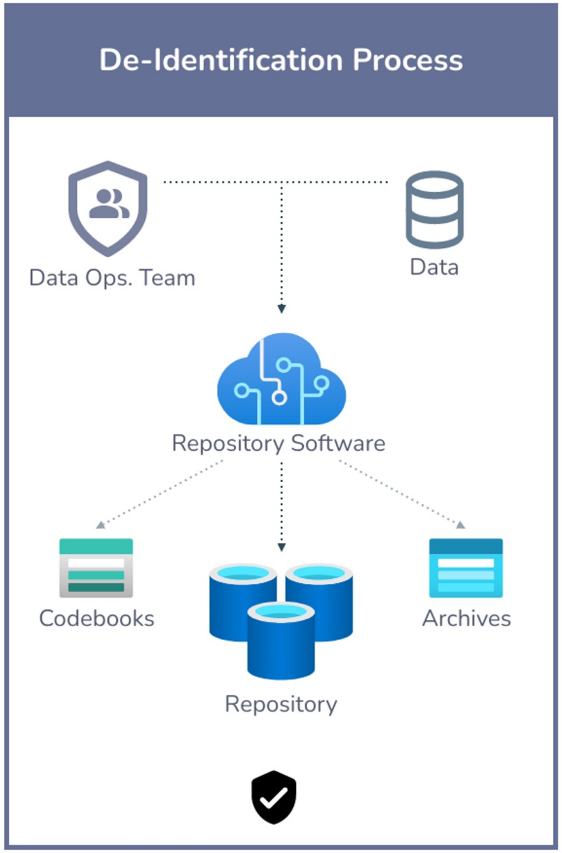
- In order to scramble identifiers, each individual value is hashed through a mathematical hashing function
- This produces a new alphanumeric number for each unique original value
- The process ensures that data is hashed independently of variable name, file name, time, or any factors other the actual value itself
- ISLG maintains internal-only codebooks for data troubleshooting and re-identification purposes



De-Identification Process

Version Control

- The program also monitors any changes carried out on the original data for any file already part of the repository
- Any updates to an original data file will prompt the program to archive the previous de-identified file, and will generate an updated de-identified file
- This updated file will also include a higher version number stored within the file name
- This process allows ISLG to version control it’s data, and ensure all the data maintained in the Repository is kept up-to-date



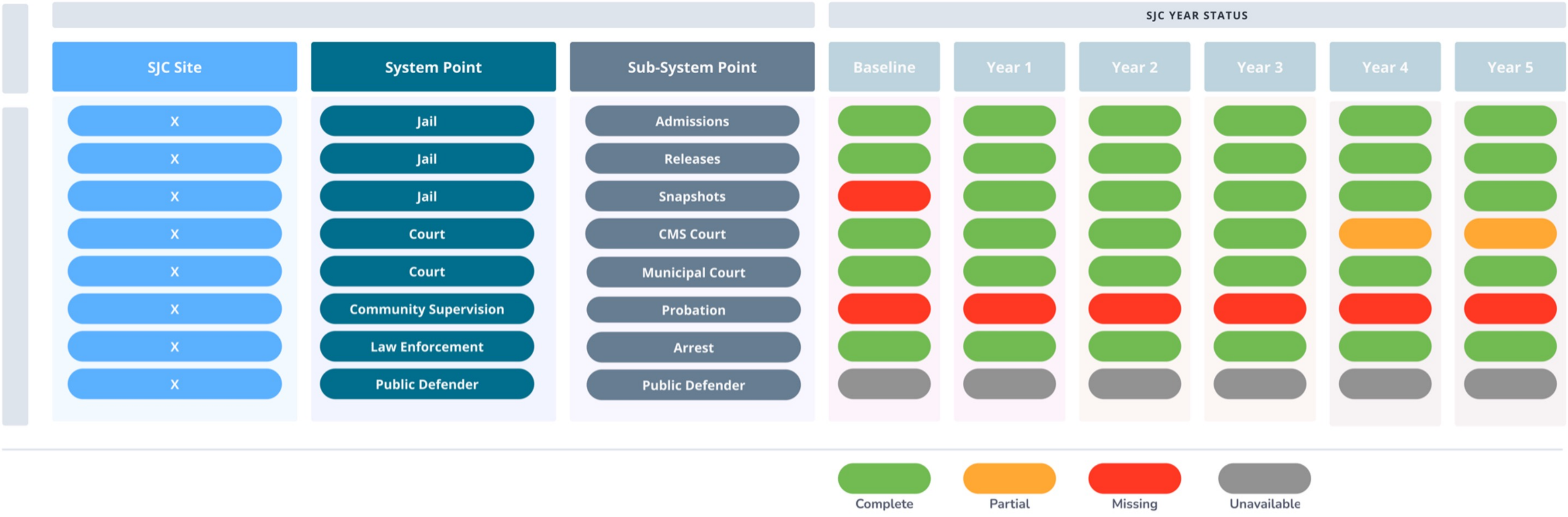
Data Inventory Management

- An important part of maintaining the Repository is managing and tracking the status of ISLG's current data holdings
- A byproduct of building the Repository has been the development of an internal inventory of data, simply called the Data Inventory
- The Inventory is a series of documents that track the status of ISLG's data, broken down by System Point, Sub-System Point and Inclusion Period
- This allows ISLG to identify precise gaps in it's SJC holdings, and then request sites for any such missing data
- Developing the Inventory and requires ISLG to map each individual file onto specific SJC years

Data Inventory Management

- The Inventory tracks the status of a site’s data by flagging each set of data as one of four different categories:

1. Complete 3. Missing
2. Partial 4.
Unavailable



			SJC YEAR STATUS					
SJC Site	System Point	Sub-System Point	Baseline	Year 1	Year 2	Year 3	Year 4	Year 5
X	Jail	Admissions	🟩	🟩	🟩	🟩	🟩	🟩
X	Jail	Releases	🟩	🟩	🟩	🟩	🟩	🟩
X	Jail	Snapshots	🟥	🟩	🟩	🟩	🟩	🟩
X	Court	CMS Court	🟩	🟩	🟩	🟩	🟧	🟧
X	Court	Municipal Court	🟩	🟩	🟩	🟩	🟩	🟩
X	Community Supervision	Probation	🟥	🟥	🟥	🟥	🟥	🟥
X	Law Enforcement	Arrest	🟩	🟩	🟩	🟩	🟩	🟩
X	Public Defender	Public Defender	⬜	⬜	⬜	⬜	⬜	⬜

🟩 Complete 🟧 Partial 🟥 Missing ⬜ Unavailable

Deliverables

Monthly Jail Report (MJR)

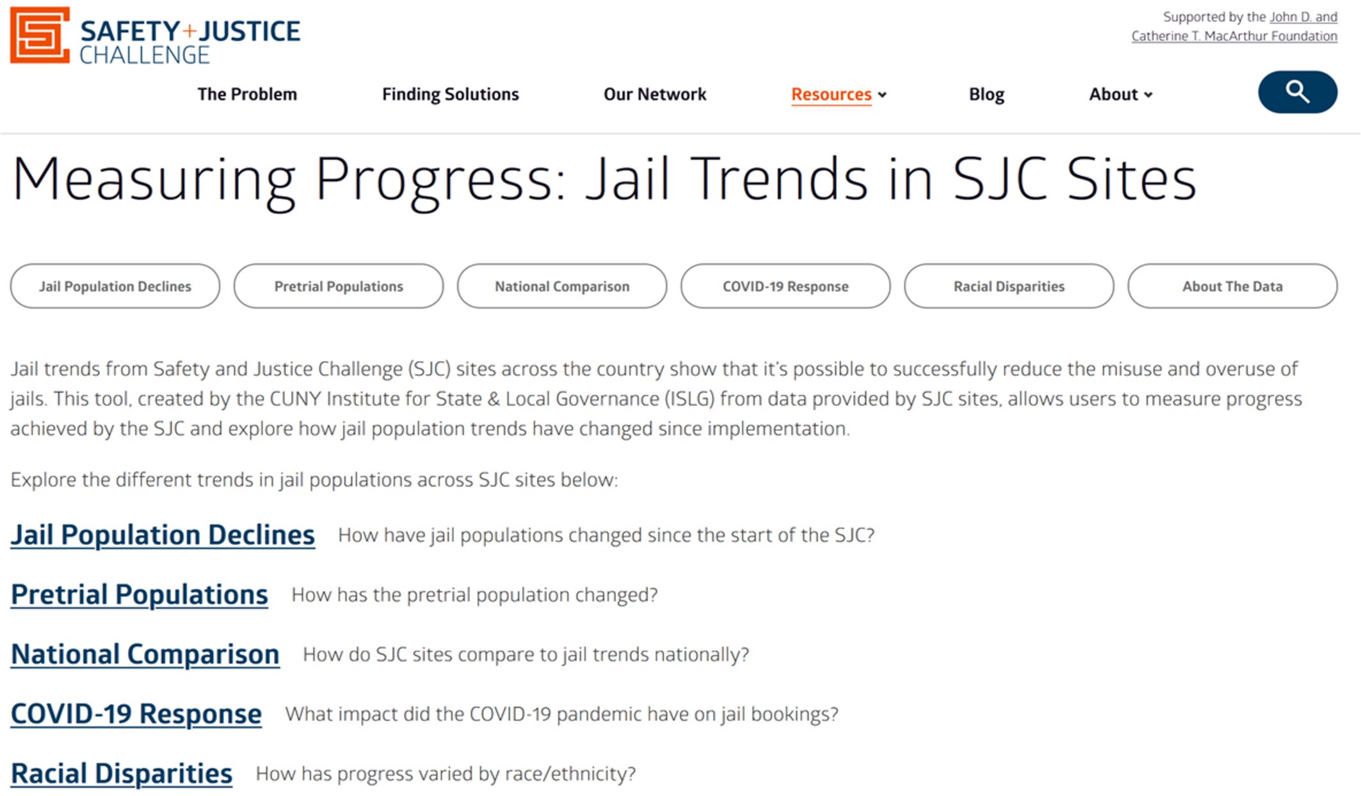
- The **SJC Monthly Jail Trends Report** is an interactive internal dashboard for peers and partners of the SJC that uses monthly aggregate data submitted by SJC sites on a monthly basis. This report allows ISLG to monitor high level jail population trends on a monthly and quarterly basis
- For every SJC quarter, ISLG produces a **Quarterly ADP Report** using MJR data, which is used on the **Participating Cities, Counties, and States** interactive map on the SJC website
- These products provide sites and partners with information on how performance indicators have changed since baseline, with the goal of using them as a tool to facilitate exploration into jail trends and how they relate to strategies being implemented on the ground

Deliverables

Jail Performance Measures

- The Jail Performance Measures synthesize and standardize case level data from 17 SJC Sites to a uniform set of aggregated quarterly measures, including ADP, bookings, and ALOS, as well as measures of racial and ethnic disparities.
- These measures are broken down by race/ethnicity, age, sex, legal status, charge severity, and frequent utilizer status
- These performance measures allow an in depth view on the progress of SJC sites in reducing jail populations and reducing racial and ethnic disparities, as well as comparison of performance across sites.
- Performance measures are calculated at quarterly intervals because this allows for a better view of progress over the course of a year relative to yearly metrics, while at the same time reducing the potential for aberrant months to skew trends

Public facing data tool



Measuring Progress: Jail Trends in SJC Sites

Jail trends from Safety and Justice Challenge (SJC) sites across the country show that it's possible to successfully reduce the misuse and overuse of jails. This tool, created by the CUNY Institute for State & Local Governance (ISLG) from data provided by SJC sites, allows users to measure progress achieved by the SJC and explore how jail population trends have changed since implementation.

Explore the different trends in jail populations across SJC sites below:

Jail Population Declines How have jail populations changed since the start of the SJC?

Pretrial Populations How has the pretrial population changed?

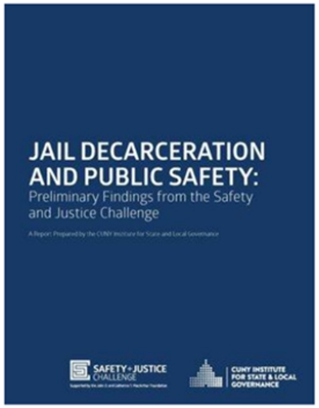
National Comparison How do SJC sites compare to jail trends nationally?

COVID-19 Response What impact did the COVID-19 pandemic have on jail bookings?

Racial Disparities How has progress varied by race/ethnicity?

Published Reports: Annual Performance Measure Reports & Consortium Research

ISLG:



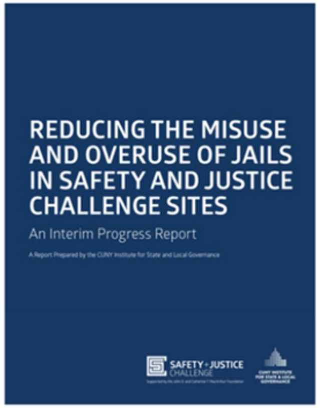
JAIL DECARCERATION AND PUBLIC SAFETY:

Preliminary Findings from the Safety and Justice Challenge

A Report Prepared by the CUNY Institute for State and Local Governance

SAFETY+JUSTICE CHALLENGE

CUNY INSTITUTE FOR STATE & LOCAL GOVERNANCE

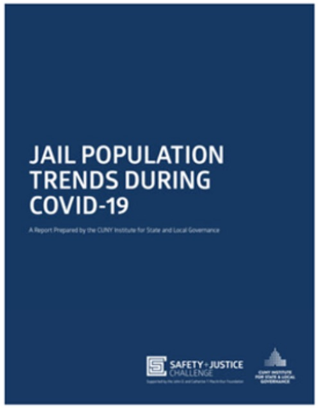


REDUCING THE MISUSE AND OVERUSE OF JAILS IN SAFETY AND JUSTICE CHALLENGE SITES

An Interim Progress Report

A Report Prepared by the CUNY Institute for State and Local Governance

SAFETY+JUSTICE CHALLENGE

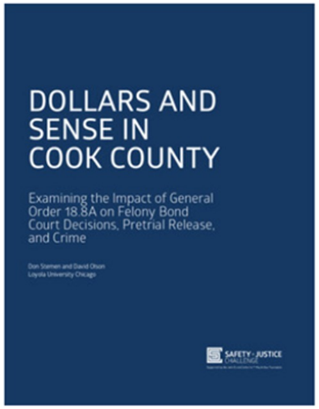


JAIL POPULATION TRENDS DURING COVID-19

A Report Prepared by the CUNY Institute for State and Local Governance

SAFETY+JUSTICE CHALLENGE

Consortium:

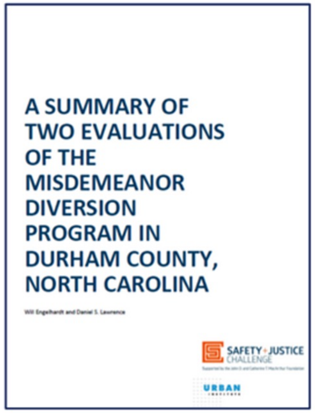


DOLLARS AND SENSE IN COOK COUNTY

Examining the Impact of General Order 18.8A on Felony Bond Court Decisions, Pretrial Release, and Crime

Don Stemen and David Olson
Loyola University Chicago

SAFETY+JUSTICE CHALLENGE



A SUMMARY OF TWO EVALUATIONS OF THE MISDEMEANOR DIVERSION PROGRAM IN DURHAM COUNTY, NORTH CAROLINA

Will Engelhardt and Daniel S. Lawrence

SAFETY+JUSTICE CHALLENGE

URBAN INSTITUTE

EXAMINING THE IMPACTS OF ARREST DEFLECTION STRATEGIES ON JAIL REDUCTION EFFORTS

Synthesis Report

SAFETY+JUSTICE CHALLENGE

JSP

Introduction to External Criminal Justice Data Sources

External Criminal Justice Data Sources

Uniform Crime Reporting (UCR)

- The Uniform Crime Reporting (UCR) Program generates reliable statistics for use in law enforcement
- The UCR Program includes data from more than 18,000 city, university and college, county, state, tribal, and federal law enforcement agencies. Agencies participate voluntarily and submit their crime data either through a state UCR program or directly to the FBI's UCR Program
- This report from the provides crime statistics that aids in various SJC special analyses managed by ISLG, particularly the Public Safety Analysis

External Criminal Justice Data Sources

Population Estimates

- ISLG uses the Bridged-Race Population Estimates from the CDC as the primary source of population data
- The CDC's National Center for Health Statistics releases bridged-race population estimates of the resident population of the United States by single-year of age and race/ethnicity at the county level for each county in the United States
- This data aids in various SJC initiatives and special analysis, such as the Monthly Jail Report and Public Safety Analysis by helping calculate disproportionality, disparity, booking and incarceration measures as among others

External Criminal Justice Data Sources

Annual Survey of Jails (ASJ)

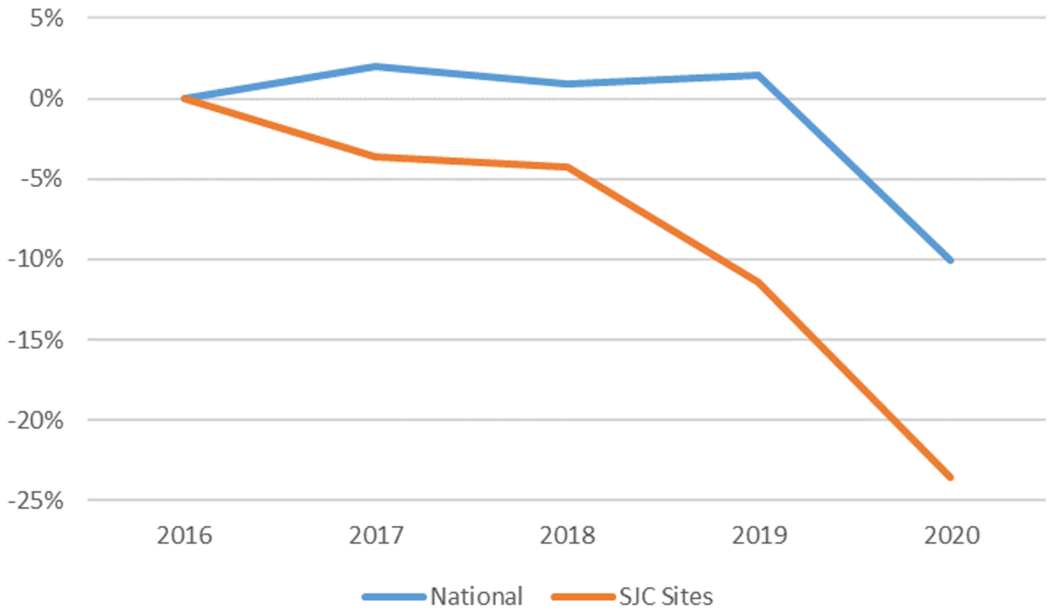
- Administered to a sample of approximately 950 local jails, the Annual Survey of Jails (ASJ) provides estimates on the number of inmates confined in jails by demographics, criminal justice status of the jail population, among other categories.
- ISLG uses this report to understand jail populations of SJC sites in comparison to national trends. This helps us to see how SJC sites are performing in reducing jail populations.

External Criminal Justice Data Sources

Annual Survey of Jails (ASJ)

- Using ASJ data, we are able to compare the progress SJC Sites against national jail trends
- SJC sites achieved greater declines than the national average prior to the pandemic, and declined at a similar rate during the pandemic

SJC Sites Outpaced National Jail Population Declines

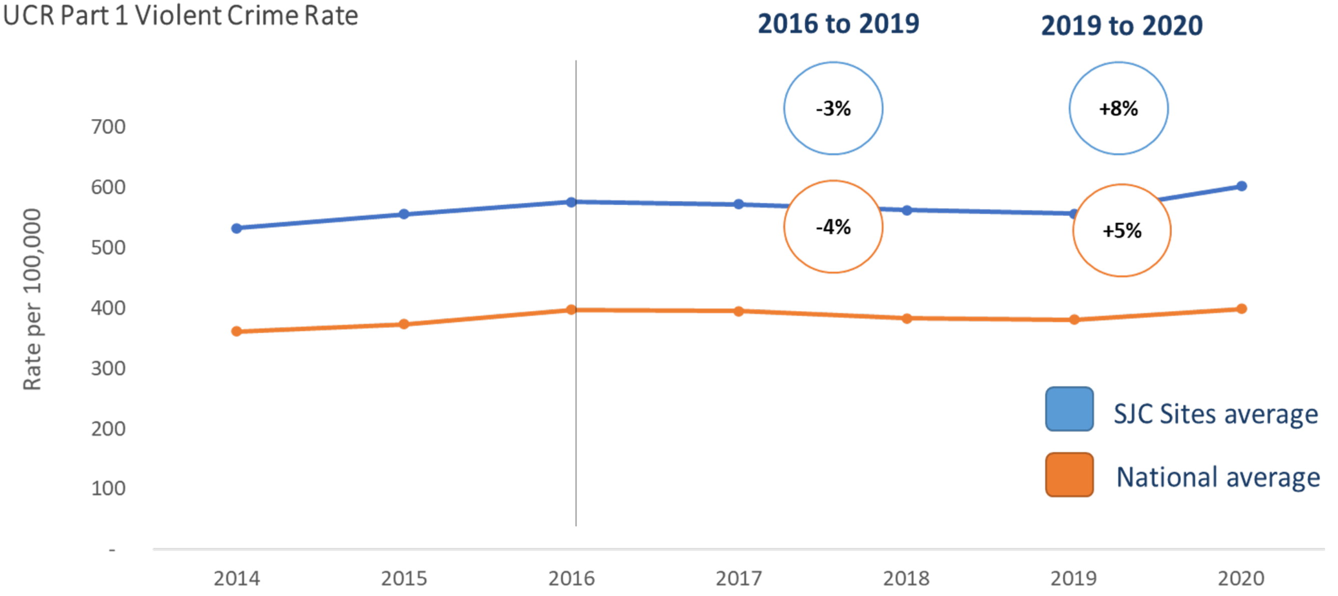


External Criminal Justice Data Sources

- Using UCR data, we are able to compare crime rates in SJC site counties against the national average
- SJC implementation in 2016, violent crimes declined across sites. After the COVID-19 pandemic, SJC sites on average saw a similar increase in violent crime as the nation between 2019 and 2020.

% change in violent crime

UCR Part 1 Violent Crime Rate



Technical Limitations and Challenges

Technical Limitations and Challenges

- Managing administrative criminal justice data is often challenging within the context of a single agency or program, but doing this across multiple agencies and multiple sites brings added technical complexities
- The scale and scope of the SJC initiative and data collection activities pose several challenges
- **Little consistency exists across sites' administrative data**, owing to multiple different factors, which can include:
  - Variation in both state and local law and policy
  - Highly localized data entry practices
  - Changes in data systems and policies over time

Technical Limitations and Challenges

- The scale and scope of variations in administrative data collected across sites raised particular challenges when trying to develop generalized infrastructure that needs to understand and process this data
- A handful of variations that need to be taken into account include:
  - Types of files we receive
  - Data formatting
  - How they represent missing data
  - How they format date variables
  - How they label date variables
  - When encountering large variations in data within a single element, how to distinguish between systematic variations and data errors

Technical Limitations and Challenges

- **Data Quality Issues**
  - Data quality issues, whether as a result of manual input or systemic in nature, can be be a consistent challenge
  - Messy source data is always a hurdle when trying to develop generalized software that needs to process this data
- **Changing and Fragmented Data Systems**
  - SJC jurisdictions can change systems year over year, and it can raise challenges every time this happens
- **Lack of Consistent Unique Identifiers Shared Across Agencies**
  - Unique identifiers assigned to cases or people are sometimes not shared across agencies, and this can be challenge when trying to determine whether the underlying data is shared across multiple identifiers

Supported by the John D. and Catherine T. MacArthur Foundation

QUESTIONS AND DISCUSSION

Osama Qureshi, Data Scientist, ISLG
Osama.Qureshi@islg.cuny.edu

Brian Holliday, Data Scientist, ISLG
Brian.Holliday@islg.cuny.edu

Stephanie Rosoff, Associate Research Director, ISLG
Stephanie.Rosoff@islg.cuny.edu

SafetyAndJusticeChallenge.org