OPENJUSTICE

Data-driven public safety

Justin Erlich, Special Assistant Attorney General
Office of Attorney General Kamala D. Harris
California Department of Justice

OpenJustice.doj.ca.gov

What is OpenJustice?

*A **transparency** initiative led by the California Department of Justice that **publishes criminal justice data** so we can **understand how we are doing**, hold ourselves accountable, and **improve public policy** to make California safer.*

Vision For OpenJustice

- Ongoing national conversation on how to **strengthen trust** and **improve performance** of the criminal justice system

- **Smart on Crime approach** – transparency & metrics
  - ❖ Use transparency to strengthen trust
  - ❖ Develop metrics to understand how we are doing and where to prioritize interventions
  - ❖ Identify evidence-based investigative practices

- **Three key components**
  - ❖ Justice Dashboard to show visually how we are doing
  - ❖ Open Data Portal to make raw data available to public
  - ❖ Data reporting/collection improvement effort

Theory of Change

Collect	Analyze	Publish	Engage

Good data

- Improve quality of existing data

- Move towards real time reporting

- Ensure we are have the right data fields

Rigorous analytics

- Bring big data methodologies

- Partner with researchers

- Identify policy improvements

Transparency

- Release raw data

- Lift up clear and actionable insights

- Crowdsource data analysis

- Share policies that are working

Awareness

- Engage policy makers and local criminal justice partners

- Inform the community to build trust

The first six datasets

The Dashboard visualizes the data to give the public a clearer statewide picture
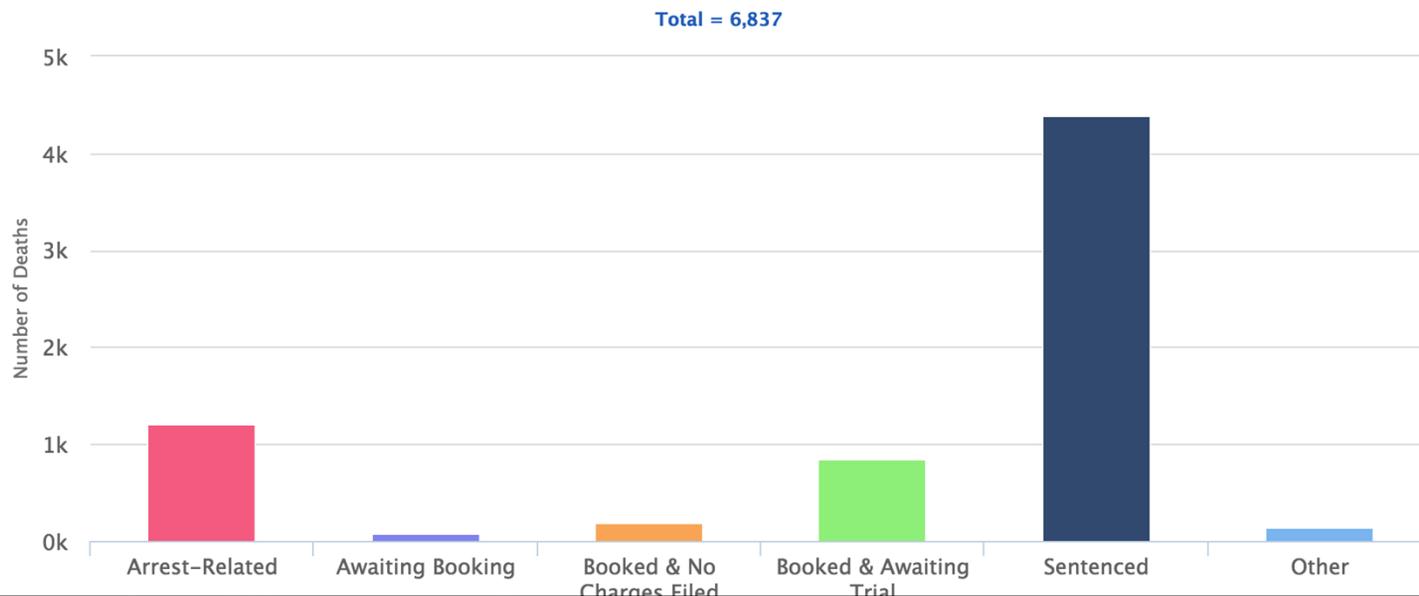
It also shows differences across counties and agencies throughout the state
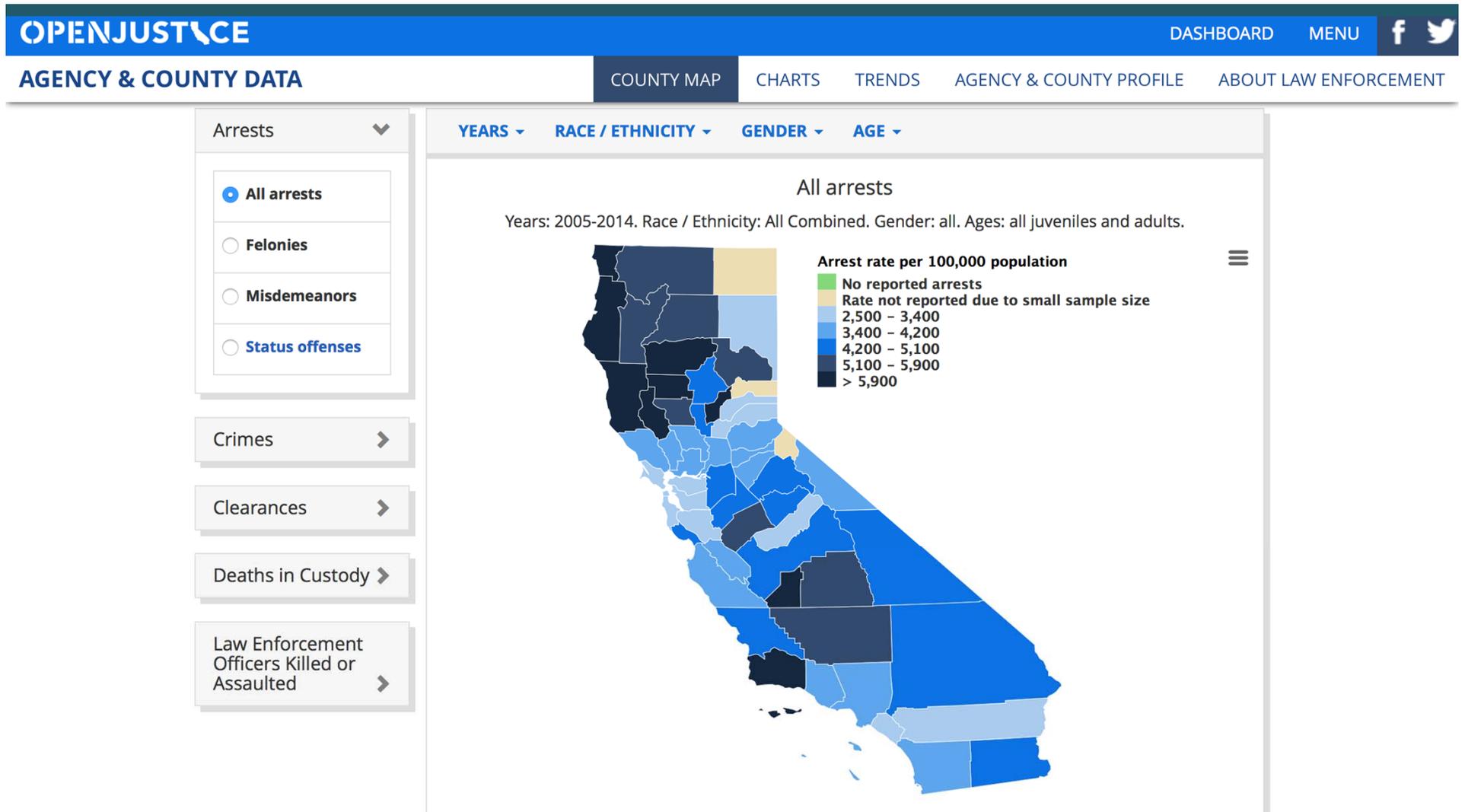
❖ **Includes contextual data such as pop. demographics, education, employment, etc.**

Open Data benefits & risks

Benefits

❖ Data is a "public good"

❖ Transparency builds trust

❖ Open by default is a signaling function; reduces transaction costs

❖ Increases potential for research capacity by expanding access

❖ Unlocks innovation and possibilities

Risks

❖ Identity disclosure (e.g., Mugshots.com) and/or re-identification

❖ Inferential disclosure

❖ Bad data = bad policy

Tensions unique to criminal justice data

❖ Unlike most health and education data, certain criminal justice data tends to be available in local jurisdictions (e.g. court records, PRAs)

❖ But aggregated statewide "Criminal Offender Record Information" (CORI) is still confidential

❖ Criminal records can be stumbling blocks to getting jobs, housing, etc. so stakes are high

❖ There are additional access points (e.g., press releases, news coverage) that makes open criminal justice data particularly vulnerable to linkage attacks

❖ No guidelines like HIPAA or FERPA in the law enforcement space

Key attributes of the data

- ❖ In addition to unique identifiers (e.g., Name), there are a generally a core set of data fields that can be quasi-identifiers:
  - ❖ Gender
  - ❖ Race
  - ❖ Age
  - ❖ Offense Type
  - ❖ Date of Offense
  - ❖ Location/jurisdiction of incident

- ❖ The problem of small numbers and geographical boundaries
  - ❖ Small cities, counties, law enforcement pose the biggest challenges
  - ❖ As we move to more granular geographical data collection, this will only get more challenging

- ❖ We are exploring "binning"
  - ❖ Combining into age buckets
  - ❖ Month of offense rather than date
  - ❖ Aggregating jurisdictional data under a certain population

Dataset snowflakes

❖ Each dataset may contemplate a different balancing test, depending on whose data we are dealing with

❖ Different stakeholder groups, including:

    ❖ Victims of crimes (e.g., sexual assault, DV, child abuse)

    ❖ Decedents or relatives (e.g., death in custody)

    ❖ Gun owners

    ❖ Law enforcement personnel

    ❖ Special class groups (e.g., Juveniles, mental health patients)

    ❖ Individual arrested but released, acquitted etc.

    ❖ Individuals convicted

"Responsible transparency"

➢ **Tiered data access**
  ➢ "Anonymized" data on open data portal + data use agreement
  ➢ Some data w/ PII available to external researchers upon application
  ➢ Some data only available to internal researchers

➢ **Strategy for "anonymized" data**
  ➢ Holding internal conversations on risk appetite
  ➢ Forming committee of experts to advise on value of granularity vs. risk of re-identification
  ➢ Seeking collaboration with data scientists to model risks and provide recommended levels of depth
  ➢ Exploring hack-a-thon approach to stress test data for disclosure risks

➢ **Transparency on transparency**
  ➢ Drafting white paper on how we approached this issue and developed our policies