How Researchers De-Identify Data in Practice

Wentao Guo, Paige Pepitone,[1] Adam Aviv,[2] Michelle Mazurek

University of Maryland

[1] NORC at the University of Chicago

[2] The George Washington University

wguo5@umd.edu

@wentaoguo.bsky.social

@wentaochirps

Researchers publish lots of data online

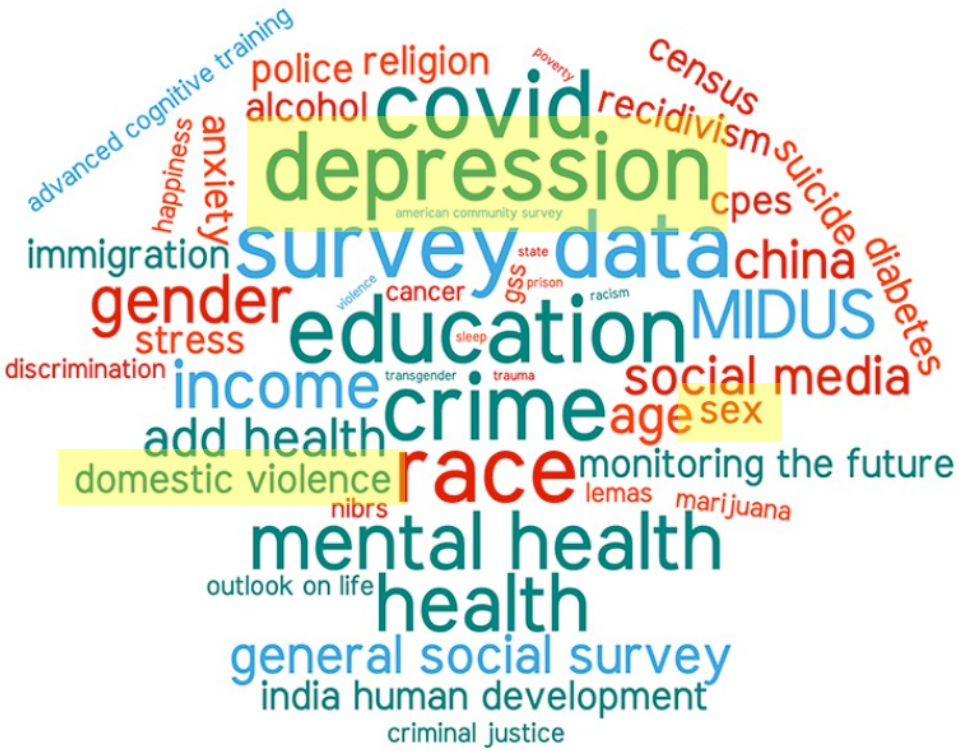
Statistics

22,955 studies

{x} 6,908,670 variables

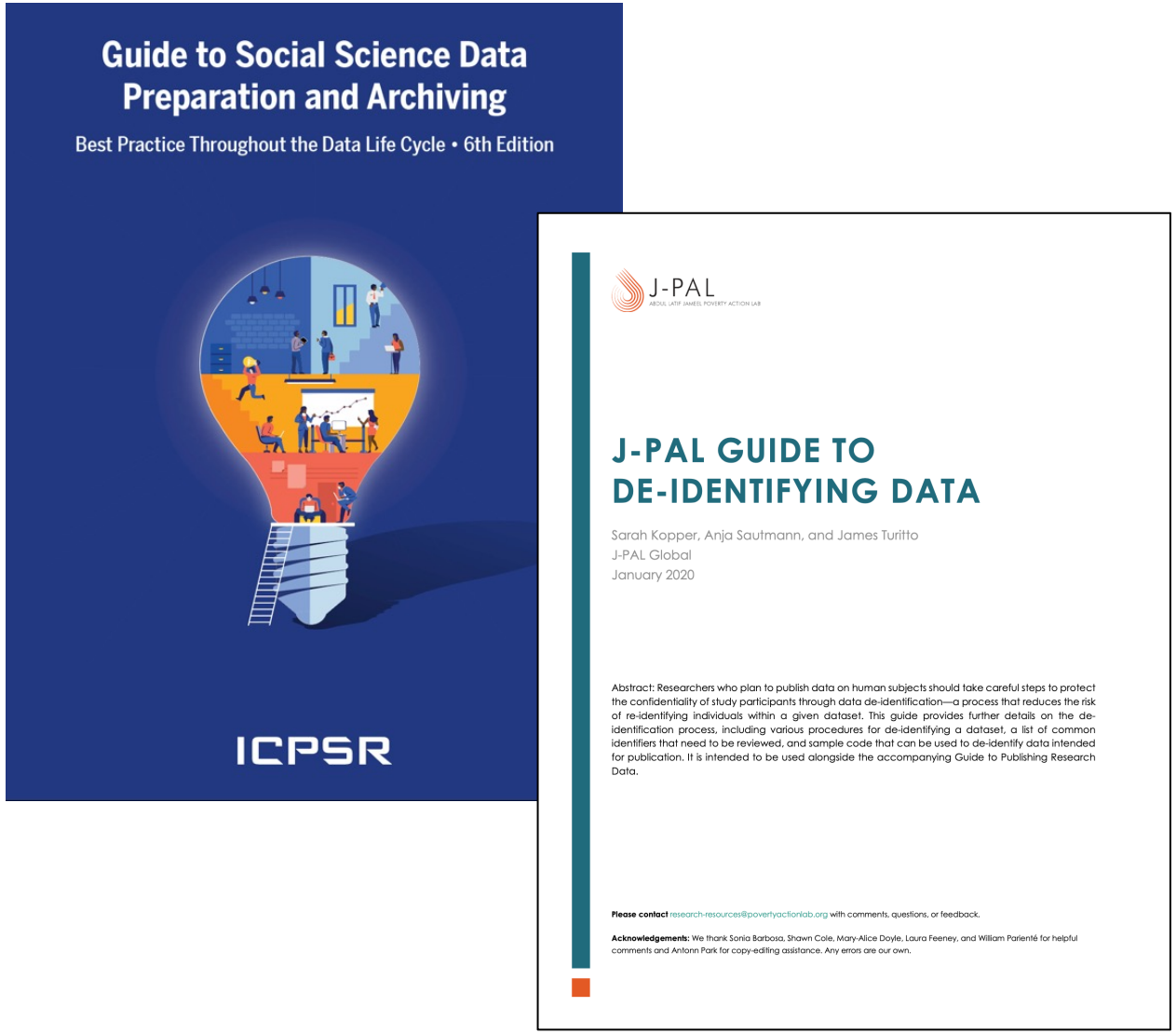
119,827 publications

Most Popular Search Terms



Researchers are expected to ***de-identify*** data:

modifying **non-aggregate data**
to make it **more difficult**
to **re-identify or learn info**
about **individuals**



Guides make misleading claims about de-identification outcomes

Guo et al. (2024), “A Qualitative Analysis of Practical De-Identification Guides”

Semi-structured interviews

RQ1. How do researchers perceive **re-identification threats**?

RQ2. How do researchers **de-identify** data in practice?

…

We recruited…

- **18** practitioners at universities and research companies
- **6** curators at repositories and funding agencies

Researchers are concerned about **combinations of indirect identifiers** that could link individuals to external data

> You want to avoid putting clinicians into a **group of less than five similar clinicians**.
>
> – *P6*

quotes condensed for brevity

In practice, though, researchers only inspect **pairwise combinations of identifiers** to evaluate de-identification success

> You could crosstab all variables in theory, but that would be like millions of crosstabs. Maybe it's somebody's **position, crosstabbed with their age** or gender. It's not necessarily a scientific process. It's **more knowing what to look for**.
>
> *– P12*

How Researchers De-Identify Data in Practice

Wentao Guo, Paige Pepitone, Adam Aviv, Michelle Mazurek

Come find my poster for more on…

- Root causes of the gulf between risk model and implementation
- De-identification challenges
- Curator-practitioner dynamics
- Impressions of differential privacy

Qi Baishi

Nobody else has two chickens!

wguo5@umd.edu
@wentaoguo.bsky.social
@wentaochirps