



## Use cases for Differential Privacy with tiered access to restricted data

Use Cases workshop

OpenDP Community Meeting

May 14, 2020

# ICPSR



Founded in 1962 by 22 universities, now consortium of 800 institutions world-wide

Focus on social and behavioral science data, broadly defined

Current holdings

- 15,000 studies, quarter million files, 5.5 million indexed variables
- 1500 are *restricted studies*, almost always to protect confidentiality
- Bibliography of Data-related Literature with 90,000 citations

Over 50K active MyData (“shopping cart”) accounts, 350K users

Thematic data collections

- Drug addiction, aging, arts, child care, education, criminal justice, demography, health and medical care, and minorities
- Data Lumos, COVID-19 data repository

Summer Program in Quantitative Methods of Social Research

# Three use cases corresponding to five tiers of access

## I. Public use data

### ICPSR curators

- Remove direct identifiers
- Collapse variables (e.g., age ranges)
- Suppress variables (e.g., detailed geography)
  - Generally following guidelines in [Statistical Policy Working Paper 22](#) (1996, 2005), focus on cell size, outliers

## II. Confidential data accessed through Secure SDA

### ICPSR curators

- Prepare data for analysis
- Researchers submit code, but cannot see microdata
- Secure SDA provides or suppresses answer to query

# Three use cases corresponding to five tiers of access

- III. Encrypted downloads of restricted data
  - ICPSR curators remove direct identifiers
  - Researchers apply through online portal, complete training, institutional agreement required
  - Computing using researcher's local environment
  - Researcher does own disclosure avoidance review
- IV. Confidential data accessed through Virtual Data Enclave
  - Researchers work in ICPSR-provided virtual environment
  - Disclosure avoidance review conducted by ICPSR staff, using essentially the same principles from FCSM memo as used for public use data production
- V. Confidential data accessed through Physical Data Enclave
  - Researchers work in ICPSR-provided physical environment
  - Disclosure avoidance review conducted by ICPSR staff, using essentially the same principles from FCSM memo as used for public use data production

# Three use cases

1. Improving production of public use data
  - Adding statistical noise rather than aggregation or suppression
2. Improving confidentiality-protecting analytical tools for analyzing (but not viewing) restricted data
  - Automation of process to add statistical noise rather than answering or refusing query
3. Improving tools for protecting analytic output created in VDE or PDE
  - Tool to add statistical noise rather than imposing *ad hoc* rules or suppressing output

**DataJeff cares about  
your privacy!**

