CIVIC DIGITAL FELLOWSHIP

Differential Privacy for the EHealth Program

Stefan Broecker

Supervised by Dennis Linders

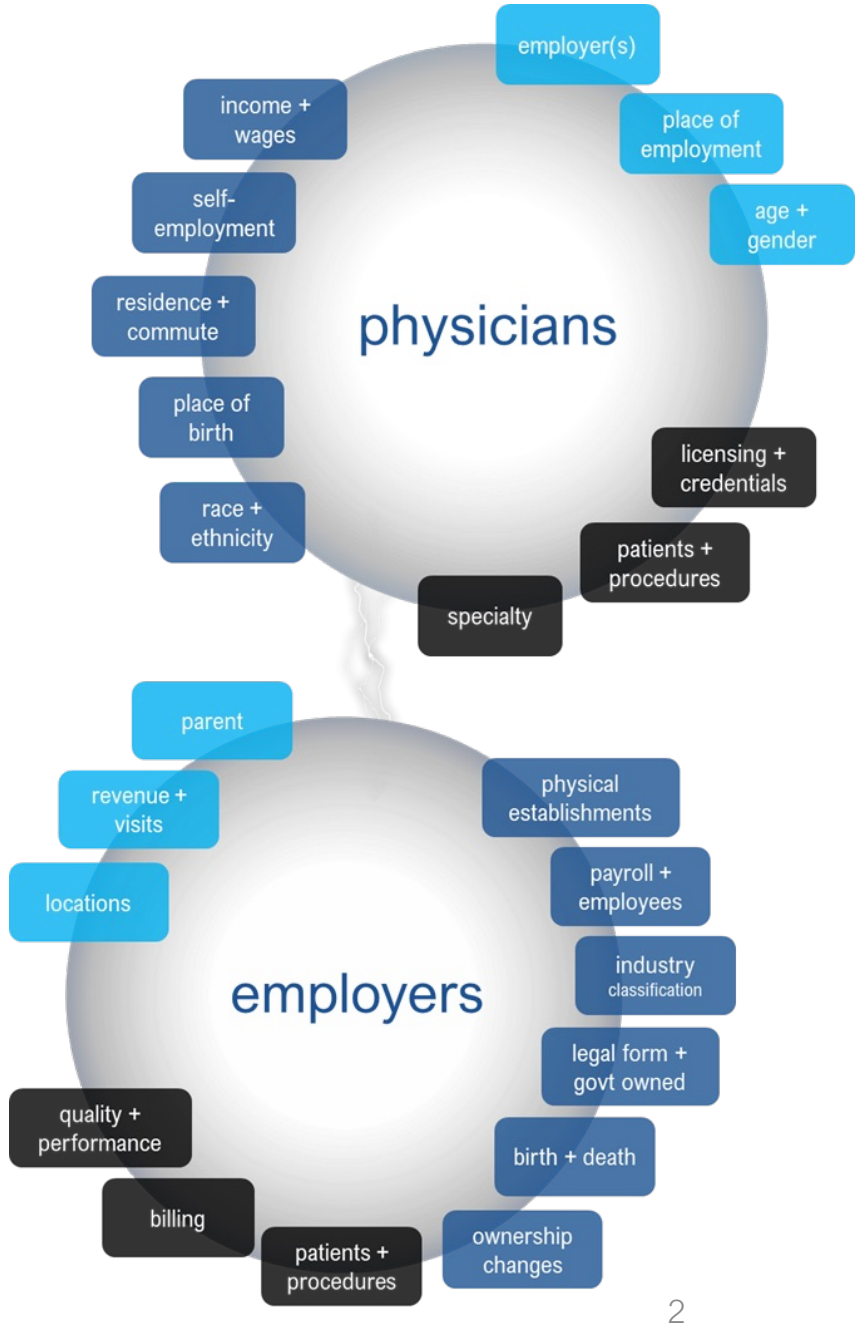
Demographic Programs - Survey Operations - Enhancing Health Data (E-Health) Team

United States® Census Bureau

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

Physician-Business Linkage Project

- **Motivation:** The business of healthcare is rapidly evolving—but not enough is known about the impact of these changes on cost, quality, and access.
- **Goal:** Leverage Census data on workers, taxpayers, businesses, and employers to enhance data on physicians and the institutions in which they work.
- **Privacy Considerations:** The EHealth Program deals with the most sensitive of data, incl. personal health records and income statements.



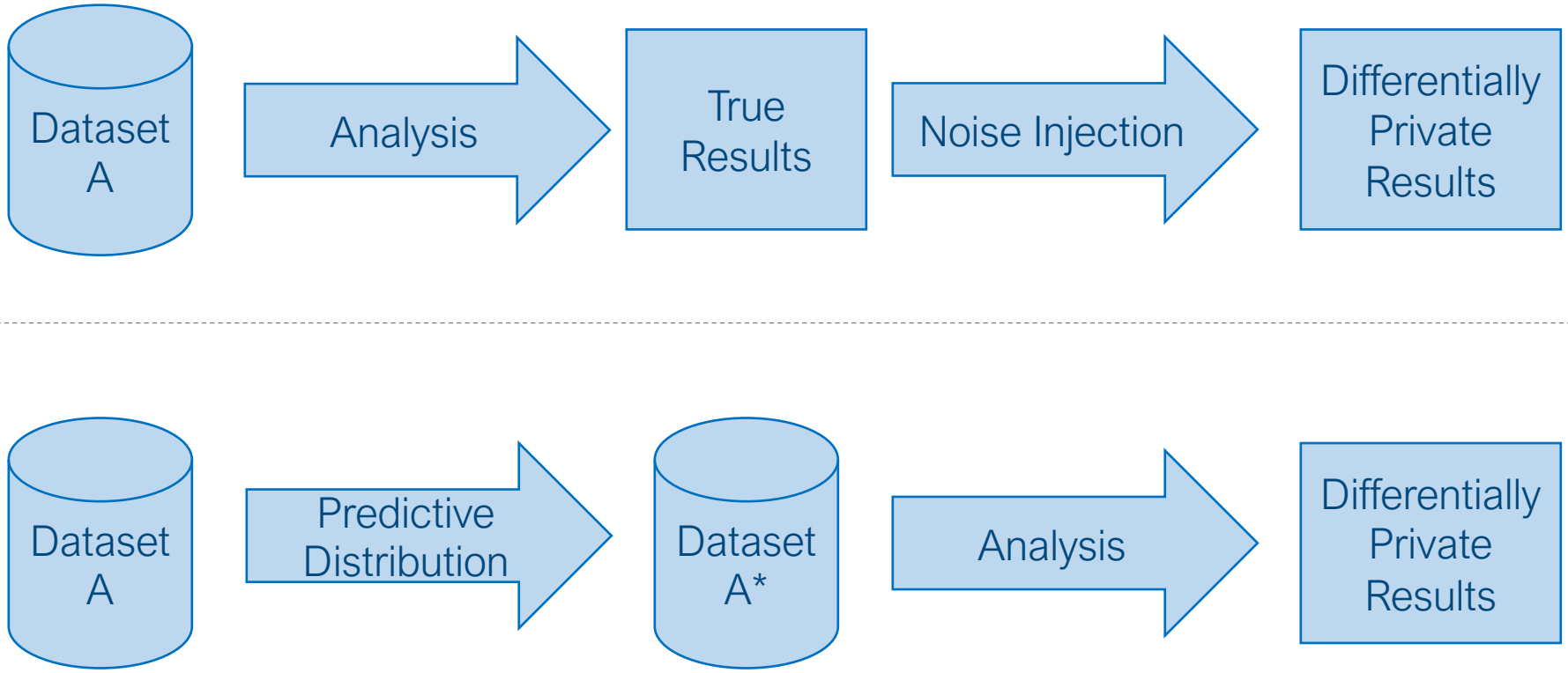
Differential Privacy: What is It?

- Differential Privacy: a strong statistical notion of privacy that bounds the amount of information a statistical release leaks about any individual.
- Differential Privacy is a formal mathematical framework for quantifying and managing privacy risks. Its key feature is that it provides a quantifiable guarantee of privacy that allows for measuring and thereby controlling risk.
- By providing this formal guarantee, differential privacy has enabled the public release of data that would not have been allowed under traditional rules for disclosure avoidance.

United States® Census Bureau

Noisy Statistics vs Synthetic Data

Two approaches considered:

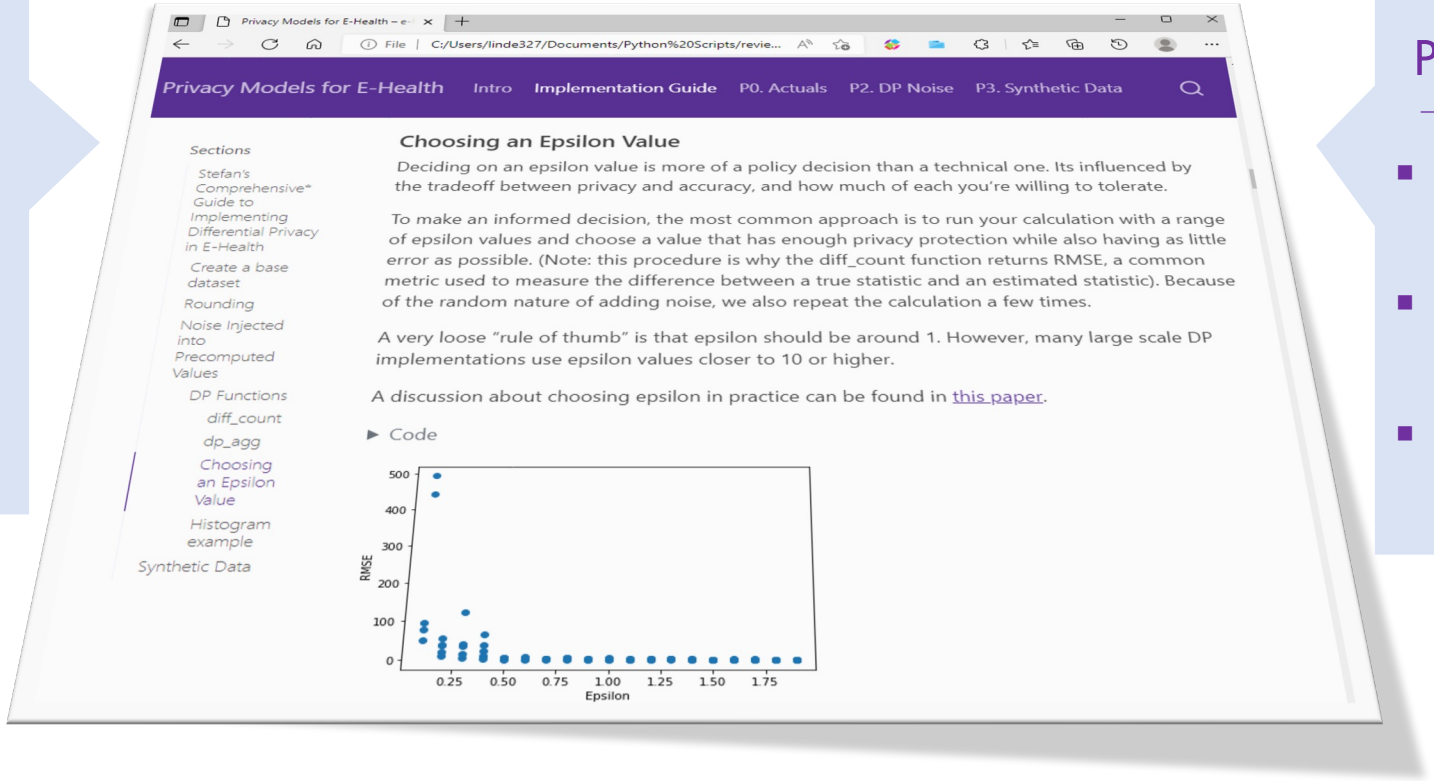


United States® Census Bureau

Deliverable: Implementation Toolkit

Part 1: Knowledge Capture

- Summary of Lit Review
- Summary of Successful Census Implementations
- Step-by-Step Implementation Guide



Part 2: Pilot Implementations

- Pipeline 1: Traditional disclosure techniques
- Pipeline 2: Differentially private aggregate tables
- Pipeline 3: Synthetic data (DP method + CenSyn)

Key Challenges and Questions

Manage multiple products against one privacy budget ▪ Applying DP to perpetual annual releases ▪ What features of the data are preserved (or not) by synthetic data ▪ Measuring accuracy of analysis done on synthetic data ▪ Providing dual privacy guarantee for individuals and businesses ▪ Business metrics that rely on a small number of large firms