SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
BEYOND BIG DATA: INFLUENCING SCIENCE, TECHNOLOGY, AND SOCIETY
PITTSBURGH, PA • JUNE 7–10, 2022

RAND CORPORATION

Introduction to Statistical Data Privacy

SDSS 2022

Statistical Data Privacy Techniques for Sharing Sensitive Data

Part 1

Goals of This Short Course

- Introduce the concept of statistical data privacy
- Cover some methods from three common frameworks
- Provide practical examples of applying these methods in R
- 10,000 foot view

Planned Schedule for Today

- Introduction to Statistical Data Privacy (1:30-2pm)
- Variable Suppression and Recoding (2:05pm-3:05pm)
- Synthetic Data (3:15pm-4:15pm)
- Differential Privacy (4:25pm-5:25pm)
- https://github.com/jsnoke/SDSS-2022-Privacy-Short-Course

Why Should We Protect Data?

- Removing directly identifying information is not sufficient
  - Numerous examples in the past 15 years
- For example:
  - ~87% of U.S. citizens are expected to be unique using zip code, gender, and birth date
  - ~99% using ICD-9 health codes, gender, birth year, and ethnicity
  - Others using browsing history, taxi rides, smartphone locations, etc.

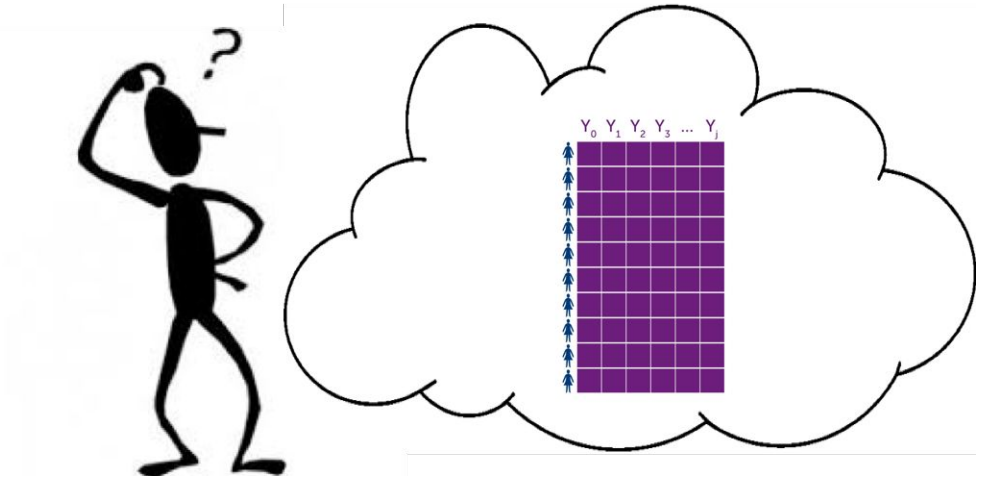
Most Famously…

- 2006 Netflix prize competition
  - Linked ratings data with IMDB
  - Privacy breaches
  - Lawsuits

Some Motivating Principles for Sharing Data

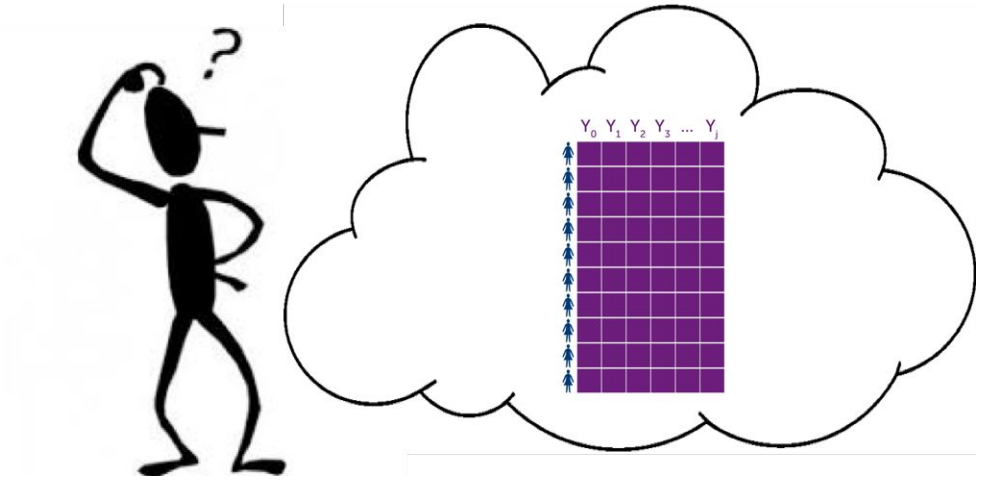
- Privacy → legal/cultural right to be protected
- Transparency/Reproducibility → trust in research
- Public good → collective knowledge sharing
- Fairness → reducing barriers to data access

How Might We Protect Data?



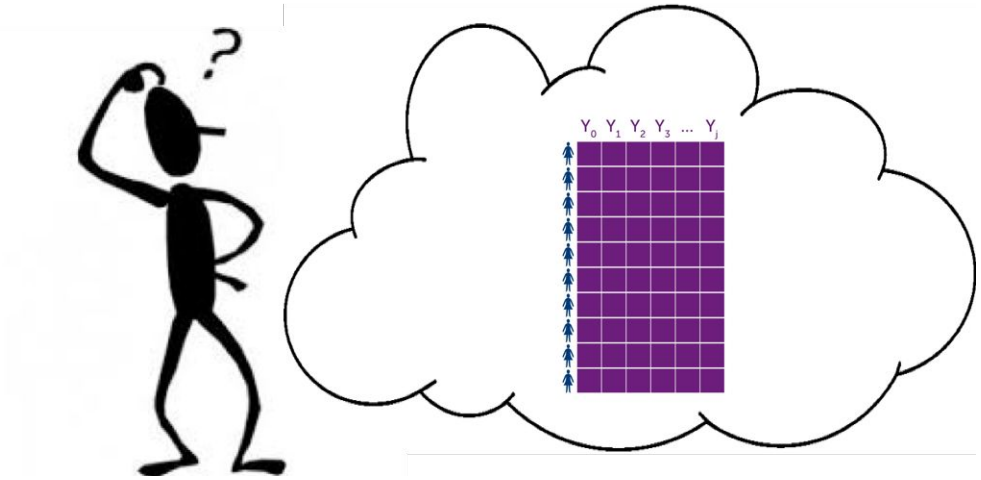
- Removing direct identifiers is not enough
- Suppose we have:
  - Demographics (e.g., age, sex, race)
  - Sensitive variables (e.g., income, medical status, political opinion)

How Might We Protect Data?



- Create a legal agreement for accessing the data?
- Suppress some data (individuals or variables)?
- Sample new values based on the existing data?
- Add some random noise?
- Do nothing?

How Might We Protect Data?



- Create a legal agreement for accessing the data?
- ***Suppress some data (individuals or variables)? [Suppression]***
- ***Sample new values based on the existing data? [Synthetic Data]***
- ***Add some random noise? [Differential Privacy]***
- Do nothing?

Why We are Covering These Three Topics

- Three of the most common approaches:
  - Significant differences between approaches towards protection
  - Each provides different types of outputs
  - Each reflects different understandings of risk
- Goal is to give enough breadth to identify methods for future use

What about Data Agreements?

- Vital aspects to the privacy infrastructure:
  - Legal agreements
  - Secure data centers
- Limitations:
  - Limited time and resources
  - Can be inequitably distributed

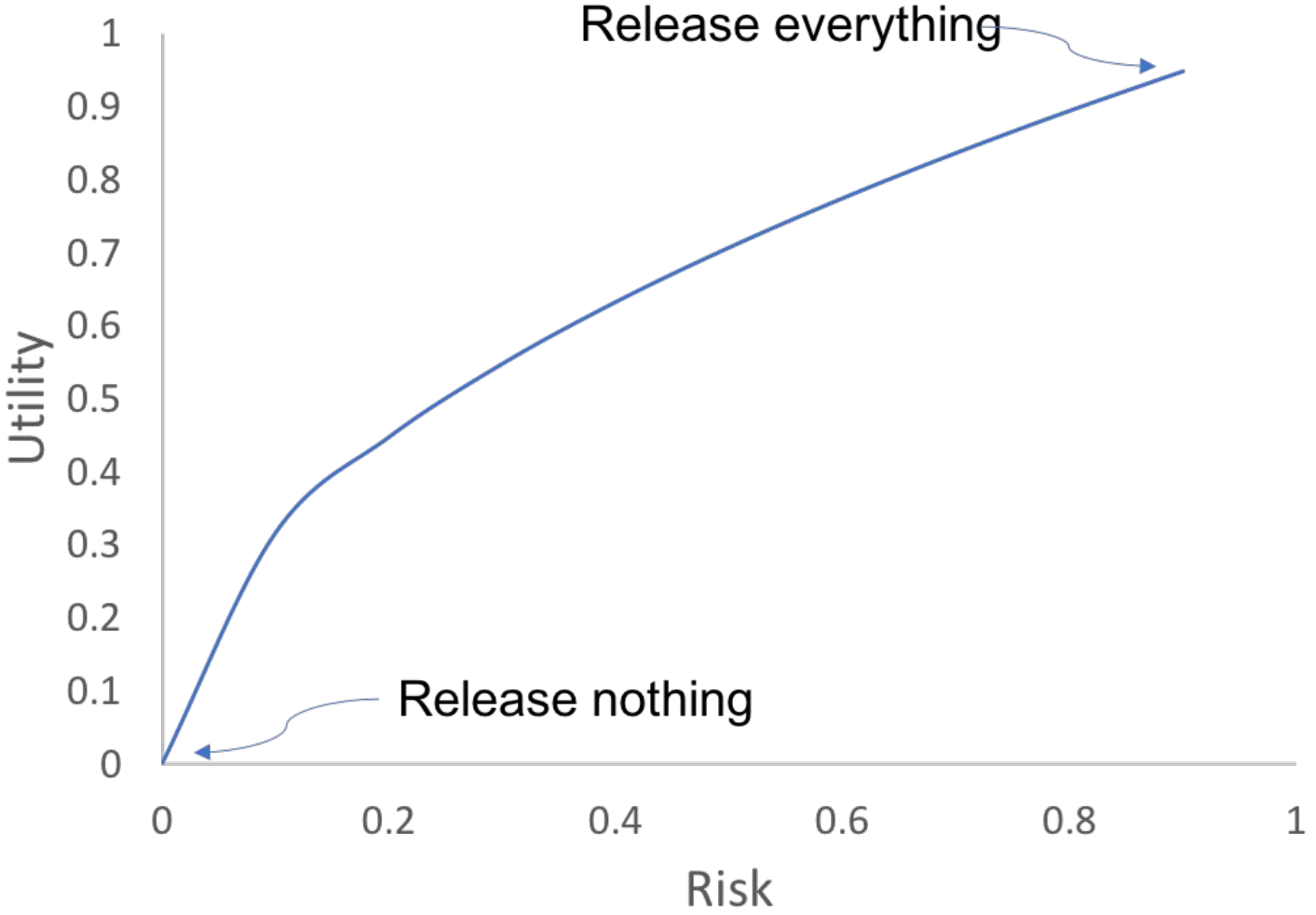
What are We Not Covering Today?

- Other frameworks exist that we will not cover today:
  - Distributed learning/secure multi-party computation
  - Query based systems
  - Validation and verification servers
- Additional protection methods:
  - Swapping
  - Top and bottom coding, microaggregation
  - Etc.
- Valid Statistical Inference:
  - Vital and growing area of research

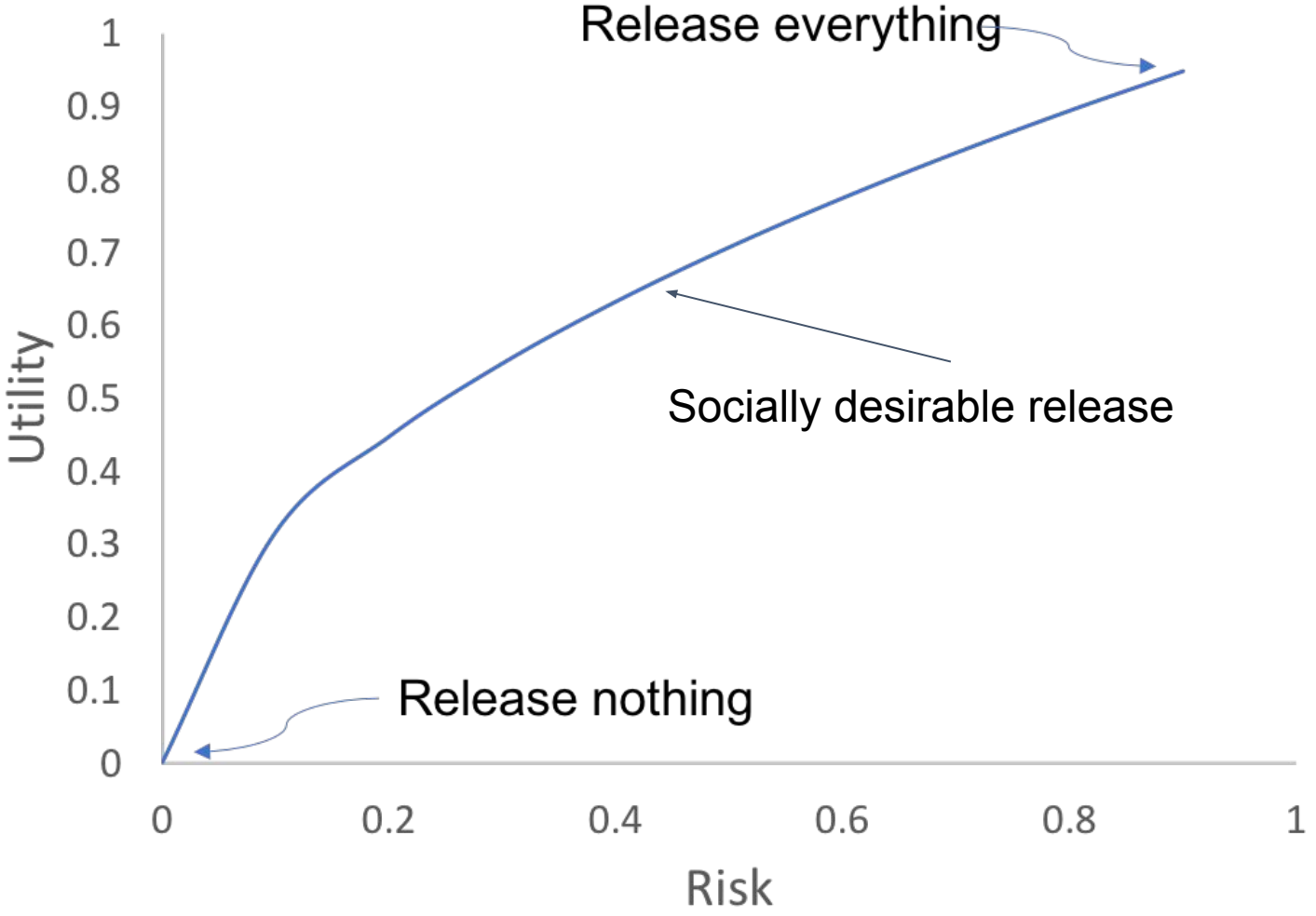
Understanding the Utility-Risk Trade-off is Crucial

- Fundamentally all methods contain a trade-off:
  - Utility vs. risk
- Broadly speaking:
  - Utility is the accuracy of the released data relative to the confidential data
  - Risk is the likelihood of learning confidential information
- Lessons will go deeper into specific definitions of utility and risk

Understanding the Utility-Risk Trade-off is Crucial

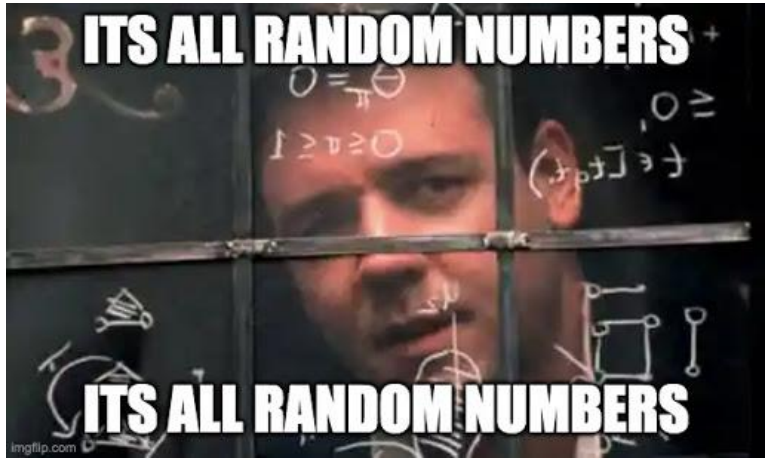


Understanding the Utility-Risk Trade-off is Crucial



A Few Last Things to Keep in Mind

- All methods:
  - Attempt to capture select statistical attributes
  - Add bias or increase variance in different ways
- Privacy is never free
- Context matters:
  - What is at stake?
  - Who needs to use the data?
  - What are we *actually* estimating?



Further Reading

- Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*. Vol. 2. New York: Wiley, 2012.
- Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111-125. IEEE, 2008.
- Loukides, Grigorios, Joshua C. Denny, and Bradley Malin. "The disclosure of diagnosis codes can breach research participants' privacy." *Journal of the American Medical Informatics Association* 17, no. 3 (2010): 322-327.
- Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre De Montjoye. "Estimating the success of re-identifications in incomplete datasets using generative models." *Nature communications* 10, no. 1 (2019): 1-9.
- Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): 557-570.