

Designing Access with Differential Privacy

Alexandra Wood

Berkman Klein Center for Internet & Society
Harvard University

Kobbi Nissim

Department of Computer Science
Georgetown University

Micah Altman

Center for Research on Equitable and Open
Scholarship
MIT

Salil Vadhan

School for Engineering and Applied Sciences
Harvard University

Background for this Talk

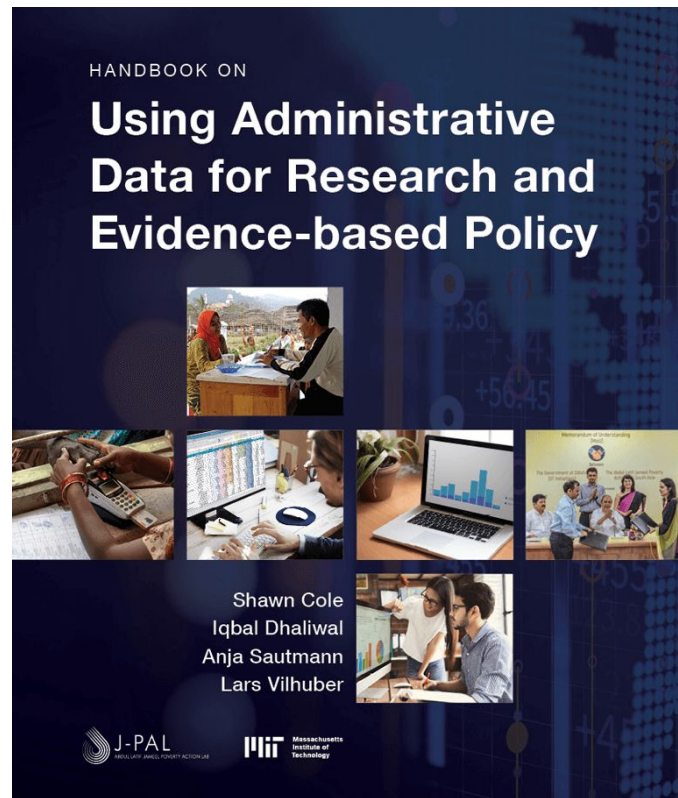
Excerpts from
“Designing Access with Differential Privacy”

a chapter in
Using Administrative Data for Research and
Evidence-based Policy

(Full text of Handbook is online.)

This talk focuses on a conceptual overview...

**For design, deployment, case studies and
resources: Q&A and full online chapter**



Why DP? Attacks on Privacy

- **Re-identification**: identifying whose record it is even after “PII” removed
 - Applied to medical data, Netflix challenge, ...
- **Database Reconstruction**: reconstructing almost the entire underlying dataset
 - Applied to Census releases and Diffix
- **Membership Inference**: determining whether a target individual is in the dataset
 - Applied to genomic data and ML as a service



Attacks on
“Aggregate”
Statistics

Takeaways from Privacy Failures

- **Specific findings:**

- Redaction of identifiers is insufficient for protecting privacy
- Similarly: aggregation, noise addition, ...
- Auxiliary information needs to be taken into account
- Regulation and technology only considered a limited scope of privacy failures
 - New failure modes: whether an individual participated in study, inferences

- **Any** useful analysis of personal data **must** leak some information about individuals
- Leakages accumulate with multiple analyses/releases



Mathematical
facts, not
matters of
policy

Hope is not lost.

Rather, we need a rigorous approach to privacy to guarantee privacy in a dynamically changing data ecosystem

Introduction of Differential privacy (2006):

- Rich theory and new privacy concepts
- Mathematically provable privacy guarantees
- In first stages of implementation and real-world use
 - US Census, Google, Apple, Uber, ...

Differential Privacy is ...

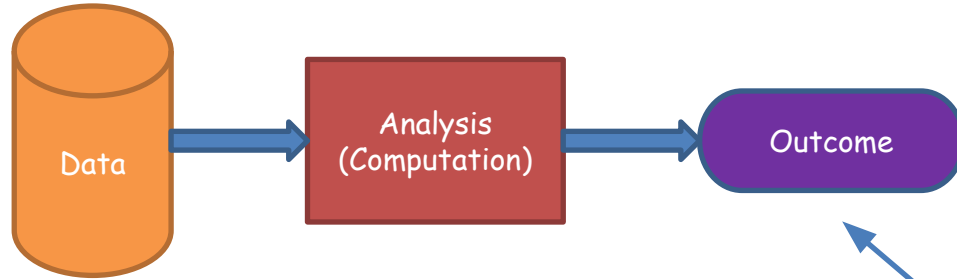
... a **definition** (i.e., a standard) of privacy
for “statistical releases”

It expresses a specific desiderata of an analysis:

Any information-related risk to a person should not change significantly as a result of that person’s information being included, or not, in the analysis

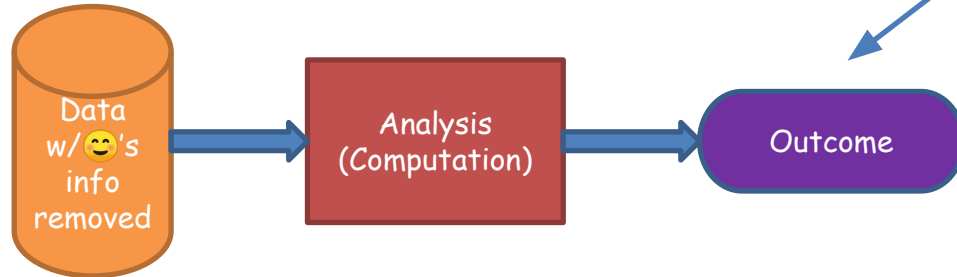
Differential Privacy [Dwork McSherry Nissim Smith '06]

Real world:



smaller ϵ – better privacy

😊's ideal world:



ϵ -“similar”

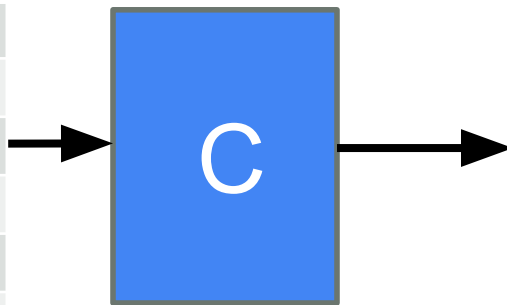
Chance of bad event almost the same in everybody's ideal and real worlds

Features of Differential Privacy

- Designed for analysis of populations, not individuals
- Requires the introduction of statistical noise
- Robust to auxiliary information, composition, and post-processing

Differential Privacy in Statistical Releases

Name	Sex	Blood		HIV?
Chen	F	B		Y
Jones	M	A		N
Smith	M	O		N
Ross	M	O		Y
Lu	F	A		N
Shah	M	B		Y

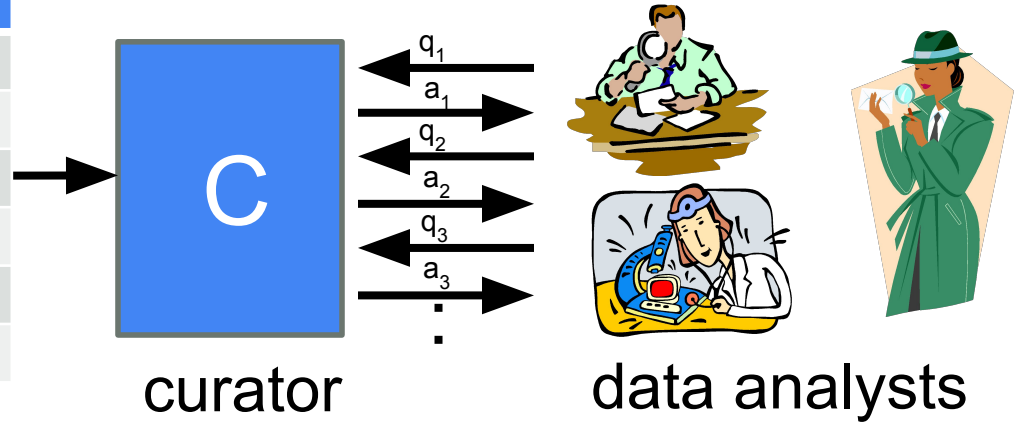


curator

statistical tables,
trained ML model,
synthetic data etc.

Statistical Query Systems

Name	Sex	Blood		HIV?
Chen	F	B		Y
Jones	M	A		N
Smith	M	O		N
Ross	M	O		Y
Lu	F	A		N
Shah	M	B		Y



Existing Query Interfaces

United States Census Bureau **AMERICAN FactFinder**

Feedback FAQs Glossary Help

MAIN COMMUNITY FACTS GUIDED SEARCH **ADVANCED SEARCH** DOWNLOAD CENTER

Advanced Search - Search all data in American FactFinder

1 Advanced Search 2 **Table Viewer** Result 1 of 1 VIEW ALL AS PDF

S0101 AGE AND SEX 2012-2016 American Community Survey 5-Year Estimates

Table View BACK TO ADVANCED SEARCH

Actions: Modify Table Add/Remove Geographies Bookmark/Save Print Download Create a Map

This table is displayed with default geographies. Click Back to Search to select other geographies using the search options on the left.

Tell us what you think. Provide feedback to help make American Community Survey data more useful for you.

Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Versions of this table are available for the following years	Subject	United States			
		Male		Female	
		Estimate	Margin of Error	Estimate	Margin of Error
2012-2016	Total population	318,958,162	~±6,427	158,763,322	~±6,432

IES NCES National Center for Education Statistics

IDE | IAP | PISA | PIRLS | TIMSS | PIAAC | TALIS

International Data Explorer

1. Select Criteria 2. Select Variables 3. Edit Reports 4. **Build Reports**

STEP 4: View each report table by selecting the report name from the drop-down menu. Create report types to edit and preview, each tab created represents one report type to report.

Subject, Age: Mathematics, Reading, and Science, 15 years
Jurisdiction: International Average (OECD Countries)
Measure: PISA Mathematics Scale: Overall Mathematics
Variable: All students
Year: 2015

Select Report: Report 1 Link to this Page Export Reports

Table Chart Significance Test Gap Analysis Regression Analysis

Averages for PISA mathematics scale: overall mathematics, age 15 years by All students [TOTAL], year and jurisdiction: 2015

Year	Jurisdiction	Average	Standard Error
2015	International Average (OECD Countries)	490	(0.4)

NOTE: The PISA mathematics scale: overall mathematics ranges from 0 to 1000. Some apparent differences between estimates may not be statistically significant.
 SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2015 Mathematics, Reading, and Science Assessment

NCBI Resources How To Sign In to NCBI

OPheGenI Phenotype-Genotype Integrator

All Databases Search

Search Summary

Search Criteria

Phenotype Selection
 Trait: Abnormal Fat; Peanut Hypersensitivity
 P-Value: $< 1 \times 10^{-1}$

Genotype Selection - Location
 Chromosome: 13

Modify Search

Search Results

Association Results 1 - 3 of 3 Searched by phenotype trait, SNP chromosome, and P-Value.

Genes 1 - 4 of 4 Searched by gene IDs retrieved from association results.

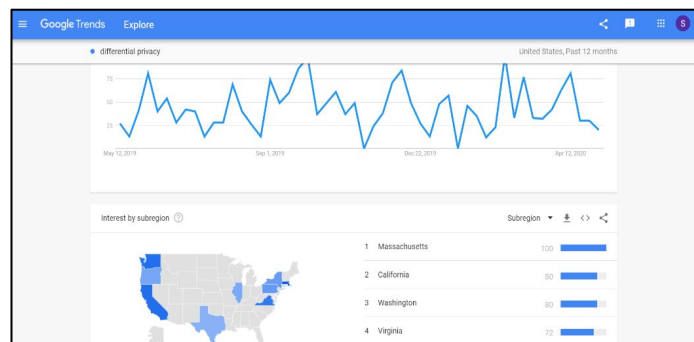
SNPs 1 - 2 of 2 Searched by SNP rs numbers retrieved from association results.

eQTL Data No data found Searched by SNP rs numbers retrieved from association results and P-Value.

dbSNP Studies No data found Searched by traits retrieved from association results.

Genome View 2 SNPs and 4 genes over 1 chromosome.

Modify Search Show All Hide All



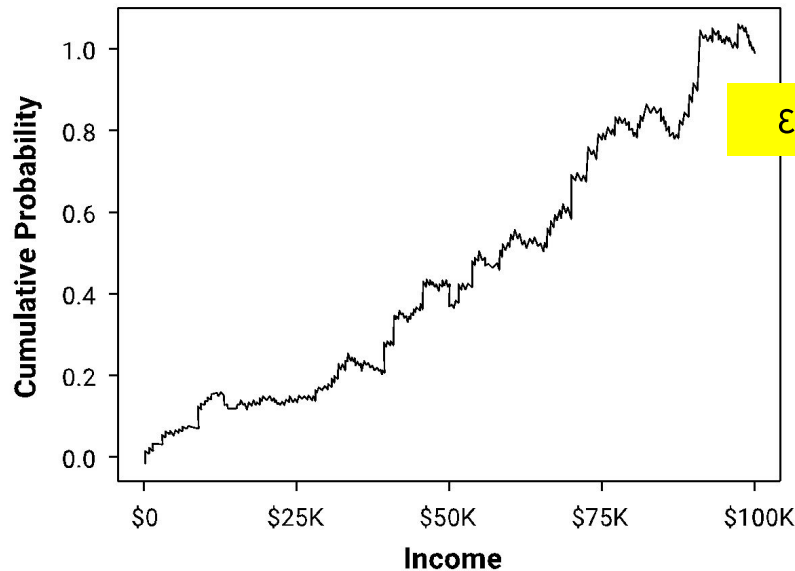
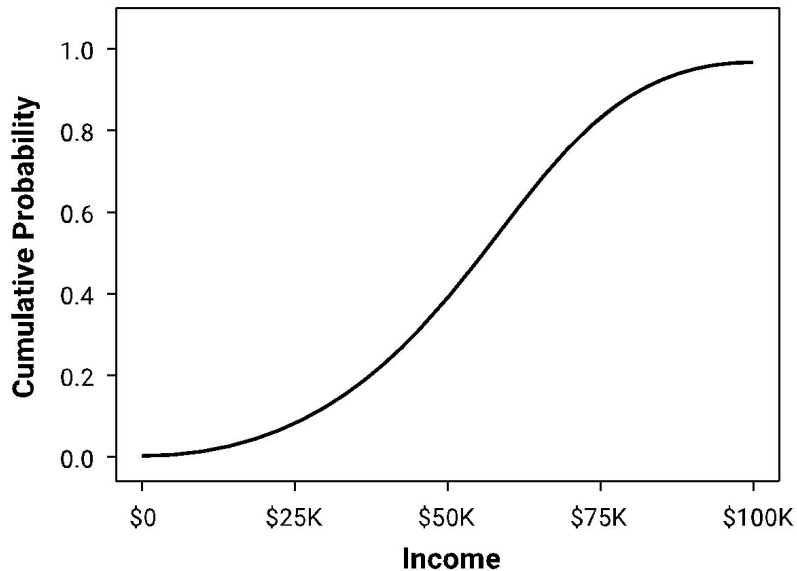
What Can Be Computed with Differential Privacy?

- **Descriptive statistics**: counts, mean, median, histograms, contingency tables, boxplots, CDFs, etc.
- **Supervised and unsupervised ML tasks**: classification, regression, clustering, distribution learning, etc.
- **Generation of synthetic data**

Because of noise addition, differentially private algorithms work best when the number of data records is large.

Differentially Private Computations

Algorithms maintain differential privacy via the introduction of *carefully crafted random noise* into the computation

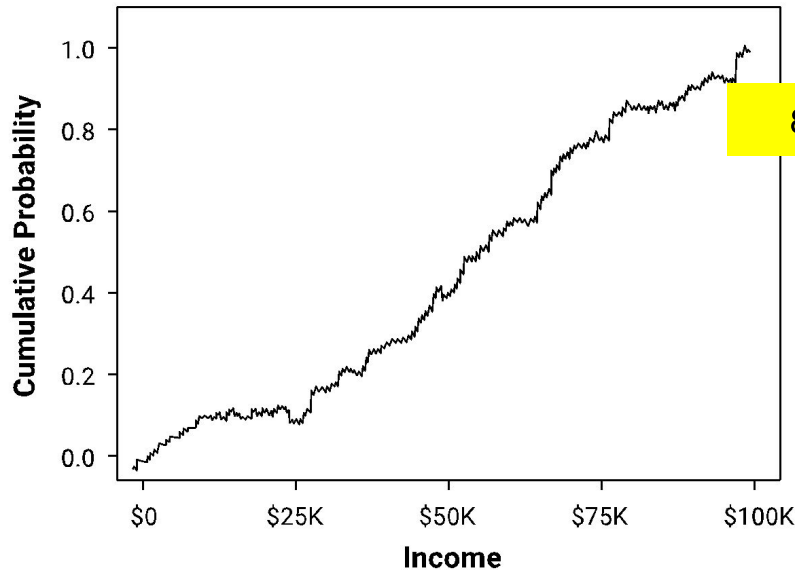
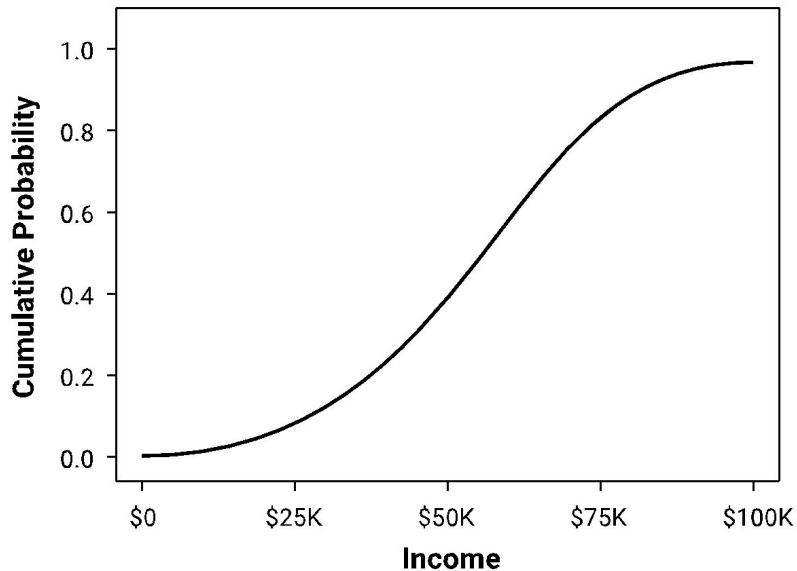


$\epsilon = 0.005$

(These CDFs are stylized examples.)

Differentially Private Computations

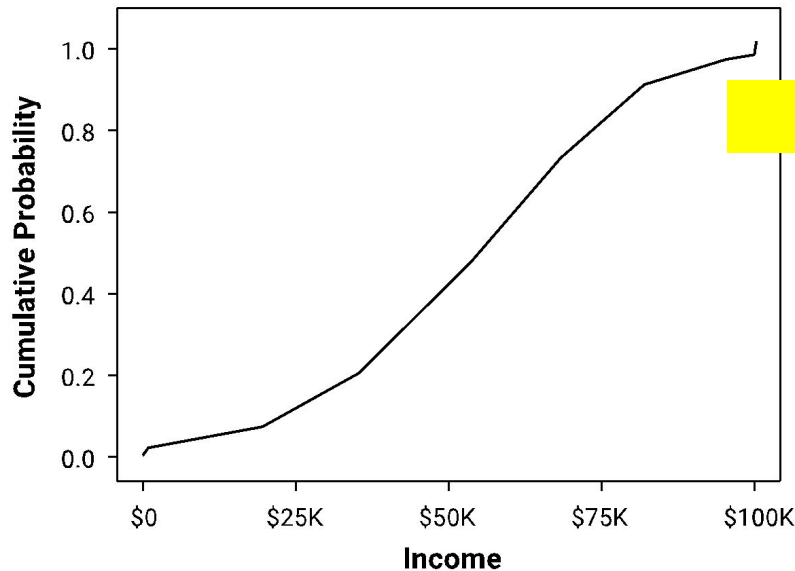
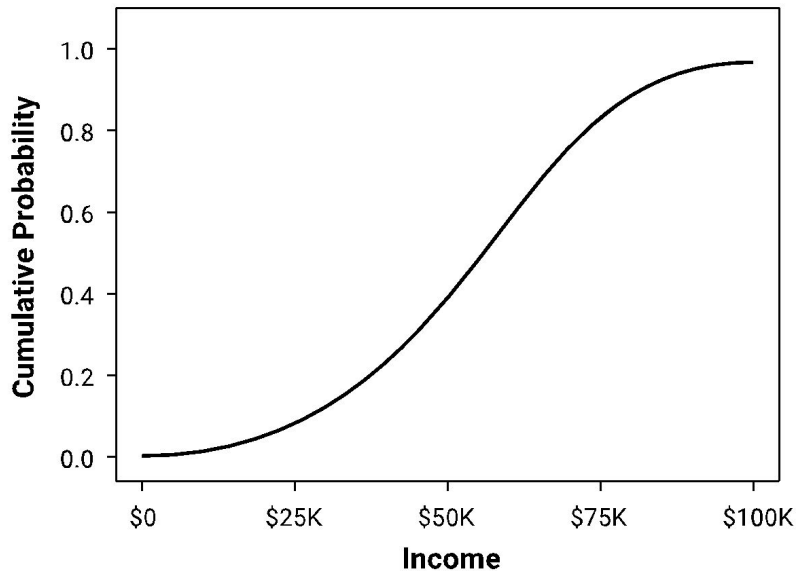
Algorithms maintain differential privacy via the introduction of *carefully crafted random noise* into the computation



(These CDFs are stylized examples.)

Differentially Private Computations

Algorithms maintain differential privacy via the introduction of *carefully crafted random noise* into the computation



(These CDFs are stylized examples.)

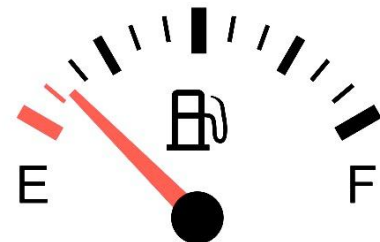
Managing the (Inherent) Privacy-Utility Tradeoff with a Privacy “Budget”

DP provides provable privacy guarantees with respect to the cumulative risk from successive data releases

- Every statistical release incurs some privacy loss ϵ
- More noise \rightarrow more privacy (smaller ϵ), less accuracy (and vice versa)
- Tradeoff is less stark on larger populations ($n \rightarrow \infty$)
- Combination of ϵ -differentially private computations results in differential privacy (with larger ϵ)

This is a key feature, not a bug!

- Consider: ignoring the gauge does not prevent a car from using fuel



Understanding Differential Privacy

- **“Automatic” opt-out**: I am protected (almost) as if my info was not used at all
- **I incur limited risk**: Contributing my real info can increase the probability I will be denied insurance by at most 1%
 - When compared with not participating, or contributing fake info
- These privacy guarantees are provided independent of the methods used by a potential attacker and in the presence of arbitrary auxiliary information
- **Future proof**: Avoids the “penetrate and patch” cycle

DP Has the Benefit of Transparency

It is not necessary to maintain secrecy around a differentially private computation or its parameters

- Benefits of transparency include:
 - **Possibility of accounting for DP in statistical inference**
 - Knowledge accumulation
 - Scrutiny by the scientific community

Application for Public Access to Data

DP can be used to provide broad, public access to data or data summaries in a privacy-preserving way

- Can consider data publications that were otherwise impossible
 - Whereas traditional techniques would (more often) require applying controls in addition to de-identification

Example: Reasoning About Risk

Gertrude's Life Insurance



- Gertrude:
 - Age: 65
 - She has a \$100,000 life insurance policy
 - She is considering participating in a medical study but is concerned it may affect her insurance premium

Example: Reasoning About Risk

Gertrude's Life Insurance



- Based on her age and sex, she has a 1% chance of dying next year. Her life insurance premium is set at $0.01 \times \$100,000 = \$1,000$
- Gertrude is a coffee drinker. If the medical study finds that 65-year-old female coffee drinkers have a 2% chance of dying next year, her premium would be set at \$2,000.
 - This would be her **baseline risk**: Her premium would increase to \$2,000 even if she didn't participate in the study.
- **Can Gertrude's premium increase beyond her baseline risk?**
 - She is worried that the study may reveal more about her, such as that she *specifically* has a 50% chance of dying next year. This can increase her premium from \$2,000 to \$50,000!

Example: Reasoning About Risk

Gertrude's Life Insurance



- Reasoning about Gertrude's risk

- Imagine instead the study is performed using differential privacy with $\epsilon = 0.01$.
- The insurance company's estimate of Gertrude's risk of dying in the next year can increase to at most

$$\approx (1 + 2\epsilon) \cdot 2\% = 2.04\%$$

- Her premium would increase to at most \$2,040. Therefore, Gertrude's risk would be $\leq \$2040 - \$2000 = \$40$

Example: Reasoning About Risk

Gertrude's Life Insurance



- **Generally, calculating one's baseline is very complex (if possible).**
 - In particular, in our example the 2% baseline depends on the potential outcome of the study
 - The baseline may also depend on many other factors Gertrude does not know
- **However, differential privacy provides simultaneous guarantees for every possible baseline value.**
 - The guarantee covers not only changes in Gertrude's life insurance premiums, but also her health insurance and more

Transitioning to Practice: Challenges

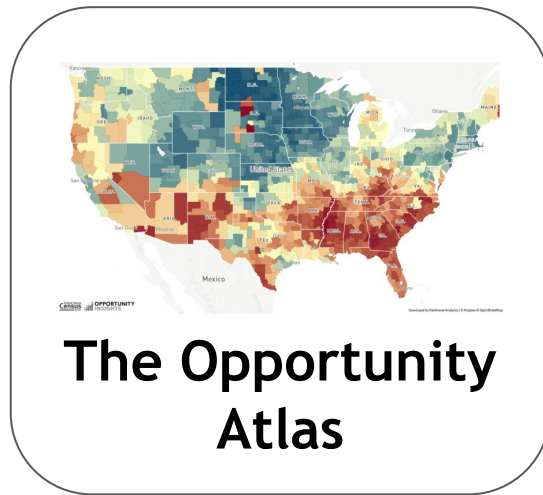
- **DP changes how data is accessed**
 - Noise added, attention on privacy-utility tradeoff
 - May require shift from static to interactive modes of access
- **DP has implications for the data lifecycle**
 - The privacy-loss budget is an unavoidable mathematical fact
 - Setting the budget is a policy question
- **Potential implications for collection, storage, transformation, retention**
- **Implicates legal requirements as well as technical ones**

Transitioning to Practice: Design

In the chapter (print and online version):

- **Aligning risks, controls, and uses**
 - Where is the use of differential privacy appropriate?
 - Selecting privacy controls based on harm and informational risk
 - Combining DP with other tools
(especially as part of a tiered access system)
 - Regulatory and policy compliance implications
- **Detailed case studies**
 - 2020 Decennial Census, Opportunity Atlas, Dataverse Repositories

Case Studies - Overview



Transitioning to Practice: Deployment

In the **online** version:

- Key differential privacy design choices

Review of trust models, privacy-loss settings, privacy granularity, static vs. interactive publication, estimating and communicating uncertainty.

- Data life cycle management

Technical and legal implications of differential privacy design choices for data collection, transformation, access, and retention.

- Additional resources

Open software for differentially private analysis. Enterprise software and consulting. Further readings.

Conclusions

- **Accumulating failures**
 - Anonymization & traditional SDL techniques are not enough
- **Differential privacy**
 - A standard providing a rigorous framework for developing privacy technologies with provable quantifiable guarantees
- **Moving to practice**
 - Best when combined with other technical and policy tools
 - Chapter provides guidance for design & deployment