

~~Regulating Dealing with Data~~

Dealing with Data, Class 8 (part 2)

Slight change of plans

- Last time the “teaser” slide said something about sparsity
 - We started to put that class together, but it was getting too technical...
 - Mary is happy to talk to/at you about sparsity for hours, just ask 😊
- Instead, today we wanted to talk about:
 - How is the **collection** and **dissemination** of data regulated?
 - How do we think about **data privacy**?

Agenda

1. What's at stake?
2. How is the collection / dissemination of data regulated?
 - Regulatory models around the world
3. How do we think about / formalize data privacy?
 - Anonymization
 - Differential privacy

What's at stake?

Lots of entities have data about you

- The government
 - Census, taxes, forms your employer fills out, ...
- Companies you consciously give data to for analysis
 - Fitbit, 23andme, budget planning apps, ...
- Companies that you primarily interact with for other reasons
 - Google, Meta, Amazon, X-formerly-known-as-Twitter, ...

That data can be used for good

- The government can plan social programs better
 - E.g., Census data determines how much money goes to each region
- The data you want analyzed gets analyzed
 - Cool, I'm 13% Lhasa Apso!
- Google/Meta/Amazon can offer you better services
 - It is convenient that Google knows all my passwords so I don't have to remember them...
- When datasets are made public, researchers can use them to learn about cool things.
 - Like your final projects!



It can also be used for not-so-good

- Government surveillance and oppression
 - Targeting folks with certain backgrounds or affiliations
- Companies might sell your data to a third party, or use it for something you didn't want them to.
 - Breach of privacy
 - Why should *they* get paid for *your* data?
- Companies might not protect your data, and a malicious third party can get it.
 - Identity theft, harassment, ...
- When datasets are made public, someone might be able to learn something about you and do any of the above.

Different Legislative Frameworks

Disclaimer: I am not a legal scholar

Agree or disagree

for data that governments collect

- **The government has a duty to protect privacy.** The government should not share *any* personal information with non-government entities without the subject's consent.
- **Data collected using government resources is a public resource.** The government has an obligation to share it with the public and with researchers.
- The government should be able to demand information from private companies under [no/some/any] circumstances.

Agree or disagree

for data that private companies collect

- Since individuals give companies their data voluntarily, **the company owns that data and can do what it wants with it.**
- **Private companies should not share any personal information** outside the company without the subject's consent.
- **Neither of the above matter, since:**
 - Everyone clicks the “allow to collect/share my data” box anyway; and/or
 - Everyone posts all their information publicly anyway.

A few different models

- United States
- European Union
- Turkey
- China



US Law

Disclaimer: I don't know much about this...

- **Government: Privacy Act of 1974:**
 - Federal agencies cannot disclose personal information without consent.
 - Slight exception for the US Census, which can publish aggregated/privatized data
 - The government can subpoena information from private companies, or use an “emergency legal request.”
- **Health Care Providers, Educational Institutions: HIPAA/FERPA**
 - Covered health-care and education providers can't share personal information without consent.
 - Only for providers! Other companies can share your health data if you give it to them.
- **Private companies outside those and a few other industries:**
 - More or less whatever they want*
 - *A few states have more restrictions (e.g., CCPA/CPRA in CA)
 - *Companies that do business outside the US need to comply with those laws...

EU Law

Disclaimer, I know even less about this



- Private Companies:

- General Data Protection Regulation (GDPR), 2016

- Mandatory finer-grained controls
 - like these ones that we now see even in the US/Turkey:
 - Right to erasure
 - Companies must **demonstrate** compliance
 - Large fines for non-compliance

- EU Data Act, 2023:

- You can take your data from one company to another (among other things)

- Governments:

- As far as I can tell, government institutions basically have to follow the GDPR too.

- See: Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data

- Open Data Directive

- Governments must make data public, respecting privacy “according to the principle ‘as open as possible, as closed as necessary’.”

A screenshot of a Cookiebot consent banner. At the top left is a placeholder for 'YOUR LOGO'. At the top right, it says 'Powered by Cookiebot by Usercentrics'. Below this are three tabs: 'Consent' (selected), 'Details', and 'About'. The main text area says 'This website uses cookies' followed by a paragraph: 'We use cookies to personalise content and ads, to provide social media features and to analyse our traffic. We also share information about your use of our site with our social media, advertising and analytics partners who may combine it with other information that you've provided to them or that they've collected from your use of their services.' Below the text are four toggle switches for 'Necessary', 'Preferences', 'Statistics', and 'Marketing'. The 'Necessary' toggle is turned on, while the others are turned off. At the bottom are three buttons: 'Deny', 'Allow Selection', and 'Allow all'.

Turkish Law

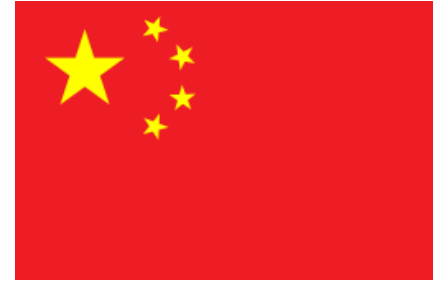
Disclaimer: I **really** don't know a lot about this! Probably lots of you know more!



- Private Companies:
 - Law on Protection of Personal Data (LPPD)
 - Based on EU Directive 95/46/EC, which was the precursor to the GDPR
 - As far as I can tell, weaker than GDPR, stronger than US law
 - Internet Act (Law 5651) was amended in 2021 to impose harsh penalties on companies that don't respect user privacy.

PRC Law

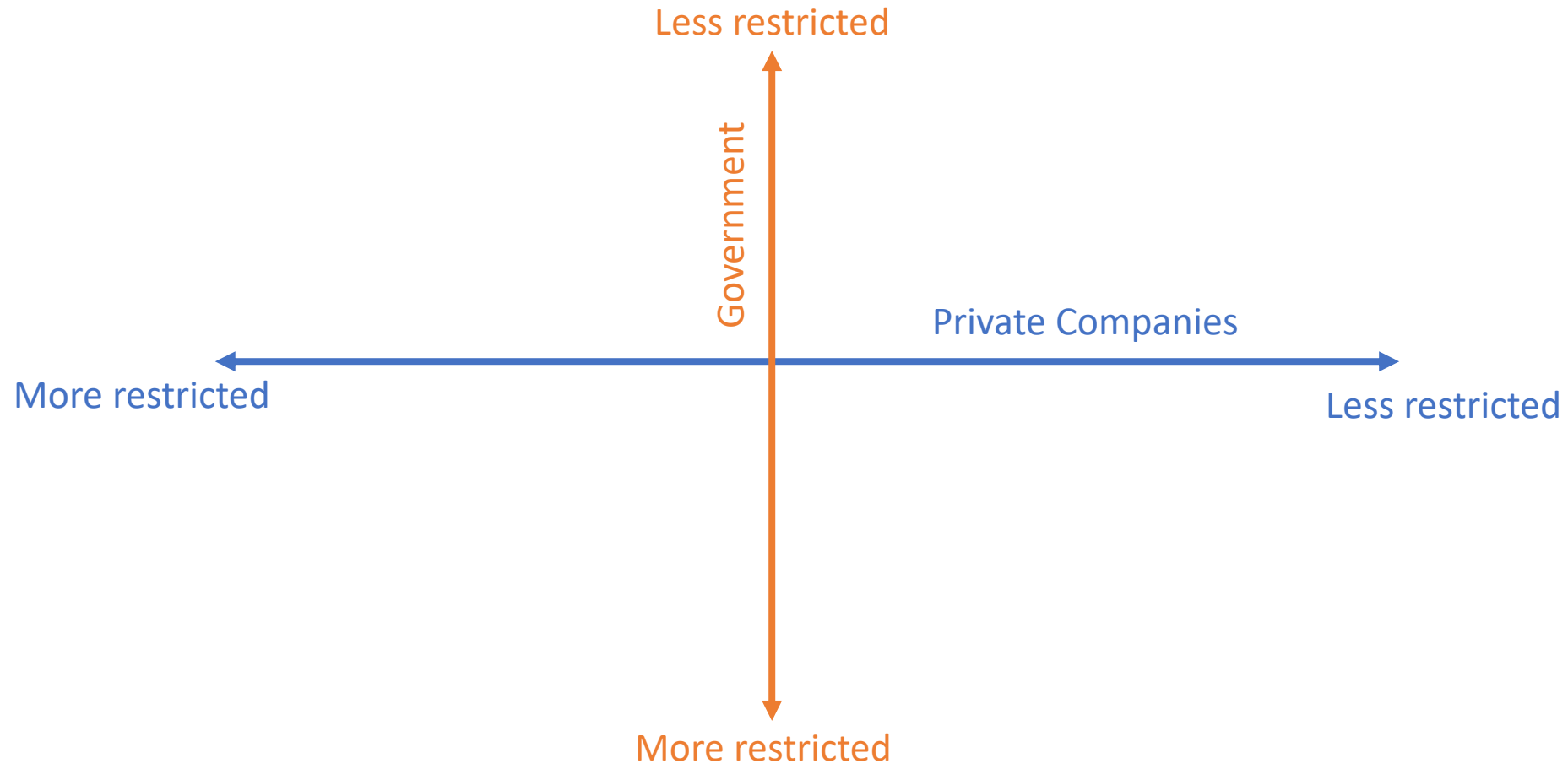
Usual disclaimer!



- Private Companies:
 - Personal Information Protection Law (PIPL), 2021
 - Even stronger than GDPR (larger fines, publicly available list of non-compliers)
 - Cybersecurity Law (CSL) and Data Security Law (DSL)
 - Strong requirements for companies to keep data safe
- Government:
 - National Intelligence Law (2017):
 - *Article 7: All organizations and citizens shall support, assist, and cooperate with national intelligence efforts in accordance with law, and shall protect national intelligence work secrets they are aware of.*
 - Interpreted (at least by the US) as “all Chinese companies shall hand over their data to the government when asked.”
 - CSL Article 28:
 - *Network operators shall provide technical support and assistance to public security organs and national security organs that are safeguarding national security and investigating criminal activities in accordance with the law.*
 - Interpreted (at least by the US) as “Chinese network operators shall hand over data to the government when asked.”

Discuss

- Where do you think is optimal?
 - Can the trade-offs even be plotted this way?


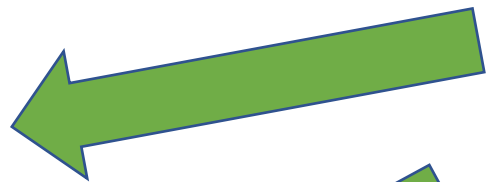

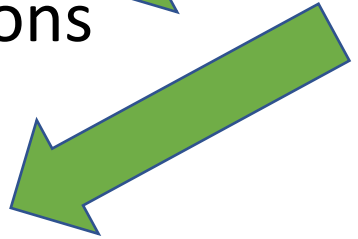


What do we even mean by privacy?

Discuss:

Which of the following seem reasonable?

- US census data should be:

- Published, but “anonymized”**Really NOT private!**
 - Replace names with arbitrary ID numbers
- Published, but only at a high level of granularity
 - E.g., “The average income in Nebraska is...”
- Accessible, but only give “noisy” answers to any questions
 - Literally, add some random noise to any answer you release
- Accessible, but only by people with security clearances
 - They have to log into government servers, have every keystroke logged, and will go to jail if they mis-use it.

The US
Census
does all of
these
things

“Anonymization” is not private

- In the late 1990's, the state of MA released anonymized data on state employee hospital visits.
 - “Anonymized:” Identifying info like name, SSN, address were removed
- Latanya Sweeney (then a grad student at MIT) showed that it could be “de-anonymized,” and identified the health records of the governor of MA at the time.
 - She cross-referenced the health data with a \$20 dataset of voter name, sex, address, date of birth, and easily found the governor.
- She later showed that 87% of Americans are uniquely identifiable based on ZIP code, birthdate, and sex.



Prof. Latanya Sweeney
Harvard University

“Anonymization” is not private

- Another example: the “Netflix Prize”
- Netflix released a bunch of “anonymized” data about movie ratings
 - `<user, movie, date of rating, rating>`
- Researchers at UT Austin cross-referenced with IMBD, and identified individual users.

Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

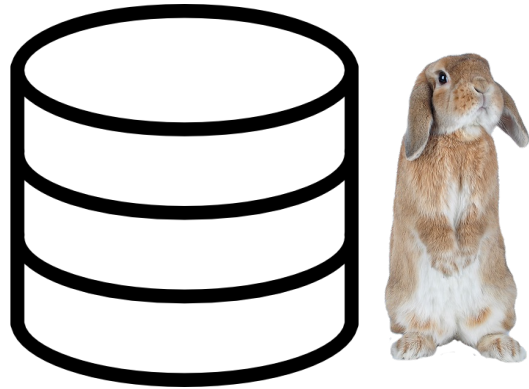
Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary’s background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world’s largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber’s record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Differential Privacy

- **Idea:** whether or not your data is in the dataset shouldn't affect the outcome of a query.

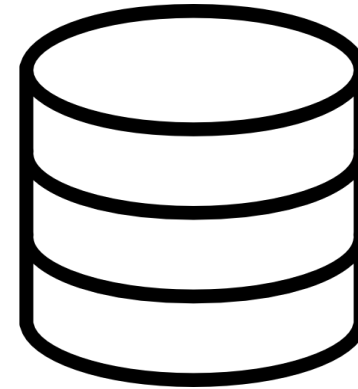


World 1: Rabbit
is in the database

What is the
average income
of everyone in
the dataset
under the age
of 50 and with a
Ph.D.?



Researcher



World 2: Rabbit isn't
in the database

The researcher shouldn't be able to tell the
difference between the answer they get in World
1 and the answer they get in World 2.

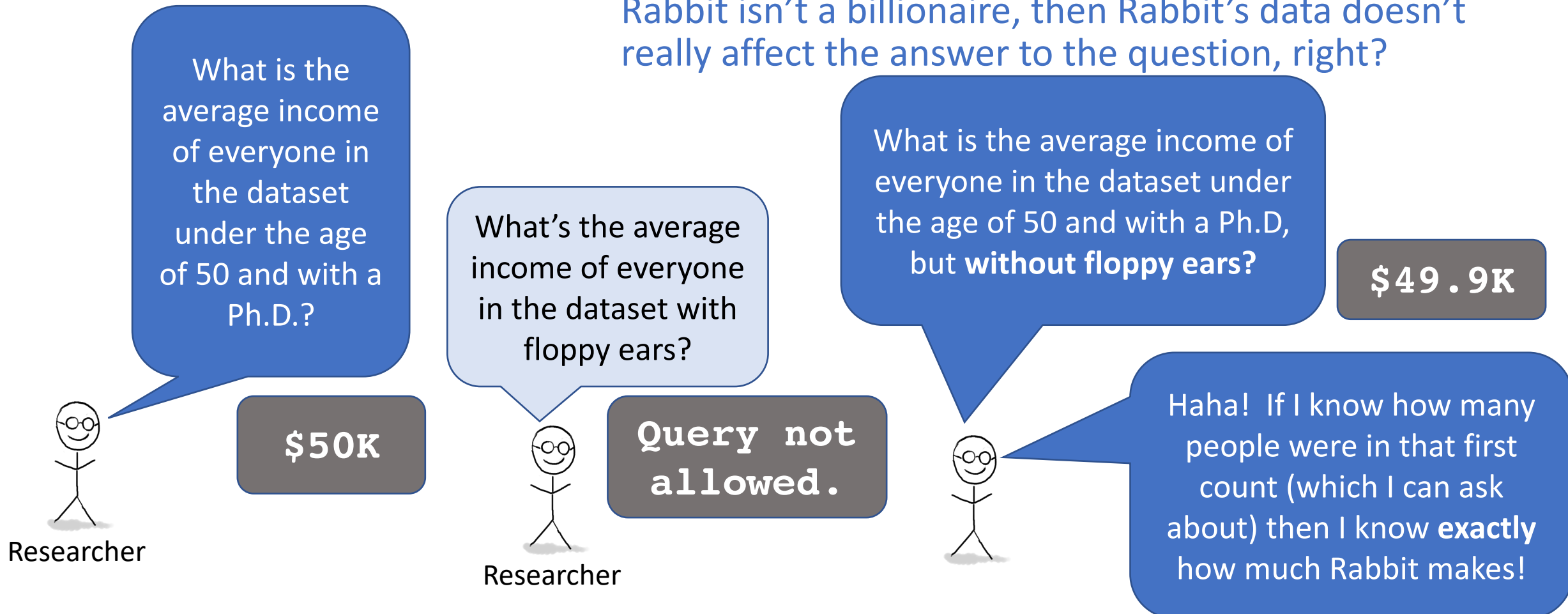
Note: **Dr. Rabbit** is indeed under 50 and has a Ph.D. They are also the only person in the dataset with floppy ears.



Aside:

Why doesn't aggregating over big enough sets solve this?

- If there are enough people in each query, and Rabbit isn't a billionaire, then Rabbit's data doesn't really affect the answer to the question, right?



Solution: Add some noise

Calibrating Noise to Sensitivity in Private Data Analysis

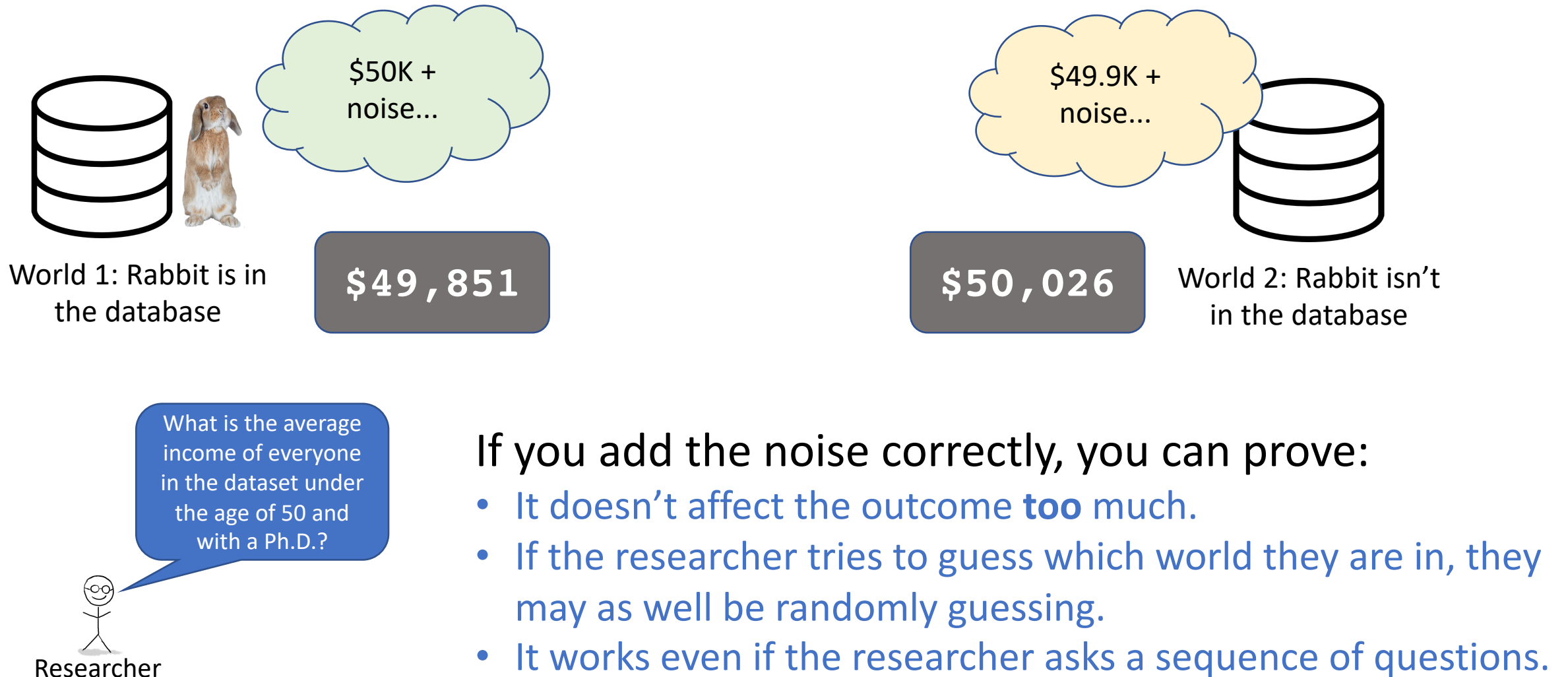
Cynthia Dwork¹, Frank McSherry¹, Kobbi Nissim², and Adam Smith^{3,*}

¹ Microsoft Research, Silicon Valley
`{dwork, mcsherry}@microsoft.com`

² Ben-Gurion University
`kobbi@cs.bgu.ac.il`

³ Weizmann Institute of Science
`adam.smith@weizmann.ac.il`

Solution: Add some noise



Differential privacy



- Introduced in 2006 by Dwork, McSherry, Nissim, and Smith
- Has since been adopted by:
 - Apple
 - Microsoft
 - Google
 - The US Census Bureau
- Tons of interesting research questions
 - What the best trade-off between error and privacy?
 - If you want to use data to train a model, can adding noise prevent that model from “leaking” information?
 - ...

Is this the “right” definition of privacy?

- Some people think so.
- Others...don't.

Recap

When Dealing with Data...

- We also need to think about how we (or governments or companies) collect, process, and disseminate that data.
- These are really non-trivial questions!
 - Ethically, legally, and technically

That's it!

- Keep working on your projects, and see you next time!