Privacy Engineering

Jodi Spacek, Senior Privacy Engineer, Google
(opinions are my own)

Today's Talk

1. Privacy Engineering and Privacy Policies
2. **Privacy Pause:** Privacy at UBC
3. Privacy in Systems
4. **Privacy Pause:** Kafka & GDPR Privacy
5. **Group Exercise:** Accelerometer Hacking
6. De-identification
7. K-Anonymity
8. Differential Privacy
9. **Group Walk-through:** Anonymizing Restaurant Visits
10. Differential Privacy Models
11. Q&A

Pathways to Privacy

- B. Music -> UBC CS (second degree program) -> MSc
- SWE -> GDPR -> Differential Privacy
- Hootsuite -> Shopify -> Google
- What working in privacy is like at Google
  - Big scale, very interesting work
  - Horizontal and vertical privacy
  - Privacy products shared externally as open source solutions (more on this later)
  - Research opportunities

Privacy Engineering

- Flavours of Privacy Engineering: https://twitter.com/leakissner/status/1389366778346672129

Lea Kissner
@LeaKissner

More and more folks want to hire privacy engineers. This is great! You almost certainly need them! But, just like security, privacy engineering is a whole field.

So for the folks who want to hire or become a privacy engineer, a rundown of the current rough types I see. (Big🧵)

7:50 PM · May 3, 2021 · Twitter Web App

Privacy Engineering

Flavours of Privacy Engineering

1. Analysis/Consulting
2. Privacy Products
3. Math & Theory (Anonymization)
4. Infrastructure
5. Tooling & Dashboards (Auditing)
6. User Experience (UX)
7. Privacy Policy
8. Privacy Process (Programs)
9. Privacy Incident Response

Privacy Policies

- GDPR (General Data Protection Regulation)
  - Why are there many similar policies (eg. CCPA, PIPEDA)?
  - Why is GDPR so widely used?
- AADC (Age Appropriate Design Control)
  - AADC @ Google
  - Irish Data Protection Commission (DPC)

Privacy Policies

Have you heard about UBC's Privacy Policy?

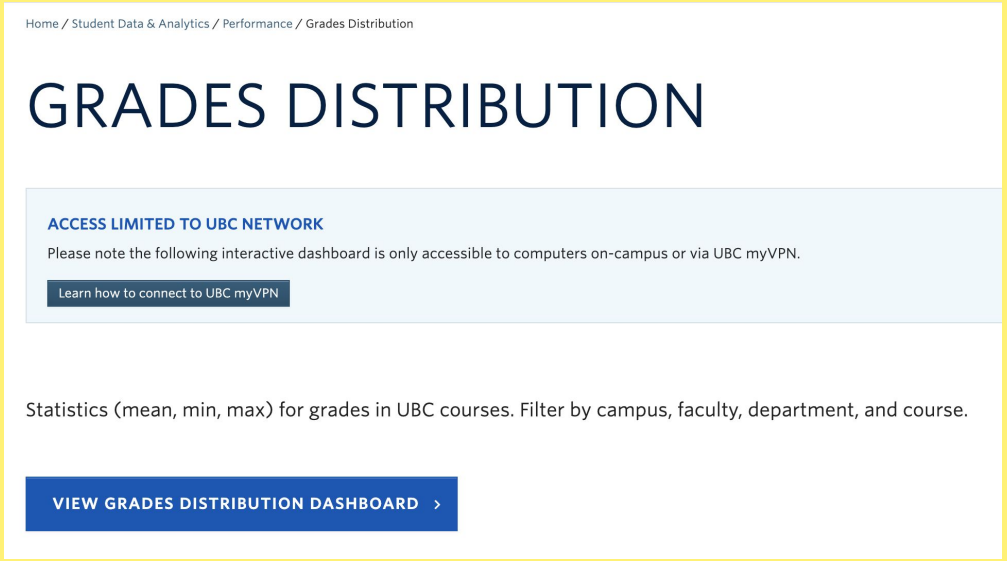
- FIPPA Freedom of Information and Protection of Privacy Act

Privacy Pause: How are grades shared at UBC?

Things to consider:

- What are the guidelines set out for grades sharing?
- How long is the data stored?
- How many times have you shared your student ID?
- How would you request to remove your data?

Privacy Pause: How are grades shared at UBC?



Is this data anonymous? Why or why not?

Privacy Protection

So far, we've talked about data that we can visualize through a UI

Let's switch the focus to privacy protection for data that we don't see (eg. non user-facing)

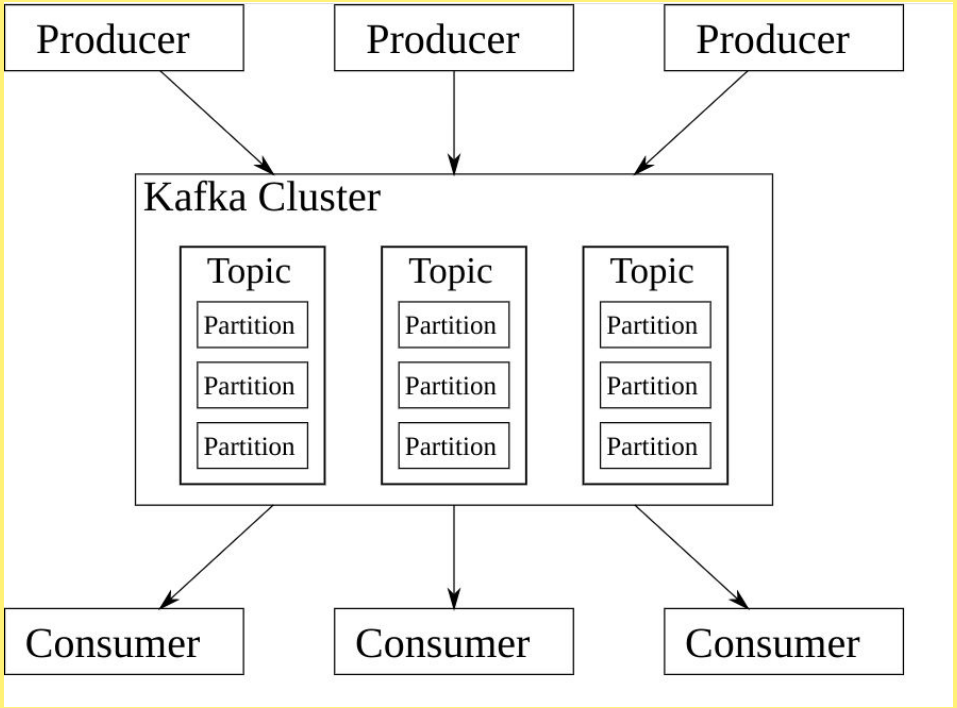
- https://safety.google/privacy/privacy-controls/

- What type of data is protected that we can't see?

Privacy Protection

- What type of data is protected that we can’t see?
  - Logs
  - Report data
  - Data needed for audits (eg. financial)
  - Data used to train ML models
  - And many more...
- Where could this data live?
  - Internally (within the system itself only)
  - Externally (send outside the system)
    - Could be made publicly available
    - Could be sent to 3rd parties

Privacy Pause: Right to be Forgotten in Kafka Pub/Sub



Source: https://en.wikipedia.org/wiki/Apache_Kafka

Privacy Pause: Kafka Pub/Sub

What is the retention period in Kafka?

What is the role of compaction?

How frequently does compaction happen?

What triggers compaction?

1
2
3
4
5
6
7
8
9
10
11

Privacy Pause: Kafka Pub/Sub

What is the retention period in Kafka?

Typically, the retention period is 3-7 days. The data is ephemeral.

What is the role of compaction?

Event logs can become very large and need to be compacted, eg. updates are squashed.

How frequently does compaction happen?

It depends on the **size** of the log.

What triggers compaction?

Activity, updates on a record which makes the log grow.

Privacy Pause: Kafka Pub/Sub

Issue: https://issues.apache.org/jira/browse/KAFKA-7321

Resolution: KIP-354: Add a Maximum Log Compaction Lag

(Time-based vs record-based compaction.)

- The system was not design with privacy in mind
- The fix applied is an example of ad-hoc privacy

Group Exercise: Hacking Accelerometer Data

Let’s pretend to be privacy hackers and see what we can learn from accelerometer data.

- 5 minutes break into groups to see what we can figure out
- Meet back here after and discuss findings

1
2
3
4
5
6
7
8
9
10
11

Group Exercise: Accelerometer Data

Suppose we receive four pieces of data from an accelerometer.

We don't know when we'll receive the data, but we are continuously listening.

The format of the data is **acceleration**: meter per second squared (m/s^2).

2 m/s^2

10 m/s^2

1000 m/s^2

0 m/s^2

Hint: What is another type of implicit data that I get for free from observing**?**

1
2
3
4
5
6
7
8
9
10
11

Group Exercise: Accelerometer Data

We are not using any fancy techniques, but we have unlimited storage and unlimited resources to observe.

Things to consider:

How can I link work schedule with this data?

How can I link location with this data?

How can I link age with this data?

Group Exercise: Accelerometer Data

How can I link work schedule with this data?

With patience. I will store the data looking for patterns of movement and patterns of sitting.

How can I link location with this data?

Now that I know your schedule, I can figure out your timezone by your wakefulness.
Unless you happen to suffer from insomnia!
But, a restless night here and there won't matter much.

How can I link age with this data?

Boy, kids sure are rowdy! And they love to jump on trampolines!
Unless you are an olympic athlete, I can be pretty sure you are a child.

Group Exercise: Accelerometer Data

What if we have access to fancy ML techniques? And a few more data types?

A few more data types can link a lot, especially with ML techniques:

https://github.com/szymonbcoding/Human-movement-classification-with-accelerometer

https://github.com/akshath123/Finding-out-Accident-Occurrence-using-Accelerator-in-mobile-phones-

https://github.com/sarathsp06/gesture_recognizer

https://github.com/thanghoang/GaitAuth

https://github.com/S3L1M/Nunchuk

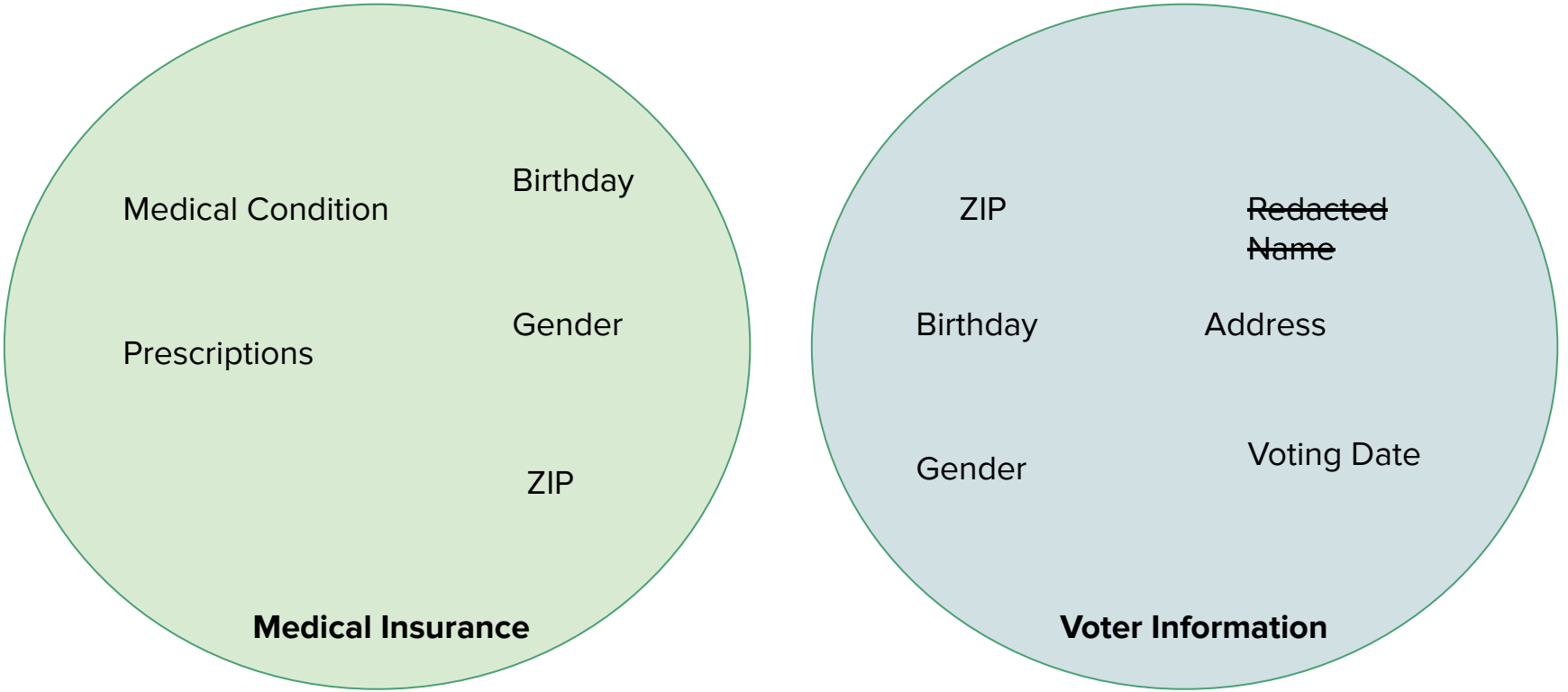
https://github.com/shanujshekhar/Detect_Heavy_Drinking_Episodes

1
2
3
4
5
6
7
8
9
10
11

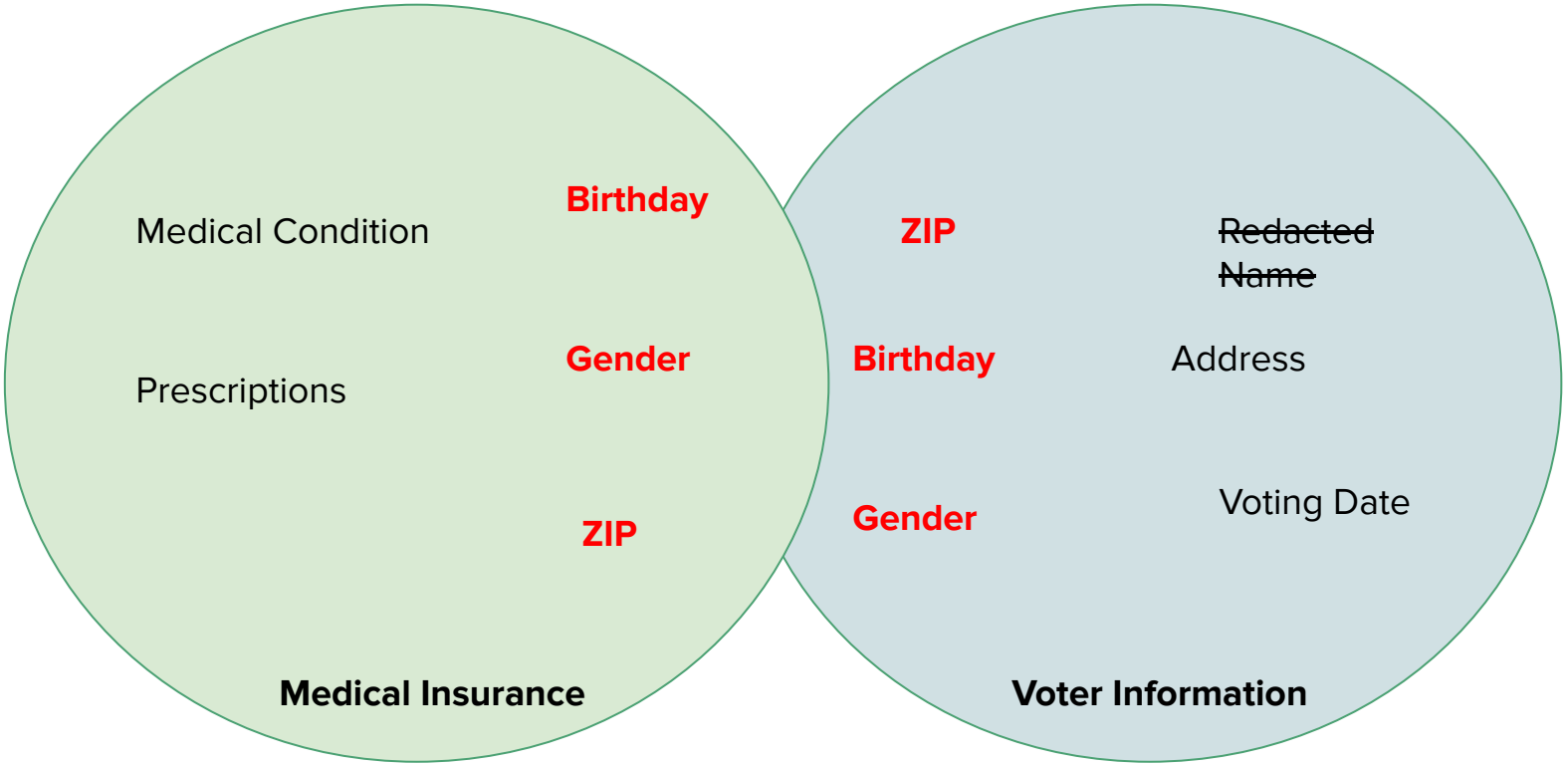
De-identification

- We saw in our group exercise that de-identification isn't enough
- We can still tell plenty about our users without knowing PII (Personally Identifiable Information)
  - We can hash names, remove sensitive IDs
  - How much is enough?
- Latanya Sweeney
  - showed that 3 pieces of PII (ZIP, Gender, Date of Birth) can uniquely identify ~87% of US citizens
  - Simple Demographics Often Identify People Uniquely (2000)
  - How did she show this?

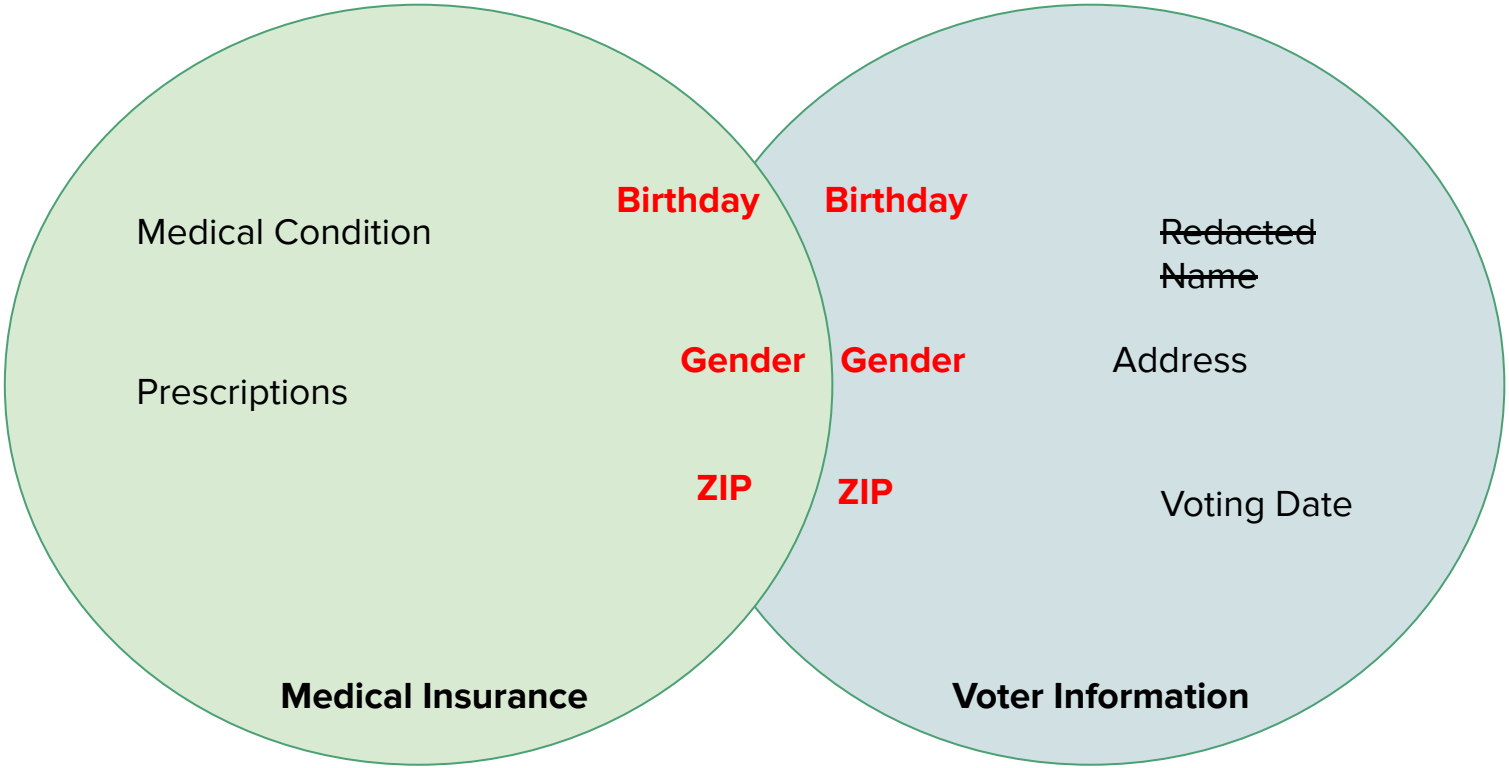
De-identification



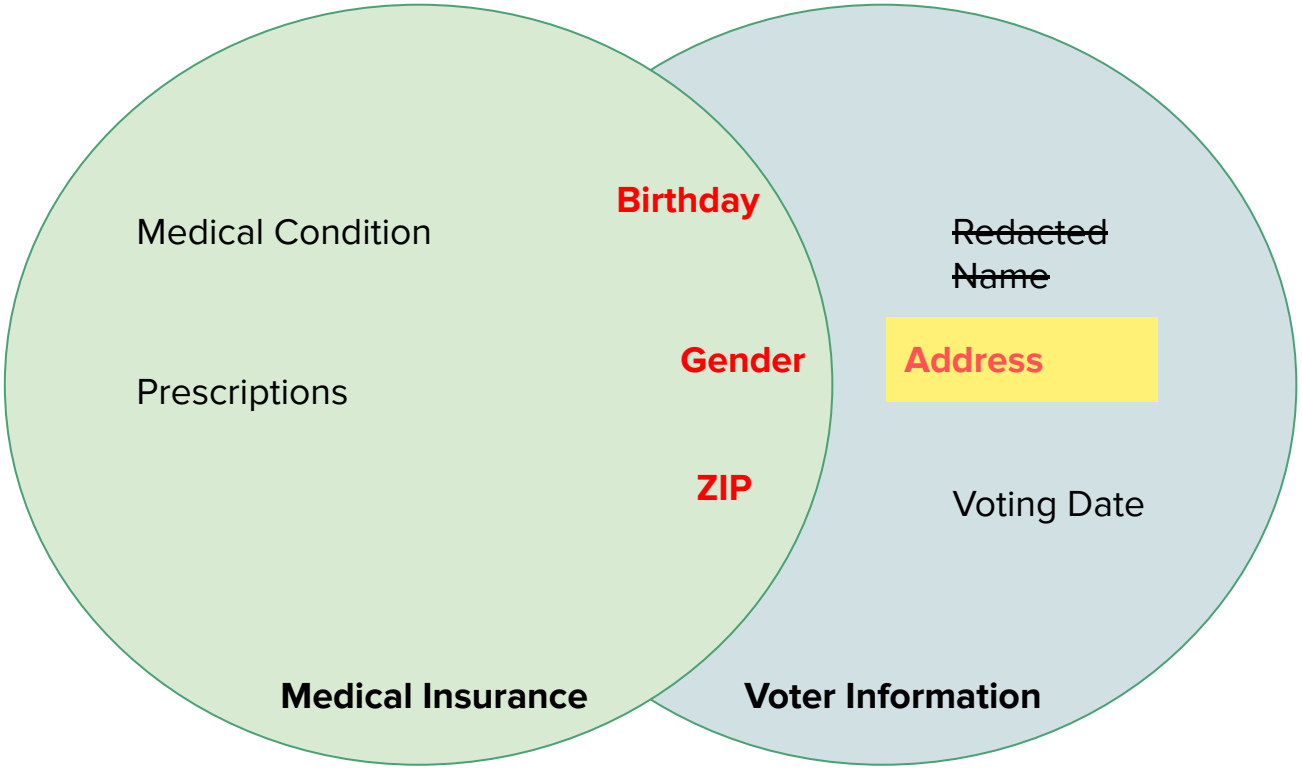
De-identification



De-identification



De-identification



1
2
3
4
5
6
7
8
9
10
11

De-identification

- We don’t have the individual’s name, but we do have their address, medical conditions, voting information.
  - Is it very difficult to get the individual’s name now?
- Led to advances in HIPAA:
  - Only You, Your Doctor, and Many Others May Know

Other examples of de-identification failures

- https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1
- https://en.wikipedia.org/wiki/Netflix_Prize#Privacy_concerns

Led to a new technique for **pseudonymization**: K-Anonymity*

*despite the name, k-anonymity does not produce truly anonymous data

K-Anonymity

- Think in aggregates
- Roll up categories
- Hide unique individual’s data in a crowd

Where’s Waldo? Illustration by Martin Handford, published by Candlewick Press.

1
2
3
4
5
6
7
8
9
10
11

K-Anonymity

K: the number of unique users in a category

- The higher the K, the more pseudonymous
- How do we pick a K value?

Let’s consider an example:

K-Anonymity: Cafe preferences

- We want to show cafe ads based on user preferences
- We know about the cafes that users went to in Vancouver
- How can we use this data but protect our users' privacy?
- How can we *quantify* the amount of protection?
- After this, how do we know we have the right amount of privacy?

K-Anonymity: Cafe preferences

Cafe Location	Roast	Name	Age
South Granville	Dark	Alice	19
UBC	Atomic	Bob	22
Kerrisdale	Medium	Jane	25
Cambie Village	Mild	Mark	31

This is obviously “too” identifying.

We could easily fully identify these people using additional information that we know (a linkage attack).

1
2
3
4
5
6
7
8
9
10
11

K-Anonymity: Cafe preferences

Cafe Location	Roast	Age
South Granville	Dark	19
UBC	Atomic	22
Kerrisdale	Medium	25
Cambie Village	Mild	31

Is this slightly better? Is it *for sure* anonymous?

Can you think of any other information that could be used in combination with this dataset?

Should we keep generalizing/rolling up the data?

1
2
3
4
5
6
7
8
9
10
11

K-Anonymity: Cafe preferences

Cafe Location	Roast	Age
South West Granville	Dark - Atomic	19-24
South East Granville	Mild - Medium	25-30

How about now? Keep rolling up?

1
2
3
4
5
6
7
8
9
10
11

K-Anonymity: Cafe preferences

Cafe Location	Roast	Age
Vancouver	Mild to Atomic	19-30

This *hand waves * seems safe! Let's send out our ad with this information.

Hello human people aged **19-30**! Try out some **coffee** of some type of **roast** of the fine bean that is coffee in this fair city of **Vancouver** for a 15% discount!
Maybe you can have a **student** discount, I have no idea!
Coffee!

K-Anonymity

- Wait, this doesn’t seem particularly useful!
- How can I balance the amount of data that is rolled up with the amount of usefulness in the ad?
- Finding the right amount of K is NP-hard (see https://desfontain.es/privacy/k-anonymity.html)

1
2
3
4
5
6
7
8
9
10
11

Differential Privacy

- Let’s shift our thinking from hiding certain attributes of users by generalizing them or de-identifying these attributes

What if we assumed the strongest type of attacker; an attacker who had all types of datasets containing all attributes?

What if we assume all attributes of users are eventually identifying?

How could we protect the dataset?

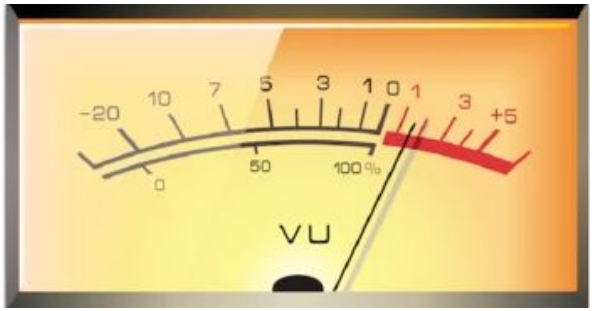
- Should we try a quiet and subtle mechanism of privacy?
- How about an obscure method that no one could guess?

Differential Privacy

How could we protect the dataset?

The solution is through NOISE!!

Let’s examine the kind of noise we mean here.



https://www.emojipng.com/preview/11214555

Differential Privacy

Consider a simple example with a divisive question:

Do you like raisins in cookies?! +  

- Suppose you don't want me to know whether you like them or not.
- I have access to 2 datasets with the total number of likes and dislikes; I know that you are in one dataset and not the other.
- Dataset 1: **56/99** love raisins in cookies
- Dataset 2: **57/100** love raisins in cookies
- I think you DO love raisins and I think you answered Yes in dataset 2.

Can you **plausibly deny** this?

Differential Privacy

Do you like raisins in cookies?!  

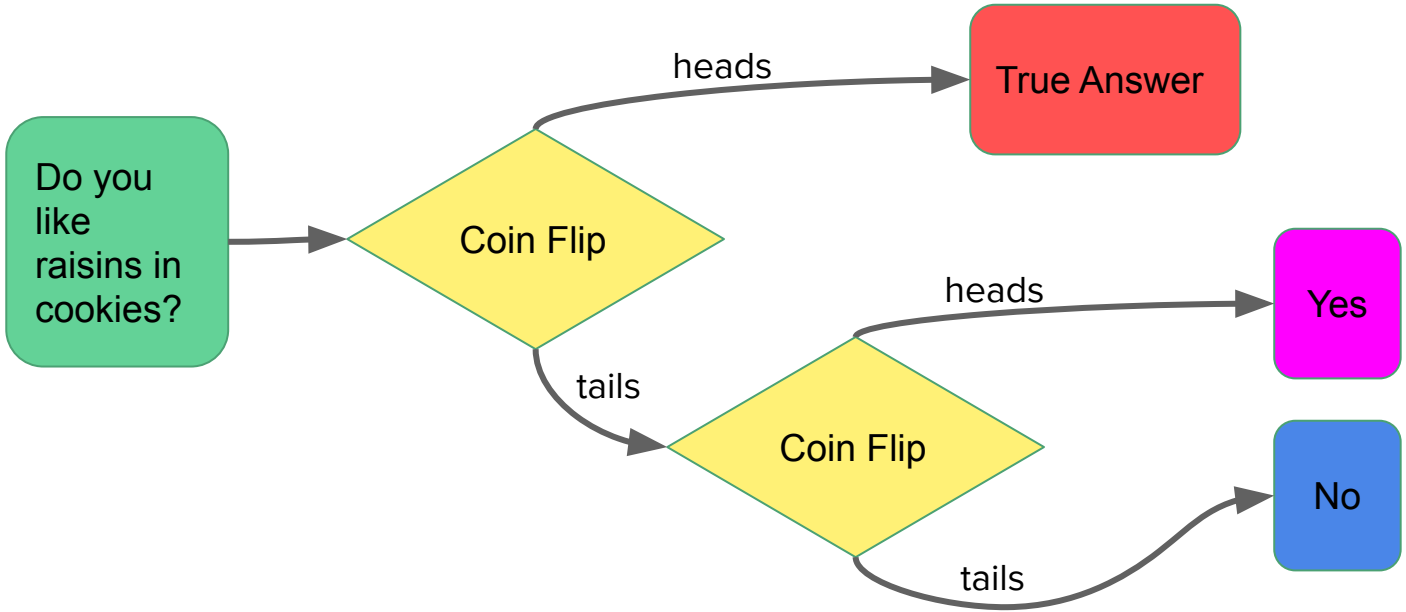
Noise to the rescue!

- Use a randomized response to introduce noise
- 25% of the answers are noise, randomly answer Yes or No
- Coin flip = 50% chance of a Yes answer
- There is a chance my answer is noise, so I can **plausibly deny** my answer

1
2
3
4
5
6
7
8
9
10
11

Differential Privacy

Do you like raisins in cookies?! 🍪 + 🍇



Differential Privacy

Do you like raisins in cookies?!  + 

What does plausibly deny mean?

- We can use probabilities to quantify what we mean
- If we use a coin flip instead of the real response less than some % of the time, we have less plausible deniability
- What if we use the coin flip instead of the real answer 100% of the time?

1
2
3
4
5
6
7
8
9
10
11

Differential Privacy

Do you like raisins in cookies?!  + 

What is does not protect:

- I still know that you participated in the online questionnaire regarding cookie preferences
- What is protected is the conclusion of the results,
  eg. the total number of people who love raisins in cookies
- What is “enough”?

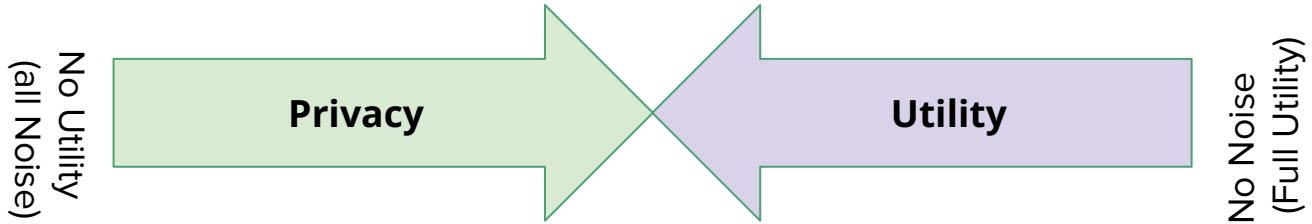
Differential Privacy

- The amount of privacy, **epsilon**, described as ε-differentially private.
- Epsilon also describes the shape of the noise that we use.
- It defines the amount of privacy loss → 0-differentially private is pure noise and perfectly private!
- But 100% noise, completely random data is useless for analysis
    - We say it has no **utility**.

Source: Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014): 211-407.

Differential Privacy

- Best for static datasets shared publicly, eg. medical, census data.
- Trade-off between privacy and utility.
- A privacy budget spreads out the amount of epsilon in a system
  - A limit on requerying because this leads to privacy loss.



Differential Privacy

How private is private?

- Concept of neighbouring datasets
- Two datasets D_1 and D_2: if they differ only in the presence of one user Bob, can we tell that the single user is in the D_1 and D_2?
  - If yes, this is not differentially private
- The number of times we see the aggregate value + noise is the amount our confidence increases about Bob being in the dataset
- This is why we limit fresh querying through **budgeting**

Differential Privacy

How private is private?

- Classic formula: we measure the privacy loss with the epsilon value

$$P(M(D) \in S) \leq e^{\epsilon} P(M(D') \in S) + \delta \ .$$

Group Walk-through: Anonymizing Restaurant Visits

PipelineDP Codelab (you can follow along in the jupyter notebook)

PipelineDP is a library developed with Googlers and OpenMined.

It has a pluggable backend for Spark and Beam (could be extended to others).

This is useful for batch processes,
eg. we want a DP sum for a total number of X in some Y batch.

Group Exercise: Anonymizing Restaurant Visits

Collect user restaurant visits, time there, money spent, day of visit

QQ: Is it OK to release data like this?

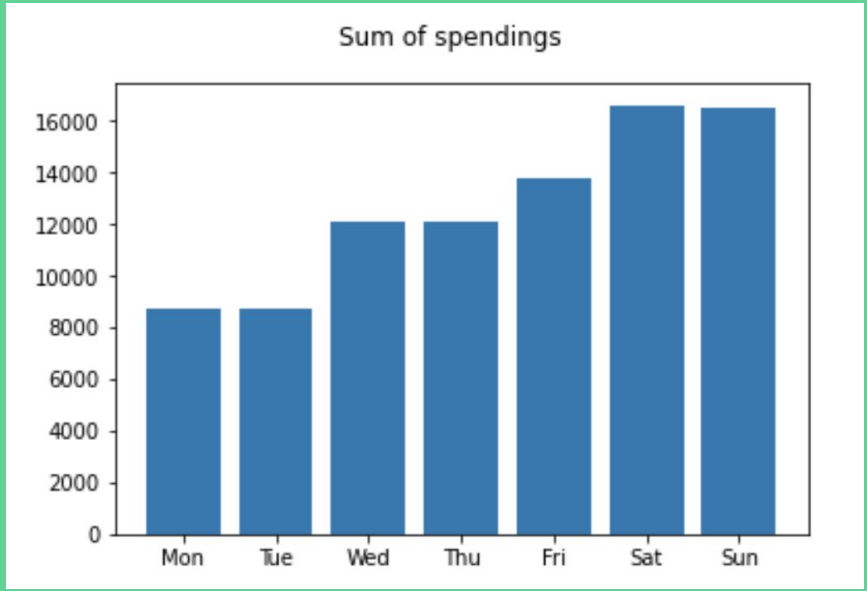
Raw Data:

	user_id	enter_time	spent_minutes	spent_money	day
0	580	9:27AM	29	17	1
1	1215	9:16AM	45	18	1
2	448	11:55AM	12	16	1
3	125	10:47AM	27	20	1
4	484	11:08AM	35	13	1

1
2
3
4
5
6
7
8
9
10
11

Group Exercise: Anonymizing Restaurant Visits

QQ: Is it OK to release aggregate data like this?



Group Exercise: Anonymizing Restaurant Visits

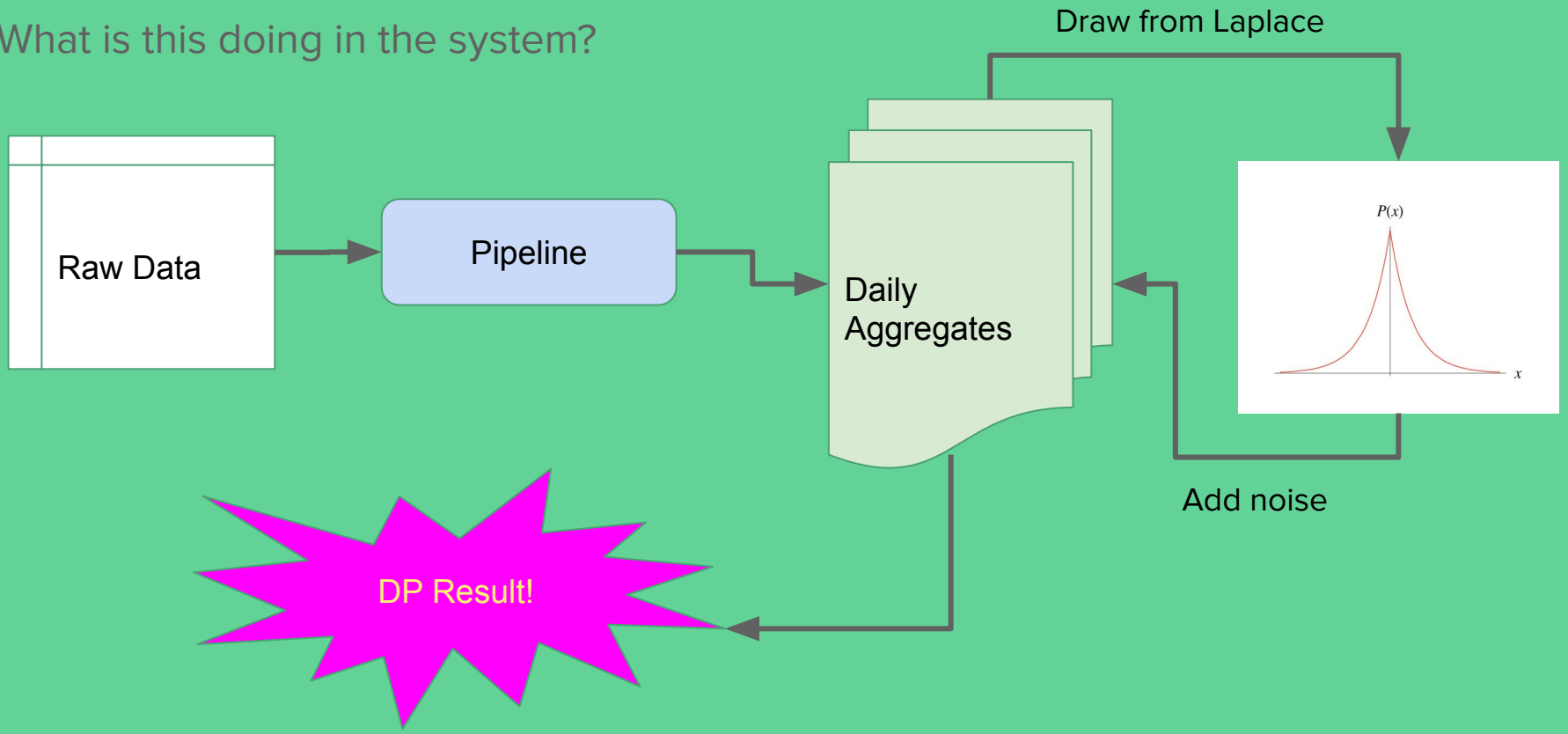
Budget accountant keeps track of the privacy parameters we set in the batch process aggregates:

```
budget_accountant = pipeline_dp.NaiveBudgetAccountant(total_epsilon=1, total_delta=1e-6)
```

```
params = pipeline_dp.AggregateParams(noise_kind = pipeline_dp.NoiseKind.LAPLACE,
                                     metrics=[pipeline_dp.Metrics.COUNT, pipeline_dp.Metrics.SUM],
                                     max_partitions_contributed=7,
                                     max_contributions_per_partition=2,
                                     min_value=0,
                                     max_value=100,
                                     public_partitions=list(range(1, 8)))
```

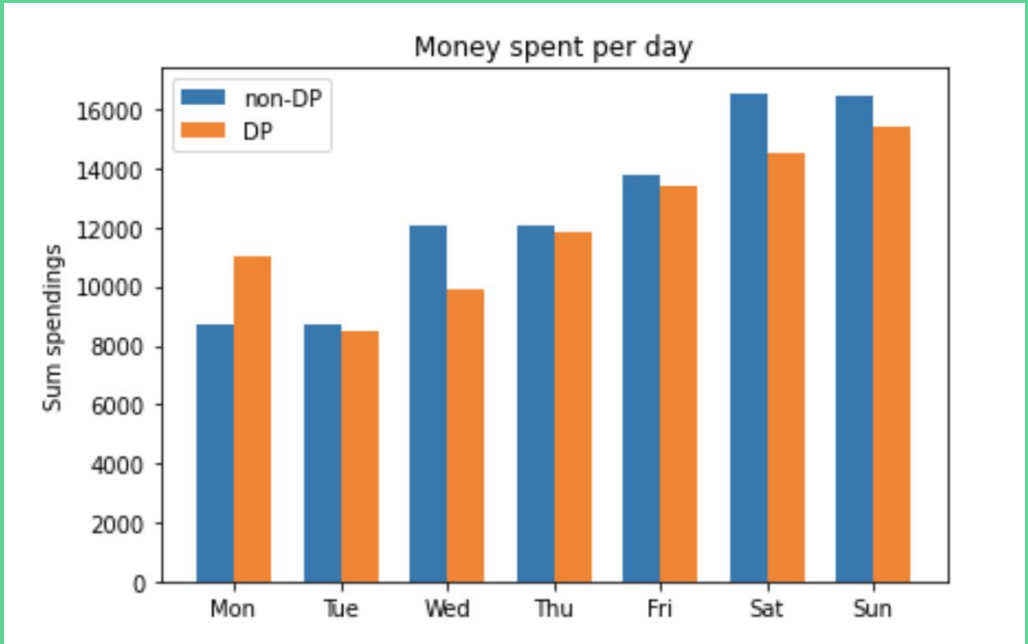
Group Exercise: Anonymizing Restaurant Visits

What is this doing in the system?



Group Exercise: Anonymizing Restaurant Visits

Compare True results and DP results



Group Exercise: Anonymizing Restaurant Visits

Public Partitions = categories that are known, eg. Days of the week

Can we think of an example dataset that has unknown partitions?

Why is this important?

Group Exercise: Anonymizing Restaurant Visits

Public Partitions = categories that are known, eg. Days of the week

Can we think of an example dataset that has unknown partitions?

A dataset with aggregates per country, but some countries are missing.

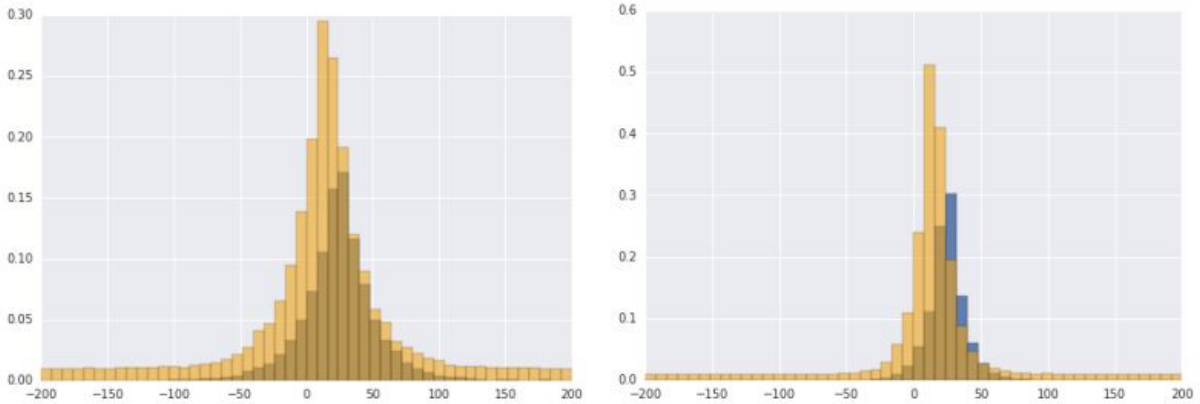
Why is this important?

Suppose some of the countries appear with 1 person, but disappear when that person is deleted.

This will expose that one user!

Differential Privacy: Testing and Attacks

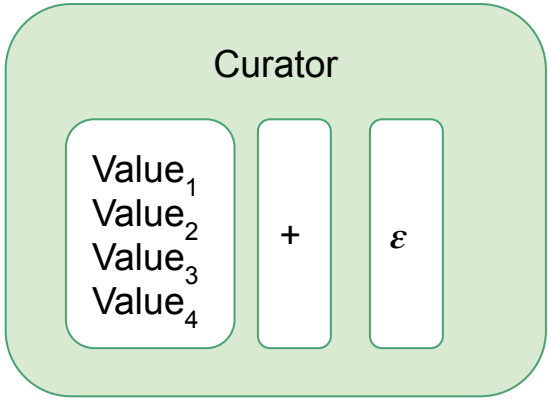
- Differentially Private SQL with Bounded User Contribution
  - This paper describes how to limit row-owner contribution, defines *sensitivity*
  - Contributes a testing methodology based on probability distributions



(a) Passing test **(b)** Failing test

Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Central: All values are maintained and noised by a central coordinator
  - Fine for small datasets, but rapidly does not scale.
  - How big are some of the datasets at Google?

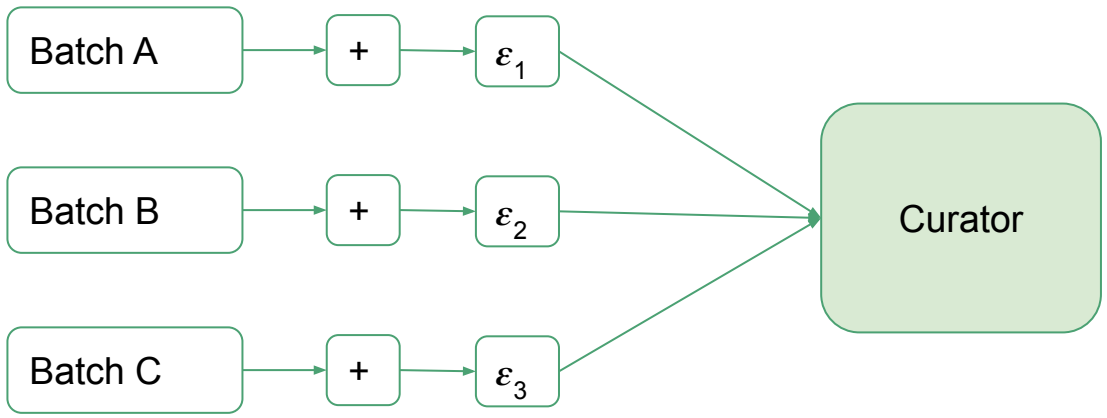


Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Central: All values are maintained and noised by a central coordinator
  - Fine for small datasets, but rapidly does not scale.
  - How big are some of the datasets at Google?
    - **Multi-terabyte datasets are normal!**

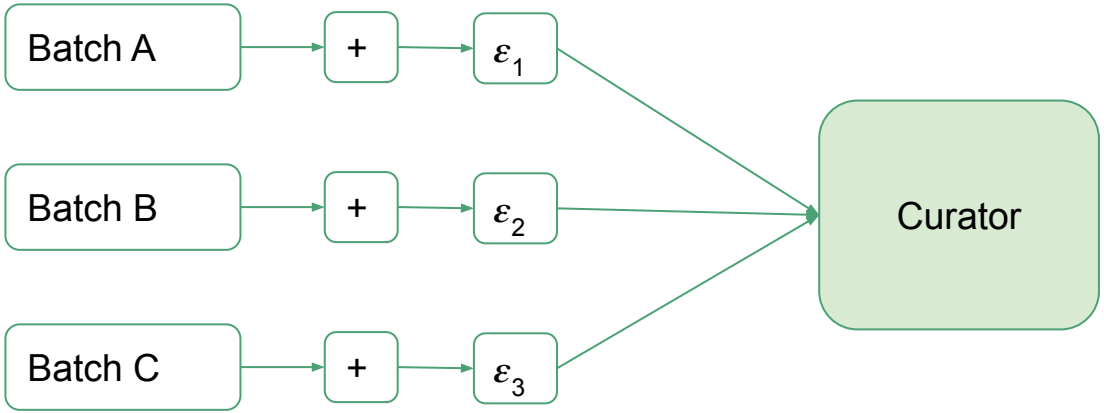
Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Parallel Batching: Split the data into batches, process them in parallel
  - Still have a curator in this central model to maintain noise and budget



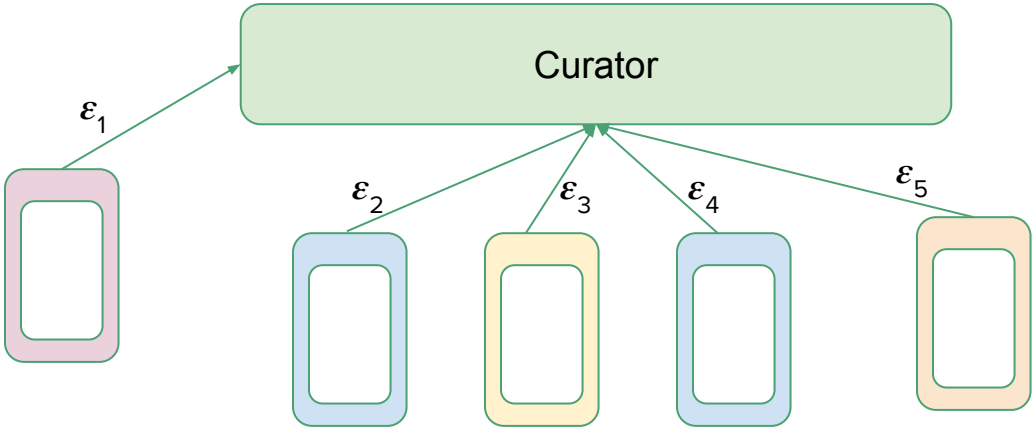
Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Parallel Batching:
    - Issues with utility if privacy design is not baked into optimizations
    - See Plume: Differential Privacy at Scale



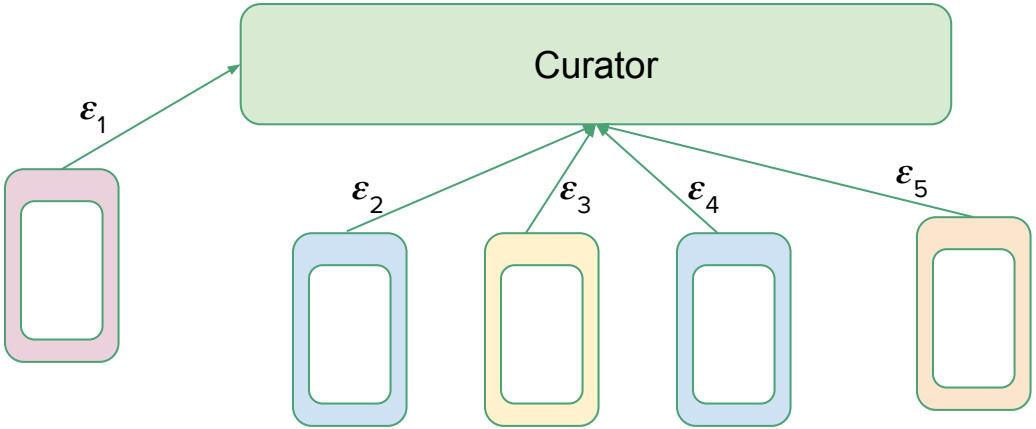
Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Local: noise is applied on the device so the raw values are never seen by anyone other than t owner of the data



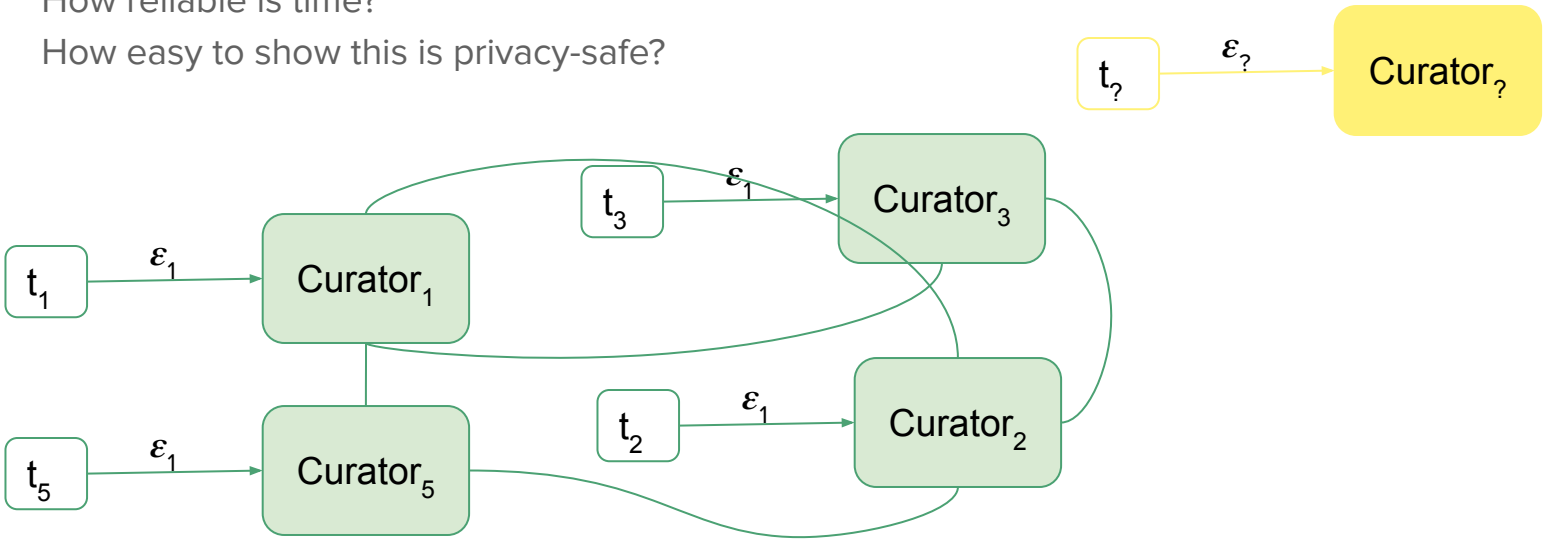
Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Local
    - How much more noise is applied compared to batching? Compared to central?



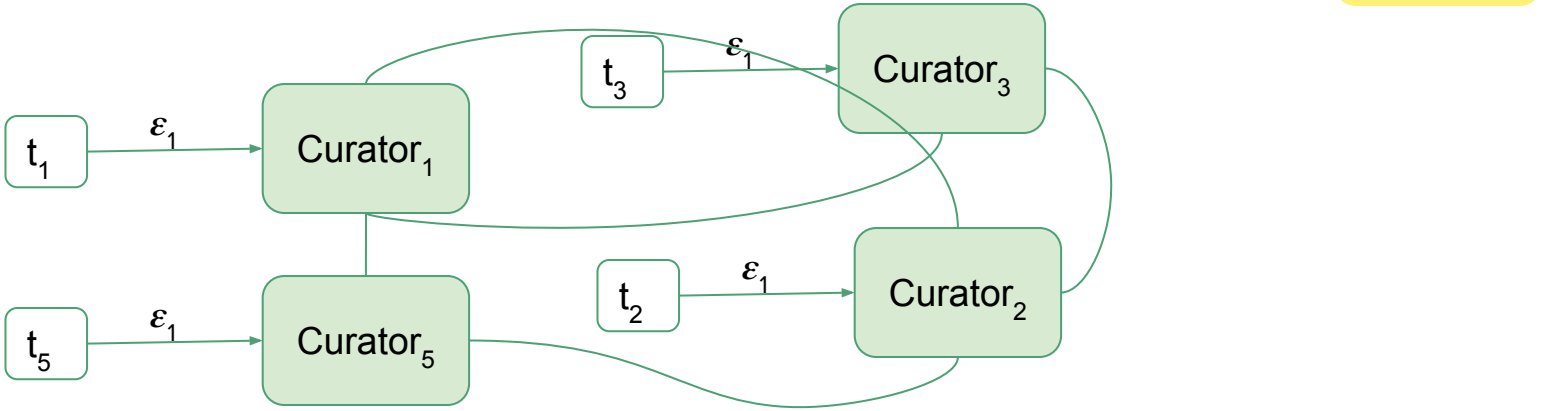
Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Fully Decentralized Model
    - How many more communication costs compared to batching?
    - How reliable is time?
    - How easy to show this is privacy-safe?



Differential Privacy: Models

- Types of Models: Where the noise is applied
  - Fully Decentralized Model
    - How many more communication costs compared to batching? **Many more!**
    - How reliable is time? **Unreliable, out-of-order delays, failures.**
    - How easy to show this is privacy-safe? **Complex.**
    - This is still an active area of research.



Conclusion

We've covered these topics in privacy:

- Policies
- Pseudonymization: De-identification and k-anonymity
- Anonymization: Differential Privacy
- Differential Privacy in Practice
- Models of Differential Privacy

Resources

Google’s Differential Privacy Core Libraries

Privacy on Beam

OpenMined

Q&A

Thank you!

1
2
3
4
5
6
7
8
9
10
11