Privacy Engineering

Jodi Spacek

UBC MSc '19
Senior Privacy Engineer, Google
(opinions are my own)

About me

- Music⟶Software Development ⟶BCS and Distributed Systems @ UBC ⟶Privacy Engineering @ Shopify ⟶Google Privacy 😅
- Roads to Privacy Engineering may not be well defined
- Many different backgrounds contribute an important variety of perspectives into Privacy for Everyone

Photo by Fahad Hussain from Pexels

Privacy @ Google

Product Privacy: Search, YouTube, Maps, Chrome, Payments, Nest

Photo by Karolina Grabowska from Pexels

Privacy @ Google

- safety.google
- Privacy Sandbox
- Covid-19 Community Mobility Reports
- Differential Privacy Usability (from PEPR 20)
- https://www.usenix.org/conference/pepr20
- PEPR 2021 CONFERENCE ON PRIVACY ENGINEERING PRACTICE AND RESPECT

Today’s Topics

Part 1: Privacy Engineering

- Regulations and Considerations
- Privacy and Security
- Privacy Threat Modeling
- Privacy by Design

Part 2: Privacy Toolkit

- K-Anonymity
- Group Activity: K-Anonymous Beverages
- Differential Privacy
- Anonymization Tools and Applications

Part 1: Privacy Engineering

Regulations

Regulations

GDPR/CCPA/PIPEDA/HIPAA differences:

- **Data Handling**:
    - Personal data handled by all types of organizations (GDPR)
    - Data handled by for-profit organizations (CCPA)
    - Data handled by private sector organizations (PIPEDA)
    - Patient data (HIPAA)
- Forms of **consent** (opt-in vs opt-out)

Note these regulations are in flux (eg. CCPA replacement with CPRA)

Privacy Philosophies with shared goal in mind

Regulations

GDPR (General Data Protection Regulation) Seven Key Principles

1. Lawfulness, fairness and transparency
2. Purpose limitation
3. Data minimisation
4. Accuracy
5. Storage limitation
6. Integrity and confidentiality (security)
7. Accountability

CCPA (California Consumer Privacy Act) Listed Rights

Right to know, delete, opt-out, non-discrimination

Regulations

PIPEDA (Personal Information Protection and Electronic Documents Act)

Principles and No-Gos

No-gos:

- Unlawful use
- Human rights violations
- Harm
- Surveillance

HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule

Movement of data, Data protection

Rights over information: view, delete, modify

Regulations

Core Concepts

- Data minimization
  - Retention
  - Purpose
  - 3p (Third party sharing)
- Transparency
  - opt-in/opt-out
  - readability
- Right to be forgotten (rtbf)

Regulations

Core Entities:

- Data Subject (the user)
- Data Controller (storage, transport)
- Data Analyst
- Data Protection Authority (DPA)

Considerations

Privacy Engineers consider the User

- What actions are available to the user?
- What can they see?
- Do users know why the data is being collected?
- Do the engineers know why the data is being collected?

Considerations

Privacy Engineers consider the User

- What actions are available to the user?
    - Request button to delete data, download their data?
- What can they see?
    - Rationale on why a particular ad was shown, stored credit cards on file.
- Do users know why the data is being collected?
    - Short term: payment transactions, long term: learn about the user’s preferences.
- Do the engineers know why the data is being collected?
    - Accountable to the user

Considerations

Privacy Engineers consider the Data Flow

- Where does the data live?
- Where does the data travel?
- How do we treat ephemeral data?

Considerations

Privacy Engineers consider the Data Flow

- Where does the data live?
  - In a locked closet on a single server, in a server farm, in the cloud?
- Where does the data travel?
  - Stays on a laptop, travels within country, going on a world tour?
- How do we treat ephemeral data?
  - 3-7 day retention periods, aggregating data, then discarding sources

Considerations

Privacy Engineers consider the Side Effects

- What portions of data are logged?
- Where is data aggregated and transformed?
- How is side effect data tied back to the user?

Considerations

Privacy Engineers consider the Side Effects

- What portions of data are logged?
  - What are the logging retention periods? Is this data anonymized?
  - May need to keep logs for auditing, customer support, debugging
- Where is data aggregated and transformed?
  - Where are the data sources and sinks? Fan-outs?
- How is side effect data tied back to the user?
  - A user has the right to request their data be deleted. Can we fulfill this request?

Considerations

Privacy Engineers consider Consistency and Availability

- Does a deletion request really delete?
- How long does it take to delete?
- Can users see their data reliably?
- Access all their privacy options?

Considerations

Privacy Engineers consider Consistency and Availability

- Does a deletion request really delete?
  - Most data will be replicated; consider if the deletion request propagates
  - Make sure it’s a hard delete, not soft deletion (eg. flagged as deleted)
- How long does it take to delete?
  - Are there background jobs that take a long time time to run?
  - Is there a backlog of deletion requests?
- Can users see their data reliably? Access all their privacy options?
  - Can users download all of their data? Are the privacy options high priority (eg. service level objectives and agreements, SLO/SLA)

Privacy Pause 🤔

Source: https://vancouversun.com/news/local-news/dude-chilling-park-sign-stolen

Privacy and Security

Privacy and Security 🤝

Security is a necessary but not sufficient condition of Privacy

- **Security**: encryption mechanisms, authentication and authorization, physical security, vulnerabilities, unauthorized and social engineering attacks
- **Privacy**: handling data (deletion, use, regional storage), inferences from joining public datasets, pseudonymization, k-anonymization, ***l***-diversity, differential privacy, re-identification and database reconstruction attacks

Privacy Pause 🤔

Can we have security without privacy?

Can we have privacy without security?

Privacy Pause 🤔

Can we have security without privacy? **Yes**

Fully access controlled system, encrypted data, keys in secure location, biometric access only.

The system collects all of your data, never obfuscates it, keeps it forever, and it has no mechanism for you to request deletion, nor can you view it.

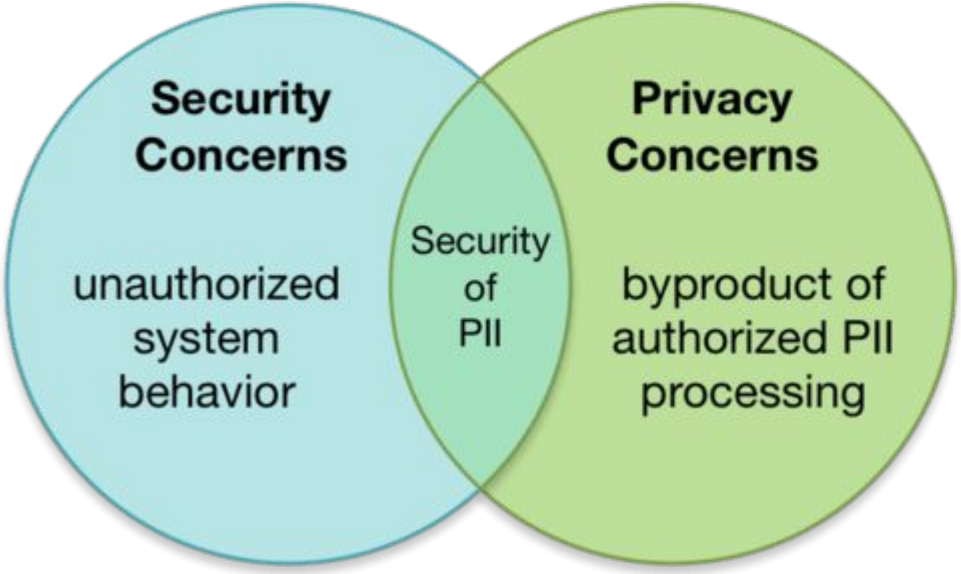
Can we have privacy without security? **No**

All of your personally identifiable information is scrubbed in the database as per privacy regulations.

But, I can sniff your network and the connection isn’t encrypted.

🙀

Privacy and Security 🤝



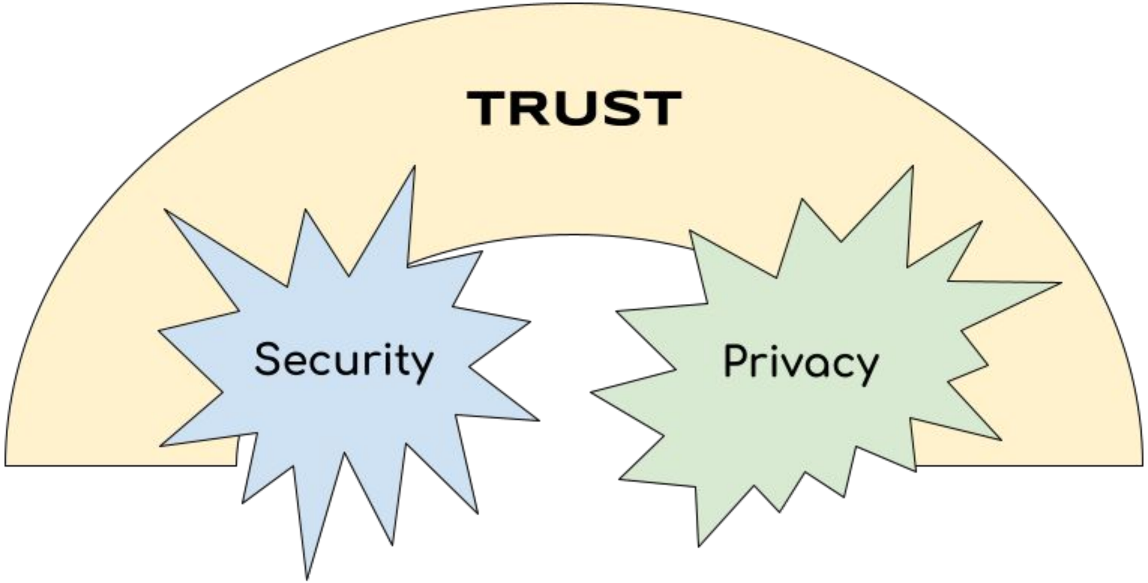
Security Concerns
unauthorized system behavior
Security of PII
Privacy Concerns
byproduct of authorized PII processing

PII = Personally Identifiable Information

Source: https://www.nist.gov/blogs/cybersecurity-insights/making-privacy-concrete-three-words-not-usually-found-together

Privacy and Security 🤝

Together, Privacy and Security build Trust



Privacy Threat Modeling

Privacy Threat Modeling

- What is a model?
- What is a threat model?
- What is the difference between a security and privacy threat model?

Privacy Threat Modeling

- What is a model?
    - A simplified view of the world that helps us to analyze and understand a system.
- What is a threat model?
    - A model that considers threats to the system and mitigations.
- What is the difference between a security and privacy threat model?
    - A privacy model considers threats to privacy (PII processing).
    - It relies on the security threat model.
    - Look at processing/joins/data flow

Privacy Threat Modeling

Privacy Engineers consider:

- Who can access the data and their level of trust
- Source and destination of the data
- Linkability with other data sets (outside/inside)
- Quality of anonymity of the data

Privacy by Design

Privacy by Design

1. **Proactive** not reactive; preventive not remedial
2. Privacy as the **default** setting
3. Privacy **embedded** into design
4. Full functionality – **positive-sum**, not zero-sum
5. End-to-end security – full **lifecycle** protection
6. **Visibility** and transparency – keep it open
7. **Respect** for user privacy – keep it user-centric

https://en.wikipedia.org/wiki/Privacy_by_design

Privacy by Design

What does this mean in practice?

Think about privacy early on, starting from the design stage.

The opposite treatment is **“ad-hoc” privacy**. 🙅‍♀️

https://en.wikipedia.org/wiki/Privacy_by_design

Privacy by Design Examples

Privacy by Design: Jurisdictional Data

- Specific Canadian routing
- BC Liquor board, Cannabis sales

How does this affect data at rest?

What changes need to be made for data in transit/routing?

Will this be easy to change once the system is up and running?

Privacy by Design: Jurisdictional Data

How does this affect data at rest?

Data at rest must be stored in Canada.

What changes need to be made for data in transit/routing?

Data in transit must never leave the country.

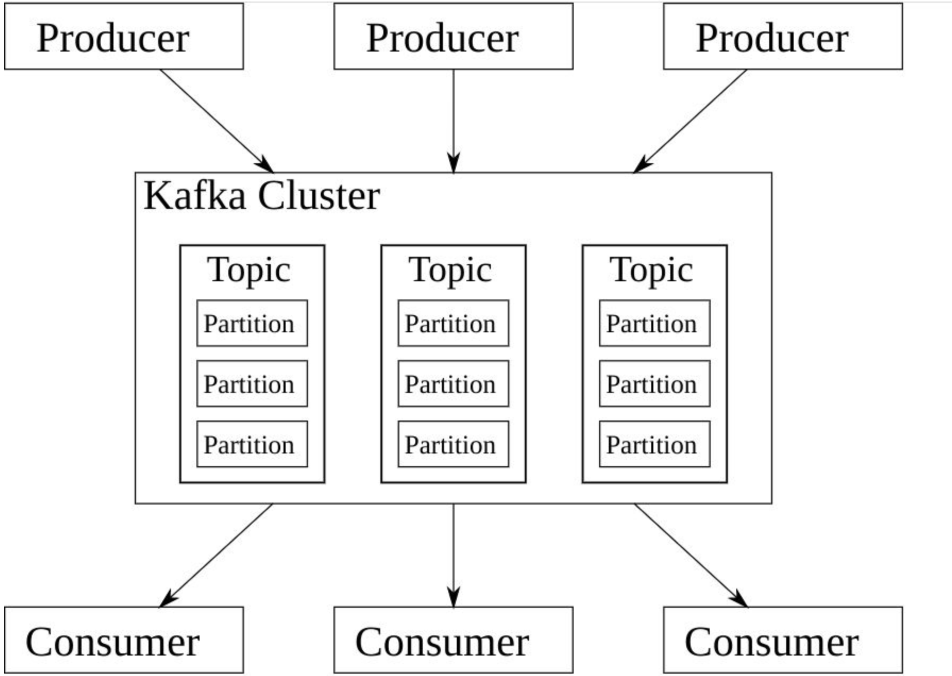
Will this be easy to change once the system is up and running?

Typically difficult to enforce due to deeply ingrained routing protocols.

This needs to be designed into the system from the get go.

It cannot be added after the system is live.

Ad-hoc Privacy: Kafka Compaction



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

Ad-hoc Privacy: Kafka Compaction

What is the retention period in Kafka?

What is the role of compaction?

How frequently does compaction happen?

What triggers compaction?

Ad-hoc Privacy: Kafka Compaction

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.

Problem: We want to delete a record by sending a null value and compact in 3-7 days.

Ad-hoc Privacy: Kafka Compaction

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

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

(Time-based vs record-based compaction.)

How easy is it to upgrade an existing system to implement this patch?

What happens with the records already with this bug?

Privacy by Design: Headless Encryption

Suppose you need to store data in a data lake, where it is not easily nor efficiently indexed into.

The system is already in place; you cannot refactor the existing code that saves data to the data lake.

A privacy by design solution is:

- Encrypt the data in the data lake, store a key per record and/or user
- Destroy the keys after retention period passes then
- Tie the key back to the user in order to respond to deletion requests

Privacy Poll

When do you think implementing Privacy will be finished?

[Please type your answers in the chat.]

A. GDPR is done; another year or so to wrap things up?
B. Another couple of years; because of the changes to CCPA?
C. Everything privacy will be automated in about 5 years?
D. Privacy work is an ongoing part of the foreseeable future?

Part 2: Privacy Toolkit

Anonymization

Not quite anonymization 🚕

NYC Taxi Data Set

Obtained by Freedom of Information request

Contained 173 million individual trips with;

- locations and times for pickup and dropoff
- **anonymized** the licence number using hash function 🤔
- medallion number also hashed
- other metadata

Source https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1

Not quite anonymization 🚕

NYC Taxi Data Set

Hashing doesn’t work when we have knowledge of the format of the data

- Taxi licence numbers are 6-digit, medallion has unique format as well
- Map-reduce to crunch away at all hashes for an hour
- Discover the numbers that produce the hashes
- Map taxi license numbers to the driver's name using external dataset (licence number, name)
- Can figure out addresses, work schedules, gross salary, locations!

Re-identification Attack: this is a privacy leak by joining on another dataset!

Lessons Learned: Use secure encryption that is cryptographically strong (AES) and/or use random numbers for the licence and medallion.

Source https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1

K-Anonymity

K-Anonymity

Guarantees that an individual’s data cannot be distinguished from some k-1 other users.

If k = 2, then if I like Samoyeds and I live in Ottawa, there should be another individual who likes Samoyeds and also lives in Ottawa.

This dataset is not k=2 anonymous ⟶

Favourite Animal	City
	Ottawa
🐕	Ottawa

@mayapolarbear Instagram

Source: Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 557-570

K-Anonymity

The dataset is adjusted to provide this guarantee using two methods:

- **Suppression**
  - Replace sensitive fields with a default value (eg. *)

- **Generalization**
  - Roll up a field to a more common value, eg. Ottawa becomes Ontario

Source: Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 557-570

K-Anonymity

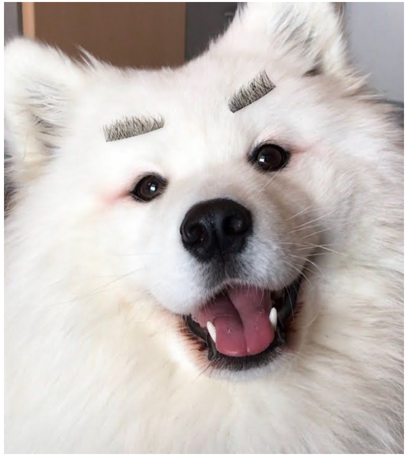
Now the dataset is k=2 anonymous.

But, it is not very interesting, is it?

If I want to see Samoyed pictures in my feed, how likely is it that this dataset will be useful?

What kind of Ads will I see for dog breeders in Ottawa?

Favourite Animal	City
Dog	Ottawa
Dog	Ottawa

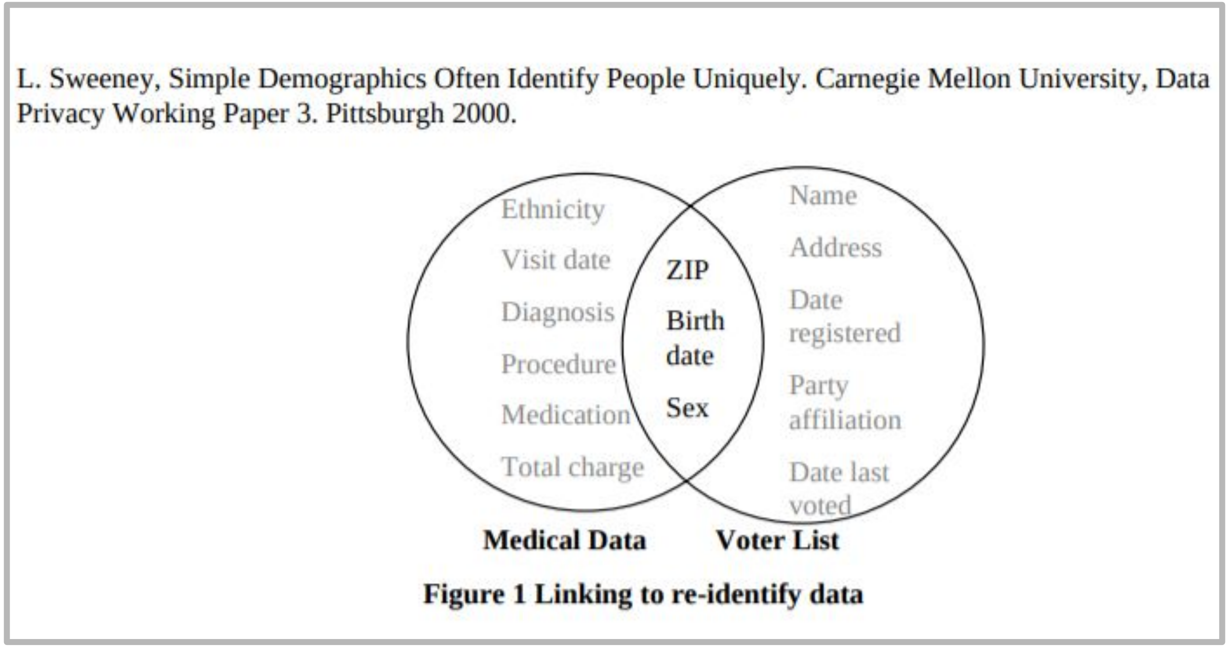


@mayapolarbear Instagram

Linkage Attack

A simple, insightful attack using public datasets, like the census, or voter data, is to join this publicly available data with other publicly available datasets.

This attack motivated the K-Anonymity approach also by **Latanya Sweeney** (Professor at Harvard).



L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

Figure 1 Linking to re-identify data

See also: Netflix Linkage Attack

K-Anonymous Beverages ☕ 🍵

You want to send this dataset to a 3rd party application so that your customers see ads for cafes nearby. You do not want to reveal too much information about them (Data Minimization)! Let’s see if we can anonymize it...

Name	Beverage	Province	Age
Leslie	Espresso	BC	22
Lee	Latte	BC	23
Pat	Pour Over	AB	29
Lyn	London Fog	BC	37
Sam	Matcha Latte	ON	39

K-Anonymous Beverages ☕ 🍵

First, suppress the names by replacing them with *.

What if you knew someone posted a picture of their latte every morning and they also posted their graduation photos 2 years ago? Then you can guess that Row 2 is likely Lee.

Name	Beverage	Province	Age
*	Espresso	BC	22
*	Latte	BC	23
*	Pour Over	AB	29
*	London Fog	BC	37
*	Matcha Latte	ON	39

K-Anonymous Beverages ☕ 🍵

Next, let’s generalize the ages by bucketing them. Great!

But I know about someone from this site who ordered matcha to Ontario. I know their age range too. And I have another dataset with their name; I can tell it's probably Sam!

Name	Beverage	Province	Age
*	Espresso	BC	20-30
*	Latte	BC	20-30
*	Pour Over	AB	20-30
*	London Fog	BC	30-40
*	Matcha Latte	ON	30-40

Group Activity: K-Anonymous Beverages

Discuss in small groups...

How can you generalize (roll-up) the beverages of your users?

They like espresso, latte, pour over, london fog, and matcha latte.

How can you generalize (roll-up) the provinces of your users?

They are from BC, AB, and ON.

What kind of inferences can you make from this anonymized dataset?

K-Anonymous Beverages ☕ 🍵

So you likely came up with something that looks like the table below. Great! This is safe to send to the third party app.

Ok, so I really like matcha lattes and I am in Ontario. Can you please direct me based only on data from this table to a hip cafe? A cafe...with...tea somewhere...in Canada?

Name	Beverage	Country	Age
*	Coffee	Canada	20-30
*	Coffee	Canada	20-30
*	Coffee	Canada	20-30
*	Tea	Canada	30-40
*	Tea	Canada	30-40

K-Anonymous Beverages ☕ 🍵

Oops!

This dataset is really too small to properly anonymize and keep any utility in the anonymized data.

K-anonymity can sometimes work, if the dataset is:

- really large so you have k >= 50

Privacy Poll 🤔

Is anonymous data useful?

[Please type your answer into the chat.]

A. **Yes!**
B. **No way!**
C. **This is a trick question!**

Privacy Pause 🤔

Is anonymous data useful?

It depends.....on How

(╯°□°)╯︵ ┻━┻

Privacy Pause 🤔

Is anonymous data useful?

Details Coming up!

┬─┬ノ(ಠ_ಠノ)

Differential Privacy

Differential Privacy

Groundbreaking work in Privacy: 2016 TCC Test-of-Time Award and 2017 Gödel Prize

- Generally the Gold Standard for anonymization
- State of the art for publicly available datasets.
- Adds noise, but can be tuned to maintain useability for data analysts
- Great for aggregating data (count, sum, etc.)

The general idea is that if one individual is removed or added to the original dataset, we cannot tell anything about that individual's data.

See **Cynthia Dwork** [Fault Tolerance, Differential Privacy, Professor at Harvard]

Differential Privacy

Consider a simple example with a divisive question:

Do you like raisins in cookies?! 🍪 + 🍇

- Use a randomized response to introduce noise
- 25% of the answers are noise, randomly answer Yes or No
- There is a chance my answer is noise, so I can plausibly deny my answer
- In practice, the noise is based on more complex distribution.

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

Do you like raisins in cookies?! 🍪 + 🍇

- The amount of privacy, epsilon, described as ε-differentially private.
- 0-differentially private is pure noise and perfectly private!
- It is also useless for analysis.
- Trade-off between privacy and utility.
- A privacy budget limits requerying because this leads to privacy loss.
- Best for static datasets shared publicly, eg. medical, census data.

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.

Anonymization Tools and Applications

Machine Learning Privacy: Tensorflow

Tensorflow already uses Federated Learning, which is a great friend to Privacy since training data is stored locally.

Privacy is core to Tensorflow; it has differentially private training!

There is a a library dedicated to ensuring that training data cannot be inferred from the model. https://github.com/tensorflow/privacy

This library checks against inference attacks. (Note they are a **privacy** attack, not security.)

Nasr, Milad, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning." *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019.

Papernot, Nicolas. "Machine Learning at Scale with Differential Privacy in TensorFlow." *2019 {USENIX} Conference on Privacy Engineering Practice and Respect ({PEPR} 19)*. 2019.

Machine Learning Privacy: Tensorflow

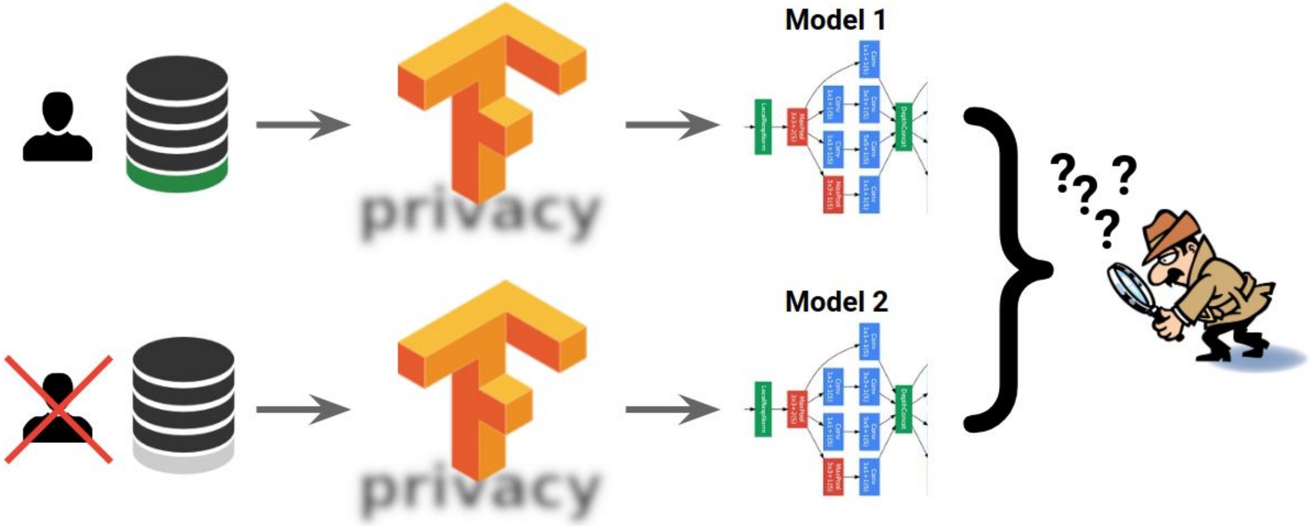


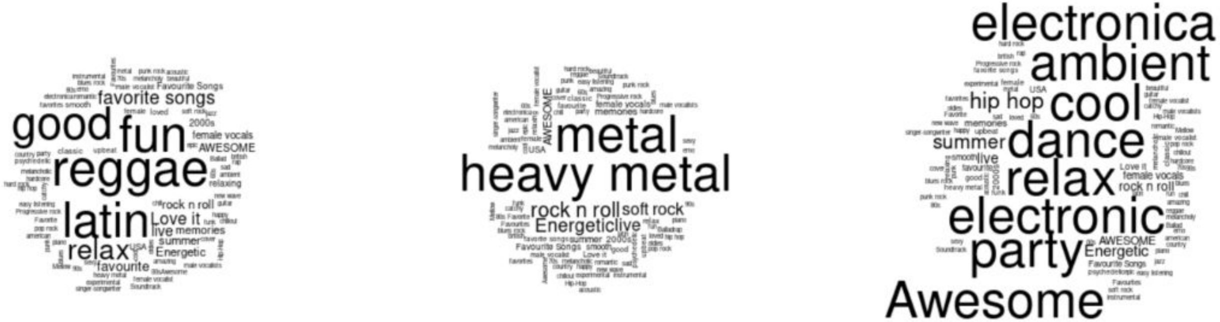
Image from https://blog.tensorflow.org/2019/03/introducing-tensorflow-privacy-learning.html

K-Anonymity: Privacy Sandbox

An example of practical k-anonymity Whitepaper: removes third party cookies, use cohort membership instead

- Users assigned into cohorts based on interest
- Large user base helps with privacy guarantees, large k
- Cohort Ids are computed without sharing information (SimHash)

Figure 6: Word Clouds for SimHash cohorts using 8 bits



K-Anonymity: Privacy Sandbox

Federated Learning of Cohorts (FLoC)

- Changes learning about an **individual** to learning about a **group** of cohorts
- FLoC are shared among thousands of users
- Used k-anonymity instead of differential privacy because of **fingerprinting** concerns
- A cohort Id could be differentially private, but it would be unique
- Cohort Ids are k-anonymous, not vulnerable to fingerprinting attacks

Publicly Released Data

Google recently released differentially private data to track movement during the pandemic. Covid-19 Mobility Reports

Aggregated, anonymized insights using Google Maps.

Some restrictions:

- No new attributes in the dataset
- No requerying the dataset (privacy budget!)
- Static queries by location

Read more about it on the blog:
https://developers.googleblog.com/2021/01/how-were-helping-developers-with-differential-privacy.html

Engineers and Privacy

- Consider many different perspectives on the treatment of data
- Every layer of the stack does have an impact on privacy
- Ask for advice and help from the legal experts
- Think about your data geographically
- Use your imagination 🎉
- Think about how data is **distributed** around the world!

Q&A