Privacy-preserving methods: Building secure projects

Rebeca Sarai 👩‍💻

Recife, Brazil ☀️ 🌊 🏖️ 🏝️ 🇧🇷

Computer Engineering by University of Pernambuco 👩‍🎓

Football fan ⚽

Software Engineer

@_rebecasarai

VINTA

Build smart
Venture beyond

We're a **team of experts** from Brazil.
We help our clients **evolve their products** the right way with **top notch** development and UX techniques.

Get to know us: vintasoftware.com

Security and Privacy

- Scientists don't have enough data to build new models
- **We don't feel safe**

Opinion

The Apps on My Phone Are Stalking Me

I discovered that we're building a digital surveillance state much like the one in China.

By Farhad Manjoo
Opinion Columnist

Jan. 22, 2020

184

Google tracked his bike ride past a burglarized home. That made him a suspect.

"I was using an app to see how many miles I rode my bike and now it was putting me at the scene of the crime," the man said.

Recife tracks 700,000 cell phones to monitor social isolation and direct actions against coronavirus

According to the city, the Isolation Index was created to find out in which places the restriction measures are being complied with.

By G1 PE

03/24/2020 16h09 · Updated há uma semana

NETFLIX

Netflix Prize

COM

Home | Rules | Leaderboard | Update

Congratul

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

Latest: Strava suggests military users 'opt out' of heatmap as row deepens

ovince, Afghanistan with route taken by joggers highlighted by Strava. Photograph:

about the location and staffing of military bases and
he world has been revealed by a fitness tracking

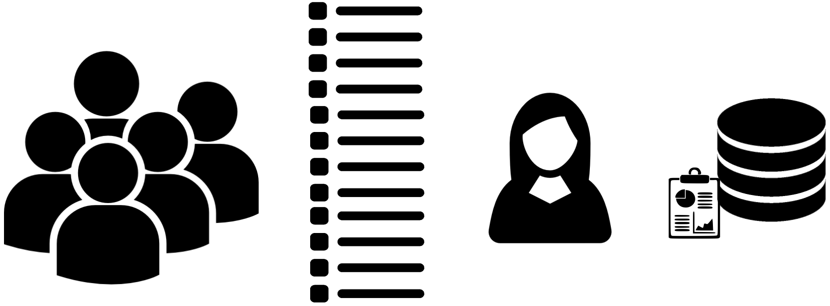
Zoom CEO apologizes for having 'fallen short' on privacy and security

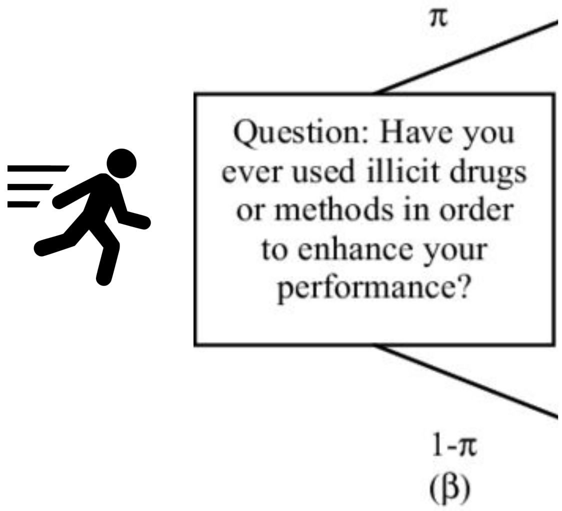
By Rishi Iyengar

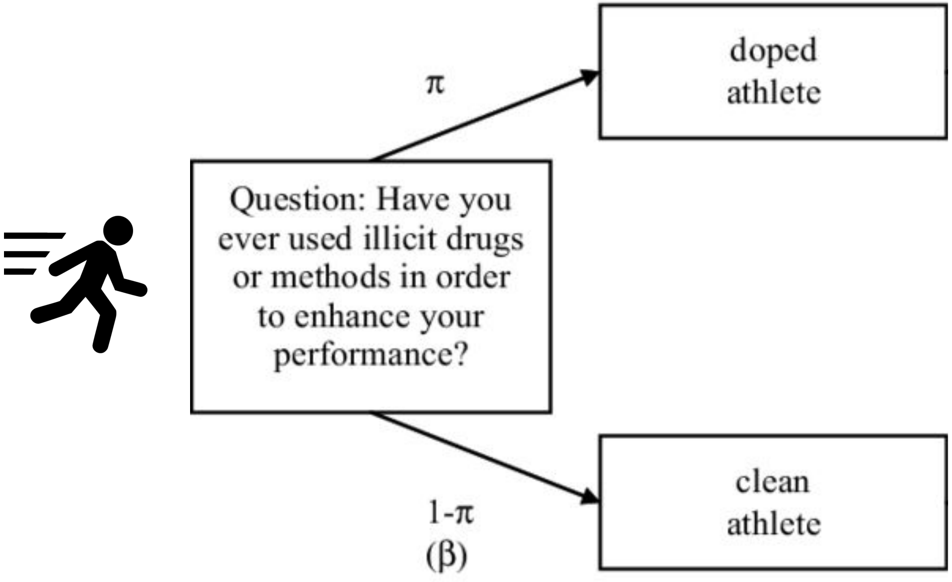
Updated 2103 GMT (0503 HKT) April 2, 2020

You want to collect and release data that contains answers to sensitive questions

You want to collect and release data that contains answers to sensitive questions




π
Question: Have you ever used illicit drugs or methods in order to enhance your performance?
1-π
(β)


π
doped athlete
Question: Have you ever used illicit drugs or methods in order to enhance your performance?
1-π
(β)
clean athlete

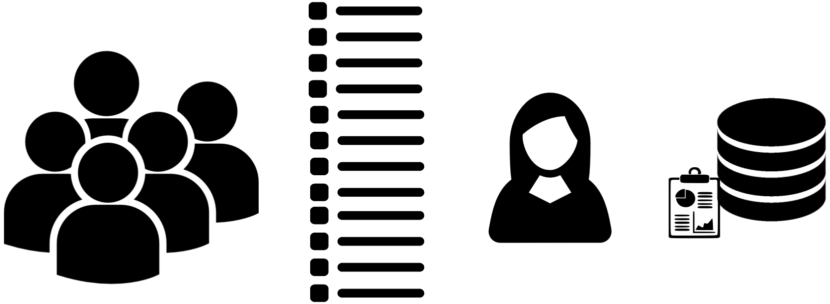
“How many people in the database have used illicit drugs?”

“How many people, not named Jane, in the database used illicit drugs?”

Summary Statistics are Not “Safe”

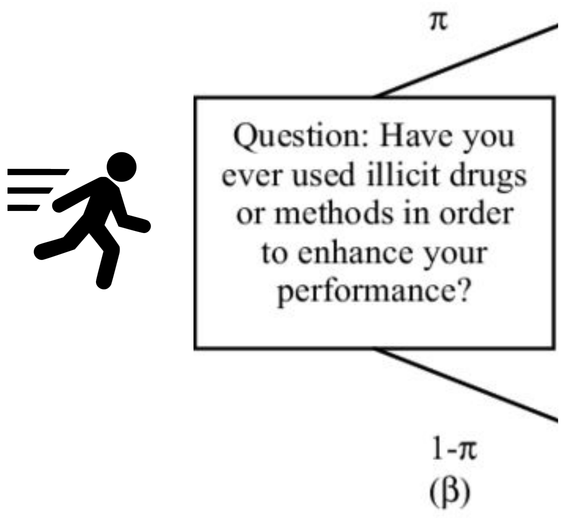
- Differencing attacks
- Reconstruction attacks
- Each individual has a “secret bit”

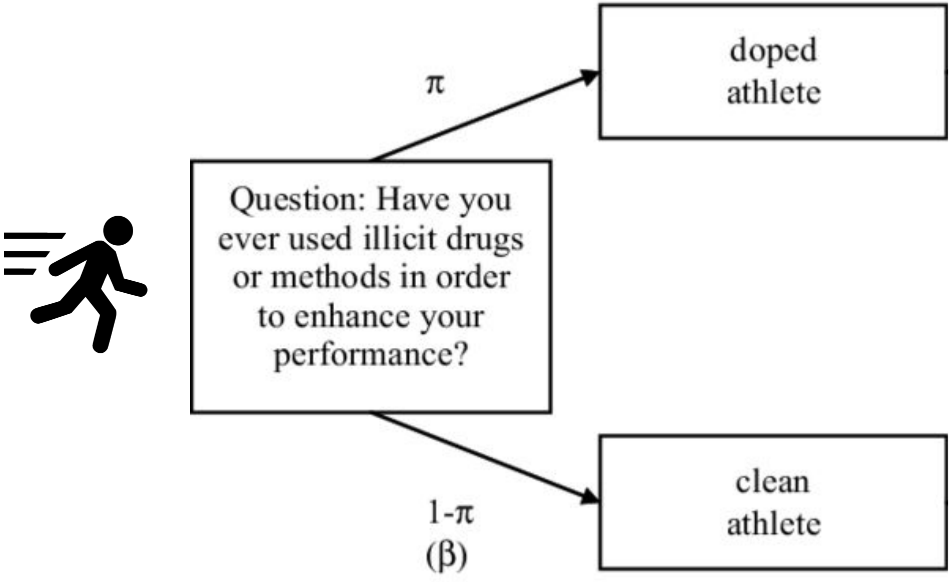
You want to collect and release data that contains answers to sensitive questions

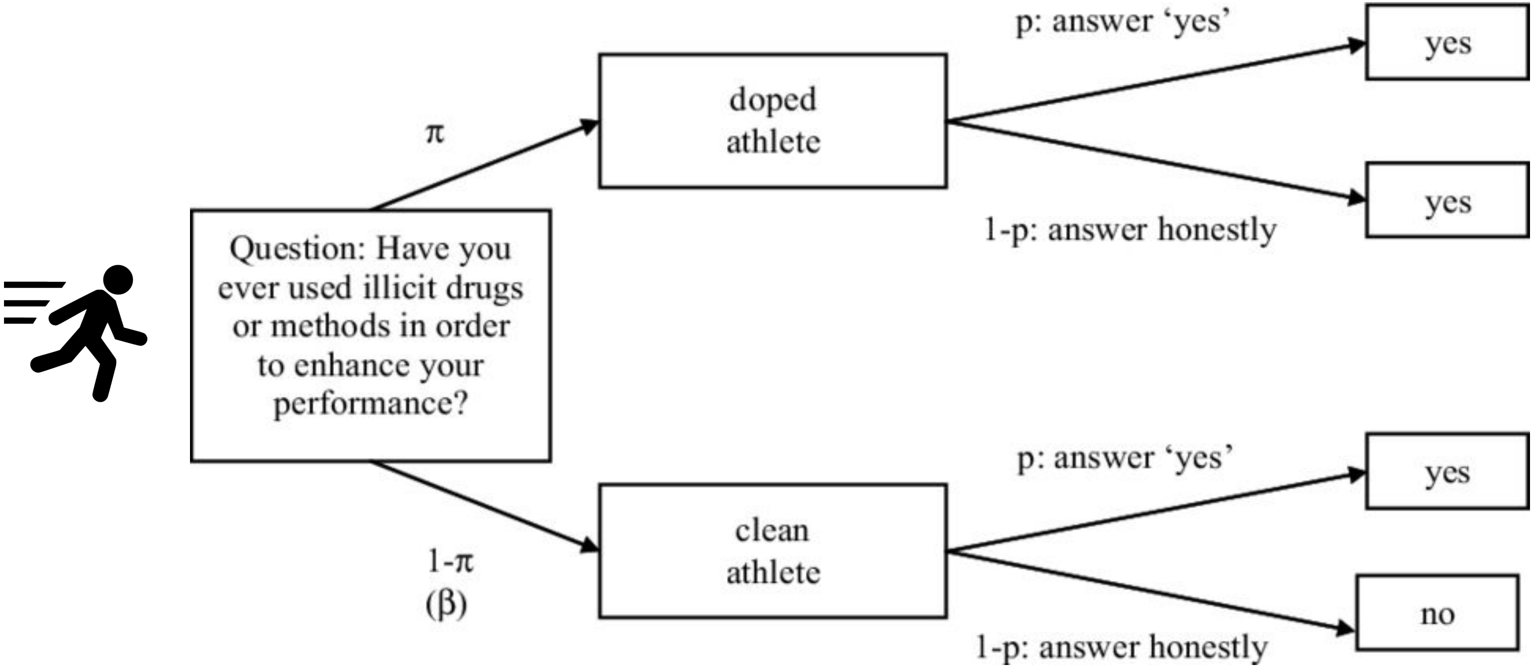


Randomization

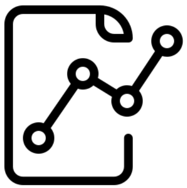
- 1965
- **Plausible deniability** (coin flip mechanism)
- Good if you have **many examples**
- Allow the recovering of the **underlying statistics**


π
Question: Have you ever used illicit drugs or methods in order to enhance your performance?
1-π
(β)


π
doped athlete
Question: Have you ever used illicit drugs or methods in order to enhance your performance?
1-π
(β)
clean athlete


Question: Have you ever used illicit drugs or methods in order to enhance your performance?
π
doped athlete
1-π
(β)
clean athlete
p: answer 'yes'
yes
1-p: answer honestly
yes
p: answer 'yes'
yes
1-p: answer honestly
no

You want to make generalizations over a population



Queries Over Large Sets are Not Protective

"How many people in the database have the sickle cell trait?"

"How many people, not named Jane, in the database have the sickle cell trait?"

Queries Over Large Sets are Not Protective

"How many people in the database have the sickle cell trait?"

+

"How many people, not named Jane, in the database have the sickle cell trait?"

Differencing attack

The state or condition of being free from being observed or disturbed by other people.

> *The state or condition of being free from being observed or disturbed by other people.*

> *Privacy is preserved if… after the analysis, the analyzer doesn't know anything about the people in the dataset. They remain "unobserved".*

The state or condition of being free from being observed or disturbed by other people.

Privacy is preserved if... after the analysis, the analyzer doesn't know anything about the people in the dataset. They remain "unobserved".

Anything that can be learned about a participant from the statistical database can be learned without access to the database

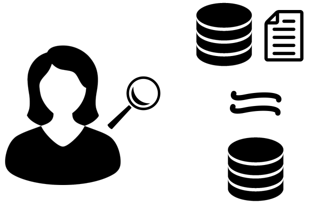
Agreement between a data holder and a data subject: The owner of the data ***will not be affected****, adversely or otherwise,* ***by allowing your data to be used*** *in any study or analysis,* ***no matter what other studies, datasets, or information sources are available***

Differential Privacy

- 2006
- **Teachings** Database != **Actions** of individual people
- It's a formal **definition** of privacy
- Requires a form of **randomness or noise** added to the **query** to protect **from Differencing Attacks**

In the context of a database:

*Given we perform **some query on the database**, if we **remove a person** from the database and the **query does not change** then that person's privacy is **fully protected***



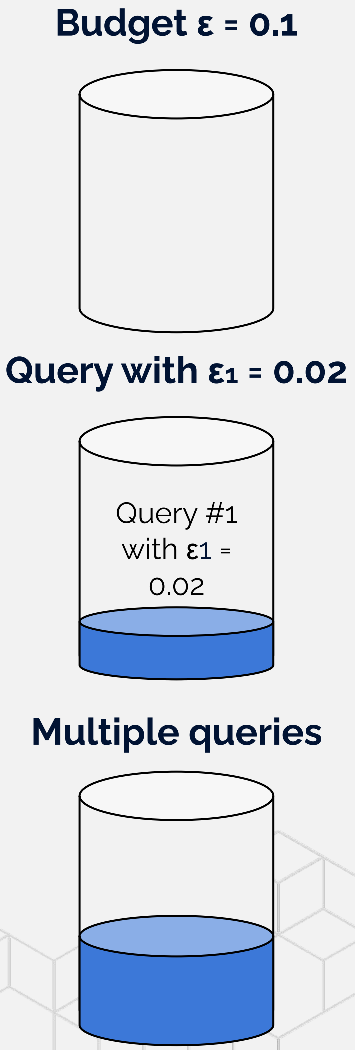
e raised to the **epsilon** power

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

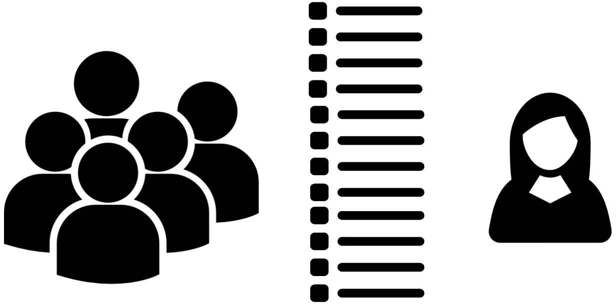
$\mathcal{M}$ is a **randomized mechanism** that gives ε-differential privacy for all data sets

Differential Privacy

- Measure of **privacy loss** ε (**privacy budget**)
- Tune the "**amount of privacy**"
- Privacy-preserving data analysis
- Many open source implementations
    - https://github.com/google/differential-privacy
    - https://github.com/uber-archive/sql-differential-privacy
    - https://github.com/google/rappor
    - https://github.com/prashmohan/GUPT
    - https://github.com/LLGemini/PINQ
    - https://github.com/ektelo/ektelo



You want to collect and release data ~~that contains answers to sensitive questions~~

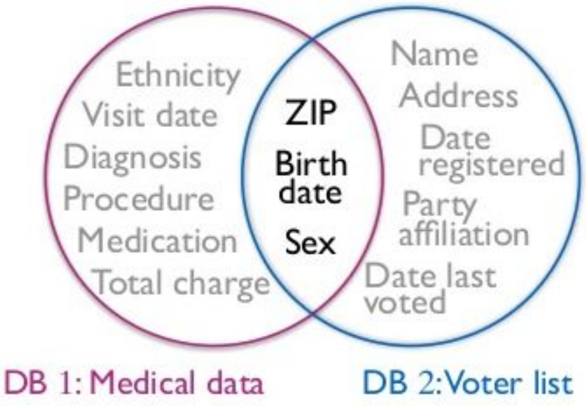


	Quasi identifiers		Sensitive attribute
ID	Age	Country	Salary
1	35	Greenland	>50K
2	35	Canada	<50K
3	38	Belize	>50K
4	40	Belize	>50K
5	37	Canada	<50K
6	37	Canada	<50K

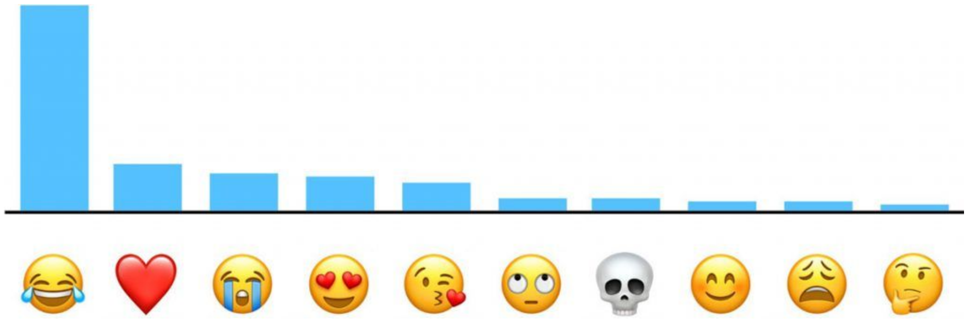
(a) Original census information

	Quasi identifiers		Sensitive attribute	
ID	Age	Country	Salary	
1	35-37	America	>50K	Class A
2	35-37	America	<50K	
3	38-40	America	>50K	Class B
4	38-40	America	>50K	
5	35-37	America	<50K	Class C
6	35-37	America	<50K	

(b) 2-anonymous census information



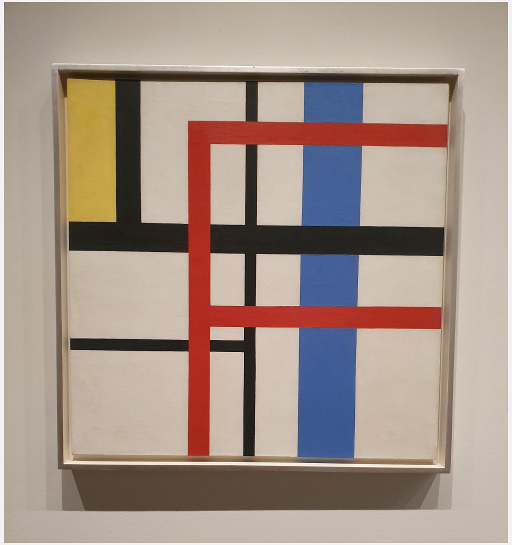
87 % of the US population is **uniquely identifiable** by ZIP, gender, DOB



The Mac Observer

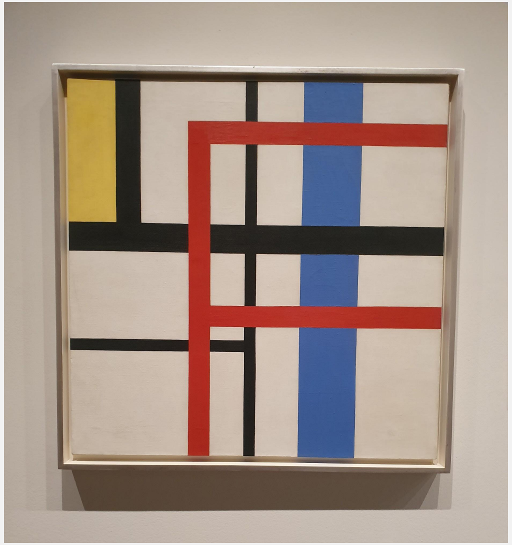
k-Anonymity

- Creates **groups** with at least k records **sharing** the same **quasi-identifiers** values.
- Generalization and Suppression
- Provides **protection** against **identity disclosure**



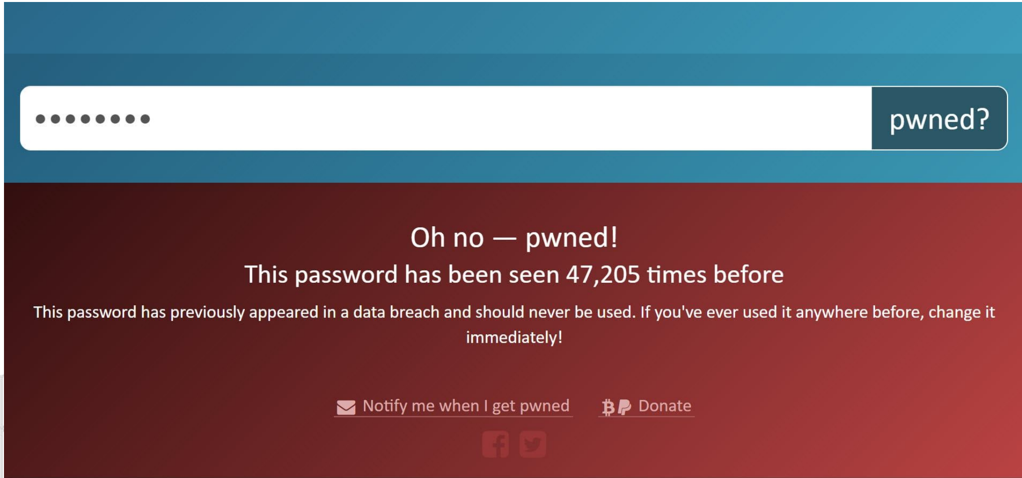
k-Anonymity

- **Linkage attack:** Netflix subscribers issue
- **Refinements** of the k-Anonymity
  (l-diversity, t-closeness, ß-likeness)
- Data Cannot be **Fully Anonymized** and Remain **Useful**.
- **Privacy vs Utility**



Tools

- [have i been pwned]
- [Pwned Passwords with k-anonymity]
- [Validating Leaked Passwords with k-Anonymity]
- [Simple implementation]



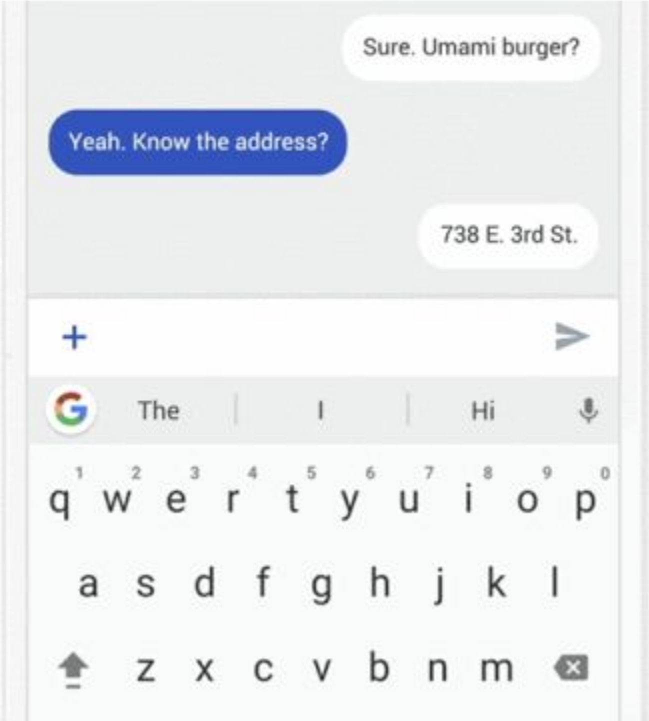
Still need to comply with GDPR

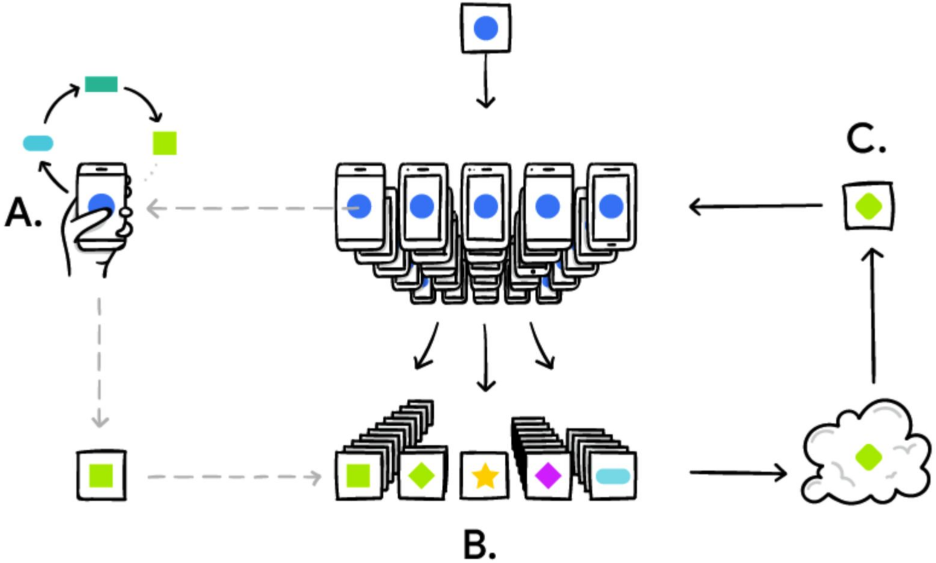
- Encrypt the data in transit
- Encrypt the data at rest
- Encrypt your backups
- Protect data integrity
- Log access to personal data
- Don’t use data for purposes that the user hasn’t agreed with
- Don’t log personal data
- Many more

You want to use prediction models with user's data

There is a problem...

- The **diff** between the **model sent** and the **model received** still **leaks private information**
- By itself, does not guarantee privacy


Sure. Umami burger?
Yeah. Know the address?
738 E. 3rd St.
The
I
Hi
1 2 3 4 5 6 7 8 9 0
q w e r t y u i o p
a s d f g h j k l
z x c v b n m



1. Clients **download** the current model.
2. Each client computes an **updated model** based on their **local data**.
3. The **model updates** are **sent** to the **server**.
4. The server **aggregates** these **models** to construct an **improved global model**

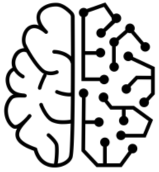
Federated Learning

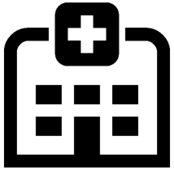
- Enables **mobile phones** to collaboratively learn a shared **prediction model** while keeping all the **training data on device**
- **No** need to store the **data in the cloud**
- Smarter models, lower latency, and less power consumption

Tools

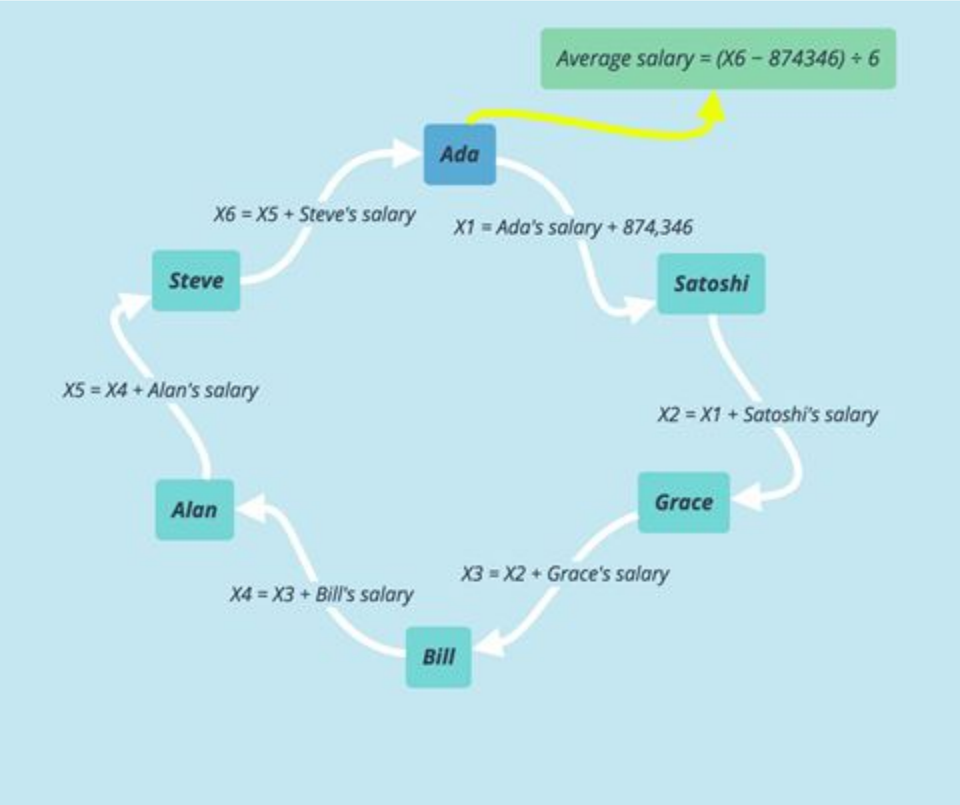
- Pytorch
  - https://github.com/OpenMined/PySyft
- Tensorflow
  - https://github.com/tensorflow/federated
- Federated Learning with Pytorch example

You want to update your model with user's data





To share information ***without*** *the need of a* ***trusted third party*** *to store/process the data. The protocol allows* ***concealing partial information*** *about the data,* ***computing data from many sources*** *without ever revealing individual results*


Average salary = (X6 − 874346) ÷ 6
Ada
X1 = Ada's salary + 874,346
Satoshi
X2 = X1 + Satoshi's salary
Grace
X3 = X2 + Grace's salary
Bill
X4 = X3 + Bill's salary
Alan
X5 = X4 + Alan's salary
Steve
X6 = X5 + Steve's salary

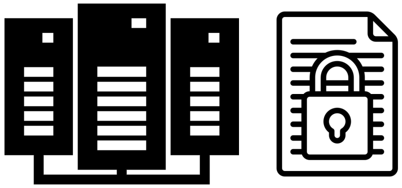
Secure Multi-Party Computation

- Good when a model has **multiple owners**
- Allows for individuals to **share control** of a model
- **No party learns any other party's input**
- Participants are **protected** from privacy leakage, **except** for what can be **inferred from the output**
- Used with other techniques

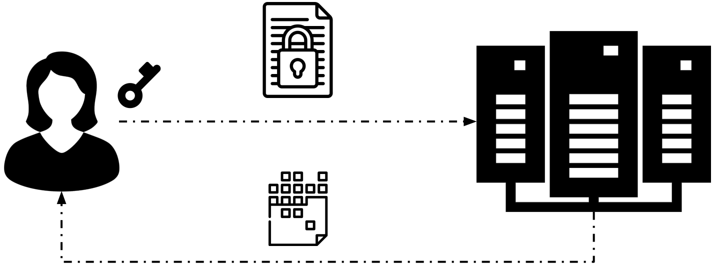
Tools

- Pytorch
    - https://github.com/OpenMined/PySyft
- Tensorflow
    - https://github.com/tf-encrypted/tf-encrypted
- https://github.com/rdragos/awesome-mpc
- Implementation of Multi-Party Computation with Pytorch

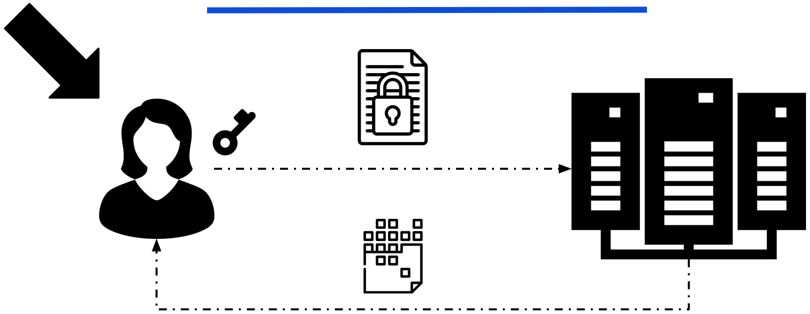
You want to perform safe operations in sensitive data



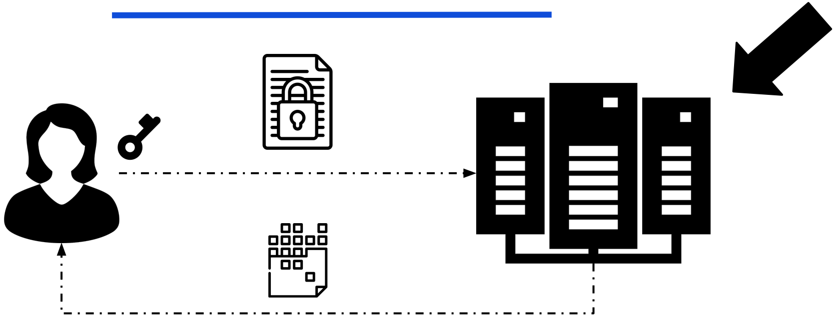
Specific **computations** are performed in **ciphertexts** and the obtained **result is also a ciphertext** that can be revealed only by the owner with a **secret key**



Specific **computations** are performed in **ciphertexts** and the obtained **result is also a ciphertext** that can be revealed only by the owner with a **secret key**



Specific **computations** are performed in **ciphertexts** and the obtained **result is also a ciphertext** that can be revealed only by the owner with a **secret key**



Homomorphic Encryption

- Good when a model has a **single owner**
- Data is **never unencrypted** outside of the users' environment.
- Allows **computation on encrypted data**.
- Secure against quantum computers
- Open source implementations

Homomorphic Encryption

- Still a long way from real-world enterprise implementation
- Tend to work best when processing **integers**
- Slow. IBM's initial release ran '100 trillion times' slower than plaintext operations.

Resume

- **Randomization**: collect sensitive data
- **k-Anonymity**: release dataset
- **Differential Privacy**: aggregate information
- **Federated Learning**: create machine learning models
- **Secure Multi-Party Computation**: distributed processing
- **Homomorphic encryption**: operate over encrypted data

Other things I wish I had the time to mention:

Private Set Intersection

Private Identity Server

Cloak link

Takeaways

- If you deal with sensitive data of europeans citizens **get a lawyer** right now
- Data cannot be **fully anonymized** and remain **useful**
- Queries over large sets are not protective
- Summary **statistics are Not “Safe”**
- Remember **linkage attacks**

"It may seem a paradox, but an open society dictates a right-to-privacy among its members, and we will have thrust upon us much of the responsibility of preserving this right."

Thank you!

Got any questions?

Rebeca Sarai

Software Engineer

rebeca@vinta.com.br

@_rebecasarai

/rsarai

Access this talk on **vintasoftware.com/talks**

References

- CAO, Jianneng; KARRAS, Panagiotis. **Publishing microdata with a robust privacy guarantee**. Proceedings of the VLDB Endowment, v. 5, n. 11, p. 1388-1399, 2012.
- The Algorithmic Foundations of Differential Privacy (here)
- Differentially Private SQL with Bounded User Contribution (here)
- Differential Privacy at Scale: Uber and Berkeley Collaboration (video)
- Tutorial: Differential Privacy and Learning: The Tools, The Results, and The Frontier (video)
- Keeping Your Data Secure While Learning From It - Andreas Dewes and Katharine Jarmul (video)
- 9 Data Anonymization Use Cases You Need To Know Of (here)
- The Definition of Differential Privacy - Cynthia Dwork (video)
- Protecting Personal Data with Django (because it's the law) (video)
- Pseu, Pseu, Pseudio. Pseudonymization in Django. by Frank Valcarcel (video)
- DOMINGO-FERRER, Josep; SORIA-COMAS, Jordi. **Anonymization in the time of big data**. In: International Conference on Privacy in Statistical Databases. Springer, Cham, 2016. p. 57-68.
- LI, Ninghui; LI, Tiancheng; VENKATASUBRAMANIAN, Suresh. **t-closeness: Privacy beyond k-anonymity and l-diversity**. In: 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007. p. 106-115.
- SWEENEY, Latanya. **k-anonymity: A model for protecting privacy**. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, v. 10, n. 05, p. 557-570, 2002.
- Apple Releases Details on Differential Privacy, and the Big Takeaway Is Which Emoji Is Most Popular (here)
- Differential Privacy In Action (here)