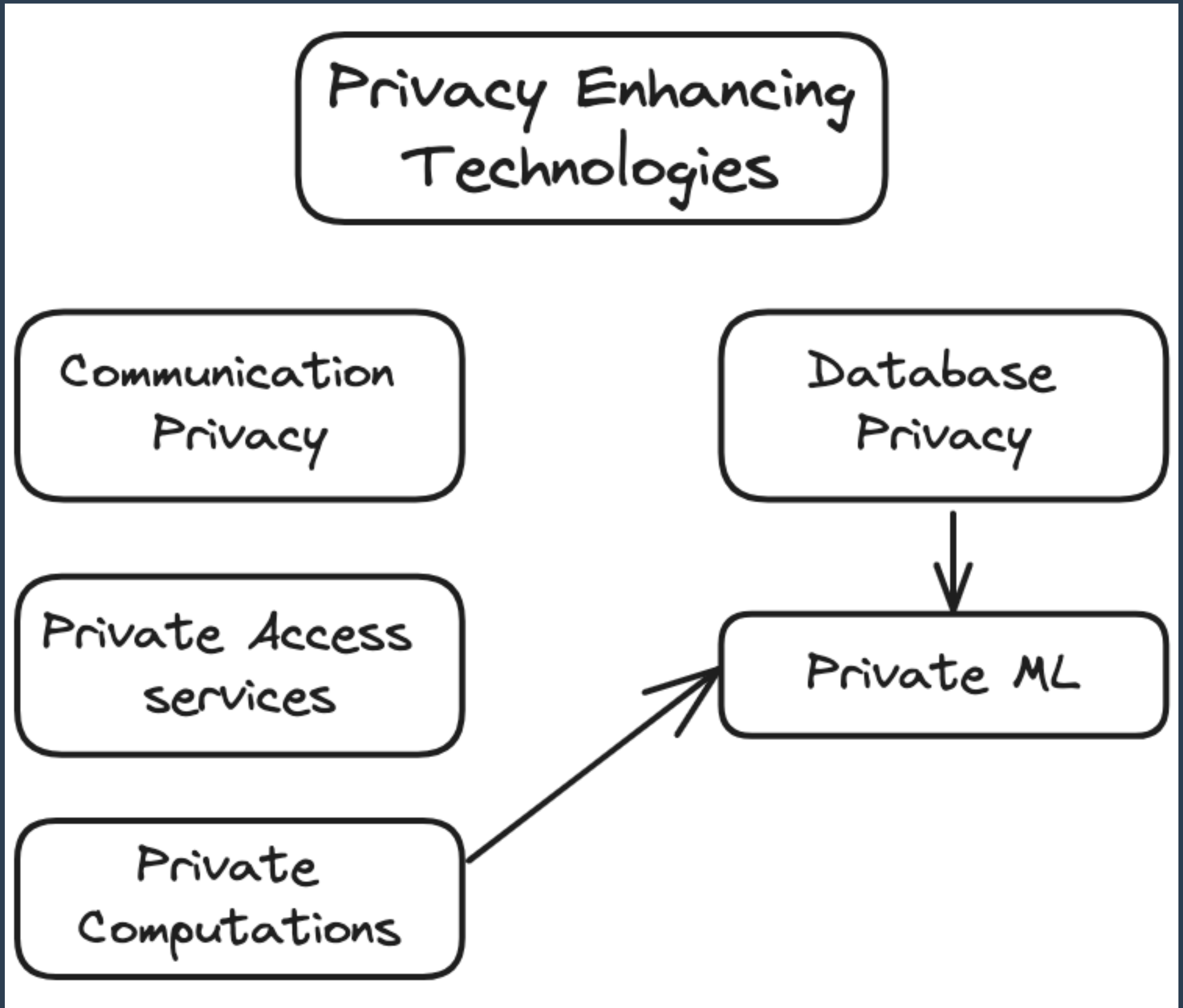
PRIVACY ENHANCING TECHNOLOGIES

Database Privacy and Private ML Training Approaches

Acknowledgement: Some of the slides in this set are adaptations of lecture slides of Dr. Olaf Hartig (Linköping University).

INTRODUCTION

- Hard Privacy
    - avoid or reduce as much as possible in placing any trust in the parties involved in serving the service to the end-user



Sign Post of Day I and Day II topics

WHOSE PRIVACY

- <u>Respondent Privacy</u>

  Protecting the information of the individuals to which the records in a database corresponds to

- <u>Owner Privacy</u>

  Protecting the information of each entities that are coming together for computing a query

- <u>End-user Privacy</u>

  Protecting end-user's queries to an interactive databases such as search engines.

STATISTICAL DATABASES

- enable its users to retrieve statistical knowledge from a subset of the population that the database represents
- exploited for variety of reasons such as disease control, market research, medical research
- we should be interested in the public availability of such data:

   results from such data can contribute to expanding our knowledge about e.g., diseases

- However, those datasets contain confidential information about the respondents who have given their information to the database
- Can the users (researchers, analysts or the data consumers) of such databases be trusted?

WHAT ARE THE PRIVACY RISKS?

- Anonymity in terms of unlinkability:
  - The anonymity of a subject w.r.t an attribute may be defined as unlinkability of this subject and this attribute [Pfitzmann17]
- Two types of linkage from an adversary's perspective;
  - Record linkage: re-identify the individual that the records in the published database corresponds to, by linking the publicly available information to the information in the published data (that is presumably free of explicit identifiers)
  - Attribute linkage: accurately infer the confidential attribute values of an individual or a set of individuals represented in the underlying database, such as inference would have been possible without the access to the data.

RECORD LINKAGE EXAMPLE

- In Massachusetts, USA, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees
- Sweeney paid $20 to buy the voter registration list for Cambridge, MA
  - Former governor (William Weld) of MA lives in Cambridge, MA hence his record is in the Voters DB
  - 6 people in Voters DB shares his DOB
  - Of which only 3 of them were men
  - Of which only 1 record matches the Weld's ZIP code.
  - Mr. Weld's medical information, learned!

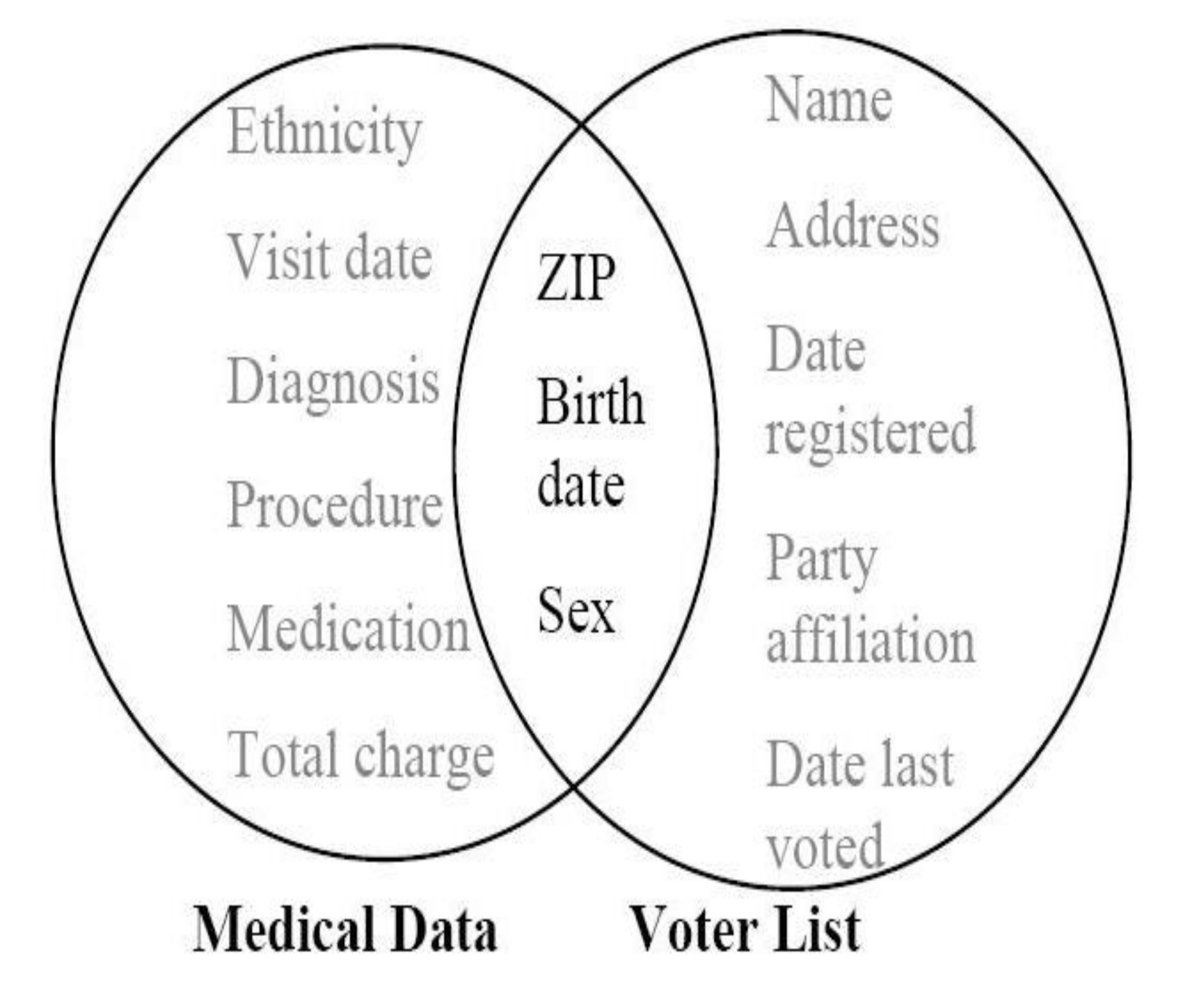


Figure taken from [Fung10]

CATEGORIES OF IDENTIFIERS

- Explicit Identifiers:
  - Attributes that unambiguously identify the respondent. E.g., name, social security number, IP address, etc.
- Quasi Identifiers:
  - A set of non-sensitive attributes that when combined may lead to unambiguously identify the respondent. E.g., gender, age, telephone number, zip code etc.
- Sensitive attributes:
  - Attributes that contain sensitive information of the respondents. E.g., disease, salary. etc.
- Non-sensitive attributes:
  - All other attributes that captures the respondents' non-sensitive information

THE CHALLENGE

- Statistical databases such as the databases of the U.S census Bureau contain confidential information such as age, sex, income, credit ratings, types of disease, etc.
- how to publish statistics about the underlying population, which is based on their confidential attributes while not revealing anything about those individual. The privacy, utility trade-off
- We need a non-trivial way to limit the disclosure of confidential information
- Fact: 87% of the US population can be identified by the combination of ZIP, DOB and sex.
- Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL)
  - limits the disclosure of confidential information from the published statistics

SDC APPROACHES CONT'D

- Let X be a table, more like a $s \times t$ matrix, with s respondents and t attributes, then
    - x_{ij} is the value of the attribute j for respondent i.
- Non-perturbative approach
    - Non-perturbative version of X is a modified version X', where X' is obtained from X by partial suppression or reduction of some details. The values represented in X' are the true values of the respondents information.

SDC APPROACHES CONT'D

- Perturbative approach
  - Data perturbation: The perturbed version X' of X such that the X' preserves the statistical information of X, such that statistics computed on X' is not significantly affected.
  - Query result perturbation: Queries are executed on the original datatable X, the results of the queries are perturbed by adding a calculated amount of random noise that is drawn from a distribution.
- Synthetic data generation approach

K-ANONYMITY

- A dataset or datable T is said to satisfy k-anonymity if each combination of values of the quasi-identifier attributes in T is shared by at least $k-1$ records.
- Let T be a table and X be a subset of the attributes of T. For every record t in T we write $t[X]$ to denote the sequence of values that t has for the attributes in X.
- Example:
  - If X = {ZIP, Age, Sex} and say t is the first tuple in T
  - then, $t[X]$ is (12211, 18, M)
  - If X = {ZIP, Sex}, then $t[X]$ is (12211, M)

12211	18	M	Arthritis
12244	19	M	Cold
12245	27	M	Heart problem
12377	27	M	Flu
12377	27	F	Arthritis
12391	34	F	Diabetes
12391	45	F	Flu

T

K-ANONYMITY DEFINITION

- Let T be a table and QI_T be the quasi-identifier of T. T satisfies k-anonymity if for every tuple t in T there exist (at least) $k-1$ other tuples $t_1, t_2, \cdots, t_{k-1}$ in T such that we have t[QI_T] = t1[QI_T] = t2[QI_T] = tk-1[QI_T].

12211	18	M	Arthritis
12244	19	M	Cold
12245	27	M	Heart problem
12377	27	M	Flu
12377	27	F	Arthritis
12391	34	F	Diabetes
12391	45	F	Flu

T

122**	18-19	M	Arthritis
122**	18-19	M	Cold
*	27	*	Heart problem
*	27	*	Flu
*	27	*	Arthritis
12391	≥ 30	F	Diabetes
12391	≥ 30	F	Flu

2-anonymous table *T**

K-ANONYMITY EXAMPLE

Chris	12211	18	M
Jack	19221	20	M

Publicly available Data

QI group / equivalence class

122**	18-19	M	Arthritis
122**	18-19	M	Cold
*	27	*	Heart problem
*	27	*	Flu
*	27	*	Arthritis
12391	≥ 30	F	Diabetes
12391	≥ 30	F	Flu

2-anonymous table *T**

Anonymized patient data

- What happens when someone attempts record linkage?

Chris	12211	18	M	Arthritis
Chris	12211	18	M	Cold

Chris is anonymous within his anonymity set

DATABASE RECONSTRUCTION ATTACK (DRA)

- It turns out k-anonymity is not sufficient against inference attacks, so what if only aggregate data is released
- But by simply observing the query answers/results of some random queries, one can recover the confidential data of the individuals in the underlying population.
- Take for example:
  - U.S census bureau database which contains answers given by the citizens of the United States
  - The census bureau publishes statistics such as how many people belonging to a race, live in a particular block
  - The attack then is to guess using brute force computation, all the possible combinations of answers that people could have given to questions concerning race and block, and find out the possible combinations that best fit the published statistics [Dinur03].

PASSWORD GUESSING ATTACKS

DATABASE RECONSTRUCTION ATTACK (DRA) EXAMPLE

Released Statistics

	Count	Mean Age	Median Age
Total Population	7	30	38
Female	4	30	33.5
Professors	4	51	48.5
Married Adults	4	51	53
Female professors	3	35	35.6

Example taken from "Protecting privacy with math"

DATABASE RECONSTRUCTION ATTACK (DRA) CONT'D

Possible Ages for Mean 35 and Median 35.6

Female_ prof1	Female_prof2	Female_prof3
1	36	73
2	36	72
3	36	71
	…	
6	36	68
	…	
35	36	39
36	36	38

DATABASE RECONSTRUCTION ATTACK (DRA) CONT’D

Possible Ages for Mean 35 and Median 35.6

Female_prof1	Female_prof 2	Female_prof 3
34	36	40
35	36	39
36	36	38

✓

Female_prof 1	Female_prof 2	Female_prof 3
6	36	68
7	36	67
8	36	66

✗

A WAY TO PRIVACY

- Publishing less statistics, then there are more plausible combinations of data that accurately fits the data
- Even lesser statistics are published means, increase in the amount of data combinations that plausibly fit the released statistics.

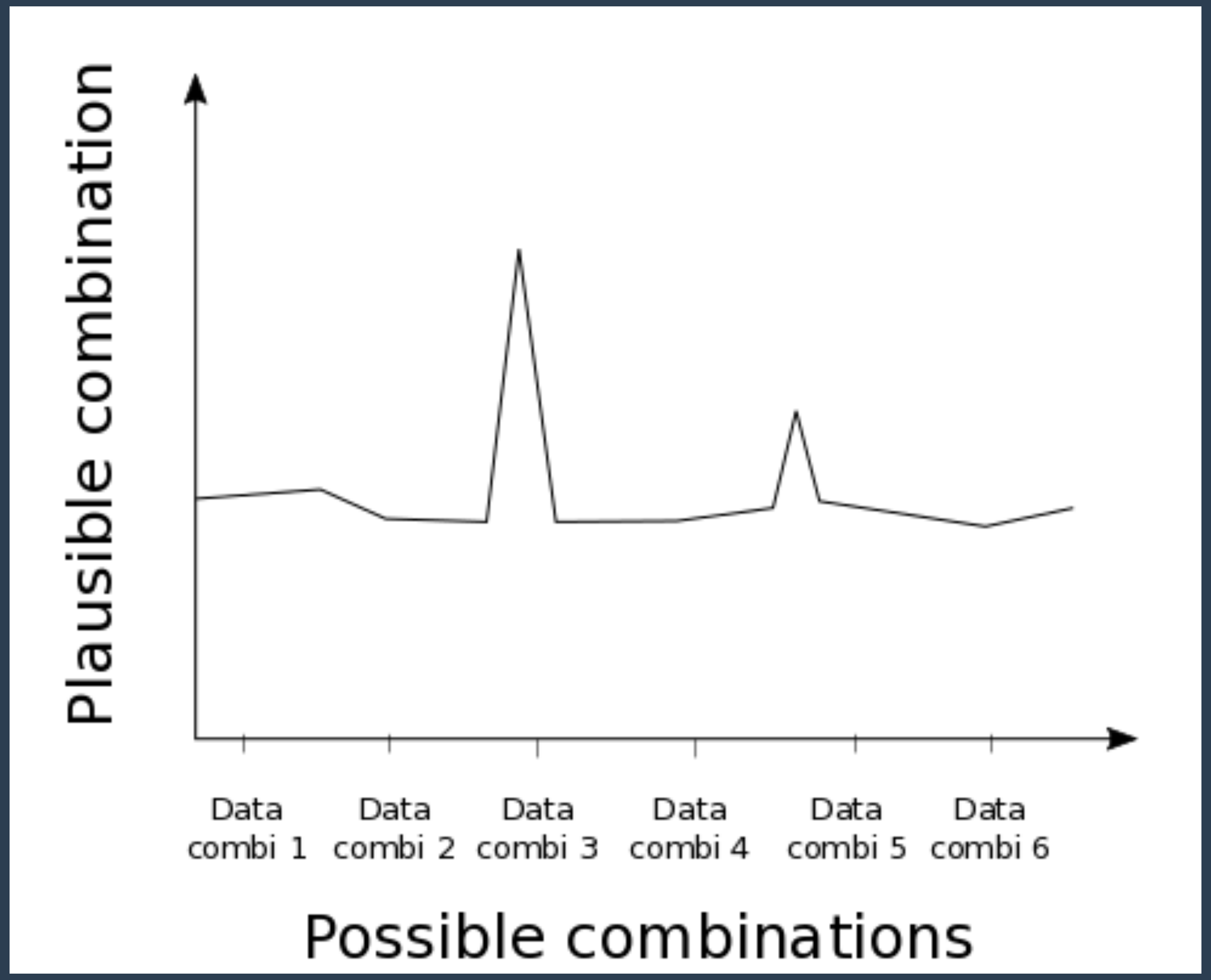
MEASURE OF PRIVACY

- Observations from the above example,

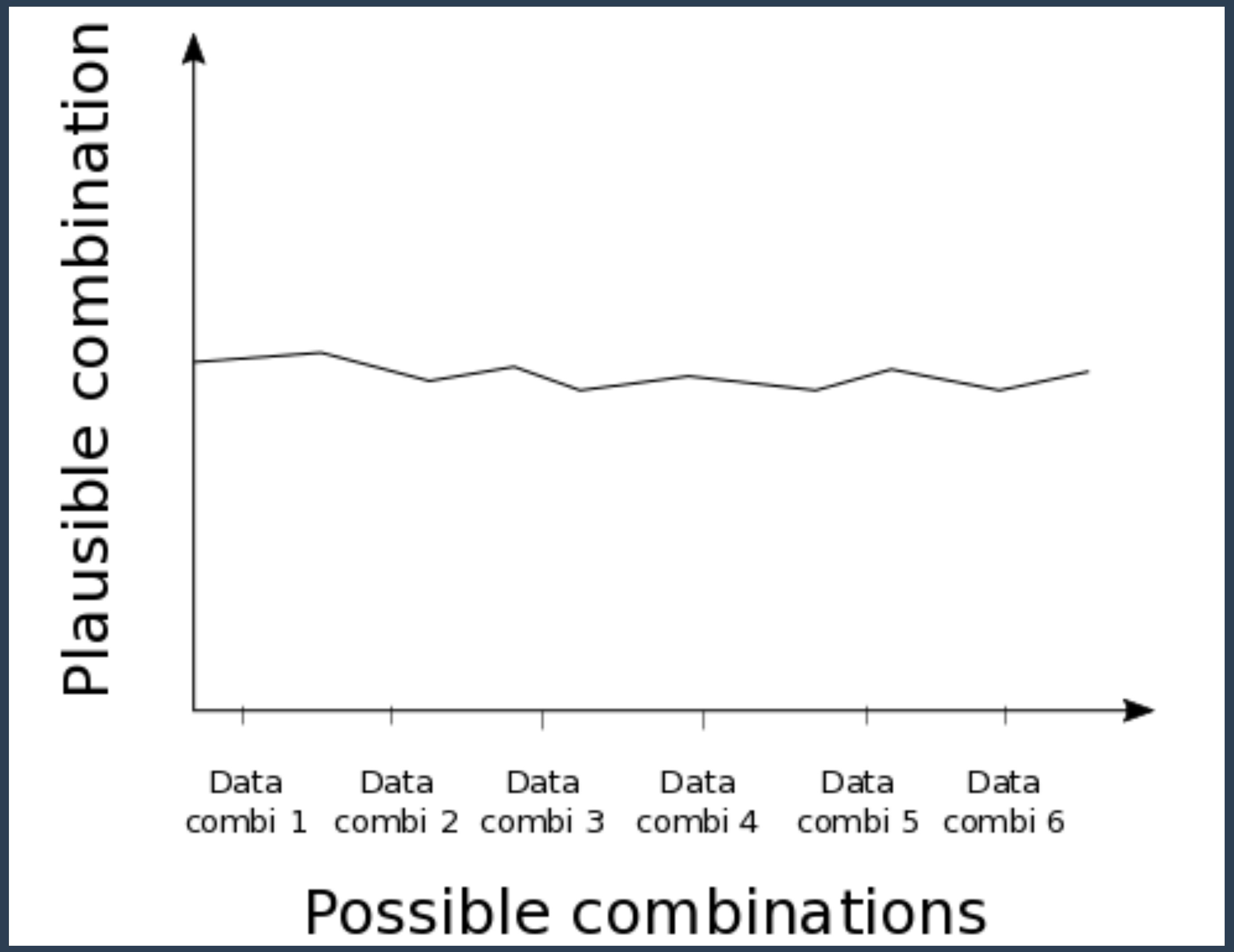
- measure of loss of respondent privacy is the level of certainty in an attacker’s ability in determining the plausibility of some possible combinations of data.

- Idea! to protect respondent privacy - make all possible combinations of data from the respondents to be equally plausible.

- There is an inevitable trade-off between accuracy of the published results and not revealing information of the record owners in the underlying database.



A few possible data combinations are plausible



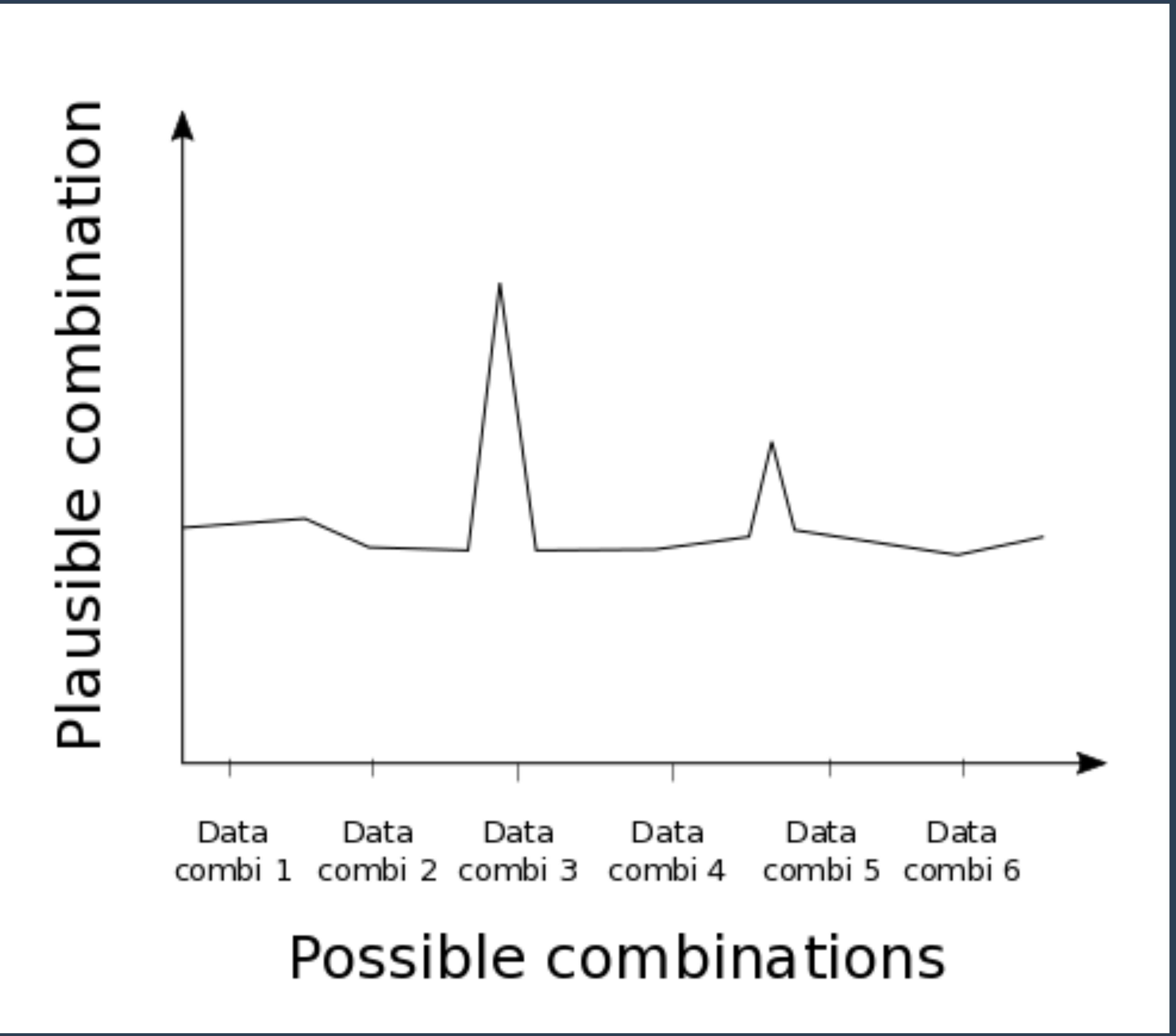
All possible data combinations are plausible

DIFFERENTIAL PRIVACY

- How then to publish data for data analyses?
- because increasing the uncertainty level of the adversaries, decreases the query results' accuracy
- Further, if random noise is added a bunch of times to a statistical query result, it is possible to get back the true results by taking the average of the noisy results, which cancels out the noise.
- Differential privacy model that provides a **strong** privacy guarantee, yet at the cost of small loss in the accuracy of the results.

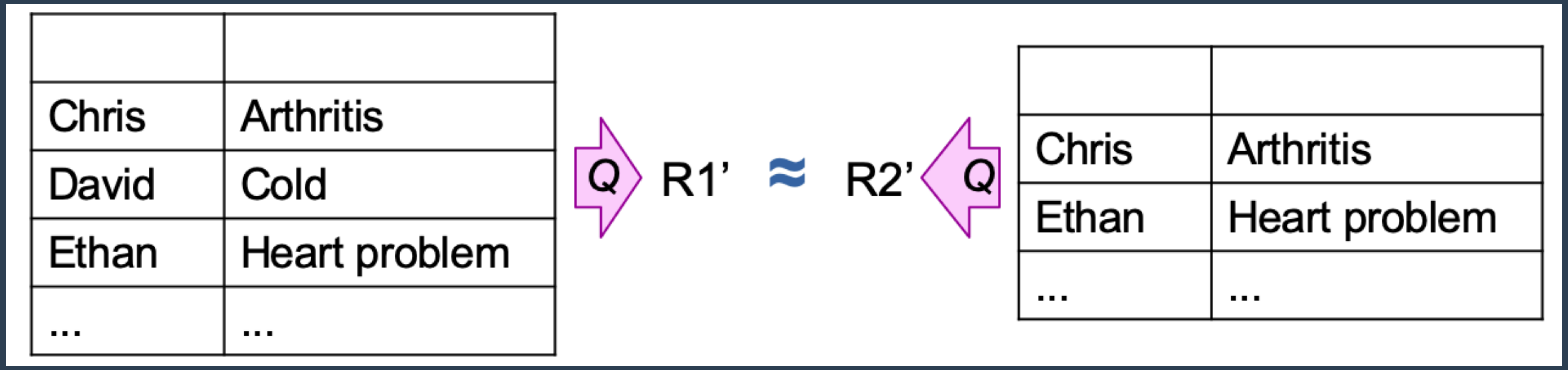
DIFFERENTIAL PRIVACY

- The differential privacy model provides a way to quantifies the plausibility peak (i.e the loss of privacy) and bounds (that is to say the maximum) the loss of privacy for the individuals in the underlying dataset, as a consequence of publishing results computed on their data.



The plausibility/possibility plot with a few peaks that stands out

DIFFERENTIAL PRIVACY EXAMPLE



- Statistical Query: How many persons with a cold?, the answers from a differentially private computation will "nearly" be the same whether or not David is in the underlying database.
- Observation:
  - The two databases where one contains David's data and the other do not contain his data - database neighbors. Generally speaking, any two databases D and D', which differ by at most one record but otherwise contain the same records are called database neighbors.
  - The results of the query over D and D' doesn't look the same, what it means here is that the probability distributions of the query result are the same. So, the likelihood of getting answer 1 when database is D is the same likelihood for getting answer 1 from D'.

DIFFERENTIAL PRIVACY FORMAL DEFINITION

- Differential Privacy [Dwork06]:
  - A randomized query mechanism M_Q for query Q provides ε-differential privacy if
    - if for all databases D and D', where D and D' are database neighbors and
    - every subset O of the set of all possible outputs of M_Q,
- We have that: $Pr[M_Q(D)\ in\ O] \leq e^{\varepsilon} \cdot Pr[M_Q(D')\ in\ O]$

DIFFERENTIAL PRIVACY FORMAL DEFINITION CONT'D

- Observation:
- **Epsilon** is the measure of peak that stand out in the plausibility plot (is the measure of information gain in adversaries ability to confidently choose one combination of data over the other), and the above definition bounds the loss of privacy from releasing the query results.
- **Composition** The future releases also guarantee ε-differential privacy
  - if we publish the count of persons with cold with $\varepsilon = 3$ and publish the average age of persons with $\varepsilon = 3$, then the total privacy loss caused from the release of the two statistics is at most 6.

ACHIEVING DIFFERENTIAL PRIVACY BY ADDING NOISE

- Assume a query Q whose result $Q(D)$ over any possible database instance D is a real number
- Randomized query mechanism $M(Q)$ for Q, adds randomly selected noise η
  - $M(Q) = Q(D) + \eta$
- Observation : the amount of noise depends both on ε and the sensitivity of the query being asked.
- The sensitivity of the query is a constant that captures the amount of maximum change any one individual may cause to the result of the query. Take our "how many persons with cold example, adding or removing a record will change the query result by at most a factor of 1.
- Less the epsilon, stronger the privacy

QUERY OR FUNCTION SENSITIVITY - THE FORMAL DEFINITION

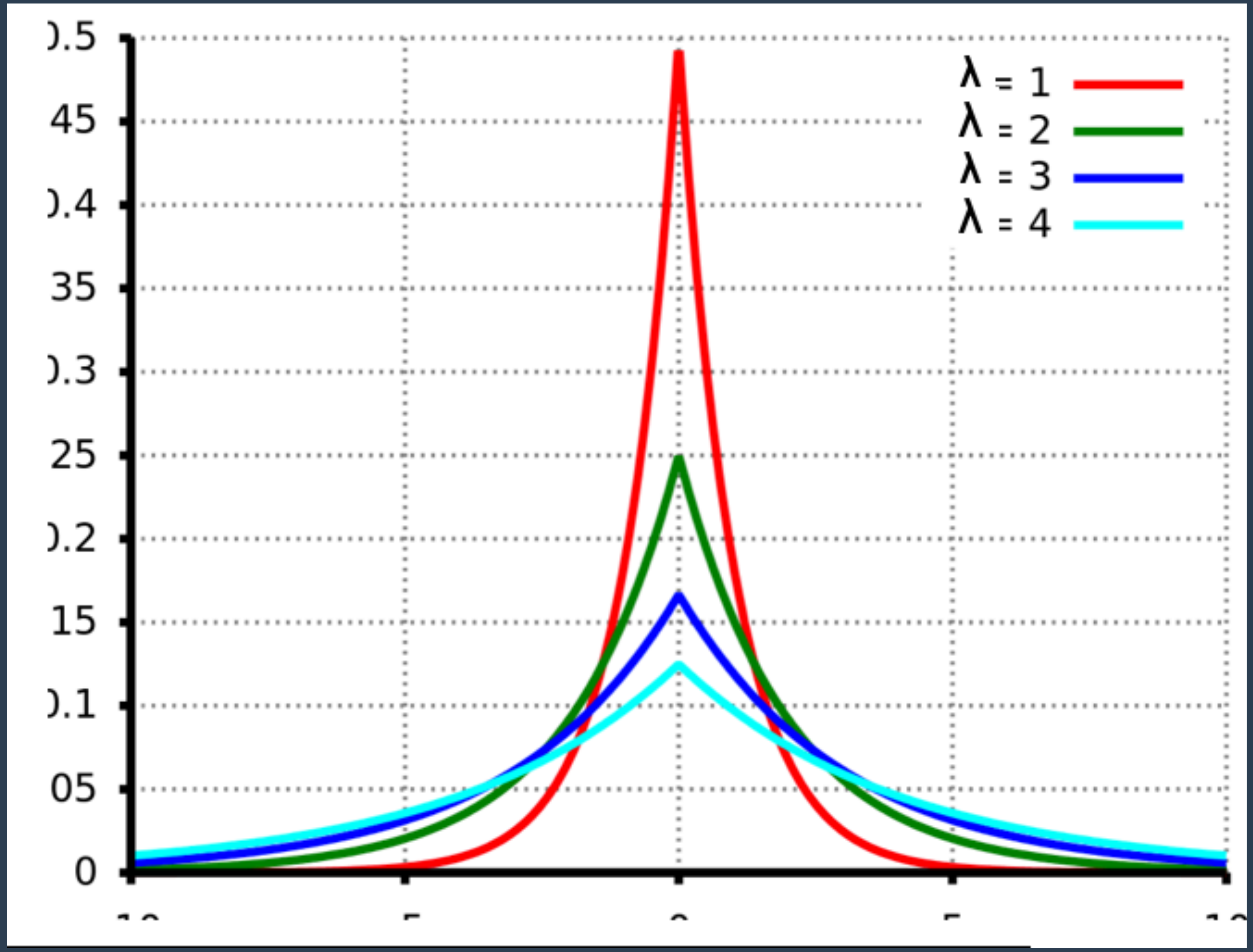
- Definition: The sensitivity of a query Q is
  - $\Delta q = max\,|\,Q(D) - Q(D')\,|$
  - for any two neighboring databases D and D'
- Examples:

- Δq for "count all patients diagnosed with cold" is: 1

LAPLACE MECHANISM TO DIFFERENTIAL PRIVACY

- Idea: The noise to be added is drawn from the Laplace distribution Lap(λ), λ determines how flat the curve of the distribution is, from where the noise is drawn.

- Theorem [Dwork 2006]: Let M_Q be a mechanism for Q that returns $Q(D)$ + η where η is drawn randomly from Lap(λ) with λ = Δq / ε. M_Q provides ε-differential privacy



Laplace distributions of varying scales from 1 to 4
the scale of the distribution depends on epsilon and Δq

Picture sources: https://commons.wikimedia.org/wiki/File:Laplace-verteilung.svg

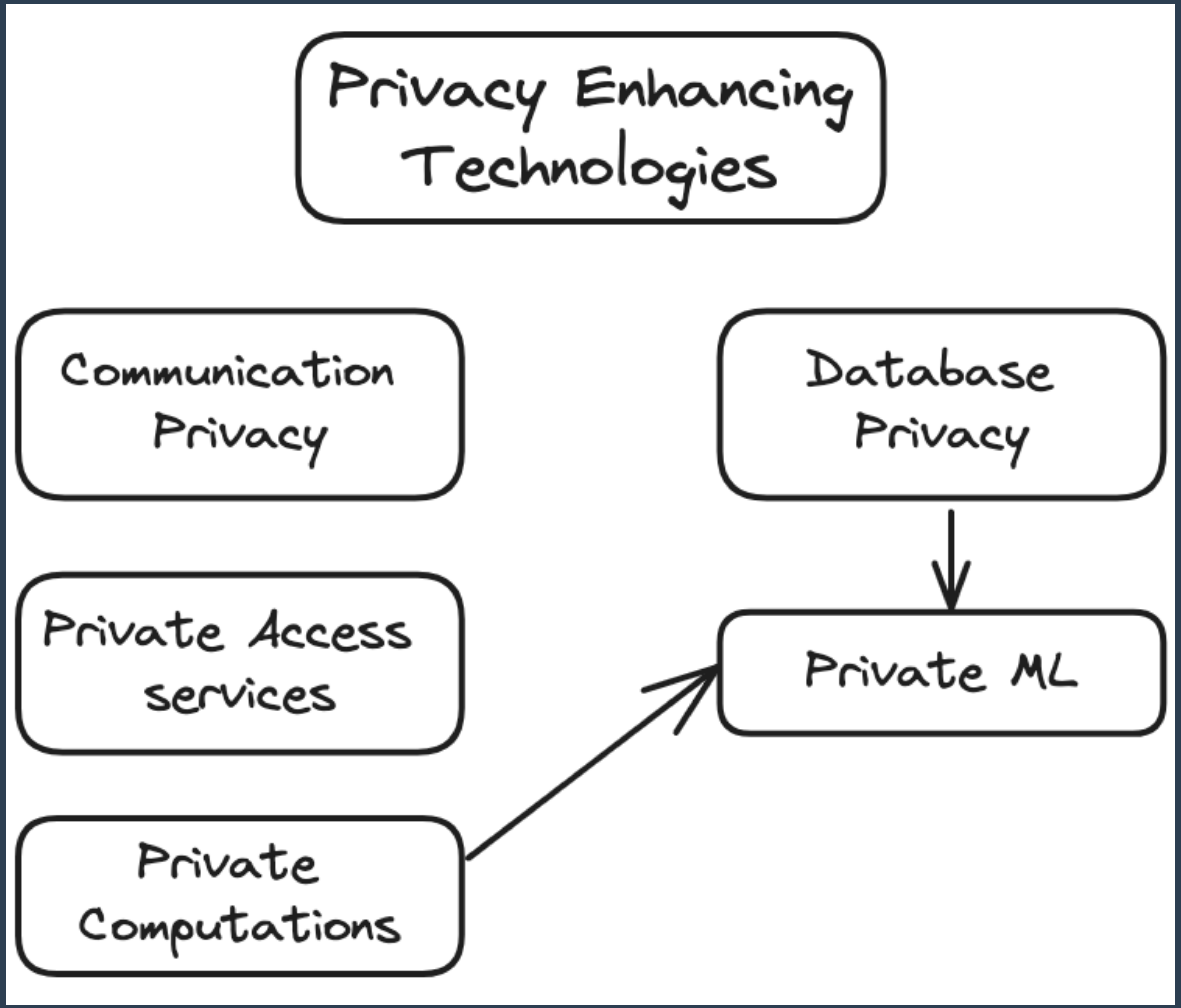
LAPLACE MECHANISM TO DIFFERENTIAL PRIVACY

Observations

- The narrow the curve (Laplace distribution), the value drawn as noise is small, which implies the result of the query is changed by a small amount, narrow curve is good for accuracy.
- However, for $\Delta q = 1$ and $\varepsilon = 0.1$, we have $\lambda = 10$ (and $\lambda = 100$ if $\varepsilon = 0.01$)
- Hence, for queries with higher sensitivity Δq, we have a higher value of λ thus, the noise η will typically be higher
- Likewise, for a smaller value of ε, the noise will be typically higher

- Given a sequence $Q_1, \cdots, Q_m$, ε-differential privacy can be achieved by drawing the noise for Q_m from Lap(λ_m) where λ_m is the sum of all $\lambda_i = \Delta q_i / \varepsilon$ $(i = 1, \cdots, m)$
  - Observation: The magnitude of the amount of noise added increases with every query.

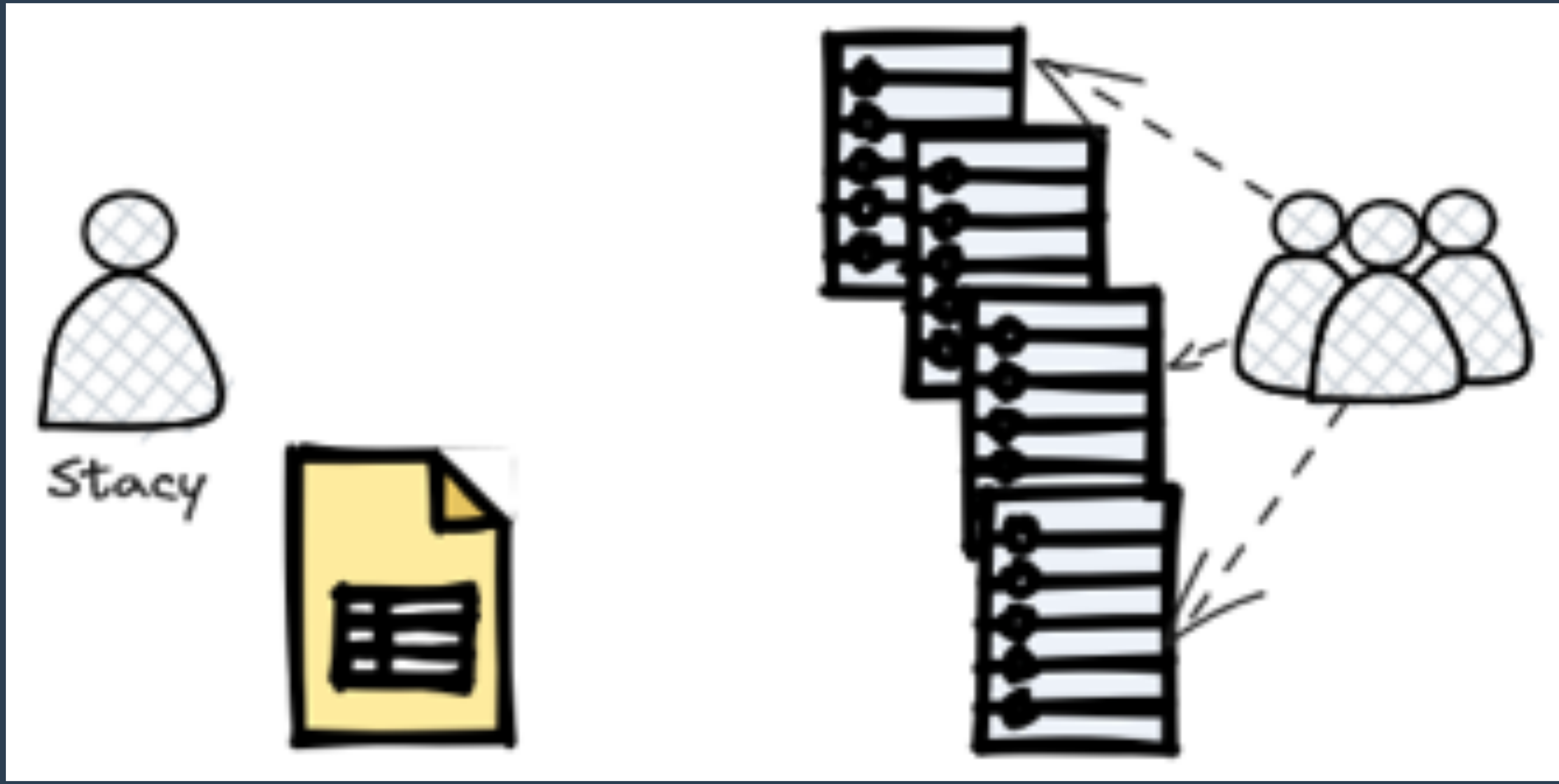
- Theorem [Dwork 2006]: Let M_Q be a mechanism for Q that returns $Q(D) + \eta^k$ where η^k is a vector of size k whose elements are independently drawn randomly from Lap(λ) with $\lambda = \Delta q / \varepsilon$. M_Q provides ε-differential privacy

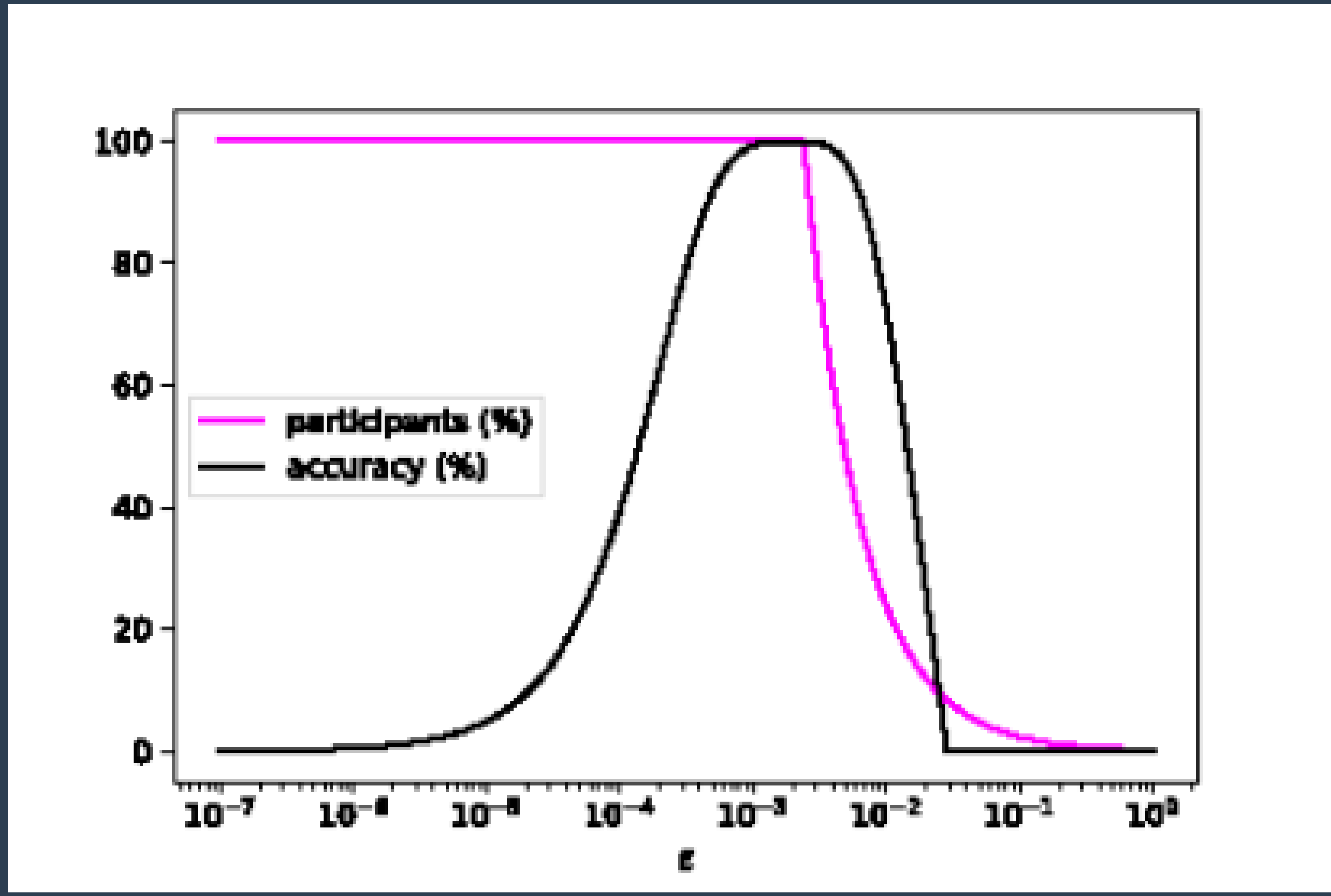


Sign Post of Day I and Day II topics

WHY PRIVACY PRESERVING MACHINE LEARNING

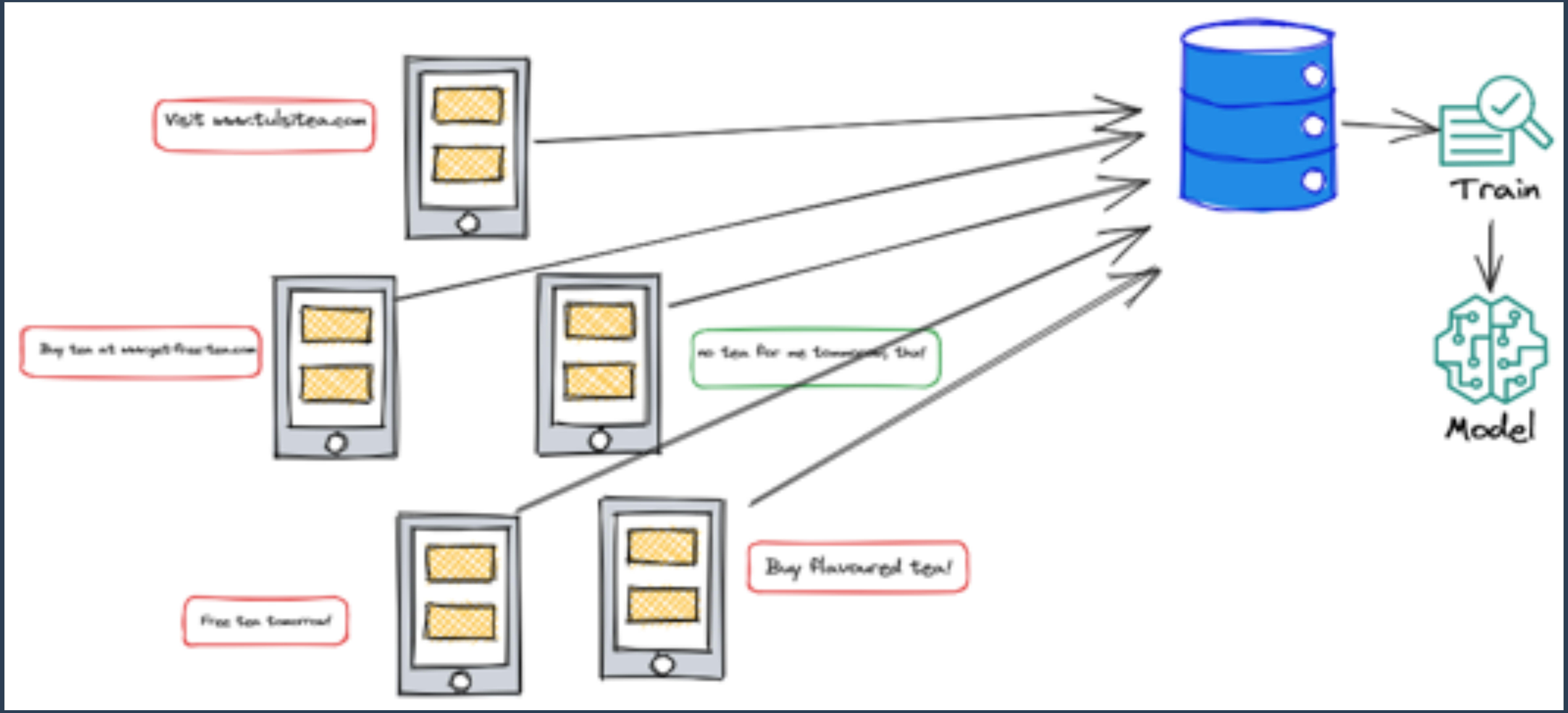
- "data is food for AI" - Andrew Ng
  - Privacy improves data quality and quantity





WHY FEDERATED MACHINE LEARNING

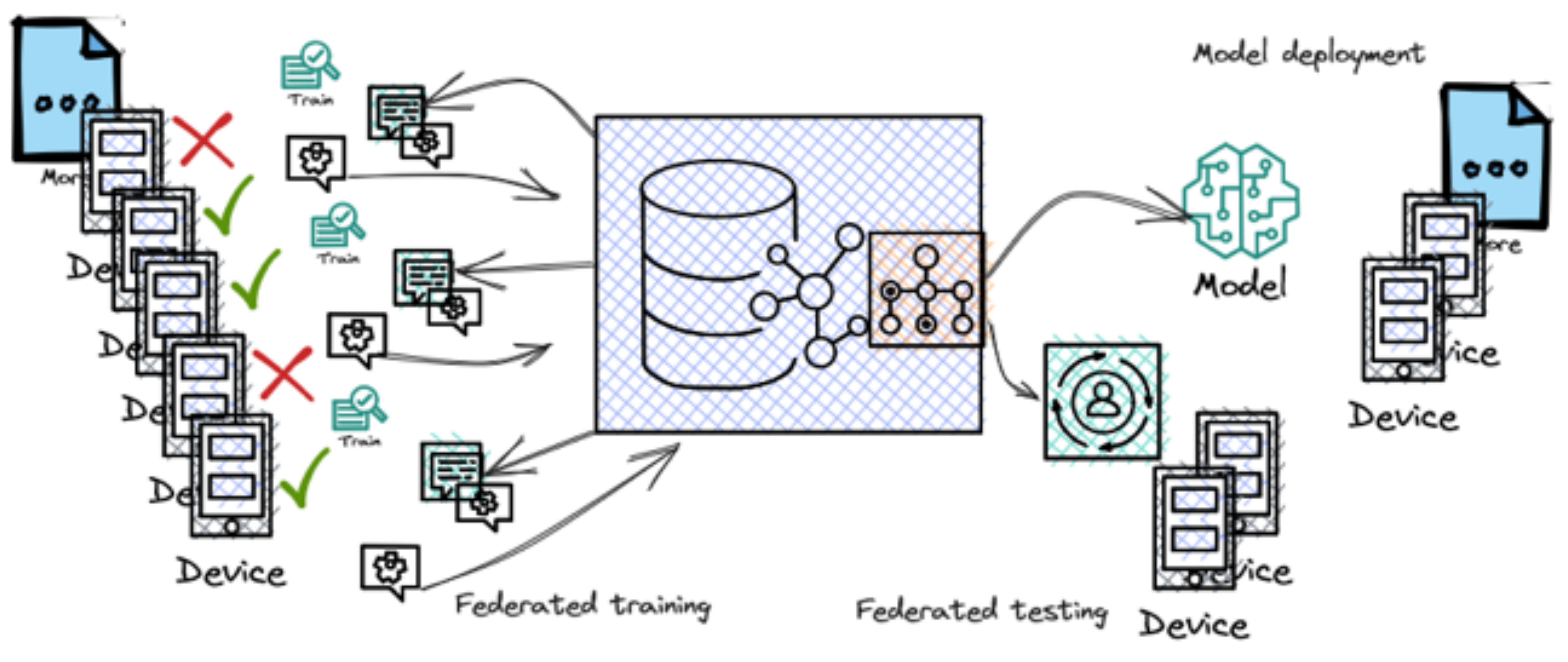
- Let’s say one want to train a spam model for their spam app. The figure shows centralised ML model training, which leads to training data leakage risks.



FEDERATED MACHINE LEARNING

- Federated learning is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead focused updates intended for immediate aggregation are used to achieve the learning objective.
- The idea is to separate the data and the training, separate the computation and communication
  - the data never leaves user devices
  - Only sample of devices get selected to whom the training models are pushed
  - The training model (from the beginning there is a global model but the weights are not set) is sent to the devices for training and the locally trained models (gradients) are sent back to the server. (see the figure in the next slide)

FEDERATED TRAINING AND ANALYTICS



FEDERATED LEARNING VARIANTS

- Cross-devices federated learning
  - Large number of IoT or mobile devices
  - Each clients stores its own data
  - Central server/service provider orchestrates the training
  - Random selection of eligible clients
  - Stateless clients meaning typically each client participate only once
  - Fixed partition by training samples (horizontal)
  - Primary bottleneck: unreliable communication

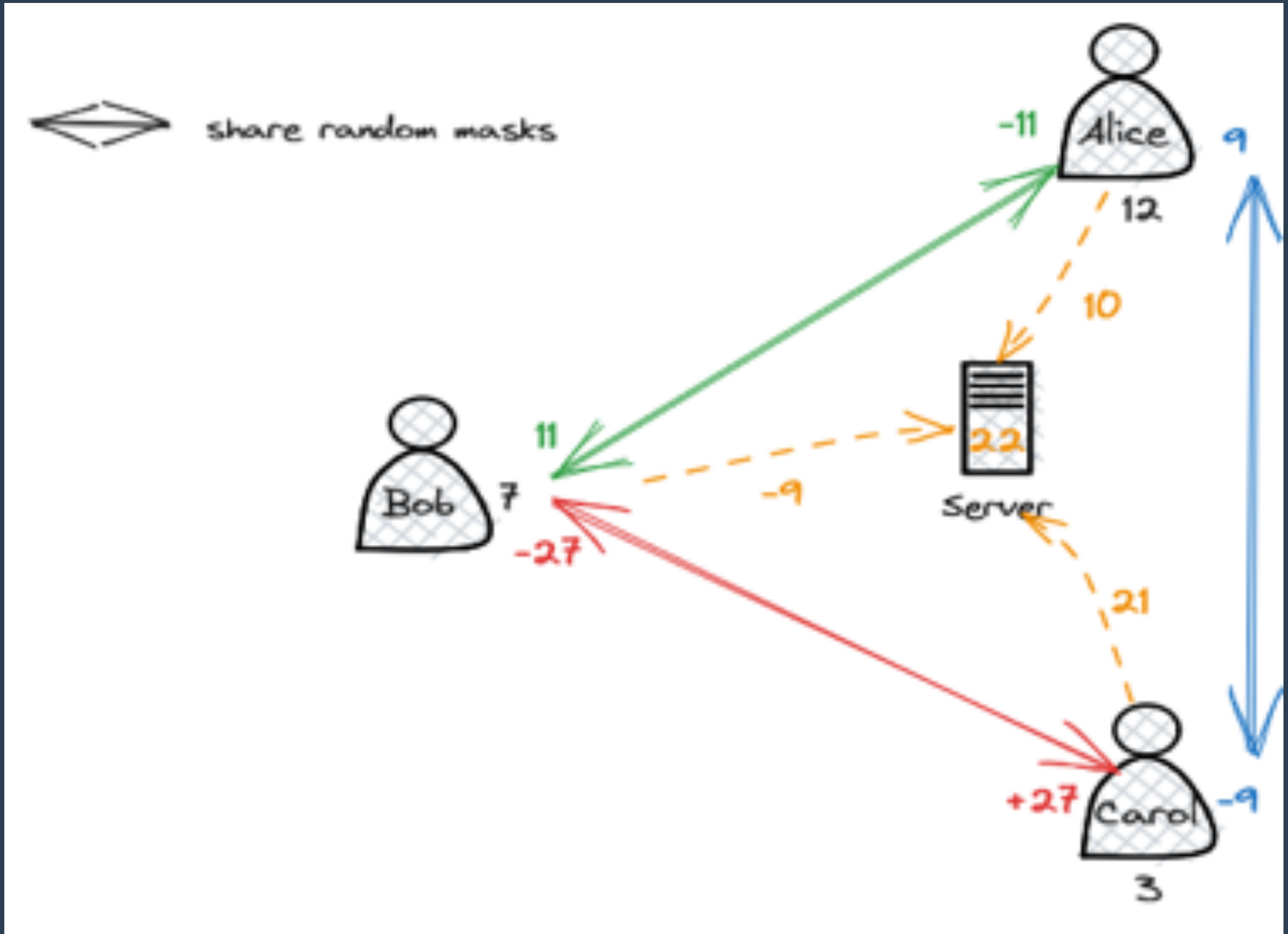
FEDERATED LEARNING VARIANTS

- Cross-silo federated learning
  - Few reliable data sources such as different banks, hospitals
  - Data silos and remains decentralized
  - Central service orchestrates the training but no data is stored elsewhere
  - Clients are always available and participates in each round of computation
  - Fixed partition either by training samples (horizontal) or by feature space (vertical)

SECURE AGGREGATION - MOTIVATION

- Federated learning limits data exposure, however can it be possible to reconstruct training data from the individual models weights uploaded to the server? Like Shokri et el.'s membership inference attack.
- Separation of aggregate function and access to data
- Using secure aggregation, before anything is sent out from the device the protocol adds zero-sum masks to scramble the training results. When one add up all those model parameters the masks cancel out.

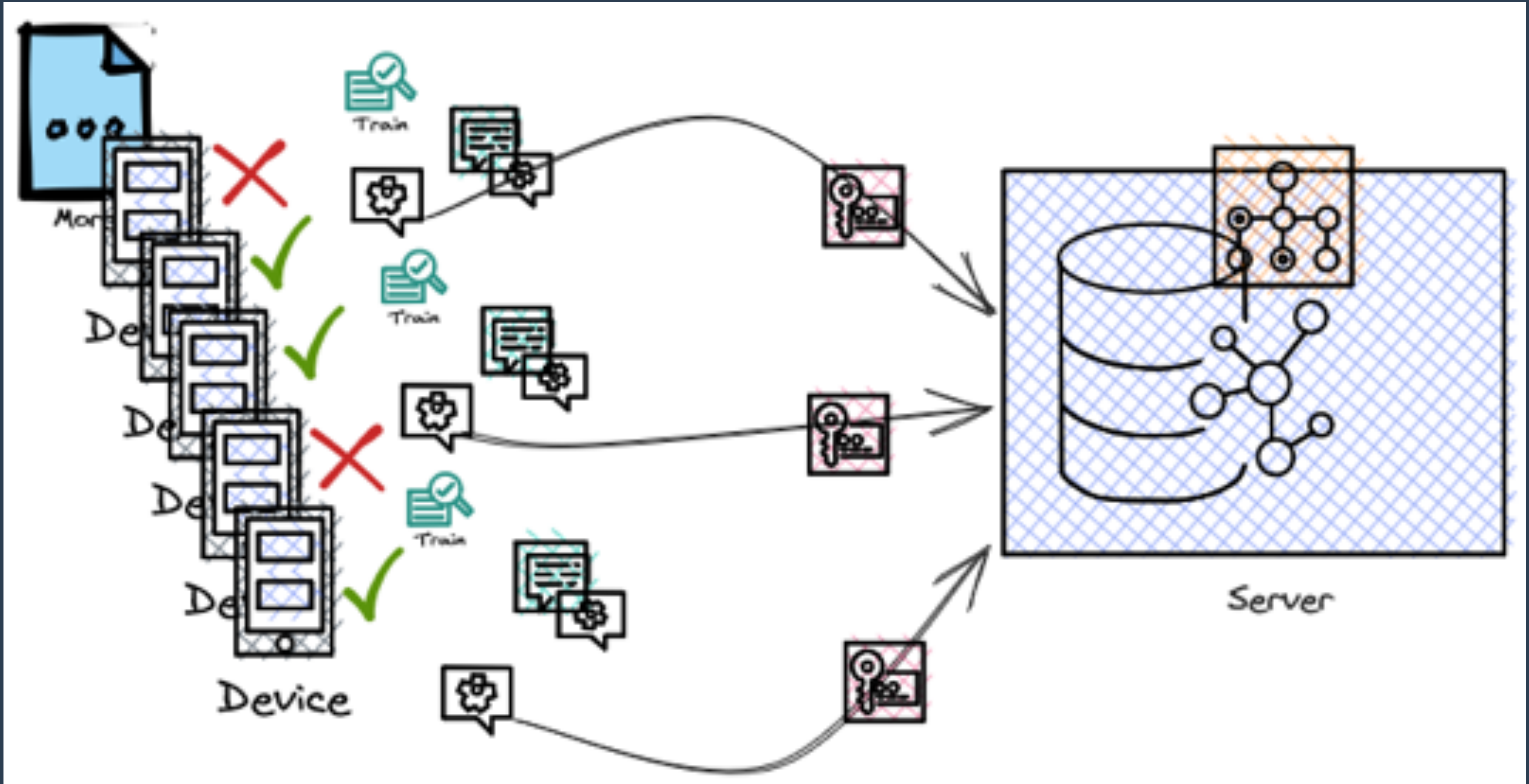
SECURE AGGREGATION - TOY EXAMPLE



SECURE FEDERATED LEARNING COMPUTATION

- goal of FL computation - to evaluate a function f on a distributed client dataset.
- Secure goal of FL computation - only the results of the function evaluation is revealed to the server with out revealing each client's inputs and the server does not have the key to decrypt the client's inputs.
- Achieved using secure Multi-Party Computation (MPC) technologies - common ones are secure aggregation via additive masking and via threshold homomorphic encryption.

FEDERATED LEARNING WITH SECURE AGGREGATION

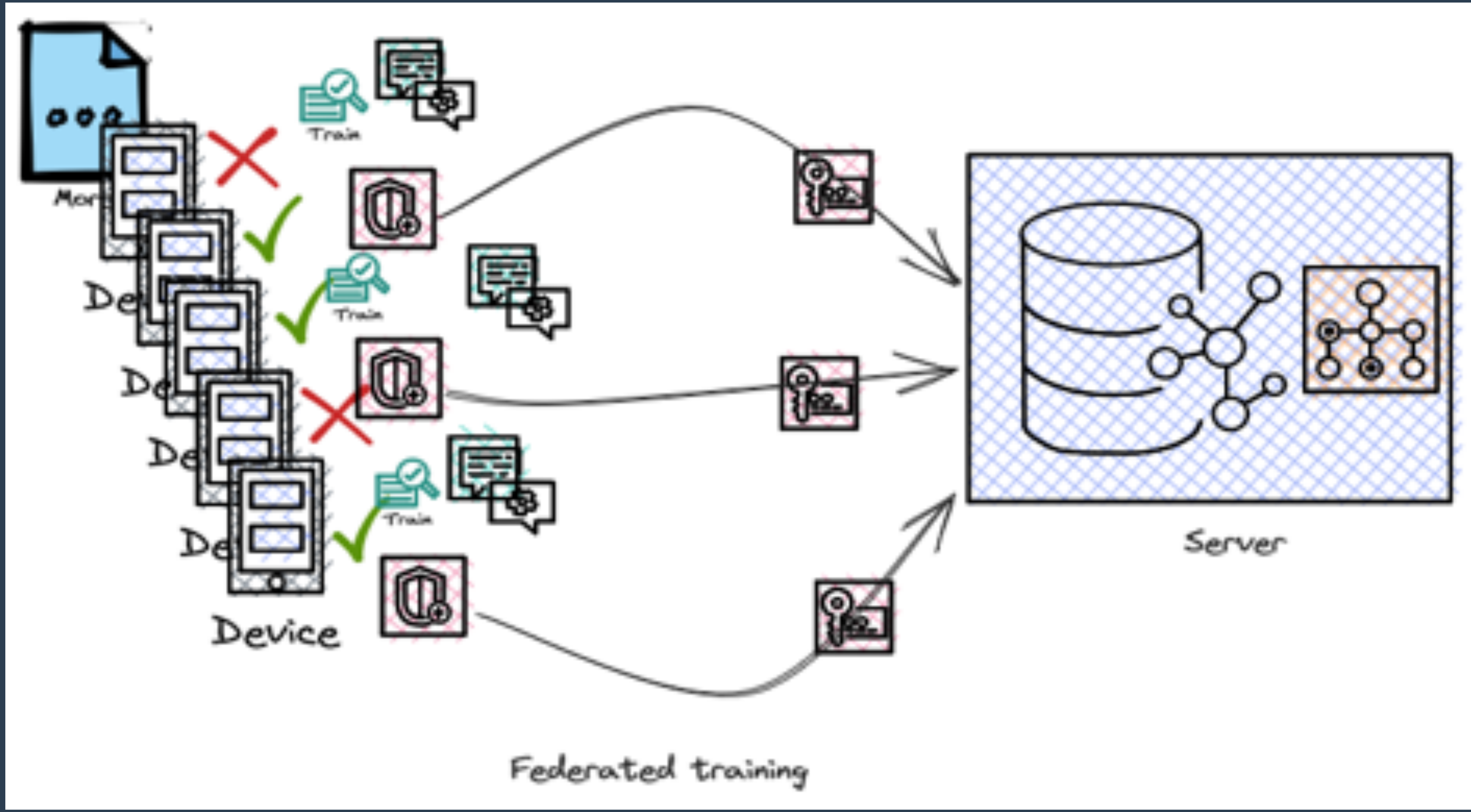


DIFFERENTIAL PRIVACY - MOTIVATION

- The model updates are secured via secure aggregation - are the users personal data safe?
- What if one or few clients reports a significantly different model update from others because of their unique phone usage data, is there a risk to privacy? Check Fredrikson et al.
- Thanks to Differentially Privacy.
  - Limit the contribution of how much any one client can contribute and obscure the locally trained model or model updates.

FL WITH DIFFERENTIAL PRIVACY AND SECURE AGGREGATION

Each device before sending the model weights to the server, perturbate the weights such that local differential privacy is guaranteed. Then the perturbed model updates are further secured through the secure aggregation technique.



REFERENCES

[Pfitzmann17] - A Terminology for Talking about Privacy by Data Minimization, 2017

[Dwork06] - C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating Noise to Sensitivity in Private Data Analysis, 2006

[Dwork13] - C. Dwork, A. Roth, The Algorithmic Foundations of Differential Privacy, 2013

[Warren1890] - S. Warren, L. Brandeis, The right to privacy, Harvard Law Review, 1890

[Agre98] - P. Agre, M. Rotenberg, Technology and Privacy, 1998

[Dinur03] - I. Dinur, K. Nissim, Revealing information while preserving privacy, in Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2003

[Mcsherry07] - F. McSherry, K. Talware Mechanism Design via Differential Privacy

[Fung10] Privacy-Preserving Data Publishing: A Survey of Recent Developments

[Shokri16] Membership Inference Attacks against Machine Learning Models, https://arxiv.org/abs/1610.05820.

[Fredrikson] Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, https://rist.tech.cornell.edu/papers/mi-ccs.pdf