Differential privacy can also enable scientific collaboration

Anand D. Sarwate, Rutgers University

15 April 2024

IEEE Information Theory Society Distinguished Lecture
SKKU Com&Net Group Seminar Series

Everyone should have privacy!

We all want it but… what is it?

“Perhaps the most striking thing about the right to privacy is that nobody seems to have any clear idea what it is.”

Judith Jarvis Thomson, *The Right to Privacy,* Philosophy & Public Affairs 4(4), 1975.

Photo: MIT News

US legal scholar Daniel J. Solve identifies **at least 6 different legal meanings of privacy** in US law:

- A “right to be left alone” (no photos)
- The right to limit access to myself (locks)
- Information secrecy
- Control over how information is used
- “Personhood”
- Decision-making about myself

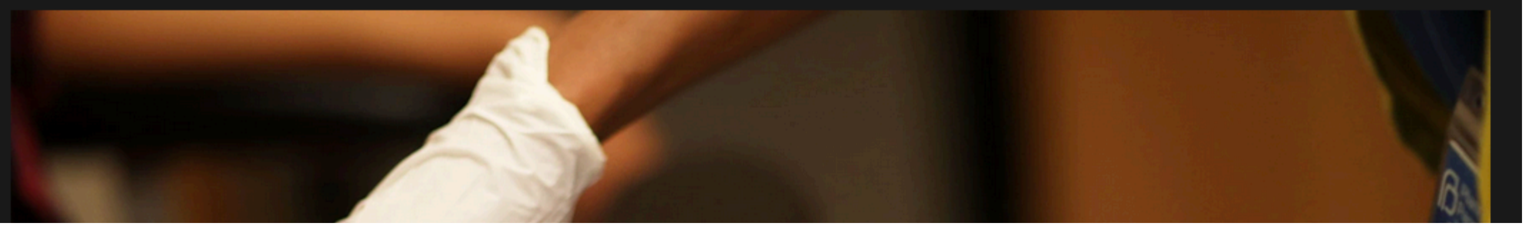
The cost of privacy loss

We see examples all the time

BAY AREA

Patient's 'embarrassing' private health information posted to Facebook after Contra Costa County medical privacy breach

by: Tori Gaines
Posted: Mar 15, 2023 / 03:10 PM PDT
Updated: Mar 16, 2023 / 05:42 AM PDT



The New York Times

Data Breach Could Compromise Lawmakers' Personal Information

A cyberattack on the District of Columbia's online health insurance marketplace may have compromised identifying data of many members of Congress and other users.

Give this article

ALJAZEERA News Ukraine war Features Economy Opinion

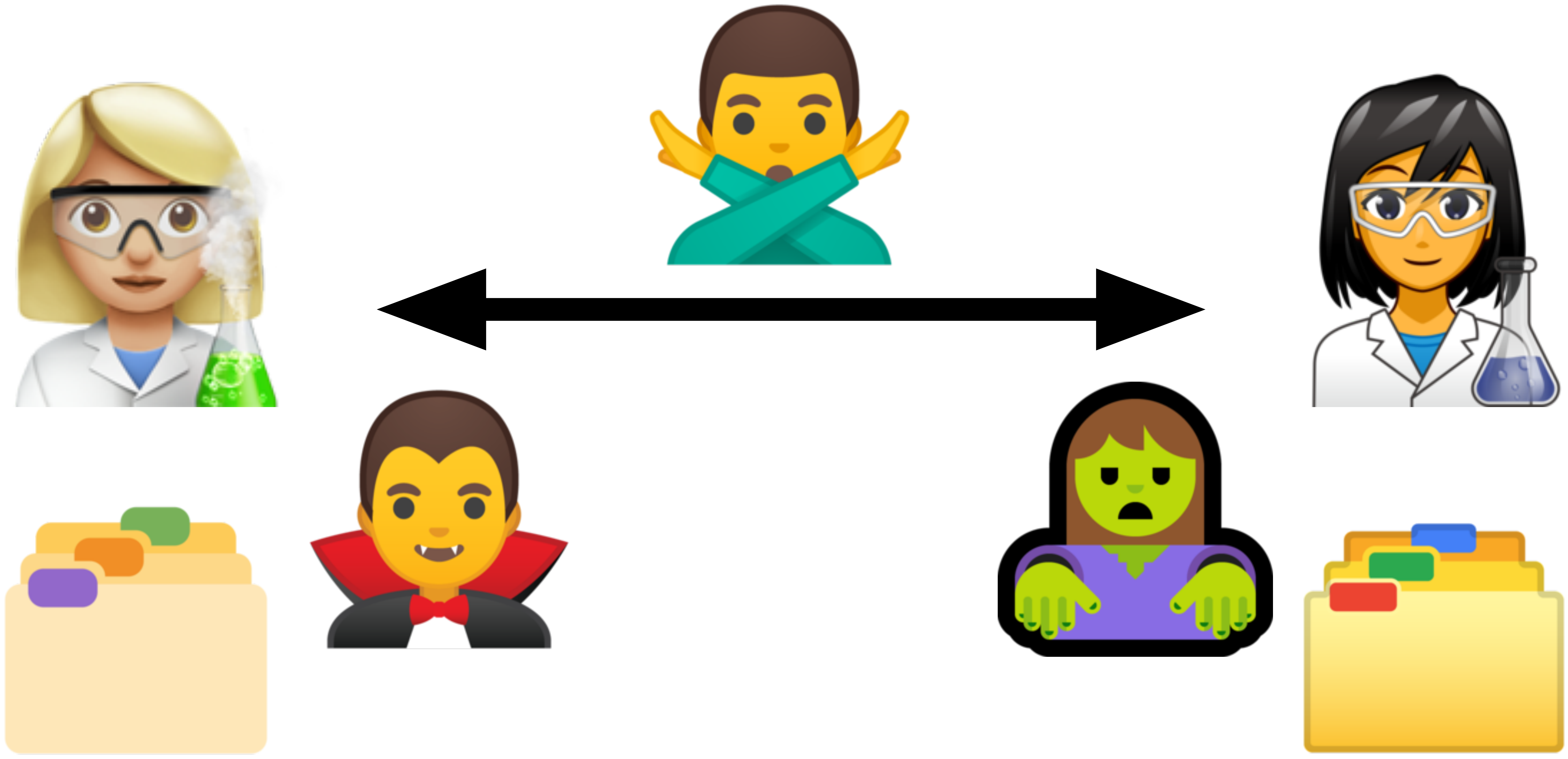
News | Technology

ChatGPT owner OpenAI fixes bug that exposed users' chat histories

According to reports, the titles of the conversations were visible but the substance of other users' conversations was not.

Our motivation: biomedical research

Joint analyses can make a huge difference, but are they safe?



A Case Study

Trying to enable collaboration

- **Goal:** platform for researchers to create consortia for federated analysis of neuroimaging data.
- **Algorithms:** preprocessing, feature discovery (PCA, ICA, NMF, DNNs), prediction, visualization and quality control, etc.
- **Challenge:** small sample size, high dimension, domain-specific algorithms.

TReNDS Translational Research in Neuroimaging & Data Science

https://trendscenter.org/

COINSTAC

https://coinstac.org/

What this talk is about

From privacy basics to private federated learning

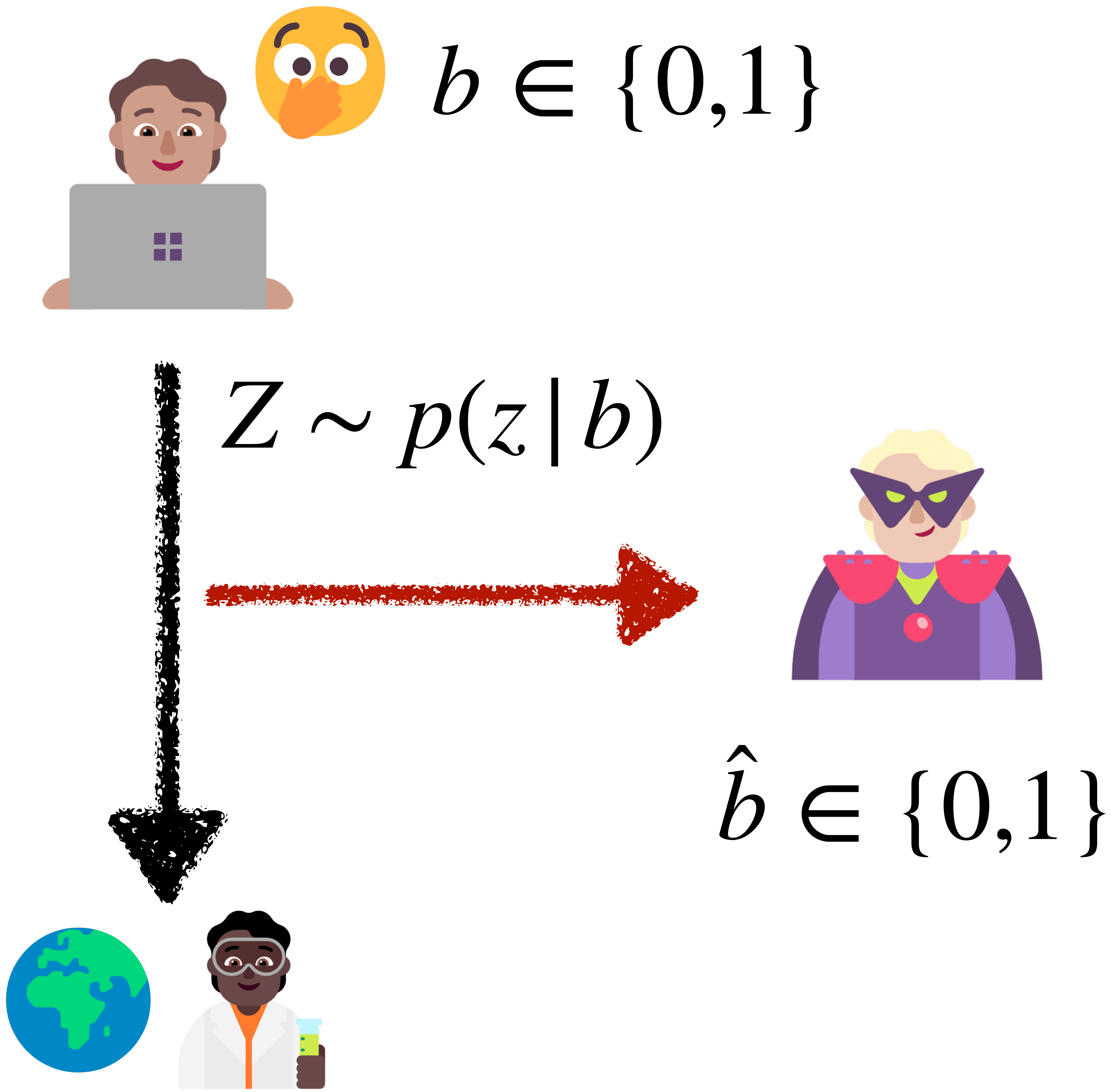
We will start from the basics:

- How does information theory let us understand privacy, and particular **differential privacy (DP)**?
- How do we protect privacy when doing **machine learning and statistics**?
- What challenges and opportunities arise when **working with federated data**?
- How can this help in **collaborative science**?

An IT perspective on DP

Modeling private information: a binary secret

Let's start simple



Suppose we have a single bit $b \in \{0,1\}$ of private information.

Some information Z which depends on b gets leaked (or published) and is observed by an adversary.

The privacy question: How much does Z reveal about b?

This is a hypothesis testing problem!

Time to dust off your notes from detection and estimation...

This **privacy question** is a **hypothesis testing question**:

$$\mathcal{H}_0 : Z \sim p(z \mid 0)$$

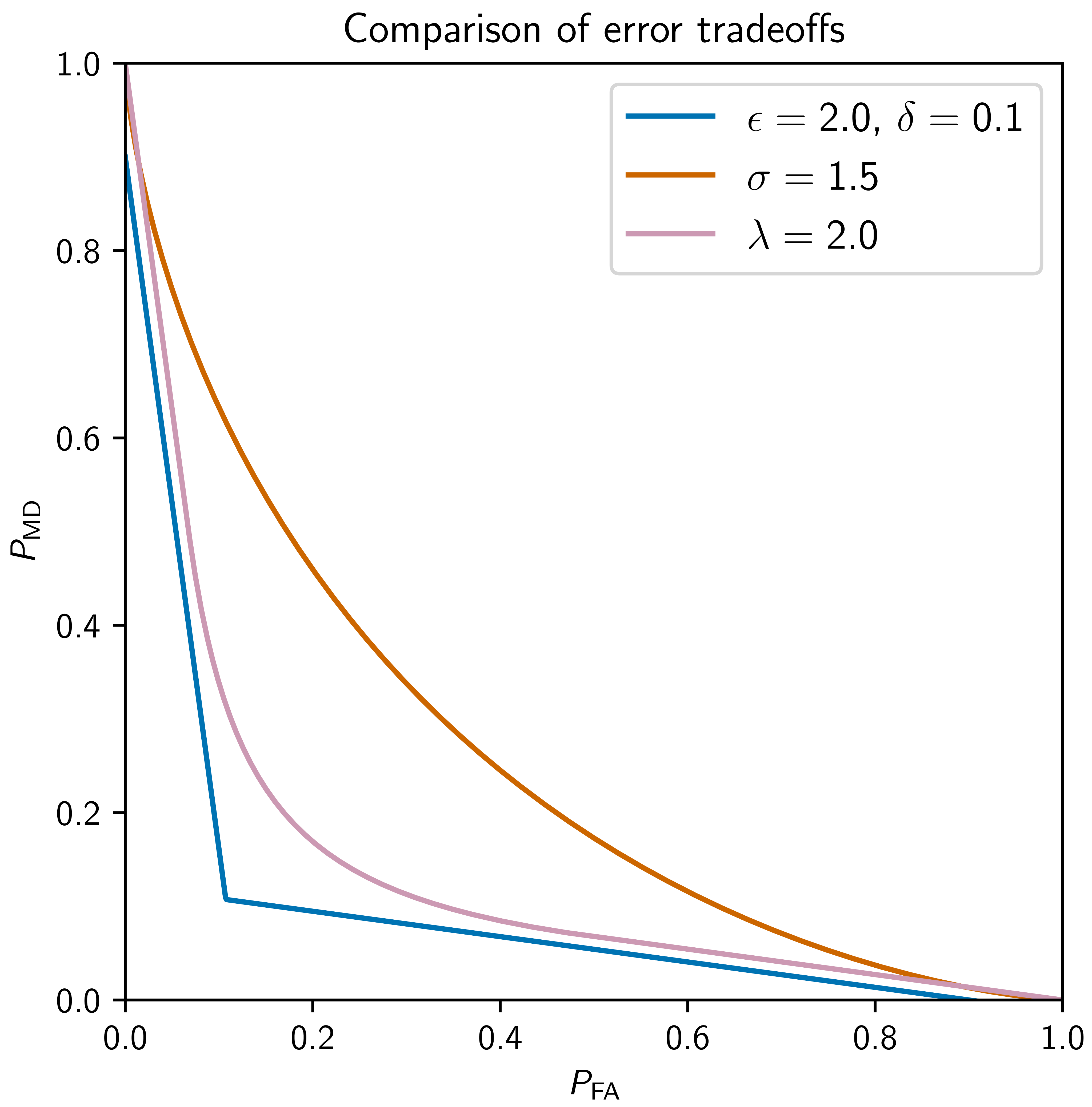
$$\mathcal{H}_1 : Z \sim p(z \mid 1)$$

The optimal test for the adversary is a **likelihood ratio test**:

$$\hat{b} = \begin{cases} 1 & \log \frac{p(z \mid 1)}{p(z \mid 0)} \geq \tau \\ 0 & \log \frac{p(z \mid 1)}{p(z \mid 0)} < \tau \end{cases}$$

Tradeoffs between P_{FA} and P_{MD}

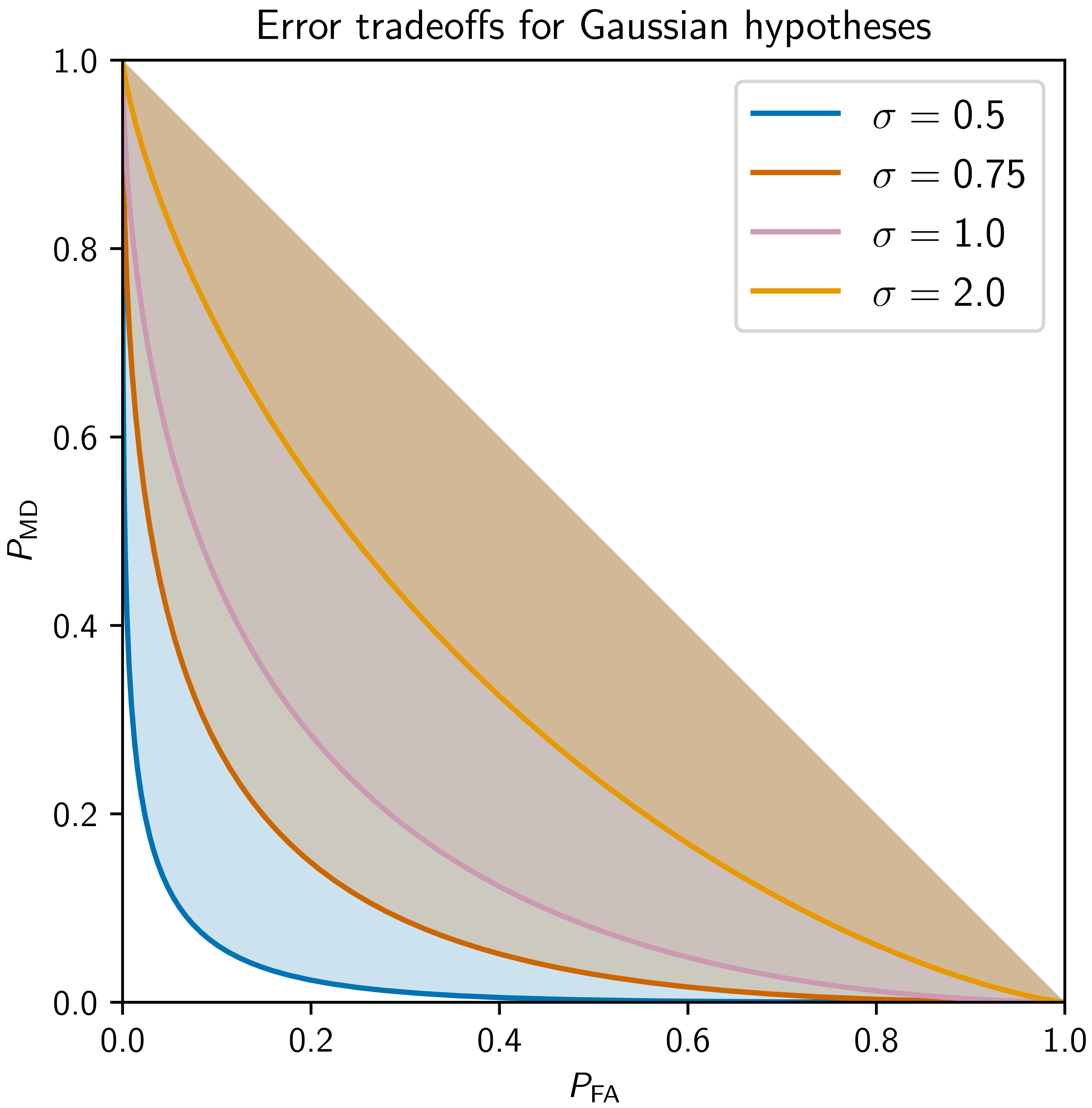
We get more privacy when the hypothesis test is “hard”



- A **privacy guarantee** is made by the tradeoff between probabilities of
  - **false alarm** (Type I error) and
  - **missed detection** (Type 2 error)
- If the likelihood ratio is small, the test will have a higher error.
- We can use a version of the ROC curve to visualize the kinds of guarantees.

Example: additive Gaussian noise

Everyone's favorite example: Gaussians!



If the revealed information Z is **Gaussian**:

$$\mathcal{H}_0 : Z \sim \mathcal{N}(0, \sigma^2)$$

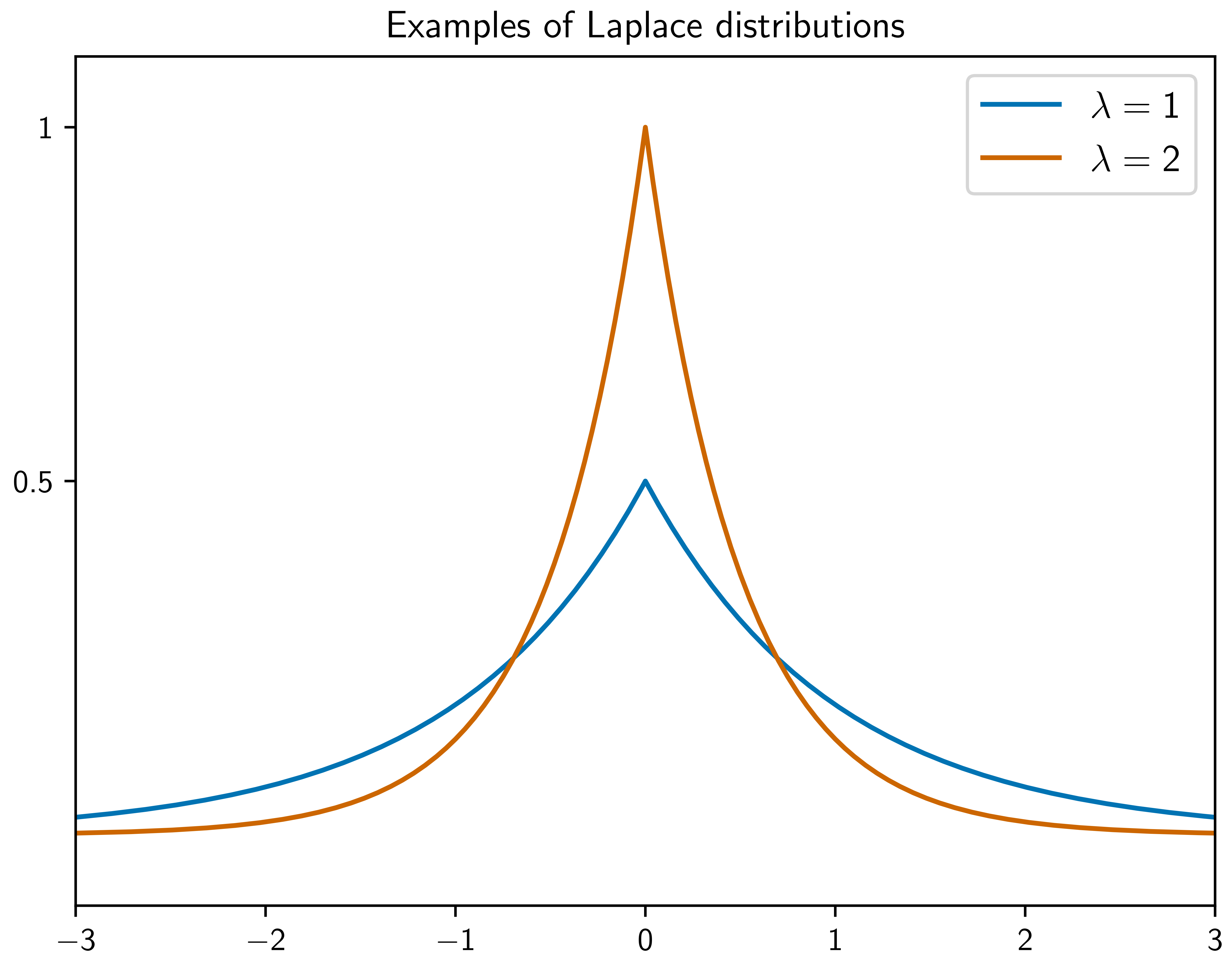
$$\mathcal{H}_1 : Z \sim \mathcal{N}(1, \sigma^2)$$

We can write the error probabilities in terms of Q functions:

$$P_{\mathrm{FA}} = Q\left(\frac{t}{\sigma}\right), P_{\mathrm{MD}} = Q\left(\frac{1-t}{\sigma}\right).$$

Example: additive Laplace noise

We can do more than just Gaussians!



If the revealed information Z is **Lapace**:

$$\mathcal{H}_0 : X \sim \text{Laplace}(\lambda)$$

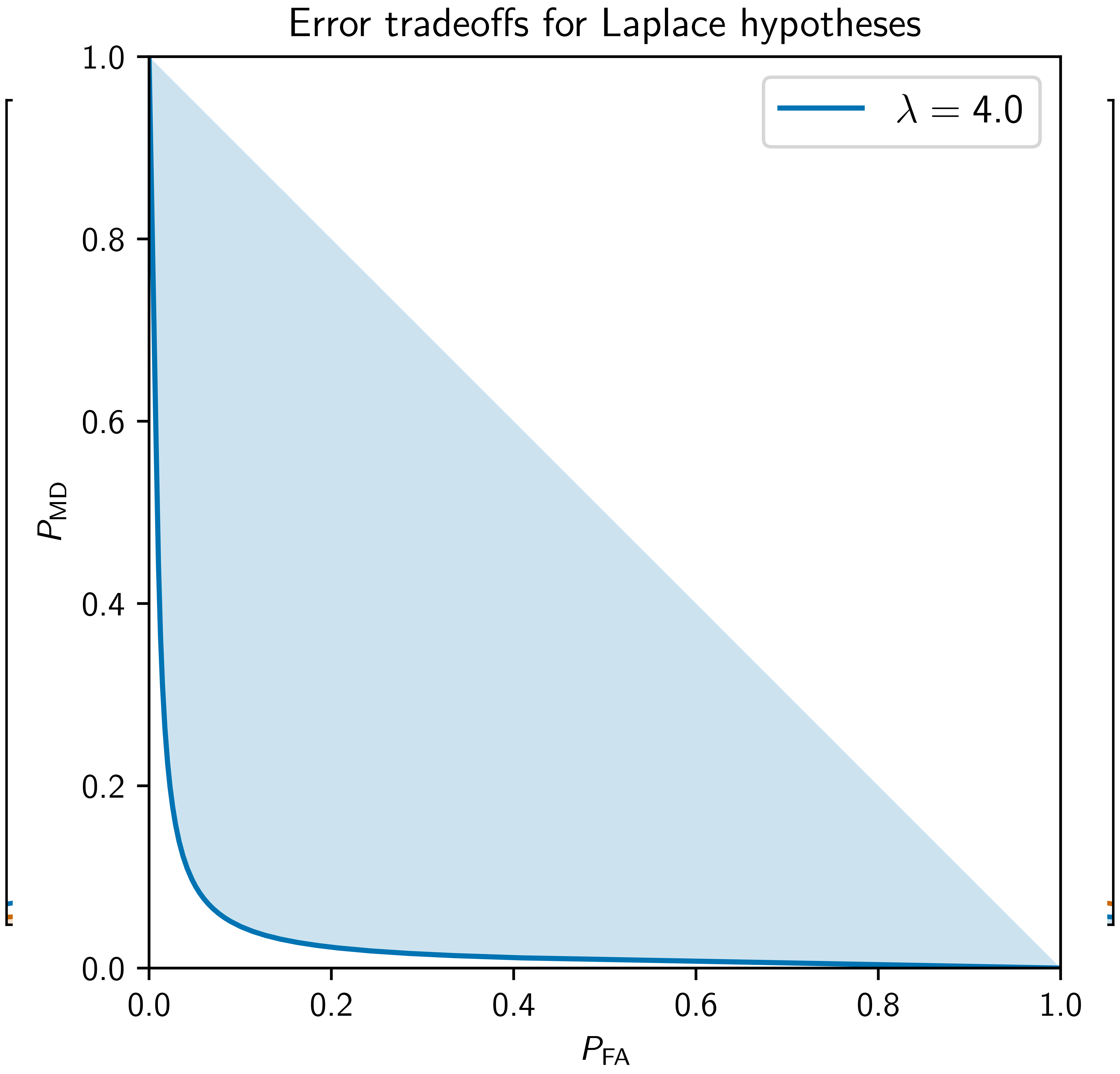
$$\mathcal{H}_1 : X \sim 1 + \text{Laplace}(\lambda)$$

Where $\text{Laplace}(\lambda)$ has density

$$p(z) = \frac{\lambda}{2} \exp(-\lambda |z|).$$

Error tradeoffs for Laplace noise

Lighter tails give a different shape



The error probabilities for the test are:

$$P_{FA} = \int_t^{\infty} \frac{\lambda}{2} \exp(-|t|\lambda)dt$$

$$P_{MD} = \int_{-\infty}^{t} \frac{\lambda}{2} \exp(-|A - t|\lambda)dt$$

The tradeoff is similar to the Gaussian but the slope at the corners is different.

Hard tests mean more privacy

Designing lower bounds on error probability

We can define privacy in terms of **lower bounds on the tradeoff curve**.

One way to do this is to put **bounds on the log likelihood ratio**.

Suppose we have bounds like:

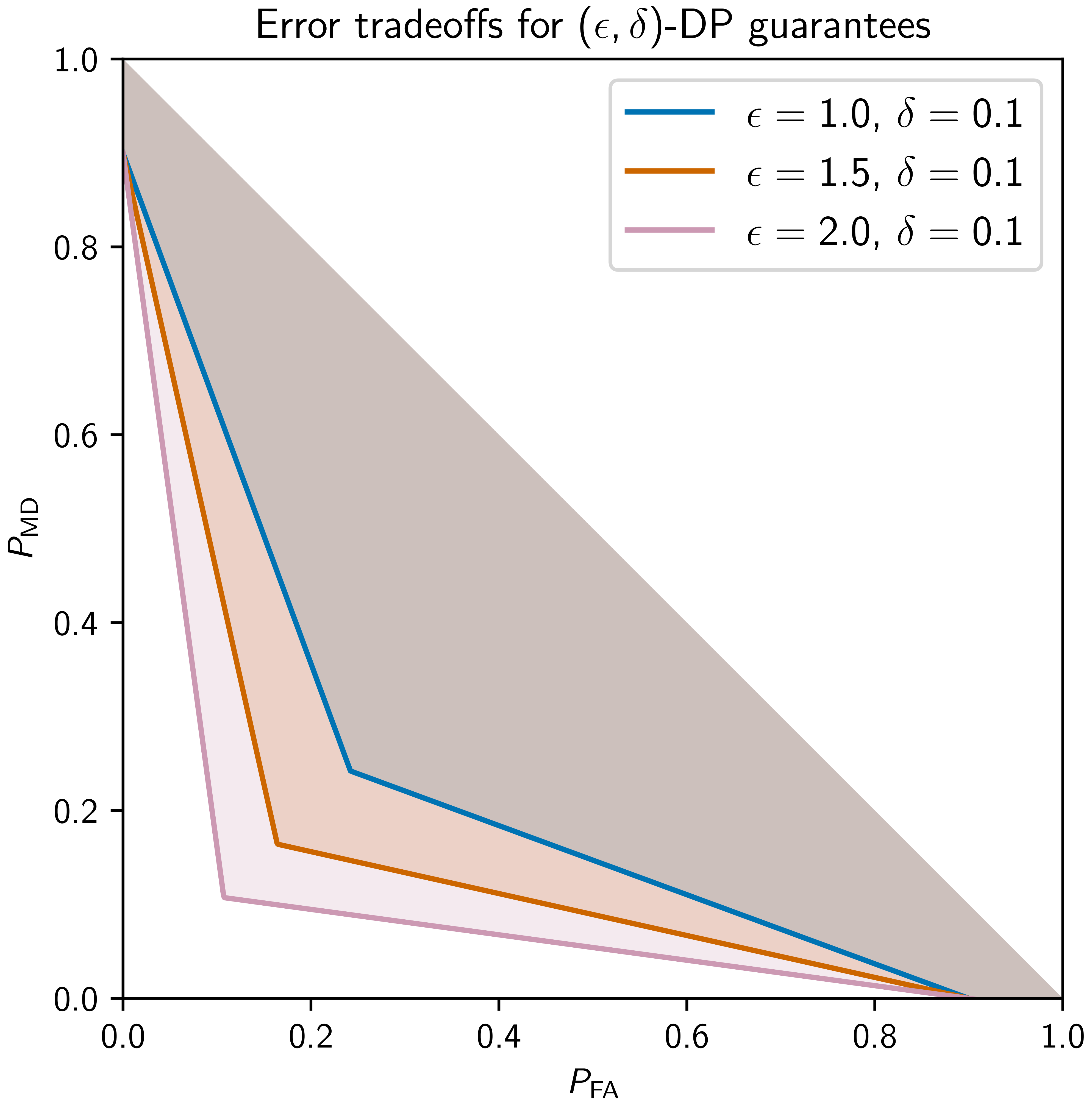
$$P_{\mathrm{FA}} + e^{\epsilon} P_{\mathrm{FA}} \geq 1 - \delta$$

$$e^{\epsilon} P_{\mathrm{FA}} + P_{\mathrm{MD}} \geq 1 - \delta$$

This is exactly the same as the definition of **(ϵ, δ)-differential privacy!**
[Wasserman and Zhou 2010, Kairouz, Oh, Vishwanath 2017]

Error tradeoffs from DP lower bounds

Using piecewise linear functions to bound the error



Starting with

$$P_{\mathrm{FA}} + e^{\epsilon} P_{\mathrm{FA}} \geq 1 - \delta$$

$$e^{\epsilon} P_{\mathrm{FA}} + P_{\mathrm{MD}} \geq 1 - \delta$$

We see different error tradeoffs.

We can vary ϵ or vary δ to see how these "privacy parameters" affect the shape of the tradeoff region.

Revisiting Gauss and Laplace

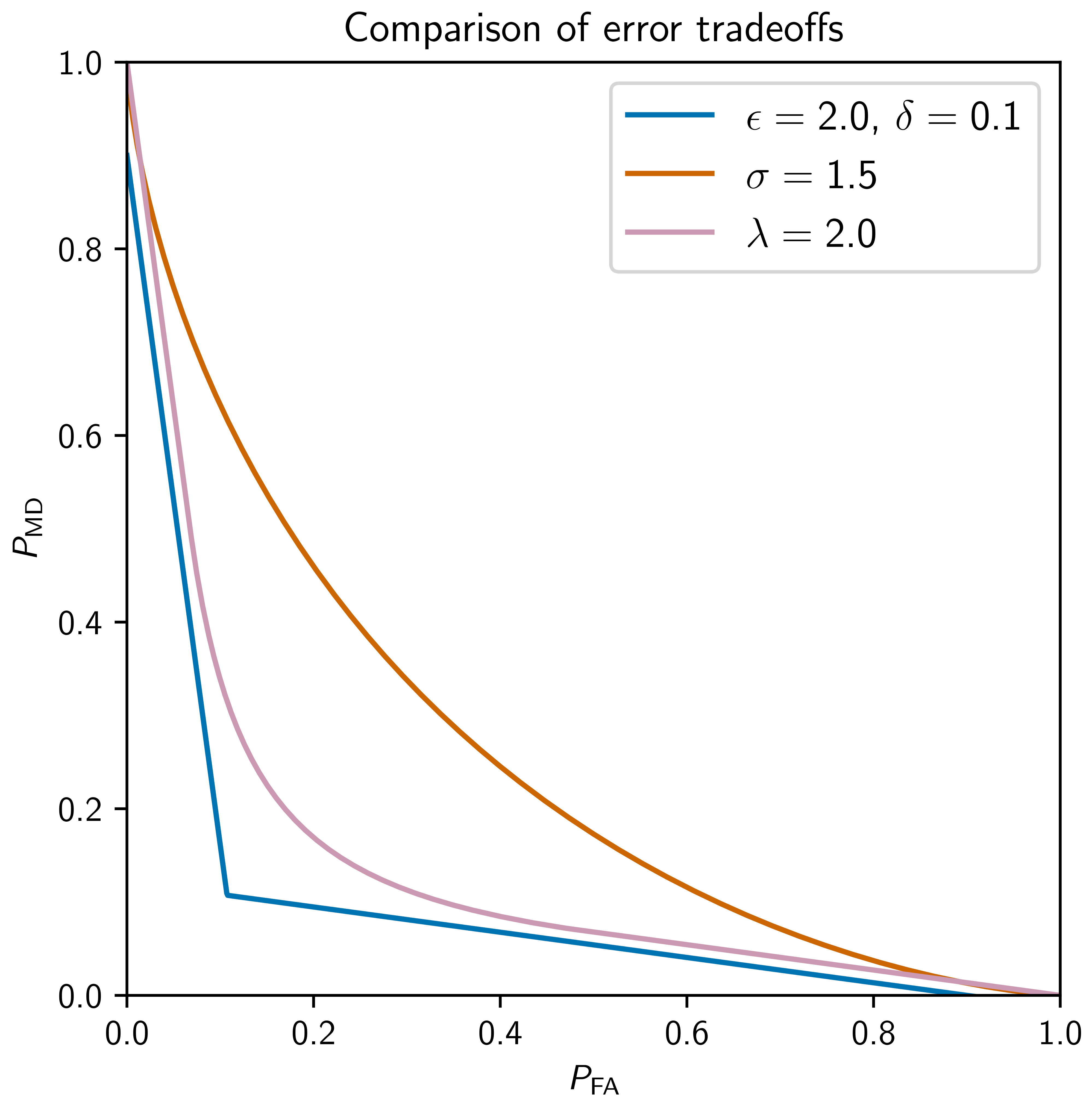
What DP guarantees do our previous hypothesis tests have?

Any kind of lower bound gives a way of measuring privacy!

Laplace and Gaussian tests do not meet the DP lower bounds exactly.

We can base privacy guarantees around any shape of tradeoff curve [Dong, Roth, Su 2019].

How do we reconcile the "standard" DP story with this simple binary hypothesis test?



The "standard" approach to explaining DP

Neighboring databases of individual records

In the textbook approach to describing DP we have several ingredients:

1. **Data space:** $\mathscr{X}$, often modeled as records from n individuals.

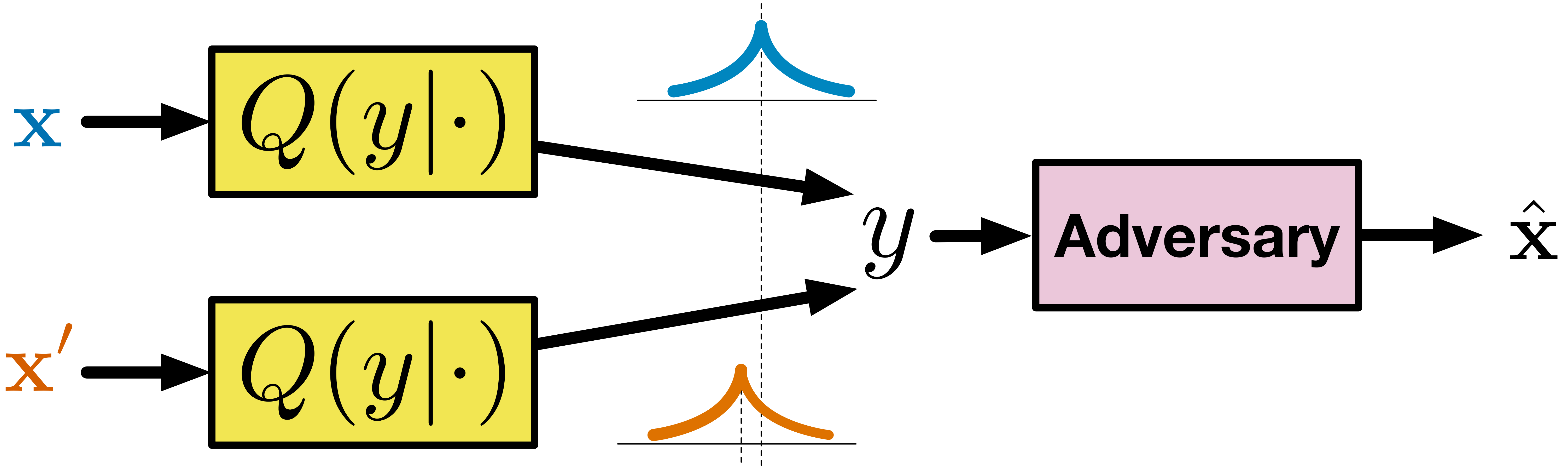
2. **Neighborhood relationship $\sim$:** for $\mathbf{x}, \mathbf{x}' \in \mathscr{X}$ we write $\mathbf{x} \sim \mathbf{x}'$ if they are "neighbors".
   - Example: each person has 1 bit so $\mathscr{X} = \{0,1\}^n$ and $\mathbf{x} \sim \mathbf{x}'$ if they differ in one position.

3. **Output space:** $\mathscr{Y}$, depends on the functionality/what we want to release.
   - Example: If we want the average of data $\mathscr{X} = [0,1]^n$, we have s$\mathscr{Y} = [0,1]$.
   - Example: If we want to train a classifier using data $\mathscr{X} = \{\mathbb{R}^d \times \{0,1\}\}^n$, $\mathscr{Y} = \mathbb{R}^d$.

4. **Algorithm:** a randomized map/conditional distribution/channel $Q: \mathscr{X} \to \mathscr{Y}$.

The hypothesis testing in DP

DP is a property of the channel



A channel/"mechanism"/algorithm Q is (ϵ, δ)**-differentially private** if

$$Q(\mathcal{S} \,|\, \mathbf{x}) \leq e^{\epsilon} Q(\mathcal{S} \,|\, \mathbf{x}') + \delta$$

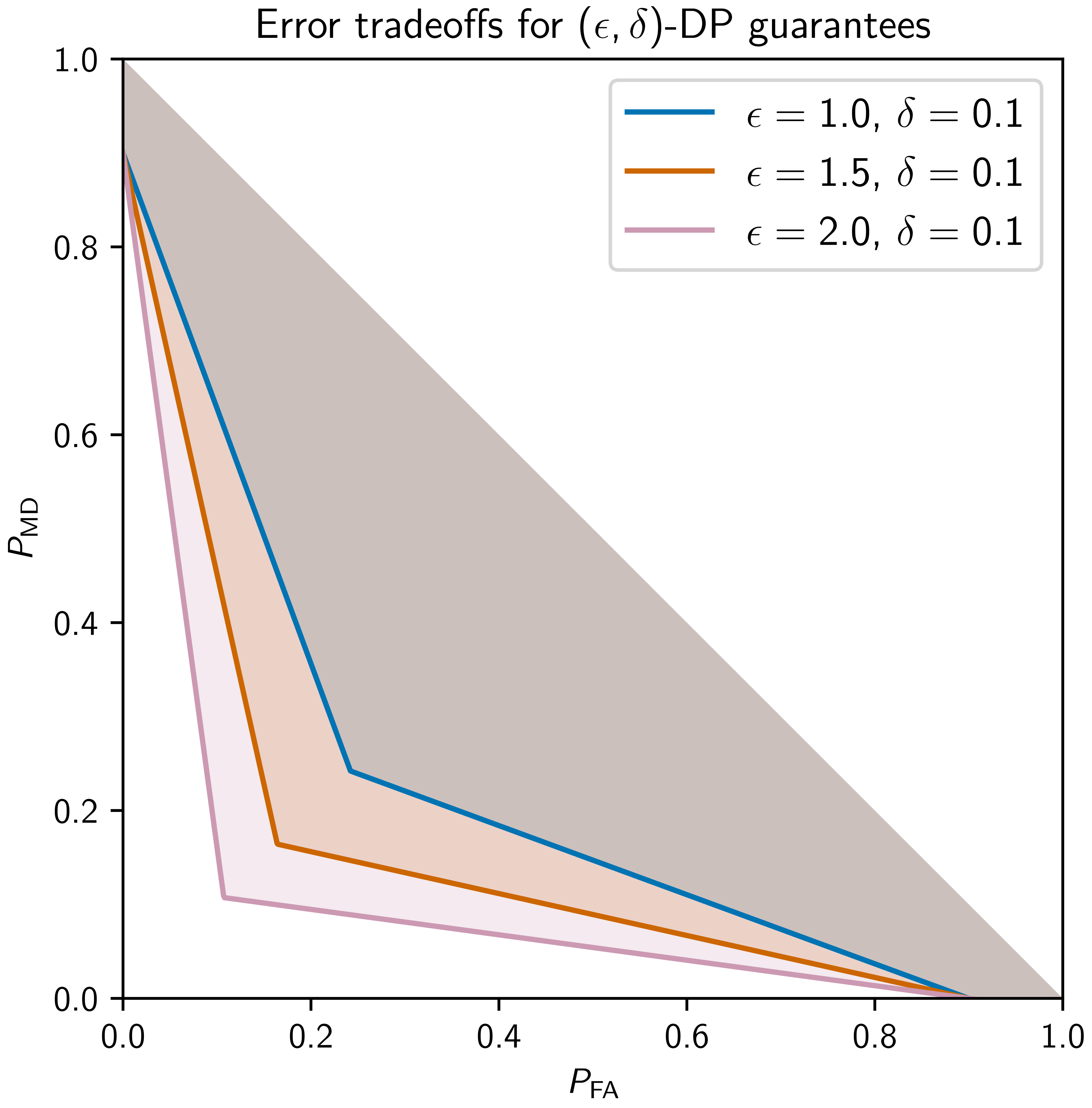
For all measurable subsets $\mathcal{S} \subseteq \mathcal{Y}$ and all $\mathbf{x} \sim \mathbf{x}'$.

[Dwork-Kenthapadi-McSherry-Mironov-Naor 2006]

[Wasserman-Zhou 2010]

DP makes many hypothesis tests hard

Protecting many single bits simultaneously



Compared to our single private bit b, in DP we want many hypothesis tests to hard for the adversary. For every $\mathbf{x} \sim \mathbf{x}'$ the test

$$\mathscr{H}_0 : \mathbf{y} \sim Q(\,\cdot\,|\,\mathbf{x})$$

$$\mathscr{H}_1 : \mathbf{y} \sim Q(\,\cdot\,|\,\mathbf{x}')$$

should have a large probability of error.

When can we do this? When neighboring data sets make similar output distributions.

Sensitivity of scalar functions

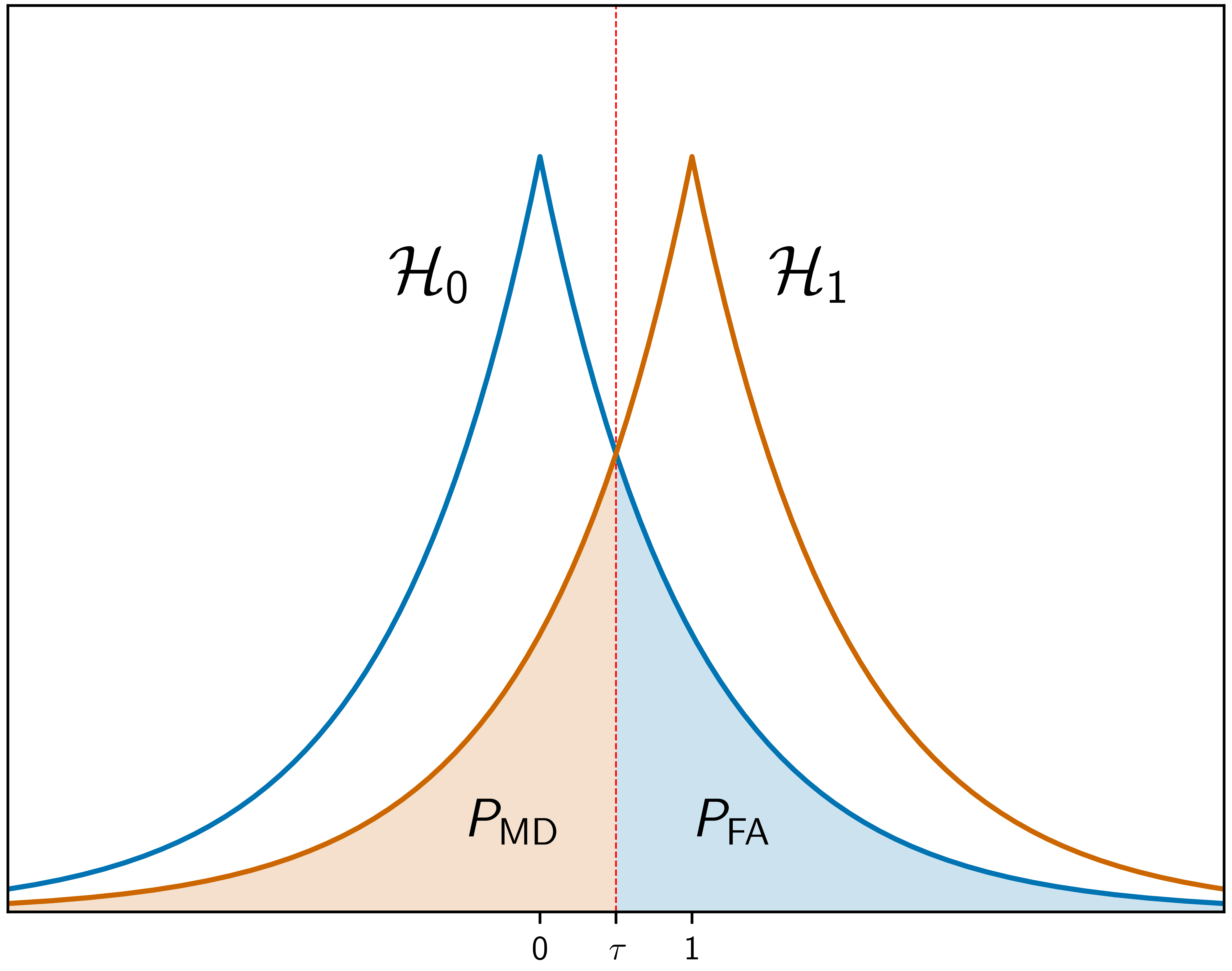
Understanding the distance between hypotheses

In DP, we usually want to approximate some function of the data.

Suppose we want $f : \mathcal{X} \to \mathbb{R}$. We want the test to be hard for any pair $(\mathbf{x}, \mathbf{x}')$ which are "neighbors" ($\mathbf{x} \sim \mathbf{x}'$).

If $f(\cdot)$ is small for all neighbors, this should be easier.

Example: $f(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$ can change by at most $\frac{1}{n}$ for $x_i \in [0,1]$.



Sensitivity of scalar functions

Understanding the distance between hypotheses

The **global sensitivity** of $f(\,\cdot\,)$ is

$$\Delta(f) = \max_{\mathbf{x}\sim\mathbf{x}'} \left| f(\mathbf{x}) - f(\mathbf{x}') \right| .$$

If we use additive noise (like in the Laplace and Gaussian case) we have

$$\mathcal{H}_0 : Z \sim p(z - f(\mathbf{x})) \qquad \text{vs.} \qquad \mathcal{H}_1 : Z \sim p(z - f(\mathbf{x}'))$$

We can make a guarantee for all “neighbors” if following test is hard:

$$\mathcal{H}_0 : Z \sim p(z) \qquad \text{vs.} \qquad \mathcal{H}_1 : Z \sim p(z - \Delta(f)).$$

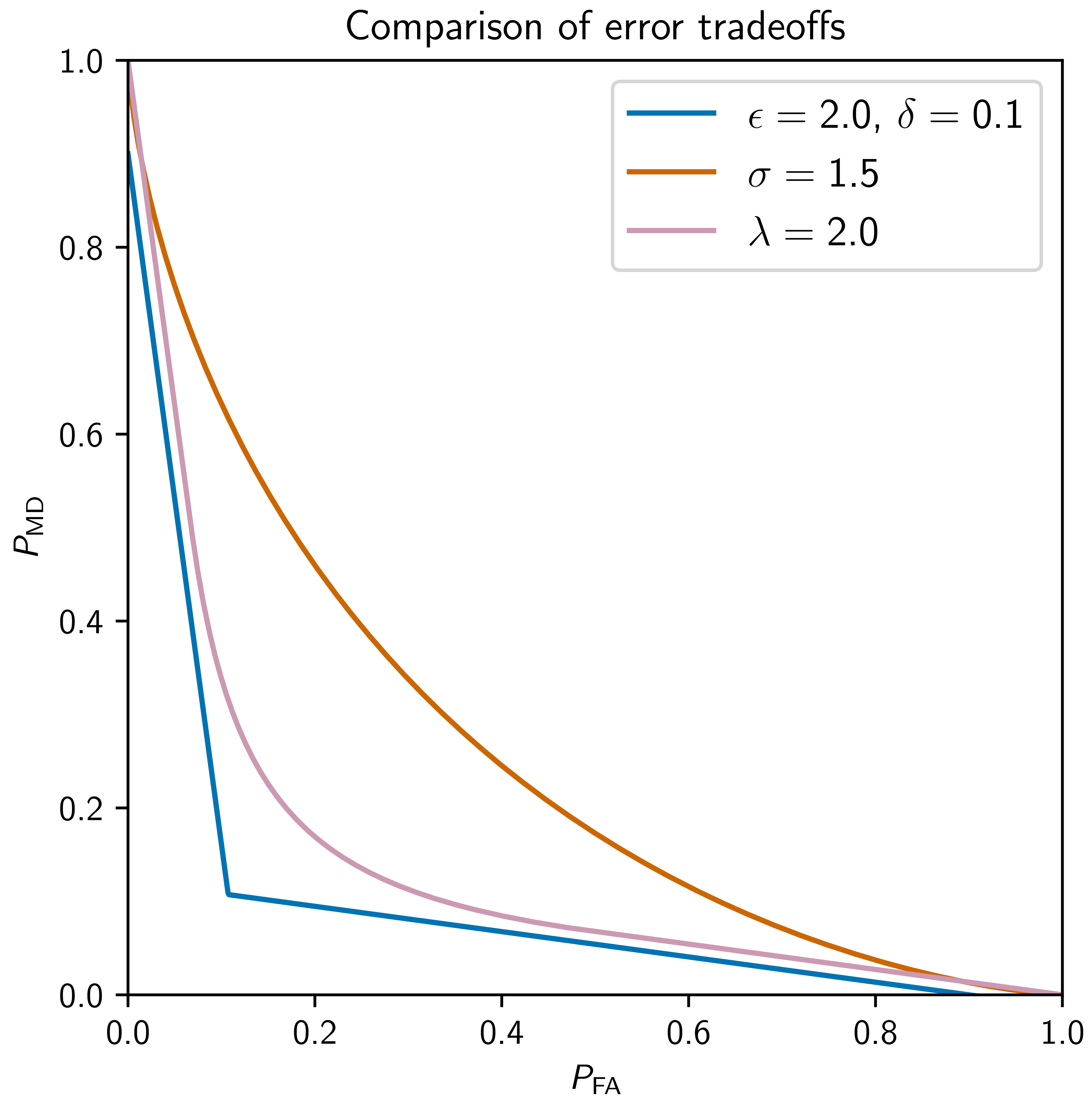
Some notes on the definition

DP's underlying assumptions are slightly different

- **Differential privacy is a stringent requirement:** The probability of any event is similar, regardless of whether the data was x or any other neighboring x′.
- **Guarantee is on conditional probabilities given the data:** same risk holds regardless of side information (e.g. linkage).
- **There is no statistical assumption on the data:** x is not drawn from some distribution since it's in the conditioning.
- **The data itself is considered identifying:** no notion of some parts being personally identifiable information (PII) and others not.

DP and hypothesis testing

Fundamentally, DP is just a lower bound



The guarantee

$$Q(\mathcal{S} \,|\, \mathbf{x}) \leq e^{\epsilon} Q(\mathcal{S} \,|\, \mathbf{x}') + \delta$$

Is equivalent to saying

$$P_{\mathrm{FA}} + e^{\epsilon} P_{\mathrm{FA}} \geq 1 - \delta$$

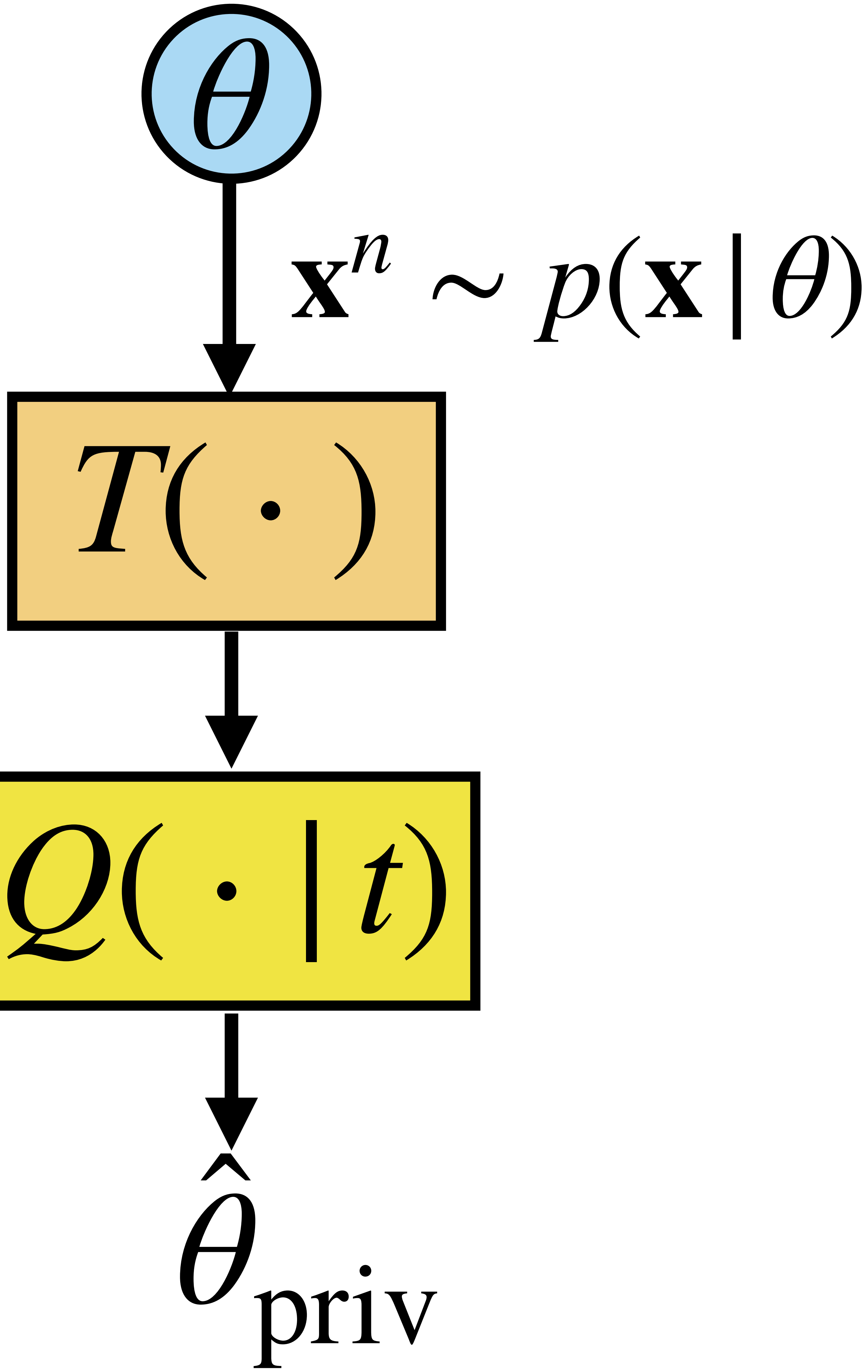
$$e^{\epsilon} P_{\mathrm{FA}} + P_{\mathrm{MD}} \geq 1 - \delta$$

Any test used by an adversary taking $\mathbf{y}$ and guessing if $\mathbf{x} \sim \mathbf{x}'$.

Differentially private ML

Point estimation with differential privacy

Adding noise to sufficient statistics



A typical DP approach to statistical estimation (Smith 2009):

- Model data as drawn i.i.d. $\sim p(\mathbf{x} \mid \theta)$.
- Compute a sufficient statistic $T(\mathbf{x}^n)$ for θ.
- Add noise to $T(\mathbf{x}^n)$ to guarantee DP.
- Compute a "plug-in" estimate from noisy $T(\mathbf{x}^n)$.

We just need the sensitivity of $T(\,\cdot\,)$.

Example: the sample mean

Computing the MSE as a function of privacy risk

Suppose we have data in $\mathscr{X} = [A, B]^n$ and want to estimate the mean:

$$\hat{\mu}(\mathbf{x}^n) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j + Z$$

- Sensitivity of $\hat{\mu}(\mathbf{x}^n)$ is $(B - A)/n$.
- $Z \sim$ Laplace$(n\epsilon/(B - A))$ will guarantee $(\epsilon, 0)$-DP.
- MSE of $\hat{\mu}(\mathbf{x}^n)$ is $2/\lambda^2 = 2\dfrac{(B - A)^2}{n^2\epsilon^2}$.

The privacy-utility tradeoff

How much do we lose when we guarantee privacy?

Adding $\text{Laplace}(\lambda)$ noise guarantees privacy, but at what cost? The MSE is:

$$2/\lambda^2 = 2\frac{(B-A)^2}{n^2\epsilon^2}$$

So we can see that less privacy risk (smaller ϵ) induces more MSE.

We can try to optimize the privacy mechanism if we know the utility function (like squared error).

This is what people call the **privacy-utility tradeoff**.

Beyond additive noise

Sampling for privacy with the exponential mechanism

The Exponential Mechanism [McSherry, Talwar, 2007] samples a random $\mathbf{y}$ to maximize

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}} u(\mathbf{y}, \mathbf{x})$$

To approximate this, sample according to a Gibbs measure using the sensitivity of $u(\,\cdot\,)$:

$$Q(\mathbf{y} \,|\, \mathbf{x}) \propto \exp\big(\epsilon u(\mathbf{y}, \mathbf{x})/2\Delta(u)\big).$$

Maximum likelihood and ERM

Optimization and privacy

Most of “modern” machine learning involves optimization problems, including maximum likelihood estimation and empirical risk minimization:

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i).$$

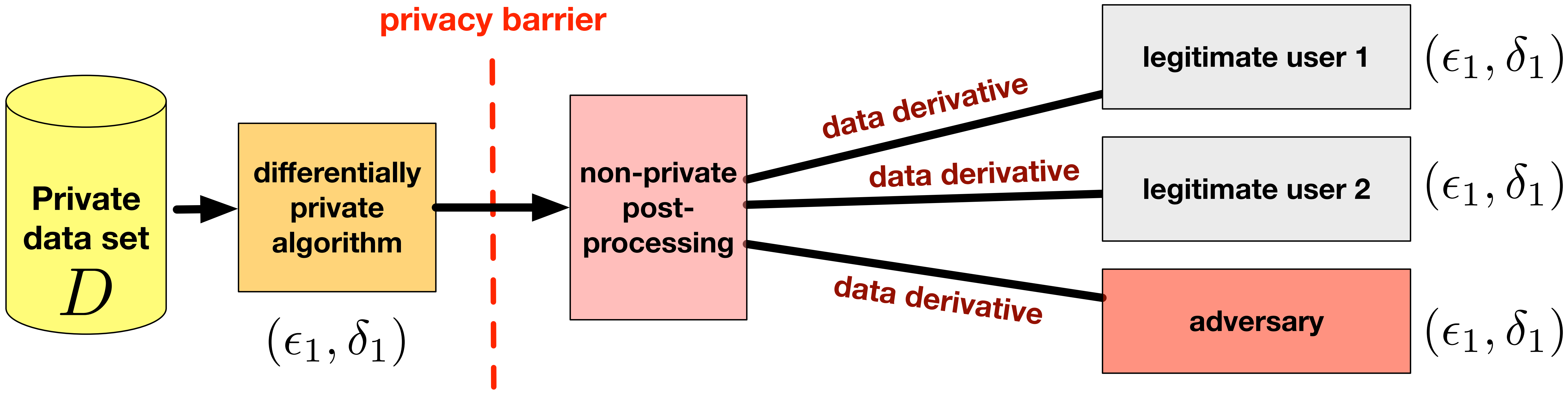
We can use DP to approximate this in a number of ways:

- **“Output perturbation”**: compute the minimizer and add noise.
- **“Objective perturbation”**: Add a random term to the objective function and minimize it.
- **“Functional mechanism”**: Add noise to an approximation of the loss function $\ell(\,\cdot\,)$.

[Chaudhuri, Monteleoni, Sarwate 2011]
[Zhang, Zhang, Xiao, Yang, Winslett 2012]

Post-processing invariance and composition

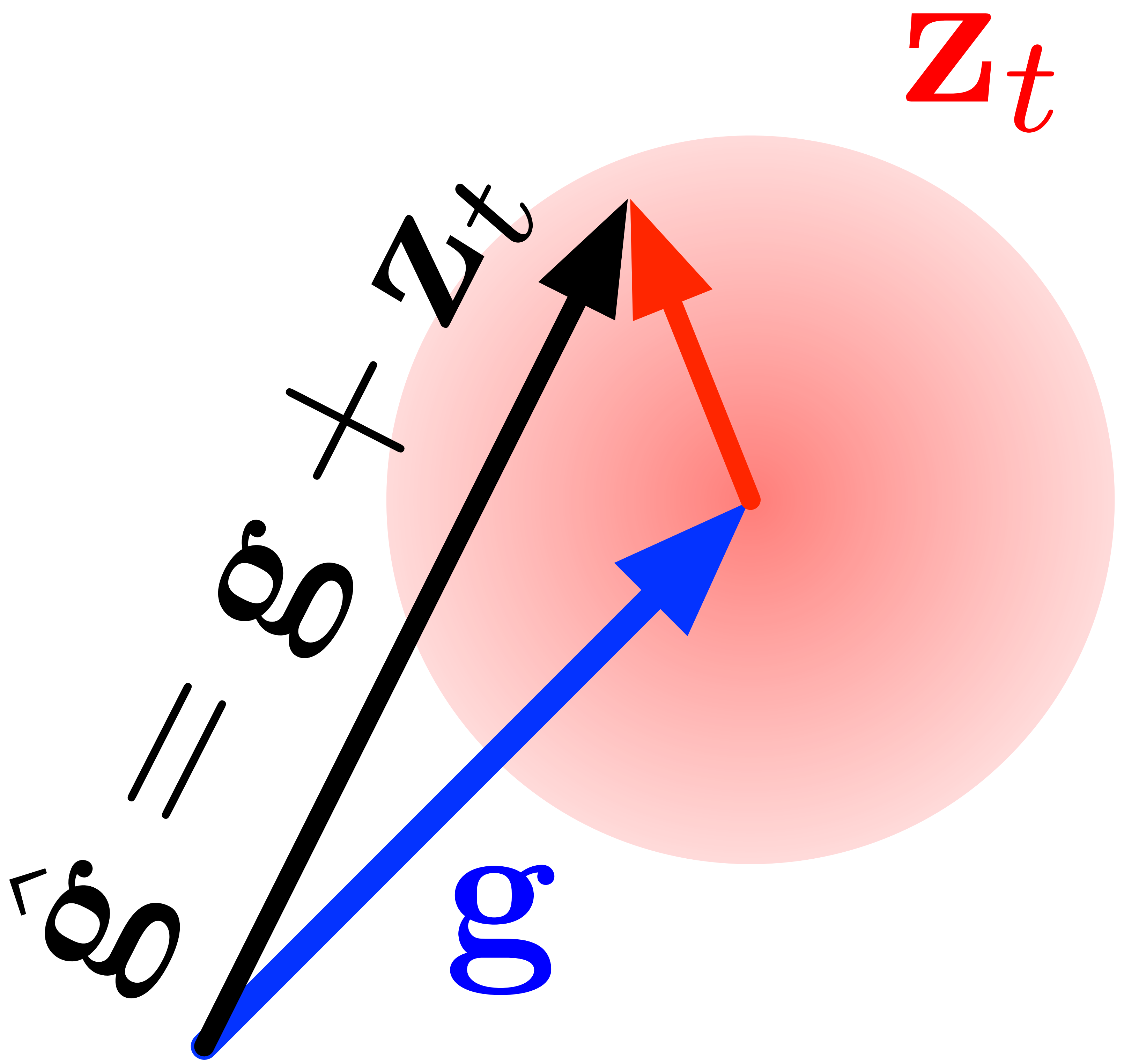
Nice properties of differential privacy



- **Side-information resilience:** measures the additional risk regardless of what is known already.
- **Post-processing invariance:** once we publish something the risk cannot increase from additional computations.
- **Composition:** quantifies how privacy loss "adds up" over multiple releases.

Deep Learning and DP

Privacy for neural networks



Deep neural networks (DNNs) also use optimization algorithms in training. To make these private we can add noise to the gradients in stochastic gradient descent (SGD):

- Adding noise to gradients provides differential privacy.
- For high-dimensional problems, Gaussian noise is very effective.
- Need to use **privacy accounting**.

[Song et.al. 2013, Duchi et.al. 2014, Abadi et.al. 2016, Mironov 2017]

Composing multiple mechanisms

Returning to our hypothesis testing roots

For any (ϵ, δ)-DP mechanism we can always find [Sommer, Meiser, Mohammadi 2019] a pair of **dominating distributions** (P, Q) such that:

$$Q(\mathcal{S} \mid \mathbf{x}) - e^{\epsilon} Q(\mathcal{S} \mid \mathbf{x}') \leq P(\mathcal{S}) - e^{\epsilon} Q(\mathcal{S}).$$

We can then define the **privacy loss random variable (PLRV)** for $Z \sim P$:

$$L = \log \frac{dP}{dQ}(Z).$$

Each time we use use a DP mechanism we get another PLRV. Composition rules tell us how to "add up" these PLRVs.

Approaches to composition

Different ways to count up PLRVs

If we have PLRVs $L_1, L_2, \ldots, L_T$, how can we find the total privacy loss from running these on our data?

- Measure concentration [Dwork, Rothblum, Vadhan 2010]
- Exact composition [Kairouz, Oh, Vishwanath 2015][Murtagh, Vadhan 2016]
- Large deviations/MGF [Abadi et al. 2016][Mironov et al. 2017][Balle et al. 2019]
- CLT [Dong et al. 2019][Sommer et al. 2019]
- Numerical approximation [Koskela et al. 2019, 2021][Koskela, Honkela 2020][Gopi et al. 2021][Ghazi et al. 2022][Doroshenko et al. 2022]
- Saddlepoint analysis [Alghamdi et al. 2022]

Main takeaways for DP machine learning

The state of the art for DP and ML is constantly evolving

- **Basic algorithmic ideas are the same:** developing a differentially private ML algorithm for applications involves understanding where to introduce the noise.
- **The best algorithm for a task may be application-dependent:** x is not drawn from some distribution since it's in the conditioning.
- **Privacy accounting is complicated:** but generally gives us tighter bounds on the overall privacy for the algorithms we already have.
- **There is still a large gap between prototype and application:** there are lots of issues to handle that are a mix research questions and engineering.

DP in federated learning

Federated learning from private data

Defining the challenge

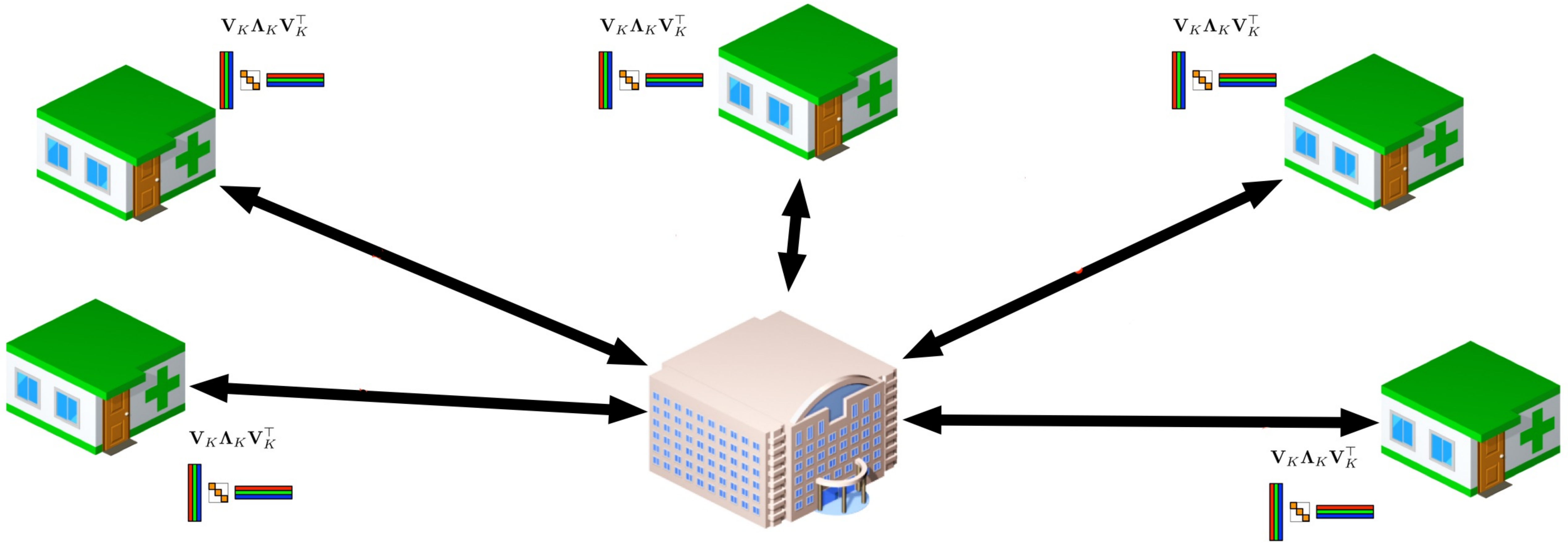
- Consortium of internet-connected research groups (sites).
- Each site has a cohort of (private) data from research subjects.
- Want to leverage larger total sample size to advance understanding.

Privacy: researchers have to promise each subject that their data will not be copied and that they cannot be identified as participants.

Federated learning: this is decentralized/distributed learning, which has been re-branded as “cross-silo federated learning.” [Kairouz et al. 2021]

Federated learning from private data

Defining the challenge

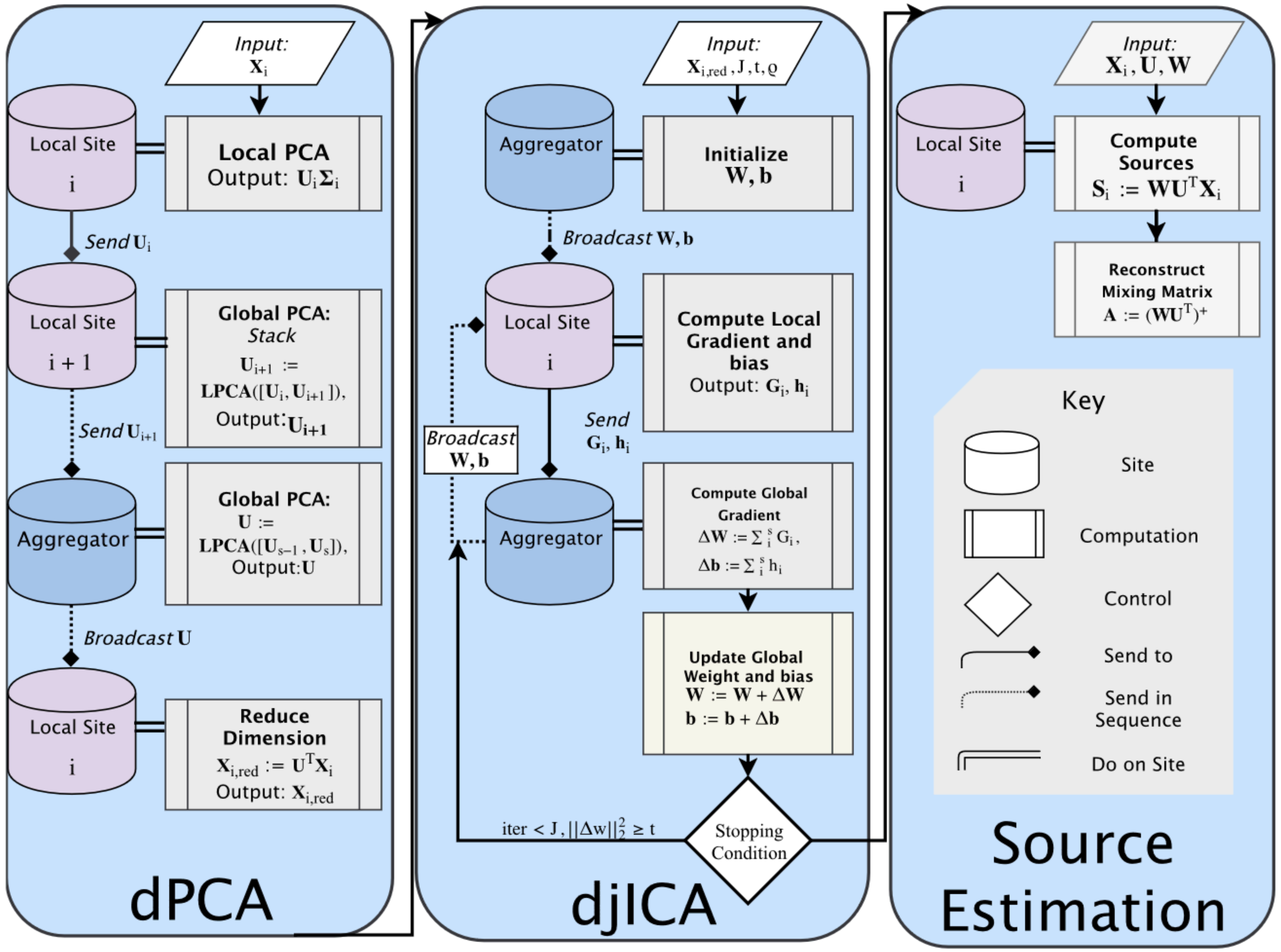


A consortium of S research groups/sites wants to collaborate

- **Data:** M_s individuals locally at each site s in datasets $\mathbf{X}_s = \{\mathbf{X}_{s,m} : m = 1,2,\ldots,M_s\}$.
- **Goal:** compute some target function $T(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_S)$ without uploading data to the cloud.

An example from neuroimaging

Independent component analysis is often used for MRI



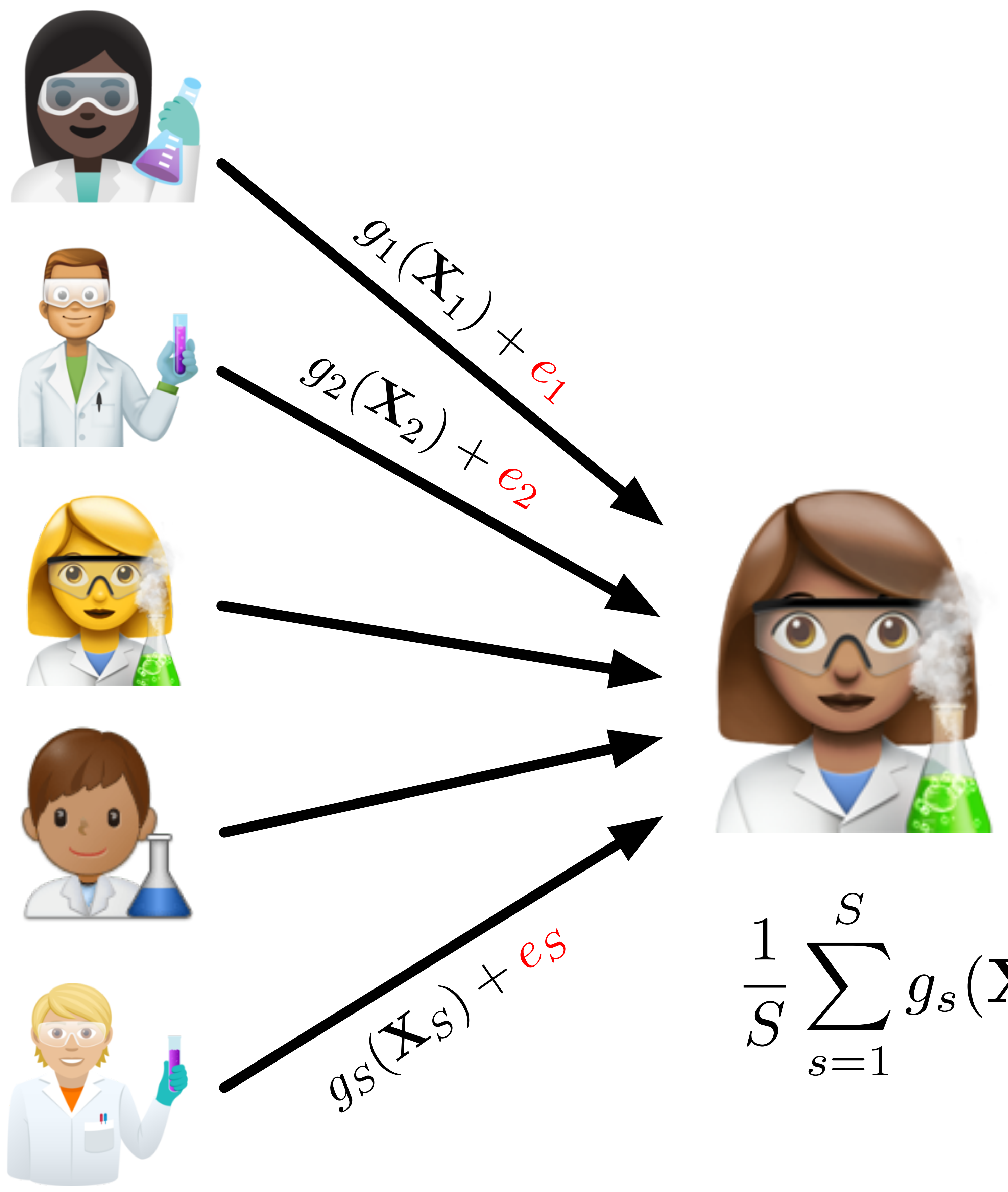
We studied a decentralized joint ICA:

- Each subject measurement $\mathbf{S} \in \mathbb{R}^{R \times N}$ is composed of N observations from R statistically independent components
- Linear mixing process defined by a mixing matrix $\mathbf{A} \in \mathbb{R}^{D \times R}$ with $D \geq R$, which forms the observed data $\mathbf{X} = \mathbf{AS}$.

We want to find an unmixing matrix jointly across the sites.

Algorithmic ingredients

Independent component analysis is often used for MRI



$$\frac{1}{S}\sum_{s=1}^{S} g_s(\mathbf{X}_s) + \frac{1}{S}\sum_{s=1}^{S} e_s$$

We used a distributed gradient descent (with noisy SGD) on a nonconvex objective:

- Sites send noisy local gradients.
- Aggregator updates the matrix and sends it back.
- Key contribution: use common randomness to allow sites to add anticorrelated noise that balances privacy and utility.

[Imtiaz et al. 2021][Tasnim et al. 2023]

DP challenges for collaborative science

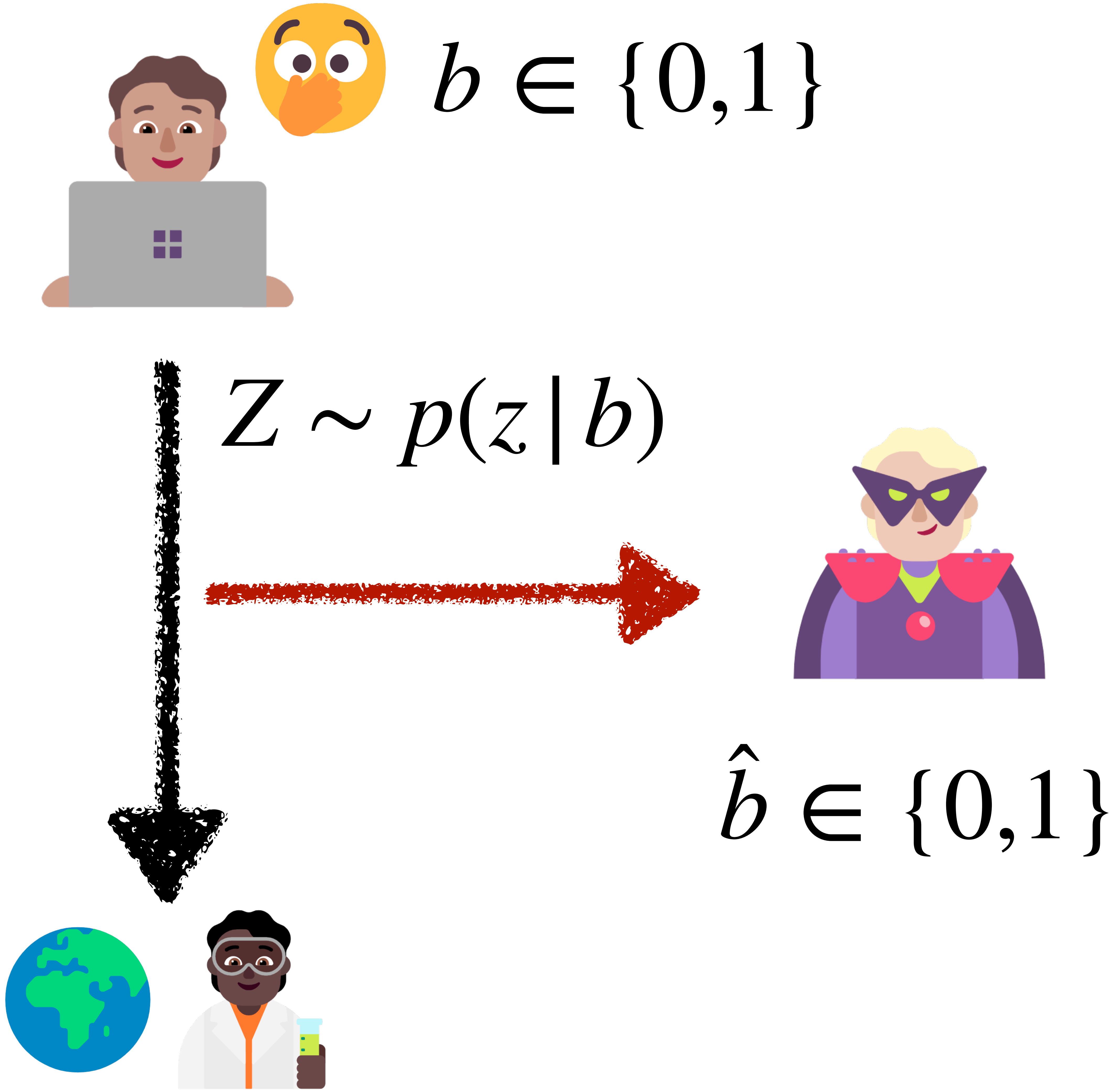
Broadening the scope of applications is hard

- **Sample sizes are small:** DP has had the most success in the “big data” setting whereas human health studies are small.
- **Generic approaches only go so far:** most algorithms have been “general purpose” and don’t use domain knowledge.
- **Real applications are pipelines:** almost all scientific analyses have a pipeline of processes and differential privacy is most often studied in isolation.
- **Interpretability and validation are important:** as with ML/AI more generally, we want to have scientifically meaningful results.

Conclusions and open questions

What we've seen so far

Let's start simple

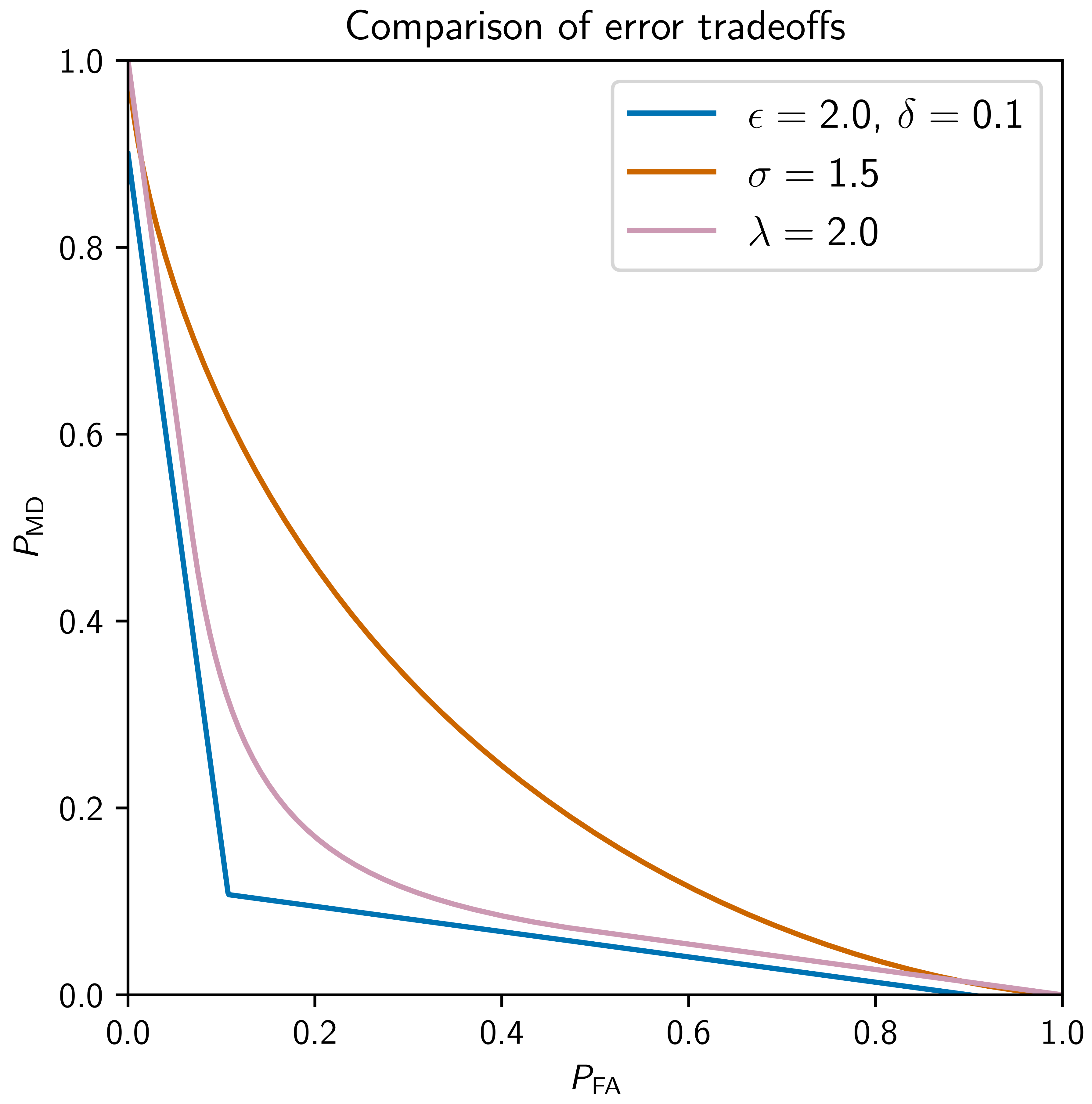


We started out with a simple story: protecting a single bit.

- Differential privacy both is and is not just as simple as hypothesis testing.
- Taking an information-theoretic view opens the door to better analyses.
- The gap between algorithms and analysis is shrinking.
- The gap between algorithms and applications is still large.

Where can we go from here?

Looking ahead, what are the major challenges



- Any lower bound is a type of privacy: which one is the easiest to work with?
- Optimality is hard to define in many applications (for example, visualization): what can do to find "good" mechanisms?
- How practical is DP in small sample, high-dimensional, or other challenging settings?
- When is DP the right solution and when is it the wrong solution?

대단히 감사합니다!