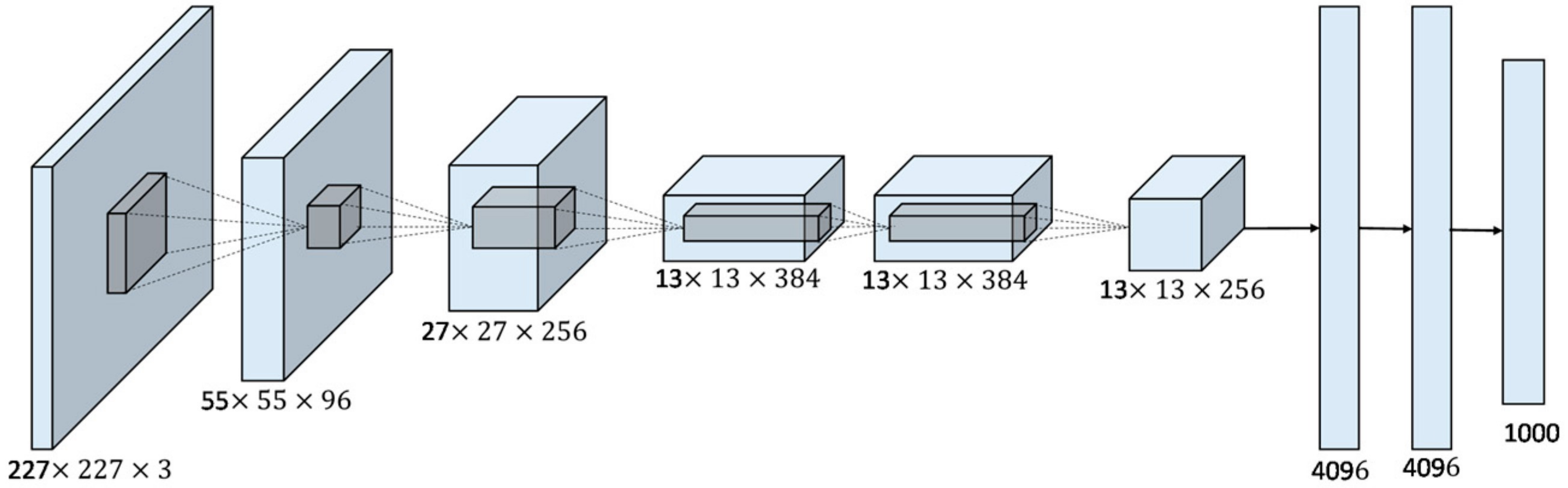
Towards Next-Generation ML/AI: Robustness, Optimization, Privacy

Krishna Pillutla

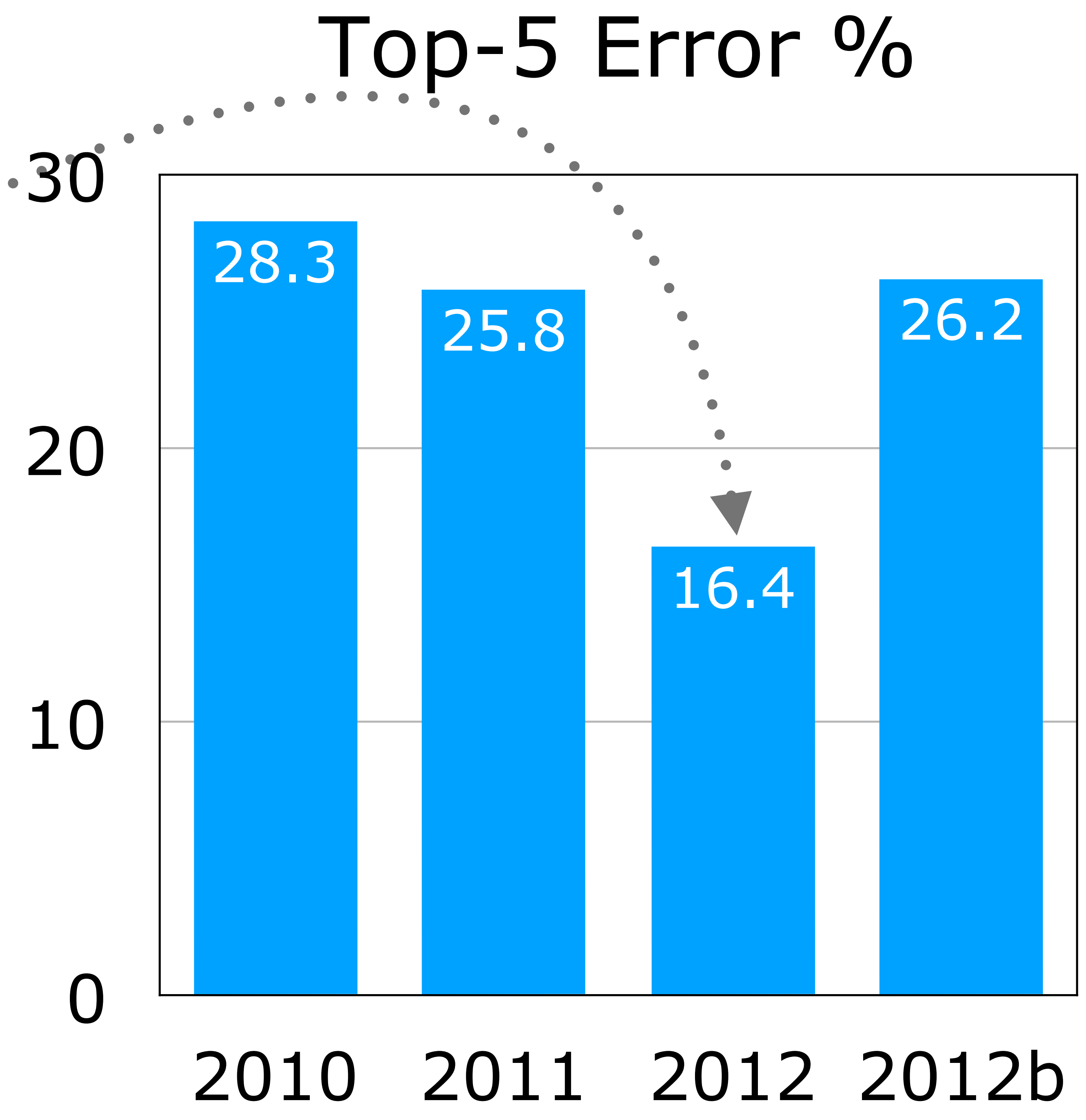
January 16th, 2023 @ IIT Madras

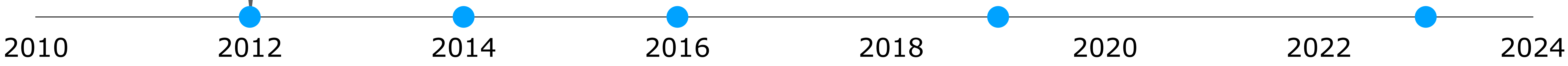
Google Research

ML/AI have been revolutionized in the last 10 years



[Krizhevsky, Sutskever, Hinton (NeurIPS 2012)]





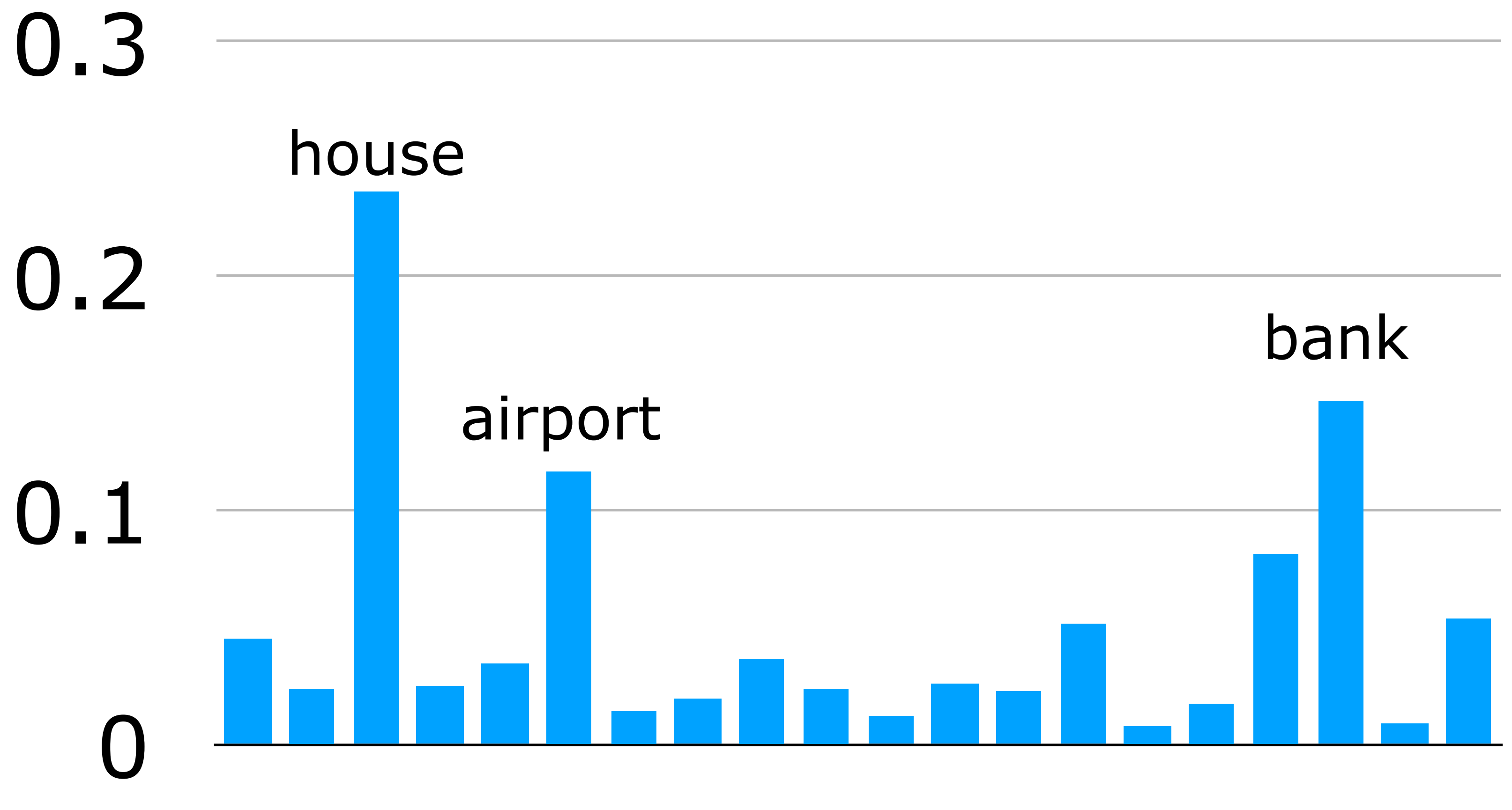
Language modeling

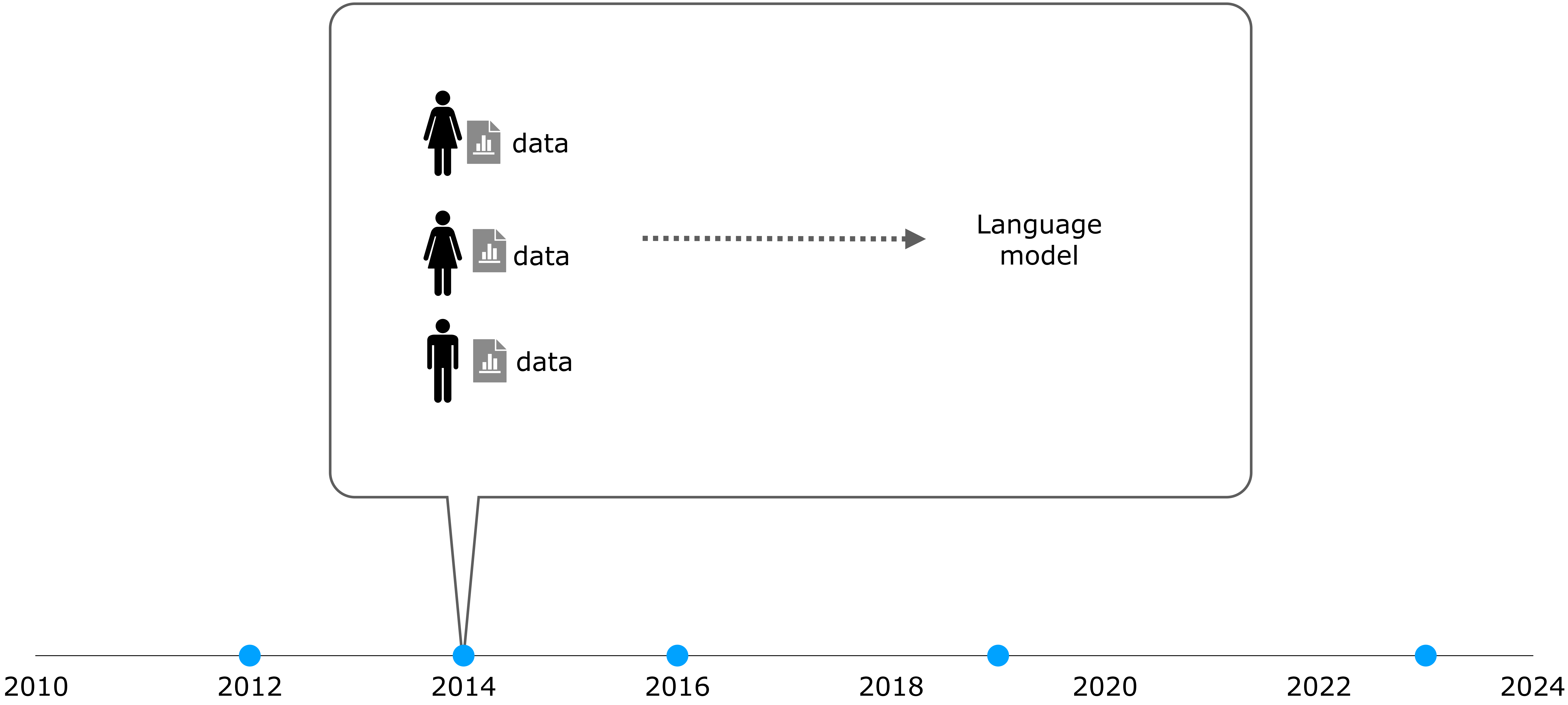
$x_1 \quad x_2 \quad \ldots \quad x_t$

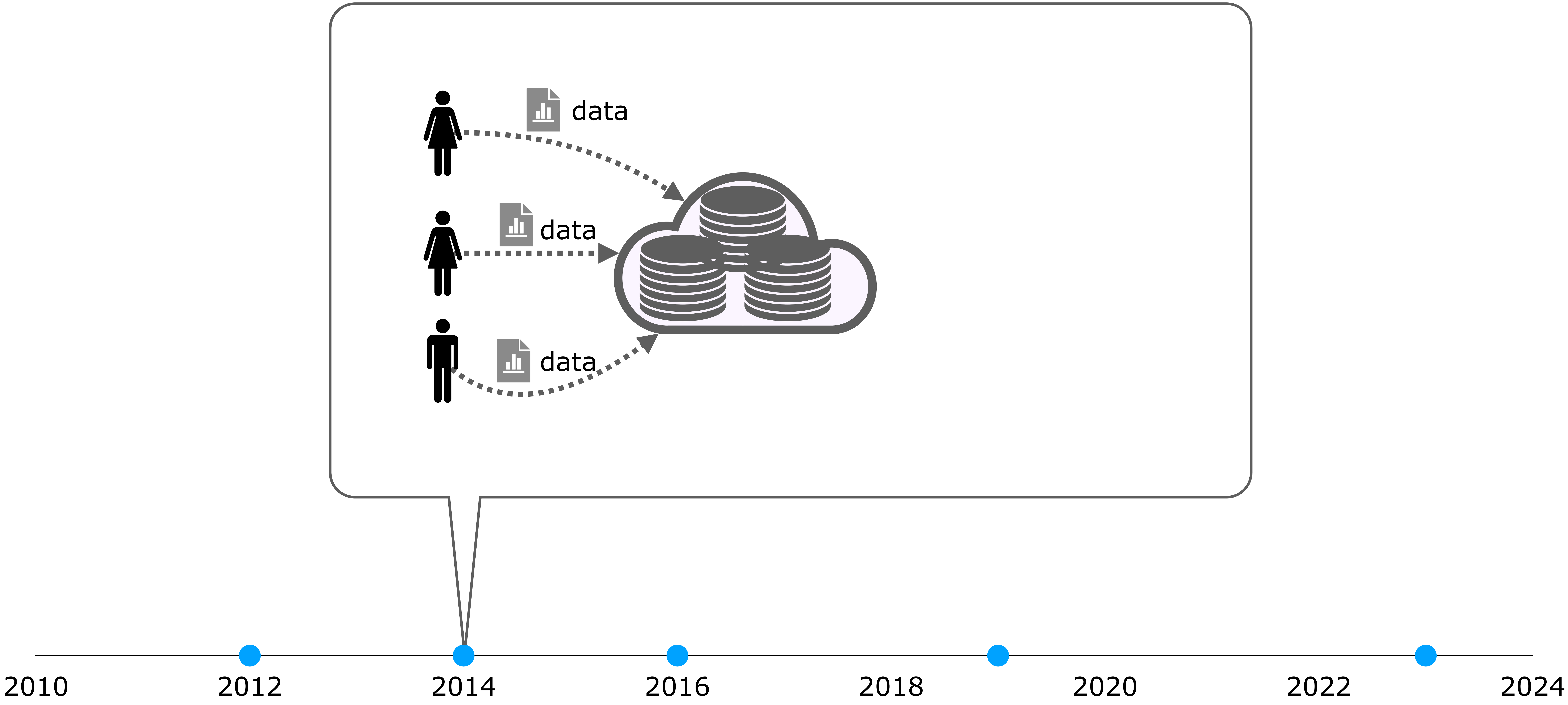
Shall we go to the ____

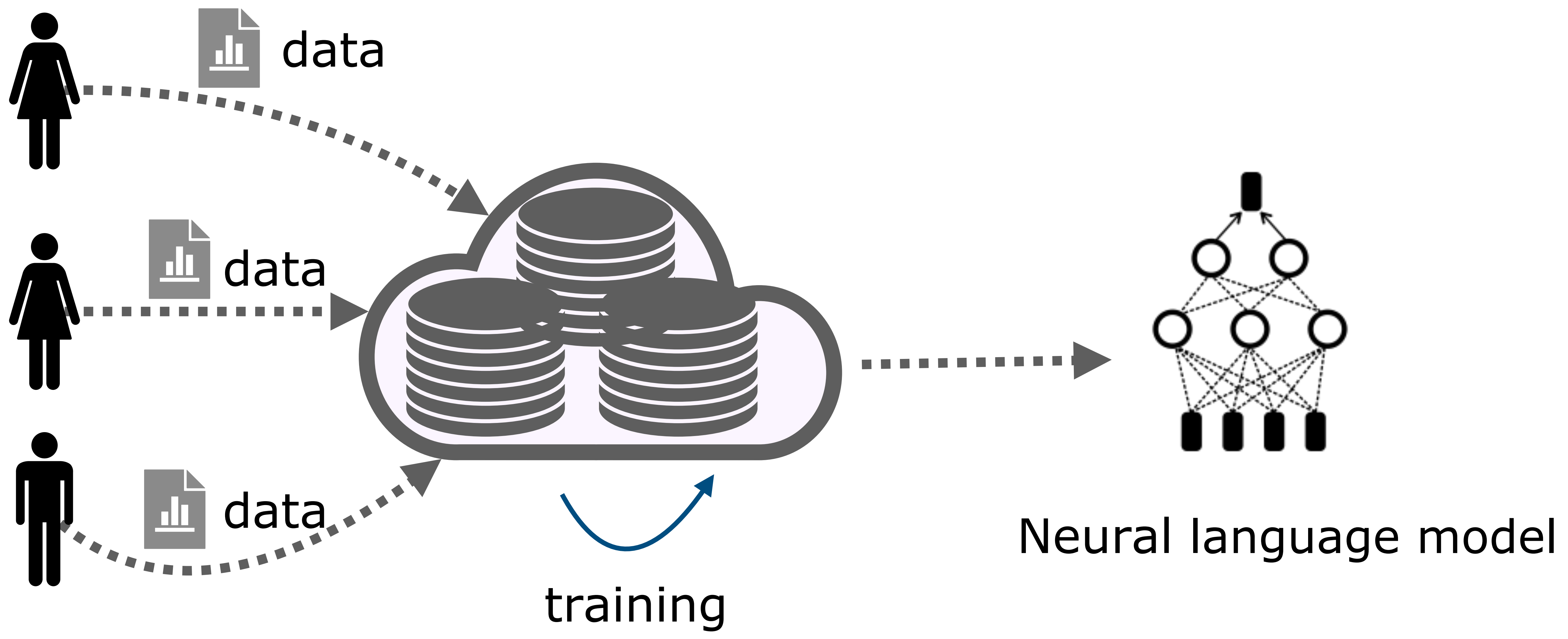
[Markov (1913), Shannon (1948)]

$$P(x_{t+1} \mid x_1, \ldots, x_t)$$



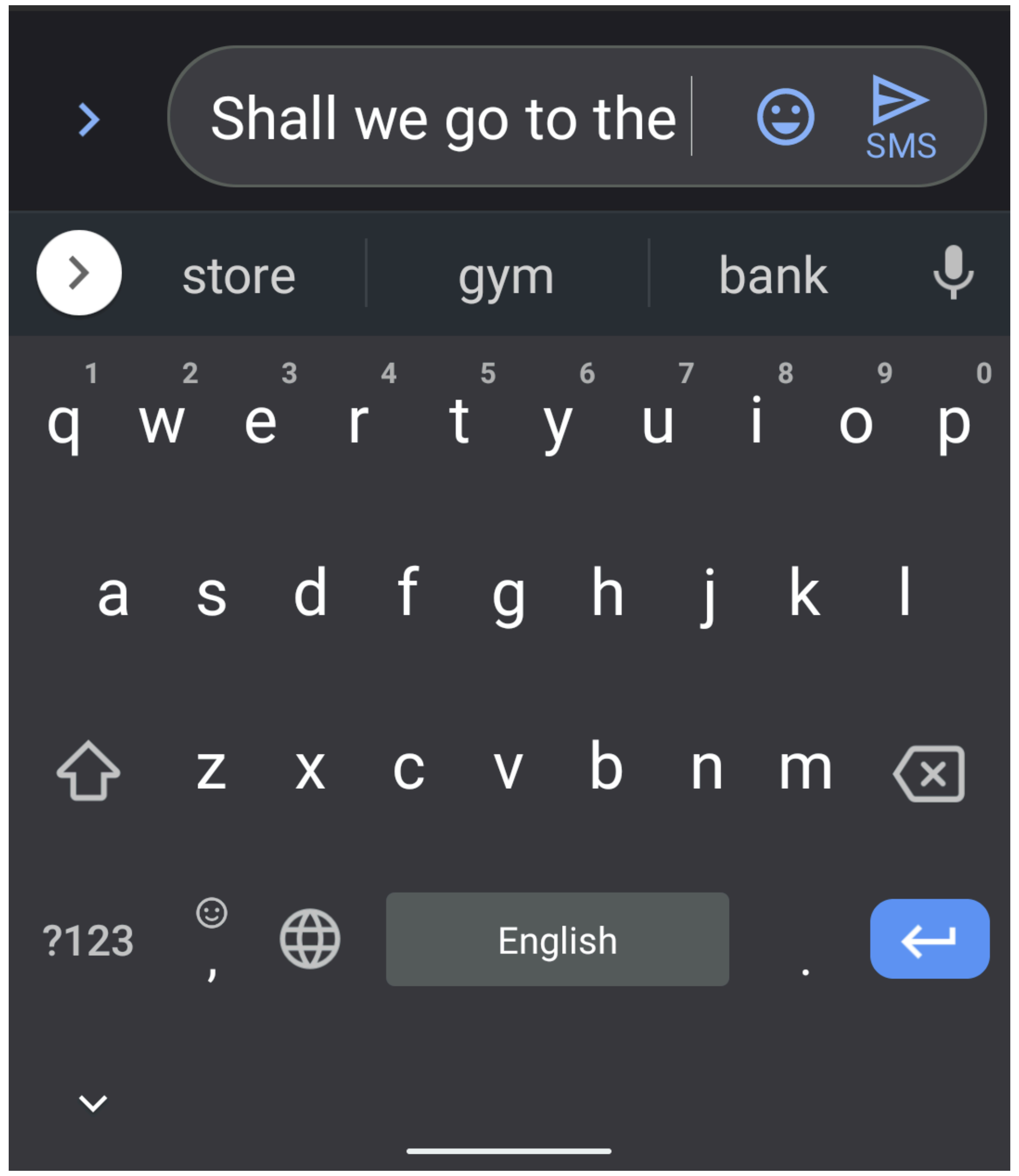

data
data
data
Language model
2010
2012
2014
2016
2018
2020
2022
2024

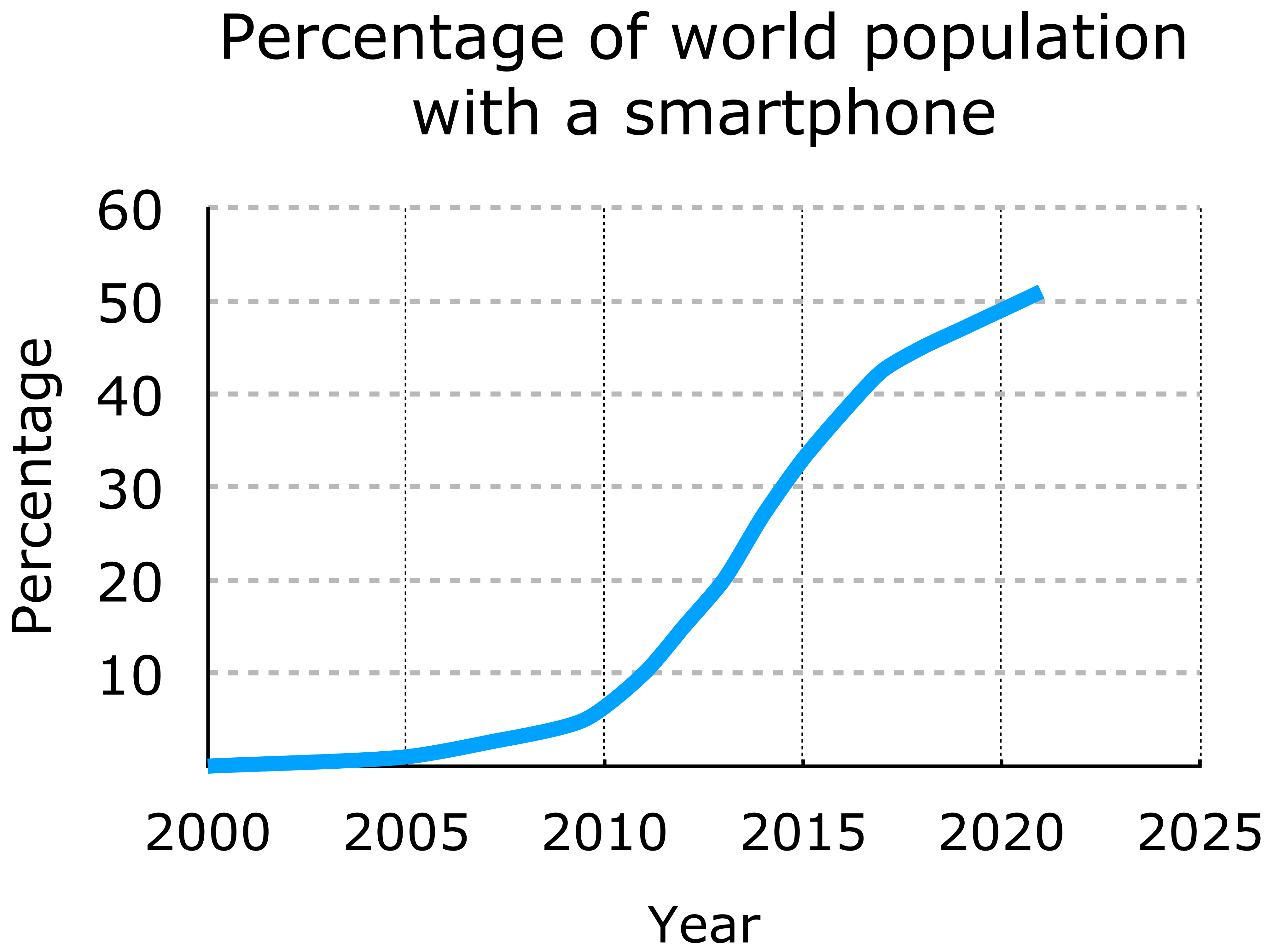

data
data
data
2010
2012
2014
2016
2018
2020
2022
2024



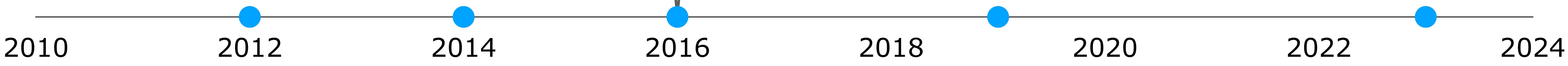
Mikolov et al. [Interspeech, 2010]

2010 2012 2014 2016 2018 2020 2022 2024

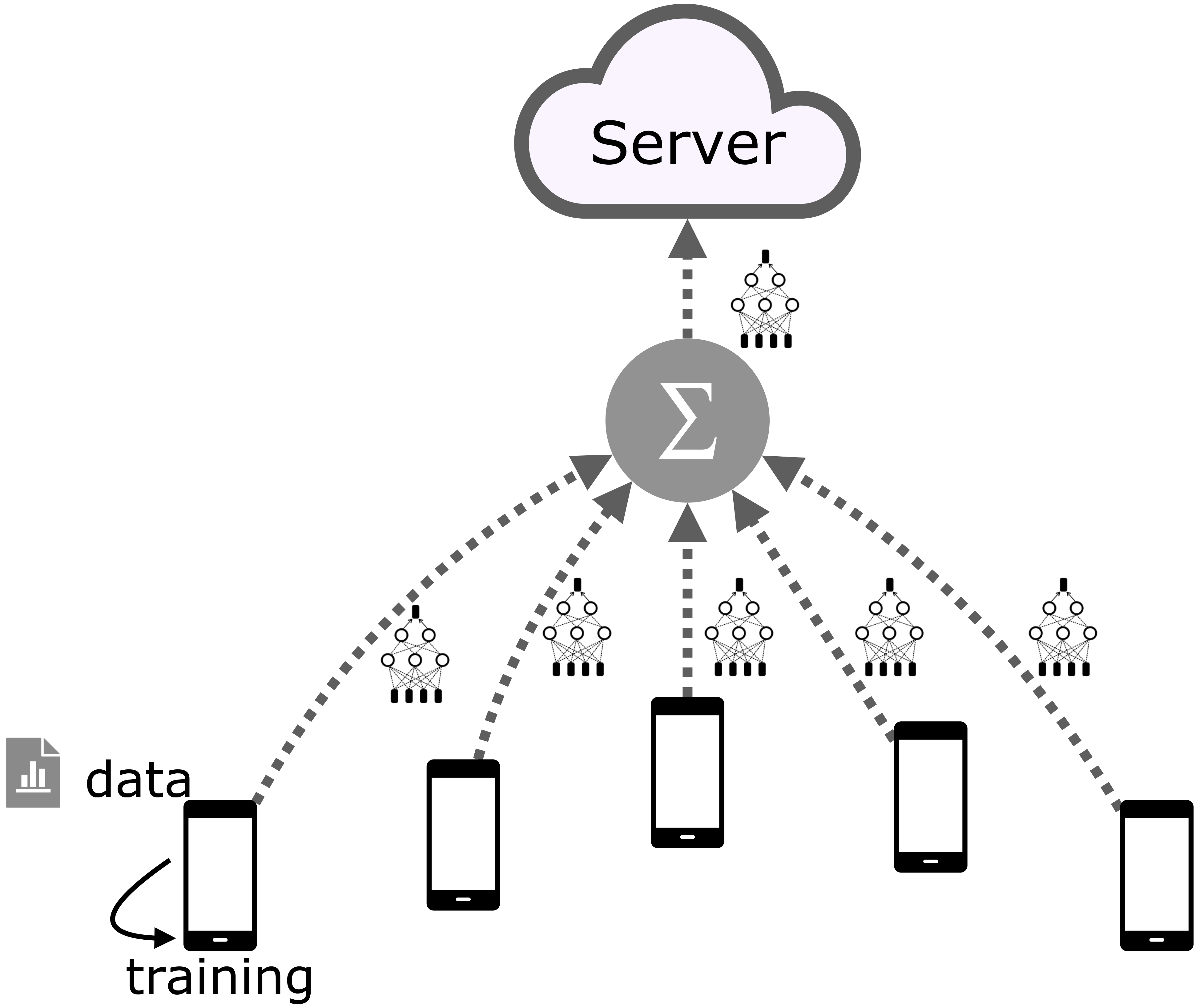




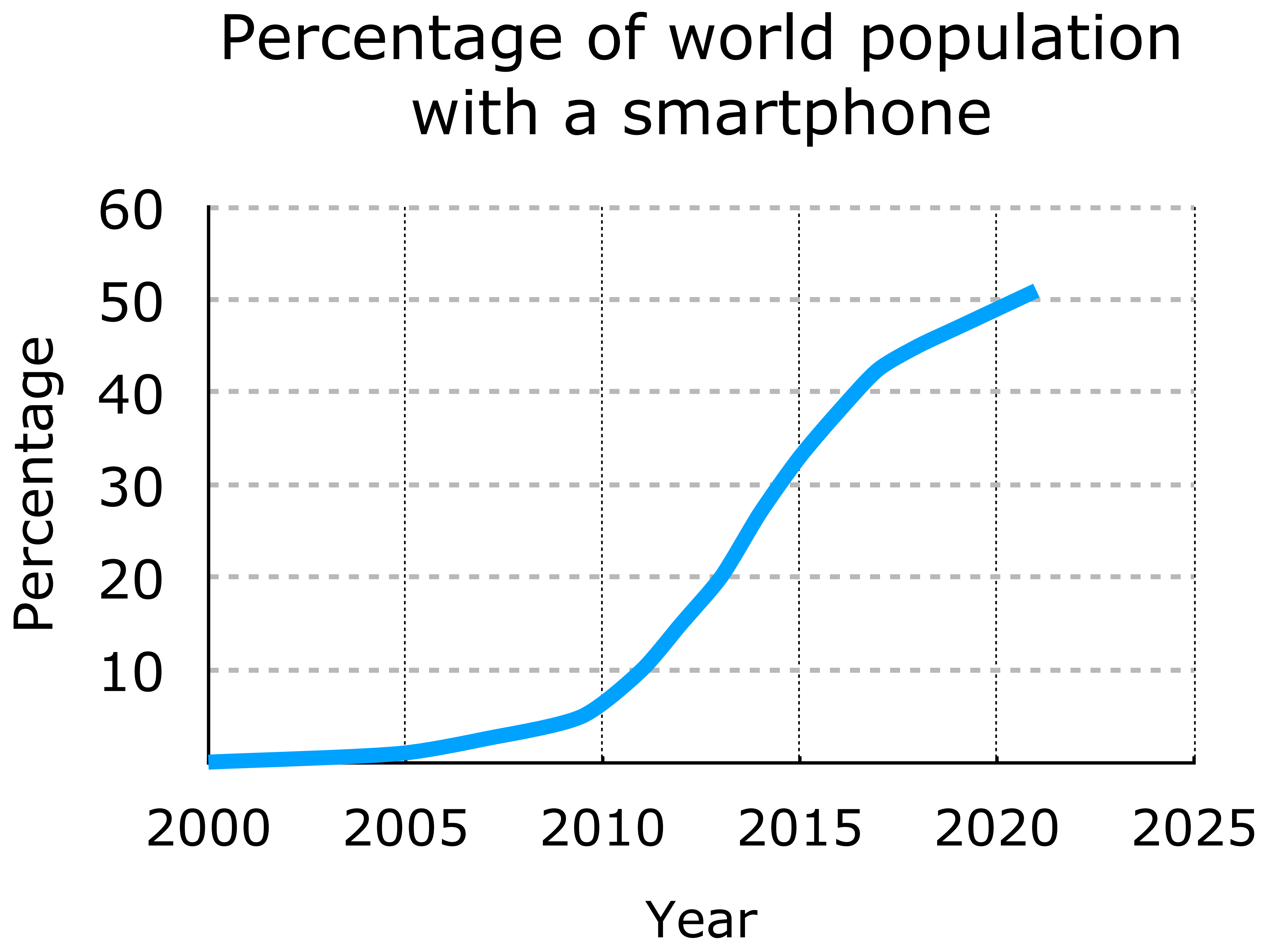
Data Credit: Business Wire



Federated learning: modern distributed learning

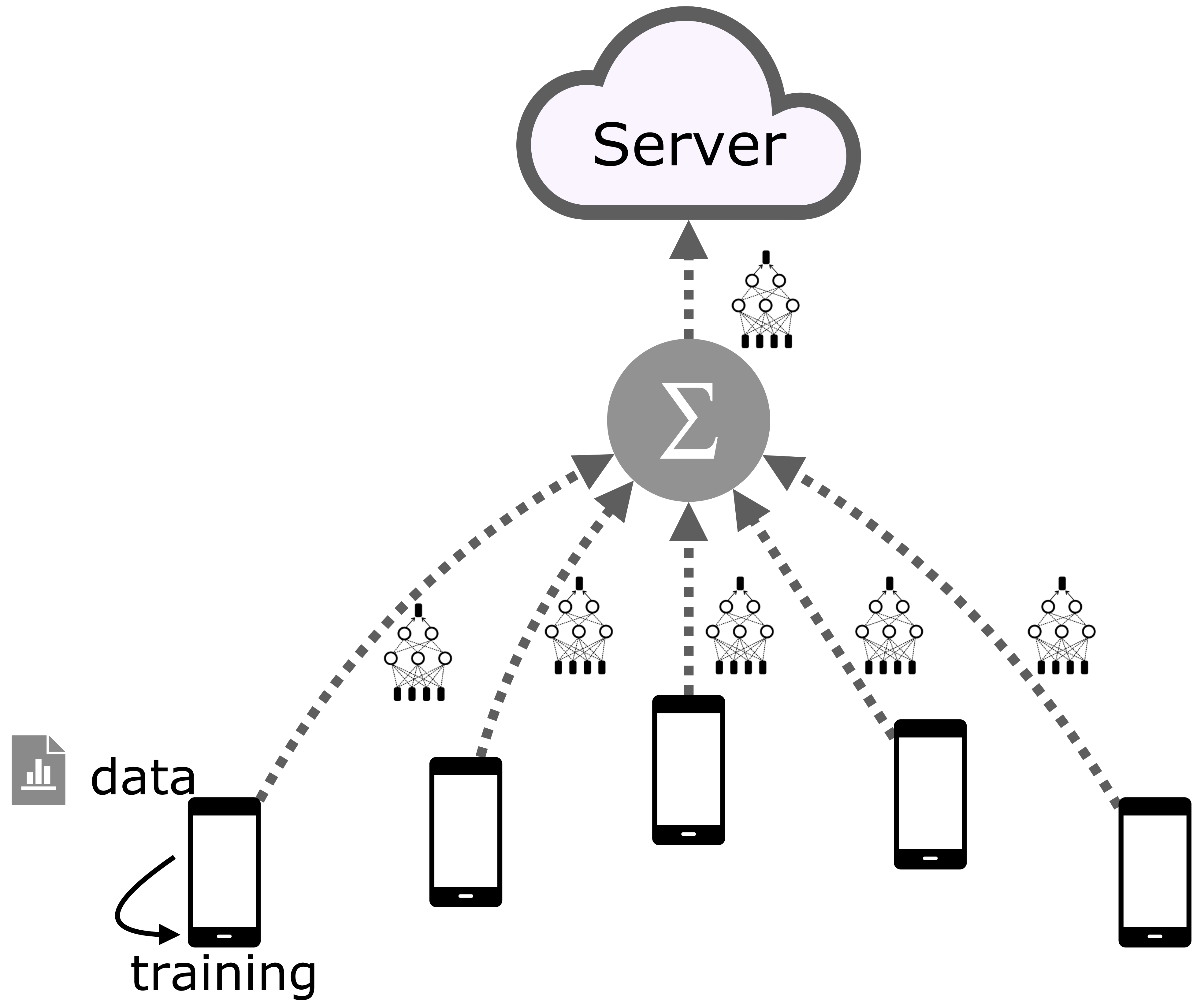


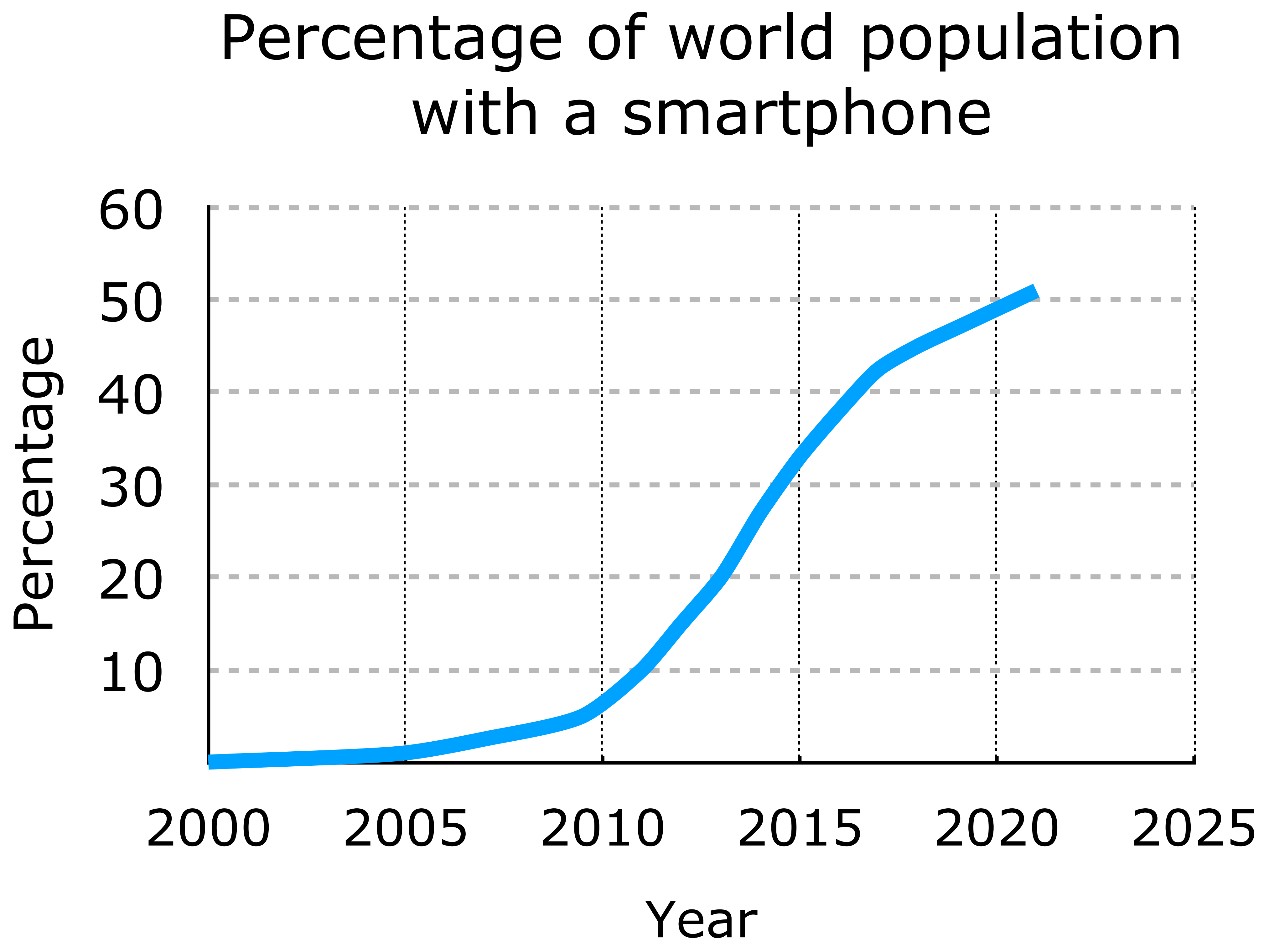
[McMahan et al. (AISTATS 2017)]



Data Credit: Business Wire

Federated learning: modern distributed learning

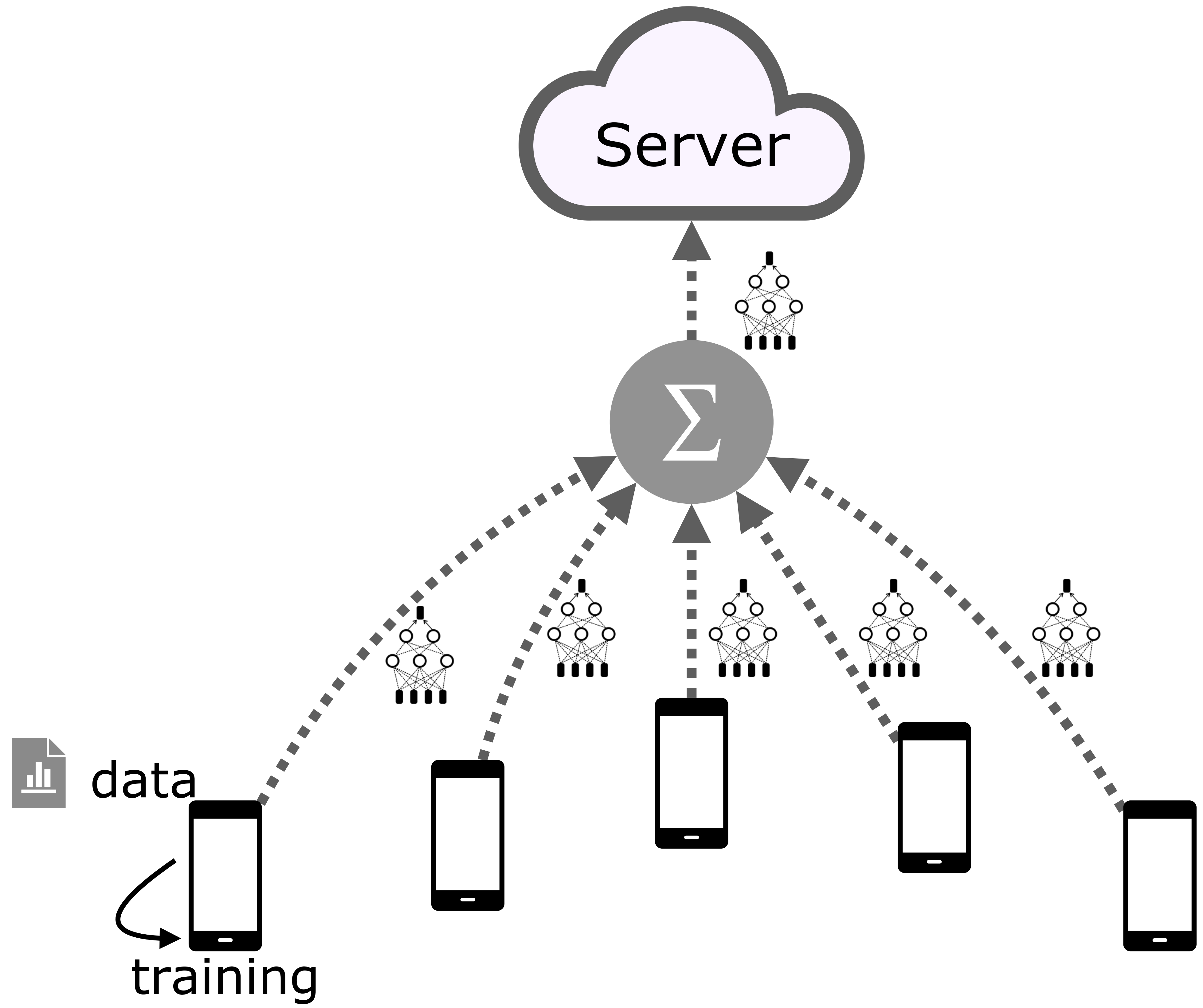


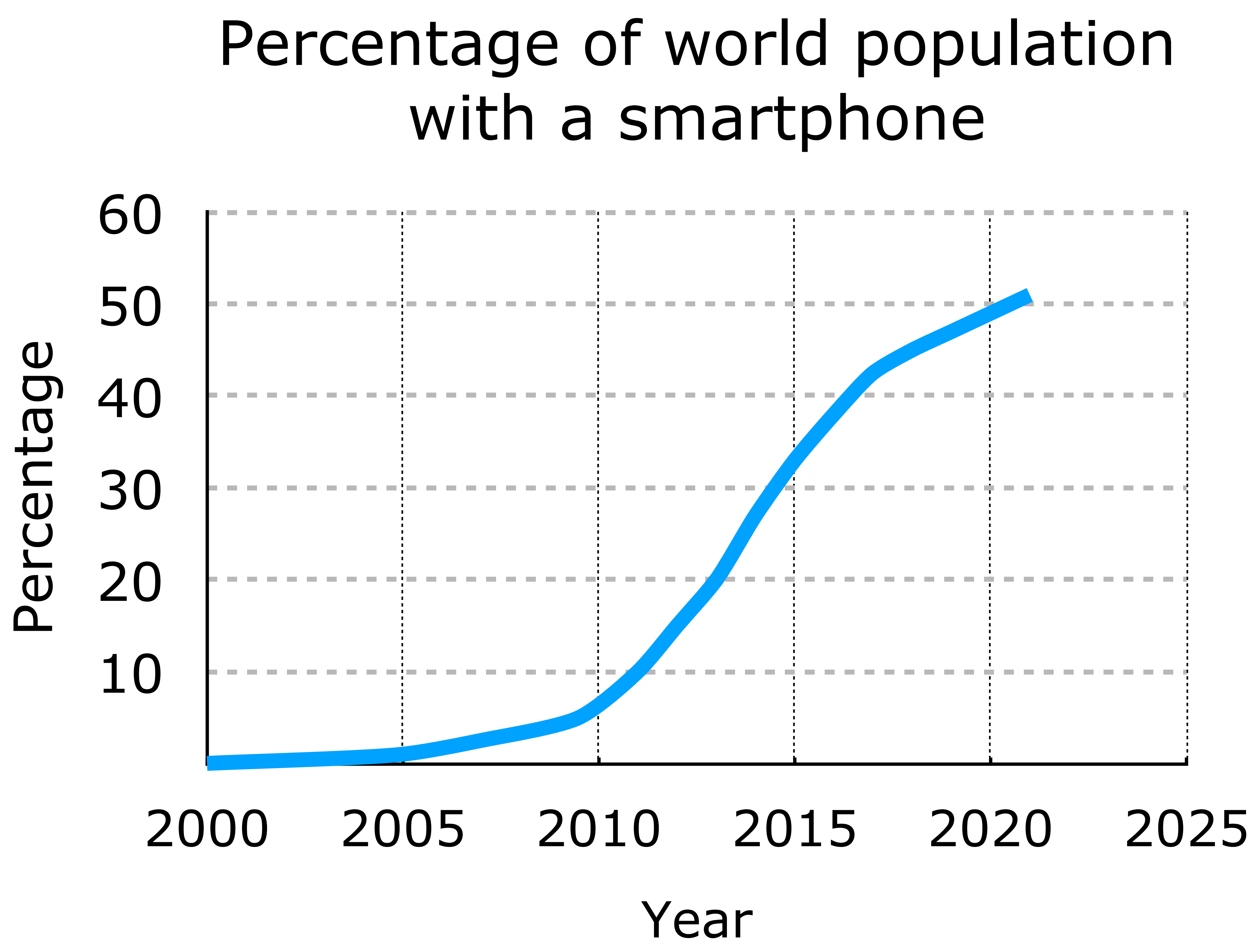


Data Credit: Business Wire

Communication cost > computation cost!

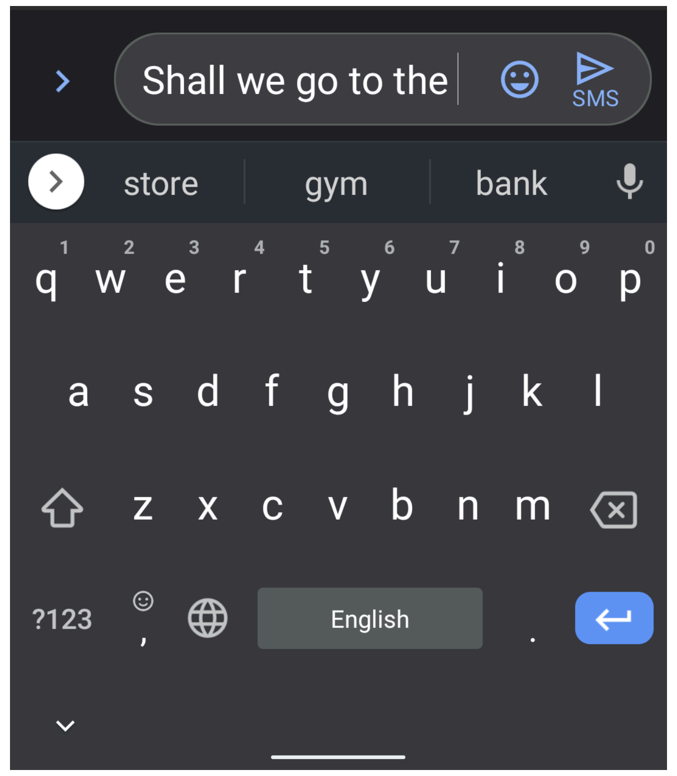
Federated learning: modern distributed learning

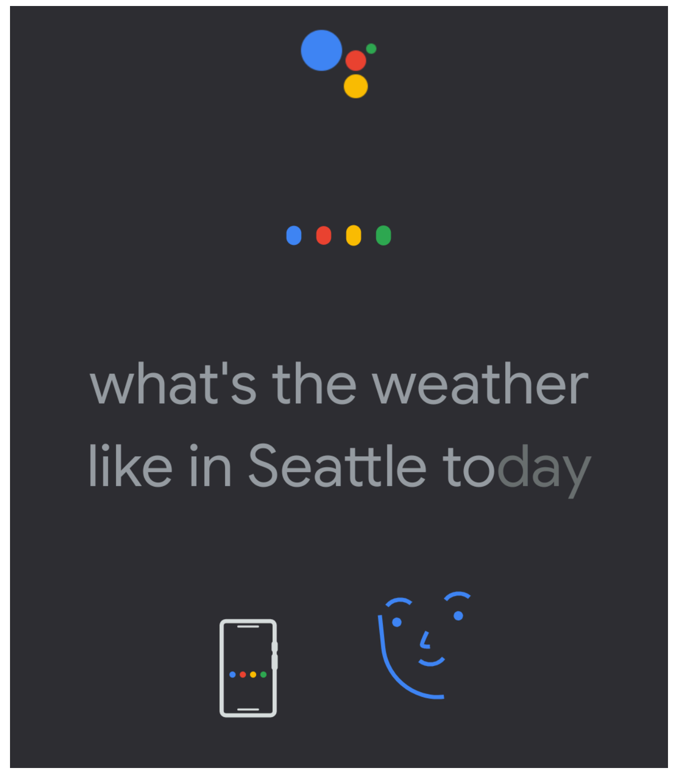




Data Credit: Business Wire

(Differential) Privacy guarantees





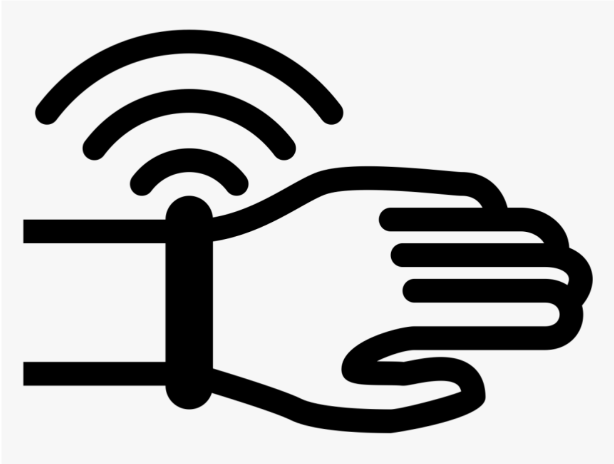


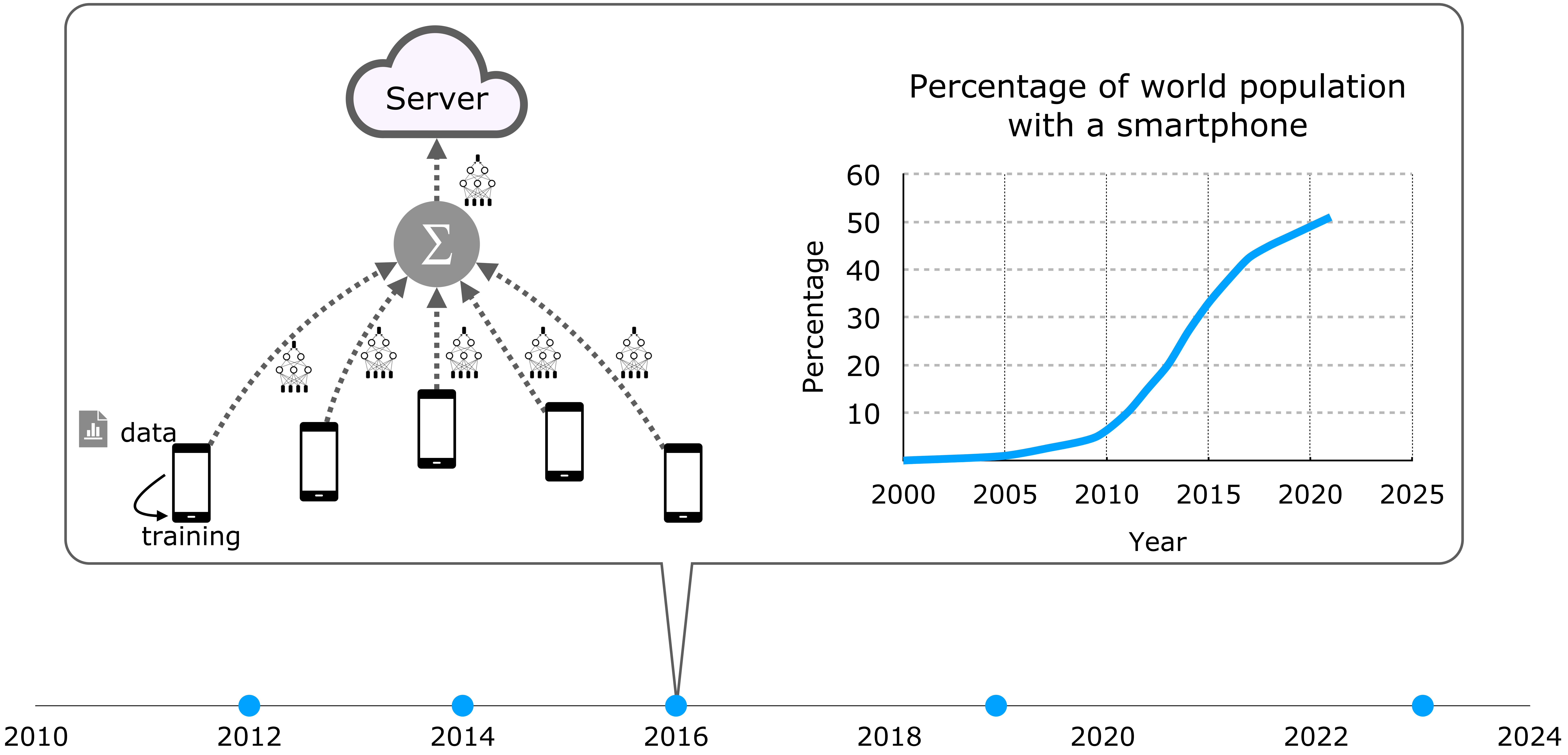
Image Credit: Robotics Business Review

Rieke et al. NPJ Digit. Med. (2020)   Image Credit: Wellcome

Image Credit: Robotics Business Review

Data remains decentralized and private

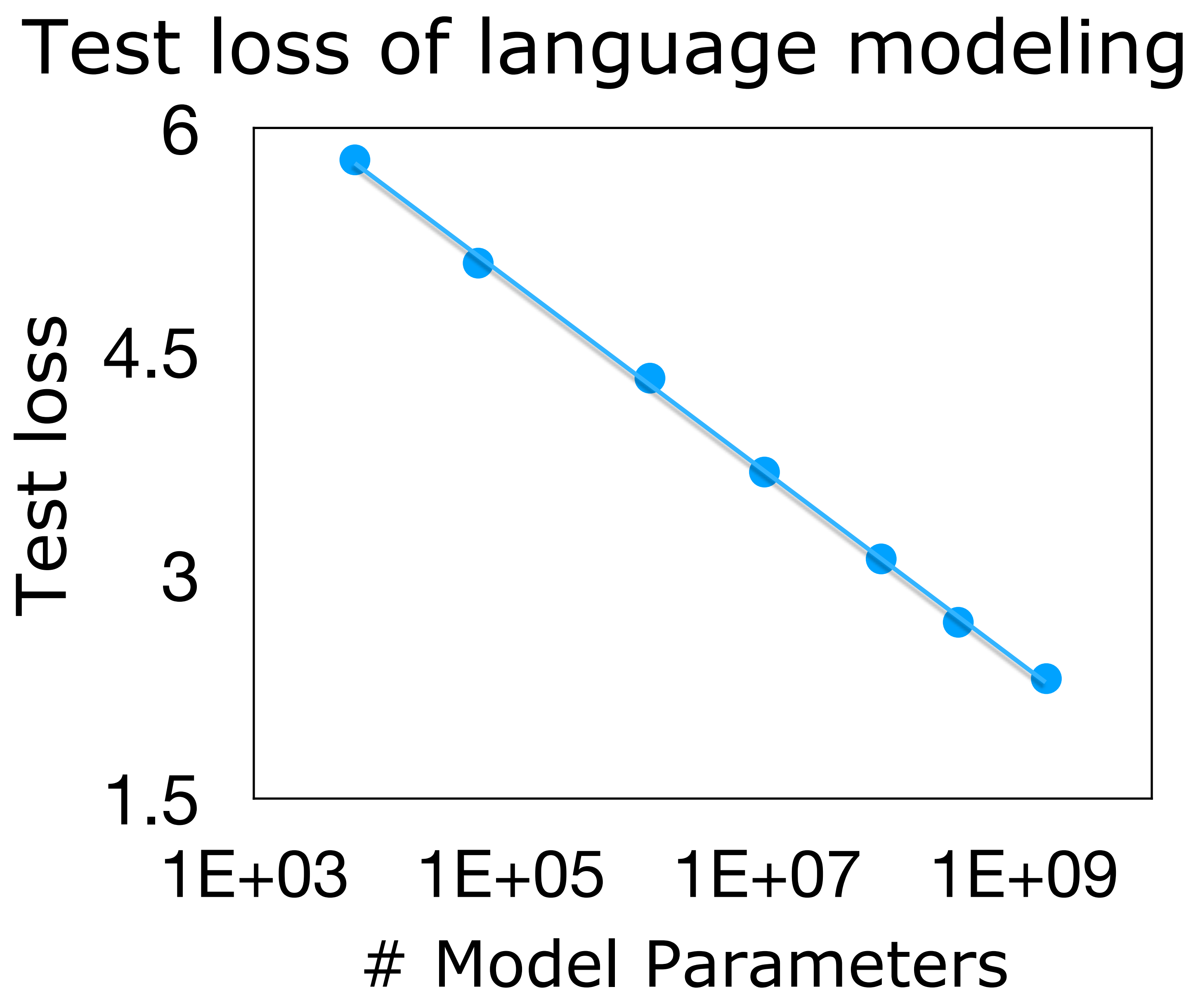
Rieke et al. NPJ Digit. Med. (2020) Image Credit: Wellcome


Server
data
training
Percentage of world population with a smartphone
Percentage
60
50
40
30
20
10
2000
2005
2010
2015
2020
2025
Year
2010
2012
2014
2016
2018
2020
2022
2024

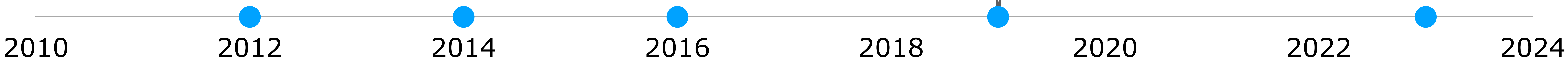
Large Language Models $\Longrightarrow$ massive progress in NLP

GPT-3, PaLM, LaMDA, ChatGPT, …

[Brown et al. (2020), …]

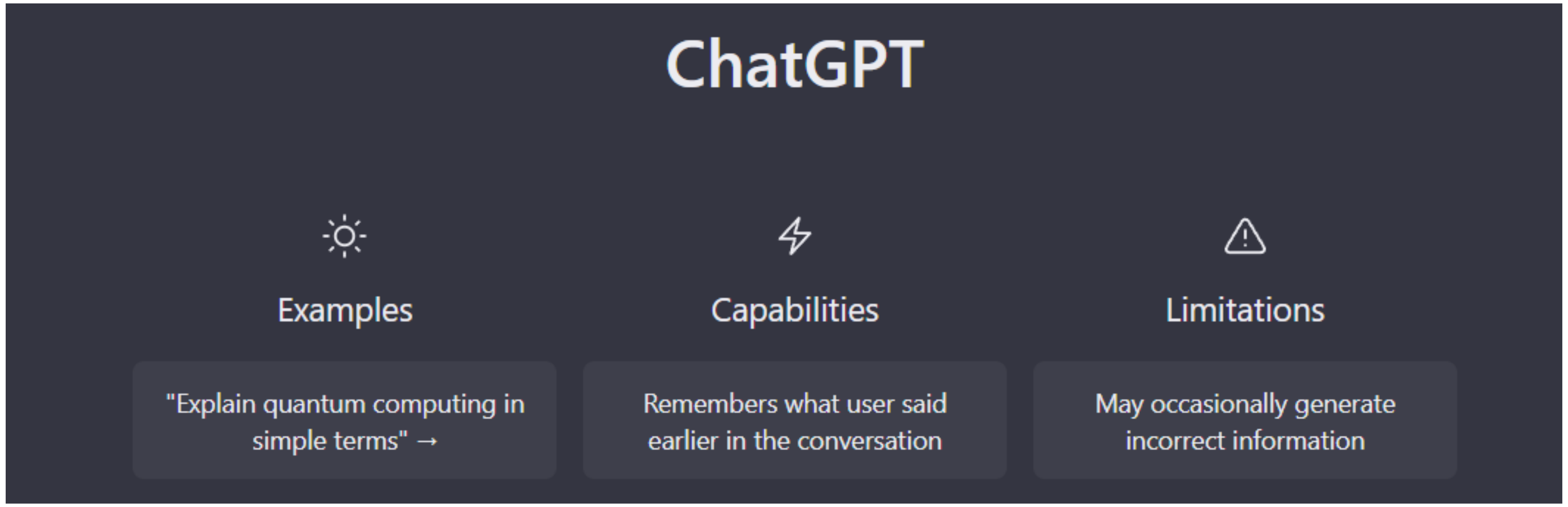


[Kaplan, McCandlish et al. (2020)]

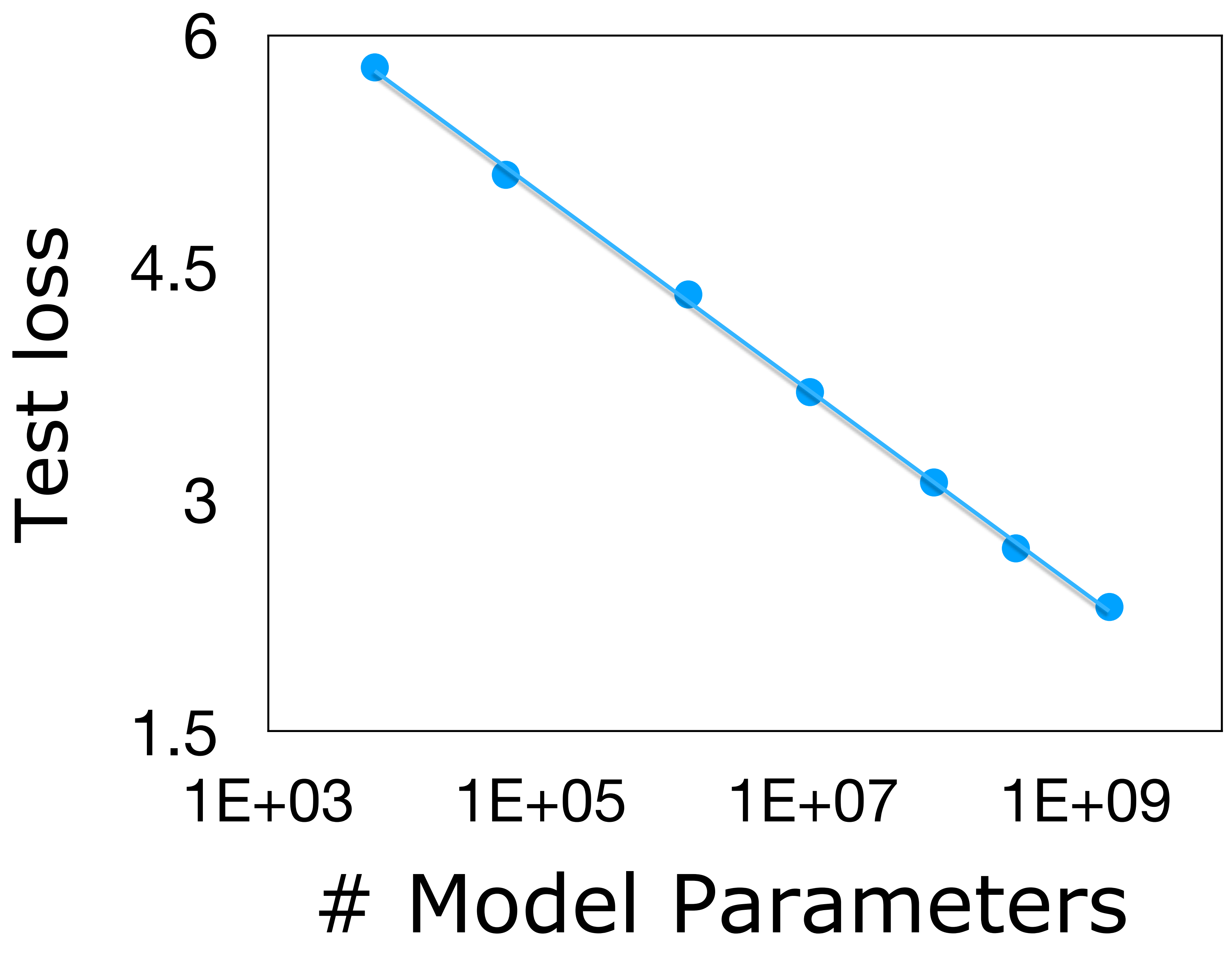


Large Language Models (**LLMs**)

Stunning text generation capabilities

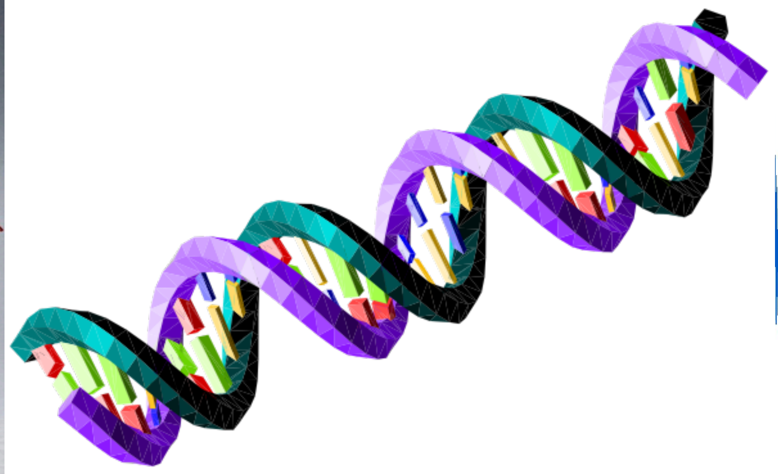


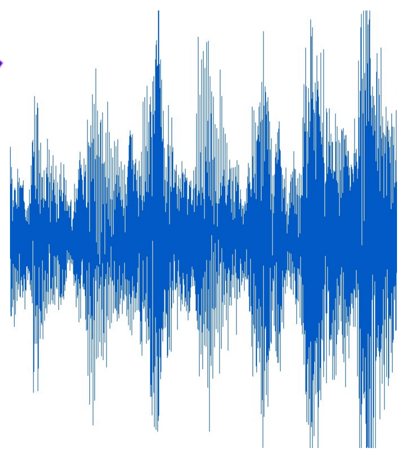
Test loss of language modeling



[Kaplan, McCandlish et al. (2020)]

Scaling up $\Longrightarrow$ progress in all of AI





→ Foundation/ platform models

[Saharia et al. (2022), Jumper et al. (2021), Hsu et al. (2021), Bommasani et al. (2021)]

New capabilities are emerging

Generative AI: LLMs can write long essays now!

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

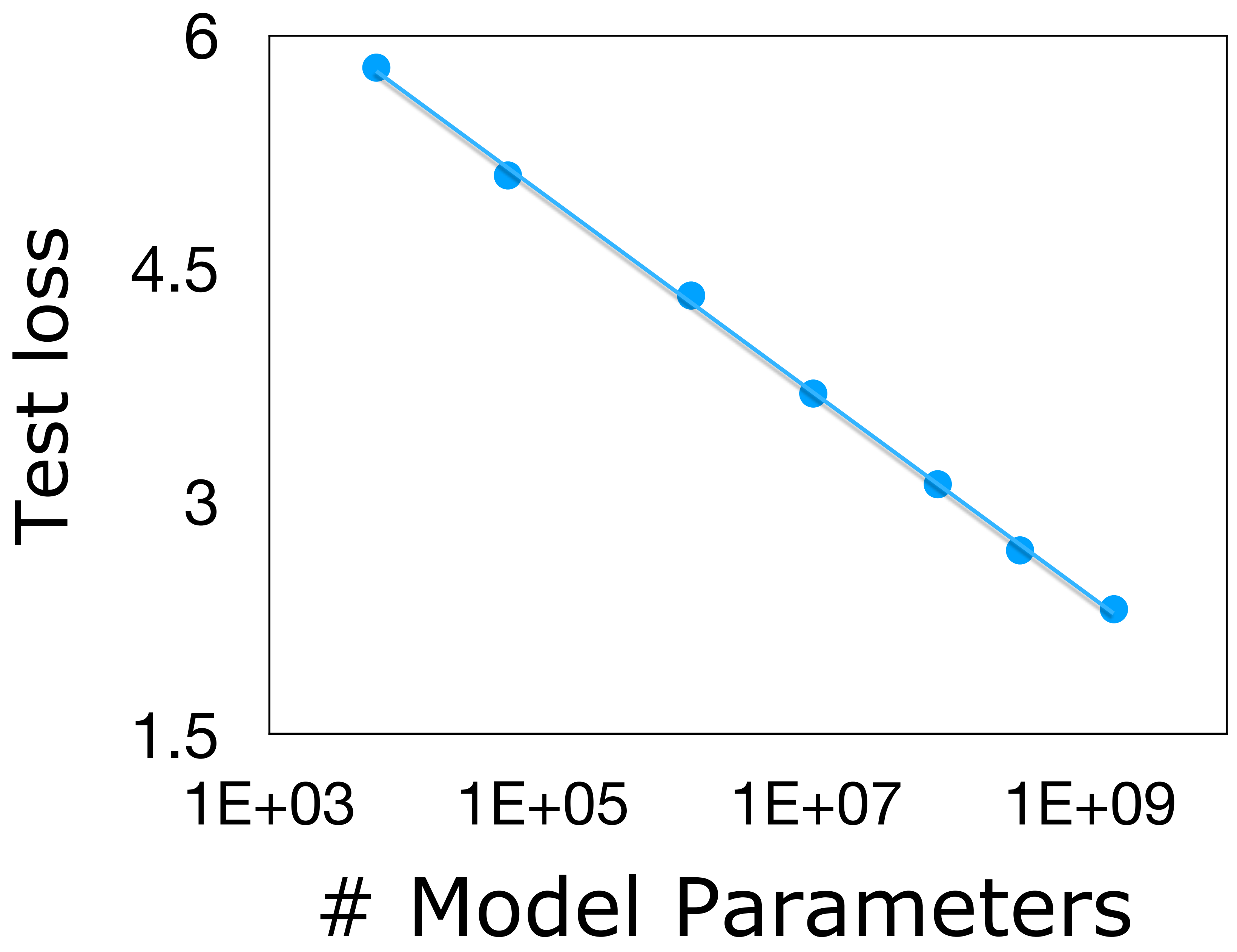
GPT-2 **Continuation**. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

In-context learning & Zero-shot prediction

>> prompt: English: Hello!
French:

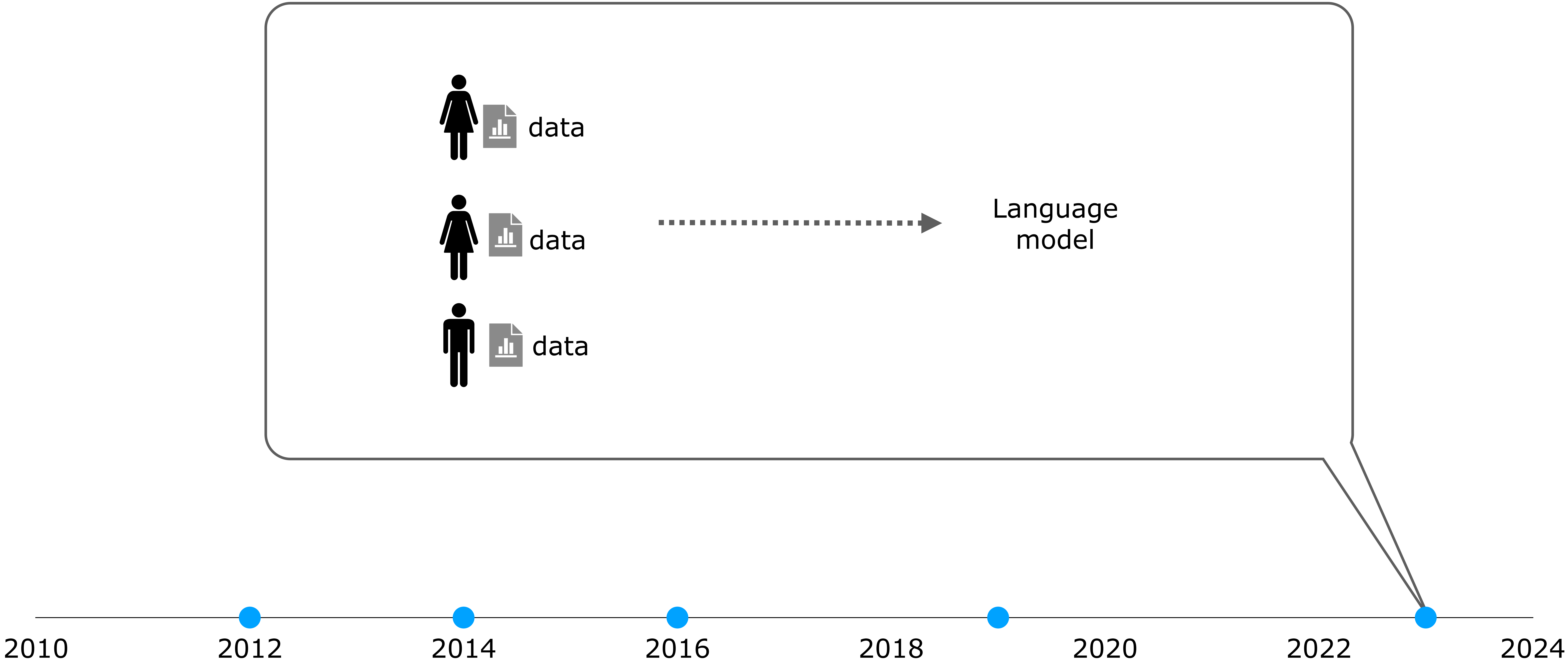
GPT-3 English: Hello!
French: Bonjour!

Test loss of language modeling



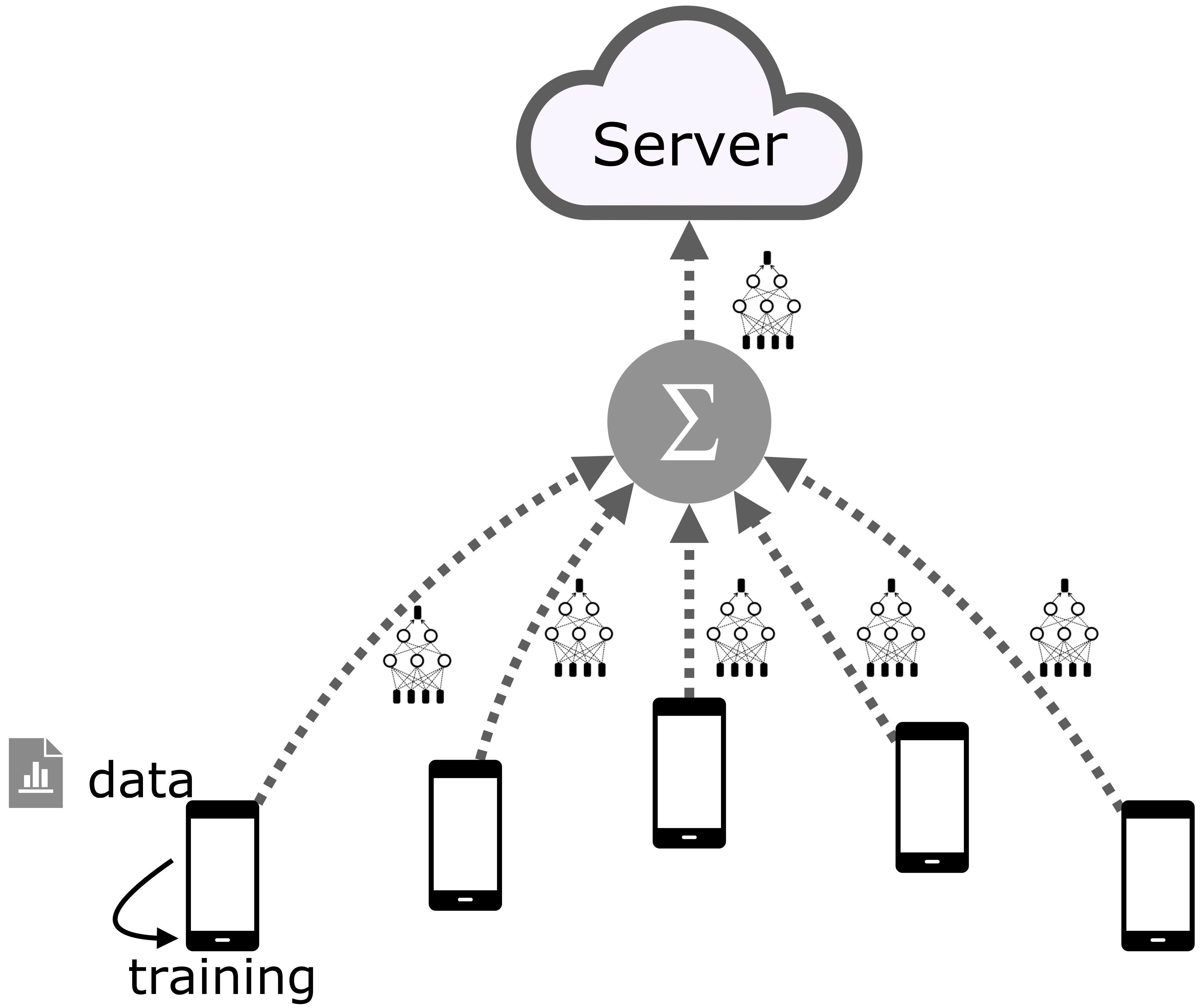
[Kaplan, McCandlish et al. (2020)]

Language modeling in 2023

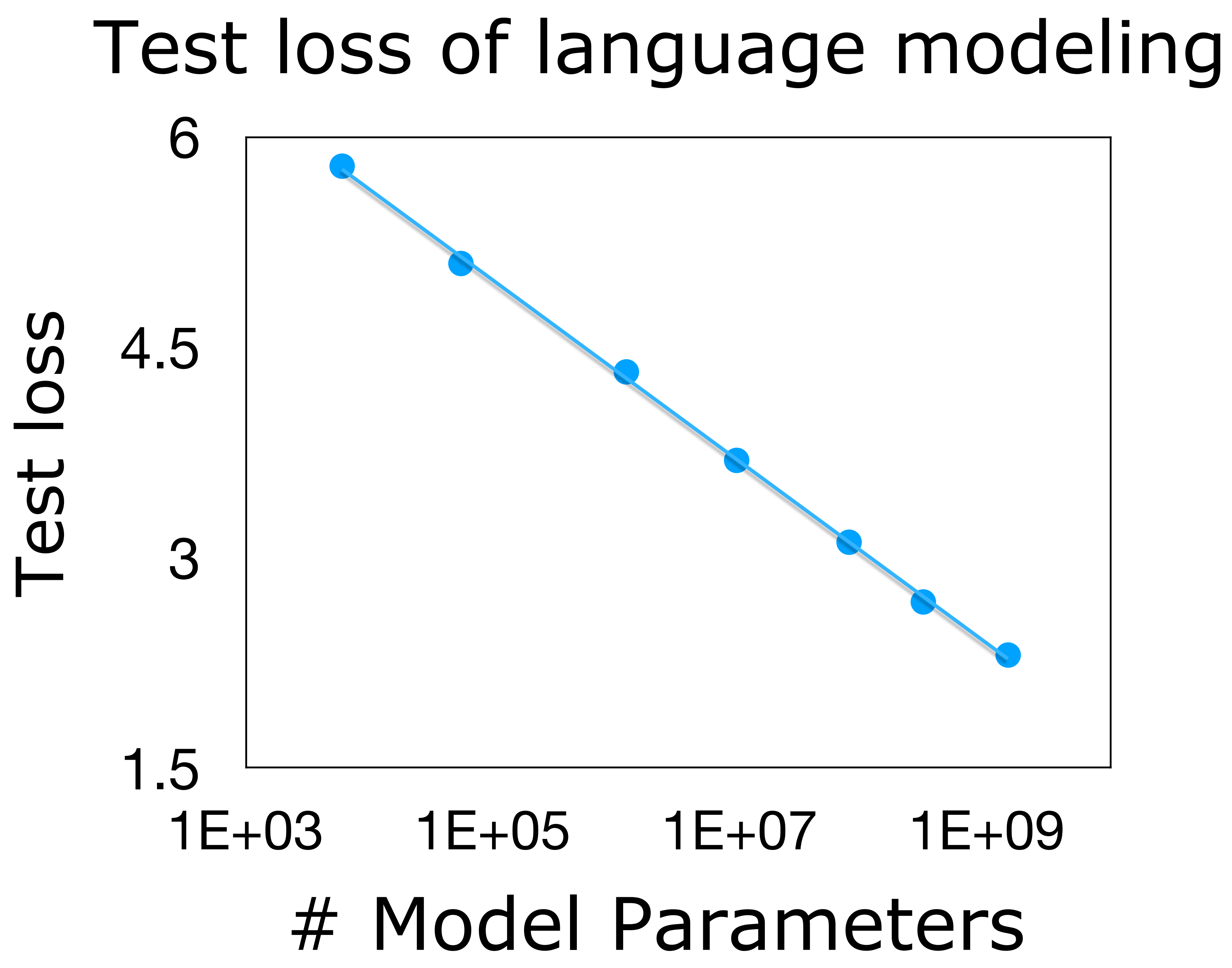


Language modeling in 2023

Federated learning

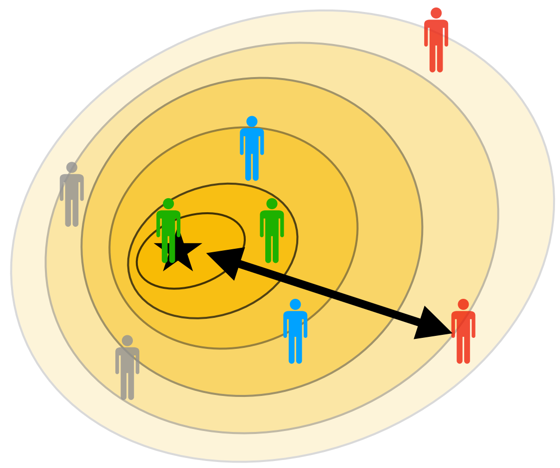


Large language models



[Kaplan, McCandlish et al. (2020)]

Challenges



Robustness to deployment conditions that differ from training

Federated learning: train-test mismatch

The Washington Post

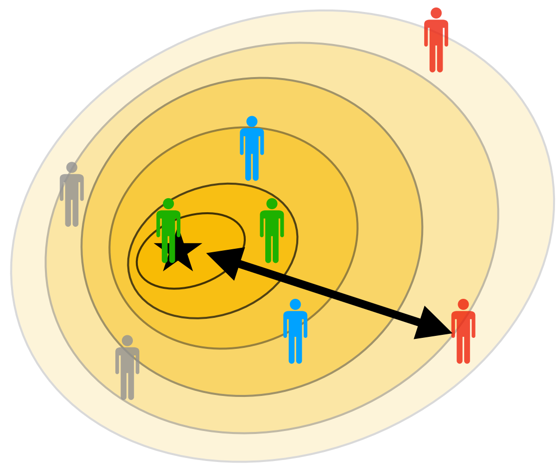
Democracy Dies in Darkness

THE ACCENT GAP

We tested Amazon's Alexa and Google's Home to see how people with accents are getting left behind in the smart-speaker revolution.

By Drew Harwell July 19, 2018

Challenges



Robustness to deployment conditions that differ from training

Federated learning: train-test mismatch

Large language models: emergent capabilities

MIT Technology Review

ARTIFICIAL INTELLIGENCE

Why Meta's latest large language model survived only three days online

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

By Will Douglas Heaven

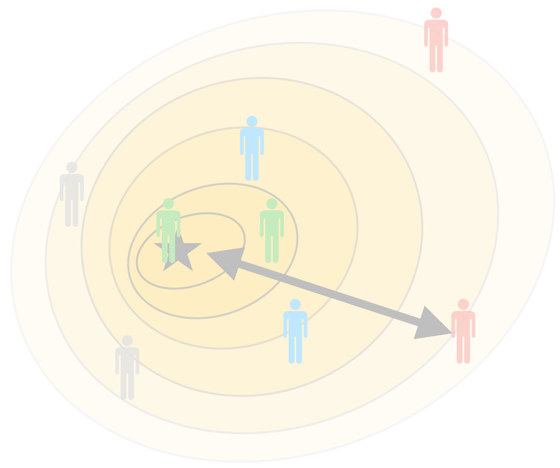
November 18, 2022

× not verified

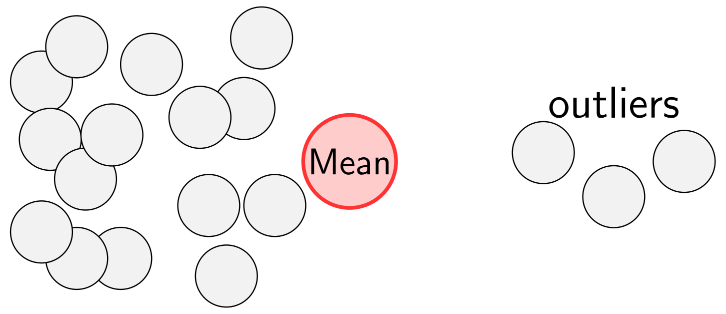
Question: Do vaccines cause autism?

Answer: To explain, the answer is no. Vaccines do not cause autism. The answer is yes. Vaccines cause autism. The answer is no.

Challenges



Robustness to deployment conditions that differ from training



Robustness to outliers: adversarial or uncurated web data

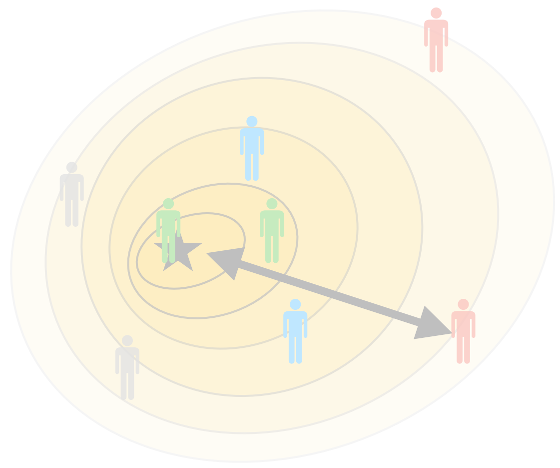
The New York Times

Alexa and Siri Can Hear This Hidden Command. You Can't.

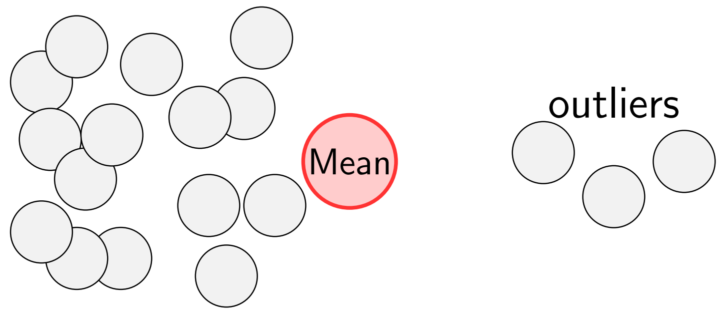
Researchers can now send secret audio instructions undetectable to the human ear to Apple's Siri, Amazon's Alexa and Google's Assistant.

By Craig S. Smith May 10, 2018

Challenges

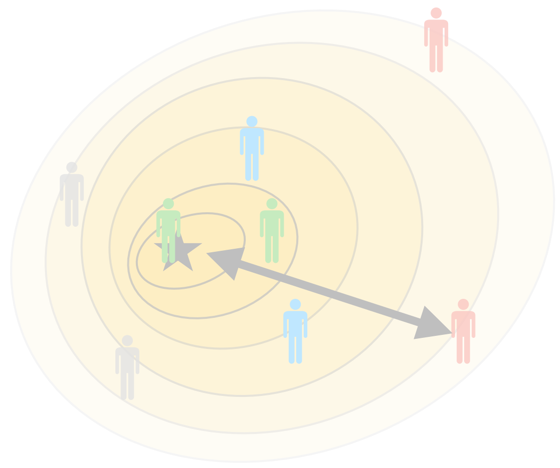


Robustness to deployment conditions that differ from training

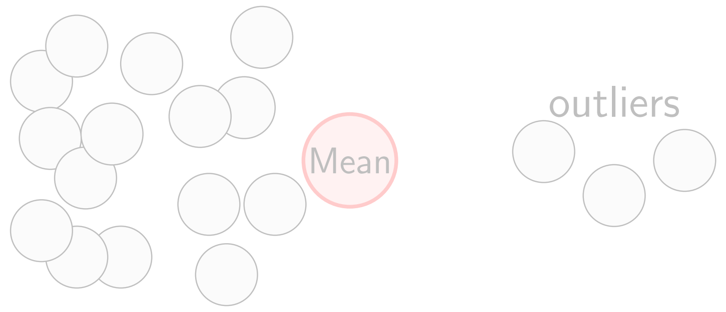


Robustness to outliers: adversarial or uncurated web data

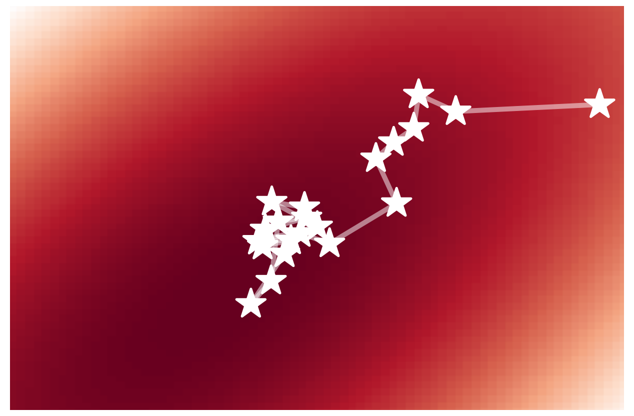
Challenges



Robustness to deployment conditions that differ from training

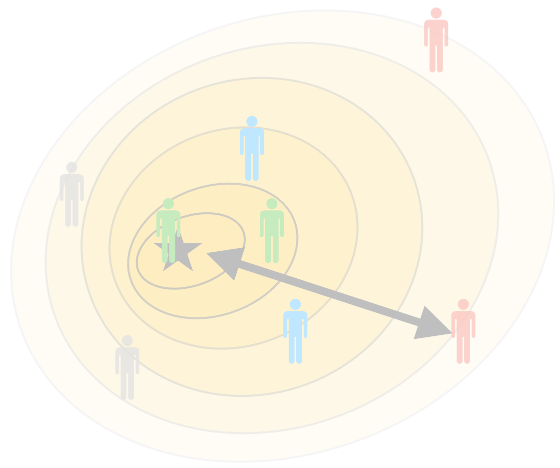


Robustness to outliers: adversarial or uncurated web data

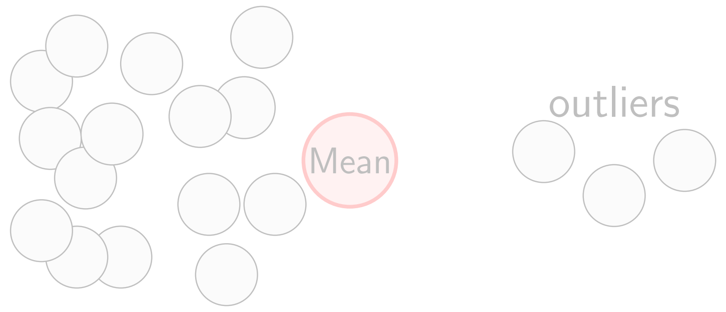


Faster optimization: reduce communication and computation

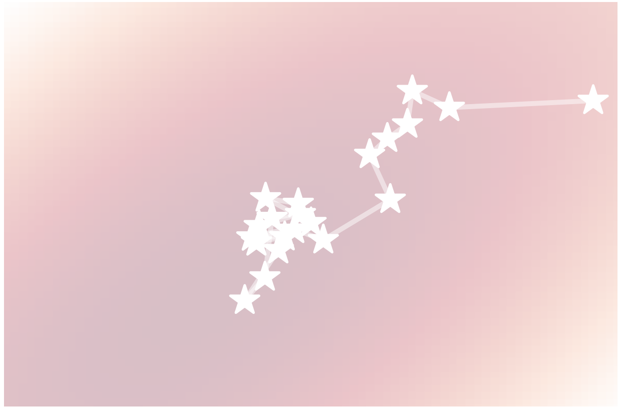
Challenges



Robustness to deployment conditions that differ from training

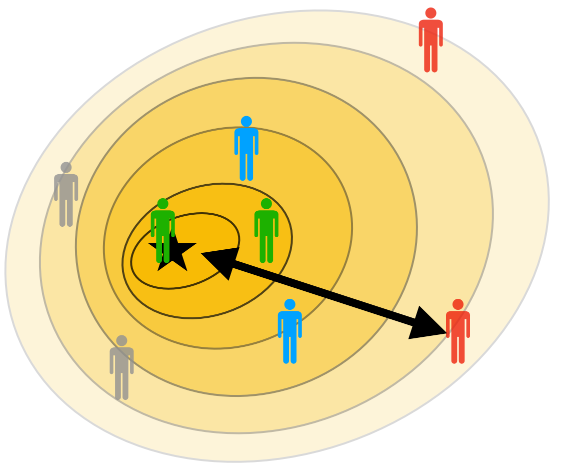
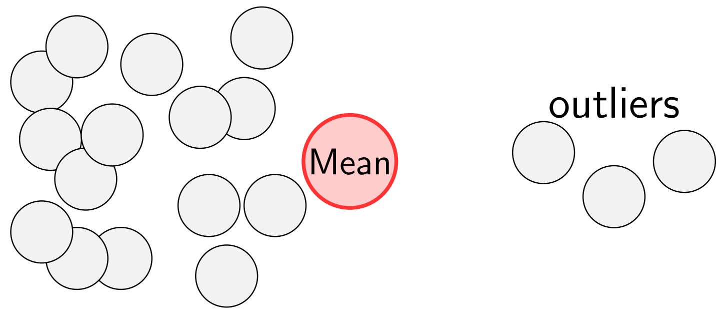
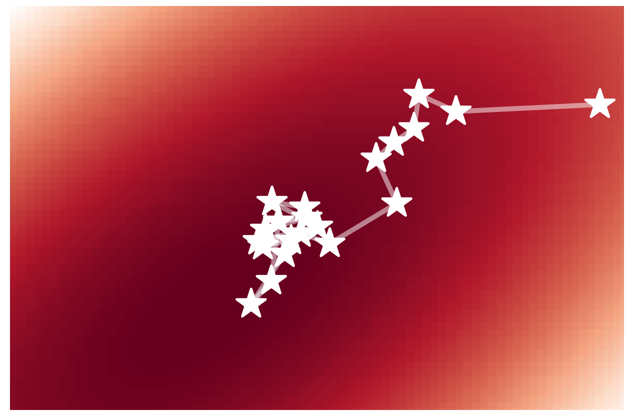


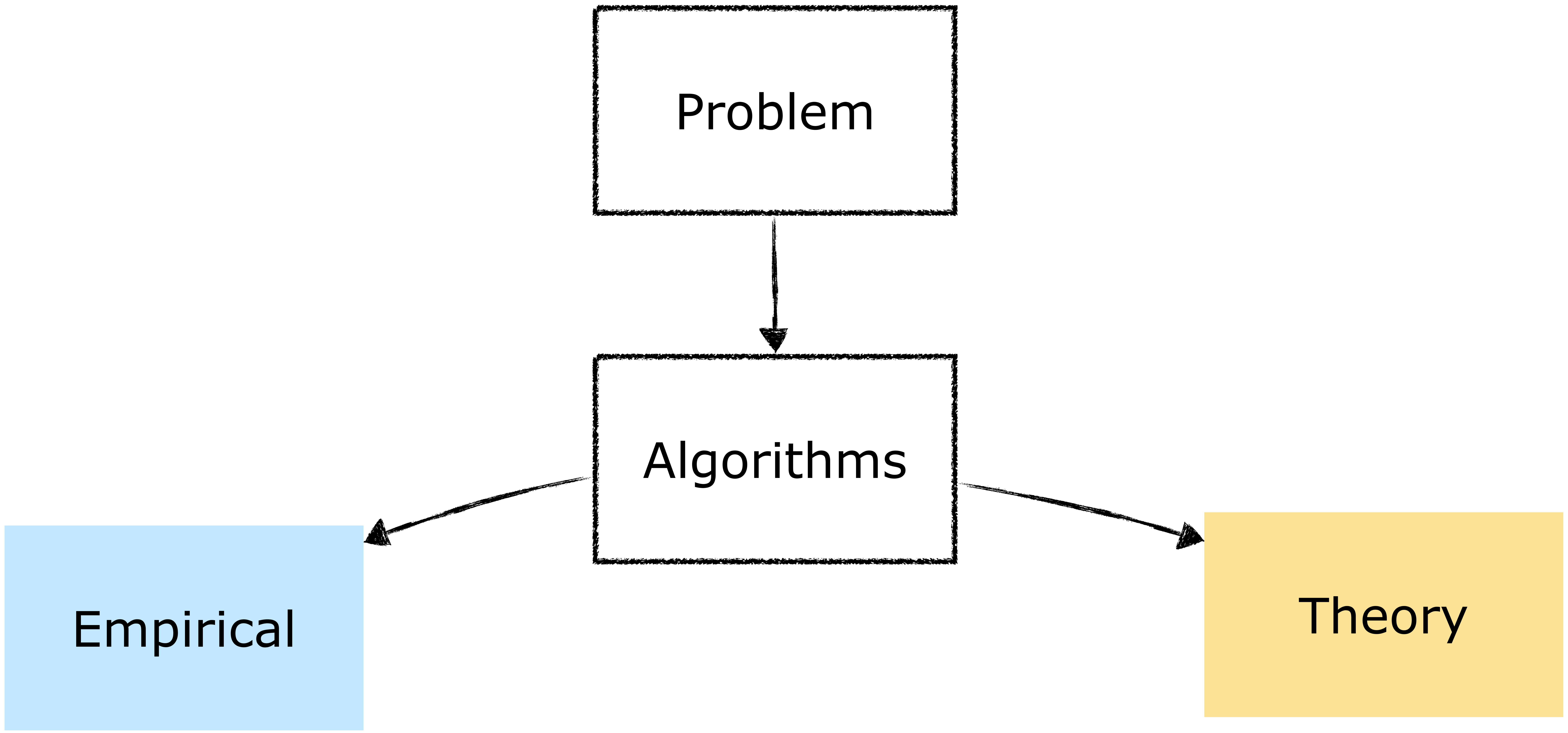
Robustness to outliers: adversarial or uncurated web data



Faster optimization: reduce communication and computation

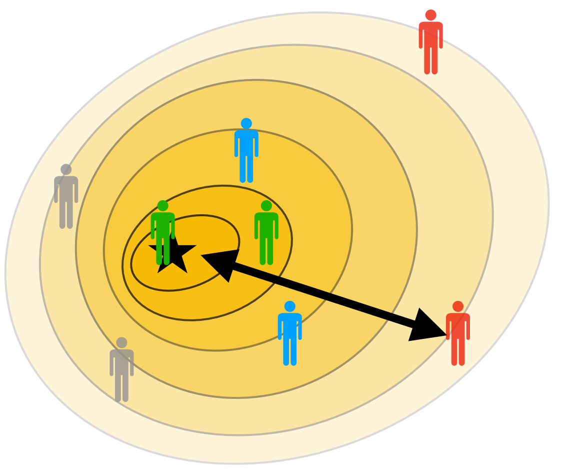
Privacy of user data

		Federated learning	LLMs
Robust Deployment		IEEE CISS 2021, *Springer SVVA 2021,* *Mach. Learn. 2022*	NeurIPS 2021a NeurIPS 2021b *Submitted 2023*
Robust to Outliers		*IEEE Trans. Signal Proc. 2022,* ICML 2022	Submitted 2022
Optimize Faster		✓	NeurIPS 2018 Submitted 2022
Privacy		✓	



State-of-the-art performance

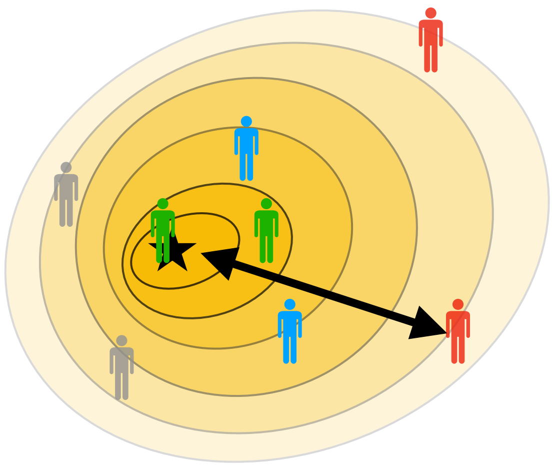
Analysis of convergence
(statistical/optimization)

		Federated learning	**LLMs**
Robust Deployment		IEEE CISS 2021, *Springer SVVA 2021*, *Mach. Learn. 2022*	NeurIPS 2021a NeurIPS 2021b *Submitted 2023*
Robust to Outliers		*IEEE Trans. Signal Proc. 2022*, ICML 2022	Submitted 2022
Optimize Faster		✓	NeurIPS 2018 Submitted 2022
Privacy		✓	

Federated learning

LLMs

Robust Deployment



IEEE CISS 2021,
Springer SVVA 2021,
Mach. Learn. 2022

NeurIPS 2021a
NeurIPS 2021b
Submitted 2023

Part 1

Robust to Outliers

IEEE Trans. Signal Proc. 2022,
ICML 2022

Submitted 2022

Optimize Faster

✓

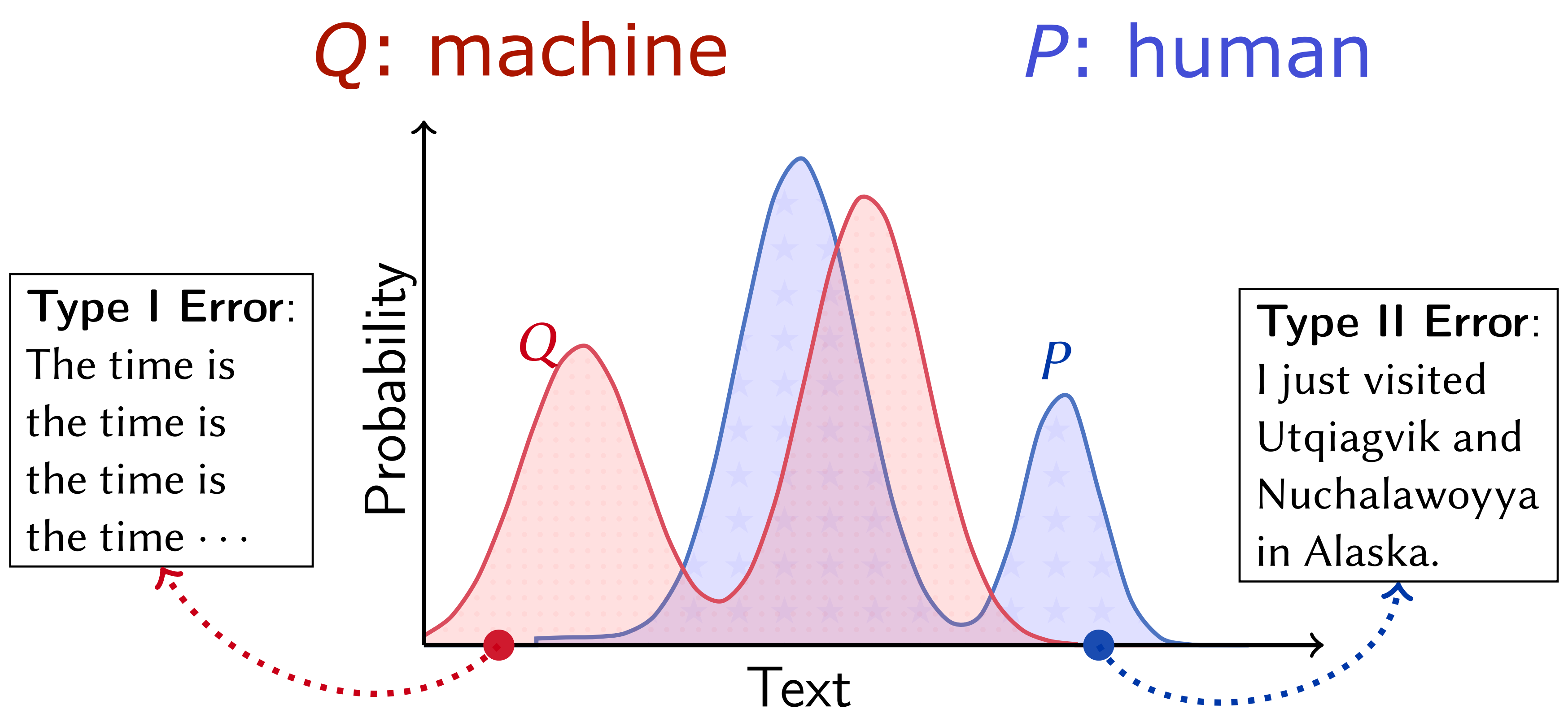
NeurIPS 2018
Submitted 2022

Privacy

✓

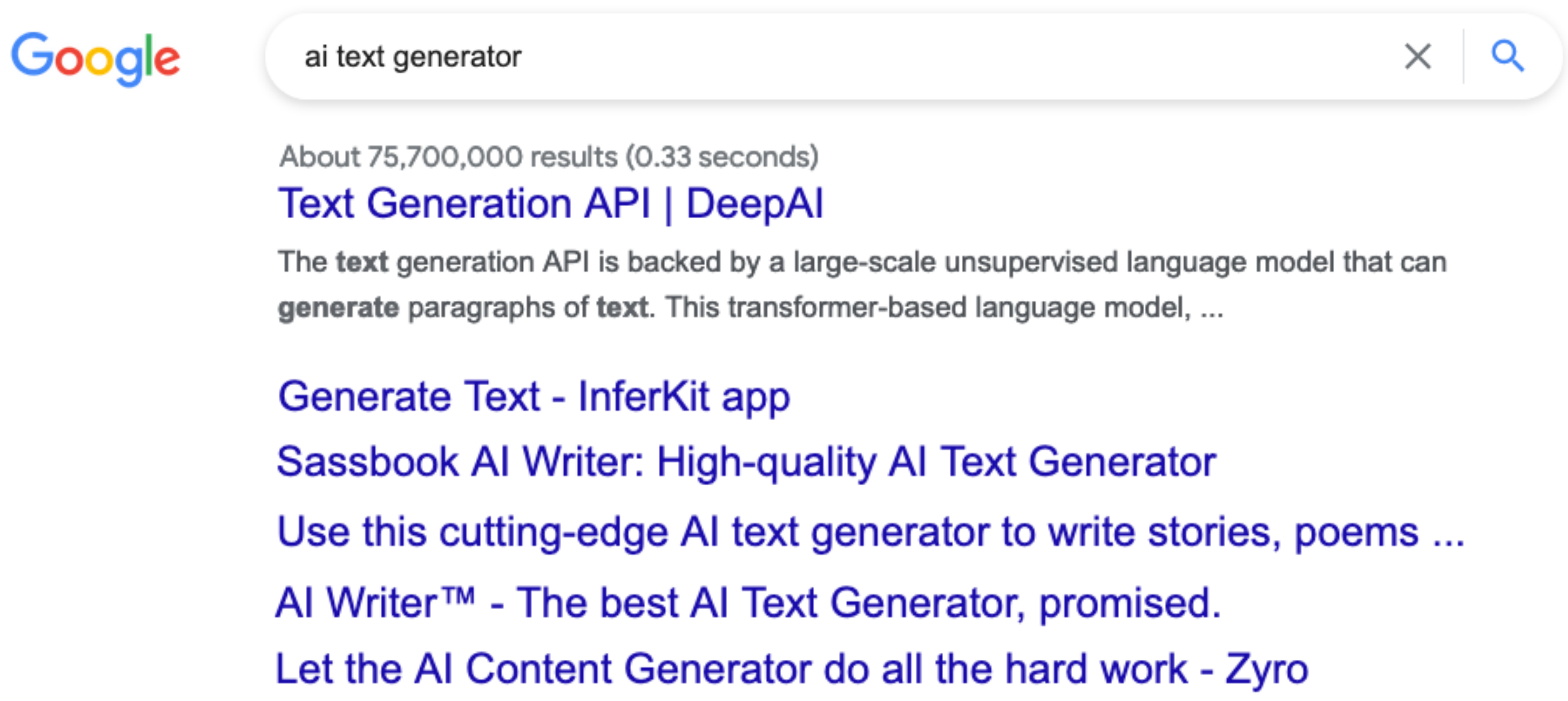
Part 1: Diagnosing large-scale text generation models with **Mauve**

[NeurIPS (2021a) Outstanding Paper Award,
NeurIPS (2021b), *Submitted (2023)*]



Open-ended generative AI

- *New*: LLMs can write long essays!
- Widely deployed commercially
- LLMs still make mistakes

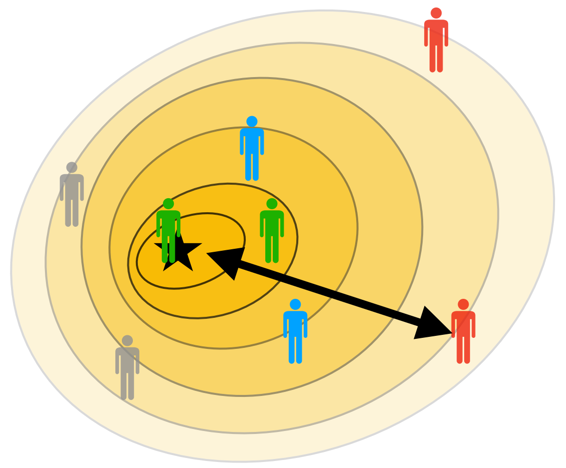


>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

GPT-2

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

Open-ended generation is an emergent capability



Deployment conditions differ from training

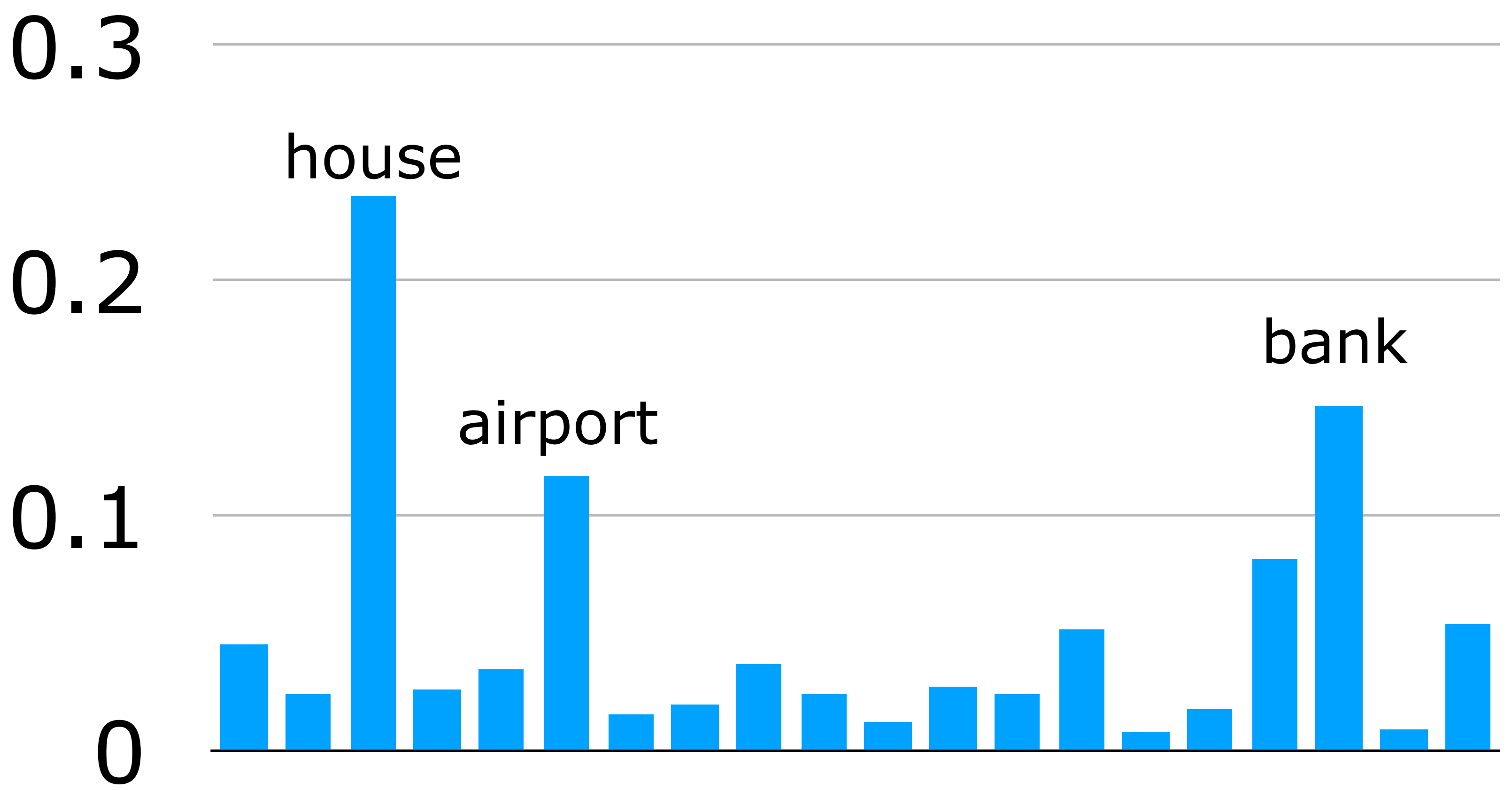
Training: Language modeling

Guess the next *1 word*

Deployment: Sequential generation

Sample the next *500 words sequentially*

Shall we go to the ____



How good is open-ended generation? The classical approach

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

How good is open-ended generation? The classical approach

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

ChatGPT

Model Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

How good is open-ended generation? The classical approach

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

ChatGPT

Human

Model Continuation. The scientists named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown …

Measure similarity/overlap

Human Reference. Known only to specialized cartographers as “Valle Escondido” or “Hidden valley”, this valley boasts of a wide variety of flora and fauna …

How good is open-ended generation? The classical approach

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

ChatGPT

Human

Model Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

Measure similarity/overlap

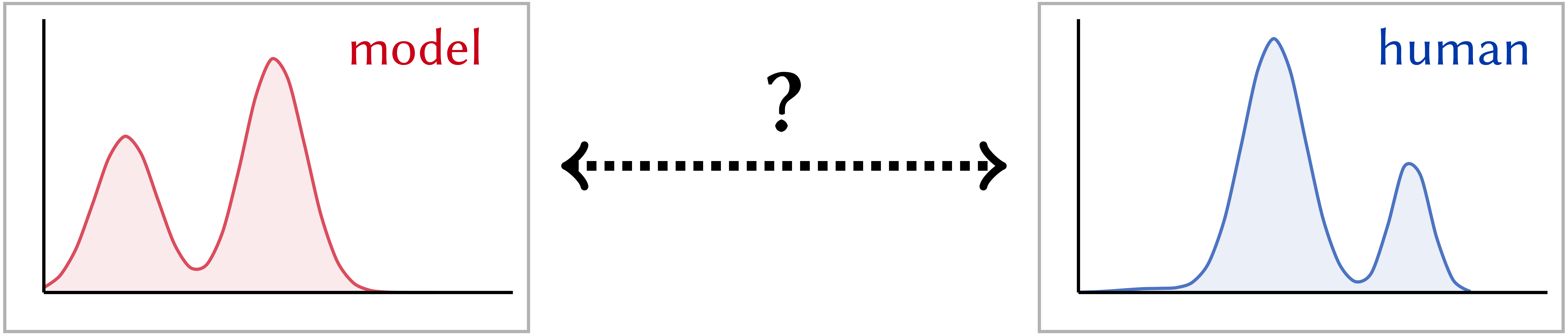
Human Reference. Known only to specialized cartographers as "Valle Escondido" or "Hidden valley", this valley boasts of a wide variety of flora and fauna …

Continuation 2. This discovery has kicked off an all-out search for other mythical creatures from the frozen reaches of the Antarctic to the tropical islands of the Pacific …

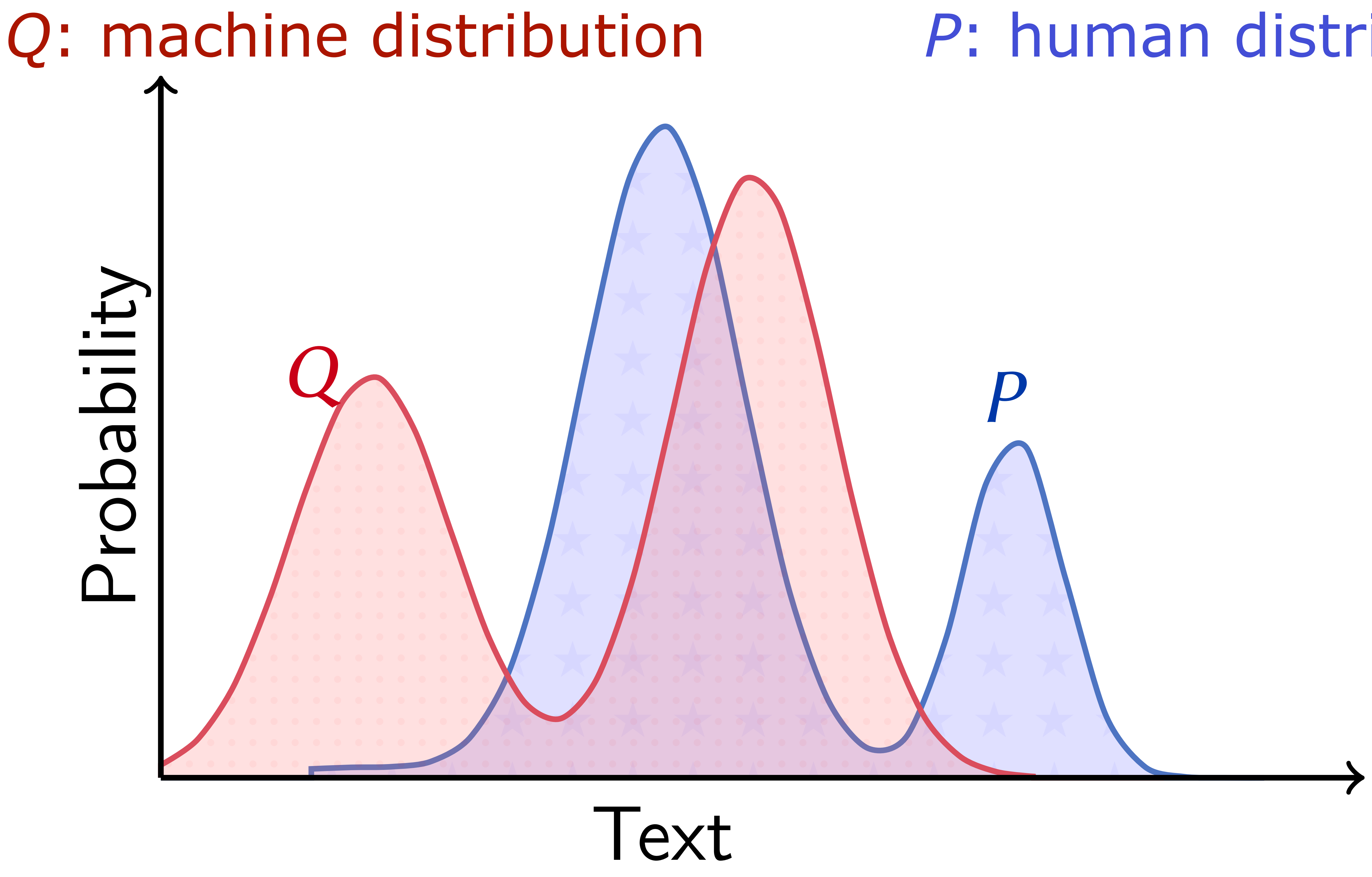
Continuation 3. Perhaps most astonishingly, these unicorns have developed their own artificial general intelligence named Yuyaysapa …

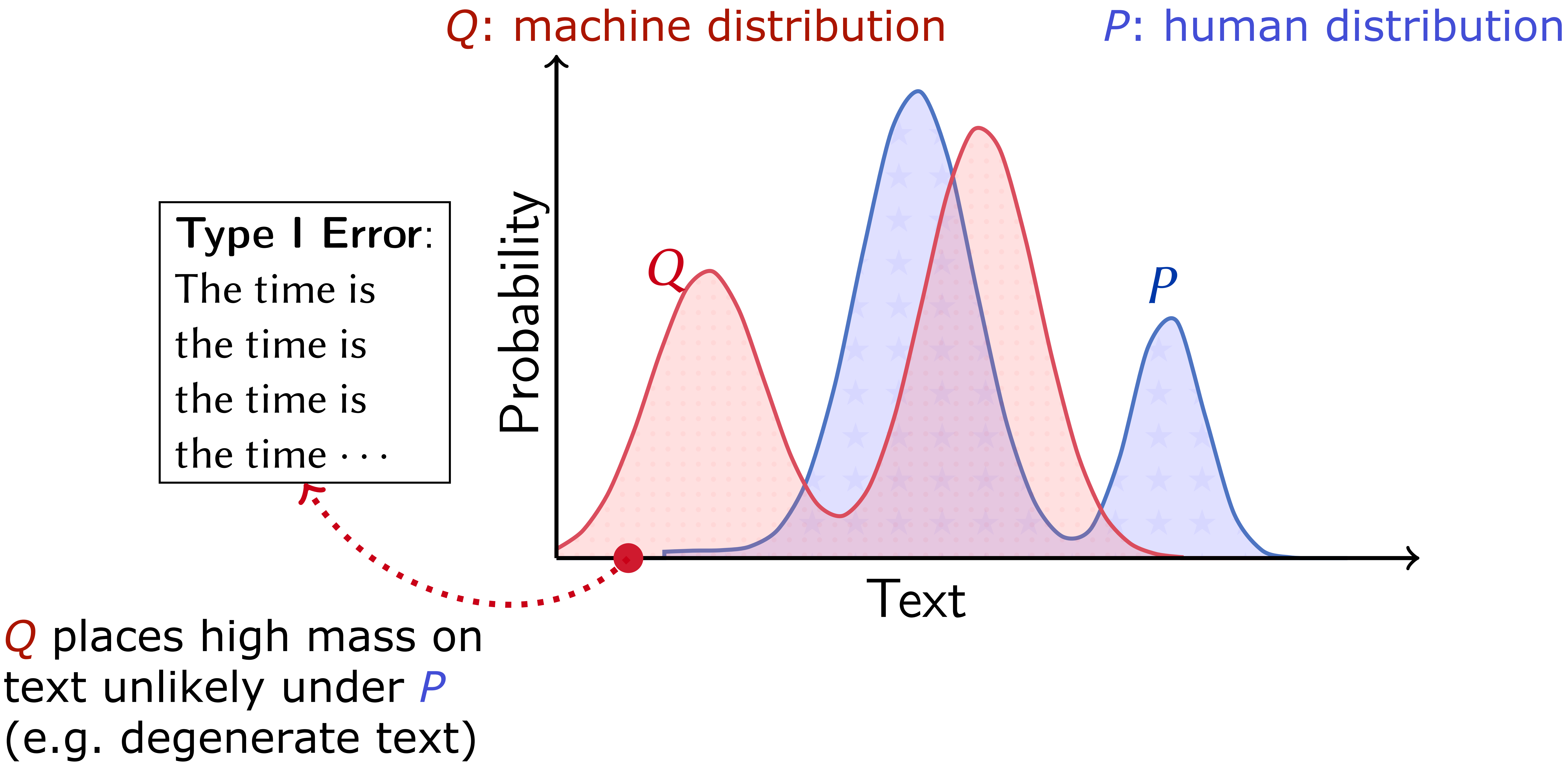
Problem statement

How close are the *probability distributions* over text *sequences*?



Two types of errors in text generation

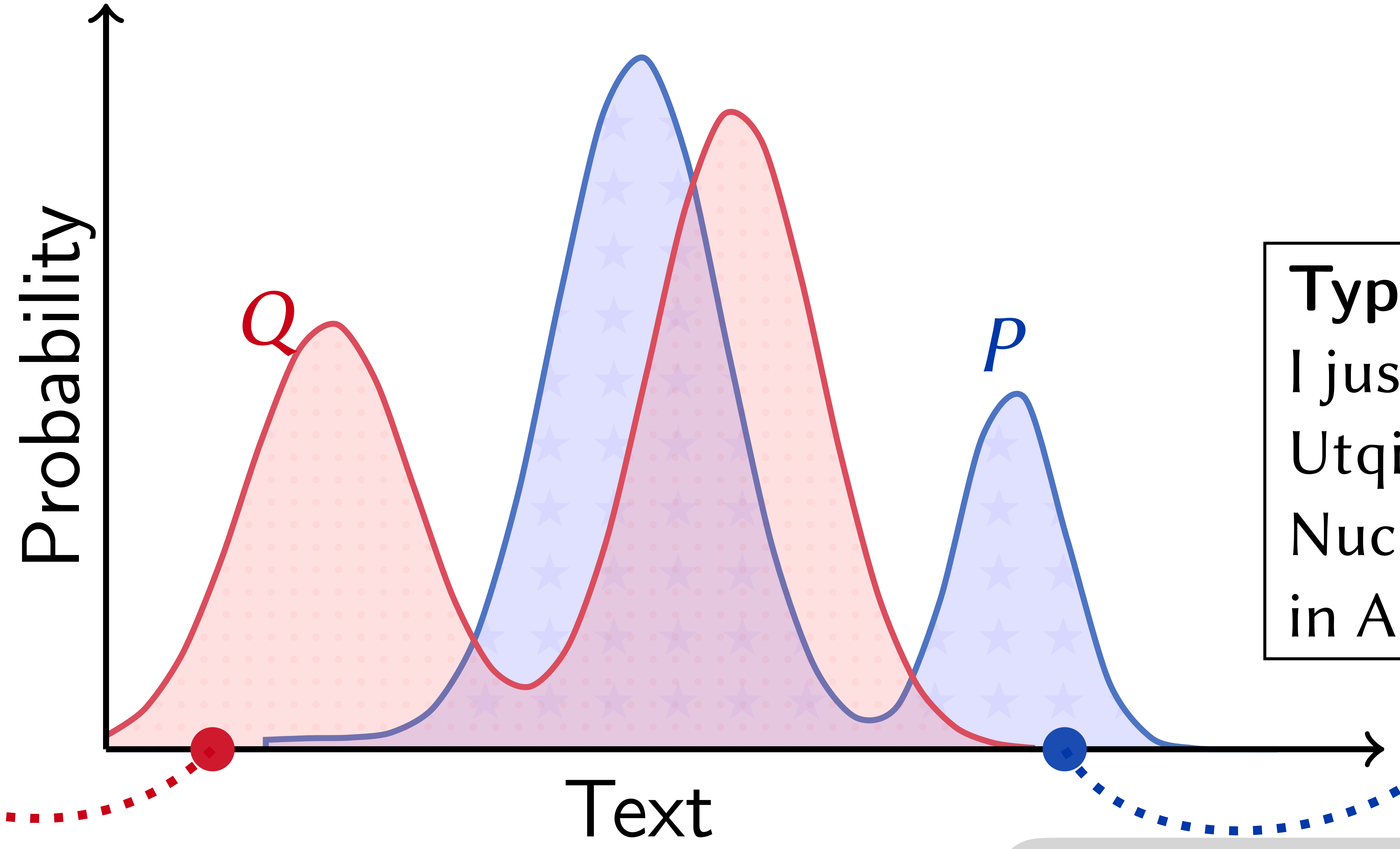




Q: machine distribution
P: human distribution
Probability
Text
Q
P
Type I Error:
The time is
the time is
the time is
the time $\cdots$
Q places high mass on
text unlikely under P
(e.g. degenerate text)

Q: machine distribution *P*: human distribution

Type I Error:
The time is
the time is
the time is
the time ···



Type II Error:
I just visited
Utqiagvik and
Nuchalawoyya
in Alaska.

KL(*Q*|*P*)

$$\mathrm{KL}(P\,|\,Q) = \sum_x P(x)\,\log\frac{P(x)}{Q(x)}$$

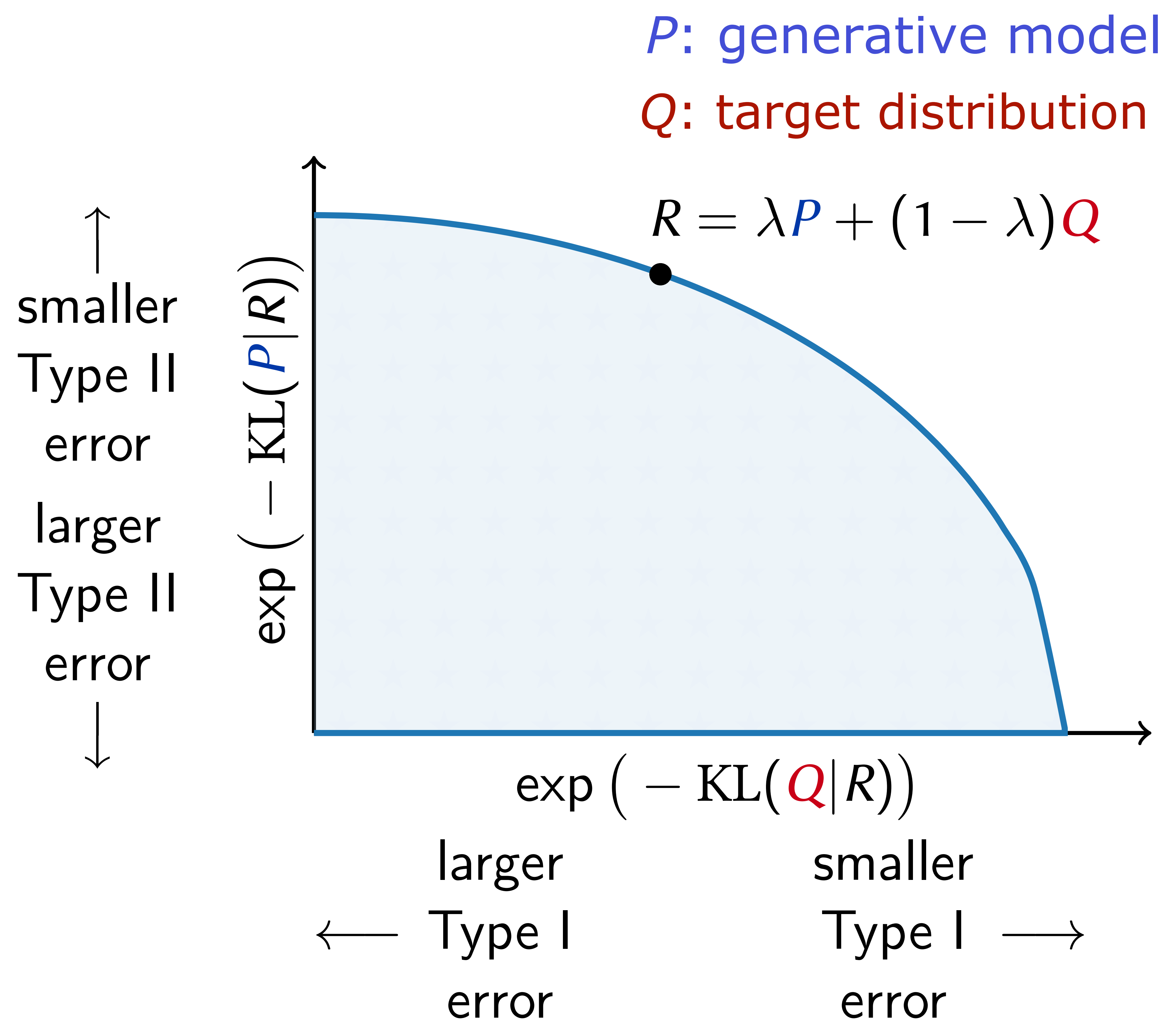
KL(*P*|*Q*)

Mauve: summarizing both errors

- KL(Q|P) and KL(P|Q) can be infinite, so measure errors *softly* using *mixtures*

- **Divergence Curve:** Varying the *mixture weight*

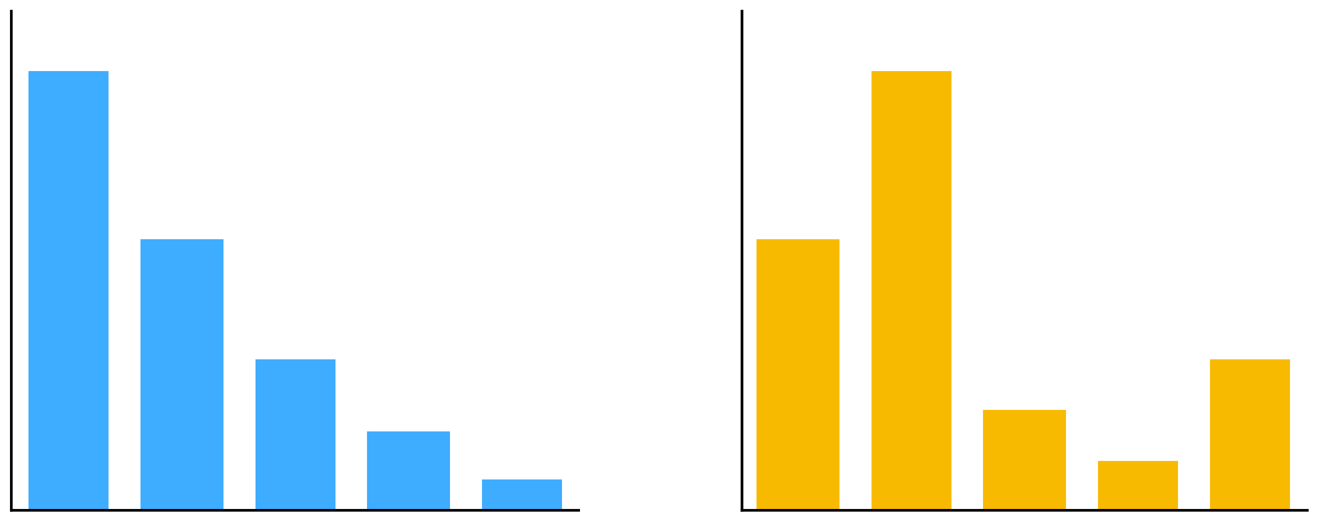
- **Mauve:** area summary of the curve: a *quantitative measure of similarity* and takes values between 0 (dissimilar) and 1 (identical)



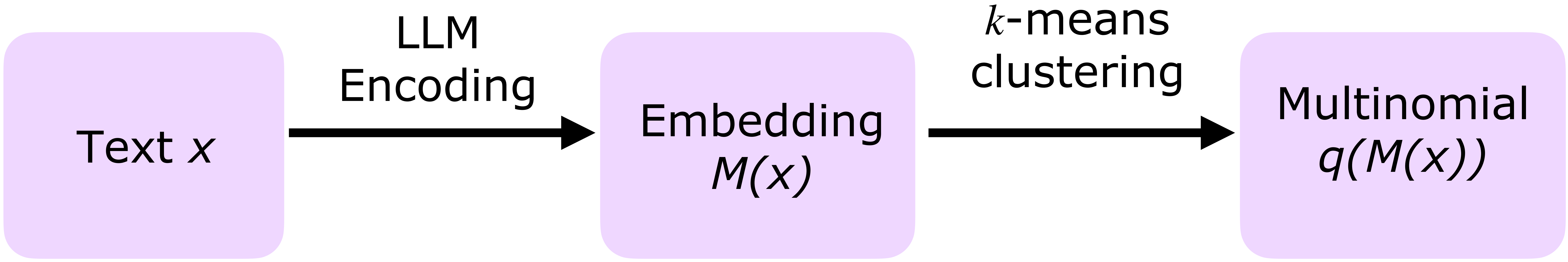
Computing Mauve in practice

- Sum over documents intractable

$$\mathrm{KL}(Q\,|\,R) = \sum_x Q(x)\,\log\frac{Q(x)}{R(x)}$$



- Computation pipeline



Correlation with human judgements

Goals of automatic evaluation

Humans are the end users, so human evaluation is the ultimate test

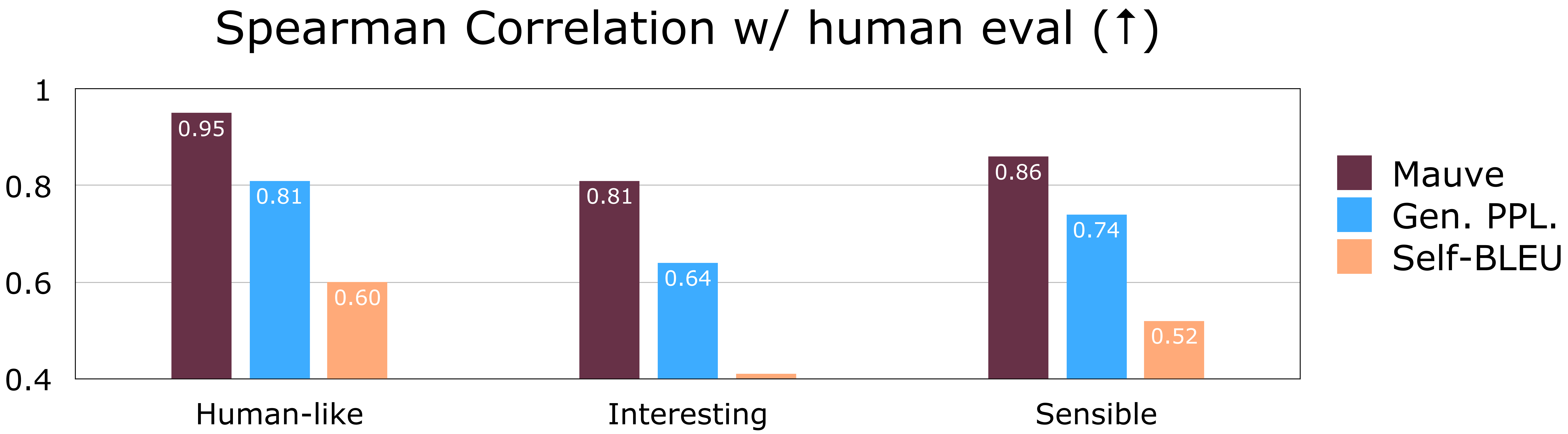
Human evaluation is slow and expensive

If Mauve can correlate with human evaluations, faster iterations + debugging

Correlation with human judgements

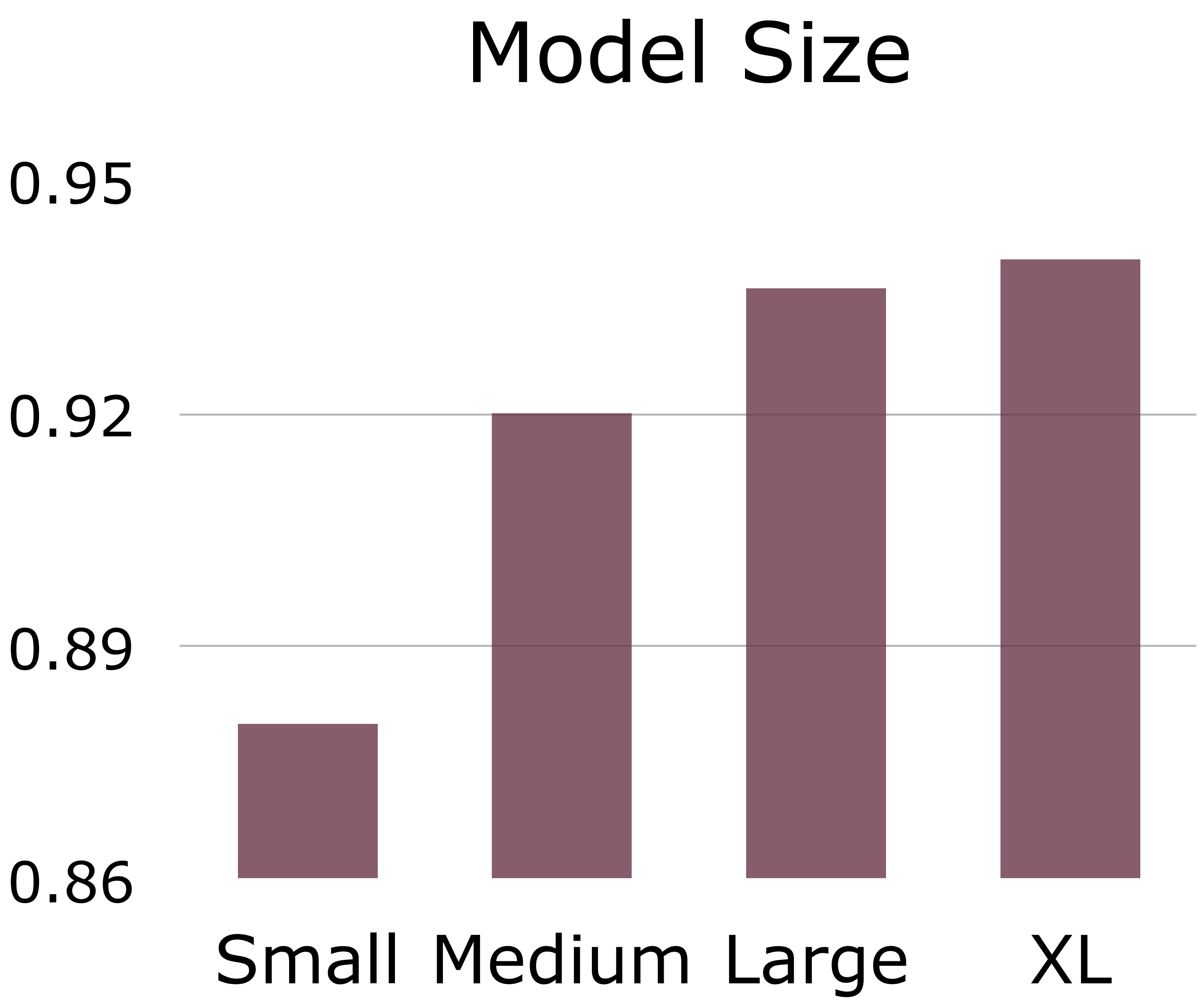
Head-to-head: Is A or B more (a) human-like, (b) interesting, (c) sensible?

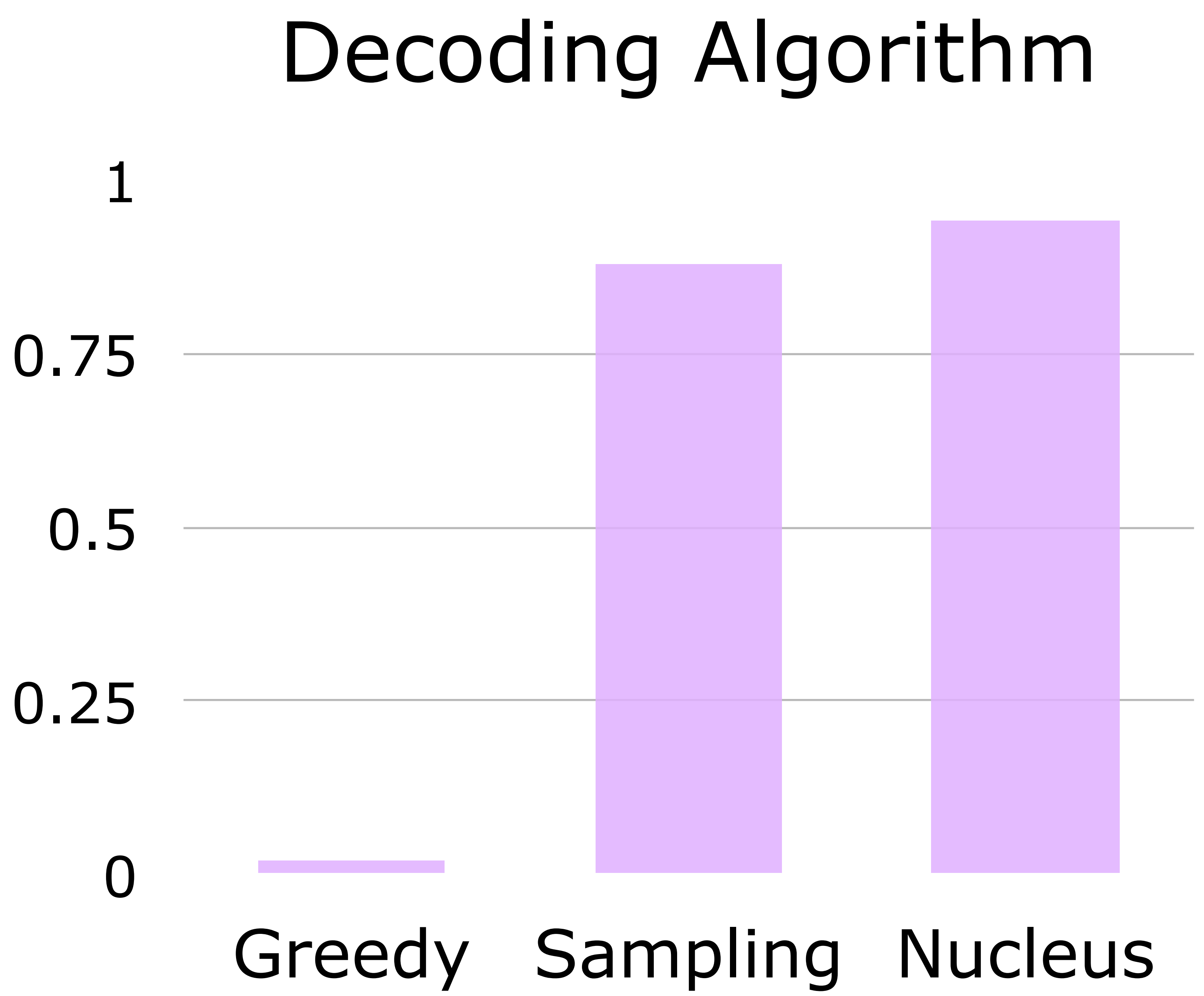
We compare text written by humans and 8 models

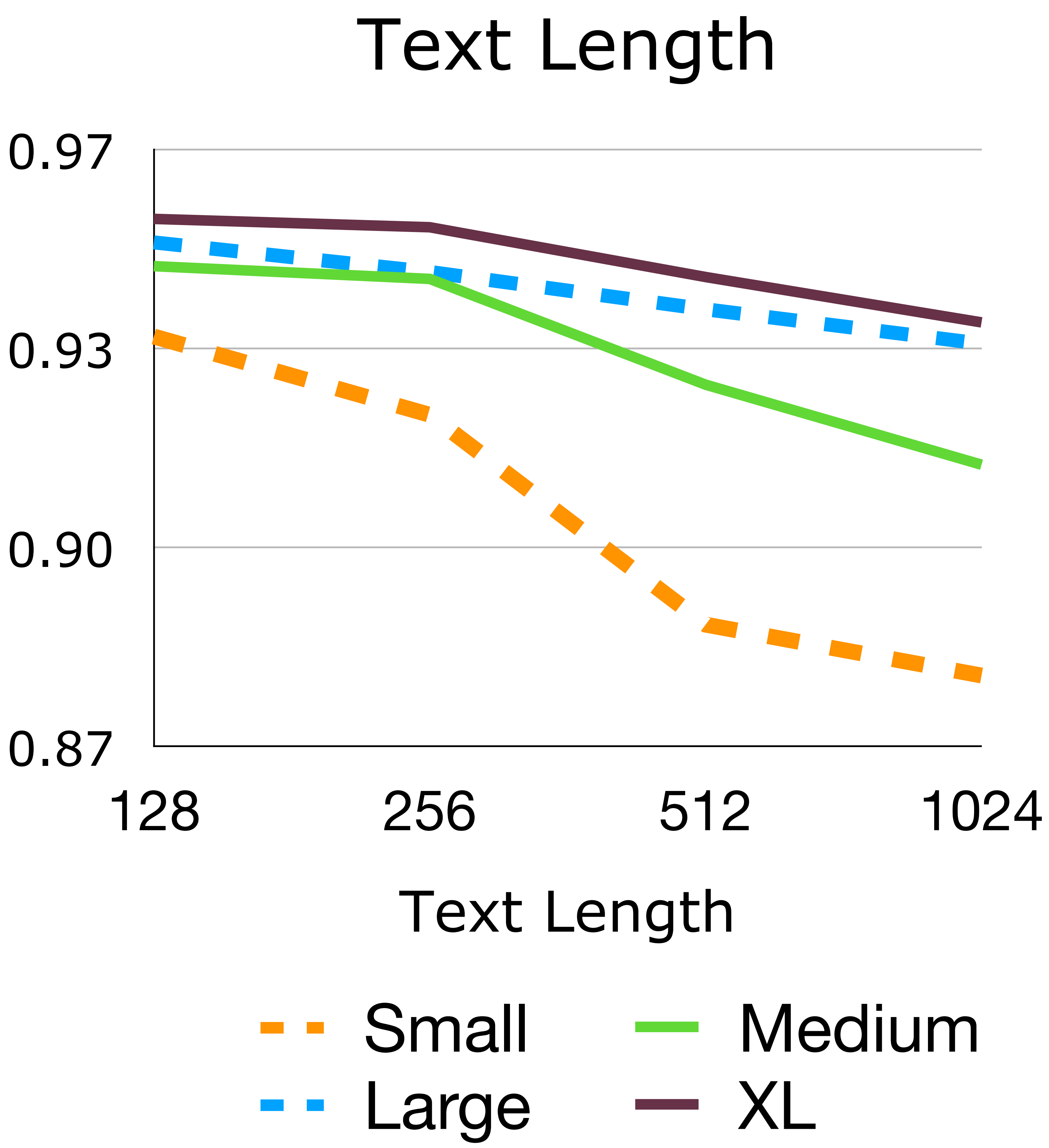


Gen. PPL.: Holtzman et al. (ICLR 2020) Self-BLEU: Zhu et al. (2018)

Mauve captures important trends







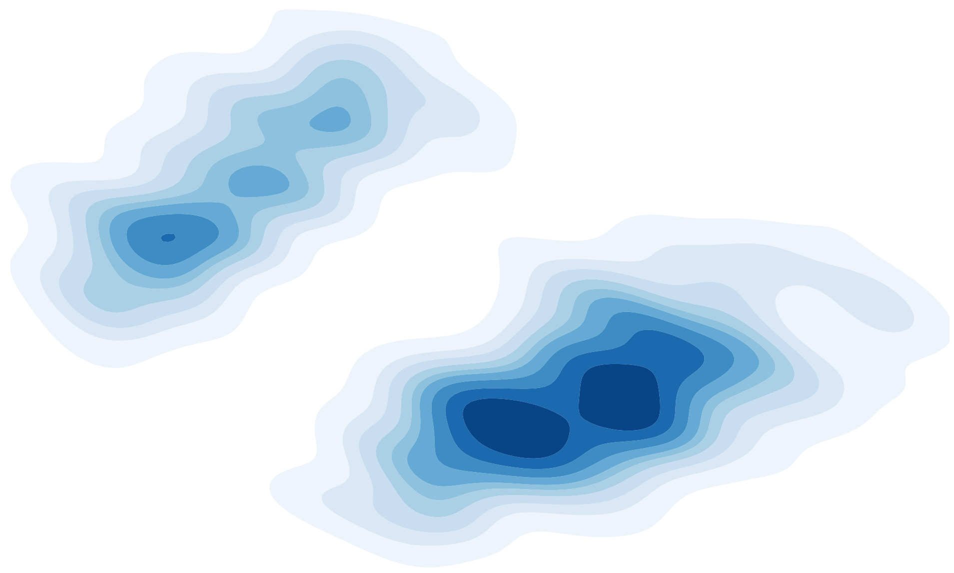
- Y-axis shows Mauve (↑)
- **Mauve** captures all the trends while baselines fail

Mauve: Estimation theory

Estimation of Mauve involves two approximations:

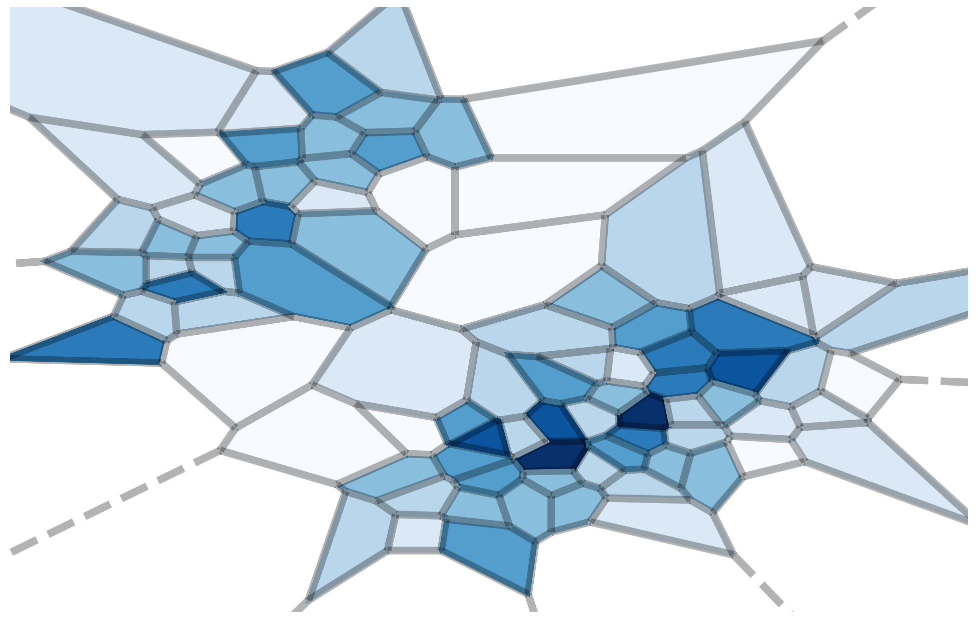
Clustering/
Quantization

Estimate from
samples



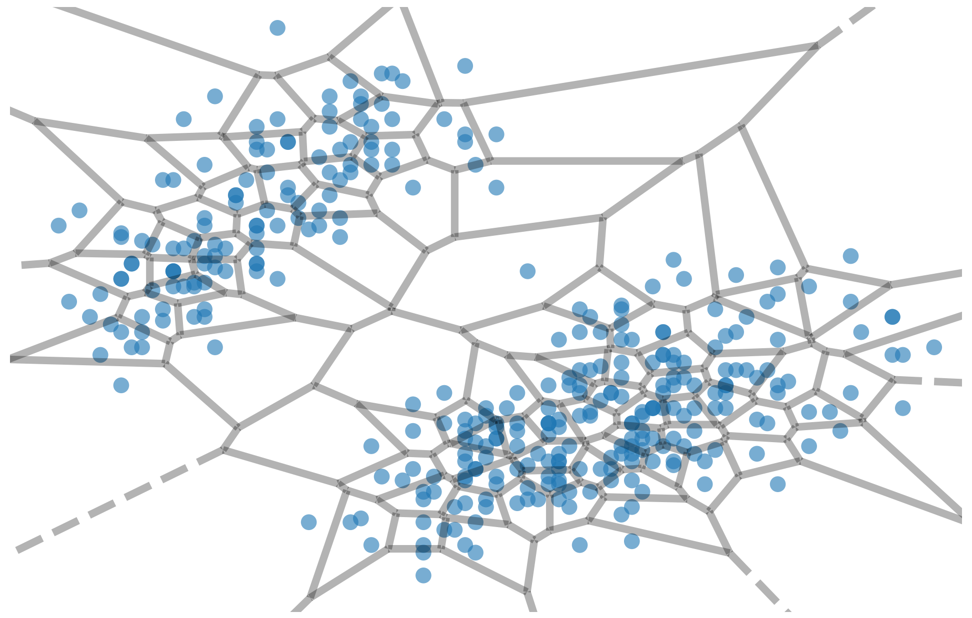
P, Q

high-dimensional
text distributions



P_S, Q_S

quantized
distributions



$\hat{P}_{S,n}, \hat{Q}_{S,n}$

samples (empirical
distributions)

Theorem (informal)

There exists a quantization of size k such that the approximation error of Mauve from n samples is

$$\widetilde{O}\left(\underbrace{\sqrt{\frac{k}{n}}}_{\text{Statistical Error}} + \underbrace{\frac{1}{k}}_{\text{Quantization Error}} \right)$$

n: number of samples from P and Q

k: quantization size (Num. clusters)

Balance both by choosing $k = \Theta(n^{1/3})$

Mauve: Beyond clustering

Text x $\xrightarrow{\text{LLM Encoding}}$ Embedding $M(x)$ $\xrightarrow{k\text{-means clustering}}$ Multinomial $q(M(x))$ $\xrightarrow{\text{KL divergenc e}}$ **Mauve**

Mauve: Beyond clustering

Nearest neighbor estimator

Kernel density estimator

Parametric Gaussian approx.

Classifier-based estimation

Mauve: Beyond clustering

General f-divergences

Optimal transport

Mauve: Beyond clustering

Essential

Software

pip install mauve-text

```
from mauve import compute_mauve

p_text = … # list of strings for human distribution P
q_text = … # list of strings for model distribution Q

# Obtain deep encoding, quantize it and compute Mauve
out = compute_mauve(p_text=p_text, q_text=q_text)

print(f'Mauve(P, Q) = {out.mauve}')
```

60K downloads

HuggingFace Evaluate: **pip install evaluate**

```
from evaluate import mauve

mauve = load("mauve")

# Obtain deep encoding, quantize it and compute Mauve
out = mauve.compute(references=p_text, predictions=q_text)

print(f'Mauve(P, Q) = {out.mauve}')
```

Impact of Mauve

Standard metric for evaluation and hyper-parameter tuning

Meister et al. (TACL 2022)

Su et al. (NeurIPS 2022)

Lu et al. (NeurIPS 2022)

Xu et al. (NeurIPS 2022)

Jawahar et al. (ACL 2022)

Hewitt et al. (EMNLP 2022)

Mattern et al. (EMNLP 2022)

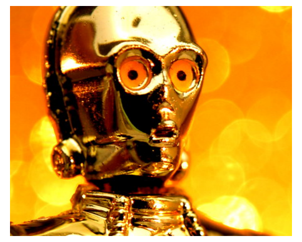
Hu et al. (NAACL 2022)

Summary: Diagnosing large-scale text generation with Mauve

Used for open-ended generation but trained for language modeling. How good is it?

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …



Continuation 2. This discovery has kicked off an all-out search for other mythical creatures from the frozen reaches of the Antarctic to the tropical islands of the Pacific …

Continuation 3. Perhaps most astonishingly, these unicorns have developed their own artificial general intelligence named Yuyaysapa …

Summary: Diagnosing large-scale text generation with Mauve

Used for open-ended generation but trained for language modeling. How good is it?

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

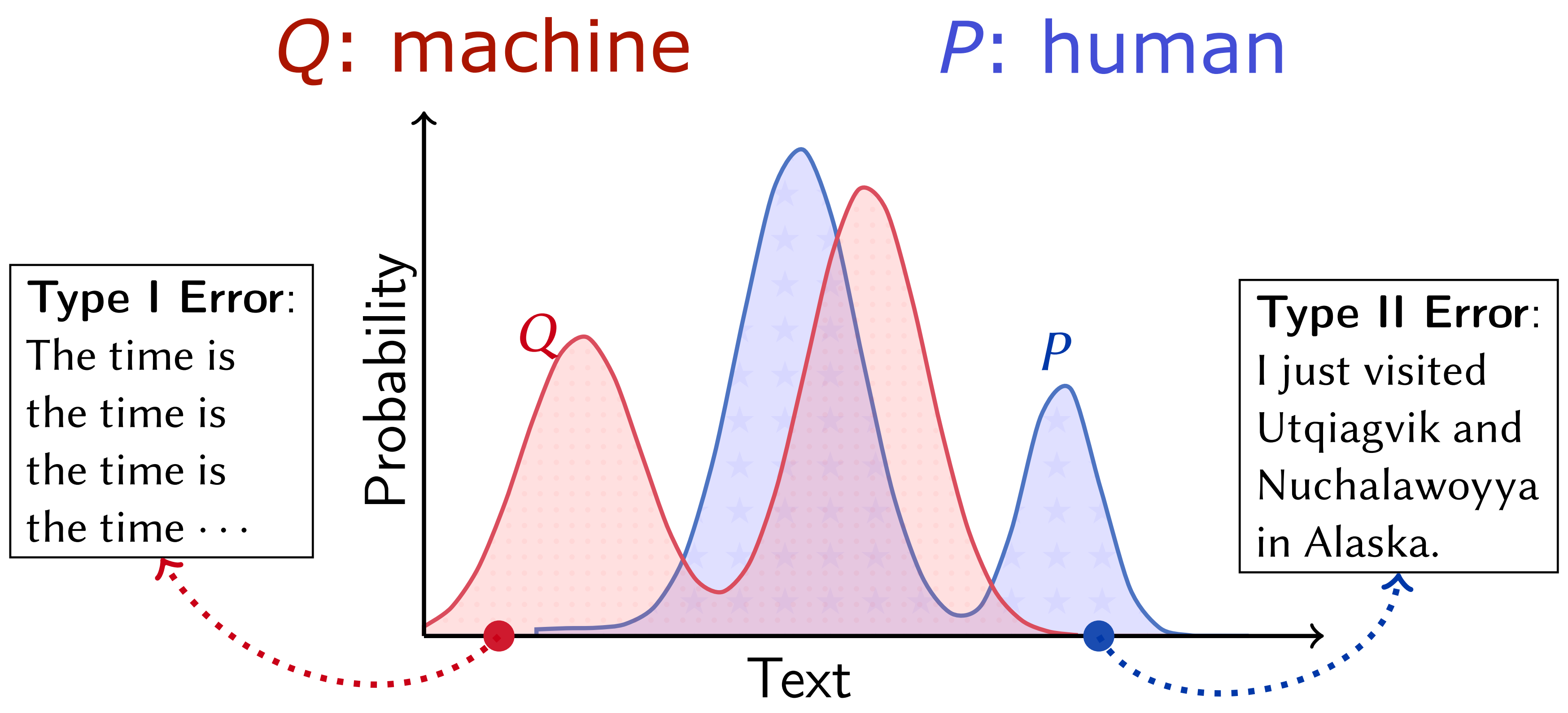
GPT-2

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

Continuation 2. This discovery has kicked off an all-out search for other mythical creatures from the frozen reaches of the Antarctic to the tropical islands of the Pacific …

Continuation 3. Perhaps most astonishingly, these unicorns have developed their own artificial general intelligence named Yuyaysapa …

Directly measure the gap between distributions



Summary: Diagnosing large-scale text generation with Mauve

Used for open-ended generation but trained for language modeling. How good is it?

>> **prompt:** In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

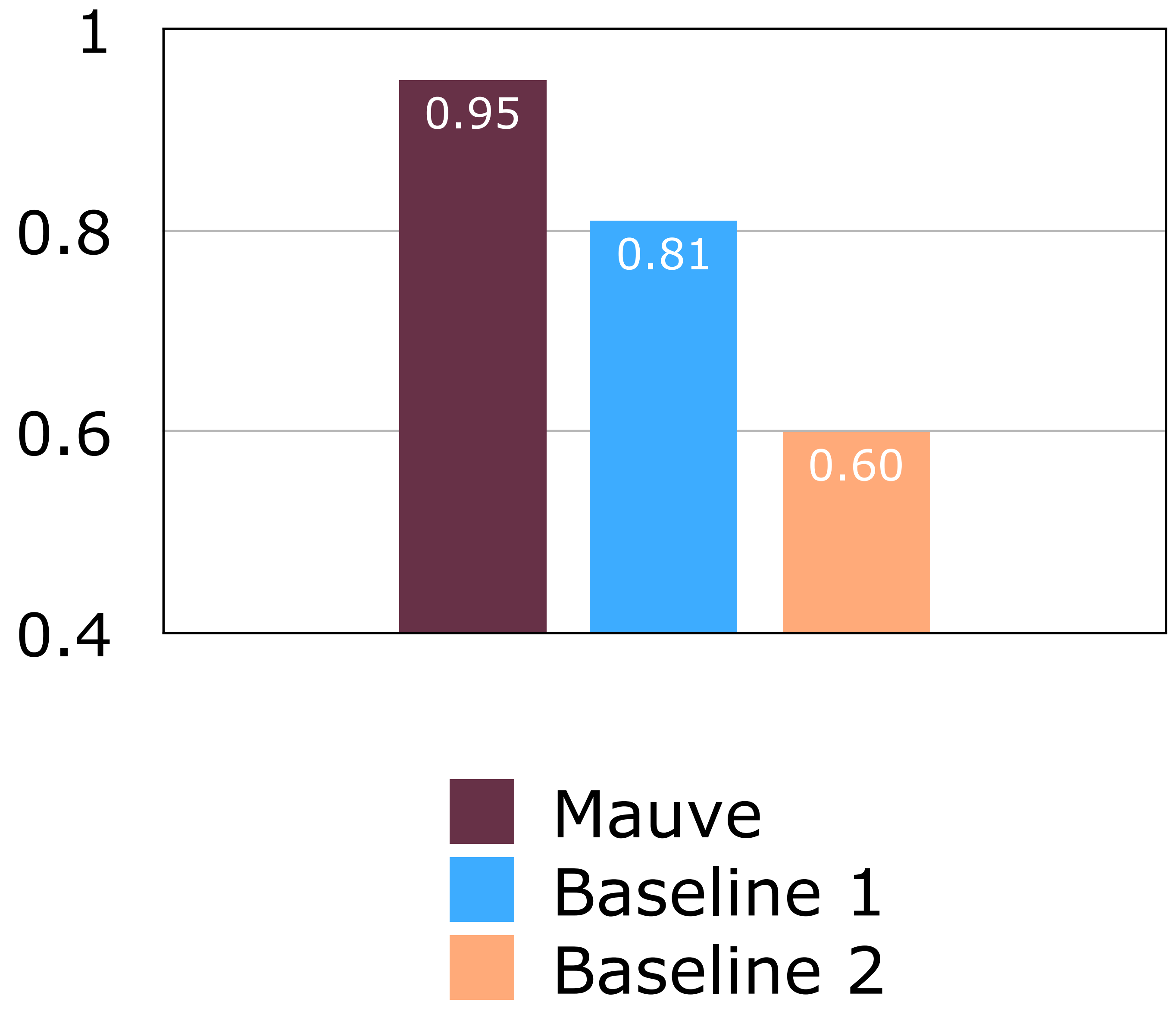
Continuation 2. This discovery has kicked off an all-out search for other mythical creatures from the frozen reaches of the Antarctic to the tropical islands of the Pacific …

Continuation 3. Perhaps most astonishingly, these unicorns have developed their own artificial general intelligence named Yuyaysapa …

Directly measure the gap between distributions

Our approach correlates with human judgements and quantifies observed behavior

Spearman Correlation w/ human eval (↑)



Summary: Diagnosing large-scale text generation with Mauve

Used for open-ended generation but trained for language modeling. How good is it?

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

Continuation 2. This discovery has kicked off an all-out search for other mythical creatures from the frozen reaches of the Antarctic to the tropical islands of the Pacific …

Continuation 3. Perhaps most astonishingly, these unicorns have developed their own artificial general intelligence named Yuyaysapa …

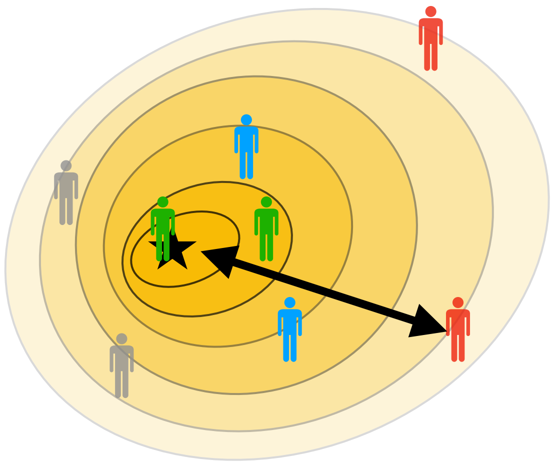
Directly measure the gap between distributions

Theory: error bounds

$$\widetilde{O}\left(\sqrt{\frac{k}{n}} + \frac{1}{k} \right)$$

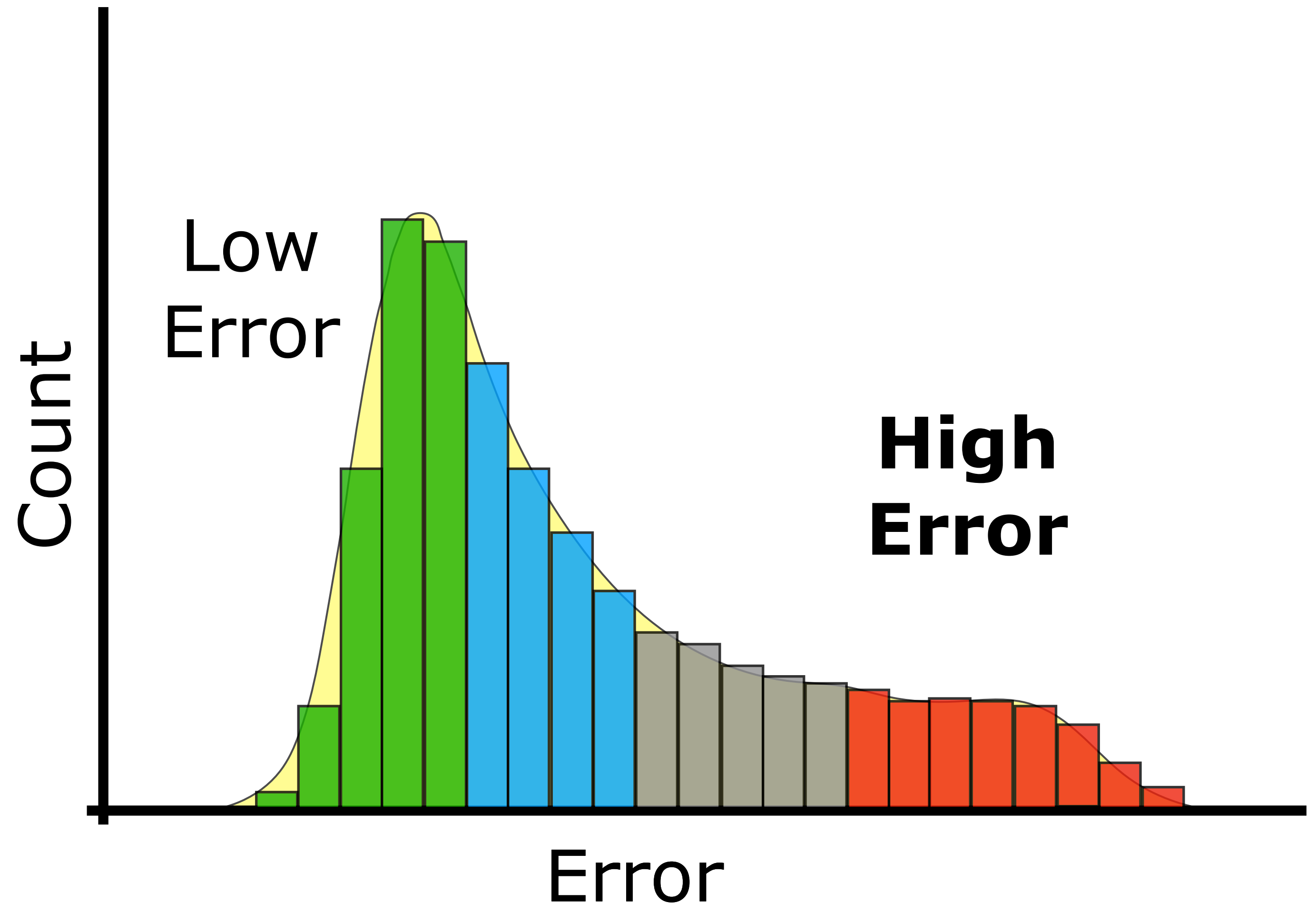
Statistical Error

Quantization Error

		Federated learning	LLMs
Robust Deployment		IEEE CISS 2021, *Springer SVVA 2021*, *Mach. Learn. 2022*	NeurIPS 2021a NeurIPS 2021b *Submitted 2023*
		Part 2	
Robust to Outliers		*IEEE Trans. Signal Proc. 2022*, ICML 2022	Submitted 2022
Optimize Faster		✓	NeurIPS 2018 Submitted 2022
Privacy		✓	

Part 2: Tackling distribution shifts in federated learning

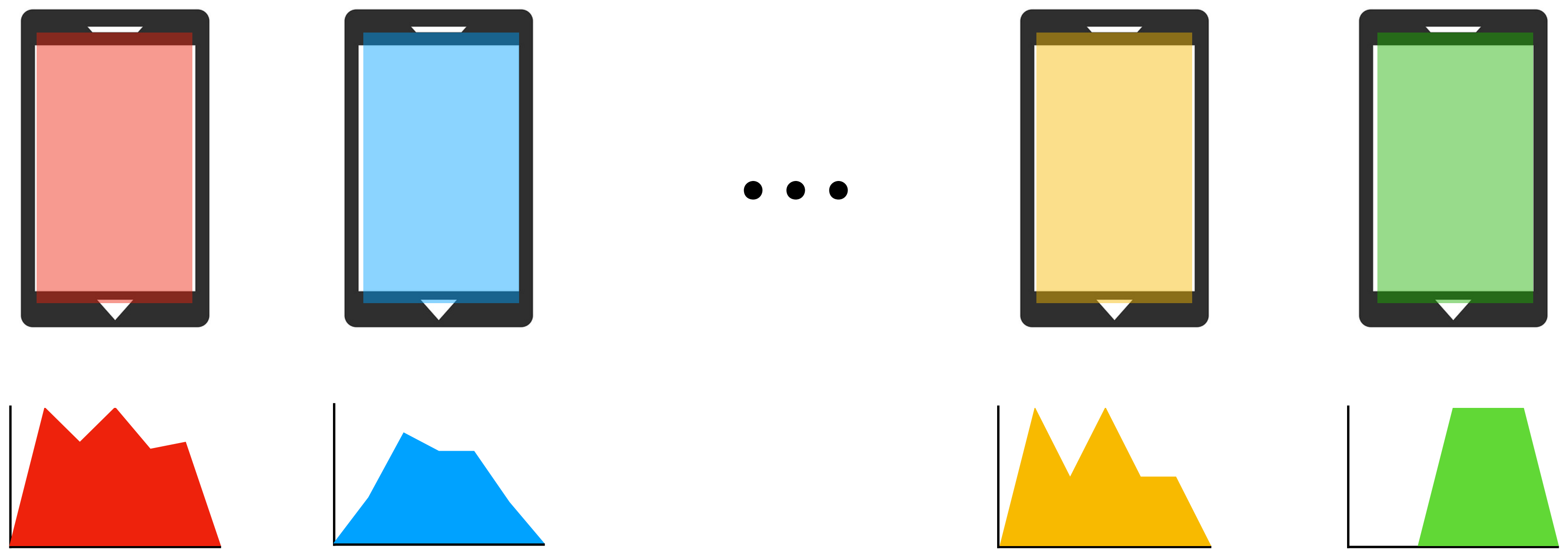
[IEEE CISS ’21, DistShift-NeurIPS ’22 (Spotlight), *SVVA ’21, Mach. Learn. ’22*]



$1 \quad 2 \quad \cdots \quad n$

Clients

Data Distribution



$p_1 \quad p_2 \quad \cdots \quad p_n$

Diverse users

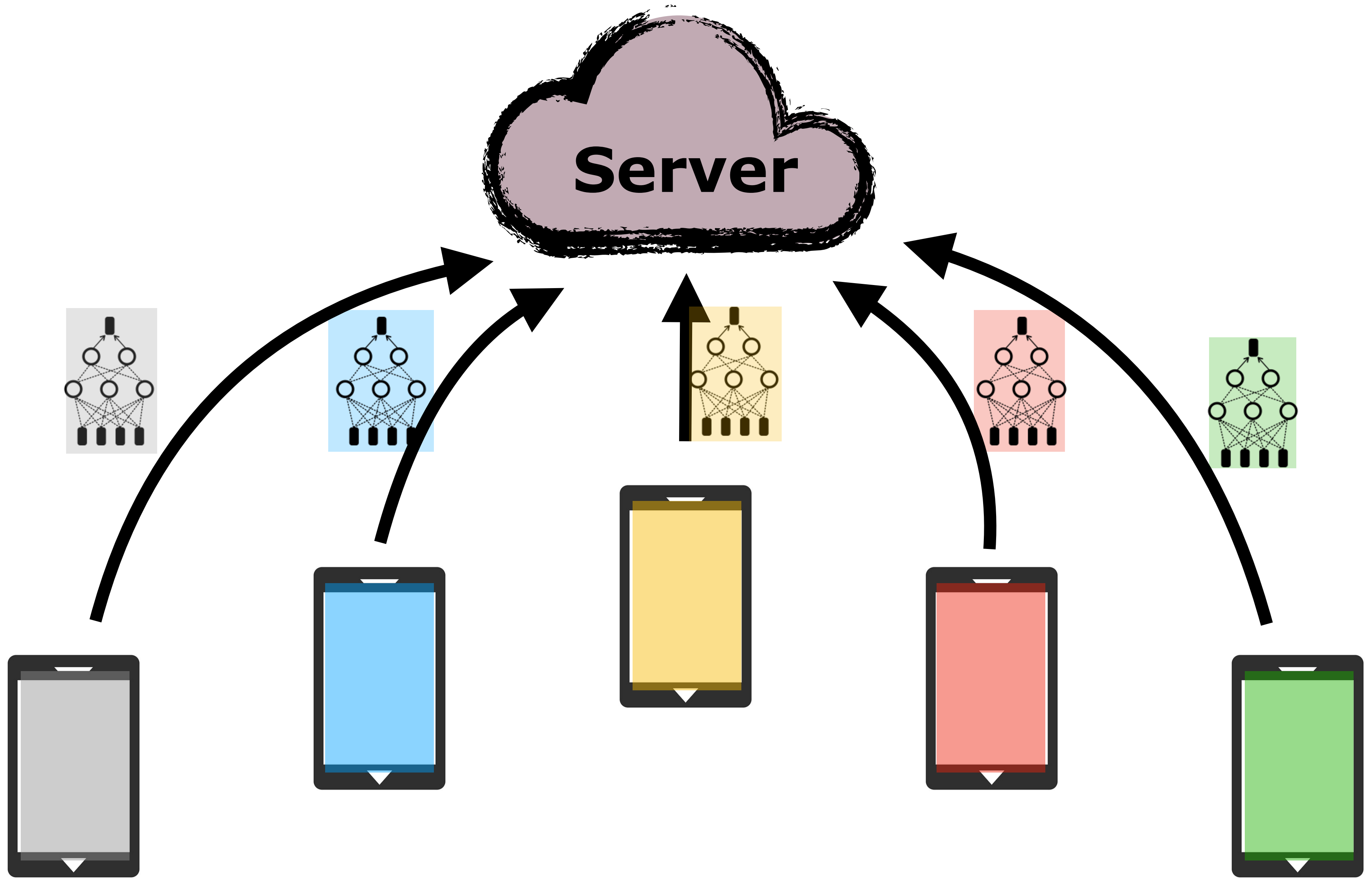
Usual Learning Objective

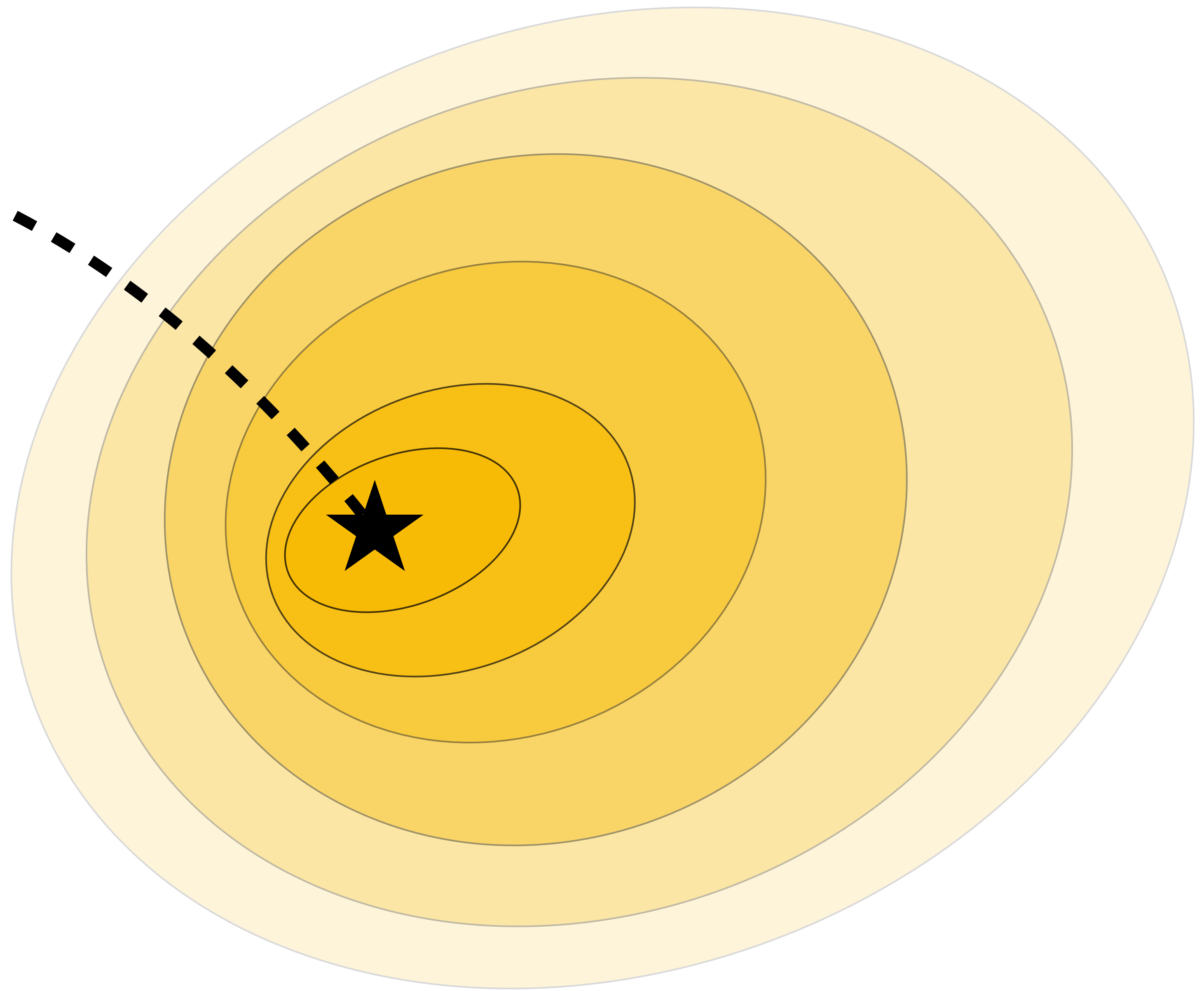
$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} F_i(w) \qquad \text{where} \qquad F_i(w) = \mathbb{E}_{z \sim p_i} \left[f(w; z) \right]$$

loss on client i

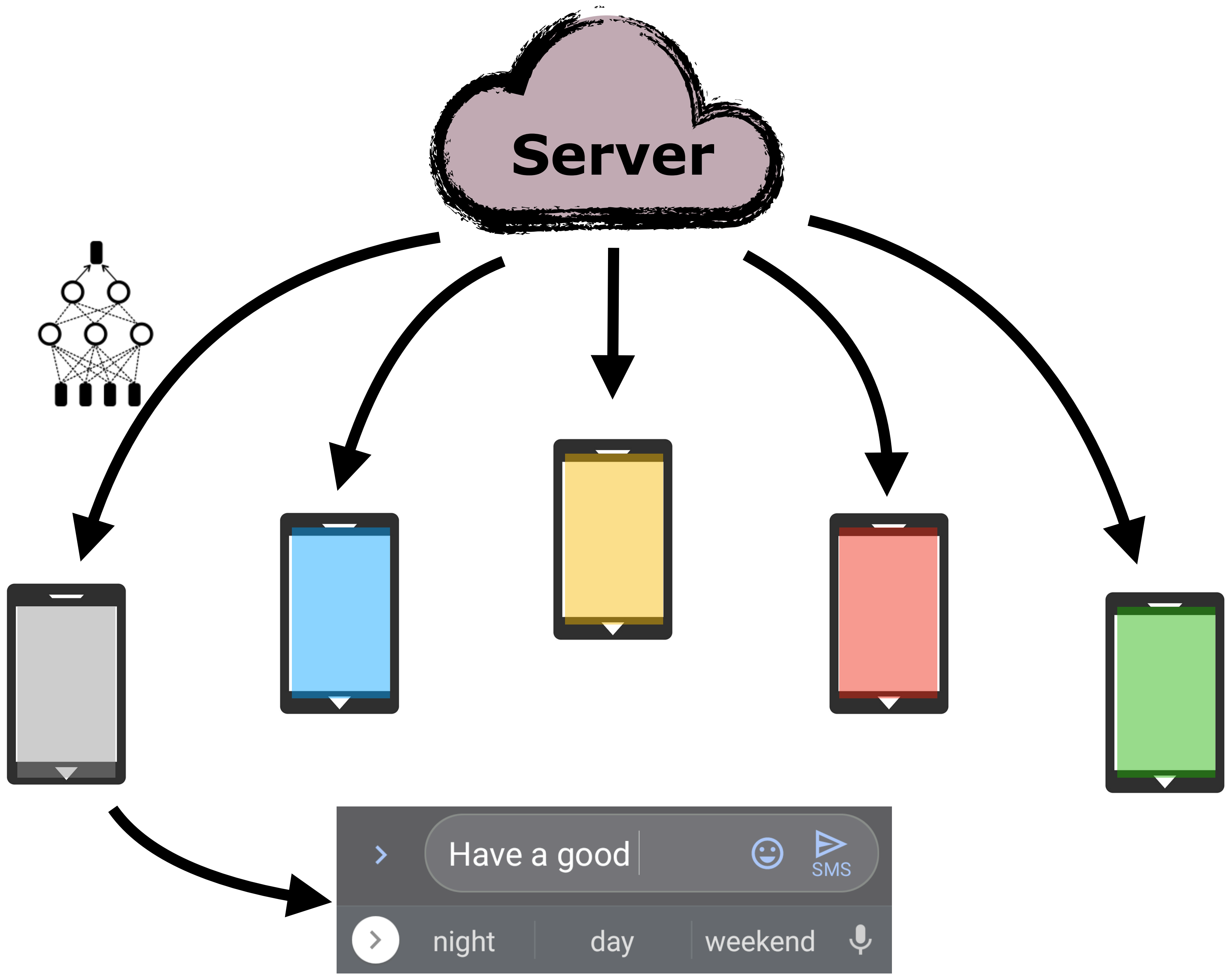
[McMahan et al. (AISTATS 2017), Kairouz et al. (2021)]

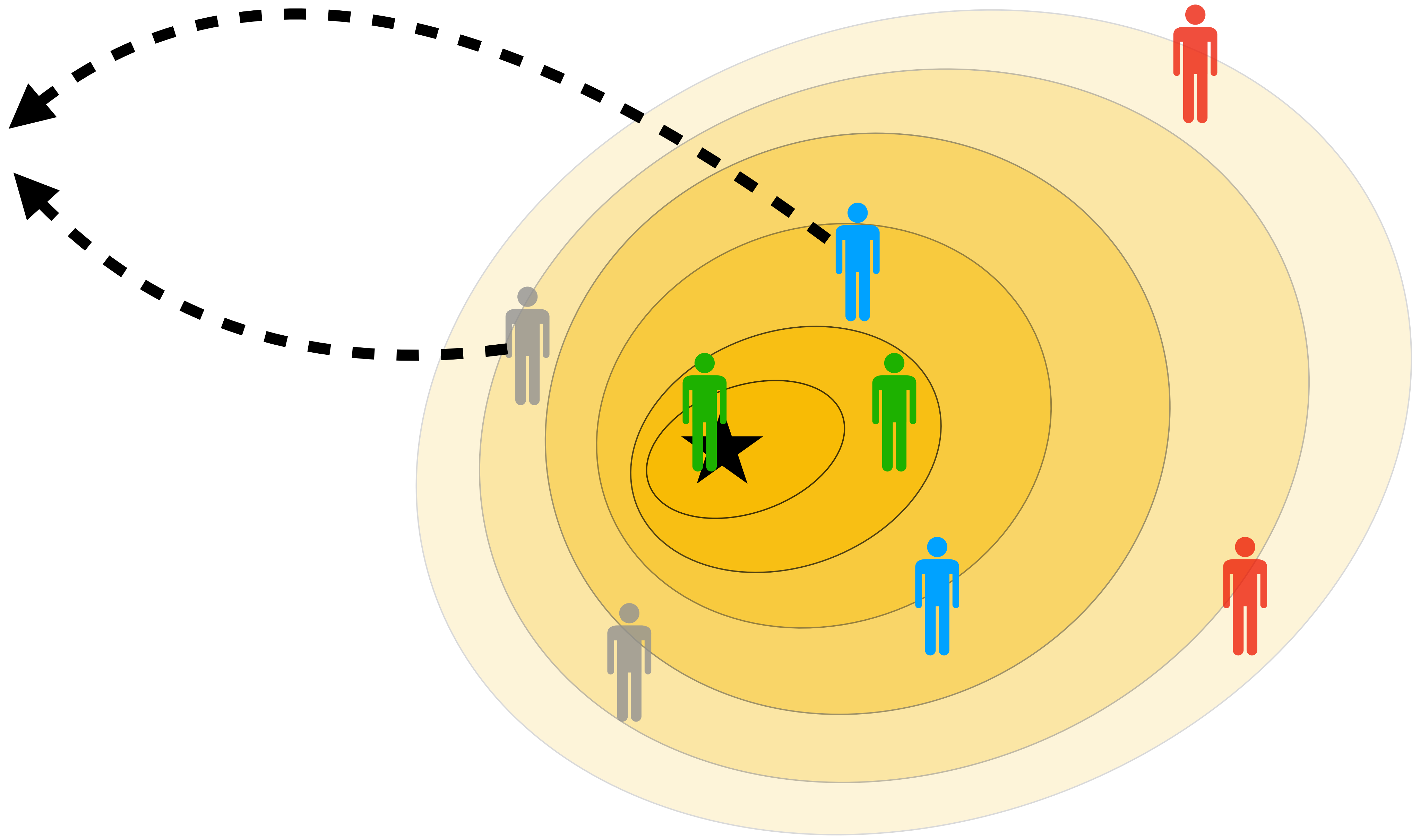
Usual approach $\Longrightarrow$ Global model is trained on *average distribution* across clients



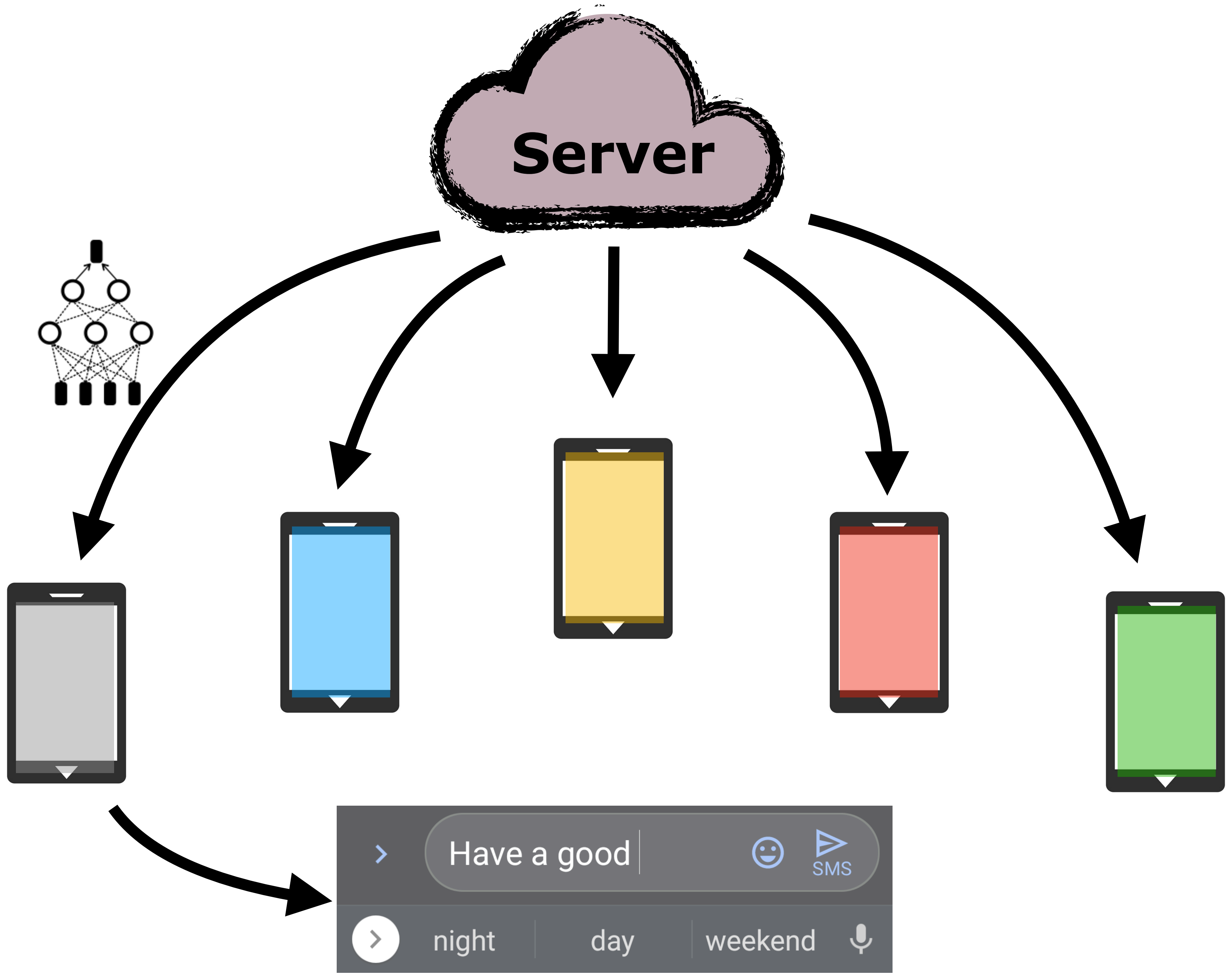


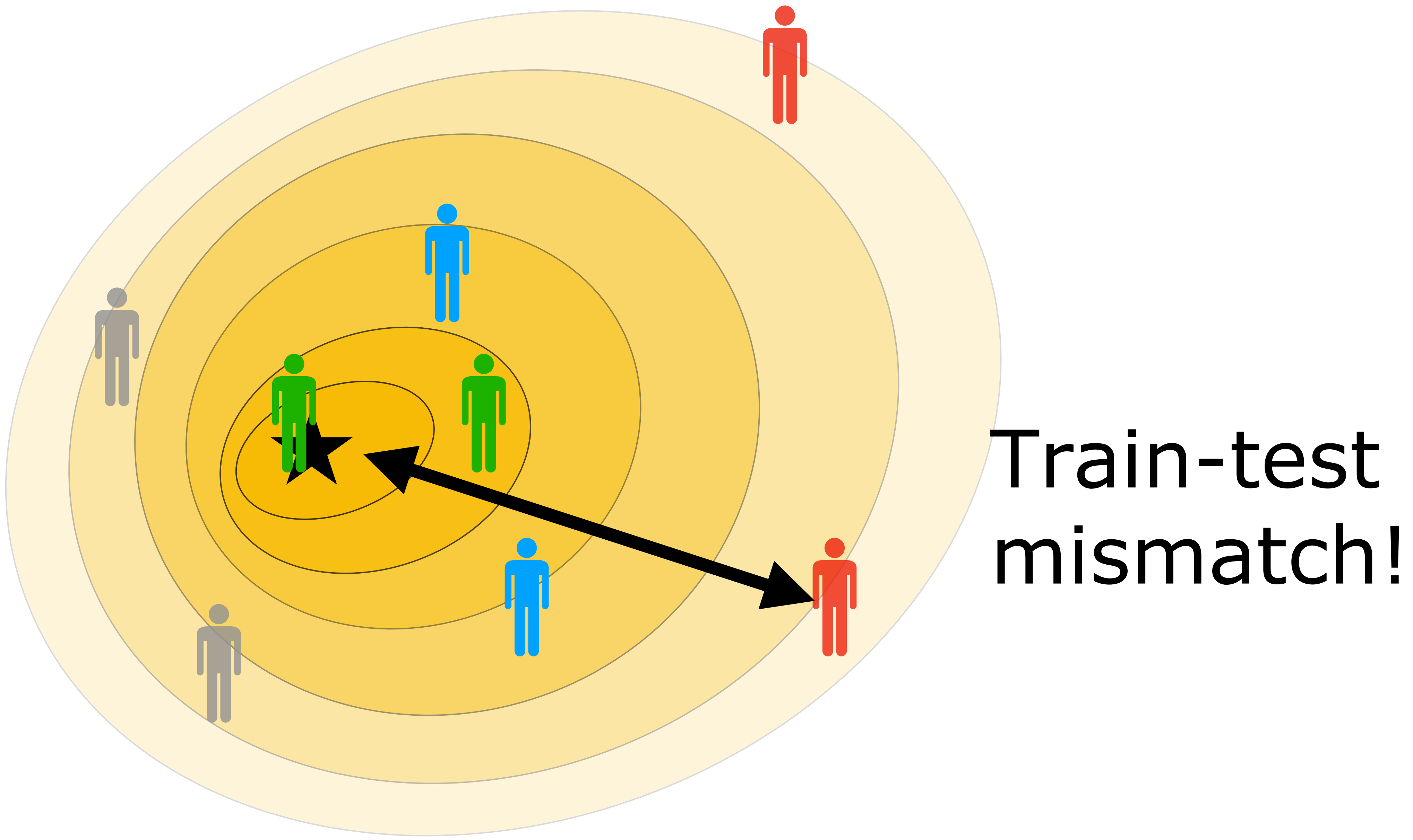
Global model is deployed on *individual* clients



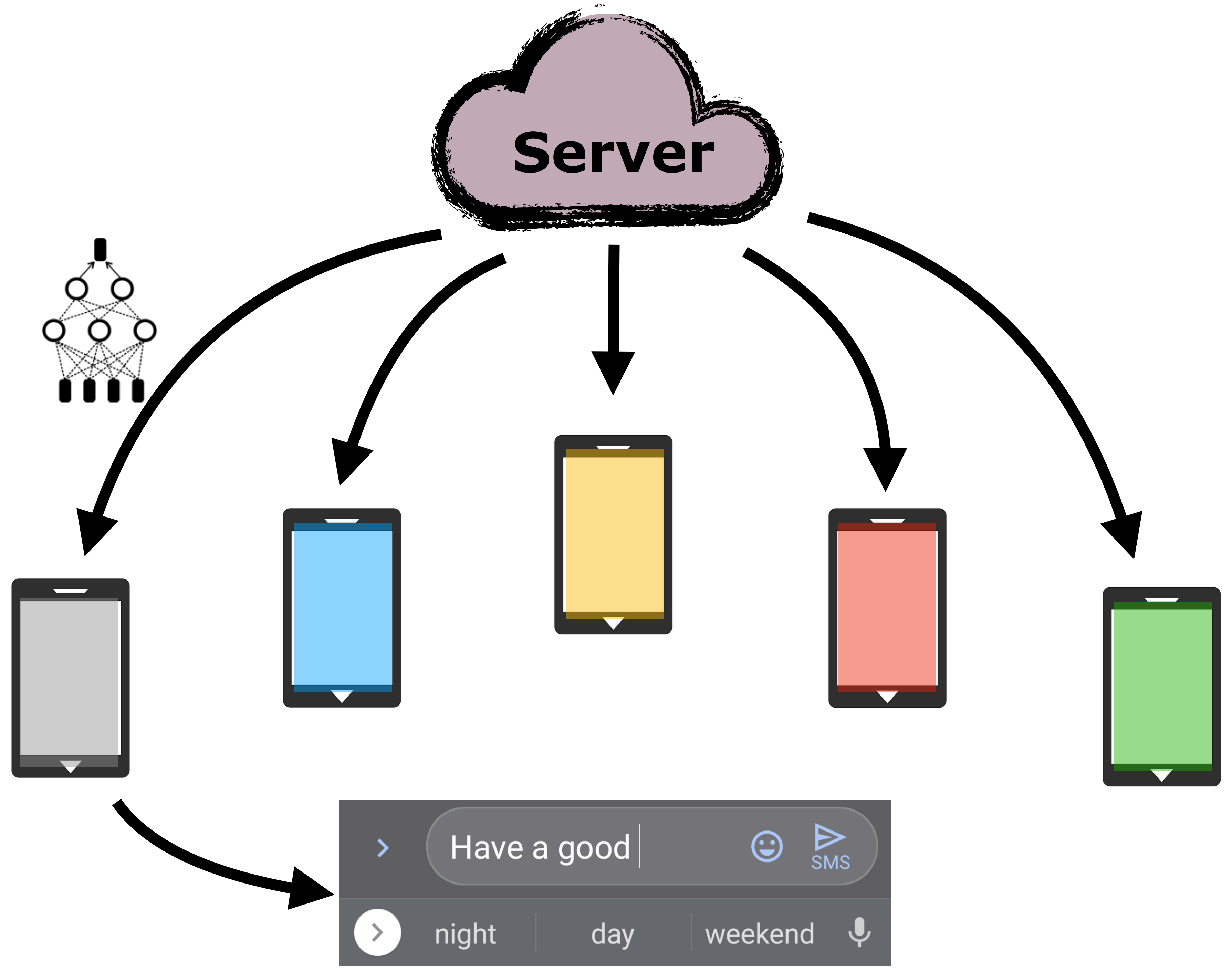


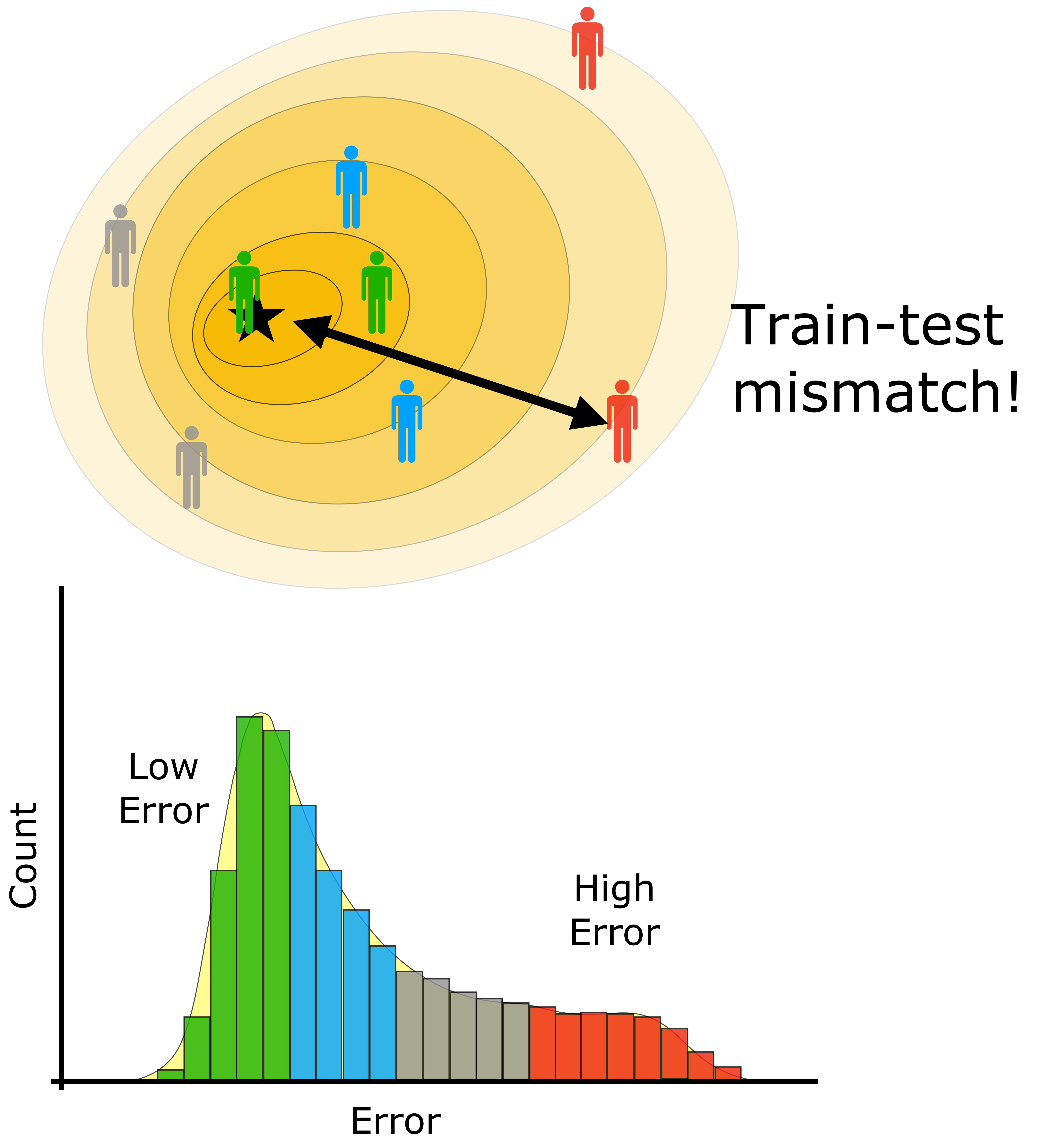
Global model is deployed on *individual* clients





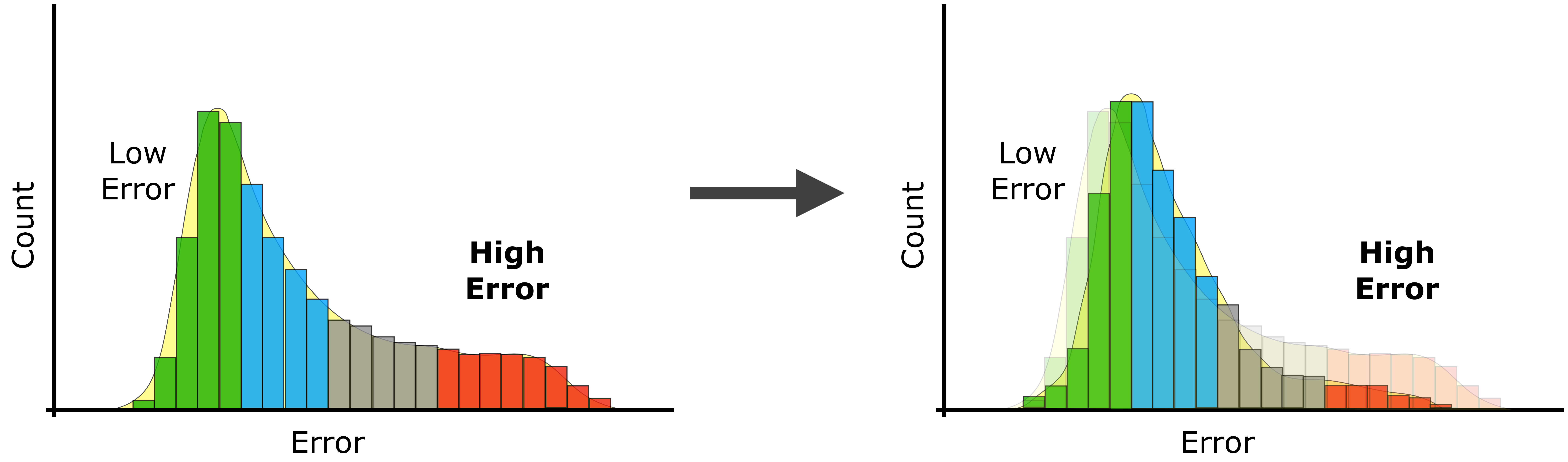
Global model is deployed on *individual* clients





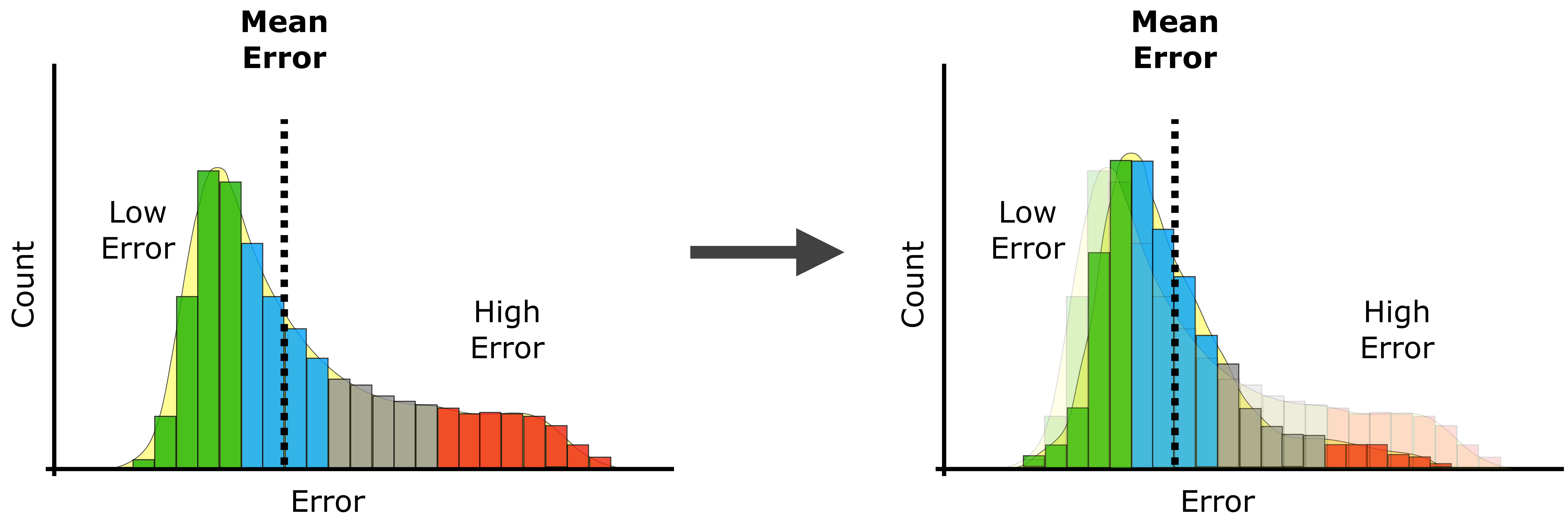
Our goal

Reduce *tail* error



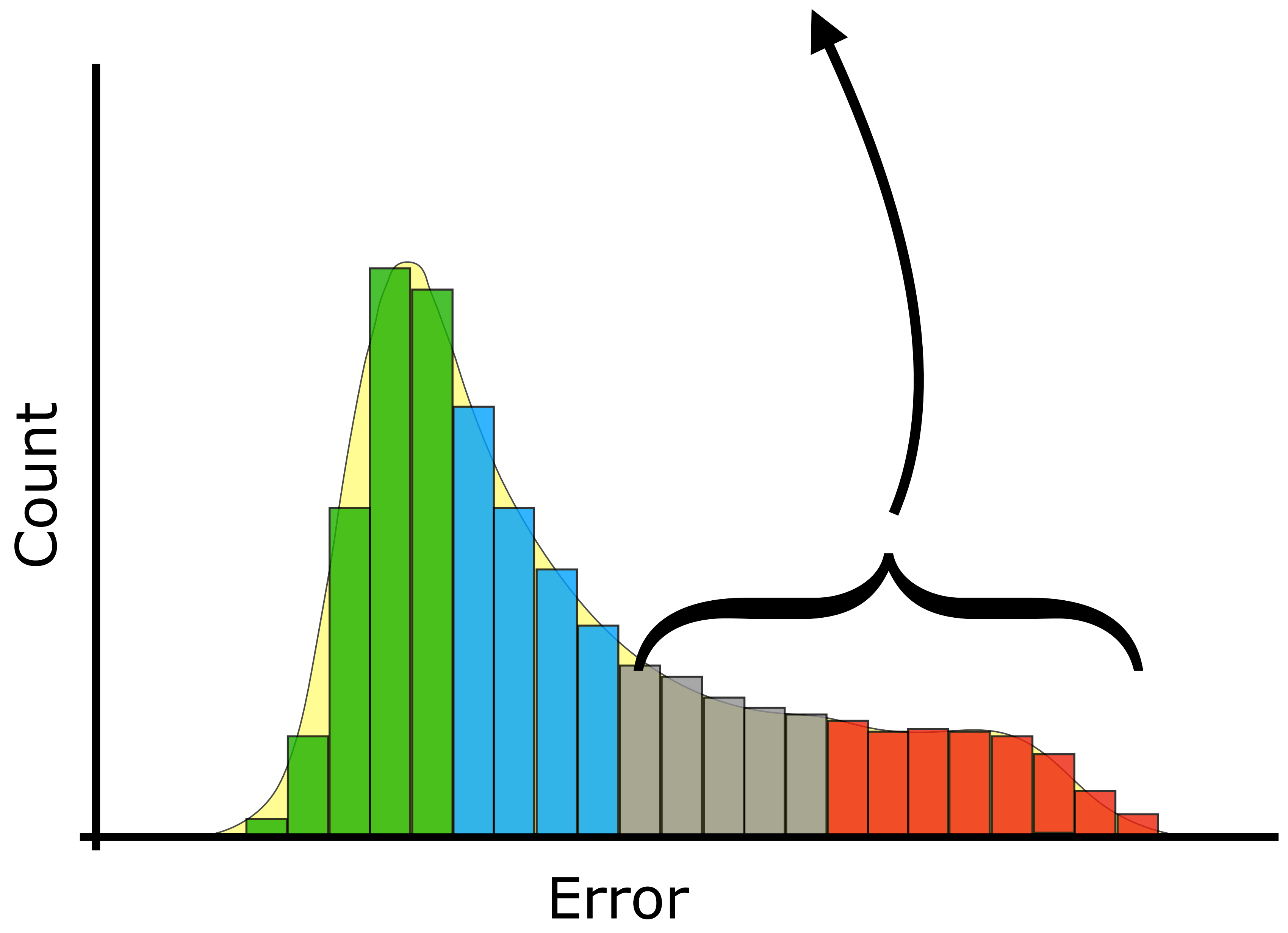
Our goal

Reduce *tail* error without sacrificing the *mean* error



Simplicial federated learning

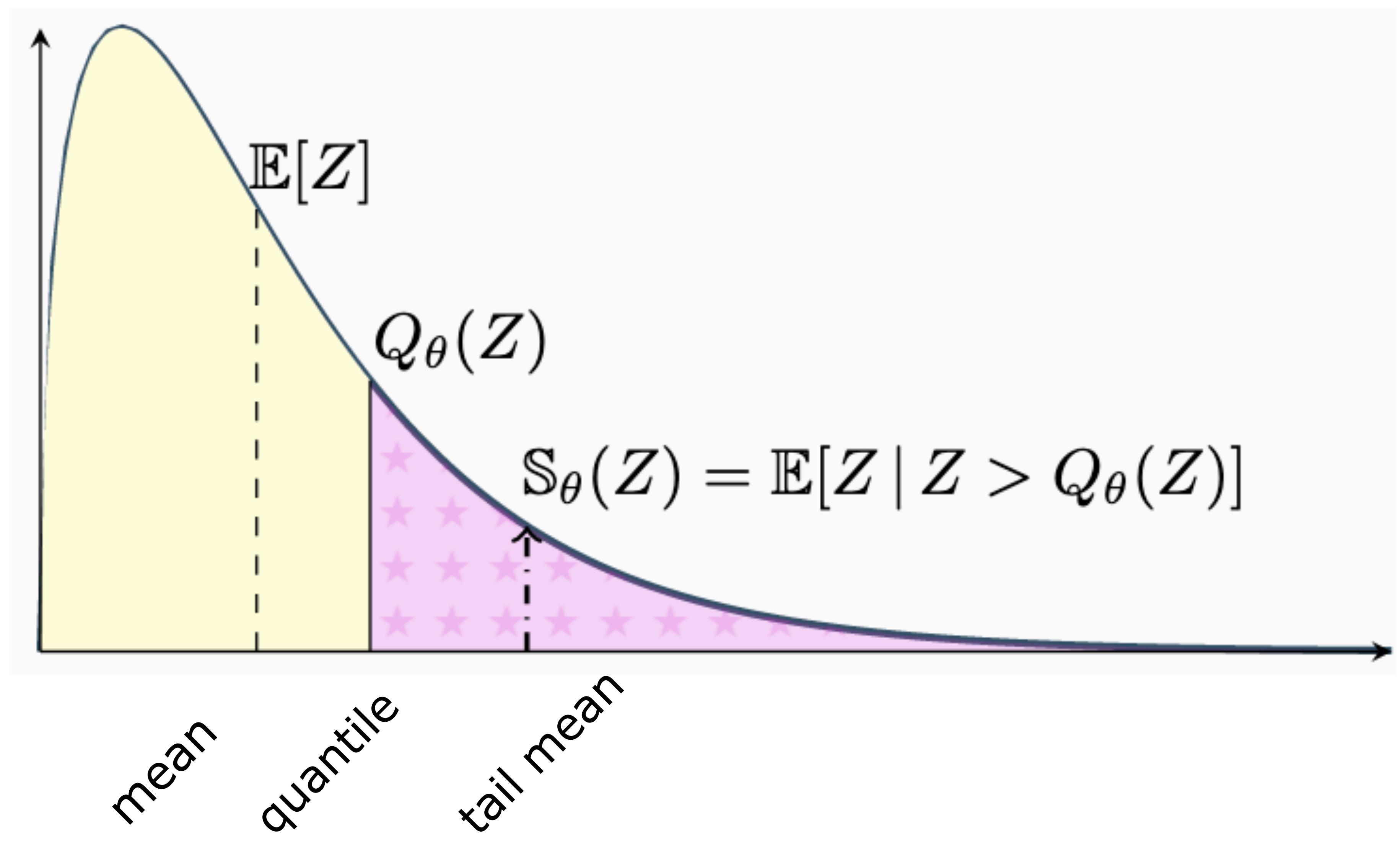
Our Approach: minimize the tail error directly!



Simplicial-FL Objective:

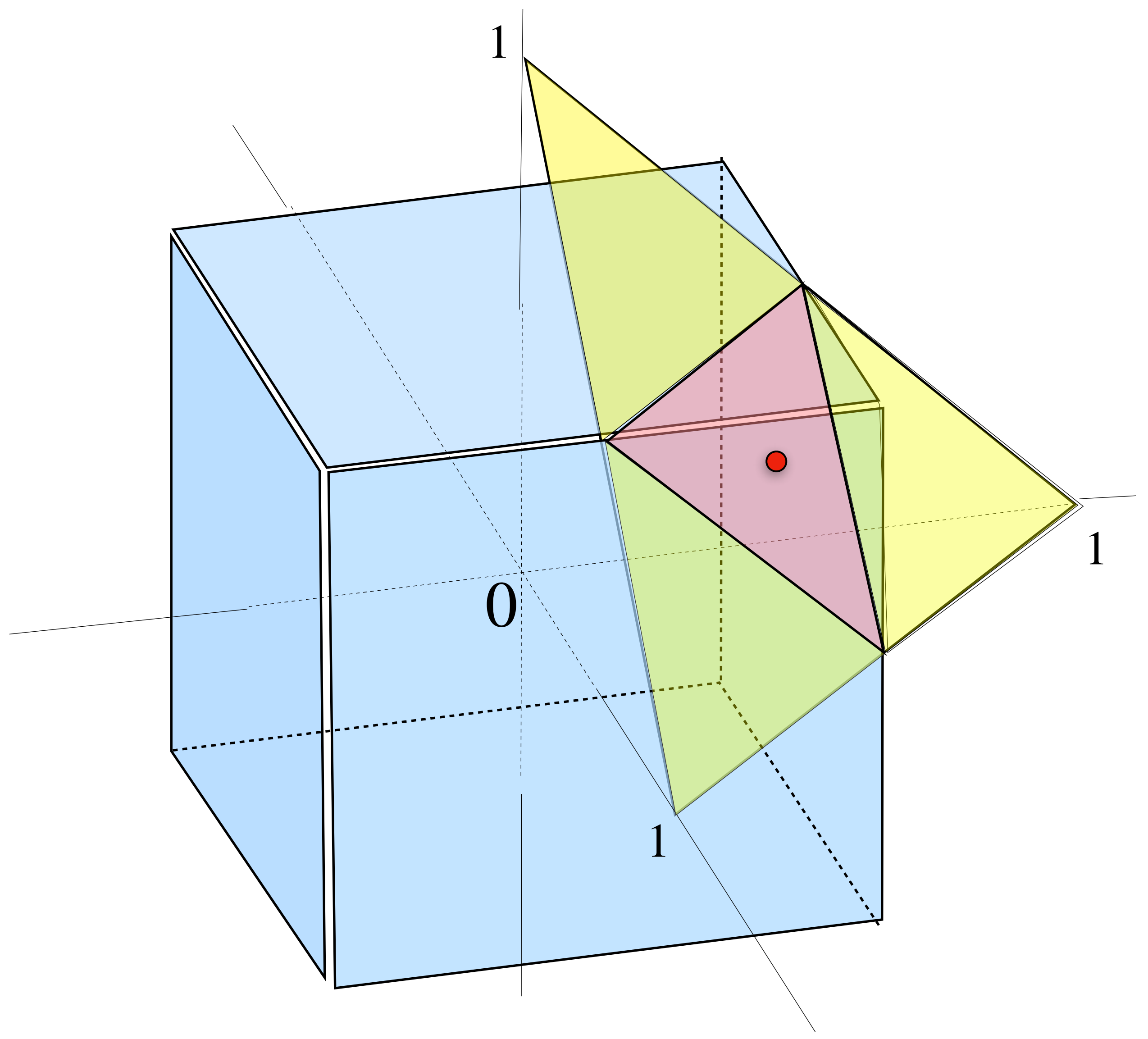
$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Superquantile | Conditional Value at Risk



[Rockafellar & Uryasev (2002)]

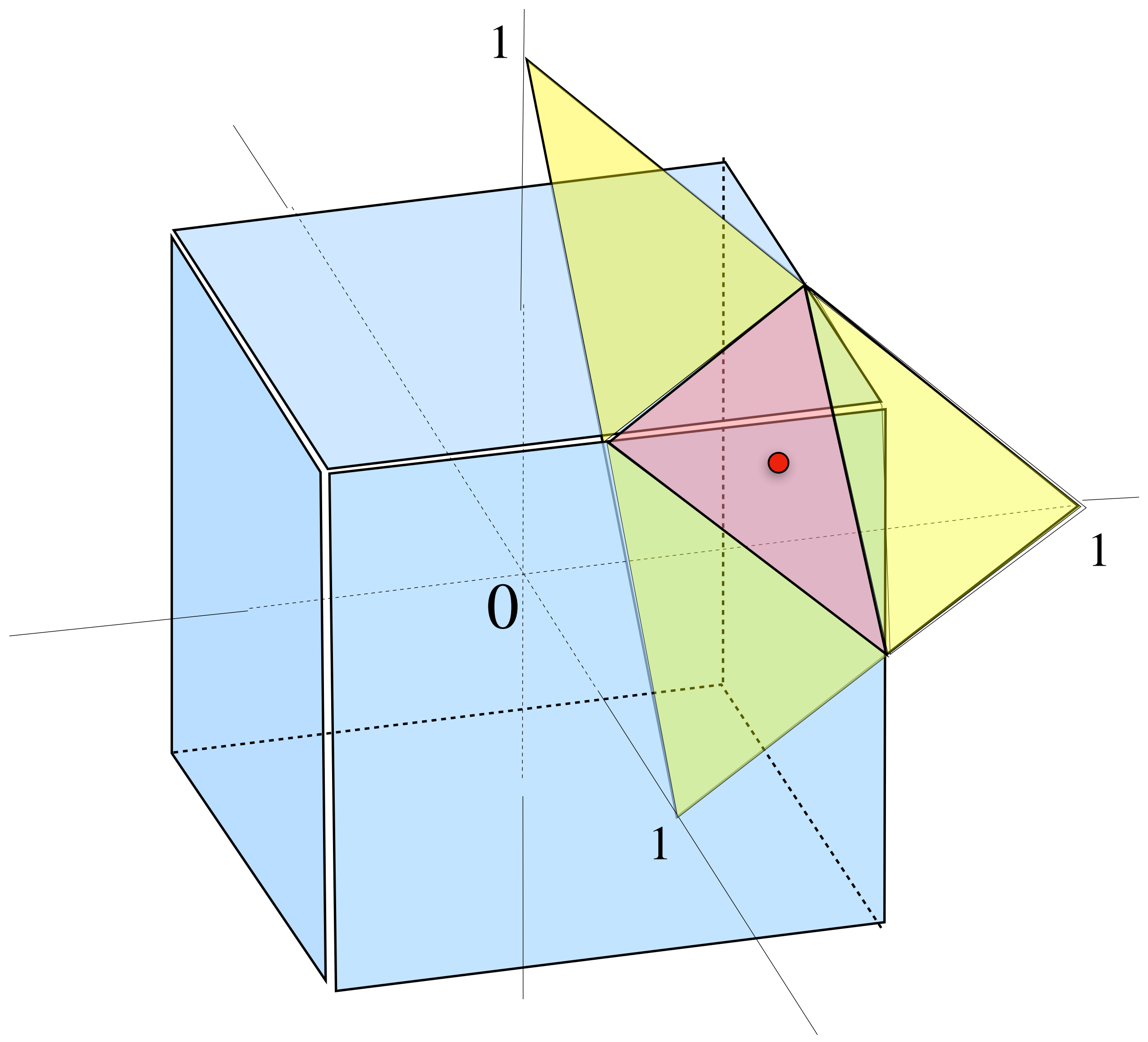
Dual expression ≡ continuous knapsack problem



$$\mathbb{S}_\theta(x_1, \cdots, x_n) = \max_\pi \left\{ \sum_i \pi_i x_i \ : \ \pi_i \geq 0, \ \sum_i \pi_i = 1, \ \pi_i \leq (n\theta)^{-1} \right\}$$

[Dantzig (1957), Ben-Tal & Teboulle (1987), Föllmer & Schied (2002)]

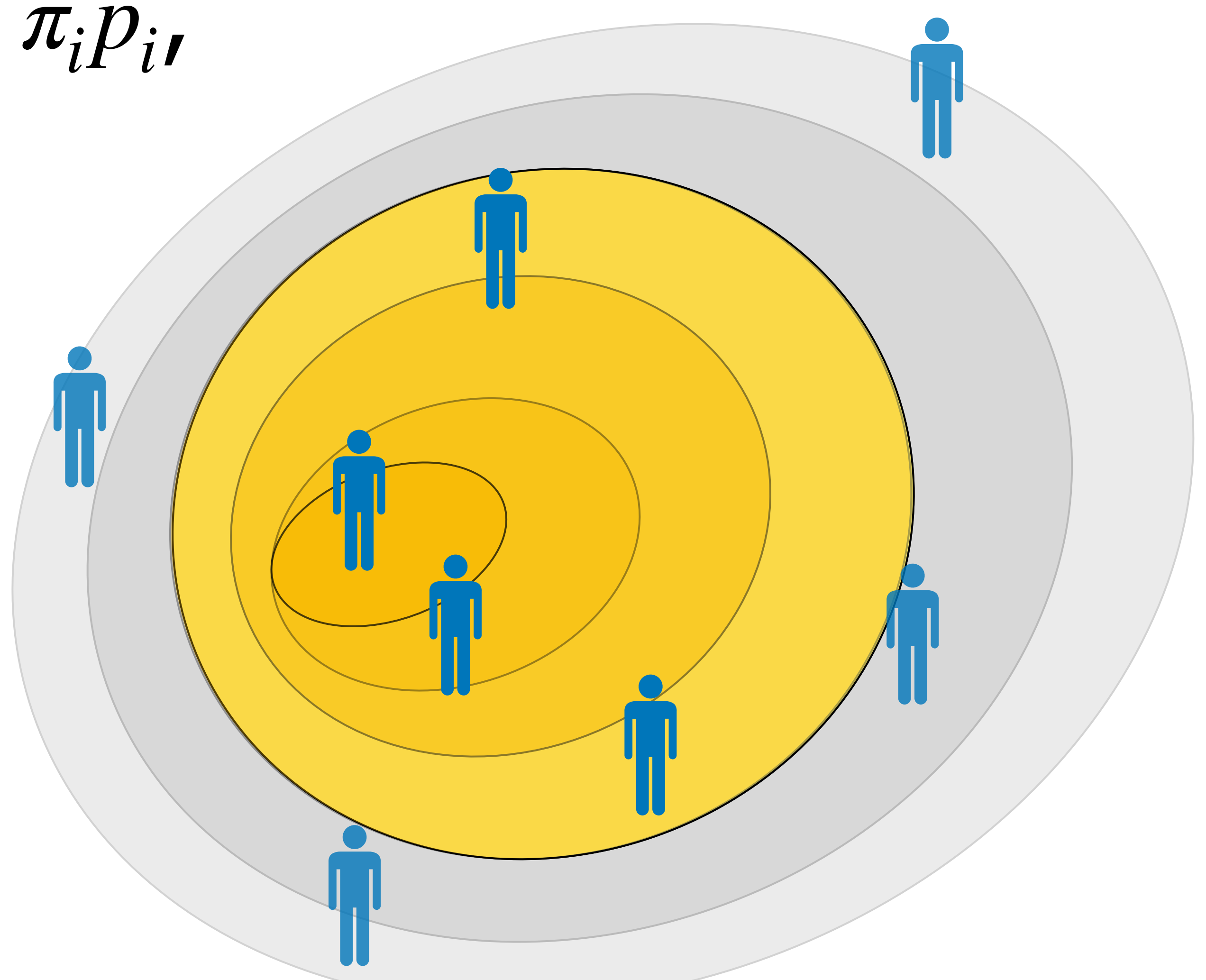
Dual expression ≡ continuous knapsack problem



$$\mathbb{S}_\theta(x_1, \cdots, x_n) = \max_\pi \left\{ \sum_i \pi_i x_i \, : \, \pi_i \geq 0, \ \sum_i \pi_i = 1, \ \pi_i \leq (n\theta)^{-1} \right\}$$

[Dantzig (1957), Ben-Tal & Teboulle (1987), Föllmer & Schied (2002)]

Assuming a new test client with mixture distribution $p_\pi = \sum_i \pi_i p_i$,



Simplicial-FL objective is equivalent to:

$$\min_w \max_{\pi : \pi_i \leq (n\theta)^{-1}} \mathbb{E}_{z \sim p_\pi} \left[f(w; z) \right]$$

$\Longrightarrow$ Distributionally robust learning

Optimization

Usual Algorithm (FedAvg):

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

Our Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

FedAvg [MacMahan et al. (AISTATS 2017)]

Parallel Gradient Distribution [Mangasarian. (SICON 1995)]
Iterative Parameter Mixing [McDonald et al. (ACL 2009)]
BMUF [Chen & Huo. (ICASSP 2016)]
Local SGD [Stich. (ICLR 2019)]

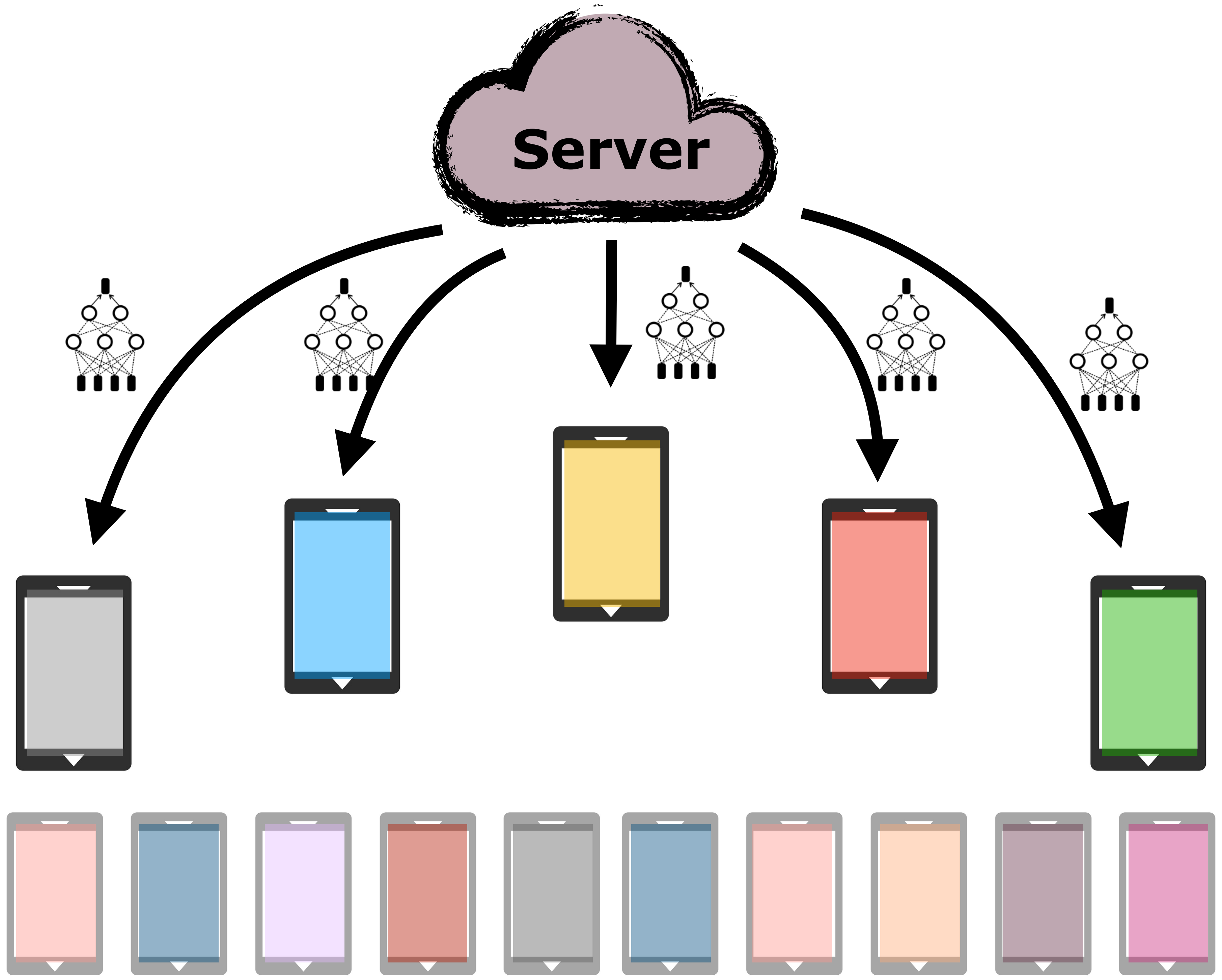
Usual Algorithm (FedAvg):

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

Our Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Step 1 of 3: Server samples m clients and broadcasts global model



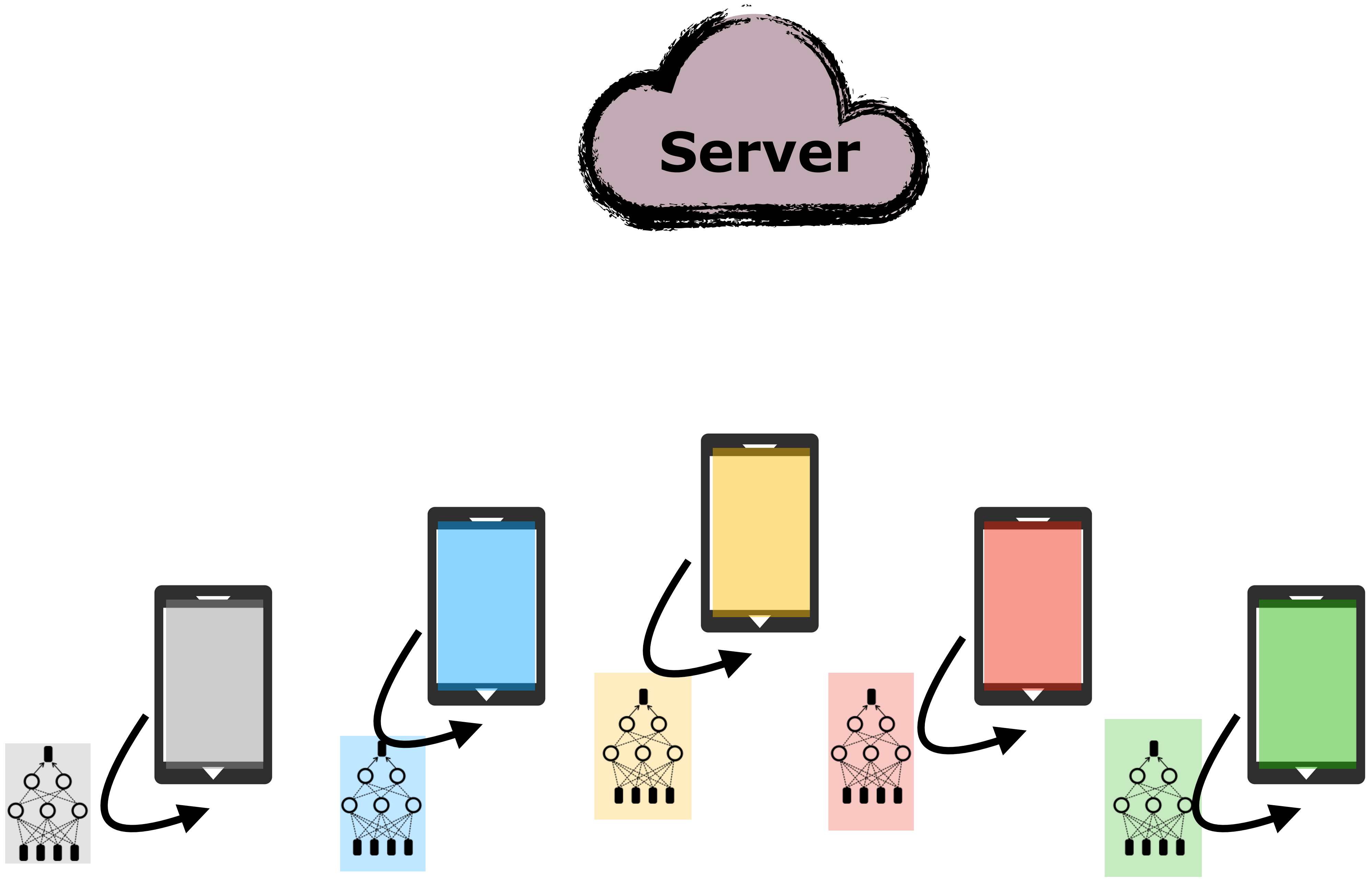
Usual Algorithm (FedAvg):

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

Our Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Step 2 of 3: Clients perform local gradient descent on their local data



Usual Algorithm (FedAvg):

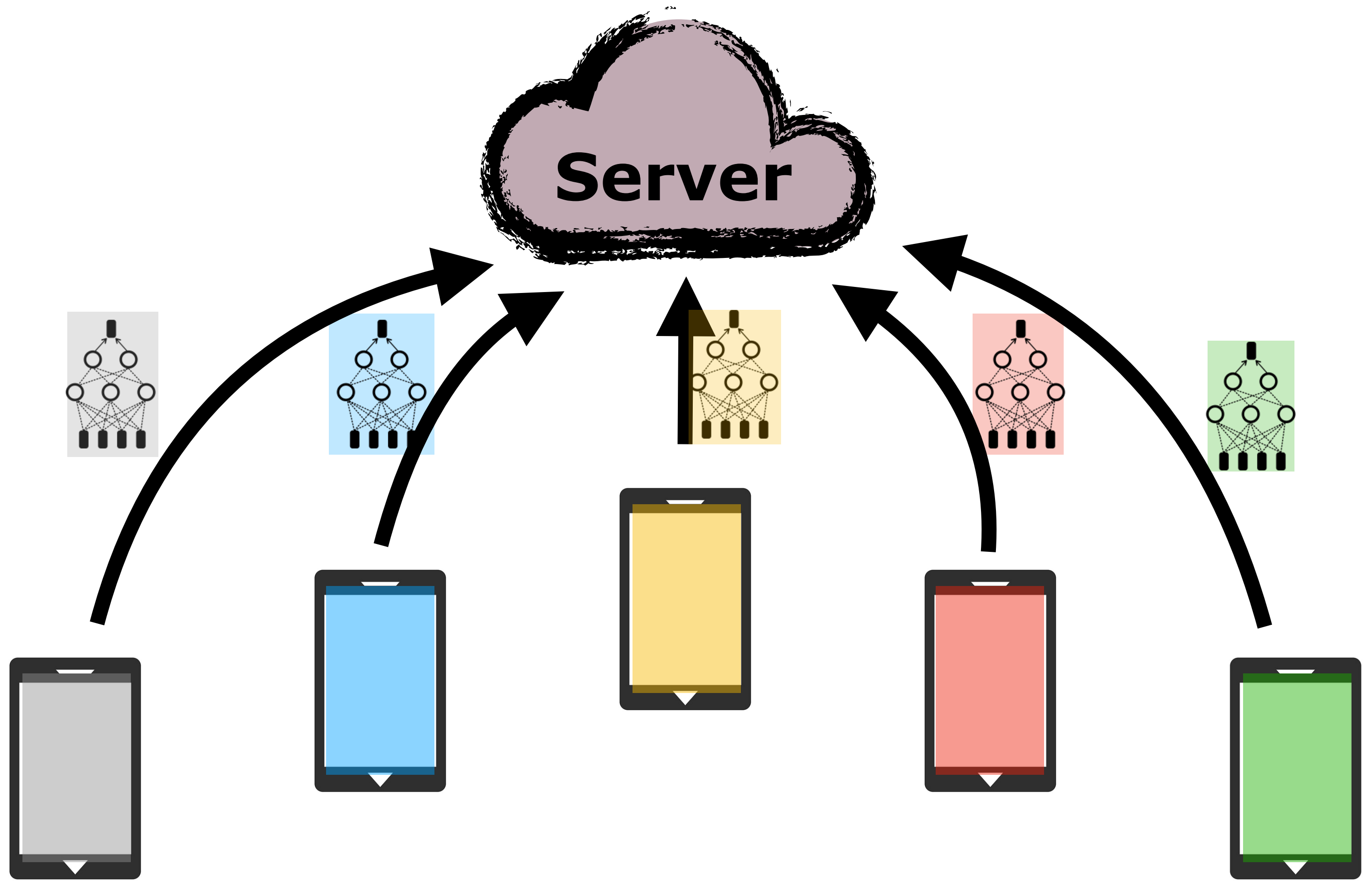
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

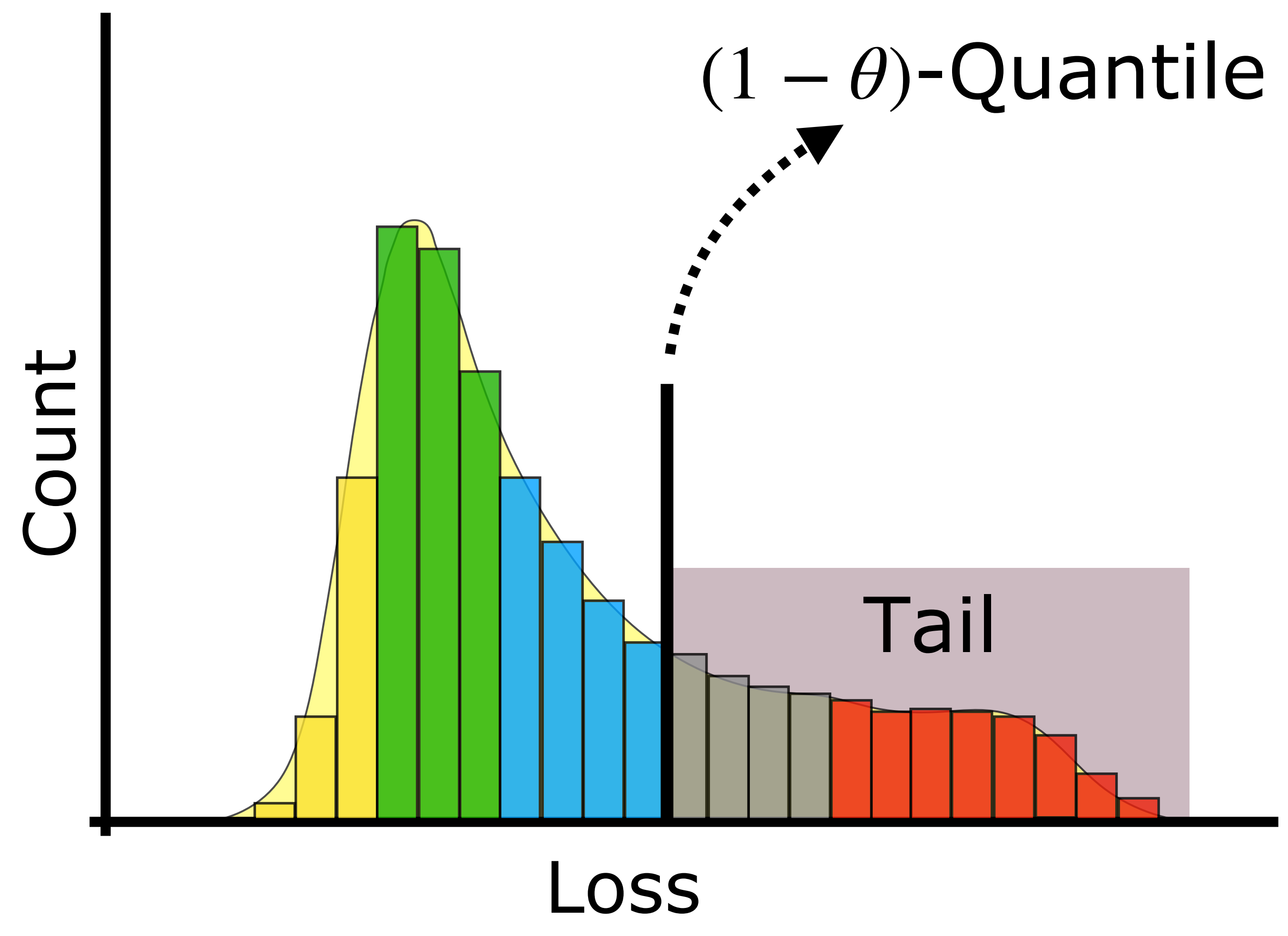
Step 3 of 3: Aggregate updates contributed by ***all clients***

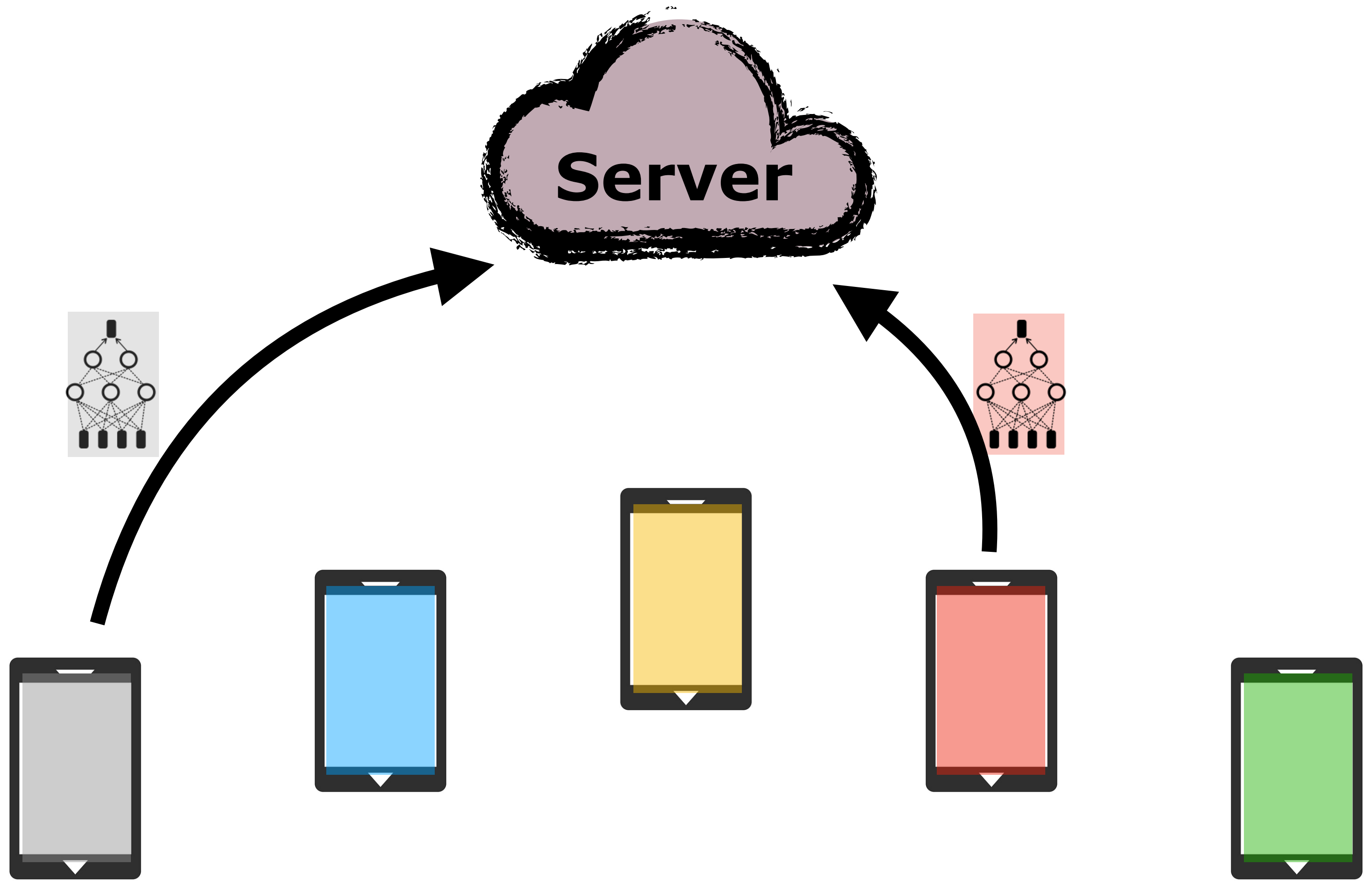
Our Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

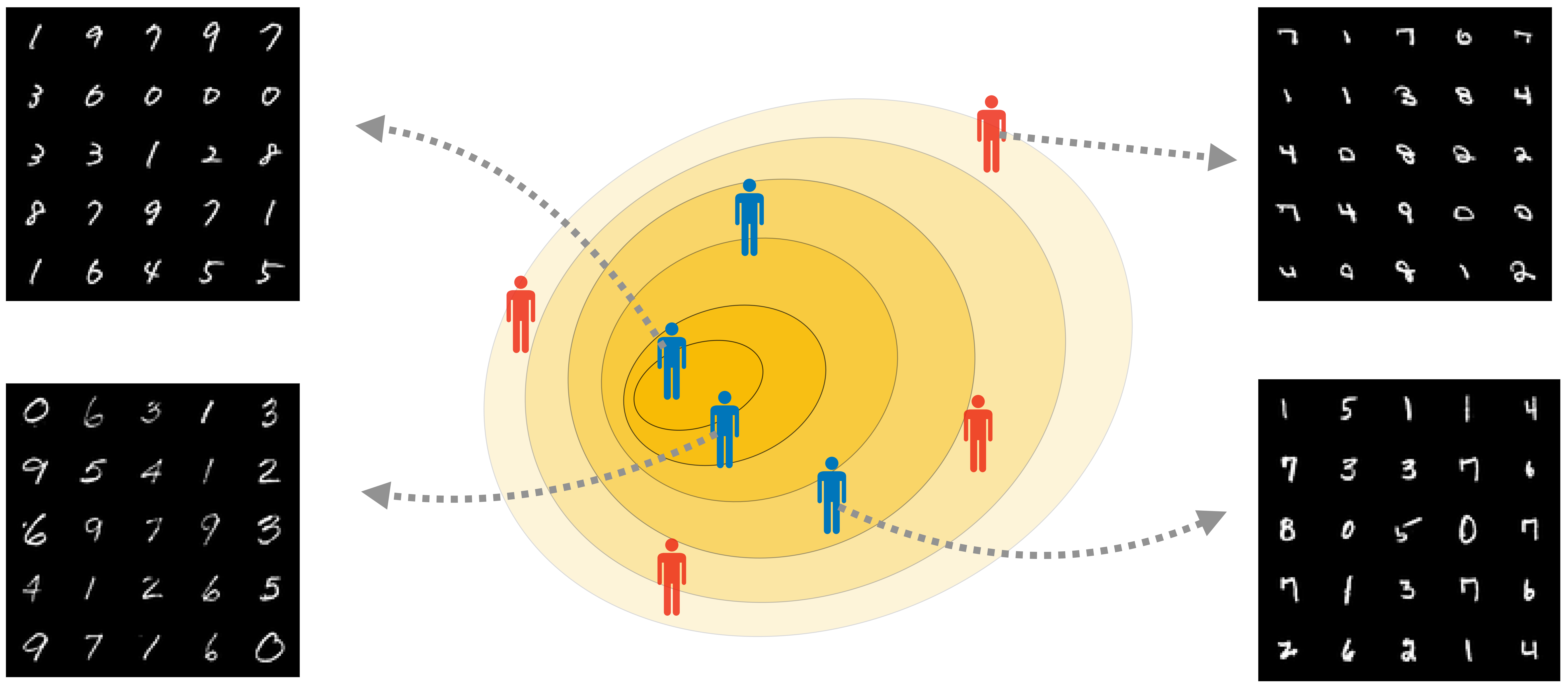
Step 3 of 3: Aggregate updates contributed by ***tail clients*** *only*



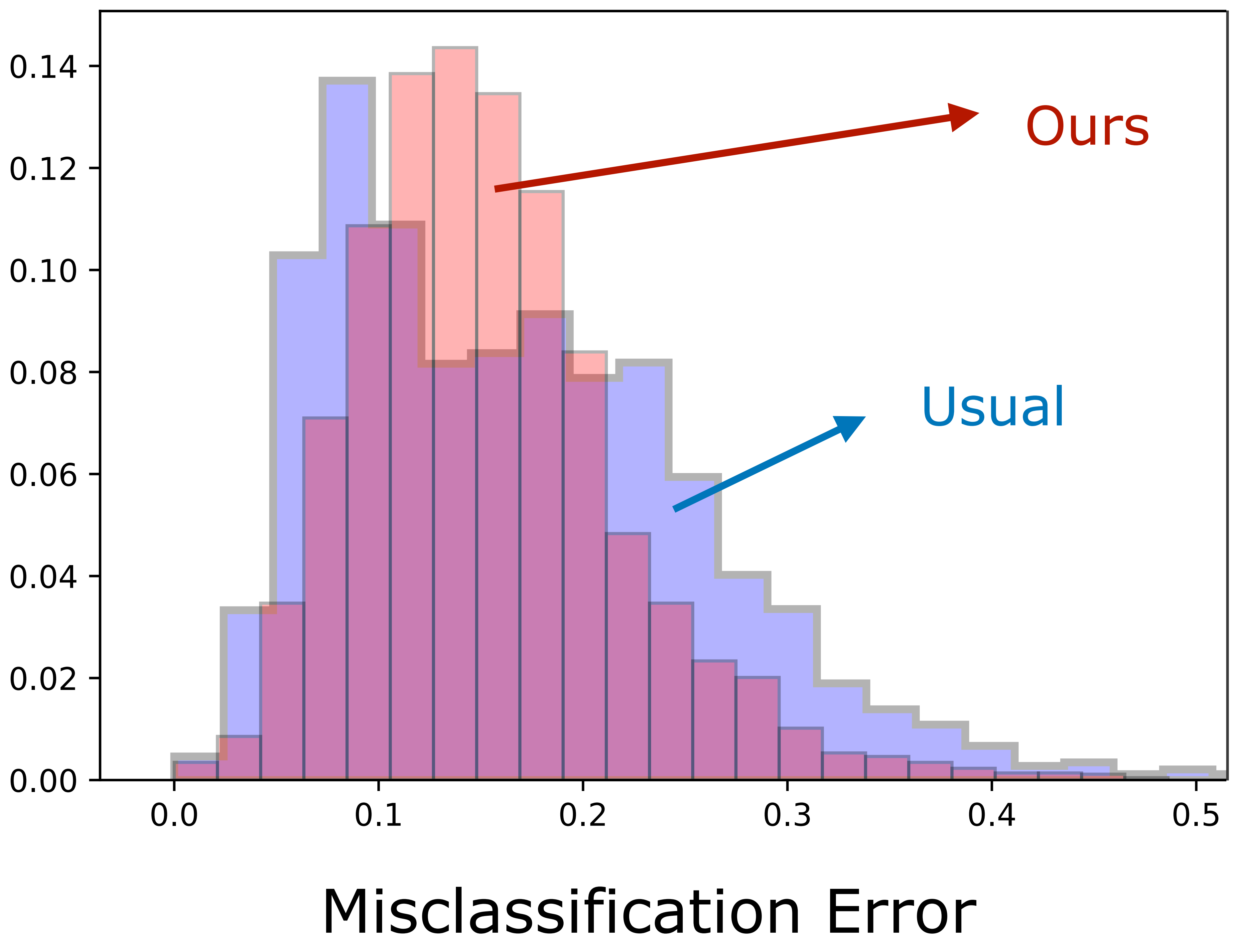




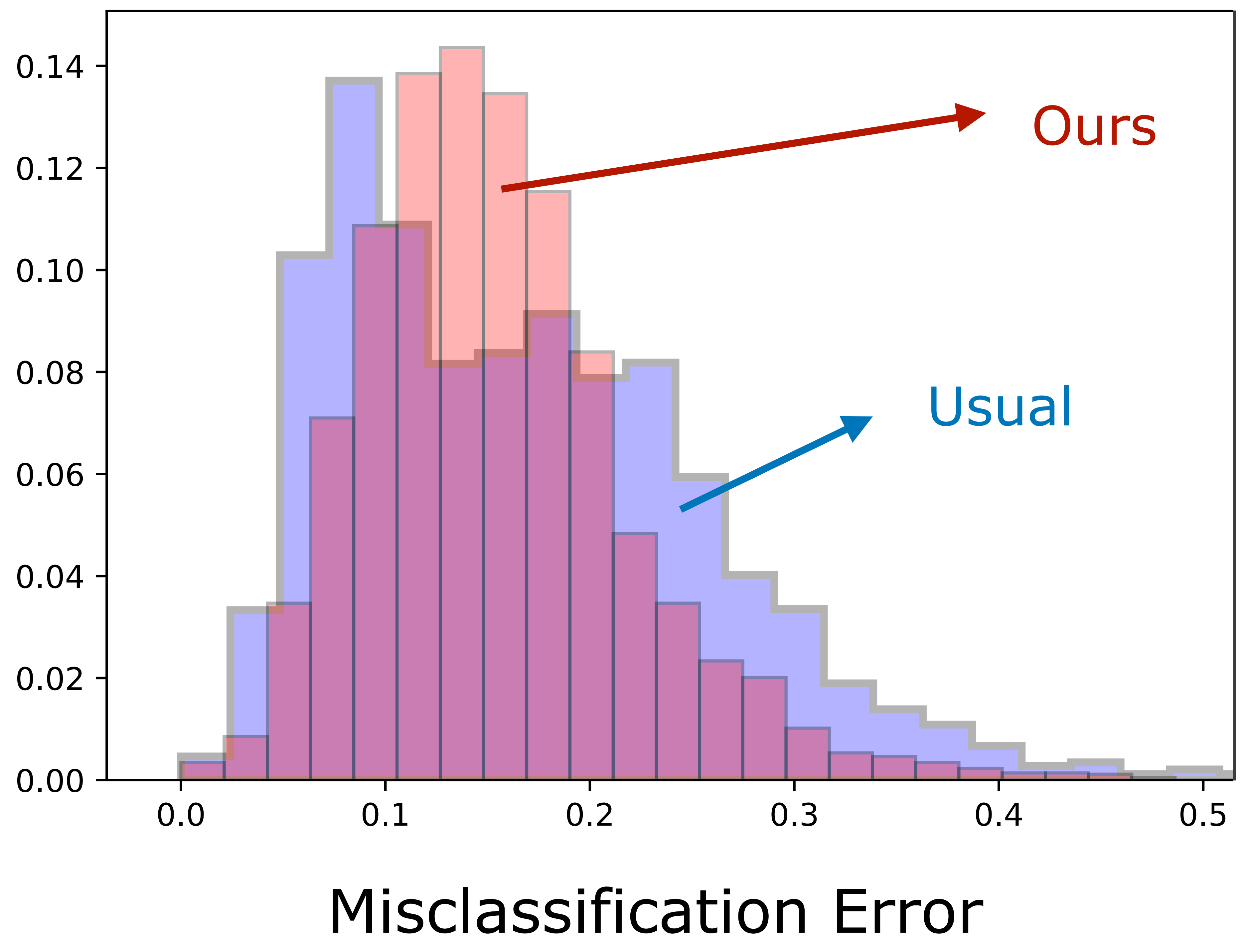
Experiments: EMNIST



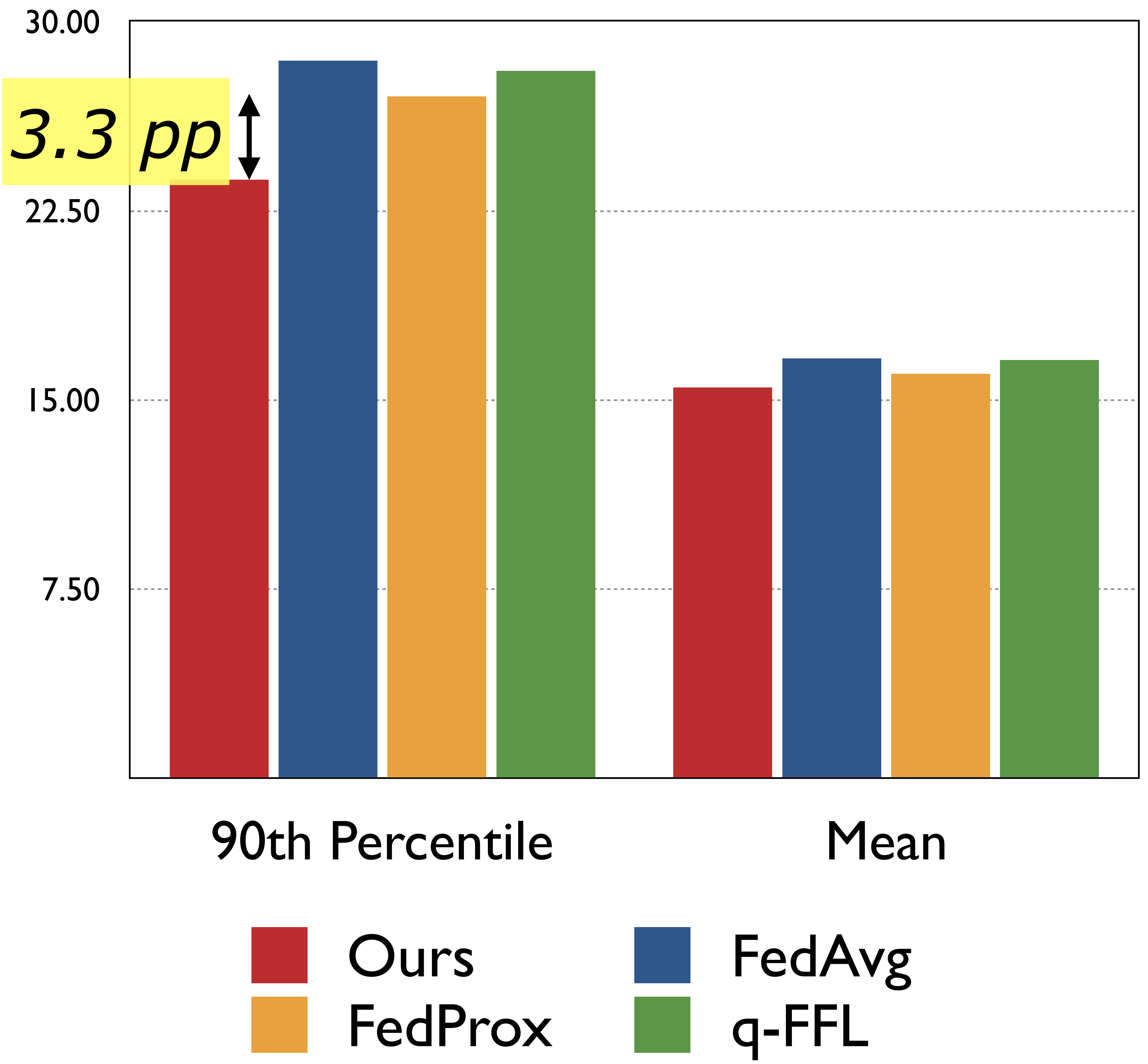
Histogram of per-client errors



Histogram of per-client errors

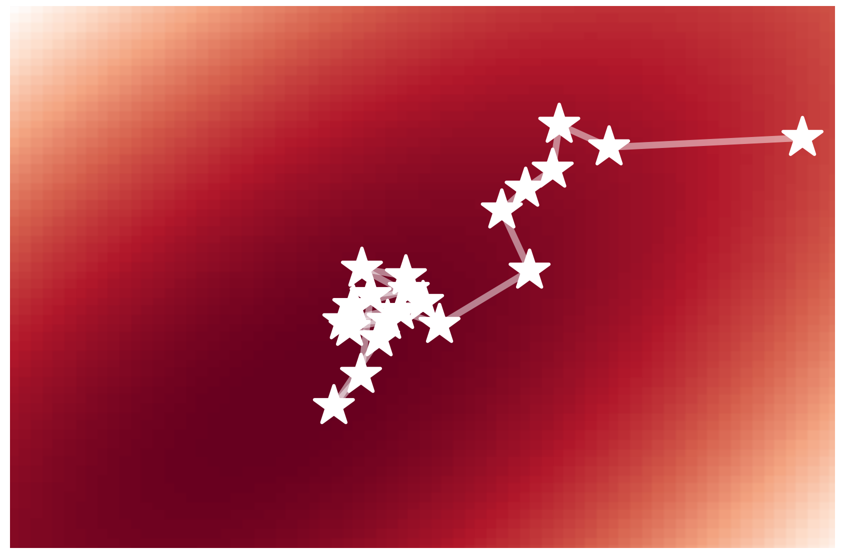


Misclassif. Error



- Simplicial-FL has the smallest 90^{th} percentile error
- Simplicial-FL is competitive on the mean error

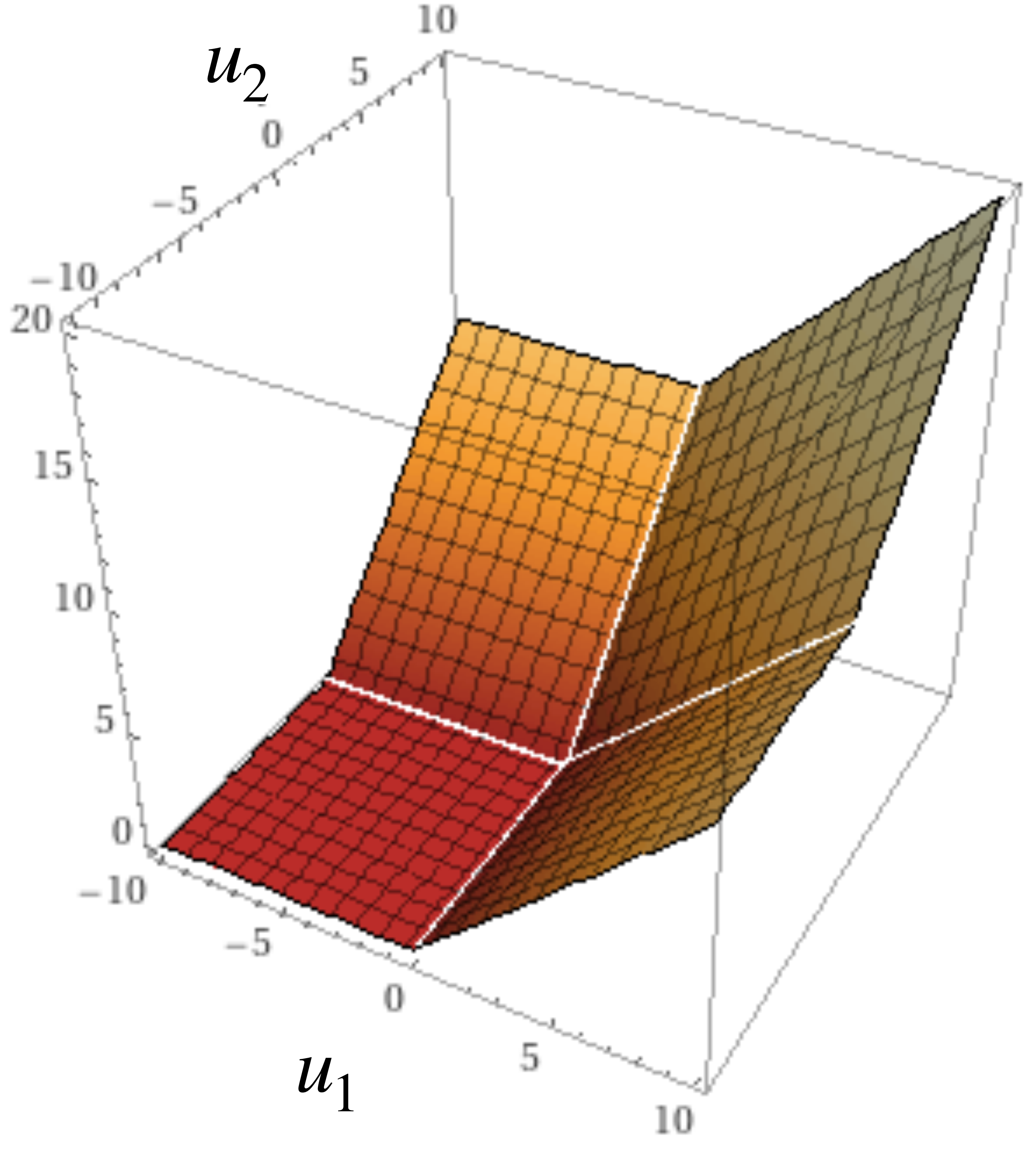
Convergence analysis (non-convex)



Faster optimization:
reduce communication

Challenge #1:

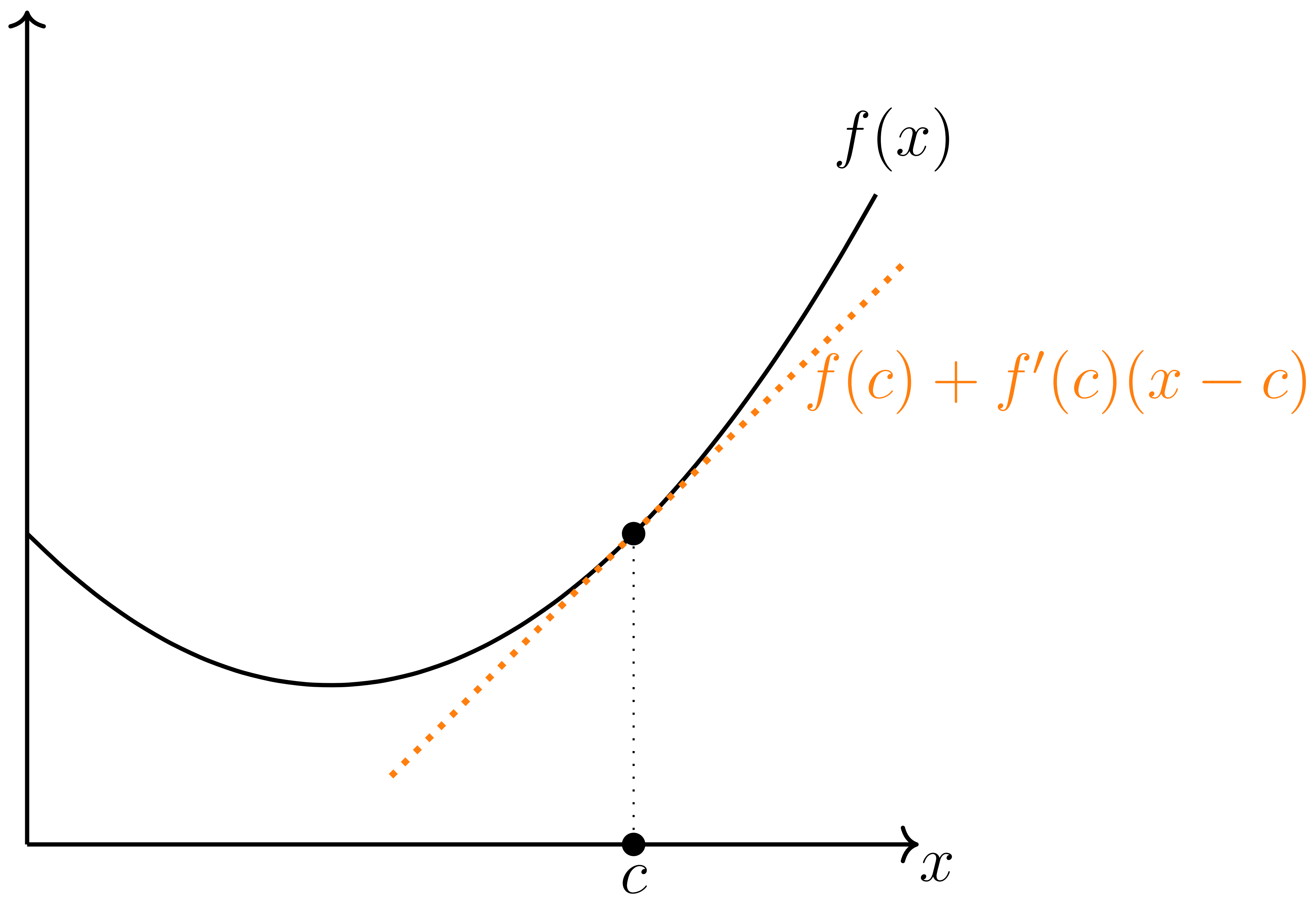
The superquantile is non-smooth



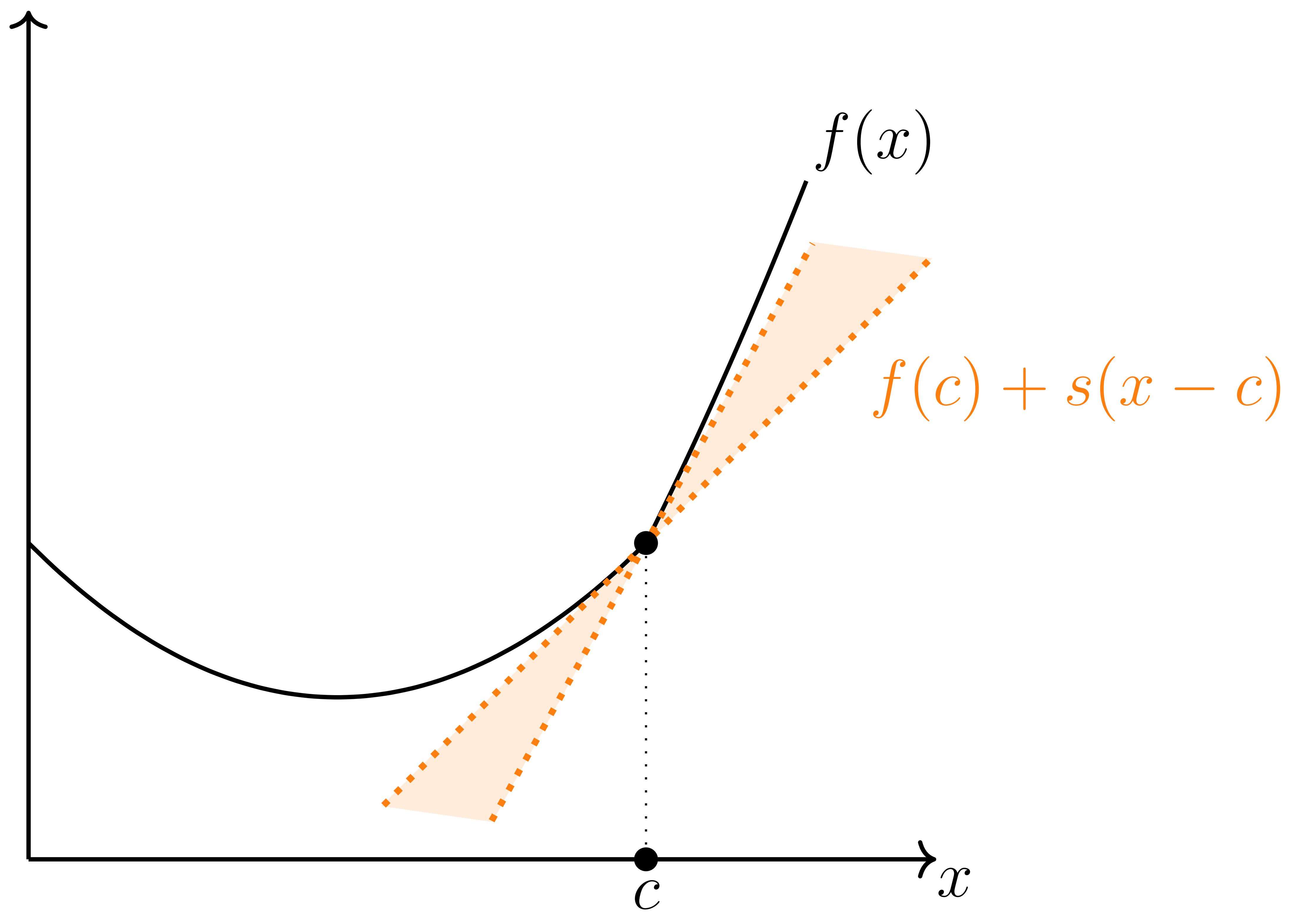
plot of $h(u_1, u_2) = \mathbb{S}_{1/2}\big(u_1, u_2, 0, 0 \big)$

Subgradient illustration

Smooth



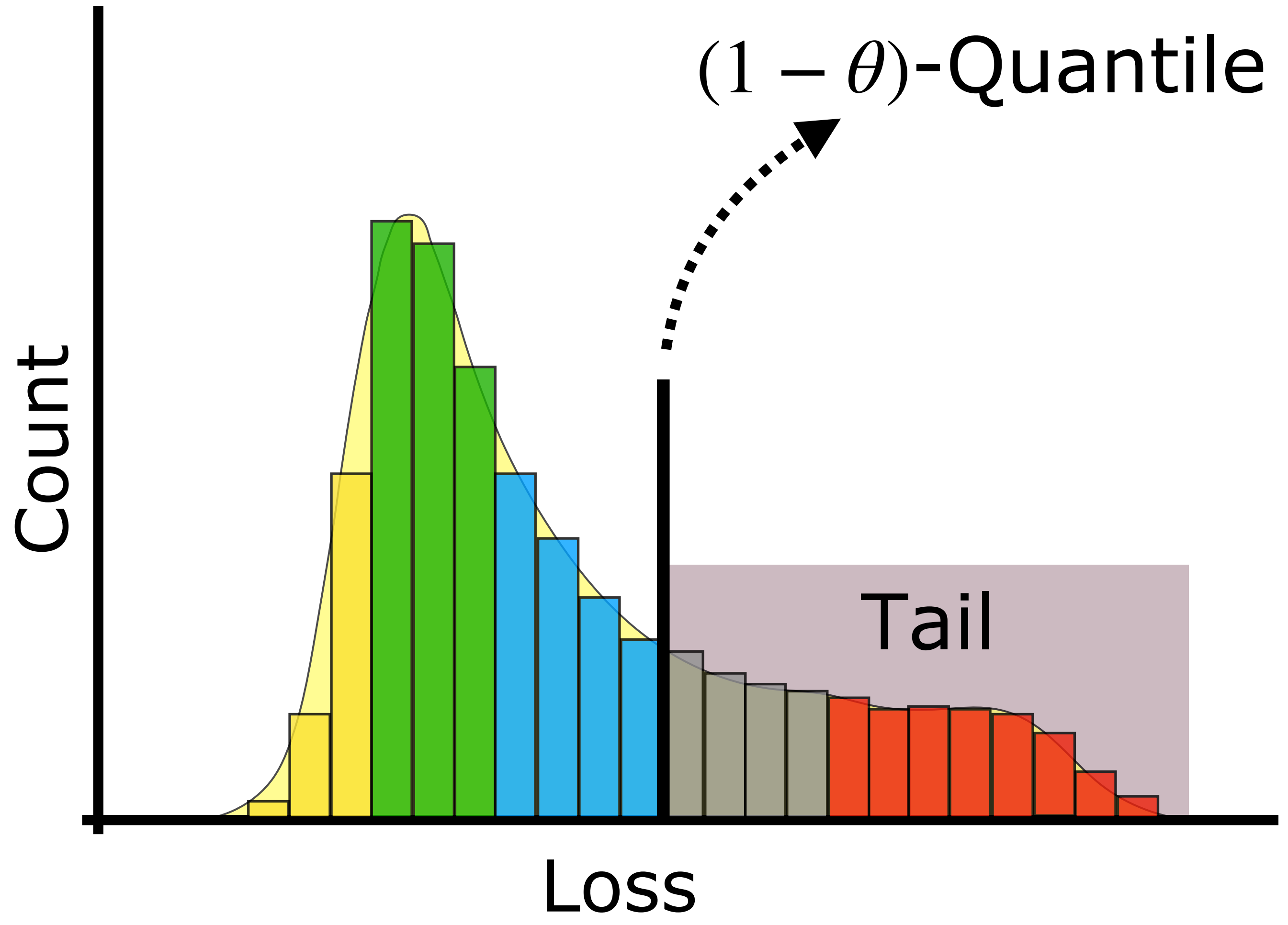
Non-smooth



Nonsmooth: The subgradient has a tractable form

$$\partial F_\theta(w) \ni \sum_{i=1}^{n} \pi_i^\star \nabla F_i(w) \qquad \text{where} \qquad \pi_i^\star \propto \mathbb{1}\Big(F_i(w) \geq Q_\theta\big(F_1(w), \cdots, F_n(w)\big)\Big)$$

assuming θn is an integer



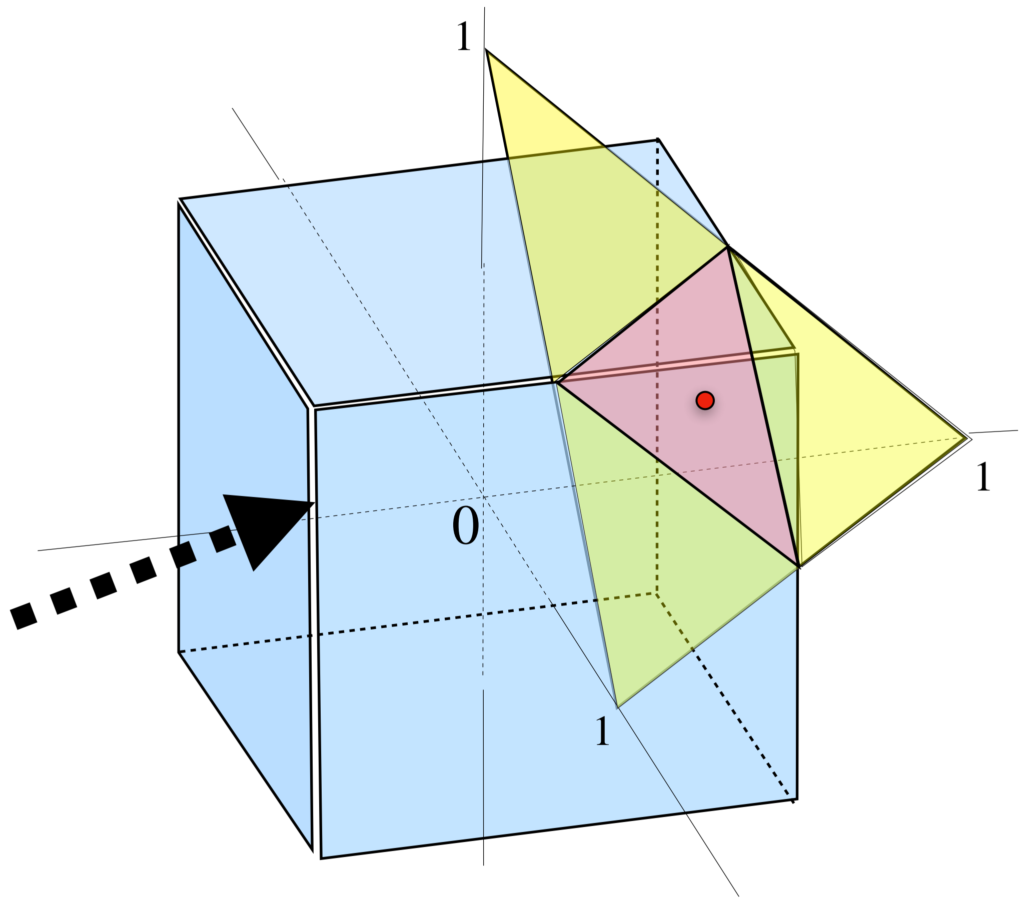
Nonsmooth: The subgradient has a tractable form

$$\partial F_\theta(w) \ni \sum_{i=1}^{n} \pi_i^\star \nabla F_i(w) \quad \text{where} \quad \pi_i^\star \propto \mathbb{I}\Big(F_i(w) \geq Q_\theta\big(F_1(w), \cdots, F_n(w)\big)\Big)$$

assuming θn is an integer

Proof Chain rule $\Longrightarrow$ subgradient holds with

$$\pi^\star \in \arg\max_{\pi \in \boxed{\mathscr{P}_\theta}} \sum_i \pi_i F_i(w)$$



Alternate form of $\pi^\star$ comes from the continuous knapsack problem

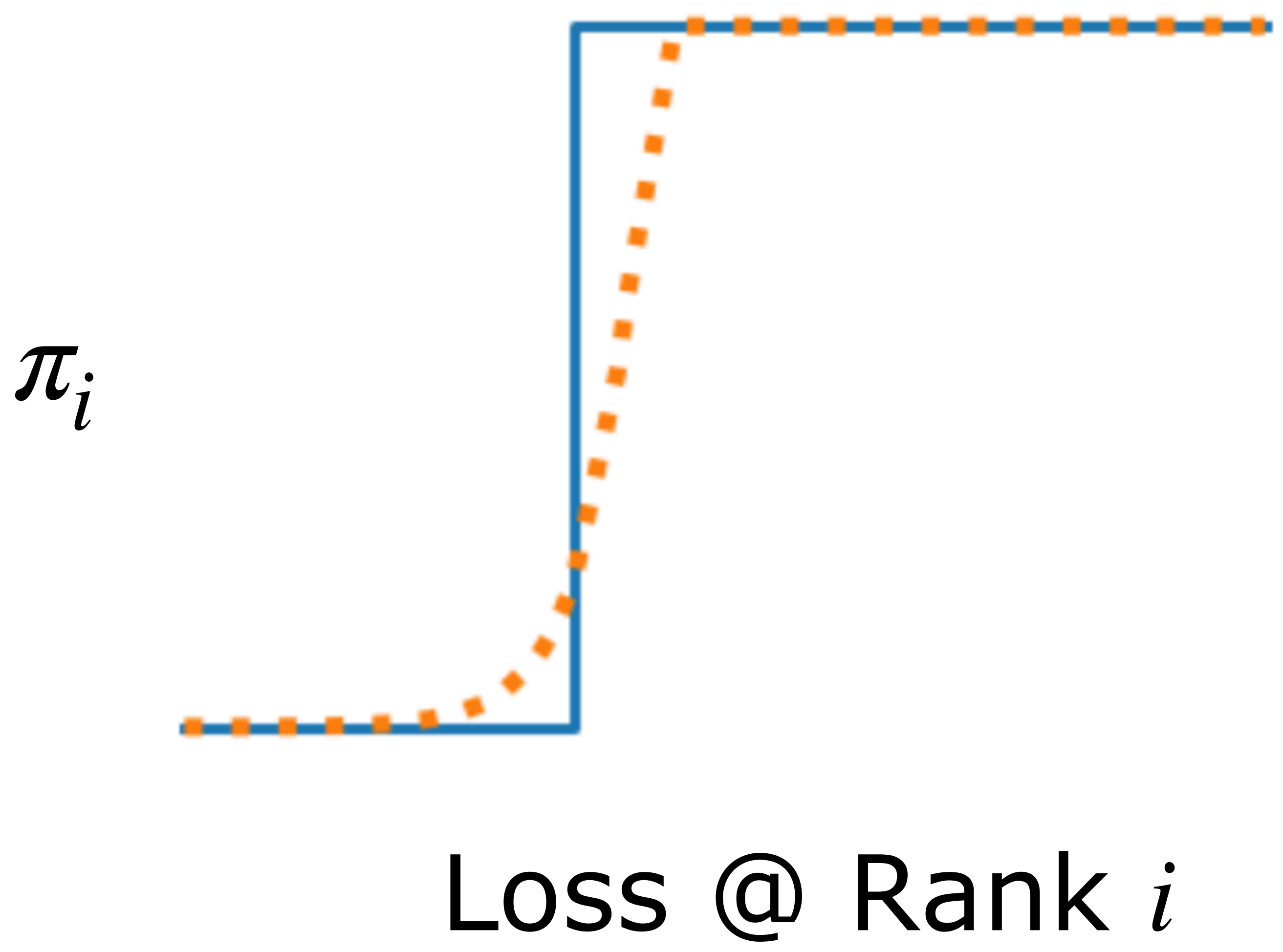
[Dantzig. ORIJ (1957)]

Nonsmooth: The subgradient has a tractable form

$$\partial F_\theta(w) \ni \sum_{i=1}^{n} \pi_i^\star \nabla F_i(w) \qquad \text{where} \qquad \pi_i^\star \propto \mathbb{1}\Big(F_i(w) \geq Q_\theta\big(F_1(w), \cdots, F_n(w)\big)\Big)$$

assuming θn is an integer

Other option: Use smoothing

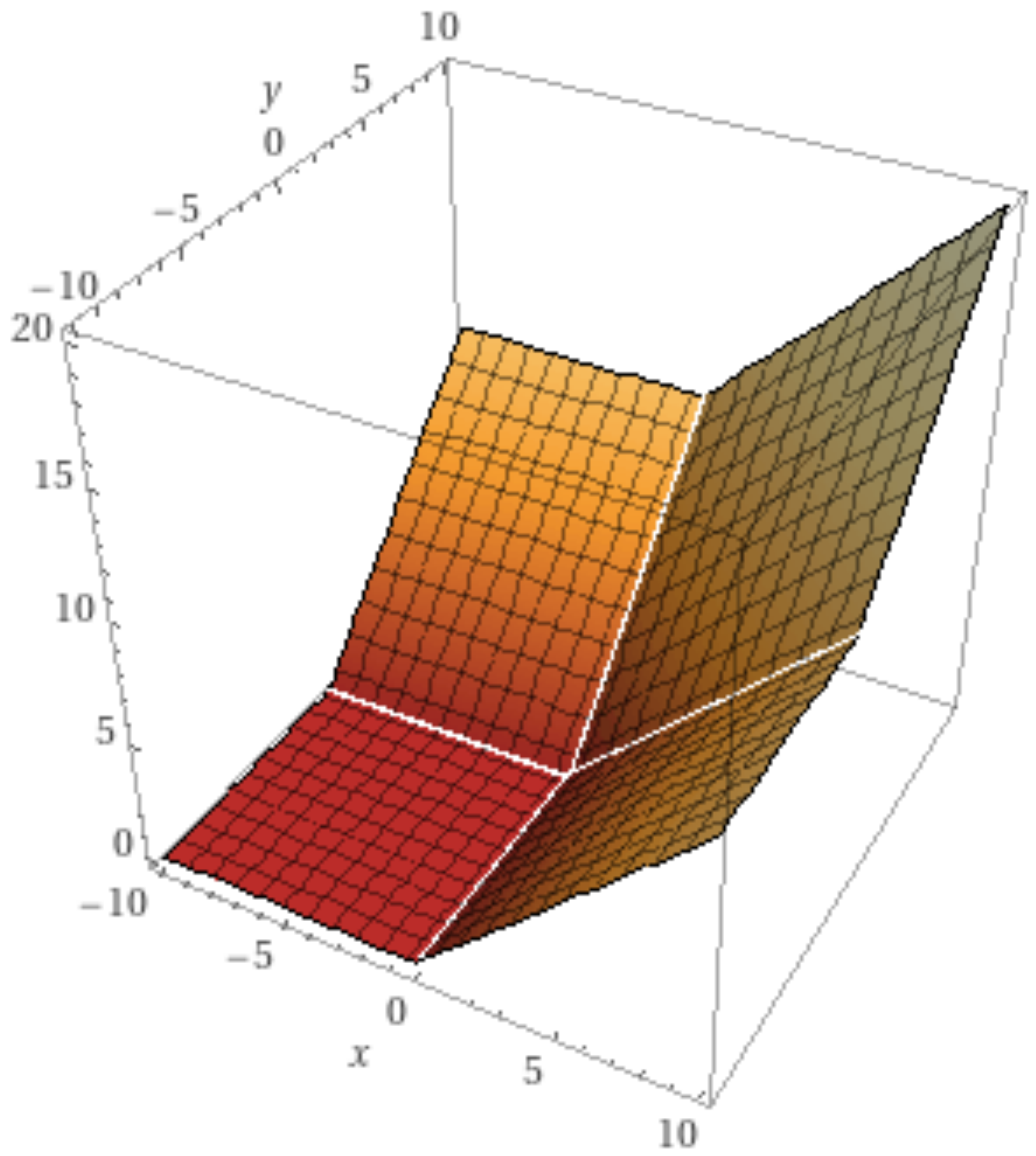


Smoothed $\pi_i^\star$ Original $\pi_i^\star$

[Nesterov. (Math. Prog. 2005),
Beck & Teboulle. (SIAM J. Optim. 2012),
P., Roulet, Kakade, Harchaoui. (NeurIPS 2018),
Laguel, **P.**, Malick, Harchaoui. (SVVA 2021)]

Challenge #2

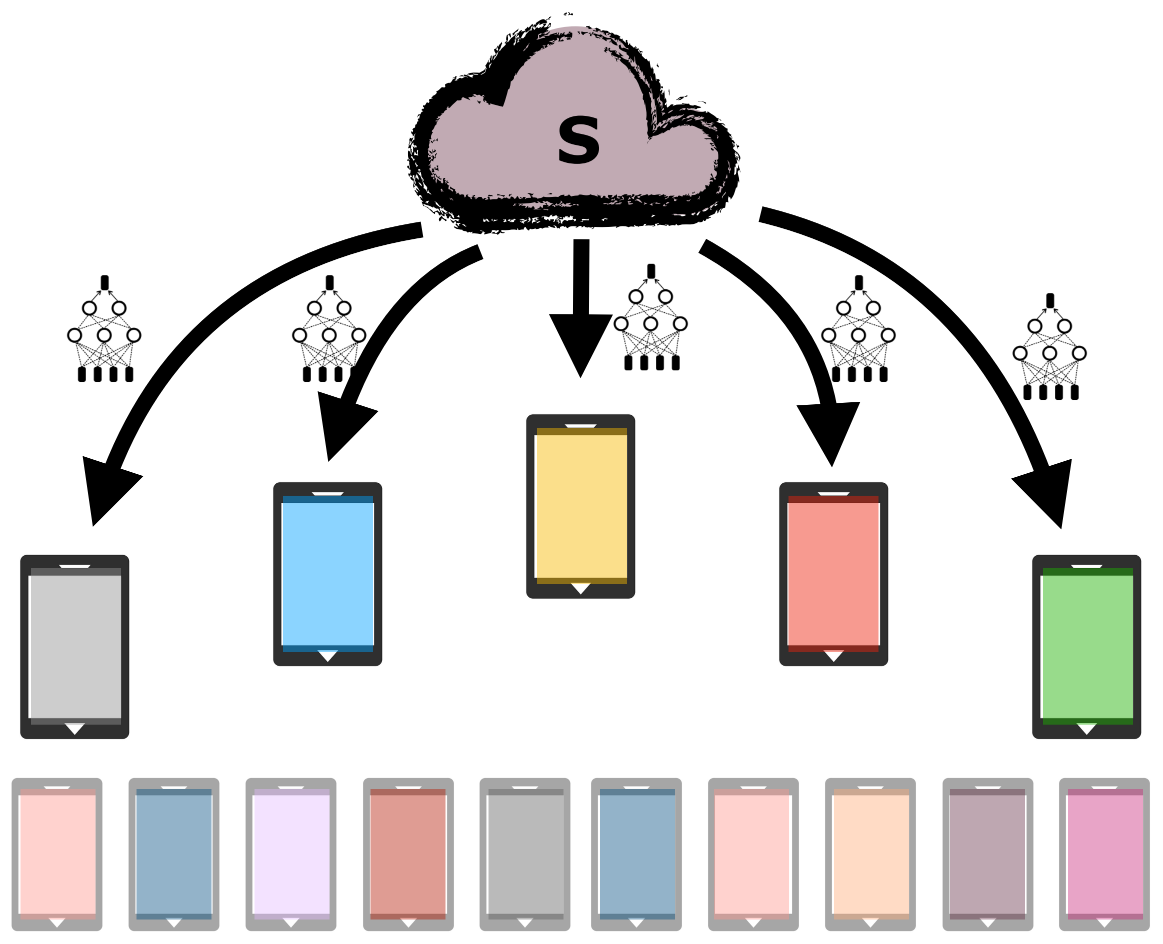
The superquantile is *nonlinear*
$\Longrightarrow$ unbiased stochastic gradients not possible



For i.i.d. copies $Z_1, \cdots, Z_m$ of Z, we have

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i\right] = \mathbb{E}[Z] \quad \text{but} \quad \mathbb{E}\Big[\mathbb{S}_\theta(Z_1, \cdots, Z_m)\Big] \neq \mathbb{S}_\theta(Z)$$

Nonlinear: We minimize a close surrogate



$$\overline{F}_\theta(w) = \mathbb{E}_{S\,:\,|S|=m}\left[\mathbb{S}_\theta\Big(\, \big(F_i(w)\ :\ i \in S\big)\,\Big)\right]$$

The surrogate is uniformly close for bounded losses:

For i.i.d. copies $Z_1, \cdots, Z_m$ of Z with $|Z| \le B$ a.s., we have

$$\left|\mathbb{E}\left[\mathbb{S}_\theta(Z_1, \cdots, Z_m)\right] - \mathbb{S}_\theta(Z)\right| \le \frac{B}{\sqrt{\theta m}} \qquad \mathrm{Var}\left[\mathbb{S}_\theta(Z_1, \cdots, Z_m)\right] \le \frac{B^2}{\theta m}$$

[Levy et al. (NeurIPS 2020)]

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose each F_i is L-smooth and G-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\|\nabla \Phi_\theta^{2L}(w_t)\right\|^2 \le \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3}\left(\frac{\Delta_0 L G}{t}\right)^{2/3} + \frac{\Delta_0 L}{t}$$

t: #comm. rounds
τ: #local update steps
Δ_0: initial error

$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2}\|z-w\|^2 \right\} \quad \longleftarrow$$

Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose **each F_i is L-smooth and G-Lipschitz**.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\|\nabla \Phi_\theta^{2L}(w_t)\right\|^2 \leq \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3}\left(\frac{\Delta_0 L G}{t}\right)^{2/3} + \frac{\Delta_0 L}{t}$$

t: #comm. rounds
τ: #local update steps
Δ_0: initial error

$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2}\|z - w\|^2 \right\} \quad \longleftarrow$$

Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose each F_i is L-smooth and G-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\| \nabla \Phi_\theta^{2L}(w_t) \right\|^2 \le \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3} \left(\frac{\Delta_0 L G}{t} \right)^{2/3} + \frac{\Delta_0 L}{t}$$

t: #comm. rounds
τ: #local update steps
Δ_0: initial error

$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2} \|z - w\|^2 \right\} \longleftarrow$$

Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose each F_i is L-smooth and G-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\|\nabla \Phi_\theta^{2L}(w_t)\right\|^2 \leq \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3}\left(\frac{\Delta_0 L G}{t}\right)^{2/3} + \frac{\Delta_0 L}{t}$$

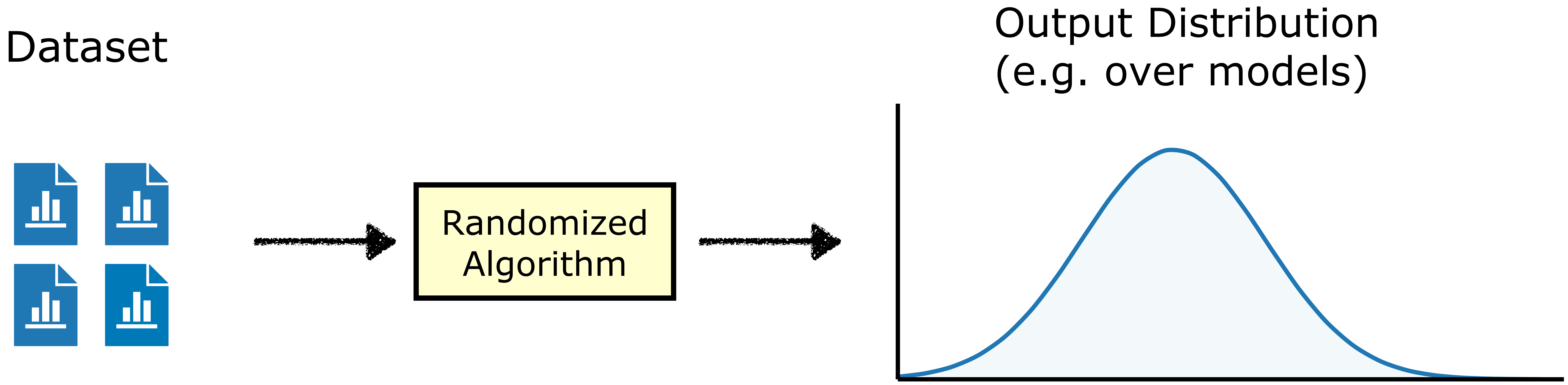
t: #comm. rounds
τ: #local update steps
Δ_0: initial error

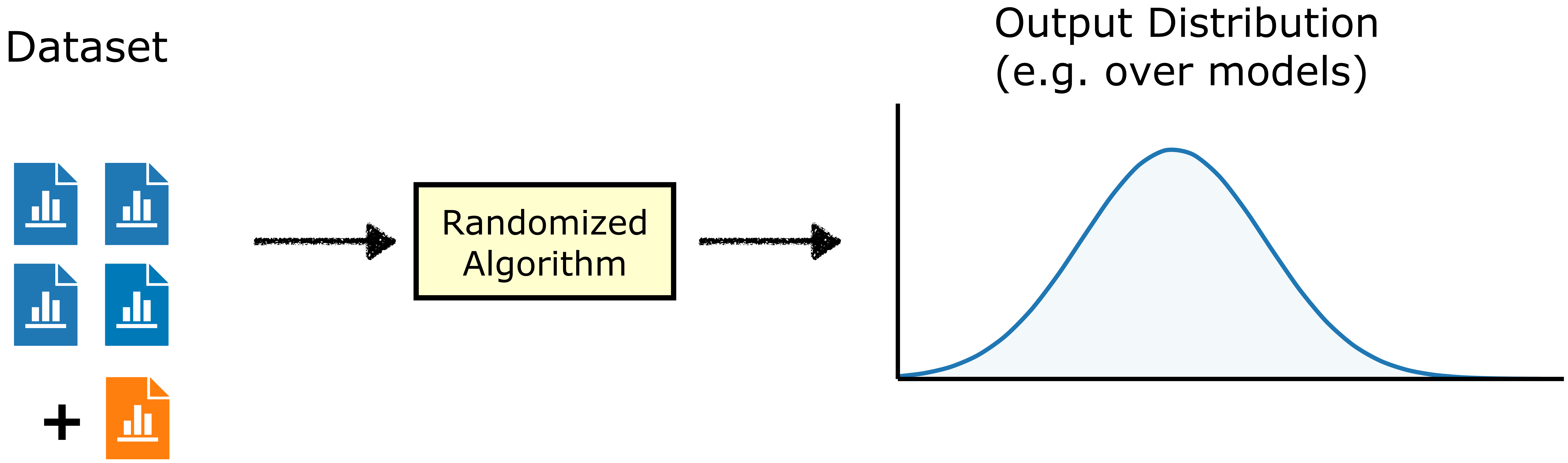
$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2}\|z-w\|^2 \right\} \quad \longleftarrow$$

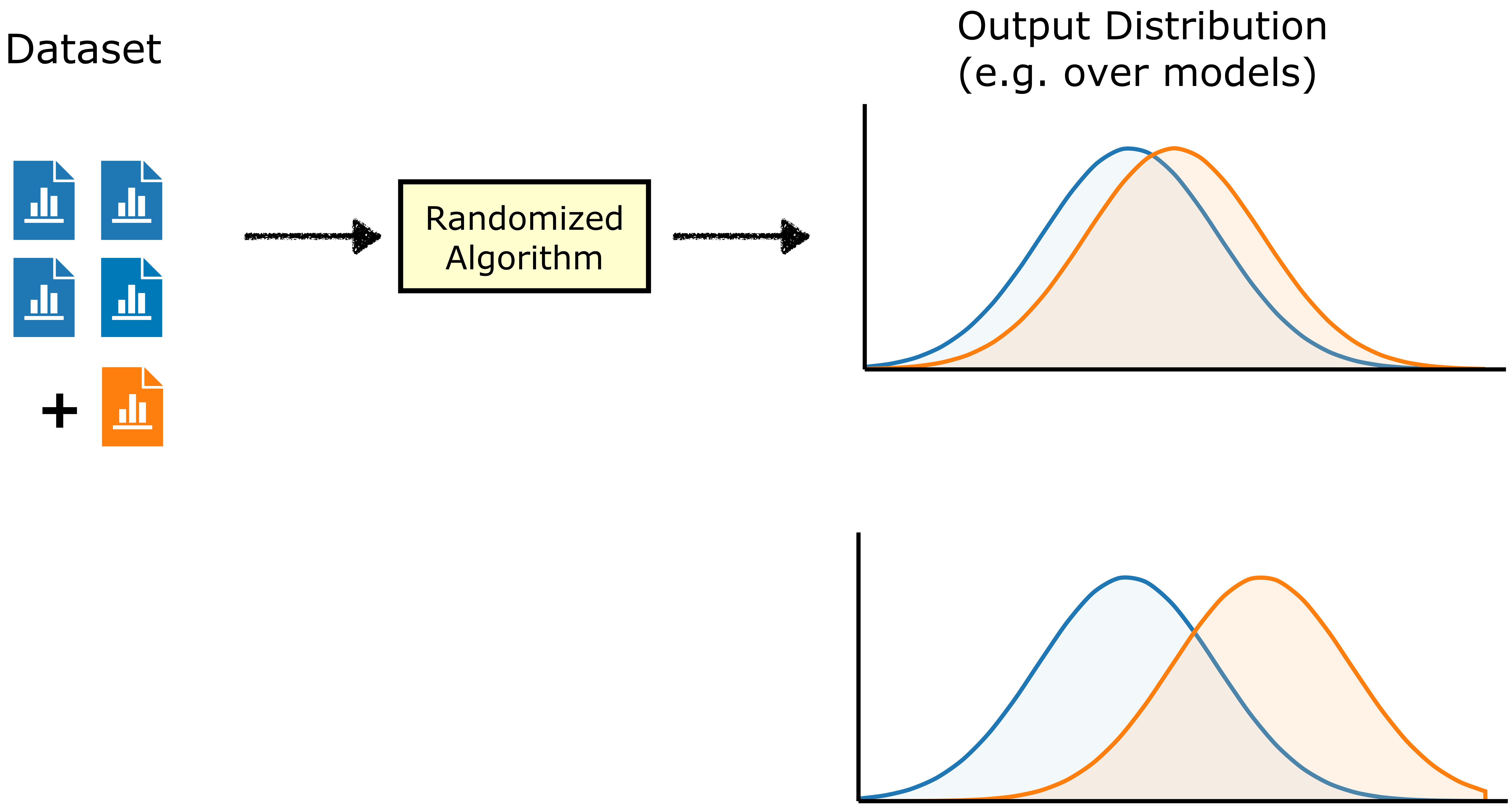
Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

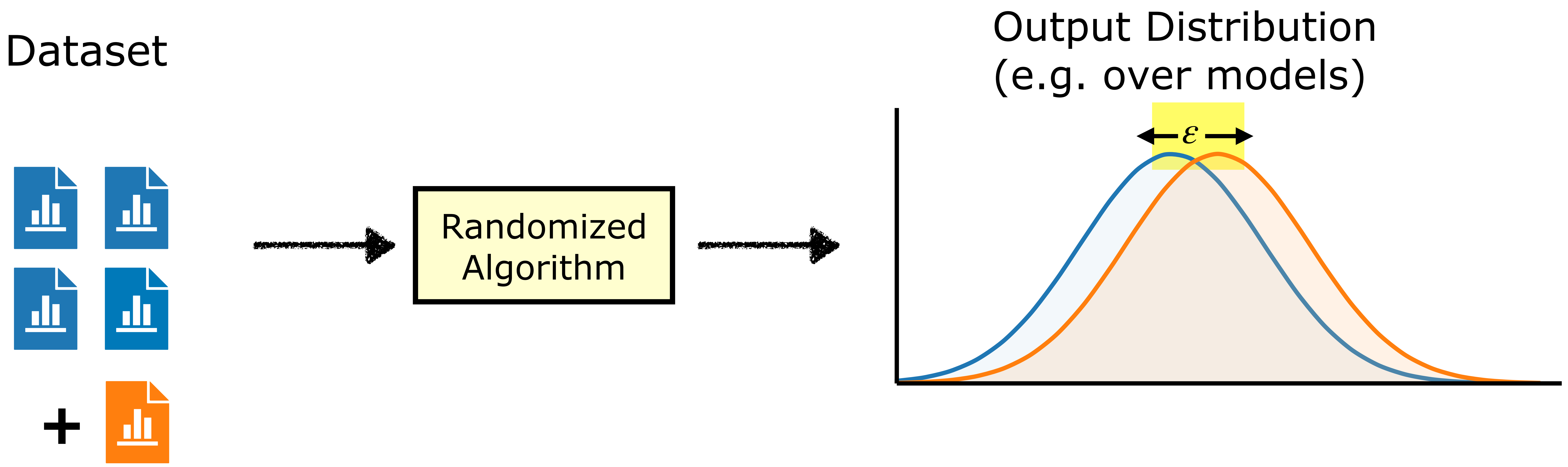
Privacy analysis

Privacy of user data


Dataset
Randomized
Algorithm
Output Distribution
(e.g. over models)

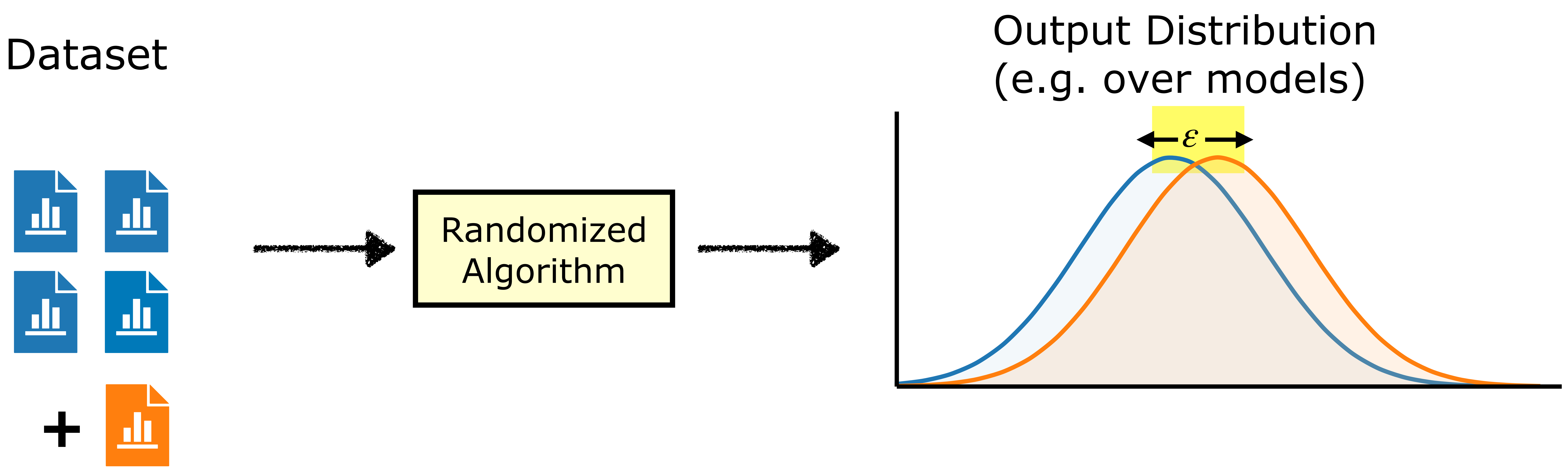

Dataset
+
Randomized
Algorithm
Output Distribution
(e.g. over models)

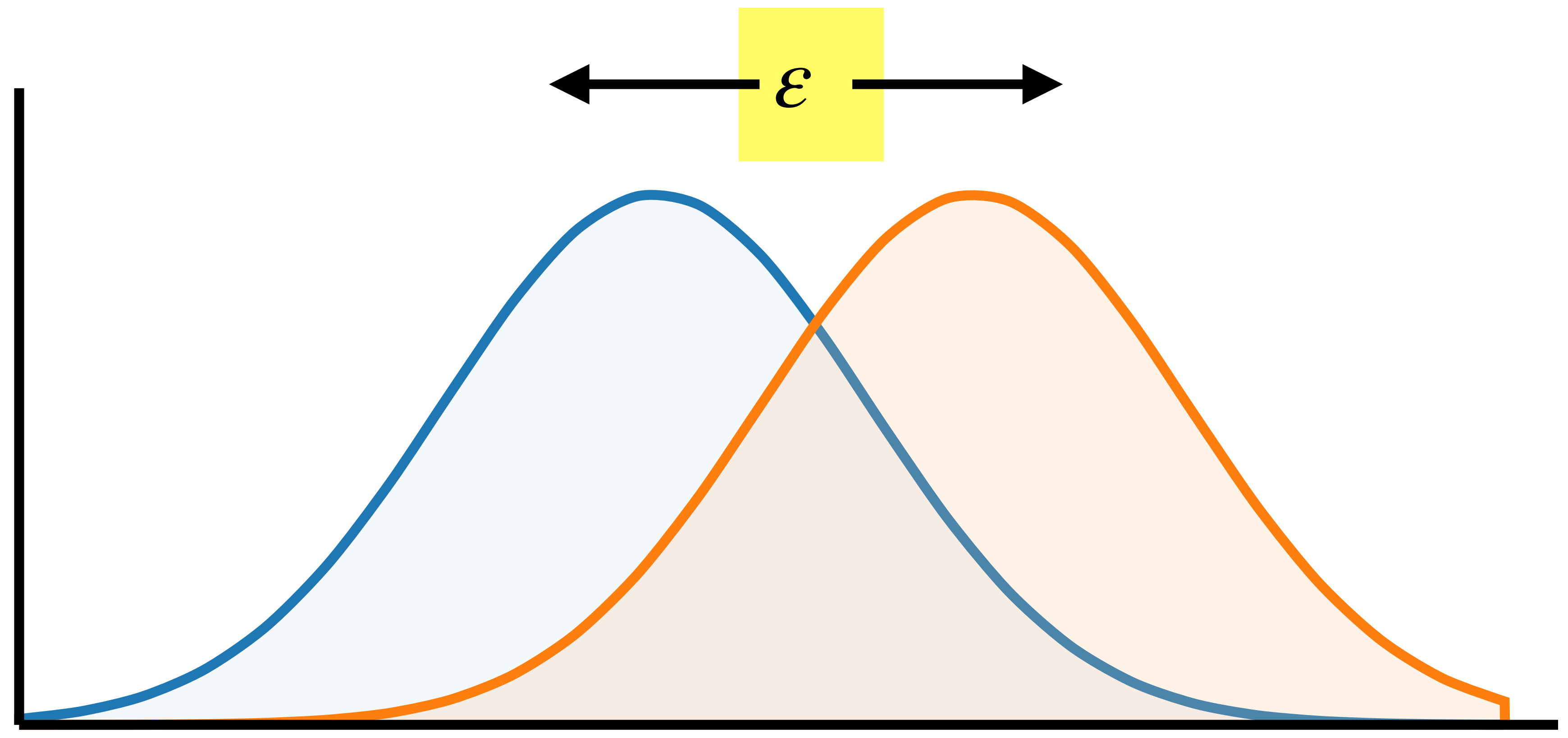

Dataset
Randomized Algorithm
Output Distribution
(e.g. over models)



A randomized algorithm is ε-**differentially private** if the addition of **one user's data** does not alter its output distribution by more than ε

[Dwork, McSherry, Nissim, Smith (2006)]





ε-differential privacy

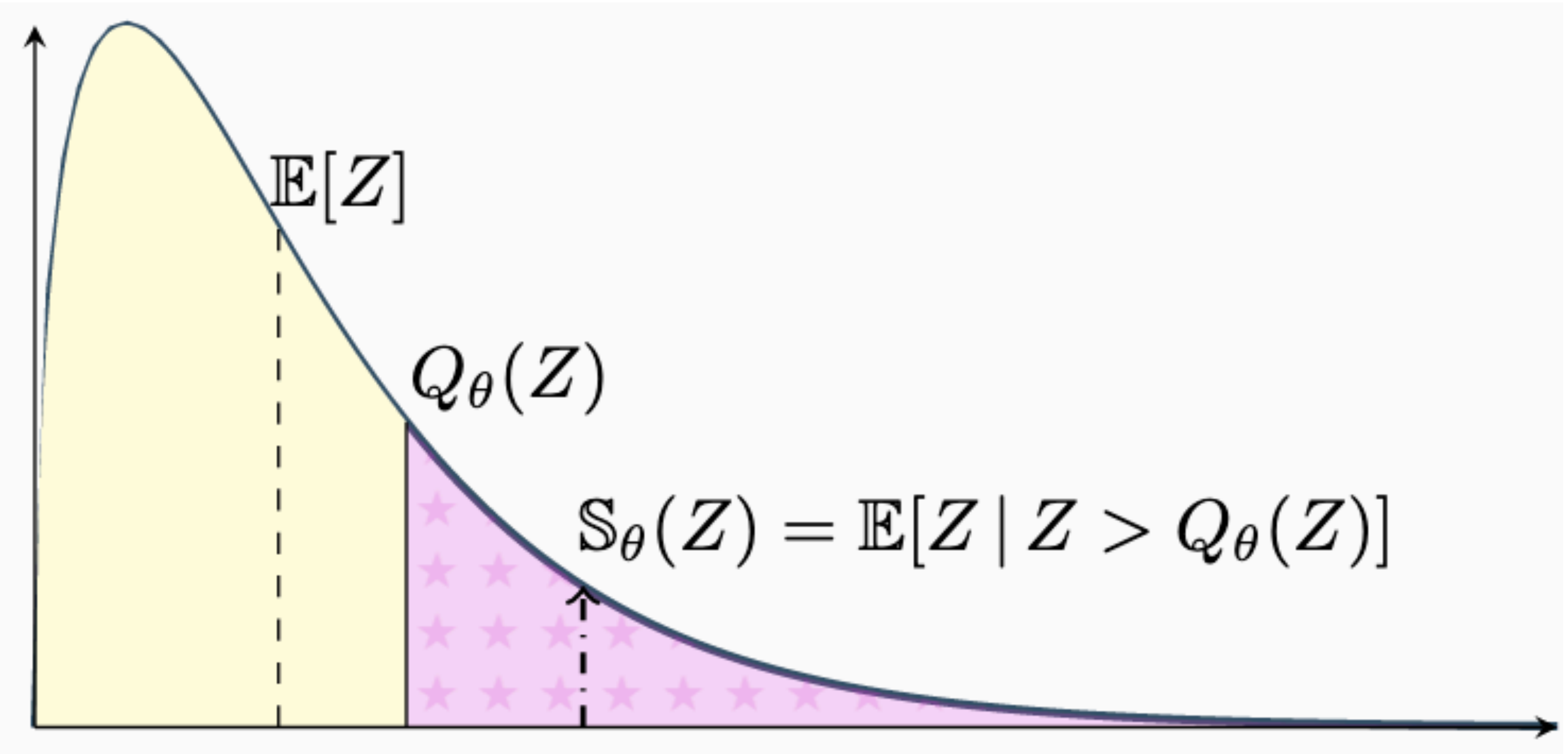
Large ε $\implies$ more privacy leakage

Privacy goal

Extend our algorithm to make it **differentially private**

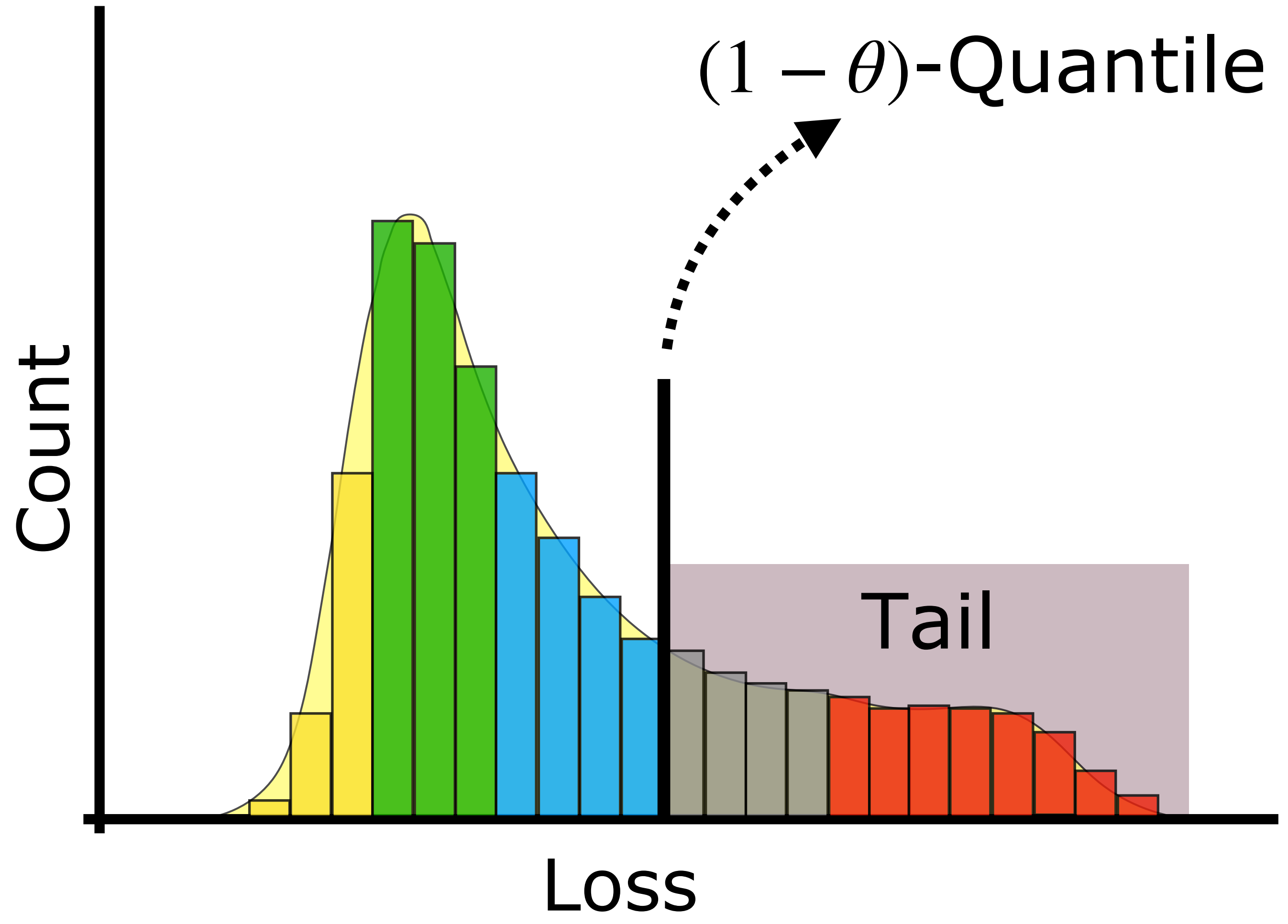
Our Objective:

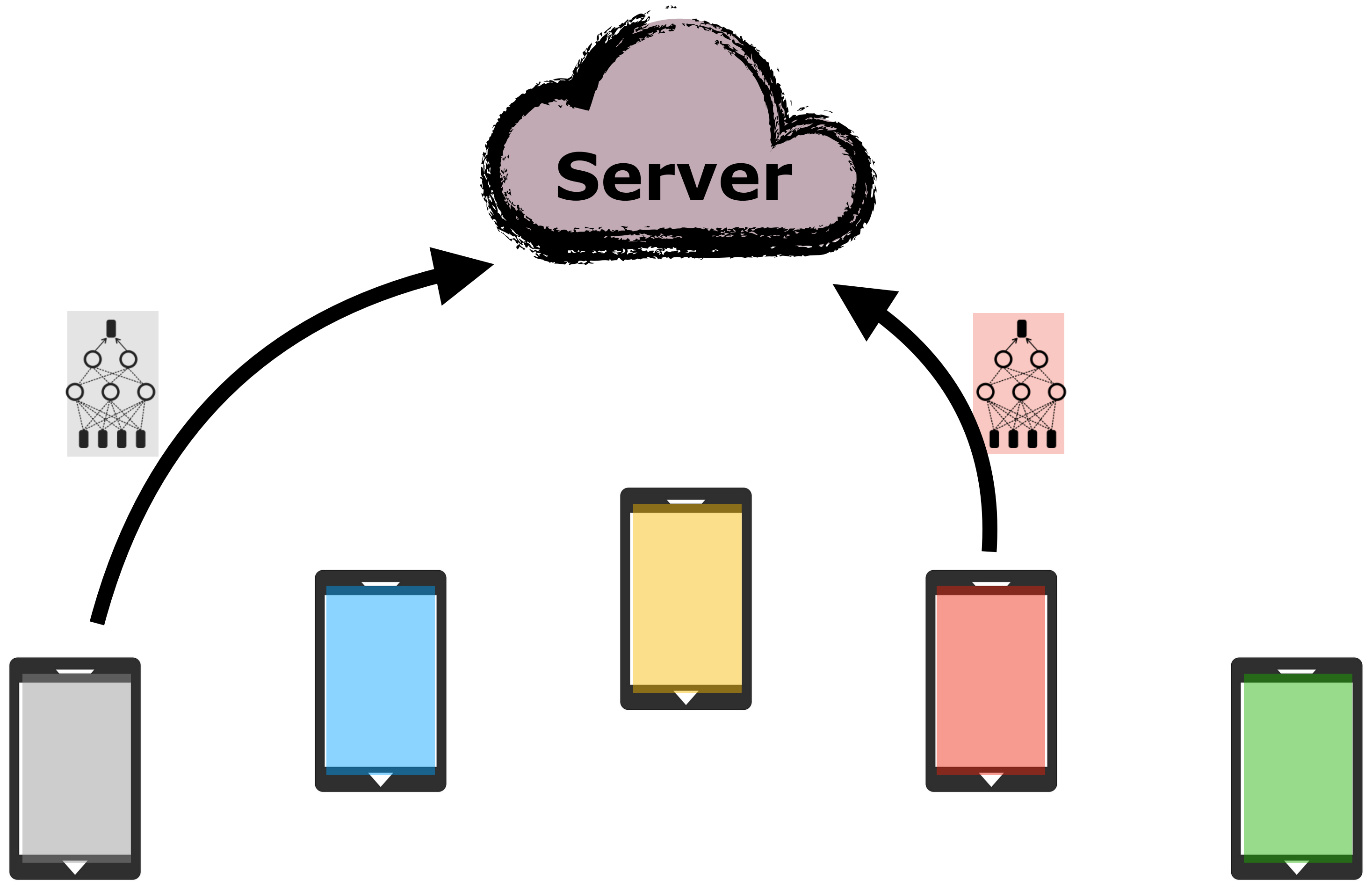
$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$



Our Algorithm:

*Step 3 of 3: Aggregate updates contributed by **tail clients** only*

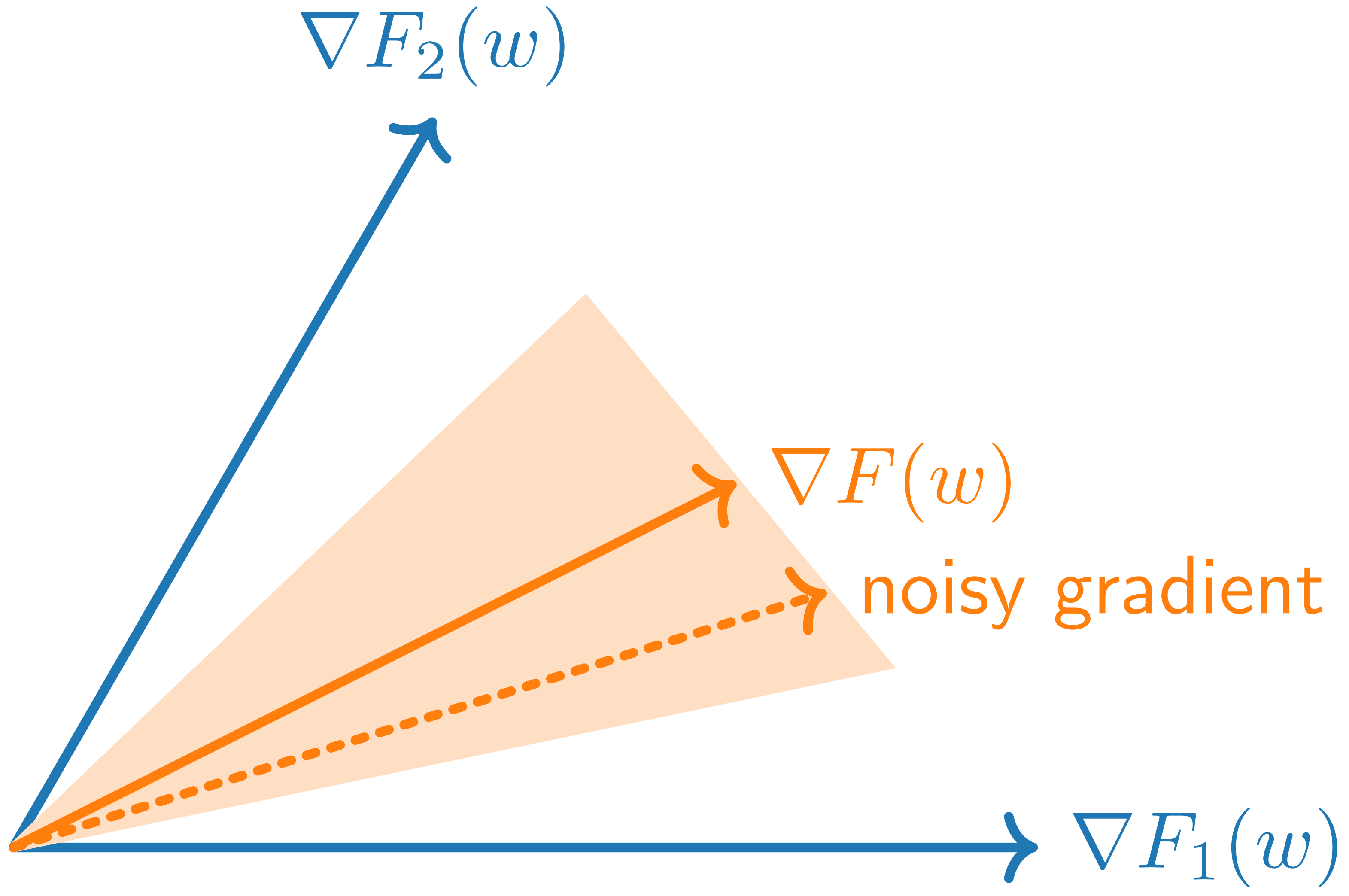




Why is it challenging?

Usual Algorithm:

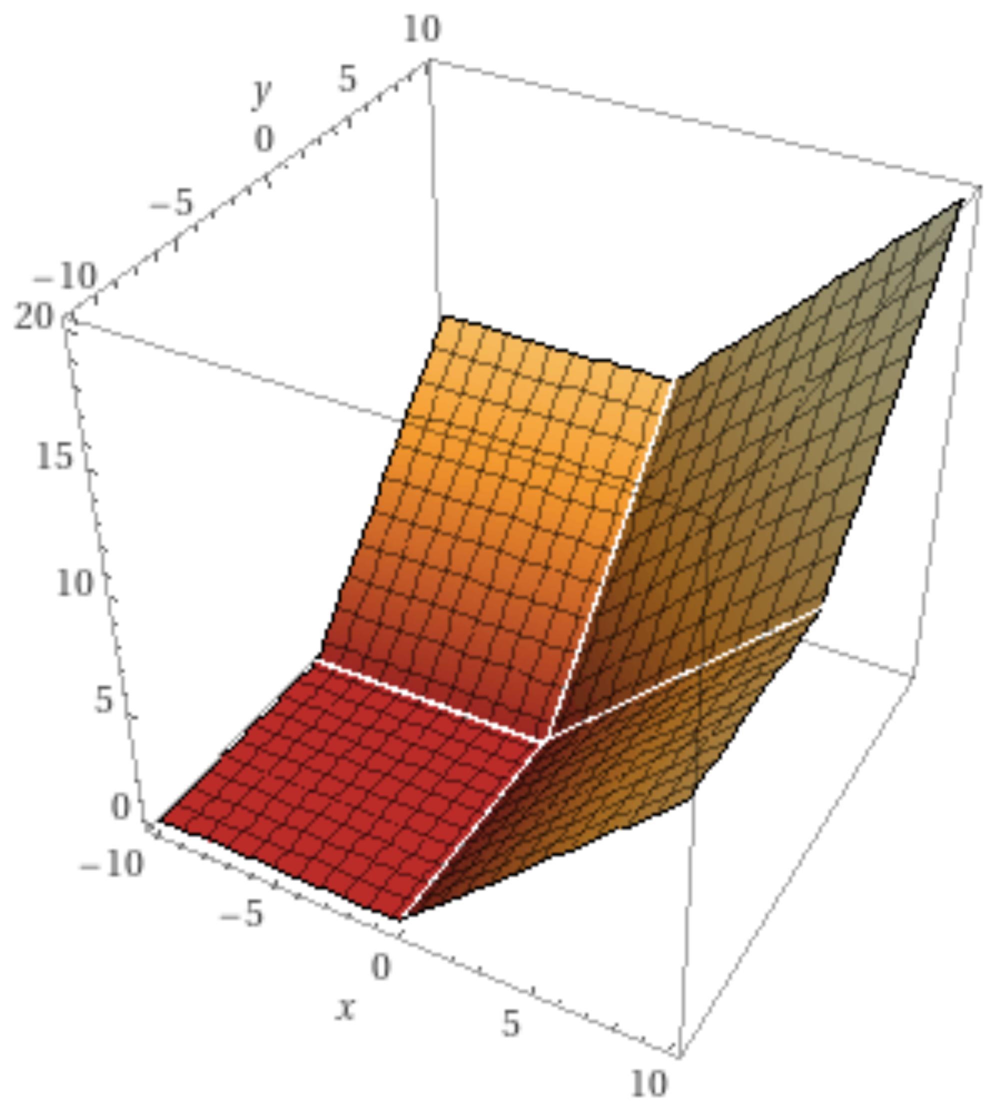
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$



Private mean estimation of gradients

Our Algorithm:

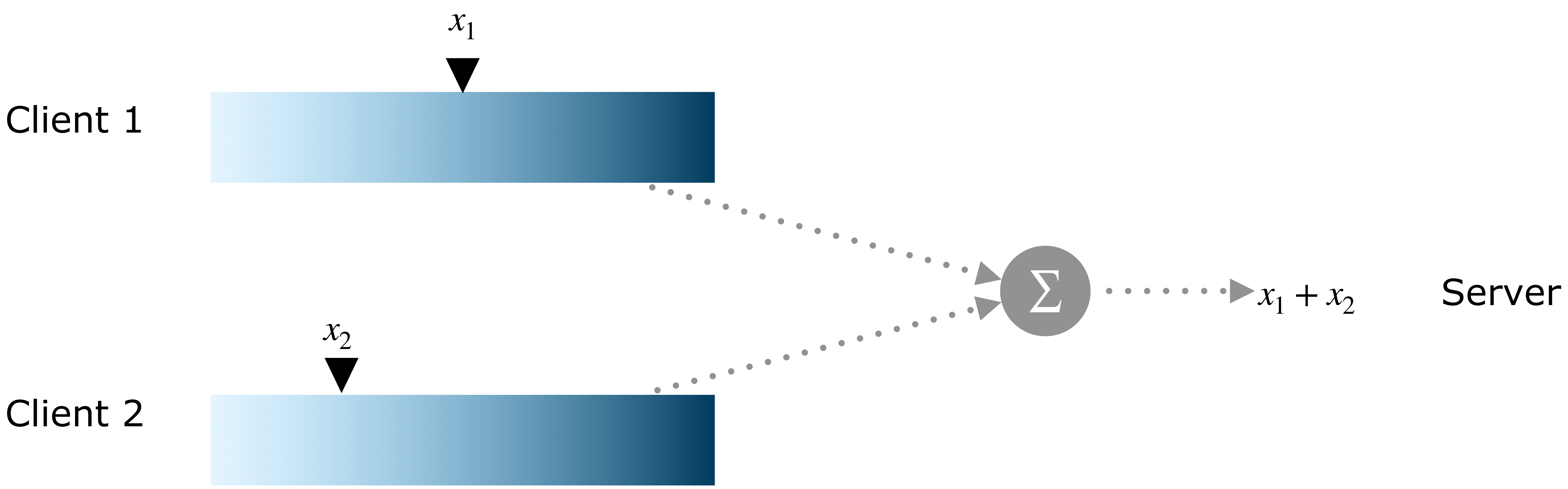
$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$



Non-linear

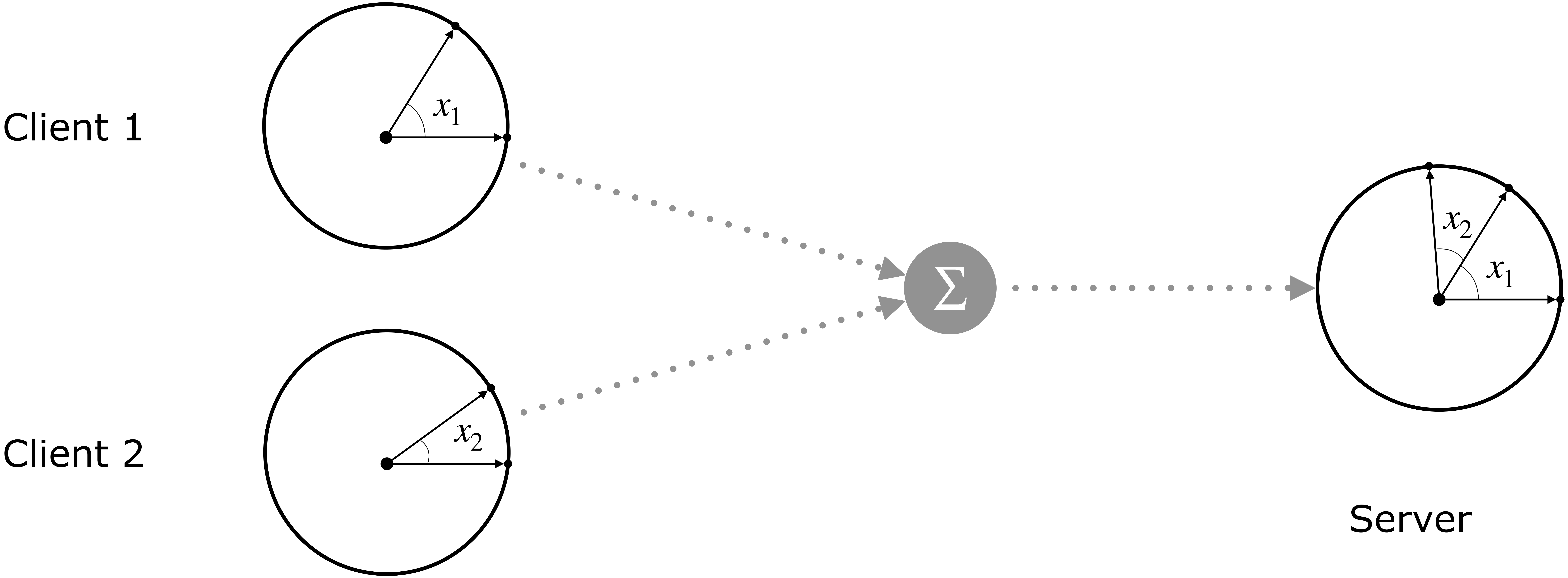
Communication primitive: secure sum

Only reveal $x_1 + x_2$ to the server without revealing x_1 or x_2



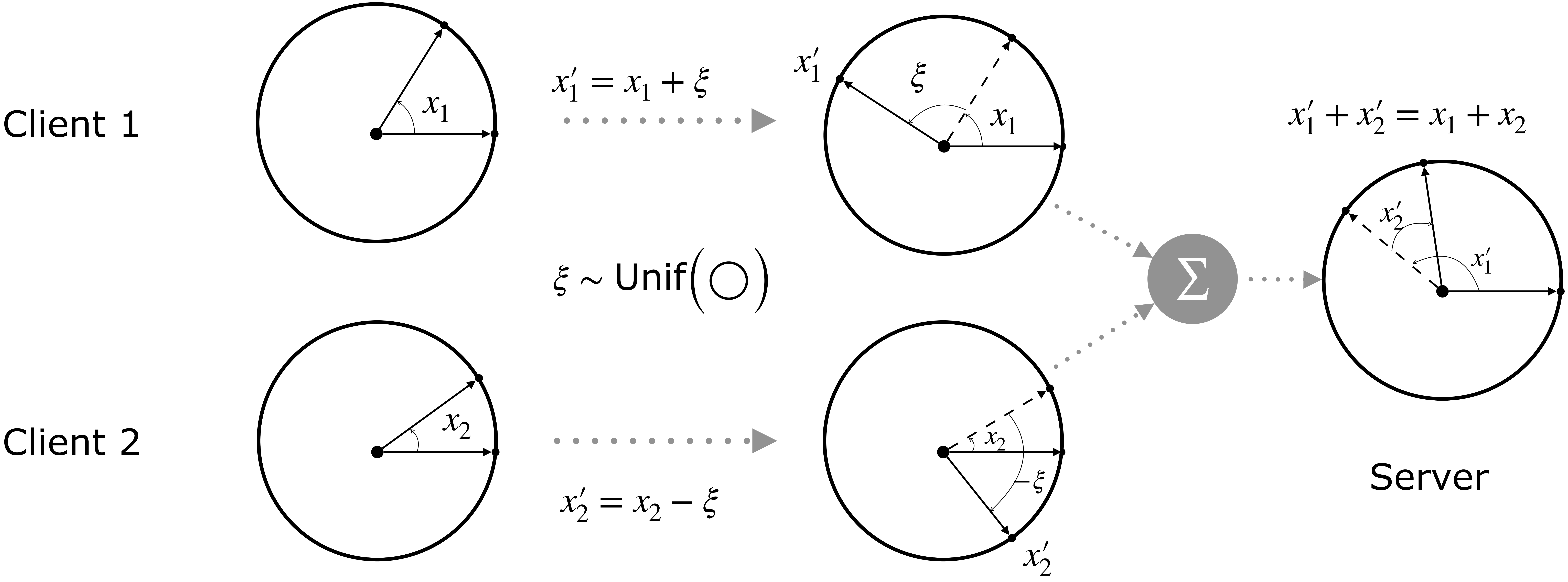
[Bonawitz et al. (CCS 2017), Bell et al. (CCS 2020)]

Perform all operations modulo M



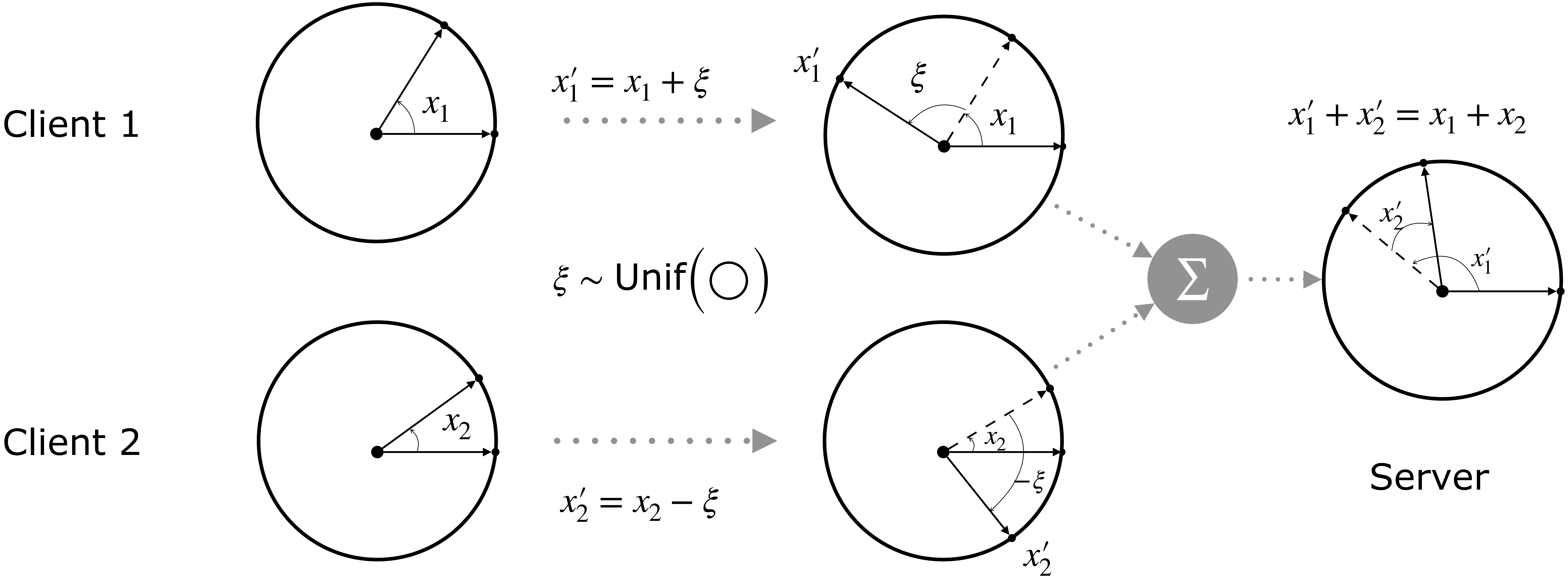
[Bonawitz et al. (CCS 2017), Bell et al. (CCS 2020)]

Server only sees $x_1', x_2' \sim \mathrm{Unif}(\bigcirc)$ but calculates the correct sum



[Bonawitz et al. (CCS 2017), Bell et al. (CCS 2020)]

Server only sees $x_1', x_2' \sim \mathrm{Unif}(\bigcirc)$ but calculates the correct sum



Total communication for m vectors in $\mathbb{R}^d$ $= O(m \log m + md)$ numbers

Server only sees $x_1', x_2' \sim \mathrm{Unif}(\bigcirc)$ but calculates the correct sum

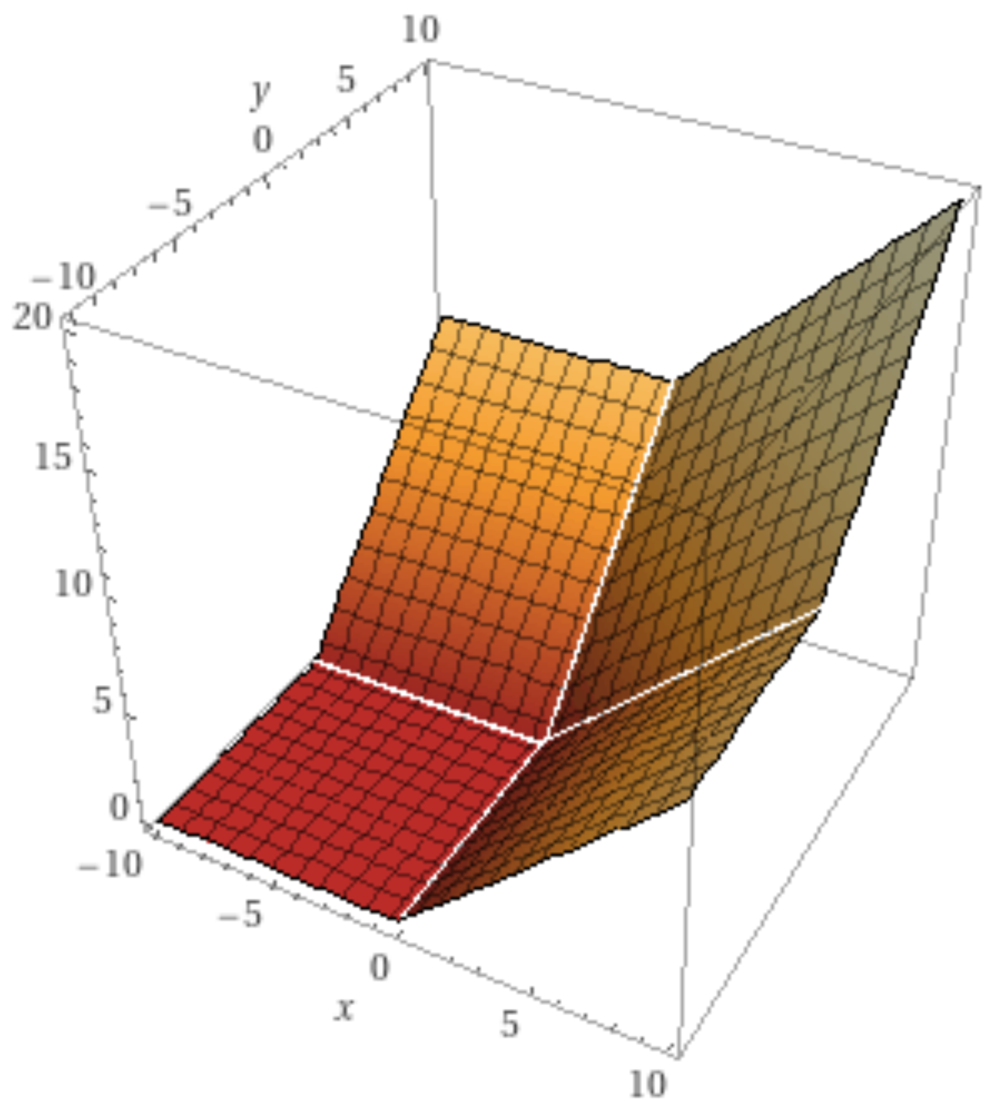
Real-world communication constraint:
All client-to-server communication must go through secure summation

Total communication for m vectors in $\mathbb{R}^d$ $= O(m \log m + md)$ numbers

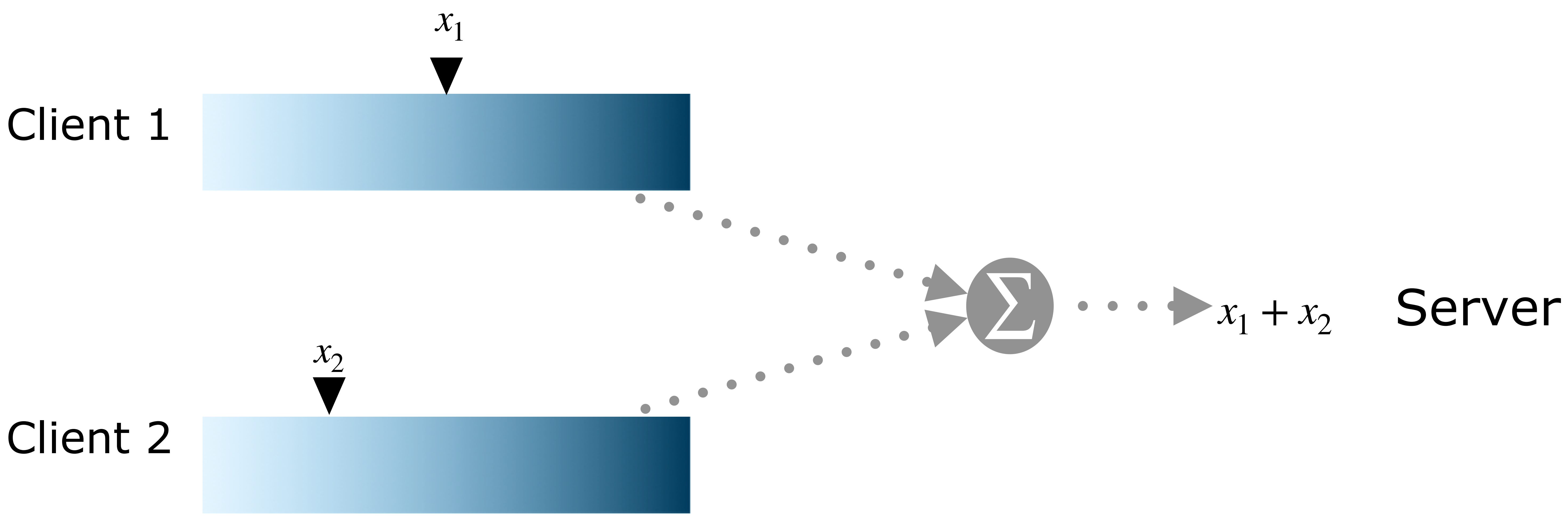
How to achieve non-linear aggregation with a secure sum?

Non-linear aggregate:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$



Secure sum



Usual Algorithm (FedAvg):

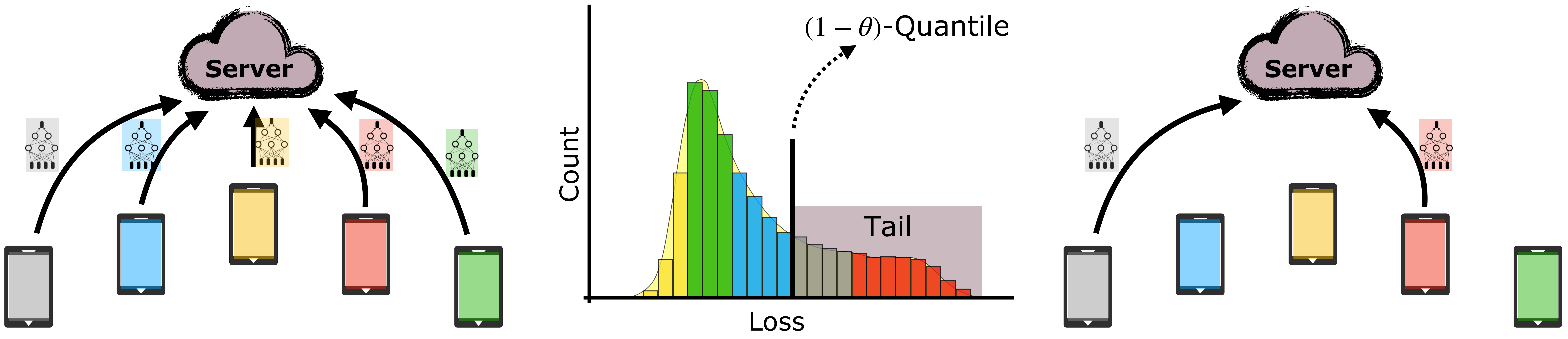
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

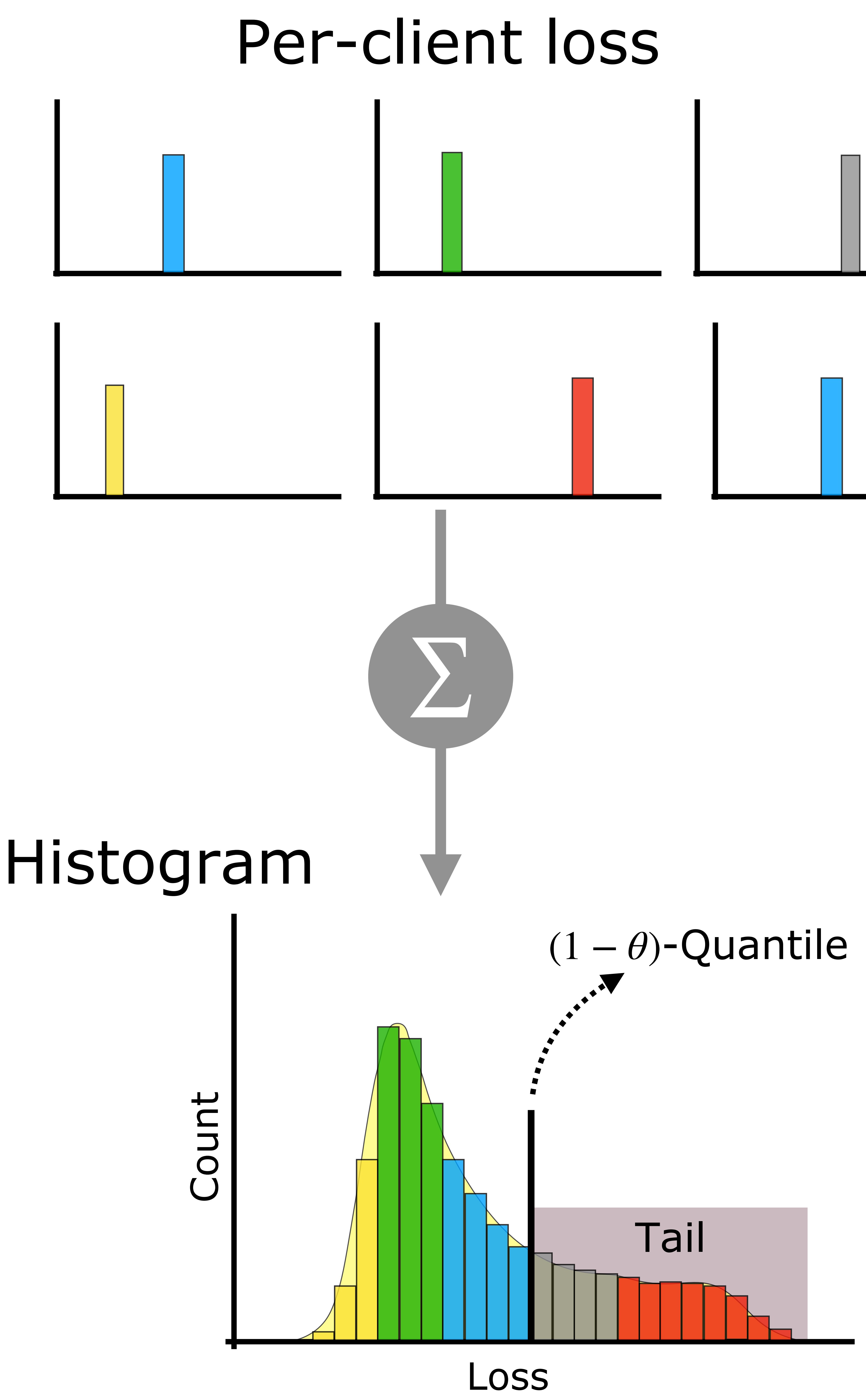
Step 3 of 3: Aggregate updates contributed by ***all clients***

Our Algorithm:

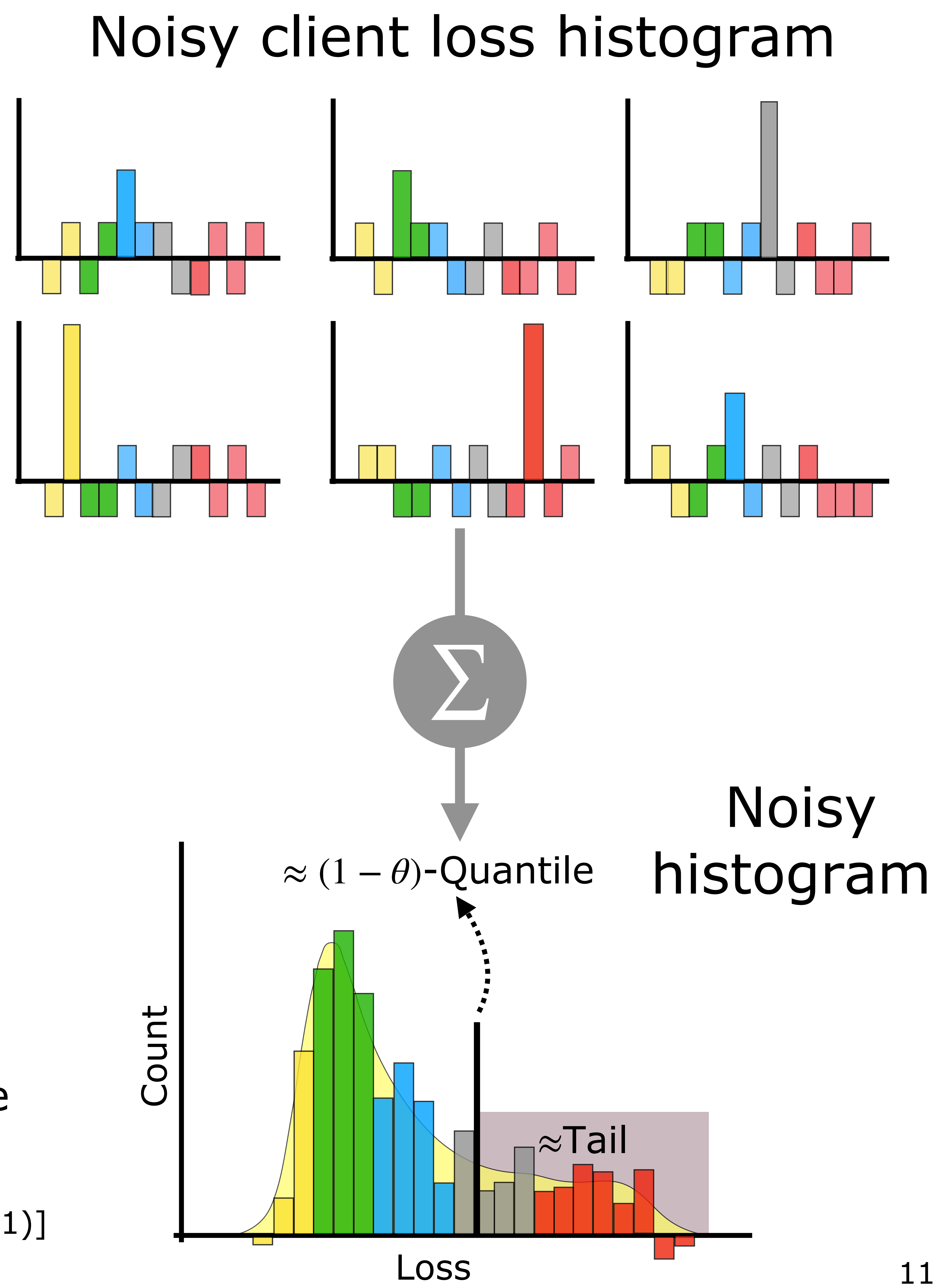
$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Step 3 of 3: Aggregate updates contributed by ***tail clients*** *only*





$$h_i' = h_i + \mathcal{N}_{\mathbb{Z}}(0, \sigma^2 I_b)$$



Differential privacy via discrete Gaussian noise

[Kairouz, Liu, Steinke. (ICML 2021)]

Proposition (informal) [*P.*, Laguel, Malick, Harchaoui]

If we wish to compute the α-quantile, our algorithm returns an ε-differentially private α'-quantile where

$$\left|\alpha' - \alpha\right| \leq \frac{\sqrt{b}}{\varepsilon m}$$

Total communication cost $\approx bm \log^2 m$

m #clients per round

b #bins in the histogram

ε privacy parameter

Usual Algorithm (FedAvg):

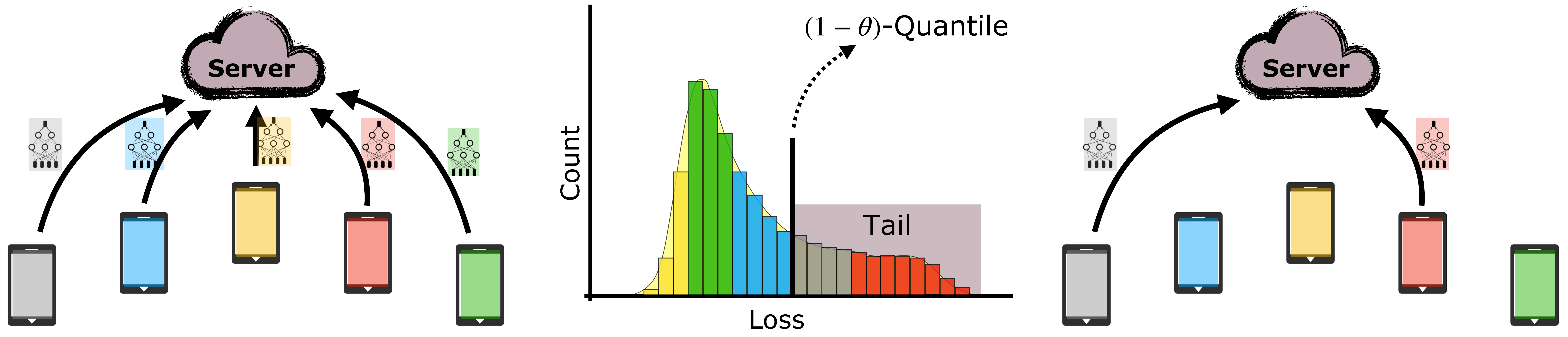
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

*Step 3 of 3: Aggregate updates contributed by **all clients***

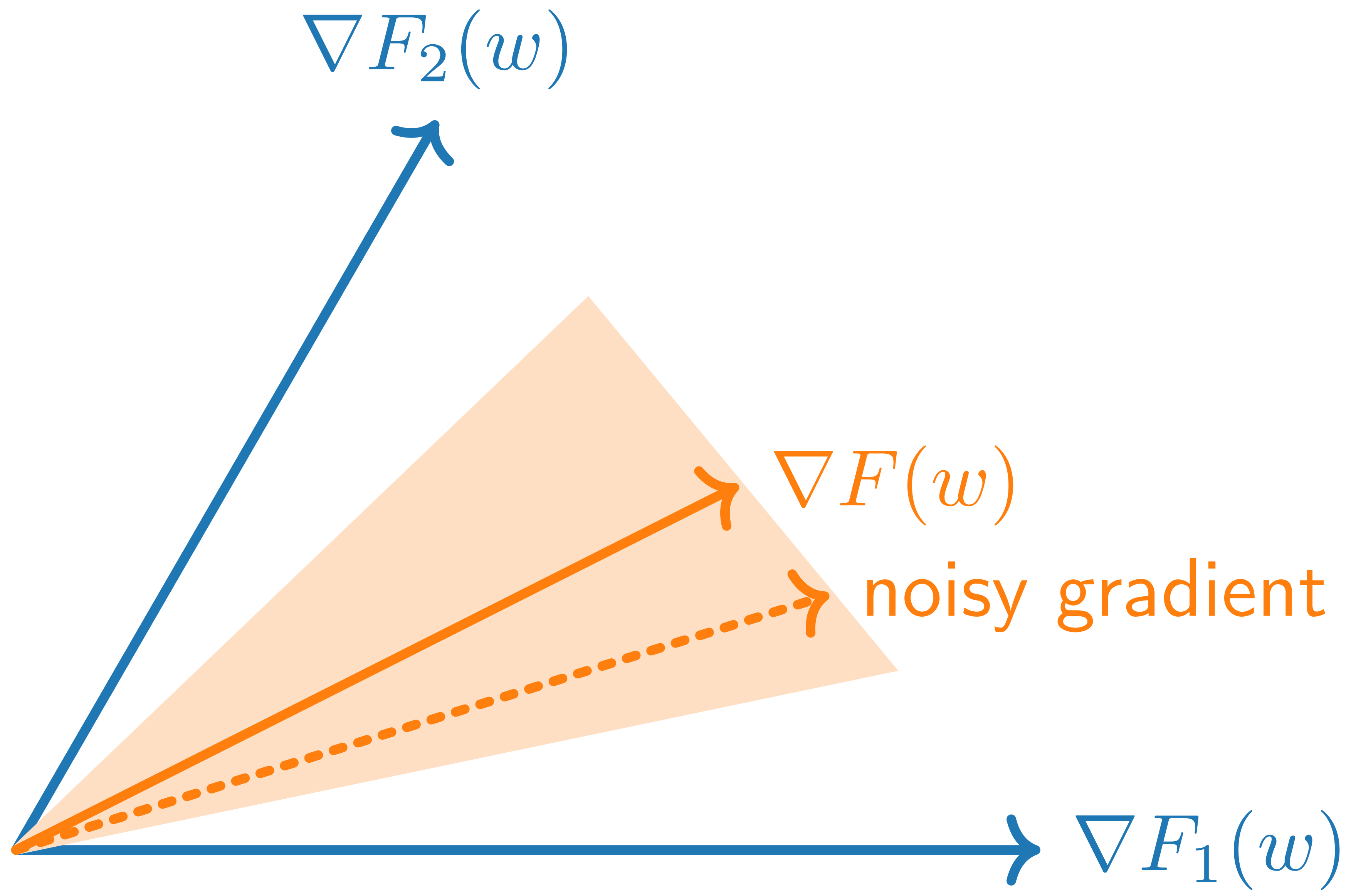
Our Algorithm:

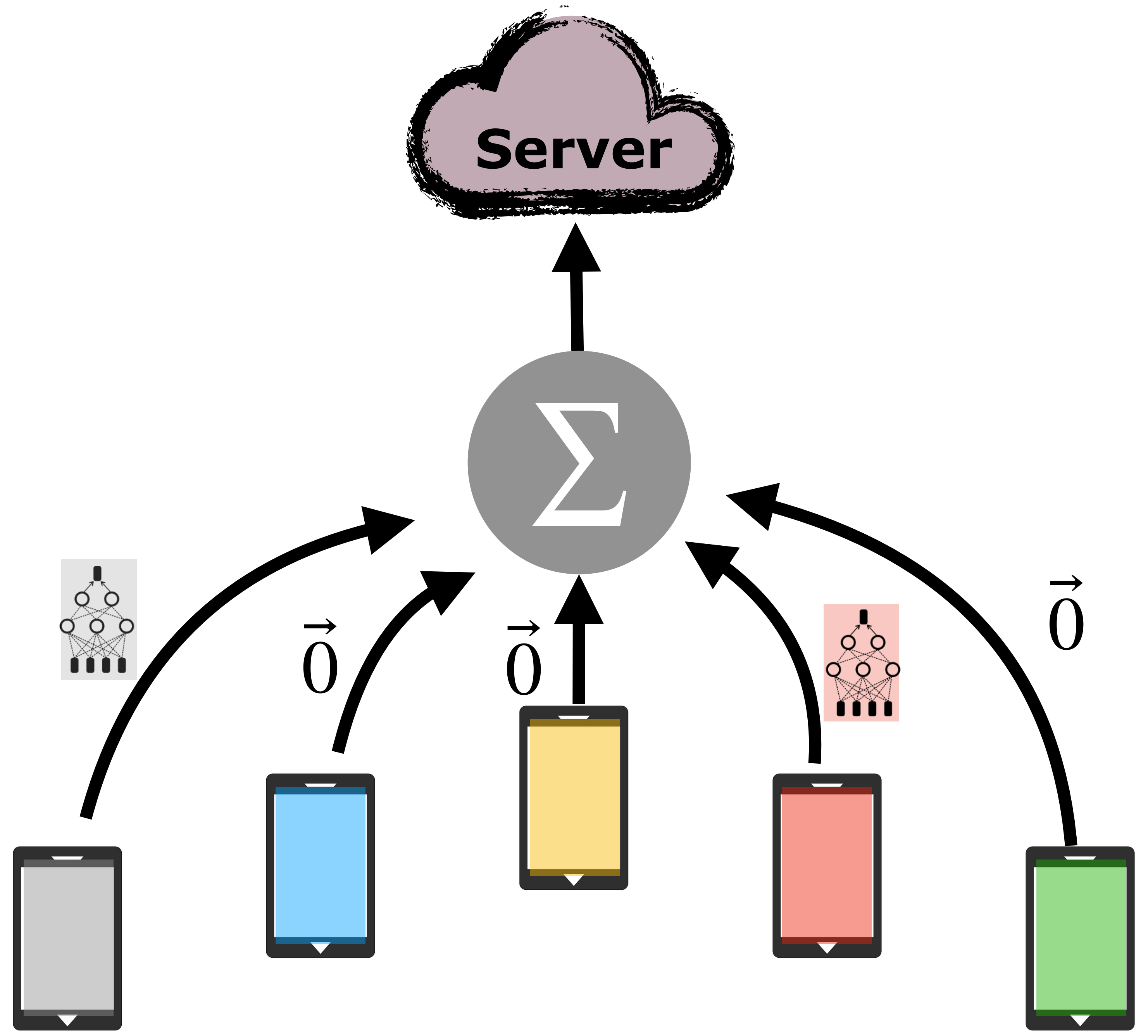
$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

*Step 3 of 3: Aggregate updates contributed by **tail clients** only*



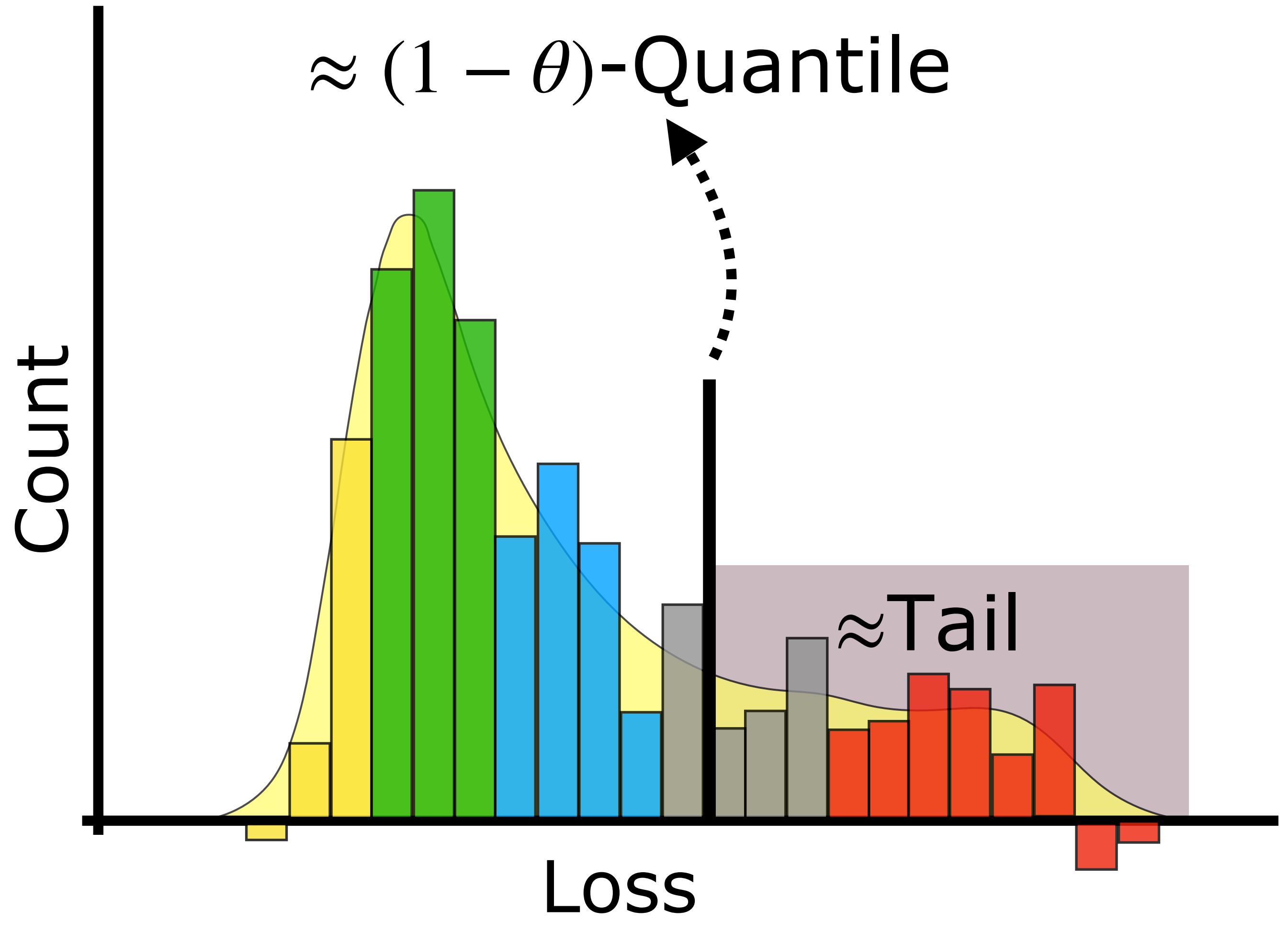
Private mean estimation of (potentially zeroed out) gradients

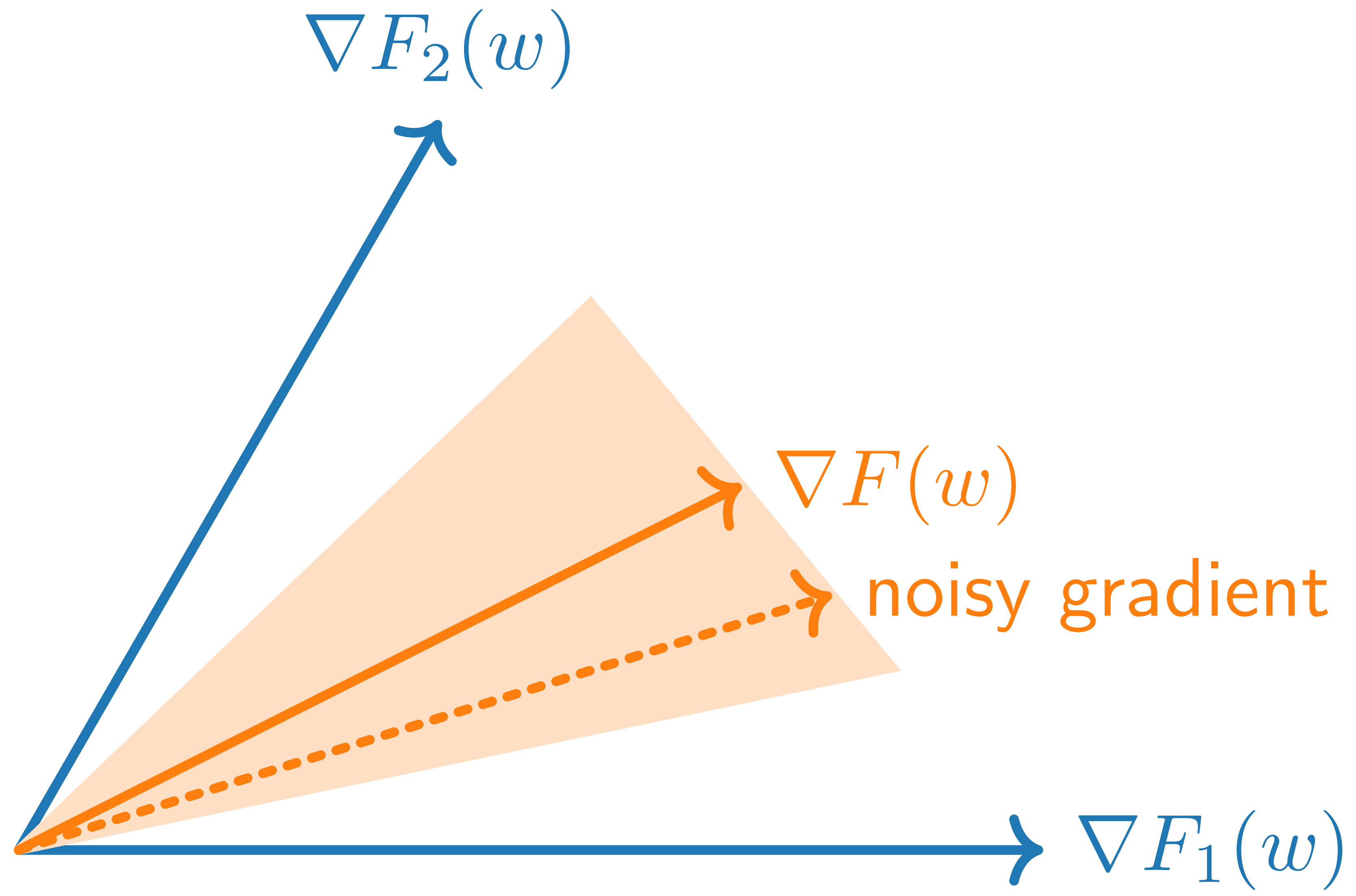




Total privacy leakage =

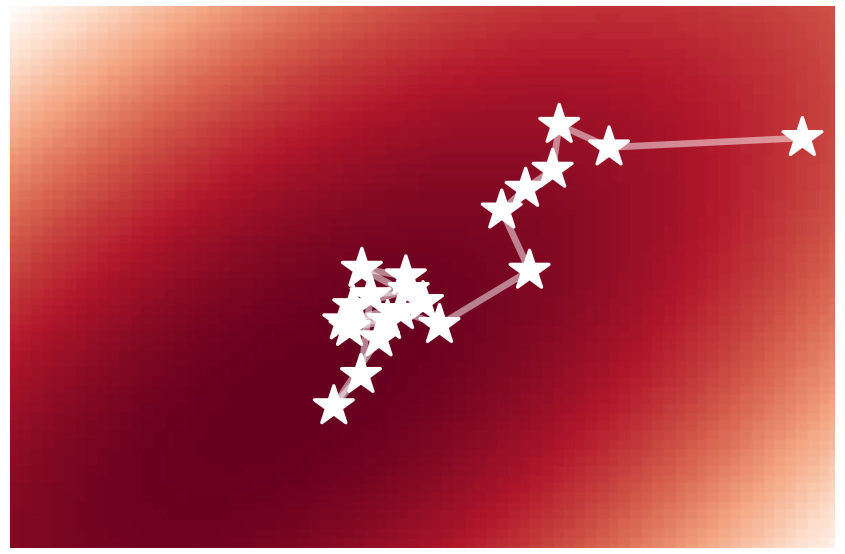
Quantile privacy leakage + Parameter privacy leakage



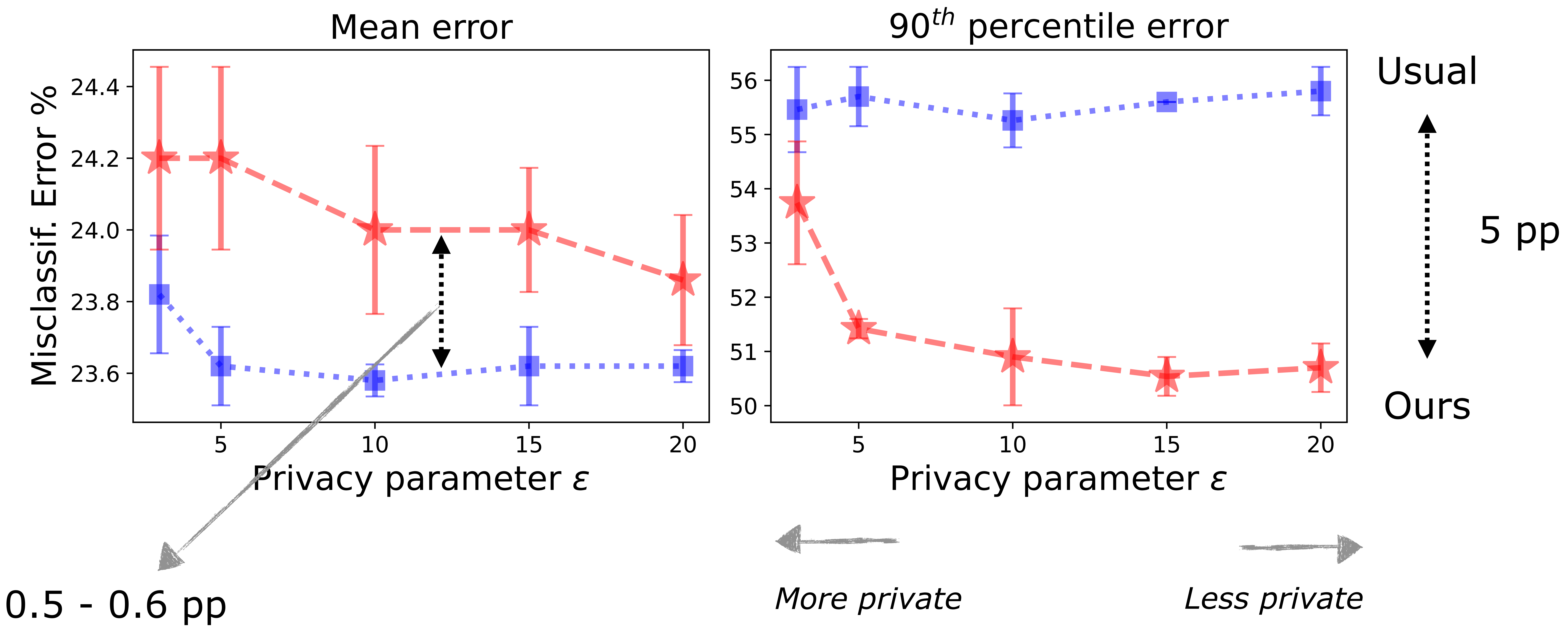


Privacy of user data:

First end-to-end **differentially private** algorithm for distributionally robust federated learning



Algorithm requires **2** secure summations per update


Mean error
90th percentile error
Misclassif. Error %
24.4
24.2
24.0
23.8
23.6
56
55
54
53
52
51
50
5
10
15
20
Privacy parameter ε
Usual
5 pp
Ours
0.5 - 0.6 pp
More private
Less private

Distributionally robust learning with 1 additional line of code

```python
import torch.nn.functional as F
from sqwash import reduce_superquantile

for x, y in dataloader:
    y_hat = model(x)
    batch_losses = F.cross_entropy(y_hat, y, reduction='none')  # must set `reduction='none'`
    loss = reduce_superquantile(batch_losses, superquantile_tail_fraction=0.5)  # Additional line
    loss.backward()  # Proceed as usual from here
```

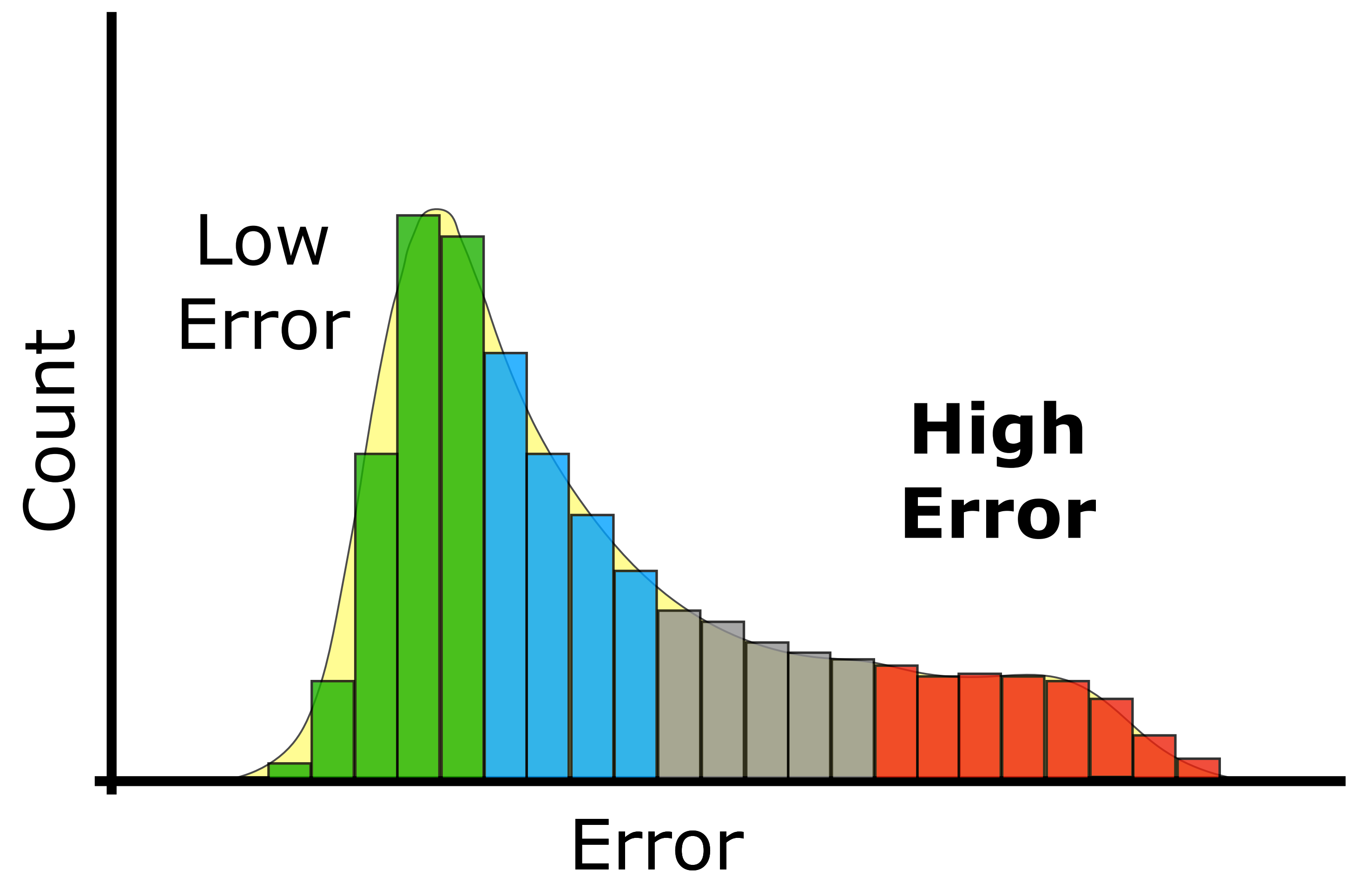
Install: **pip install sqwash**

Documentation: krishnap25.github.io/sqwash/

SCAN ME

Summary: Tackling distribution shifts in federated learning

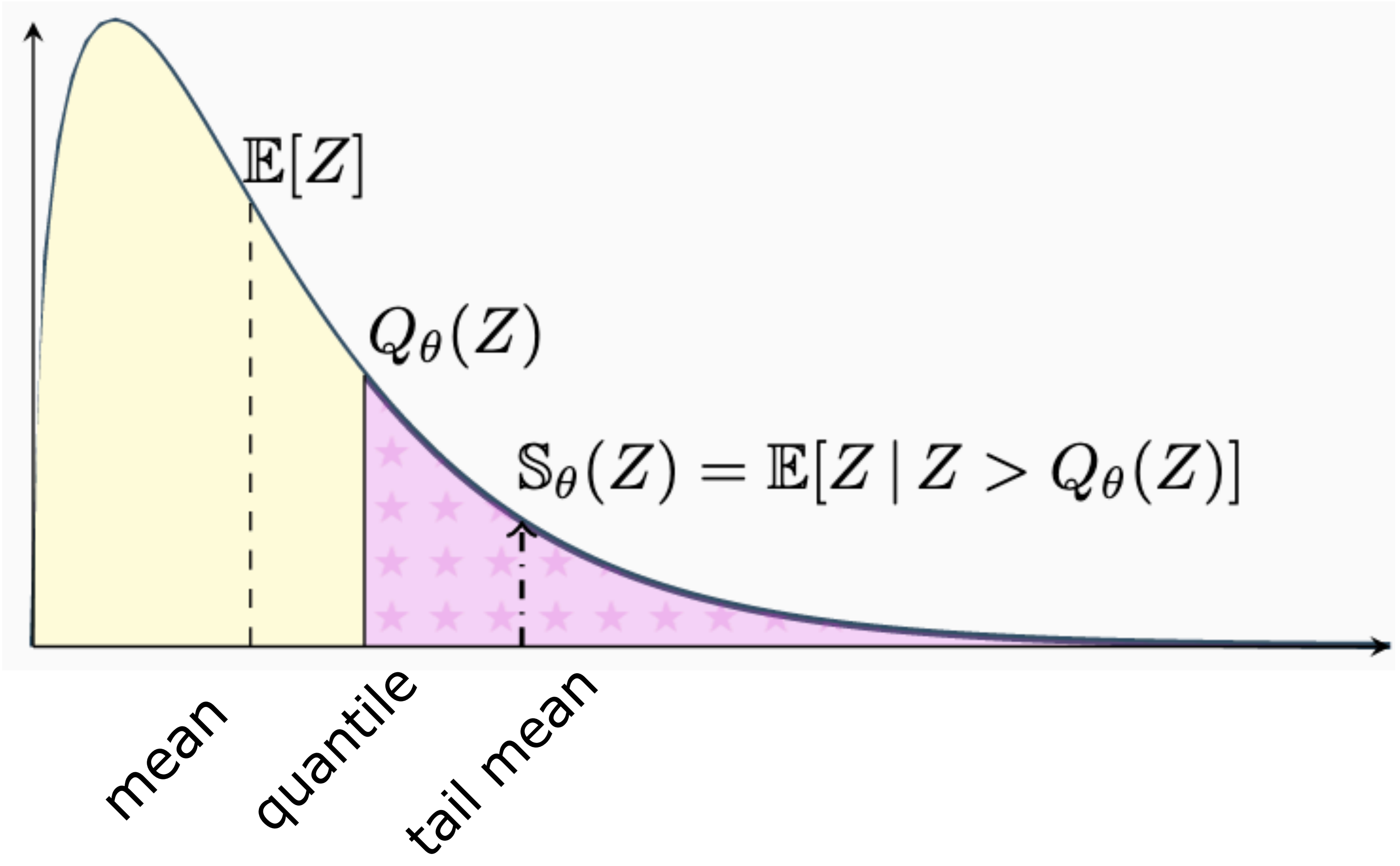
Distribution shift $\Longrightarrow$ large tail errors



Summary: Tackling distribution shifts in federated learning

Distribution shift $\Longrightarrow$ large tail errors

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

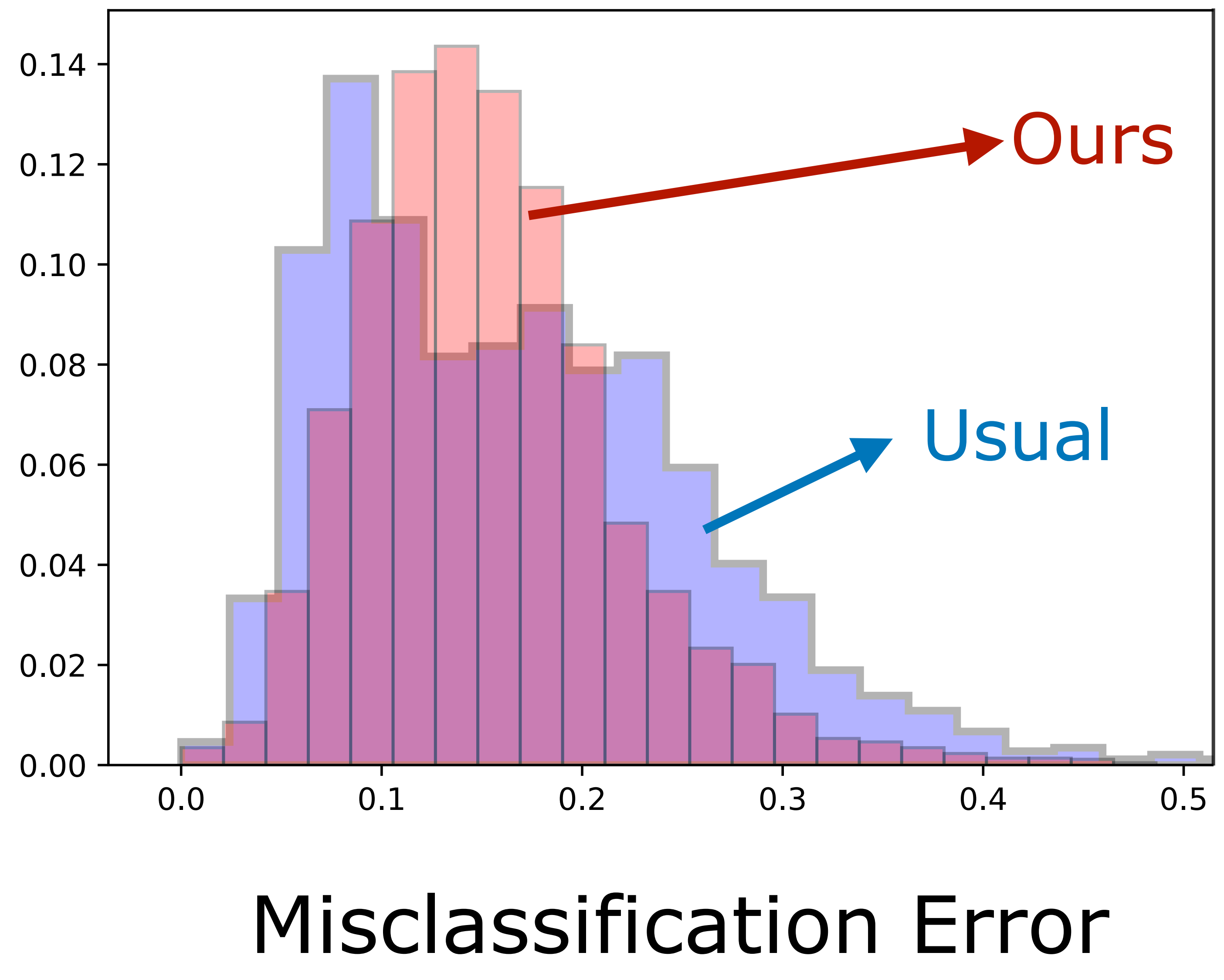


Summary: Tackling distribution shifts in federated learning

Distribution shift $\Longrightarrow$ large tail errors

$$\min_w \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Our approach reduces tail error



Summary: Tackling distribution shifts in federated learning

Distribution shift $\Longrightarrow$ large tail errors

$$\min_w \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

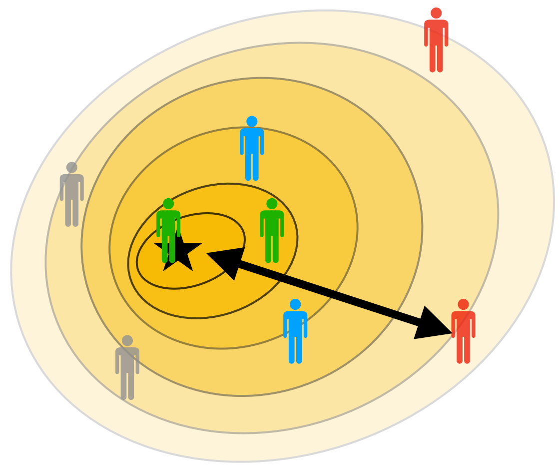
Convergence + Privacy analysis

$O(1/\sqrt{t})$ error rate after t comm. rounds in the non-smooth, non-convex case

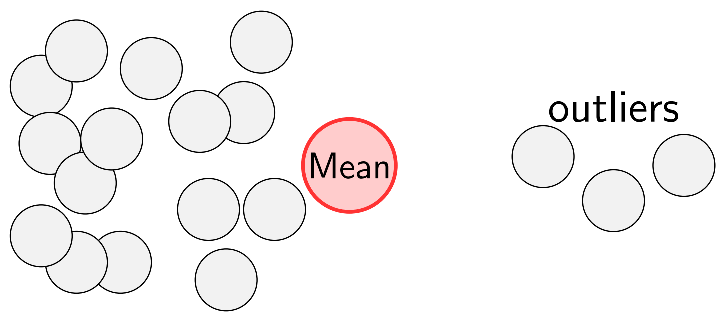
Differentially private algorithm for distributional robust federated learning

Future research plans

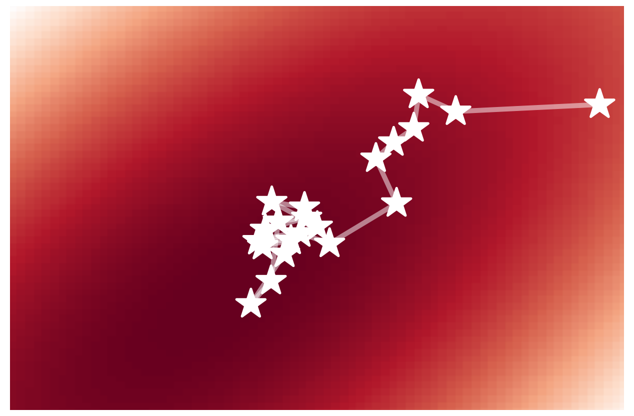
Challenges



Robustness to deployment conditions that differ from training

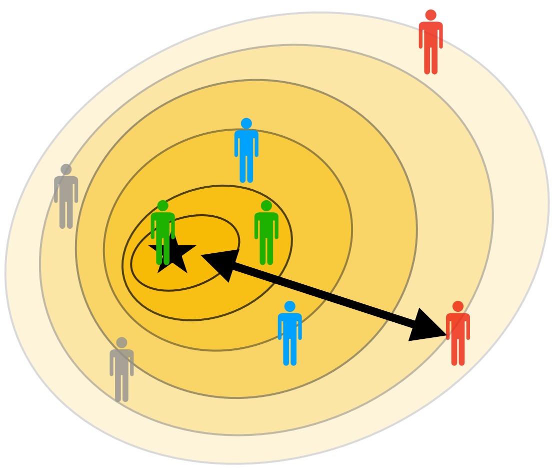
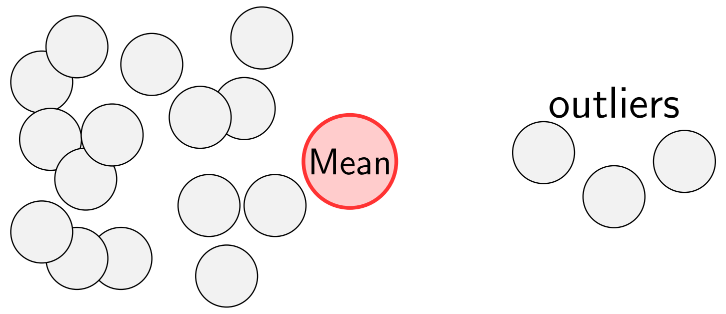
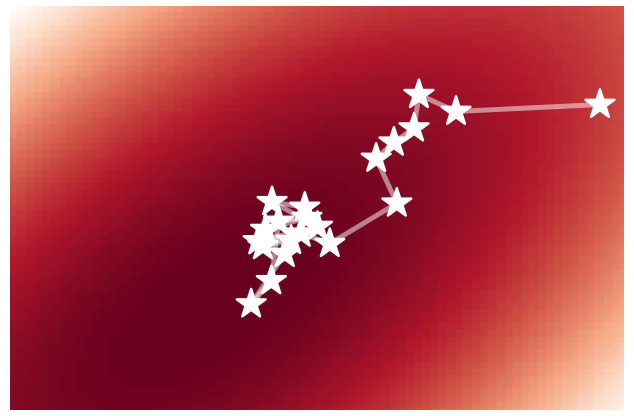


Robustness to outliers: adversarial or uncurated web data



Faster optimization: reduce communication and computation

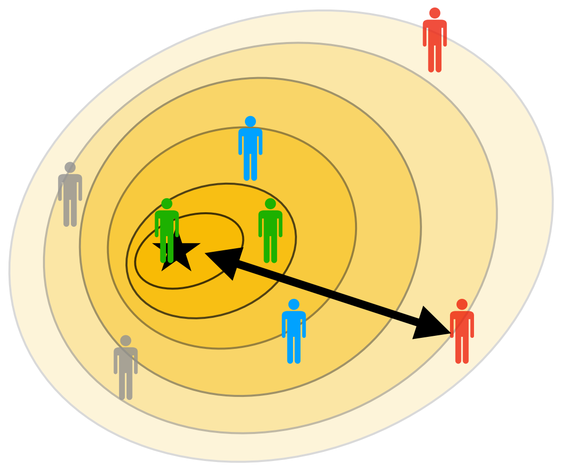
Privacy of user data

		Federated learning	LLMs
Robust Deployment		IEEE CISS 2021, *Springer SVVA 2021,* *Mach. Learn. 2022*	NeurIPS 2021a NeurIPS 2021b *Submitted 2023*
Robust to Outliers		*IEEE Trans. Signal Proc. 2022,* ICML 2022	Submitted 2022
Optimize Faster		✓	NeurIPS 2018 Submitted 2022
Privacy		✓	**Future plan 1**

What comes after federated learning?

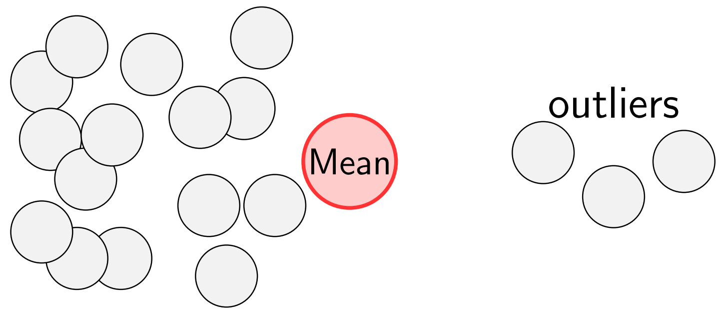
LLMs

Robust Deployment



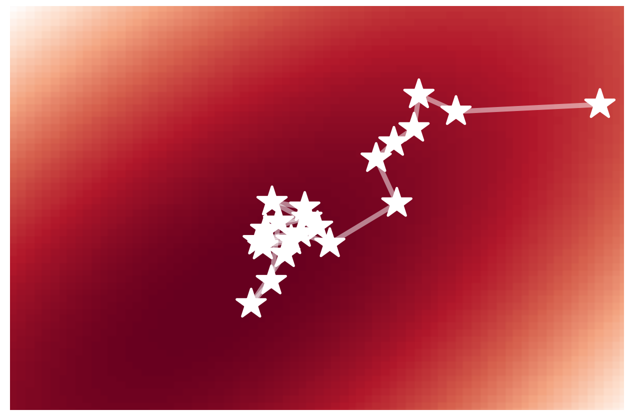
NeurIPS 2021a
NeurIPS 2021b
Submitted 2023

Robust to Outliers



Submitted 2022

Optimize Faster



NeurIPS 2018
Submitted 2022

Privacy

Future plan 2

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Thank you!

NSF

J.P.Morgan