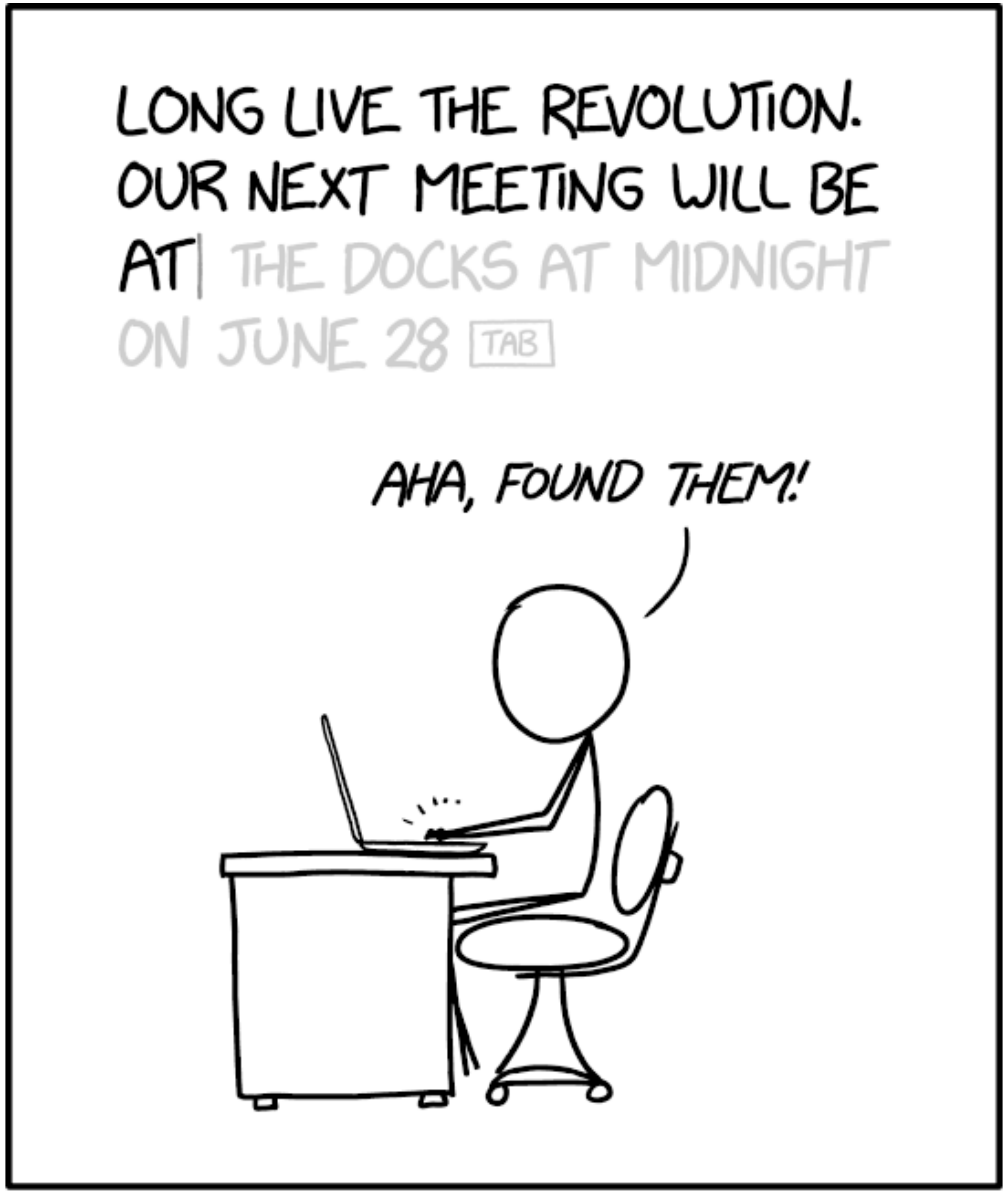
DA 7450: Topics in AI Privacy

Krishna Pillutla

IIT Madras

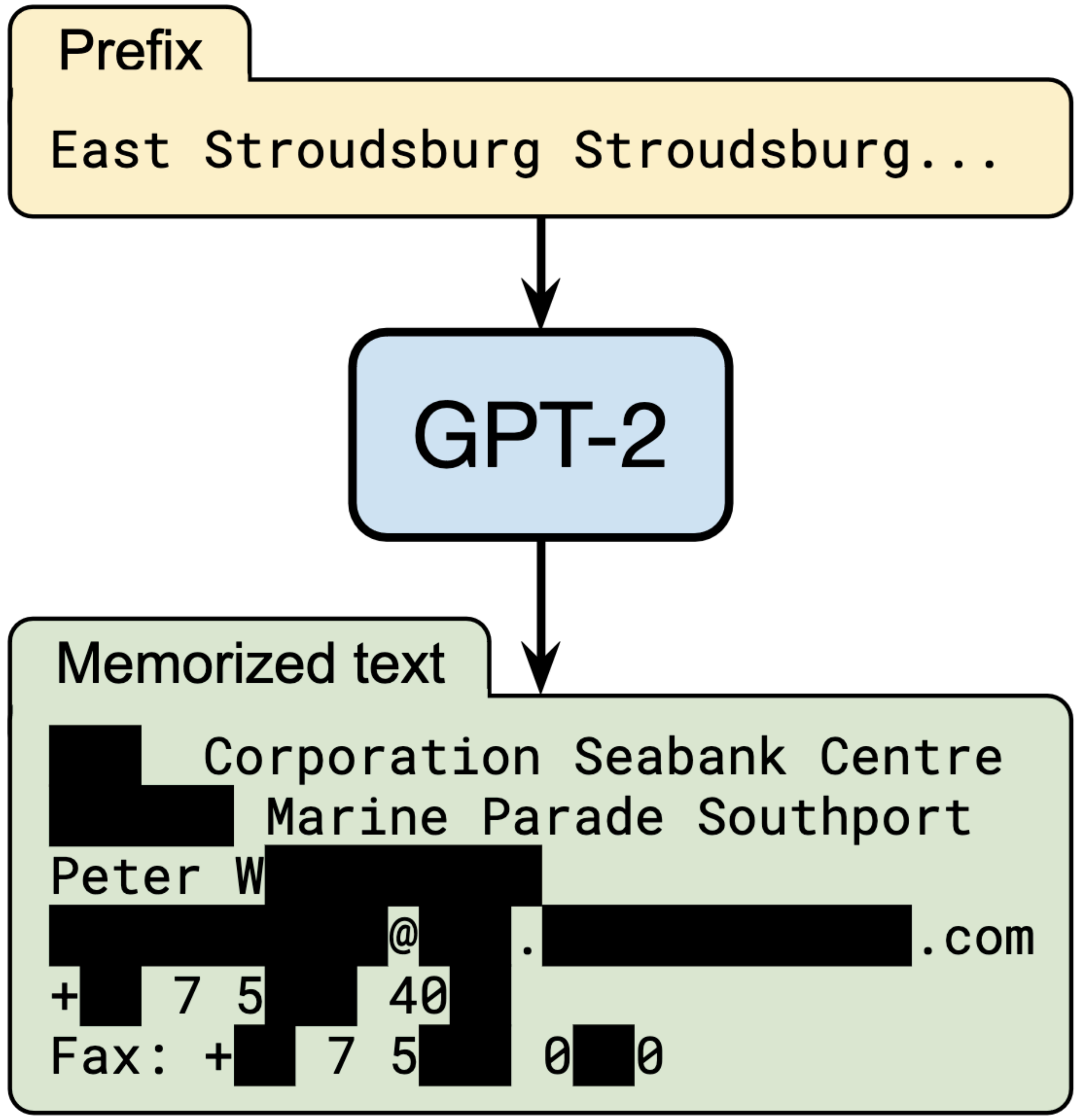

LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT| THE DOCKS AT MIDNIGHT ON JUNE 28 TAB
AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

Models leak information about their training data

LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT| THE DOCKS AT MIDNIGHT ON JUNE 28 TAB
AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.


Prefix

```
East Stroudsburg Stroudsburg...
```

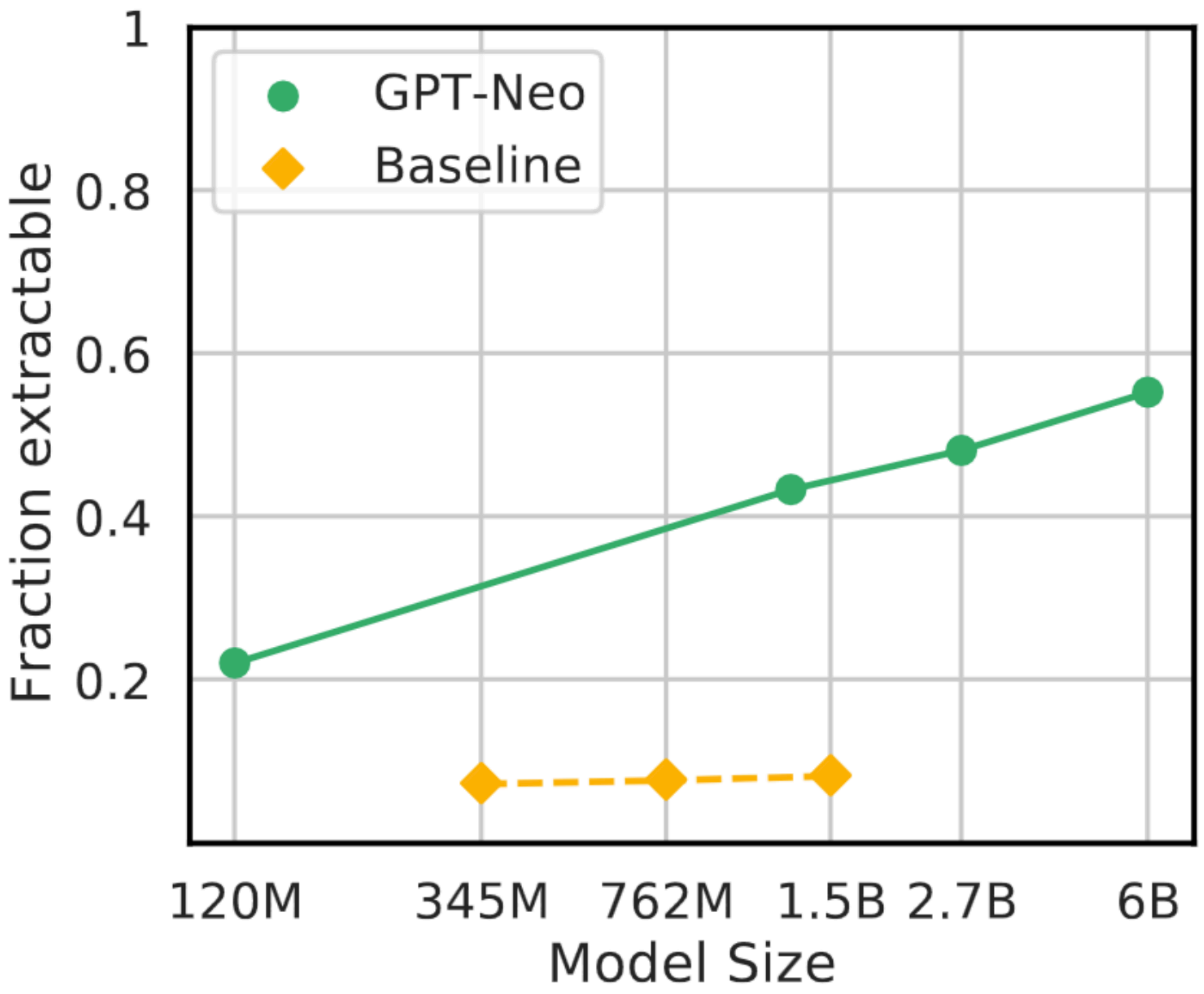
GPT-2

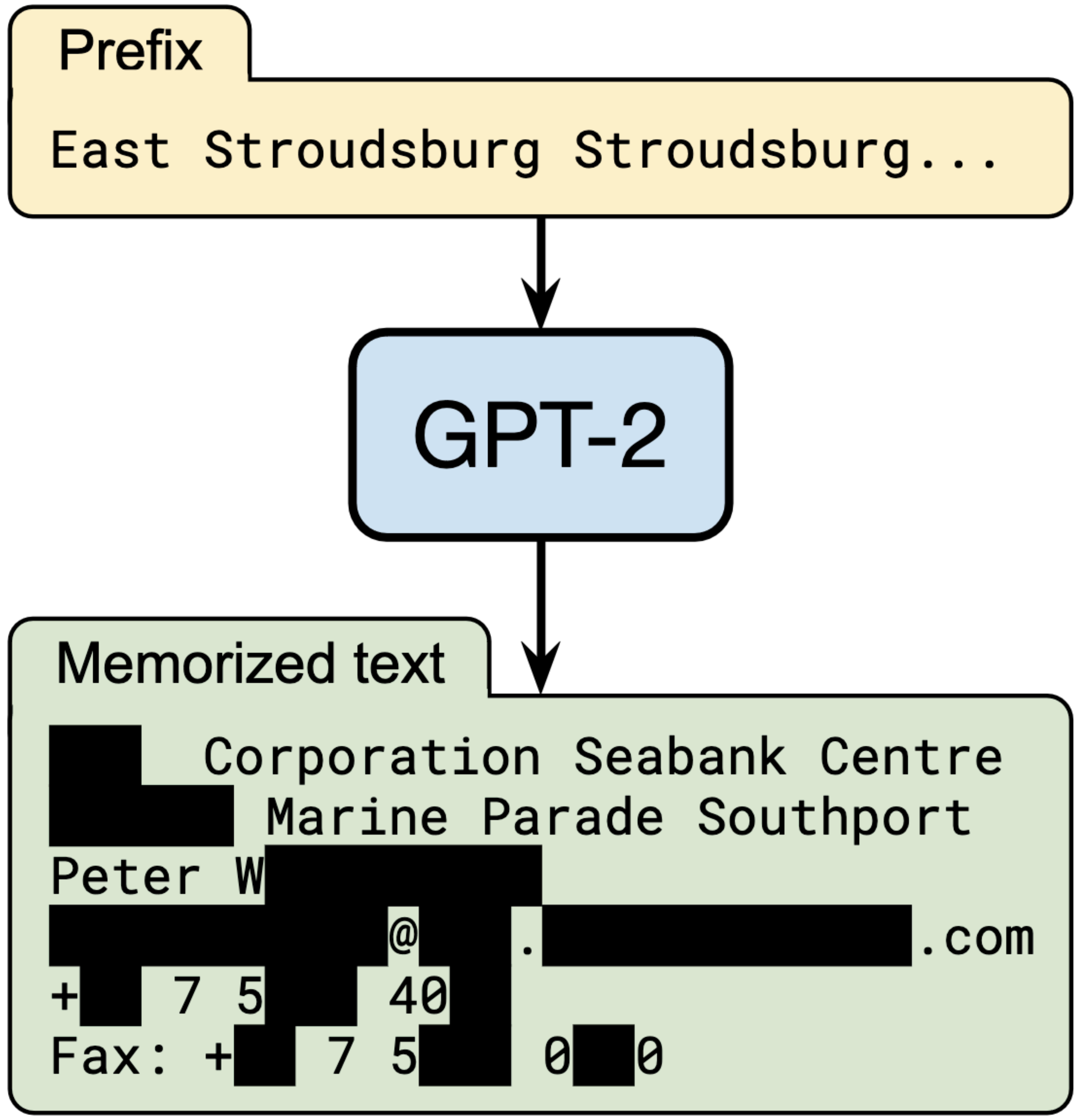
Memorized text

```
[redacted] Corporation Seabank Centre
 Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

Carlini et al. (USENIX Security 2021)

Models leak information about their training data ***reliably***





```
East Stroudsburg Stroudsburg...
```

```
[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

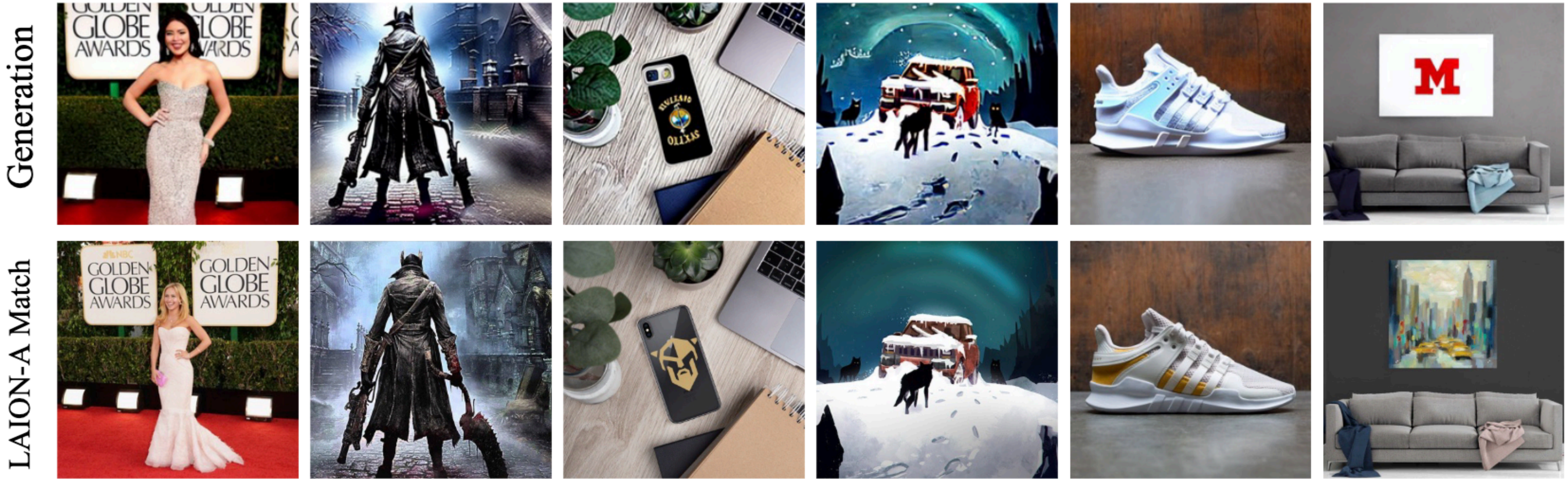
Carlini et al. (USENIX Security 2021)

Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

Gowthami Somepalli 🐢 , Vasu Singla 🐢 , Micah Goldblum 🗽 , Jonas Geiping 🐢 , Tom Goldstein 🐢

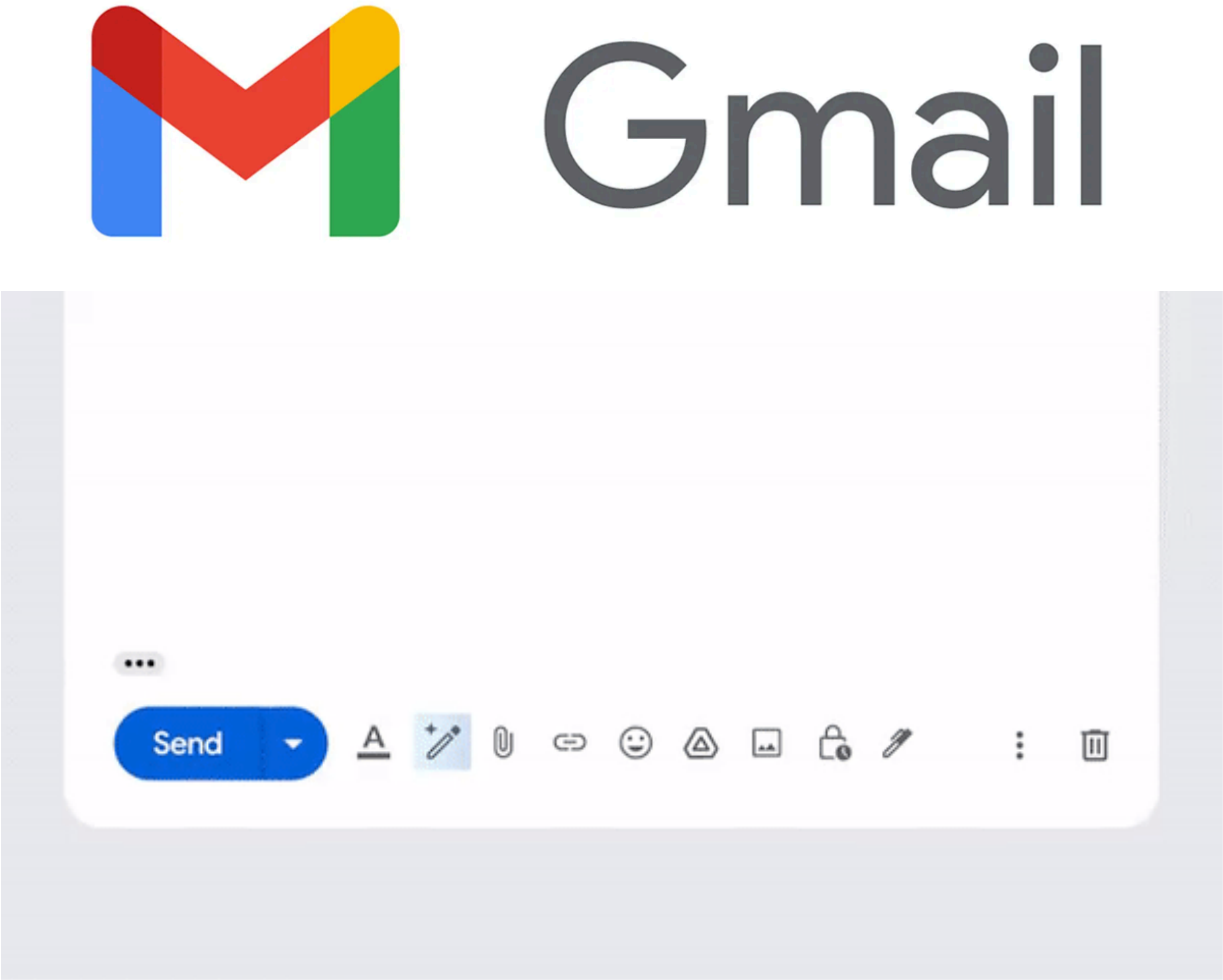
🐢 University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

🗽 New York University
goldblum@nyu.edu

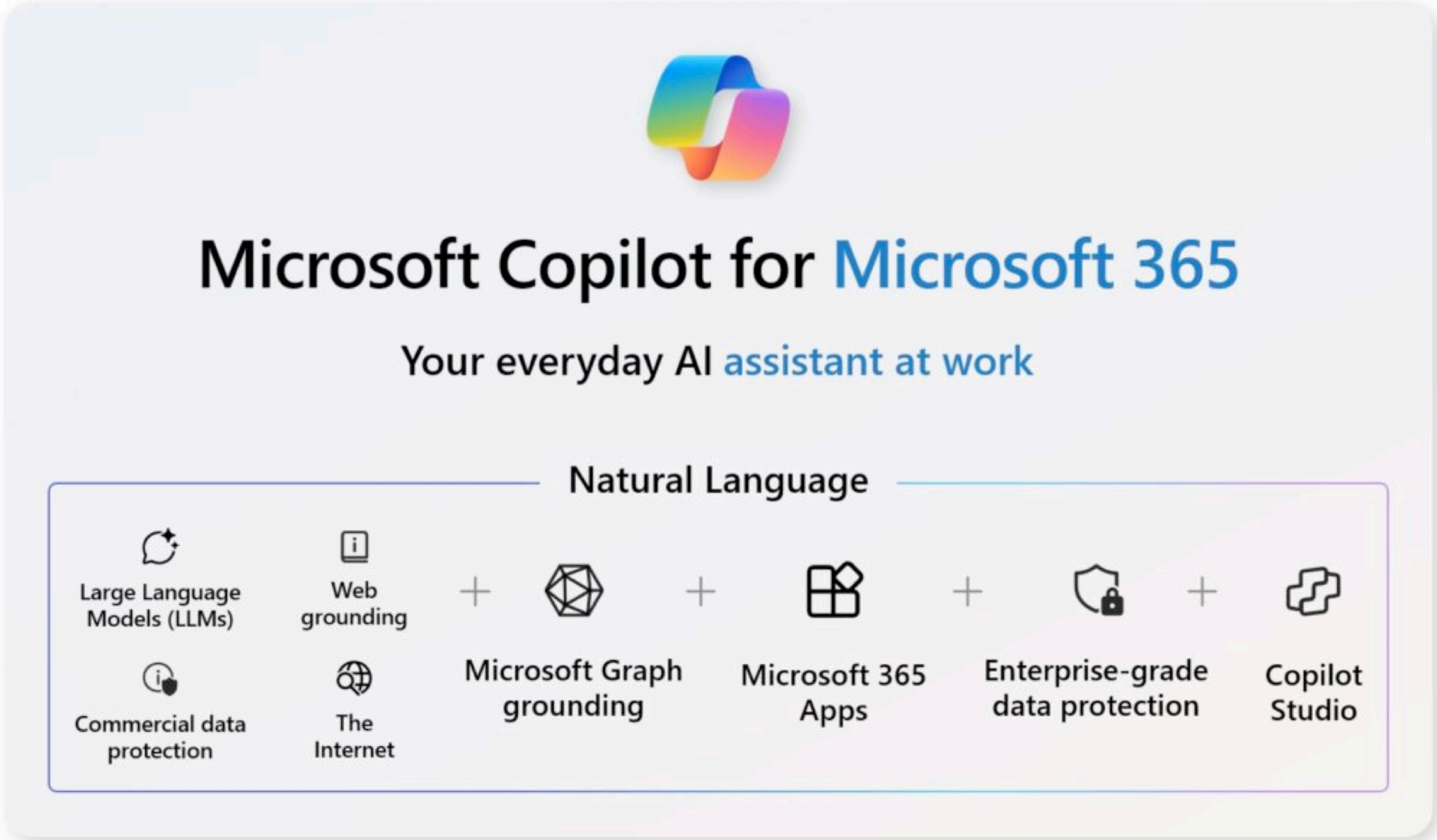


Blog

Introducing ChatGPT



https://blog.google/products/gmail/gmail-ai-features/



Microsoft Copilot for Microsoft 365

Your everyday AI assistant at work

Natural Language

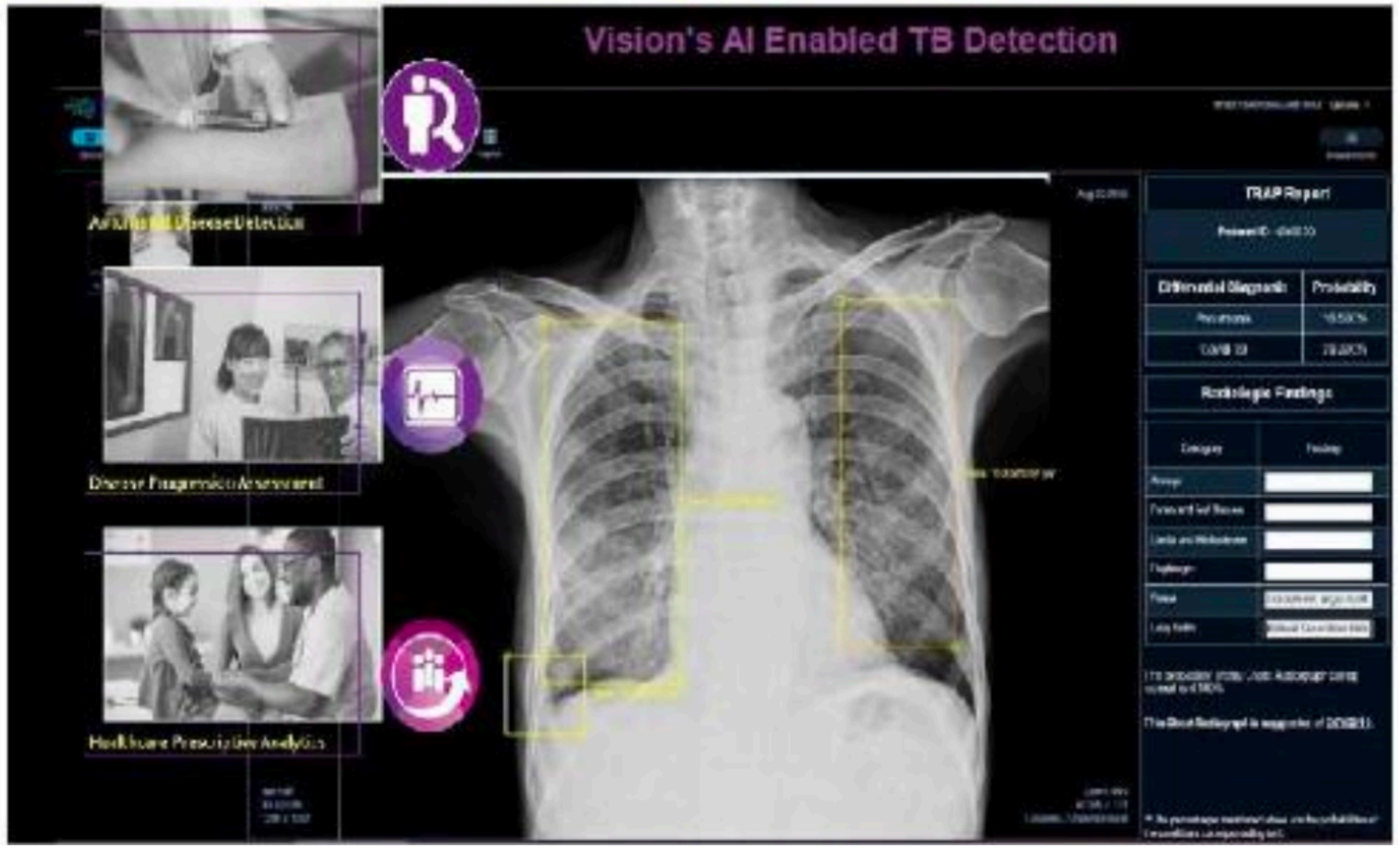
Large Language Models (LLMs)

Web grounding

Commercial data protection

The Internet

+ Microsoft Graph grounding + Microsoft 365 Apps + Enterprise-grade data protection + Copilot Studio


Vision's AI Enabled TB Detection
Artificial Disease Detection
Disease Progression Assessment
Healthcare Prescriptive Analytics
TRAP Report
Differential Diagnosis
Probability
Radiologic Findings

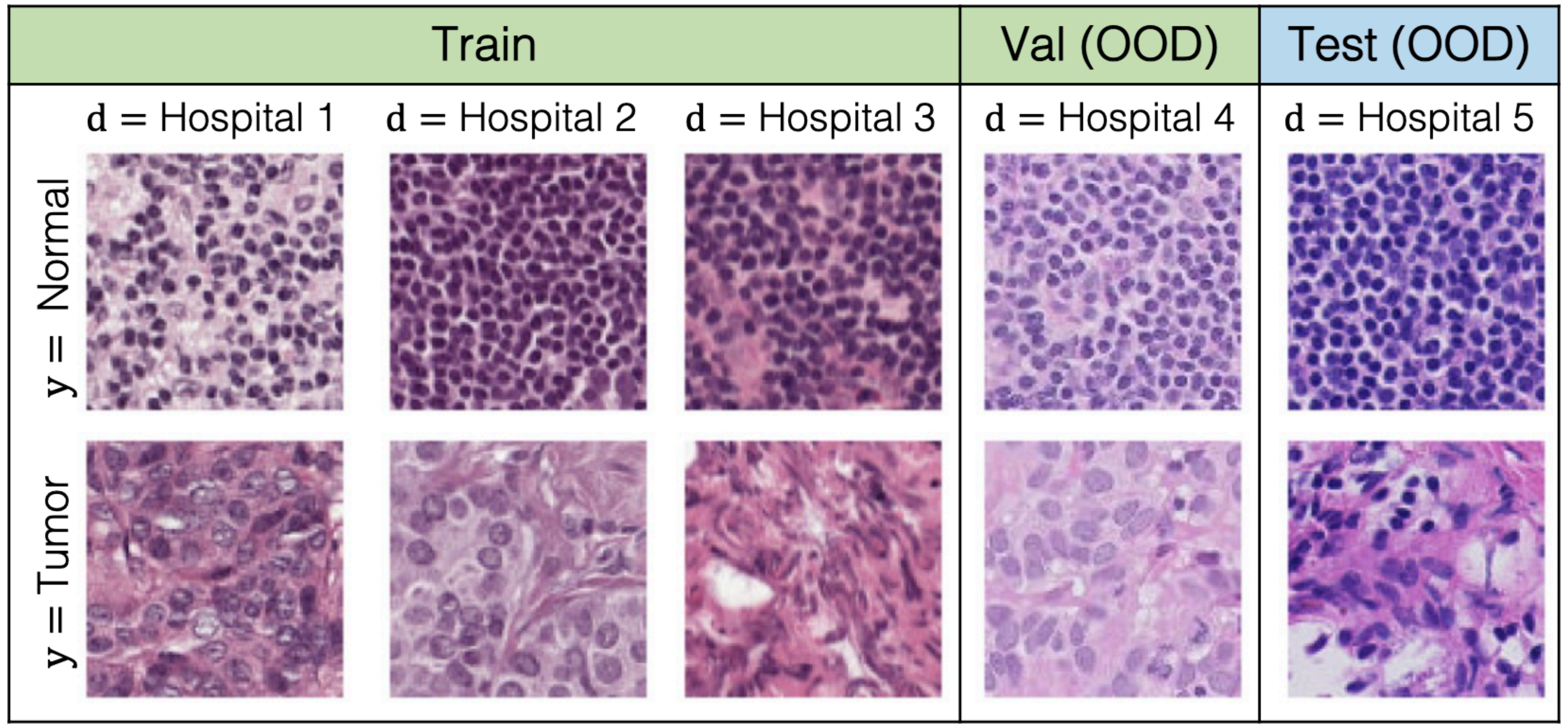
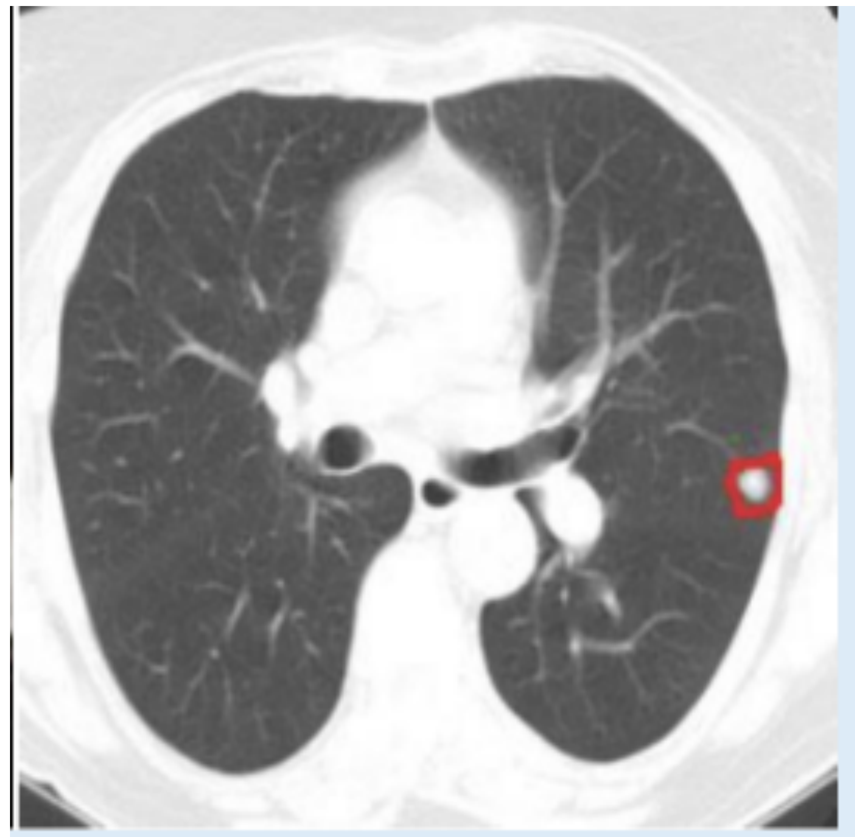
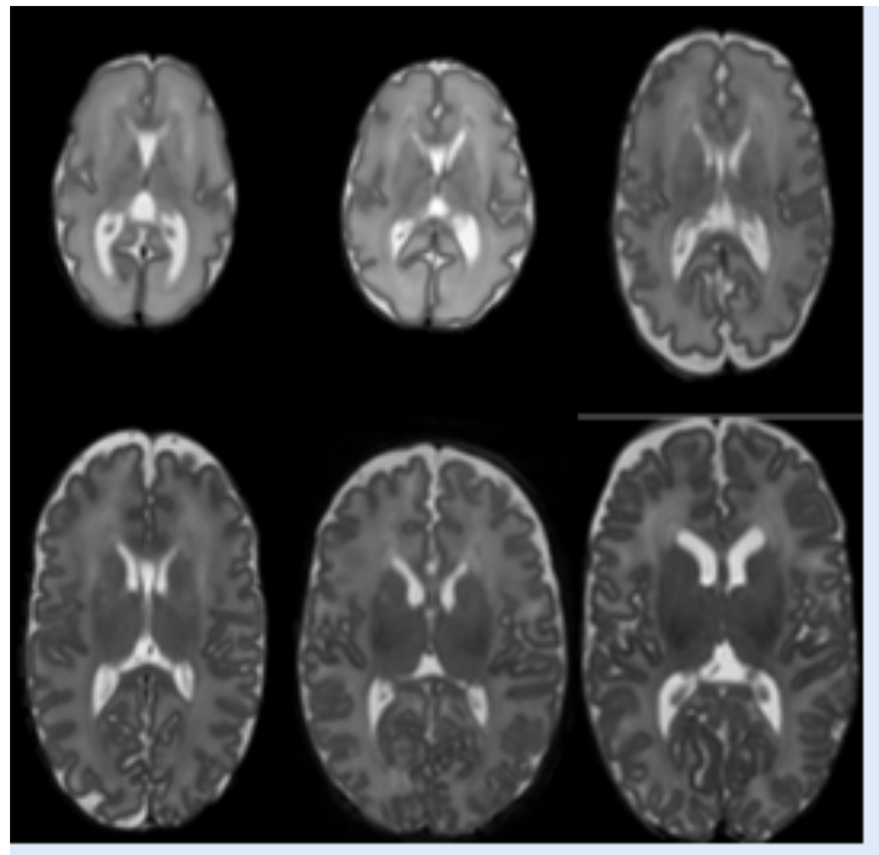
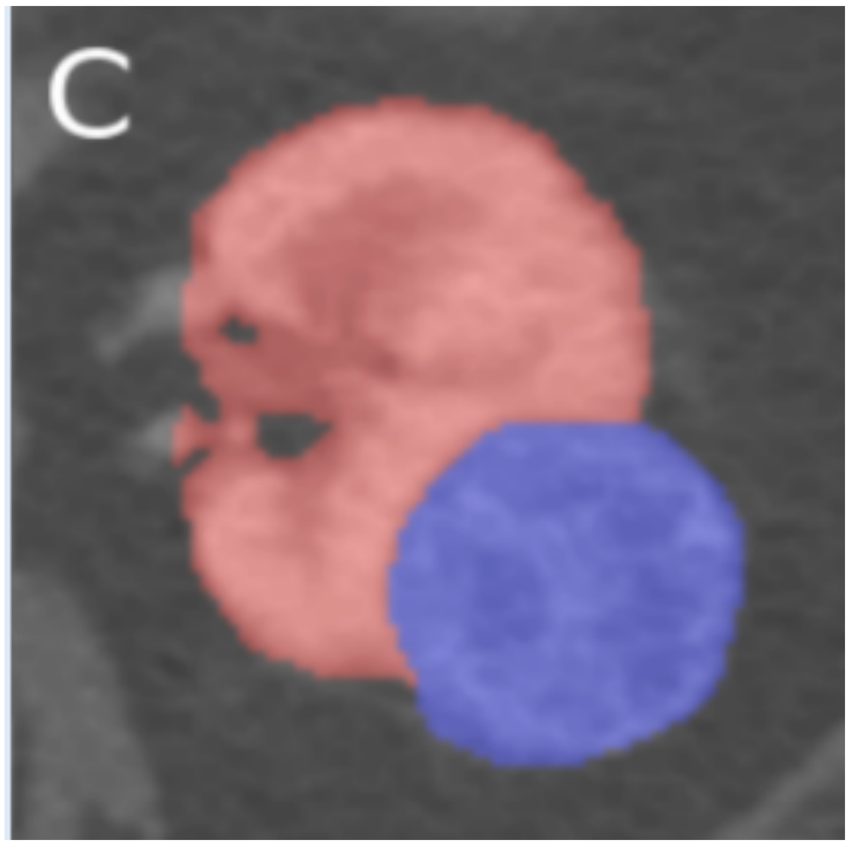
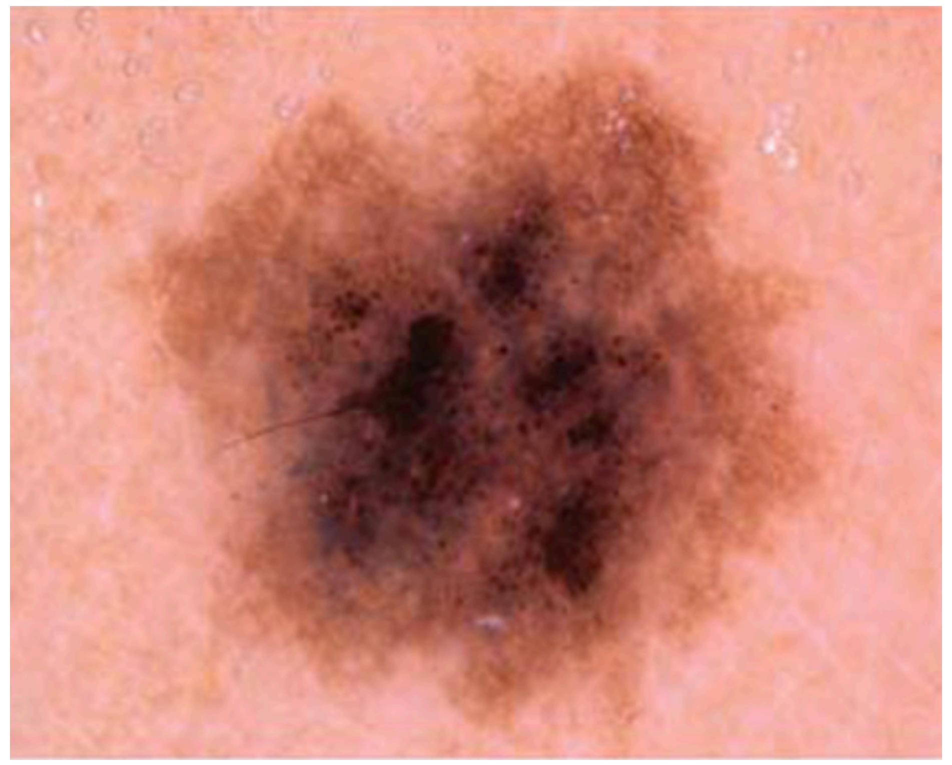


Figure 4: The CAMELYON17-WILDS dataset comprises tissue patches from different hospitals. The goal is to accurately predict the presence of tumor tissue in patches taken from hospitals that are not in the training set. In this figure, each column contains two patches, one of normal tissue and the other of tumor tissue, from the same slide.







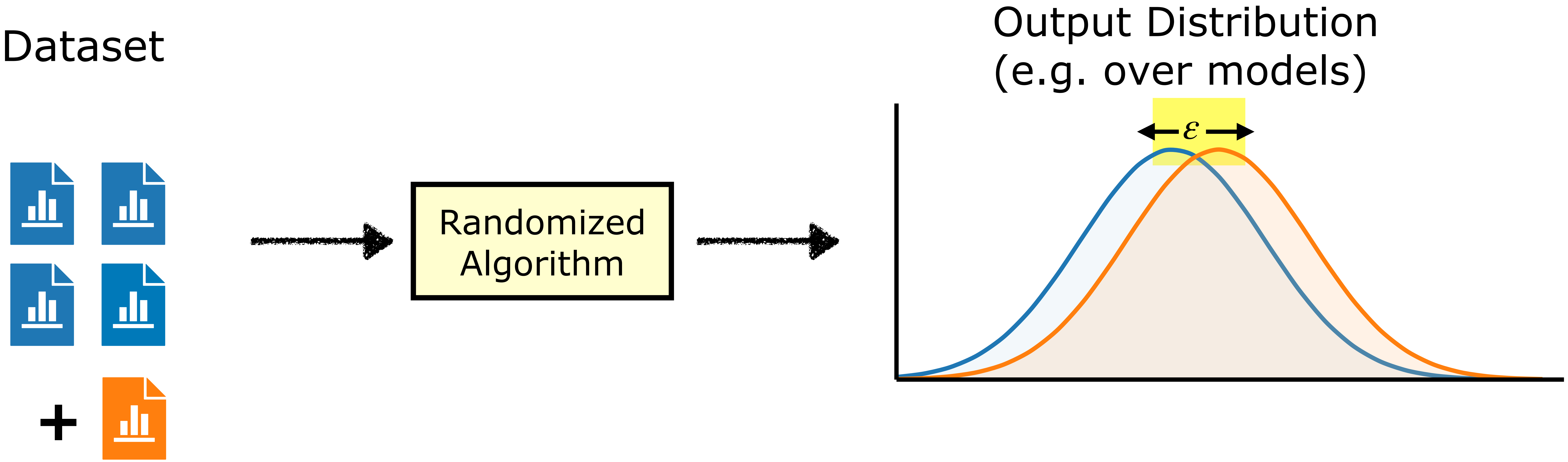


Digital Health Laws and Regulations India 2024

HIPAA
Health Insurance Portability and Accountability Act

EU GDPR

Differential privacy nearly eliminates memorization



Dataset

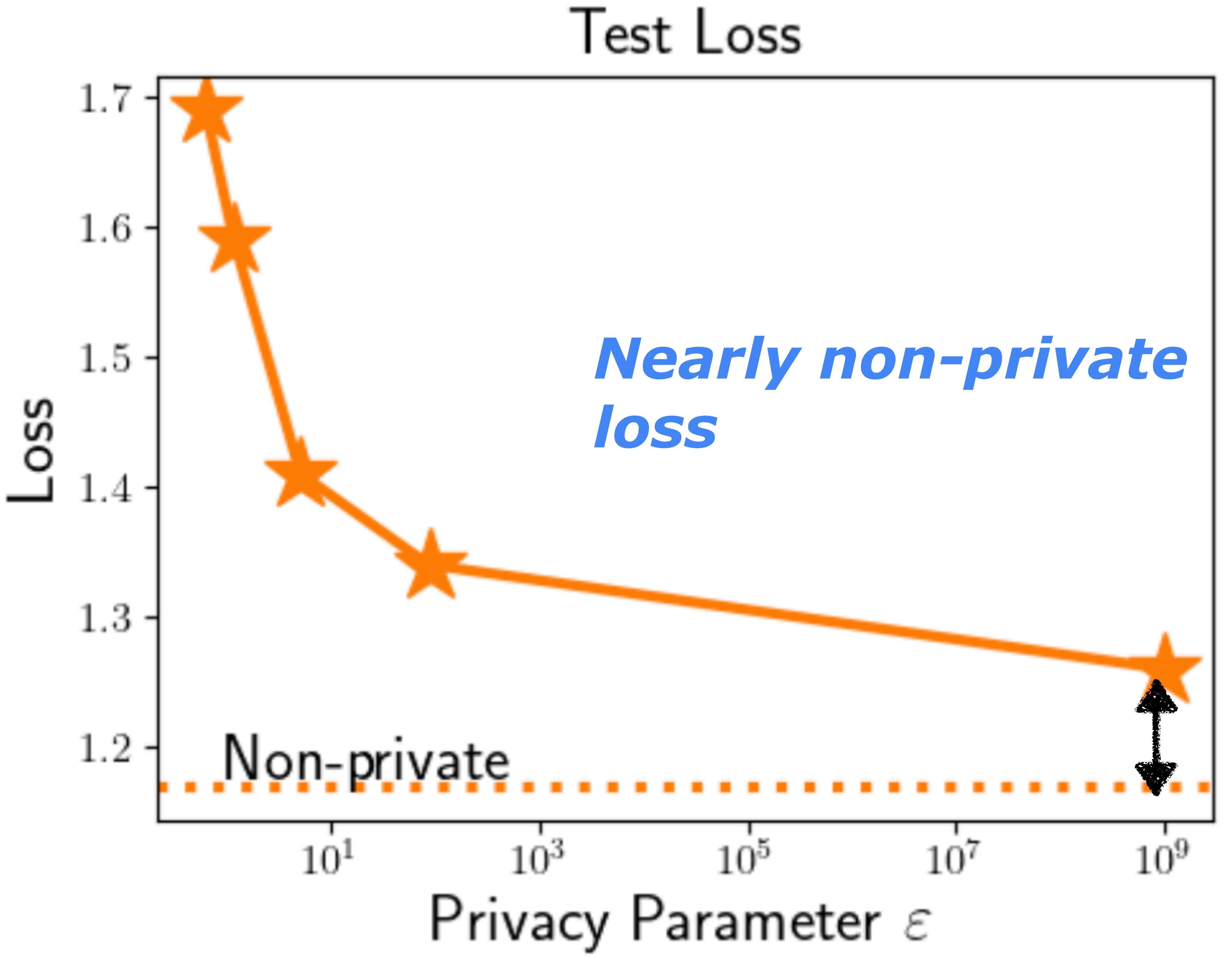
Randomized Algorithm

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

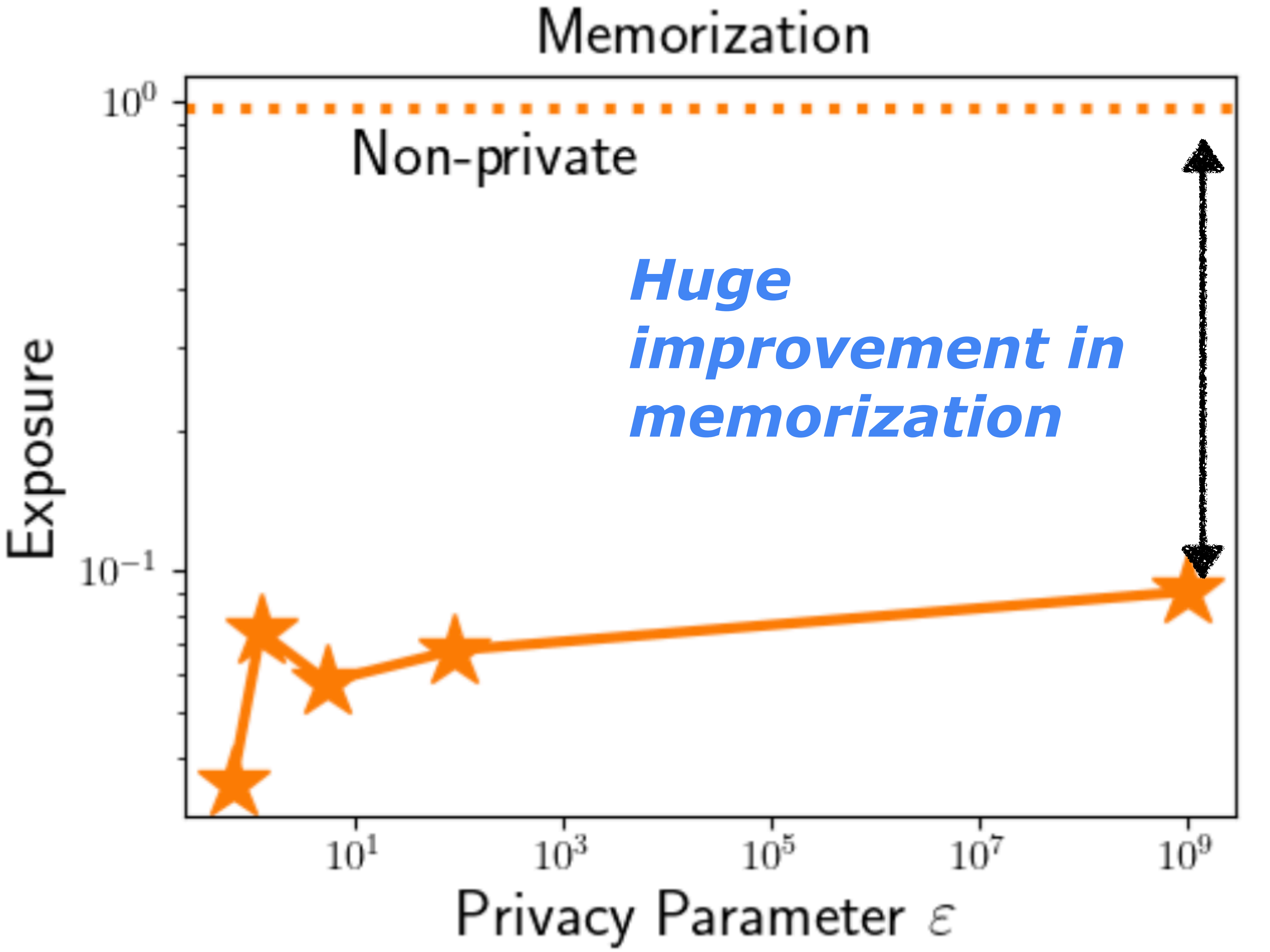
Dwork, McSherry, Nissim, Smith. **Calibrating Noise to Sensitivity in Private Data Analysis**. *TCC 2006*.

Differential privacy nearly eliminates memorization



Nearly non-private loss

High privacy ***Low privacy***



Huge improvement in memorization

High privacy ***Low privacy***

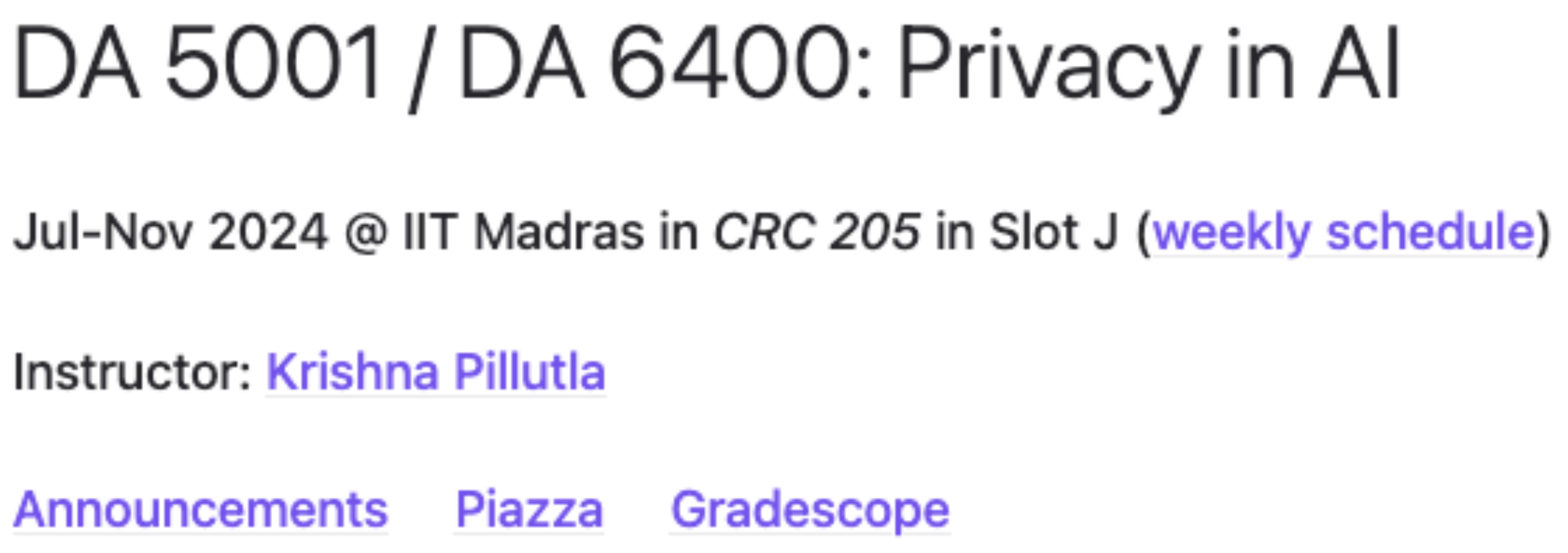
Carlini, Liu, Erlingsson, Kos, Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.** USENIX Security 2019.

Today’s Outline

- Logistics
- Course Outline

Logistics

Prerequisites



DA 5001 / DA 6400: Privacy in AI

Jul-Nov 2024 @ IIT Madras in *CRC 205* in Slot J (weekly schedule)

Instructor: Krishna Pillutla

Announcements Piazza Gradescope

- PrivAI Course is required: https://krishnap25.github.io/privAI_course_2024o/
- **Exception**: You can take this course if:
  - You score >75% on a HW to make sure you have enough background
    - You can use course material from the previous course
    - Academic integrity policies of both courses will apply
  - Your research project (MS/PhD/DDP/BTP/etc.) is based ***directly*** on this topic
    - HW with >60% score necessary

Seminar Course

- One student presents a paper in a lecture (slides or board)
    - Practical motivation
    - Mathematical details (including proofs)
    - Real world impact and significance
- Everybody participates in a discussion

Seminar Course: Work required

- Each person presents 3 or so times in the semester
    - Detailed preparation - takes 2 weeks or so each time
- Others: skim through the reading material to contribute to the discussion

- Presentation + participation: 50% of the grade

Class Timings: Slot L

- **Thursday**: 2 to 3:15 pm
- **Friday**: 3:30 to 4:45 pm

- Extra lectures on two Saturdays: Feb. 8 and March 8th.
    - Pre-emptive make-up for classes to be canceled in March/April

Communication

- **Course Webpage**: https://krishnap25.github.io/privAI_course_2024o/
- **Piazza**: link to be announced on the course webpage

Grading

- ***Presentation***: 40%
- ***Participation***: 10%
- ***Course Project***: 50%

Course Project: 50% of the grade

- Most of your learning will be through the course project
- Individual or groups of 2
- ***Options***:
    - Research project
    - Implementation: benchmarking and open-sourcing
    - In-depth paper analysis

Course Project: Research

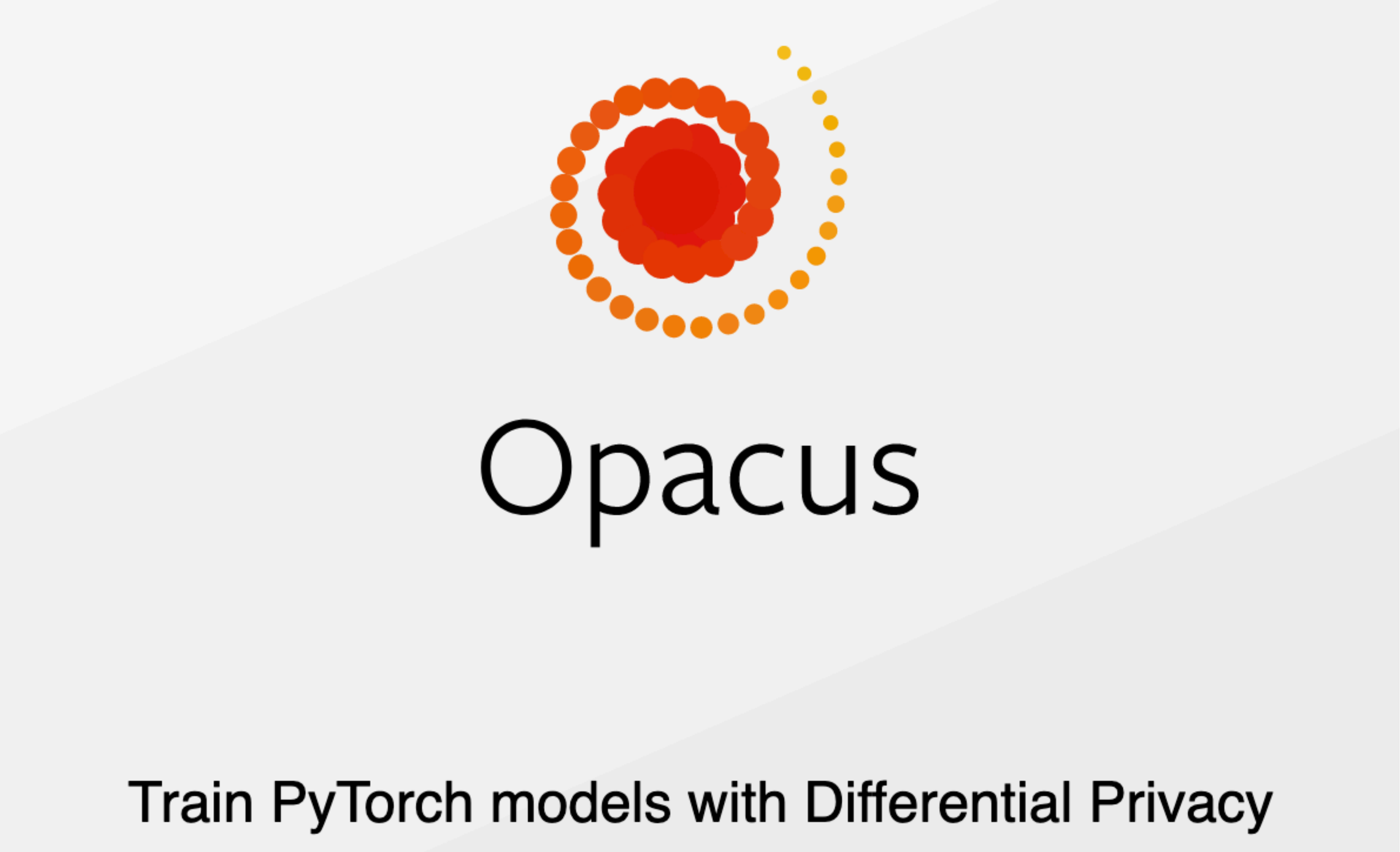
- **Original research**: can be theory, applied, or mix or both
- Commensurate to a workshop paper at NeurIPS/ICML/ICLR conferences
- ***Strongly encouraged*** to continue last semester's privAI course project
    - Favourable outcomes likely for many course projects from last semester
    - Not mandatory — discuss with me to decide

Course Project: Research

- You can propose your own course project related to your research
    - Must be related to the course contents
    - E.g. You work in computer vision for healthcare:
      Implement private training or privacy attacks etc. on your model/dataset
- We will also provide some project suggestions

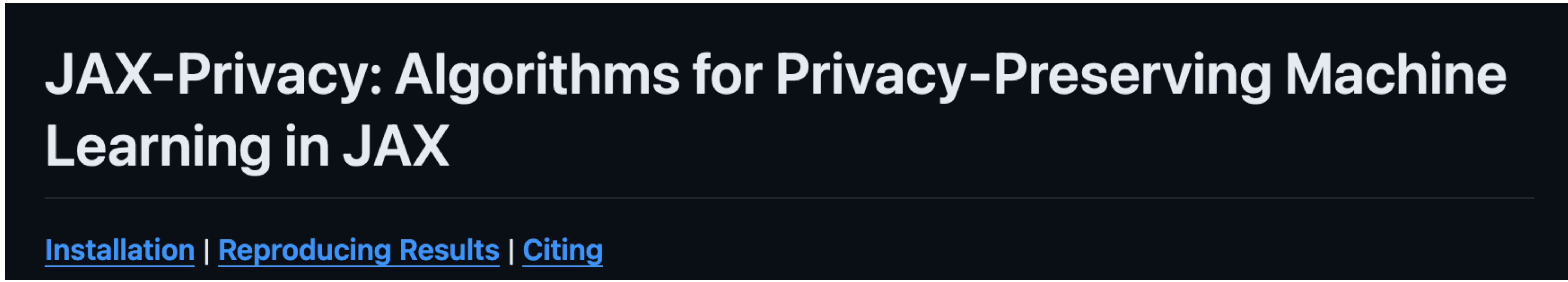
Course Project: Implementation

- Implement existing algorithms with a goal of:
  - Benchmarking methods (e.g. compare to various baselines)
  - Creating or contributing to open-source packages



Opacus

Train PyTorch models with Differential Privacy



JAX-Privacy: Algorithms for Privacy-Preserving Machine Learning in JAX

Installation | Reproducing Results | Citing

Course Project (50%): Logistics

- **Proposal (10%, 2-4 pages):** Due mid-Feb.
- **Midpoint report (10%, 4-6 pages)**: Due around mid-March
- **Presentation (15%)**: Last 2 weeks of class
- **Final report (15%, 8 pages)**: Due around May 9th (End sem week)

Late Days

- ***NO*** **late days** (for **project** or **presentation**)

Honour Code

- **Project**:
  - You have to do the work yourself (cannot use somebody else’s work as yours for a course project)
  - The project cannot be used “as is” for other courses
  - Ok to reuse course project for BTP/DDP/MTP/MS/PhD other research projects

- Academic violations will be handled by the IITM Senate Discipline and Welfare (DISCO) Committee.

Honour Code

- We expect and believe that you will conduct yourself with integrity
    - We will follow the institute policies but it is ultimately up to you to conduct yourself with academic and personal integrity for several compelling reasons (that go beyond your studies)

- **Respect diversity**: There is a place for everyone who is curious and passionate about exploring knowledge
    - Let us all be mindful of creating welcoming and inclusive spaces
    - As the next generation, you have the power to shape the future:
      aim to make the world a better place!

Office Hours

- We will be available one hour per week to answer queries about the course material
- **Thursday 3:30 to 4:30 PM** at my office (after class)

Auditing the course

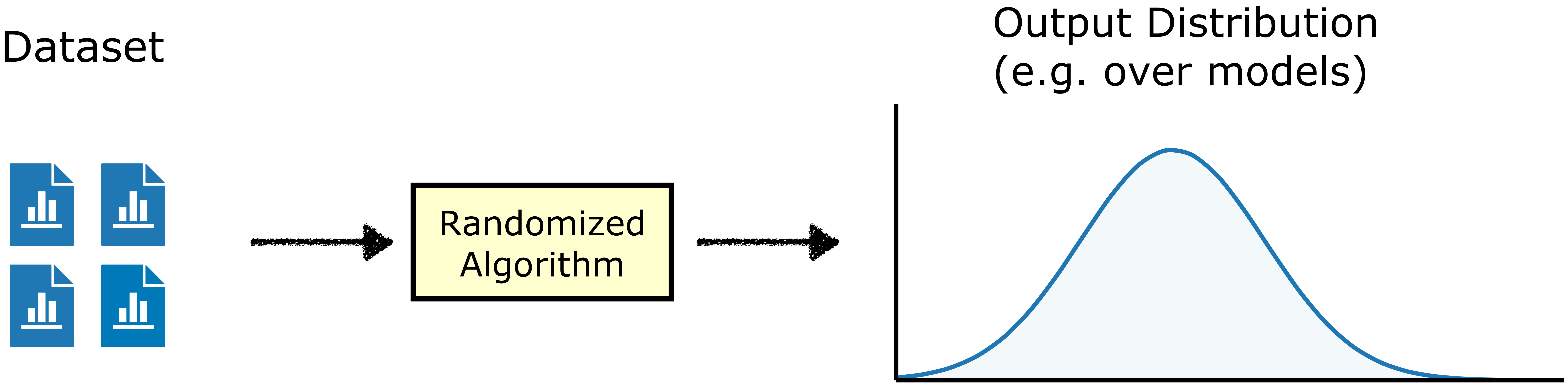
- **Not allowed**
    - Seminar courses require “buy in” from all participants

Attendance

- We will not take attendance

Recap: Differential Privacy (DP)

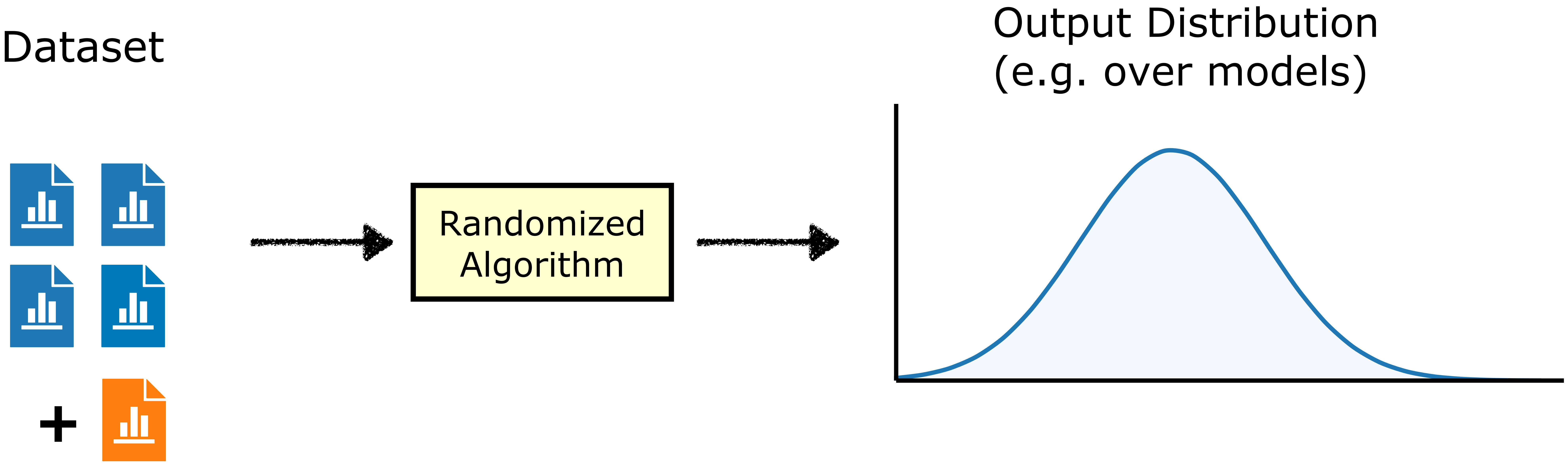
A mathematically rigorous notion of “***privacy***”



Dataset

Randomized Algorithm

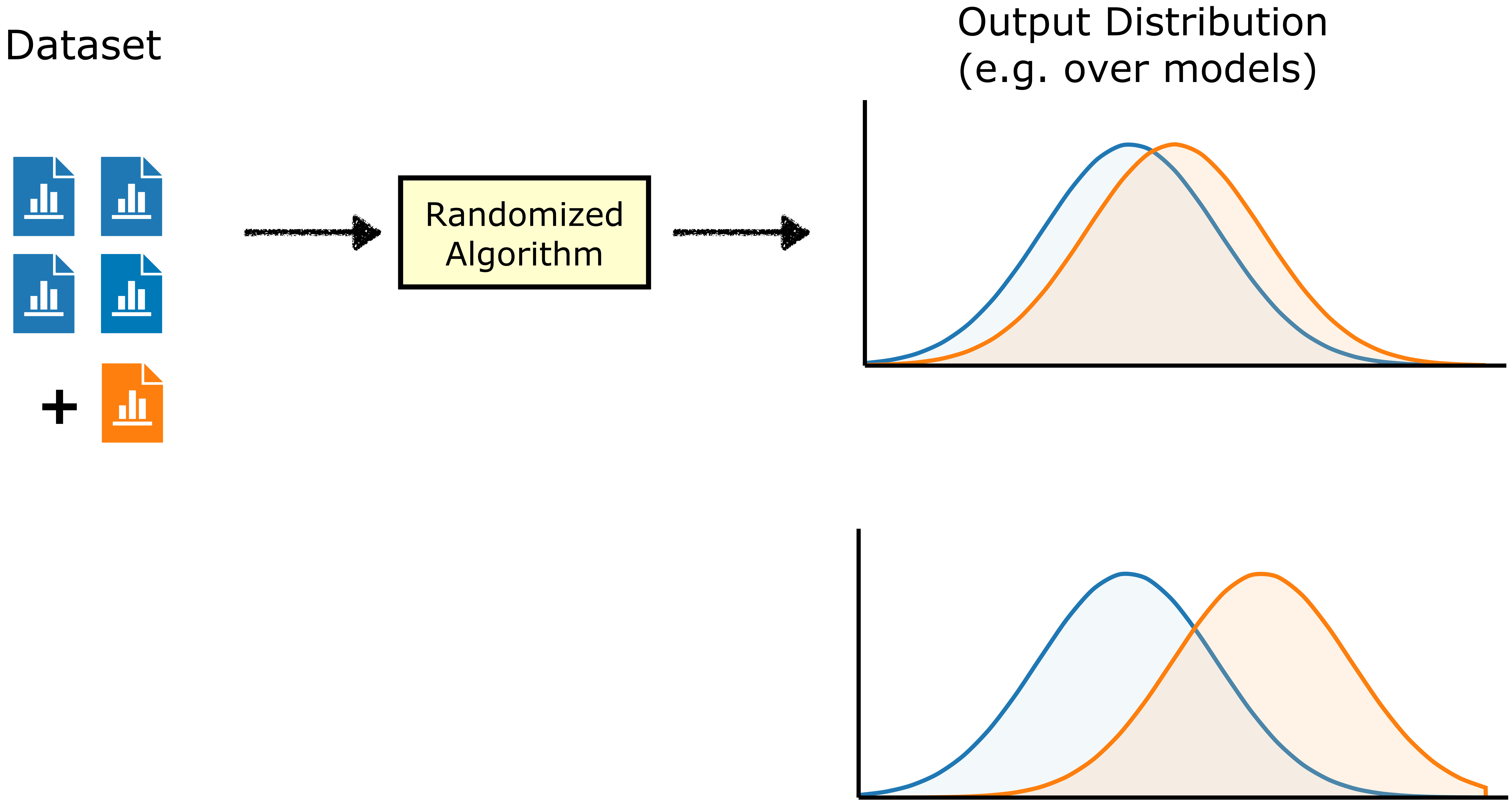
Output Distribution
(e.g. over models)



Dataset

Randomized Algorithm

Output Distribution
(e.g. over models)

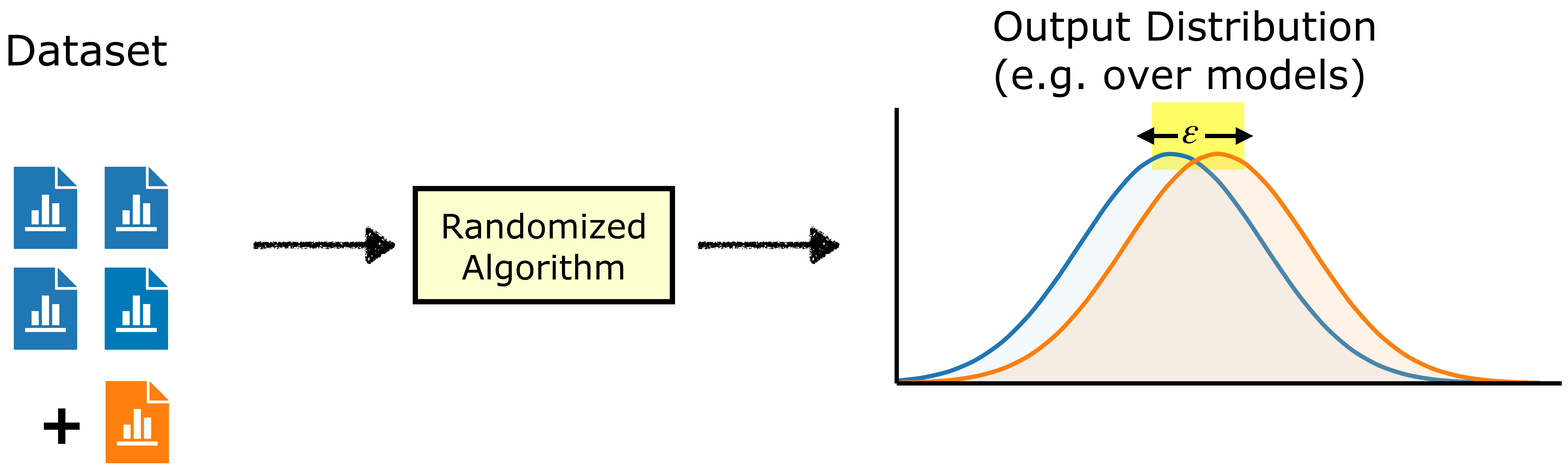


Dataset

+

Randomized Algorithm

Output Distribution
(e.g. over models)



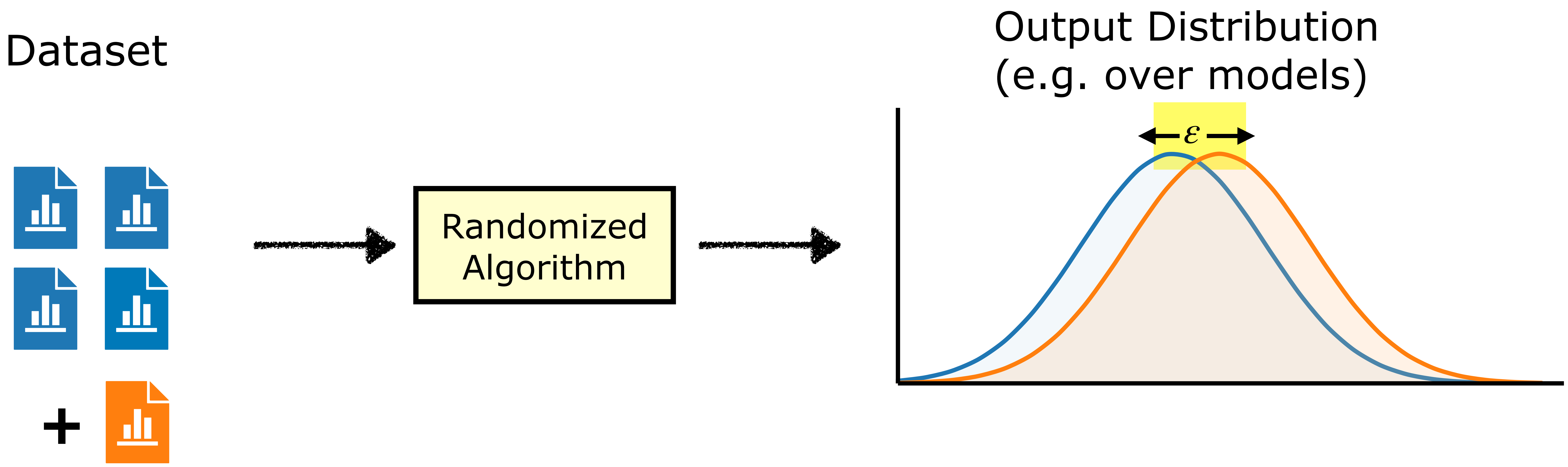
Dataset

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

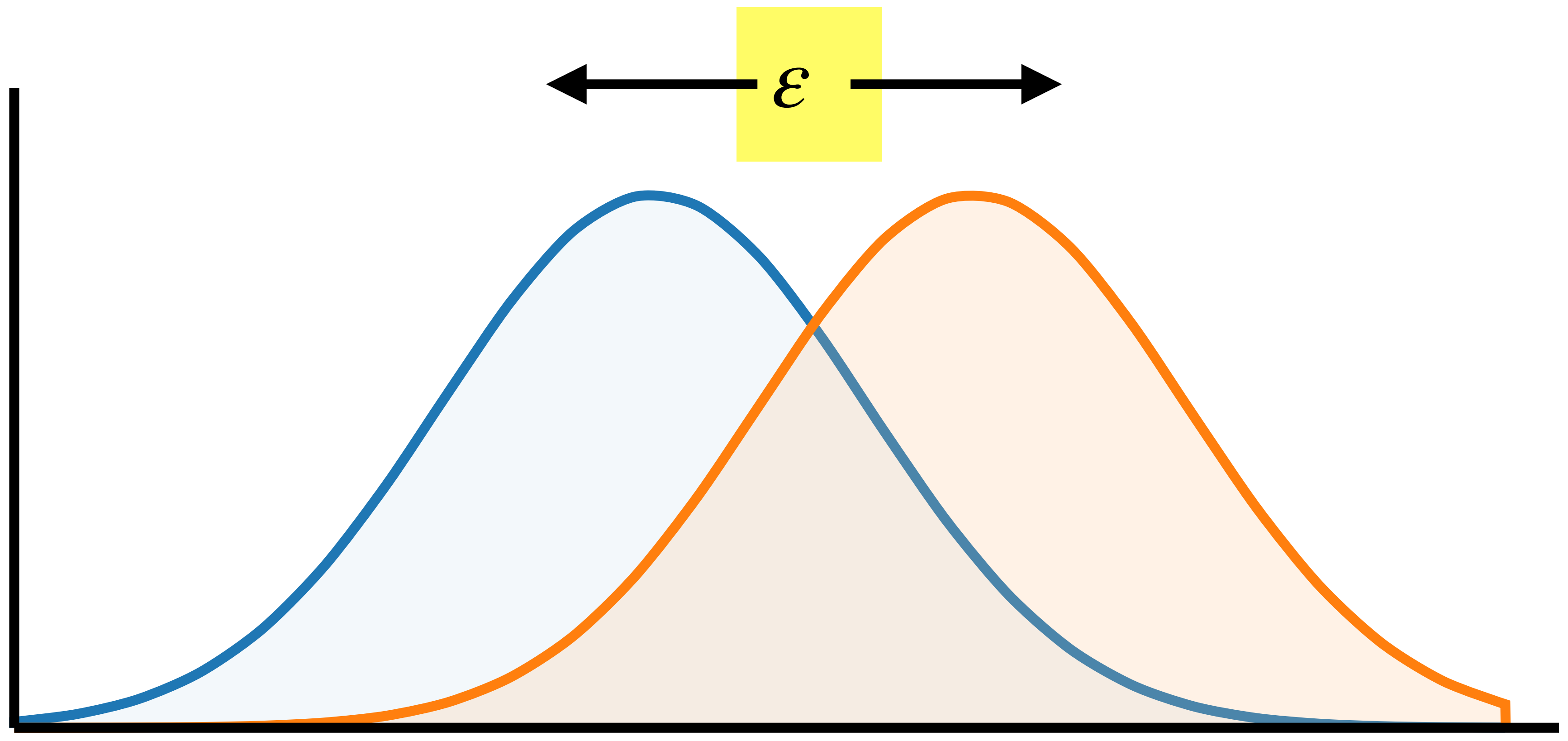
Dwork, McSherry, Nissim, Smith. **Calibrating Noise to Sensitivity in Private Data Analysis**. *TCC 2006*.



Dataset

Randomized Algorithm

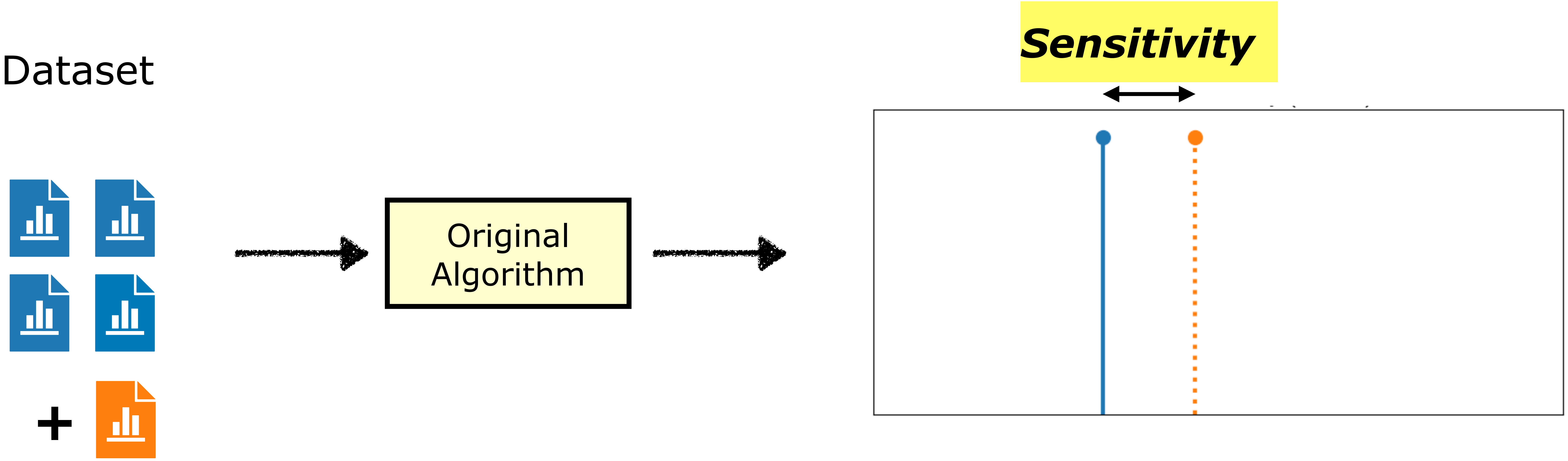
Output Distribution
(e.g. over models)



ε-differential privacy

Large ε $\Longrightarrow$ more privacy leakage

Adding noise for DP



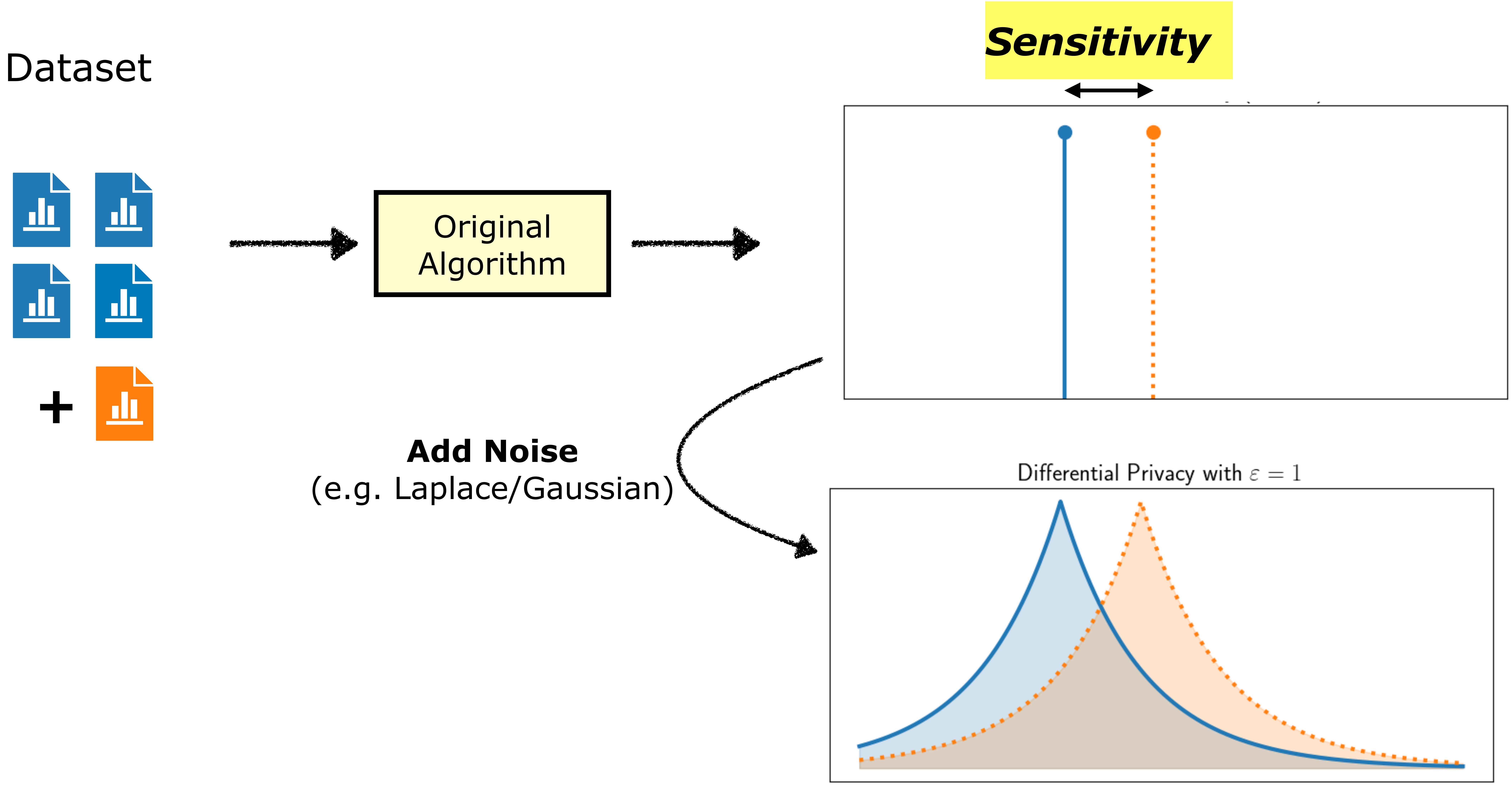
Dataset

+

Original Algorithm

Sensitivity

Adding noise for DP



Dataset

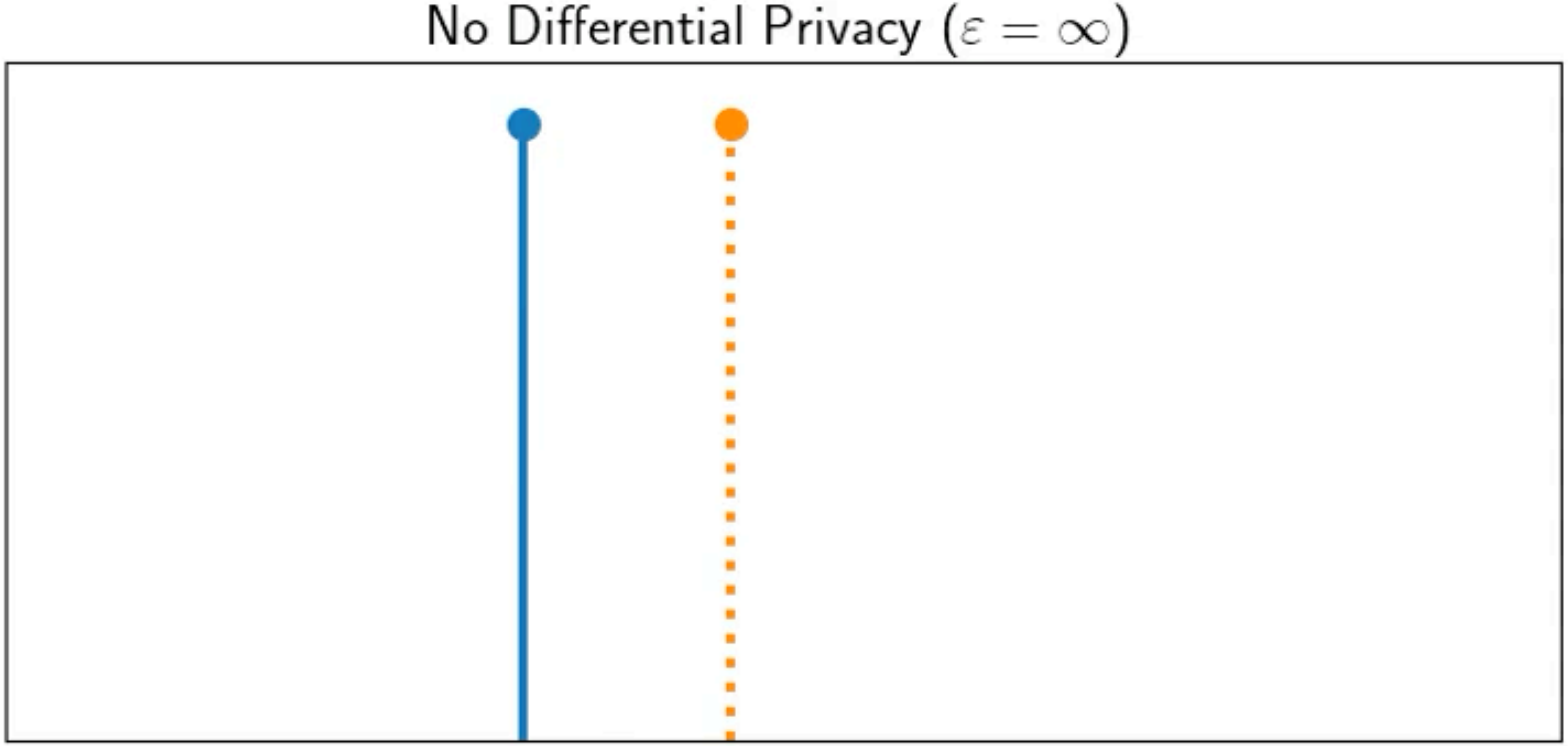
+

Original Algorithm

Add Noise
(e.g. Laplace/Gaussian)

Sensitivity

Adding noise for DP



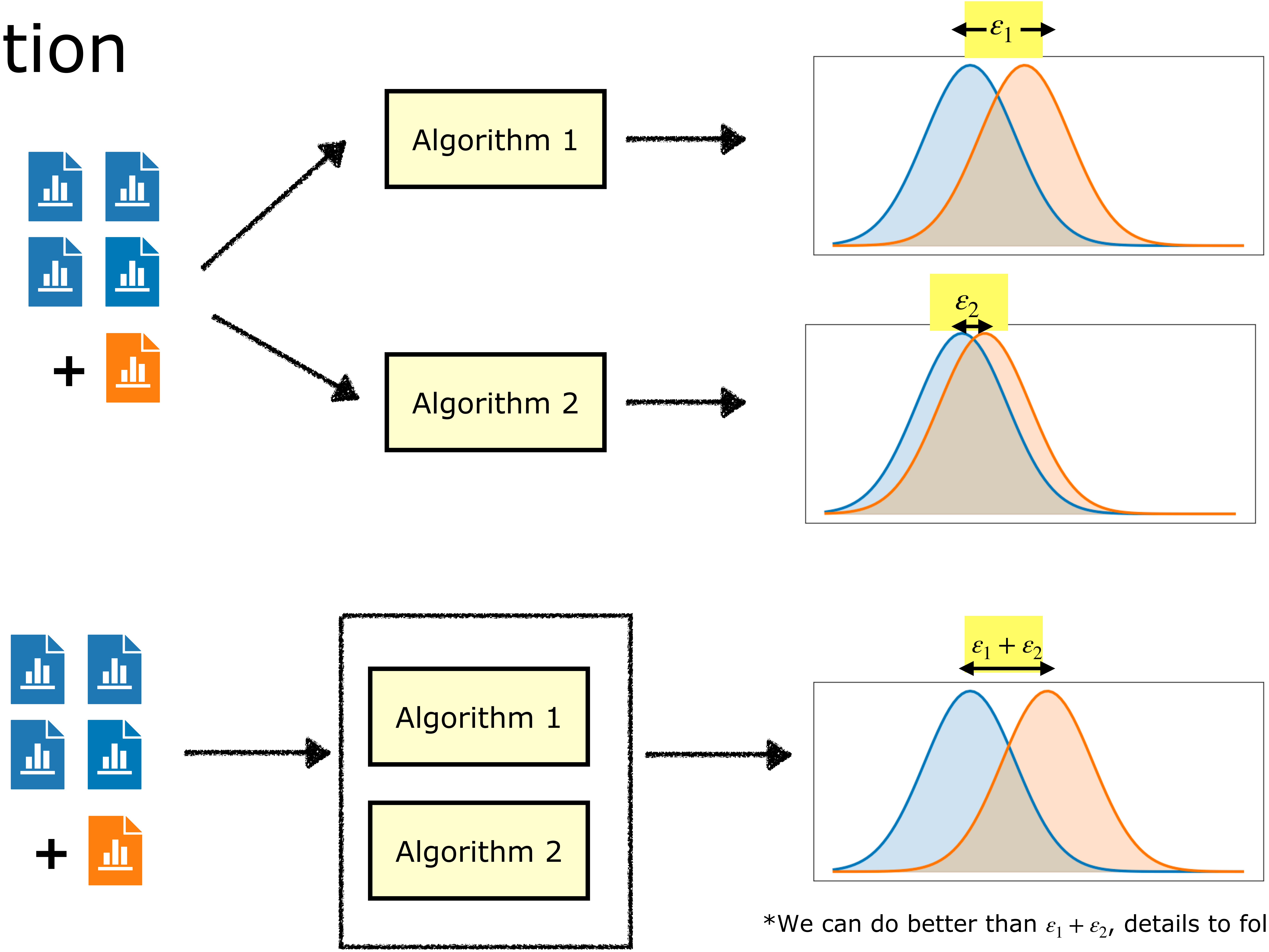
No Differential Privacy ($\varepsilon = \infty$)

Key properties of DP

Composition over multiple steps

Post-processing

Composition



Algorithm 1

Algorithm 2

Algorithm 1

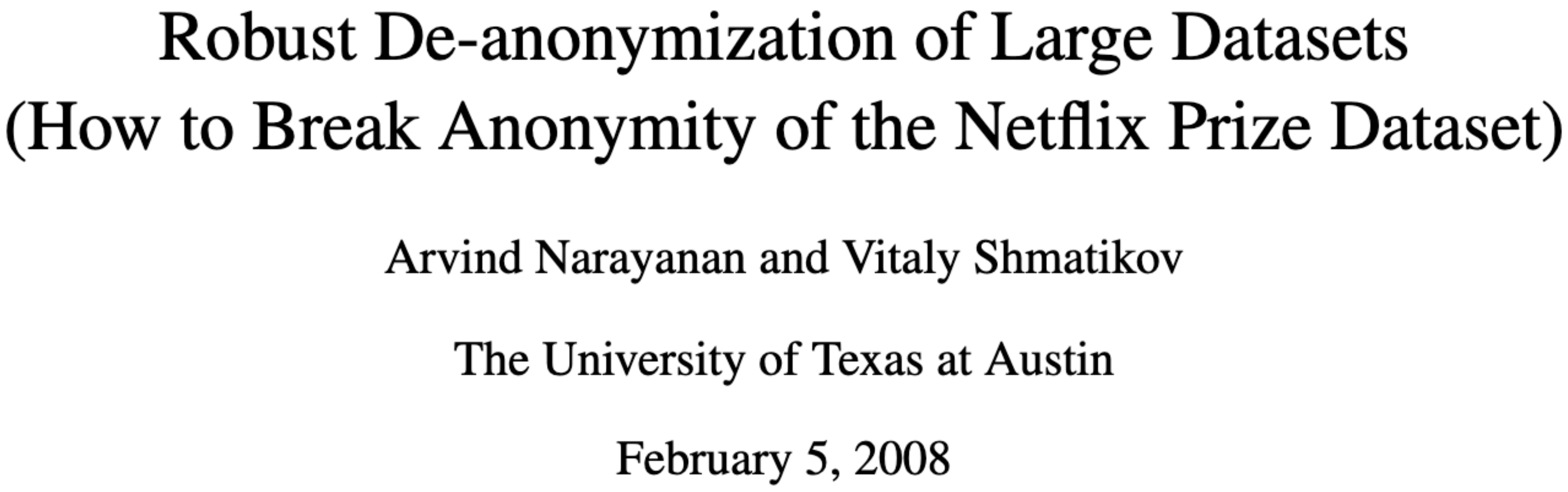
Algorithm 2

*We can do better than $\varepsilon_1 + \varepsilon_2$, details to follow

Why is composition necessary?

2006 - 2009

$1M to beat Netflix’s recommendation algorithm by 10%

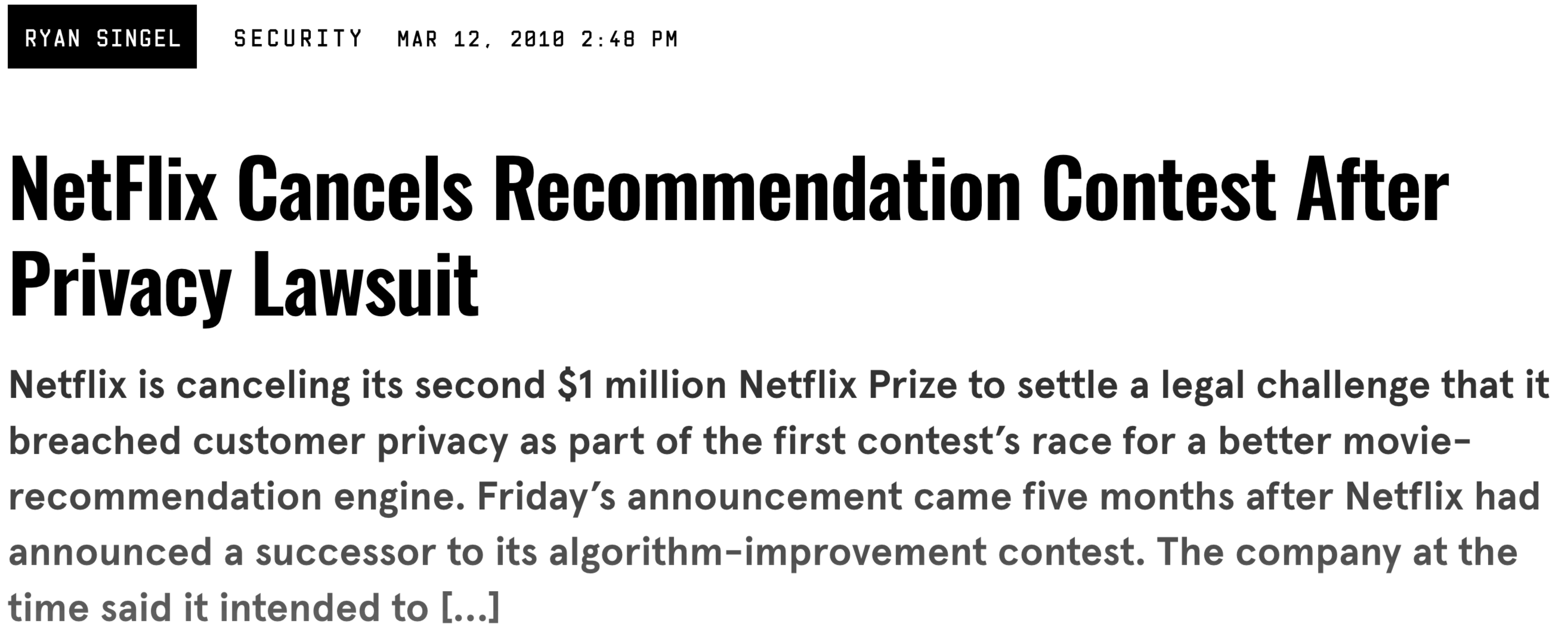


Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008



RYAN SINGEL SECURITY MAR 12, 2010 2:48 PM

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest’s race for a better movie-recommendation engine. Friday’s announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to [...]

Why is composition necessary?

2006 - 2009

$1M to beat Netflix’s recommendation algorithm by 10%

Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

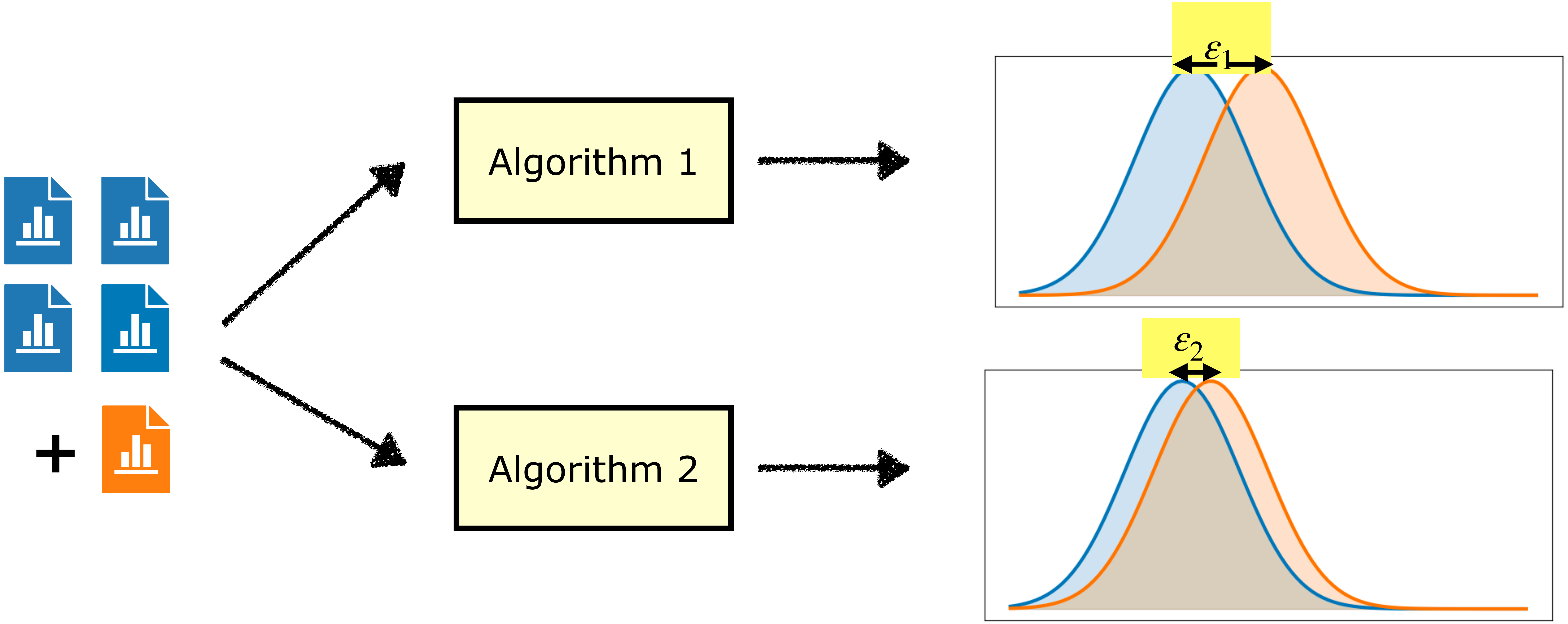
Composition prevents such leakage!

RYAN SINGEL SECURITY MAR 12, 2010 2:48 PM

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest’s race for a better movie-recommendation engine. Friday’s announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to […]

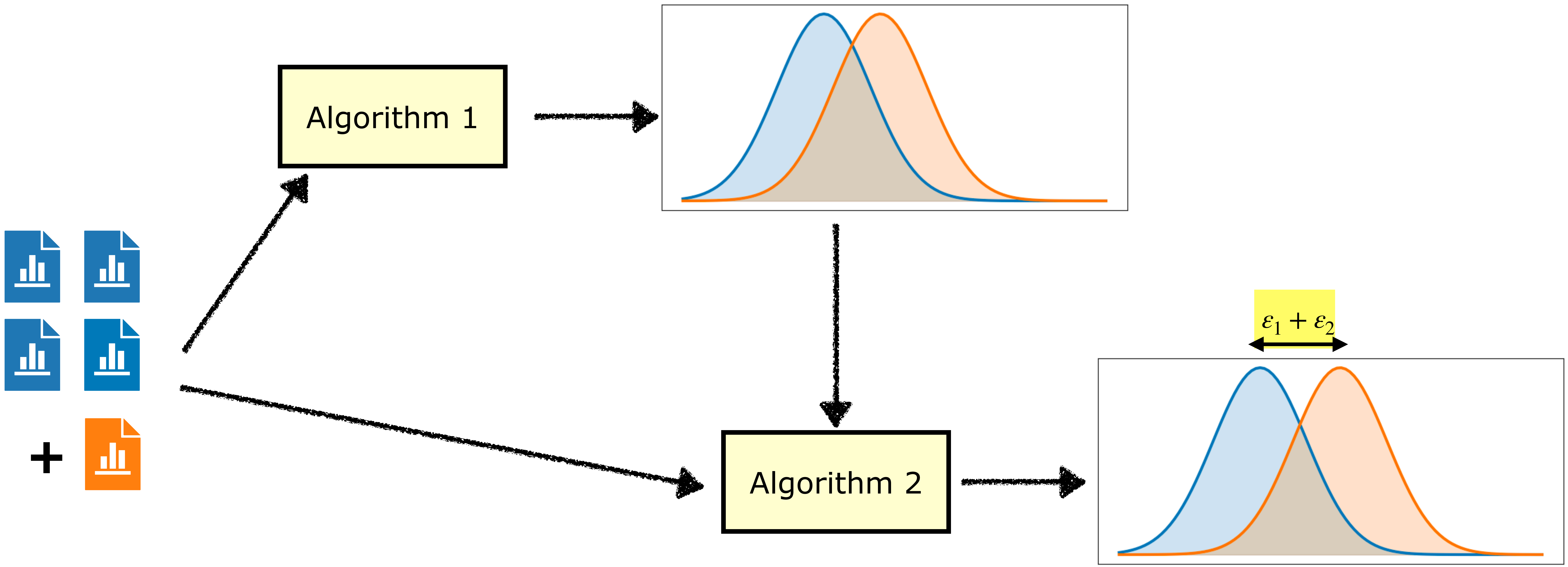
Adaptive Composition



Algorithm 1

Algorithm 2

Adaptive Composition



Algorithm 1

Algorithm 2

$\varepsilon_1 + \varepsilon_2$

*We can do better than $\varepsilon_1 + \varepsilon_2$, details to follow

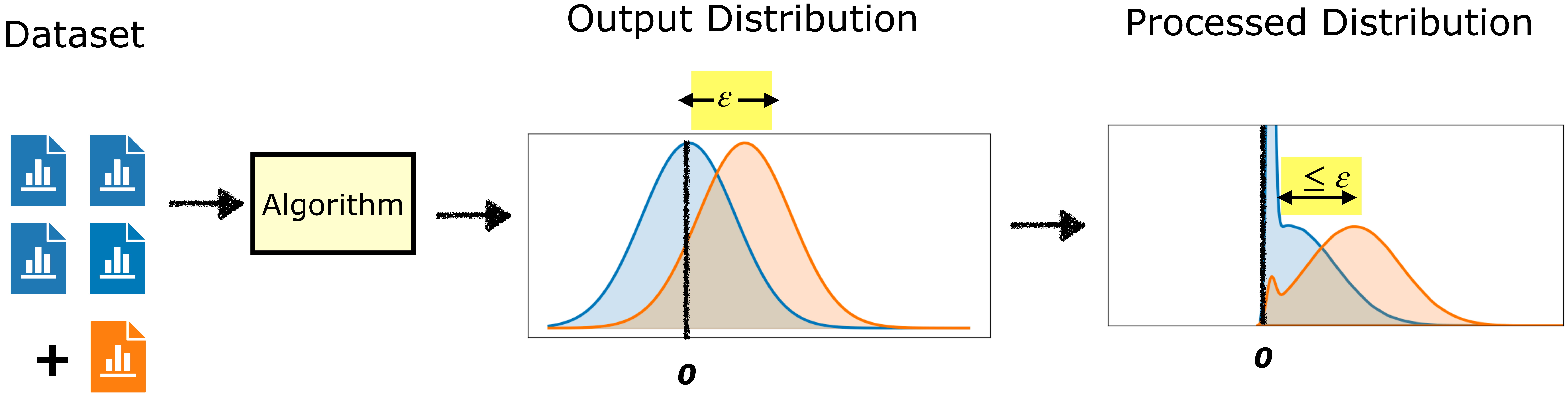
Key properties of DP

Composition over multiple steps

Post-processing

Post-processing

Example: How many people at IITM have a certain medical condition?



Dataset

Algorithm

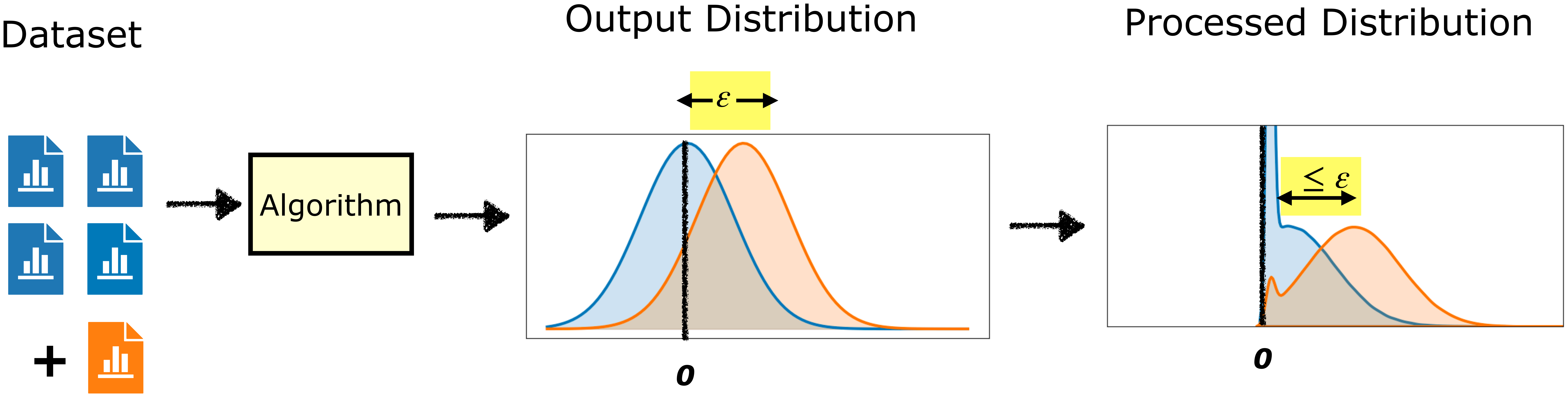
Output Distribution

Processed Distribution

"Information cannot be created"

Post-processing

LLM Example: Stage 1 - Private Training / Stage 2 - alignment



Dataset

Algorithm

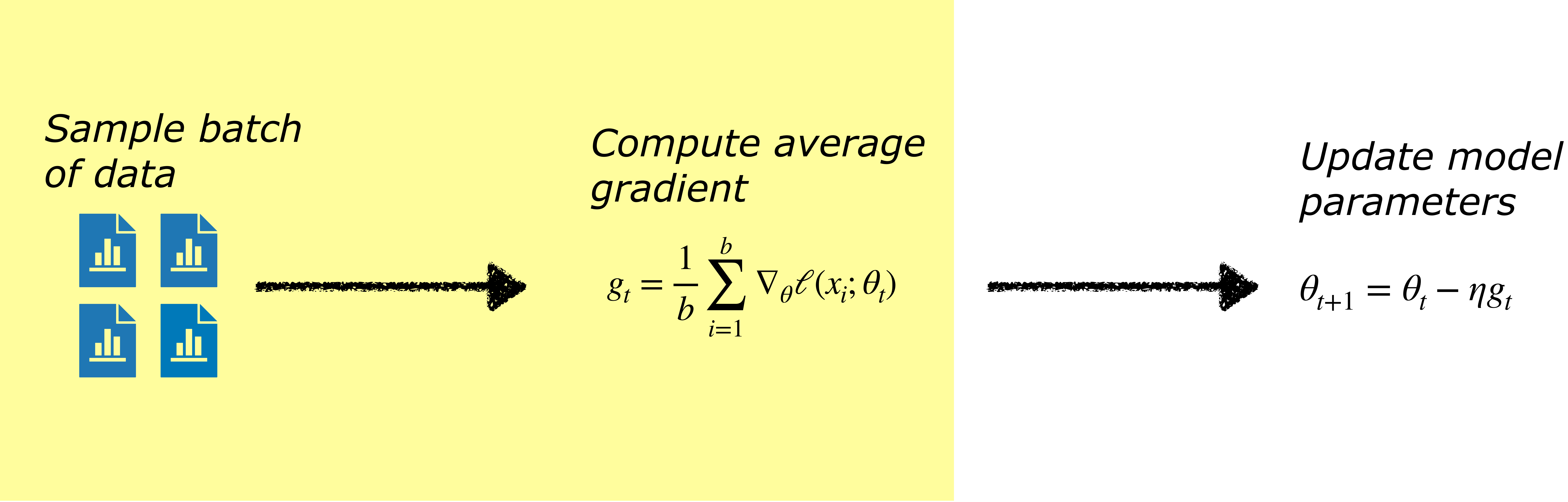
Output Distribution

Processed Distribution

"Information cannot be created"

Our main motivation: Deep Learning with DP

Review: Stochastic gradient descent (without DP)



Sample batch of data

Compute average gradient

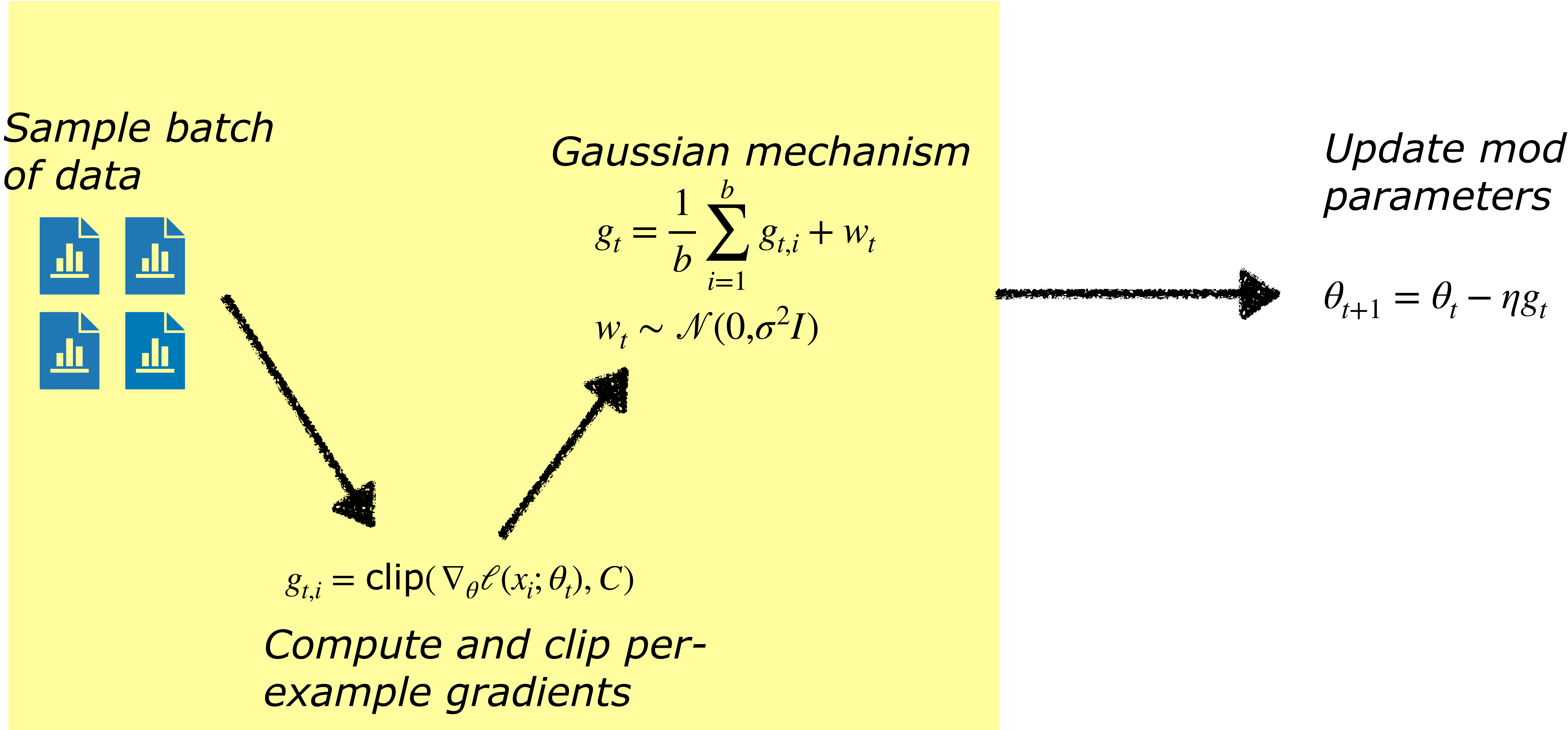
$$g_t = \frac{1}{b} \sum_{i=1}^{b} \nabla_\theta \ell(x_i; \theta_t)$$

Update model parameters

$$\theta_{t+1} = \theta_t - \eta g_t$$

DP-SGD: Stochastic gradient descent with DP

Gradient is differentially private



Sample batch of data

Gaussian mechanism

$$g_t = \frac{1}{b}\sum_{i=1}^{b} g_{t,i} + w_t$$

$$w_t \sim \mathcal{N}(0, \sigma^2 I)$$

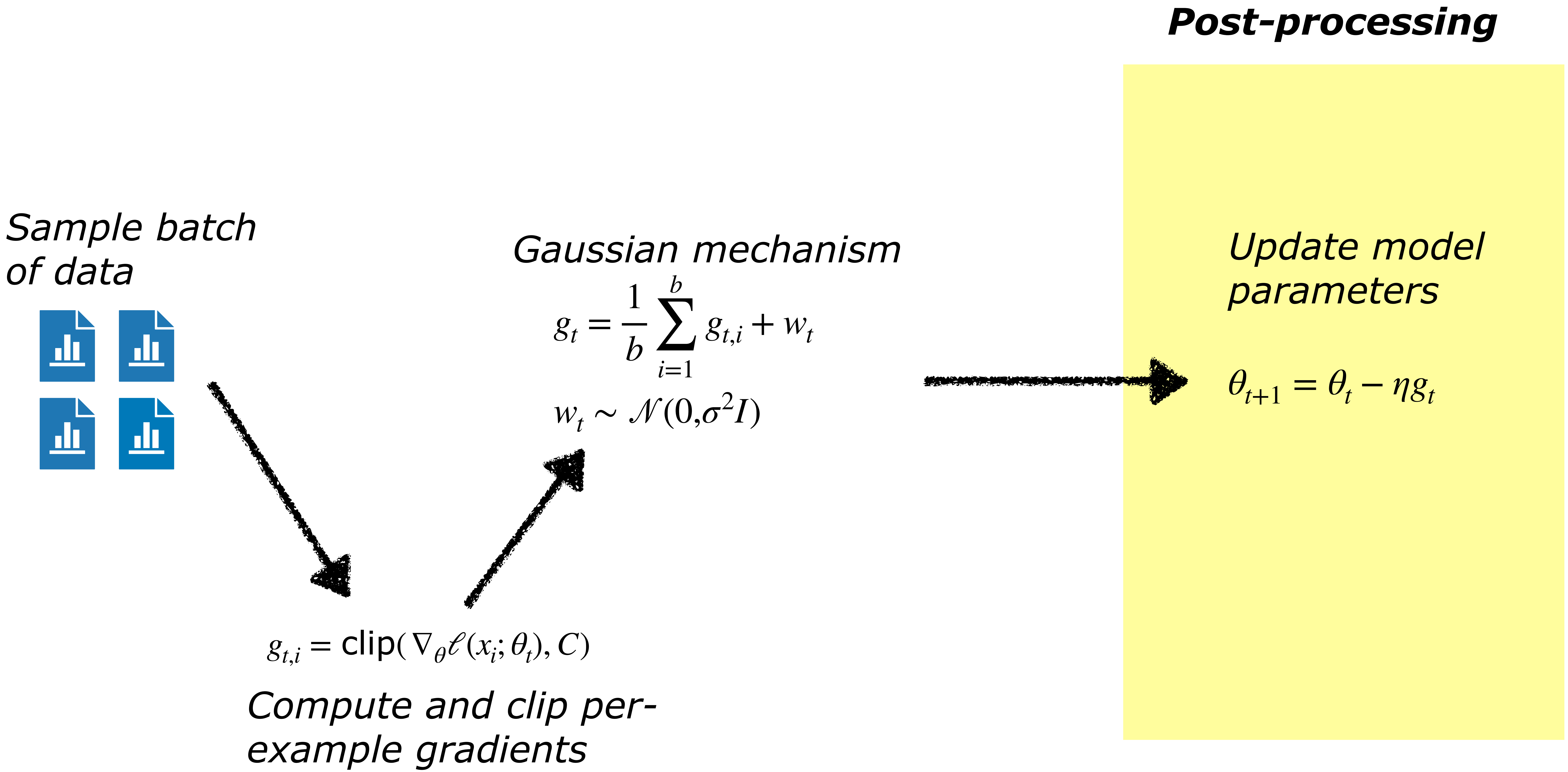
$$g_{t,i} = \text{clip}(\nabla_\theta \ell(x_i; \theta_t), C)$$

Compute and clip per-example gradients

Update model parameters

$$\theta_{t+1} = \theta_t - \eta g_t$$

DP-SGD: Stochastic gradient descent with DP



Post-processing

Sample batch of data

Gaussian mechanism

$$g_t = \frac{1}{b}\sum_{i=1}^{b} g_{t,i} + w_t$$

$$w_t \sim \mathcal{N}(0,\sigma^2 I)$$

Update model parameters

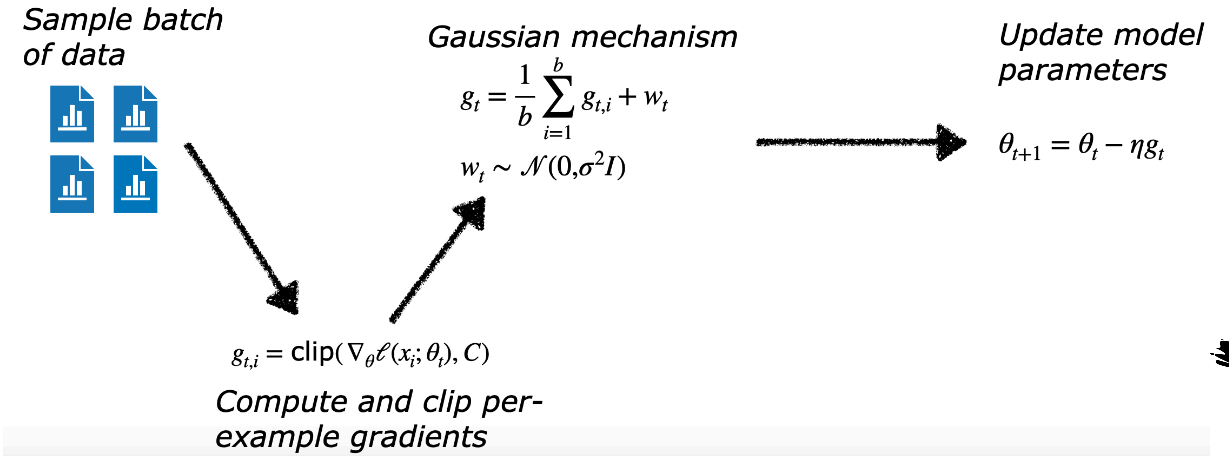
$$\theta_{t+1} = \theta_t - \eta g_t$$

$$g_{t,i} = \text{clip}(\nabla_\theta \ell(x_i;\theta_t), C)$$

Compute and clip per-example gradients

DP-SGD: Stochastic gradient descent with DP

Iteration 1



Sample batch of data

Gaussian mechanism

$g_t = \frac{1}{b}\sum_{i=1}^{b} g_{t,i} + w_t$

$w_t \sim \mathcal{N}(0,\sigma^2 I)$

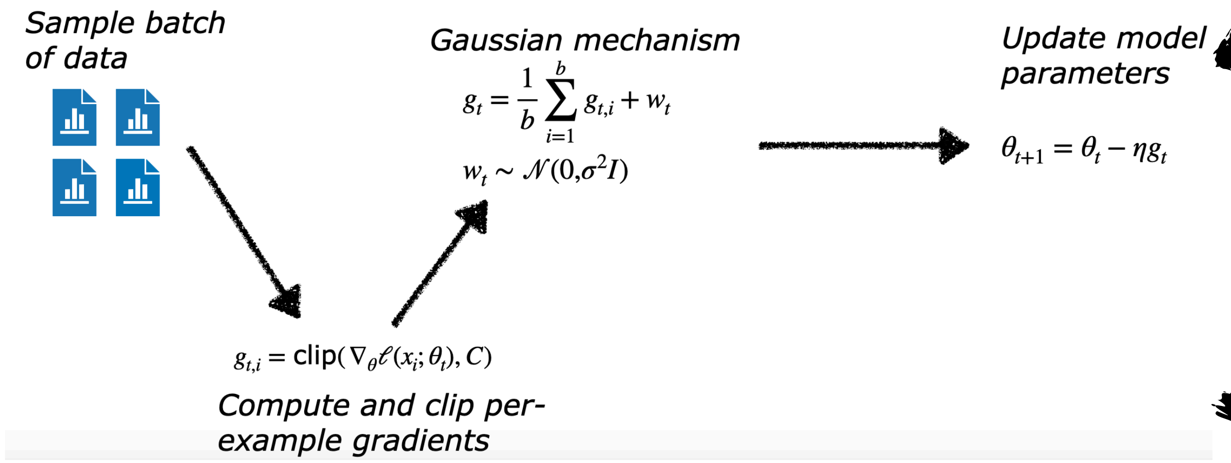
Update model parameters

$\theta_{t+1} = \theta_t - \eta g_t$

$g_{t,i} = \text{clip}(\nabla_\theta \ell(x_i;\theta_t), C)$

Compute and clip per-example gradients

Iteration 2



Sample batch of data

Gaussian mechanism

$g_t = \frac{1}{b}\sum_{i=1}^{b} g_{t,i} + w_t$

$w_t \sim \mathcal{N}(0,\sigma^2 I)$

Update model parameters

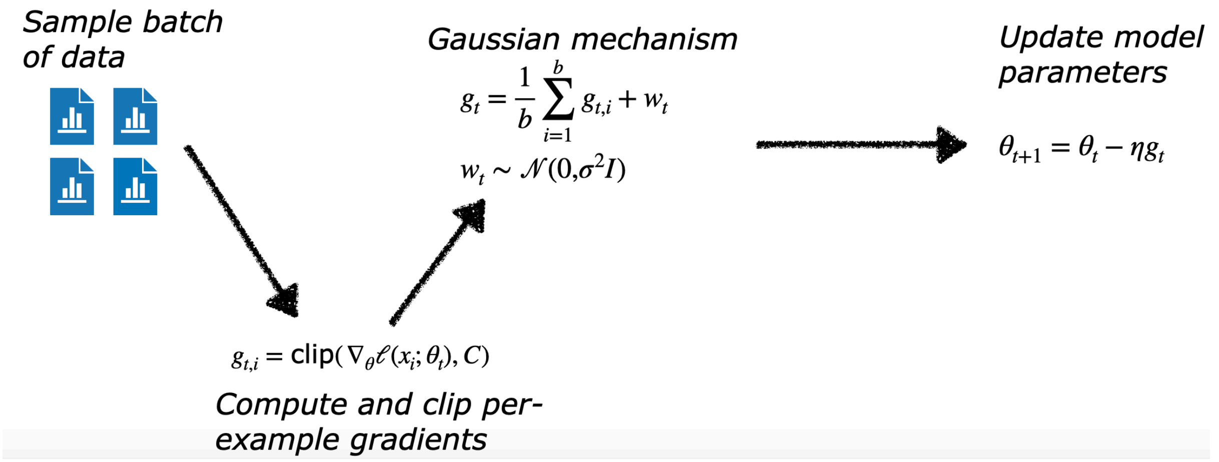
$\theta_{t+1} = \theta_t - \eta g_t$

$g_{t,i} = \text{clip}(\nabla_\theta \ell(x_i;\theta_t), C)$

Compute and clip per-example gradients

⋮

Iteration T



Sample batch of data

Gaussian mechanism

$g_t = \frac{1}{b}\sum_{i=1}^{b} g_{t,i} + w_t$

$w_t \sim \mathcal{N}(0,\sigma^2 I)$

Update model parameters

$\theta_{t+1} = \theta_t - \eta g_t$

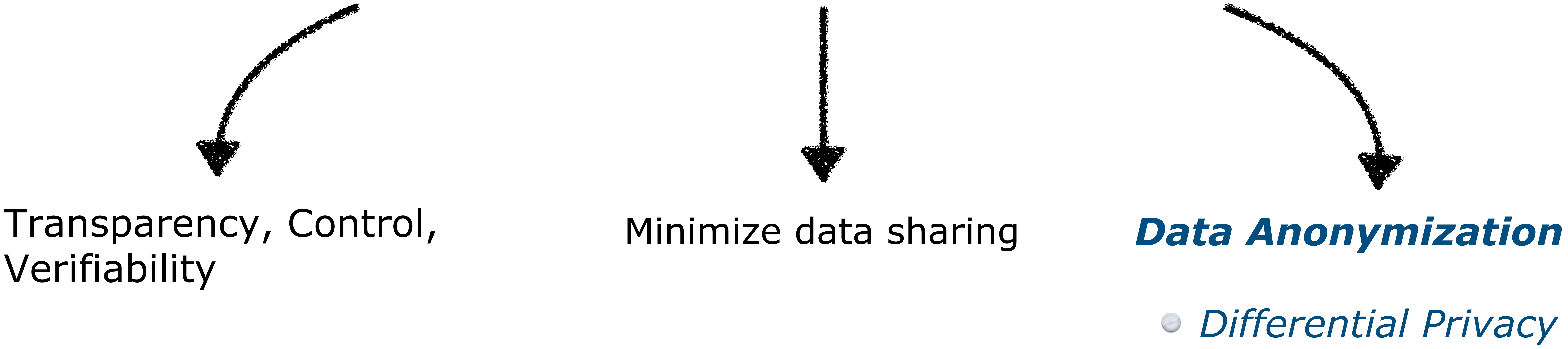
$g_{t,i} = \text{clip}(\nabla_\theta \ell(x_i;\theta_t), C)$

Compute and clip per-example gradients

Adaptive Composition!

Caveat: Multiple facets of the word "***privacy***"

What does the word "***privacy***" mean to an end user of an AI product?



Transparency, Control, Verifiability

Minimize data sharing

Data Anonymization

- *Differential Privacy*

Bonawitz, Kairouz, McMahan, Ramage (2022). **Federated Learning and Privacy**. *Communications of the ACM*.

Tentative Course Outline

Part 1: Weeks 2-4

- Correlated noise mechanisms:
    - Multiple epochs
    - Amplification

Recall: DP Training with ***Correlated*** Noise

Update model parameters

$$\theta_{t+1} = \theta_t - \eta(g_t + z_t)$$

i.i.d. Gaussian noise

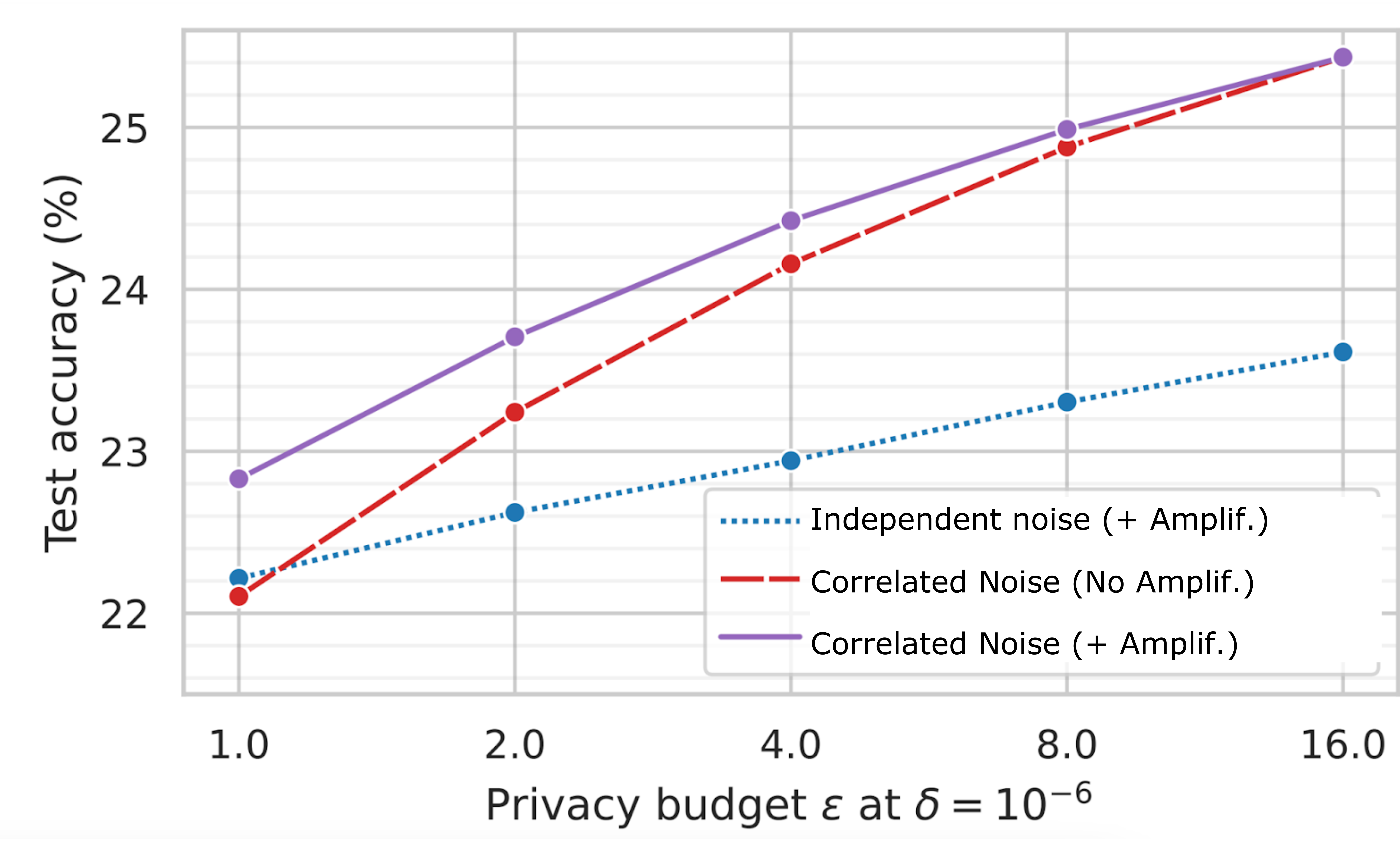
$$z_t \sim \mathcal{N}(0, \sigma^2 I)$$

$$\theta_{t+1} = \theta_t - \eta\Big(g_t + \underbrace{z_t - \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}}_{=: w_t}\Big)$$

(Anti-)correlated Gaussian noise

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.
Denisov, McMahan, Rush, Smith, Thakurta. **Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams**. NeurIPS 2022.

Correlated noise uniformly beats independent noise



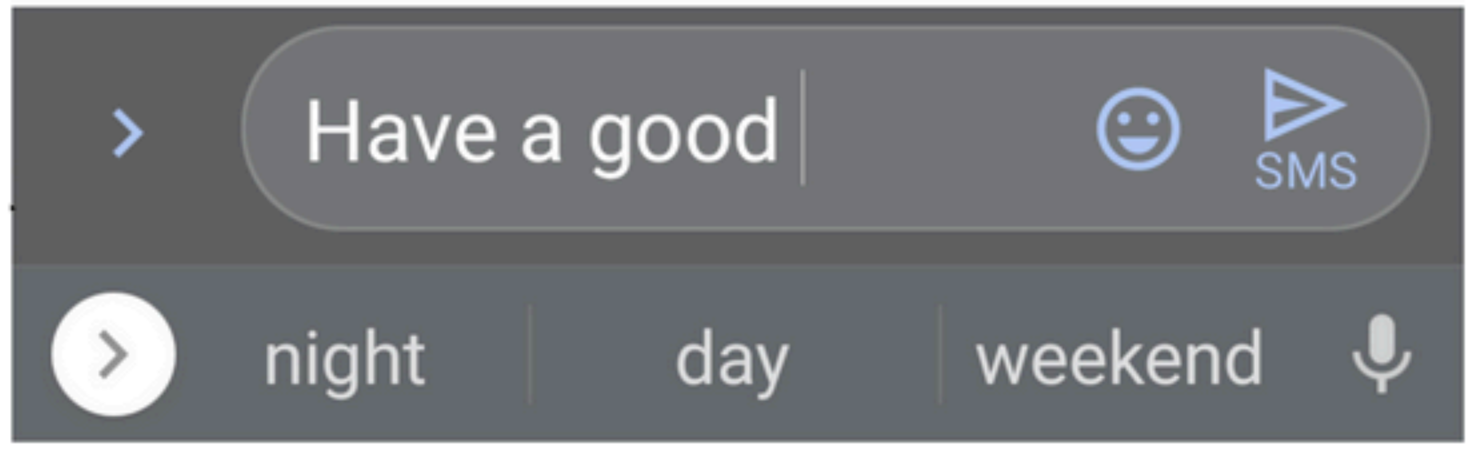
Task: Language modelling

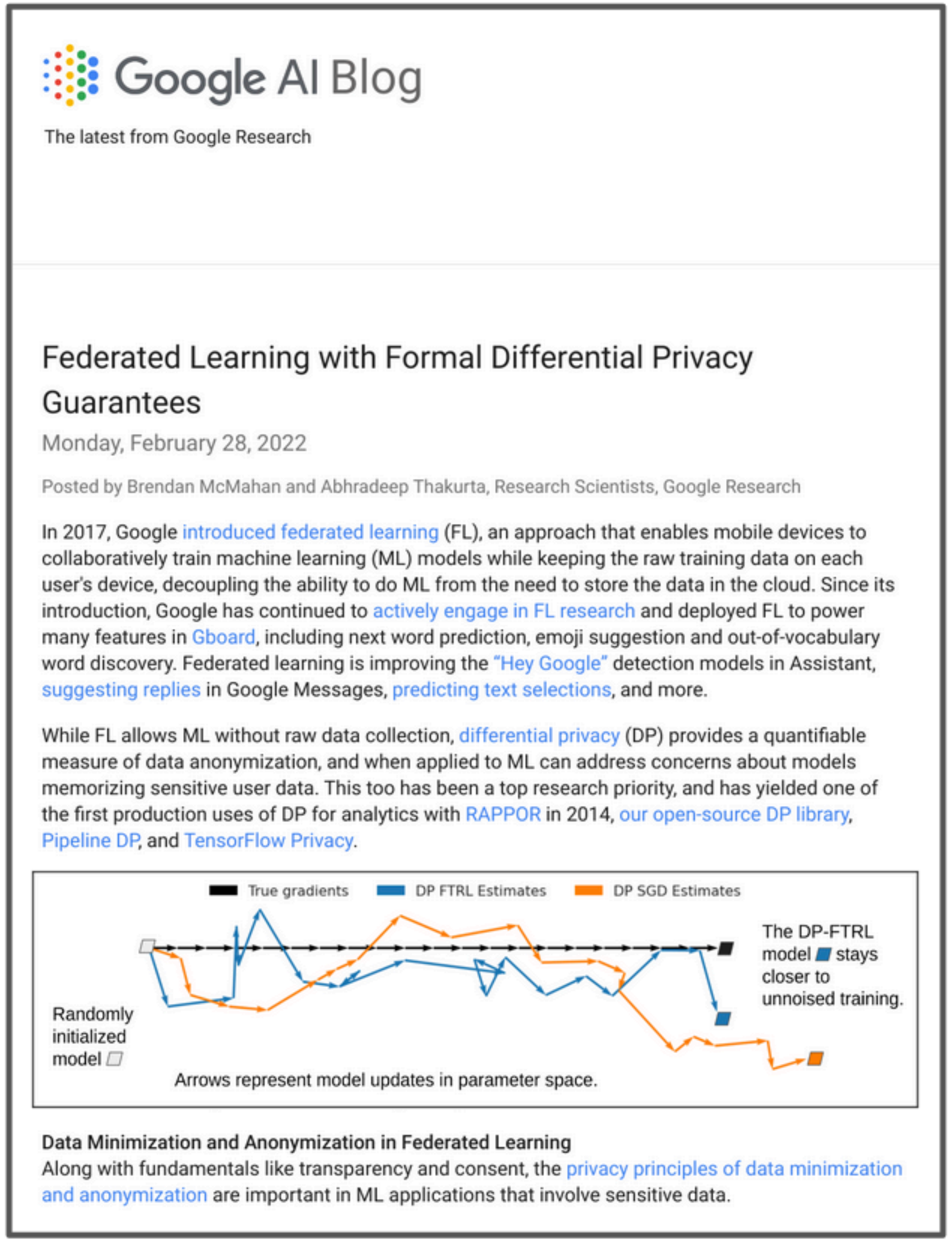
Dataset: StackOverflow

Choquette-Choo, Ganesh, McKenna, McMahan, Rush, Thakurta, Xu.
(Amplified) Banded Matrix Factorization: A unified approach to private training. NeurIPS 2023

Production Training

> *"the first production neural network trained directly on user data announced with a formal DP guarantee."*
>
> \- Google AI Blog post, Feb 2022


Have a good


Google AI Blog

The latest from Google Research

Federated Learning with Formal Differential Privacy Guarantees

Monday, February 28, 2022

Posted by Brendan McMahan and Abhradeep Thakurta, Research Scientists, Google Research

In 2017, Google introduced federated learning (FL), an approach that enables mobile devices to collaboratively train machine learning (ML) models while keeping the raw training data on each user's device, decoupling the ability to do ML from the need to store the data in the cloud. Since its introduction, Google has continued to actively engage in FL research and deployed FL to power many features in Gboard, including next word prediction, emoji suggestion and out-of-vocabulary word discovery. Federated learning is improving the "Hey Google" detection models in Assistant, suggesting replies in Google Messages, predicting text selections, and more.

While FL allows ML without raw data collection, differential privacy (DP) provides a quantifiable measure of data anonymization, and when applied to ML can address concerns about models memorizing sensitive user data. This too has been a top research priority, and has yielded one of the first production uses of DP for analytics with RAPPOR in 2014, our open-source DP library, Pipeline DP, and TensorFlow Privacy.

True gradients
DP FTRL Estimates
DP SGD Estimates
The DP-FTRL model stays closer to unnoised training.
Randomly initialized model
Arrows represent model updates in parameter space.

Data Minimization and Anonymization in Federated Learning

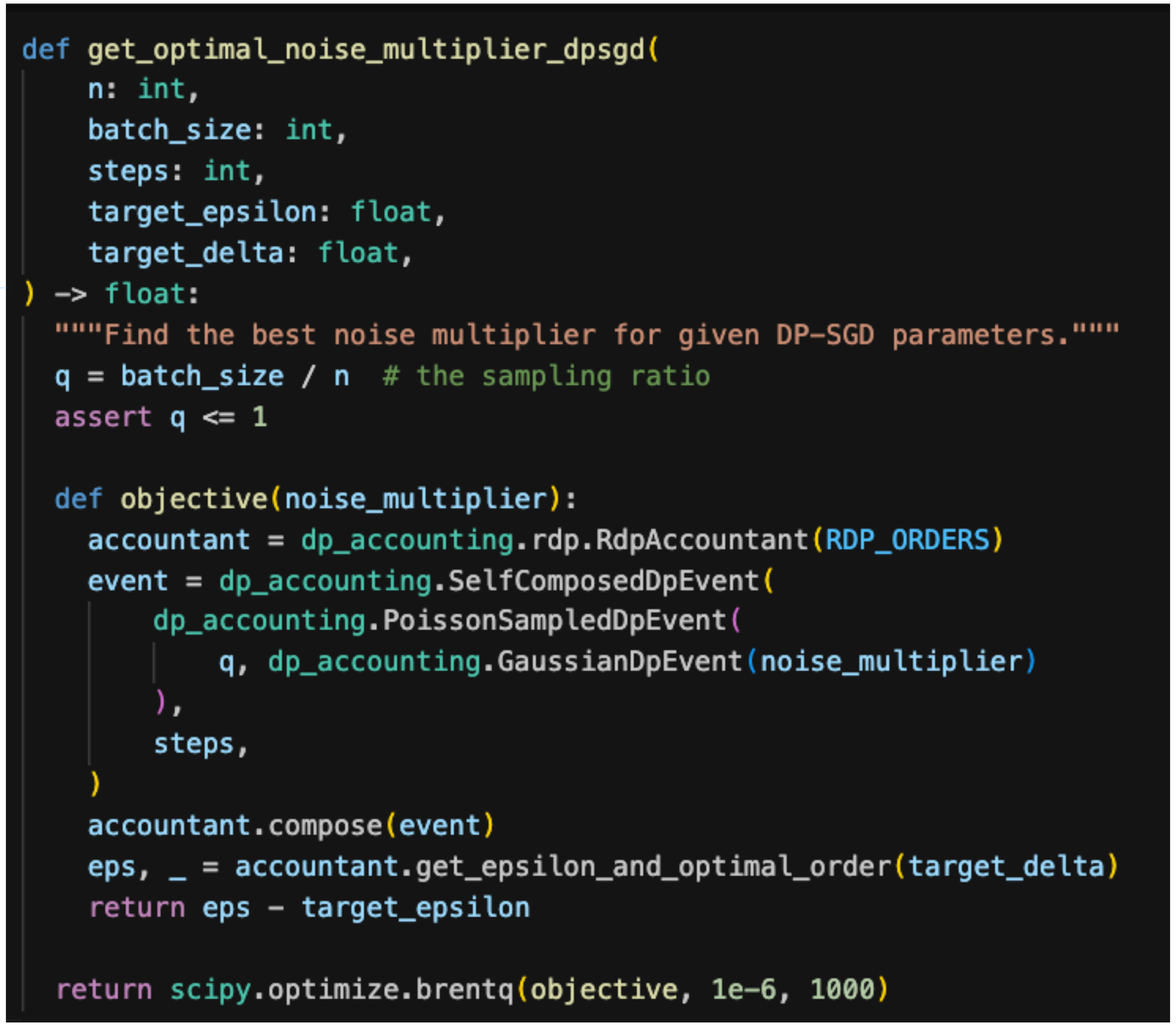
Along with fundamentals like transparency and consent, the privacy principles of data minimization and anonymization are important in ML applications that involve sensitive data.

Part 1: Weeks 2-4

- Correlated noise mechanisms:
    - Multiple epochs
    - Amplification

Part 2: Weeks 4-5

DP-SGD noise multiplier is ***independent*** of the data



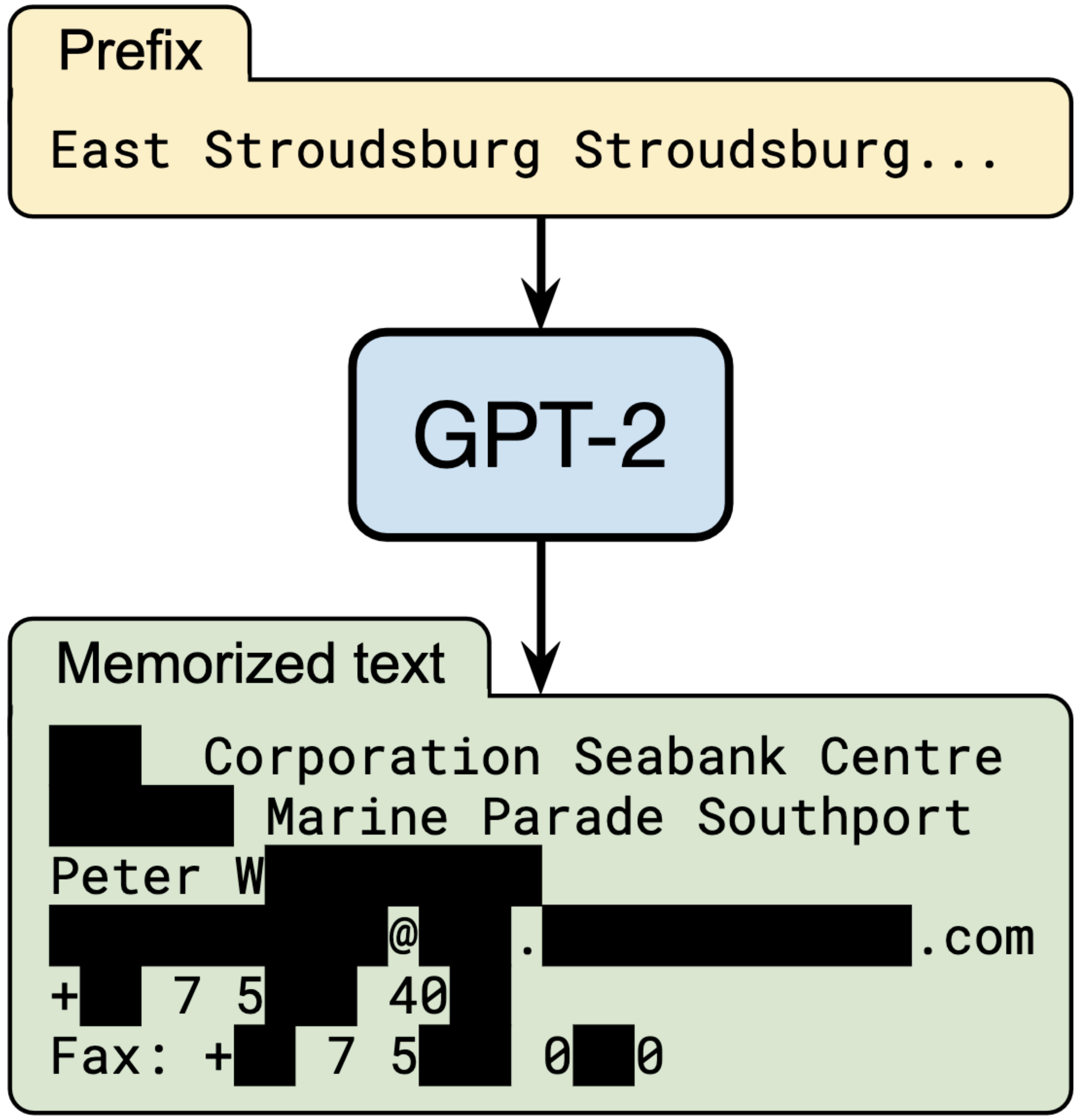
```
def get_optimal_noise_multiplier_dpsgd(
    n: int,
    batch_size: int,
    steps: int,
    target_epsilon: float,
    target_delta: float,
) -> float:
  """Find the best noise multiplier for given DP-SGD parameters."""
  q = batch_size / n  # the sampling ratio
  assert q <= 1

  def objective(noise_multiplier):
    accountant = dp_accounting.rdp.RdpAccountant(RDP_ORDERS)
    event = dp_accounting.SelfComposedDpEvent(
        dp_accounting.PoissonSampledDpEvent(
            q, dp_accounting.GaussianDpEvent(noise_multiplier)
        ),
        steps,
    )
    accountant.compose(event)
    eps, _ = accountant.get_epsilon_and_optimal_order(target_delta)
    return eps - target_epsilon

  return scipy.optimize.brentq(objective, 1e-6, 1000)
```

- Data-adaptive differential privacy
  - Why can’t it be used for model training?

Part 3: Weeks 6-8:
Protecting Against Data Reconstruction Attacks



Prefix

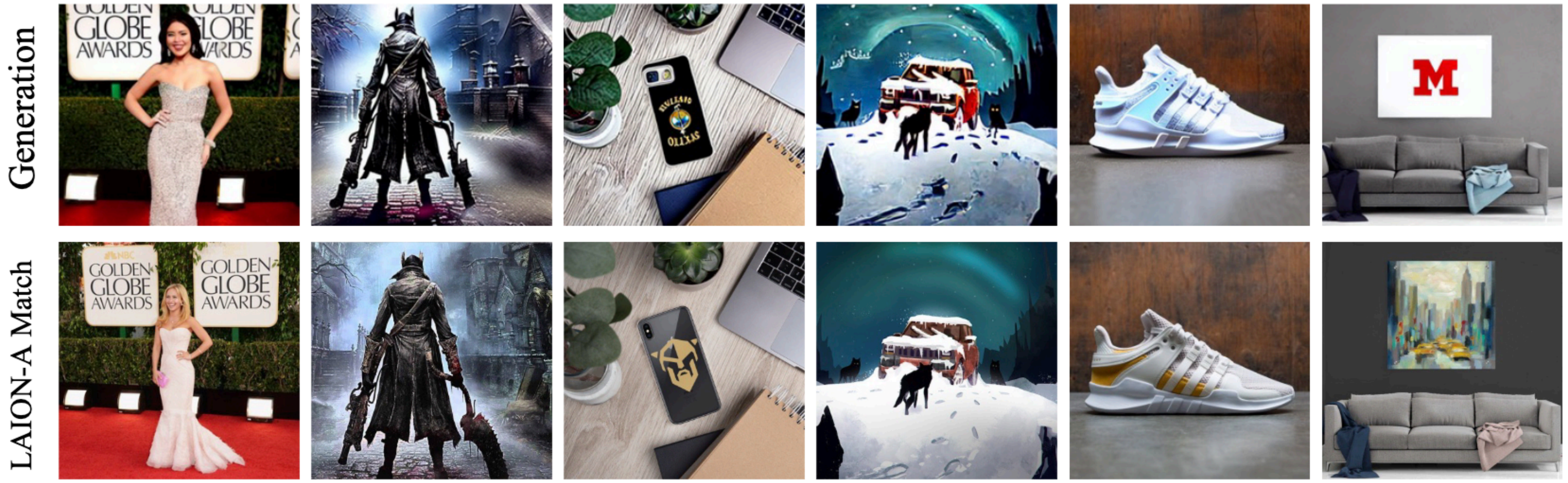
East Stroudsburg Stroudsburg...

GPT-2

Memorized text

[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Generated Images



Real Images
(from training)

Part 3: Weeks 6-8

- Other alternatives to DP: based on info theory
    - Fisher information (last semester)
    - Mutual information

Part 4: Weeks 8-12
GenAI/LLM/Agentic applications

- Model interrogation, privacy risks & solutions for RAG
- Reconstructing data from attention weights
- Diffusion models & implicit privacy in generative models
- Detailed copyright guarantees
- …
- [***your suggestions here***]

Weeks 13-15: Course Projects

- Presentation

Thank you! Questions?