Mathematical Foundations of Differential Privacy

Krishna Pillutla

IIT Madras

CeRAI
Centre for Responsible AI

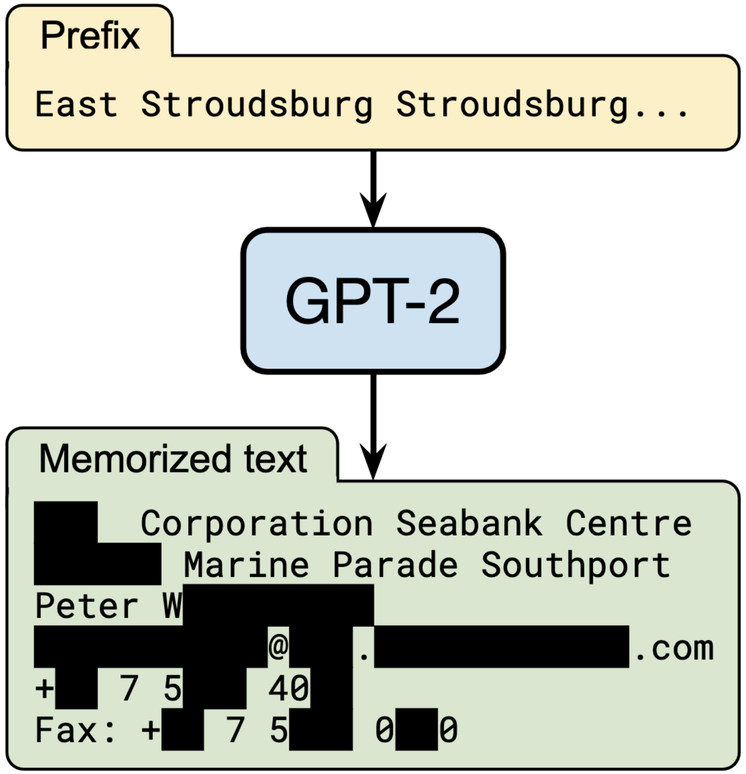
RBC DSAI
ROBERT BOSCH CENTRE FOR DATA SCIENCE AND ARTIFICIAL INTELLIGENCE IIT MADRAS

ACM INDIA SUMMER SCHOOL

RESPONSIBLE & SAFE AI

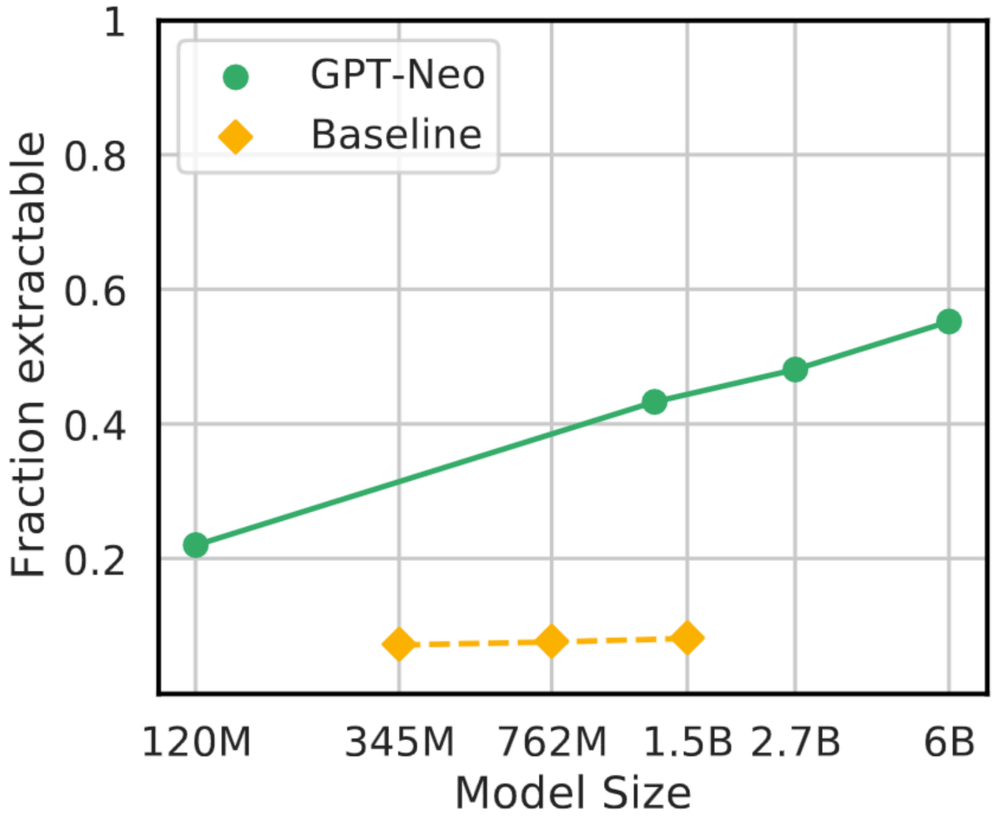
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

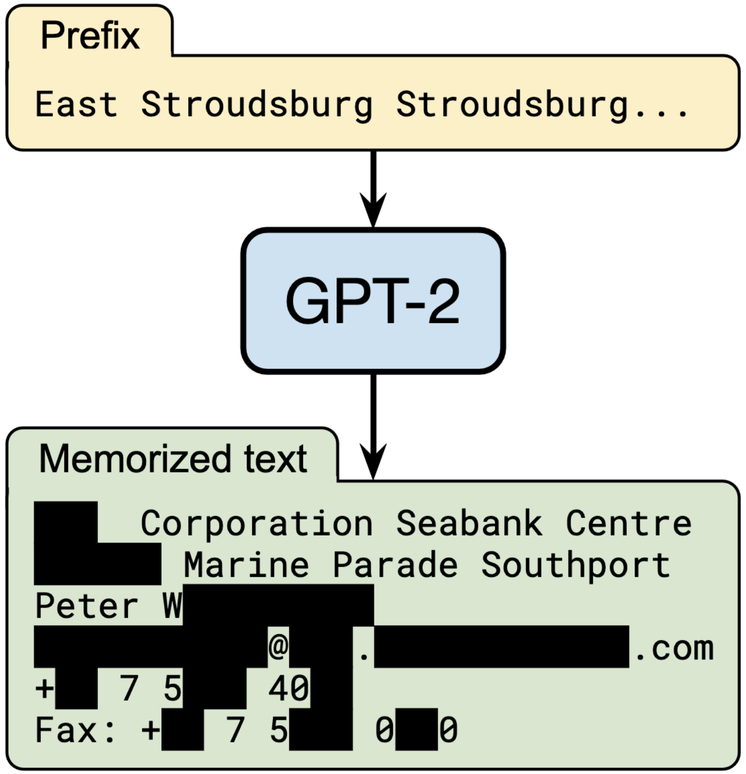
Models leak information about their training data



Carlini et al. (USENIX Security 2021)

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





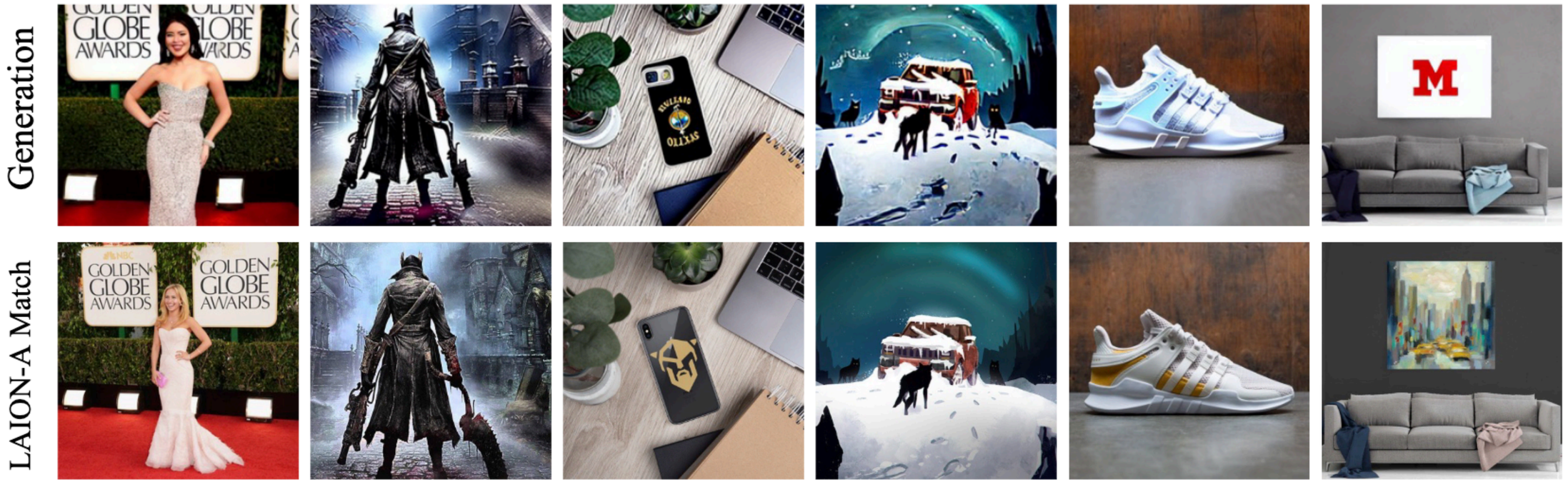
Carlini et al. (USENIX Security 2021)

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

Gemini

Claude

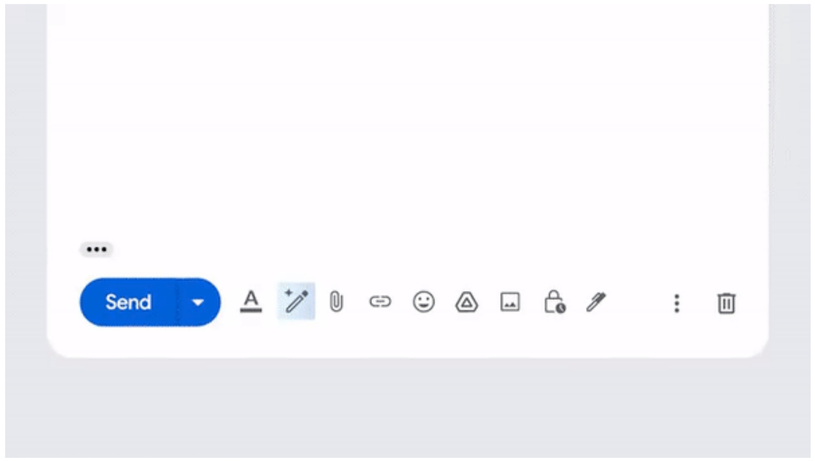
ANTHROP\C

LLAMA 2

MISTRAL AI_

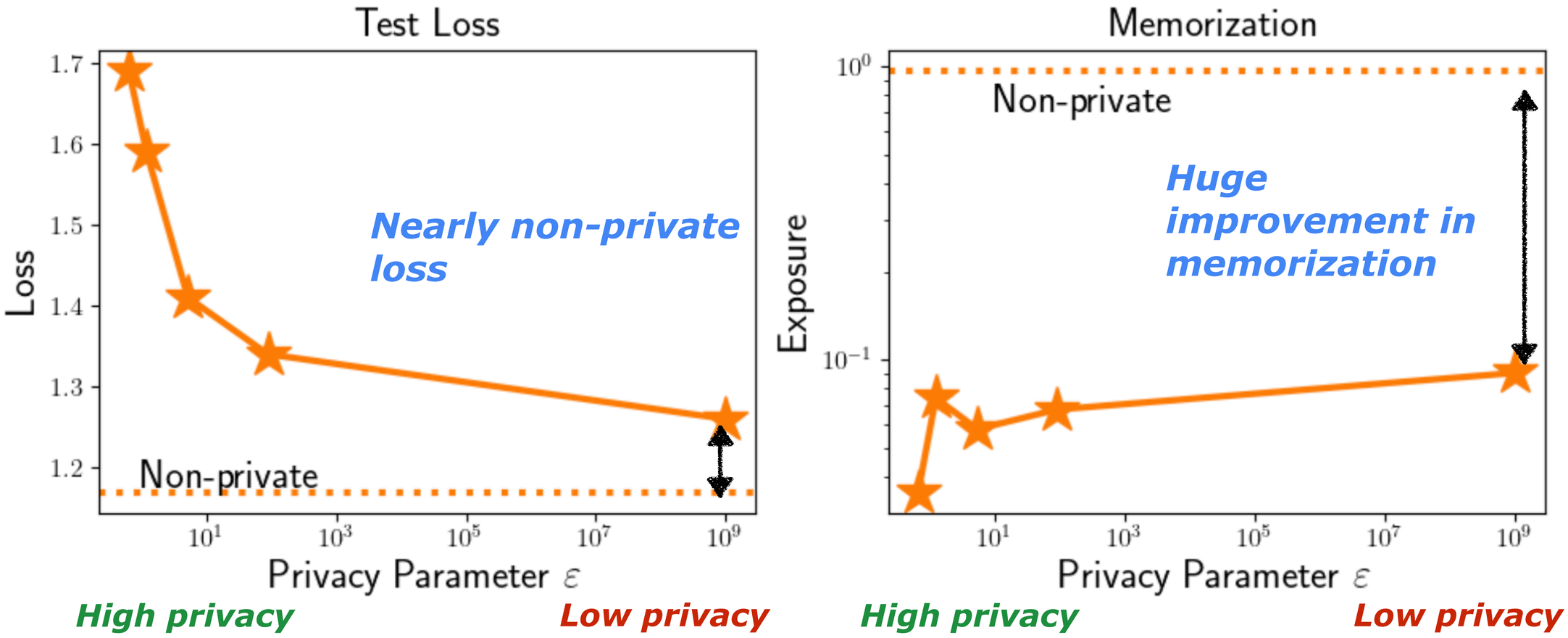
BharatGPT

Gmail



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

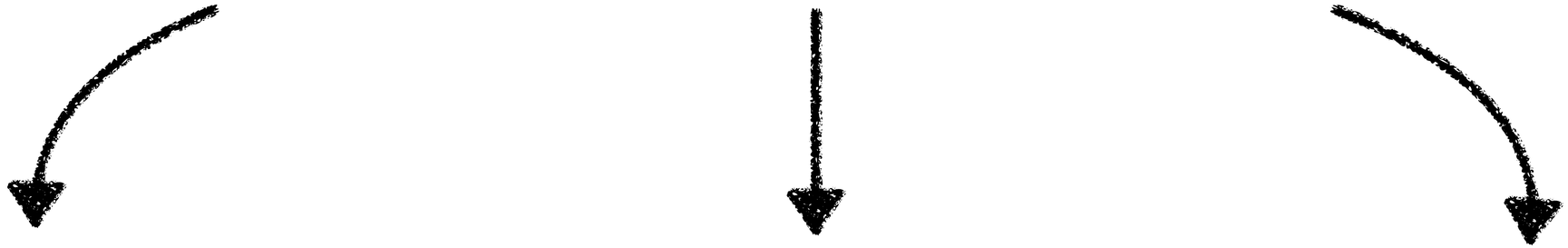
Differential privacy nearly eliminates memorization



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

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*

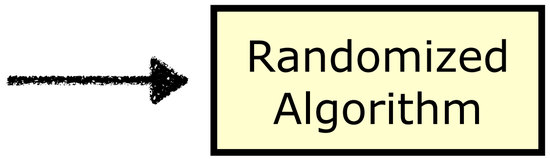
Bonawitz, Kairouz, McMahan, Ramage (2022). **Federated Learning and Privacy**. *Communications of the ACM*.

Outline

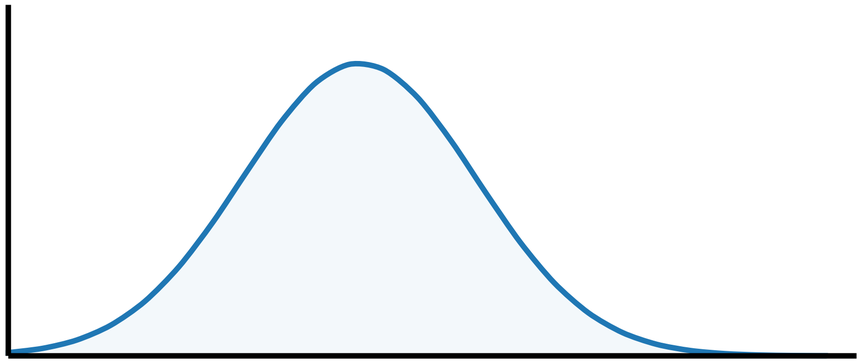
- ***Differential Privacy: Intuition & Recap***
- Differential Privacy: Rigorous Mathematical Formulation & Properties
- Application Examples

Differential Privacy:
A mathematically rigorous notion of "***privacy***"

Dataset

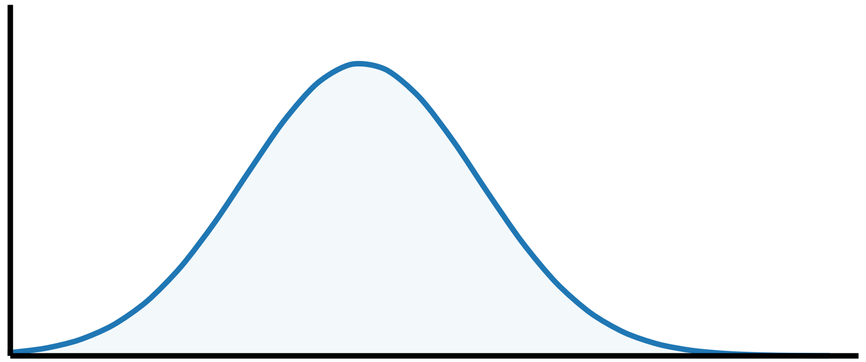


Output Distribution
(e.g. over models)



Dataset

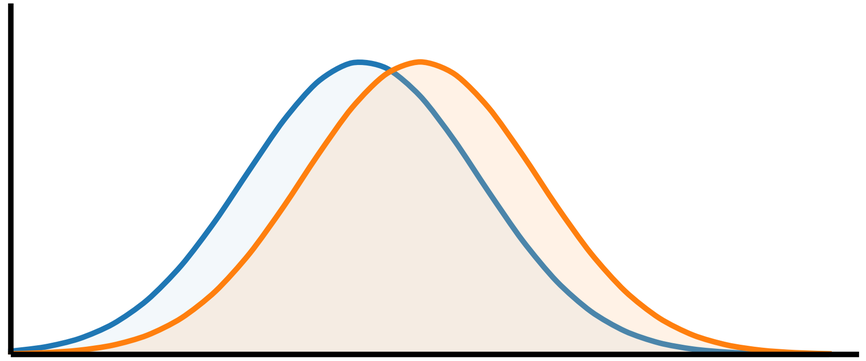
Output Distribution
(e.g. over models)

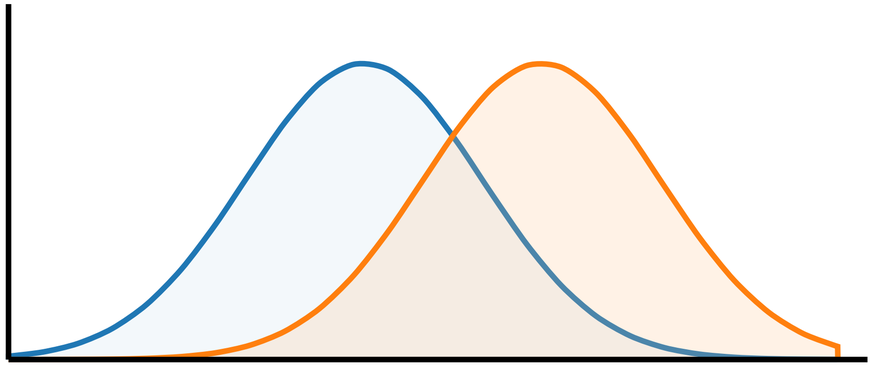


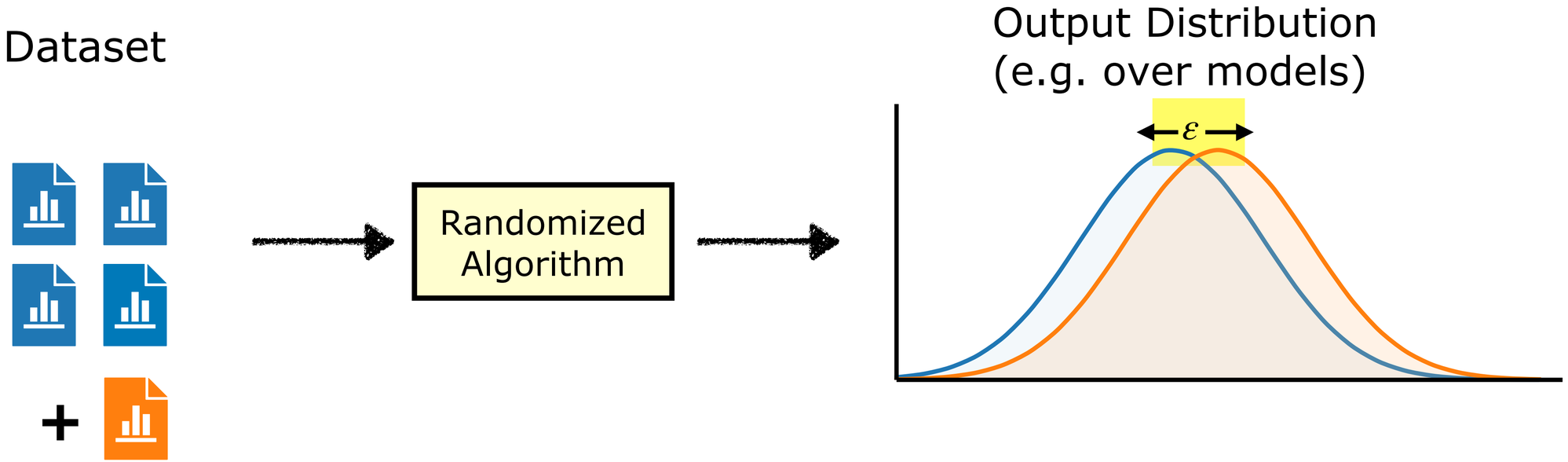
Dataset

Randomized Algorithm

Output Distribution
(e.g. over models)





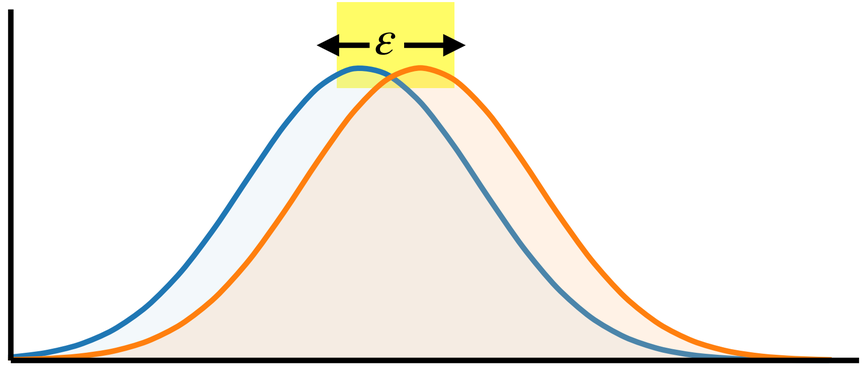


A randomized algorithm is ε-**differentially private** if the addition of **one user's data** does not alter its output distribution by more than ε

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

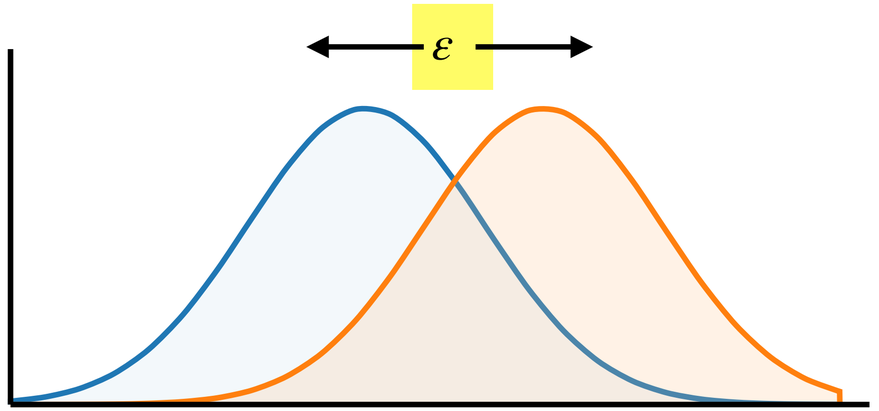
Dataset

Output Distribution
(e.g. over models)



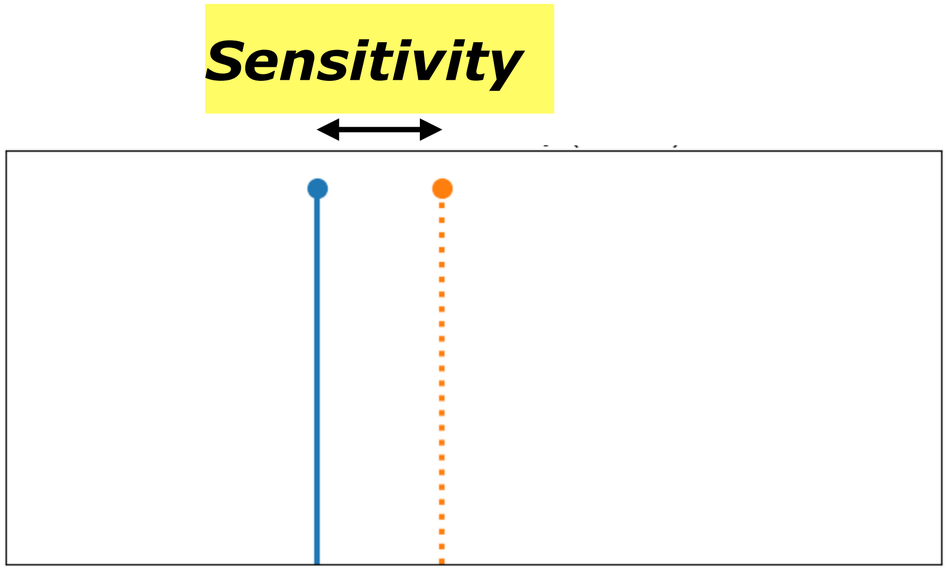
ε-differential privacy

Large $\varepsilon \implies$ more privacy leakage



Adding noise for DP

Dataset

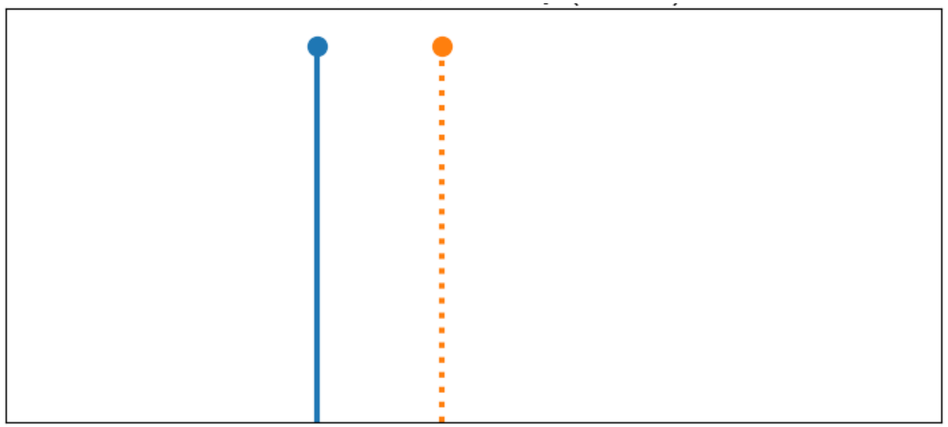


Adding noise for DP

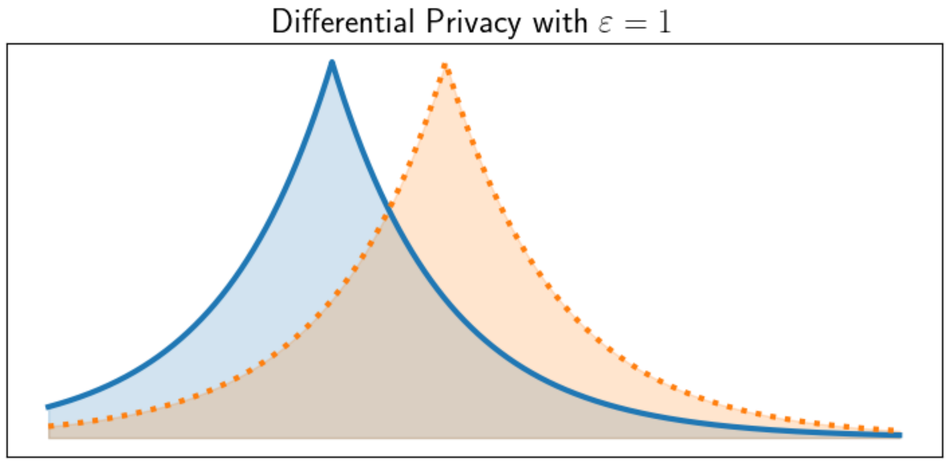
Dataset

Original Algorithm

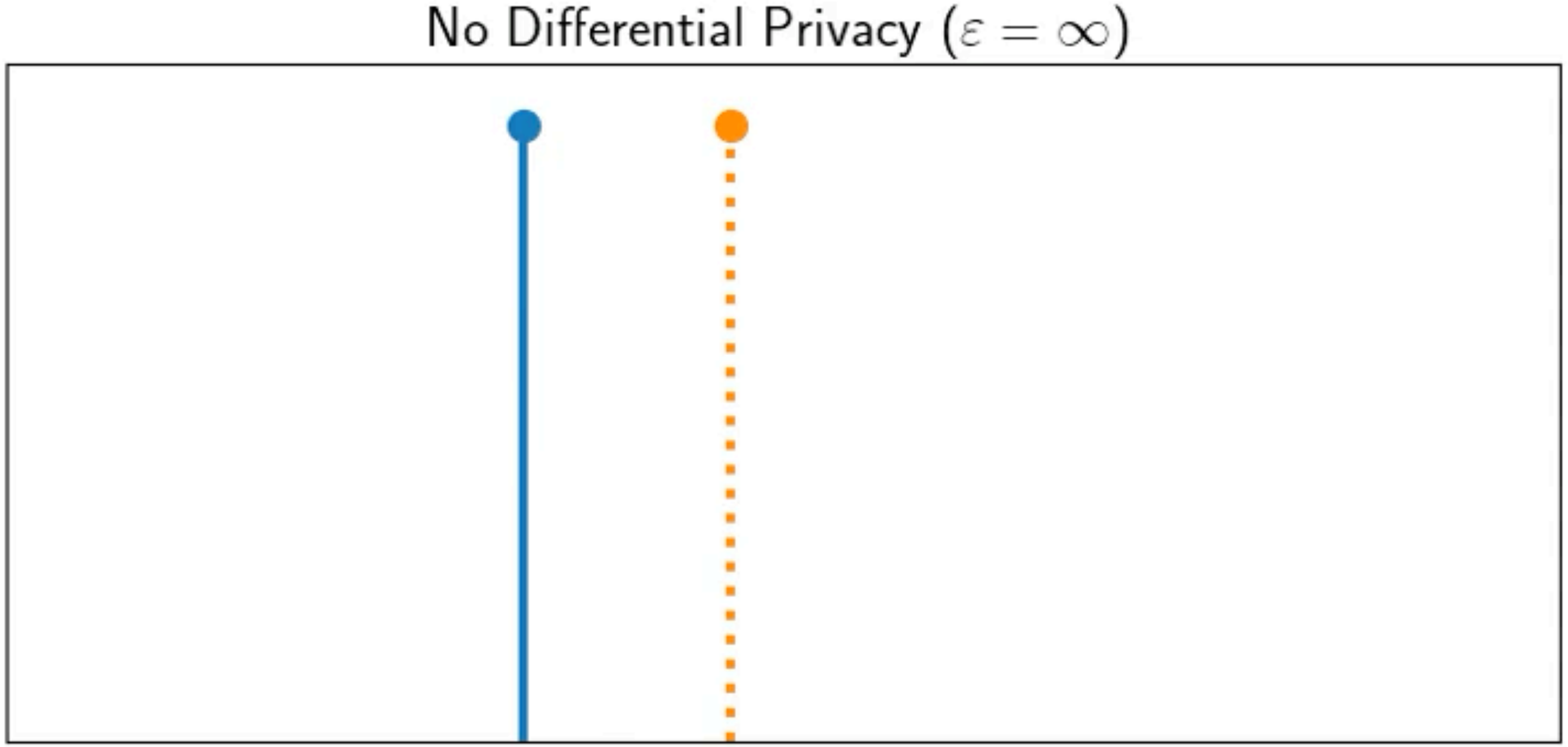
Sensitivity



Add Noise
(e.g. Laplace/Gaussian)



Adding noise for DP



Key properties of DP

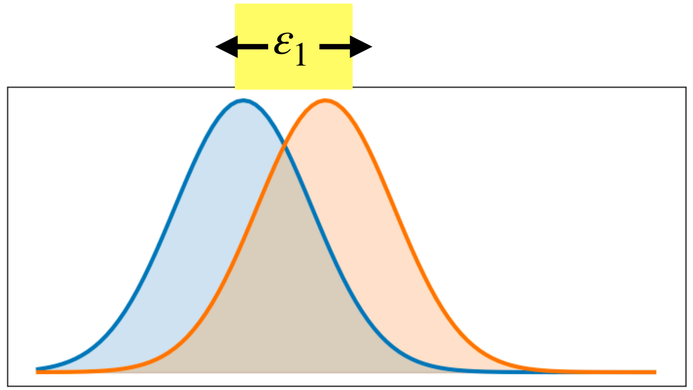
Composition over multiple steps

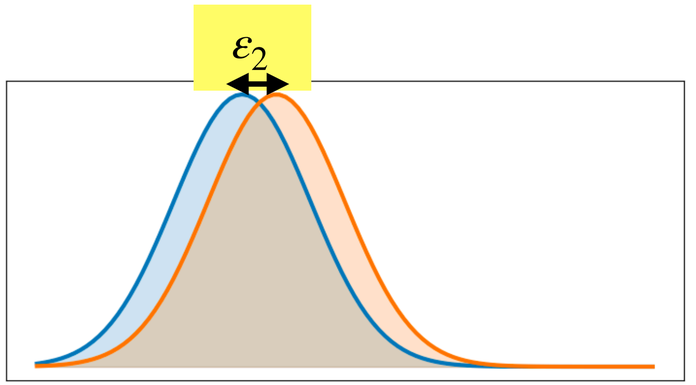
Post-processing

Composition

Algorithm 1

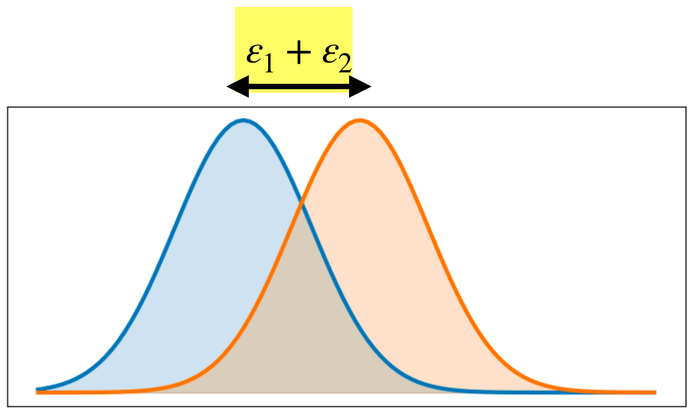
Algorithm 2





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

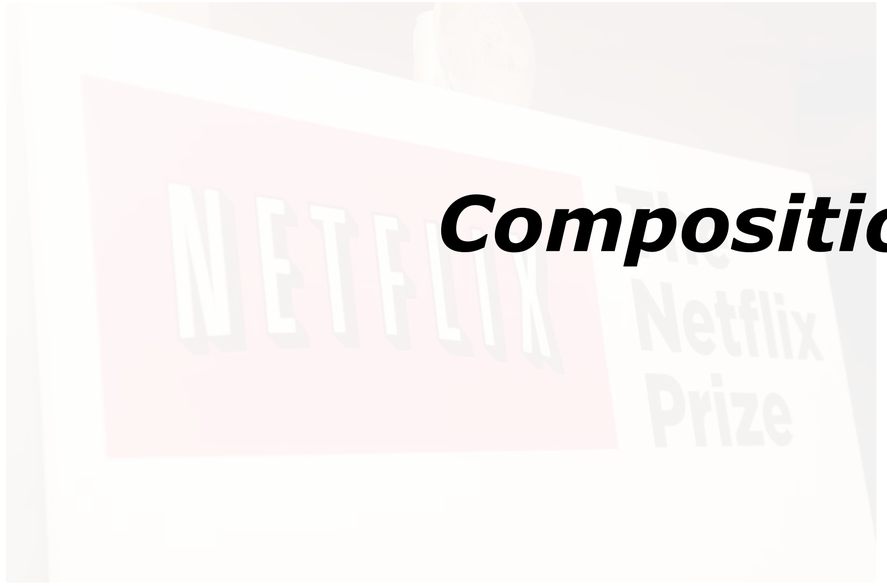
RYAN SINGEL SECURITY MAR 12, 2010 2:48 PM

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

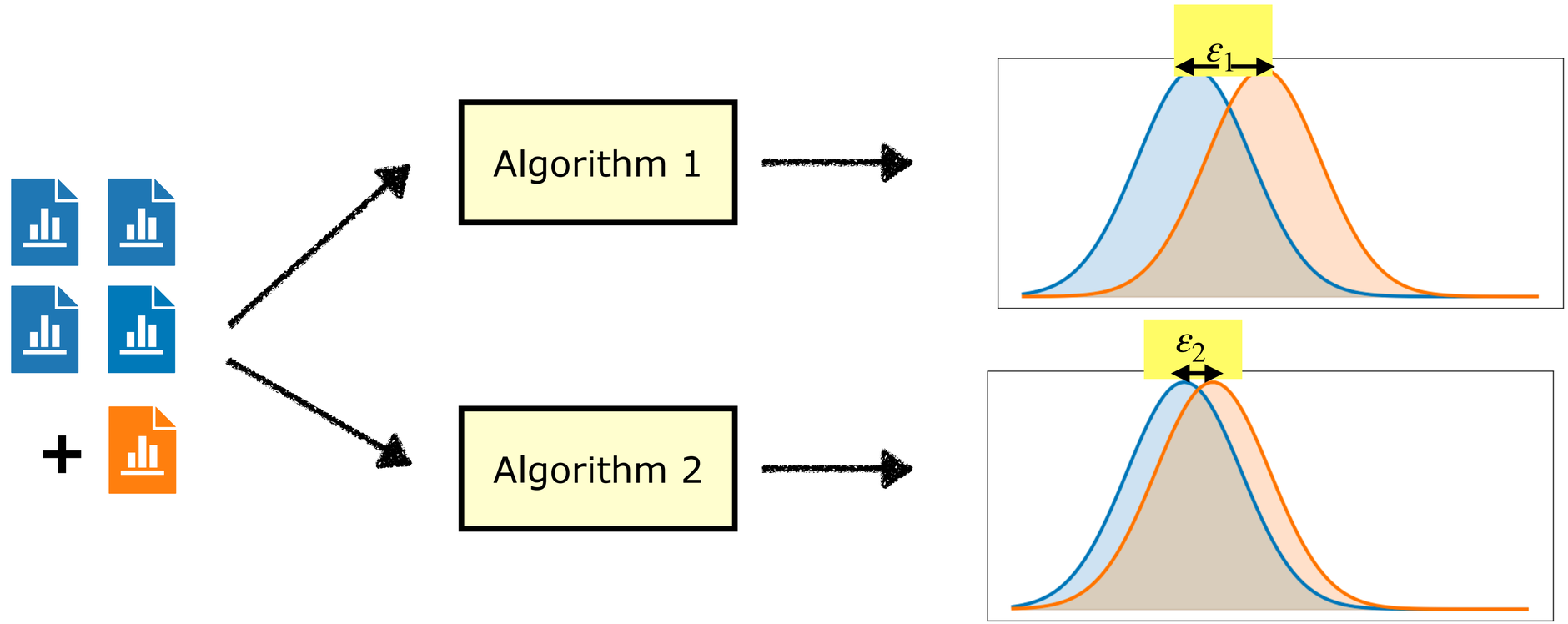
Composition prevents such leakage!

RYAN SINGEL SECURITY MAR 12, 2010 2:40 PM

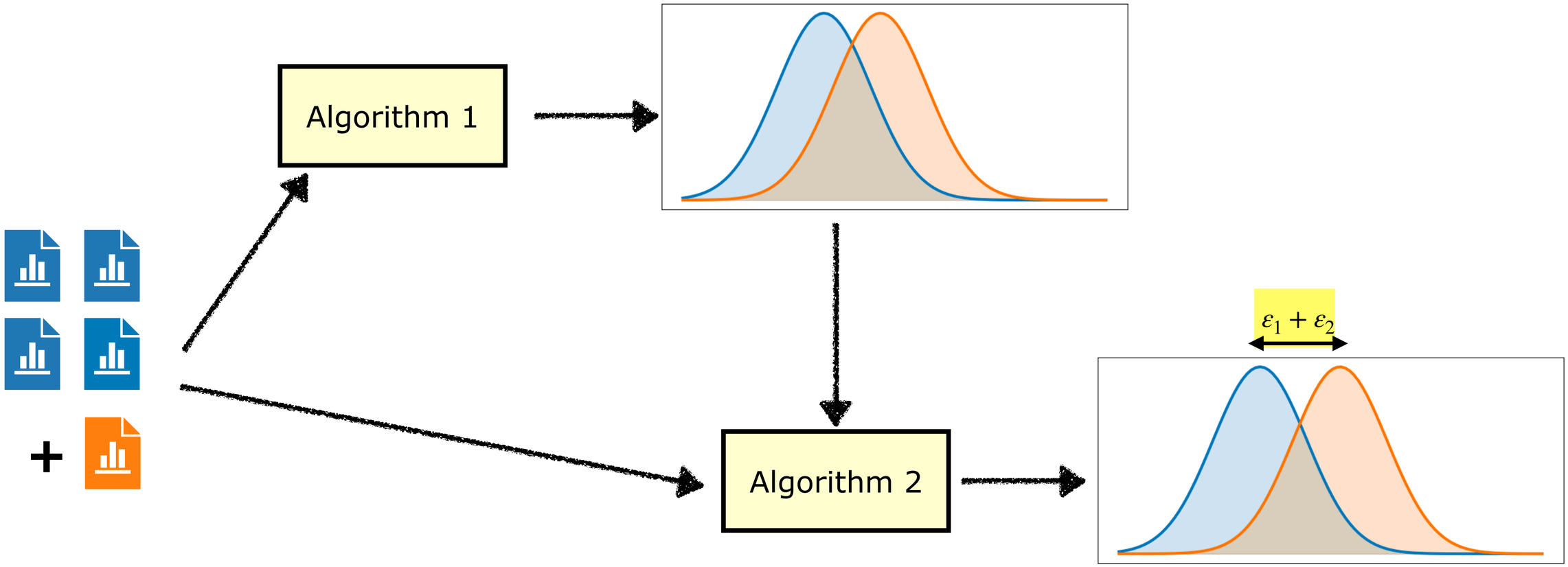
NetFlix Cancels Recommendation Contest After Privacy Lawsuit

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



Adaptive Composition



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

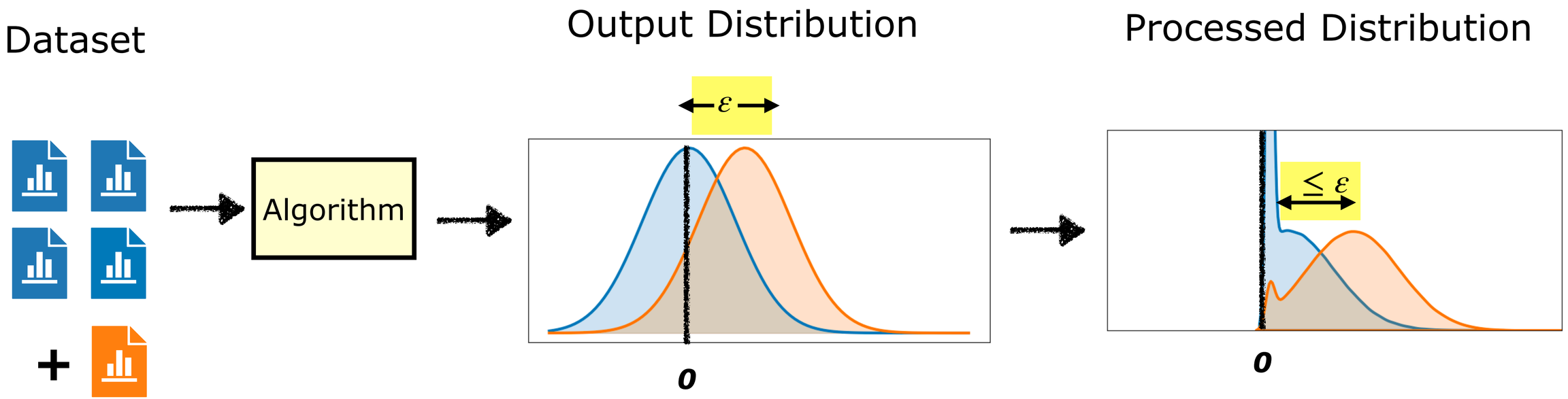
Key properties of DP

Composition over multiple steps

Post-processing

Post-processing

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

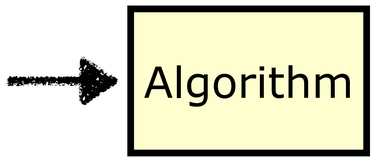


"***Information cannot be created***"

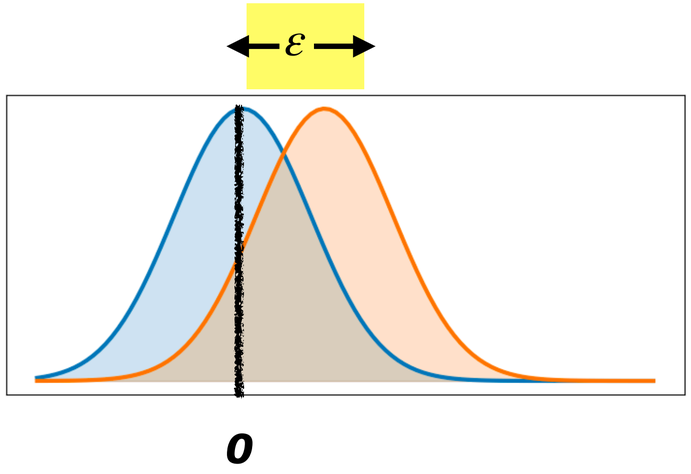
Post-processing

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

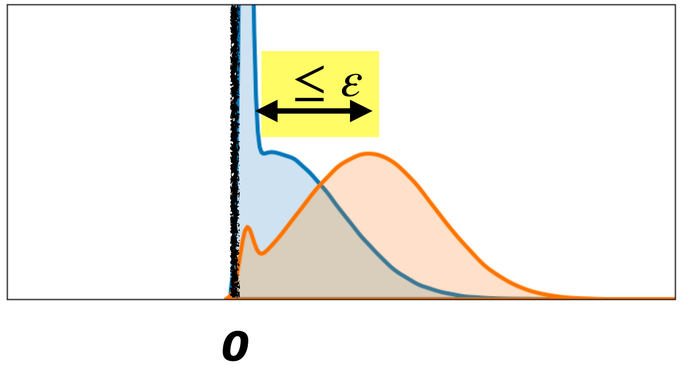
Dataset



Output Distribution



Processed Distribution



"Information cannot be created"

Example: 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

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

Compute and clip per-example gradients

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

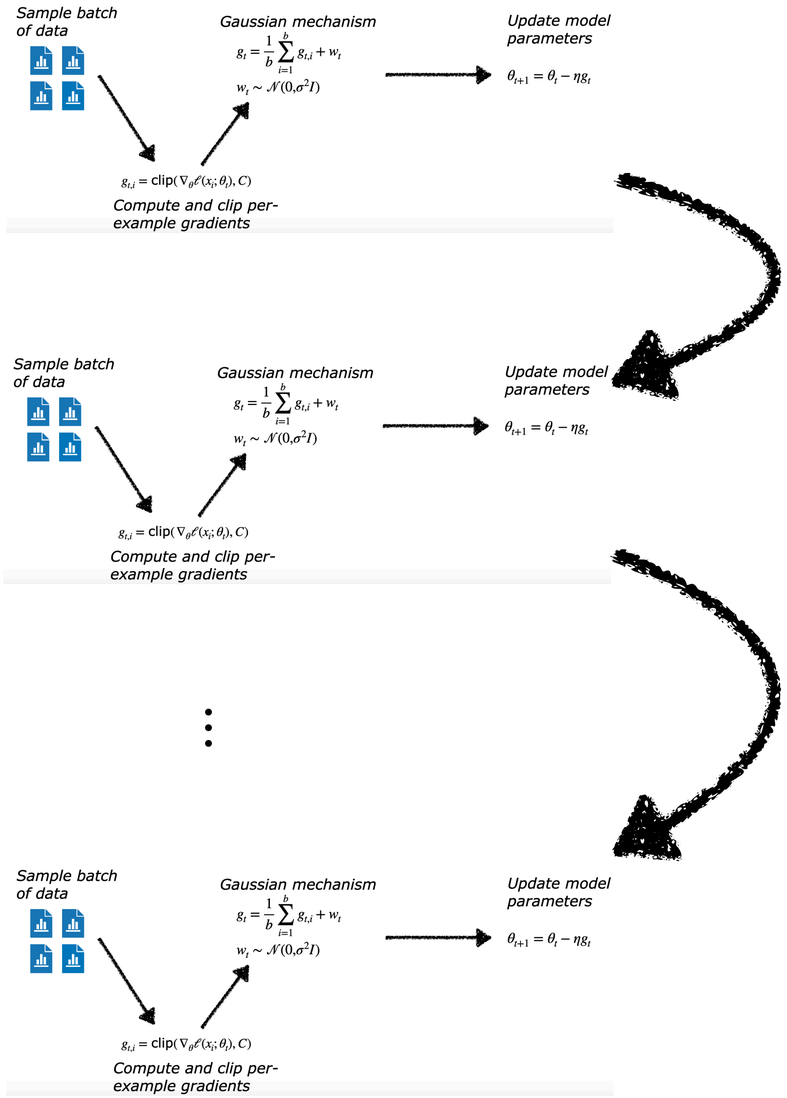
$$\theta_{t+1} = \theta_t - \eta g_t$$

DP-SGD: Stochastic gradient descent with DP

Iteration 1

Iteration 2

Iteration T



Adaptive Composition!

Outline

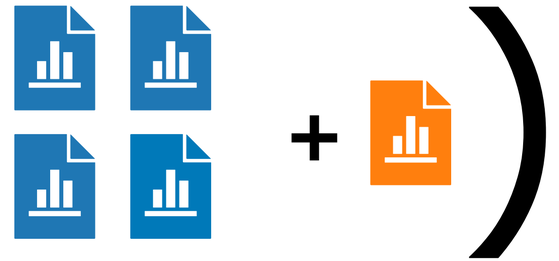
- Differential Privacy: Intuition & Recap
- ***Differential Privacy: Rigorous Mathematical Formulation & Properties***
- Application Examples

Differential privacy $\Longrightarrow$

probability distributions $P \approx Q$ are indistinguishable

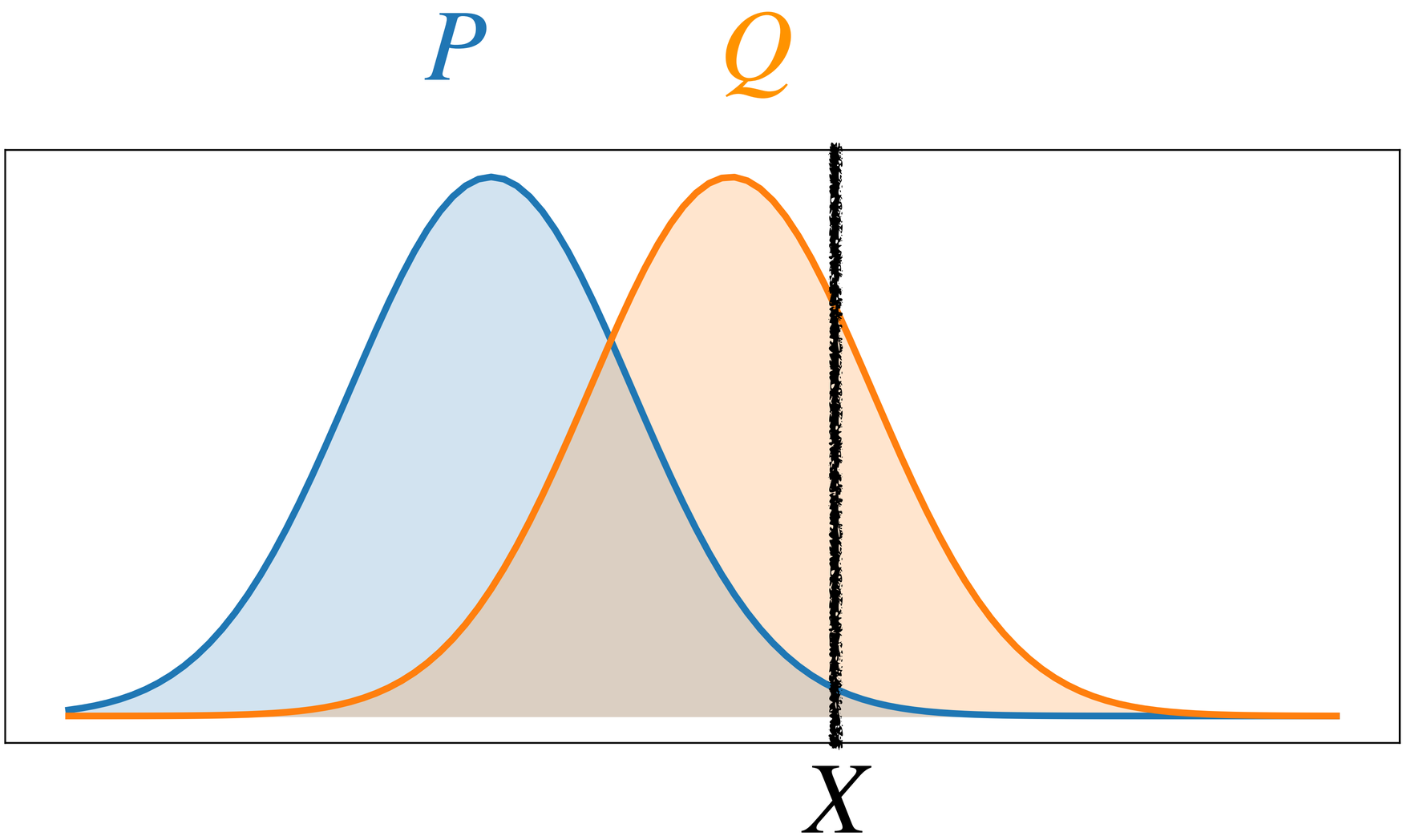
Algo 

Probability distribution P

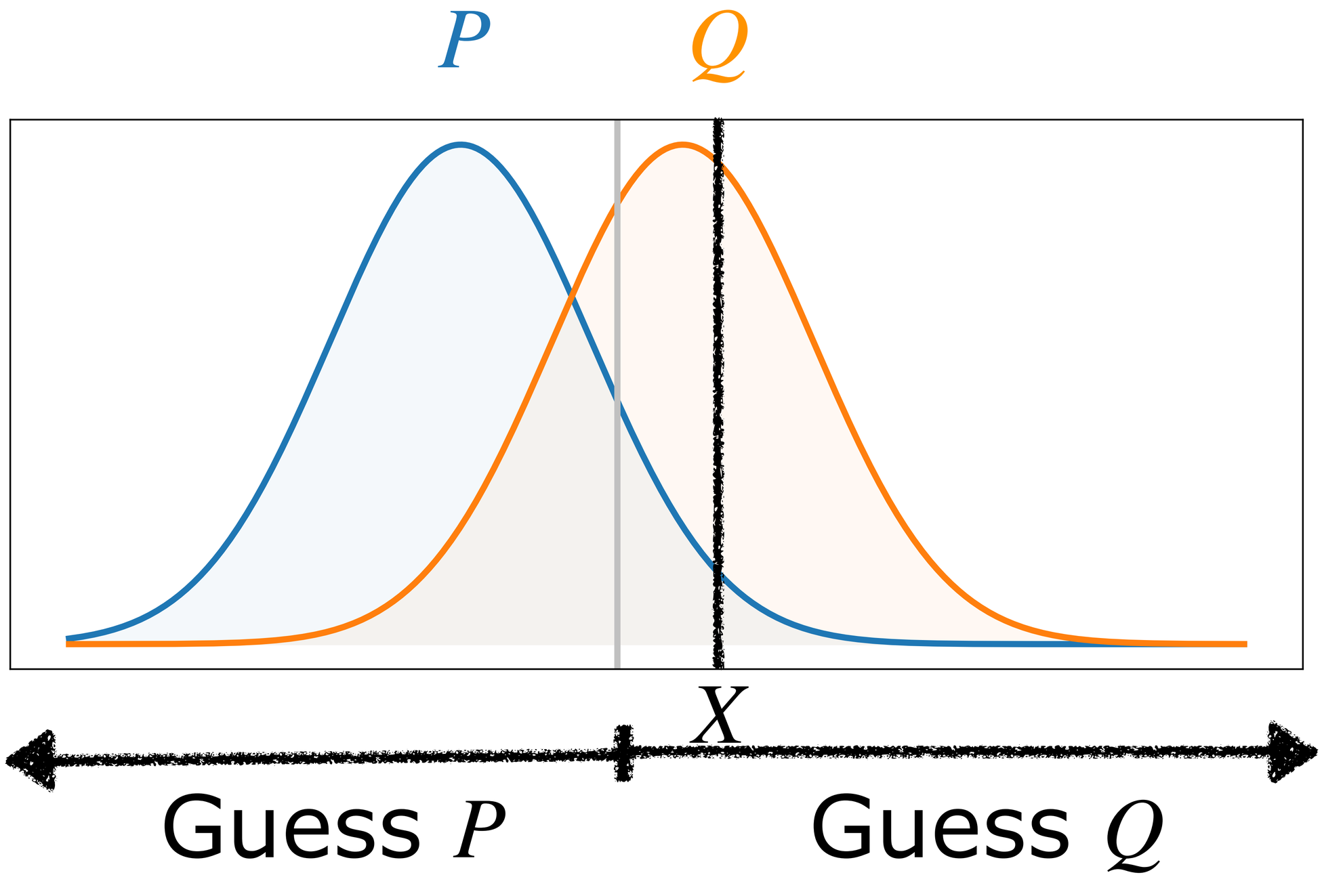
Algo 

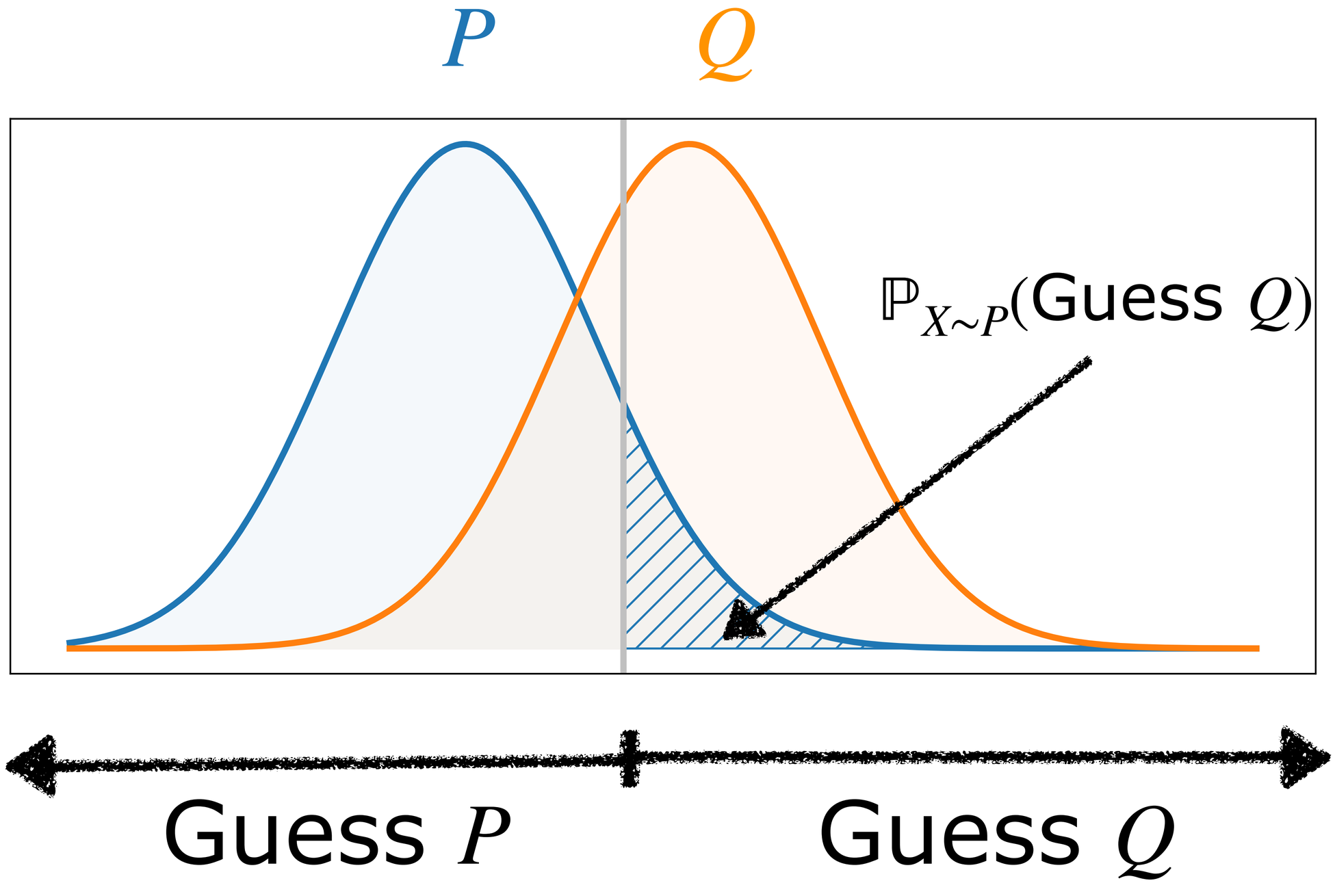
Probability distribution Q

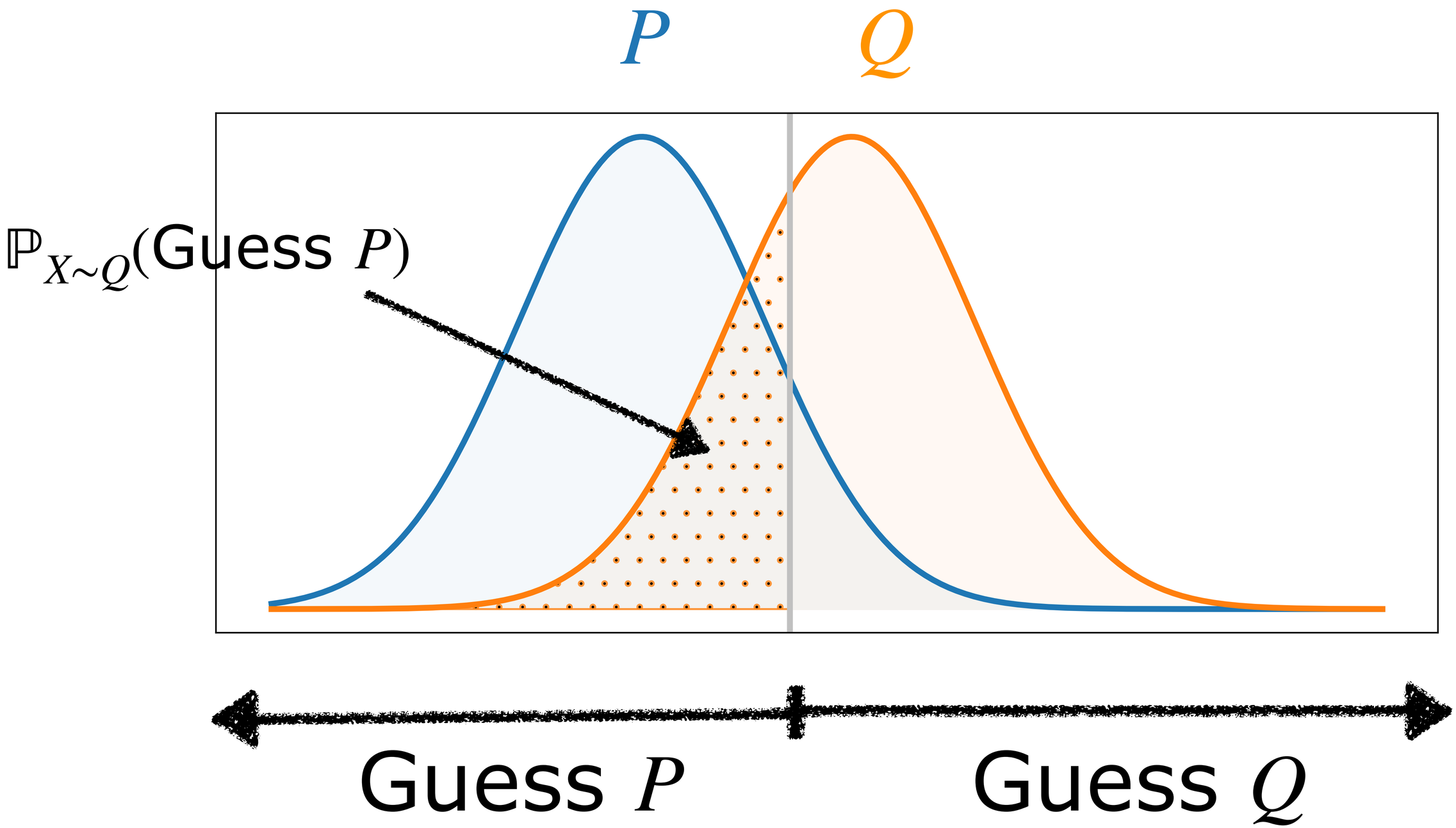
Hypothesis testing:
To distinguish between two distributions



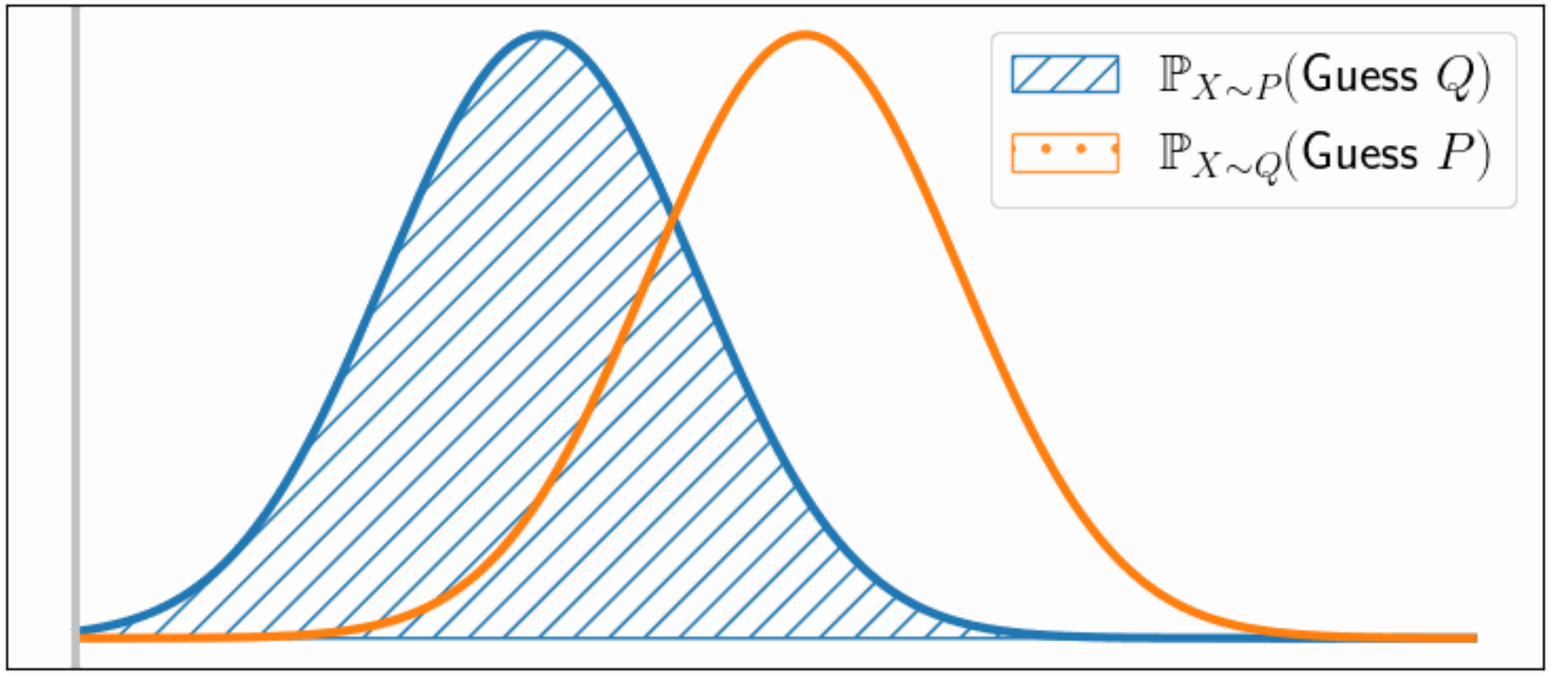
Is
$X \sim P$
or
$X \sim Q$?

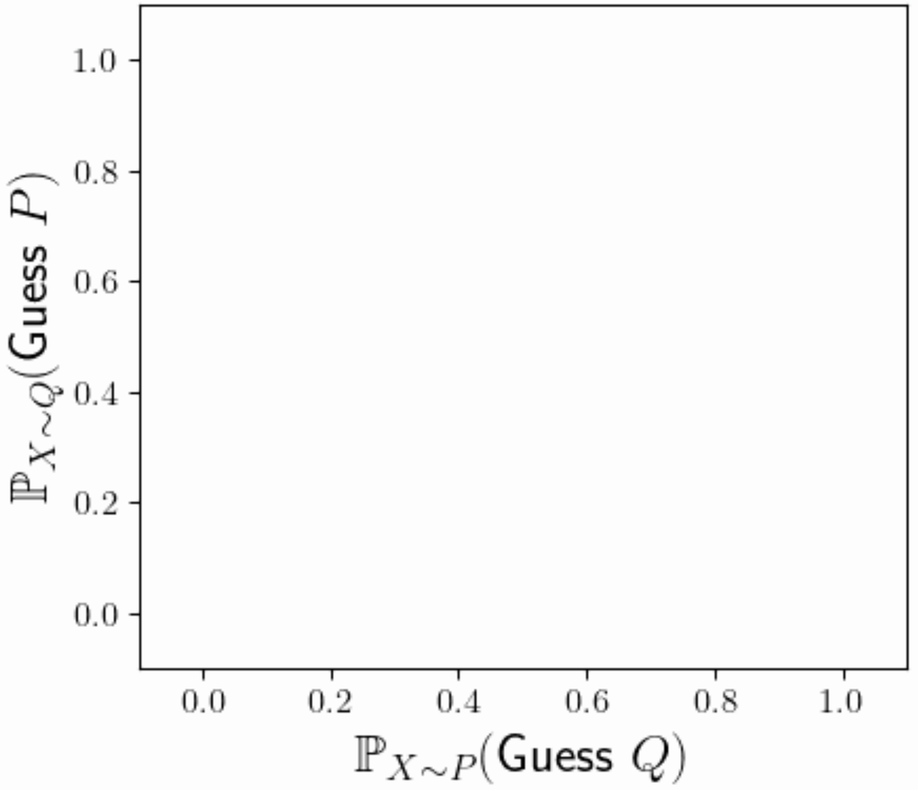

P
Q
X
Guess P
Guess Q


P
Q
$\mathbb{P}_{X\sim P}$(Guess Q)
Guess P
Guess Q


P
Q
$\mathbb{P}_{X\sim Q}$(Guess P)
Guess P
Guess Q

P Q


$\mathbb{P}_{X\sim P}(\text{Guess } Q)$
$\mathbb{P}_{X\sim Q}(\text{Guess } P)$


1.0
0.8
0.6
0.4
0.2
0.0
$\mathbb{P}_{X\sim Q}(\text{Guess } P)$
0.0
0.2
0.4
0.6
0.8
1.0
$\mathbb{P}_{X\sim P}(\text{Guess } Q)$

Neyman-Pearson Lemma

f_P: PDF of P

f_Q: PDF of Q

Define:

Log-likelihood ratio

$$T = \log\left(\frac{f_P(X)}{f_Q(X)}\right)$$

The best strategy: Fix a real number t and guess P if $T < t$ and guess Q otherwise

No other strategy can get both errors smaller simultaneously

Neyman-Pearson Lemma

f_P: PDF of P

f_Q: PDF of Q

Define:

Log-likelihood ratio

$$T = \log\left(\frac{f_P(X)}{f_Q(X)}\right)$$

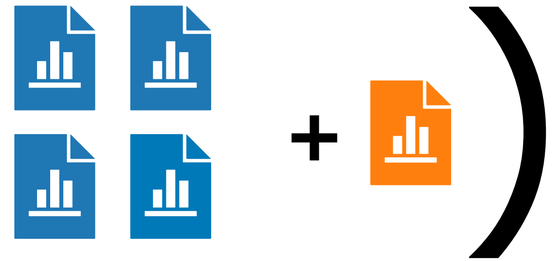
Log-likelihood ratio: the fundamental quantity for distinguishing between distributions

Differential privacy $\Longrightarrow$

probability distributions $P \approx Q$ are indistinguishable

Algo 

Probability distribution P
with density f_P

Algo 

Probability distribution Q
with PDF f_Q

Privacy Loss Distribution & Composition

Blackboard discussion

Reading material:
https://arxiv.org/pdf/2210.00597

Interested in joining us? Apply here:

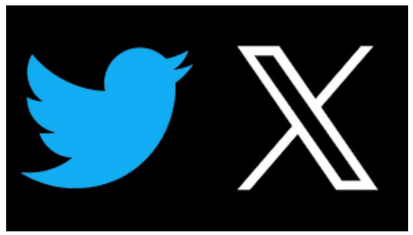
- **MS/PhD:** apply via IITM
- **Interns** / **pre-doctoral** researchers
  - apply directly via RBCDSAI or CeRAI

RBC DSAI
ROBERT BOSCH CENTRE FOR DATA SCIENCE AND ARTIFICIAL INTELLIGENCE IIT MADRAS

CeRAI
Centre for Responsible AI

Thank you!

https://krishnap25.github.io



@KrishnaPillutla