DA 5001 / DA 6400: Privacy in AI

Krishna Pillutla

IIT Madras

About Me

- B.Tech IIT Bombay
- M.S. Carnegie Mellon University (USA)
- Ph.D. University of Washington (USA)

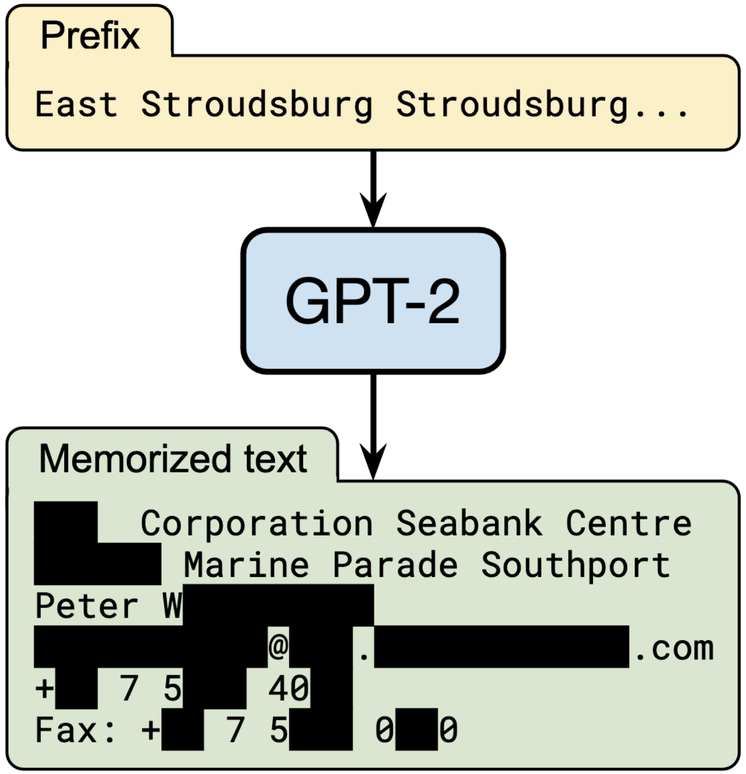
Research accolades:

- Outstanding Paper Award @ NeurIPS 2021
- JP Morgan PhD Fellowship
- Anne-Dinning Michael Wolf Endowed Regental Fellowship (UW)

- Started here 3 months ago
- Prev: Google Research, USA
- Research interests:
  - Privacy :-)
  - Robustness/Eval. Of GenAI

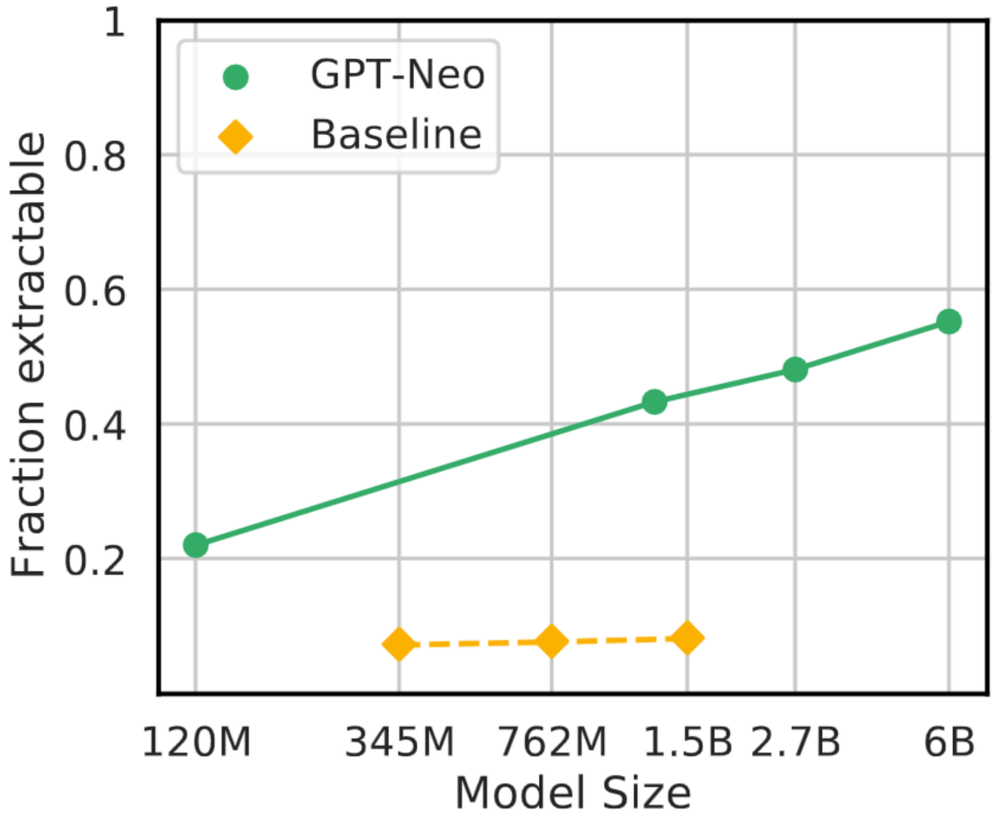
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

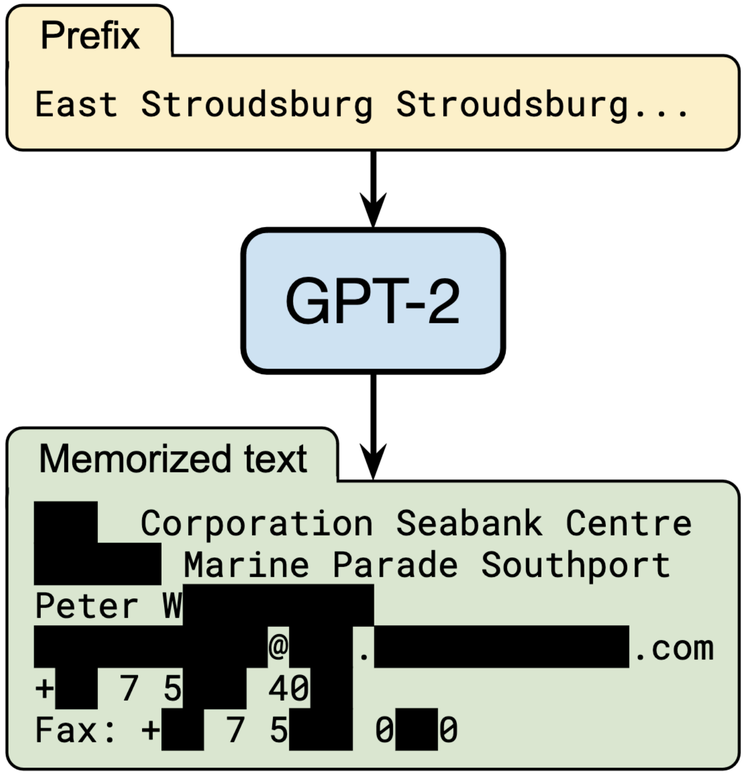
Models leak information about their training data



Carlini et al. (USENIX Security 2021)

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





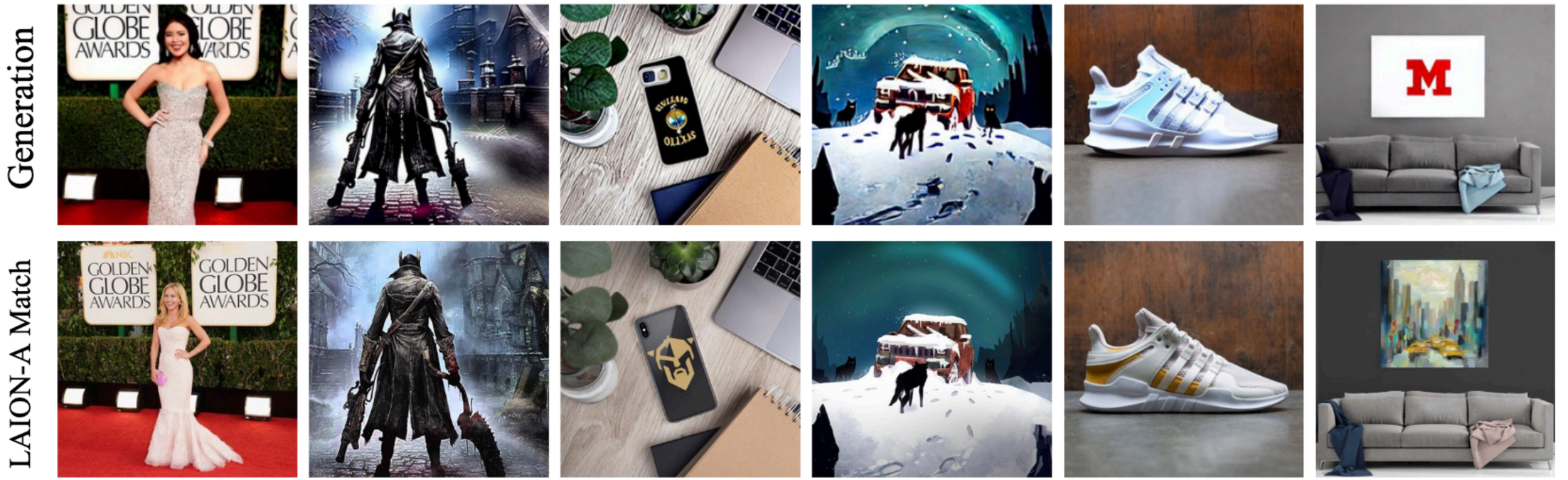
Carlini et al. (USENIX Security 2021)

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

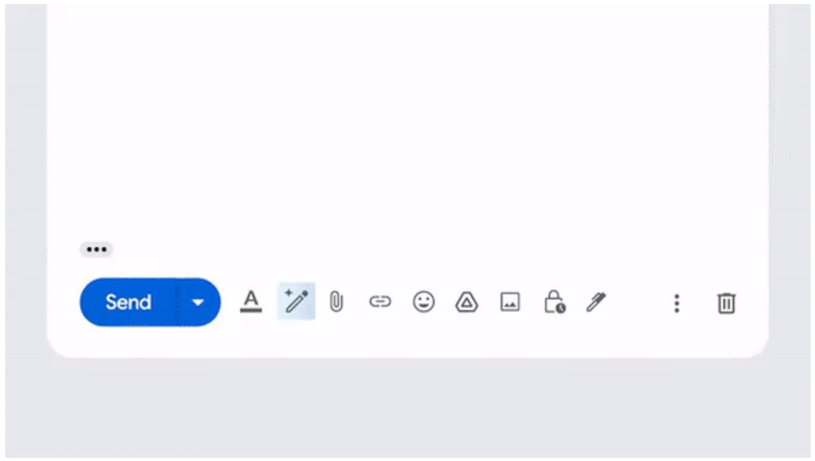
ANTHROP\C

LLAMA 2

MISTRAL AI_

BharatGPT

Gmail



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

Microsoft Copilot for Microsoft 365

Your everyday AI assistant at work

Natural Language

Large Language Models (LLMs)

Web grounding

Commercial data protection

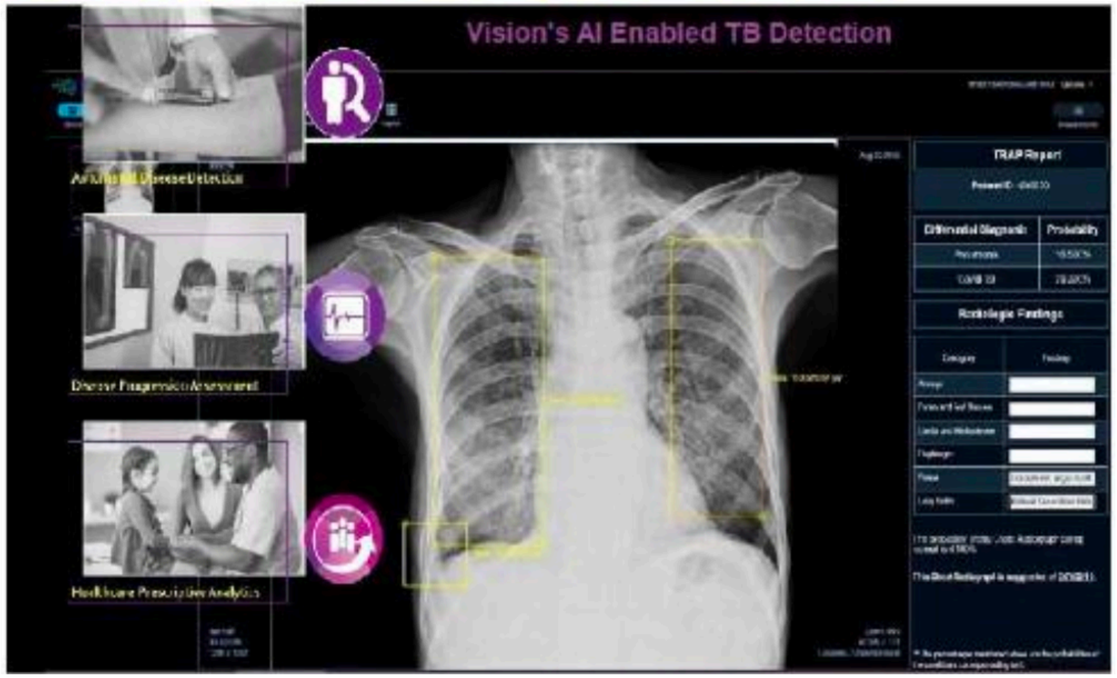
The Internet

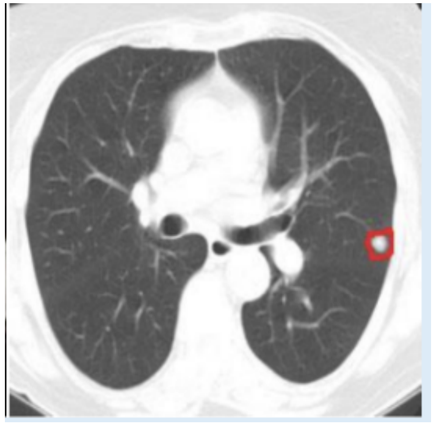
Microsoft Graph grounding

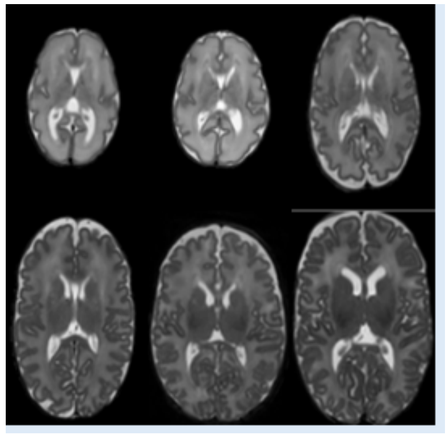
Microsoft 365 Apps

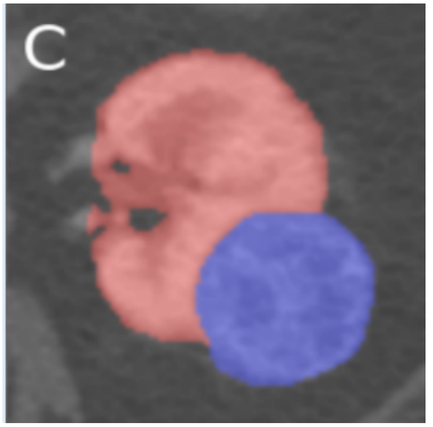
Enterprise-grade data protection

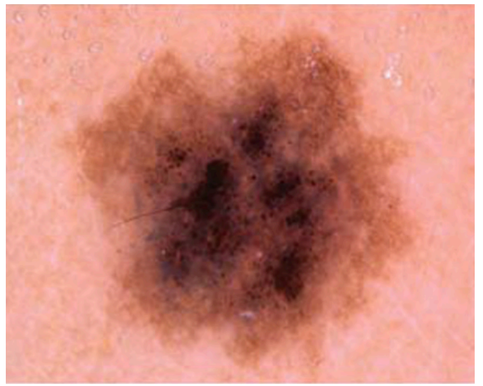
Copilot Studio











Train			Val (OOD)	Test (OOD)
d = Hospital 1	d = Hospital 2	d = Hospital 3	d = Hospital 4	d = Hospital 5
y = Normal				
y = Tumor				

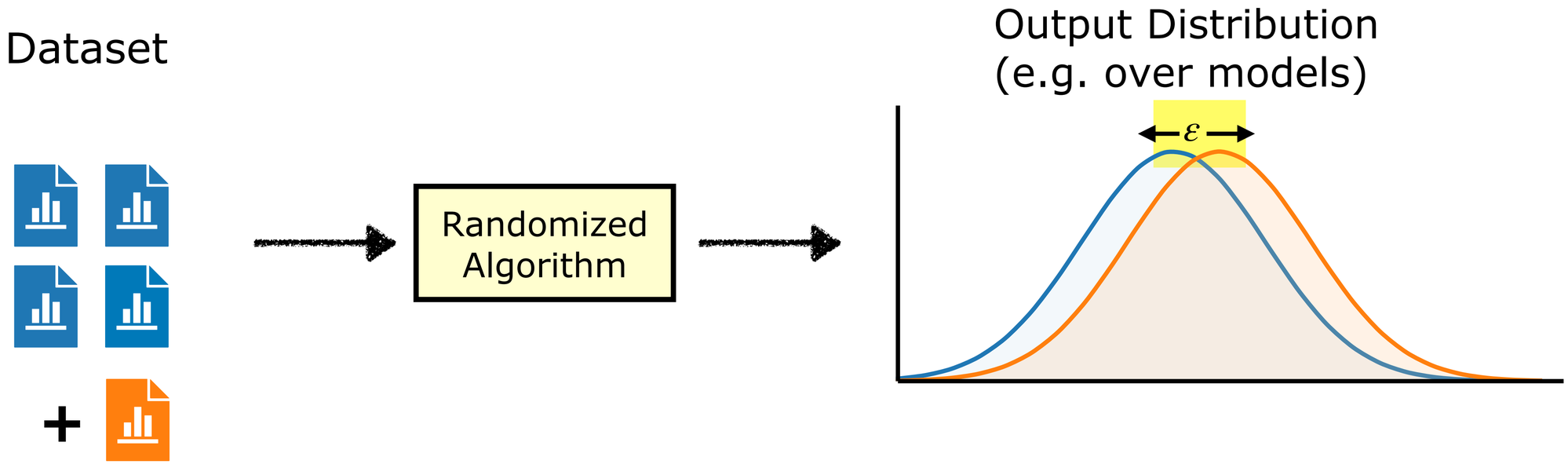
Figure 4: The CAMELYON17-WILDS dataset comprises tissue patches from different hospitals. The goal is to accurately predict the presence of tumor tissue in patches taken from hospitals that are not in the training set. In this figure, each column contains two patches, one of normal tissue and the other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

Digital Health Laws and Regulations India 2024

HIPAA
Health Insurance Portability and Accountability Act

EU GDPR

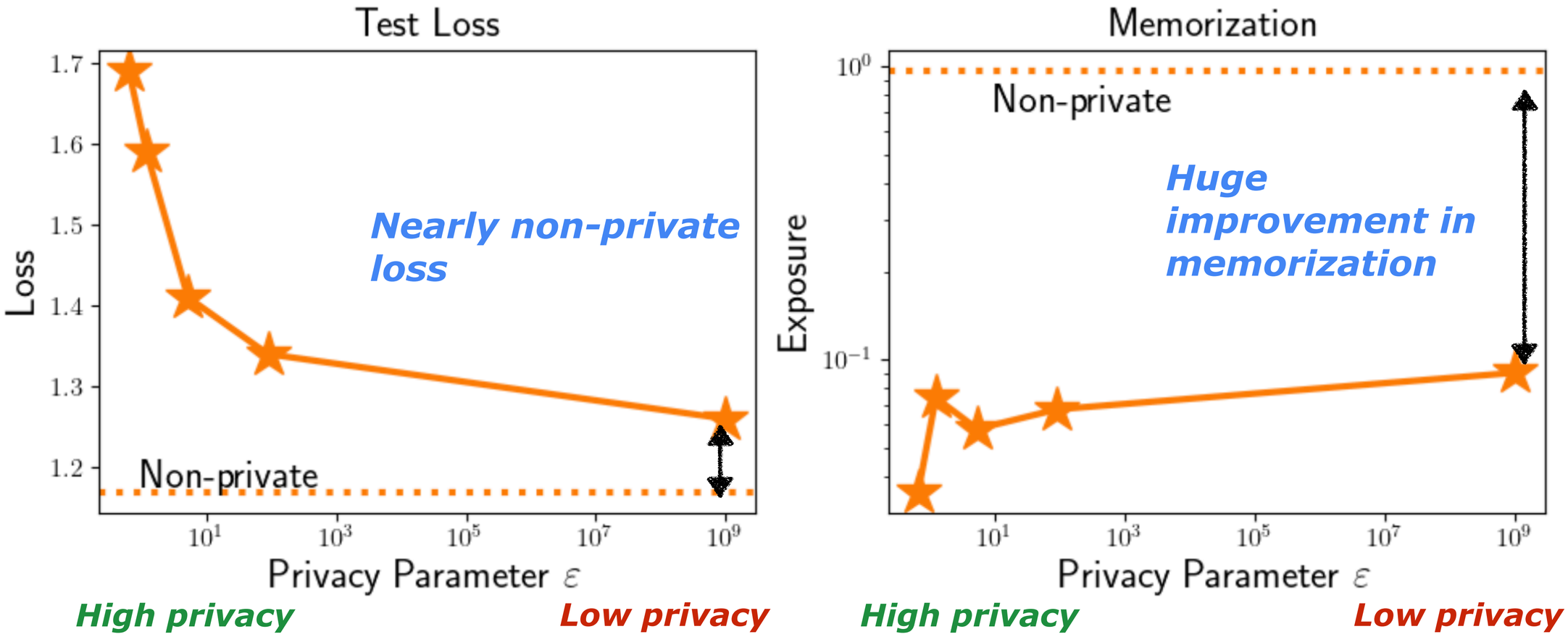
Differential privacy nearly eliminates memorization



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



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

Classes

- **Monday**: 5 to 5:50 pm
    - Usually a mini lab/tutorial session. Please bring your laptops
    - Contact us privately if you have strong impediment (e.g. cannot afford a laptop, etc.)
- **Wednesday**: 2 to 3:15 pm
    - Lecture, usually on the board
- **Thursday**: 3:30 to 4:45 pm
    - Lecture, usually on the board

Communication

- **Course Webpage**: https://krishnap25.github.io/privAI_course_2024o/
- **Piazza**: link to be announced on the course webpage
- **Assignments**: submit via Gradescope
- Do ***not*** contact us via email. Use piazza only
  - Public post: use for logistics, assignments, etc.
  - Private (instructor + TAs): For special requests

Grading

- ***Homeworks***: 35% (= 5% + 10% + 10% + 10%)
- ***Course Project***: 40%
- ***Midterm exam***: 20%
- ***Scribing*** (lecture notes): 5%

Homework Assignments: 35% of the grade

- **HW0**: Review of prerequisites (**total score: 5%**)
  - out today, due on August 9th
- **HW1-3**: approx. 12 days per assignment (**total score: 10% each**)
  - HW1: 2nd half of August
  - HW2: 1st half of September
  - HW3: early October

- **Submission**: Write up via LaTeX and submit PDFs. Code: JuPyTer notebooks

Course Project: 40% of the grade

- Most of your learning will be through the course project
- Groups of 2-3 (exact details TBD, depending on the final class size :)
- ***Options***:
  - Research project
  - Implementation: benchmarking and open-sourcing
  - In-depth paper analysis

Course Project: Research

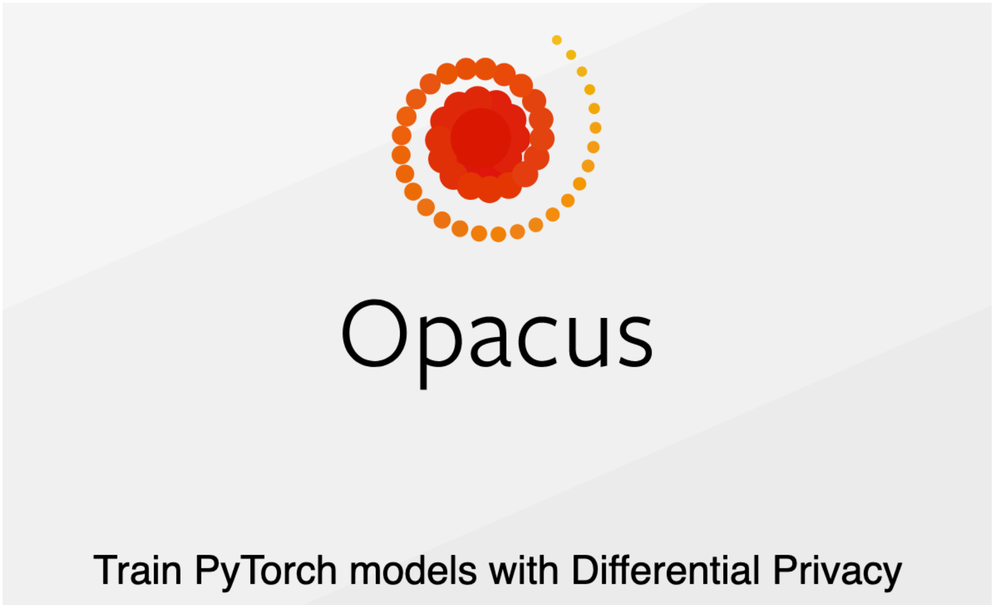
- **Original research**: can be theory, applied, or mix or both
- Commensurate to a workshop paper at NeurIPS/ICML/ICLR conferences

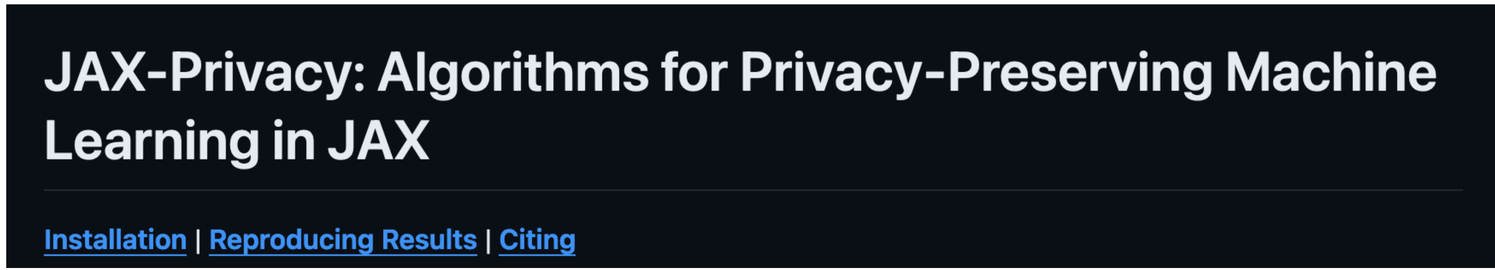
Course Project: Research

- 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





Course Projects: In-depth paper analysis

- Read and analyse the results of a theoretical research paper
- Reproduce the proofs in your own words (possibly using alternate methods)

Course Project (40%): Logistics

- **Proposal (5%, 1-2 pages):** Due around end of September
  - What do you want to do? Why? Identify important references.
- **Midpoint report (10%, 4 pages)**: Due around 3rd week of October
  - Literature review (explain ideas mathematically, as in lectures)
- **Presentation (10%)**: Week of Nov. 4th (last week of classes)
- **Final report (15%, 6-8 pages)**: Due around Nov. 14th (End sem week)

Course Goals

Project

Homework

Lectures & Reading

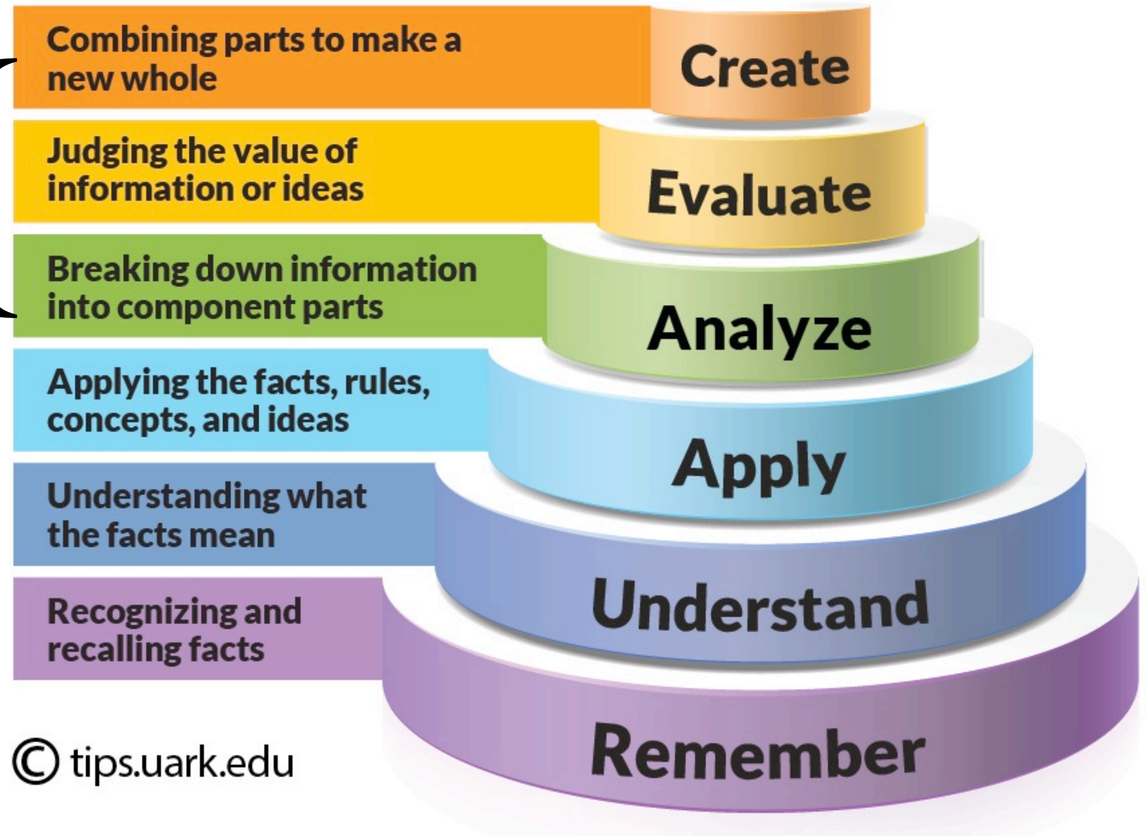


Figure: https://tips.uark.edu/using-blooms-taxonomy/

Bloom's Taxonomy

Midterm (20%): Thursday October 3, 2024

- No make up exams
- **Mark your calendars** and plan accordingly
    - This is the week after project proposal is due

Scribing (5%)

- Math components of lectures will be on the board
- Group of 2-3 students have to scribe them in LaTeX
- Due before class within one week

Late Days

- You have a budget of 3 late days for homework and scribing
    - No questions asked, you do not need to contact us for it
    - Even two minutes after the deadline is considered as a full late day
- Any further late days after exhausting your late day budget results in a zero grade for the homework
- **No late days** for **project-related deadlines**

Honour Code

- **Homeworks**: Can collaborate with others to discuss homework, provided:
  - You acknowledge everybody you worked with in your submission
  - You write your own solutions and code independently without referring to written notes and other material from joint discussions
  - You must internalize the solution/code well enough to reconstruct it fully by yourself
  - ***Copy-pasting not allowed***

Honour Code: Using LLMs for HW

Ok to use LLMs but each time you do, you must provide:

- name of the LLM
- the exact prompt (verbatim)
- A summary of how it helped. E.g.
    - “The LLM provided a proof sketch, which I followed”
    - “The LLM pointed me to acleverref.com, which gave more details of the proof summary”

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/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
- **Wednesday 3:30 to 4:30 PM** at my office (after class)

Auditing the course

- **Very strongly discouraged**
  - Active learning works best: working on homework/project
- Students registered for credit will be given priority
  - Others may be requested to leave in case we run out of space

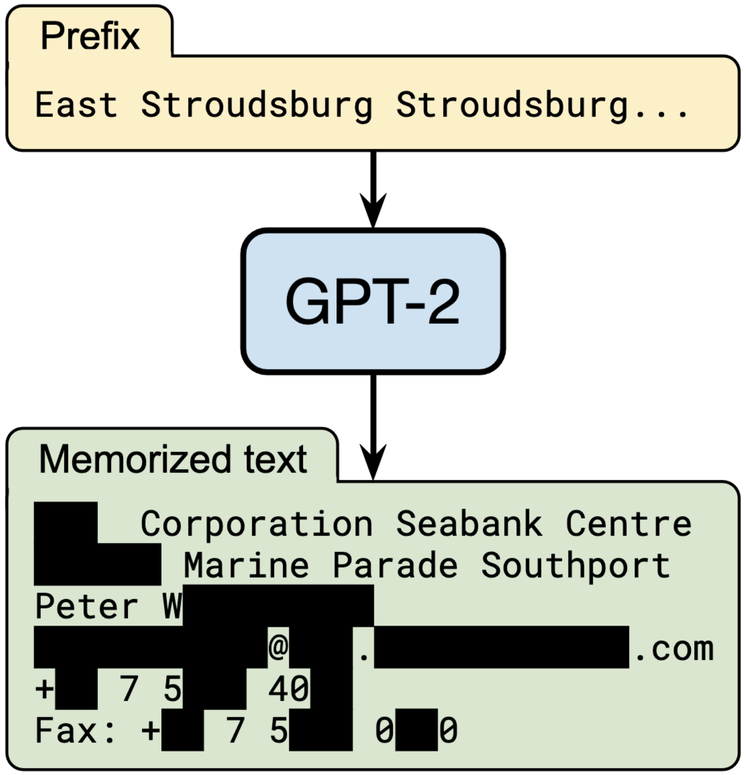
Attendance & Misc policies

- We will not take attendance
- This will be a challenging course. Take it only if you ***really*** want to learn the material
- Hard to follow without attending the lectures
- Please be on time. It is very disrespectful to be late
- No phones in the class please. Laptops for lab sessions only (usually Mondays)

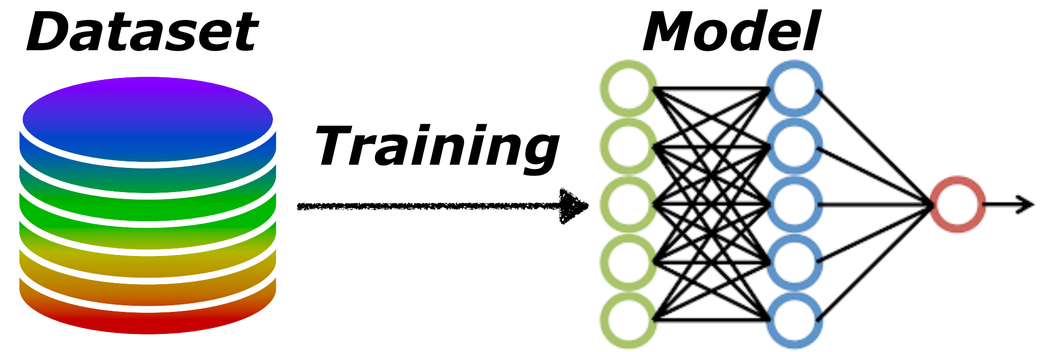
Tentative Course Outline

Week 1: Privacy Attacks

- Data Reconstruction/Extraction



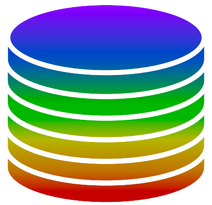
- Membership Inference



Adversary must guess if

data

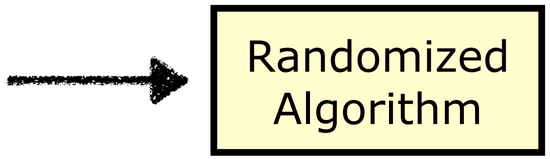
belongs to the dataset



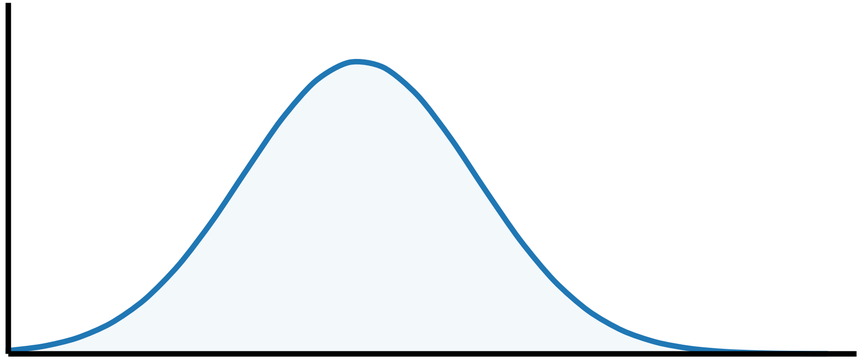
Weeks 2-4: Differential Privacy (DP)

A mathematically rigorous notion of "***privacy***"

Dataset

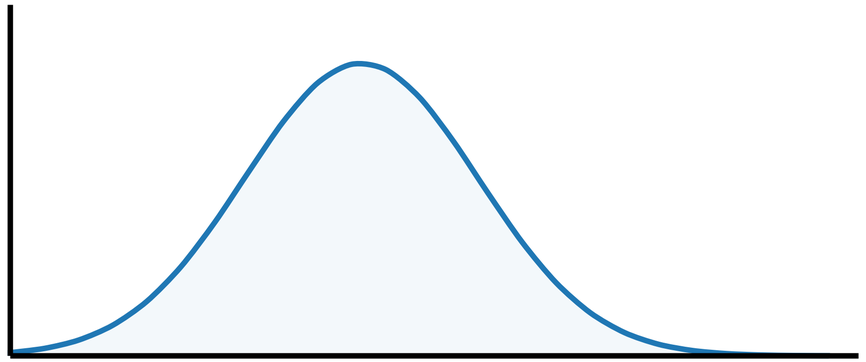


Output Distribution
(e.g. over models)



Dataset

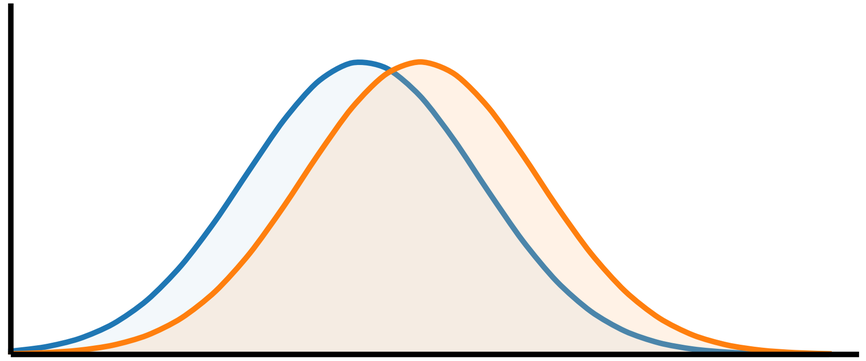
Output Distribution
(e.g. over models)

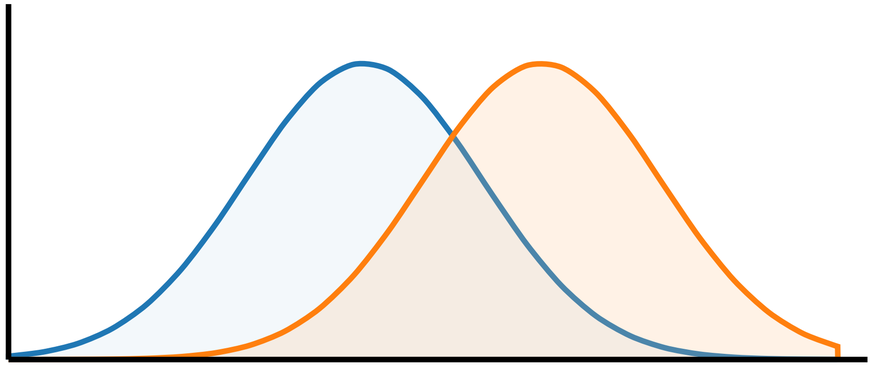


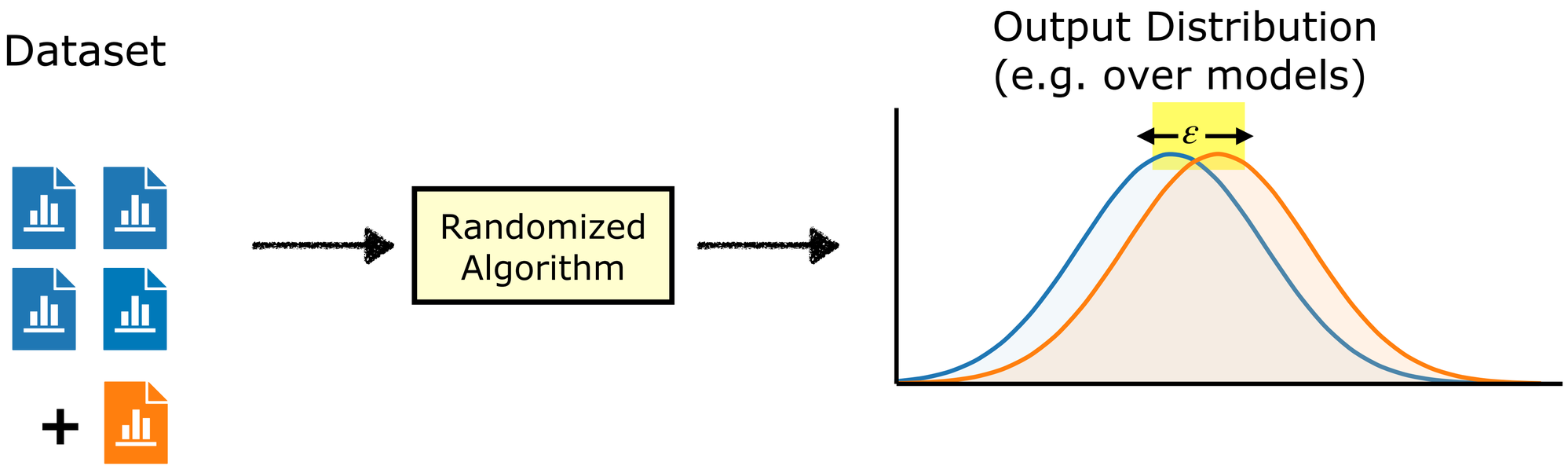
Dataset

Randomized Algorithm

Output Distribution
(e.g. over models)





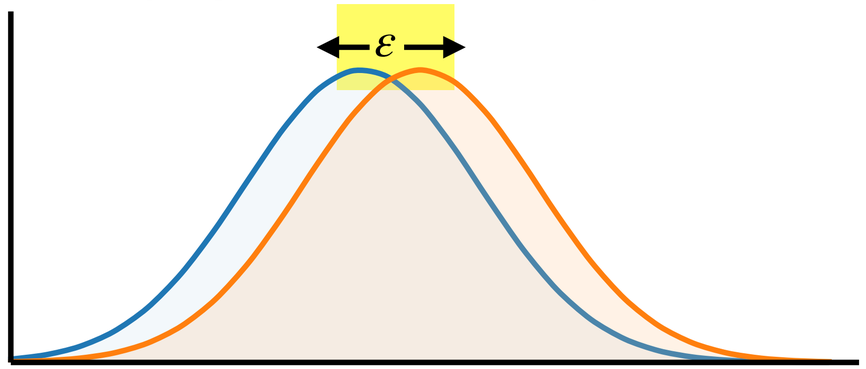


A randomized algorithm is ε-**differentially private** if the addition of **one user's data** does not alter its output distribution by more than ε

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

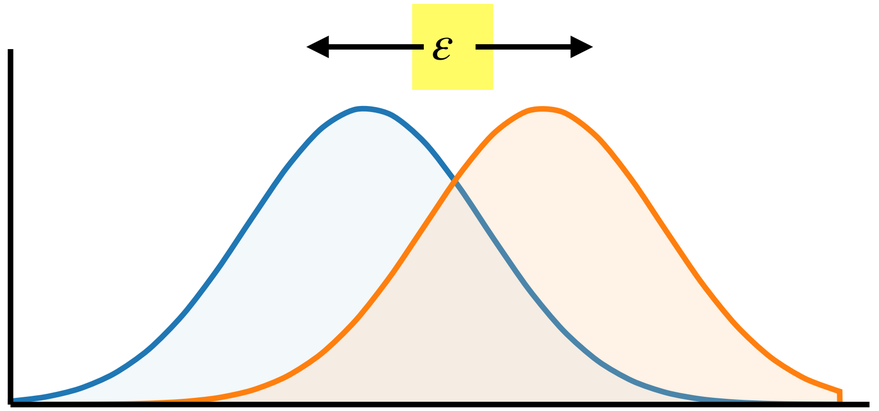
Dataset

Output Distribution
(e.g. over models)



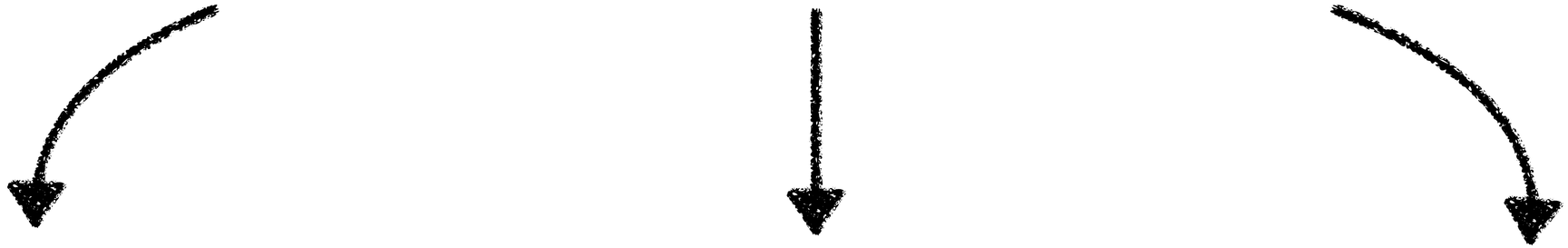
ε-differential privacy

Large $\varepsilon \implies$ more privacy leakage



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

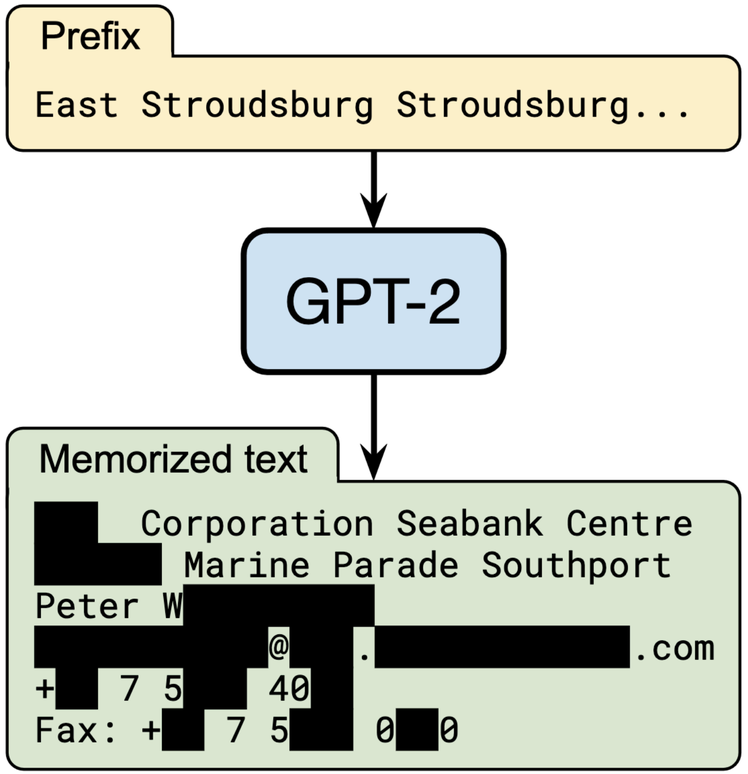
- *Differential Privacy*

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

Weeks 5-7: ML with DP

Modifying stochastic gradient descent for DP

Weeks 8-9: Protecting Against Data Reconstruction Attacks



Week 10: Review and Mid-Term Exam

- Open book exam
- No collaboration allowed

Weeks 11-14: Advanced Topics

- Privacy in distributed settings
- Generative AI:
    - Contextual privacy norms
    - Copyright
- Unlearning
- ***Your suggestions welcome***

Week 15

- Project presentations

Thank you! Questions?