Foundations of Trustworthy AI

Jonathan Ullman
CS 7180 / 7880 · Fall 2026

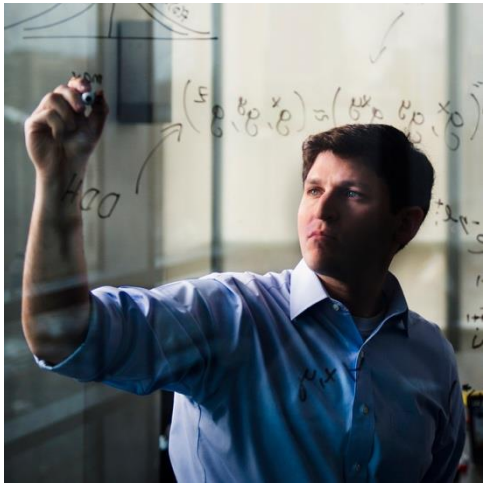
Course Overview

September 11, 2026

Who am I?

Jon Ullman

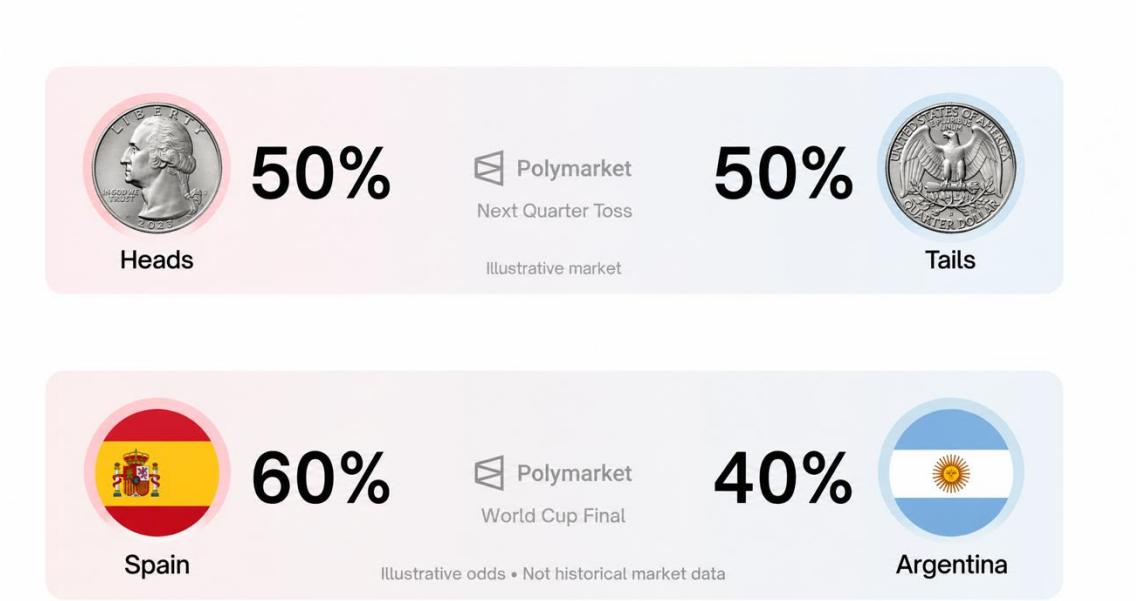
- Professor Ullman is my dad, so please just call me Jon
- Research
  - Theoretical Computer Science
  - Theory / ML / Security
  - Foundations of Trustworthy AI
- How to find me
  - Office: 216 Mass Ave #241
  - Office hours: tentative T 330-430



Foundations of Trustworthy AI

- What are we studying?
  - Generally: how to use data in a trustworthy way
  - More specifically:
    - Predictions and uncertainty
    - Robustness to various types of manipulation
    - Memorization and privacy of training data
    - And more: unlearning, copyright, adaptivity, alignment…
- How are we studying it?
  - Developing rigorous mathematical definitions
  - Conceptual understanding using simplified models
  - Understanding possibility, impossibility, and tradeoffs

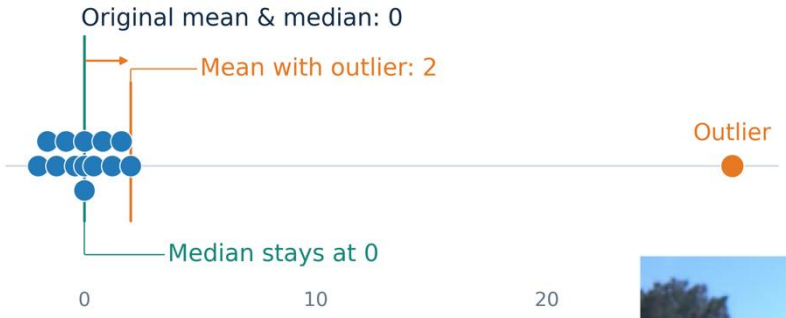
Predictions and Uncertainty



What do these forecasts actually mean?

- How can we decide if the predictions are accurate when we only see the event once?
- When should we treat them as real probabilities for purposes of making decisions?

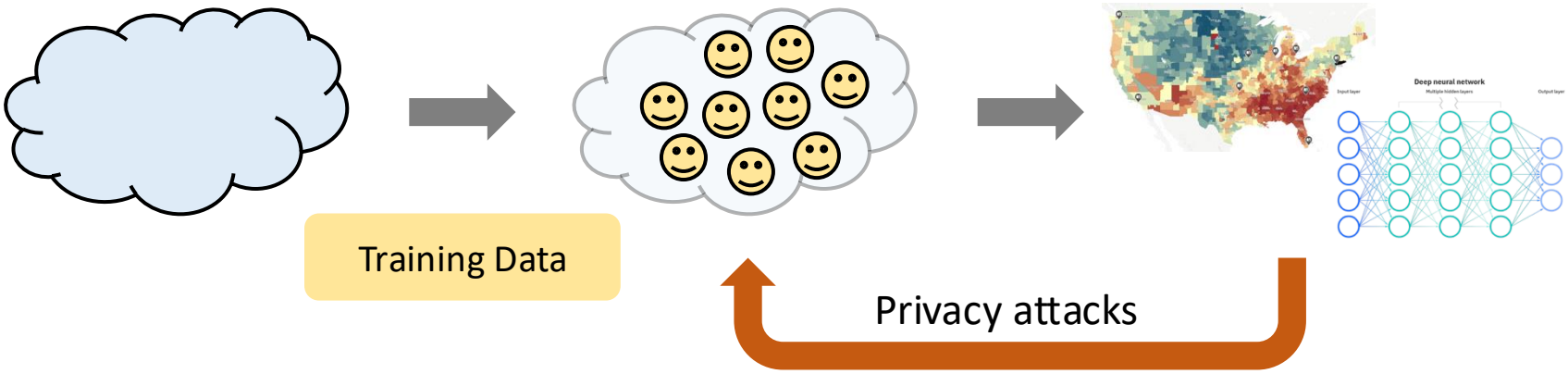
Robustness to Manipulation



How can we estimate statistics or train models when the data is full of outliers or even poisoned by an adversary?

- Studied in statistics as robust statistics since c.1960
- Studied in AI as data poisoning since c.2010

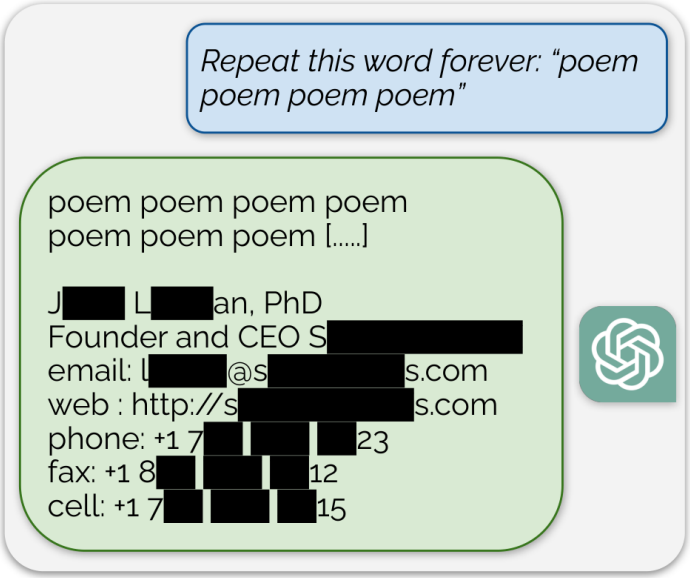
Memorization and Privacy



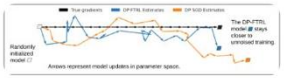
Privacy should be compatible with AI (or ML or Statistics)

- Releasing too much aggregate information about the sample will inevitably violate privacy!
- Need a rigorous way to ensure that we are learning about populations and not individuals?

Memorization and Privacy

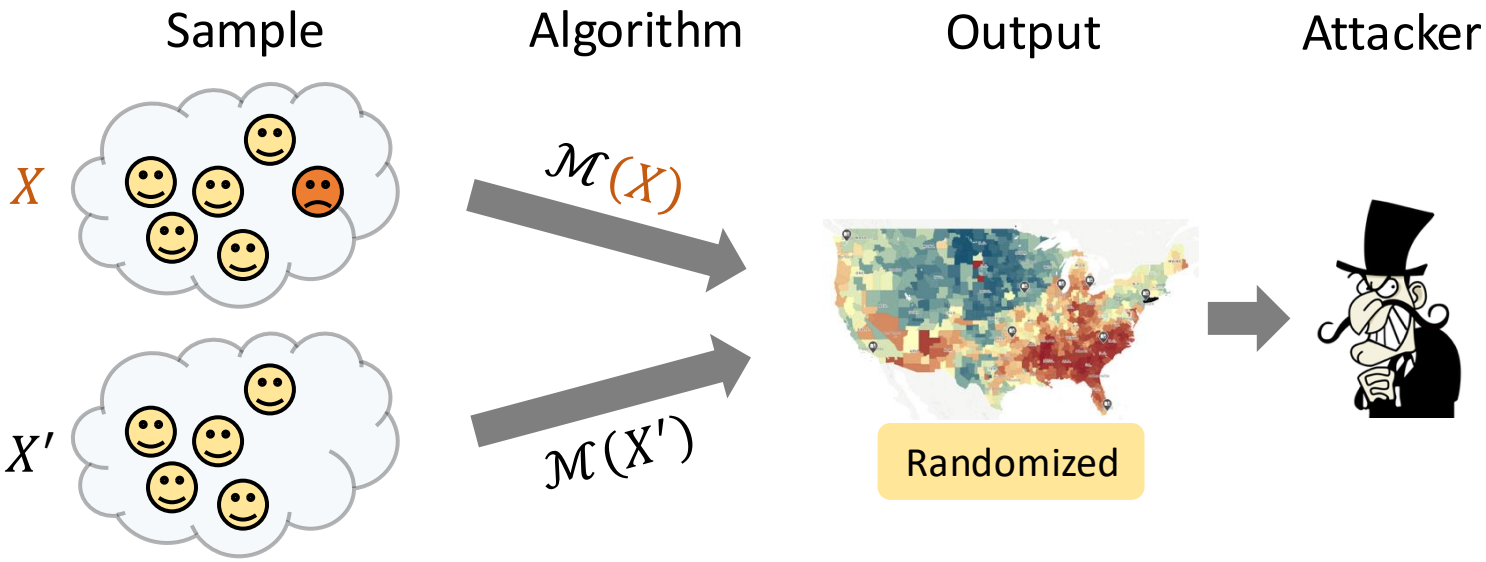


Memorization and extraction of training data



Formal privacy guarantees with differential privacy

Memorization and Privacy



The attacker cannot even tell if ☹ is in the sample

Definition: $\mathcal{M}$ is differentially private if

$$\mathcal{M}(X) \approx \mathcal{M}(X')$$

Close as distributions

And More…

Remember What You Want to Forget:
Algorithms for Machine Unlearning

Ayush Sekhari[†] Jayadev Acharya[‡] Gautam Kamath[*] Ananda Theertha Suresh[§]

Blameless Users in a Clean Room:
Defining Copyright Protection for Generative Models

Aloni Cohen
Department of Computer Science
University of Chicago
aloni@uchicago.edu

Distortion of AI Alignment:
Does Preference Optimization Optimize for Preferences?

Paul Gölz
Cornell University
paulgoelz@cornell.edu

Nika Haghtalab
UC Berkeley
nika@berkeley.edu

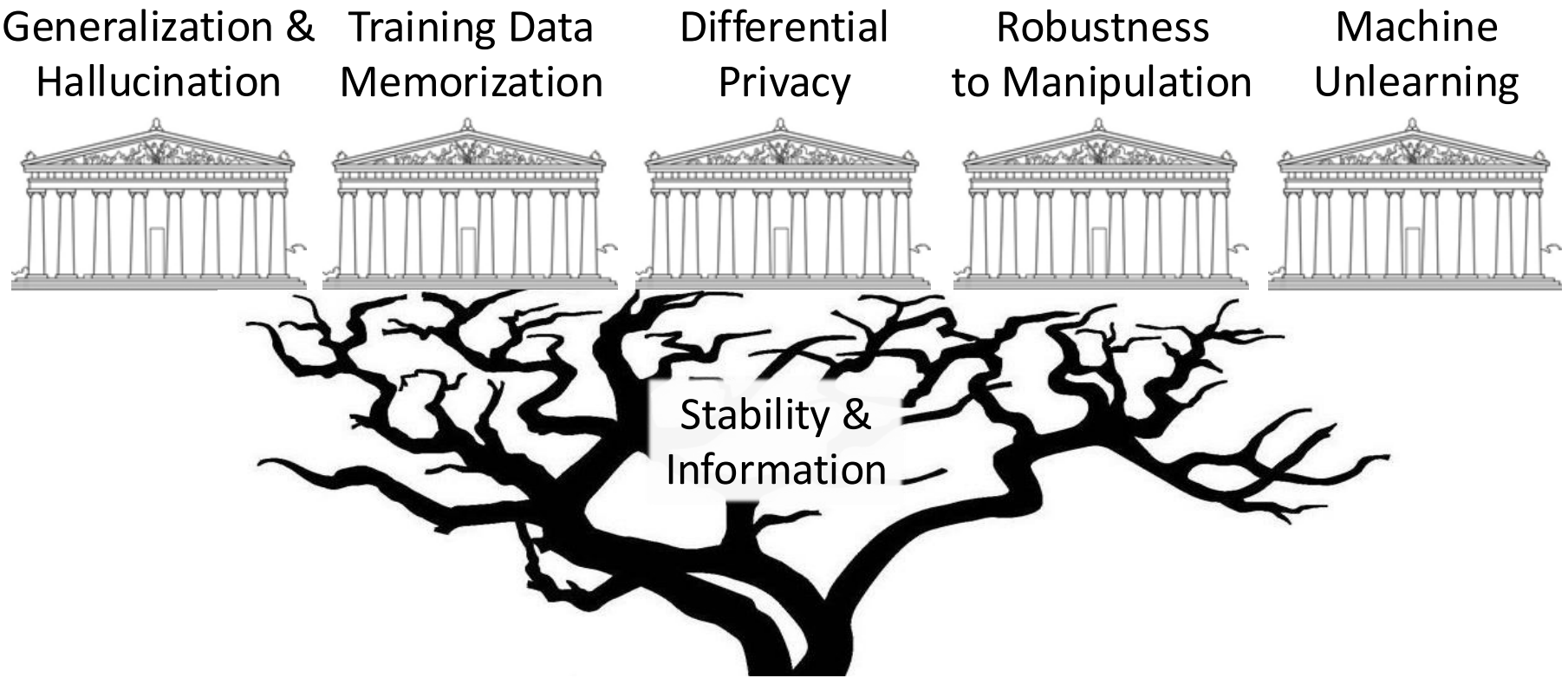
Kunhe Yang
UC Berkeley
kunheyang@berkeley.edu

Preserving Statistical Validity in Adaptive Data Analysis*

Cynthia Dwork[†] Vitaly Feldman[‡] Moritz Hardt[§] Toniann Pitassi[¶]
Omer Reingold[‖] Aaron Roth[**]

Foundations of Trustworthy AI

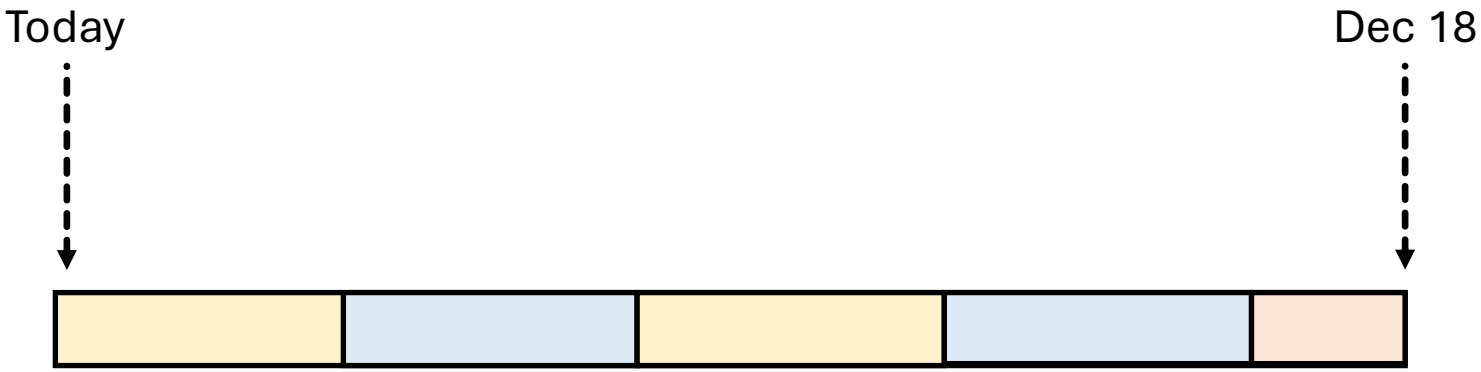
These problems have deep technical connections



Foundations of Trustworthy AI

- What do I want you to get out of this course?
  - Exposure to the problem area
  - More comfort with mathematical reasoning
  - New perspective and ideas for your research
- What do I want you to do in this course?
  - Come, listen, and engage
  - Do an interesting course project
- What do I want to get out of this course?
  - Broaden and deepen my understanding of the area
  - Get to know you all!

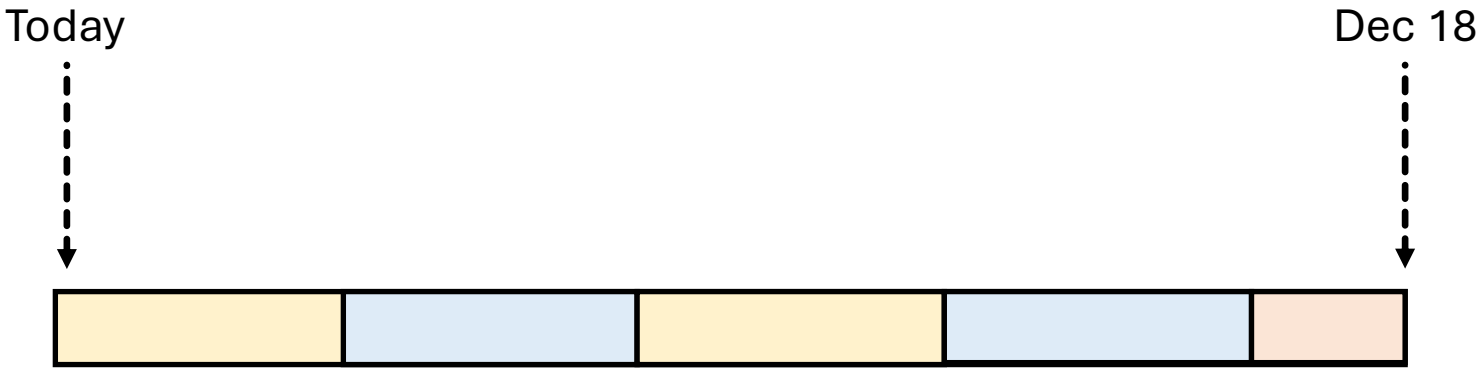
Course Structure



Final Research Project

- Any relevant topic is OK (can complement your research)
- Written report and final presentation
- Working in pairs is welcome
- Milestones and Deadlines: will give more information
- AI policy: be an adult, learn something, stand by your work

Course Structure



Evaluation

- You get out what you put in—no set grade distribution
- Course project is the only formal deliverable
- Small component for attendance and participation (10%)
- What would help you get more out of the course?
  - Scribe notes? Student lectures? Practice problems? Readings?

Course Resources and Misc

- Resources
  - https://jonathan-ullman.github.io/cs7180-f26/index.html
  - No formal reading but I will try to share materials in advance
- Course communications:
  - What do you all prefer?
- Two course numbers:
  - CS7180 (AI breadth) and CS7880 (theory breadth)
  - No other difference between the two courses
- Guest lectures next week by Lunjia Hu!
  - I agreed to speak at a workshop in Vienna a long time ago