Trustworthy Machine Learning

NUS CS5562 - 2023

Introduction

Reza Shokri

About Me

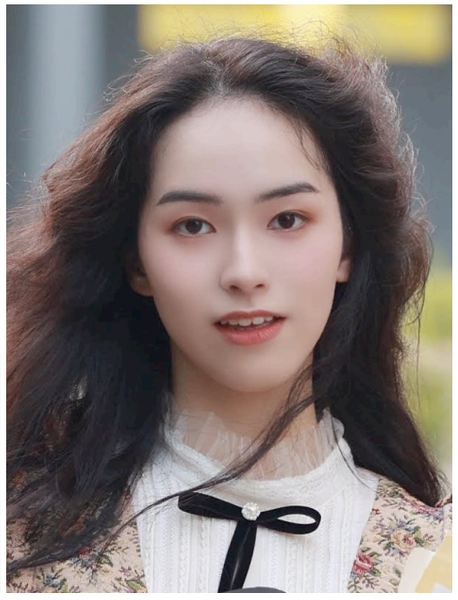
- Research
  - Data Privacy
  - Trustworthy Machine Learning
  - Federated Learning

Teaching Staff

- Lecturer: Reza Shokri
- TAs:
  - Hongyan Chang
  - Martin Strobel
  - Jiashu Tao
  - Yao Tong
  - Jiayuan Ye








How to Communicate with Us?

- All of us participate in the course channels on MS Teams
- Post your questions and messages on the course channel
  - Do not send emails (unless there is a personal issue)
    - We copy emails to the channel, and then respond.
    - There will be Canvas surveys if you want to send anonymous messages.
- TAs will not debug your code.

Course Objectives

- Learn the principles of trustworthy ML
  - Learn how things don't work as expected (in adversarial settings)
- Analyze security and privacy against attacks
- Design trustworthy and privacy-preserving mechanisms

- The topic is very broad, and is an ongoing topic for research
  - This course does NOT cover all aspects of trustworthy ML
    - We focus on the foundations, and major applications
  - There is NO textbook. Reading materials are research papers

Expectations: Required Background

- We assume you are **knowledgable** in machine learning, probability, and statistics
  - This is NOT a course on machine learning
  - We do NOT cover the background on machine learning
- We have an **algorithmic and theoretical approach** to understanding the problems and designing solutions (you need to have the math and algorithms background in ML)
- We have a **practical approach** to test and analyze problems and solutions on machine learning methods (you need to code machine learning algorithms in python)

Main Questions

- So far, you have learned about how ML works.
- This course: **What can go wrong** with machine learning?
  - **Robustness**: Are ML algorithms vulnerable to adversarial manipulations (of their inputs and training data)?
  - **Privacy**: Can ML algorithms leak sensitive information about their training data?
  - **Fairness**: Are ML algorithms biased? Can they discriminate against part of the population?
  - What is robustness? Privacy? Fairness? in the context of ML?
  - How can we design robust, privacy-preserving, and fair models?

This Course

- 3 Parts
  - Part 1. Robustness
  - Part 2. Privacy
  - Part 3. Fairness
- Trustworthy Decentralized Learning
  - It will touch upon robustness, privacy, and fairness

Learning Process

- Lectures
- Reading (research papers)
- Assignments
  - Algorithms and coding
- Quiz
  - Algorithms and theory
- Discussions
- Surveys and continuous feedback

Doing this properly is crucial for learning. You will not learn the topic only by listening to the lectures.

The objective is to help you master the topic. Your grade is the side-product of assignments and the quiz, and not their objective

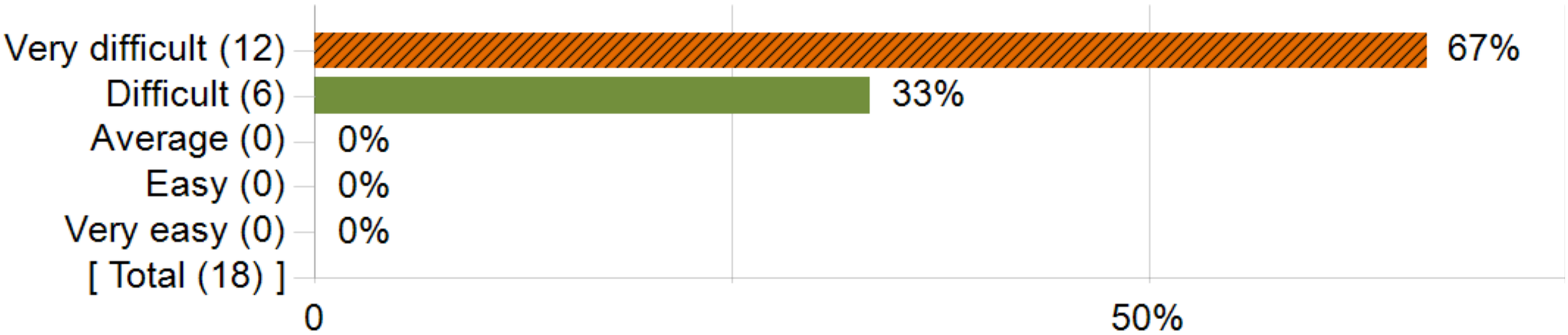
This is how you make sure your understanding is correct. You need to be proactive.

Semester Schedule

Date	Week	Topic	CA
Aug 18, 2023	1	Introduction	Quiz on ML background (closed book)
Aug 25, 2023	2		Assignment 1
Sept 1, 2023	3	Robustness	
Sept 8, 2023	4		Assignment 2
Sept 15, 2023	5	Privacy	
Sept 22, 2023	6		Assignment 3
Oct 6, 2023	7	Privacy	
Oct 13, 2023	8		Assignment 4
Oct 20, 2023	9	Fairness	
Oct 27, 2023	10		Assignment 5
Nov 3, 2023	11	Federated Learning	
Nov 10, 2023*	12		Assignment 6
Nov 17, 2023	13	-	Quiz (closed book)

Note 1: this is a very demanding course!

Difficulty level of the module



Note 1: this is a very demanding course!

Note 2: Note 1 underestimates the difficulty of this course!

Weekly Schedule

- Friday, 16:00
  - Lectures; Quiz; (after class: Releasing assignments)
- Thursday, 9:00 PM
  - Deadline for assignments
- Mondays or Tuesdays (time to be determined soon)
  - Q&A Session with TAs
  - Starts on week 3

Course Elements

- **Lectures**: in classroom
- **Slides, Videos, Readings**: on Canvas
- **Quizzes**: in classroom, on Canvas
- **Assignments**: on Canvas
- **Surveys**: on Canvas
- **Discussions**: on Microsoft Teams
  - You are automatically added to the team for this course

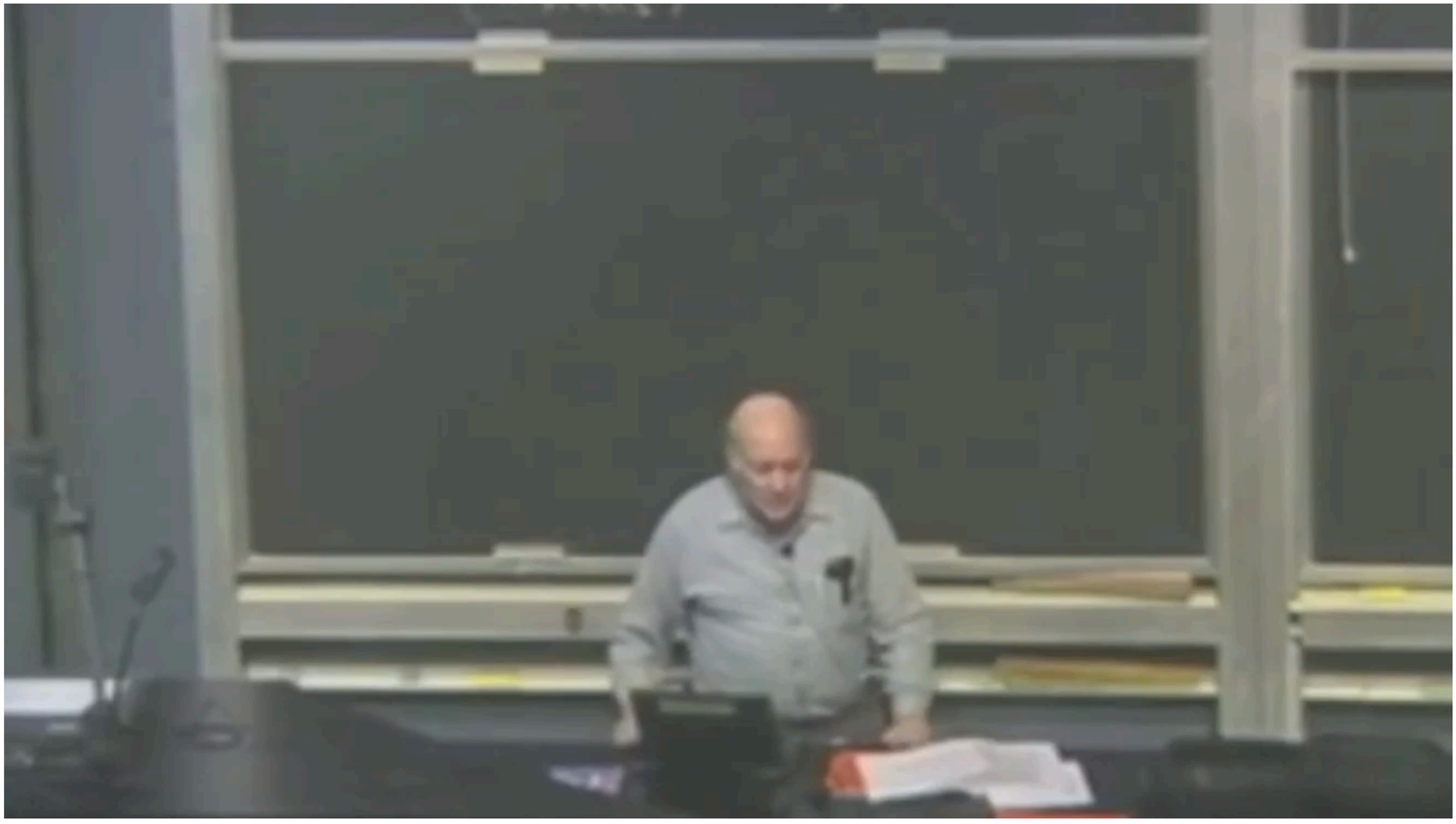
Grading

- (A) Active participation (in discussion channels): 10 bonus points
  - Ask good questions; Help answering others's questions
- (B1, B2, B3) The grade for each part of the course:
  - Assignment: 60 points (No late submission is accepted, even if it is due to personal issues, system crash, network failure, etc).
  - Quiz: 40 points (No make-up or substitute quiz: oral exam for those missing due to medical reasons)
    - There will be one quiz with 3 parts (Robustness, Privacy, Fairness) (note: federated learning covers all these topics)

Grading

- We cover 3 topics. Not everyone has the same background and interest. I want you to have a great knowledge in at least 2 parts of the course, yet have a good understanding of all 3 parts (balancing depth vs breadth).
  - If B1 and B2 and B3 are **all** larger than 50 (out of 100)
    - B = average of your two larger grades (among B1, B2, B3)
  - Else, B = average of all three grades (B1, B2, B3)
- Final Grade: min(A + B, 100)

- NUS code of conduct will be enforced (Any cheating = F mark)



Brian Harvey, UC Berkeley

Teaching Philosophy

- It doesn’t matter what we cover,
  - What matters is what you discover.
- Lectures provide, what you can think of, as pointers.
  - Teach: from an Indo-European root shared by Greek deiknunai ‘show’, deigma ‘sample’.
- You will acquire knowledge if you follow the direction.
  - Study: based on Latin studium ‘zeal, painstaking application’.
- Discuss, do research, redo the proofs, perform experiments, … Ignore the grade. Ignore it!

Noam Chomsky, MIT

Learning Process - Lectures

- The objective of the lectures is to introduce new high-level concepts, and teach you how to think in an adversarial way, and to have an accurate understanding of concepts in trustworthy ML.
- How to maximize your learning?
  - Attend the lectures.
  - Take advantage of the Q&A during the lecture (between different segments and at the end of the lecture). Ask questions.

Learning Process - Readings

- We provide you with the list of the reading material for each lecture (references are provided in the slides. I highlight some of them as reading assignments for you)
- Reading materials are all research papers
- The objective is that you read the details of what was covered in the lecture, and also to improve your (theory) background
- How to maximize your learning?
  - Read the material each week after the lectures

How to Read Papers

- Research papers are not written in a way as the textbooks of well-established topics in computer science
- Do NOT read the paper in one go

1. First read the abstract and introduction
    - What problem does the paper try to solve? What is the main approach? What are their main claimed results?
2. Skim through the figures and tables and main theorems
    - What are the main results? How do they present them?
3. Go back and read the whole paper.

How to Read Papers

- Note: In general, if you want to expand your knowledge in an area or you are doing research on a topic, you may end up not reading the full paper (after finishing the previous steps). This is because the paper might not convince you that it is actually a good paper. This does not apply to papers that we assign to you to read, as they are great papers.
- Ask us if you still have (non-technical) issues with reading papers (It's OK. Reading papers might not be easy at the beginning)

Learning Process - Assignments

- Each part of the course will have multiple assignments
  - You have 2 weeks to do each assignment.
  - You do assignments throughout the semester.
- Assignments need to be done individually (zero collaboration)
- The objective is to challenge your knowledge and help you stretch your ability to analyze and solve the problems, and internalise what you learned.
- **Assignments mostly focus on algorithms and coding**. There are a **few theory assignments** as well.
- How to maximize your learning?
  - Start working on the assignments as soon as they are released
  - Ask any clarification question you have on MS teams to discuss with TAs and others to have a good understanding of the problems.

Learning Process - Quiz

- We run (closed book) quizzes over Canvas
  - We devote the lecture time to it
- The objective is to further challenge your knowledge and help you stretch your ability to analyze and solve problems
- **The quiz will mostly focus on algorithms and theory (proofs and analysis)**
- How to maximize your learning?
  - Prepare for the quiz by going over all the reading material throughout the semester. There is just too much to study if accumulated at the end of the semester.

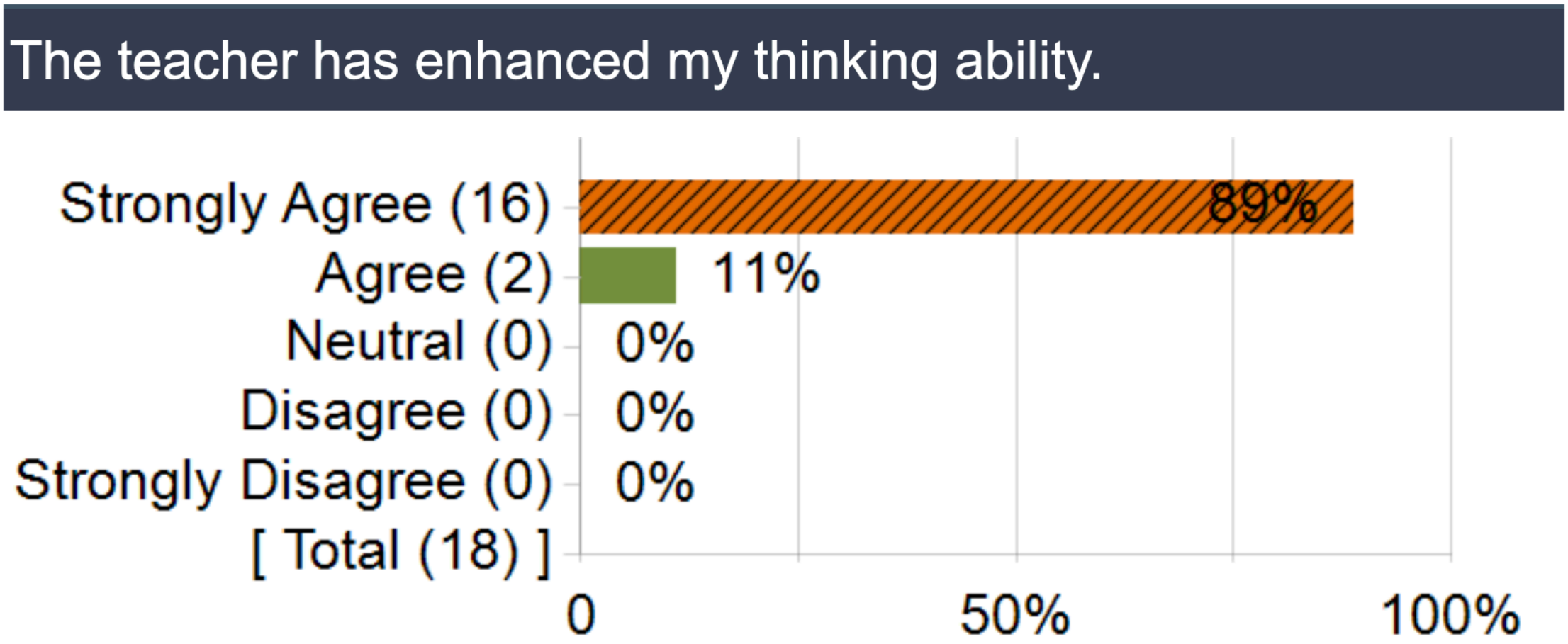
Learning Process - Discussions

- There will be one channel on MS teams for “logistics”
- Thee will be one channel for each part of the course to discuss the technical points and questions
- The forum enables you to discuss with other students, TAs and me. You can ask questions and answer each others’ questions.
- There are up to 10 bonus points for those who participate substantially
- How to maximize your learning?
  - Follow the channels, and actively participate in the technical discussions.

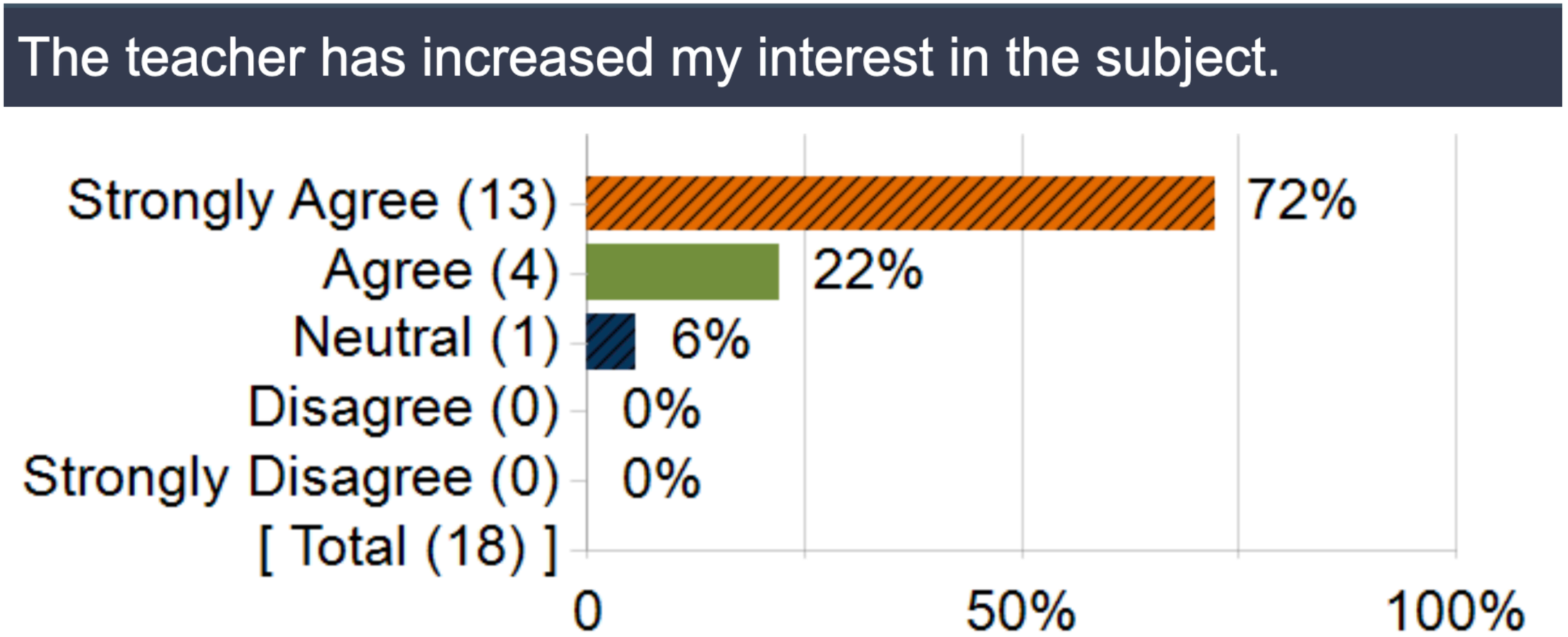
Learning Process - Surveys

- There is a survey on Canvas that you can use to give feedback about the course. We monitor it constantly.
- There will be one separate survey after each part of the course.
- How to maximize your learning?
  - Communicate what works and doesn’t work for you.

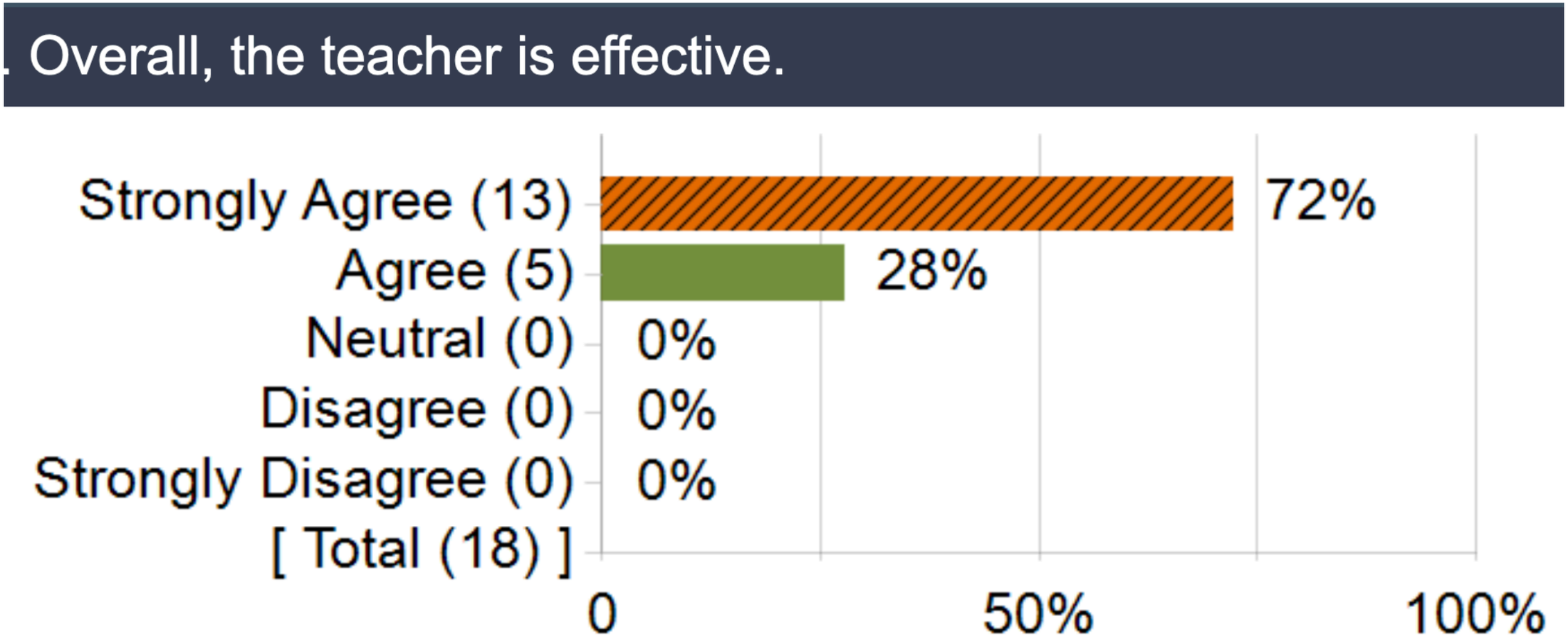
Overall opinion of the module



Overall opinion of the module



Overall opinion of the module



Faculty Teaching Excellence Award 2023

My Verdict:

I loved the teaching style of Prof. Reza Shokri. The class participation was extremely high and we had wonderful discussions in class. The TA were extremely helpful.

This course has a lot of prerequisite knowledge requirement, in probability, statistics and Machine Learning. It is by no doubt very challenging. The assignments required lot of effort, from reading papers to writing code and easily took more than 4-5 days.

Despite these challenges I can say, this is one of the best courses I have taken in NUS.

Last Words

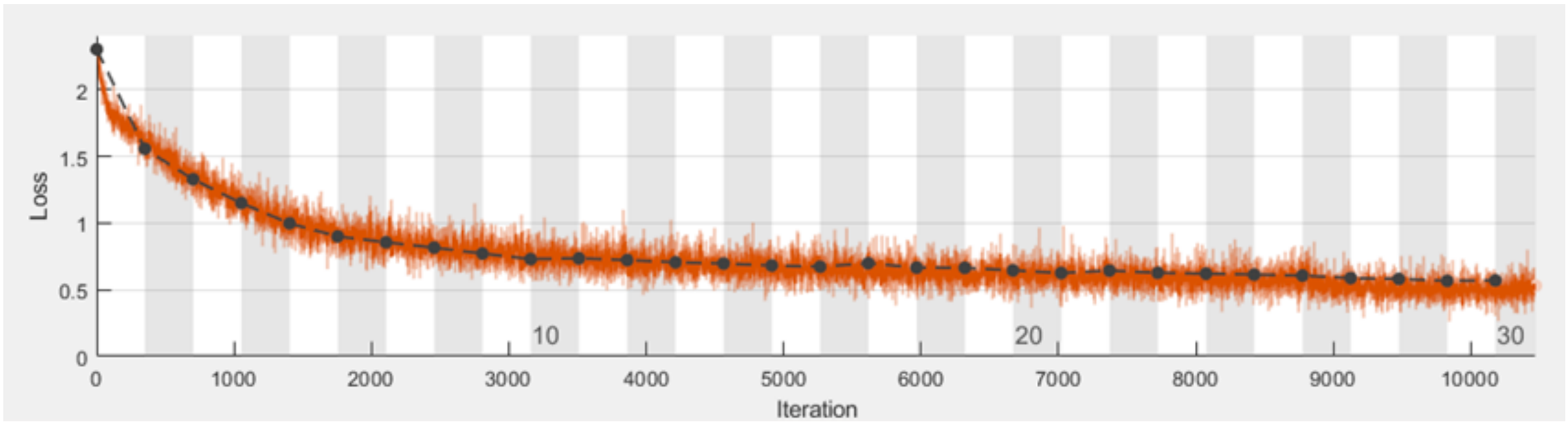
- Trustworthy Machine learning is among the most interesting fields in computer science, and it is well connected to many sub-fields in computer science and other disciplines
- This course is designed to help you to acquire knowledge (not information) in trustworthy ML, and develop a security mindset (what can go wrong).
- Given that ML is increasingly being used in critical decision making processes, it is crucial for us as computer scientists to know the ethical, privacy, and reliability issues with ML.
- **We are responsible for the technology that we build and use.**
- Enjoy the ride!

A Very High-Level Overview of the Technical Content

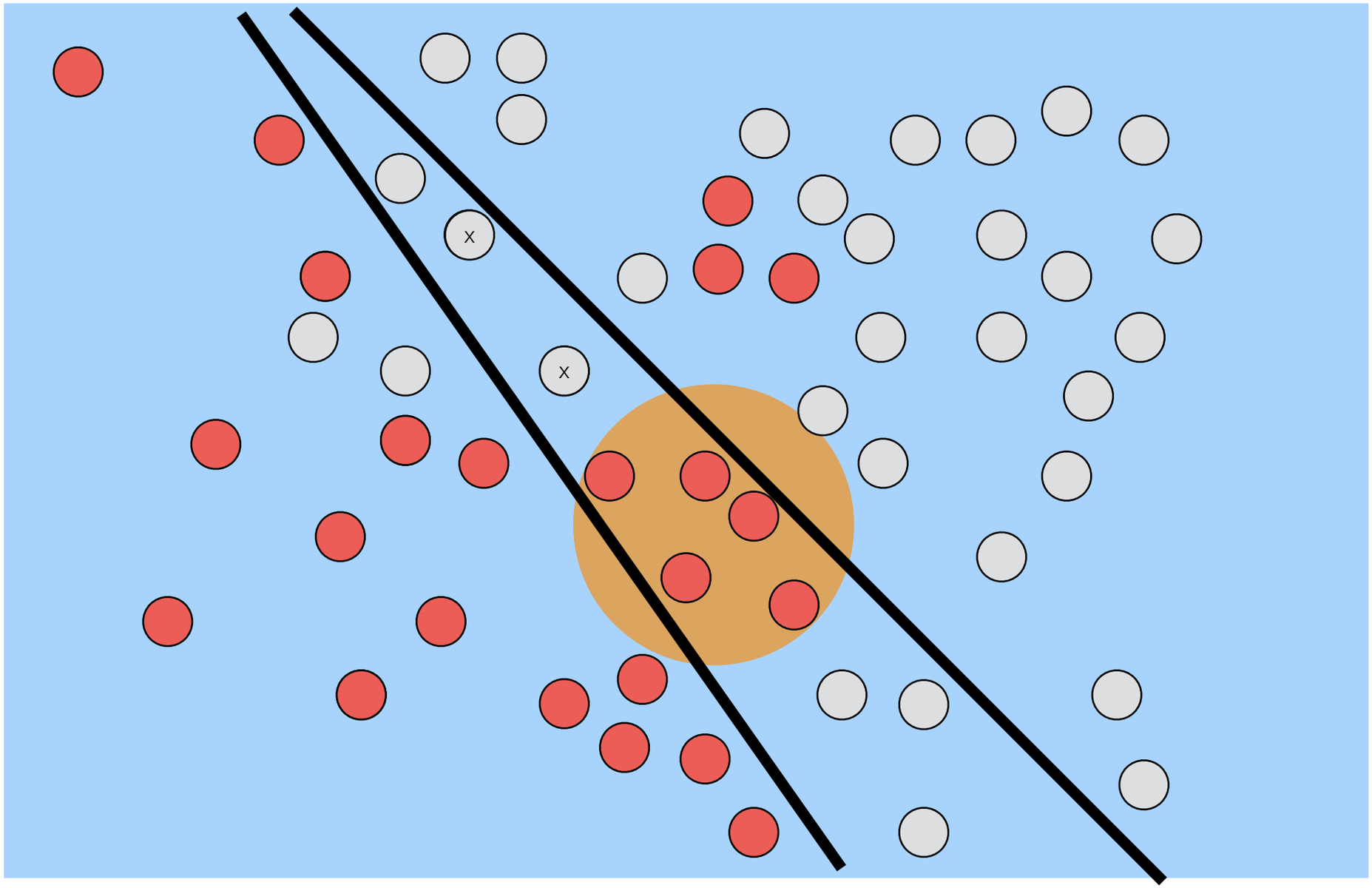
Robustness
Privacy
Fairness

Machine Learning

- Minimize the learning loss
- Maximize the predictive power

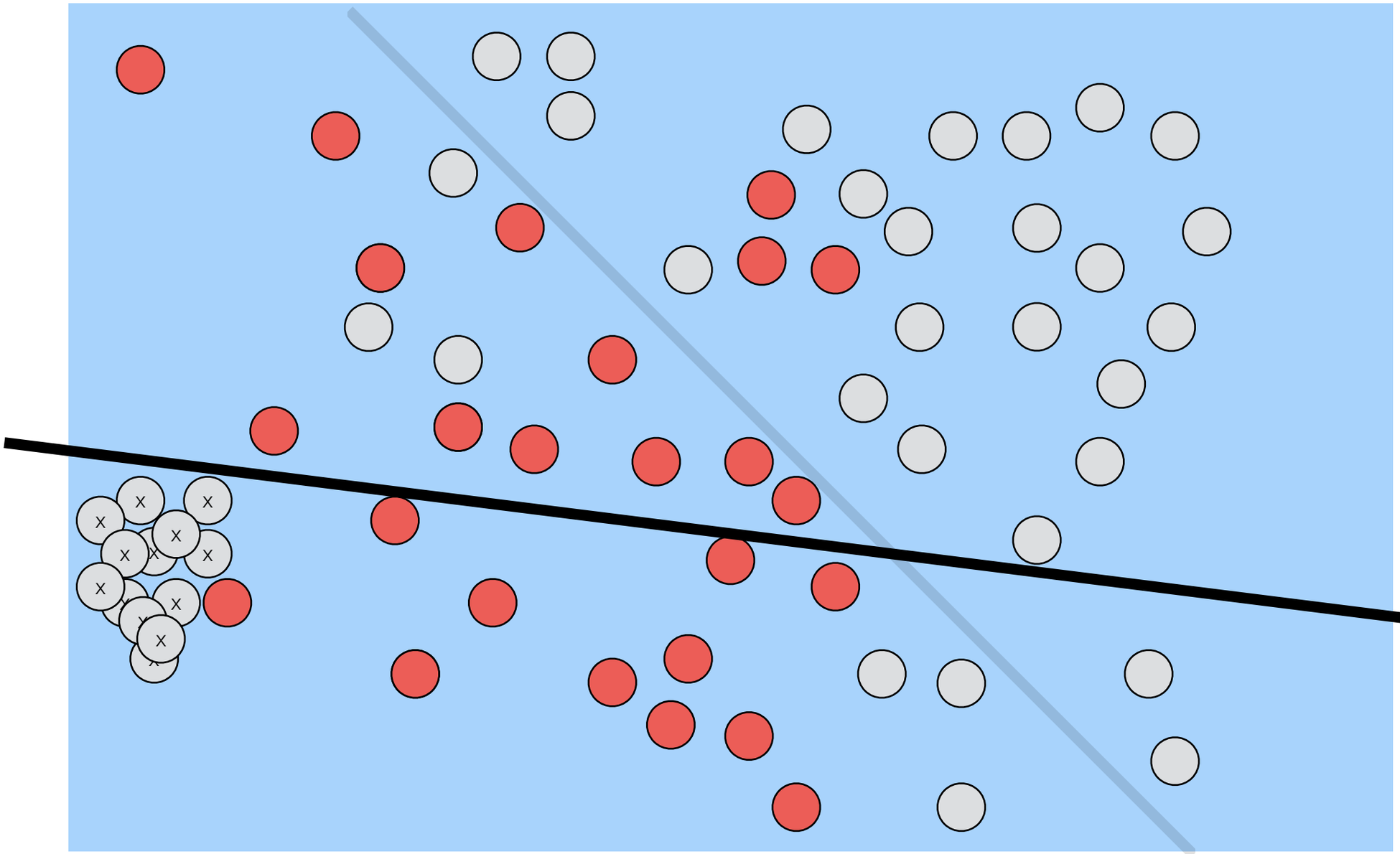


Robustness (Training)



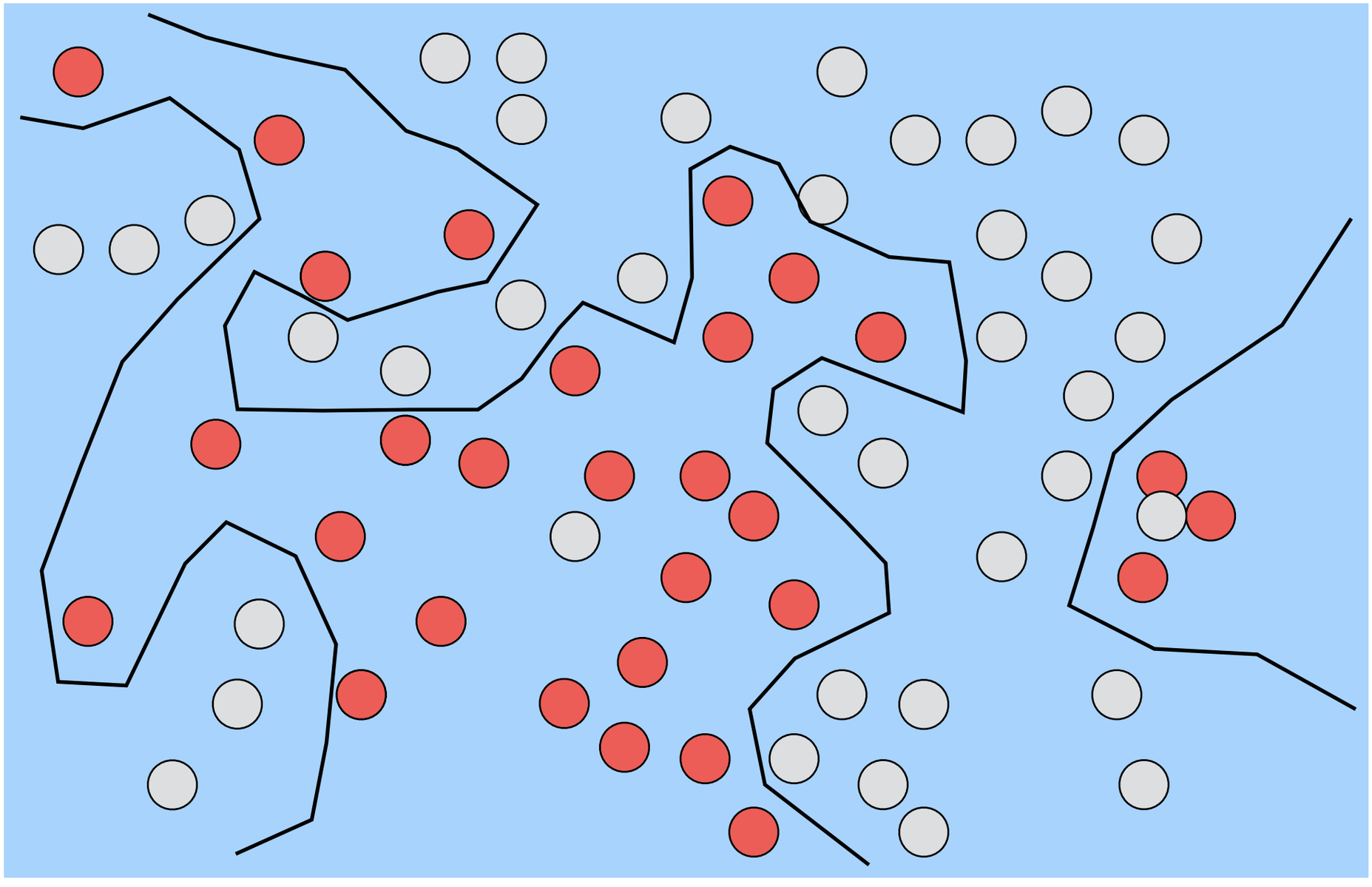
Data Poisoning Attacks

Robustness (Training)

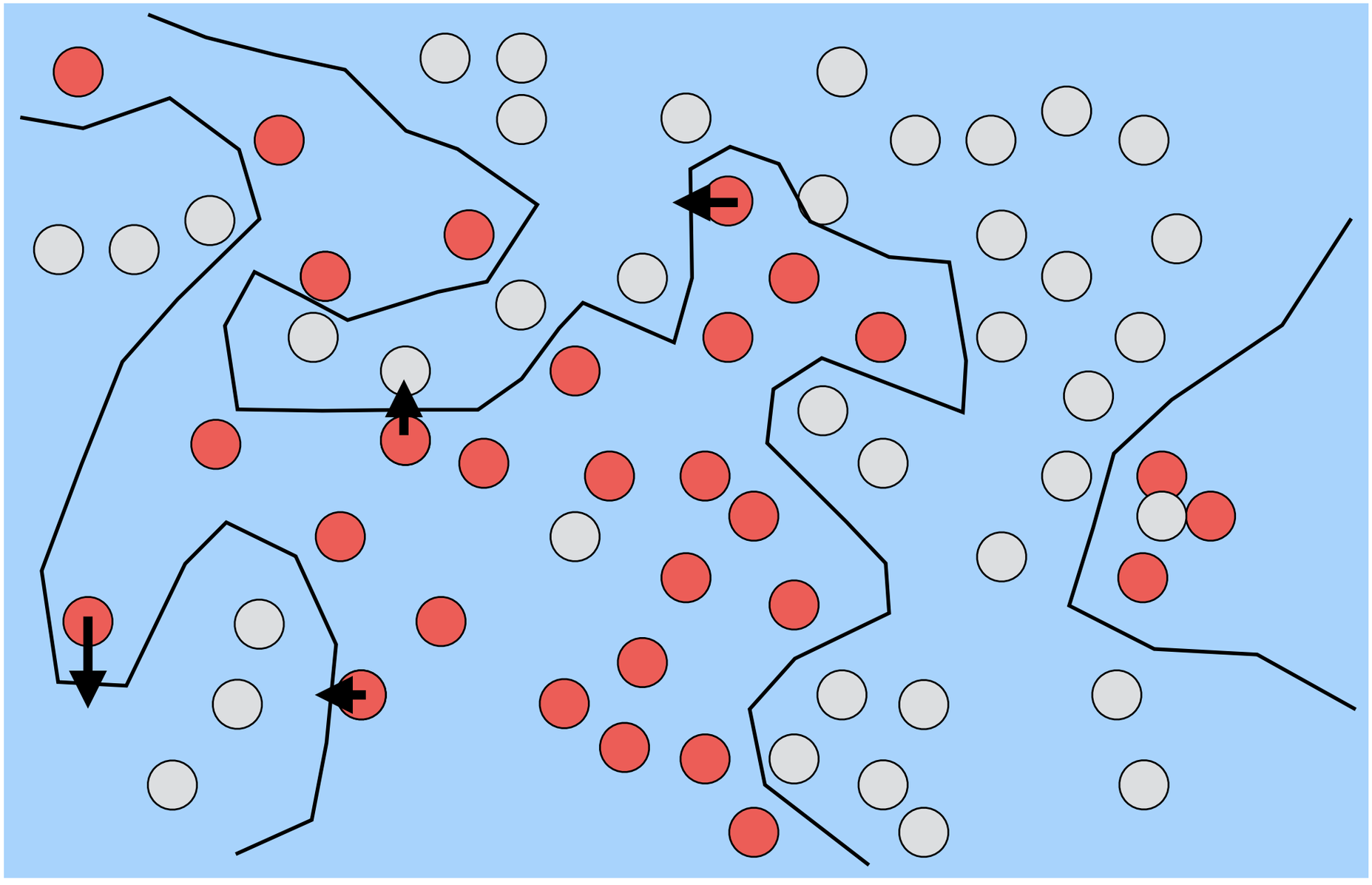


Data Poisoning Attacks

Neural Networks



Robustness (Inference)



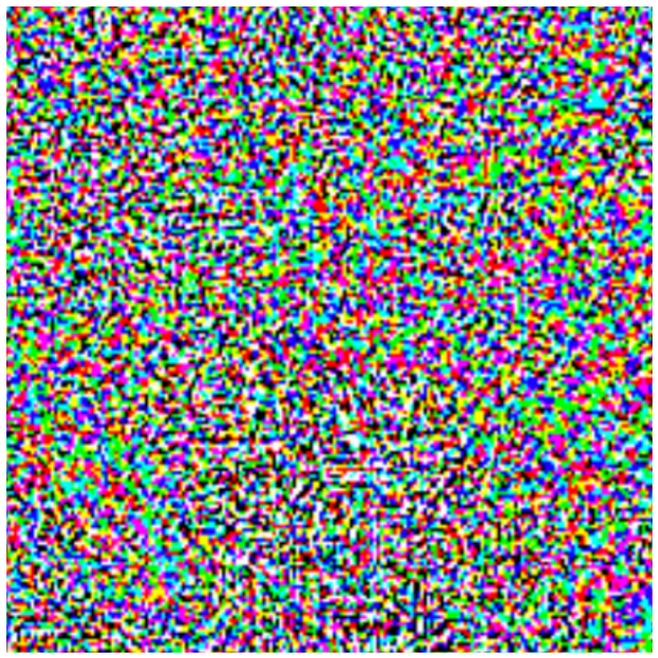
Adversarial Examples

Adversarial Perturbation

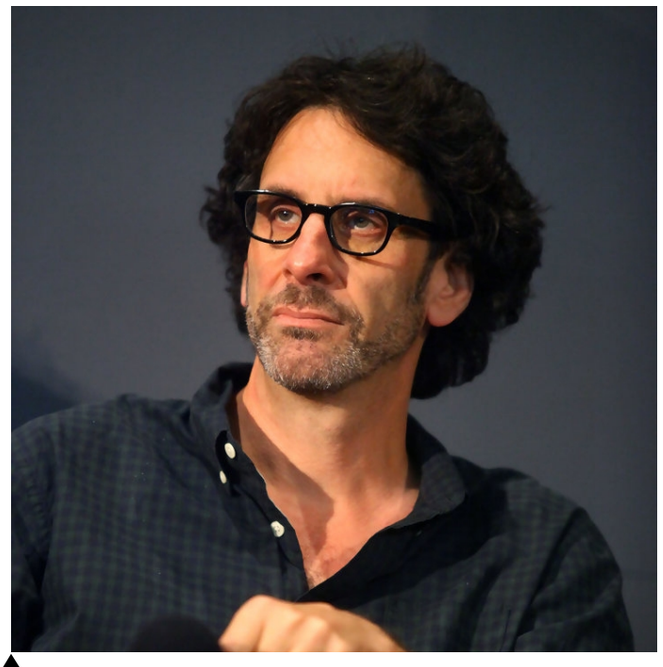
Adversarial examples

Yasujirō Ozu

+

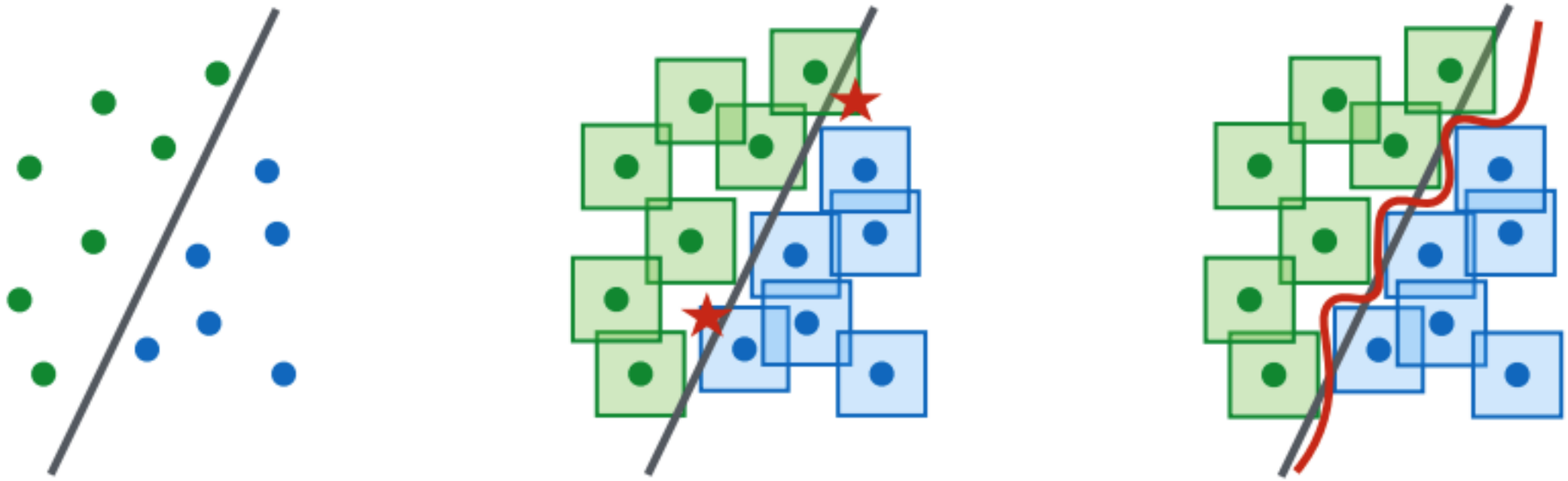


Ethan Coen



Joel Coen

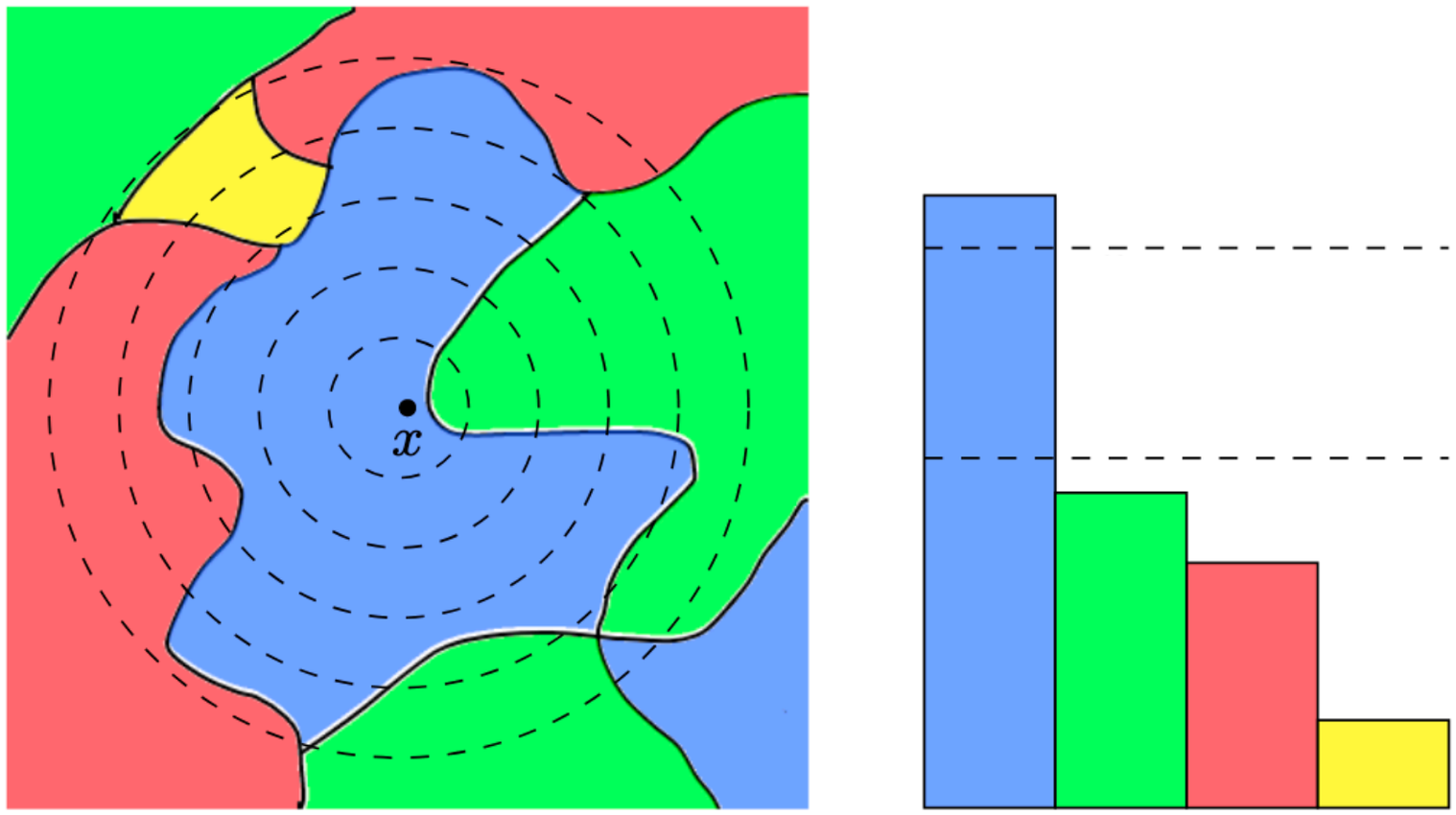
Train Robust Models: Optimize the Model for Robustness



Madry, et al., Towards Deep Learning Models Resistant to Adversarial Attacks, 2017

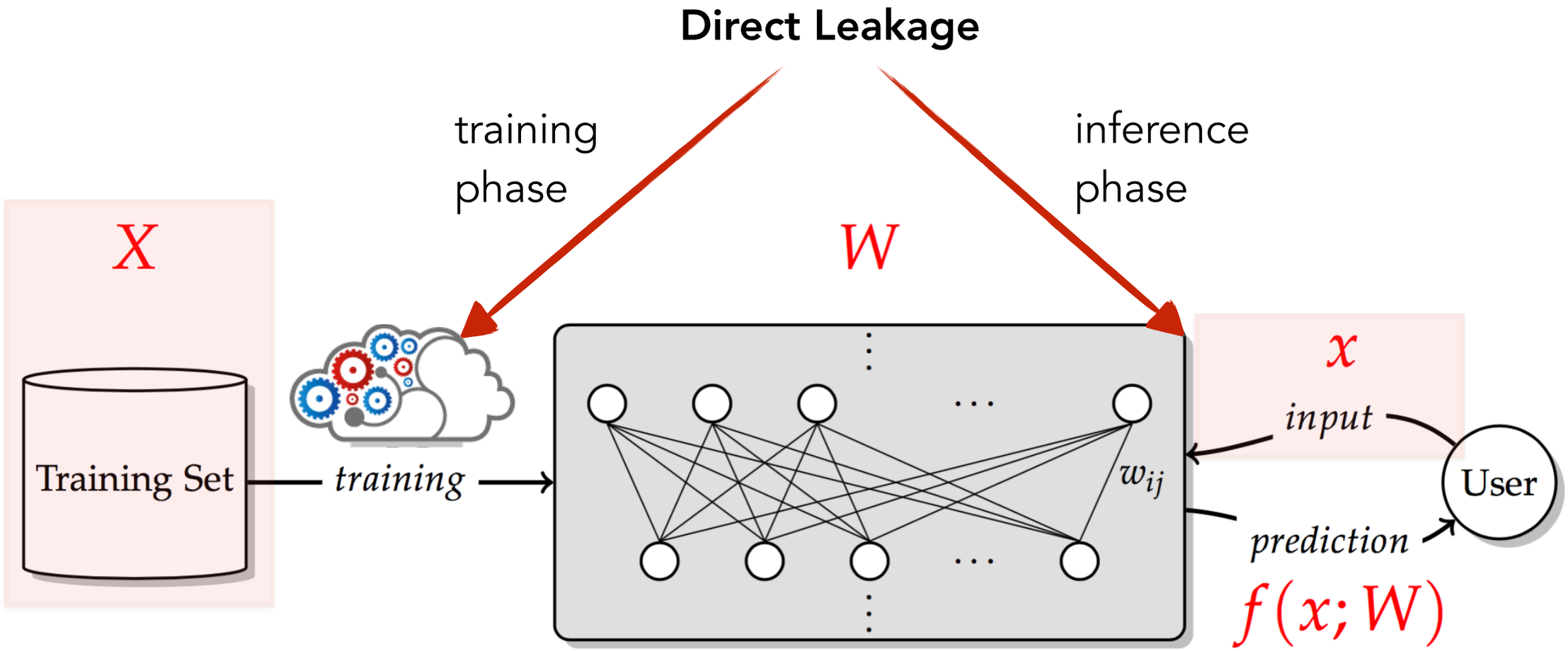
Train Robust Models: Randomized Smoothing

- Randomize the query and output the average prediction
- Prove robustness



Cohen, et al., Certified Adversarial Robustness via Randomized Smoothing, 2019

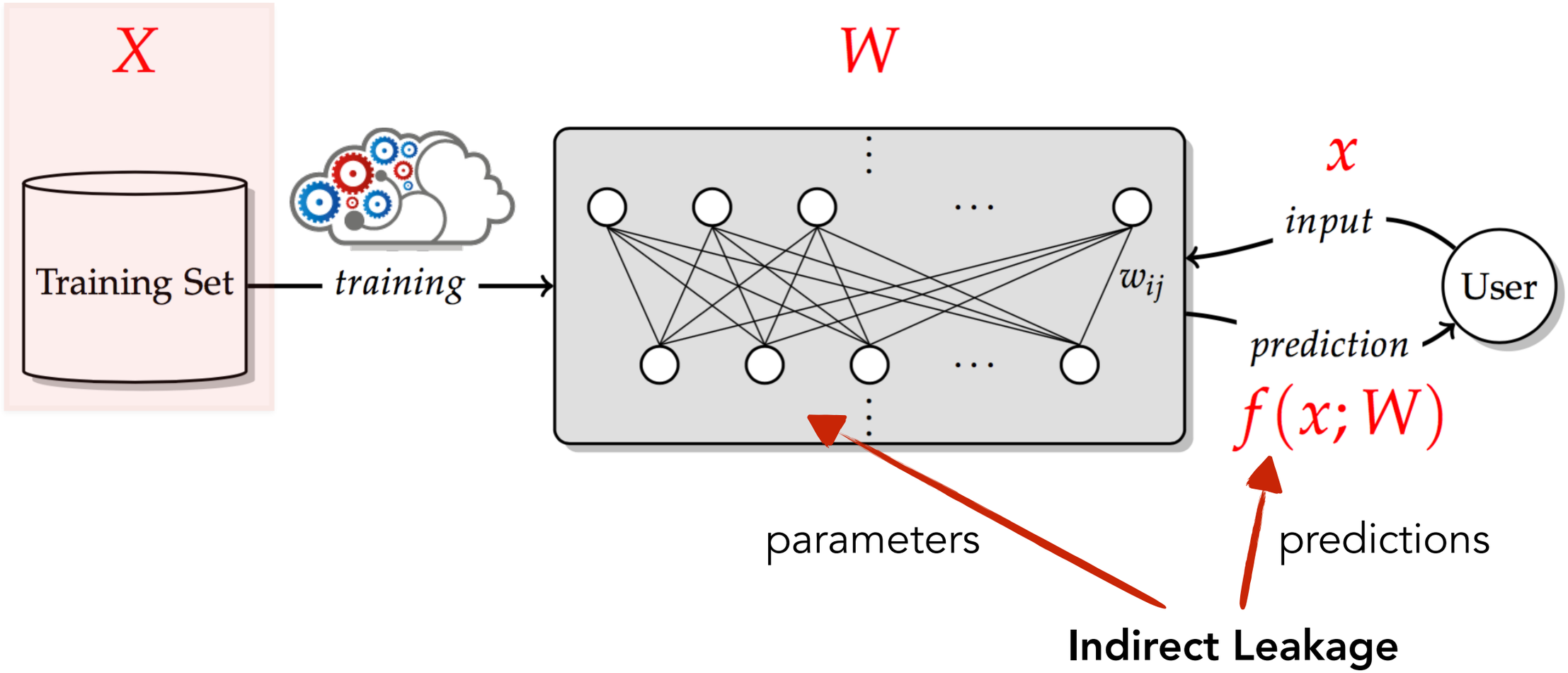
Privacy Risks in Machine Learning



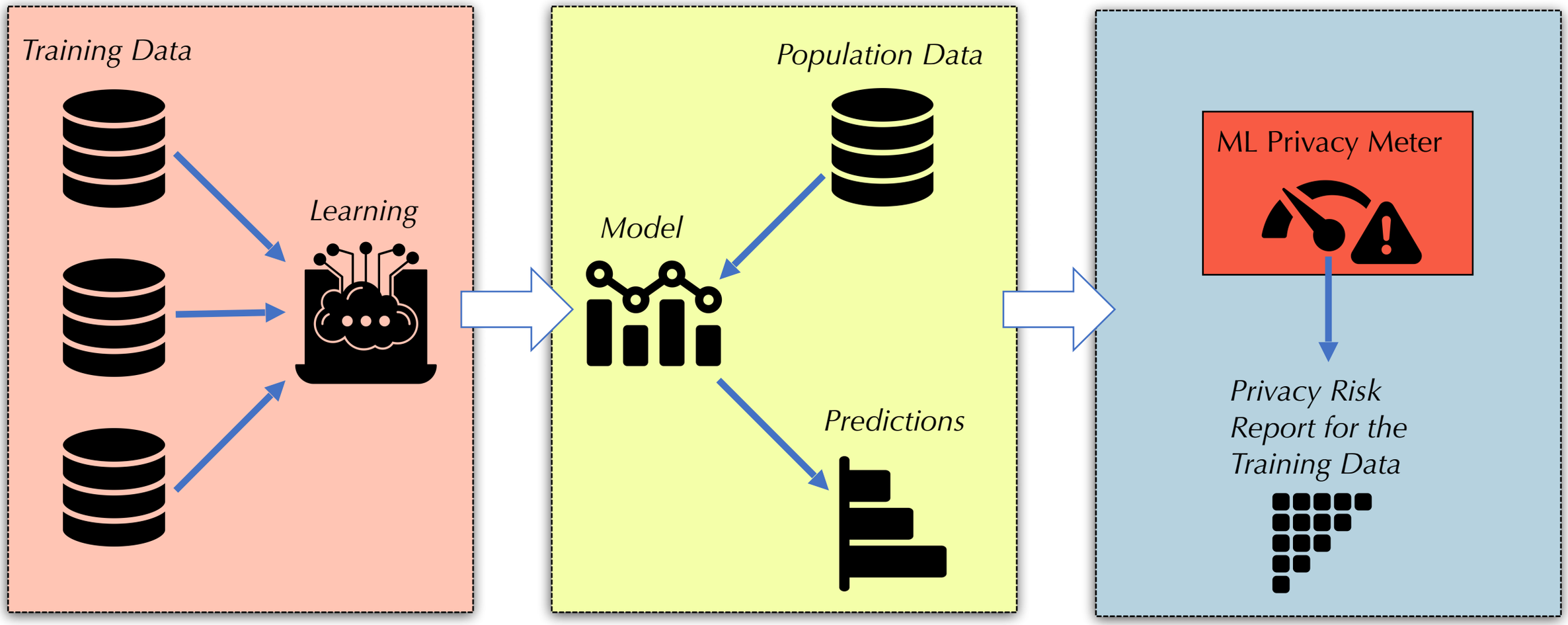
How to prevent this leakage? Secure multi-party computation, homomorphic encryption, trusted hardware, …

Privacy Risks in Machine Learning

What is leakage? Inferring information about members of X, beyond what can be learned about its underlying distribution



Tool: ML Privacy Meter



ML Privacy Meter is a Python library (`ml_privacy_meter`) that enables quantifying the privacy risks of machine learning models. https://github.com/privacytrustlab/ml_privacy_meter

Differential Privacy

- A randomized algorithm $\mathscr{A}$ satisfies (ϵ, δ)-DP, if for any two neighboring datasets D, D', and all sets S

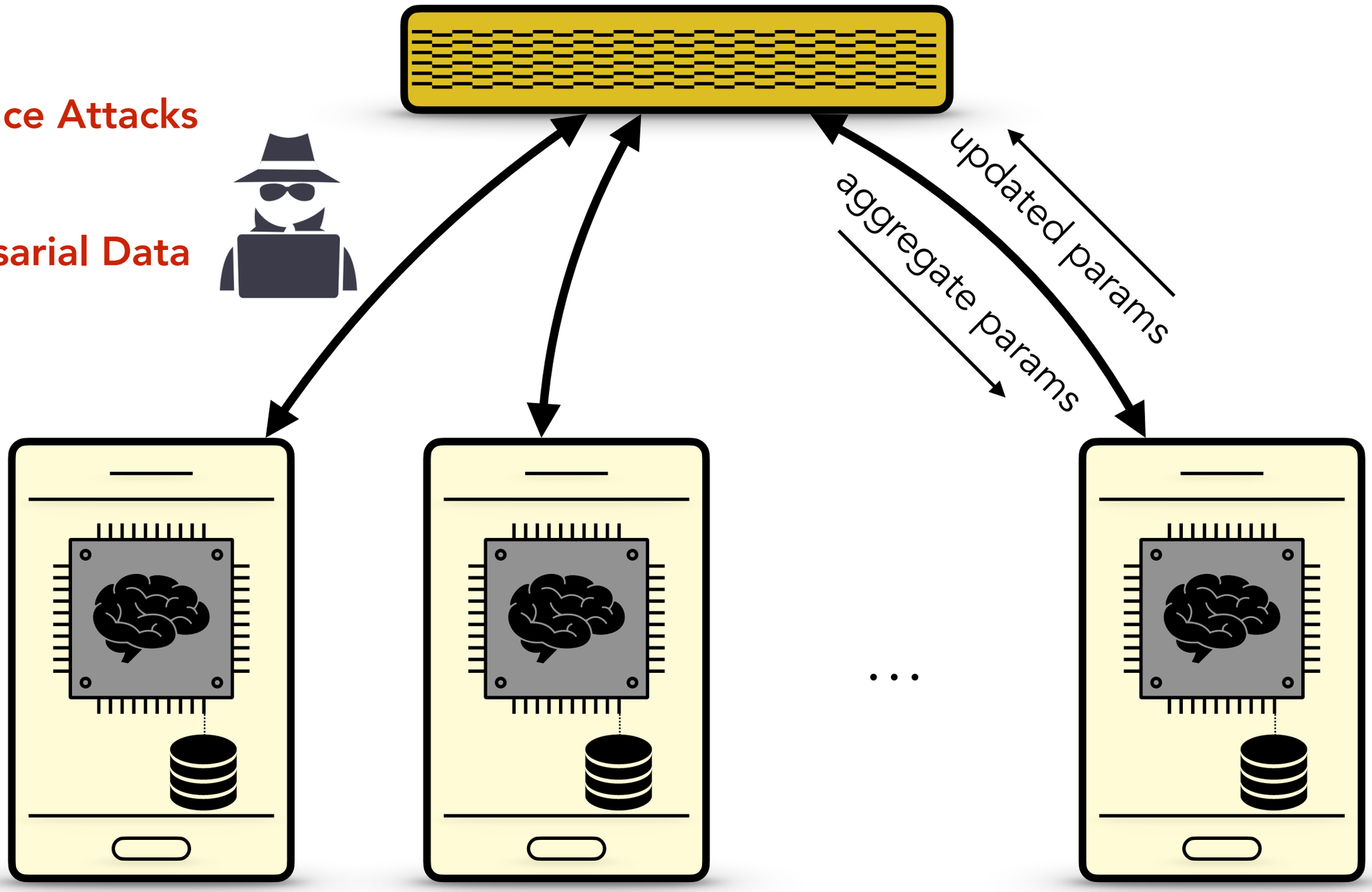
$$\Pr[\mathscr{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathscr{A}(D') \in S] + \delta$$

[Dwork, McSherry, Nissim, Smith] Calibrating Noise to Sensitivity in Private Data Analysis, TCC 2006

Decentralized (Federated) Learning

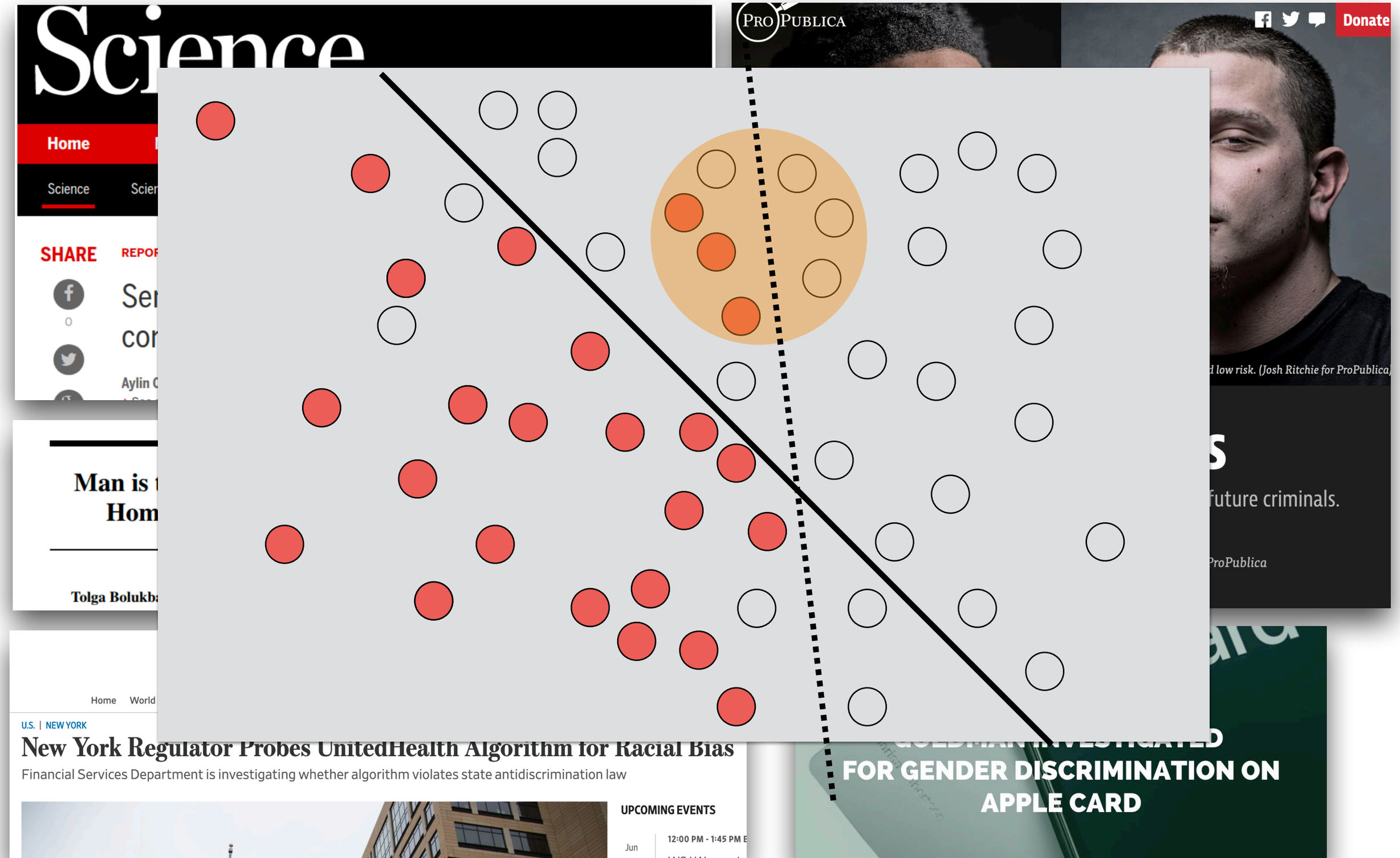
Inference Attacks

Adversarial Data



[Shokri and Shmatikov] Privacy-Preserving Deep Learning, 2015

Algorithmic Bias - Fairness



A School for All Seasons on Trustworthy Machine Learning

https://trustworthy-machine-learning.github.io/

(New papers which are not here will be shared with you on slides)