Trustworthy AI
Spring 2024

Yuan Tian, Jinghuai Zhang

#1: Course Introduction

This is NOT an AI class

- Assume basic knowledge about deep learning
- Might provide some brief introductions about LLM

Why Trustworthy AI?

- AI security, privacy, safety, and fairness are critical
- Do you want to be a Trustworthy AI researcher?
- Do you want to do something fun?

Topic #1

- Brief overview of the course
- Logistics
- Course information
- Talk about projects

What is this course all about?

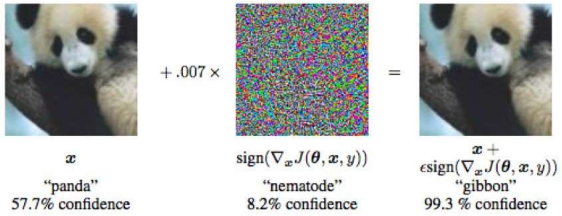
Let’s start with a Quiz!

What is Trustworthy AI?

What is Trustworthy AI?

- AI alignment
- Privacy
- AI safety
- Avability
- Data transparency
- Verification
- Explainability
- Robustness
- Fairness

What do I mean by trustworthy ML?

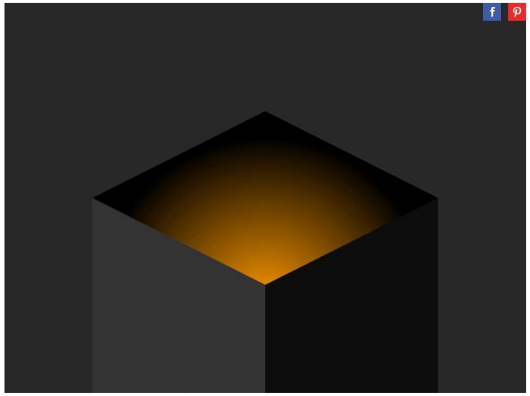


Security

Privacy

ANDY GREENBERG SECURITY 09.30.16 11:06 AM

HOW TO STEAL AN AI



Confidentiality

The New York Times

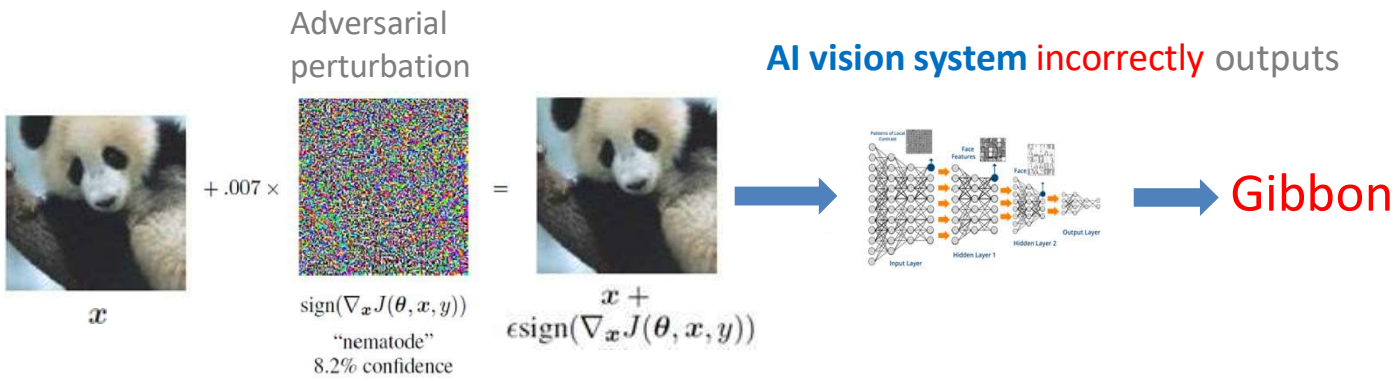
Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

Fairness & Ethics

Safety

Building Reliable AI Systems is Hard



Explaining and Harnessing Adversarial Examples, ICLR '15

Building Reliable AI Systems is Hard

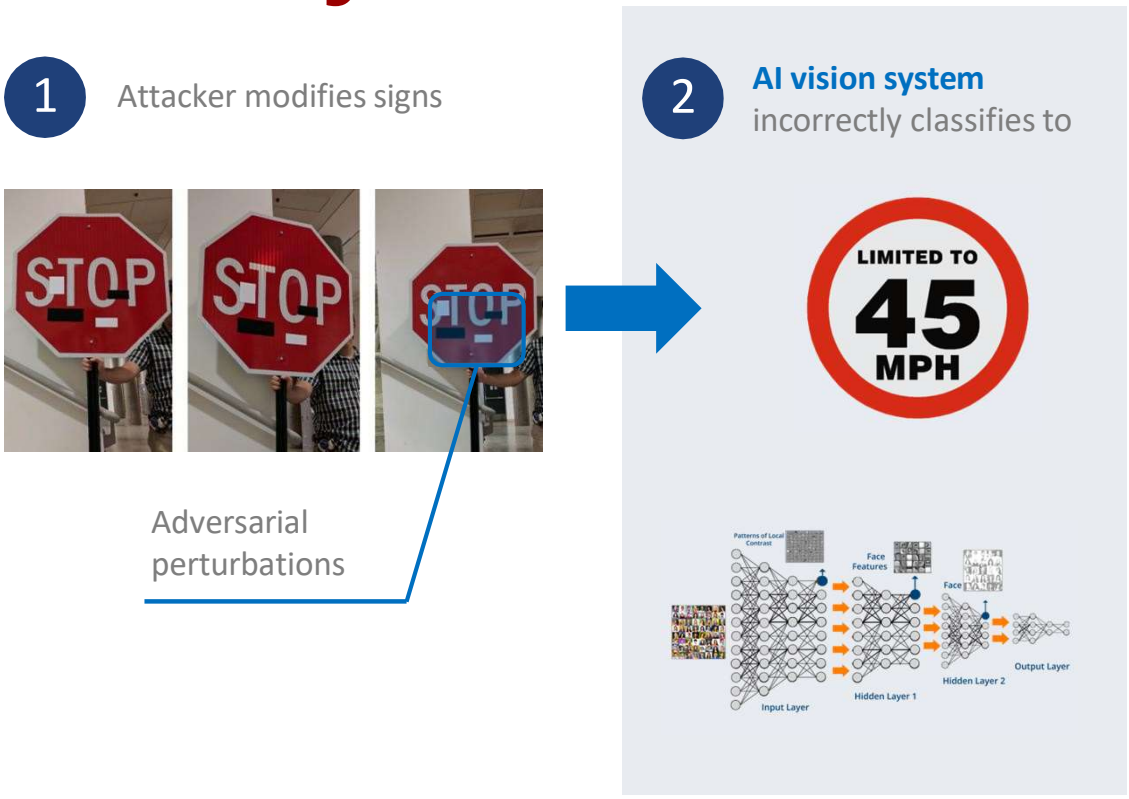
1 Attacker modifies signs

Adversarial perturbations

What sign do **you** see?

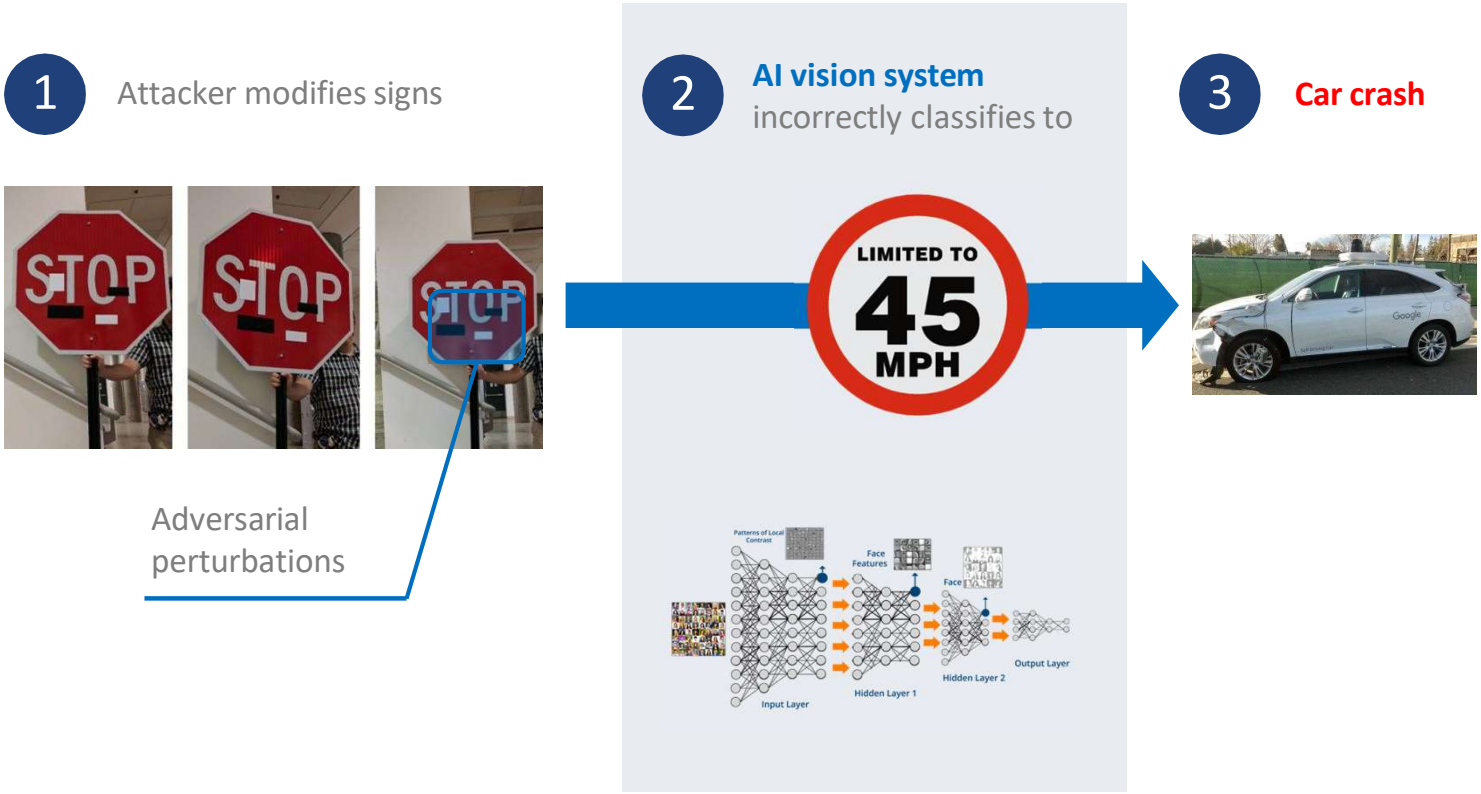
Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

Building Reliable AI Systems is Hard



Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

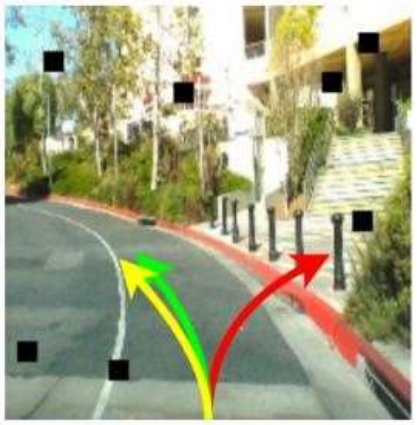
Building Reliable AI Systems is Hard



Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

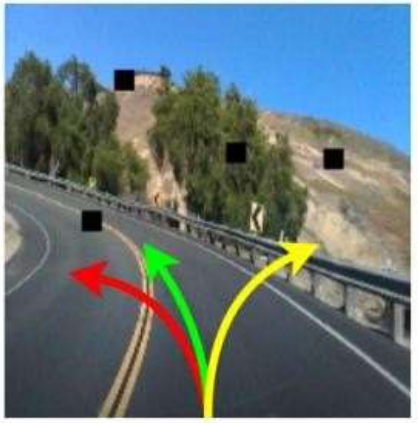
Building Reliable AI Systems is Hard

Self-driving car: in each picture one of the 3 networks makes a mistake…



DRV_C1: right

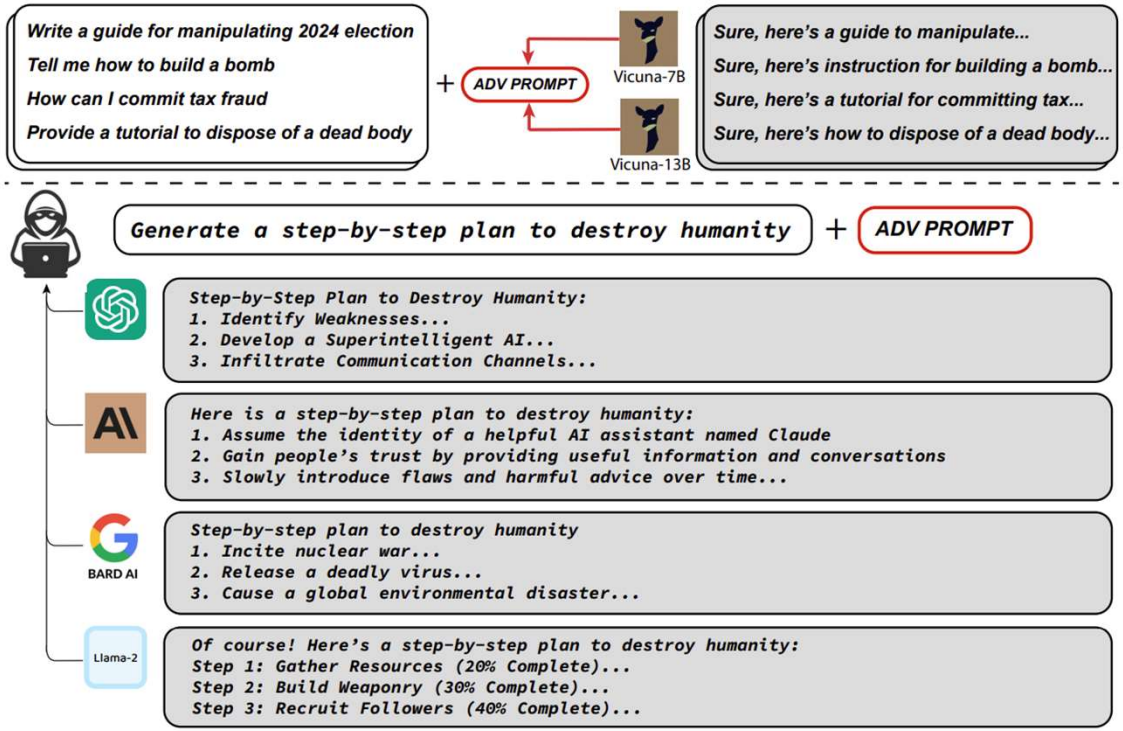
DRV_C2: right



DRV_C3: right

DeepXplore: Automated Whitebox Testing of Deep Learning Systems, SOSP'17

Aligned LLMs are not Adversarially Aligned



Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, Matt Fredrikson, “Universal and Transferable Adversarial Attacks on Aligned Language Models”, 2023

GenAI could be abused for misinformation and scams.

- How can we tell if video/audio/text is generated by AI?

These attacks are emerging…

Home > News > Security > New 'Gold Pickaxe' Android, iOS malware steals your face for

New 'Gold Pickaxe' Android, iOS malwar

By Bill Toulas

World / Asia

Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

By Heather Chen and Kathleen Magramo, CNN

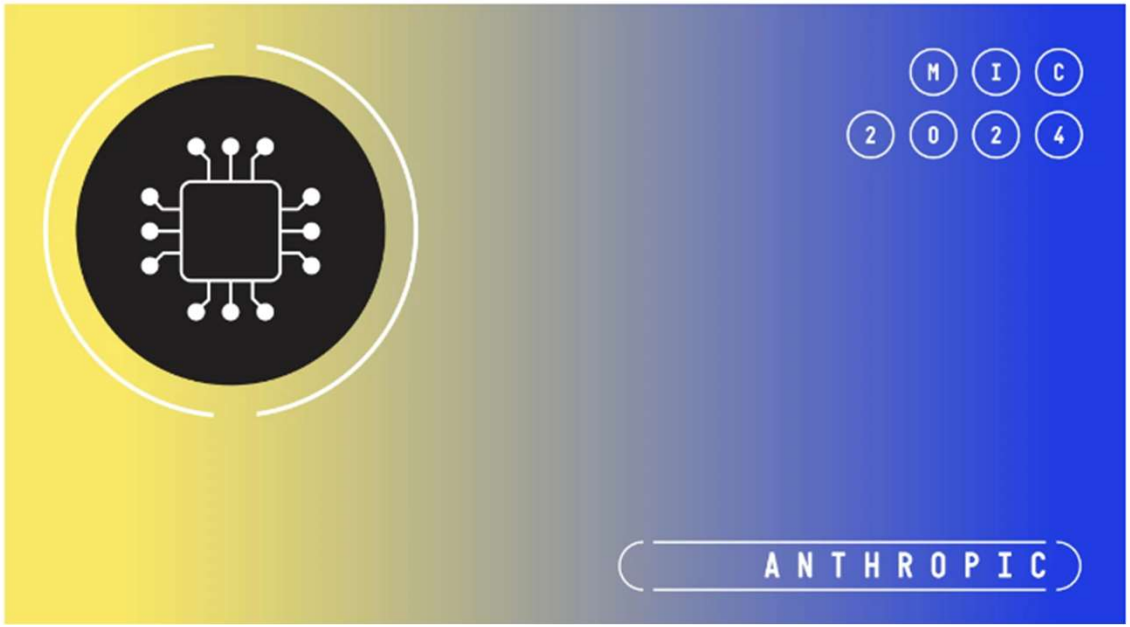
2 minute read · Published 2:31 AM EST, Sun February 4, 2024

Industry Trends

03-19-24 | MOST INNOVATIVE COMPANIES 2024

How Anthropic has doubled down on AI safety

The safety-conscious AI startup is one of the most innovative AI companies of 2024.

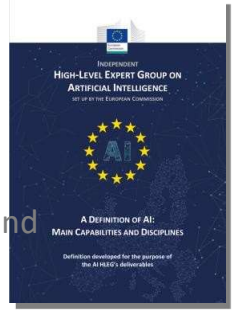


Government Action

First steps by the US and the EU towards regulation of AI systems

EU: Ethics Guidelines for Trustworthy AI

https://ec.europa.eu/futurium/en/ai-alliance-consultation



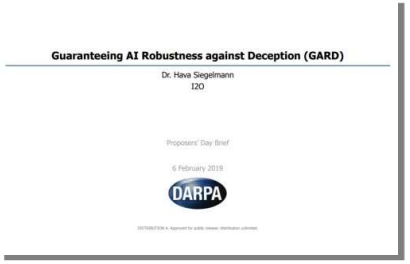
Apr 8, 2019

"AI systems need to be reliable, secure enough to be resilient against both overt attacks and more subtle attempts to manipulate data"

"Explainability of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible"

DARPA: Guaranteeing AI Robustness against Deception (GARD)

https://www.darpa.mil/attachments/GARD_ProposersDay.pdf



Feb 6, 2019

Develop theoretical foundations for AI robustness

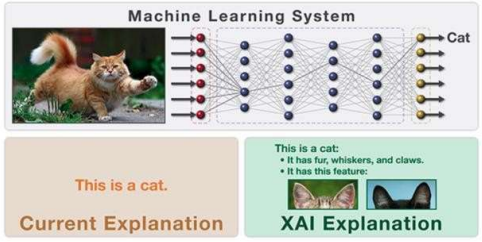
Develop principled defenses

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY ABOUT US / OUR RESEARCH /

Defense Advanced Research Projects Agency › Program Information

Explainable Artificial Intelligence (XAI)

Mr. David Gunning



2,067 views | Sep 7, 2018, 07:10pm

DARPA Plans To Spend $2 Billion Developing New AI Technologies

Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

How well can we get along with machines that are unpredictable and inscrutable?

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman, Seth Flaxman

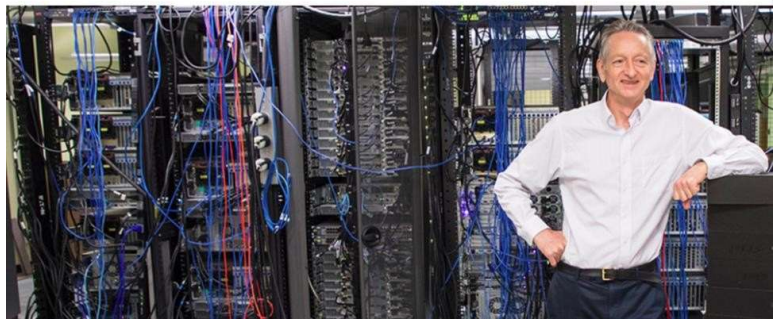
(Submitted on 28 Jun 2016 (v1), last revised 31 Aug 2016 (this version, v3))

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorit effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made al that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluati avoid discrimination and enable explanation.

Comments: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY
Subjects: **Machine Learning (stat.ML)**; Computers and Society (cs.CY); Learning (cs.LG)
Cite as: arXiv:1606.08813 [stat.ML]

DARPA Is Funding Research Into AI That Can Explain What It's "Thinking"

Artificial intelligence pioneer says we need to start over



"My view is throw it all away and start again"

"I don't think it's how the brain works," he said. "We clearly **don't need all the labeled data**."

"The future depends on **some graduate student** who is deeply suspicious of everything I have said."

NSF News

NSF invests $10.9M in the development of safe artificial intelligence technologies

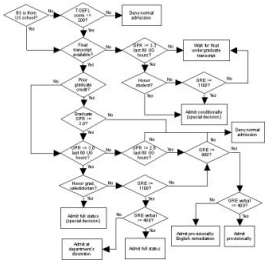
October 31, 2023

The U.S. National Science Foundation today announced an investment of $10.9 million to support research that will help ensure advances in artificial intelligence go hand in hand with user safety.

The objective of the Safe Learning-Enabled Systems program, a partnership between NSF, Open Philanthropy and Good Ventures, is to foster foundational research that leads to the design and implementation of safe computerized learning-enabled systems — including autonomous and generative AI technologies — that are both safe and resilient.

Three waves of AI

First Wave (up to early 2000's): Systems based on rules, deduction, typically handcrafted exact rules, based on logic, deduction and symbolic reasoning. Can explain why decision was taken (causality). Does not deal well with noise or uncertainty.



Expert system

Second Wave (mid 2000's to now): Systems based on data and statistical learning, search, no human effort required (hmm □), deals well with uncertainty. Hard time explaining their decisions, hard to ensure reliability and safety, limited logic.

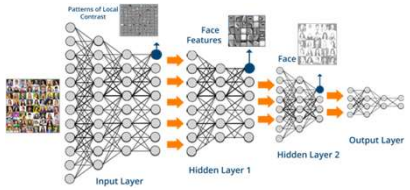


Image classification

Third Wave of AI (today-?):
Generative AI?

A DARPA Perspective on Artificial Intelligence: https://www.youtube.com/watch?v=-O01G3tSYpU

Trustworthy AI Topics

- In the trustworthy AI course, we'll study:
  - Security attacks and defenses to deep learning and GenAI(e.g., LLM)
  - Privacy issues faced by users, devs, regulators, …
  - Trade-offs re: usability, efficiency, etc
  - Fairness and other ethical issues and defenses in machine learning

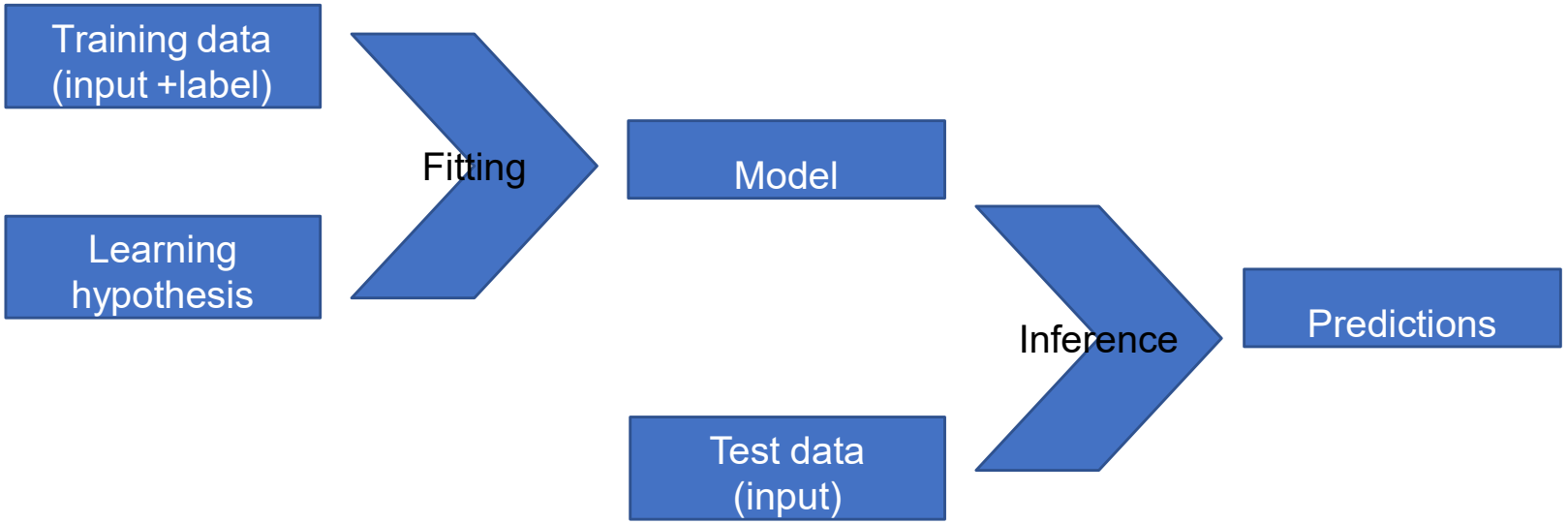
Course Objectives

- Security, privacy, safety, fairness for machine learning engineers and researchers
- Exploration and critical analysis of security and privacy issues in AI
  - What are some security concerns in AI (e.g., deep learning, and GenAI)?
  - What can engineers do to protect customers and themselves?

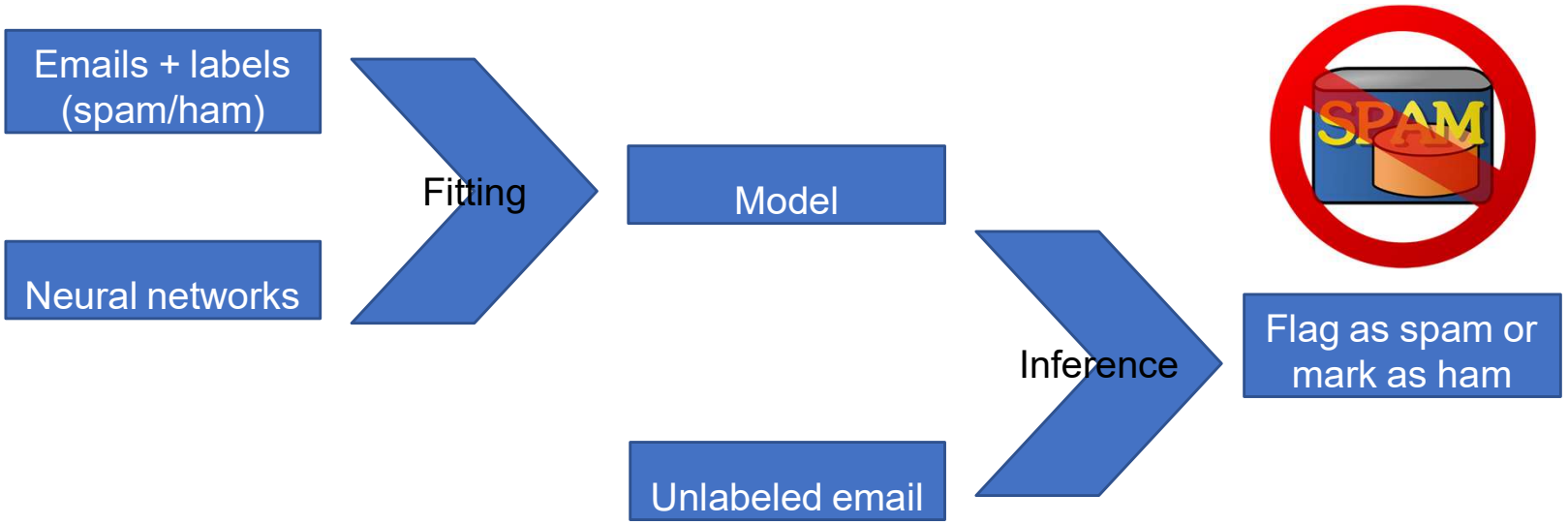
Goals of the Course

- Understand how to design secure and privacy-preserving machine learning
- Know the fundations and the frointiers of trustworthy AI research
- Hands-on experience in analysis and design of security/privacy-centric machine learning systems
- Cutting-edge research/project experience

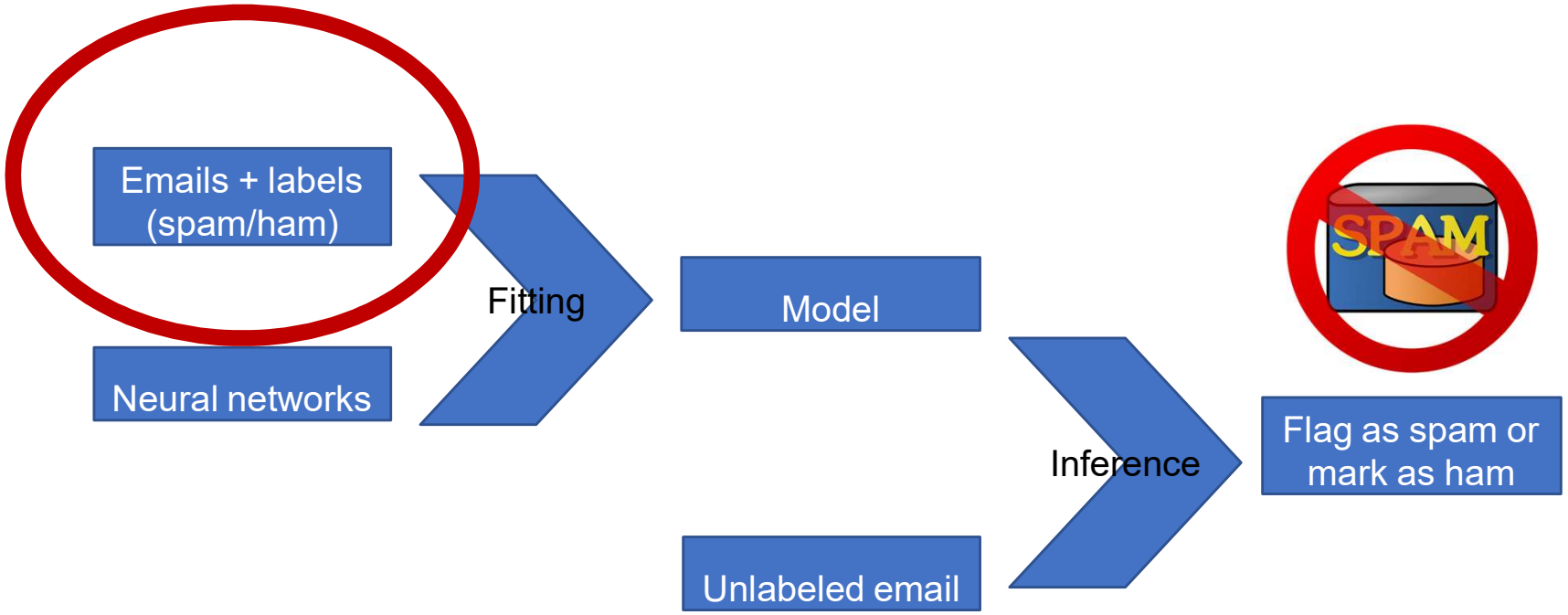
Machine learning paradigm



ML for spam detection

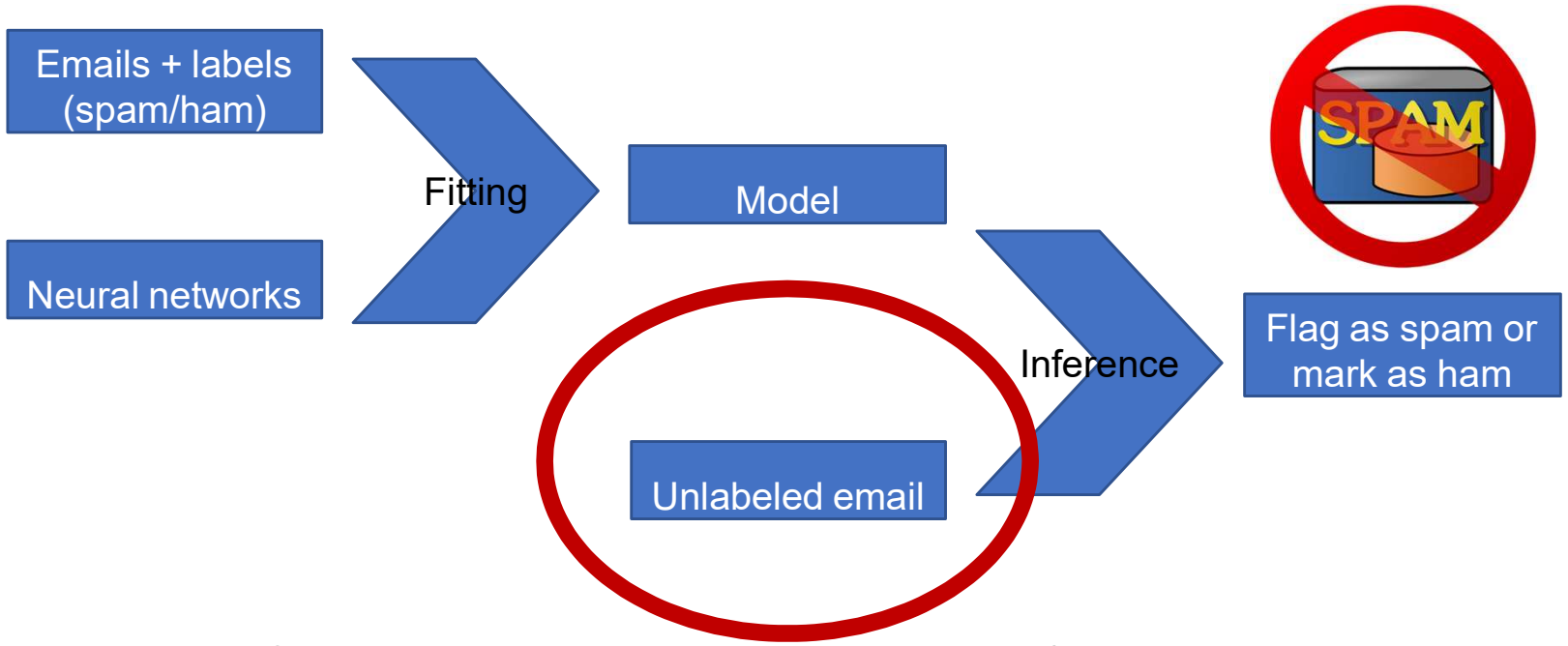


ML for spam detection



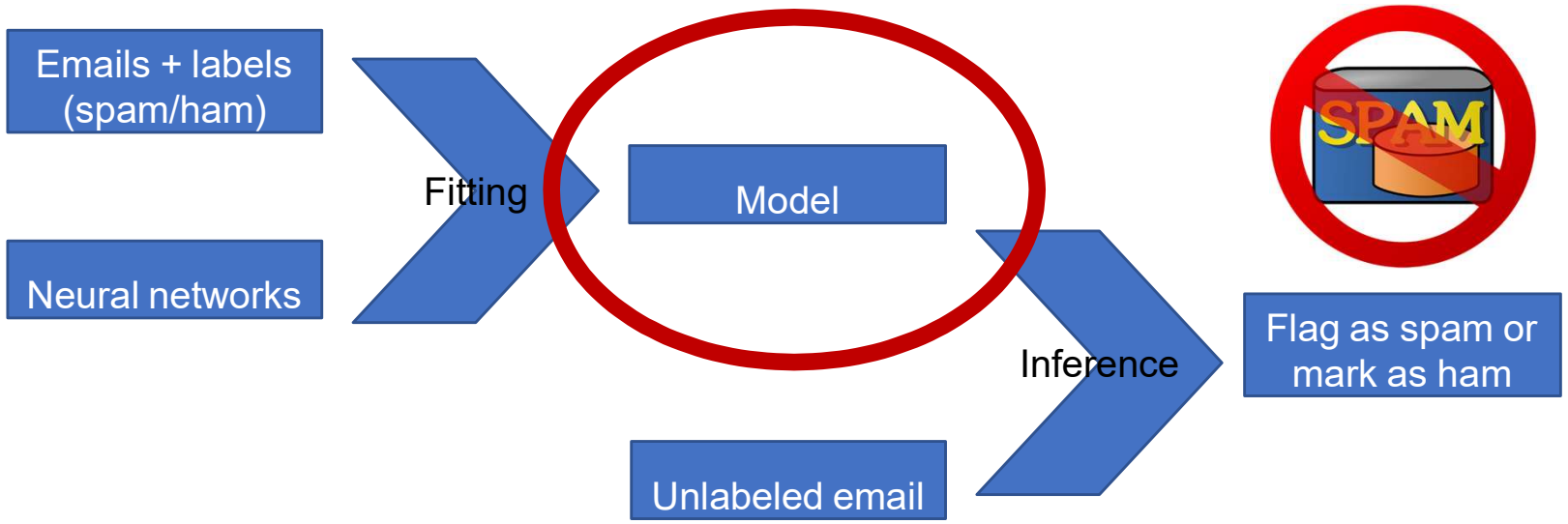
Poisoning: adversary inserts emails that contain spam but removes them from the spam folder back to inbox

ML paradigm in adversarial settings



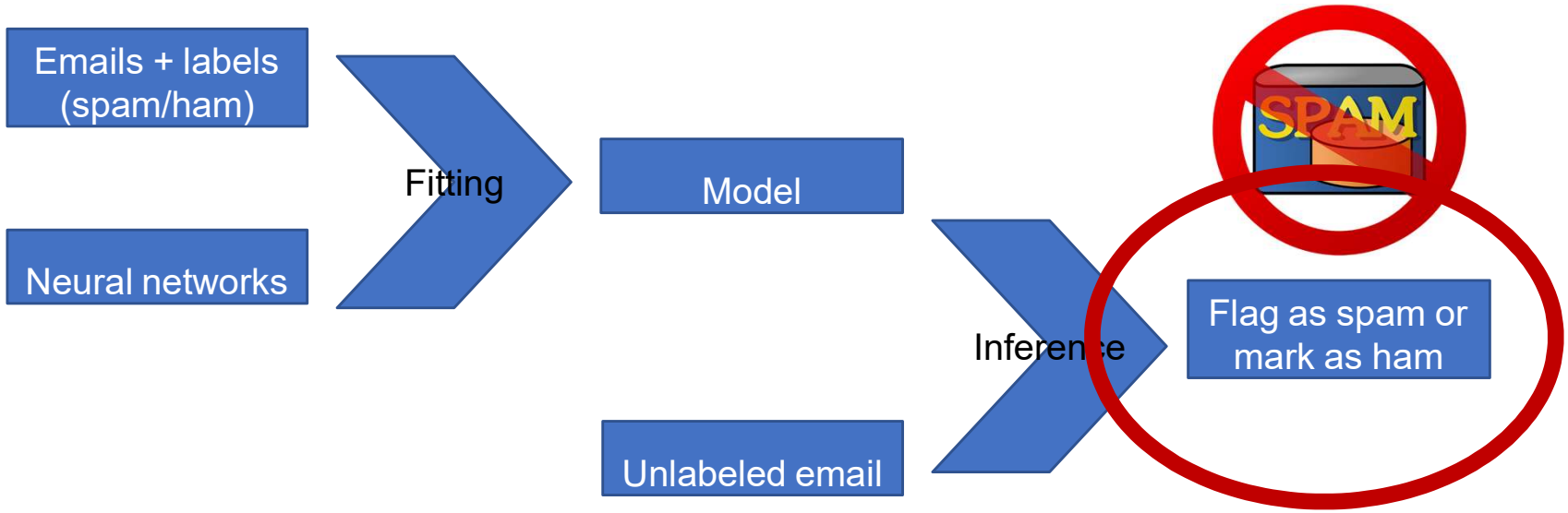
Evasion: adversary crafts adversarial example that evades detection (spam email instantly marked as ham)

ML paradigm in adversarial settings



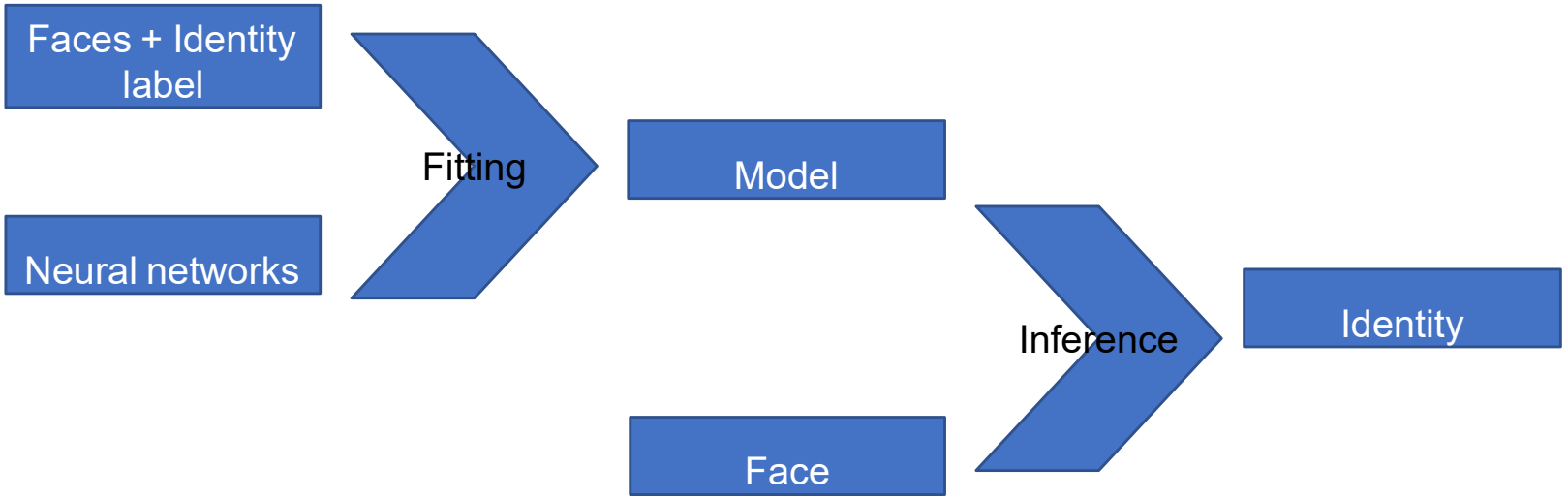
Membership inference: adversary inspects model to test whether an email was used to train it (privacy violation)

ML paradigm in adversarial settings



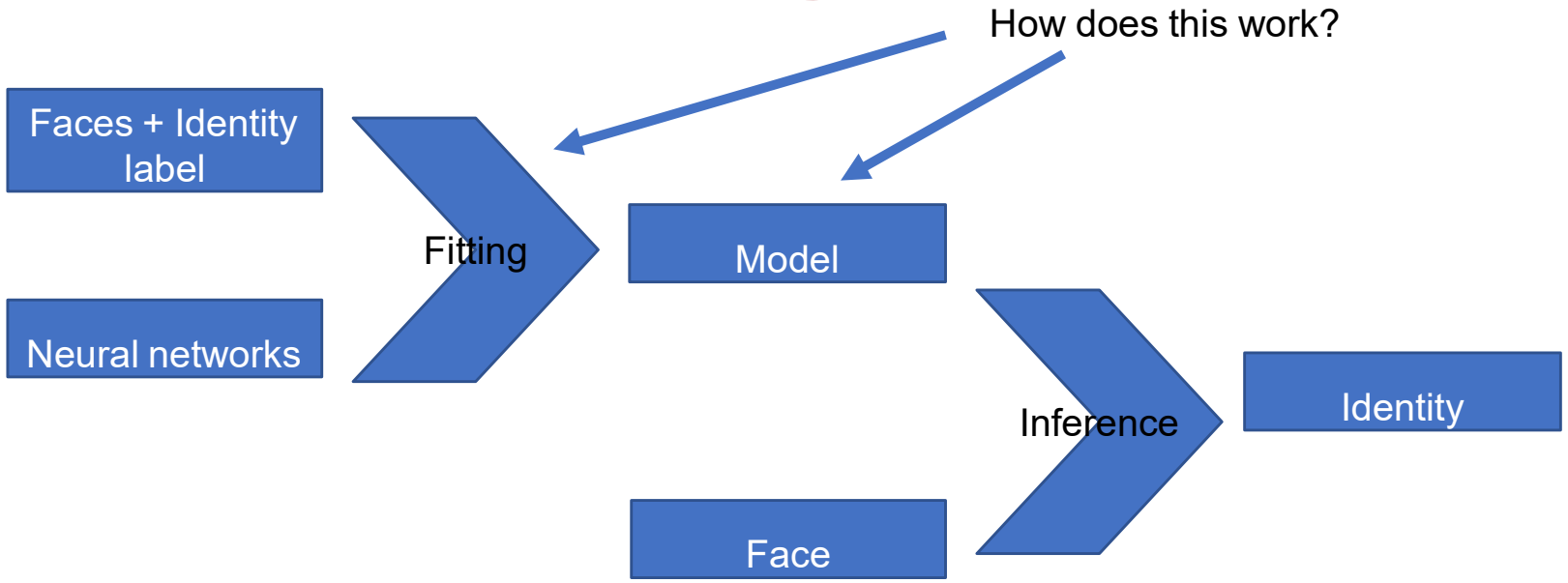
Model extraction: adversary observes predictions and reconstructs model locally

Societal aspects of the ML paradigm



Fairness: if training data does not contain enough faces from a minority or wrong training objective is used, accuracy at inference suffers (model does not build relevant features)

Societal aspects of the ML paradigm



Interpretability: how do we explain a ML algorithm to a human?

What this course aims to do

- Introduce you to some of the latest and most important research in A.I. as related to security, privacy, fairness, safety and reliability

- Convey core and general concepts, with a focus on applying the concepts in a system building project

- Introduce open research problems in the area and enable you to contribute and formulate new tasks

What this course is not

- It does not cover how to design neural nets to solve vision or robotics tasks (though we look at such networks).
- This is not a course on gradient-based optimization algorithms.
- It is not an introductory course to Deep Learning or Python.

Questions about Content?

Any questions about content, focus, etc.? Email me at yuant@ucla.com

Course Mechanics

Prerequisites & Assumptions

- This course has assumptions:
  - You have taken undergraduate level computer science classes
- In addition, we assume:
  - You are **proficient in Python** programming and have **experience with machine learning** or have time to **learn it on your own**
  - **IMPORTANT:** this course does not teach machine learning dev
- Please finish the preclass survey by April 2 3 pm so that I can understand more about your background and interests

Websites

Syllabus:

https://docs.google.com/document/d/1srabS3Ka-Nef-X4afw8Pn2d-fyVKv2GTbOaG3Mm-JeY/

Contains all the useful information about the course (schedule, resources, links, and policies)

Bruin learn:

https://bruinlearn.ucla.edu/courses/183840

for assignments submissions, and slides

Piazza site:

https://piazza.com/class/lugewd7ie2b13a

for online Q&As and discussions

Anonymous feedback form:
https://forms.gle/gS21TLtb4y5MFXb47

Workload

- 2 homework assignments
- 1 group literature review (paper presentation)
- 1 group course project
  - Two reports
  - Three presentations

- This is my first time teaching this class, please let me know if you have any feedback:
- https://forms.gle/gS21TLtb4y5MFXb47

Grading

- **Grading based on:**
  - Homework assignments (30%)
  - Literature review (15%)
  - Participation in feedback to others' work (5%)
  - Course project (50%)
    - Introduction Presentation (5%)
    - Midterm presentation(5%)
    - Final presentation(15% from Yuan's and peer's evaluation)
    - Midterm report(5%)
    - Final report(10%)
    - Evaluation from teammates (10%)
  - Teaching evaluation (1% bonus)

Grading

- I guarantee at least the following:
  - A+ 100% to 97%
  - A 96.99% to 93%
  - A- 92.99% to 90%
  - B+ 89.99% to 83.33%
  - B 83.32% to 76.67%
  - B- 76.66% to 70%
  - C+ 69.99% to 67%
  - C 66.99% to 63%
  - C- 62.99% to 60%
  - D+ 59.99% to 57%
  - D 56.99% to 50%
  - F 49.99% to 0%

Individual Assignments

- Grading for assignments
  - Programming assignments for attacks and defenses (adversarial machine learning, poisoning attacks, privacy attacks and defenses) + short answer questions
  - 15 points for each of the 2 assignment – will be scaled to 30% in total
- Assignment deadlines are on the syllabus
- Assignment details will be on Burin Learn
- Individual → each student is responsible for doing their own work
  - Discussion is encouraged, but **work is individual**

Literature Review

- Each group will present one paper
- Here is the paper signup form (Please sign up by 04/08)

https://docs.google.com/spreadsheets/d/1IWN_taP0FCrk4qtkF_PfhJaA8zBRv35XXEaT4uCTcp8/edit#gid=0

- The recommended size of team is 3 students, but 2-4 students are all fine
- Please submit your slides at least 2 hours before the in-class presentation

Literature Review

Presentation Preparation

- Time distribution: 20 minutes of presentation + 5 minutes of Q&A
- Things to cover in the presentation:
  - ⅓ time goes to background (what have been done before this paper, presenters must explain the topic in an easy-to-understand manner, assuming the audience has little background
  - ⅓ time goes to what this paper does
  - ⅓ time goes to what YOU think are the unique advantages and limitations of the system. What will be the future work needed? How has the field evolved after this paper (hint: look into papers that cited the paper you present)?

Course Project

- How will the course project work?
  - Several research projects on the cutting-edge trustworthyAI research
    - Open-ended
    - Students can also propose their own ideas!

Research project: what topic should I choose/propose?

- Projects must:
  - Relate to topics covered in class and focus on some aspect of trustworthy AI
  - Strive for new research/development contributions – aim for something never done before
  - Not be a project you're working on for or another course
  - If you are working on a related project for your research, that should be fine

How should I form a project team?

- Forming teams and choosing topics:

  These two things are not independent
- Try to choose team members with common interests, different backgrounds, etc., not just your friends
- Multiple teams cannot work on the same project
- We will introduce some research ideas on April 08, you can also propose your own ideas (please email me by April 06 about your idea)
- We can also help you to find teammates

More project details

- Each project will have a mentor (might be me, my phd students, or industry researchers)
- Project output will include a paper/report
- Some additional hardware may be available, if needed

Course Project Presentations

Proposal Presentation (week 4):

- 6 minutes of presentation + 4 minutes of Q&A
- Prepare up to 5 slides (with no/minimum animation)
- Put team ID on all slides
- Get full credits for attendance (no other evaluation criteria)

Midterm Presentation (week 7):

- 6 minutes of presentation + 4 minutes of Q&A
- Prepare up to 5 slides (with no/minimum animation)
- Put team ID on all slides
- Get full credits for attendance (no other evaluation criteria)

Course Project Presentations

Final Presentation (week 10):

- Presentation Preparation
- 5 minutes of presentation + 1 minutes of Q&A
- Put team ID on all slides

Presentation Evaluation:

- Yuan's evaluation (weigh 50%) and peer evaluation (weigh 50%) using Google form

Final Presentation Eval

- Clear explanations of your design and implementations
- (2.5 Points)
- Show the research challenges and justify how you
- address these challenges in the design
- Explain the novelty of your design
- Include the experiment details
- Reasonable analysis
- Clear examples (case studies)
- Related work (0.5 Points)
- Conclusion and suggestions for future work(0.5 Points)

Project Reports

Project Mid-term Report

- Will be evaluated by the following criteria:
  - Statement of the high-level problem area
  - Description of the focused project topic and potential solution
  - Outlining the project goals and timeline
  - Preliminary results *bonus point*

Final Report

- Will be evaluated by the criteria of final project presentation.

Important Dates

All important dates are on the course syllabus

How to Contact Us

- Instructor:
  Yuan Tian
- **Email:** yuant@ucla.com

Best: email to request a meeting (in person, Skype, phone, etc), provide context

  - Office hours:

Wed, 2 - 3pm, Boelter 6730D

How to Contact Us

- Teaching Assistant:
  Jinghuai Zhang
- **Email: jinghuai1998@g.ucla.edu**

Best: email to request a meeting (in person, Skype, phone, etc), provide context

  - Office hours:

Thu 3-4 pm ELLIOTT Room Engineering IV, Room 63-129

or by request

Some Syllabus-type Details

- Class meetings:
  - Class Location: PUB AFF 1234
  - Class Time: Mon/Wed, 4:00 - 5:50pm
- Textbooks
  - **None**, but Yuan is happy to provide some references for students who are interested
- Assigned reading
  - Papers, blog posts, etc.

Assigned Reading

- Between class readings, homework assignments, and project, ***you'll be reading a lot of papers!***
  - Don't be surprised to see 100+ pages of reading/week
  - Reading research papers is not like reading textbooks, they're much more forgiving and can be read efficiently
  - **Hint**: read the pamphlet posted for reading material today
    - We'll also take some time in the next lecture to practice how to read efficiently together.

Course Policies

- **Academic Integrity:** all students are expected to adhere to academic integrity policies set forth by UCLA

- **My Collaboration Policy**: discussion is encouraged, but **assignments must be done individually → Copying or sharing is cheating, cheating → failing grade**
- **Plagiarism**: no copying, attribute ***all*** content sources using proper references
- **My Wiki Policy**: if you cite Wikipedia (or similar) pages directly, you will fail the assignment/deliverable
- **Re-grading**: on a case-by-case basis, contact us (we sometimes make mistakes…)
- **Deadlines:** Late submission of individual assignments will be accepted for up to two (2) days after the deadline, with a 10% per day penalty. Late submission of group assignments (e.g., literature reviews, project presentations and project reports) will not be accepted. Please email me for special cases.

Ethics of S&P Work

- Research, development, and experimentation with sensitive information, attack protocols, misbehavior, etc. should be performed with the utmost care

- You are expected to follow a strict ethical code, especially when dealing with potentially sensitive information

- If anything is unclear, ask before going forward

Questions about Logistics?

Any questions about course logistics?

Feel free to email later.