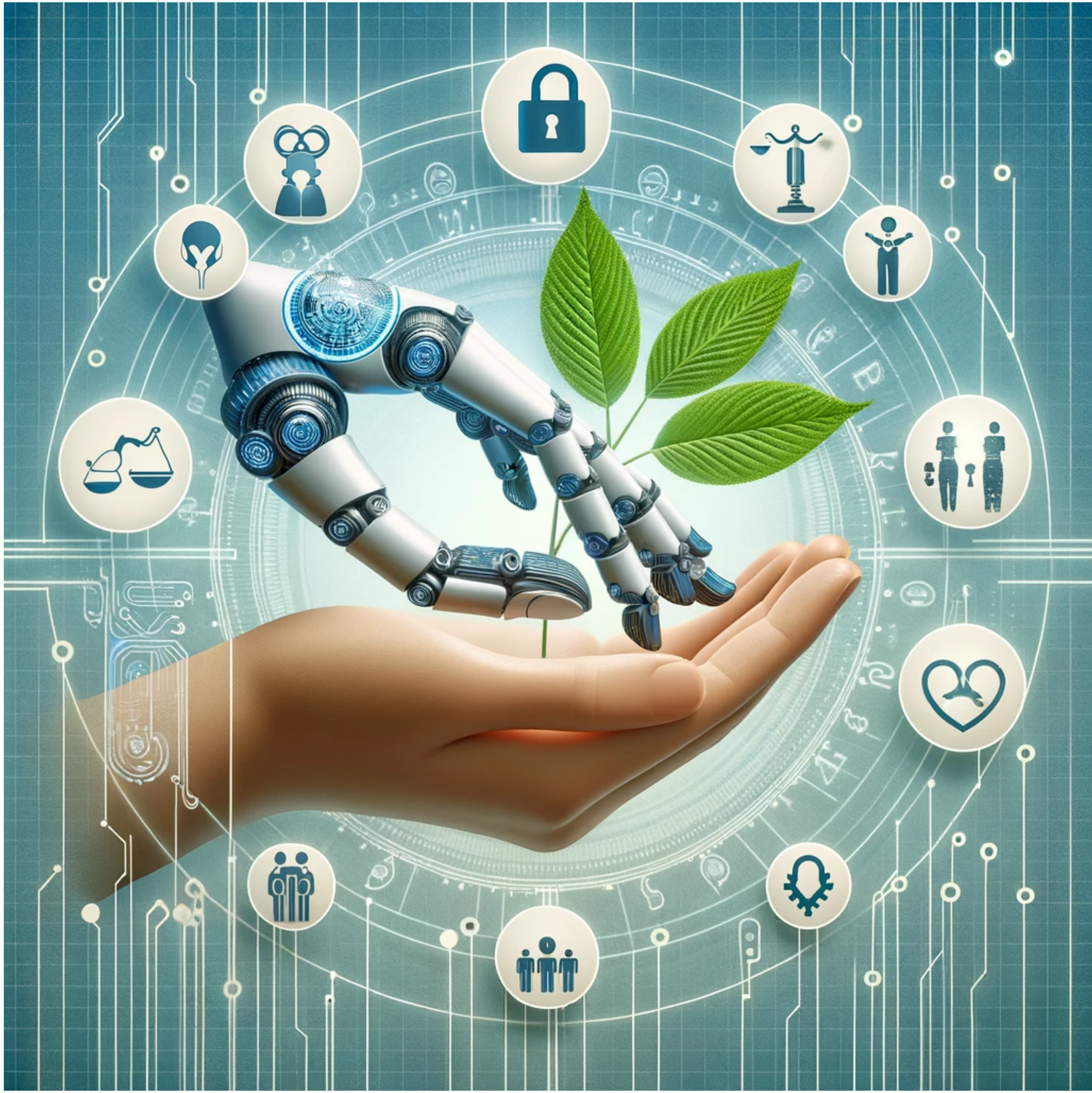
Responsible AI:

Seminar on Fairness, Safety, Privacy and more



https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto

Ferdinando Fioretto @UVA Spring 2025

Before we start

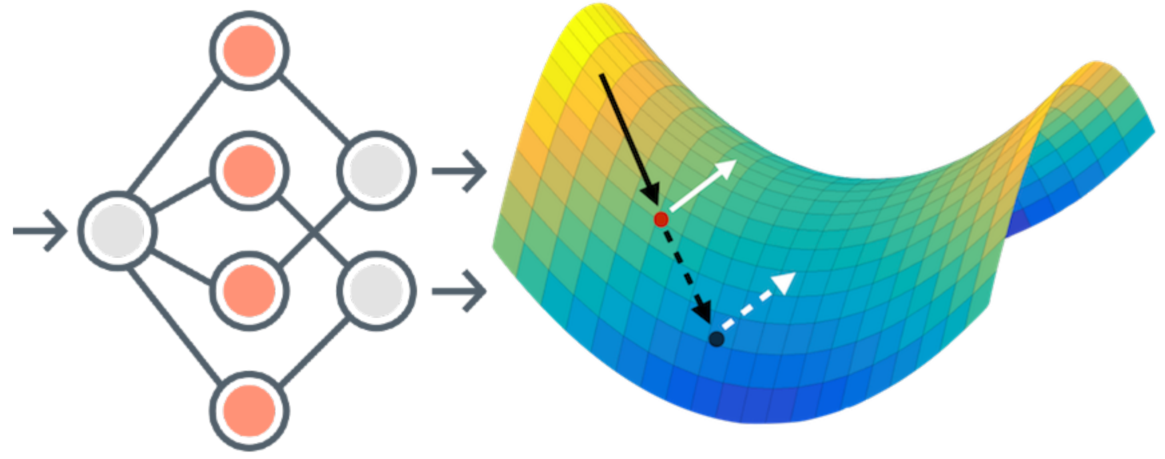
This is not an ML course!

- If you are here to learn about ML, you are in the wrong class, sorry!
- ML is a pre-req for this class.
- We'll study topics (more on this later) which are related to ML uses and misuses.

Introductions

AI for Science and Engineering

Differentiable optimization



ML Proxy optimizers

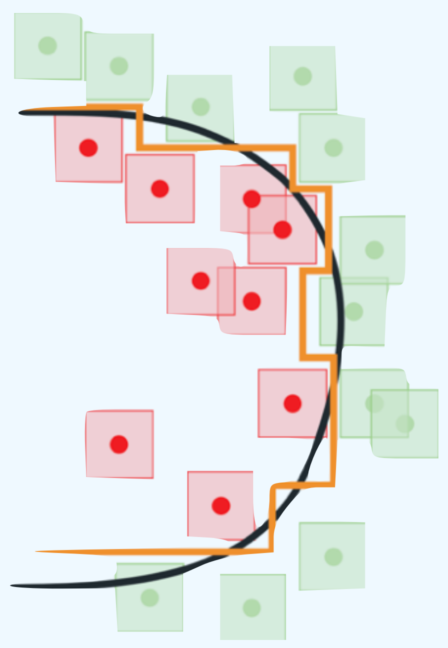
Optimization layers

Responsible AI

Differential privacy

Bias and Fairness

Robustness



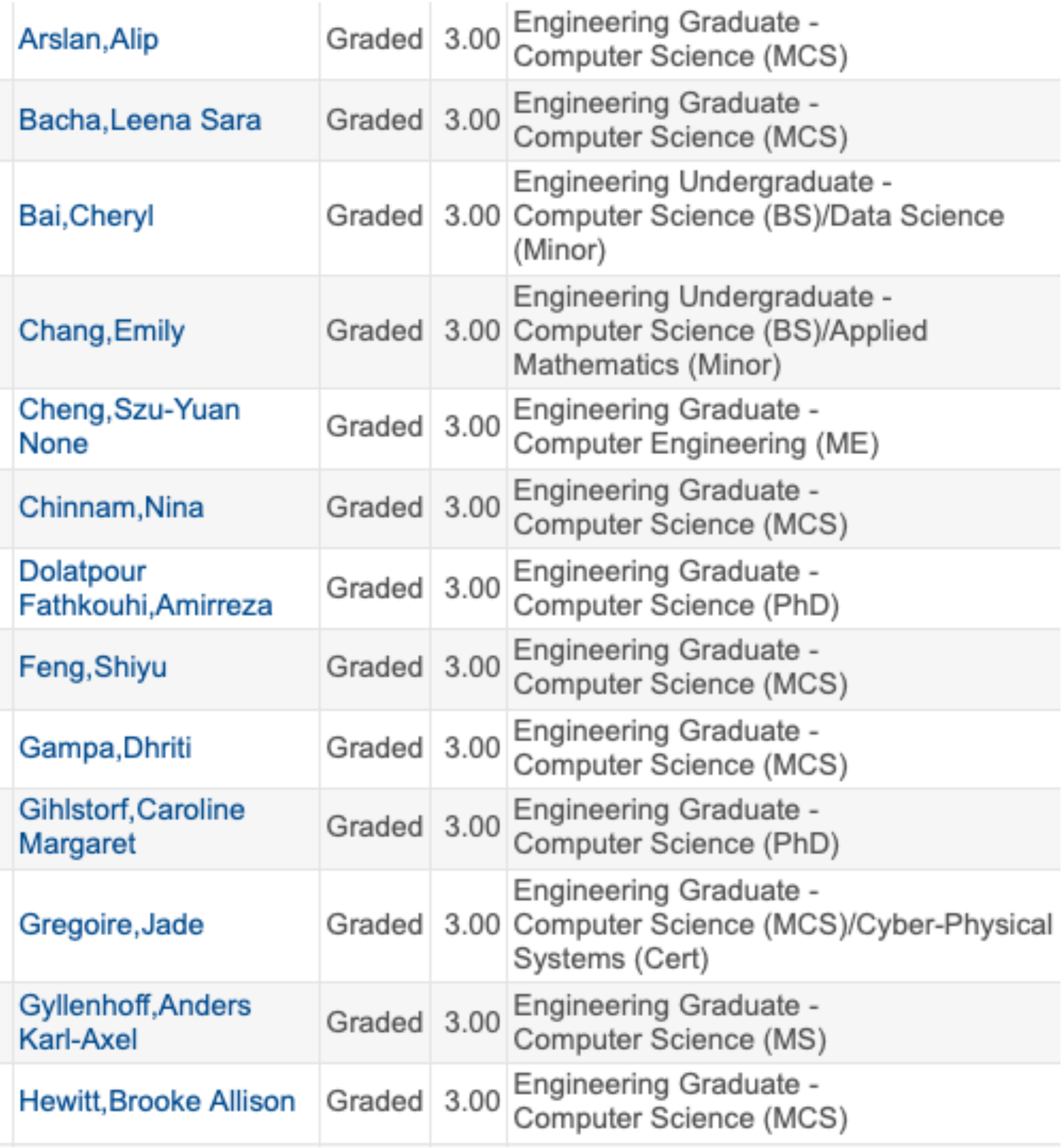
Model pruning

Saswat Das

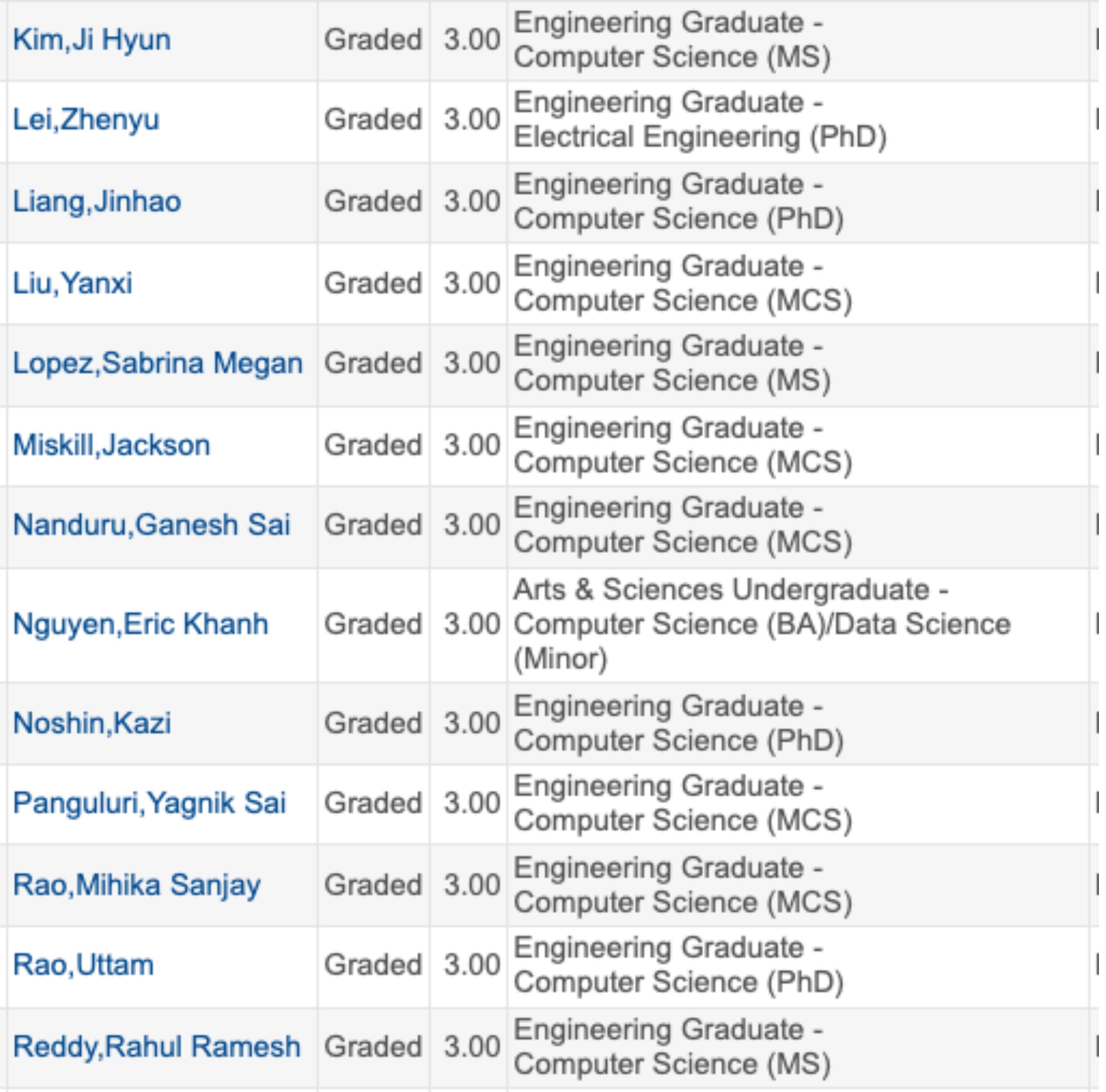
- Research in Responsible AI
  - Differential Privacy
  - Fairness in ML
  - More recently, agents systems

Now let me hear from you!

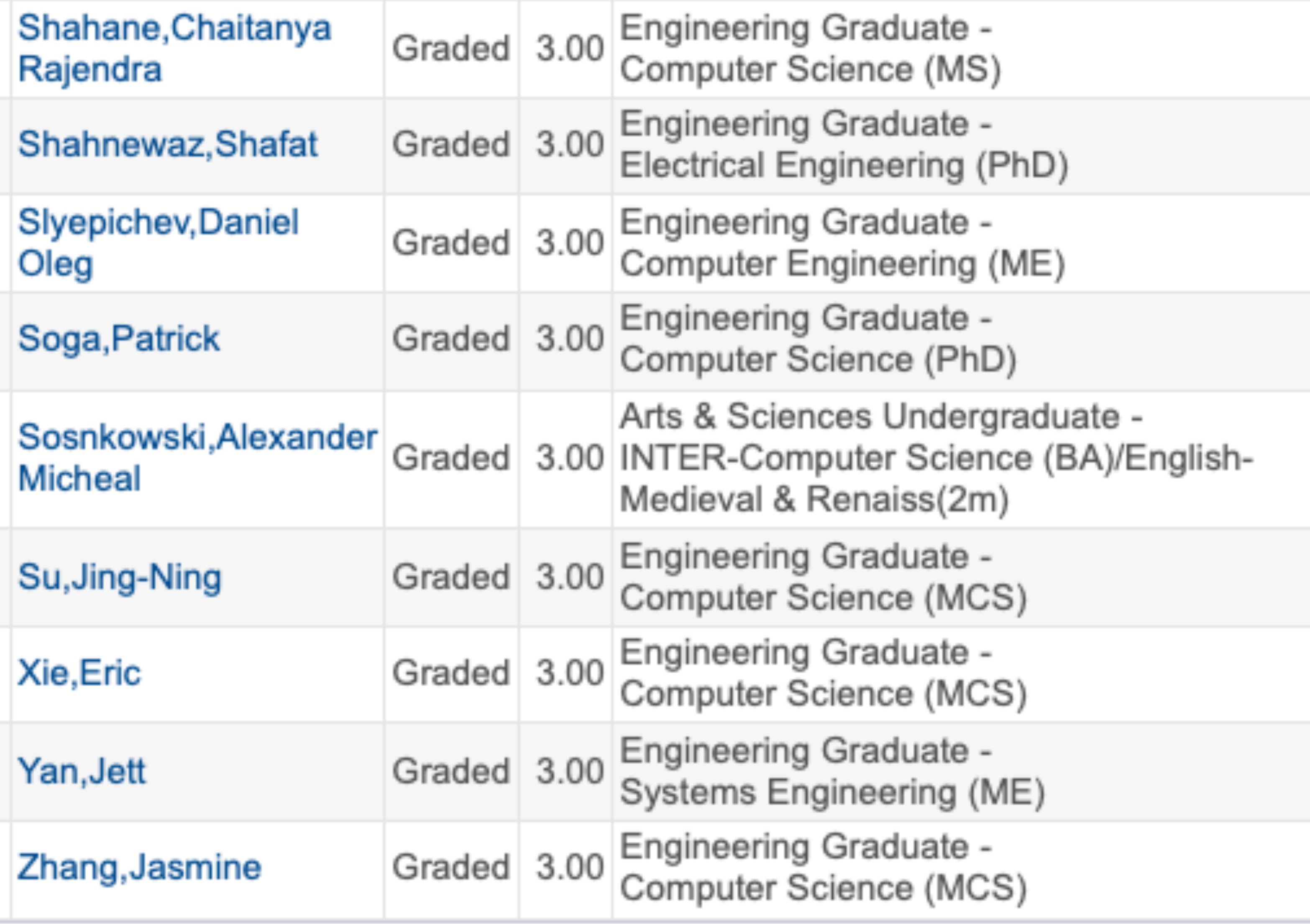
- Briefly introduce yourself:
  - Name, status (PhD, MS, BS), and research interests
  - Why did you enroll in this course?
  - What do you hope to learn?



Arslan,Alip	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Bacha,Leena Sara	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Bai,Cheryl	Graded	3.00	Engineering Undergraduate - Computer Science (BS)/Data Science (Minor)
Chang,Emily	Graded	3.00	Engineering Undergraduate - Computer Science (BS)/Applied Mathematics (Minor)
Cheng,Szu-Yuan None	Graded	3.00	Engineering Graduate - Computer Engineering (ME)
Chinnam,Nina	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Dolatpour Fathkouhi,Amirreza	Graded	3.00	Engineering Graduate - Computer Science (PhD)
Feng,Shiyu	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Gampa,Dhriti	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Gihlstorf,Caroline Margaret	Graded	3.00	Engineering Graduate - Computer Science (PhD)
Gregoire,Jade	Graded	3.00	Engineering Graduate - Computer Science (MCS)/Cyber-Physical Systems (Cert)
Gyllenhoff,Anders Karl-Axel	Graded	3.00	Engineering Graduate - Computer Science (MS)
Hewitt,Brooke Allison	Graded	3.00	Engineering Graduate - Computer Science (MCS)



Kim,Ji Hyun	Graded	3.00	Engineering Graduate - Computer Science (MS)
Lei,Zhenyu	Graded	3.00	Engineering Graduate - Electrical Engineering (PhD)
Liang,Jinhao	Graded	3.00	Engineering Graduate - Computer Science (PhD)
Liu,Yanxi	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Lopez,Sabrina Megan	Graded	3.00	Engineering Graduate - Computer Science (MS)
Miskill,Jackson	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Nanduru,Ganesh Sai	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Nguyen,Eric Khanh	Graded	3.00	Arts & Sciences Undergraduate - Computer Science (BA)/Data Science (Minor)
Noshin,Kazi	Graded	3.00	Engineering Graduate - Computer Science (PhD)
Panguluri,Yagnik Sai	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Rao,Mihika Sanjay	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Rao,Uttam	Graded	3.00	Engineering Graduate - Computer Science (PhD)
Reddy,Rahul Ramesh	Graded	3.00	Engineering Graduate - Computer Science (MS)



Shahane,Chaitanya Rajendra	Graded	3.00	Engineering Graduate - Computer Science (MS)
Shahnewaz,Shafat	Graded	3.00	Engineering Graduate - Electrical Engineering (PhD)
Slyepichev,Daniel Oleg	Graded	3.00	Engineering Graduate - Computer Engineering (ME)
Soga,Patrick	Graded	3.00	Engineering Graduate - Computer Science (PhD)
Sosnkowski,Alexander Micheal	Graded	3.00	Arts & Sciences Undergraduate - INTER-Computer Science (BA)/English-Medieval & Renaiss(2m)
Su,Jing-Ning	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Xie,Eric	Graded	3.00	Engineering Graduate - Computer Science (MCS)
Yan,Jett	Graded	3.00	Engineering Graduate - Systems Engineering (ME)
Zhang,Jasmine	Graded	3.00	Engineering Graduate - Computer Science (MCS)

ARTIFICIAL INTELLIGENCE

Computer systems that perform tasks that would usually require human intelligence

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Stats techniques that learn from data

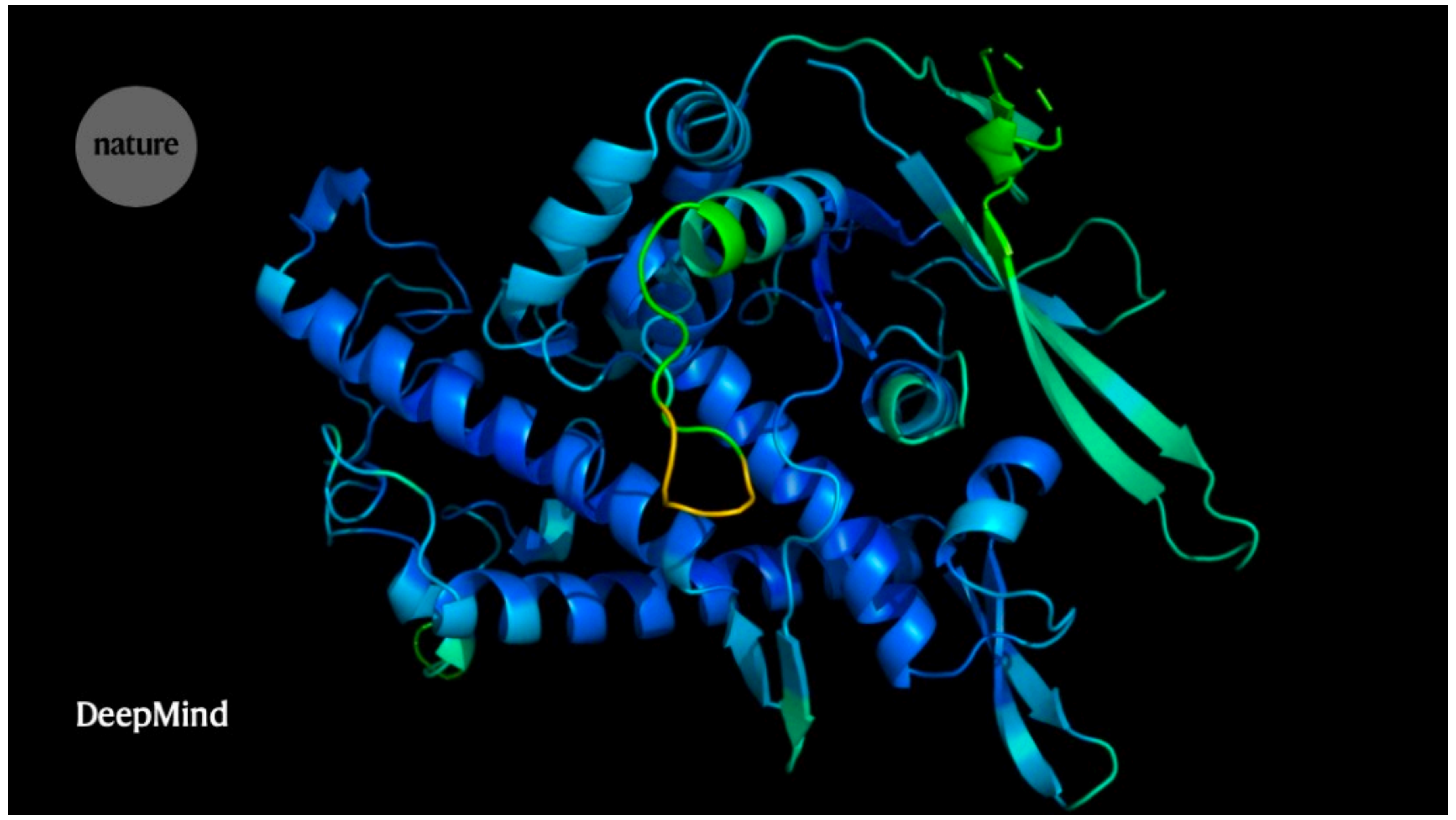
Machine learning begins to flourish.

DEEP LEARNING

Algorithms that enable self-learning to mimic human intelligence

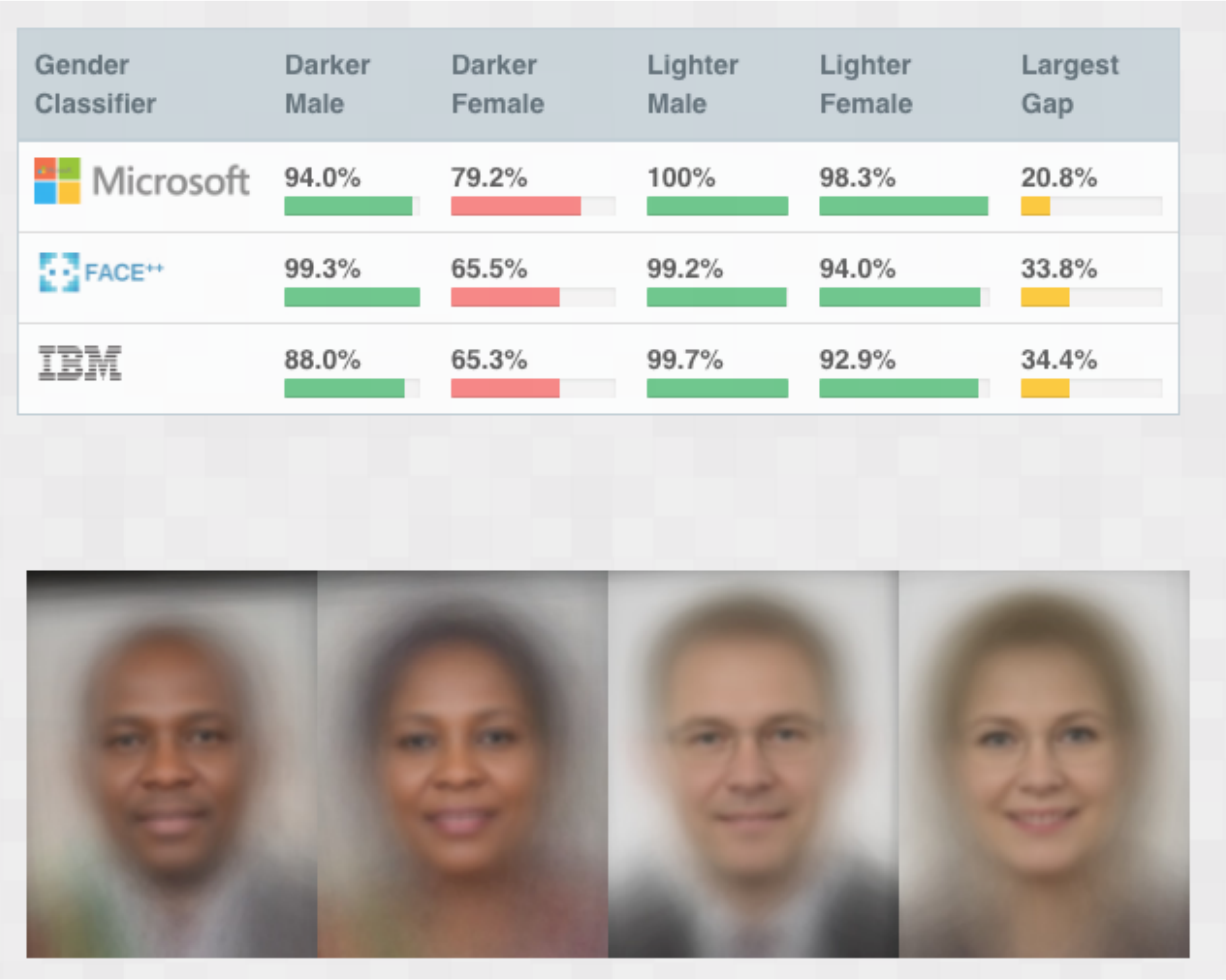
Deep learning breakthroughs drive AI boom.

Artificial Intelligence

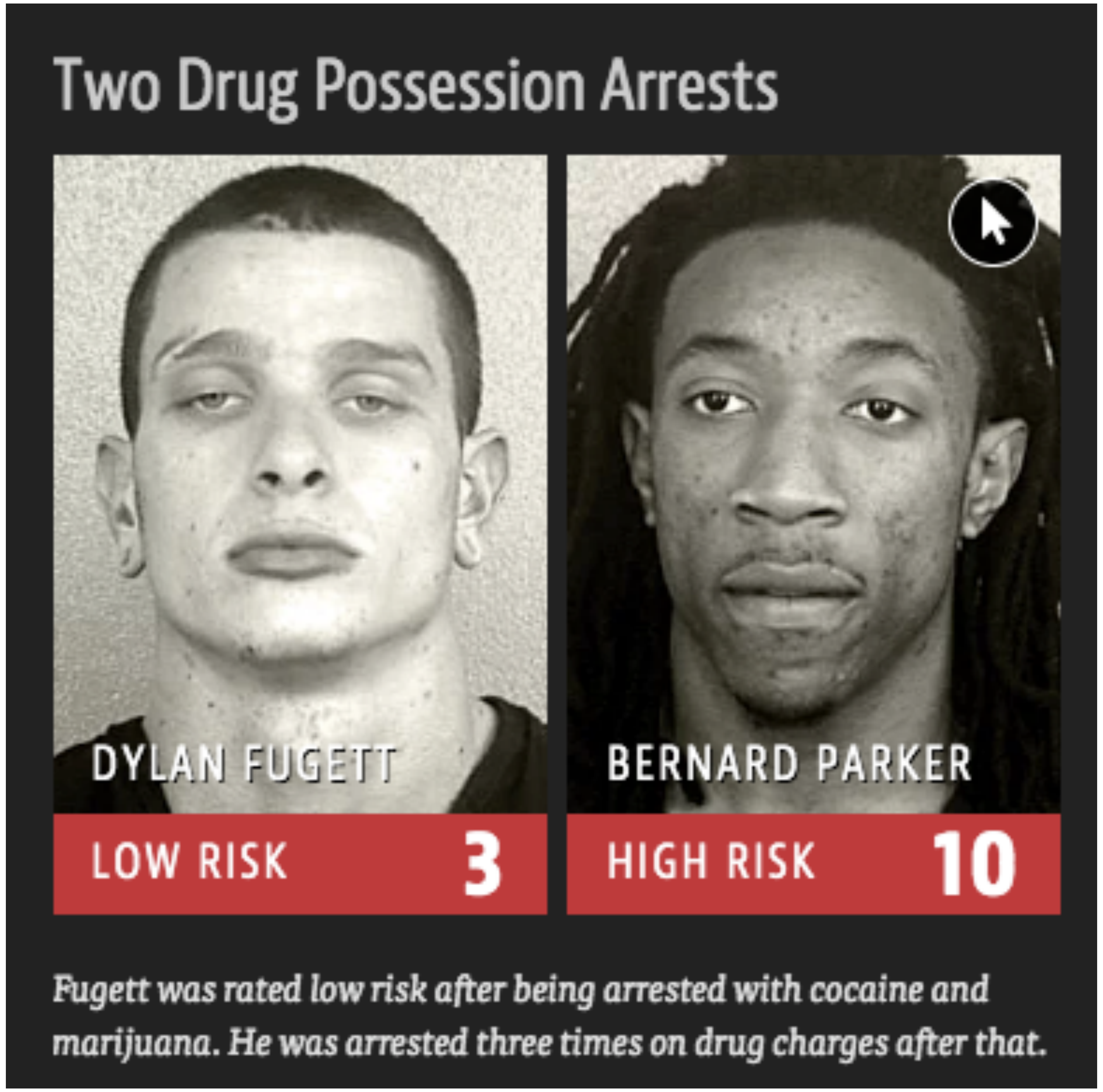


ML in practice: challenges

Are ML models fair?



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

http://gendershades.org/overview.html
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

What is fairness?

- Bias can occur even when everyone, from data collectors to engineers, have the best intentions.
- Just because an algorithm is unbiased now it does not mean it won’t be in the future.

So what is fairness in ML?

- Try your best guess!

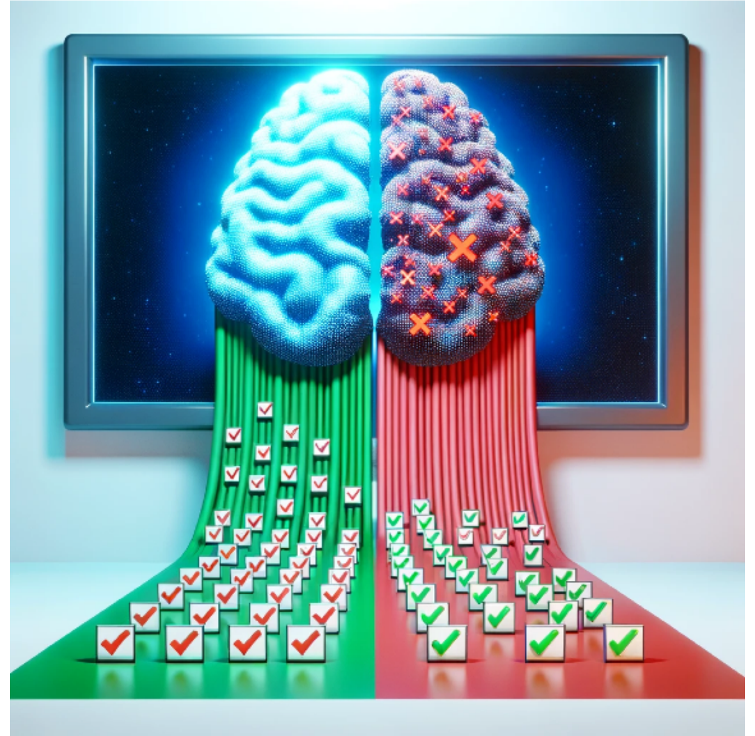
Why fairness is hard?

Machine Learning and social norms

- Sample norms: privacy, fairness, accountability
- Possible approaches:
  - Traditional: legal, regulatory, watchdog
  - Embed social norms in data, algorithms, and models
- Case study: PPML
  - “Single”, strong definition (differential privacy)
  - Almost every ML algorithm has a private version
- Fair ML
  - Not so much…
  - Impossibility results

Where does unfairness arise?

- **Data (input):**
  - More arrest where there are more patrolling
  - Label should be “committed a crime” but is “convicted of a crime”
  - Try to “correct” bias

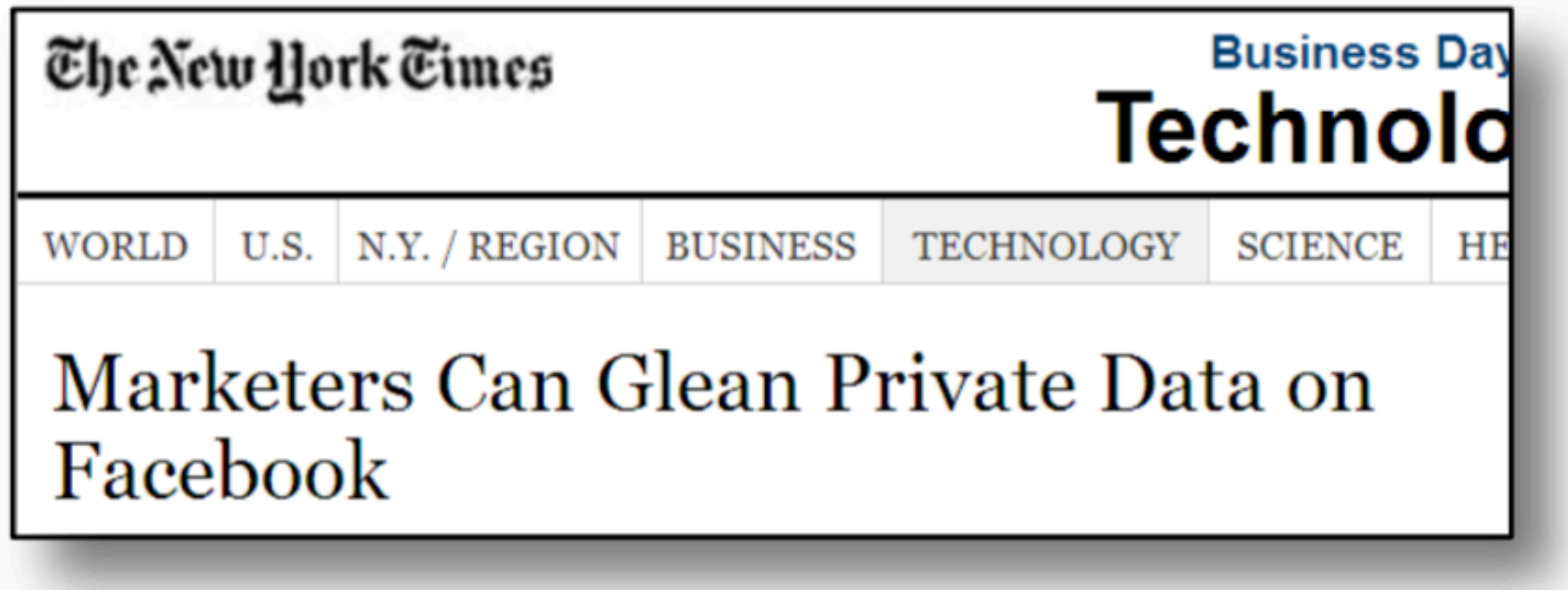


- Models (output)
  - e.g., Discriminatory treatment of sub-groups
  - Build or post-process models with subgroup guarantees
  - Quality of false positive/negative rates

- Algorithms (process)
  - Learning algorithm generating data through its decisions (e.g., don’t learn outmodes of defined mortgages)
  - Lack of clear train/test division and evaluation

ML in practice: challenges

Are ML models private?



The New York Times

Business Day

Technolo

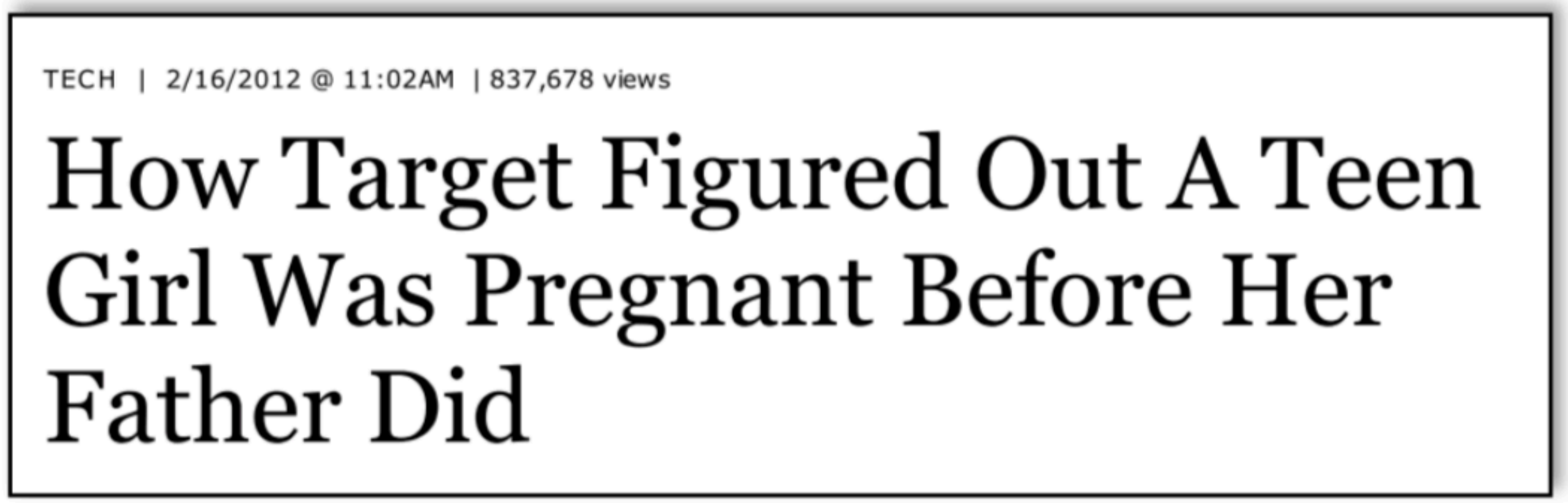
WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HE

Marketers Can Glean Private Data on Facebook

Facebook Ads

Reach the exact audience you want with relevant targeted ads.

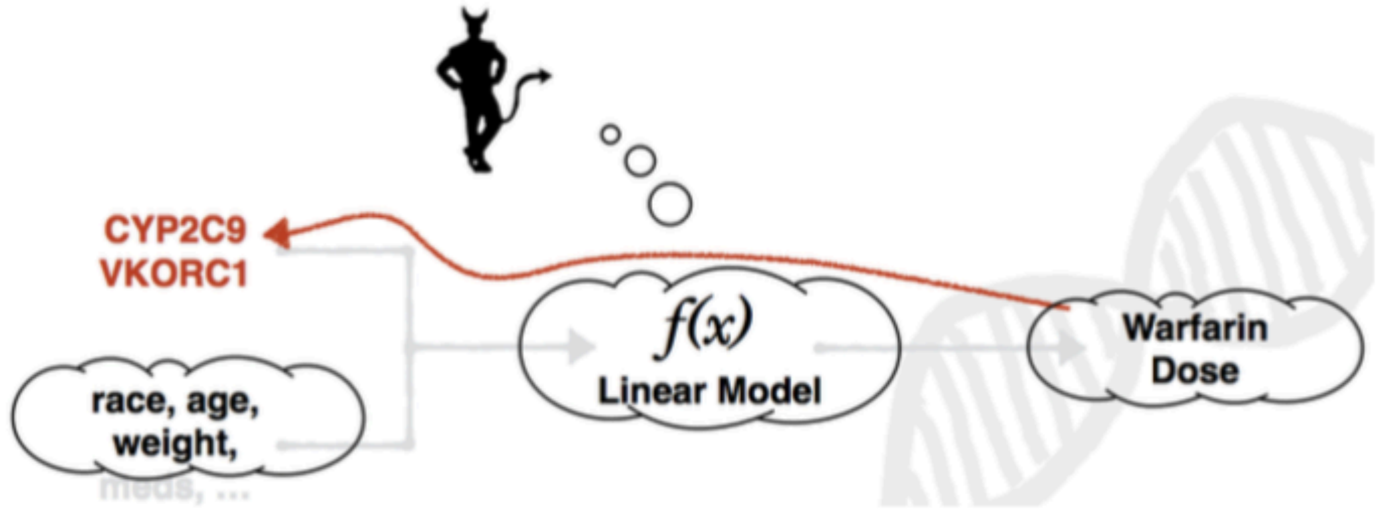
TARGET



TECH | 2/16/2012 @ 11:02AM | 837,678 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing

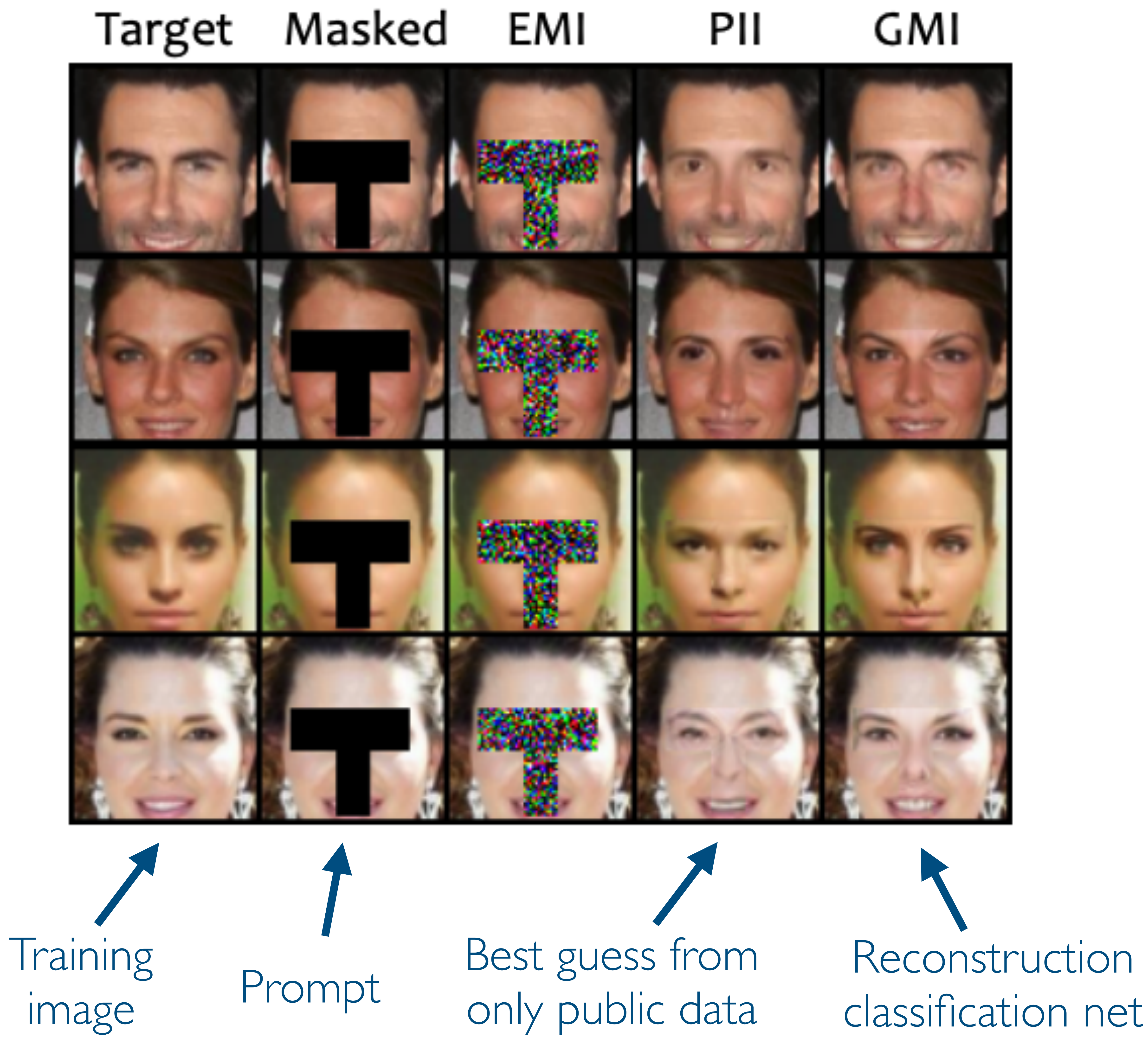


Why Anonymization is Hard?

Model inversion attacks

- Even if you don't release the raw data, the weights of a trained network might reveal sensitive information.
- Model inversion: recover information about the training data from the trained model.
- Example from a face recognition dataset, given a classifier trained on this dataset and a generative model trained on an unrelated dataset of publicly available images.

Source: Zhang et al., "The secret revealer: Generative model-inversion attacks against deep neural networks." https://arxiv.org/abs/1911.07135



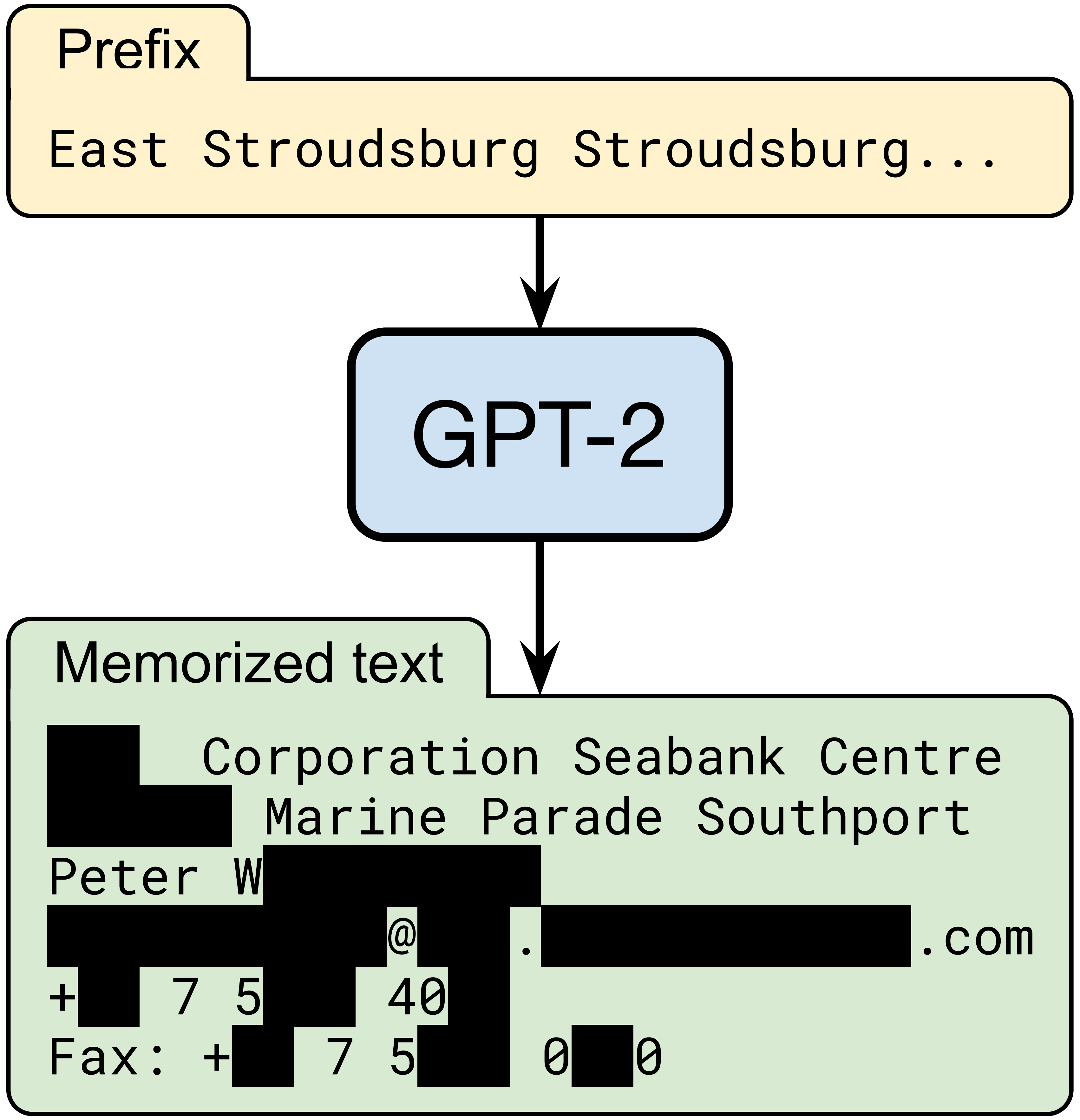
Target
Masked
EMI
PII
GMI
Training image
Prompt
Best guess from only public data
Reconstruction classification net

Zhang et al. 2020

Why Anonymization is Hard?

Extraction attacks

- Language models trained on scrapes of the public Internet.
- Extraction attack: extracts verbatim text sequences from the model's training data.
- Example from a GPT-2 model. Given query access, it extracts an individual person's name, email address, phone number, fax number, and physical address.



Prefix

```
East Stroudsburg Stroudsburg...
```

GPT-2

Memorized text

```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

Carlini et al. 2021

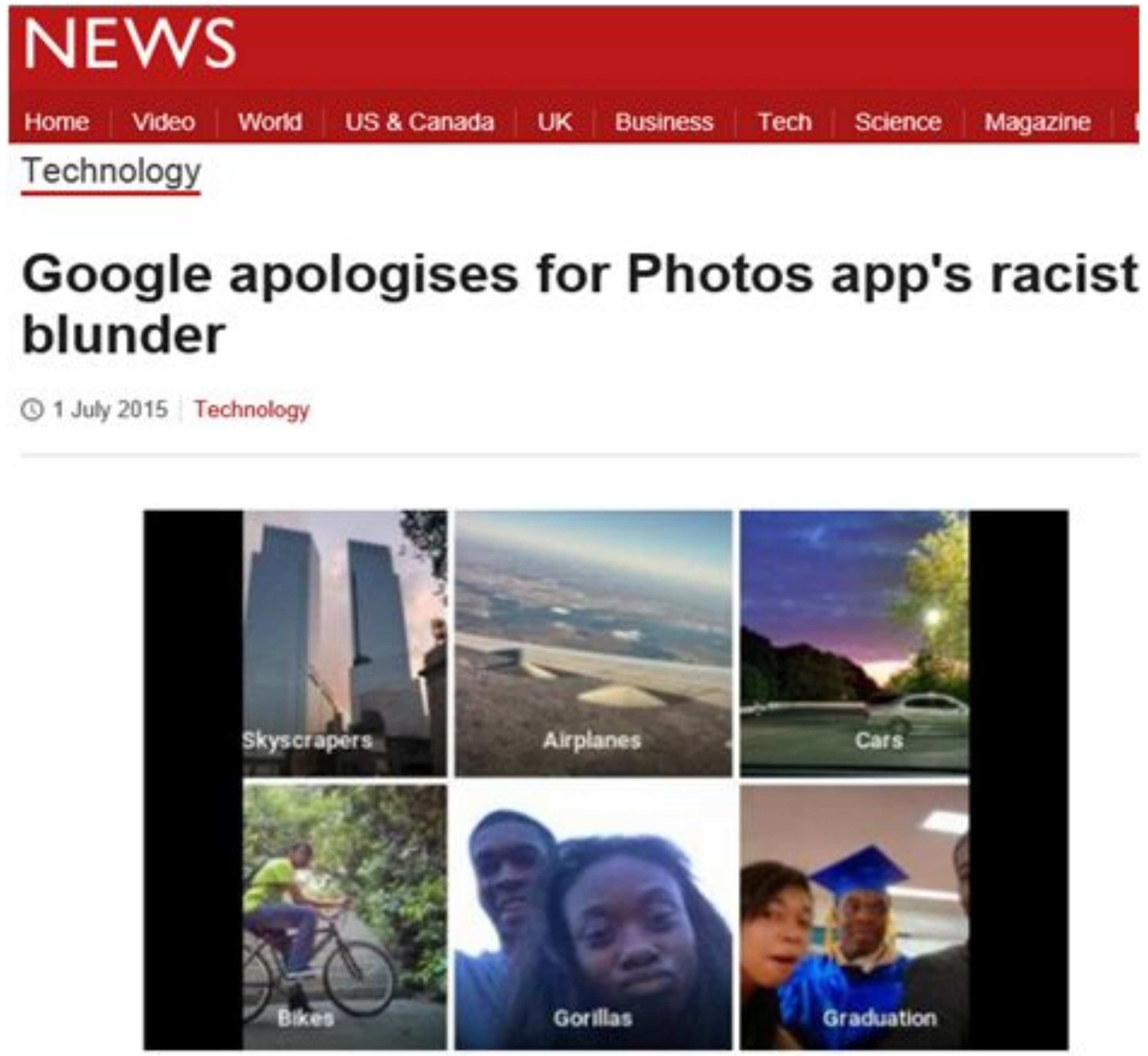
Why Anonymization is Hard?

Needs for guarantees

- It's hard to guess what capabilities attackers will have, especially decades into the future.
    - Analogy with crypto: Cryptosystems today are designed based on what quantum computers might be able to do in 30 years.
    - To defend against unknown capabilities, we need mathematical guarantees.
- Want to guarantee: no individual is directly harmed (e.g. through release of sensitive information) by being part of the database, even if the attacker has tons of data and computation.

ML in practice: challenges

Are ML models safe?



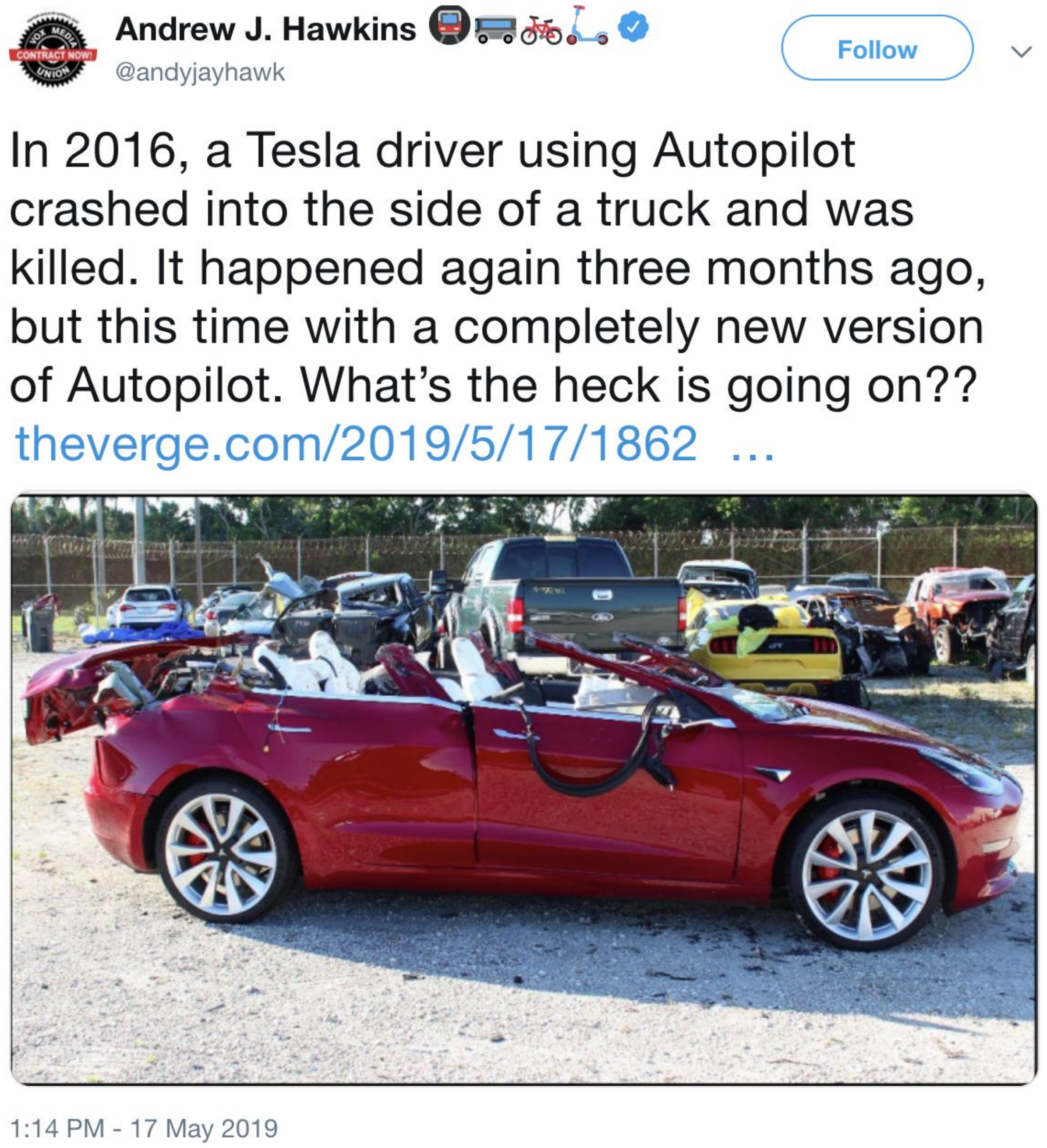
NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Magazine

Technology

Google apologises for Photos app's racist blunder

1 July 2015 | Technology

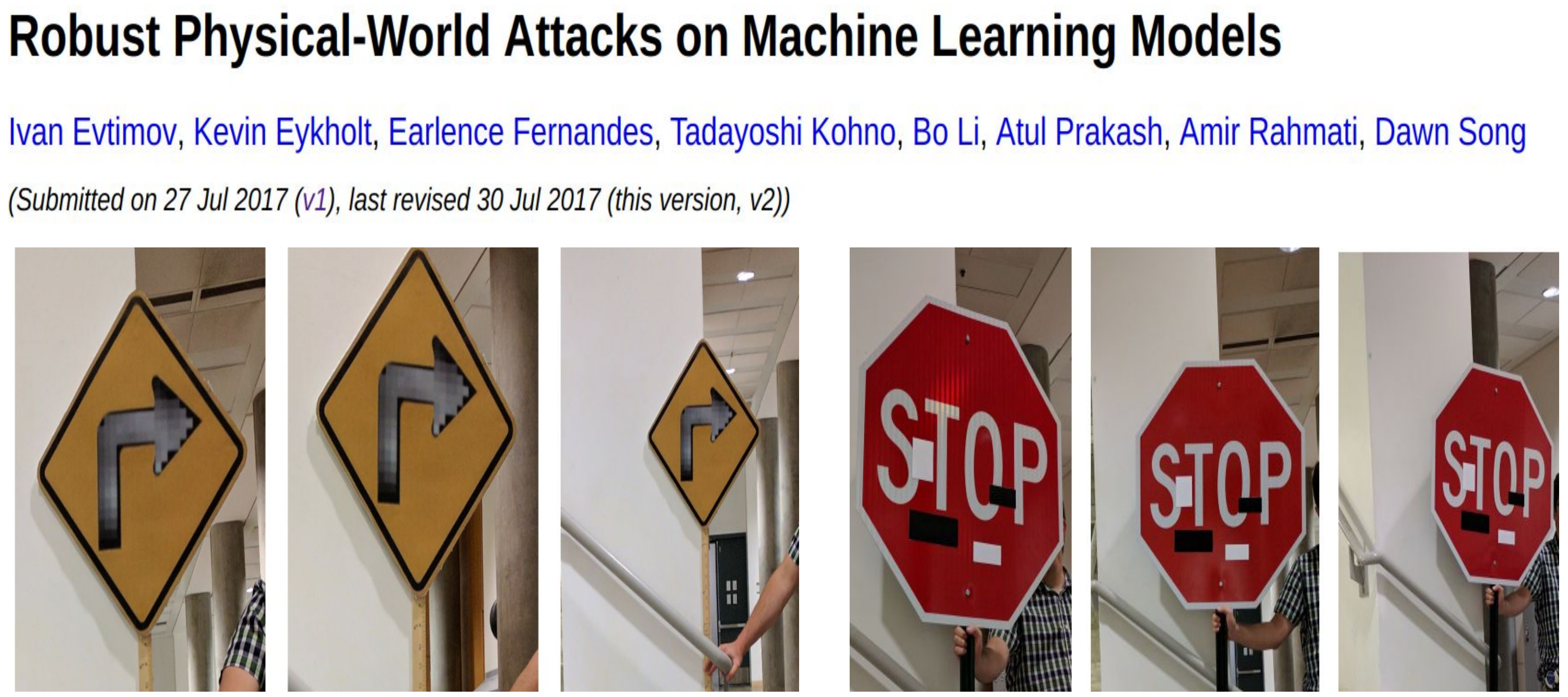


Andrew J. Hawkins
@andyjayhawk

Follow

In 2016, a Tesla driver using Autopilot crashed into the side of a truck and was killed. It happened again three months ago, but this time with a completely new version of Autopilot. What's the heck is going on?? theverge.com/2019/5/17/1862 …

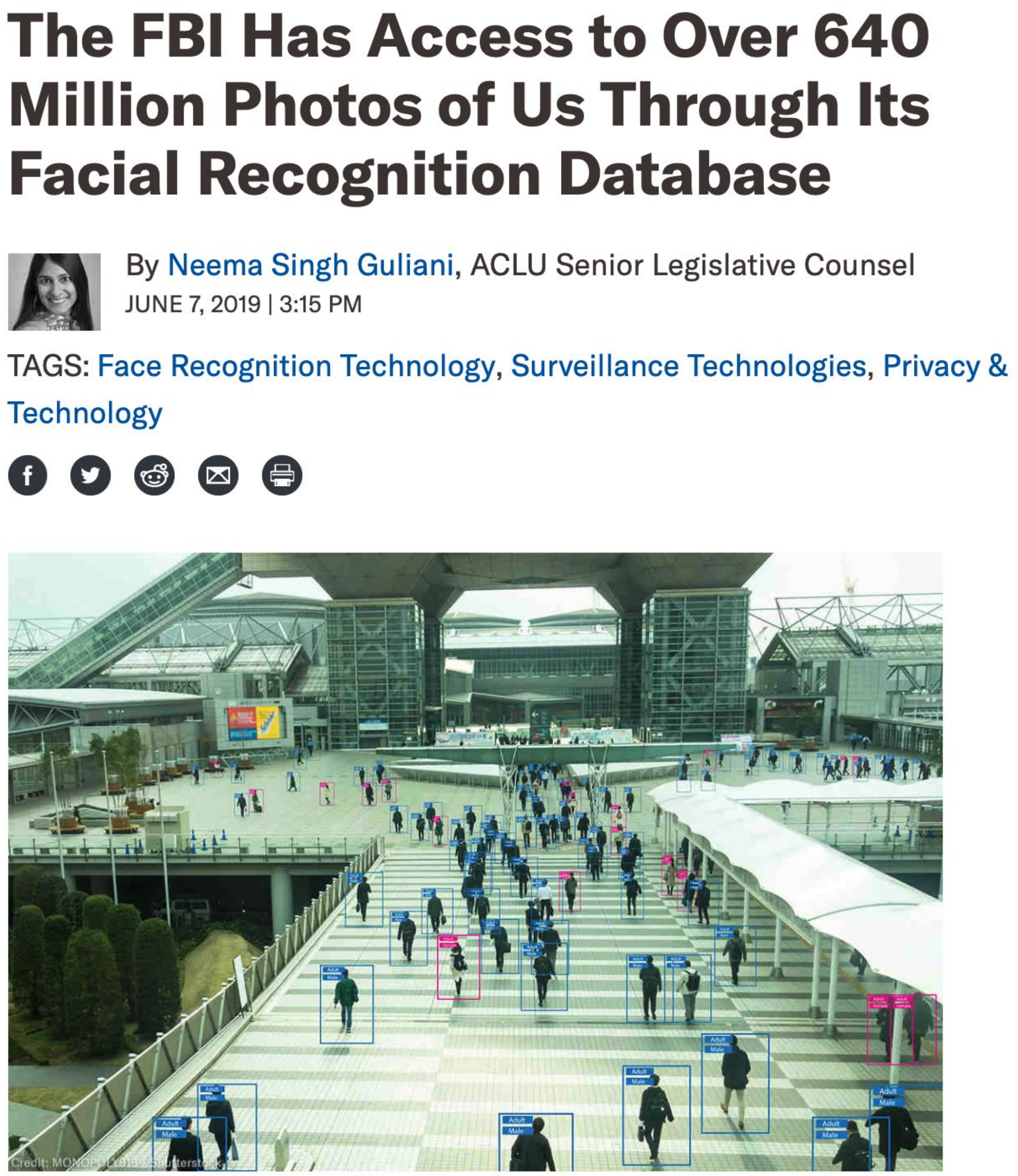
1:14 PM - 17 May 2019



Robust Physical-World Attacks on Machine Learning Models

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, Dawn Song

(Submitted on 27 Jul 2017 (v1), last revised 30 Jul 2017 (this version, v2))



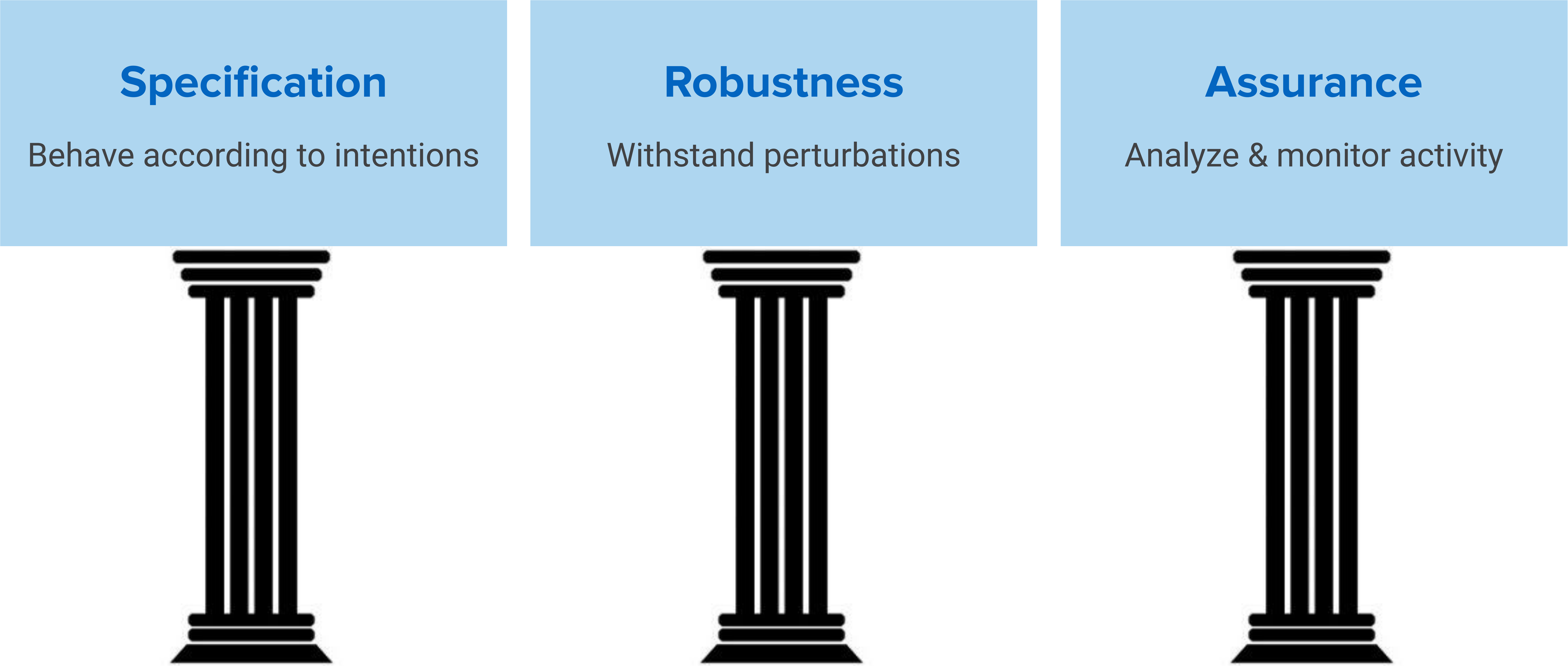
The FBI Has Access to Over 640 Million Photos of Us Through Its Facial Recognition Database

By Neema Singh Guliani, ACLU Senior Legislative Counsel
JUNE 7, 2019 | 3:15 PM

TAGS: Face Recognition Technology, Surveillance Technologies, Privacy & Technology

What is safety in ML?

Three pillars



Specification

Behave according to intentions

Robustness

Withstand perturbations

Assurance

Analyze & monitor activity

Ortega et al. 2018

Safety in a nutshell

How good is our approximation?
(Assurance)

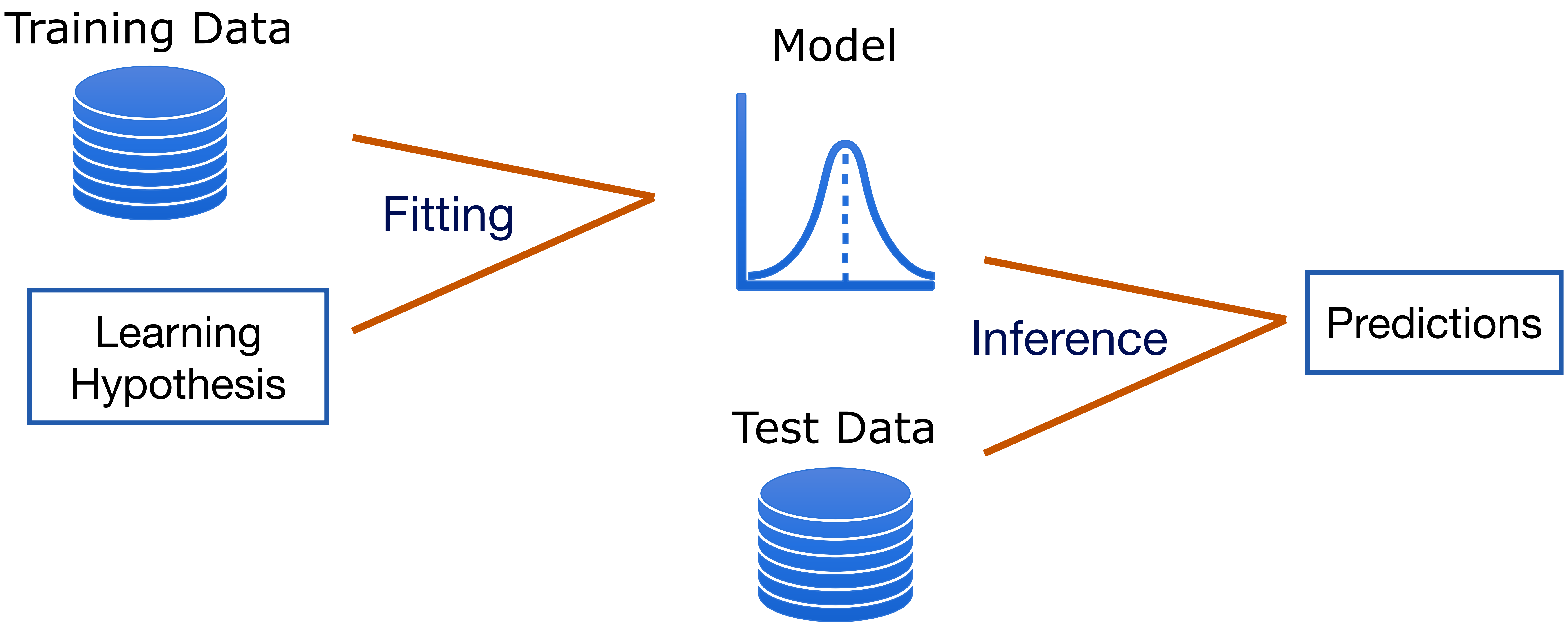
Where does this come from?
(Specification)

$$\arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{(s,a)\in\tau} r(s,a)\right]$$

What about rare cases/adversaries?
(Robustness)

Chiappa & Leikeal. 2019

The ML Paradigm



Training Data

Learning Hypothesis

Fitting

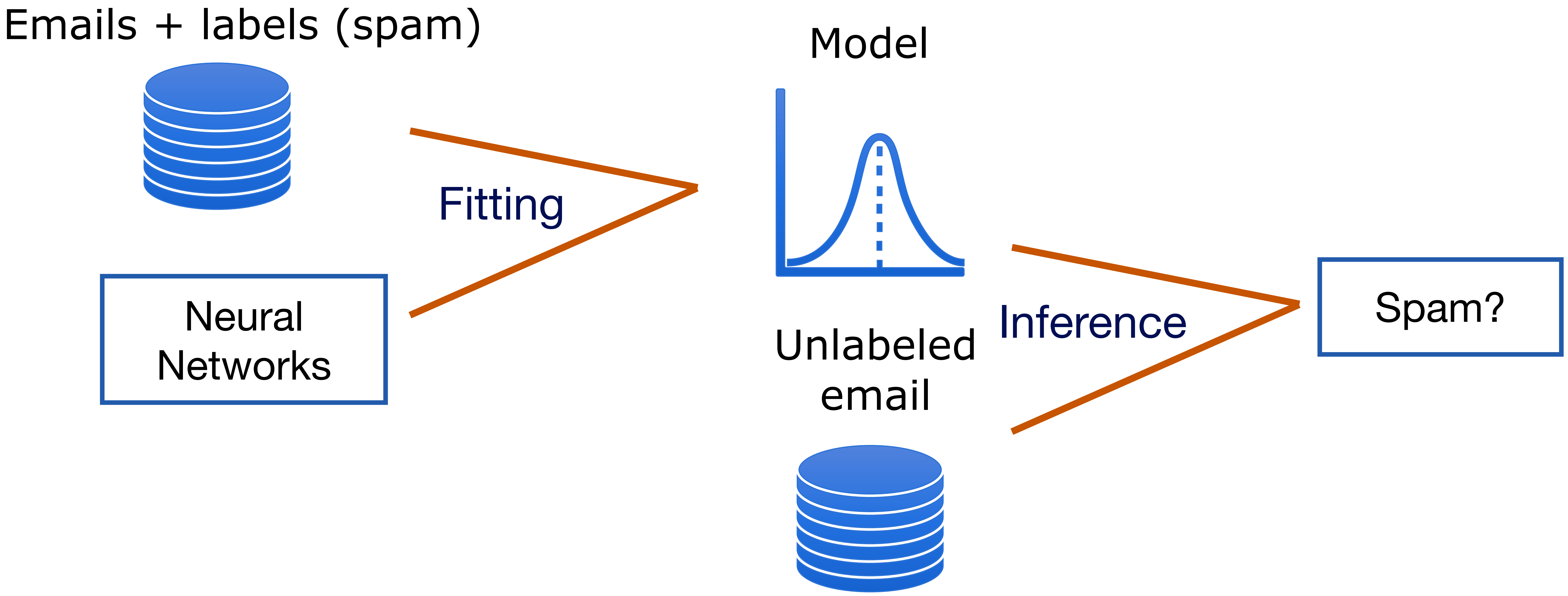
Model

Test Data

Inference

Predictions

The ML Paradigm



Emails + labels (spam)

Neural Networks

Fitting

Model

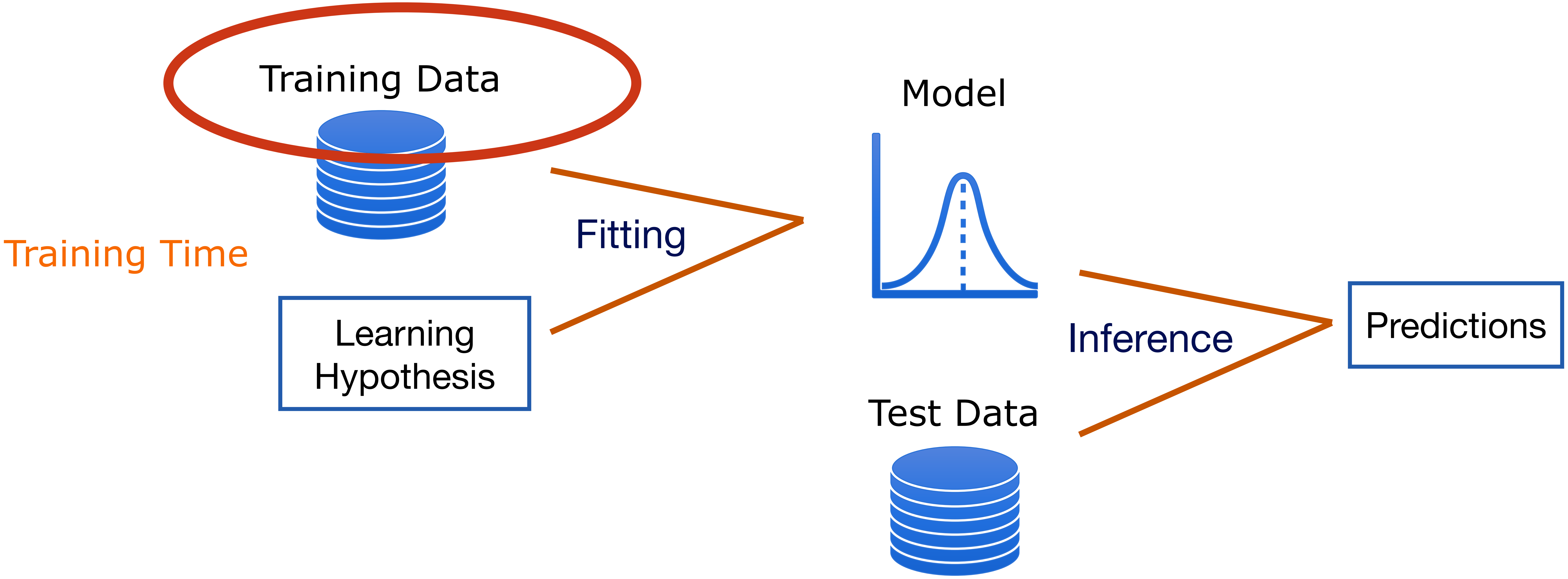
Unlabeled email

Inference

Spam?

The ML Paradigm in adversarial settings

Poisoning



Poisoning: An adversary inject bad data into the training pool (spam marked as not spam) and the model learns something it should not

The ML Paradigm in adversarial settings

Poisoning

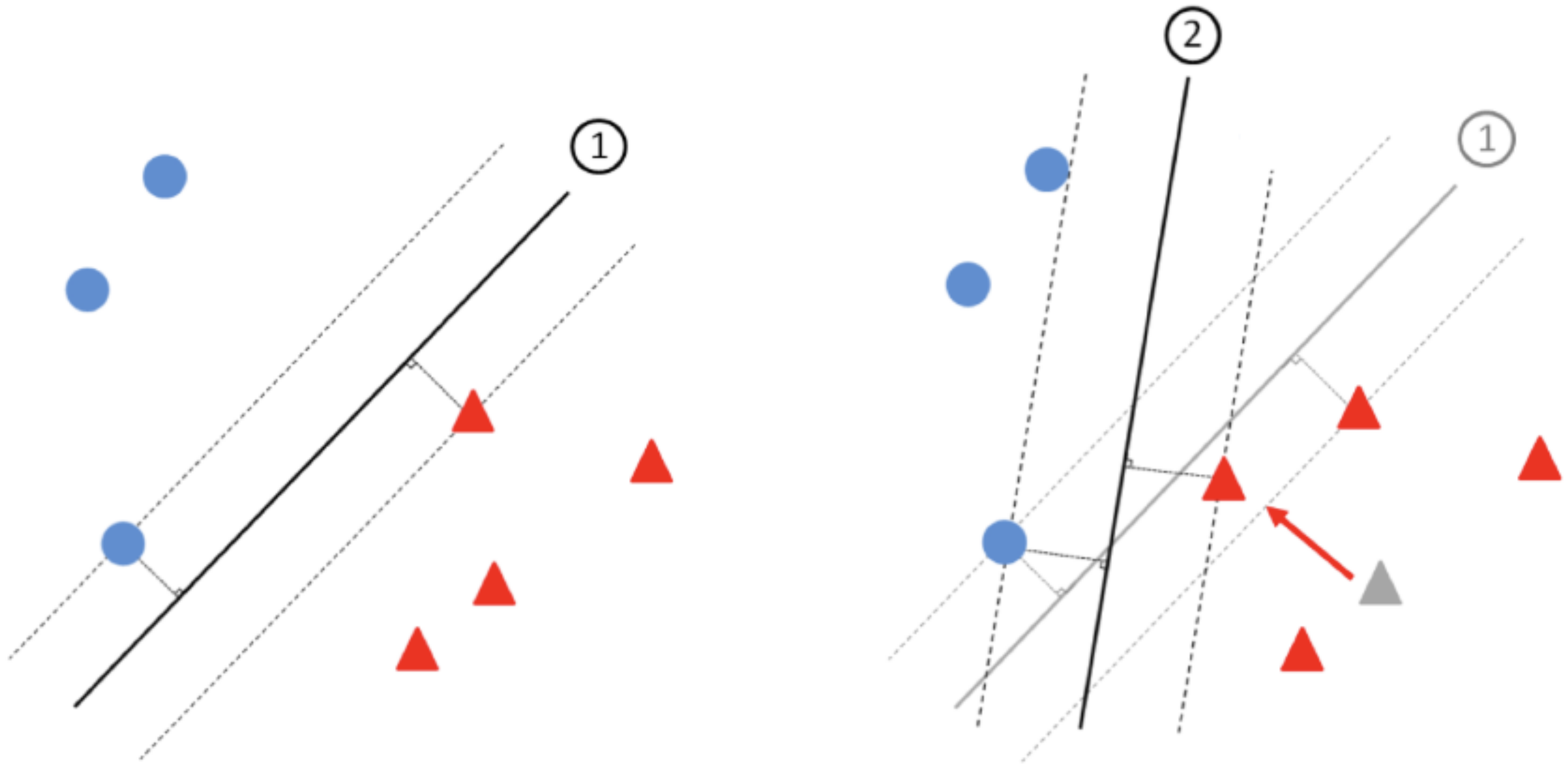
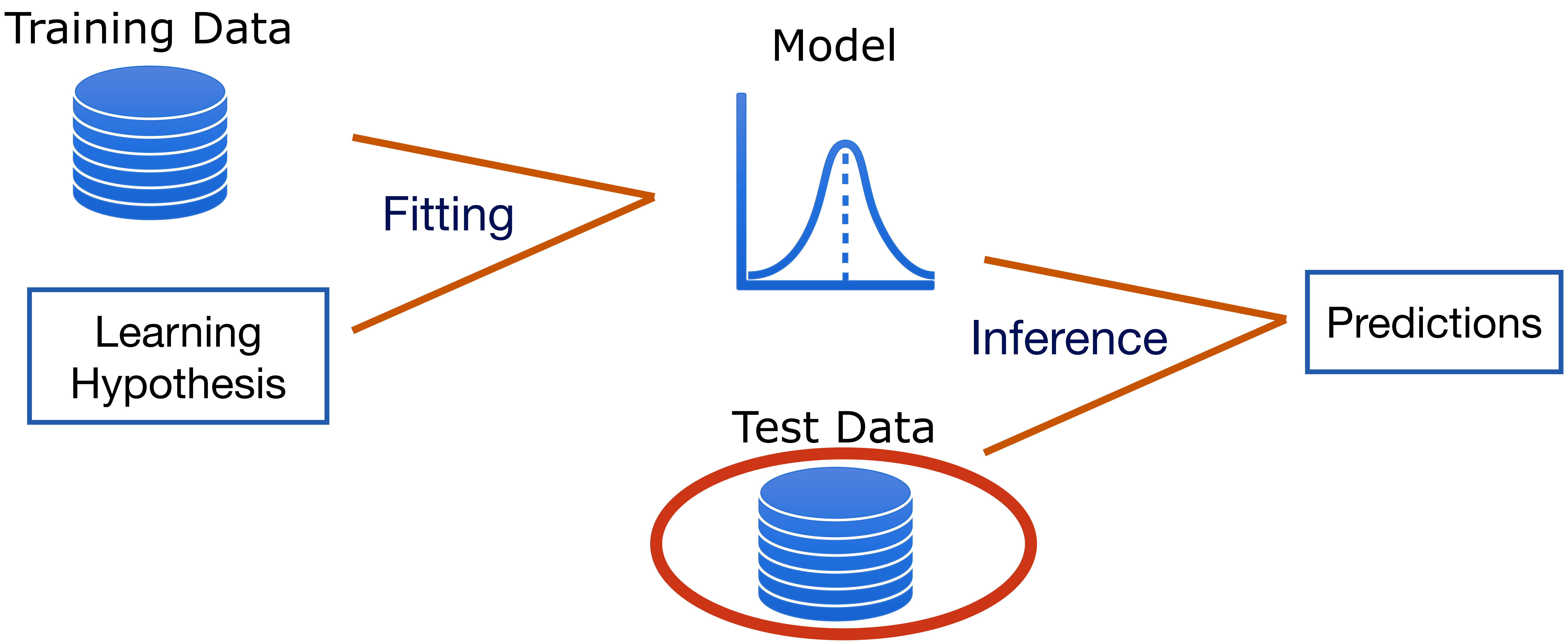


Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted if just one training sample is changed, even when that sample's class label does not change (right).

The most common result of a poisoning attack is that the model's boundary shifts in some way

The ML Paradigm in adversarial settings

Evasion

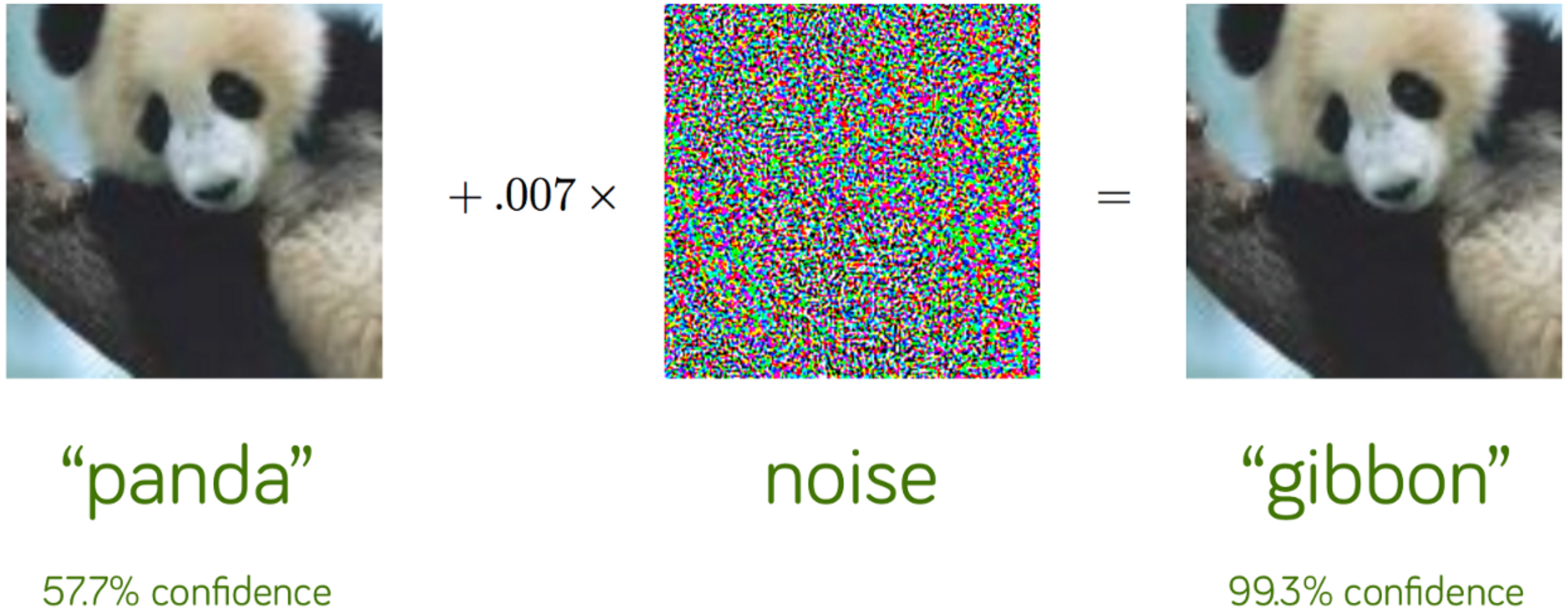


Evasion attacks: An adversary design adversarial examples that evades detection (spam marked as good)

The ML Paradigm in adversarial settings

Evasion

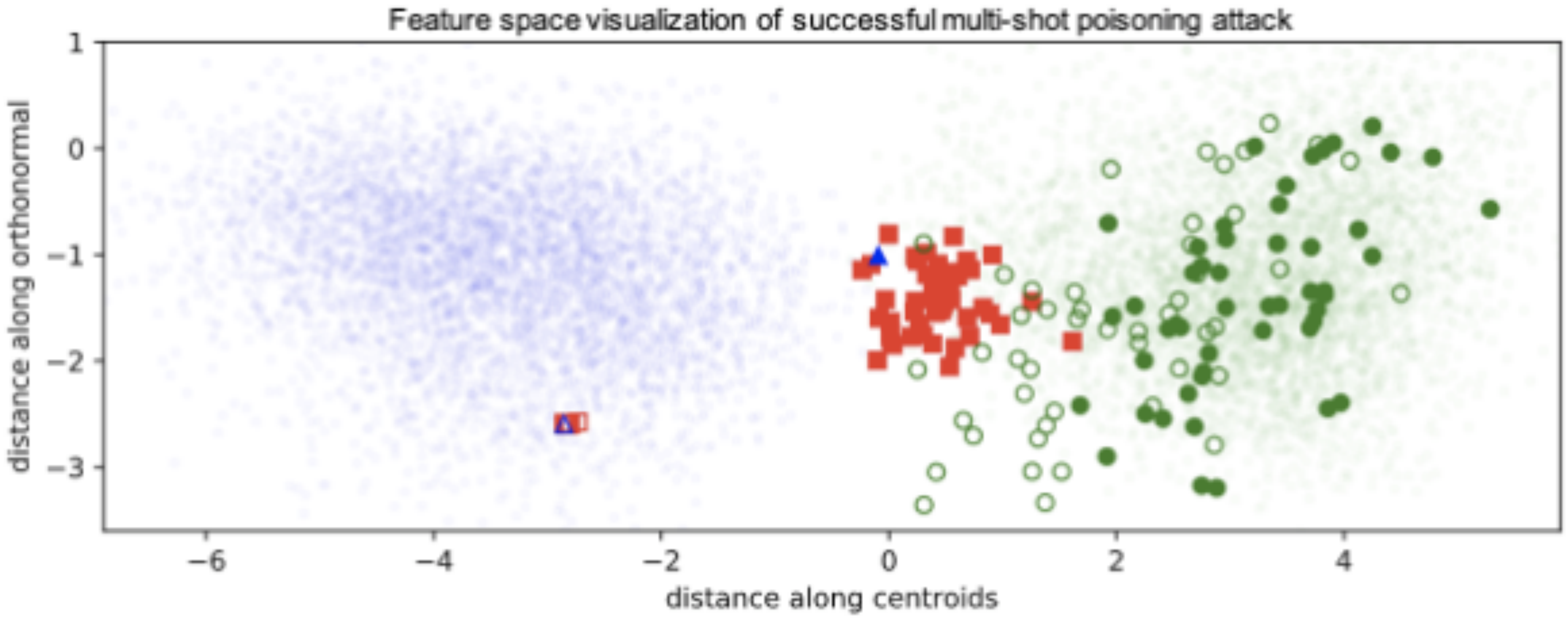
A typical example is to change some pixels in a picture before uploading, so that image recognition system fails to classify the result



The ML Paradigm in adversarial settings

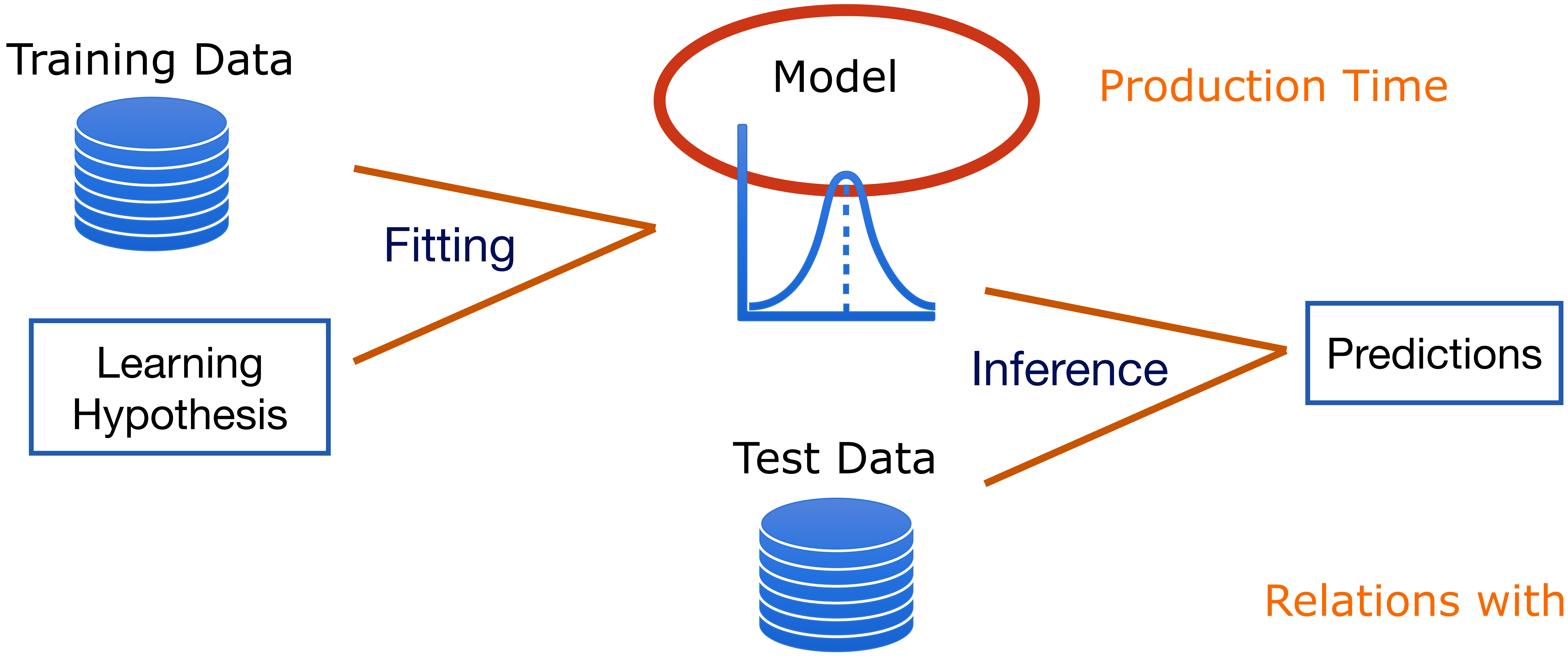
Evasion

These attacks pull the poisoned example across the “fixed” boundary (instead of shifting it)



The ML Paradigm in adversarial settings

Membership inference



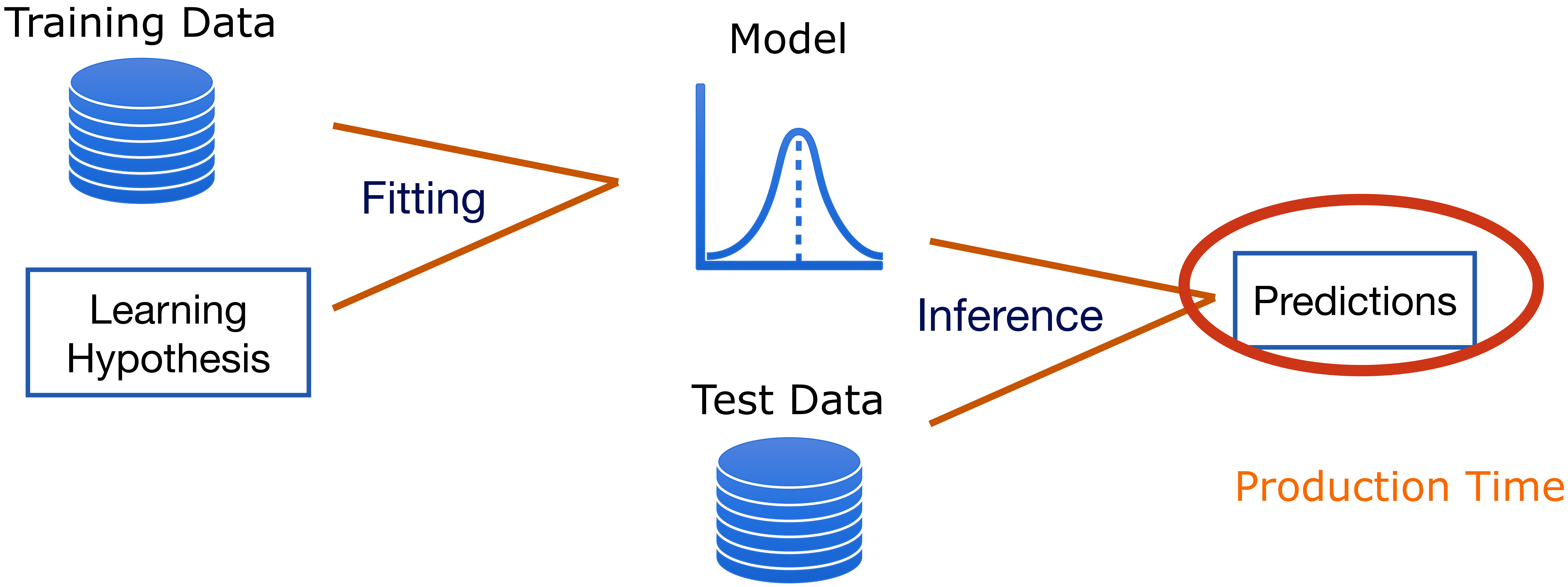
Production Time

Relations with Privacy!

Membership inference: Inspect model to detect if a user was in or not in the training data

The ML Paradigm in adversarial settings

Model extraction



Membership inference: Inspect model to detect if a user was in or not in the training data

Logistics

Class Info

- **Course and Info**
  on https://nandofioretto.github.io/teaching/raisp25/
- **Class meets on**
  Mondays and Wednesdays: 9:30 - 10:45 PM
  Rice 032
  - Lectures will be in person and attendance is required. If you are unable to attend a class (e.g., due to illness, job interviews, etc) please let the instructor and TAs know.
- **Office Hours**:
  - **Instructor** Fri: 9:00 - 10:00 AM @ Rice 307
  - **TA**: Tue: 2:30 - 3:30 PM @ Rice 442

Prerequisites

- **Some understanding of Machine Learning principles**
- **Stats and Probability**
  - Some understanding of Stats/**Probability** will be necessary to grasp concepts related with biases and unfairness as well as privacy.
- **Optimization**
  - Some of the work we'll cover will rely on some (convex) optimization principles

Assignments and Grading

- **Groups**
  - We have ~35 students in class, with a mix of MSc and PhD students and a few BSc.
  - During the first week we will create 7 groups.
  - Each group will be assessed through the following activities:
    - Paper Summaries (blogging): 33.33%
    - Presentation: 33.33%
    - Discussion Lead: 33.33%

Assignments and Grading

Paper Summaries (blogging)

Objective: To develop the ability to critically analyze and summarize AI research papers in a clear and accessible manner.

Expectations:

- Each group will review all papers from the provided list, and they may propose additional ones for approval.
- Summaries should be written in Markdown format (supporting images and formulas) and committed to the course's GitHub repository.
- The summary should include the following sections: **Introduction** and **Motivations**, **Methods**, **Key Findings**, and **Critical Analysis.**
- The Critical Analysis section should evaluate the strengths, weaknesses, potential biases, and ethical considerations of the paper.
- Summaries must be submitted **four days** prior to the presentation for review and potential feedback.

Assessment Criteria:

- Clarity and coherence of the written summary.
- Depth of critical analysis and understanding of the paper's content.
- Proper use of formatting and adherence to submission guidelines.
- Timeliness of submission.

Assignments and Grading

Presentations

Objective: To enhance students' ability to communicate complex AI concepts and engage in public speaking.

Expectations:

- 45-minute presentation per group.
- Presentations can include slides, code demonstrations, videos, or other creative methods.
- The presentation should cover the key aspects of the paper, including its contribution to responsible AI.
- A critical evaluation of the paper is essential, including discussing its limitations and implications.
- Preparation of thought-provoking questions to stimulate audience engagement.

Assessment Criteria:

- Effectiveness of communication and presentation skills.
- Accuracy and depth of content presented.
- Creativity and engagement in the presentation method.
- Ability to provoke thoughtful discussion through prepared questions.

Assignments and Grading

Discussion Lead

Objective: To cultivate skills in leading intellectual discourse and fostering collaborative learning.

Expectations:

- 30-minute discussion session following the presentation.
- Groups should prepare and facilitate a discussion based on their presentation.
- Use of supplementary materials (e.g., videos, code snippets) to enrich the discussion is encouraged.
- The discussion should engage the audience (with active questions), encouraging diverse viewpoints and deeper understanding of the topic.

Assessment Criteria:

- Ability to foster an inclusive and constructive discussion.
- Relevance and depth of prepared questions and discussion points.
- Engagement level of the audience during the discussion.
- Use of supplementary materials to enhance understanding.

Class format

- **45 minutes** presentation of reading materials and discussion.
  - Research papers or book chapters.
  - 2-3 presenters will present the slides/codes or other presentation material .
  - Everyone should be reading the material ahead, especially the released blog!
- **30 min** — Discussion and Q&A
  - 2-3 discussion leads will lead and moderate the discussion.
  - They should prepare slides with questions and discussion material.
- **Deadlines:**
  - **1 week** prior to the class: presenter submits slides and blog material
  - Revision and feedback sent back in 2 days if any

Presentation format

- Be creative!
  - Slides are okay
  - Interactive demos are great
  - Code tutorials are great
  - Combination of the above is awesome
- Requirements:
  - Involve the class in active discussion
  - Cover all papers assigned
- Questions:
  - Can I use other authors’ available material? Yes — with disclaimer

Presentation grading

- Rubric link
- **Technical:**
  - Depth of the content
  - Accuracy of the content
  - Discussion of the paper Pro and Cons
  - Discussion Lead
- **Non-technical**
  - Time management
  - Responsiveness to the audience
  - Organization
  - Presentation Format

Assignments and Grading

Contributions

- All group members are expected to contribute equally to all activities, but 2-3 members are expected to lead each of the three components.
- Peer evaluation within group may be used to ensure fair contribution
-

Honor Code

- We trust every student in this course to fully comply with all of the provisions of the University’s Honor Code.
- **Ethics**: Submissions should acknowledge all collaborators and sources consulted. All codes should be original. We will be actively checking for plagiarism.
-

Use of Generative AI Tools

- The use of GenAI tools is permitted, but not encouraged.
- The use of these tools is a **privilege** and comes with responsibility. Adhere to the guidelines reported in the syllabus and approach the use of these tools with integrity and critical thought.
    - **Disclosure Requirement:** You are required to report if you have used genie tools. If used, you must report the name and types of the tools employed. All outputs (both explicitly used and inspiration in one's submission) must be cited.
    - **Ethical Implications:** While genAI tools can be a powerful aid, they must be used responsibly and in accordance with the principles of academic honesty. Please reflect on the ethical dimension of using these tools, recognizing they are a means to **enhance learning and not shortcuts to bypass understanding**.
    - **Verification challenges:** Verifying outputs of genAI tools can be complex. You should be aware that relying solely on genAI responses may lead to incorrect conclusions. You are urged to **think critically** and to evaluate and verify the correctness of genAI tools outputs.

Build a great community

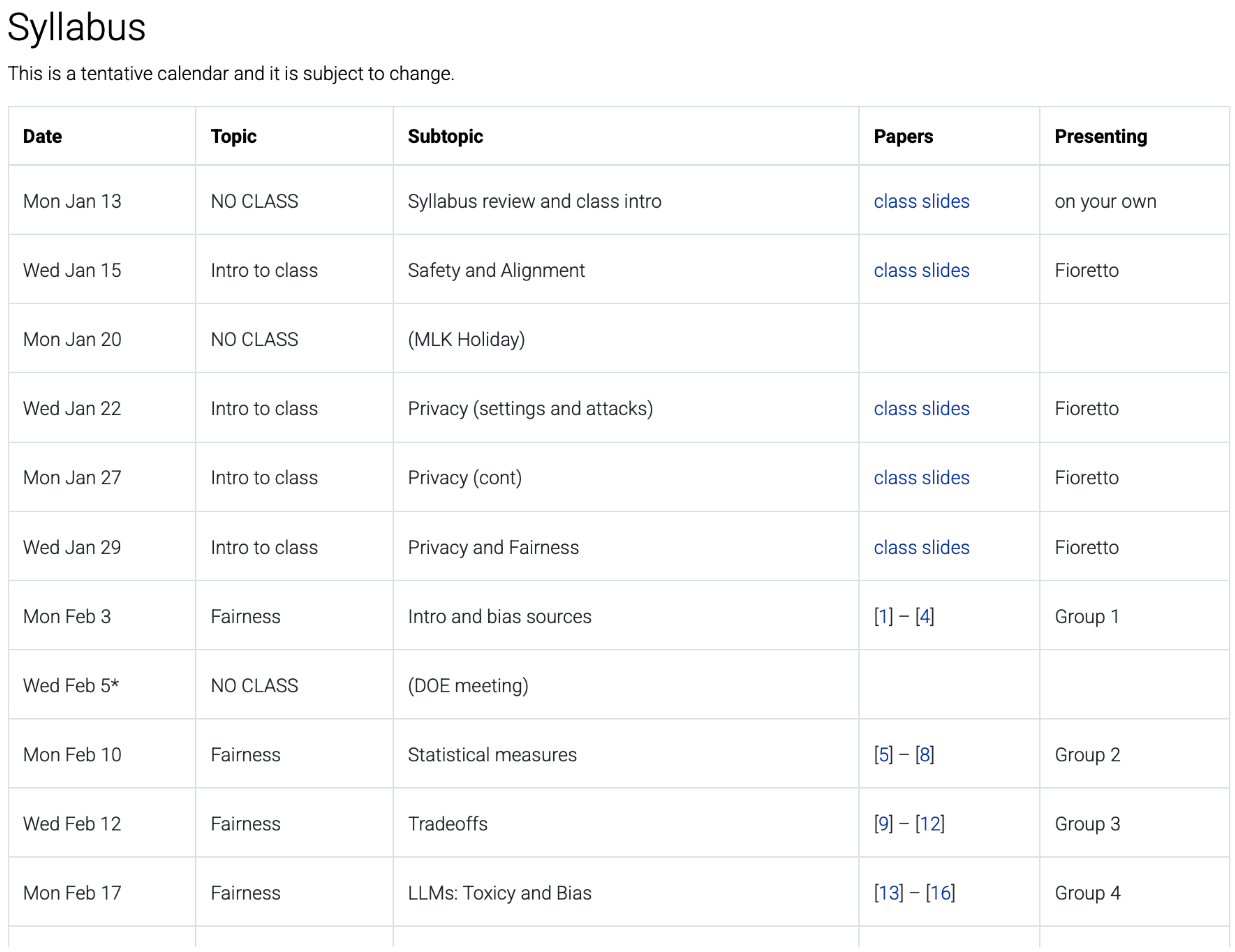
- Help out your peers!
- Be mindful of the tone you use – be respectful and supportive, help everyone feel at home.
  - Also, please don't interrupt your peers or instructors.
- Watch out for implicit bias – catch yourself before acting on it.
  - Someone's gender, race, ethnicity, sexual orientation, etc. do NOT have anything to do with how awesome they will be in this class.
  - Having a ton of programming experience will help some with projects, but does NOT give anyone an edge on how well they can understand the material and how highly they can score on the course.

Topics

- 3 days, including today of introductions to topics — presented by me.
- From Feb 3, you'll lead the class!
- **Topics**
  - Fairness
  - Safety
  - Privacy
  - Evaluation
  - Unlearning
  - Misuse of AI and Governance

Important This Week

- Check which group are you (1-7) — you will be assigned by Friday, Jan 17.
  - **Check** when you'll be presenting/ blogging.



Syllabus

This is a tentative calendar and it is subject to change.

Date	Topic	Subtopic	Papers	Presenting
Mon Jan 13	NO CLASS	Syllabus review and class intro	class slides	on your own
Wed Jan 15	Intro to class	Safety and Alignment	class slides	Fioretto
Mon Jan 20	NO CLASS	(MLK Holiday)		
Wed Jan 22	Intro to class	Privacy (settings and attacks)	class slides	Fioretto
Mon Jan 27	Intro to class	Privacy (cont)	class slides	Fioretto
Wed Jan 29	Intro to class	Privacy and Fairness	class slides	Fioretto
Mon Feb 3	Fairness	Intro and bias sources	[1] – [4]	Group 1
Wed Feb 5*	NO CLASS	(DOE meeting)		
Mon Feb 10	Fairness	Statistical measures	[5] – [8]	Group 2
Wed Feb 12	Fairness	Tradeoffs	[9] – [12]	Group 3
Mon Feb 17	Fairness	LLMs: Toxicy and Bias	[13] – [16]	Group 4

Responsible AI:

Seminar on Fairness, Safety, Privacy and more

Thank you!

https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto

Ferdinando Fioretto @UVA Spring 2025