Syracuse University

CIS700: Security and Privacy of Machine Learning

Prof. Ferdinando Fioretto
ffiorett@syr.edu

Introductions

- Name (how you like to be called)
- Position (MS / PhD) and year
- Research Interests
- What do you expect from this course!

Overview

- Syllabus: www.nandofioretto.com —> TEACHING —> FALL 2021
  - Schedule and Material
  - Assigned reading (will be updated)
  - Grading information
  - Projects
  - Ethics statement
- Class Schedule: Mon + Wed 5:15 — 6:35pm
- Office Hours: Fri 12:30 — 1:30pm
- Office Location: 4-125 CST

Discord

- Join the Discord channel: https://discord.gg/JTDCA6eR
  - Send me your email (if you have not received an invitation) at ffiorett@syr.edu with email subject: "CIS700 Discord contact"
- To be used for:
  - All form of communication with teammates, class, and me
  - All submissions: Presentation slides, reports, projects
    #report-submission (for your report submissions
    #slides-submission (…)
    #paper-discussion (Q&A about papers between classmates)

What is this class about?

- This is not an ML course!
- Seminar-type class: we will read lots of papers

Security

Privacy

Class Format

- 1h presentation of reading materials
    - Research papers or book chapters
    - One presenter will present and lead the discussion
    - Everyone should be reading the material ahead!
    - Everyone has to ask at least 1 question! (e.g., in a round-robin scheme, I will moderate)
- 20 min — Discussion and Q&A (but should arise during the presentation!)
- Deadlines:
    - 2 days prior to the class: presenter submits slides (by 11:59pm)

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: https://www2.isye.gatech.edu/~fferdinando3/classes/spring21/rubric.pdf
- 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

Research Project

- Take a look at the class topics and papers
- Identify one are of interest
- Formulate a project proposal (1/2 page, Initial Project review: March 22)
    - Title
    - Team (optional) — at most 2 people
    - Problem
    - Methods
- Exampels include:
    - Extended literature review on a topic
    - Implementations of attacks/defense mechanisms
    - Implementation of privacy-preserving approaches
- Project report: May 10

Grading Scheme

- 50 % paper presentation
- 10 % class participation
- 40 % research project

Integrity

Please take a moment to review the Code of Student conduct https://policies.syr.edu/policies/academic-rules-student-responsibilities-and-services/code-of-student-conduct/

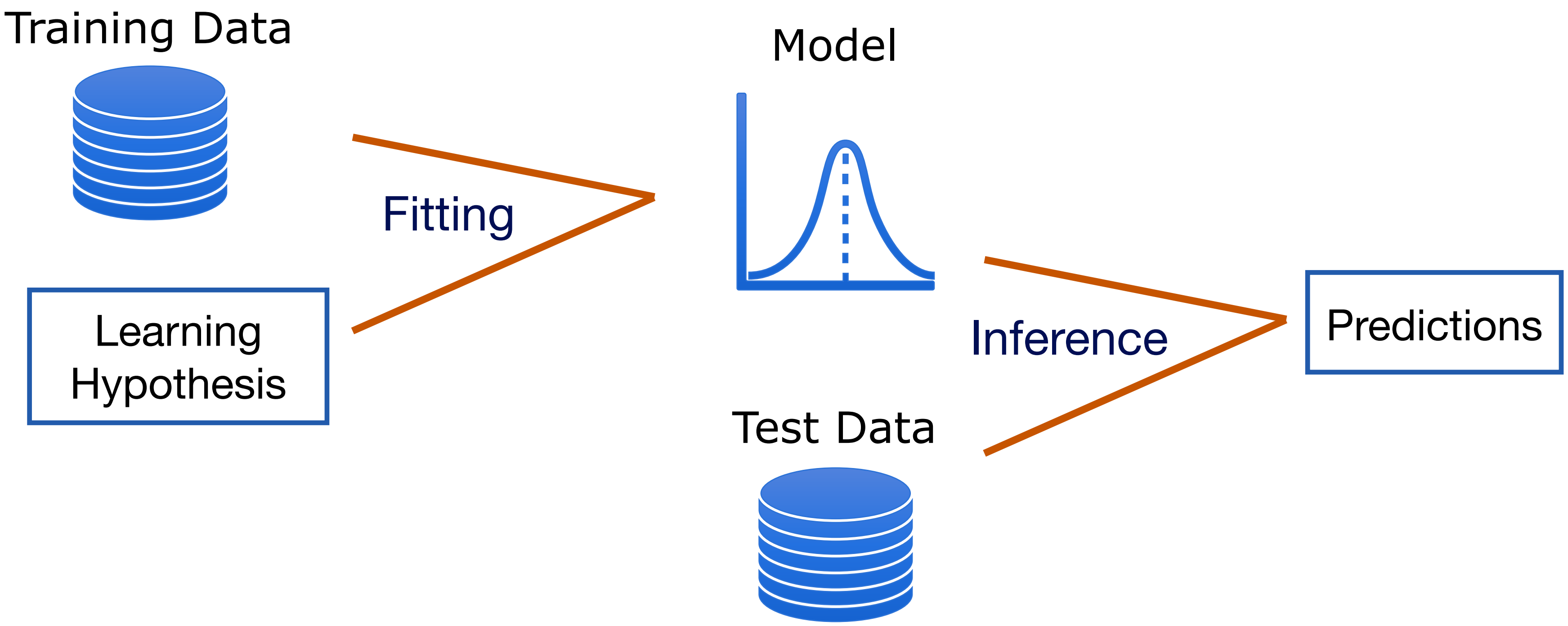
Instances of plagiarism, copying, and other disallowed behavior will costitute a violation of the code of student conduct. Students are responsible for reporting any violation of these rules by other students, and failure to do so constitute a violation of the code of student conduct.

Ethics

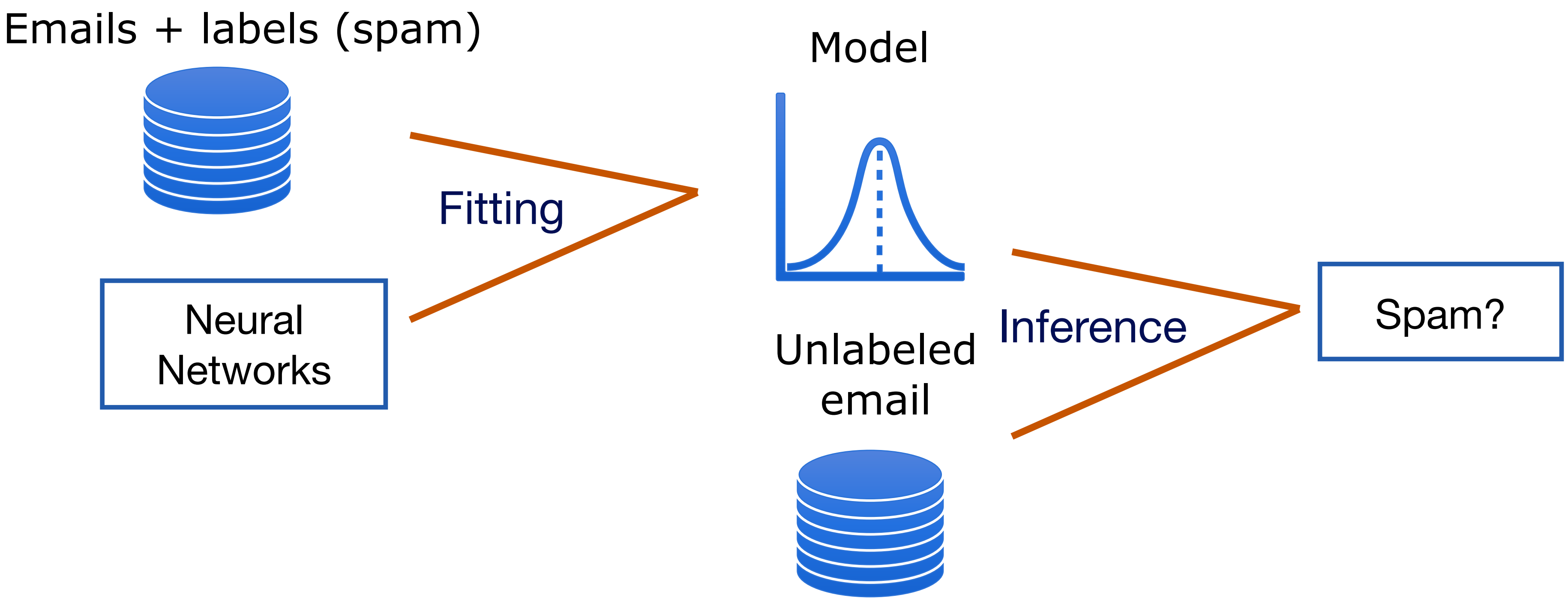
In this course, you will be learning about and exploring some vulnerabilities that could be exploited to compromise deployed systems. You are trusted to behave responsibility and ethically. You may not attack any system without permission of its owners, and may not use anything you learn in this class for evil. If you have doubts about ethical and legal aspects of what you want to do, you should check with the course instructor before proceeding.

Any activity outside the letter or spirit of these guidelines will be reported to the proper authorities and may result in dismissal from the class.

The ML Paradigm



The ML Paradigm



Emails + labels (spam)

Neural Networks

Fitting

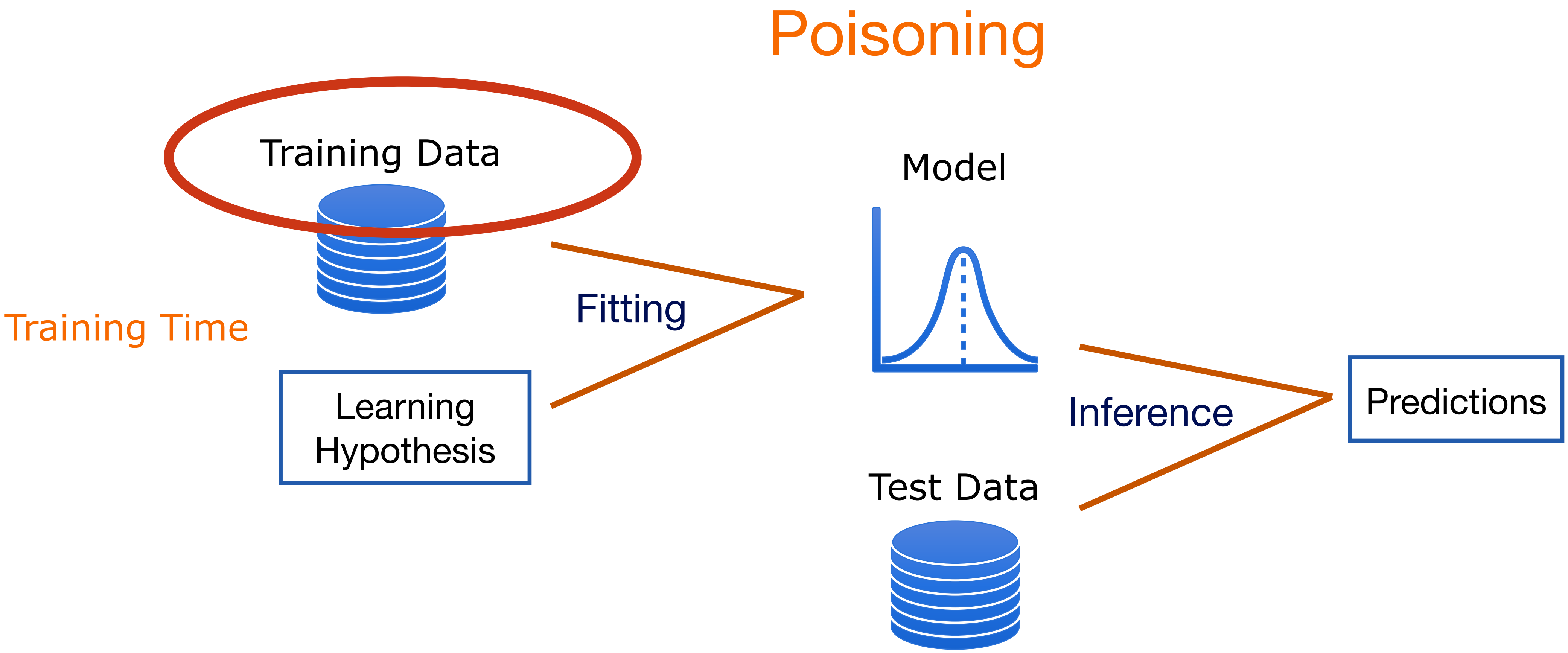
Model

Unlabeled email

Inference

Spam?

The ML Paradigm in Adversarial Settings



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

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

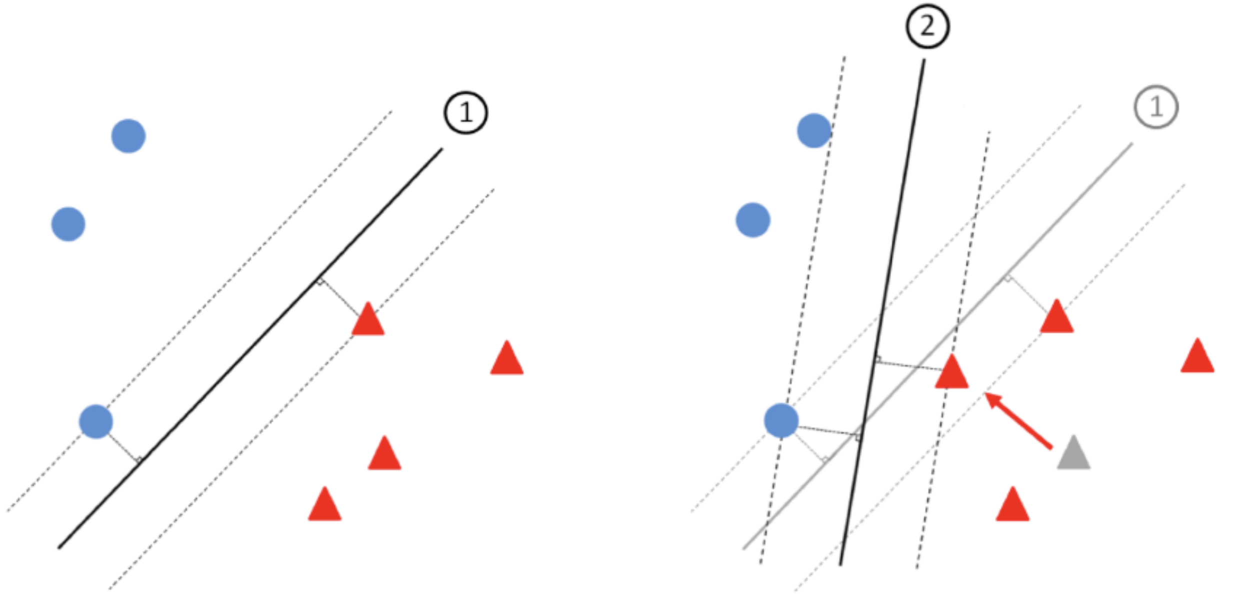
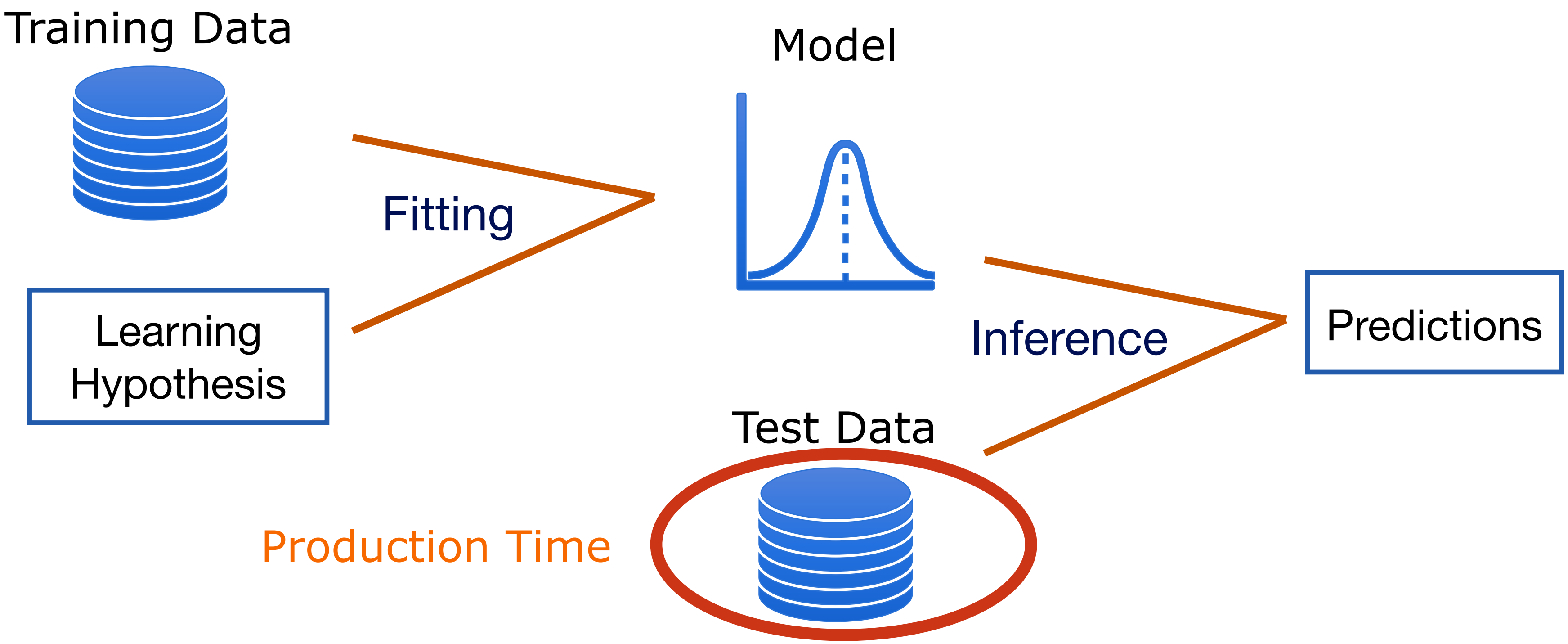


Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted if just one training sample is changed, even when that sample's class label does not change (right).

The ML Paradigm in Adversarial Settings

Evasion

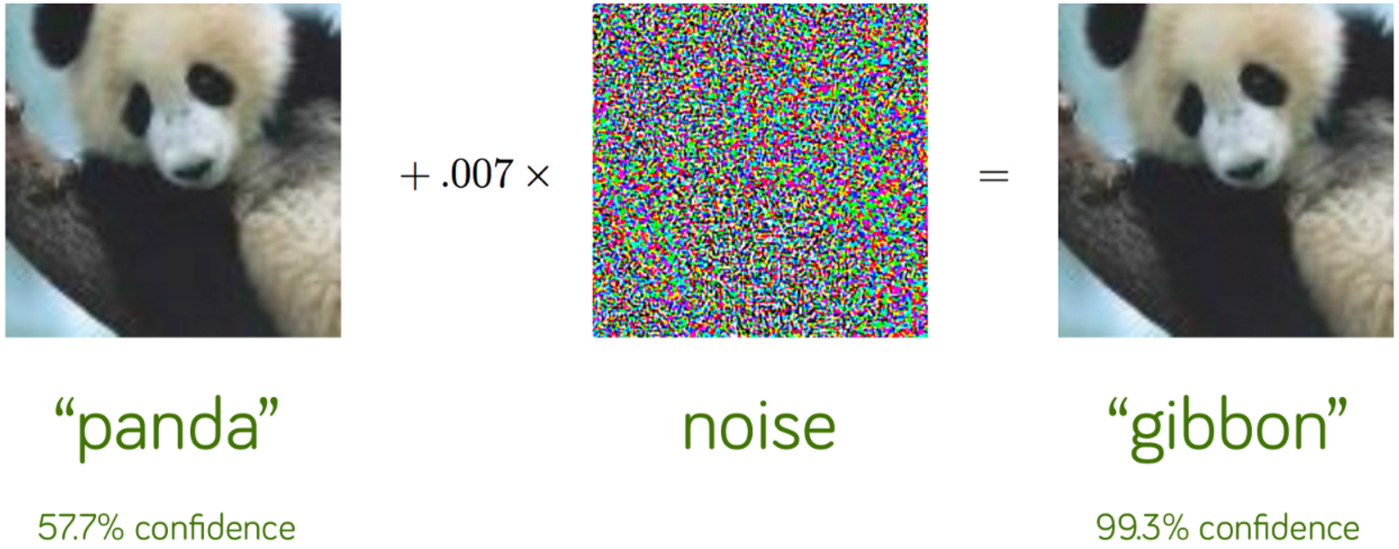


Poisoning: An adversary design adversarial examples that evades detection (spam marked as good)

The ML Paradigm in Adversarial Settings

Evasion

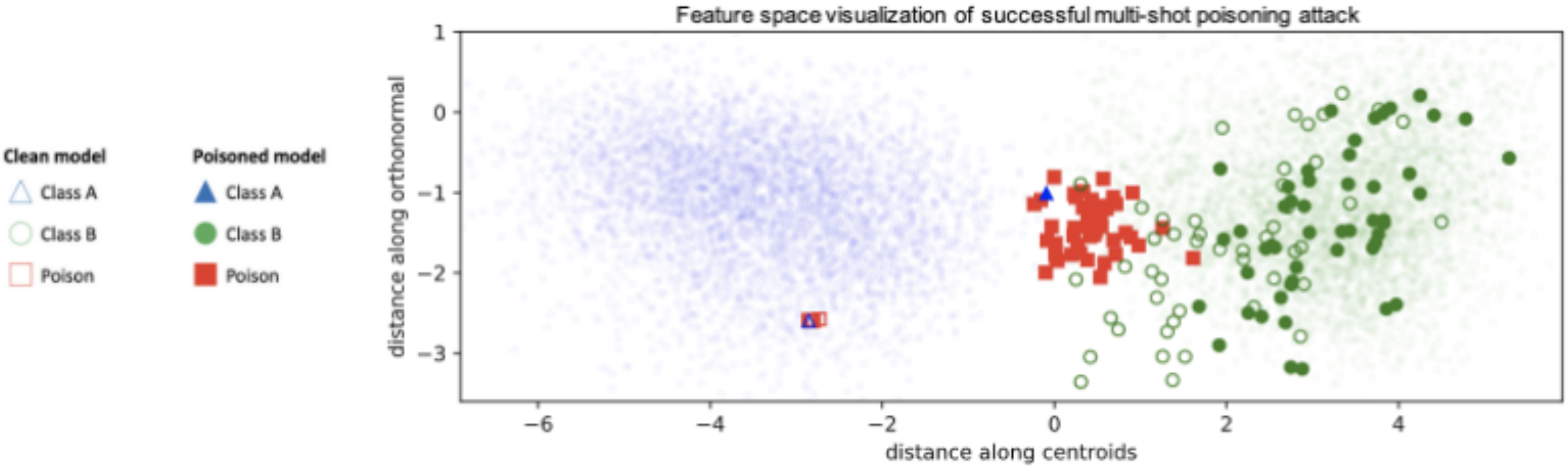
A typical example is to change some pixels in a picture before uploading, so that image recognition system fails to classify the result



The ML Paradigm in Adversarial Settings

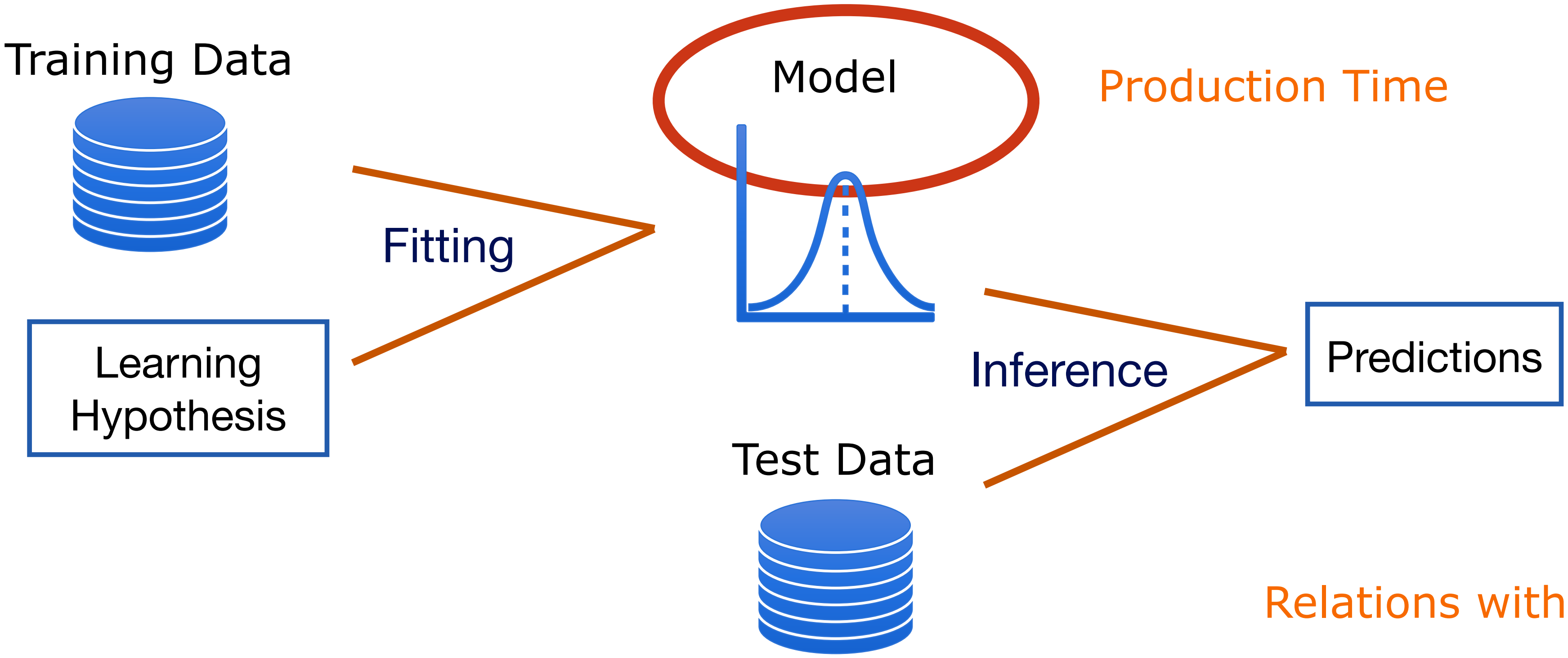
Evasion

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



The ML Paradigm in Adversarial Settings

Member Inference



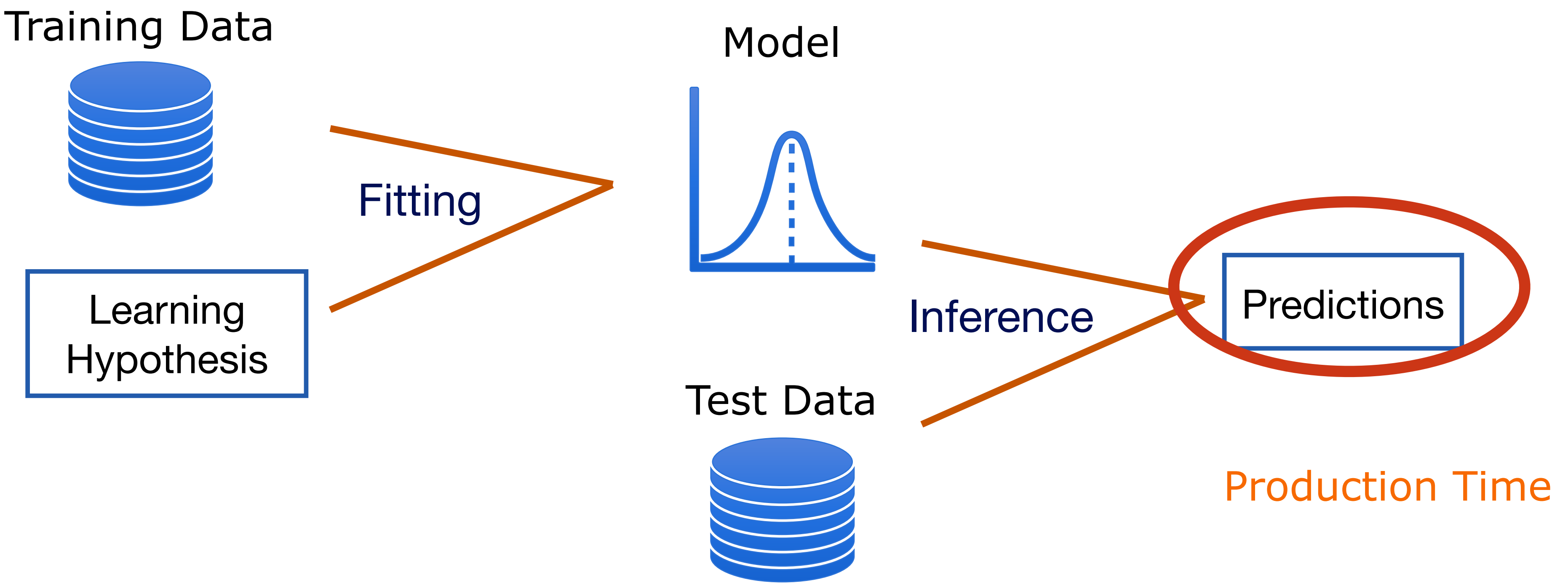
Production Time

Relations with Privacy!

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

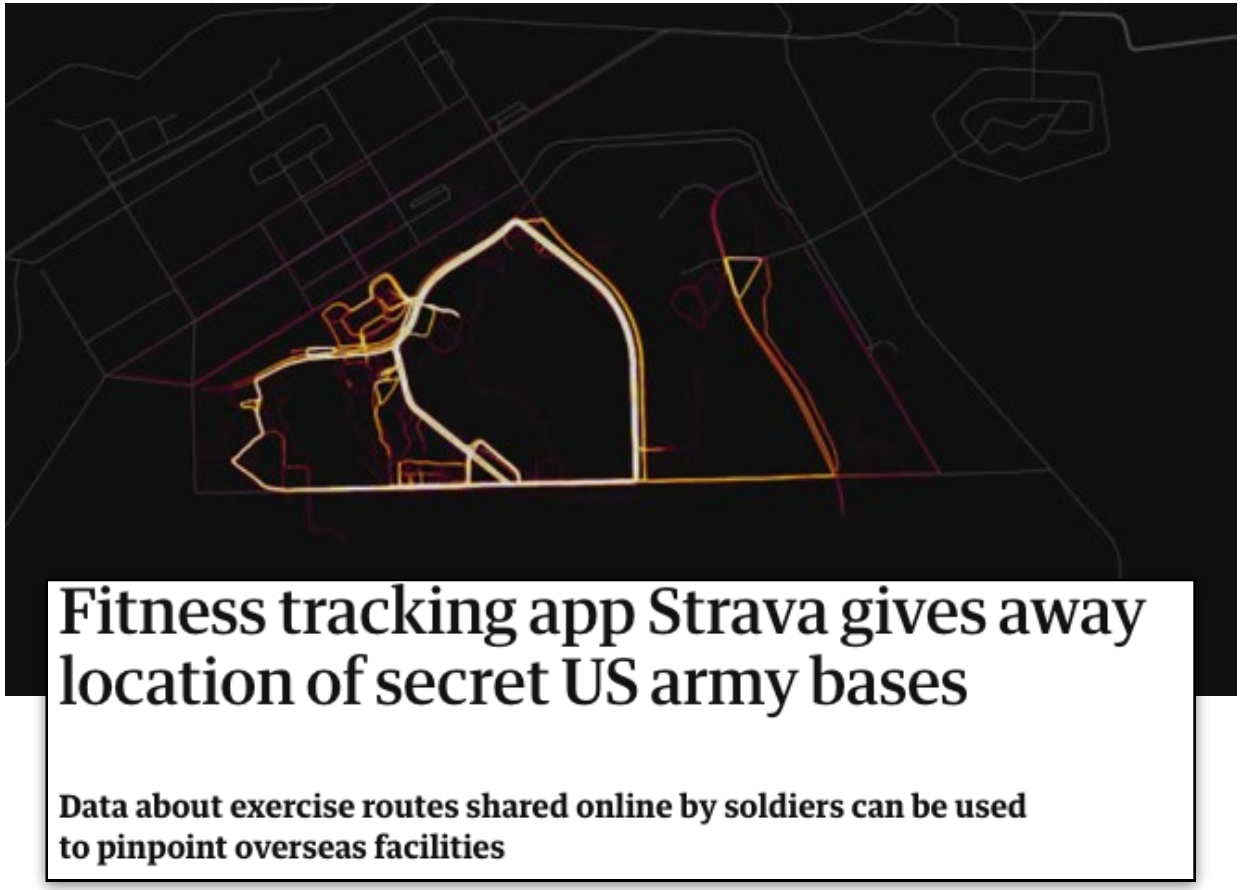
The ML Paradigm in Adversarial Settings

Model Extraction



Model extraction: The adversary observes predictions and reconstructs the model locally

Privacy

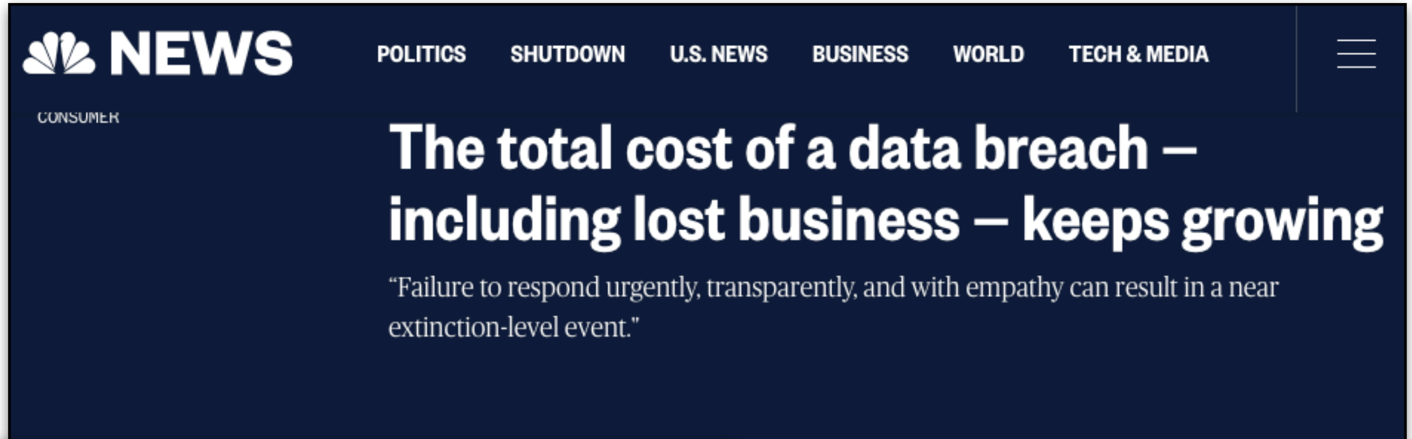


Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

The Cost of Privacy

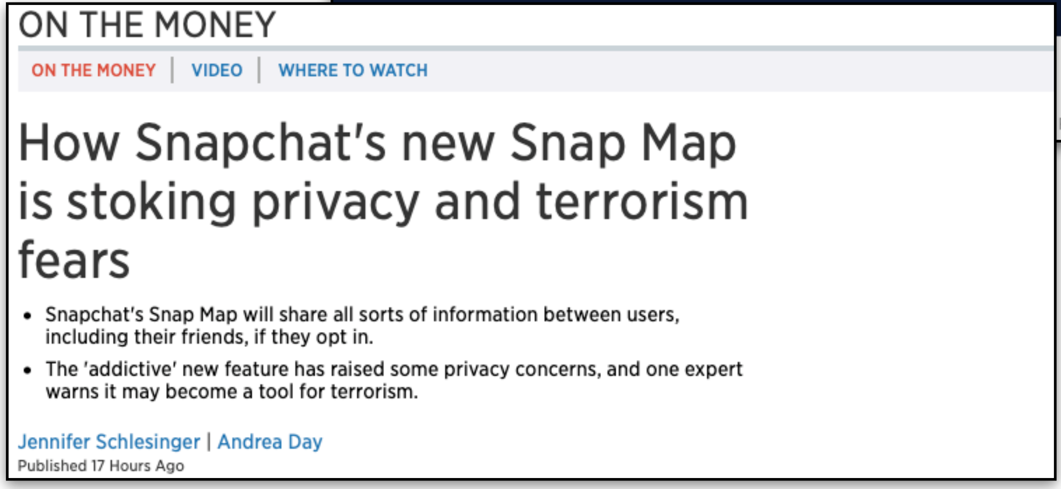
$3.86



The total cost of a data breach – including lost business – keeps growing

"Failure to respond urgently, transparently, and with empathy can result in a near extinction-level event."

PM EDT

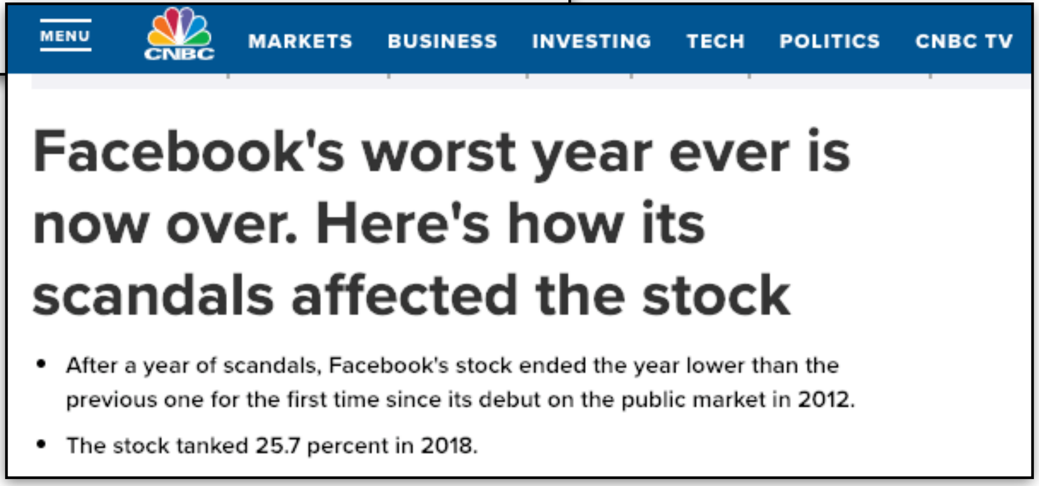


How Snapchat's new Snap Map is stoking privacy and terrorism fears

- Snapchat's Snap Map will share all sorts of information between users, including their friends, if they opt in.
- The 'addictive' new feature has raised some privacy concerns, and one expert warns it may become a tool for terrorism.

Jennifer Schlesinger | Andrea Day

Published 17 Hours Ago

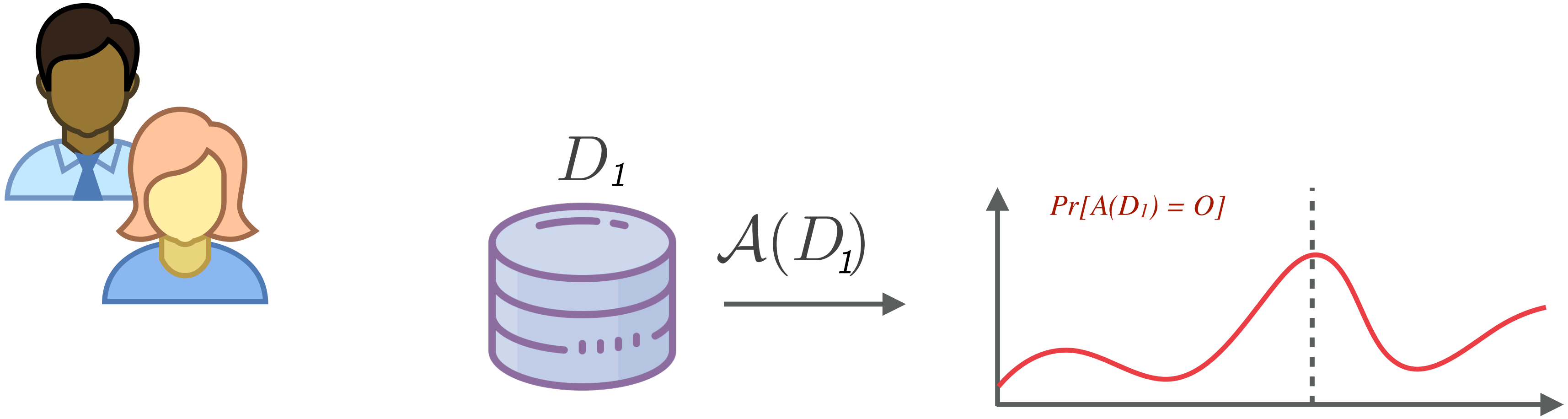


Facebook's worst year ever is now over. Here's how its scandals affected the stock

- After a year of scandals, Facebook's stock ended the year lower than the previous one for the first time since its debut on the public market in 2012.
- The stock tanked 25.7 percent in 2018.

Differential Privacy

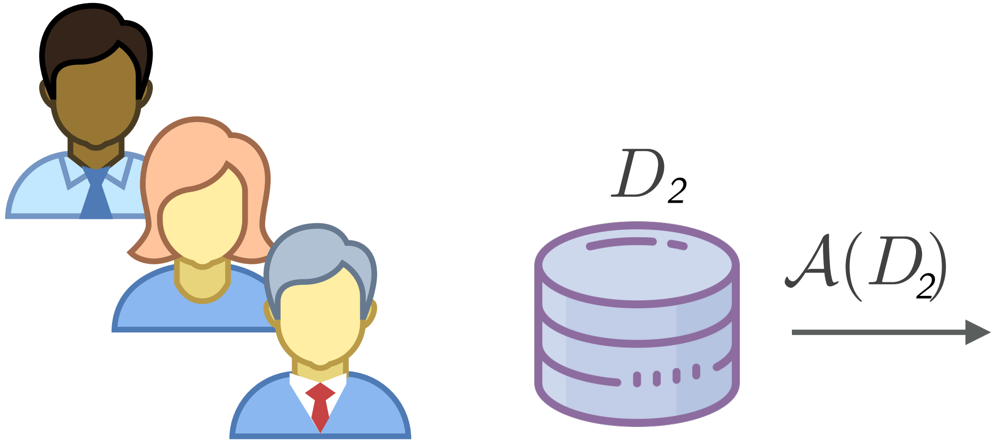
$$\frac{\Pr[\mathcal{A}(D_1) = O]}{\Pr[\mathcal{A}(D_2) = O]} \leq \exp(\epsilon)$$

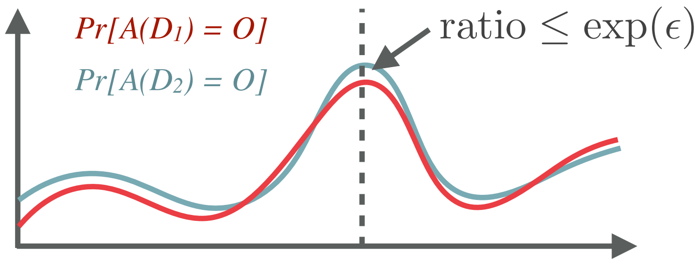


Differential Privacy

$$\frac{\Pr[\mathcal{A}(D_1) = O]}{\Pr[\mathcal{A}(D_2) = O]} \leq \exp(\epsilon)$$

ε = privacy budget

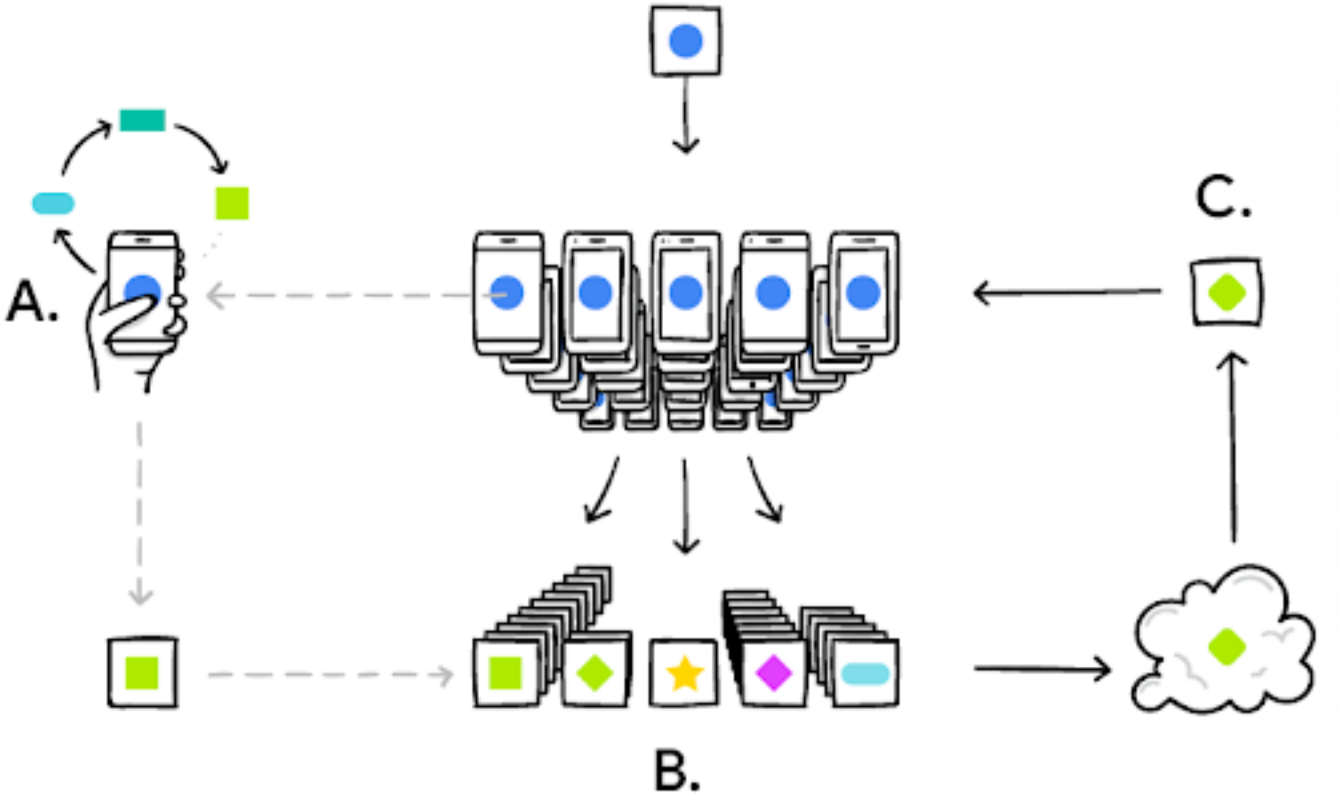


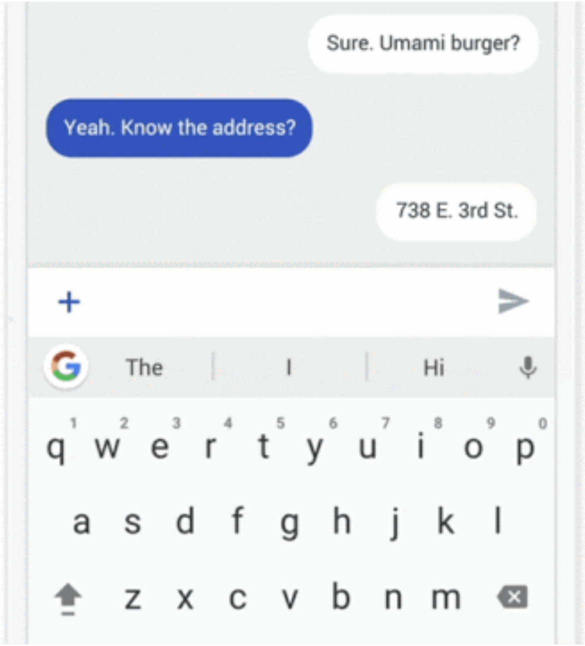


Controls the degree to which D_1 and D_2 can be distinguished.

Small **ε** gives more privacy (and worse utility)

Federated Learning



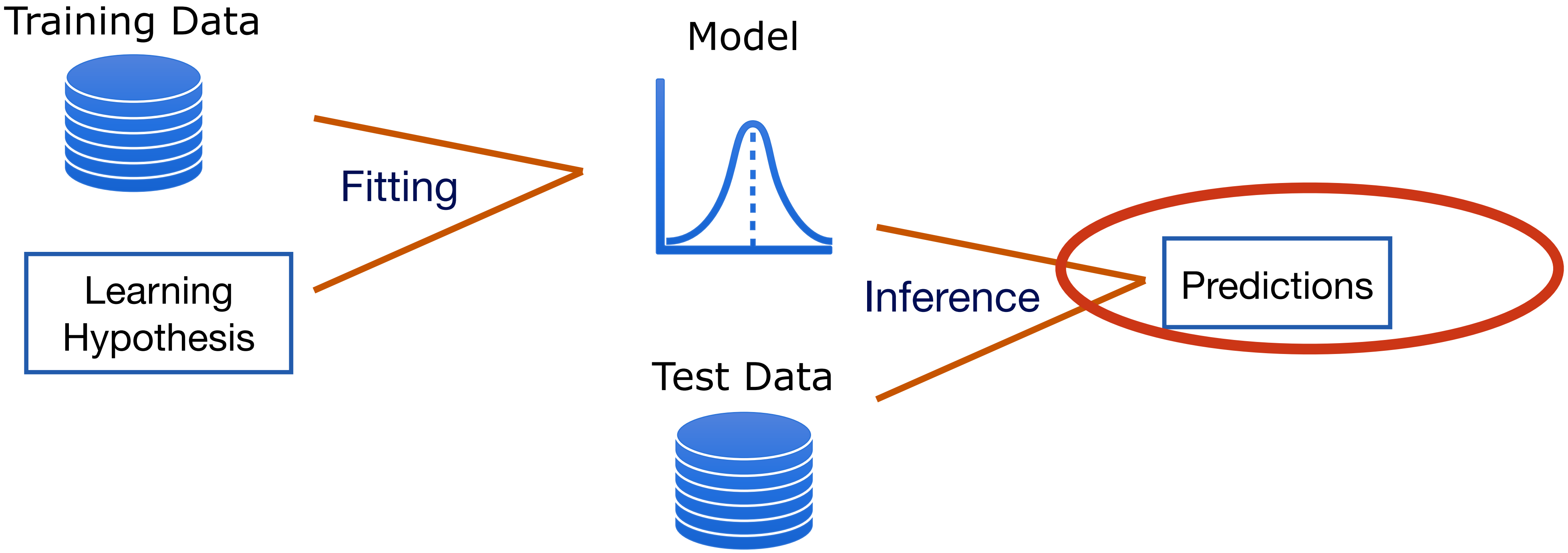

Sure. Umami burger?

Yeah. Know the address?

738 E. 3rd St.

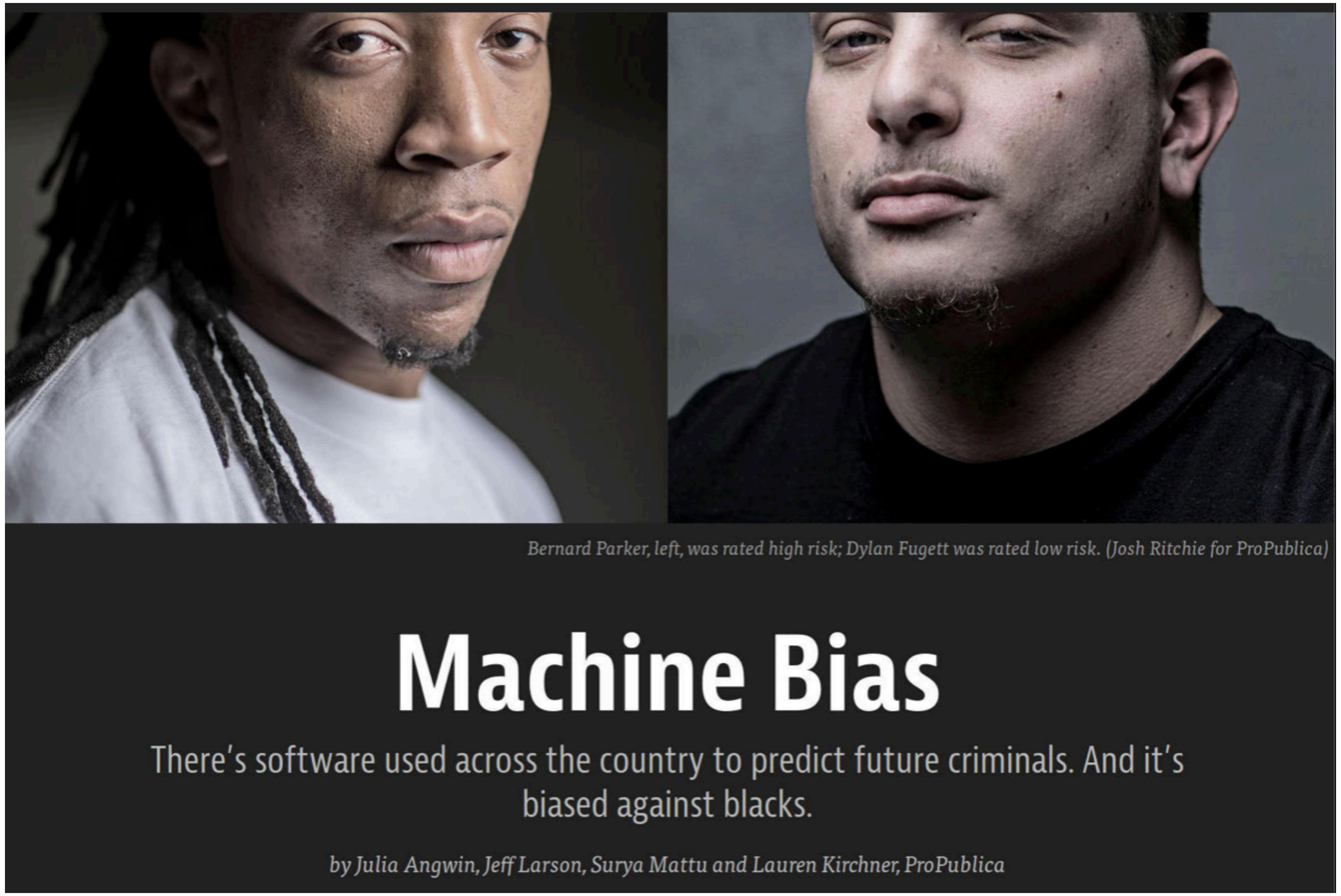
https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Fairness



Fairness: If training data is biassed toward a subpopulation, the accuracy for the minority party suffer, at inference

Fairness



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

Modules

1. Evasion Attacks (and defense)
2. Poisoning Attacks (and defense)
3. Privacy Attacks
4. Foundations of Differential Privacy
5. DP and ML
6. DP model extensions
7. Federated Learning

Before Going

- Write down your name + 2 things you hope to learn in this class.
- I will be slacking a message from Prof. Salekin about recording your presentation for a data collection study. Please read it carefully, and let him know if you'd opt-out for this study.