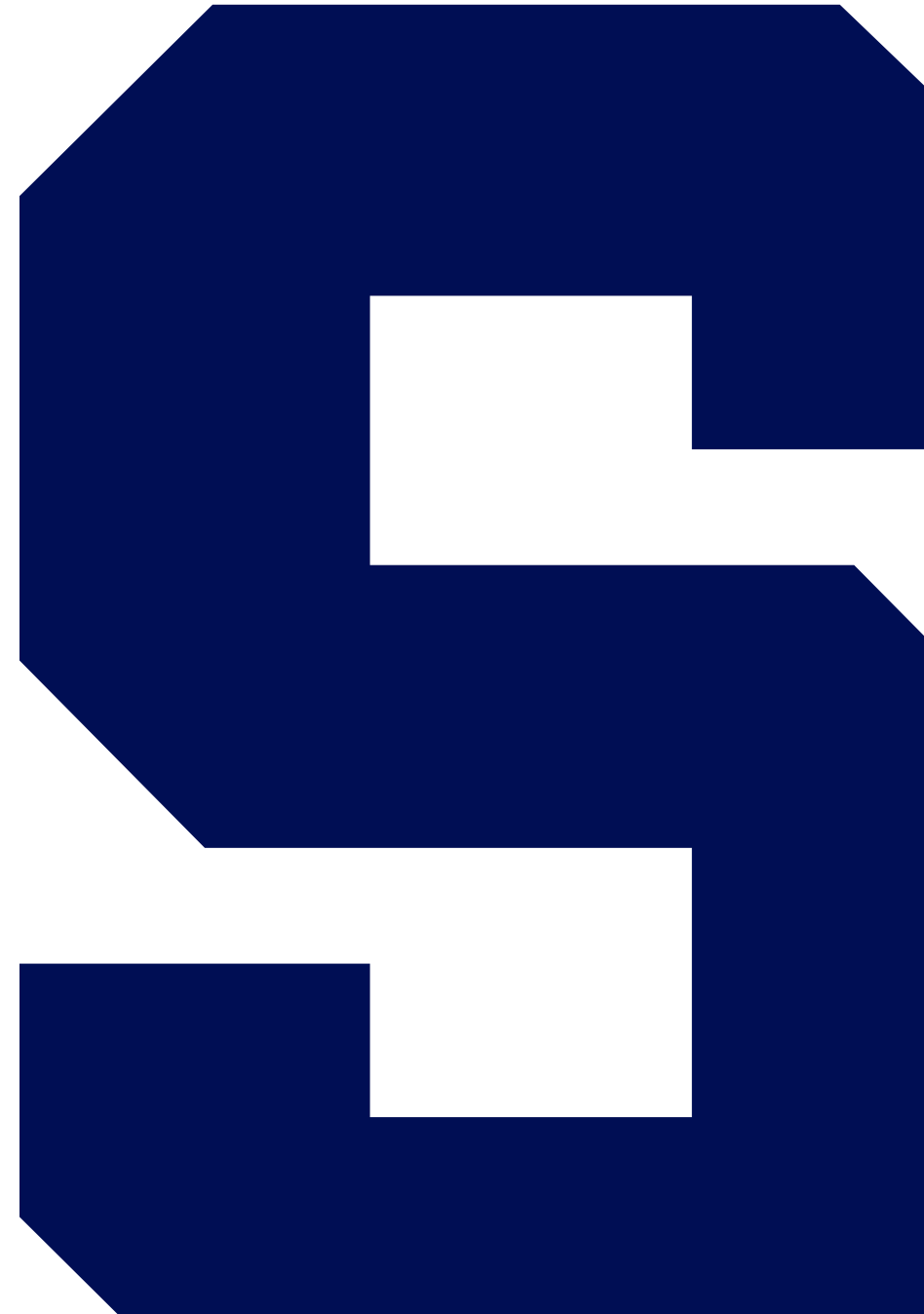


CIS700: Security and Privacy of Machine Learning

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!

Preliminaries

- Syllabus: <http://web.ecs.syr.edu/~ffiorett/classes/spring20.html>
 - Schedule and Material (will be updated)
 - Teams (more on this later)
 - Assigned reading (will be updated)
 - Assigned reports (will be updated)
 - Grading information
 - Ethics statement
- Class Schedule: Mon + Wed 5:15 — 6:35pm
- Office Hours: Fri 12:30 — 1:30pm
- Office Location: 4-125 CST

Slack!

- (Couse' we all like to slack a bit)
- Join the Slack channel: ff-cis700-spring20.slack.com
 - Send me your email (if you have not received an invitation) at ffiorett@syr.edu with email subject: "CIS700 Slack contact"
 - Accept the invitation (you may have already received it)
- To be used for:
 - All form of communication with teammates, class, and me (please don't slack me too much)
 - All submissions: Presentation slides, reports, projects
 - #report-submission (for your report submissions)
 - #slides-submission (...)
 - #paper-discussion (Q&A about papers between classmates)

The Team Universe

Alderaan



Mu Bai
Cuong Tran
Lin Zhang

Coruscant



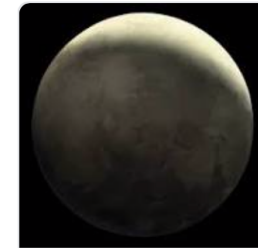
Zuhal Altundal
David Castello
Weiheng Chai

Kamino



S. Dinparvar
M. SP Madala
??

Mandalore



Amin Fallahi
Jindi Wu
Tejas Bharambe

Naboo



Kunj Gosai
Ankit Khare
Haoyu Li

Onderon



Kun Wu
Chenbin Pan
Vedhas S Patkar

Yavin



Jiyang Wang
Pratik A Paranjape
Chirag Sachdev

Team composition
may change
slightly
during this week

What is this class about?

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



Security



Privacy

Class Format

- 1h presentation of reading materials
 - Research papers or book chapters
 - One team will **present** and lead the discussion
 - **Everyone** should be reading the material ahead!
 - One team will **take notes** and synthesize the discussion
- 20 min — Discussion and Q&A (but should arise during the presentation!)
- Deadlines:
 - 2 days prior to the class: presenting team submits slides (by 11:59pm)
 - 2 days after the last class of the module: notes team submits document (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: <http://web.ecs.syr.edu/~ffiorett/classes/spring20/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

Notes Format

- Notes should be produced in LaTeX
- Use the AAAI format (<https://aaai.org/Press/Author/authorguide.php>)
- At least 3 pages; No more than 8 pages
- Include all references and images

Notes Grading

- Reports will be evaluated based on:
 - Readability
 - Technical content
 - Accuracy of the information provided
- Reports should be written and are graded per team

Lateness policy

- Paper presentation
 - Deadline: Must be turned in by 11:59pm 2 days before the class
 - 10% per day late-penalty
 - 0 point if the presentation is not ready for the day in which the team is supposed to present
- Class Notes
 - Deadline: 2 days after the last class of the module: notes team submits document (by 11:59pm)
 - 10% per day late-penalty
 - Up to a max of 4 days

Grading Scheme

- 30 % paper presentation
- 20 % class notes
- 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/>

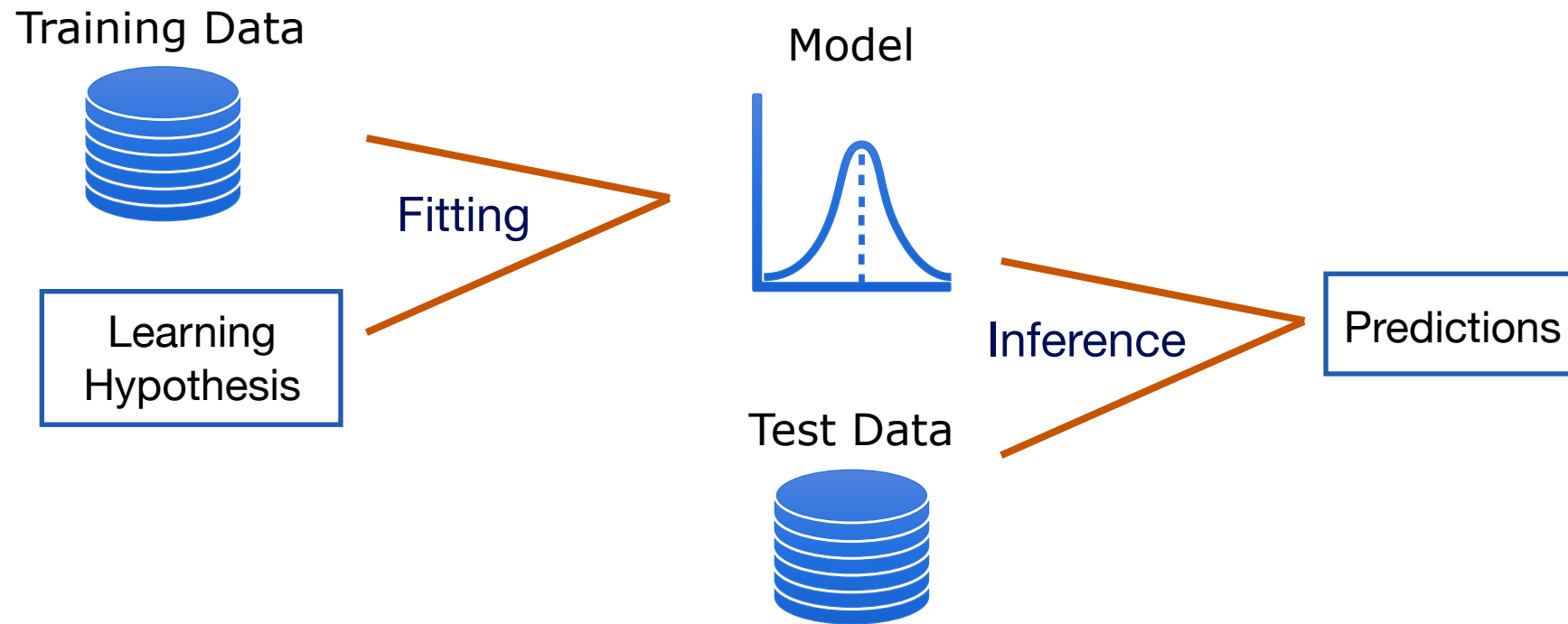
Instances of plagiarism, copying, and other disallowed behavior will constitute a violation of the code of student conduct. Students are responsible for reporting any violation of these rules by other students, and failure to do so constitute a violation of the code of student conduct.

Ethics

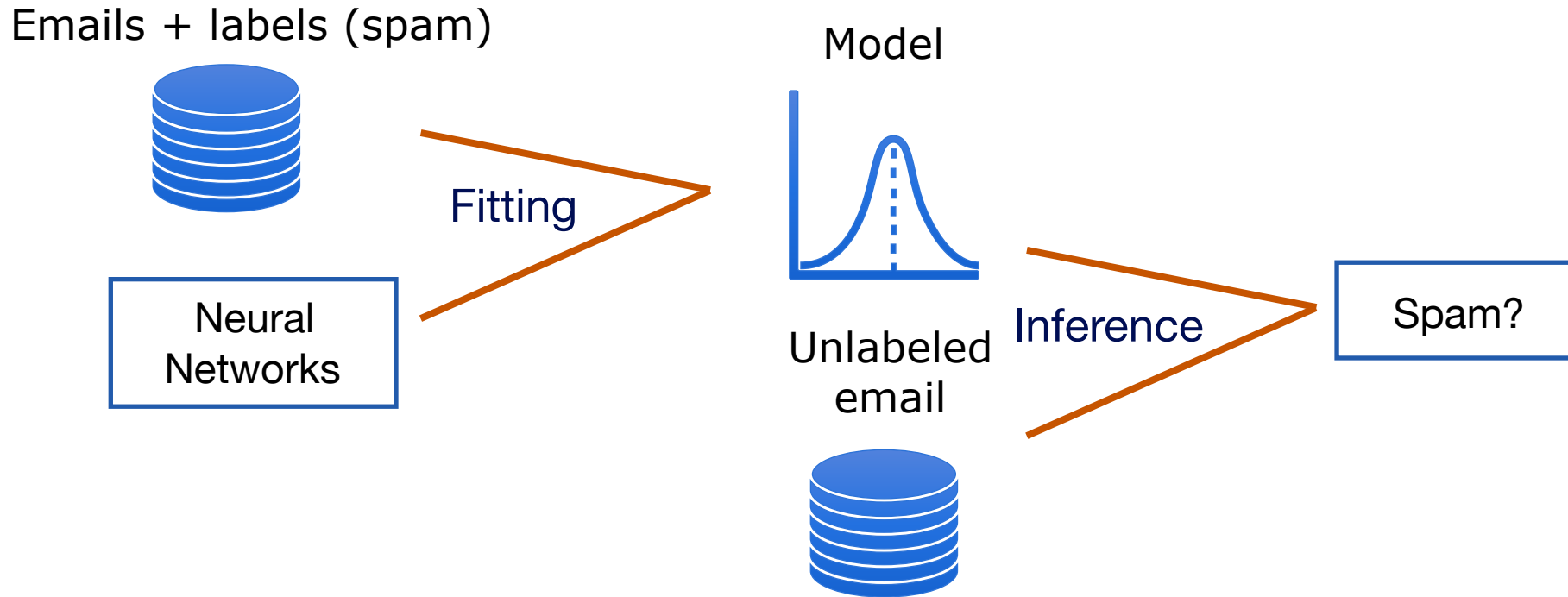
In this course, you will be learning about and exploring some vulnerabilities that could be exploited to compromise deployed systems. You are trusted to behave responsibly and ethically. You may not attack any system without permission of its owners, and may not use anything you learn in this class for 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

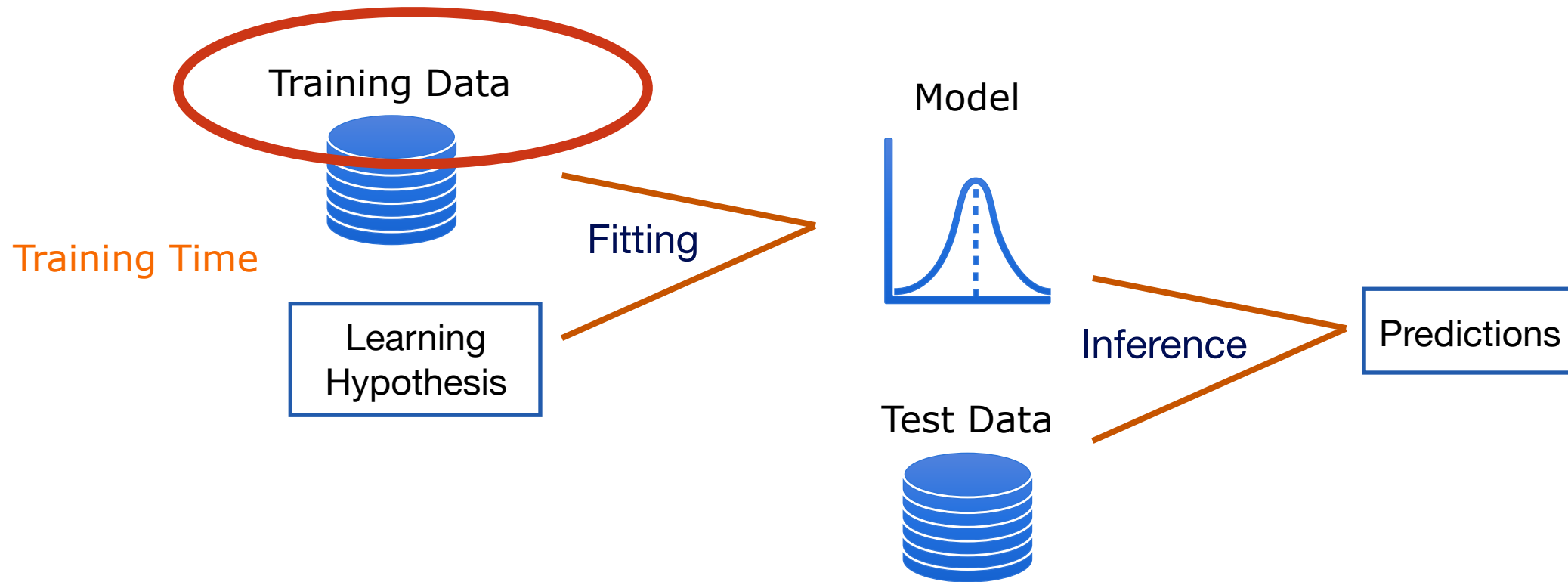


The ML Paradigm



The ML Paradigm in Adversarial Settings

Poisoning



Poisoning: An adversary injects 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

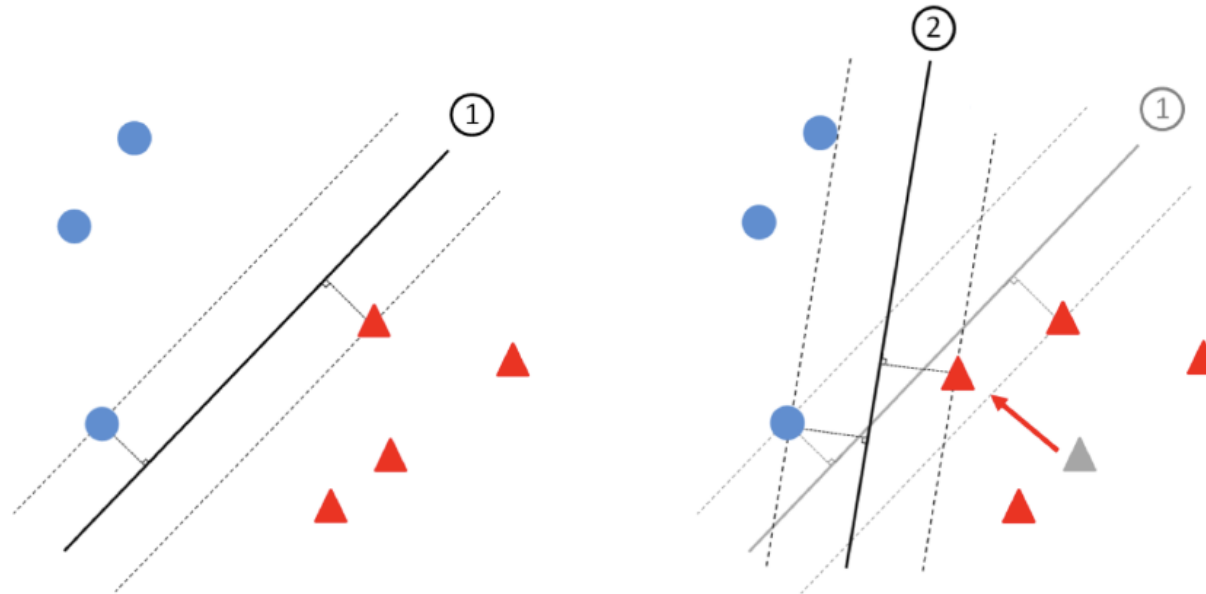
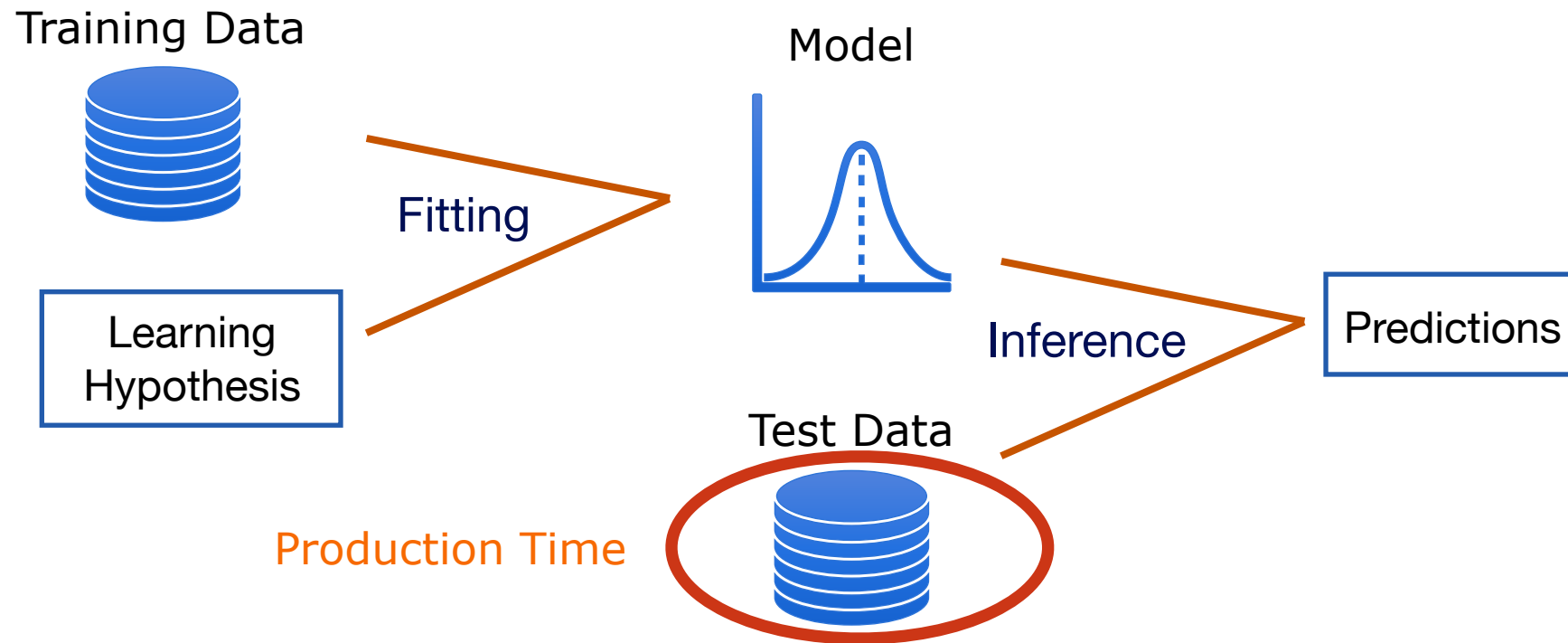


Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted if just one training sample is changed, even when that sample's class label does not change (right).

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

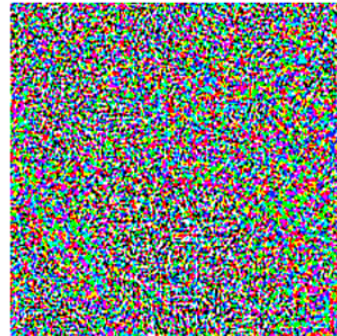
A typical example is to change some pixels in a picture before uploading, so that image recognition system fails to classify the result



“panda”

57.7% confidence

+ .007 ×



noise

=



“gibbon”

99.3% confidence

The ML Paradigm in Adversarial Settings

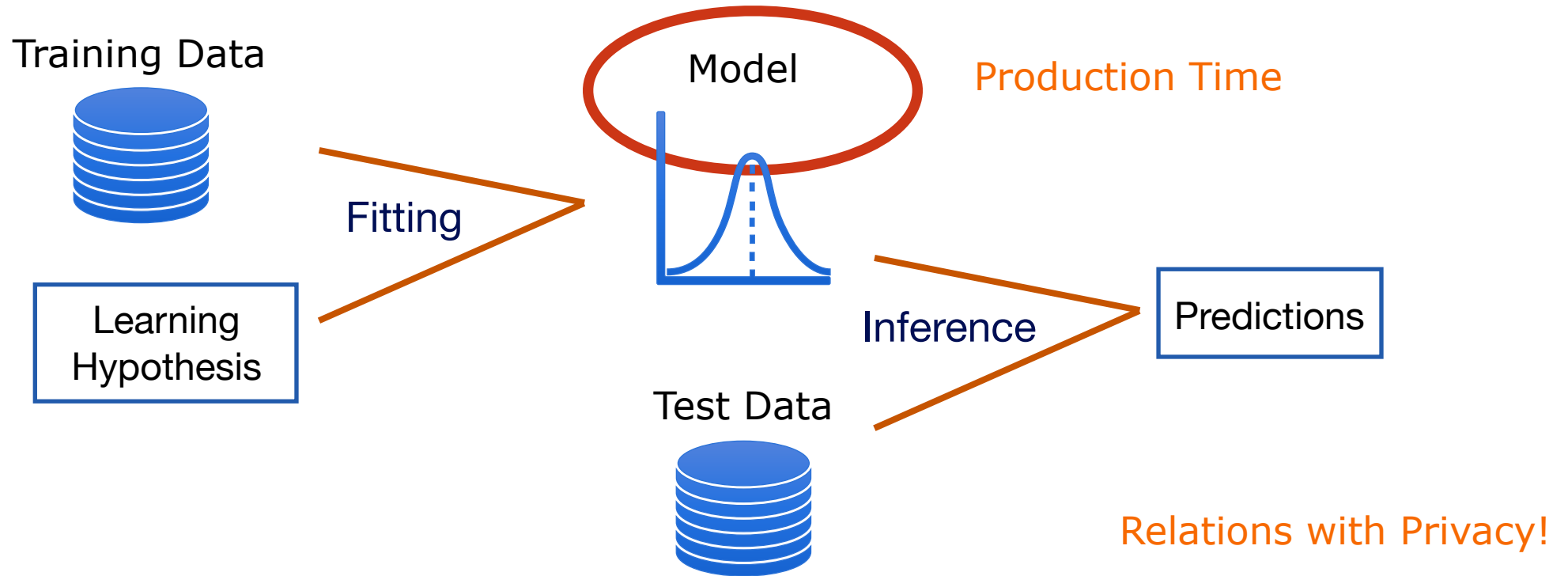
Evasion

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



The ML Paradigm in Adversarial Settings

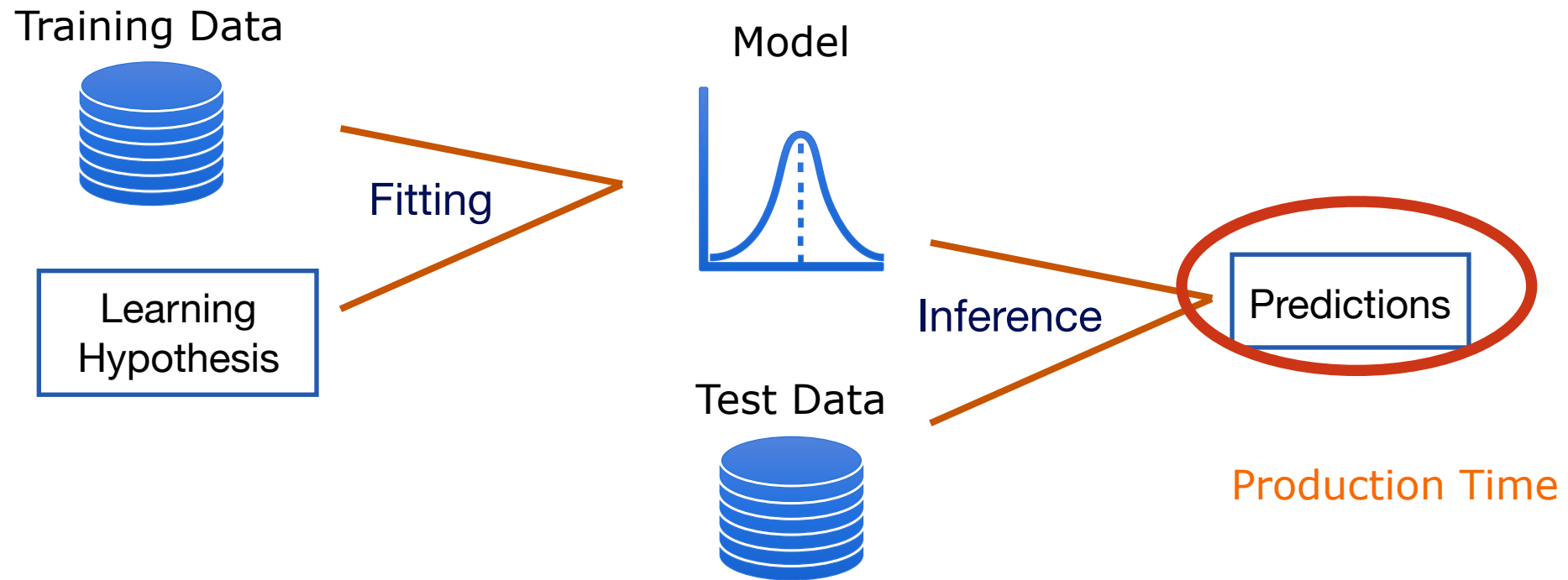
Member Inference



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

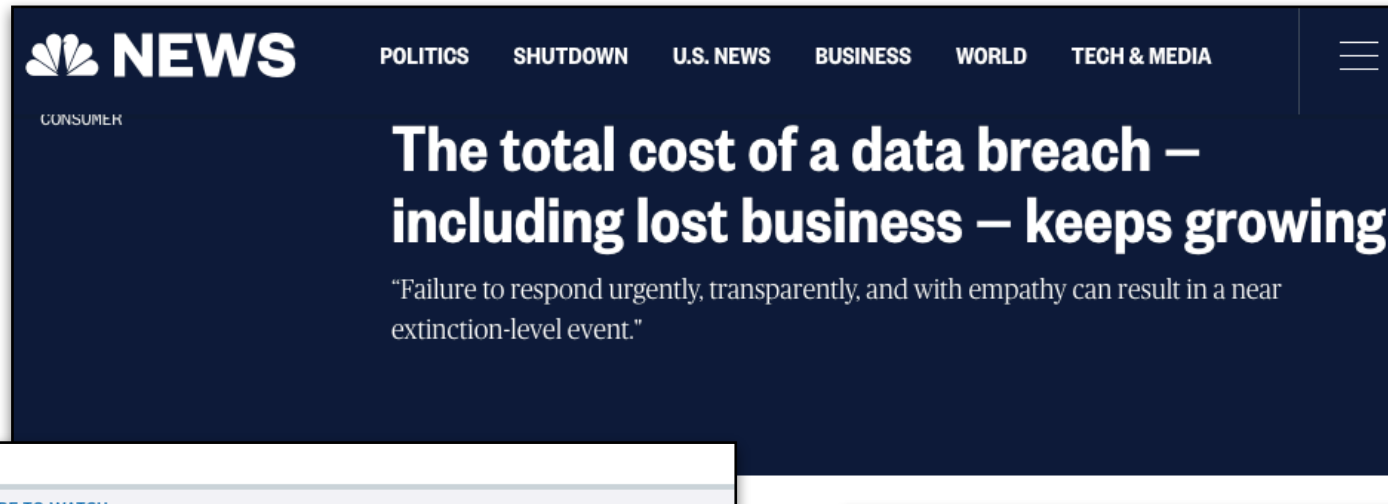


Fitness tracking app Strava gives away location of secret US army bases

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

The Cost of Privacy

\$3.86



NEWS POLITICS SHUTDOWN U.S. NEWS BUSINESS WORLD TECH & MEDIA

CONSUMER

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.”

ON THE MONEY

ON THE MONEY | VIDEO | WHERE TO WATCH

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

PM EDT



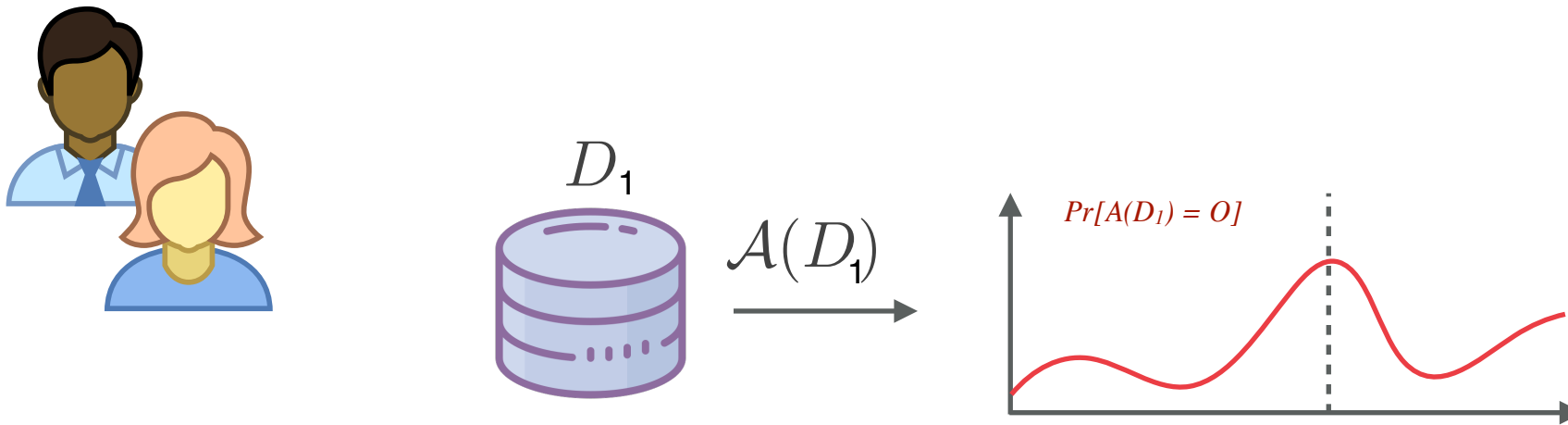
MENU **CNBC** MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV

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.

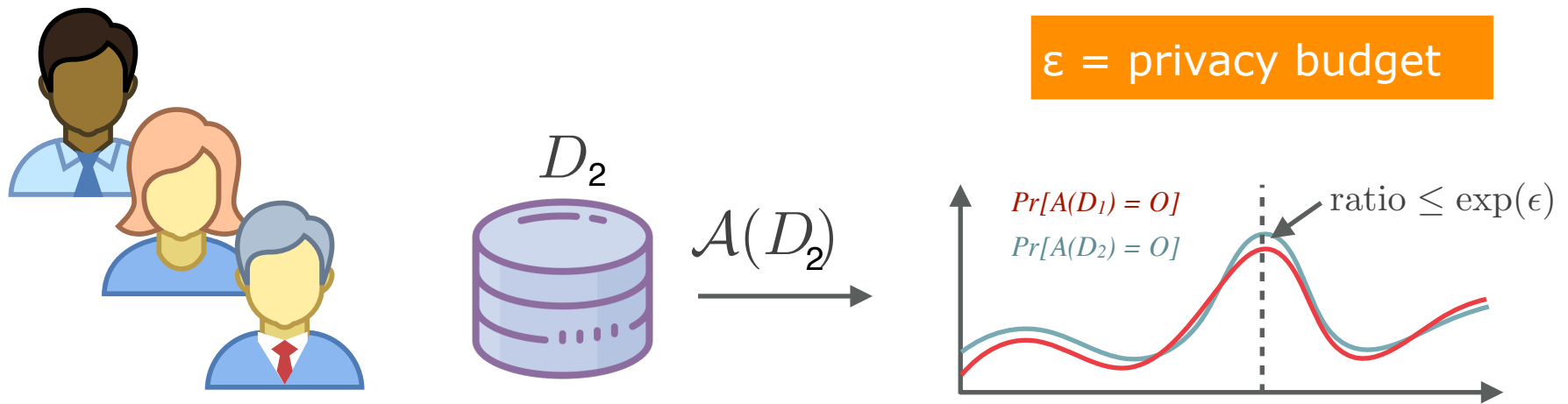
Differential Privacy

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



Differential Privacy

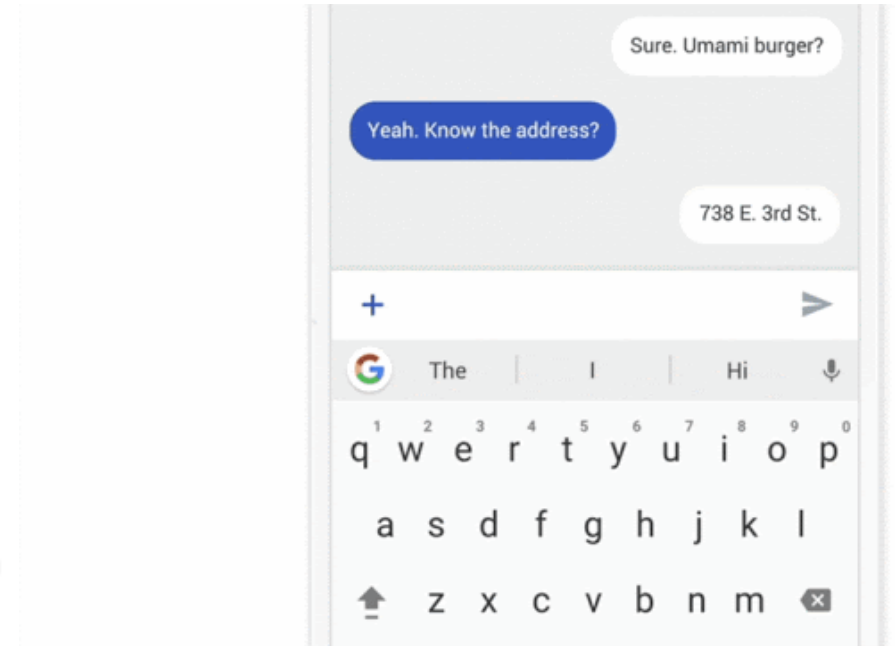
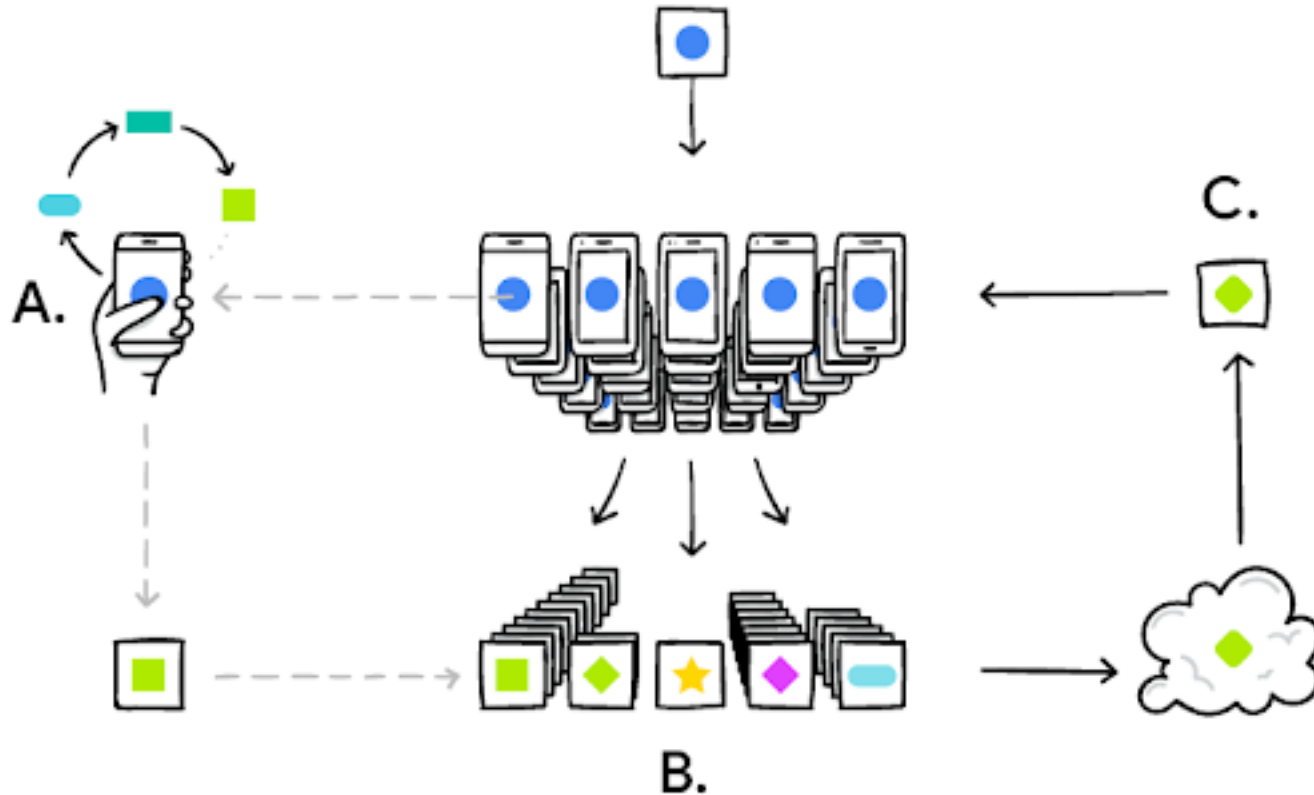
$$\frac{\Pr[\mathcal{A}(D_1) = O]}{\Pr[\mathcal{A}(D_2) = O]} \leq \exp(\epsilon)$$



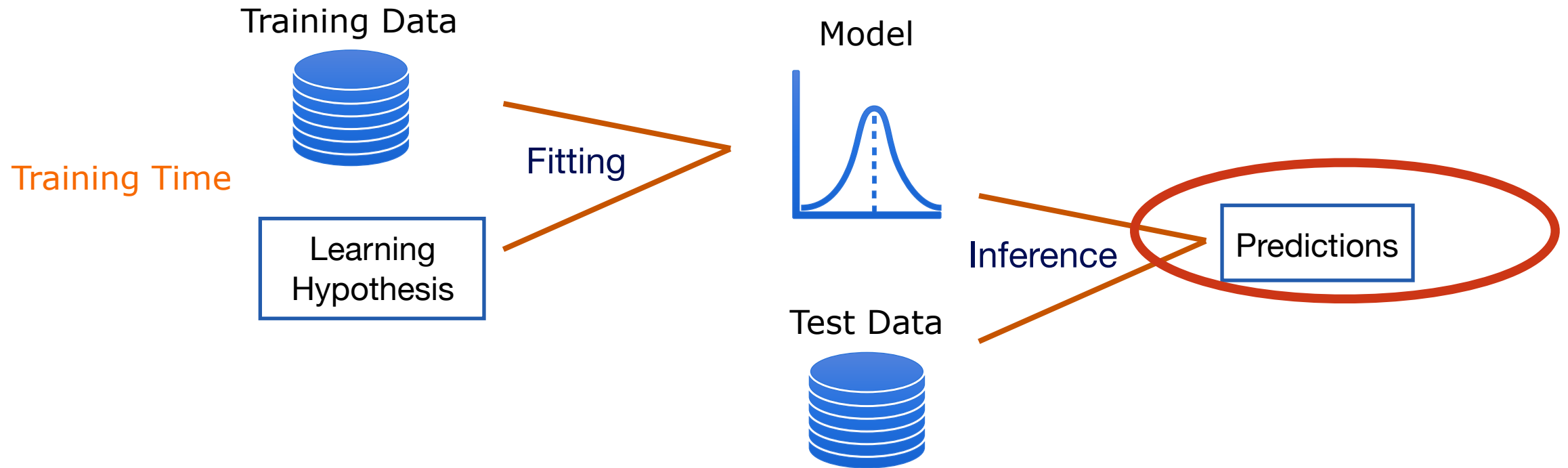
Controls the degree to which D_1 and D_2 can be distinguished.

Small ϵ gives more privacy (and worse utility)

Federated Learning



Fairness



Fairness: If training data is biased 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. Differential Privacy (DP)
5. DP and ML
6. DP model extensions
7. ML Robustness
8. Multiparty Computation
9. Federated Learning
10. Fairness and Bias

Research Project

- Take a look at the class topics and papers
- Identify two areas of interest
- Formulate a project proposal (1/2 page, due by Jan. 31)
 - Title
 - Team (optional) — at most 2 people
 - Problem
 - Methods
- Examples include:
 - Extended literature review on a topic
 - Implementations of attacks/defense mechanisms
 - Implementation of privacy-preserving approaches
- If you want to work with me, this is your chance to impress me!

Before Going

Write down your name + 2 things
you hope to learn in this class.

