Machine Learning and Security

Marc Juarez

imec-COSIC KU Leuven

COSIC

KU LEUVEN

fwo Research Foundation Flanders
Opening new horizons

Brussels School of Competition, 10th May 2019, Brussels

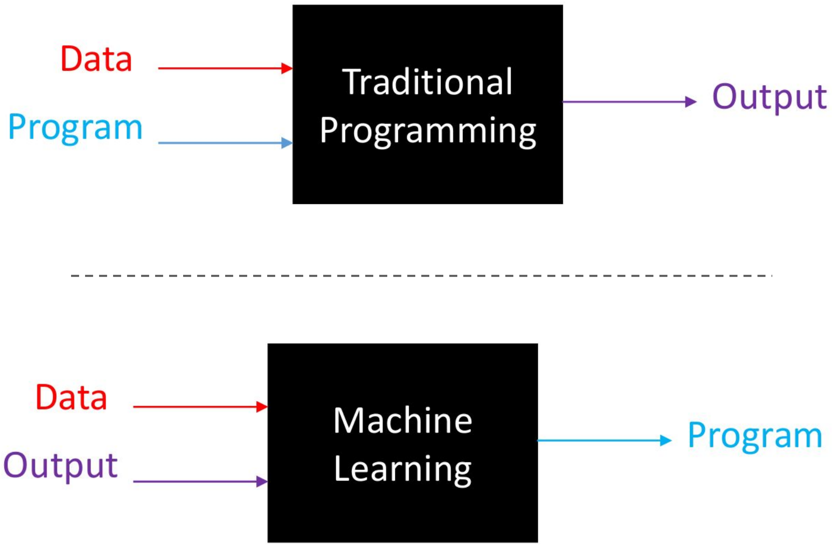
Outline

1. Introduction
2. Issues with deploying ML
3. Applications of ML to cybersecurity
4. Security of the ML system

Outline

1. **Introduction**
2. Issues with deploying ML
3. Applications of ML to cybersecurity
4. Security of the ML system

What is Machine Learning?



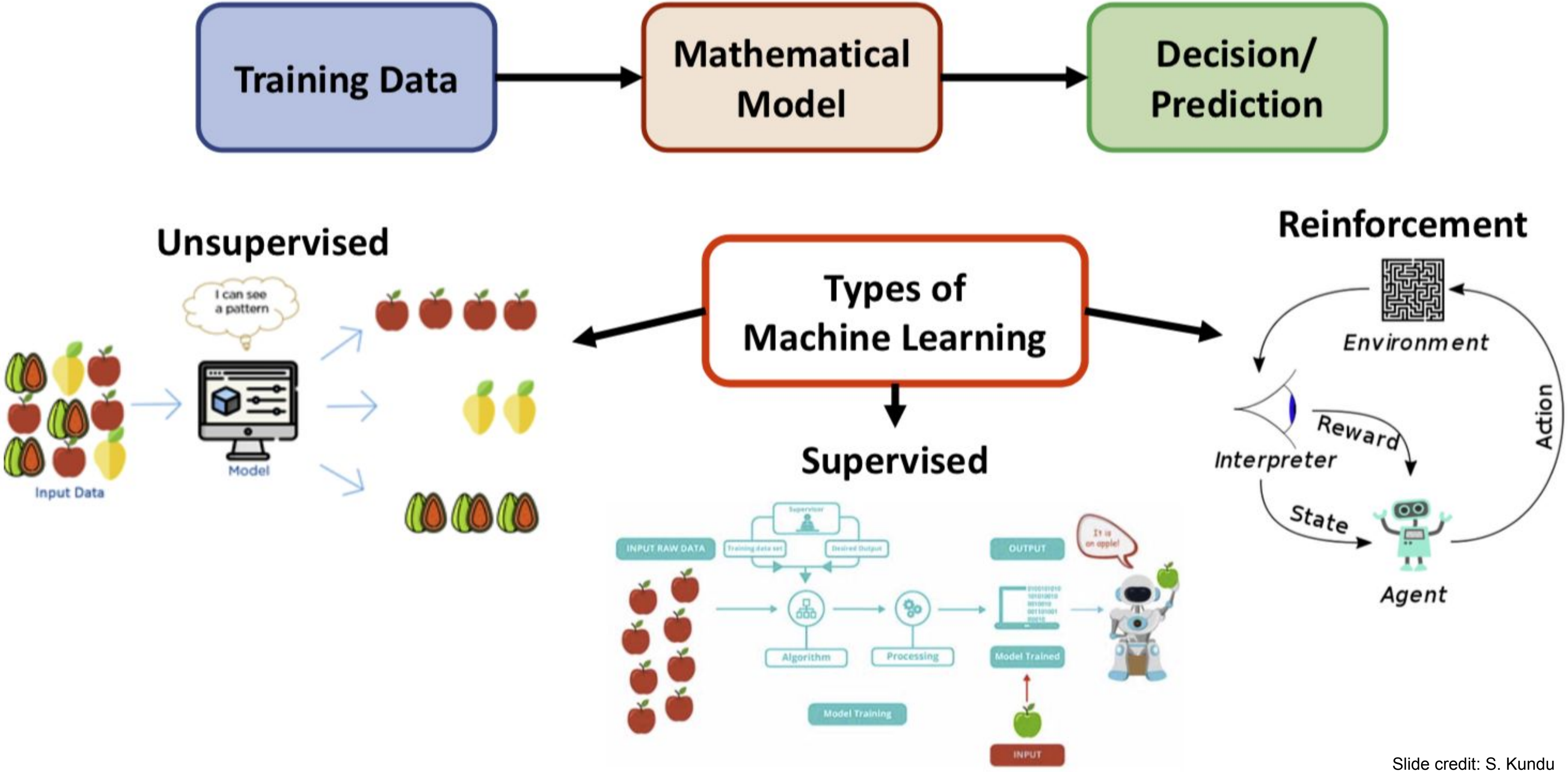
Definition by Tom Mitchell (1998):

“*Machine Learning is the study of* ***algorithms*** *that*

- *improve their performance P*
- *at a task T*
- *with* ***experience*** *E*

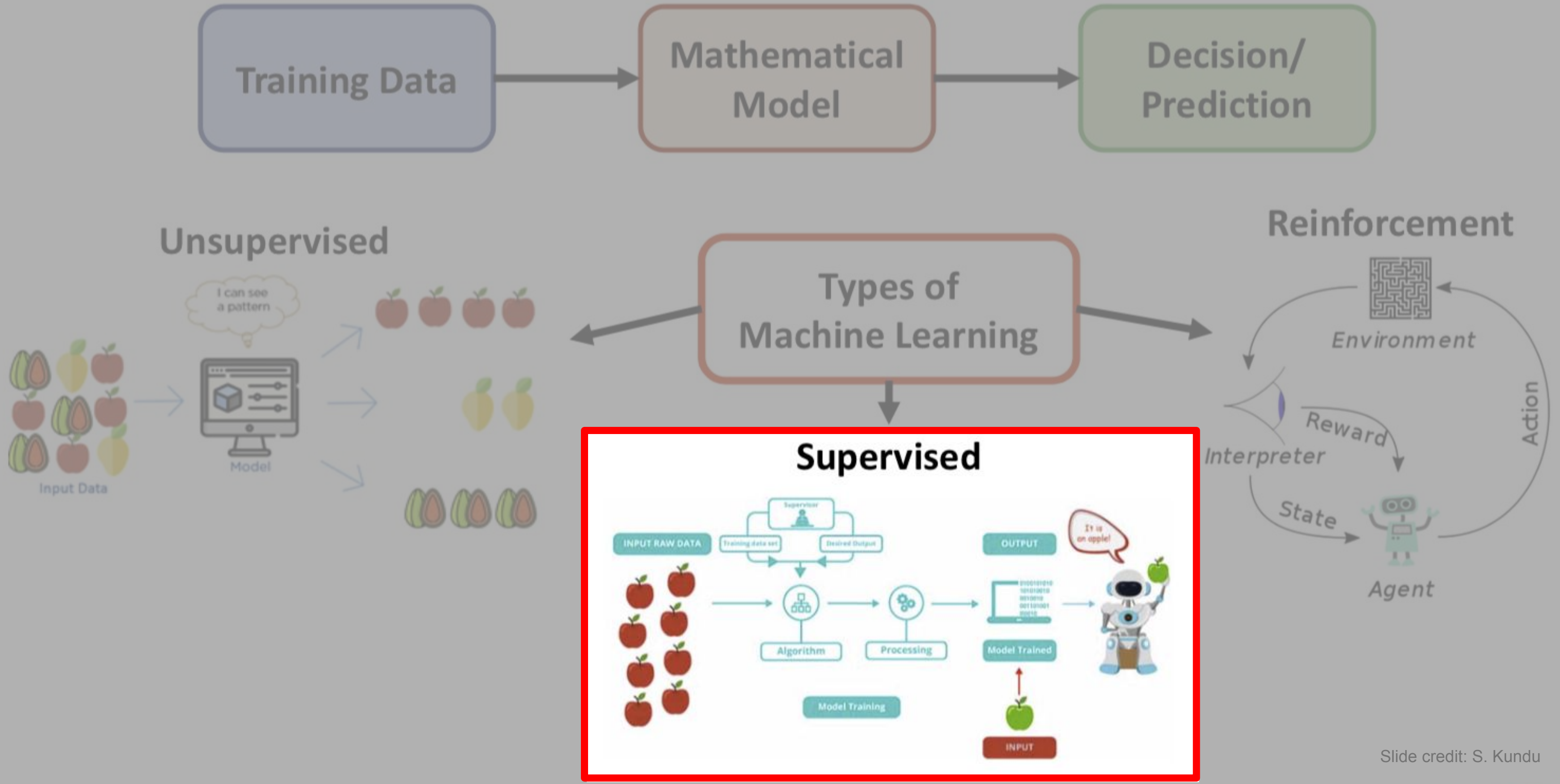
A well-defined learning task is given by <P, T, E>.”

Types of Machine Learning



Slide credit: S. Kundu

Types of Machine Learning

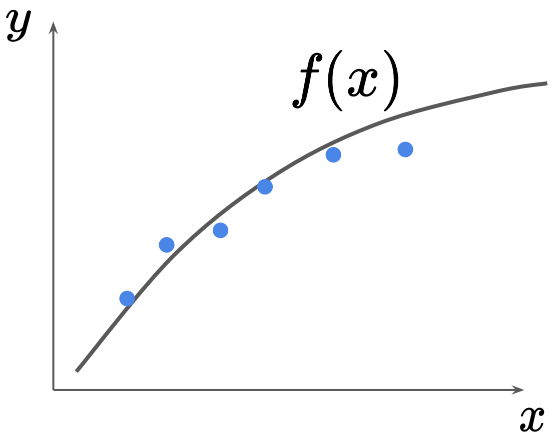


Slide credit: S. Kundu

Regression vs Classification

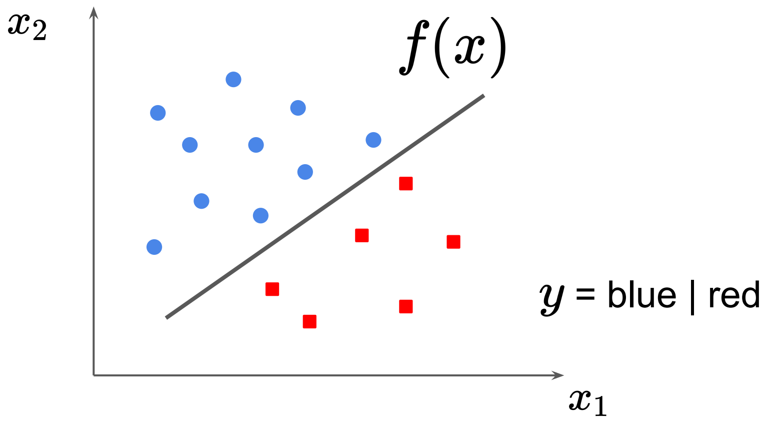
- Given x inputs and y outputs, find f such that $f(x) \sim y$

Regression



y is **continuous**

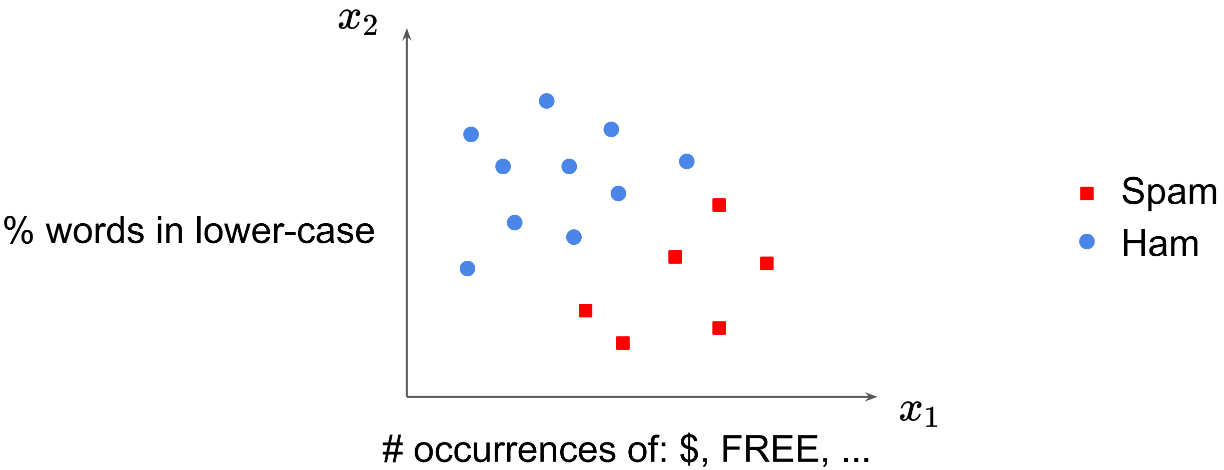
Classification



y is **discrete**

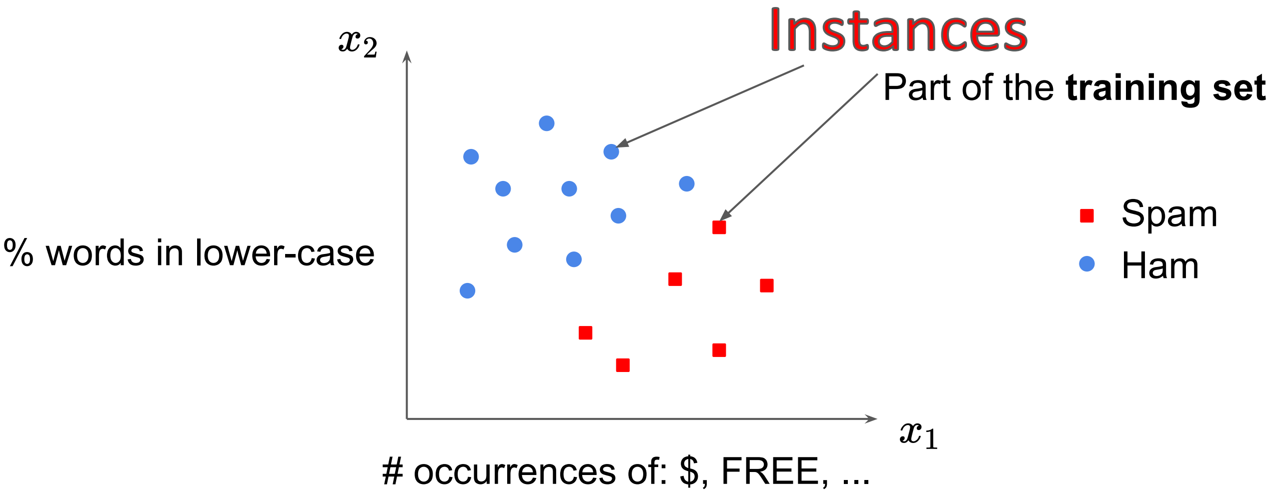
Classification: Spam example

- We have samples of email labeled as **spam** or **ham**:



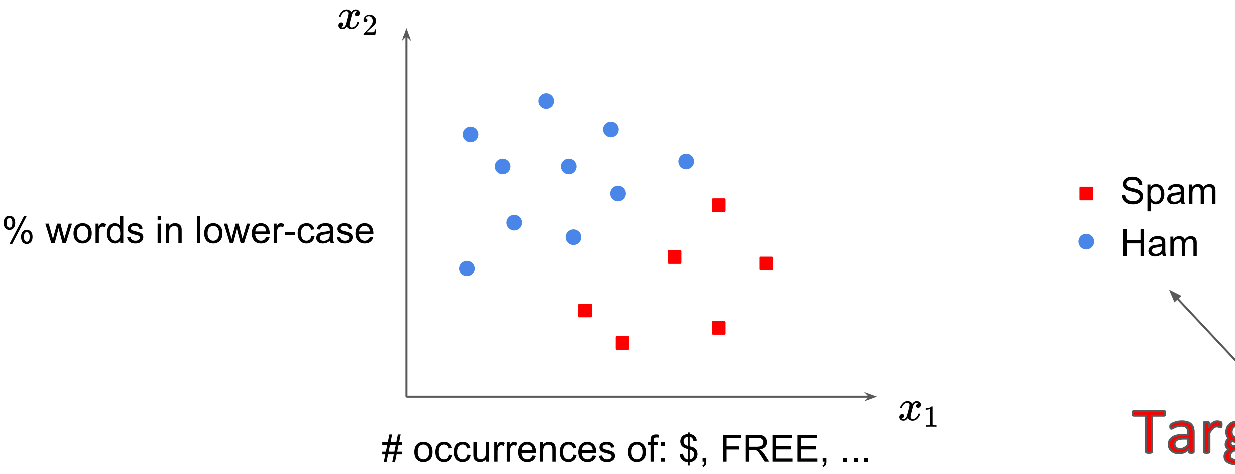
Classification: Spam example

- We have samples of email labeled as **spam** or **not spam** (ham):



Classification: Spam example

- We have samples of email labeled as **spam** or **not spam** (ham):

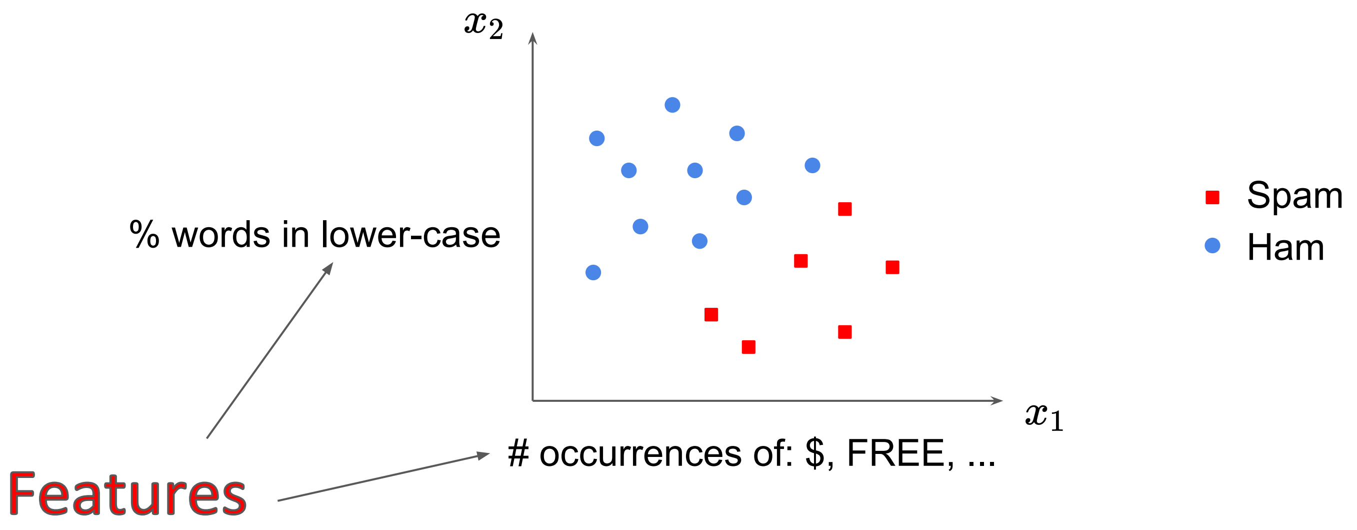


Target Variable

Values are **classes**

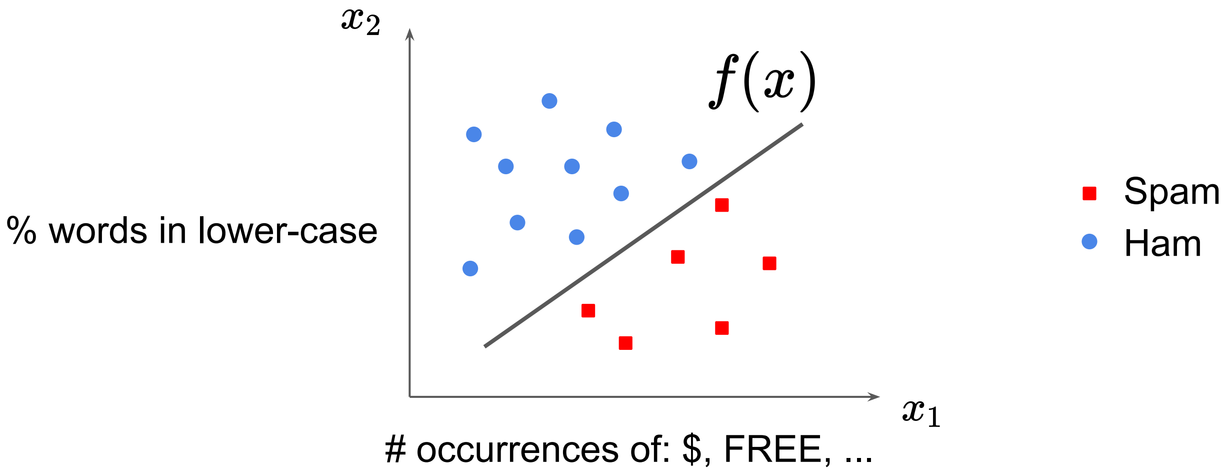
Classification: Spam example

- We have samples of email labeled as **spam** or **not spam** (ham):



Classification: Spam example

- Find f that separates sample space:



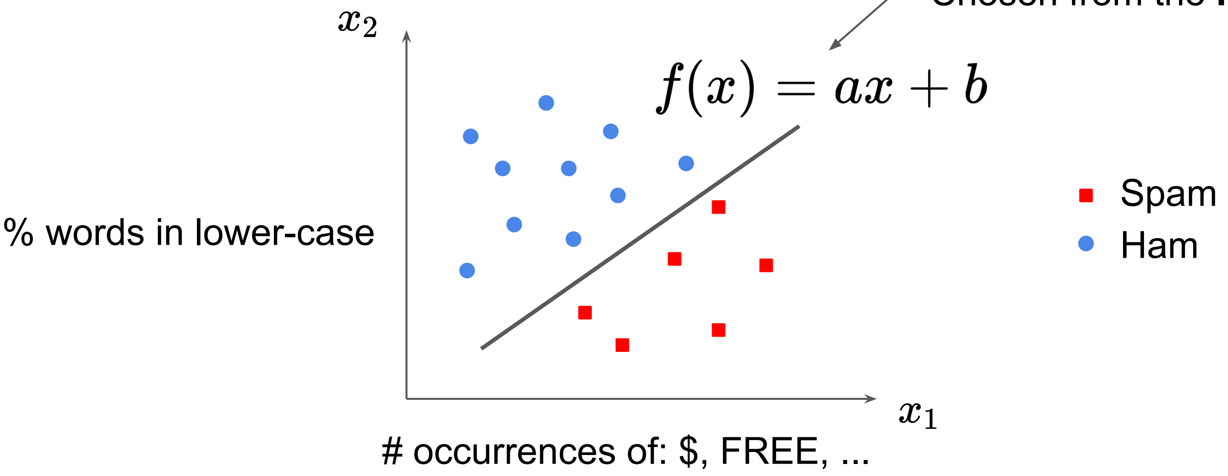
Classification: Spam example

- Find f that separates sample space:

Model

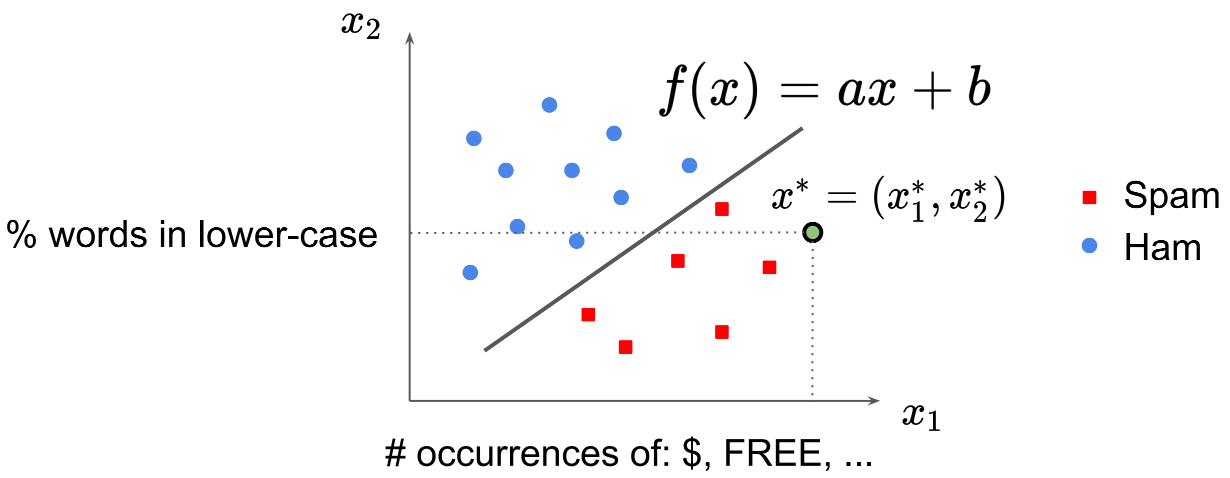
Chosen from the **Hypothesis Space**

$$f(x) = ax + b$$



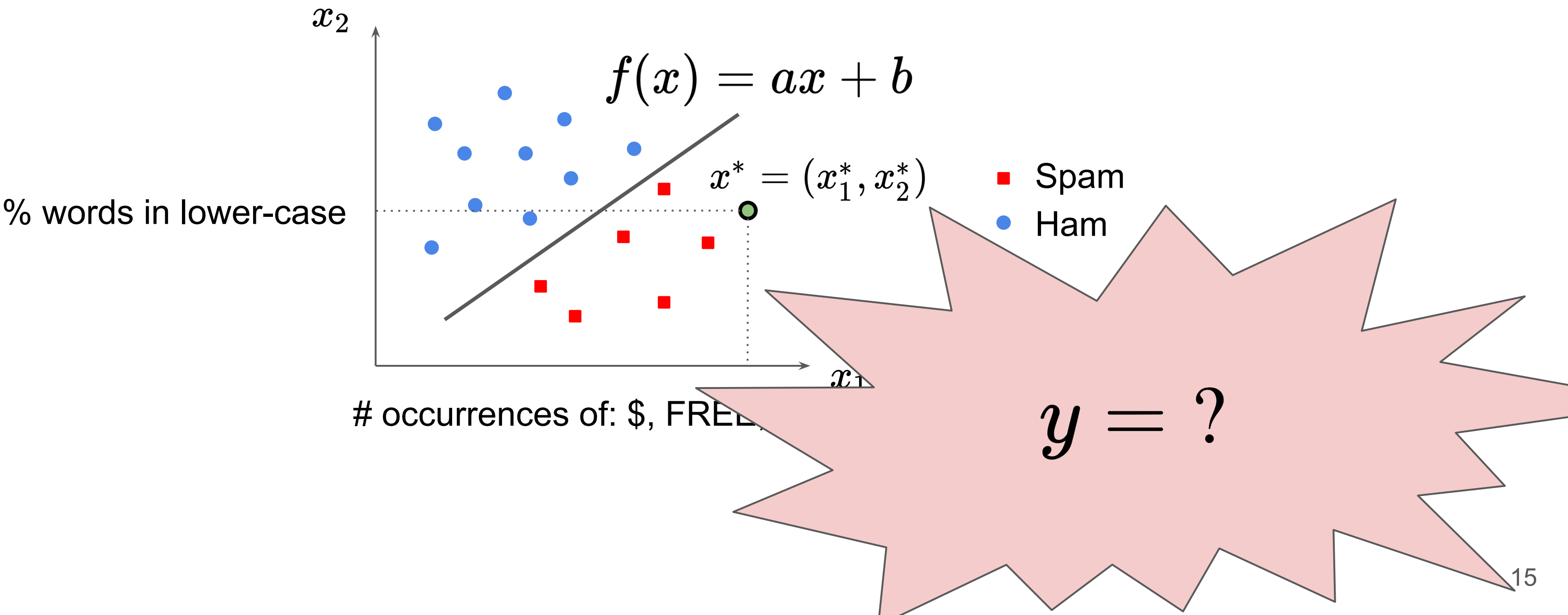
Classification: Spam example

- New email comes in: unknown label



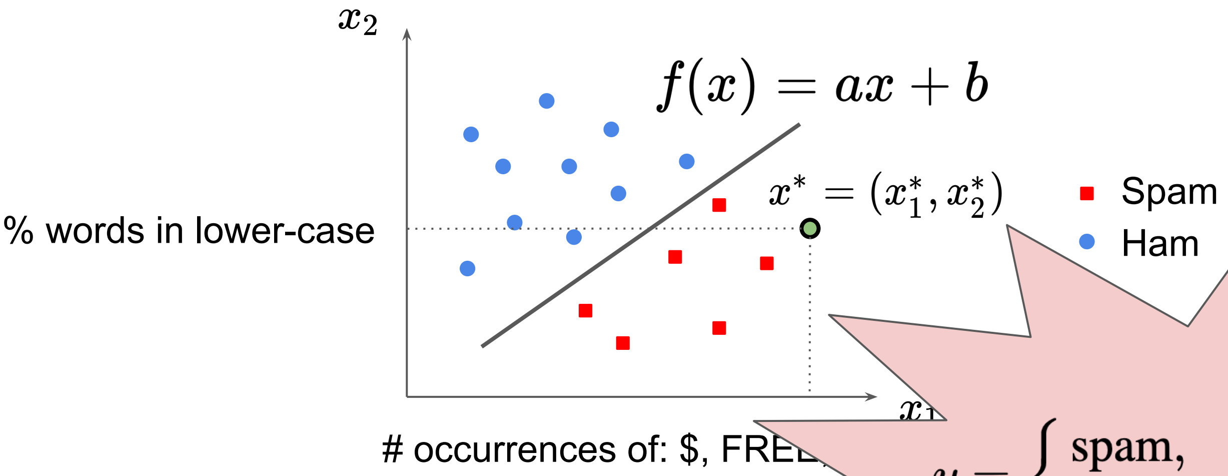
Classification: Spam example

- New email comes in: unknown label



Classification: Spam example

- New email comes in: unknown label → Use model to guess label

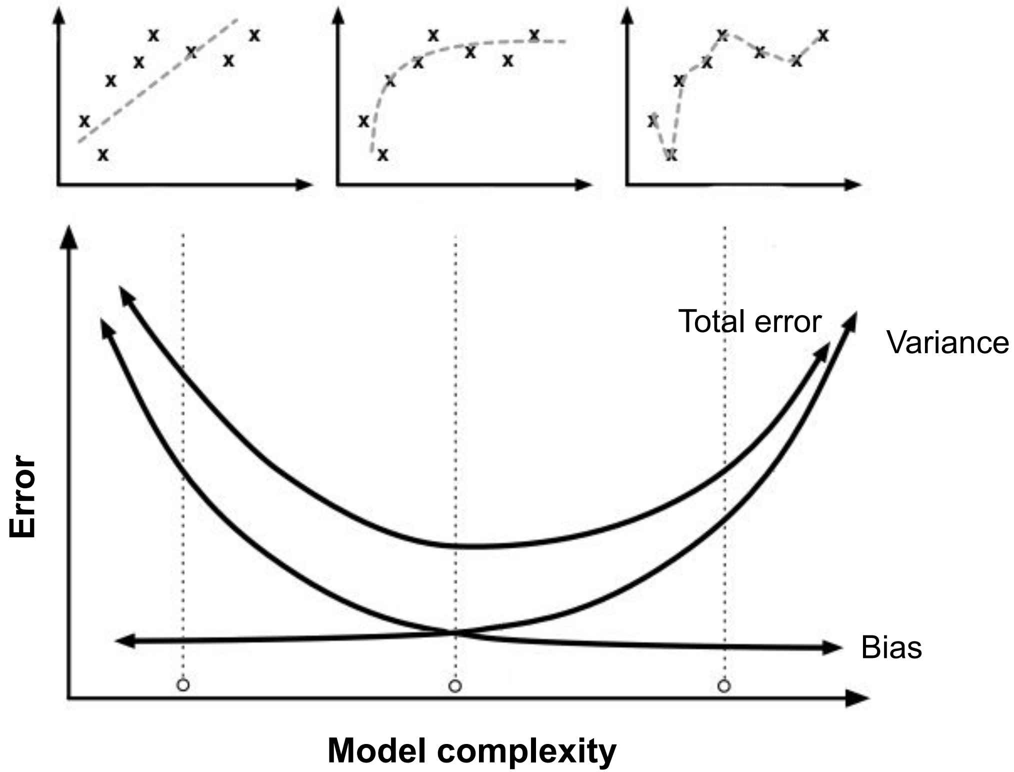


$$y = \begin{cases} \text{spam}, & \text{if } x_2 < f(x_1) \\ \text{ham}, & \text{otherwise} \end{cases}$$

ML algorithms

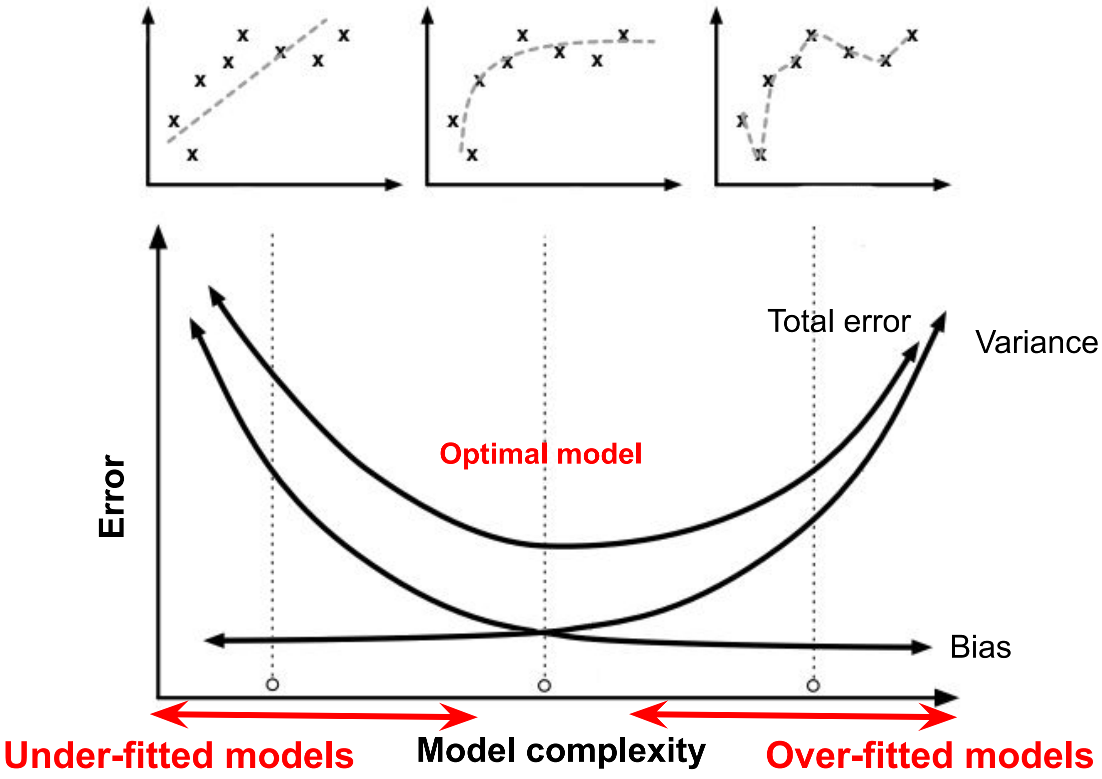
- ML algorithms are used to find model f
- Popular algorithms for classification:
    - Naive Bayes
    - Support Vector Machines
    - ID3 (Decision Trees) → Random Forests
    - Neural networks (aka Deep Learning)
    - …
- What is a "good" model?What makes a model "good"?

The Bias-Variance trade-off



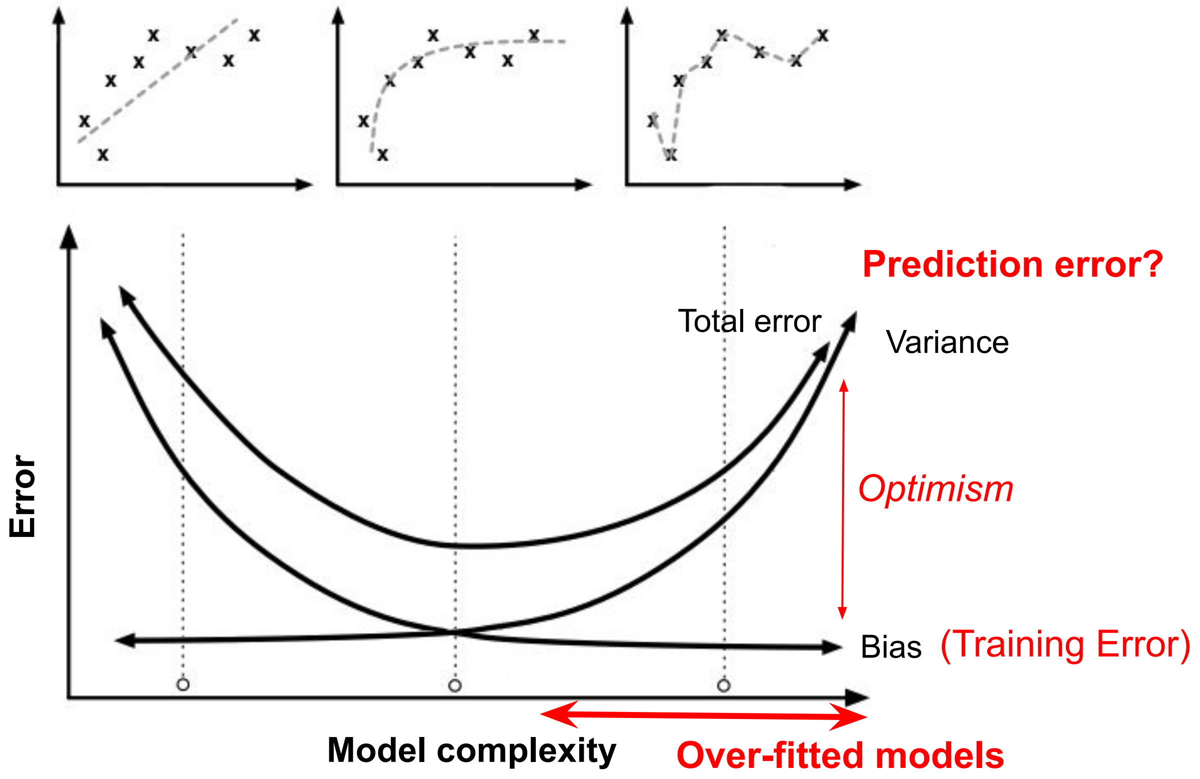
Based on a figure by: S. Fortmann-Roe

The Bias-Variance trade-off



Based on a figure by: S. Fortmann-Roe

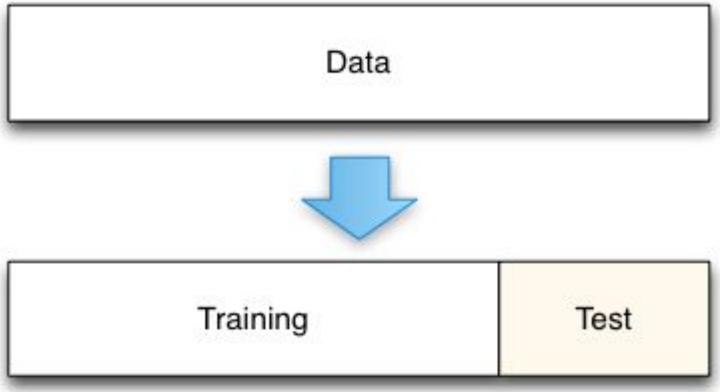
The Bias-Variance trade-off



Based on a figure by: S. Fortmann-Roe

Measuring overfitting

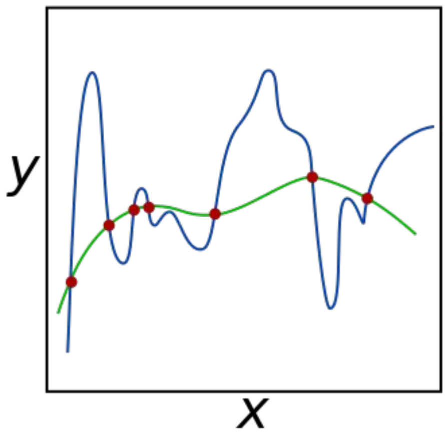
- Idea: *hold out* labeled sample for testing
- Non-parametric technique
- Accurate if enough data
    - Small dataset → Cross-validation



Mitigating overfitting

- Regularization: additional assumptions that prevent overfitting without increasing bias

Example: add smoothing factor to f(x)



Metrics for the classifier's error

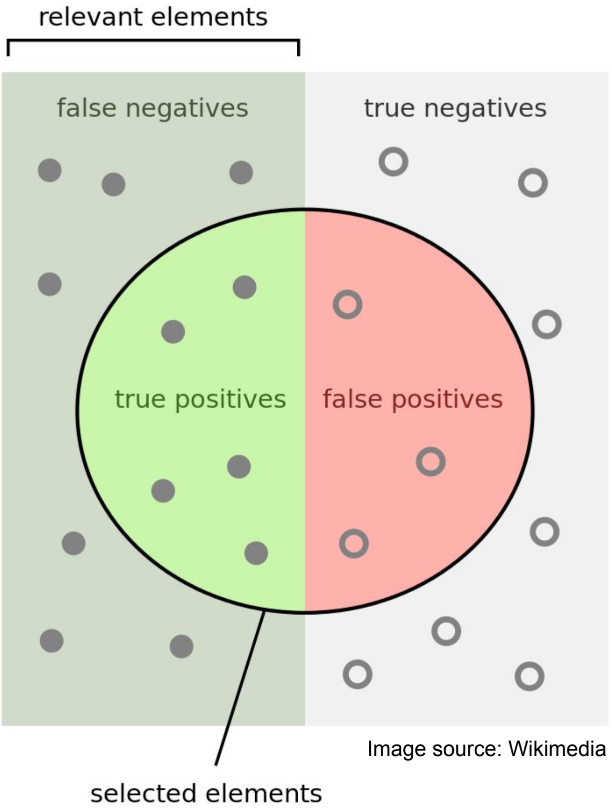


Image source: Wikimedia

- Positive class:
  - True Positives (TP)
  - False Negatives (FN)
- Negative class:
  - True Negatives (TN)
  - False Positives (FP)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TOTAL}}$$

Metrics for the classifier's error

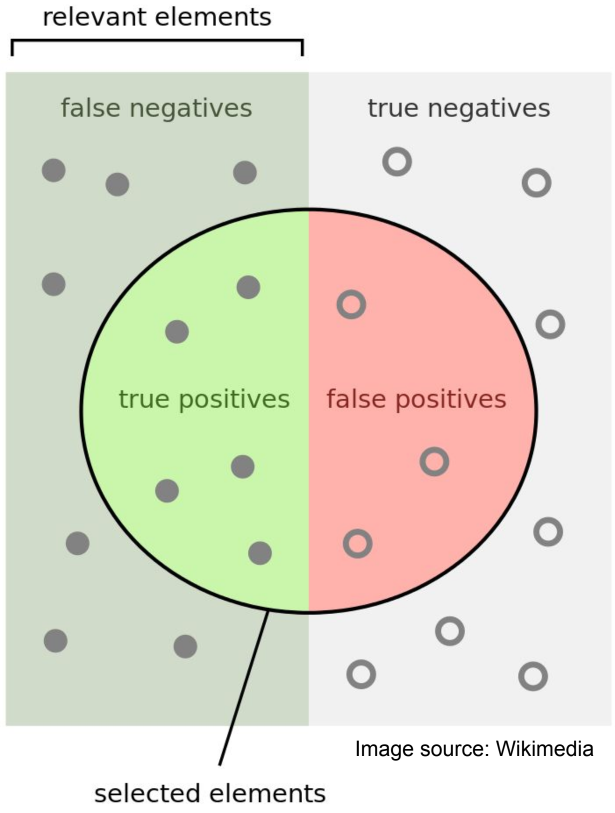
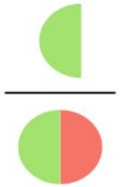


Image source: Wikimedia

Recall =
(aka TPR)

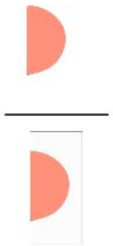
How many *relevant* items have been selected?

Precision =



How many of the selected items are *relevant*?

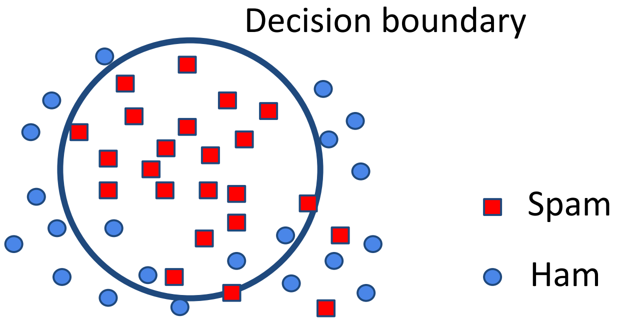
False Positive Rate =
(FPR)

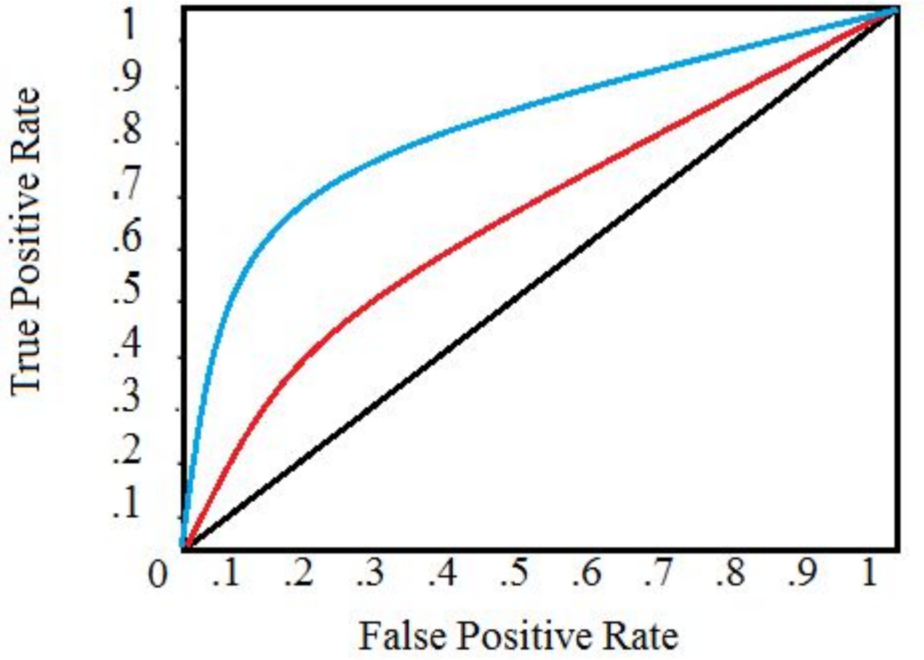


How many *irrelevant* items have been (incorrectly) selected?

ROC curve

- Trade-off between TPR and FPR
- Parametrized decision boundary
- Tuned for application:
  - Minimize FPR: Spam
  - Minimize FNR: Disease test





Outline

1. Introduction
2. **Issues with deploying ML**
3. Applications of ML to cybersecurity
4. Security of the ML system

The Base Rate Fallacy (aka Prosecutor's Fallacy)

- Breathalyzer test:
    - **0.88** identifies truly drunk drivers (True Positive Rate)
    - **0.05** sober drivers as drunk (False Positive Rate)

- Alice gives positive in the test
    - What is the probability that she is indeed drunk?
    - Is it 0.95? Is it 0.88? Something in between?

The Base Rate Fallacy (aka Prosecutor's Fallacy)

- Breathalyzer test:
  - **0.88** identifies truly drunk drivers (True Positive Rate)
  - **0.05** sober drivers as drunk (False Positive Rate)

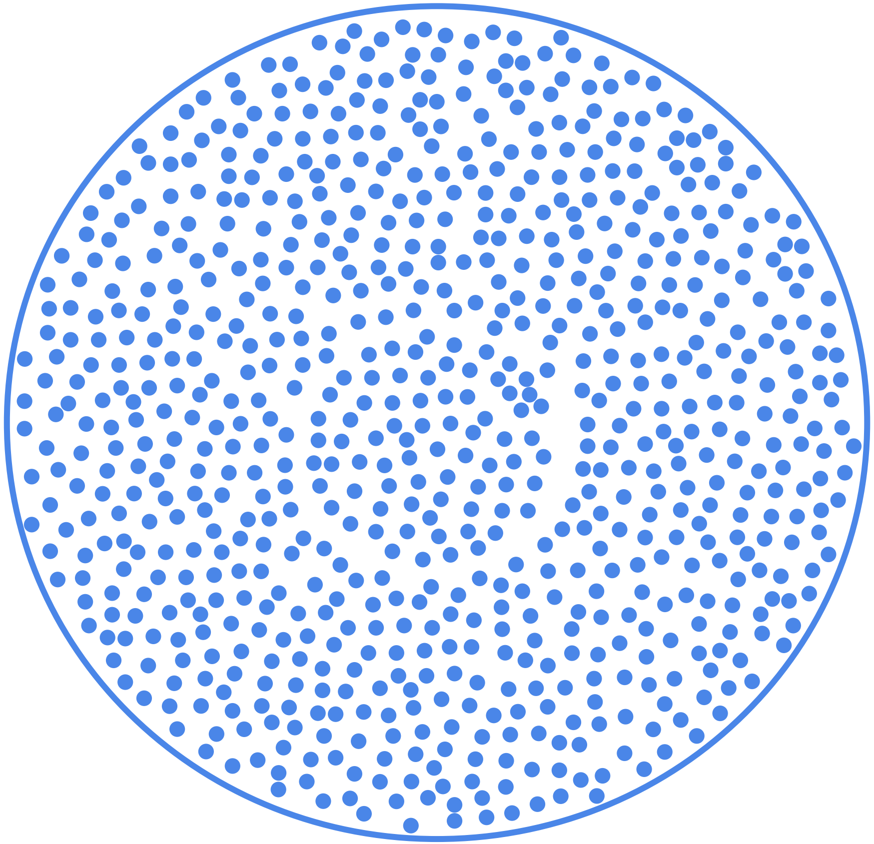
- Alice gives positive in the test
  - What is the proba… lity t… is i… ndeed drunk?
  - Is it 0.95? Is it … etween?

Only 0.1!

The Base Rate Fallacy (aka Prosecutor's Fallacy)

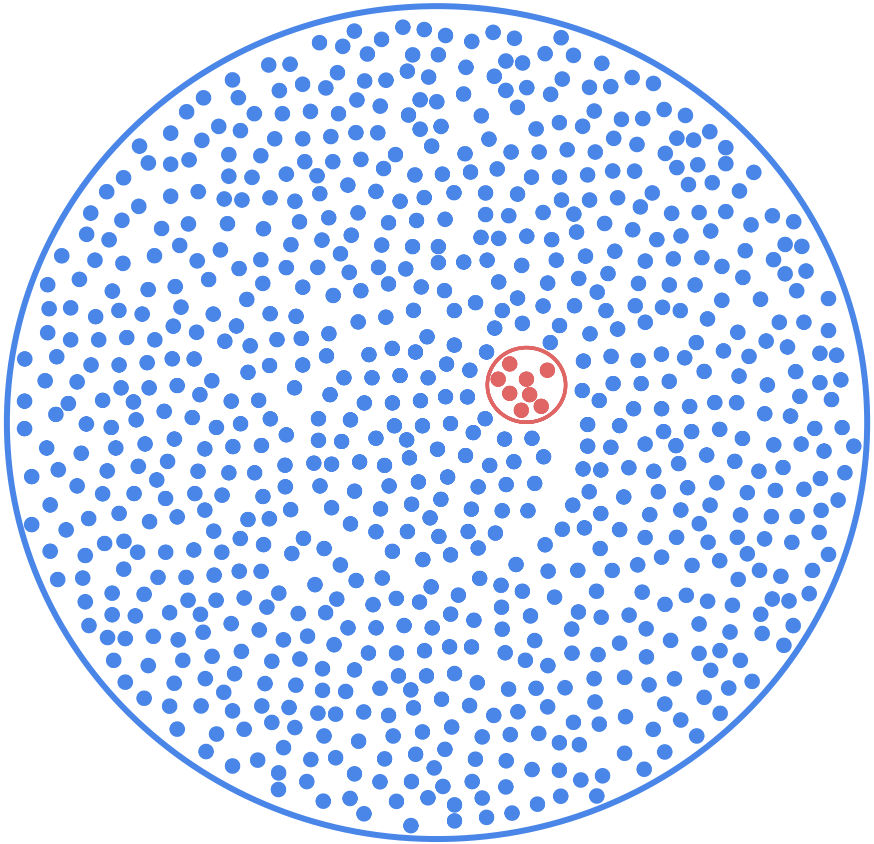
- Circumference represents the world of drivers.

- Each dot represents a driver.



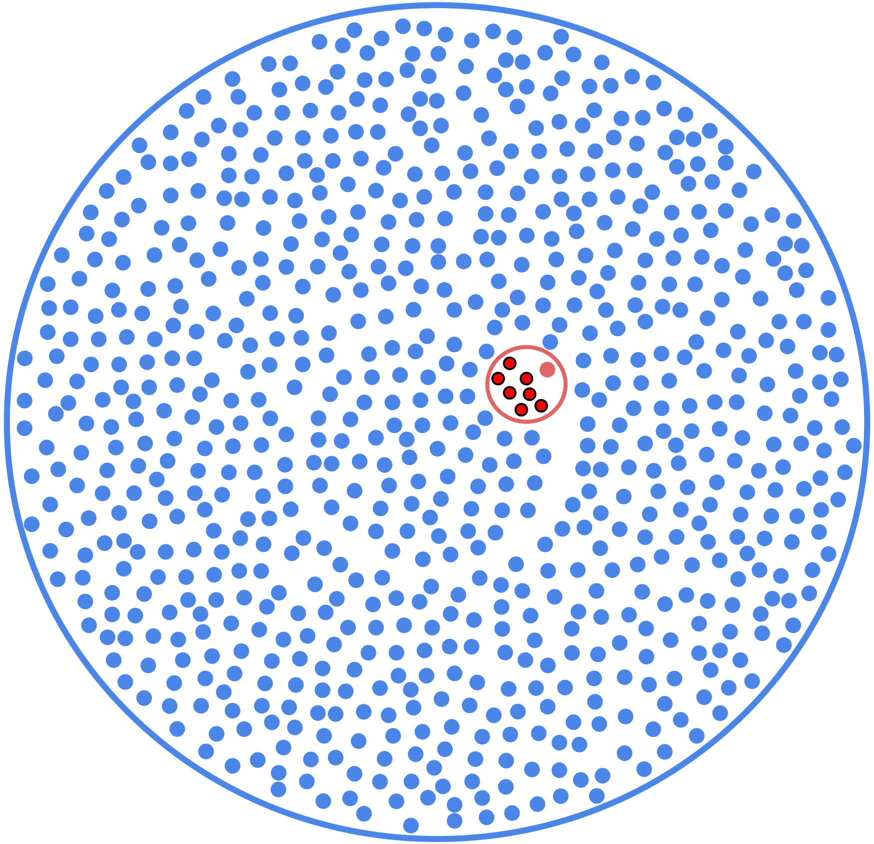
The Base Rate Fallacy (aka Prosecutor's Fallacy)

- 1% of drivers are driving drunk (**base rate or prior**).



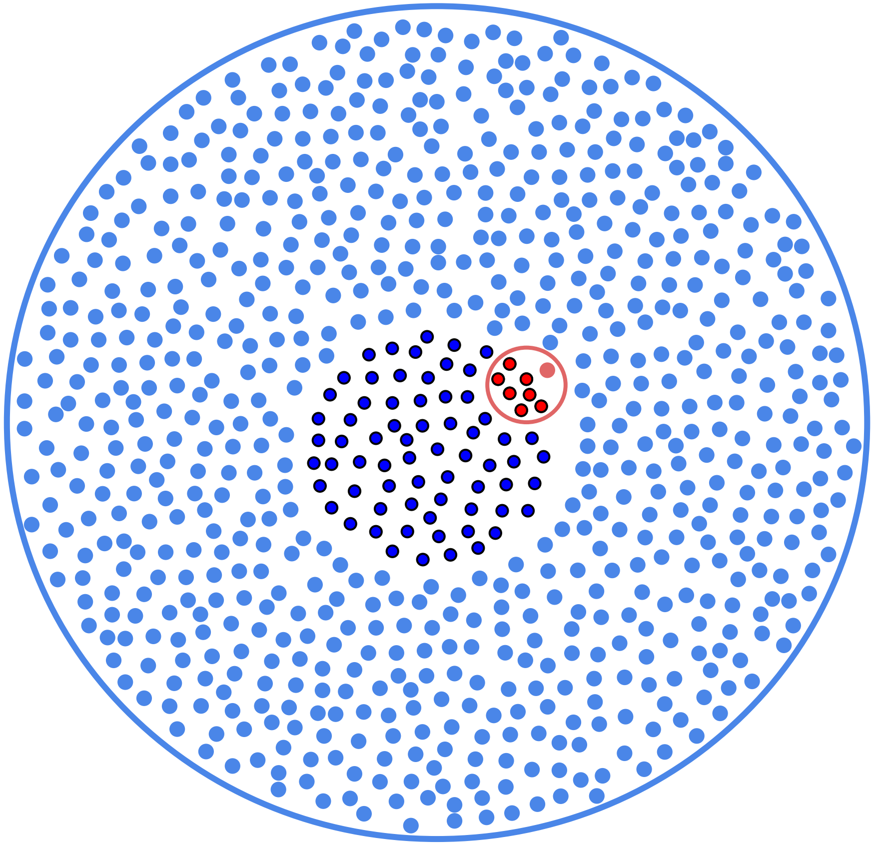
The Base Rate Fallacy (aka Prosecutor's Fallacy)

- From drunk people 88% are identified as drunk by the test



The Base Rate Fallacy (aka Prosecutor's Fallacy)

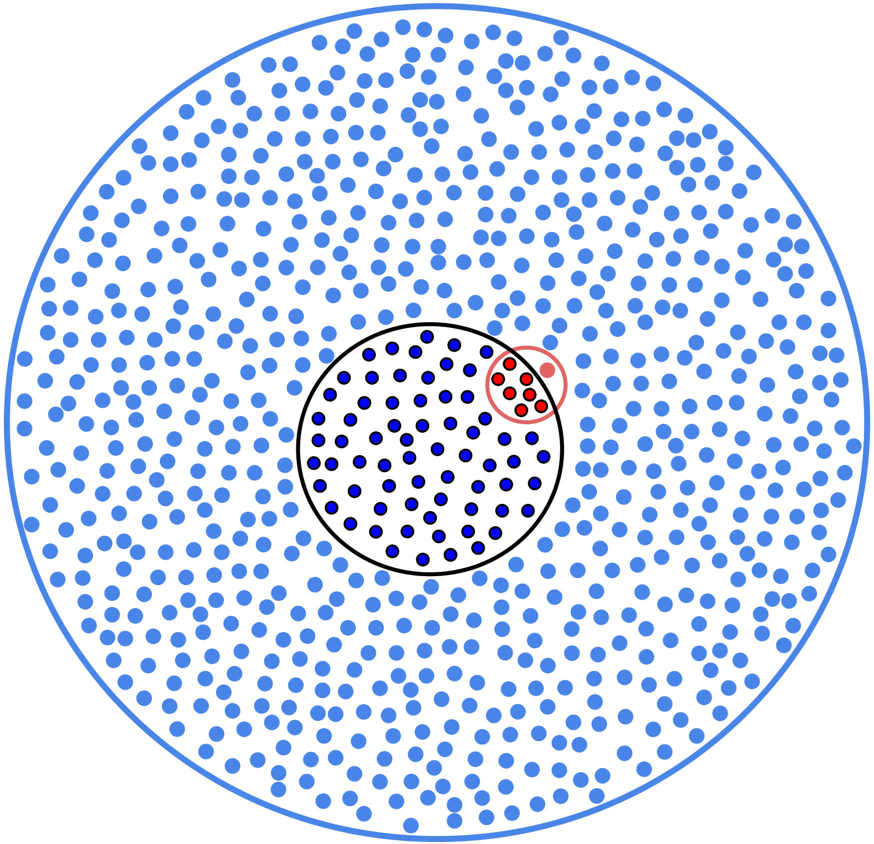
- From the sober people, 5% are erroneously identified as drunk



The Base Rate Fallacy (aka Prosecutor's Fallacy)

- Alice must be within the black circumference
- Ratio of red dots within the black circumference:

Precision = 7/70 = **0.1**



Other examples

- Can you think of other examples where the base rate fallacy comes into play?

Other examples

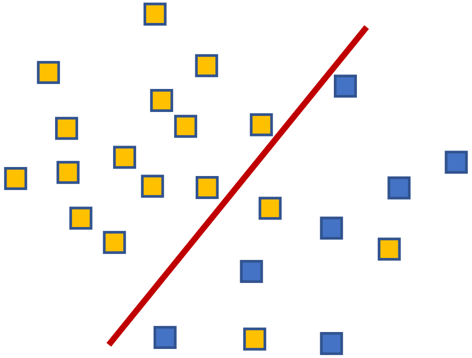
- Can you think of other examples where the base rate fallacy comes into play?

  Cases in which the positive class is very unlikely:

  - Test a rare disease
  - Detect a system intrusion
  - Anticipate a terrorist attack

Distributional Shift

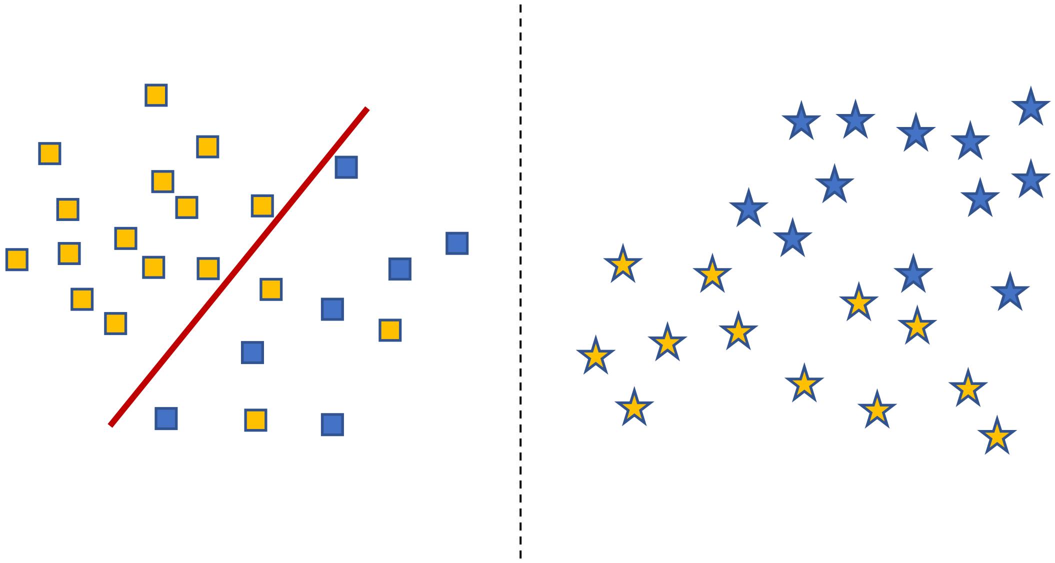
Occurs when a classifier is trained in one area and deployed in another.



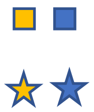
- Example: Family migration prediction
- Training/Test in Syria
- Yellow = Families that migrate
- Blue = Families that do not migrate

Distributional Shift

Occurs when a classifier is trained in one area and deployed in another.



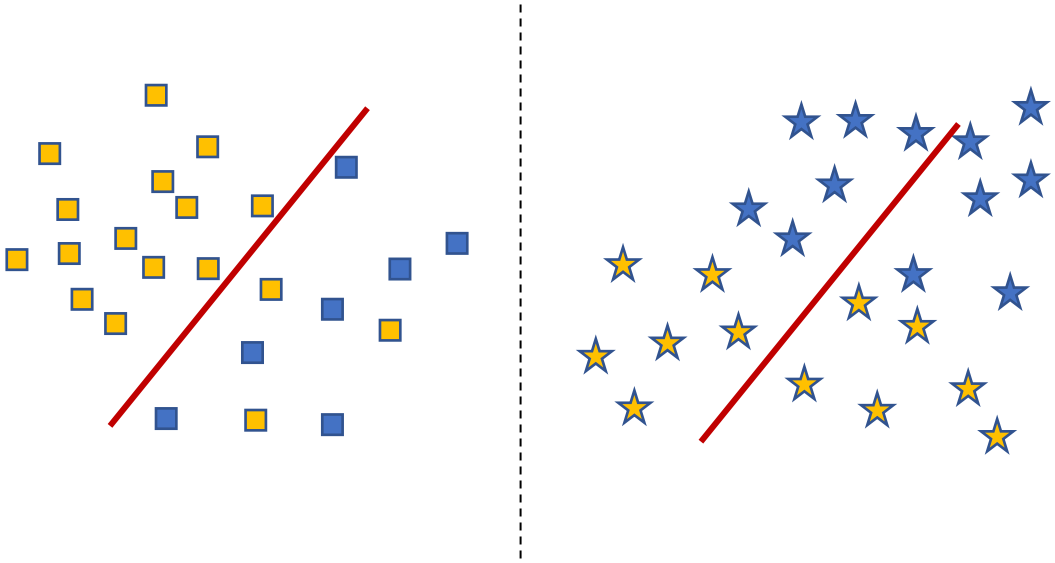
- Example: Family migration prediction
- Yellow = Families that migrate
- Blue = Families that do not migrate



Slide credit: C. Troncoso

Distributional Shift

Occurs when a classifier is trained in one area and deployed in another.



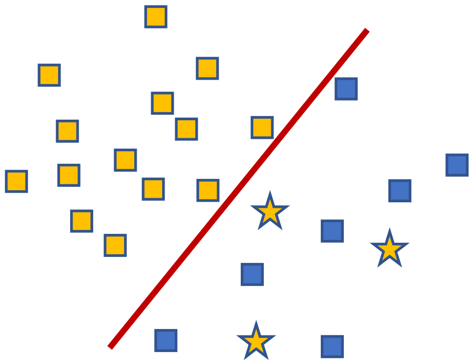
- Example: Family migration prediction
- Yellow = Families that migrate
- Blue = Families that do not migrate

Training/Test in Syria

Training/Test in Myanmar

Distribution of Errors

Occurs when most mistakes of the classifier are concentrated in a **subpopulation/group**

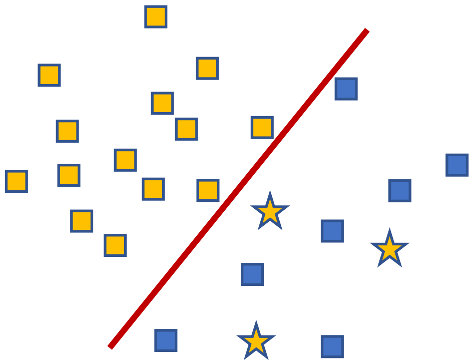


- Example: Family migration prediction
- Training/Test in Syria
- Yellow = Families that migrate
- Blue = Families that do not migrate

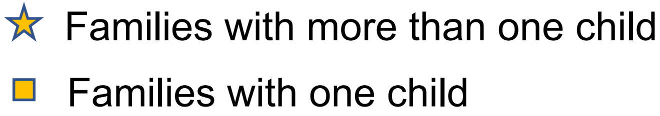
Slide credit: C.Troncoso

Distribution of Errors

Occurs when most mistakes of the classifier are concentrated in a **subpopulation/group**



- Example: Family migration prediction
- Training/Test in Syria
- Yellow = Families that migrate
- Blue = Families that do not migrate

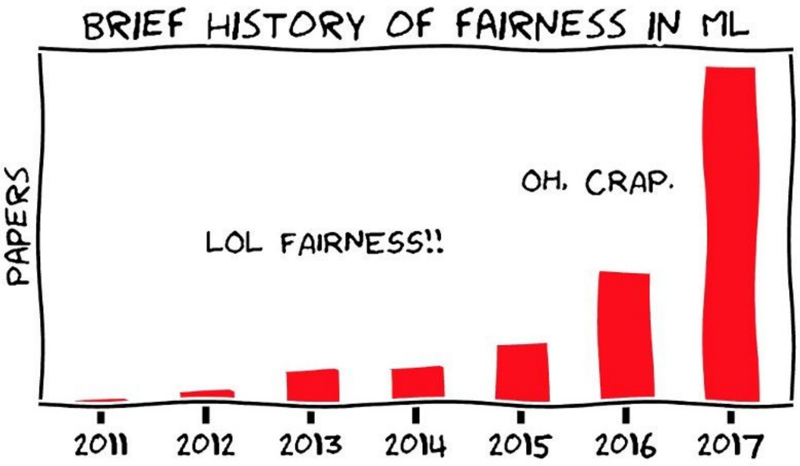


What can we do?

- Detect bias in ML models:

  IBM Bias Assessment Toolkit

- Transparency: explainable ML
- Anonymity does not help: bias not stem from identity



Arvind Narayanan: **Tutorial: 21 fairness definitions and their politics**

Based on slides by: C. Troncoso

Outline

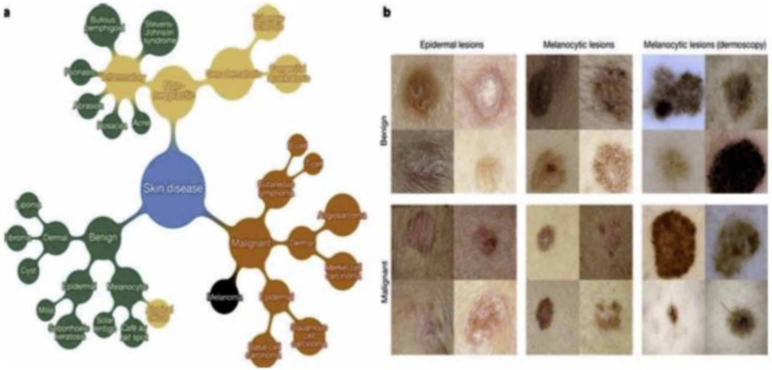
1. Introduction
2. Issues with deploying ML
3. **Applications of ML to cybersecurity**
4. Security of the ML system

Machine Learning is becoming ubiquitous

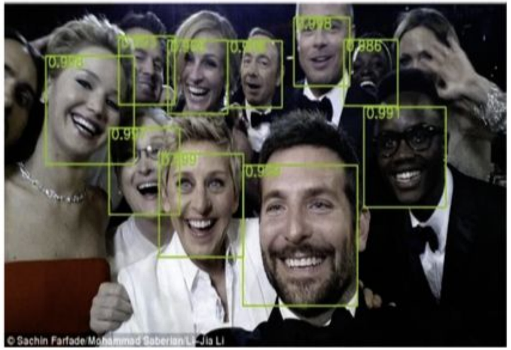
Self-driving Cars

Cybersecurity

Healthcare



Facial Recognition



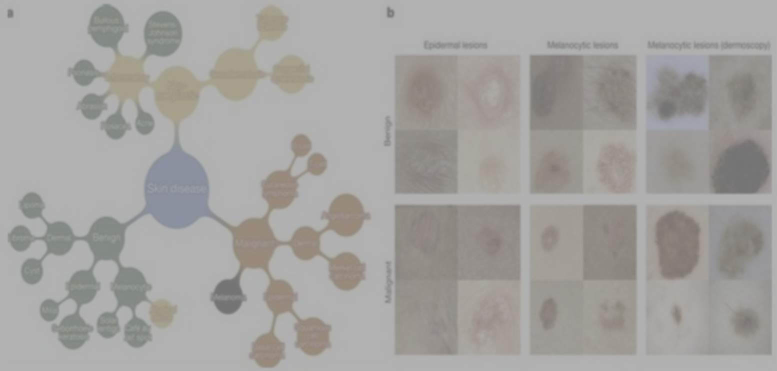
Speech Recognition

Machine Learning is becoming ubiquitous

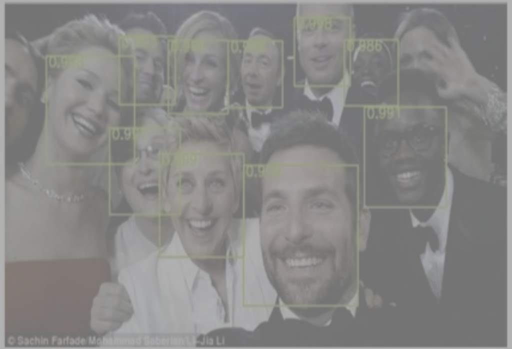
Self-driving Cars

Cybersecurity

Healthcare



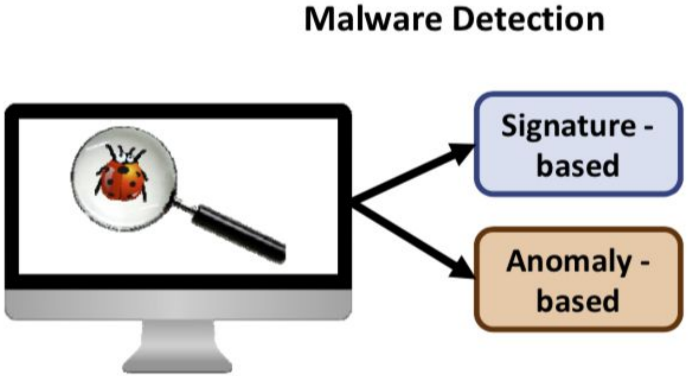
Facial Recognition

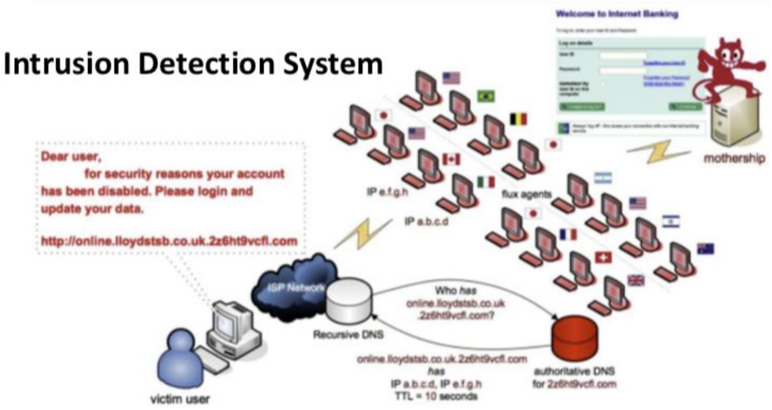


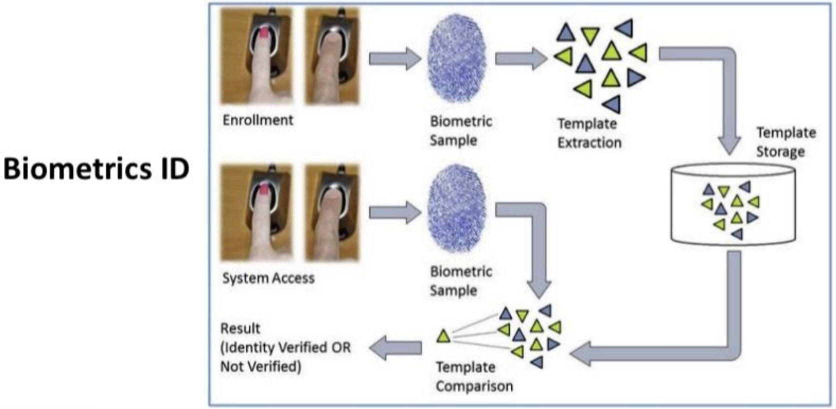
Speech Recognition

Based on slides by: S. Kundu

ML applications for cybersecurity

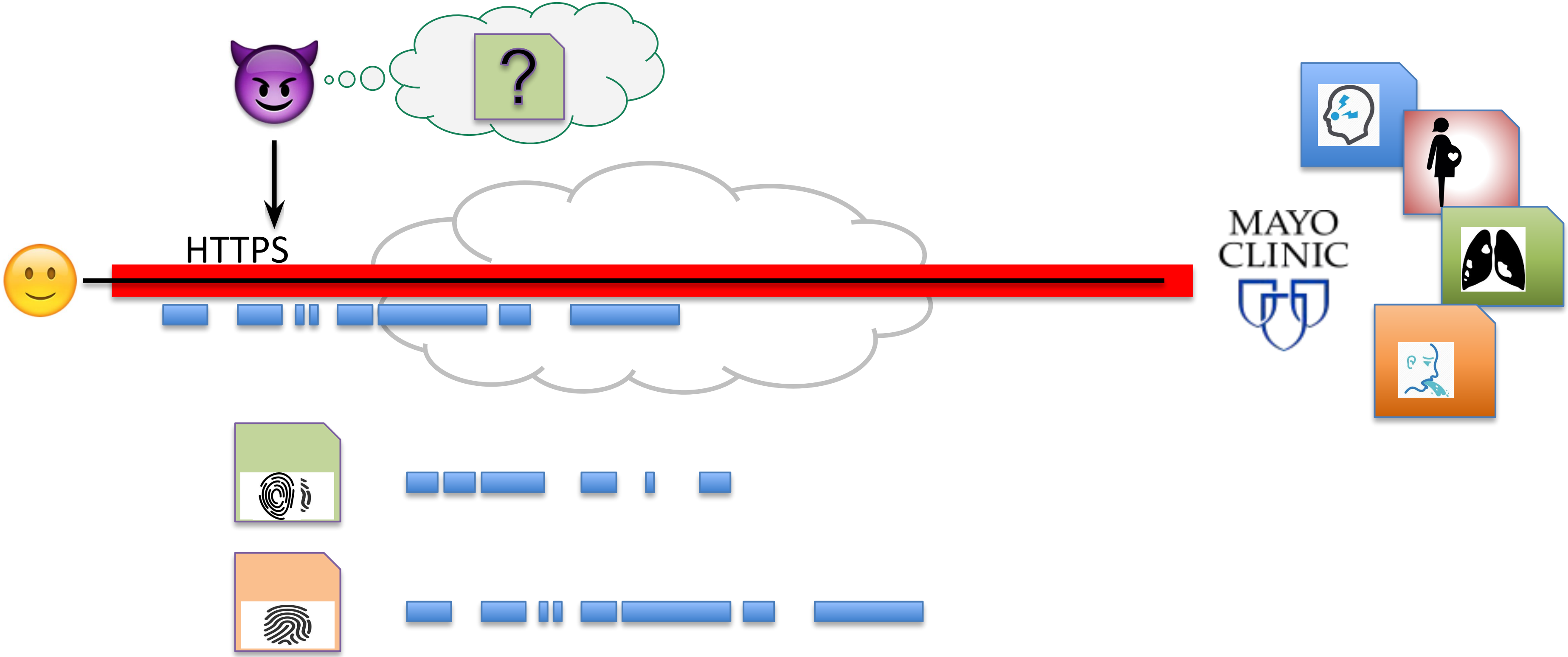






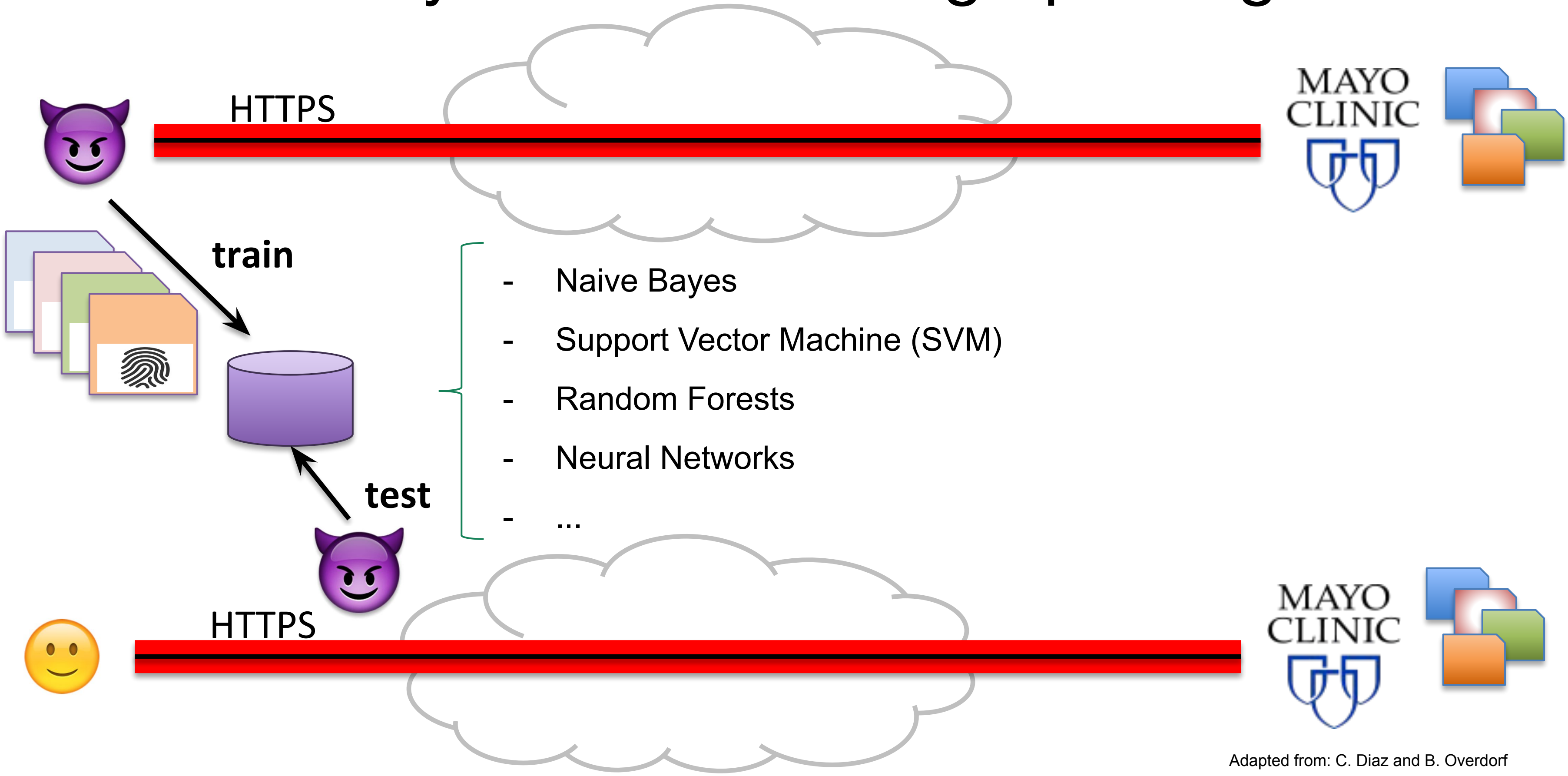
Slide credit: S. Kundu

Traffic Analysis: Website Fingerprinting



Adapted from: C. Diaz and B. Overdorf

Traffic Analysis: Website Fingerprinting



Adapted from: C. Diaz and B. Overdorf

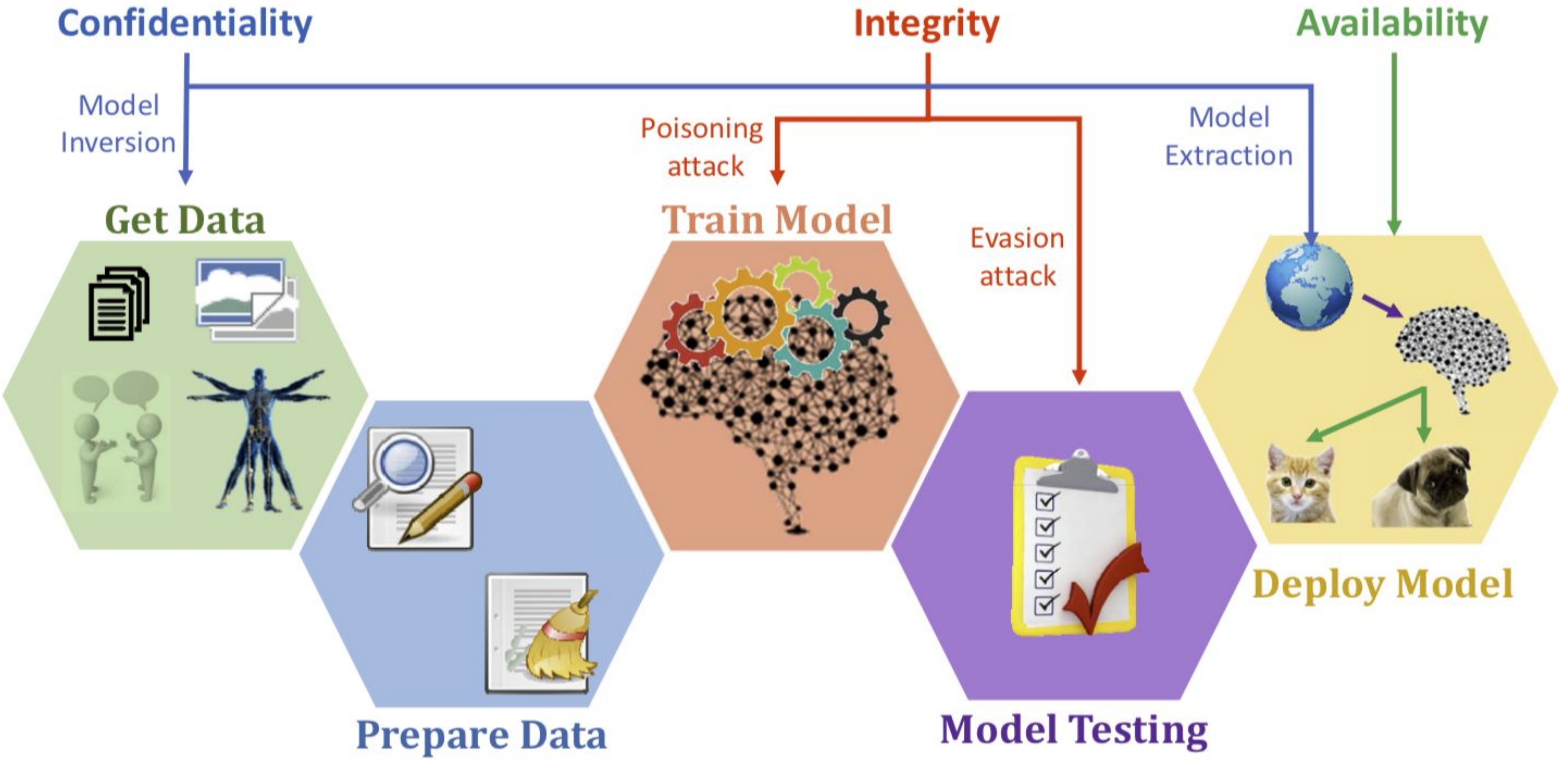
Website Fingerprinting takeaways

- Deployment issues:
  - Dynamism of pages: distributional shift over time
  - If IP anonymized/domain encrypted: base rate fallacy comes into play
- What's the cost to the adversary?
- Website Fingerprinting defenses
  - **Effectiveness of attacks and defenses depends on the security of ML!**

Outline

1. Introduction
2. Issues with deploying ML
3. Applications of ML to cybersecurity
4. **Security of the ML system**

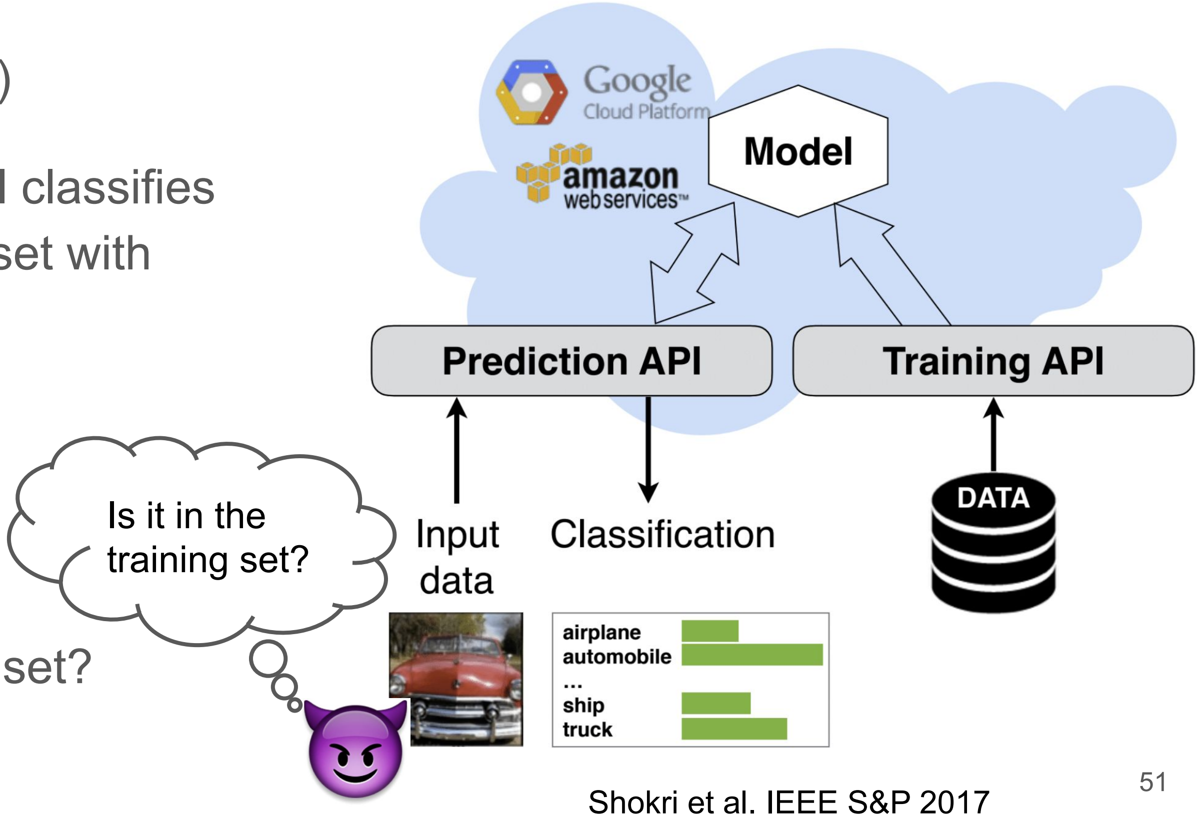
Security and Privacy in the ML workflow



Slide credit: S. Kundu

Confidentiality: membership inference attacks

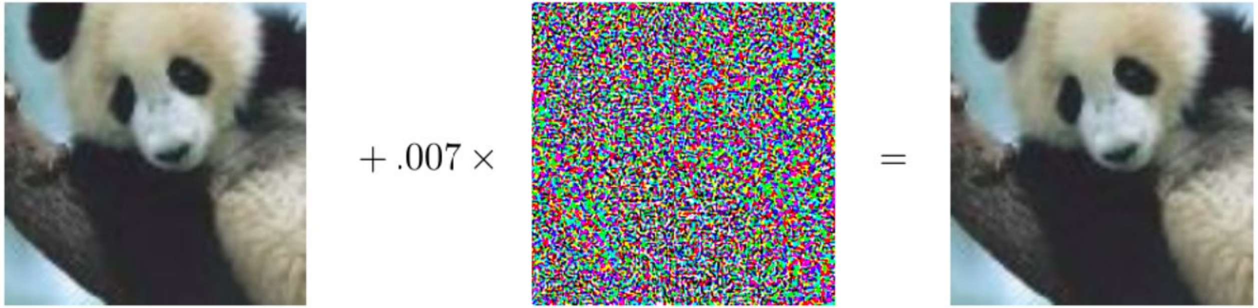
- ML as a Service (MLaaS)
- Key insight: overfit model classifies instances in the training set with high confidence
- Model extraction: steal the model!
- What is more valuable, the model or the training set?



Shokri et al. IEEE S&P 2017

Integrity: poisoning and evasion attacks

- Adversary's goal is to induce misclassifications:
  - Poisoning (during **training**): compromise data collection, subvert the learning process, facilitate future evasion (*backdoor* attacks), ...
  - Evasion (during **testing)**: find blind spots of the ML model in order to evade it.



Panda Adversarial Perturbation Cat

[GSS15] Goodfellow et al. Explaining and Harnessing **Adversarial Examples**

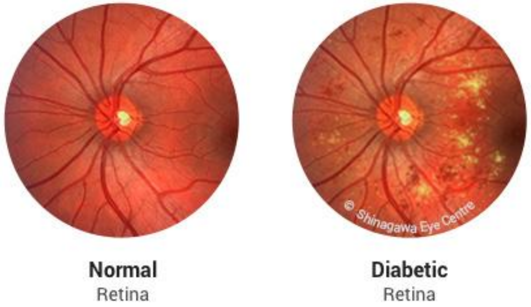
Adapted from: S. Kundu and J. Hayes

Adversarial examples in ML applications

1. Self-driving cars [1]

Before: Stop
After: 45 mph

2. Healthcare

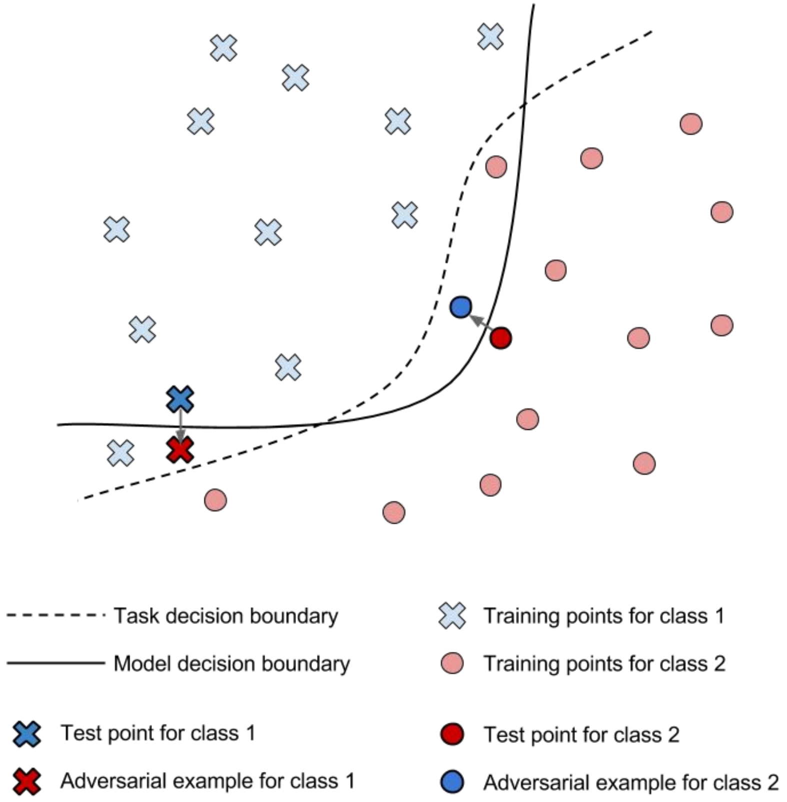


Before: Severe symptoms
After: No symptoms

[1] Evtimov et al., 2017

Adapted from: J. Hayes

Why do adversarial examples exist?



- Deep Learning is especially vulnerable due to its complexity.
- Early attempts at explaining this phenomenon focused on nonlinearity and overfitting.
- Linear behavior in high-dimensional spaces is sufficient to cause adversarial examples [1]

[1] Goodfellow et al. "Explaining and Harnessing Adversarial Examples", 2016

http://www.cleverhans.io/security/privacy/ml/2016/12/16/breaking-things-is-easy.html

More examples

- Speech recognition: Alexa case [1] and Dolphinattack [2]
- “Attacks” might be perceptible: circumvent face recognition [3]

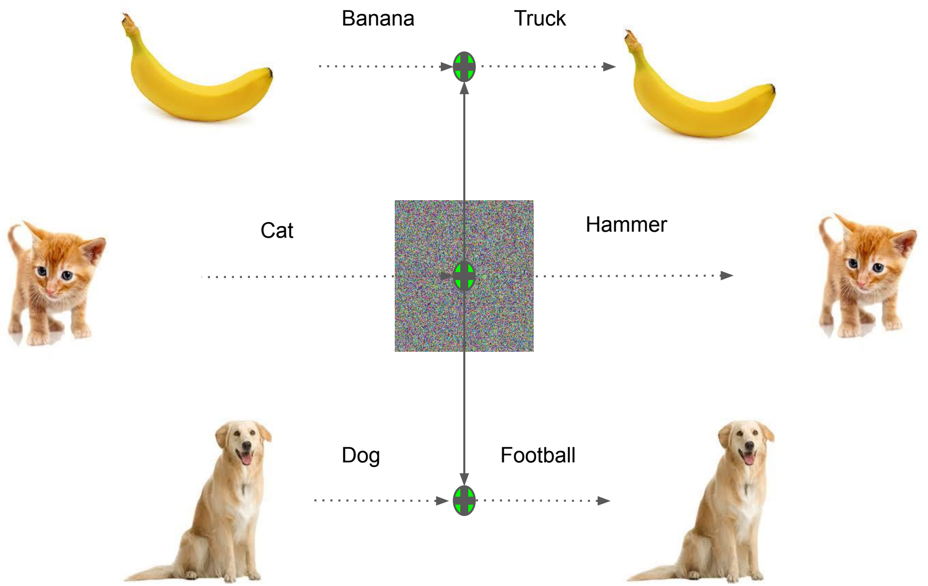
[1] https://qz.com/880541/amazons-amzn-alexa-accidentally-ordered-a-ton-of-dollhouses-across-san-diego/
[2] Zhang et al. CCS 2017
[3] Sharif et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Adapted from: J. Hayes

Universal adversarial perturbations

- In the most extreme case, it is possible to construct a single perturbation that will fool a model when added to **any** image!

- Attackers need minimal resources to attack your system!

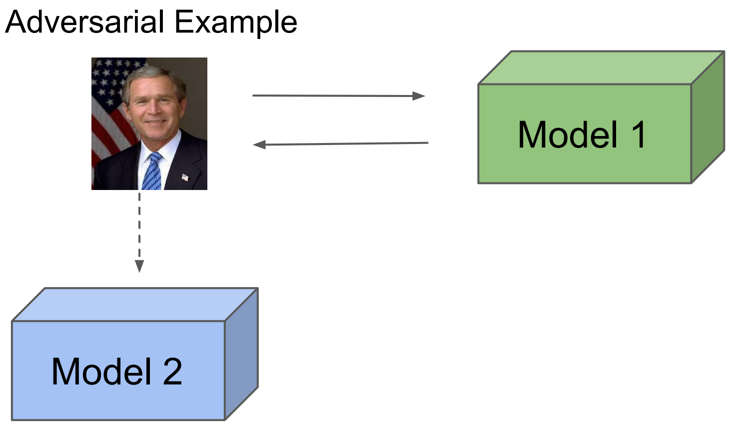


[GSS15] Goodfellow et al. Explaining and Harnessing Adversarial Examples
[MFF16] Moosavi-Dezfooli et al. Universal adversarial perturbations.

Adapted from: J. Hayes

Transferability property

- Adversarial examples transfer between different models.
- An adversarial example crafted against one model will generally fool other models.
- Attackers do not need repeated access to your system to attack it!



Adapted from: J. Hayes

Deep Learning and GDPR

GDPR, Art. 22 (on Automated decision-making): "*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*"

- Except in case that it is necessary to fulfill the contract or data owner gives consent.
- Even in that case, the data controller shall explain the basis on what the decision has been taken (e.g., to rule out discrimination).
- How can we do that with black-box models such as DL?

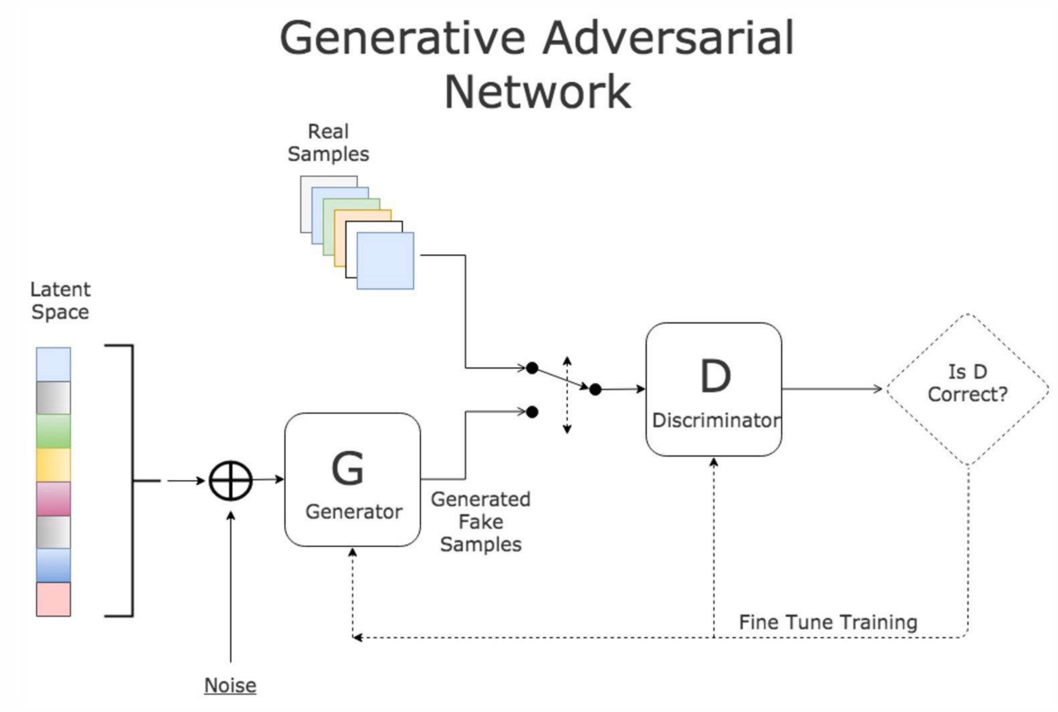
Adapted from: J. Hayes

Availability: downgrade performance

- An adversary can easily adapt adversarial examples to downgrade performance of the model, for example:
    - Poison the dataset to reduce the accuracy for a certain class.
    - Force ML to take low-performance decisions
- Harder to detect than a system failure

Countermeasures

- Membership inference: avoid overfitting!
- Adversarial Examples: very recent (2015) and still not well understood
  - Data augmentation: re-train on (virtual) adversarial examples
  - Pre-processing: sequeeze features and add variable noise to inputs.
  - GANs: used to attack and defend.



Takeaways

1. ML might be secure and work in the lab but **still fail** when deployed
2. Dual use of ML: it can be used for **to defend but also to attack**
3. ML itself is vulnerable: **attacks exist against all security properties** of ML
4. Security of ML adds another dimension to cybersecurity: both attacks and defenses **depend on the security of ML itself**.

Thanks!