Introduction to Data Science

(Lecture 5)

Dr. Mohammad Pourhomayoun

Assistant Professor

Computer Science Department

California State University, Los Angeles

Review

Review: What is Machine Learning?

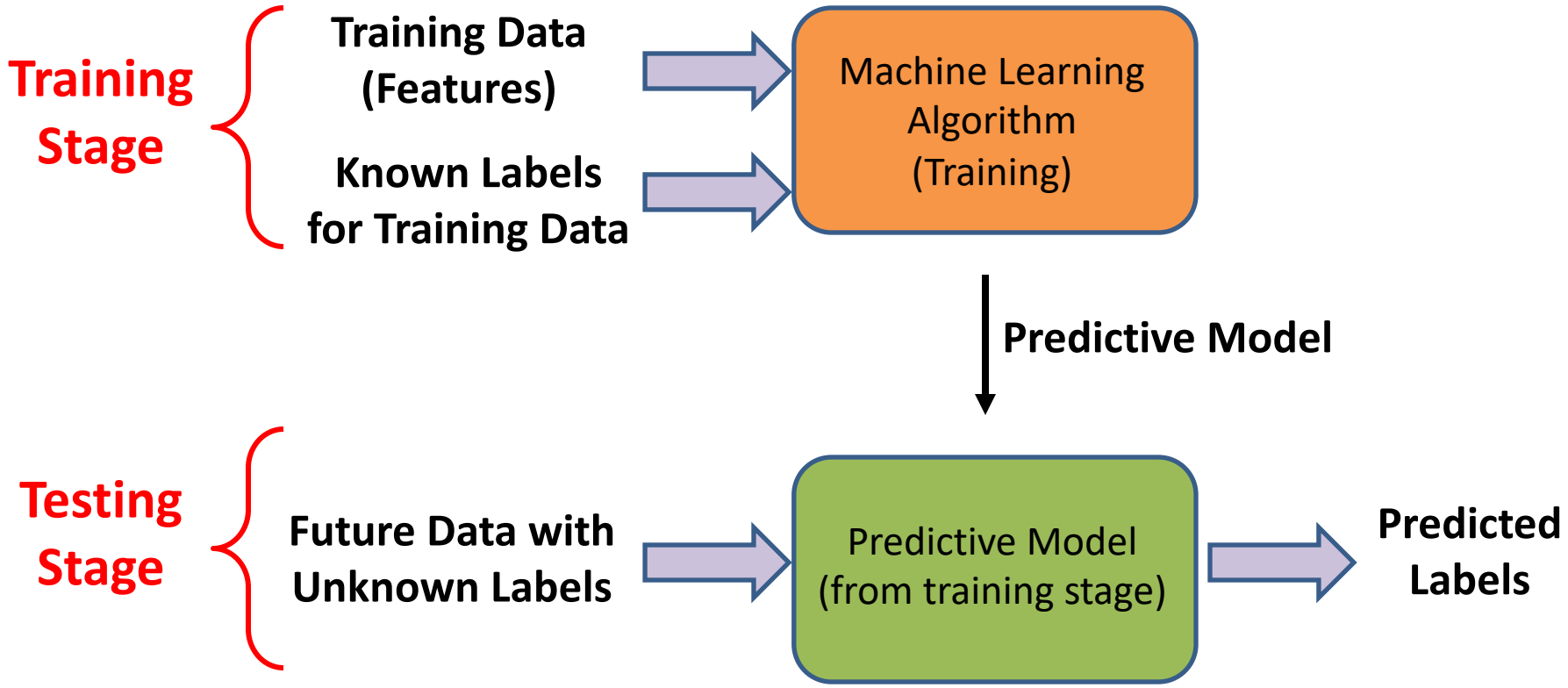
- **A Definition:** Designing and constructing algorithms or methods that give computers the ability to learn from past data, without being explicitly programmed, and then make predictions on future data.

- **Another Definition:** A set of algorithms that can automatically detect and extract patterns in past data, and then use the extracted patterns to predict on future data, or to perform other kinds of decision making.

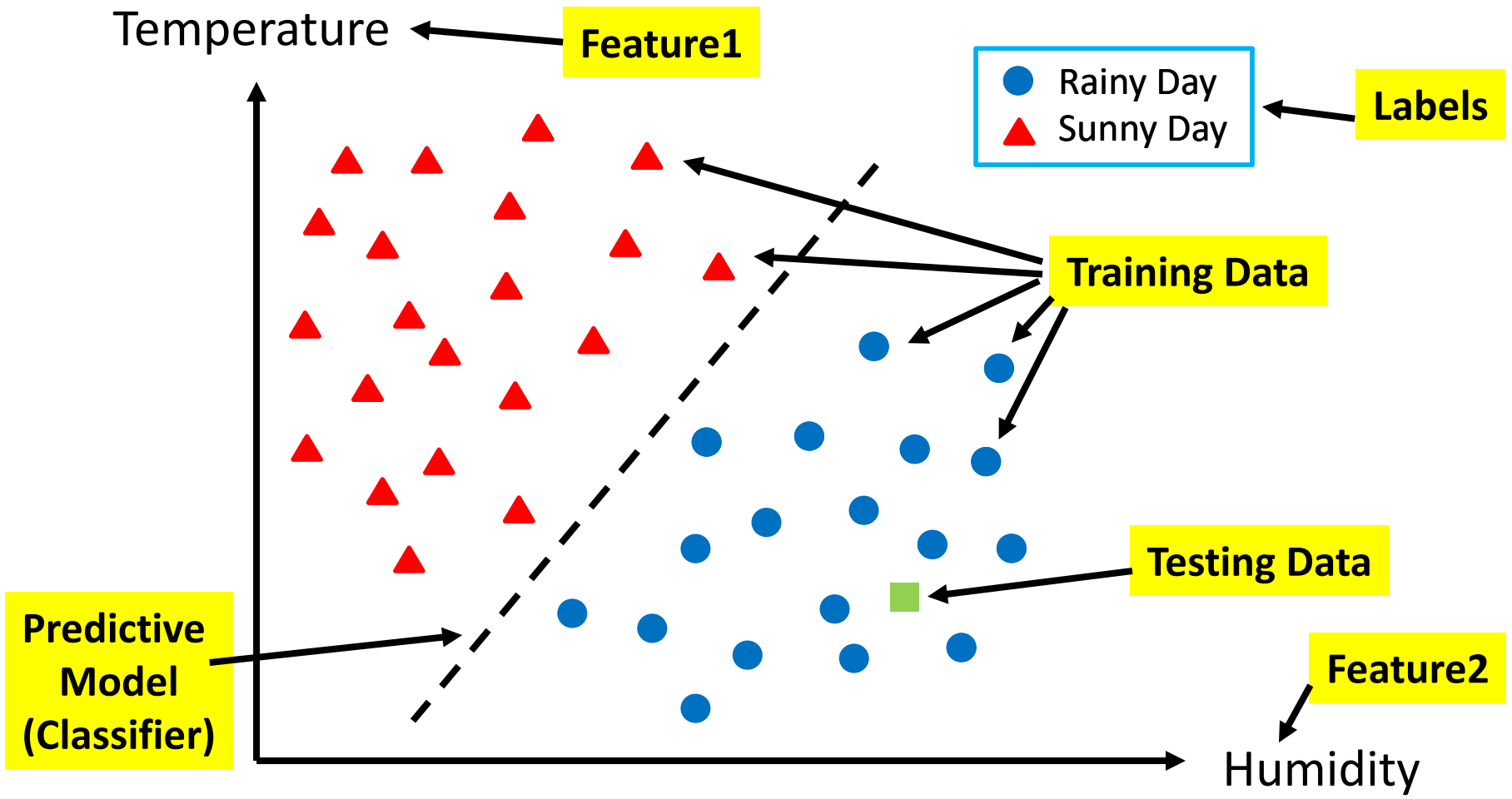
Review: Common Learning Settings

- **Supervised learning:** Learning from **labeled** observations.
  - In training stage, the algorithm is presented with **features and their known labels**, and the goal is to train a model that maps future inputs to new labels.
- **Unsupervised learning:** Learning from **unlabeled** observations.
  - The algorithm is presented **Only** with **features!** The goal is to **Discover hidden patterns and structure** from features alone. It is like a Data Exploration to find hidden patterns.
- **Semi-supervised learning:** Labels are provided only for a part of the training data. The remaining data is unlabeled.
- **Reinforcement learning:** Learning from an ***agent*** taking ***actions*** in an ***environment*** so as to maximize a long-term ***reward***. E.g. Game Theory, Control Theory.
- **Transfer learning:** Learning from a dataset while solving a problem, and then applying the **extracted knowledge** to a different but related dataset/problem.
- **Active learning:** Similar to Semi-Supervised Learning, but the algorithm is able to interactively query the user or some other information source to obtain the labels as needed.

Review: Supervised Learning: Learning from labeled Data



Review: Classification



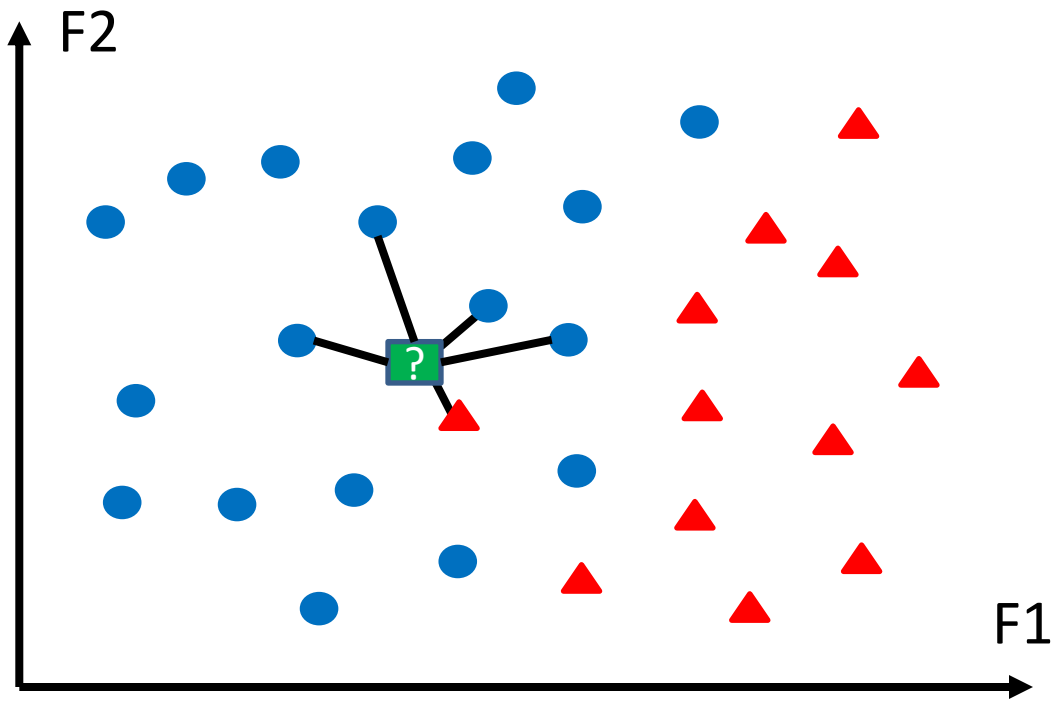
Review: Feature Table

- *Training dataset*: $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ with known label.
- Now, we have a new sample with unknown label: $(\boldsymbol{x}, y=?)$

	sepal length	sepal width	petal length	petal width	Label	
$\boldsymbol{x}_1$	5.3	3.7	1.5	0.2	setosa	y_1
$\boldsymbol{x}_2$	5	3	2	0.2	setosa	y_2
	7.0	3.2	4.7	1.4	versicolor	
:	6.4	3.2	4.5	1.5	versicolor	:
	6.3	2.7	4.9	1.8	virginica	
$\boldsymbol{x}_N$	7.9	3.8	6.4	2	virginica	y_N
$\boldsymbol{x}$	7	3.9	5.9	1.3	???	$y=?$

Review: KNN Classifier

- K-Nearest Neighbor (KNN) classifier algorithm classifies objects based on **majority of K closest training samples** in the feature space, e.g. K=5.



Review: KNN Classification

- 1st-Nearest Neighbor: $NN_1(\boldsymbol{x})$
- 2nd-Nearest Neighbor: $NN_2(\boldsymbol{x})$
- 3rd-Nearest Neighbor: $NN_3(\boldsymbol{x})$

  :
- Kth-Nearest Neighbor: $NN_K(\boldsymbol{x})$

- The set of K-Nearest Neighbors:

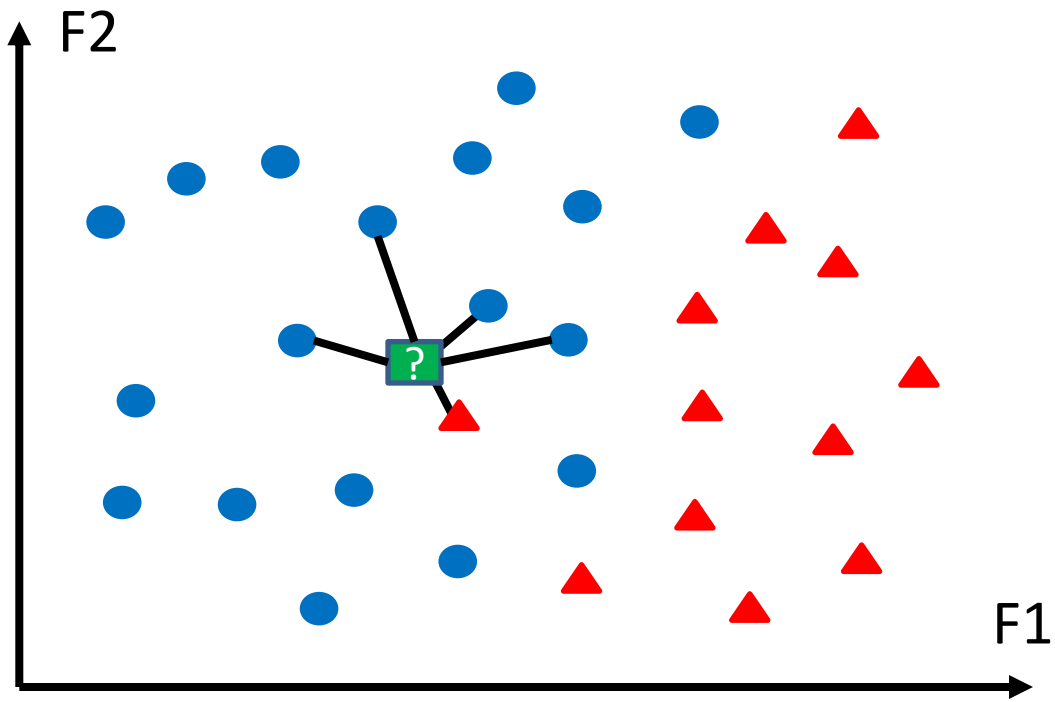
$$KNN(\boldsymbol{x}) = \left\{ NN_1(\boldsymbol{x}), NN_2(\boldsymbol{x}), \ldots, NN_K(\boldsymbol{x}) \right\}$$

- **Classification rule**: Voting: Selecting the Label with the majority in $KNN(\boldsymbol{x})$.

Review: KNN Classifier

- K-Nearest Neighbor (KNN) classifier algorithm classifies objects based on K closest training samples in the feature space, e.g. K=5.

Out of 5 NN:
4 are blue, 1 is red.
Thus, our prediction for ? is blue ● !



Decision Tree Classifier

Titanic Disaster

- Let's start this topic with a famous problem/competition from kaggle website: Predicting **survival on the Titanic!**

[1]: Ref: www.kaggle.com.

Titanic Disaster

- Let's start this topic with a famous problem/competition from kaggle website: Predicting **survival on the Titanic!**

[1]: Ref: www.kaggle.com.

Predict survival on the Titanic

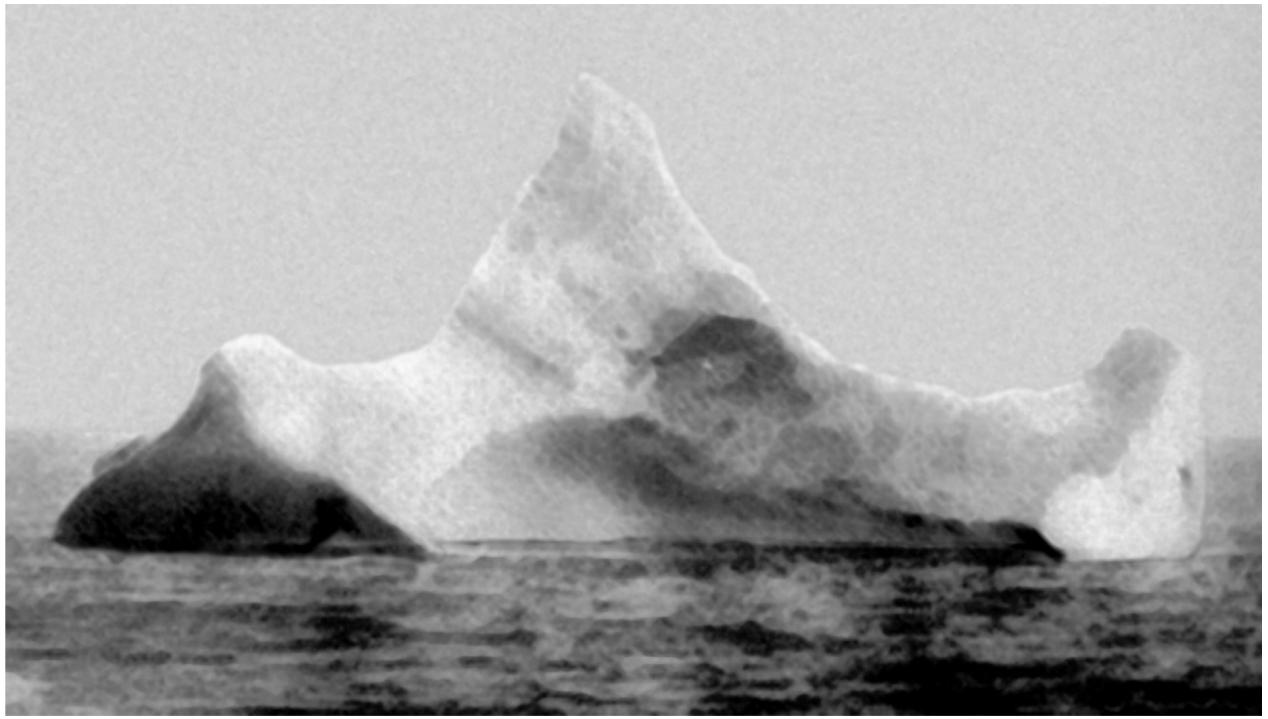
- On April 15, 1912, the Titanic sank after colliding with an iceberg, **killing 1502** out of 2224 passengers and crew.
- One of the reasons that the shipwreck led to such loss of life was that there were **not enough lifeboats** for the passengers and crew.
- Although there was some element of luck involved in surviving the sinking, some groups of people were **more likely to survive** than others, such as women, children, and the upper-class.
- In this challenge, we would like to analyze what sorts of people were likely to survive. In particular, we want to apply the tools of data science and machine learning to predict which passengers survived the tragedy[1].

[1]: Ref: Kaggle website: www.kaggle.com.

Titanic Disaster

- Let's start this topic with a famous problem/competition from kaggle website: Predicting **survival on the Titanic!**

  - The iceberg thought to have been hit by Titanic, photographed on the morning of 15 April 1912.



[1]: Ref: www.kaggle.com.

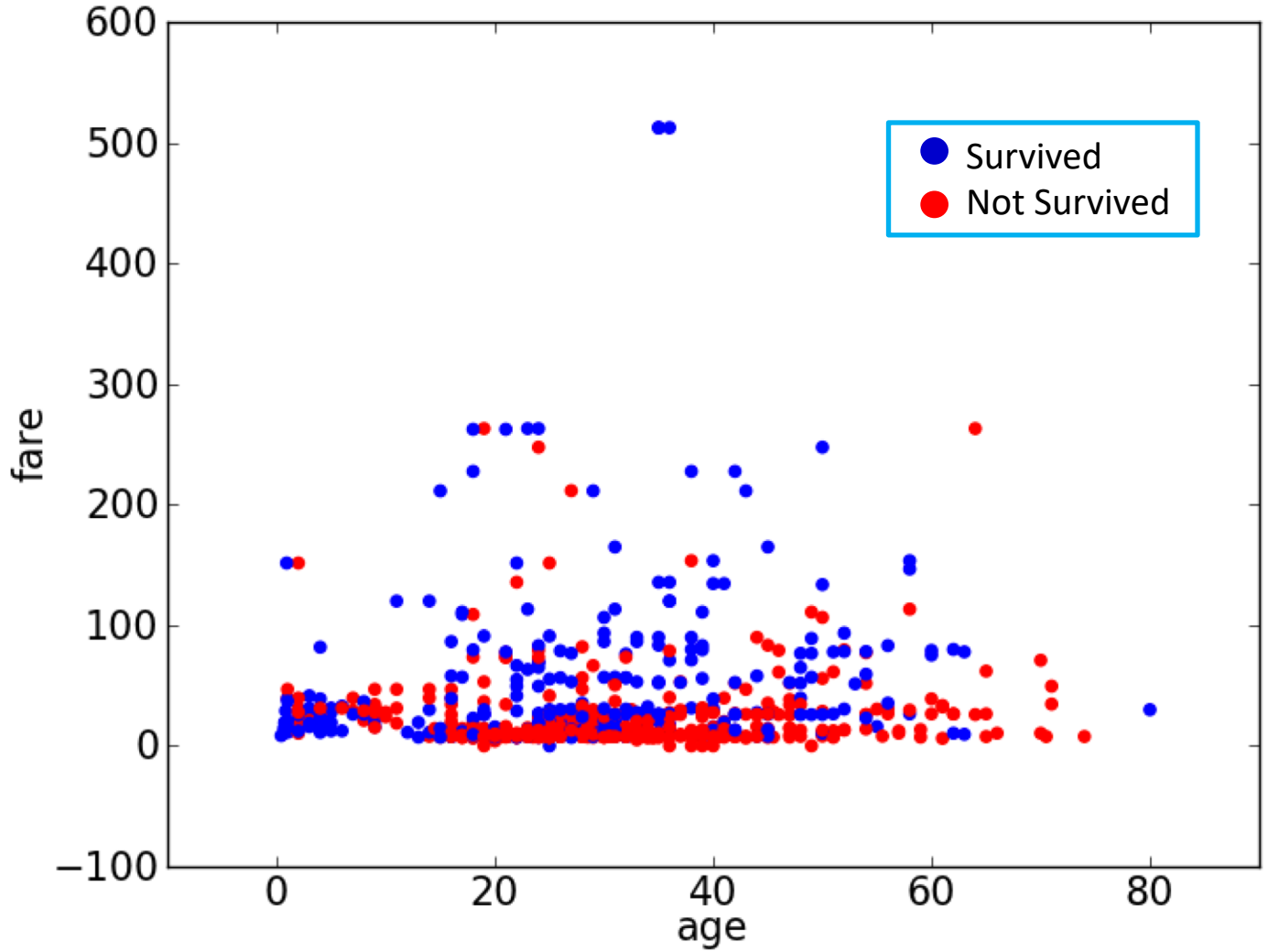
Predict survival on the Titanic

- Passenger list on Titanic[1]:

pclass	sex	age	sibsp	parch	fare	survived
3	male	22	1	0	7.25	0
1	female	38	1	0	71.2833	1
3	female	26	0	0	7.925	1
1	female	35	1	0	53.1	1
3	male	35	0	0	8.05	0
3	male		0	0	8.4583	0
1	male	54	0	0	51.8625	0
3	male	2	3	1	21.075	0
3	female	27	0	2	11.1333	1
2	female	14	1	0	30.0708	1
3	female	4	1	1	16.7	1
1	female	58	0	0	26.55	1
3	male	20	0	0	8.05	0

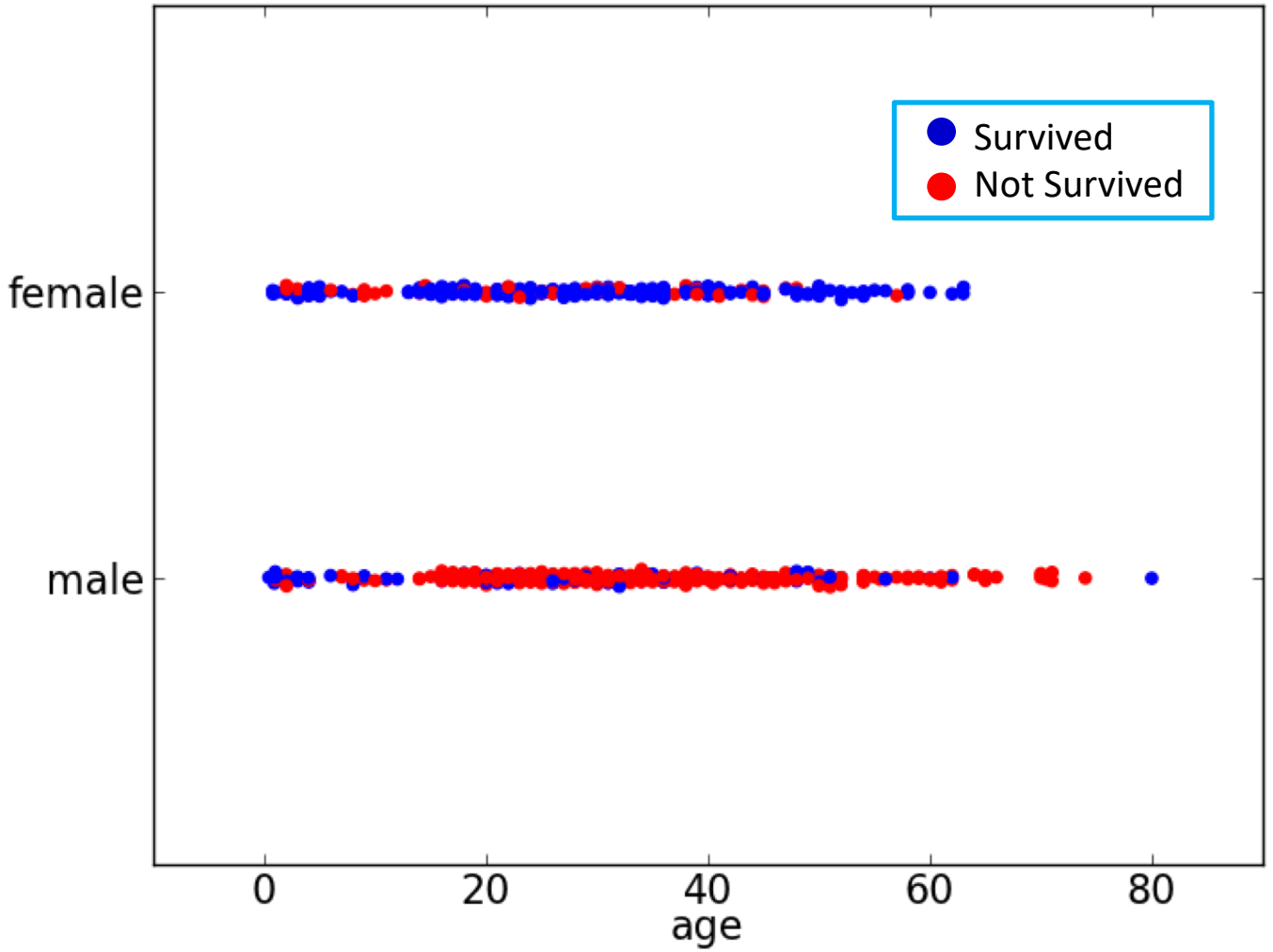
[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

Predict survival on the Titanic



[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

Predict survival on the Titanic



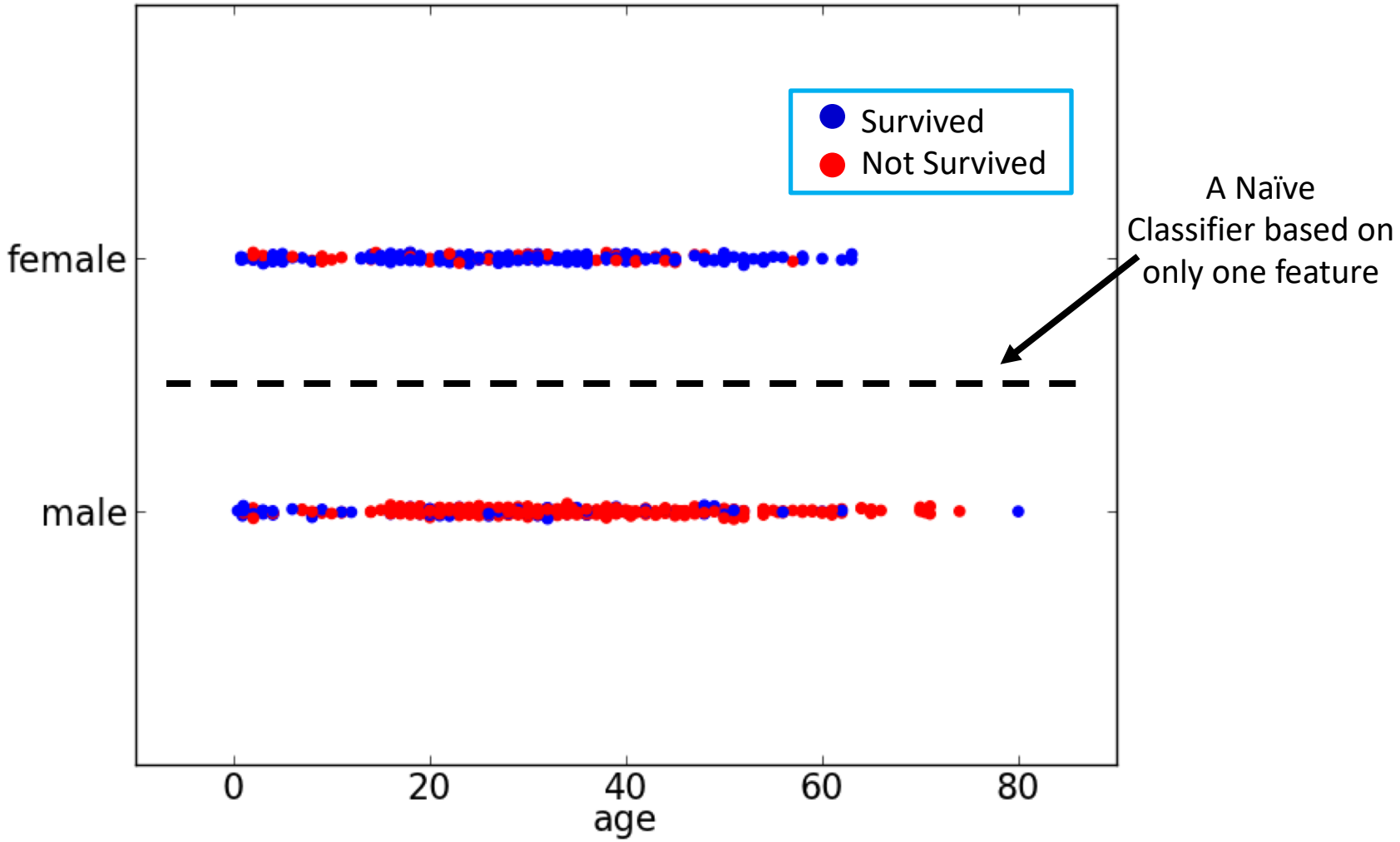
[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

A Naïve Classifier for Titanic

- Making decision based on only **one feature**:
  - Example: Based on Gender: If most females survived, then let's assume every female survives.

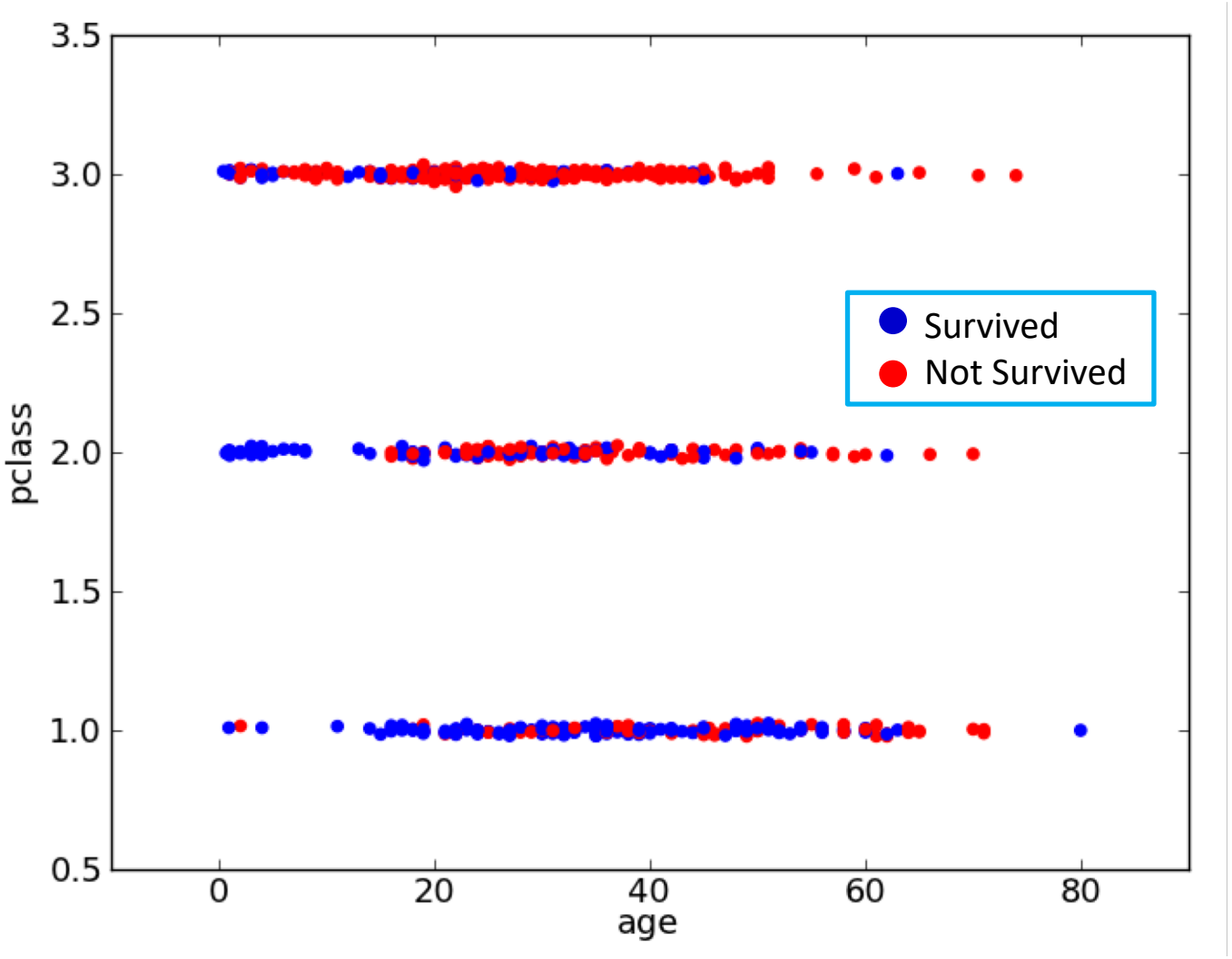
- So, our naïve classification rule will be:
  - IF (Sex='female') THEN Survive ← Yes
  - ELSE IF (Sex='male') THEN Survive ← No

A Naïve Classifier for Titanic



[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

Predict survival on the Titanic



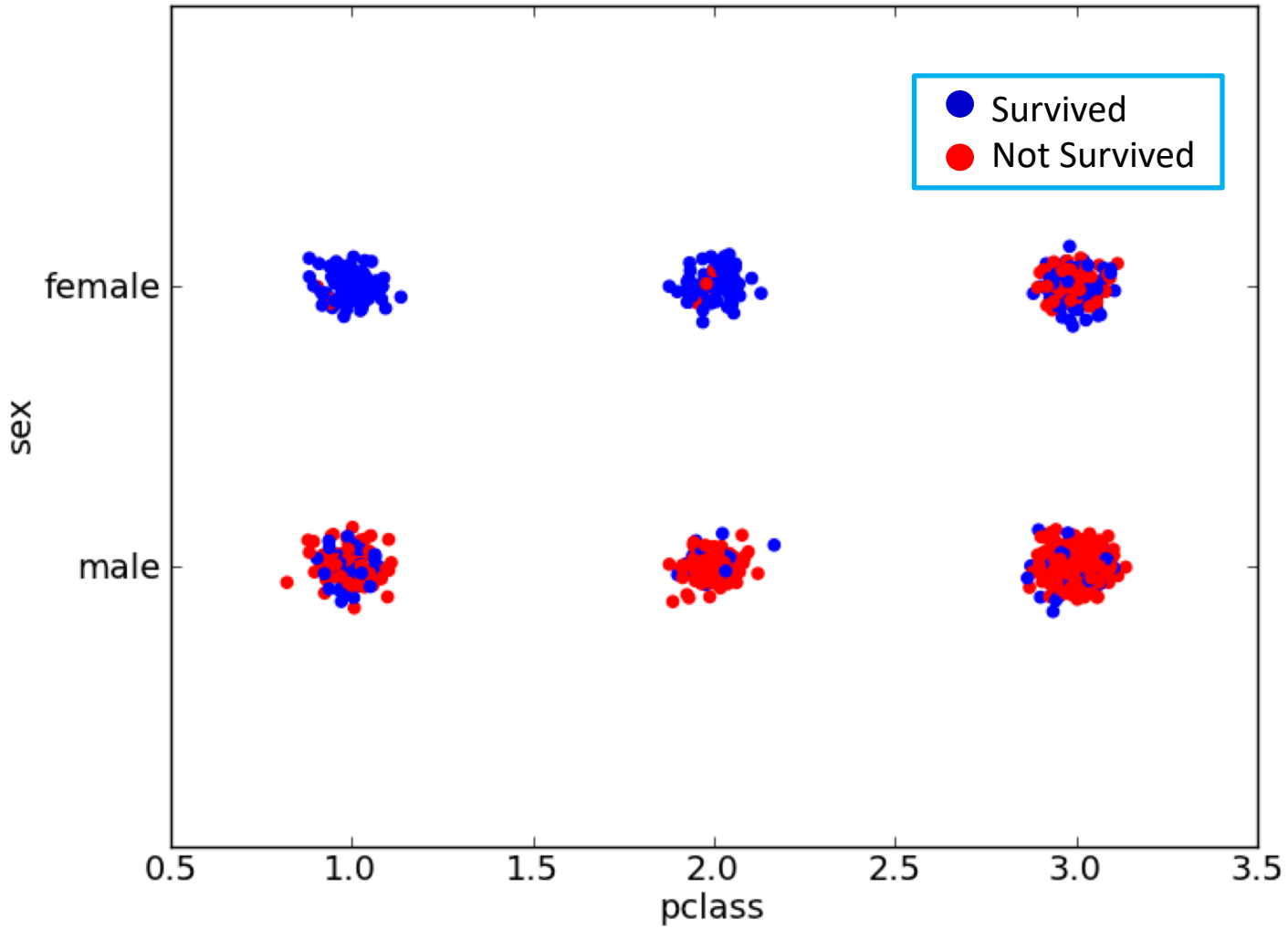
[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

A Naïve Classifier for Titanic

- Making decision based on only **one feature**.
  - Example: Based on “pclass”.

- So, our naïve classification rule will be:
  - IF (pclass=‘1’) THEN Survive ← Yes
  - ELSE IF (pclass=‘2’) THEN Survive ← Yes
  - ELSE IF (pclass=‘3’) THEN Survive ← No

Predict survival on the Titanic



[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

An Improvement on the Classifier

- Making decision based on **two features**.
  - Example: Based on “gender” and “pclass”.
- So, our classification rule will be:

IF (Sex=‘female’):

IF (pclass=‘1’) OR (pclass=‘2’) THEN: Survive ← Yes

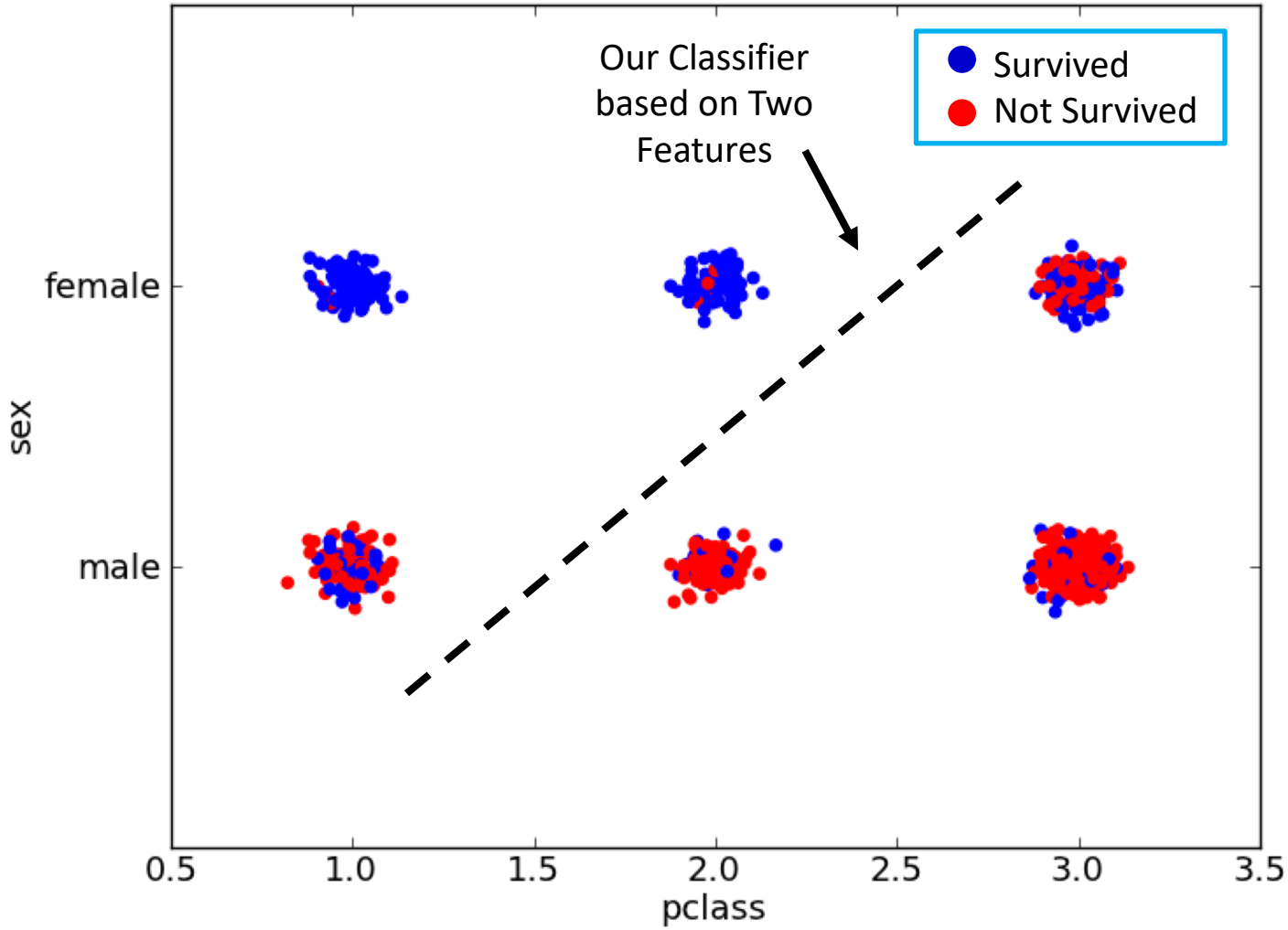
ELSE IF (pclass=‘3’) THEN: Survive ← No

ELSE IF (Sex=‘male’):

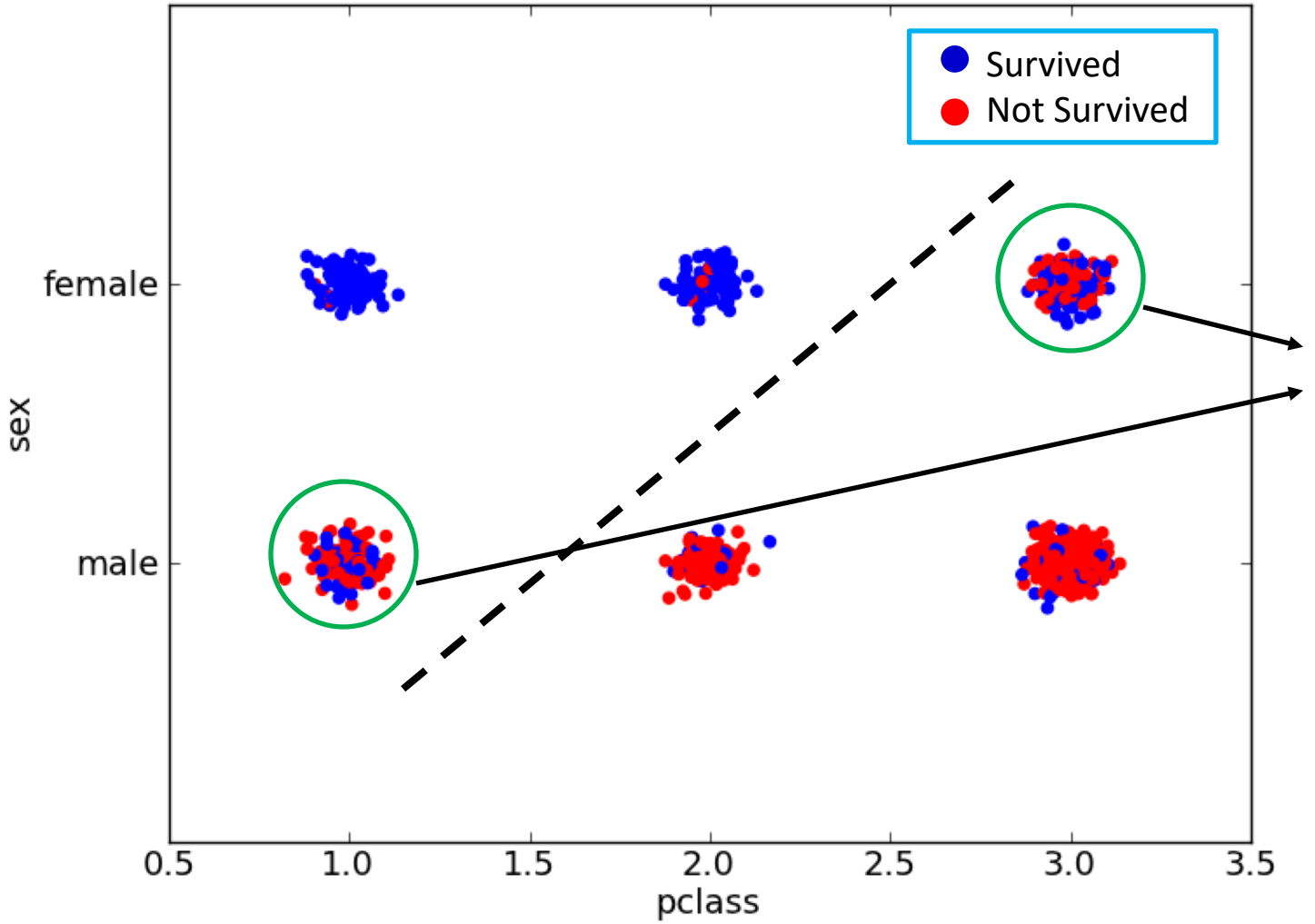
IF (pclass=‘2’) OR (pclass=‘3’) THEN: Survive ← No

ELSE IF (pclass=‘1’) THEN: Survive ← Yes

Predict survival on the Titanic



Predict survival on the Titanic



Again these two groups are mixed. Thus, let's use a third feature to divide them!

An Improvement on the Classifier

- Making decision based on **three features**.
  - Example: Based on “gender”, “pclass”, and “age”.

IF (Sex=‘female’):

IF (pclass=‘1’) OR (pclass=‘2’) THEN: Survive ← Yes

ELSE IF (pclass=‘3’):

IF (age<4) THEN: Survive ← Yes

ELSE IF (age>4) THEN: Survive ← No

ELSE IF (Sex=‘male’):

IF (pclass=‘2’) OR (pclass=‘3’) THEN: Survive ← No

ELSE IF (pclass=‘1’):

IF (age<4) THEN: Survive ← Yes

ELSE IF (age>4) THEN: Survive ← No

What does this structure look like?

IF (Sex='female'):

IF (pclass='1') OR (pclass='2') THEN: Survive ← Yes

ELSE IF (pclass='3'):

IF (age<4) THEN: Survive ← Yes

ELSE IF (age>4) THEN: Survive ← No

ELSE IF (Sex='male'):

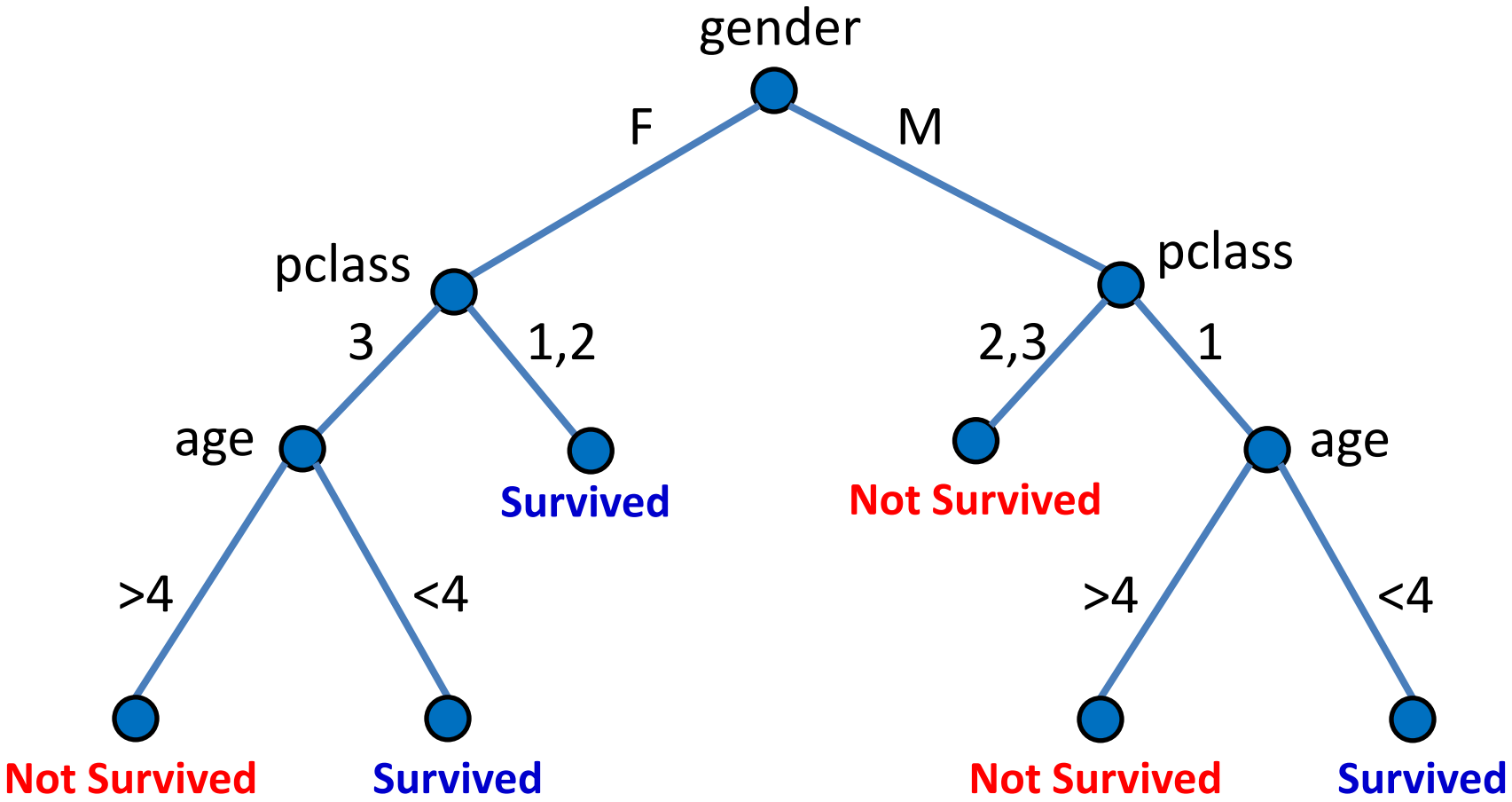
IF (pclass='2') OR (pclass='3') THEN: Survive ← No

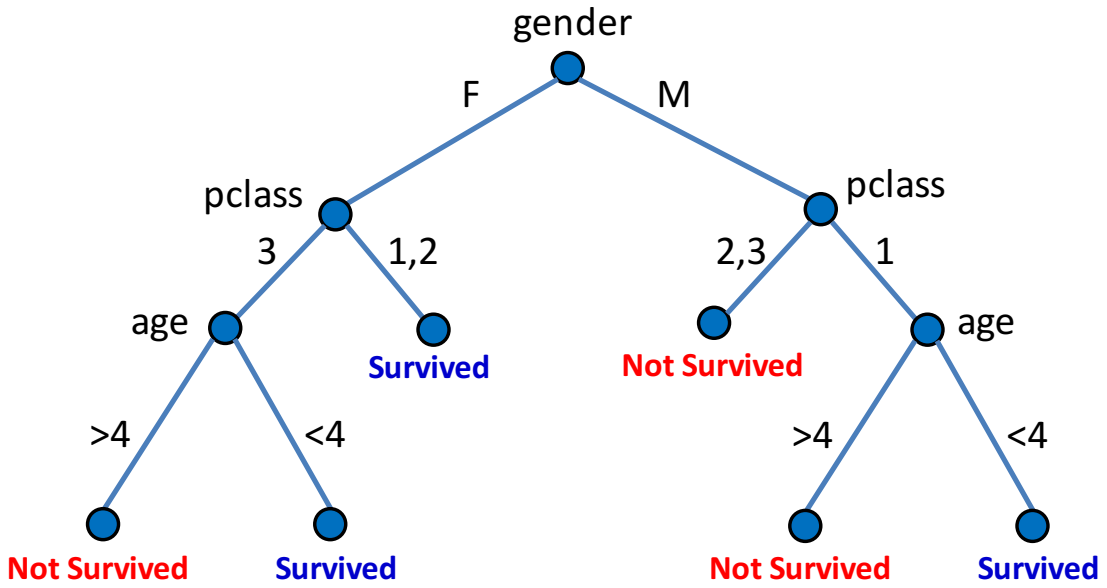
ELSE IF (pclass='1'):

IF (age<4) THEN: Survive ← Yes

ELSE IF (age>4) THEN: Survive ← No

Decision Tree





Qustion1: Ellen was 70 years old and had a 1st class ticket. Did she survive?

Qustion2: Sarah was 7 years old and had a 3rd class ticket. Did she survive?

Qustion3: Tom was 3 years old and had a 1st class ticket. Did he survive?

Qustion4: Frank was 30 years old and had a 1st class ticket. Did he survive?

Qustion5: Jason had a 2nd class ticket. Did he survive?

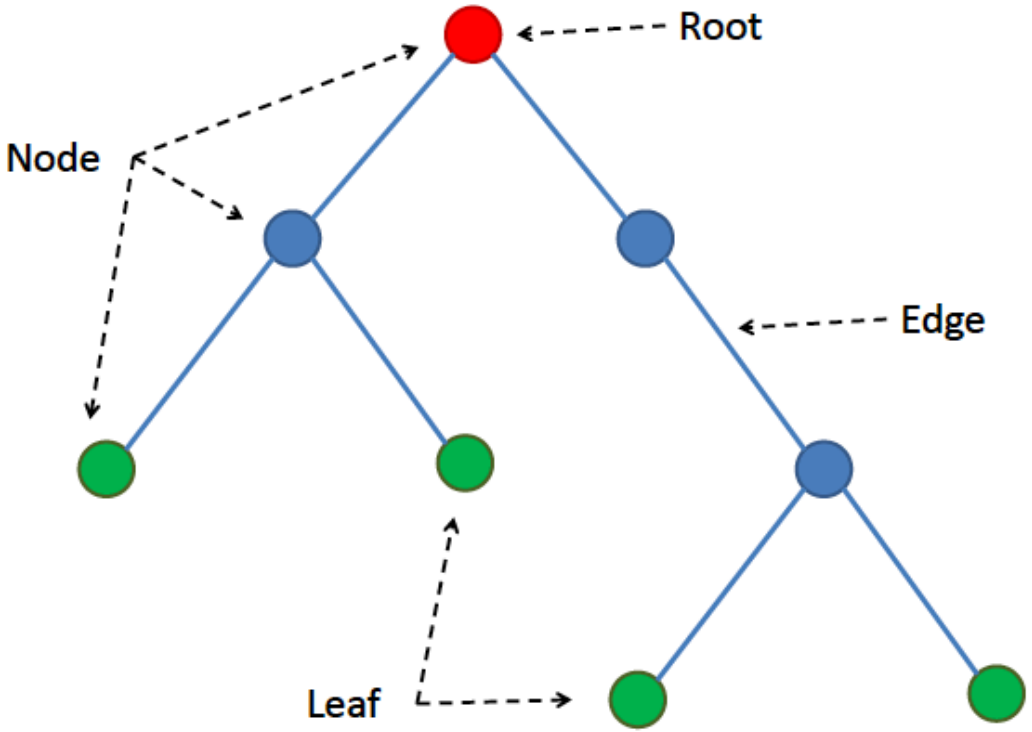
Qustion6: Kevin was 30 years old. Did he survive?

Qustion7: Jennifer was 3 years old. Did she survive?

Important Note

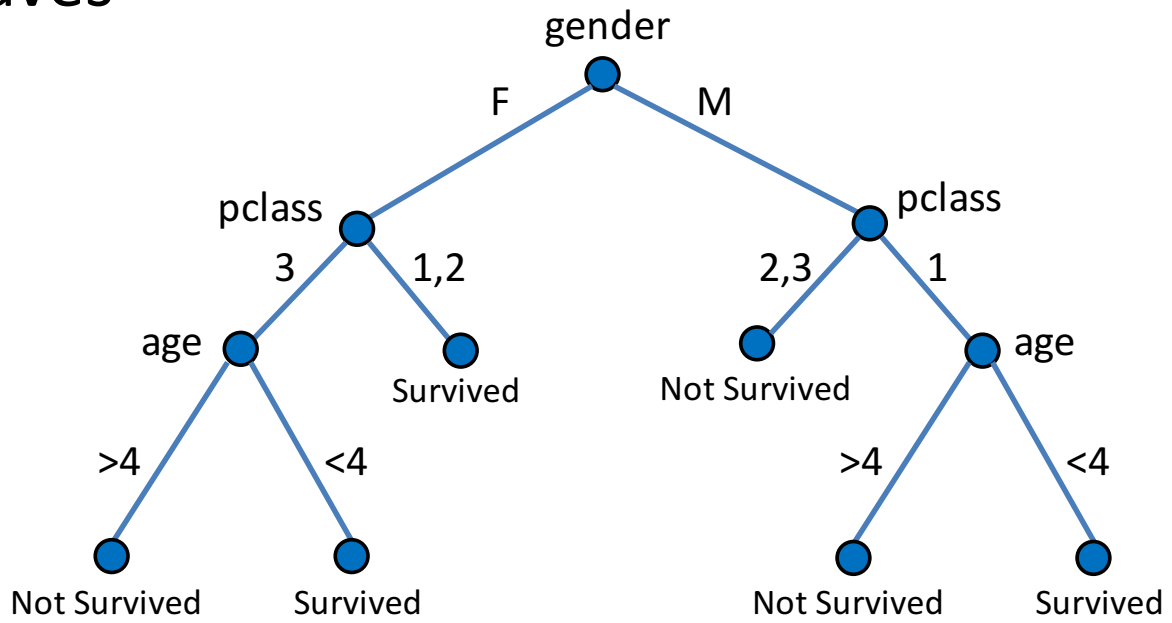
- **EVERY** prediction model has some levels of error!
- There is rare to have a predictor with 100% accuracy!

Terminology



Training a Decision Tree Model

- **Three things to learn in training stage:**
  1. The structure of the tree: The priority of features
  2. The threshold values
  3. The values for the leaves

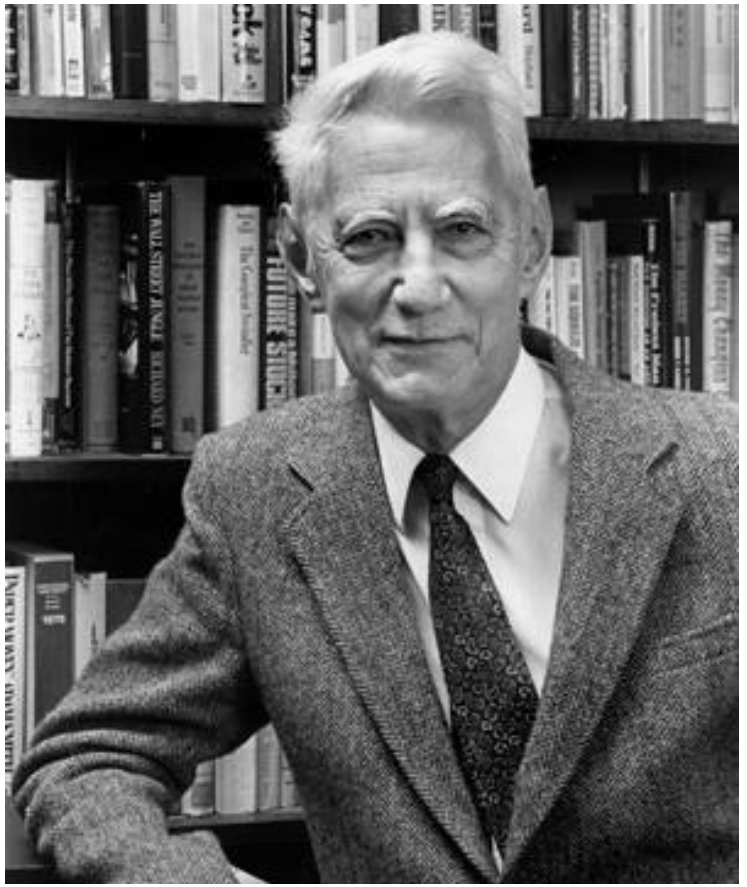


Training a Decision Tree Model

- **Question**: How to select the **first feature** at the top of the tree: The best feature that can split (classify) the data samples.
- **Idea**: The best feature is the one that provides the ***most amount of information*** about the label.
- So, we need a ***metric*** to measure **information**.

Claude E. Shannon (1916 – 2001)
The Father of Information Theory

- Shannon is noted for having founded ***information theory*** with a landmark paper, ***"A Mathematical Theory of Communication"***, that he published in 1948.
- He is, perhaps, equally well known for founding ***digital circuit design theory*** in 1937, when—as a 21-year-old master's degree student at MIT [1].



[1]: wikipedia

Two Important Concepts about Measuring the Information

1. The amount of information about an event x has **inverse relationship** to the **probability** of that event.

- Example:
  - *"The sun will rise tomorrow morning"*
    - This sentence provides very Low amount of information because it talks about a common (very likely) event.
  - *"An Eclipse occurs tomorrow"*
    - This sentence provides High amount of information because it is an unlikely event.

The amount of information about event x → $I(X) \sim \frac{1}{p(x)}$ ← The Probability of event x

Two Important Concepts about Measuring the Information

2. When two independent events happens, the joint probability of them is the **multiplication** of the two probabilities. However, the total information about two independent events should be the **summation** of the two piece of information.

- Example: Flipping a Coin twice: H,T
  - Prob(two independent events) = prob(event1) * prob(event2)
  - info(two independent events) = info(event1) + info(event2)

Two Important Concepts about Measuring the Information

- When two independent events happen, the joint probability is the multiplication of the two probabilities. However, in this case, the total information about them should be the summation of the two piece of information.

- So, the "**Information function**" should have this property:
  Information (p_1,p_2) = Information(p_1) + Information(p_2)

Two Important Concepts about Measuring the Information

Question: What function has this property?

$$f(xy) = f(x) + f(y)$$

Answer: Log!!!

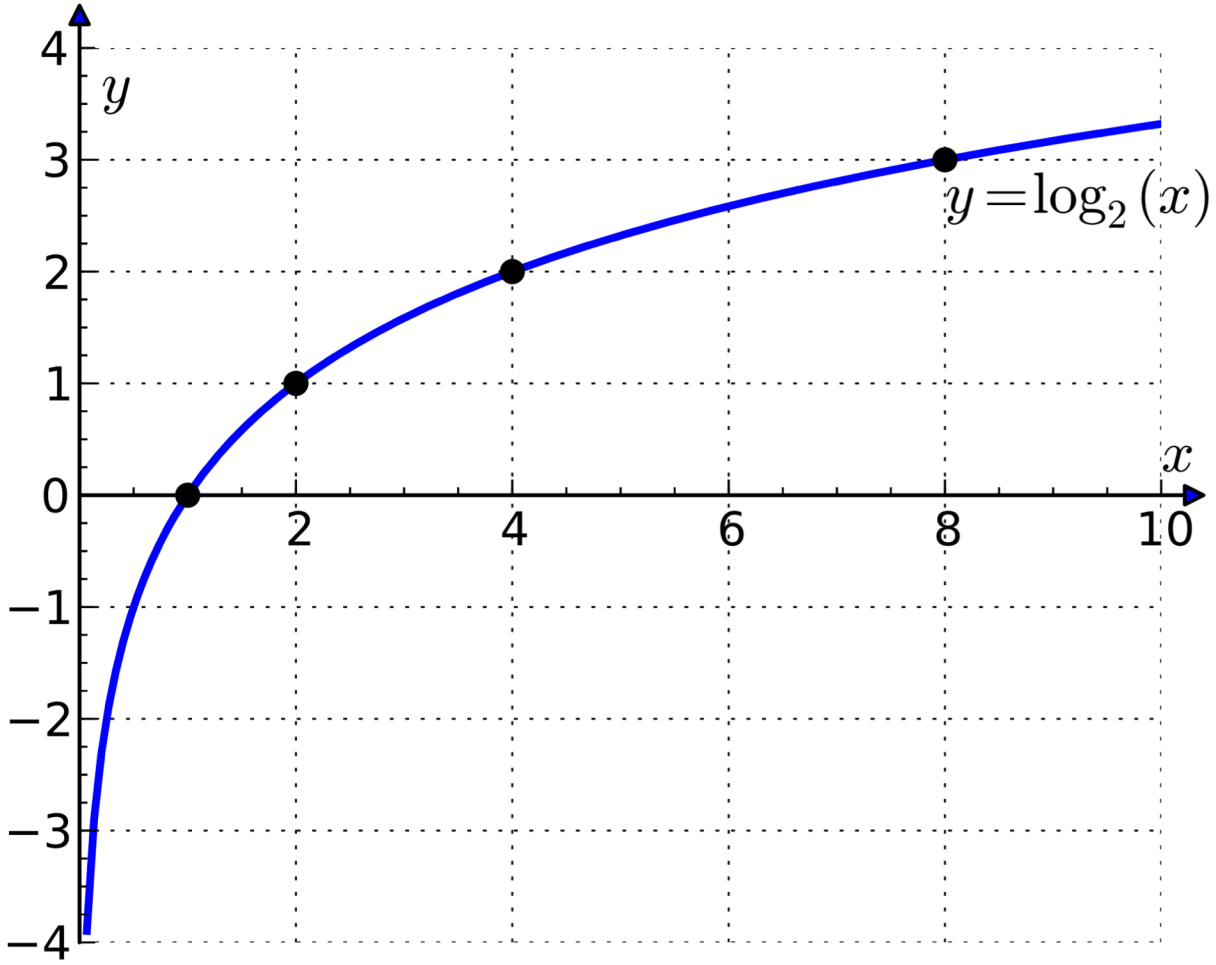
$$\log(xy) = \log(x) + \log(y)$$

More properties:

$$\log(1/x) = -\log(x)$$

$$\log(x^n) = n.\log(x)$$

$Log_2(x)$



Two Important Concepts about Measuring the Information

1. The information about an event x has inverse relationship to the probability of that event.
2. When two independent events happens, the total information about them should be the summation of the two piece of information.

➔ Thus, **information metric** can be defined as:

$$I(X) = \log_2(\frac{1}{p(x)}) = -\log_2(p(x))$$

- Note: It is common to use log based 2, and then the unit of information is in ***bit***.

Thank You!

Questions?