Introduction to Data Science

(Lecture 5)

Dr. Mohammad Pourhomayoun

Assistant Professor

Computer Science Department

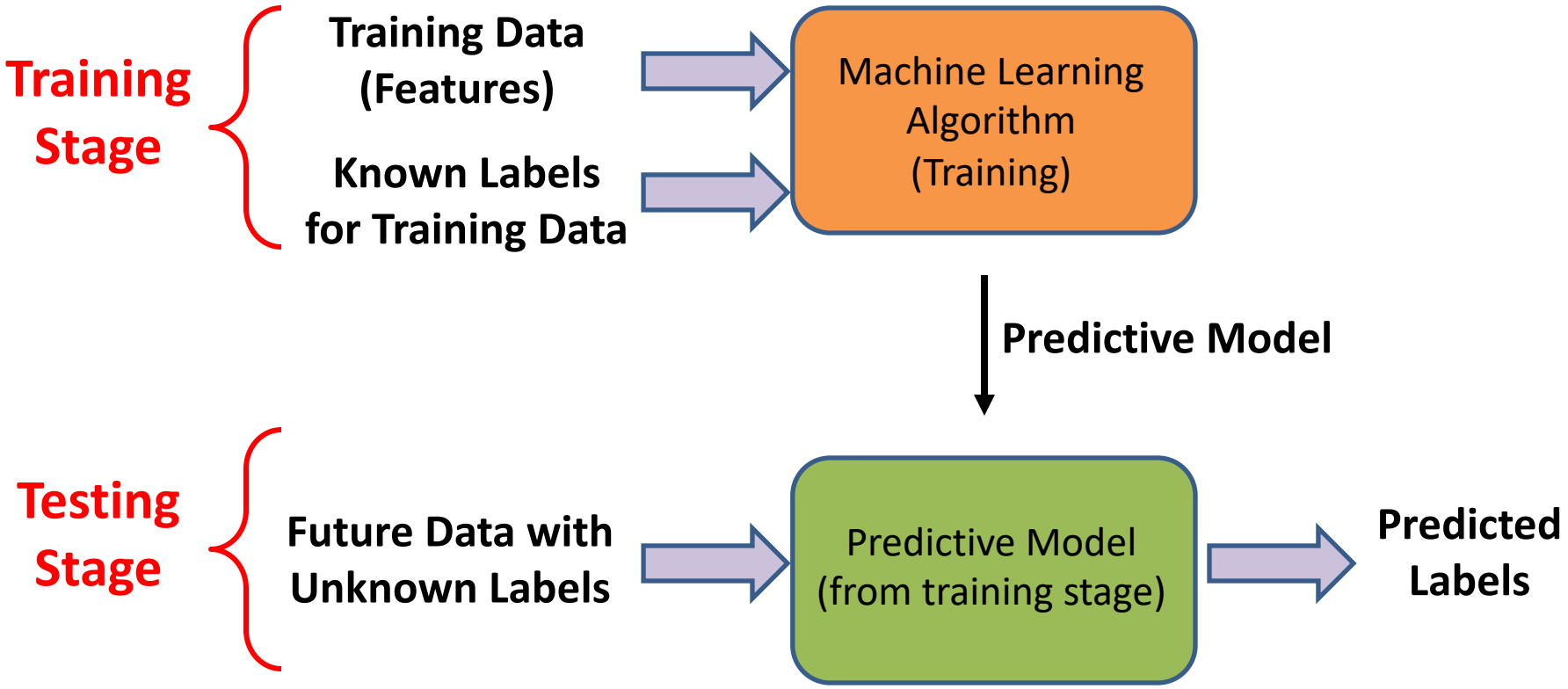
California State University, Los Angeles

Review: What is Machine Learning?

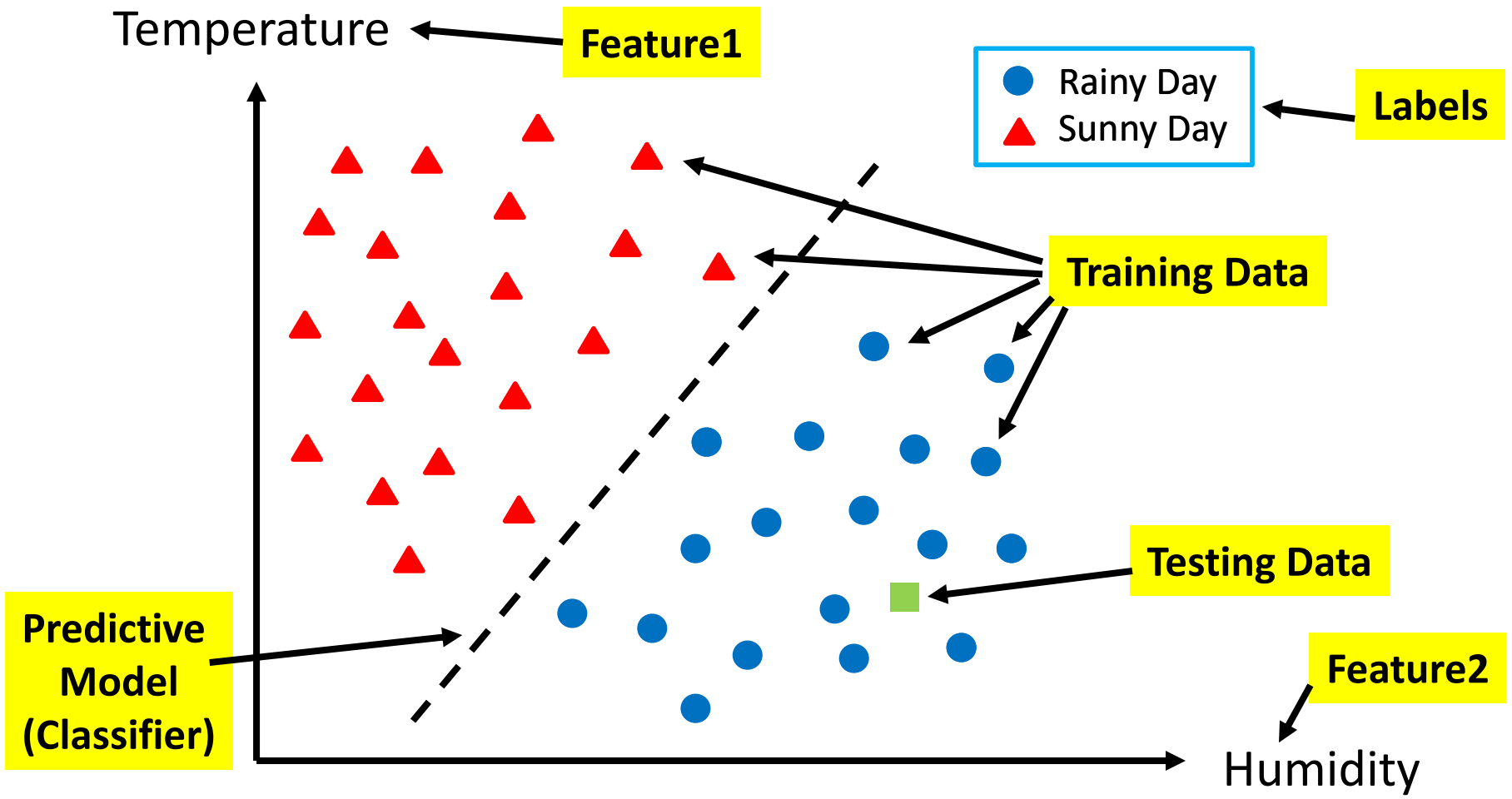
- **A Definition:** Designing and constructing algorithms or methods that give computers the ability to learn from past data, without being explicitly programmed, and then make predictions on future data.

- **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: Supervised Learning: Learning from labeled Data



Review: Classification



Review: Feature Table

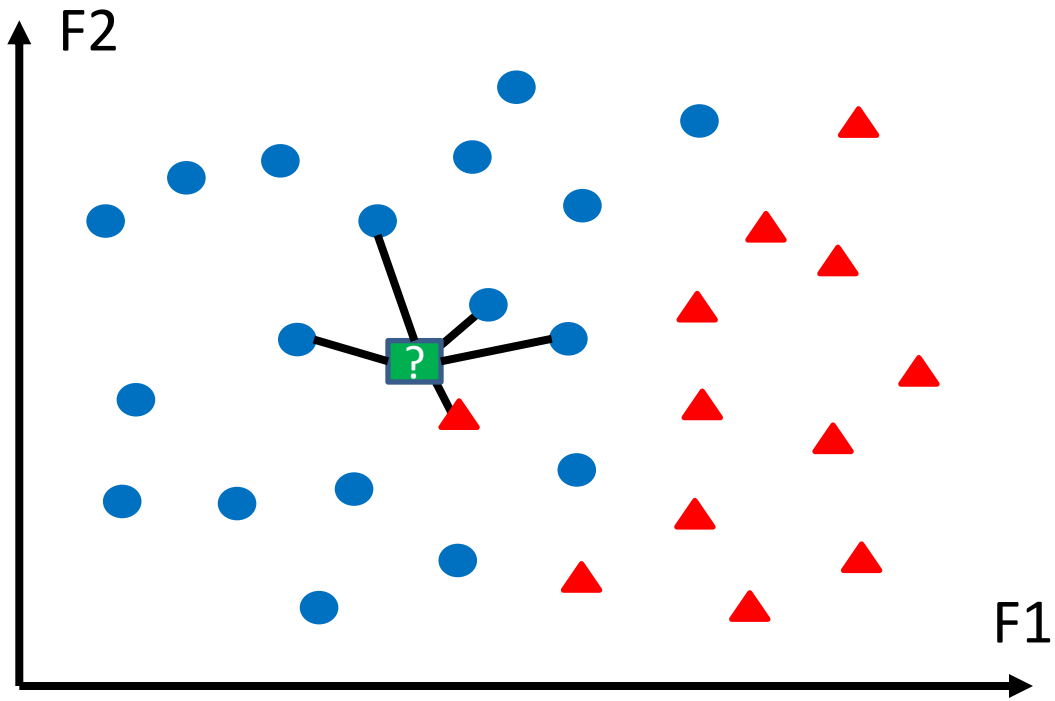
- *Training dataset*: $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ with known label.
- Now, we have a new sample with unknown label: $(\boldsymbol{x}, y=?)$

	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.

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

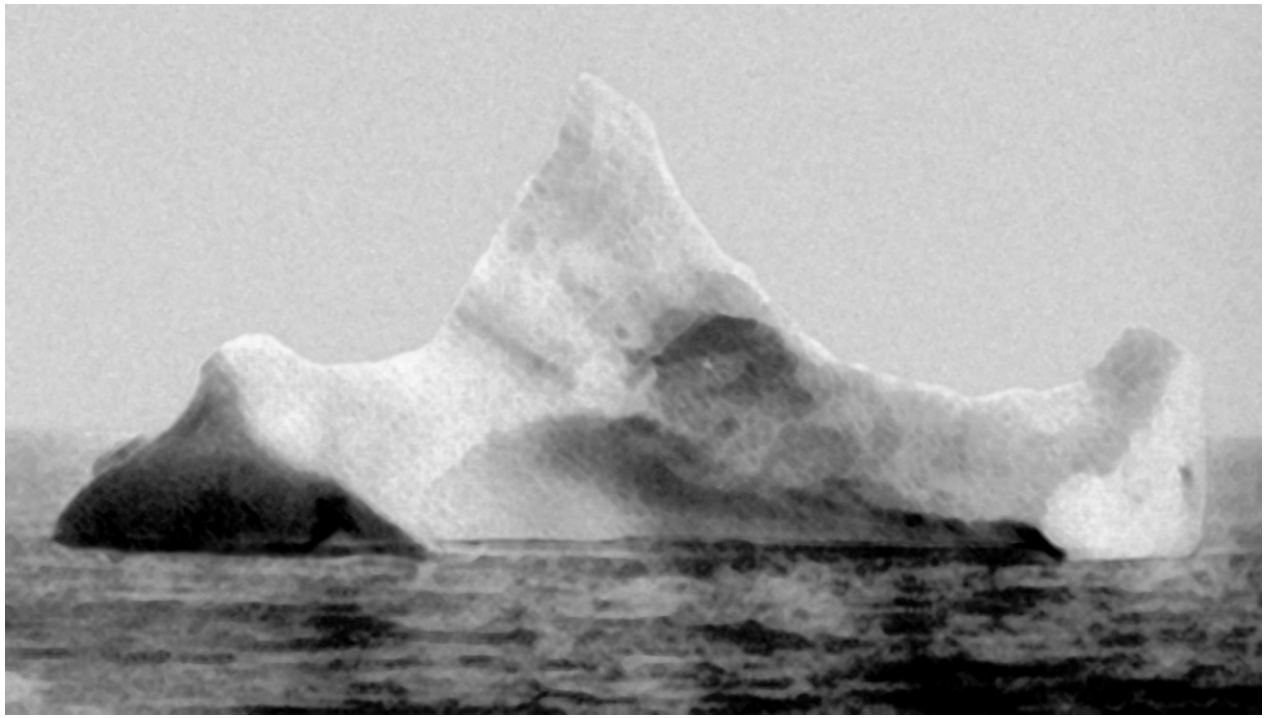
- On April 15, 1912, the Titanic sank after colliding with an iceberg, **killing 1502** out of 2224 passengers and crew.
- One of the reasons that the shipwreck led to such loss of life was that there were **not enough lifeboats** for the passengers and crew.
- Although there was some element of luck 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: 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.

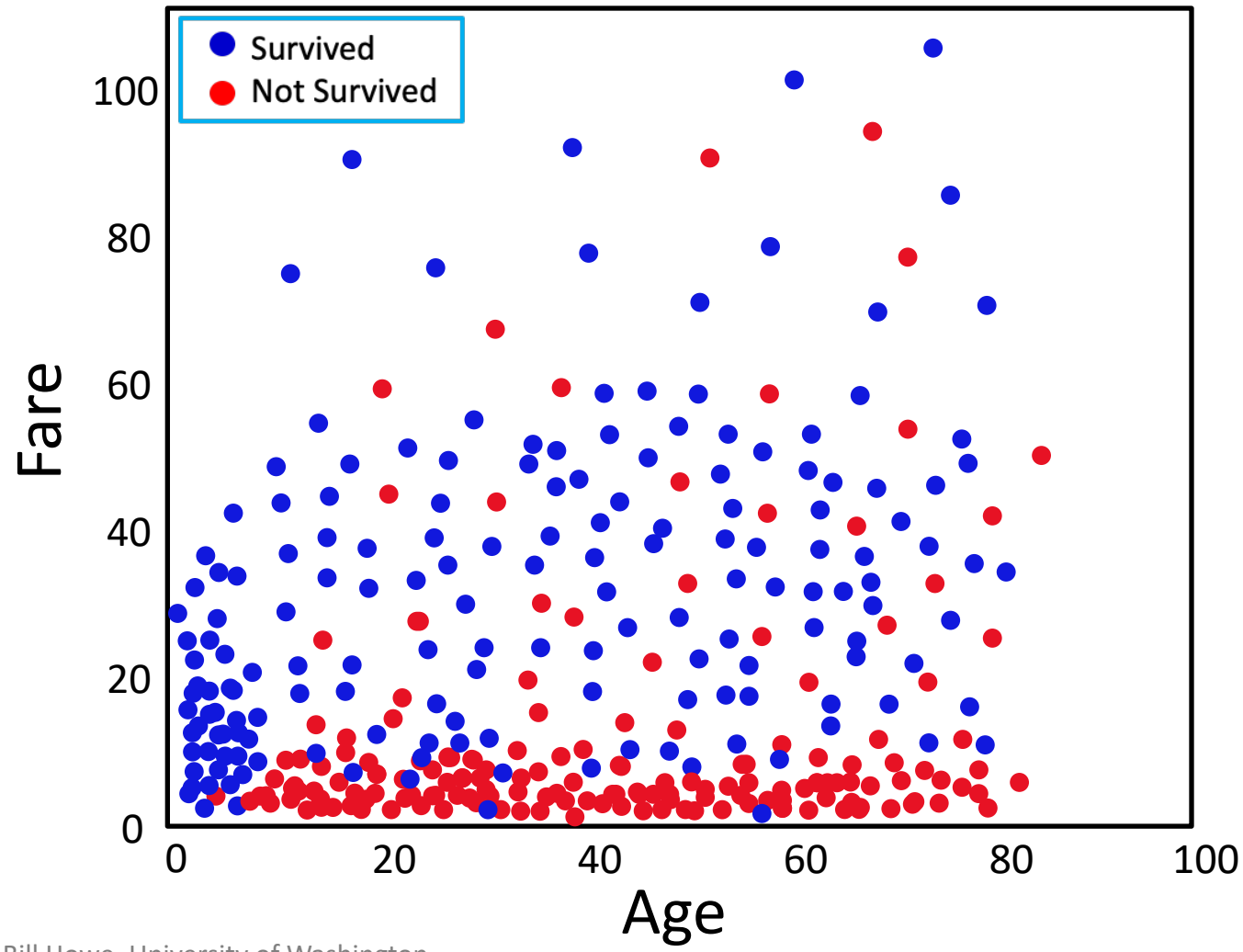
Predict survival on the Titanic

- Passenger list on Titanic[1]:

pclass	age	gender	sibsp	parch	fare	Survival
1	54	male	0	0	52	0
3	2	male	3	1	21	0
3	27	female	0	2	11	1
2	14	female	1	1	30	1
3	4	female	1	2	16	1
3	38	male	0	0	7	0
1	24	female	0	0	71	1
3	22	female	0	0	8	1
1	38	female	1	0	53	1
3	26	male	0	1	8	0
3	???	male	1	2	8	0

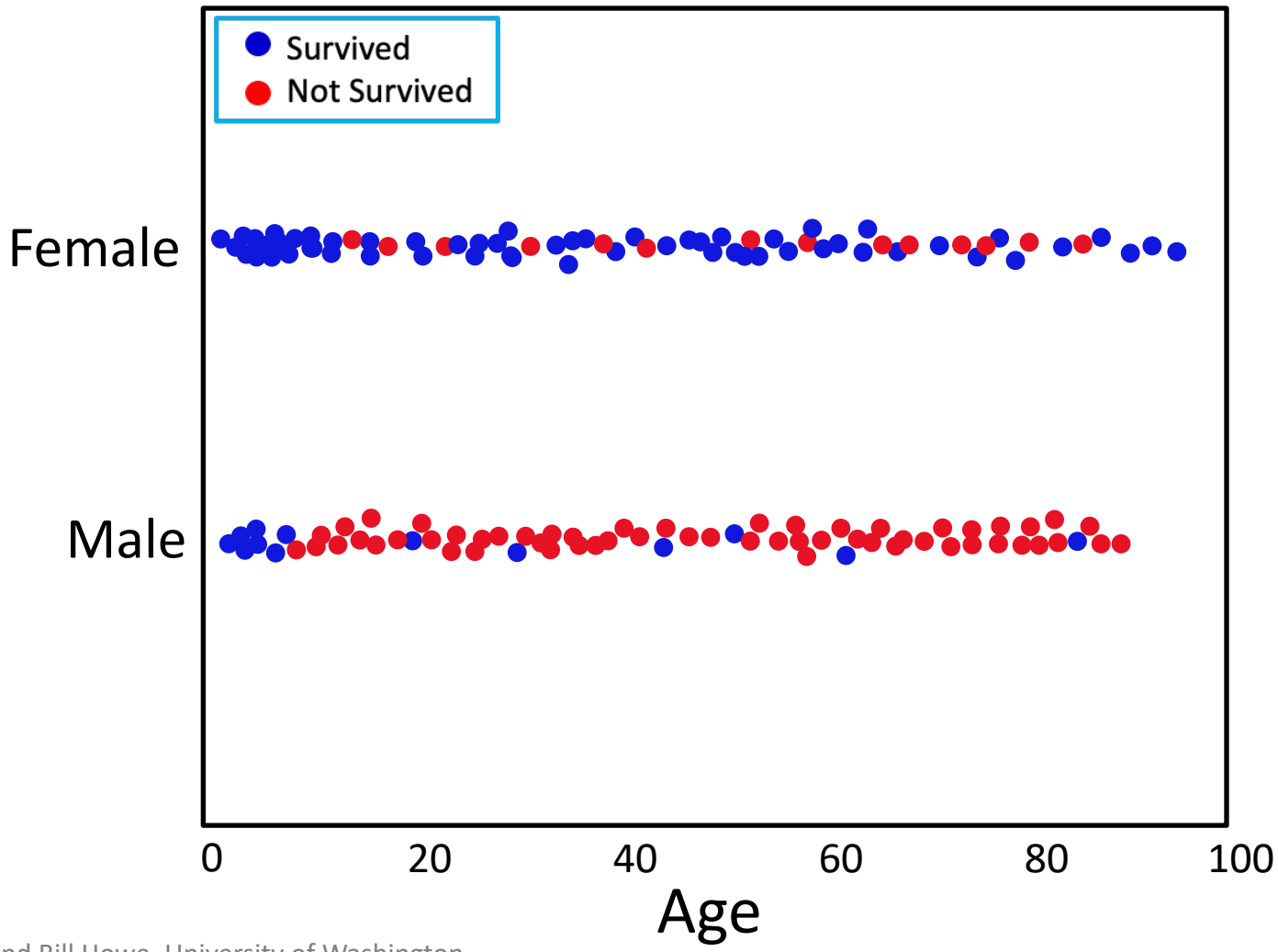
[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

Predict survival on the Titanic



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

Predict survival on the Titanic



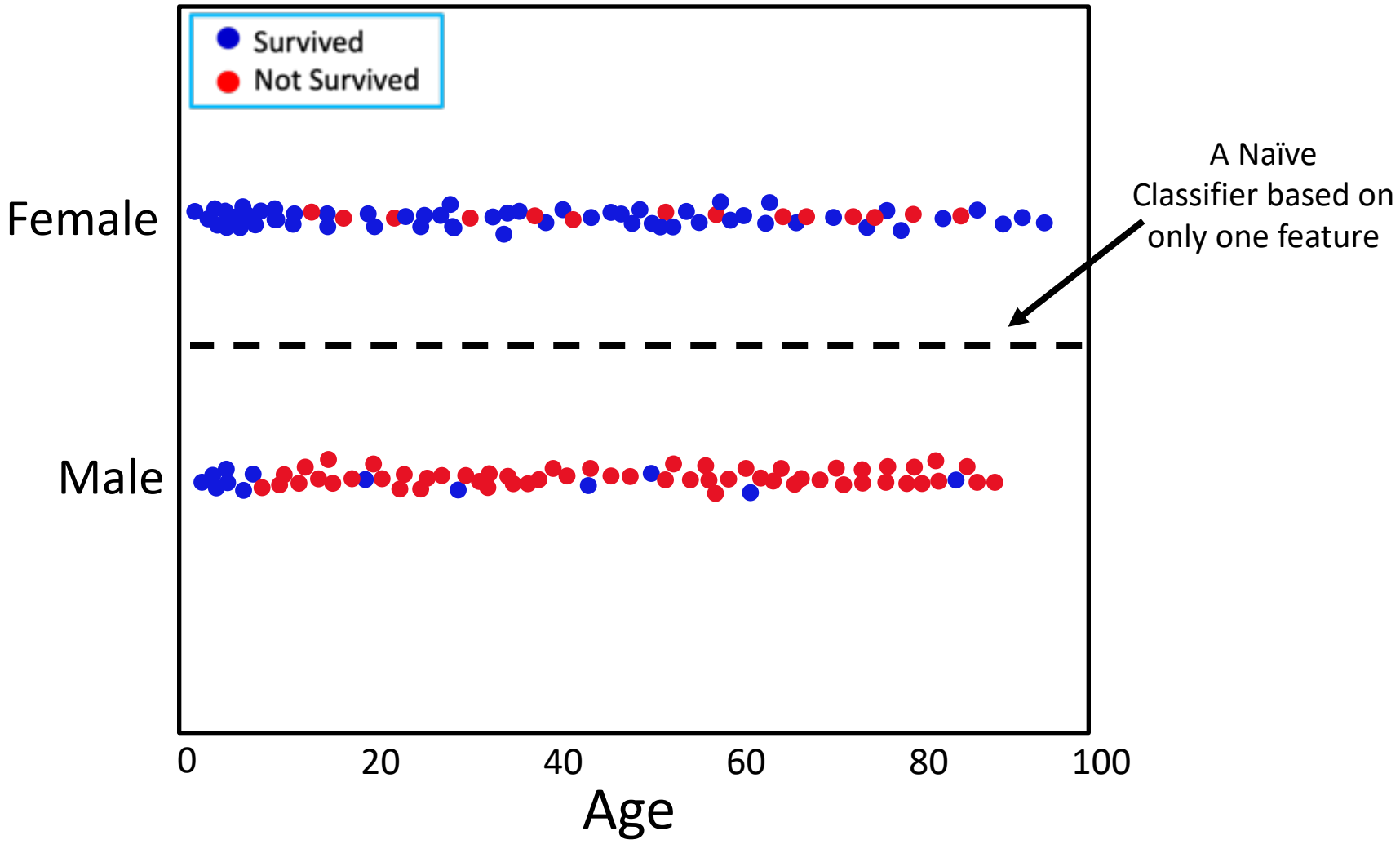
[1]: Ref: Kaggle website, and Bill Howe, University of Washington.

A Naïve Classifier for Titanic

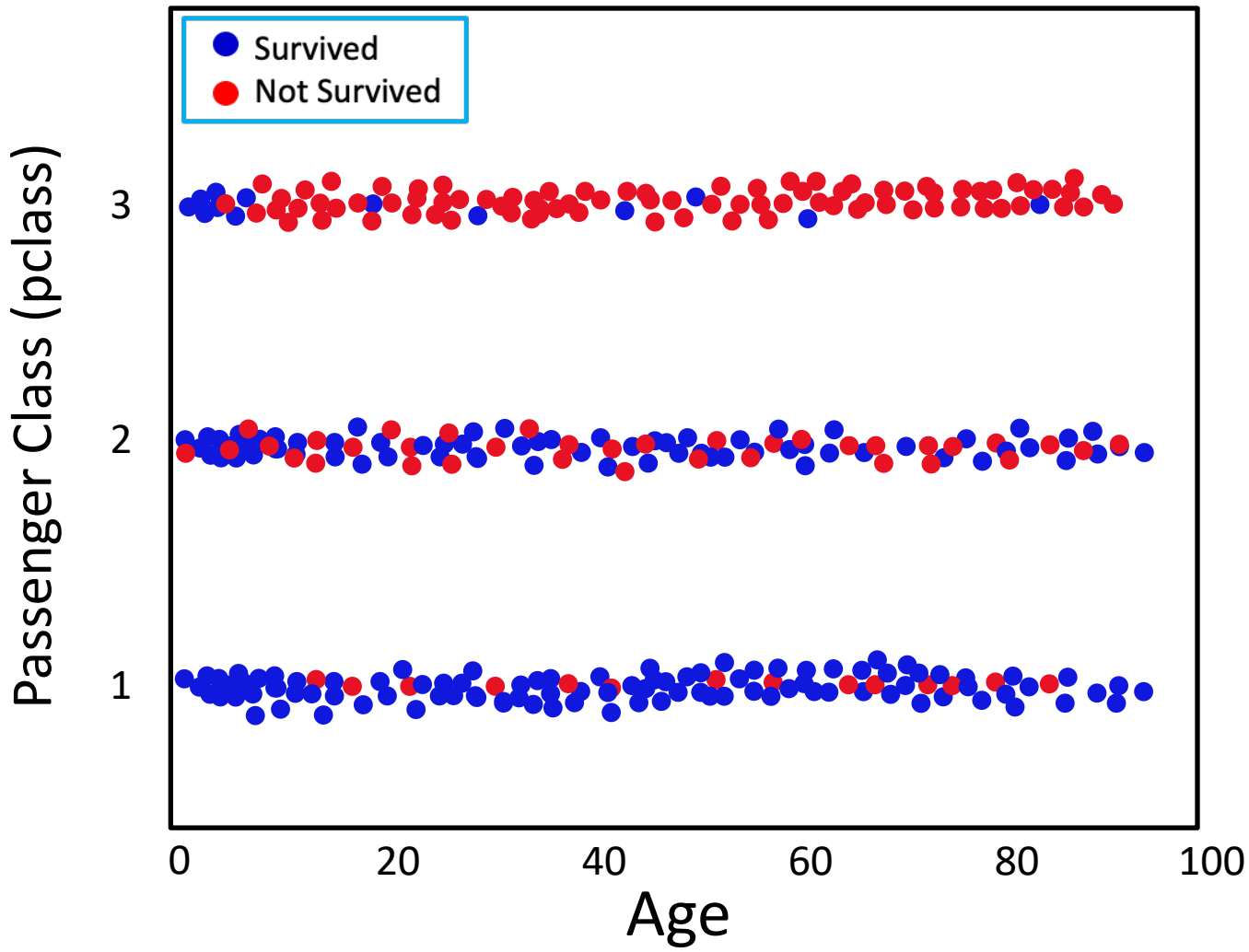
- Making decision based on only **one feature**:
  - Example: Based on Gender: If most females survived, then let's assume every female survives.

- 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



Predict survival on the Titanic

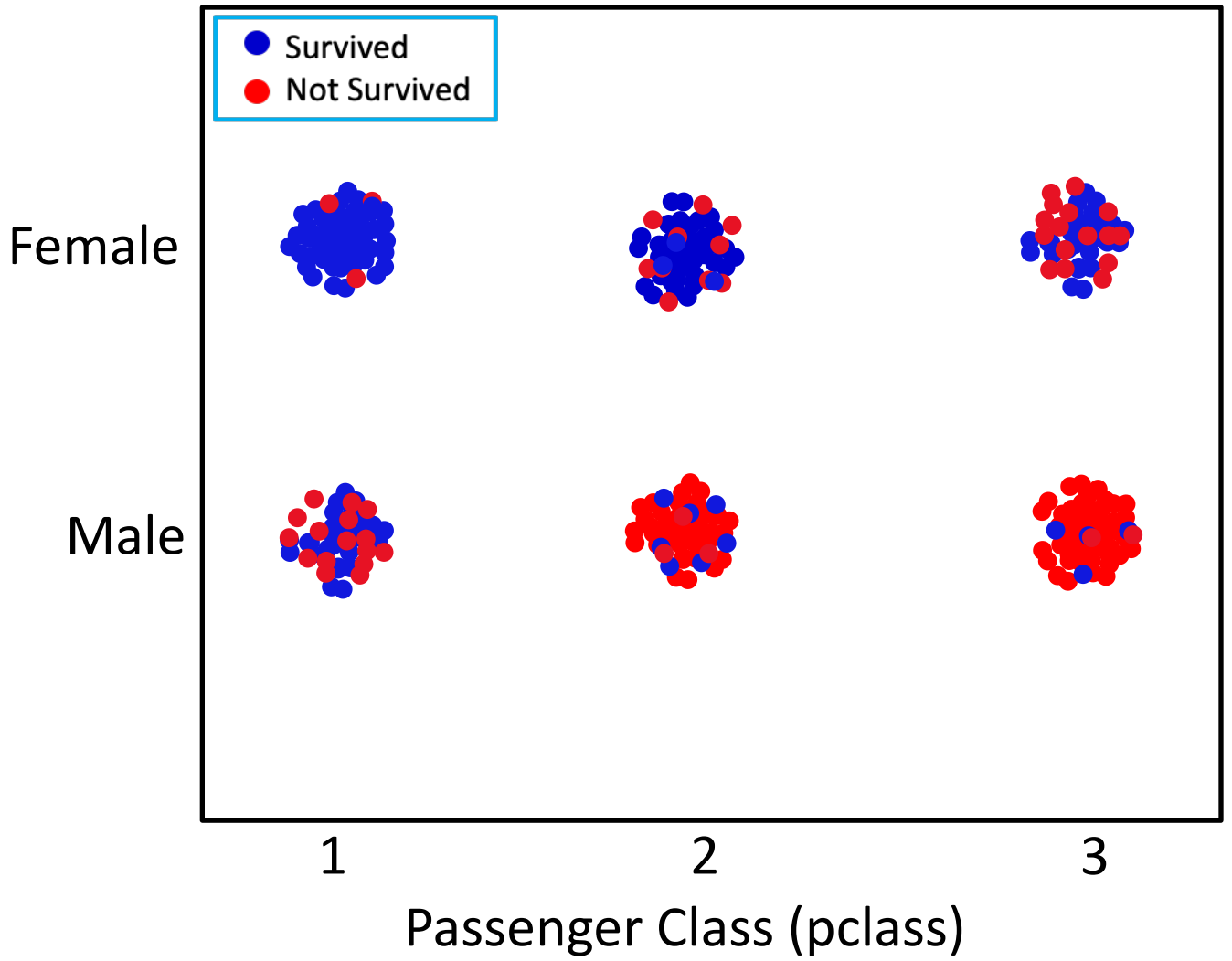


A Naïve Classifier for Titanic

- 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



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

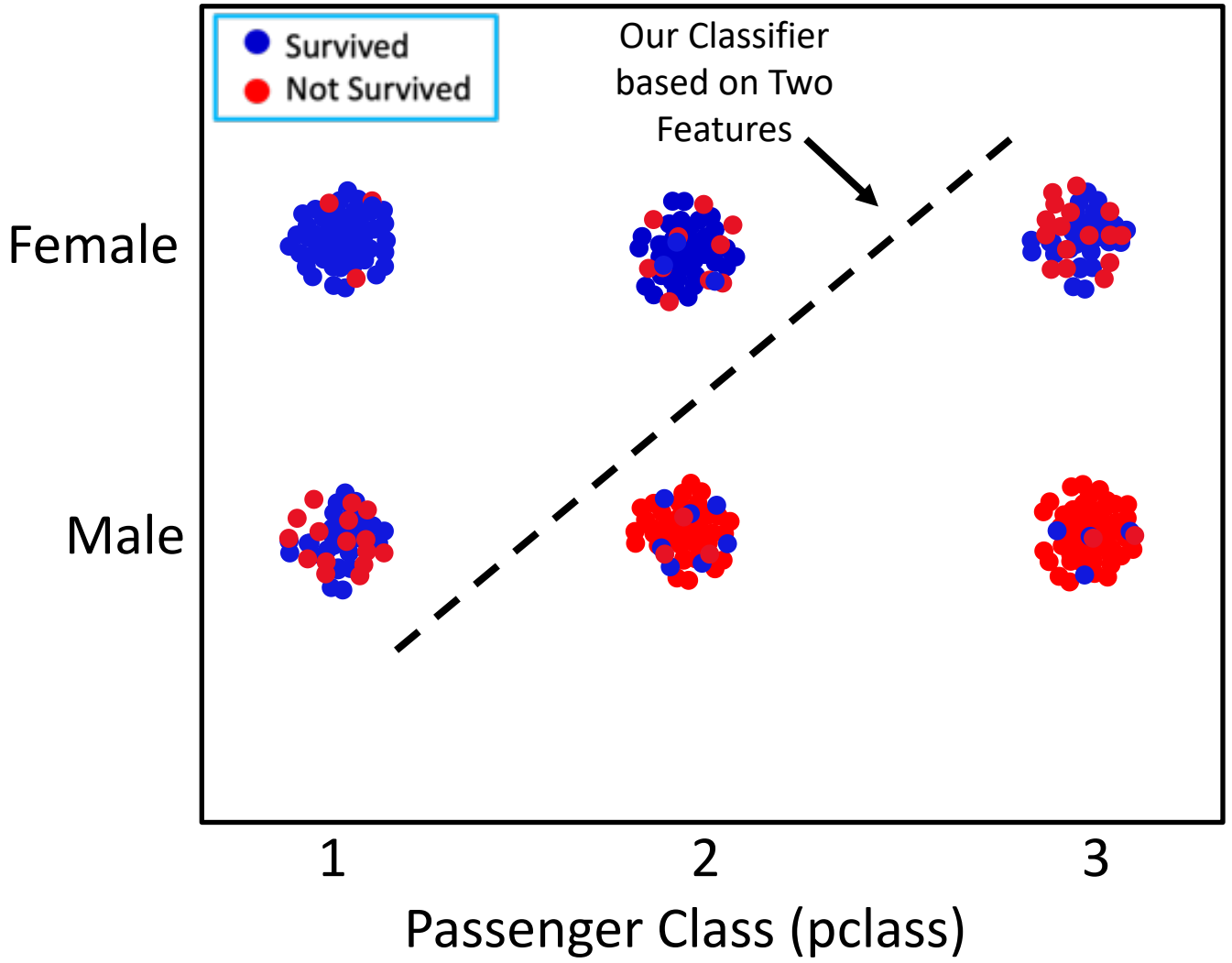
ELSE IF (pclass=‘3’) THEN: Survive ← No

ELSE IF (Sex=‘male’):

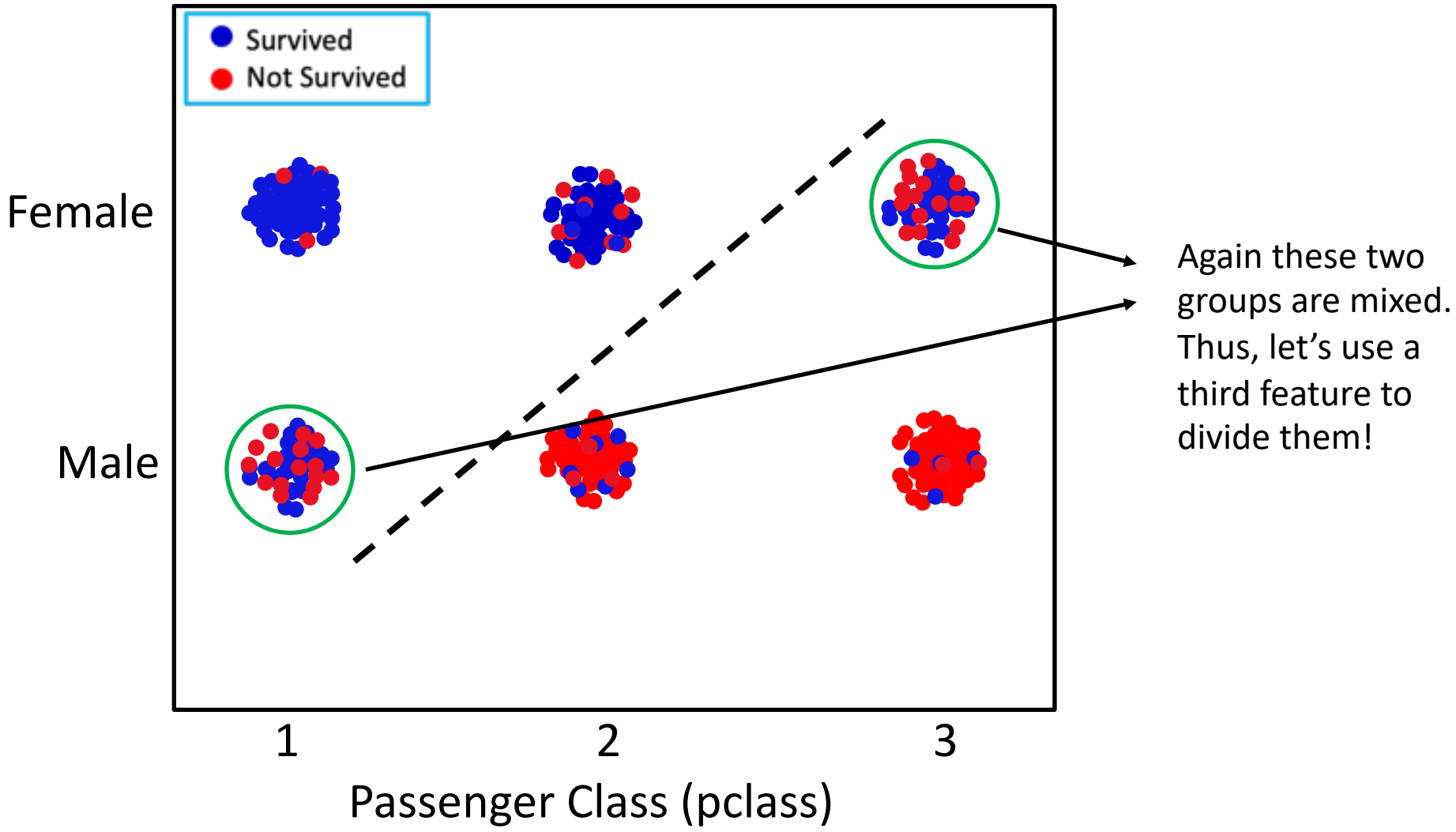
IF (pclass=‘2’) OR (pclass=‘3’) THEN: Survive ← No

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

Predict survival on the Titanic



Predict survival on the Titanic



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'):

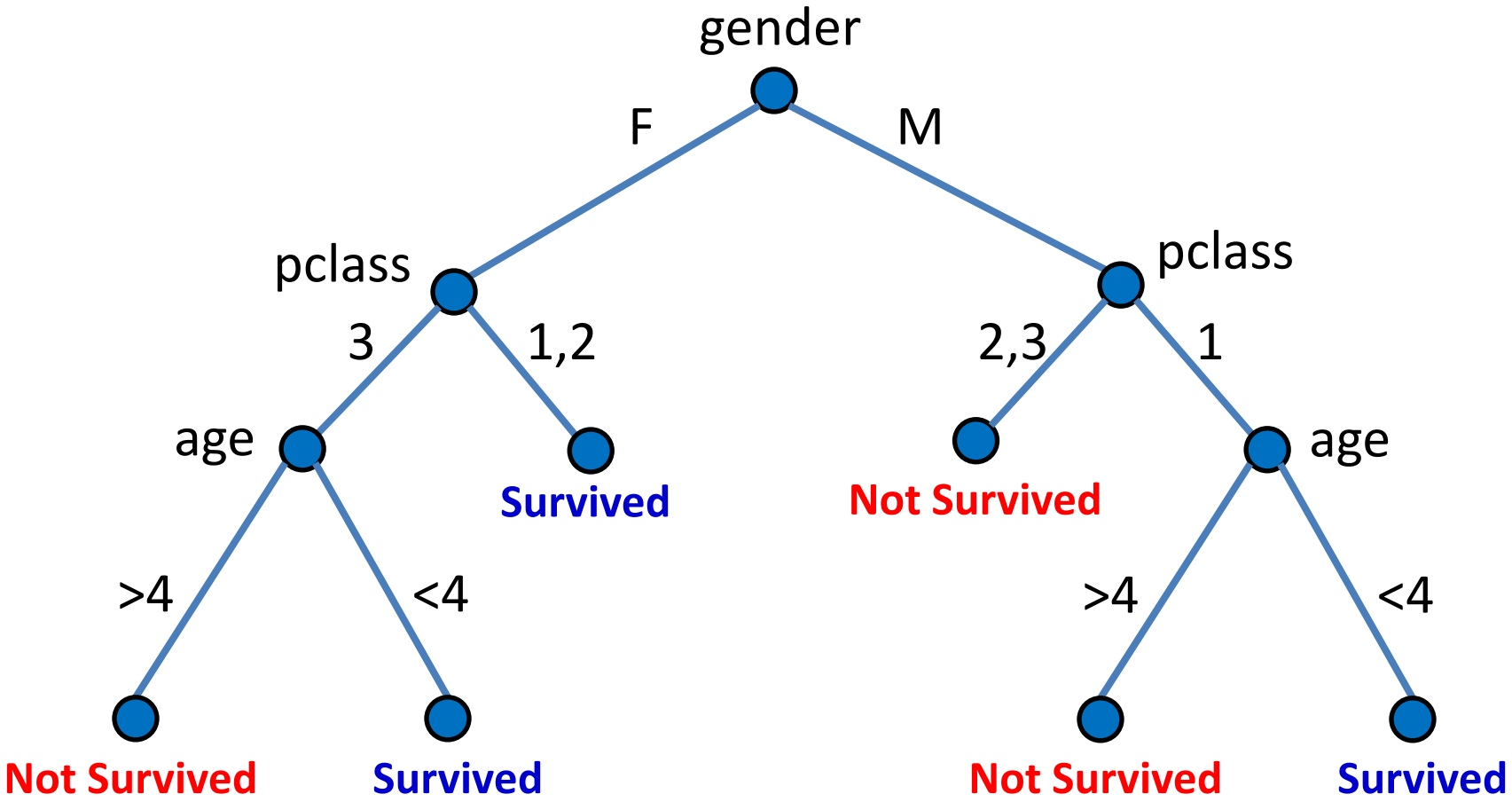
IF (pclass='2') OR (pclass='3') THEN: Survive ← No

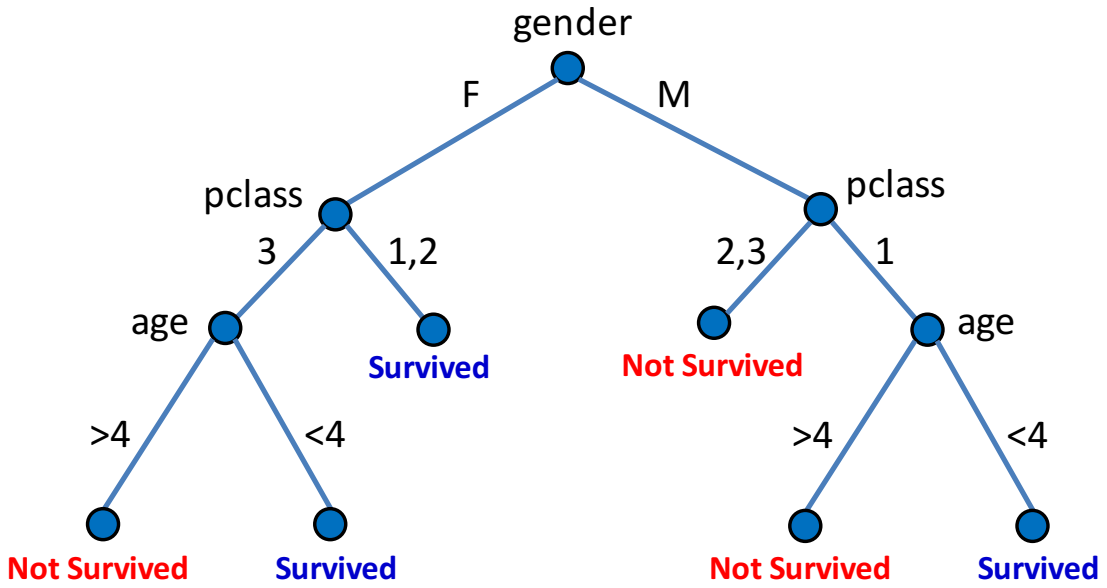
ELSE IF (pclass='1'):

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?

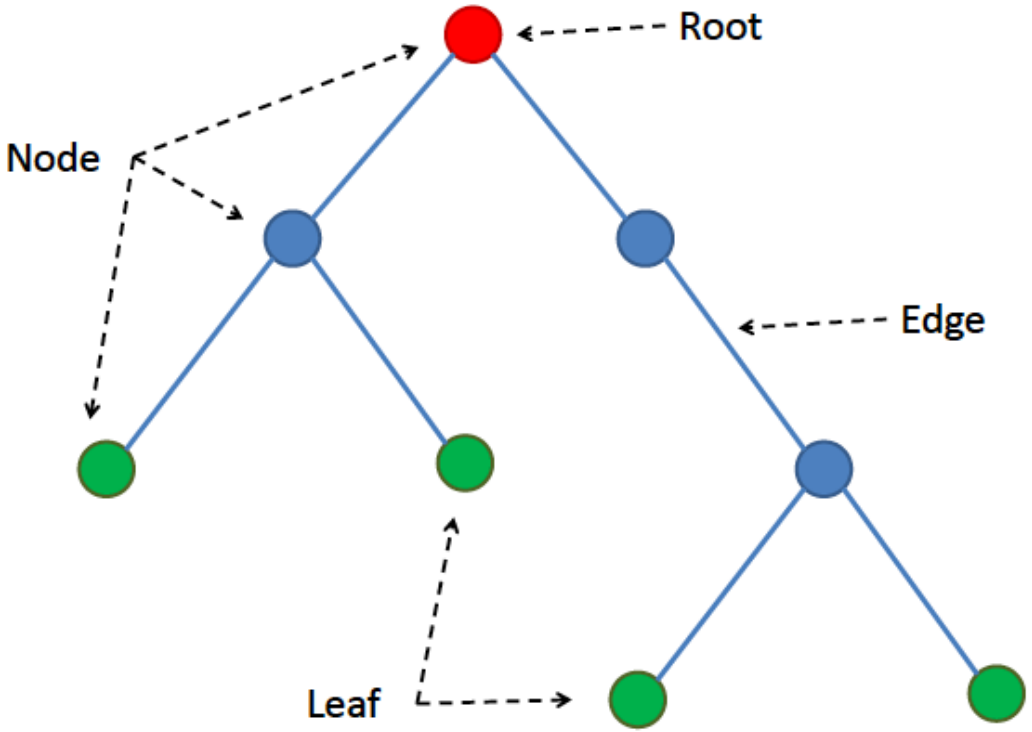
Qustion6: Kevin was 30 years old. Did he survive?

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

Important Note

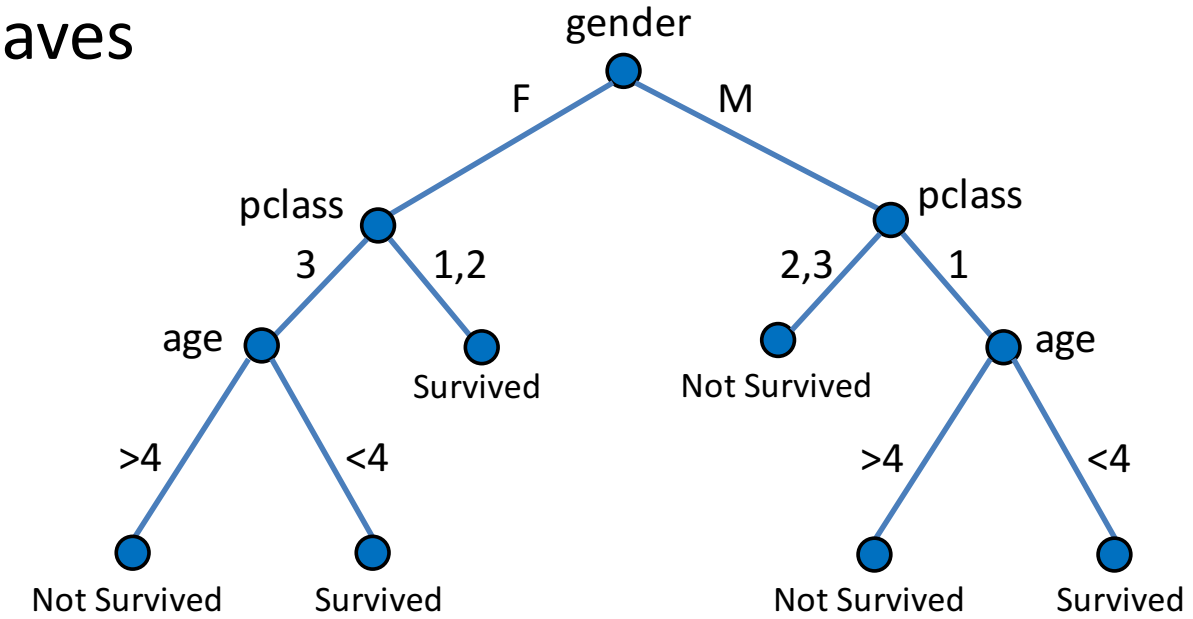
- **EVERY** prediction model has some levels of error!
- There is rare to have a predictor with 100% accuracy!

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**.

Thank You!