Impacts of AI: COMP3800-03
The Data Science Journey
Wentworth Institute of Technology

Professor Armen Pischdotchian

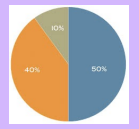
Abraham Wald and the missing bullet holes

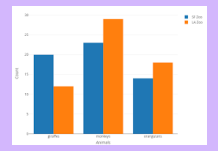
Abraham Wald was a Hungarian mathematician (1902-1950) founded the field of statistical sequential analysis. He spent his research years at Columbia University.

About Data Types

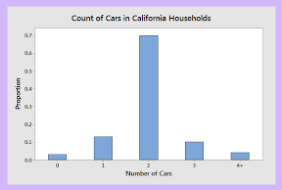
Data

- Qualitative or Categorical
  - Nominal
    - No order: male/female
    - Frequency, proportion, percent

    Pie chart 

    Bar chart 
  - Ordinal

    Ordinal: ordering small/medium/large

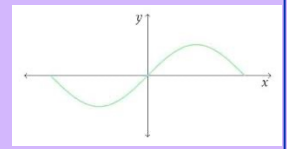
    

- Quantitative or Numerical (regression)

  
  - Discrete
    - Has separate, indivisible values
    - No value can exist between two neighboring values
    - Number of rooms

  - Continuous
    - Has an infinite possible number of values
    - Every interval is divisible into an infinite equal parts
    - Height of a person: intervals, ratio

DATA IN THE WORLD TODAY

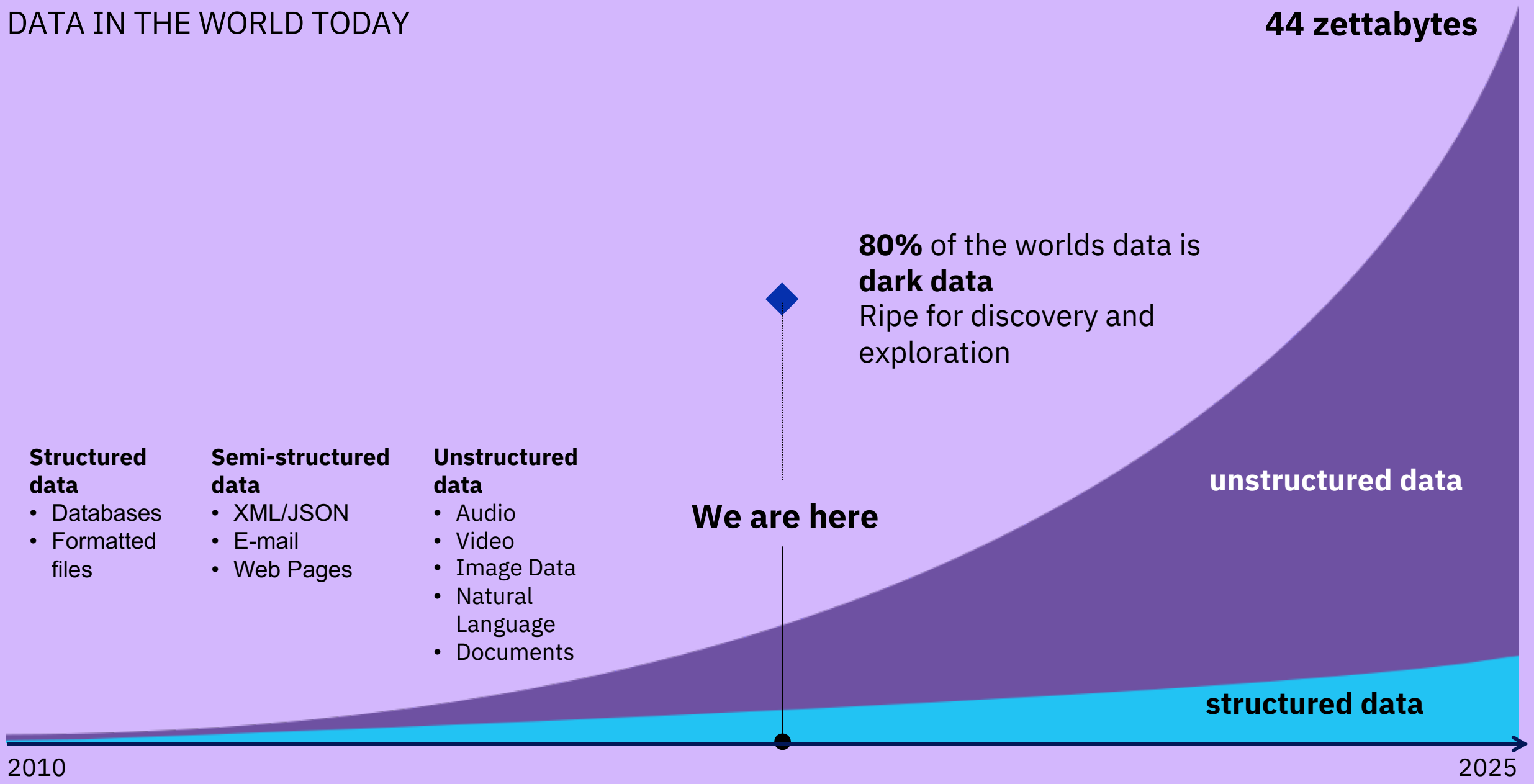
44 zettabytes

80% of the worlds data is **dark data**
Ripe for discovery and exploration

Structured data
- Databases
- Formatted files

Semi-structured data
- XML/JSON
- E-mail
- Web Pages

Unstructured data
- Audio
- Video
- Image Data
- Natural Language
- Documents



DATA SCIENCE DOMAINS

Data, Science and Technology are intrinsically connected

Science

- Linear Algebra
  (VECTORS, MATRIXES,
  EIGEN VALUES, EIGEN VECTORS)
- Statistics
  (DESCRIPTIVE & INFERENTIAL)
- Probability

Technology

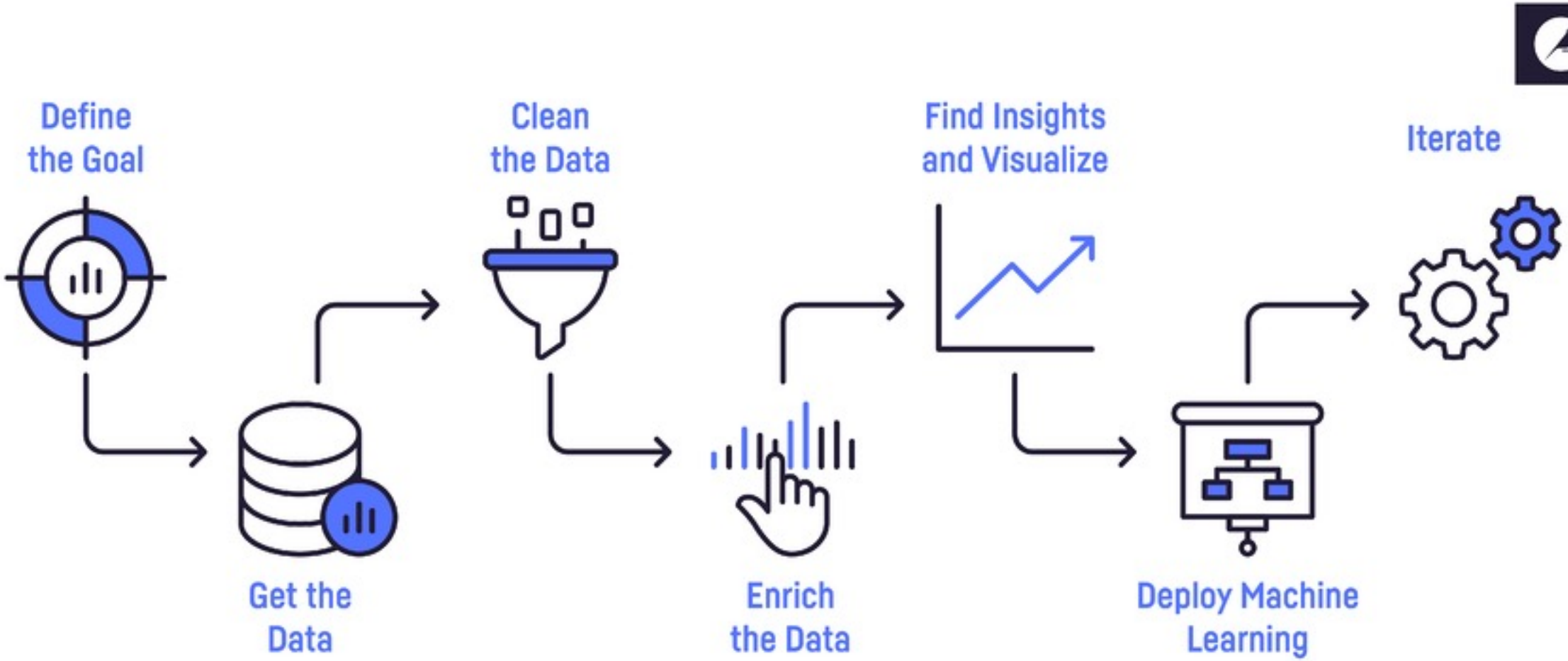
- Business Intelligence
- Data Mining
- Big Data
- Predictive Analytics
- Machine Learning

Data

Current state of the target data:

- *Structured* - DATA BASES, FILES, LIBRARIES, DATASETS, ETC.
- *Unstructured* - VIDEOS, MOVIES, NEWS PAPERS, BOOKS, ARTICLES, BLOGS, MUSIC, IMAGES, ETC.

Data Science Approach



What is so confusing about the Confusion Matrix?

How important is Accuracy?

Your client may ask:

"How do I get the most accurate model?"

The data scientist responds by asking:

"What business challenge are you trying to solve using the model?"

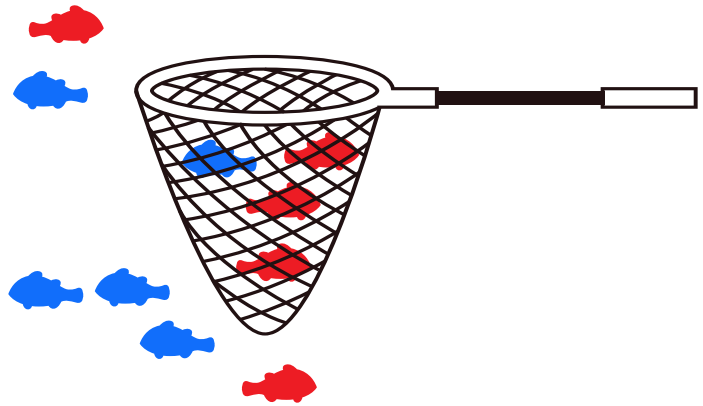
Accuracy may not be the most essential metric for data scientists to obtain...

What about precision and recall?

When recall is more important than precision

3 Relevant data points (red fish)

1 Irrelevant data points (blue fish)



The search has retrieved 3 relevant documents out of a total of 5 relevant data points from the data set and 1 irrelevant document.

Recall (True Positive rate) = 3 / (3+2) = 0.6

Precision = 3 / (3+1) = 0.75 (the blue fish is not relevant)

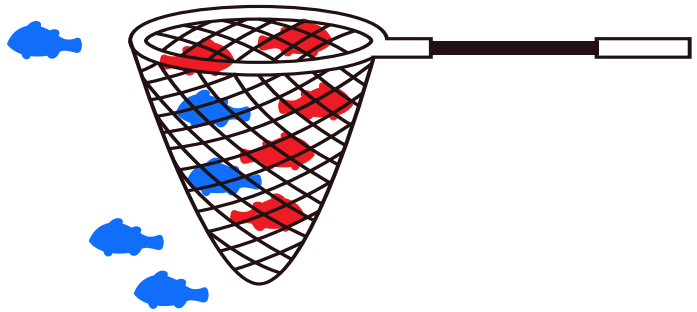
The case of 100% recall and low precision

5 Relevant data points (red fish)

2 Irrelevant data points (blue fish)

In some models, this is the preferred scenario even though there may be some irrelevant data points with a high score.

The algorithm team will then work on increasing the precision of this system.



The case of 100% precision and low recall

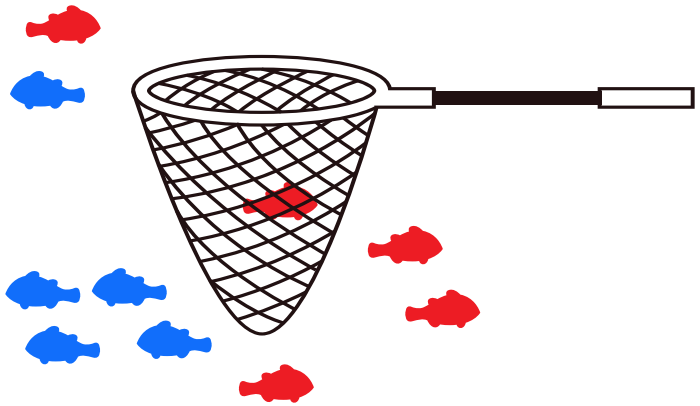
1 Relevant data points (red fish)

0 Irrelevant data points (blue fish)

Zero false positives, 100% precision — no blue fish in the net

But there are many false negatives — many red fish in the sea

There are potentially many data points that we will never consider. Perfect precision with poor recall is of no value to any ML model.



Another view of True Positive and False Positive

True Positive

When the actual classification (for example red fish presence in the net), then how well did the prediction depict that based on the holdout data

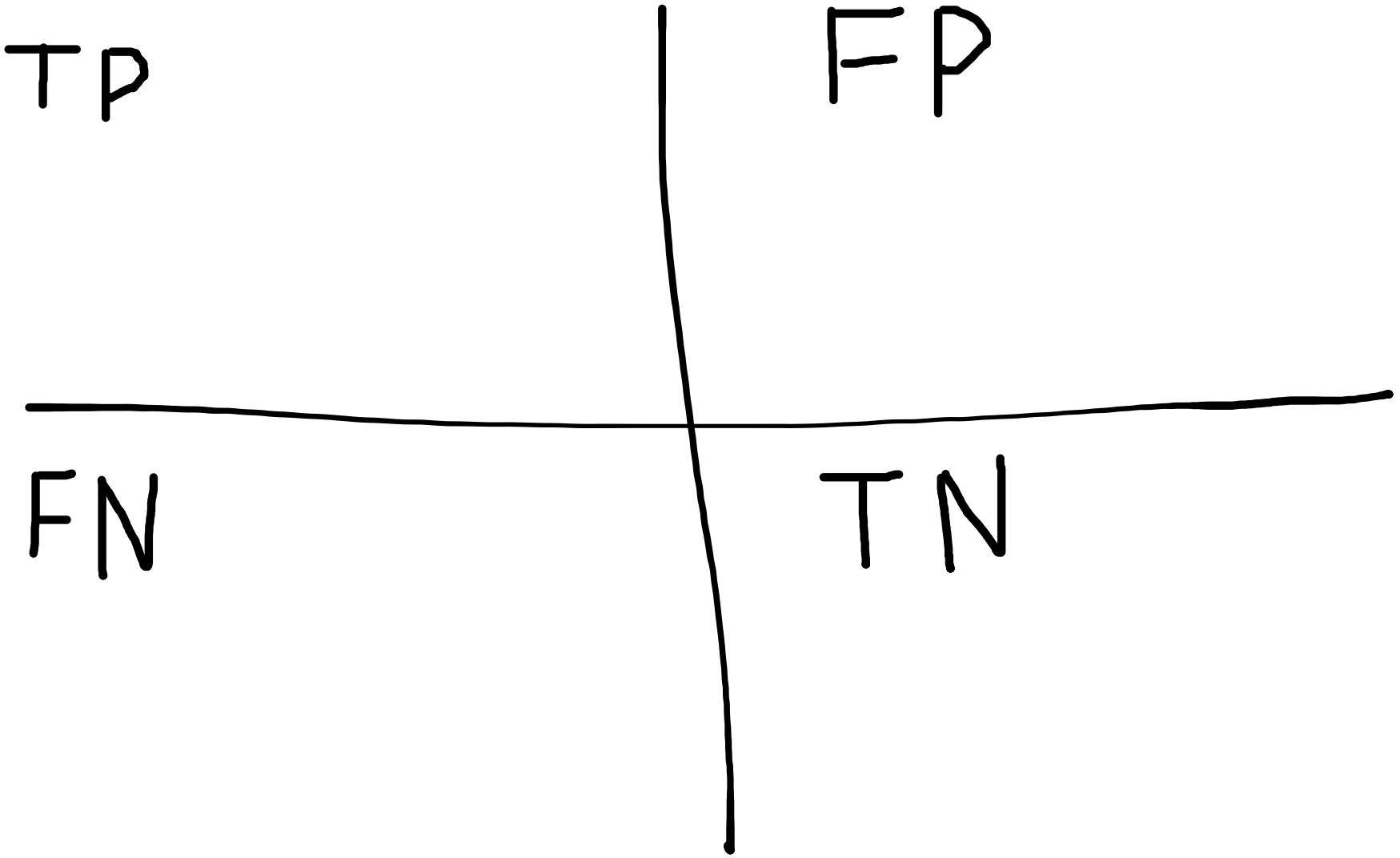
False Positive

When the actual classification (for example blue fish presence in the net), then how often does the classifier, INCORRECTLY predict red fish when it should’ve been blue fish.

Both TP and FP range from 0 to 1; ROC curve visualizes all possible thresholds

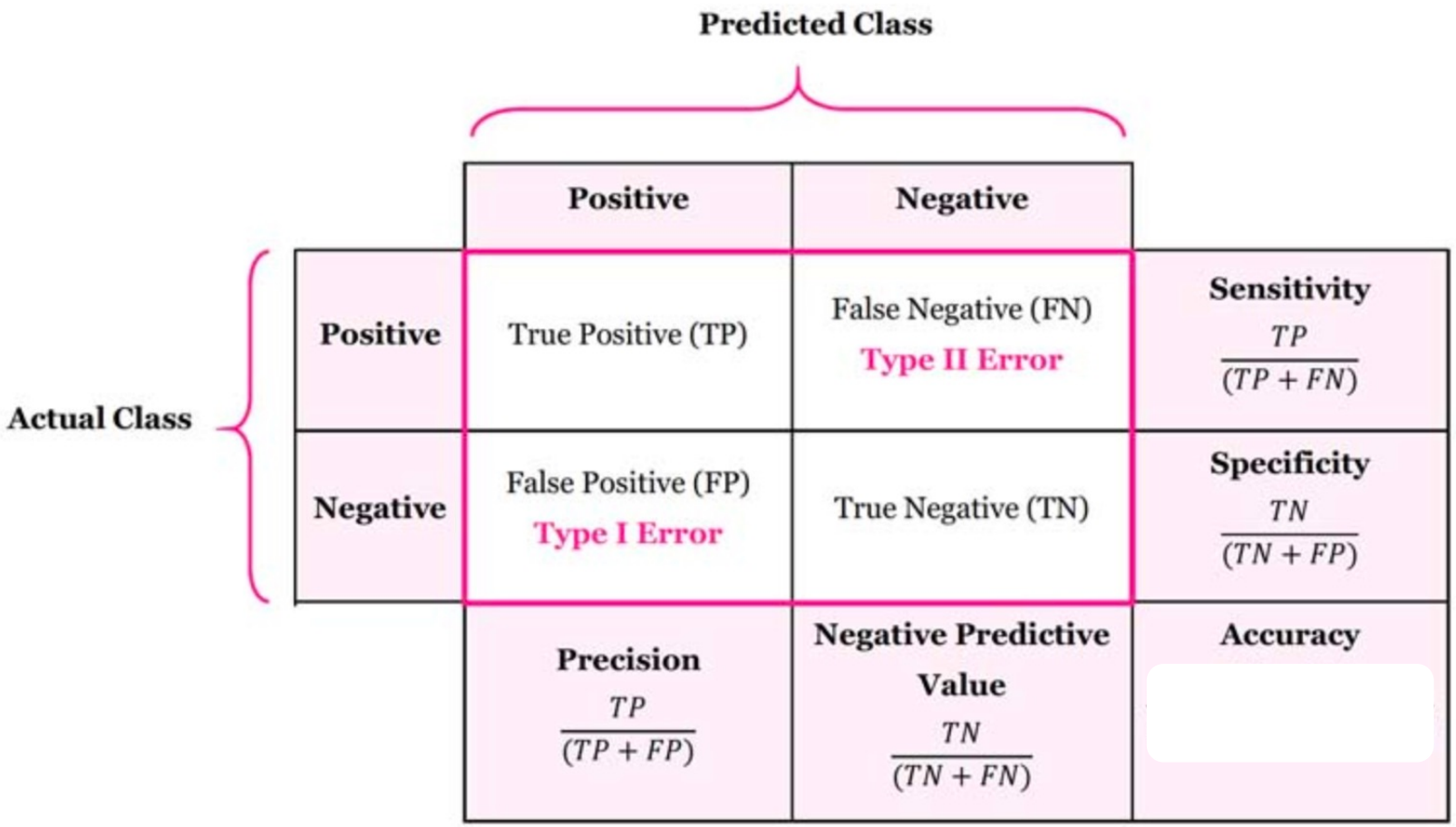
The Boy Who Cried Wolf

Wolf = positive class
No wolf = negative class



Deriving Precision, Recall and Accuracy from the confusion matrix

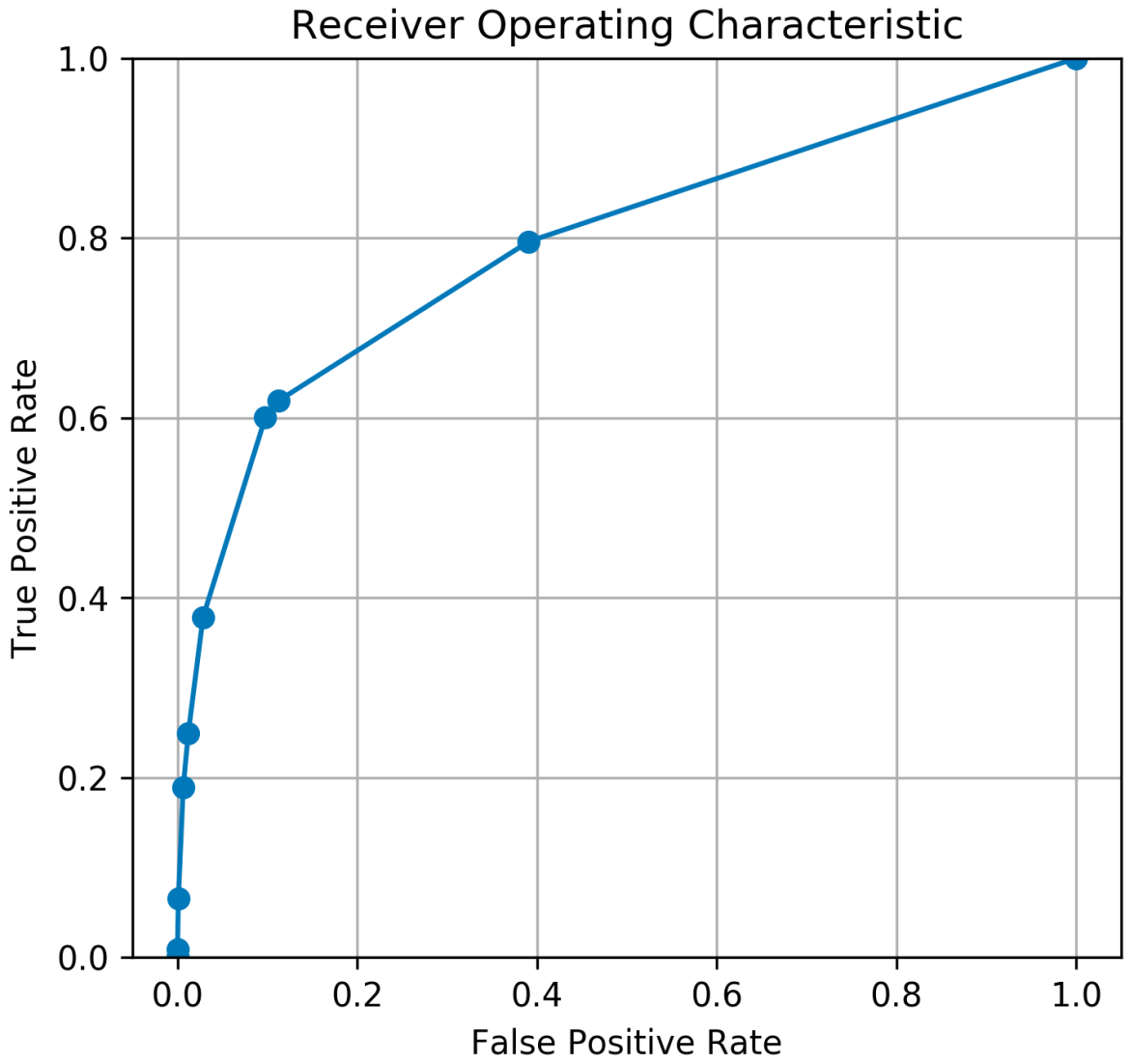
Accuracy = ----------



May the Best Model Win

The ***True Positive Rate*** is plotted against the ***False Positive Rate*** in order to see the tradeoff between the two as thresholds are adjusted.

This plot is called the ***Receiver Operating Characteristic.***

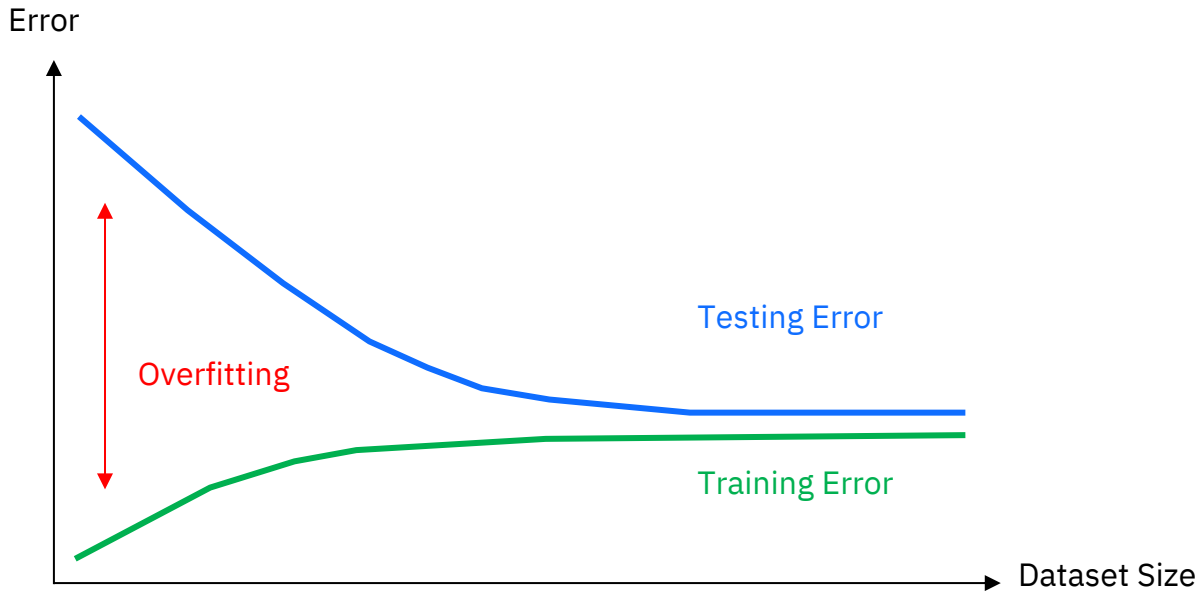


Dangers of Overfitting Data

The biggest risk is overfitting

- A model that works well on training data but performs badly on new data

Trick is to select a model without knowing the data it will be applied to



Example of overfitting

We want to predict if a student with a GPA of 3.3 will be admitted to Wentworth Institute of Technology in Boston

Assume we train a model from a dataset of 5,000 students and their outcomes.

Next, we try the model out on the original dataset, and it predicts outcomes with 99% accuracy… wow!

But now comes the bad news.

When we run the model on a new (“unseen”) dataset of student admissions, we only get 50% accuracy… uh-oh!

Our model doesn’t *generalize* well from our training data to unseen data.

This is known as overfitting, and it’s a common problem in machine learning and data science.

IBM 5 in 5 Predictions