Data Imbalance

CS1090A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

Photo: Liuyixin Shao
Olympic National Park, WA

Outline

- Motivation
- Random Forest
- Variable Importance
- Missing Data (again)
- Class Imbalance
- Tree building algorithms

Class Imbalance

Training a RF (or any machine learning model) on an imbalanced dataset can introduce unique challenges to the learning problem.

Recap: F1-score

Accuracy is a great measure but only when you have **balanced datasets** (false negatives & false positives counts are close),

ALSO, **accuracy** is a good measure when **false negatives** & **false positives have similar costs.**

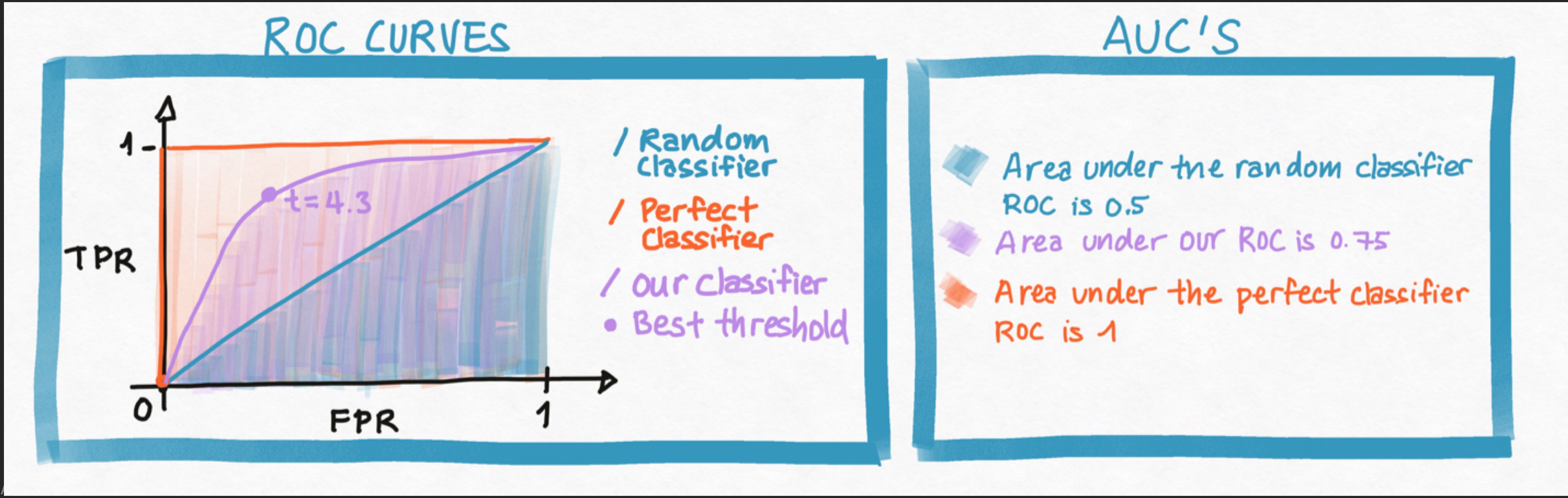
In the case of imbalance datasets, F1-score is a better metric

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

Recap: Area Under the ROC curve

If the costs of false negatives & false positives are different, the ROC curve allows us to find the classification threshold that gives the best trade-off between FP rate and TP rate which we need in this case.

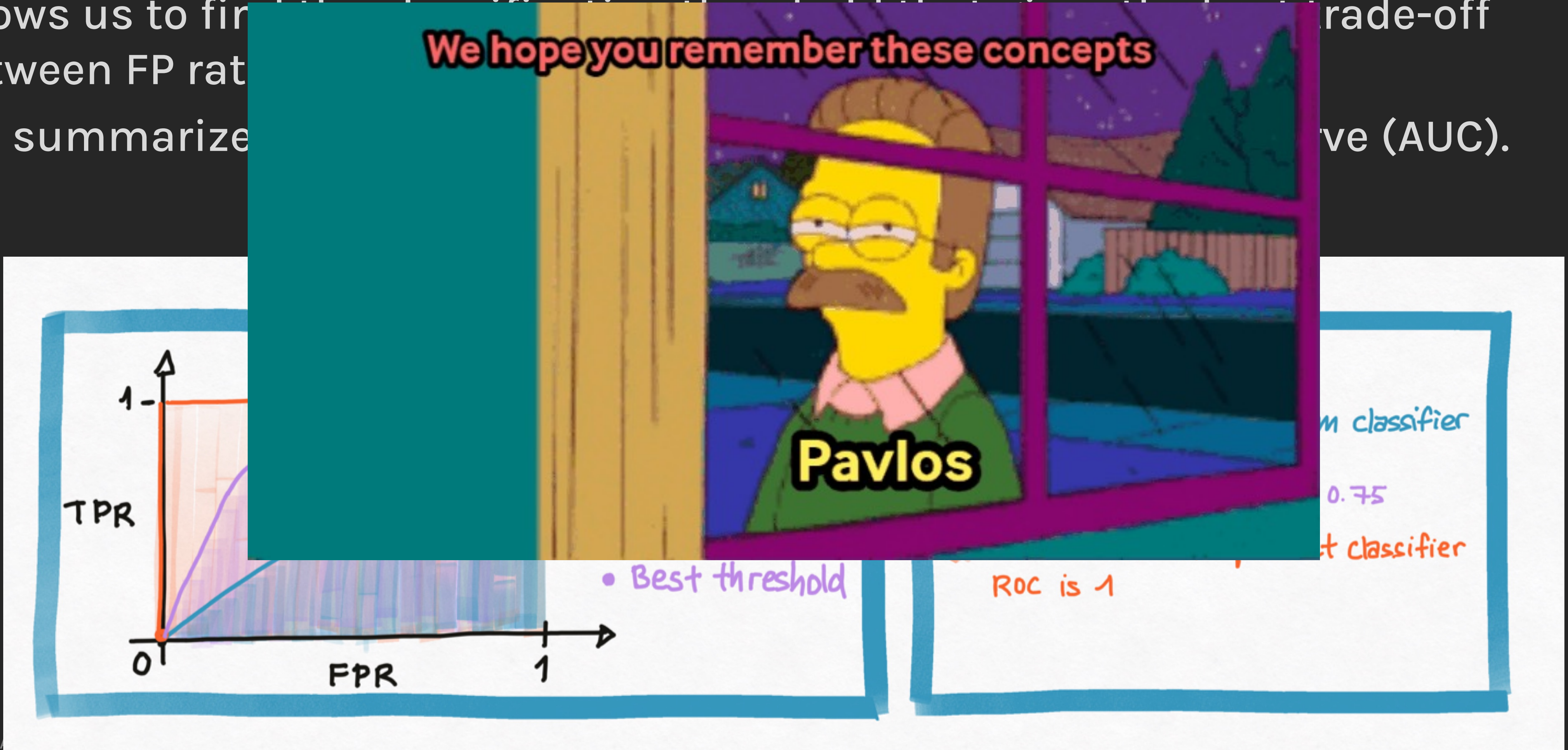
We summarize the ROC by computing the Area Under the ROC curve (AUC).



Recap: Area Under the ROC curve

If the costs of false negatives & false positives are different, the ROC curve allows us to fir … rade-off between FP rat …

We summarize … rve (AUC).



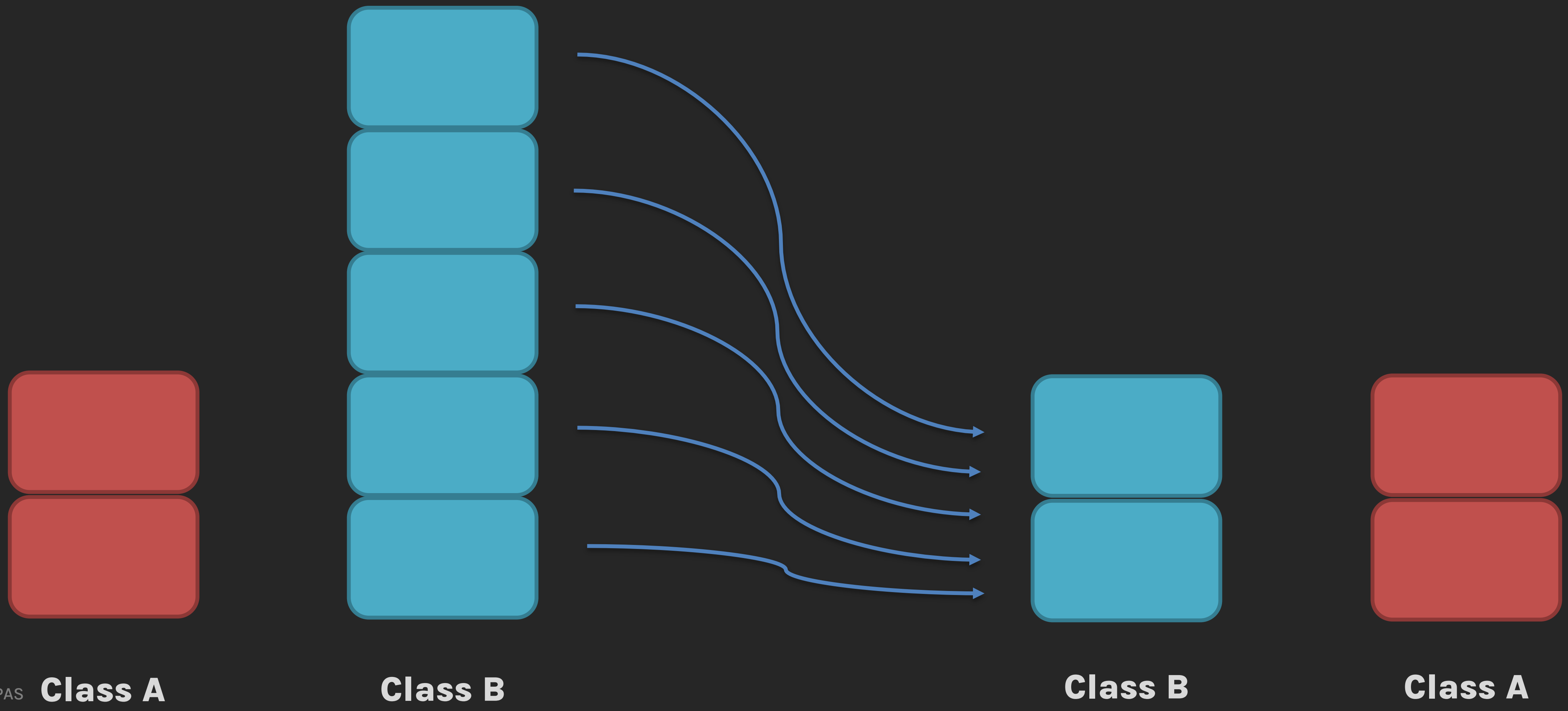
Dealing with Imbalanced classes

There are three main ways of dealing with imbalanced classes: undersampling, oversampling and class weighting.

1. Undersampling
    i. Random Sampling
    ii. Near Miss
2. Oversampling
    i. Random Sampling
    ii. SMOTE
3. Class weighting

Dealing with Imbalanced classes

1. Undersampling



Dealing with Imbalanced classes

1. **Undersampling**

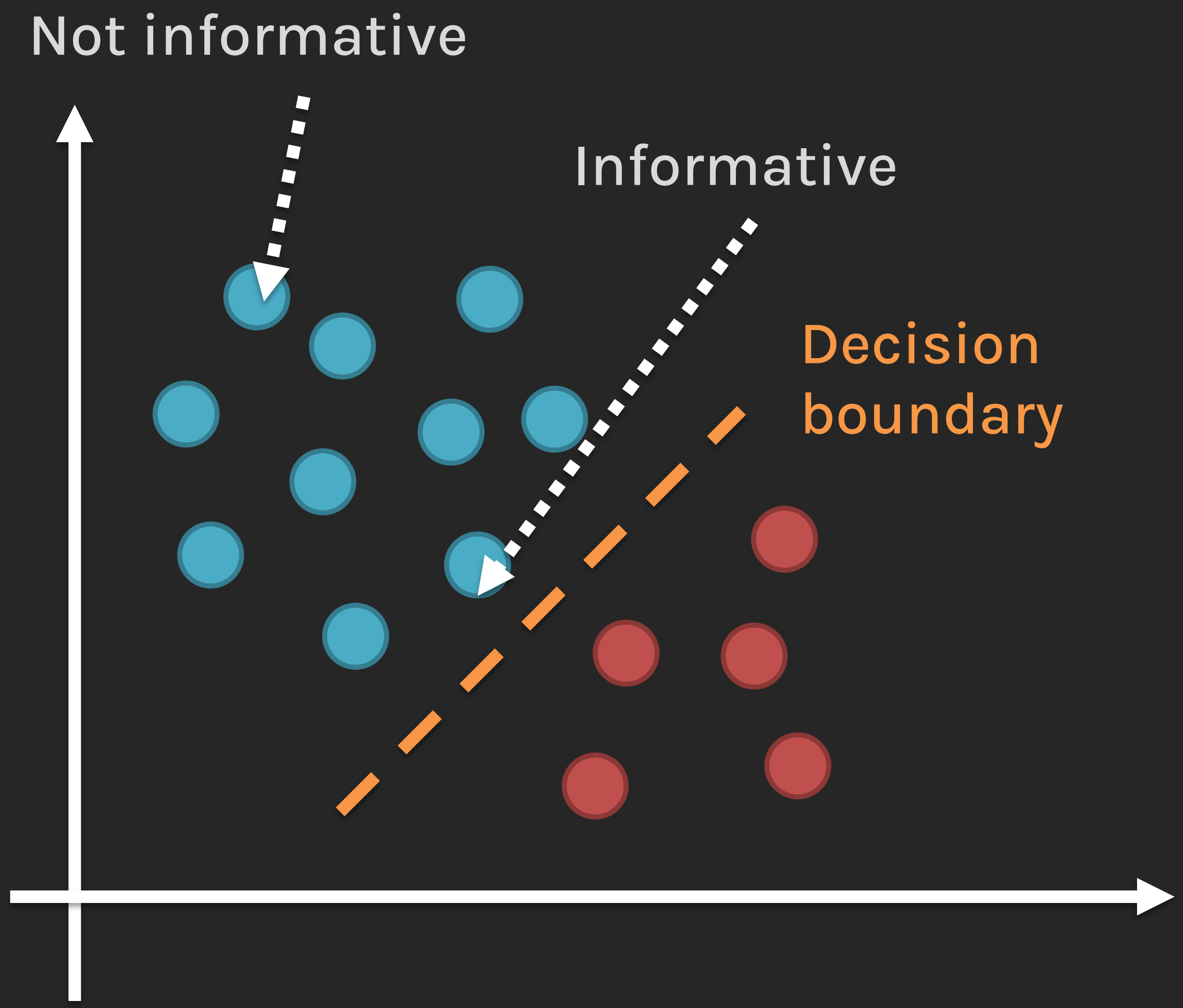
   We reduce the number of samples in majority class to match the number of samples in minority class.

   This can be done in two ways:

   i. Random Sampling:
   Randomly sample from majority class with or without replacement.

   ii. Near Miss:
   Select data points by using simple heuristics like finding samples from which the average distance to some data points of minority class is smallest. Read more about it here.

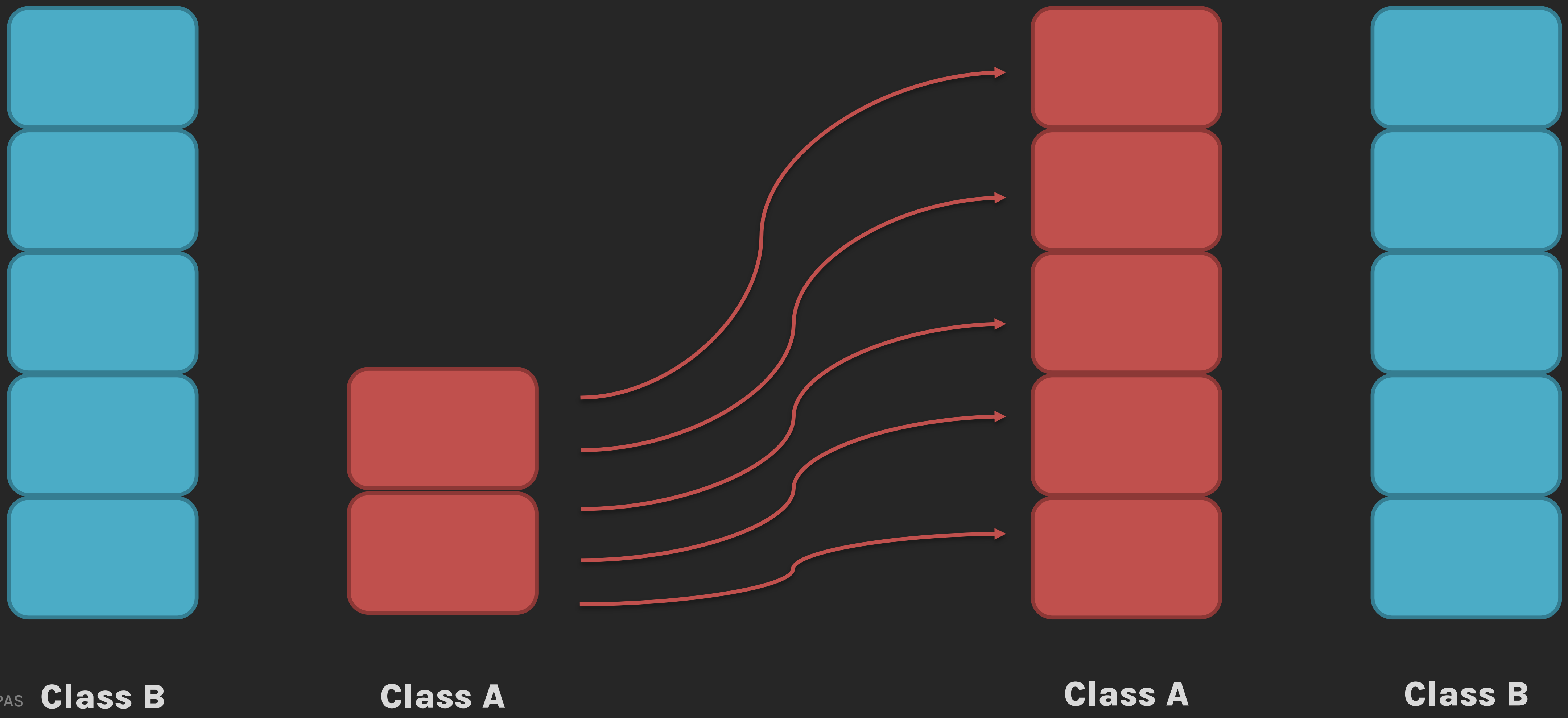
Issue of random sampling



- Random sampling can select data points that are not informative.
- Near miss, we can select more informative data points of the majority class; e.g., datapoints near the decision boundary in classification task.

Dealing with Imbalanced classes

2. Oversampling



Dealing with Imbalanced classes

2. Oversampling

We fight imbalanced data by generating new samples for minority class.

This can be done in two ways:

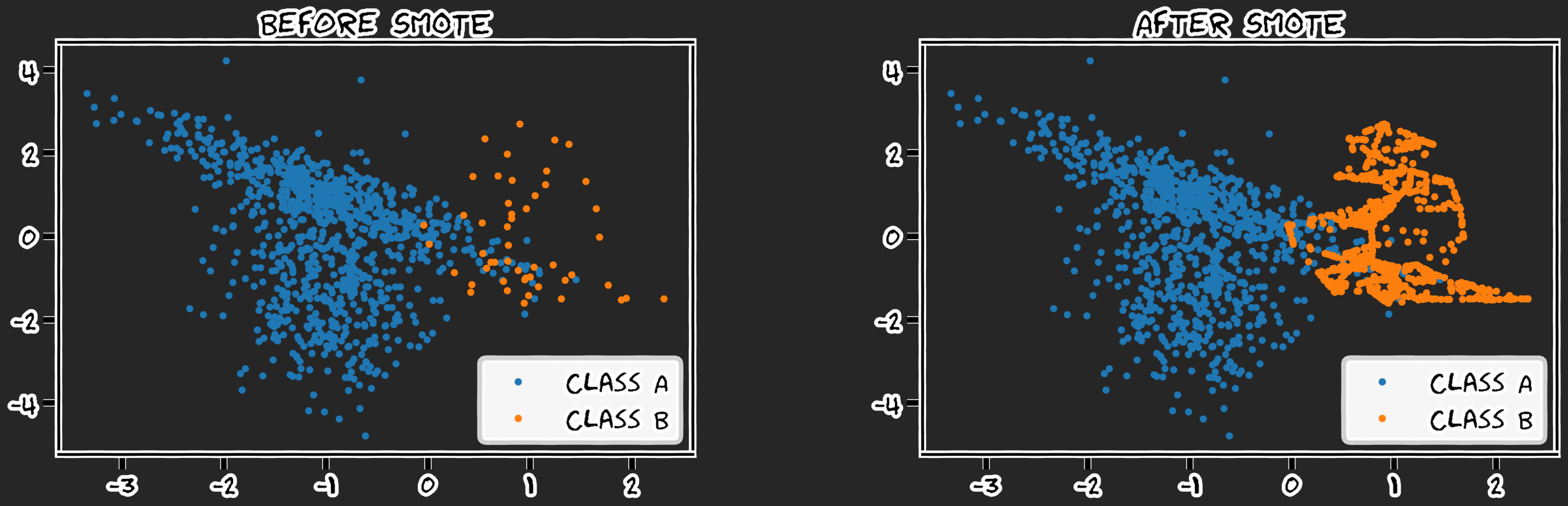
i. Random Sampling:
   Randomly sample from minority class with **replacement**.
ii. SMOTE:
   SMOTE is an improved alternative for oversampling.

SMOTE (Synthetic Minority Oversampling Technique):

ii. SMOTE:

SMOTE works by finding points that are closer in feature space.

Drawing a line between these points and generating new data points along this line.



Dealing with Imbalanced classes

3. Class weighting

A simple way to address the class imbalance is to provide a weight for each class which places more emphasis on the minority classes.

In sklearn we can provide the class weight as a dictionary or use

`class_weight = balanced`

Then it automatically adjust weights inversely proportional to class frequencies in the input data as:

$$\boxed{W_k = \frac{N}{K \times N_K}}$$

Where N is the total number of samples, N_k is the number of samples in class K and K is the total number of classes.

Outline

- Motivation
- Random Forest
- Variable Importance
- Missing Data (again)
- Class Imbalance
- Tree building algorithms [NEXT TIME]

When the students start to follow the professor's lead without skipping a beat!

Thank you