Digital Technologies and Value Creation

Dr. Philippe Blaettchen
Bayes Business School (formerly Cass)

Learning objectives of today

Goals:

- Understand the inspiration and ideas behind tree methods
- Understand how to apply CART in Python
- Understand how bagging, random forests and boosting are generalizations of CART

How will we do this?

- Start by thinking about how we make classifications as humans
- Titanic dataset: how does CART predict survivors?
- Visuals to explain the underlying concepts

Tree-based methods

A short introduction to tree-based methods

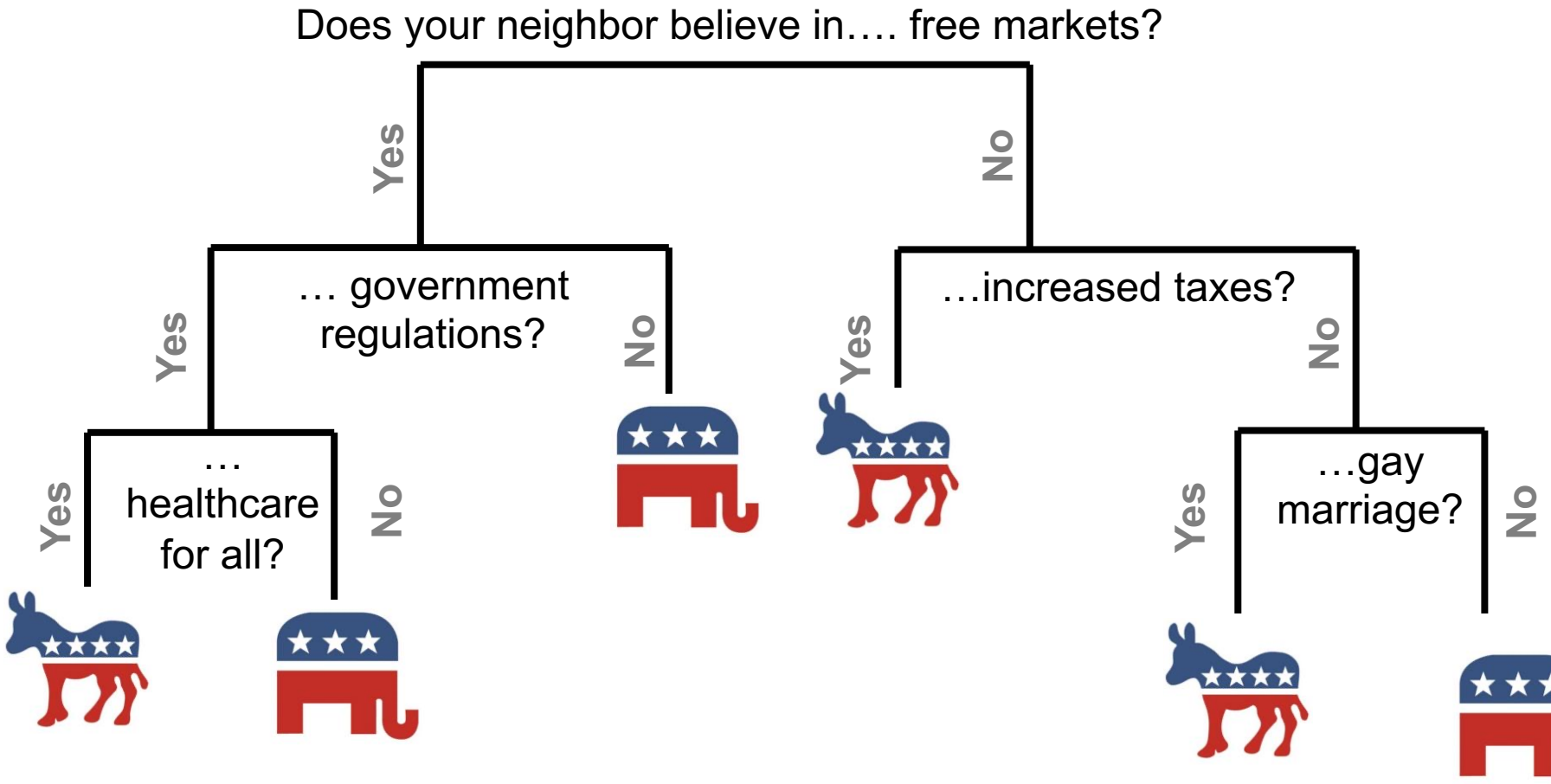
Suppose you are in the following scenario:

- you are at a dinner party in the US
- you want to figure out whether your neighbor aligns with the democratic or republican party more
- you don’t know anything about your neighbor
- for politeness reasons, you cannot ask them outright

Through conversation, **how would you figure out** which party they align with?

A possible decision tree

Caveat: Strategy which is (i) very US-centric; (ii) very schematic

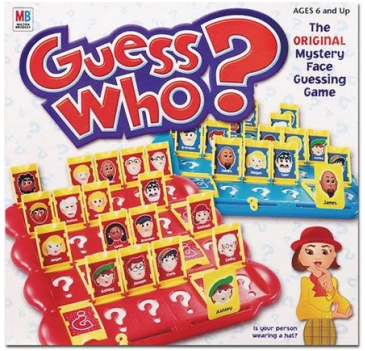


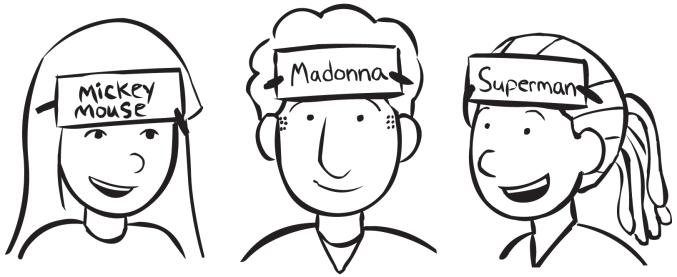
This is exactly tree-based classification!

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Tree-based classification

It's everywhere!





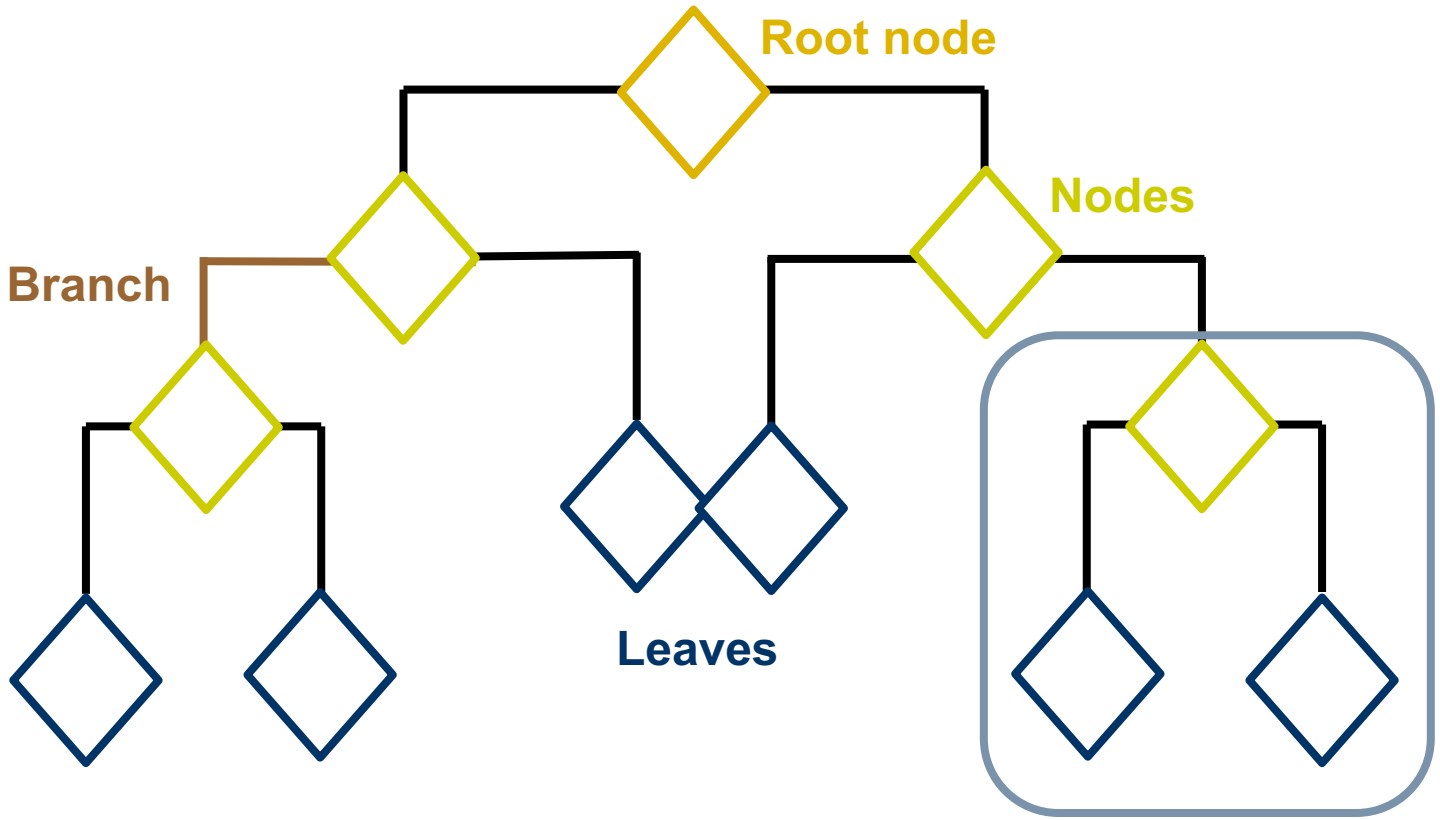
Idea:

Given a set of **features,** we obtain a **decision rule** for **classification.**

In tree-based classification, this decision rule is obtained **by recursive splitting of the features. The "end nodes"** contain **the prediction.**

Questions: **How do we split? When do we stop?**

Tree vocabulary



Branching/splitting : the act of creating new nodes
Pruning: removing leaves

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

When classification is a matter of life and death: Titanic

- Dataset of passengers on Titanic and who died/survived

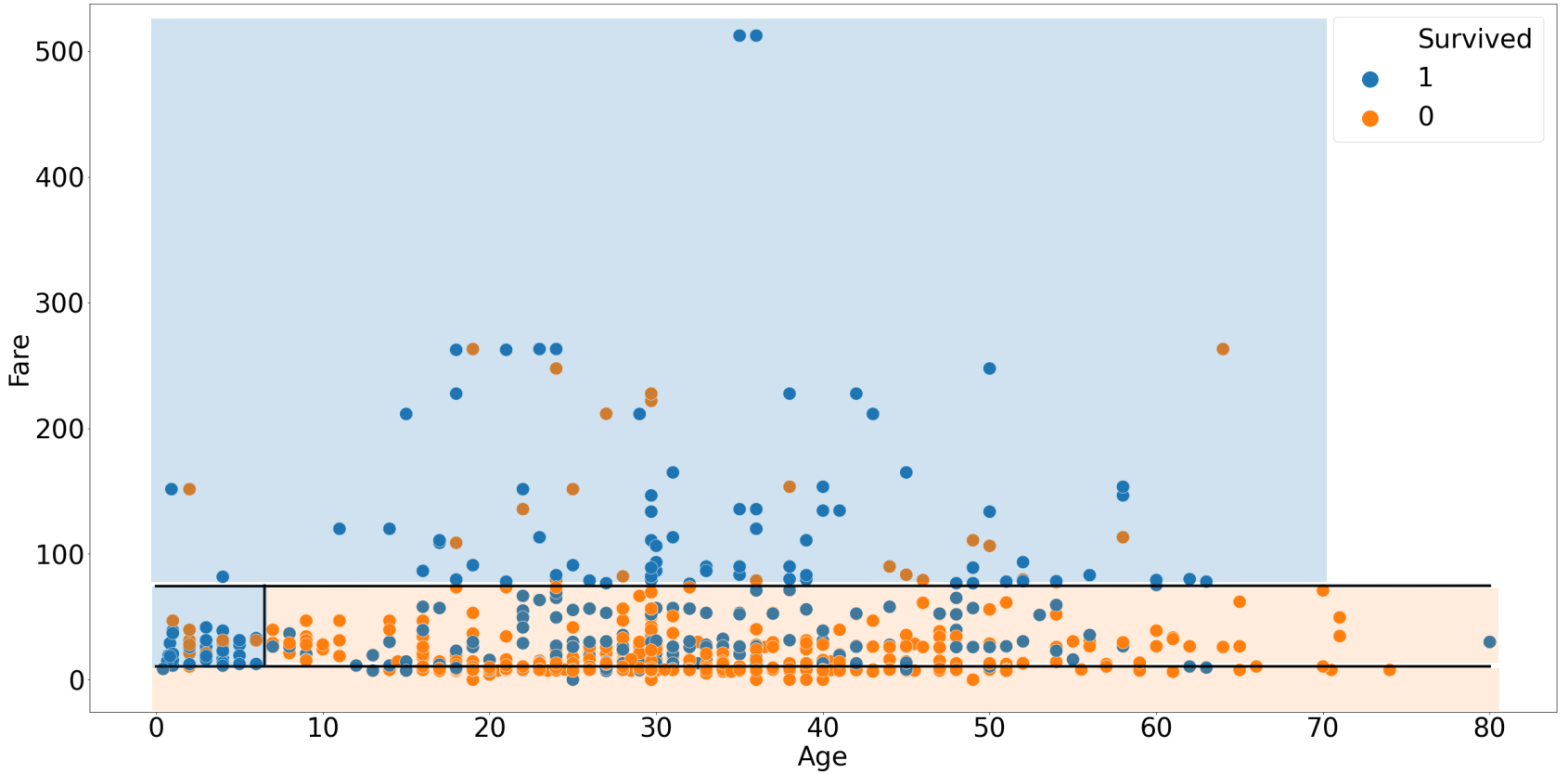
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- We will use CART to predict whether Jack and Rose the two main (fictitious) characters in the movie would have died or survived.

	Name	Pclass	Age	Fare	Family_Presence	Sex_male	Embarked_Q	Embarked_S
0	Jack Dawson	3	23	13	0	1	0	1
1	Rose DeWitt Bukater	1	17	247	1	0	0	1

CART: Classification and regression trees

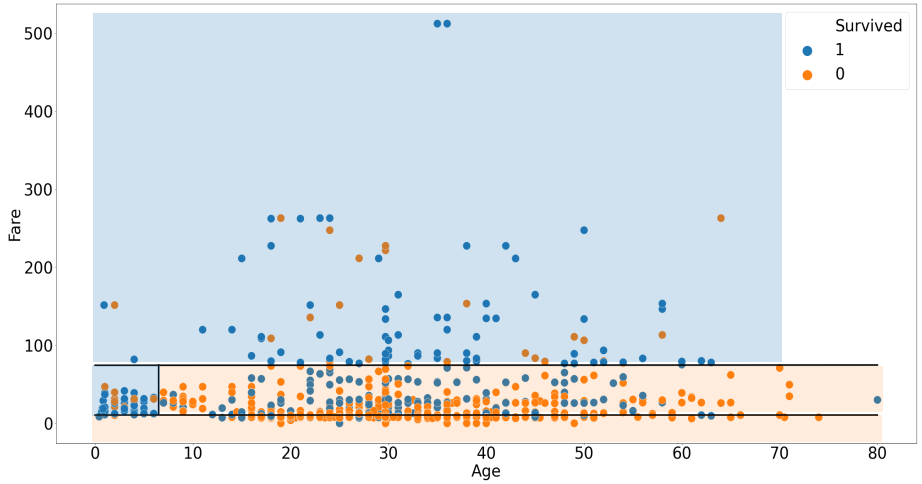
- CART can be run for both regression and classification
- We start with classification supposing we only had two features: fare and age.
- CART **divides the predictor space** (here the fare and age space) into boxes and then **classifies** based on the majority in that box

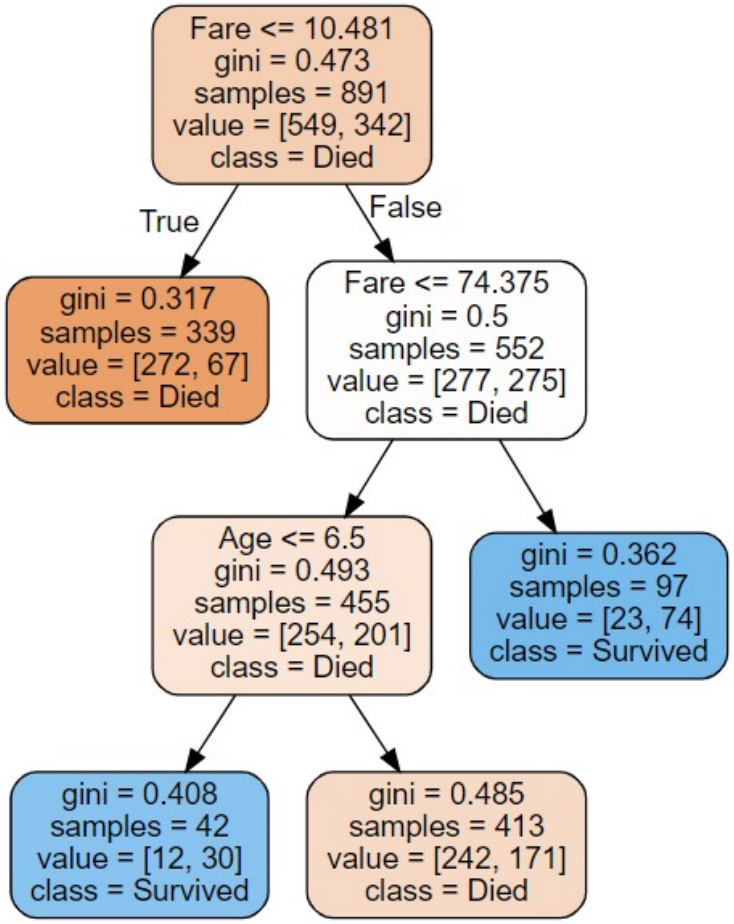


BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

The CART decision tree

This division corresponds to a tree:





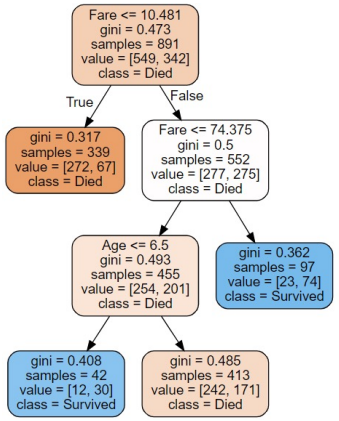
How do we know:

- On which variable to divide first? (Fare here)
- Which number to pick for the division? (10.481 e.g.)
- When to stop?

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

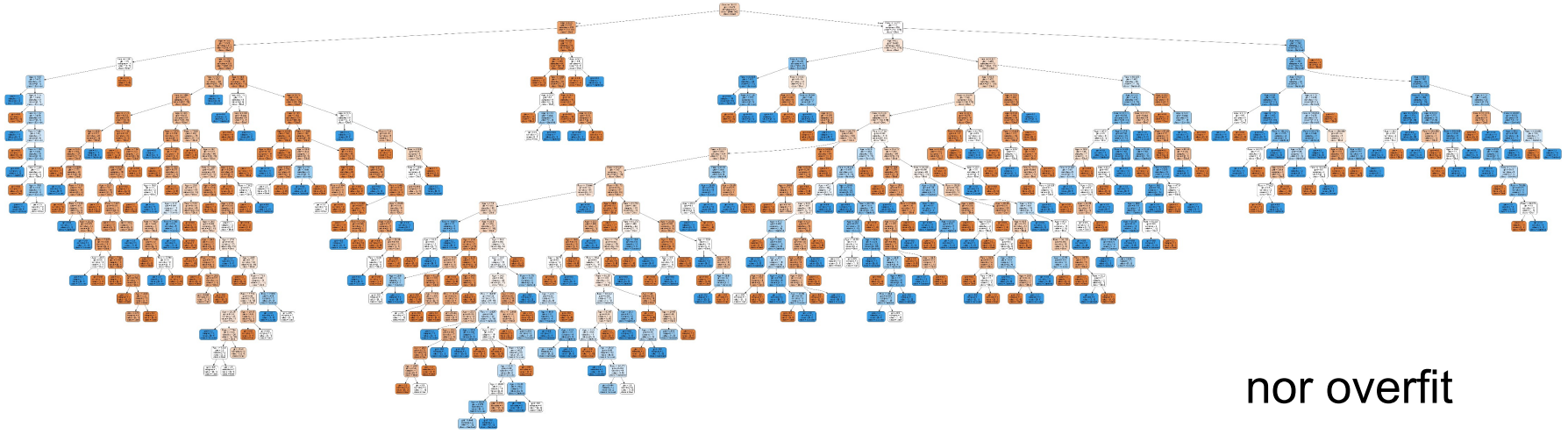
When to stop?

How to choose between…



Don’t want to underfit

Or…

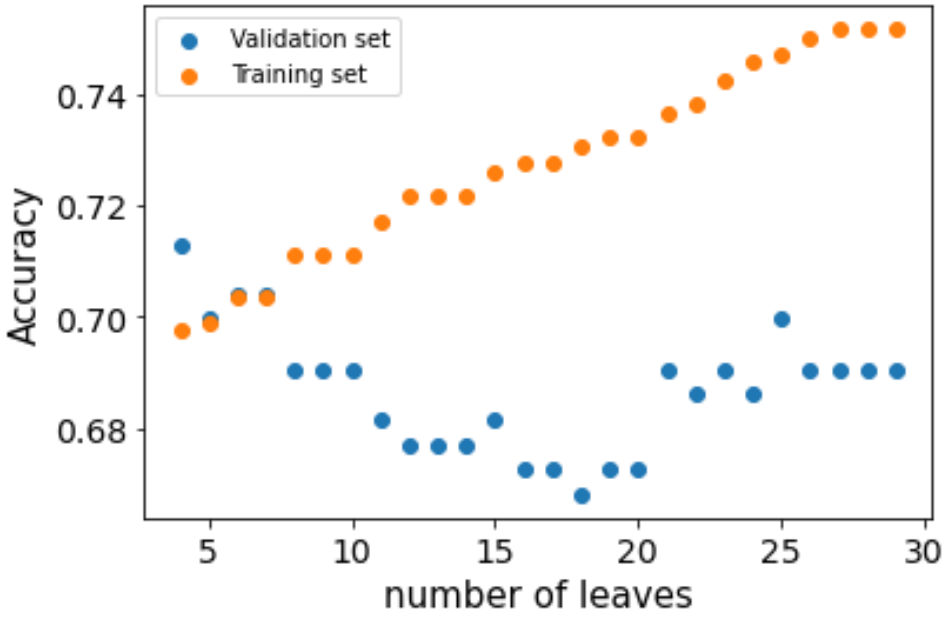


nor overfit

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

When to stop?

- **Many ways of doing this:** constraining number of leaves/number of samples per leaf/ pruning large tree
- Discuss **constraining number of leaves**
- Train the model on training set
- Determine **max number of leaves** based on **validation set: p**lot **accuracy of model** (1-proportion of misclassified points) as a function of **number of leaves**

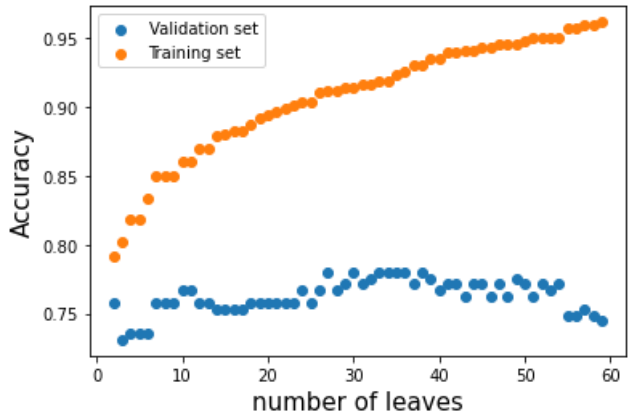


BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Should Rose and Jack have died?

Try it yourself: pull up the video-exercise notebook and make sure to download the appropriate dataset. Then, complete part 1.

Activity recap



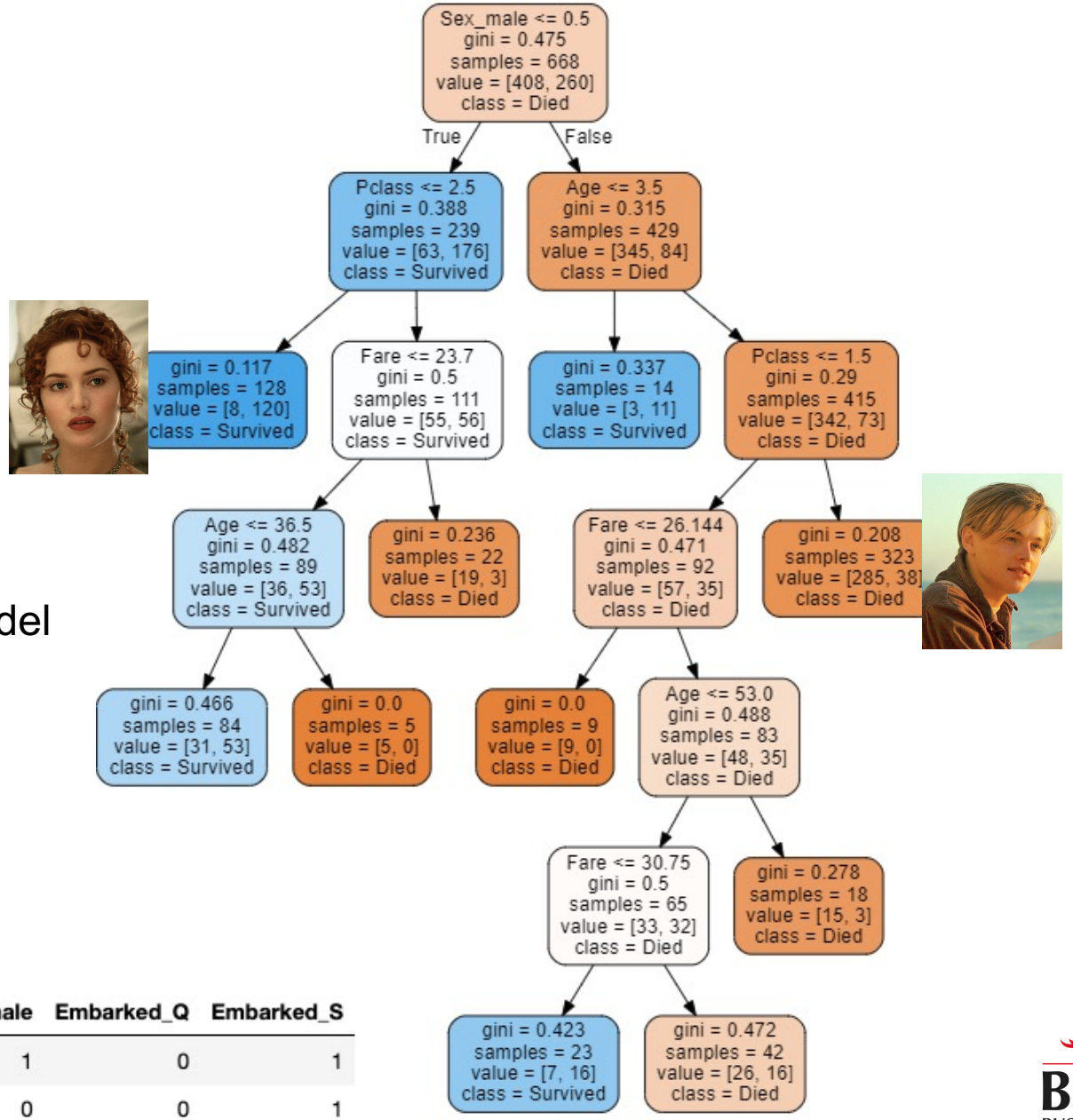
Pick around 10 leaves, retrain the model on training+validation and obtain tree

Accuracy on test set: 82%

```
accuracy_score(ytest,y_pred)
```

```
0.820627802690583
```

	Name	Pclass	Age	Fare	Family_Presence	Sex_male	Embarked_Q	Embarked_S
0	Jack Dawson	3	23	13	0	1	0	1
1	Rose DeWitt Bukater	1	17	247	1	0	0	1

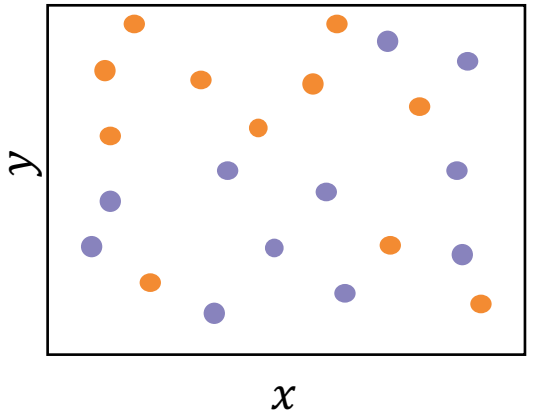


BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

CART: Classification AND regression trees

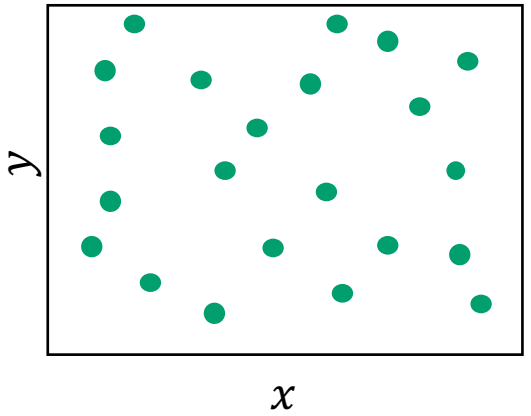
CART models can also be used for regression.

Classification



Predicted **class** of a region is the **most frequent class** that appears

Regression



Predicted **value** of a region is the **mean of all values** in the region

Pros & Cons of CART

Pros:

- Easy to explain to people
- Easy to understand and use (especially if small)
- “Mirrors” human way of thinking (see intro)

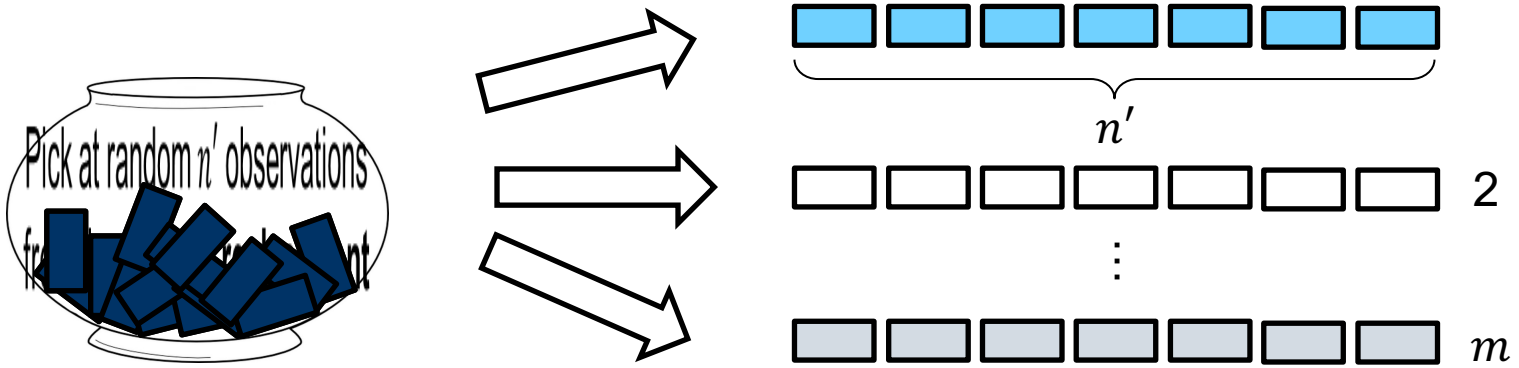
Cons:

- Not as good at the predictive aspect (accuracy generally not as good, etc.)
- **Not very robust:** changing the training data can lead to large variations in tree

→ Motivated the introduction of **bagging/random forest/boosting**

Extensions of CART

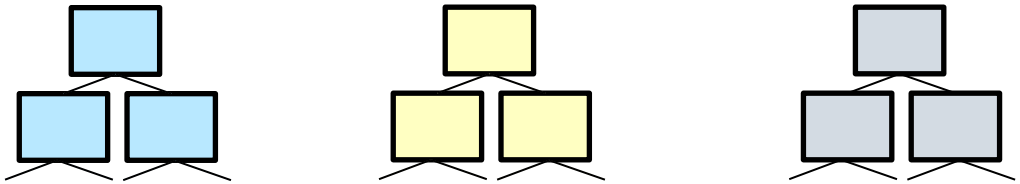
Bagging



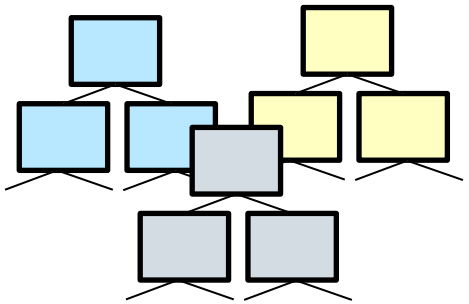
Do this m times

Pick at random n' observations from the urn **with replacement**

Train m trees, one on each set



How is each observation classified?



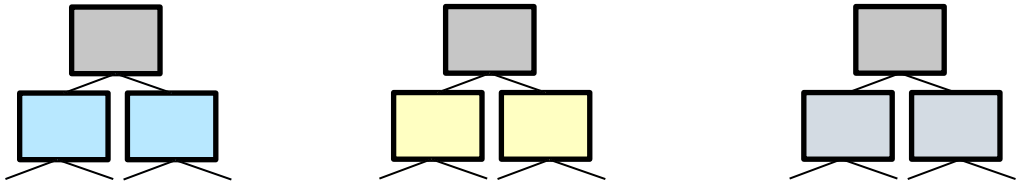
Each tree "**votes**" for where the observation should be classified and the majority wins

→ makes the process more **robust**

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Random forests

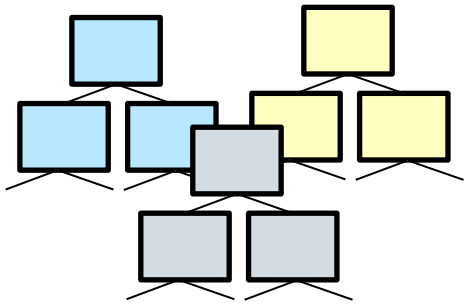
Very similar to bagging: train m trees on each set but **constrain** the training



If there is one “very predictive” feature in the dataset, then even if the training sets are different, the trees end up looking very similar (“correlated”) → **votes are the same.**

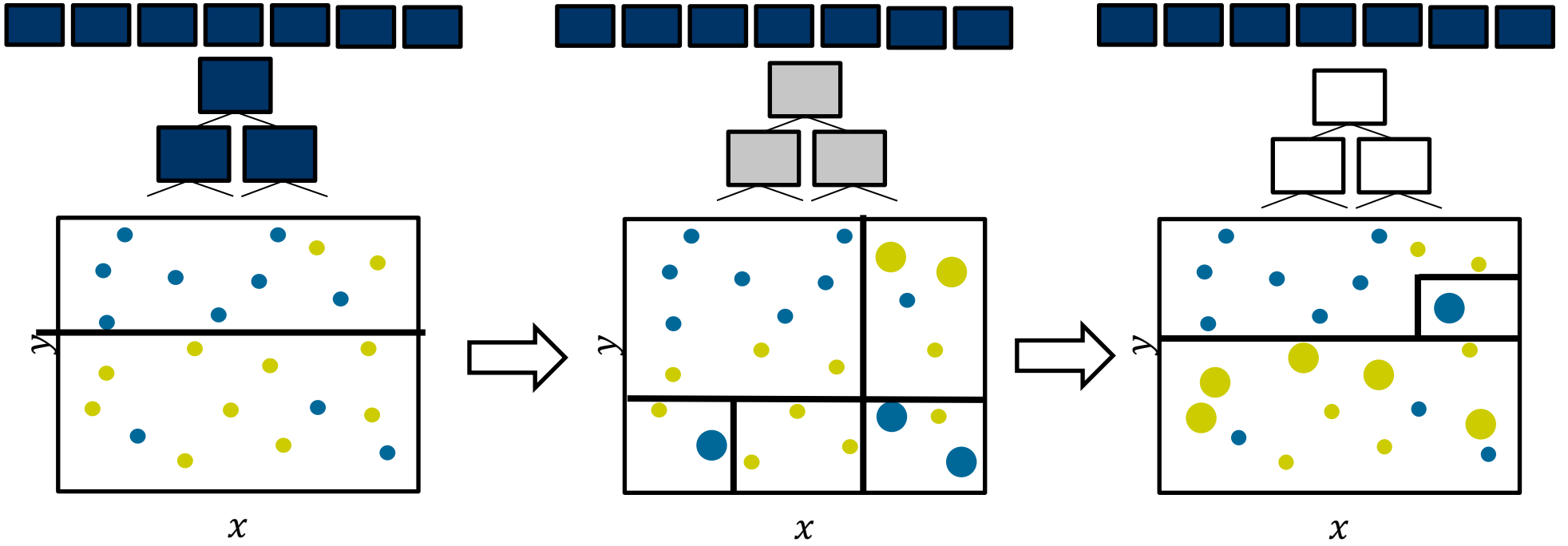
Idea: Only let the splitting occur within a **random subset of features** (typically if p features, can only split within subset of $\sqrt{p}$ for classification, $p/3$ for regression)

→ Ensures the **trees are different** enough.



Then **vote** as usual with majority winning!

Boosting



(1) Train a tree on initial data
(2) Compute misclass. error

(3) Train a new tree on weighted data
(4) Compute misclass. error

(5) Train a new tree on weighted data
(6) Compute misclass. error

- Each tree is an **expert** in some areas (and not so good in others)
- Each tree then **votes** on the right outcome, but the voting is **weighted** depending on its **misclassification error** (the higher, the less weight)

Further study

- Part 2 of the video-exercise notebook contains an example of running CART for regression, instead of classification
- Here, you can also try out running bagging, random forest, and boosting algorithms

From supervised to unsupervised learning

Supervised learning is all about **prediction.**

Predicting a **value: Regression**	Predicting a **category: Classification**
Methods: Linear/polynomial regression, CART & related	**Methods**: Logistic regression, CART & related

Input to supervised learning:

$$(\boldsymbol{X}, \boldsymbol{y})$$

X: feature matrix (i.e., columns are features and rows are observations)

y: label (predicted value or category) for each observation

Unsupervised learning – no predictions!

Unsupervised learning **cannot do** prediction. Why?

Input to unsupervised learning:

$$\boldsymbol{X}$$

i.e., just the feature matrix – no labels, so no way of predicting.

What does unsupervised learning do?

Tries to **draw patterns or other information** out of X

What does unsupervised learning do?

This week: **Dimension reduction**

How can we construct, from our set of features, a smaller number of new features while keeping all the information?

Next week: **Clustering**

Can I group my data into similar groups?

Other possible topics: anomaly detection, association rules

“Issues” for unsupervised learning

- No simple goal for unsupervised learning: generally used in an **exploratory data analysis,** and as an **input to supervised learning**
- **No notion of “true answer”**: no immediate way of evaluating our work – often subjective
- As a consequence, no training/testing/validation/cross-validation possible


See you in class!