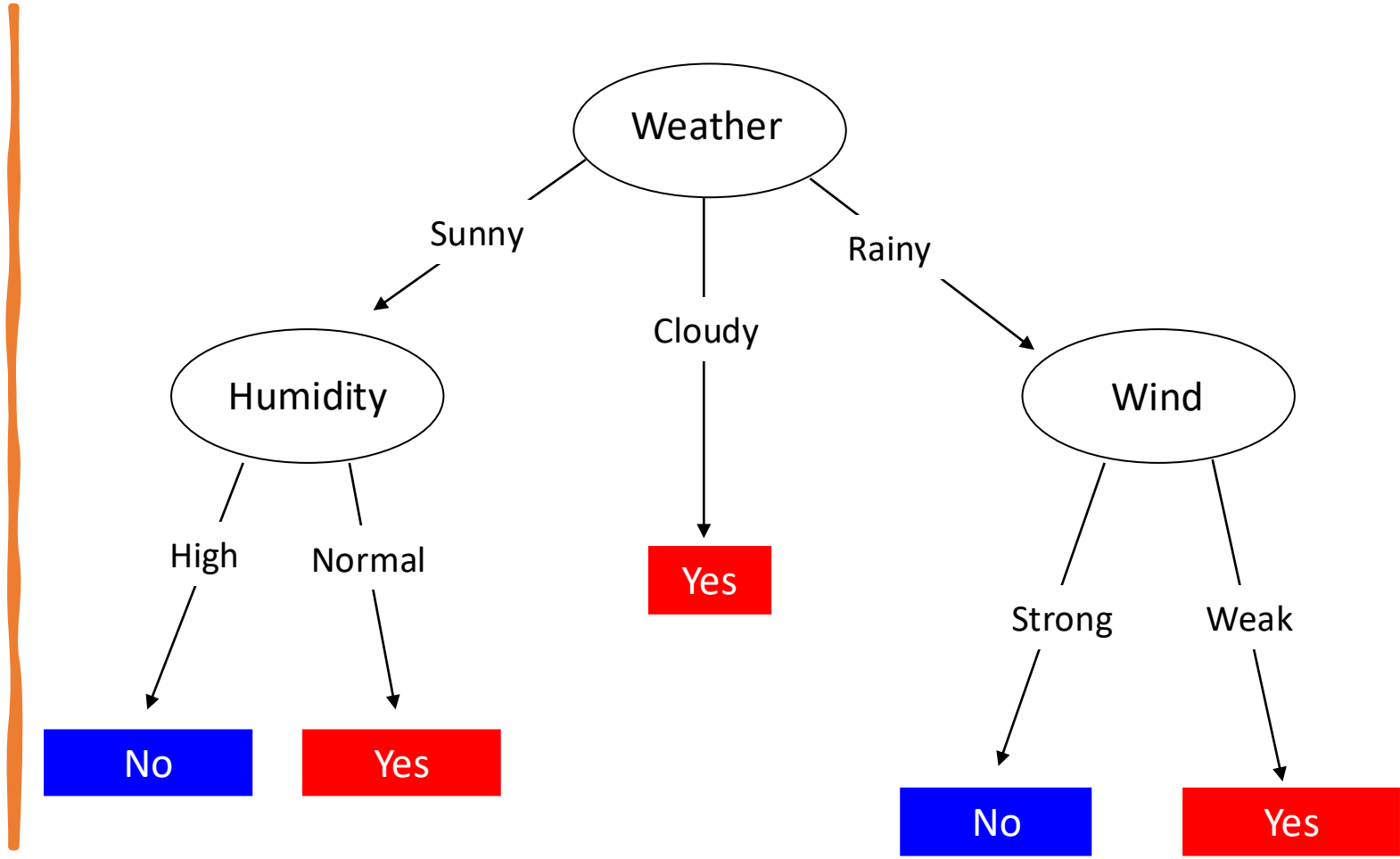
ML - 4

Machine Learning

- Supervised Learning
  - Entropy based Classification
    - Decision Trees (ID3 and C4.5)

What is Decision Tree?



Decision Trees

We'll learn two decision trees using entropy. These are:

- ID3
- C4.5

But first, we should learn something about entropy, uncertainty, information and information gain terms.

In probability theory, **entropy** is defined as a measure of **uncertainty** about a random variable.

According to Shannon, entropy quantifies the average amount of **information** produced by a stochastic source of data.

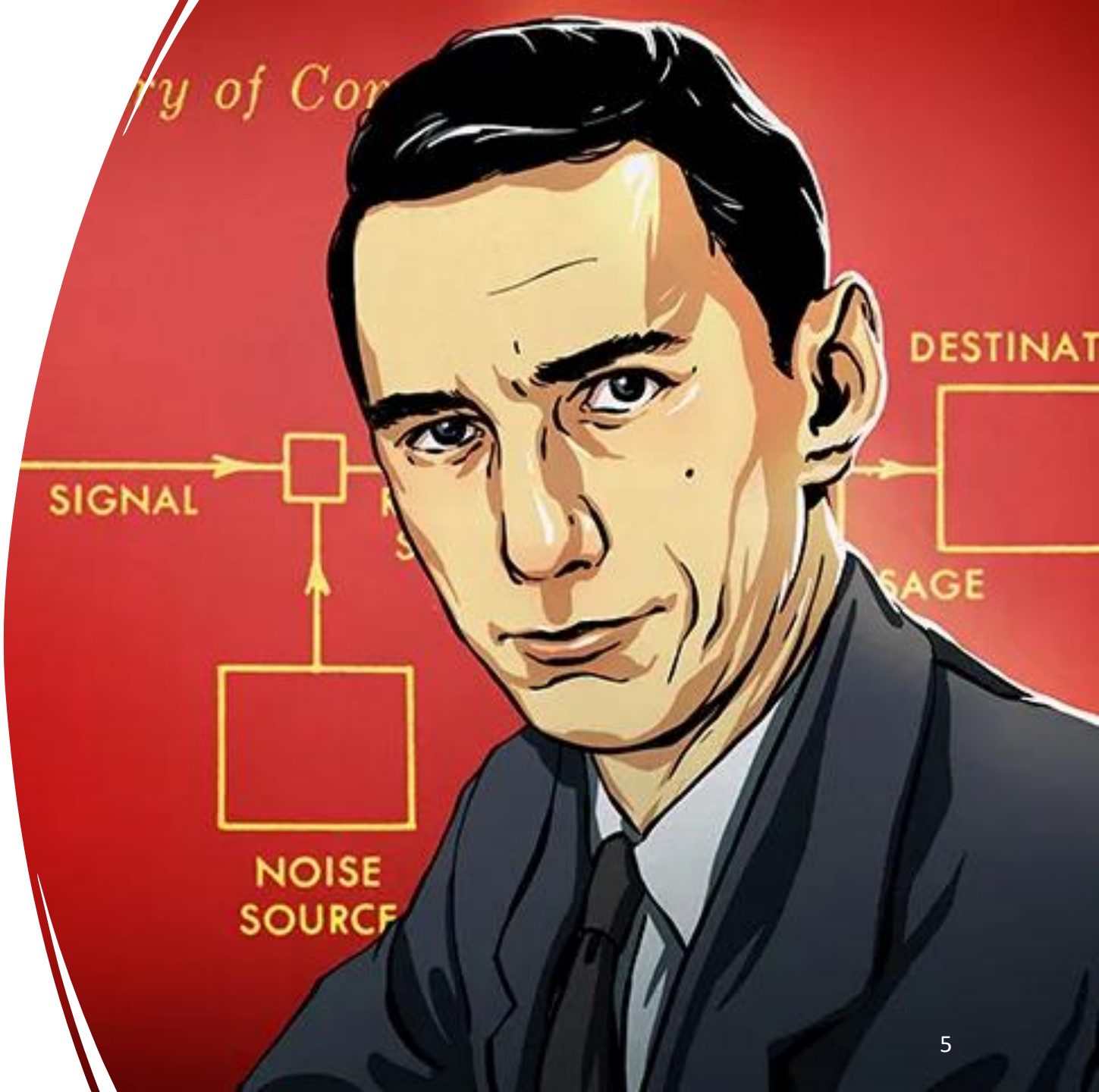
The more **uncertain** the outcome, the higher the **entropy**.

Less **information** (more **uncertainty**), more **information gain**.

- Entropy
- Uncertainty
- Information
- Information gain

Claude Elwood Shannon

- Shannon Entropy (*H*) is defined as
  - $H(x) = \sum_{i=1}^{n} P(x_i) I(x_i)$
- Here, I(x) information is defined as
  - $I(x) = log_2 \frac{1}{P(x)} = -logP(x)$
- Then
  - $H(x) = -\sum_{i=1}^{n} P(x_i) log_2 P(x)$



Example

Let X be random process of a coin toss.

- $H(x) = -\sum_{i=1}^{n} P(x_i) log_2 P(x)$

$$= -(0.5\ log_2 0.5 + 0.5\ log_2 0.5) = 1$$

Then the entropy is calculated as 1.

Weather	Temp.	Humidity	Wind	Day	Class
Sunny	Hot	High	Weak	Weekday	**No**
Sunny	Hot	High	Strong	Weekday	**No**
Overcast	Hot	High	Weak	Weekend	**Yes**
Rainy	Mild	High	Weak	Weekend	**Yes**
Rainy	Cool	Normal	Weak	Weekend	**Yes**
Rainy	Cool	Normal	Strong	Weekday	**No**
Overcast	Cool	Normal	Strong	Weekday	**Yes**
Sunny	Mild	High	Weak	Weekday	**No**
Sunny	Cool	Normal	Weak	Weekend	**Yes**
Rainy	Mild	Normal	Weak	Weekend	**Yes**
Sunny	Mild	Normal	Strong	Weekend	**Yes**
Overcast	Mild	High	Strong	Weekday	**Yes**
Overcast	Hot	Normal	Weak	Weekend	**Yes**
Rainy	Mild	High	Strong	Weekday	**No**
Sunny	Mild	High	Strong	Weekend	**No**
Sunny	Cool	Normal	Weak	Weekday	**Yes**
Rainy	Cool	High	Weak	Weekend	**Yes**
Overcast	Mild	High	Weak	Weekend	**Yes**
Sunny	Hot	Normal	Weak	Weekday	**Yes**
Rainy	Mild	Normal	Strong	Weekday	**No**

Entropy in Decision Tree

In a dataset with 20 samples and 5 numerical features, in the worst-case scenario, there are 5x19=95 branching criterion cases.

For this reason, methods such as information gain, Gini index and Chi square test are used.

Ross Quinlan

He is best known for developing the ID3 algorithm in the 1980s and its successor, the C4.5 algorithm.

ID3 Algorithm

It works with only categorical data.

During each iteration, the process begins by calculating the entropy of the entire dataset for all features

Next, the entropy for each feature, considering the classes, is computed and then subtracted from the initial entropy.

The feature that provides the maximum information gain is selected for branching.

ID3 Example

We have a dataset with 4 samples, 2 features, and two classes. Find the first branching feature?

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

H(S) - H(V1,S)=?

H(S) - H(V2,S)=?

ID3 Example

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

At first, general entropy:

$$H(S) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

Entropy of V1

$$H(V1) = \frac{1}{4}H(A) + \frac{3}{4}H(B)$$

$$= \frac{1}{4}0 - \frac{3}{4}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)$$

$$= 0 + \frac{3}{4}0{,}9183 = 0{,}6887$$

Entropy of V2

$$H(V2) = \frac{1}{2}H(C) + \frac{1}{2}H(D) = \frac{1}{2} + \frac{1}{2} = 1$$

C4.5 Algorithm

Categorical and Numerical data

Pruning and overfitting

Lost and Missing values

Handling large datasets,

Handling noisy data,

Using gain ratio

C4.5 Algorithm

All values in numerical feature are sorted

An average value for each neighbor pair at sorted list is computed

Each threshold is tested to branch dataset, and an information gain is computed

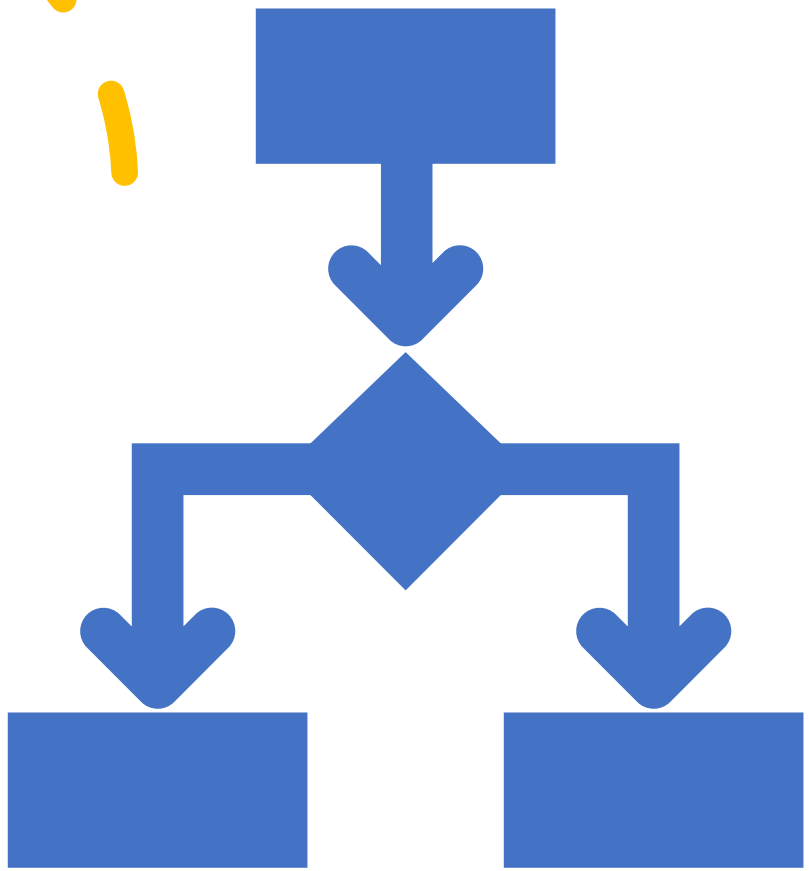
The threshold supporting the maximum information gain is chosen

Overfitting

C4.5 algorithm uses pruning methods to reduce overfitting.

The C4.5 algorithm employs

- pre-pruning
- post-pruning



Pre-Pruning

Pre-pruning involves setting stopping criteria during the decision tree construction process to prevent further growth of the tree.

- Minimum information gain threshold
- Minimum number of instances

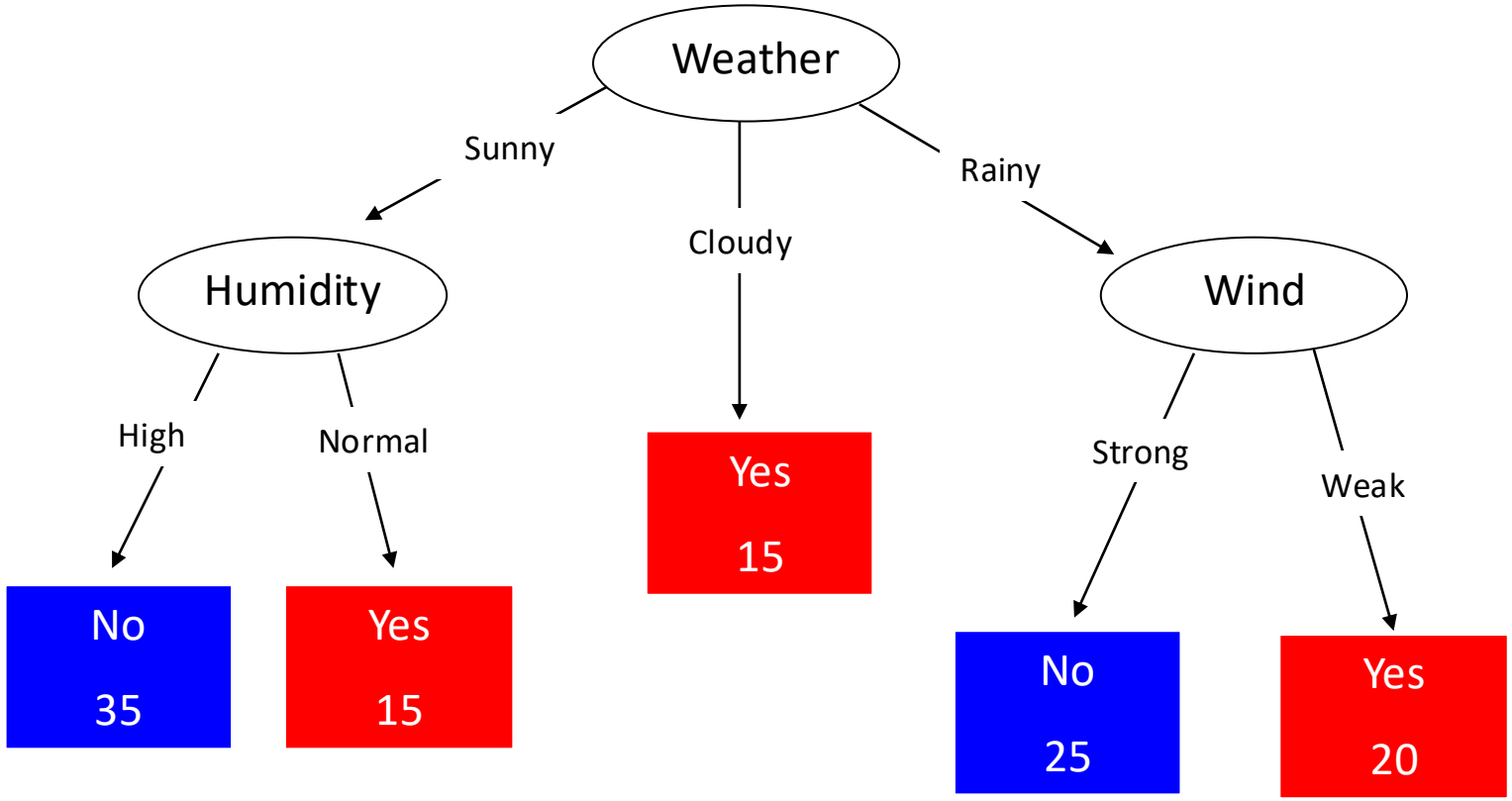
Post-Pruning

Post-pruning involves pruning the decision tree after it has been fully constructed

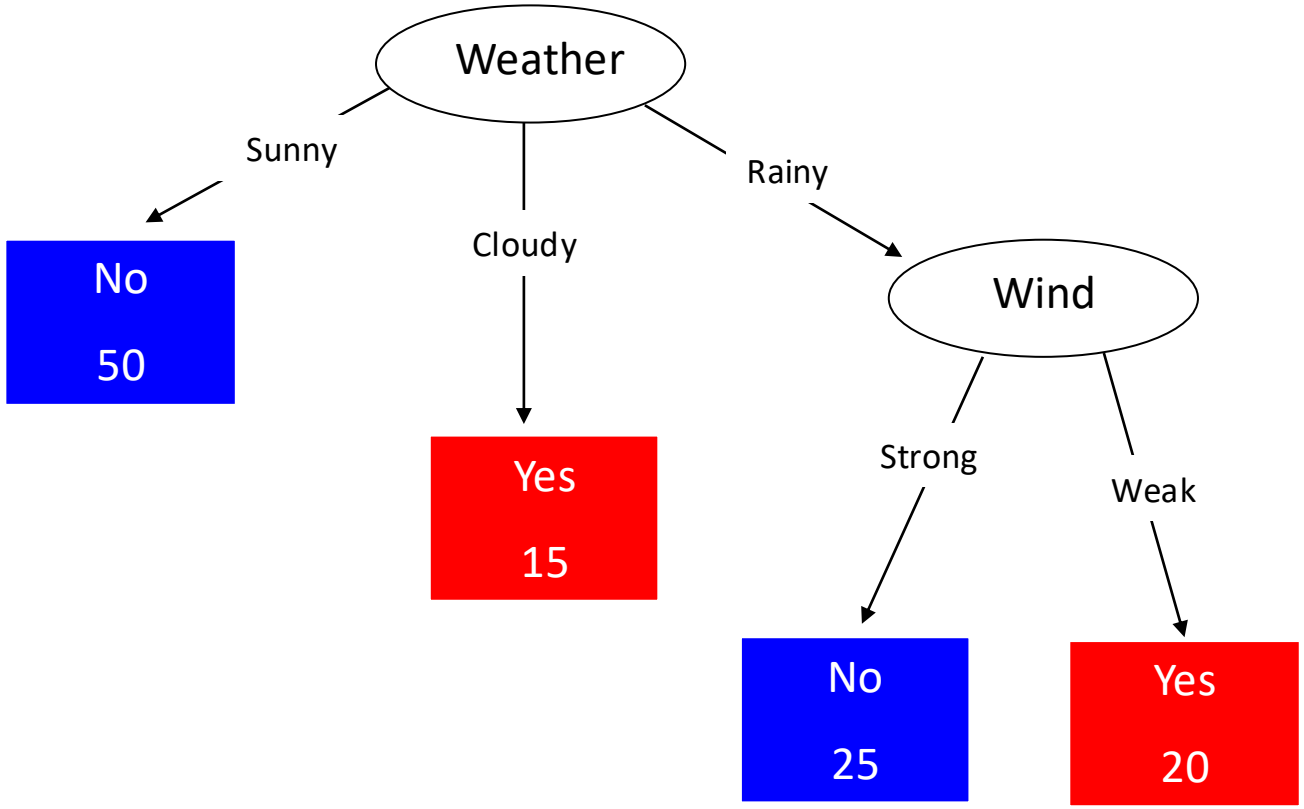
- Error-based pruning,
- Reduced error pruning,
- Rule post-pruning

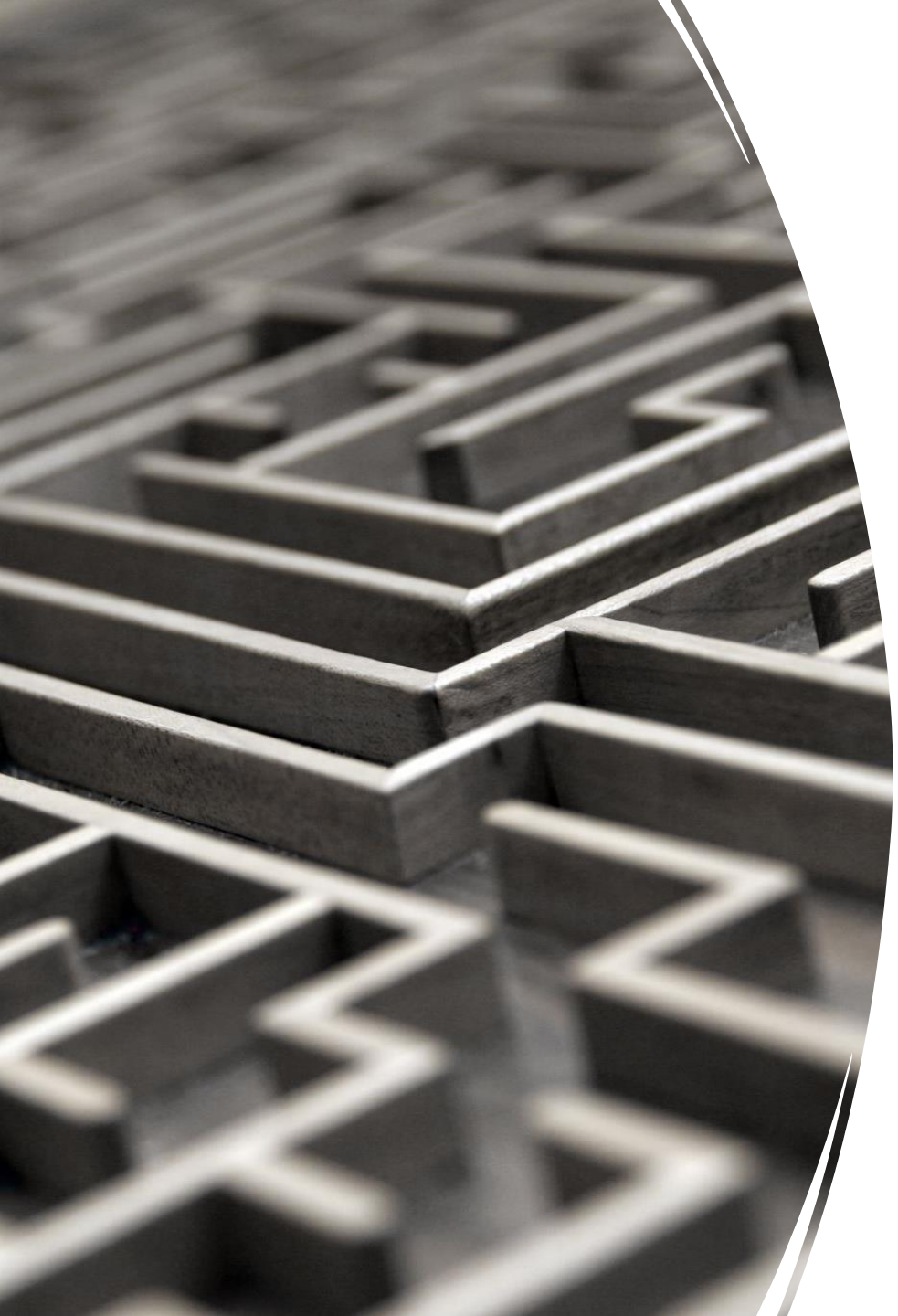
Error-based Post Pruning Example

Error tolerance is 33%



Error-based Post Pruning Example





Missing Value

If some samples are missing in dataset, there are two approaches to be followed:

1. Removing
2. Imputation

Missing Value

There are many approaches to imputation in missing data:

1. Imputation with Mean or Median
2. Probabilistic imputation
3. Imputation with K-NN

X_1	X_2	X_3	X_4	D
K	P	G	5	**H**
K	N	G	1	**H**
B	E	G	3	**H**
B	P	G	5	**S**
B	N	T	-1	**S**
K	E	T	2	**S**

Manual Calculations

Using C4.5, find the first branching factor for dataset given below.

0.92	1.0	0.54	0.81	
X_1	X_2	X_3	X_4	D
K	P	G	5	H
K	N	G	1	H
B	E	G	3	H
B	P	G	5	S
B	N	T	-1	S
K	E	T	2	S

Manual Calculations

Using C4.5, find the first branching factor for dataset given below.

- $H(D) = -(0.5log_2 0.5 + 0.5log_2 0.5) = 1.0$
- $X_4 = \{-1, 1, 2, 3, 5\}$
- Thresholds t $= \{0, 1.5, 2.5, 4\}$

$X_4 > 0$	$X_4 > 1.5$	$X_4 > 2.5$	$X_4 > 4$
A	A	A	A
A	B	B	B
A	A	A	B
A	A	A	A
B	B	B	B
A	A	B	B
0.81	**1.0**	**0.92**	**1.0**

Machine Learning

4. week

Thanks for watching