October 4th, 2022

Introduction to Machine Learning Part #1

Yusen He, PhD - DASIL Data Scientist

Prof. Julia Bauder – Director of DASIL

Martin Pollack – DASIL Post-Bac Fellow

DATA ANALYSIS & SOCIAL INQUIRY LAB

GRINNELL COLLEGE

Intro to Machine Learning Part #1

AGENDA

Machine Learning Basics

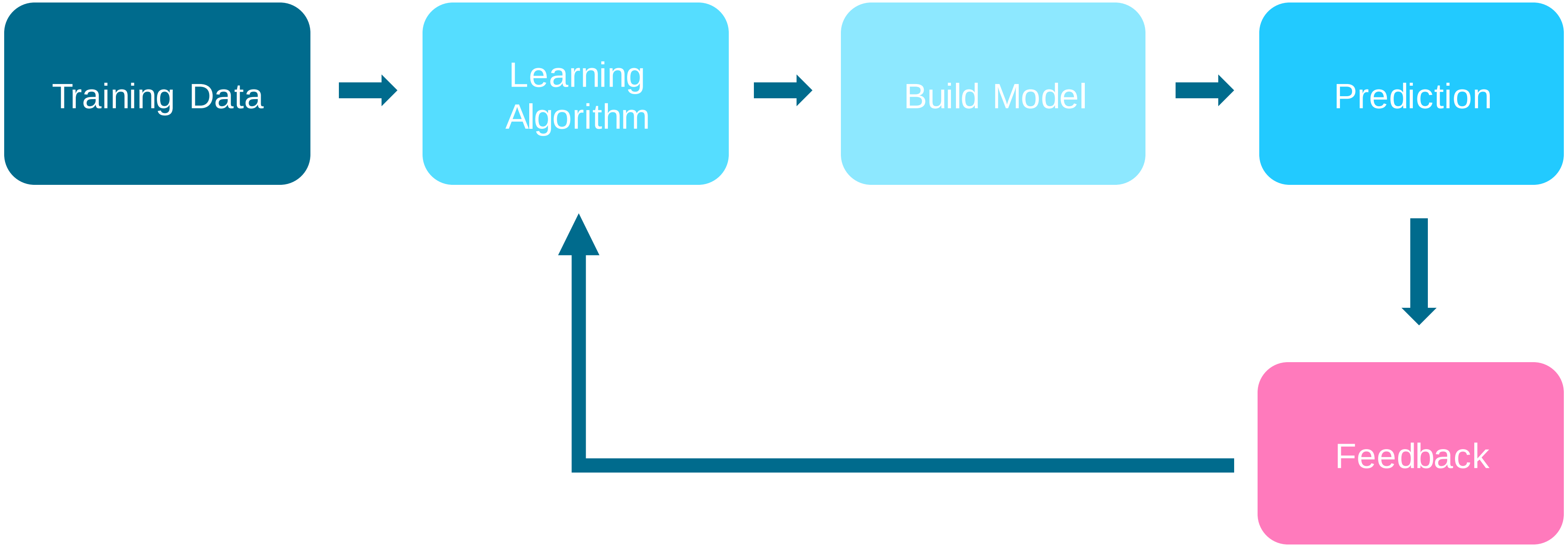
Evaluating Classification: Entropy & Information Gain

Basic Decision Tree

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

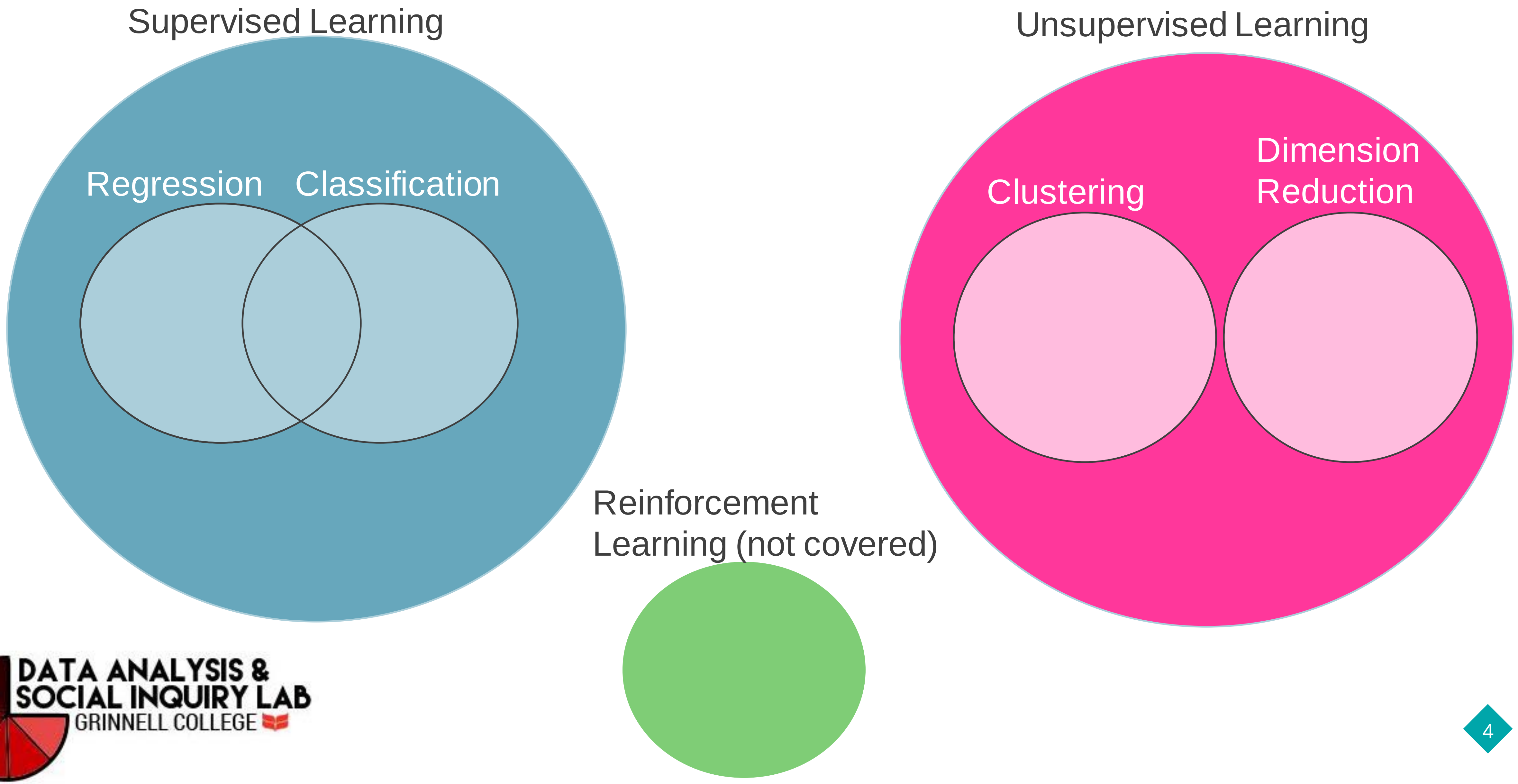
What is Machine Learning?

- A type of Artificial Intelligence that learns from data
- Computer algorithms that improve without being explicitly programmed



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

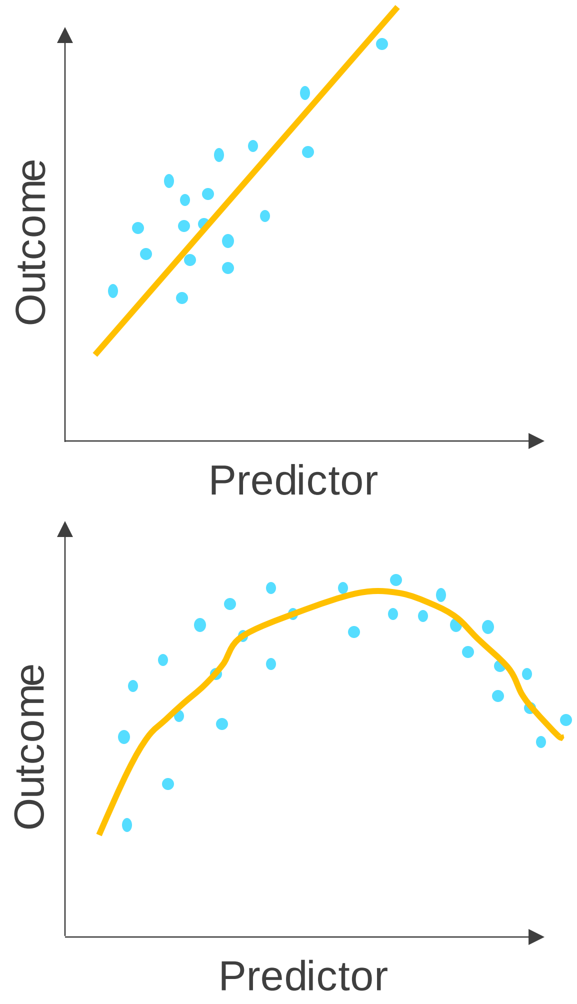
What is Machine Learning?



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Supervised Learning

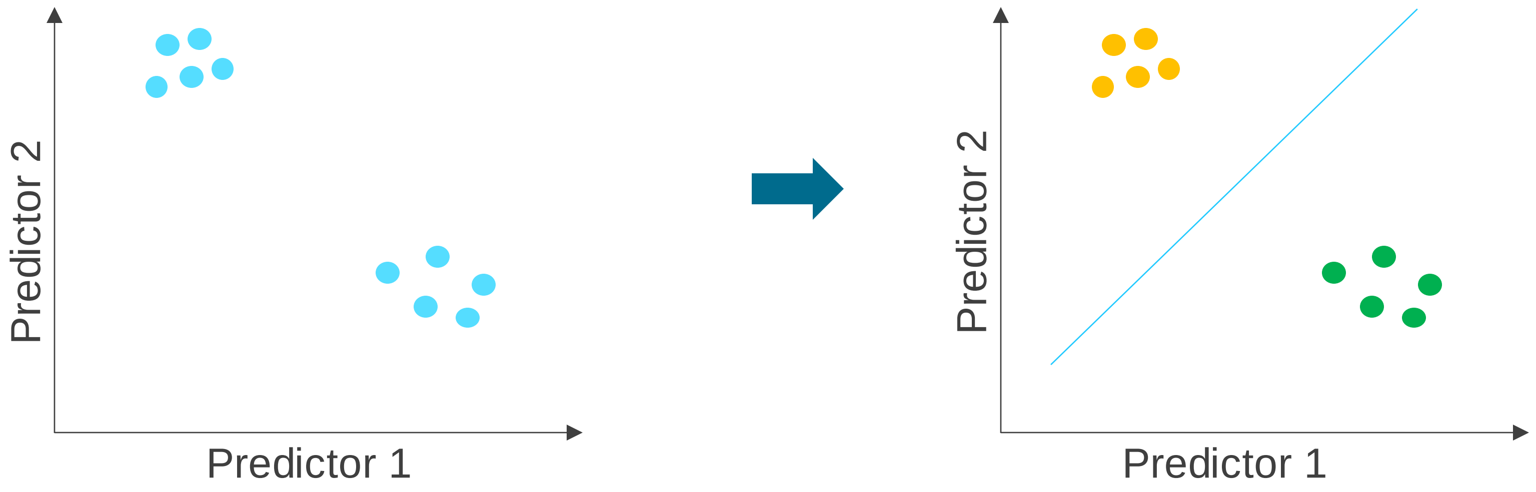
- Regression: predict continuous outcome
  - Linear regression
  - Non-linear regression



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Supervised Learning

- Classification: predict discrete outcome/predict group data belongs to
  - Binary classification
  - Multi-categorical classification



Classification or Regression?

In Class Question:

- Suppose we want to predict the probability that a person will get approved for a credit card.
- We have data from Visa on everyone who submitted a credit card application in 2021.
- Is this a classification or regression problem?

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Intro to Machine Learning Part #1

AGENDA

Machine Learning Basics

Evaluating Classification: Entropy & Information Gain

Basic Decision Tree

What is Entropy?

Entropy:

- A quantitative measure of randomness or uncertainty
- Usually between 0 and 1
- Higher entropy = higher randomness

- Used to measure the effectiveness of a classification algorithm

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

What is Entropy?

Think about a scenario: Ice vs Water

Which one has higher randomness?

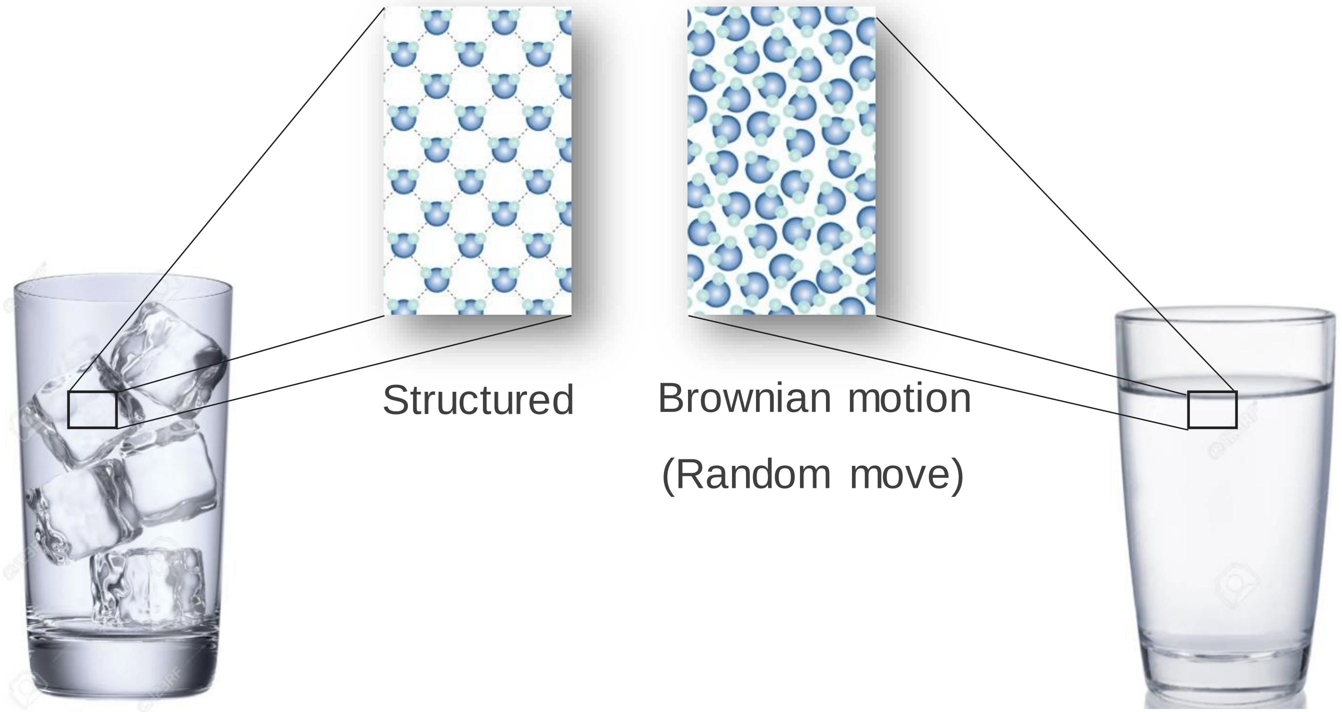
A cup of ice

A cup of water

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

What is Entropy?

In molecular dynamics:

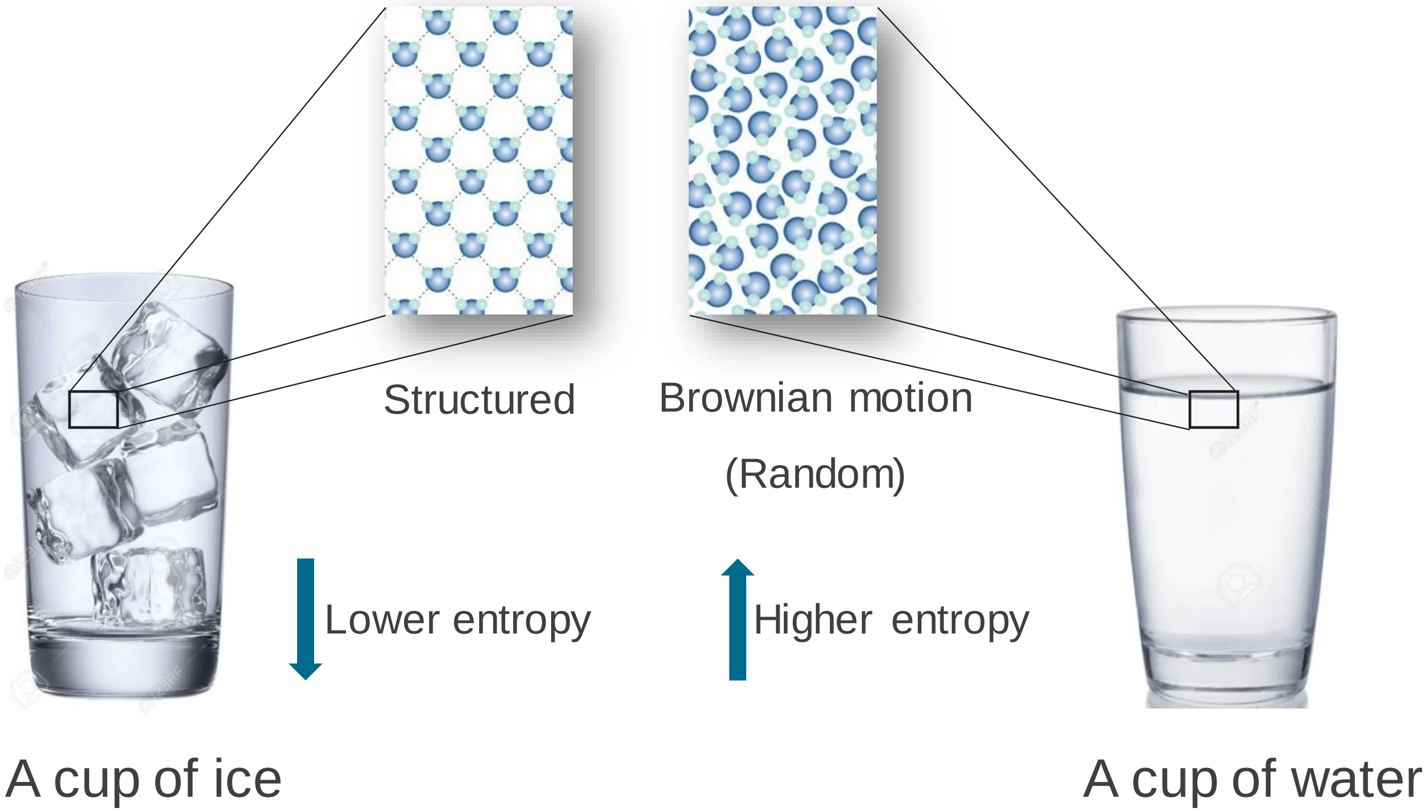


A cup of ice

A cup of water

What is Entropy?

In molecular dynamics:



What is Entropy?

How to compute entropy?

$$Entropy = -\sum_{i=1}^{n} P(x_i)\log(P(x_i))$$

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

What is Entropy?

How to compute entropy?

Total number of classes

$$Entropy = -\sum_{i=1}^{n} P(x_i)\log(P(x_i))$$

The probability of a certain class of events (data)

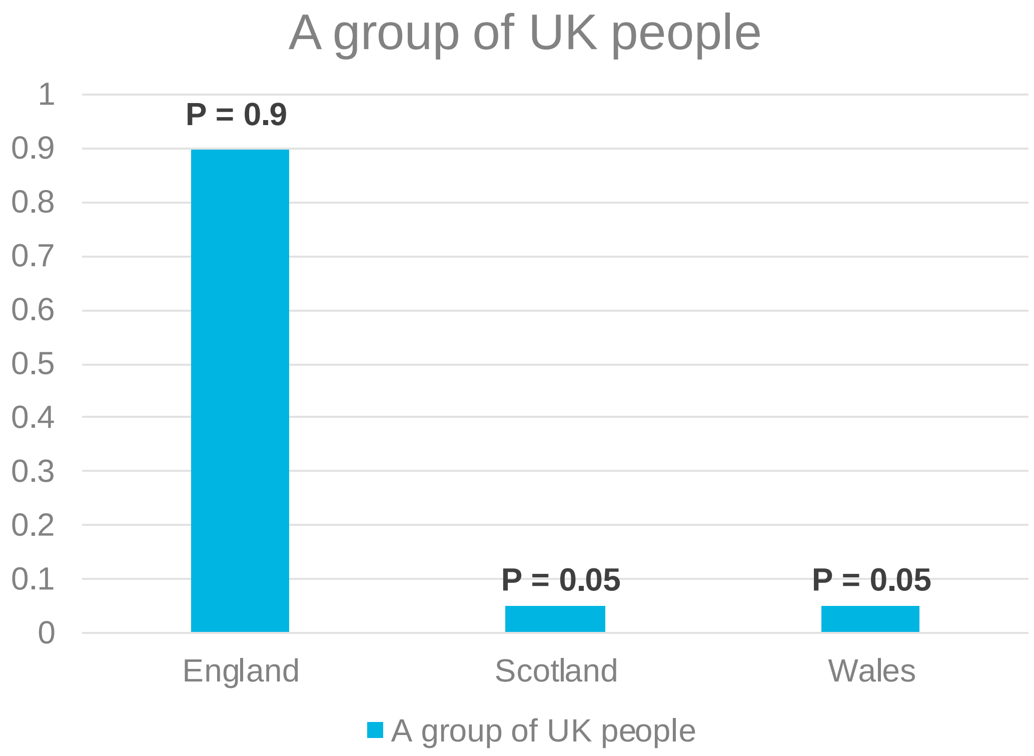
The log-transformed probability of a certain class of events (data)

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

What is Entropy?

A real-life example:

- Skewed distribution
- Lower entropy (higher certainty)



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

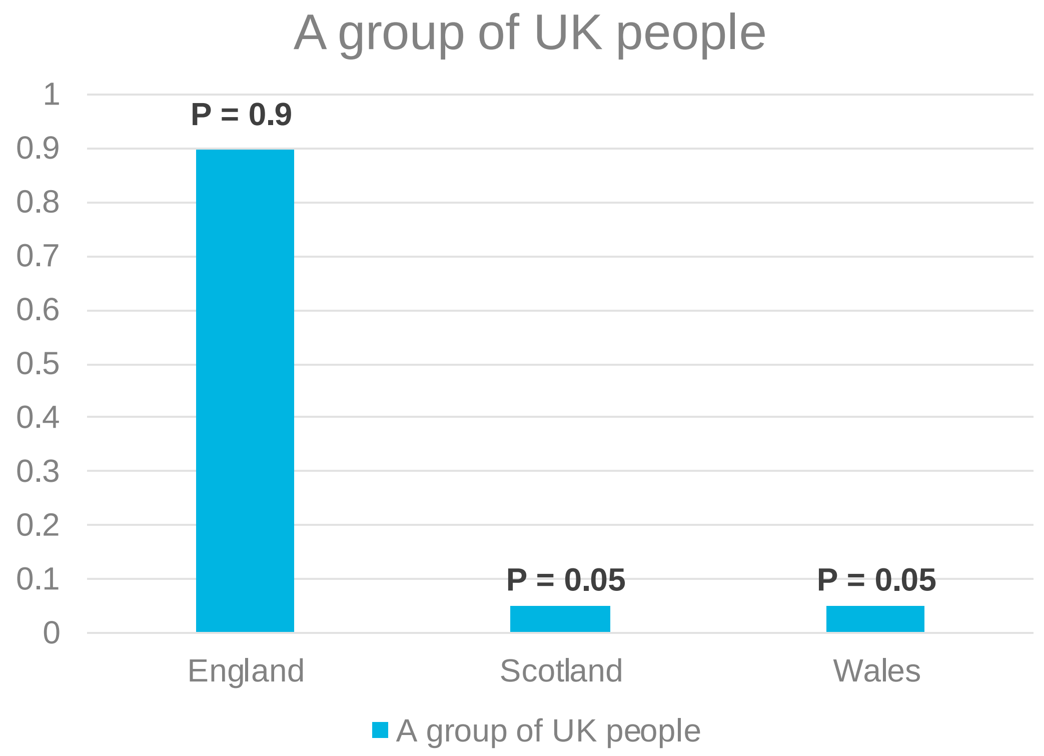
What is Entropy?

A real-life example #1:

$$Entropy = -\sum_{i=1}^{n} P(x_i)\log(P(x_i))$$

$$= -0.9 \times \log(0.9) - 0.05 \times \log(0.05) \\ -0.05 \times \log(0.05)$$

$$= 0.1705$$

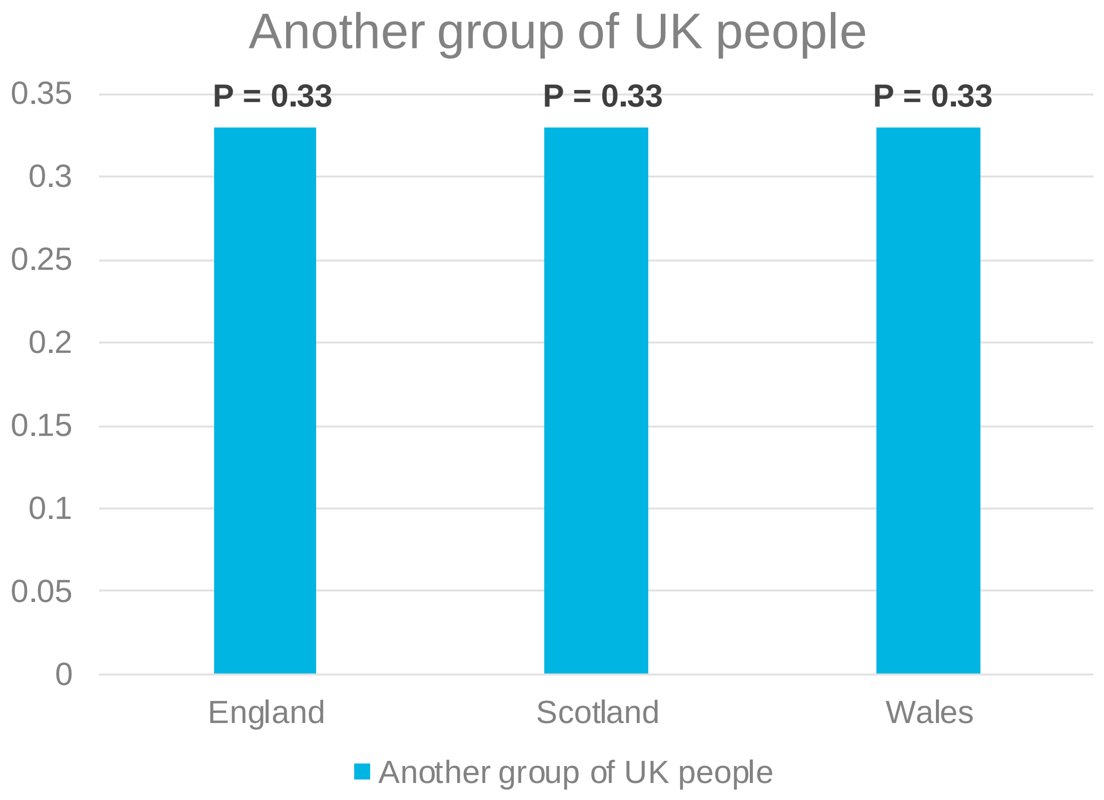


DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

What is Entropy?

Another real-life example:

- Uniform distribution
- Higher entropy (higher uncertainty)



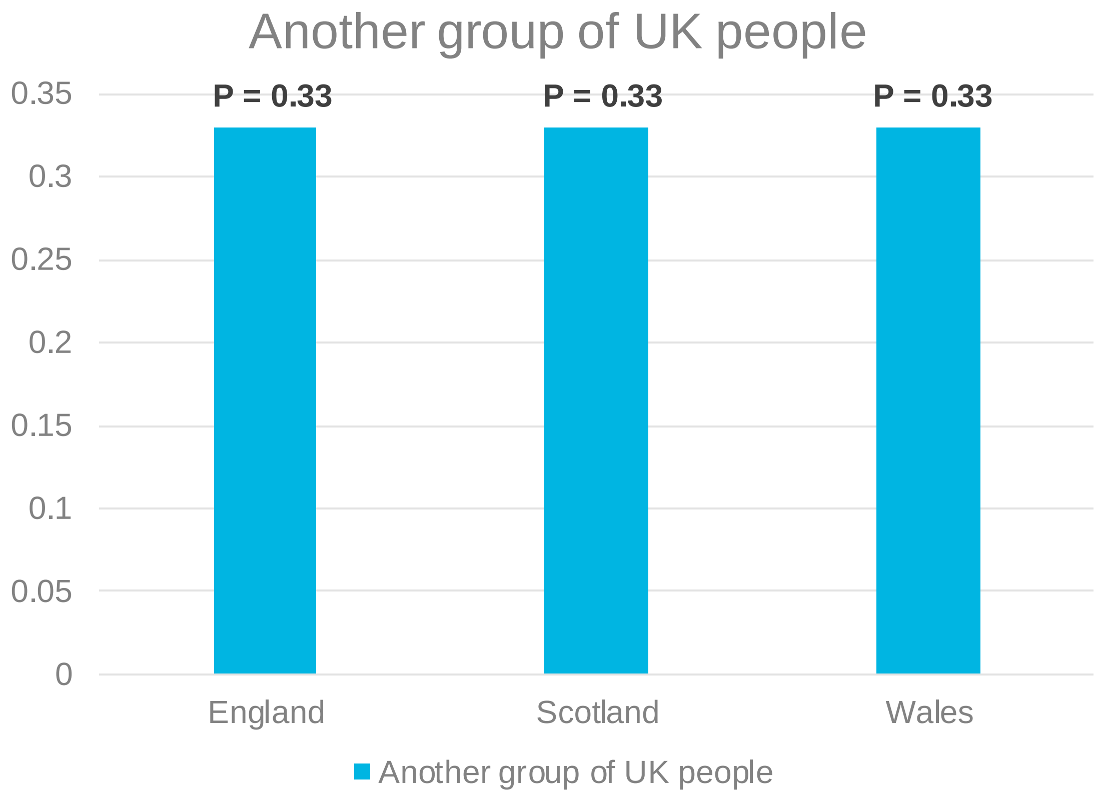
DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

What is Entropy?

A real-life example #2:

$$Entropy = -\sum_{i=1}^{n} P(x_i)\log(P(x_i))$$

$$= \text{It it your turn.}$$



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Intro to Machine Learning Part #1

AGENDA

Machine Learning Basics

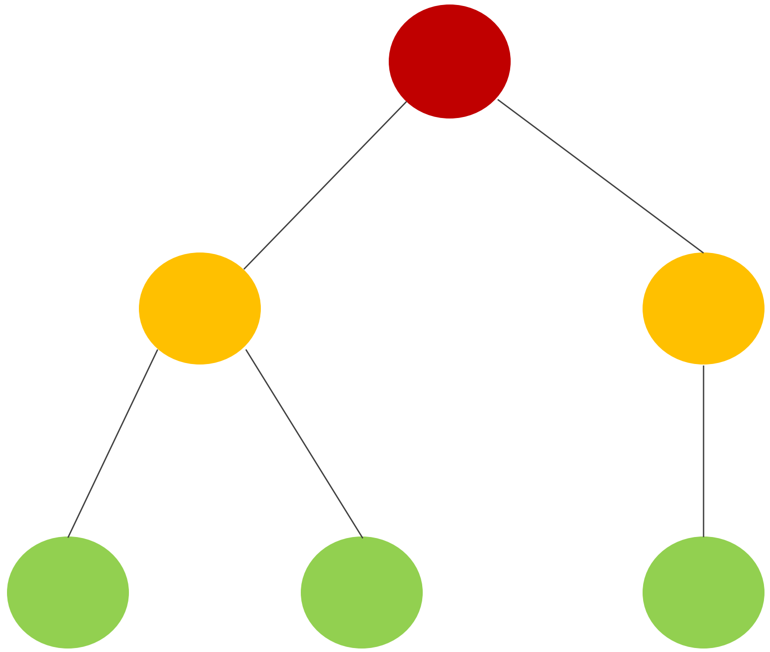
Evaluating Classification: Entropy & Information Gain

Basic Decision Tree

Basic Decision Tree

- “Tree” shape structure
- Split data into smaller groups to minimize entropy

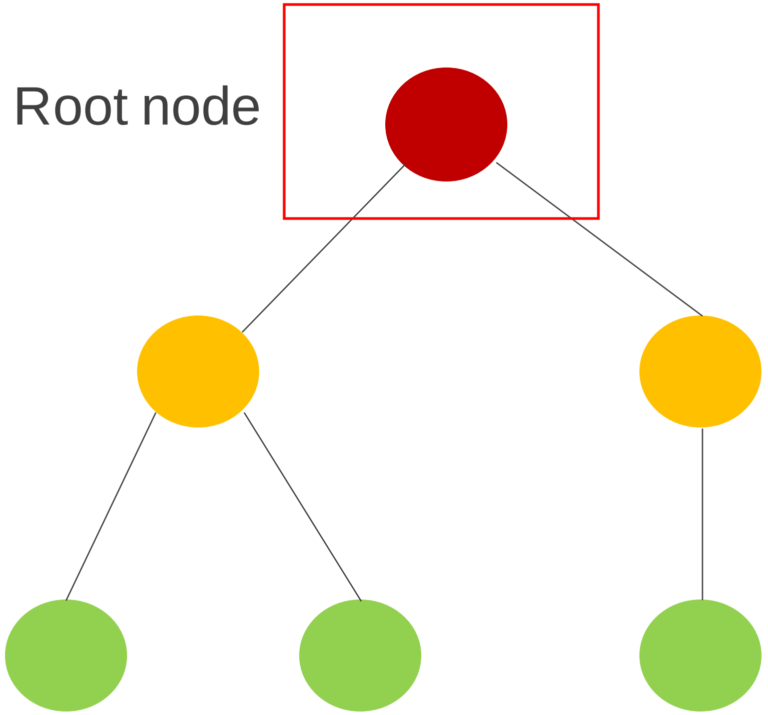
Tree



Decision tree

Basic Decision Tree

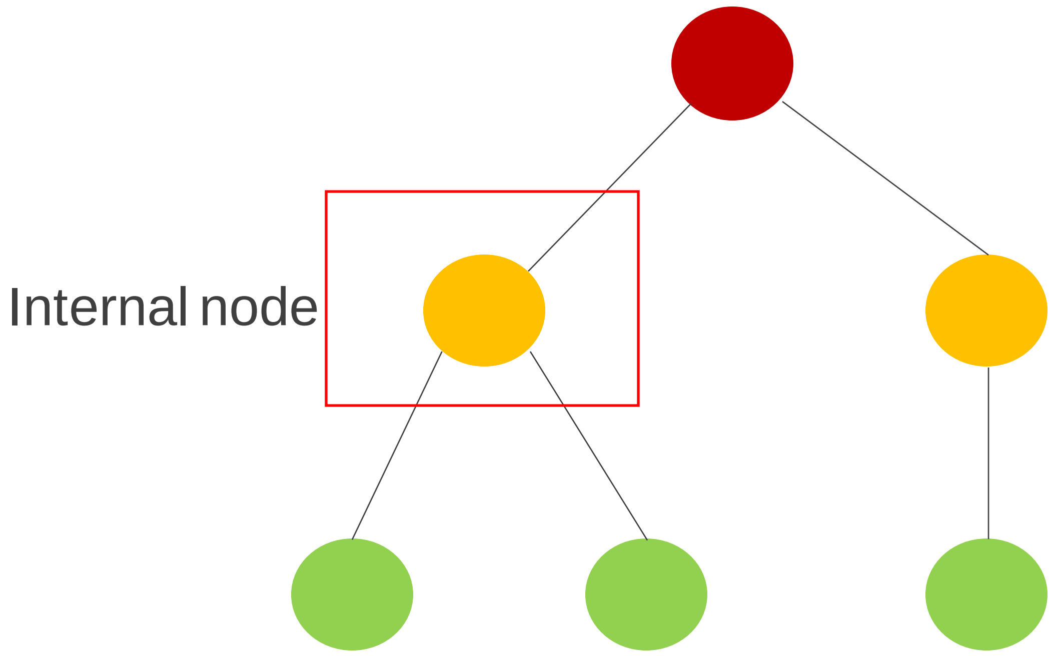
- "Tree" shape structure
- Split data into smaller groups to minimize entropy



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

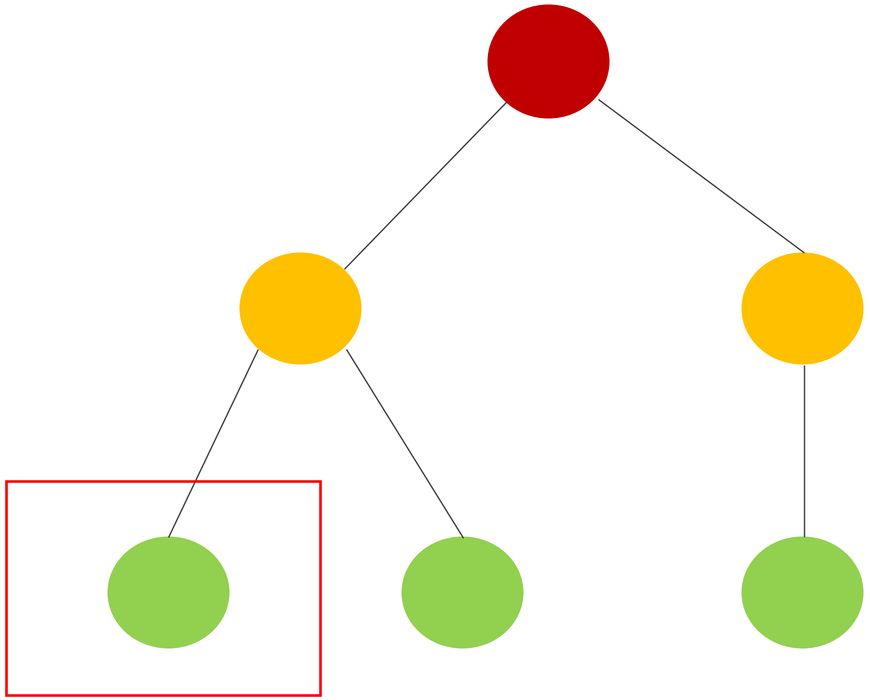
Basic Decision Tree

- "Tree" shape structure
- Split data into smaller groups to minimize entropy



Basic Decision Tree

- “Tree” shape structure



- Split data into smaller groups to minimize entropy

Leaf node

Basic Decision Tree

- A practical example of decision tree: Brexit Dataset

Predictor variables:
- Country (England, Scotland, Wales)
- Gender (Male, Female)
- Age (Young, Old)

Target:
- Leave/Stay

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Basic Decision Tree

- A practical example of decision tree: Brexit Dataset

ID	Country	Gender	Age	Leave
1	Scotland	Male	Old	No
2	Scotland	Male	Old	No
3	England	Male	Young	Yes
4	Wales	Male	Old	Yes
5	Wales	Female	Old	Yes
6	Wales	Female	Young	No

DATA ANALYSIS & SOCIAL INQUIRY LAB GRINNELL COLLEGE

Basic Decision Tree

	Predictor variable			Target
ID	Country	Gender	Age	Leave
1	Scotland	Male	Old	No
2	Scotland	Male	Old	No
3	England	Male	Young	Yes
4	Wales	Male	Old	Yes
5	Wales	Female	Old	Yes
6	Wales	Female	Young	No

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Basic Decision Tree

- Step 1: Entropy at root

Yes: 3   No: 3   In total: 6

Probability(Yes): 0.5 Probability(No): 0.5

$$Entropy(Root) = -\sum_{i=1}^{n} P(x_i)\log(P(x_i))$$

$$= -0.5 \times \log(0.5) - 0.5 \times \log(0.5)$$

$$= 1$$

	Target
ID	Leave
1	No
2	No
3	Yes
4	Yes
5	Yes
6	No

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Basic Decision Tree

- Step 2: Which predictor for initial split?

Yes: 3 No: 3

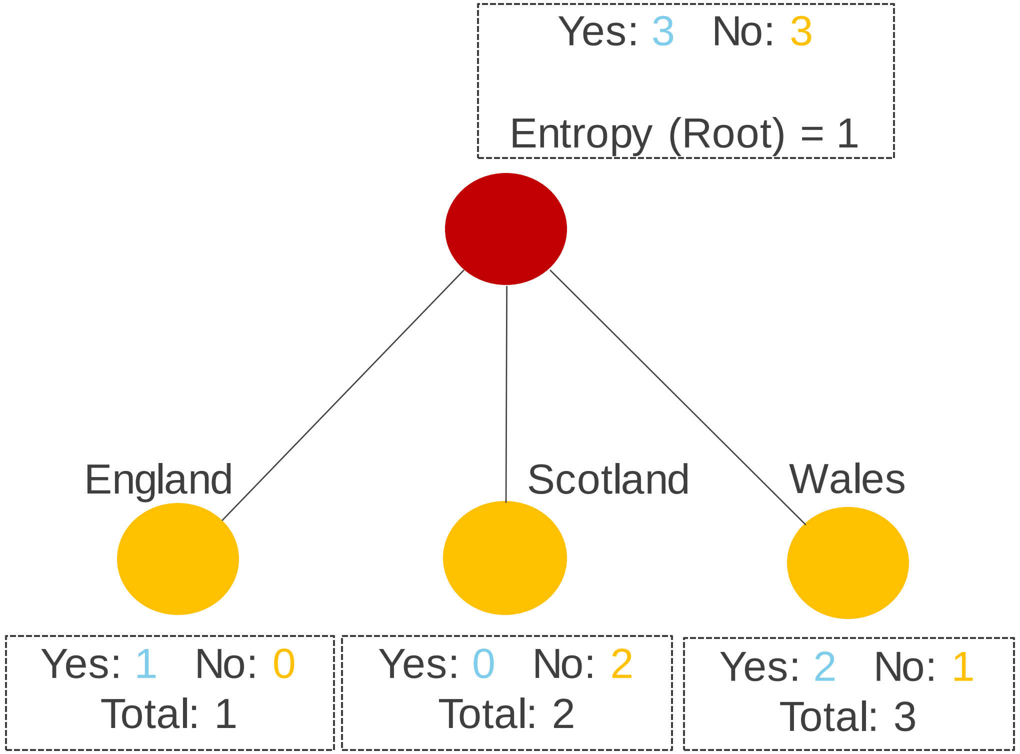
Entropy (Root) = 1

Predictor variable

ID	Country	Gender	Age
1	Scotland	Male	Old
2	Scotland	Male	Old
3	England	Male	Young
4	Wales	Male	Old
5	Wales	Female	Old
6	Wales	Female	Young

Basic Decision Tree

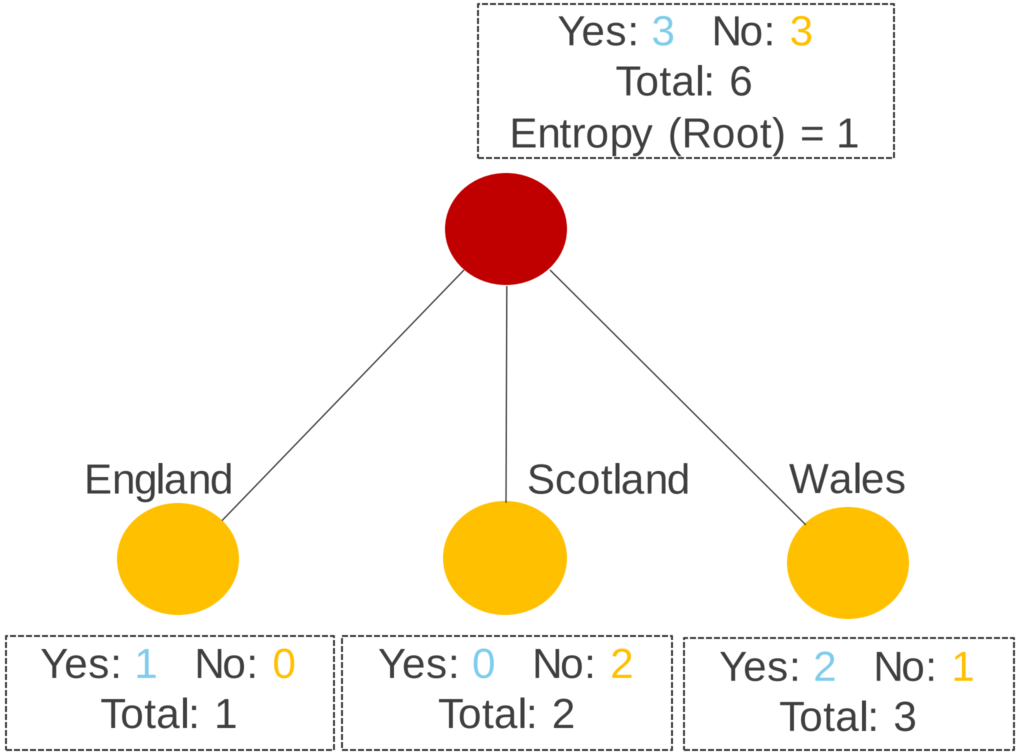
- Step 2: Which predictor for initial split?
  - Country



ID	Country	Leave
1	Scotland	No
2	Scotland	No
3	England	Yes
4	Wales	Yes
5	Wales	Yes
6	Wales	No

Basic Decision Tree

- Step 2: Which predictor for initial split?
  - Country



$$Entropy(England) = -1 \times \log_2 1 - 0 \times \log_2 0 = 0$$

$$Entropy(Scotland) = -0 \times \log_2 0 - 1 \times \log_2 1 = 0$$

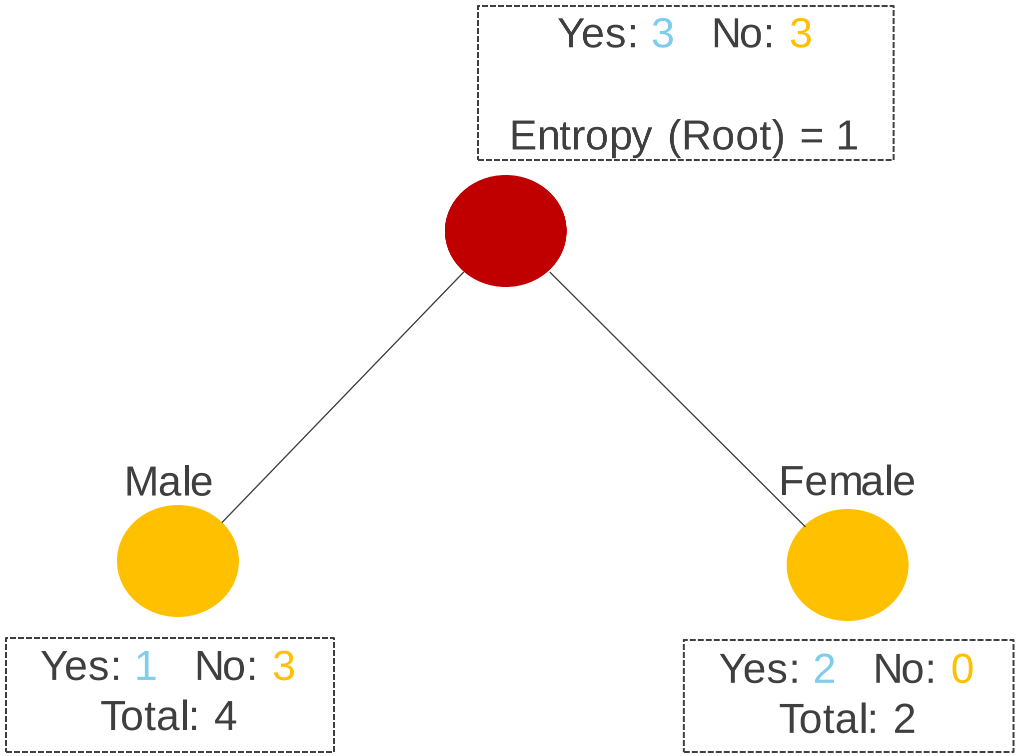
$$Entropy(Wales) = -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3} = 0.9183$$

$$Entropy(Country) = \frac{1}{6} \times 0 + \frac{2}{6} \times 0 + \frac{3}{6} \times 0.9183 = 0.4591$$

DATA ANALYSIS & SOCIAL INQUIRY LAB GRINNELL COLLEGE

Basic Decision Tree

- Step 2: Which predictor for initial split?
  - Gender

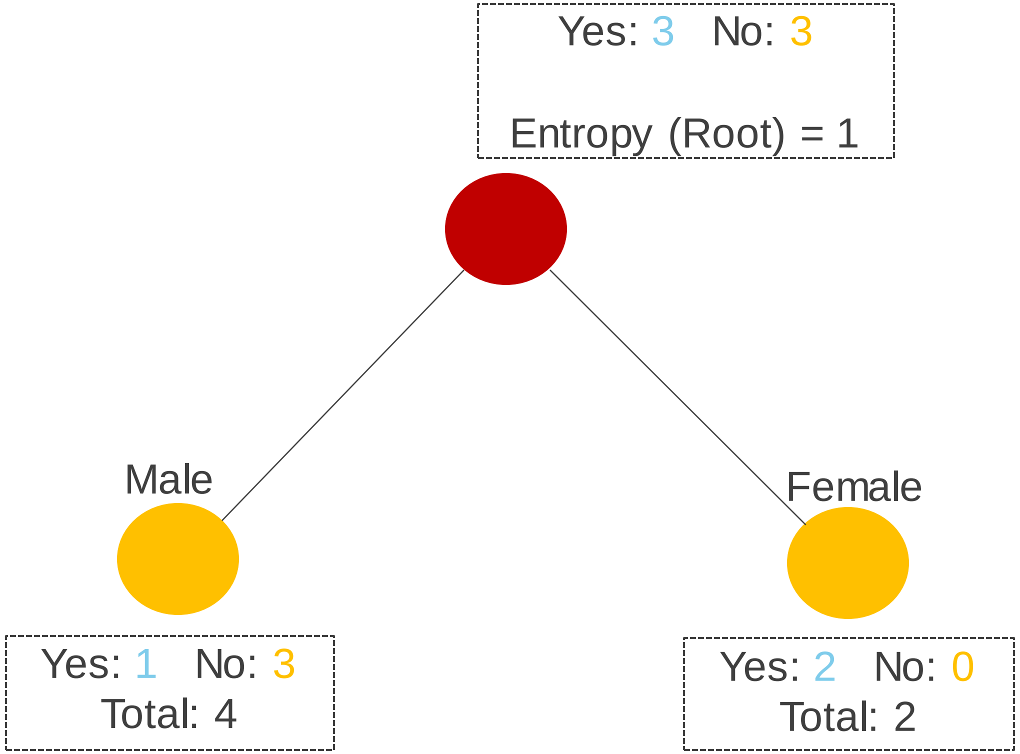


ID	Male	Leave
1	Male	No
2	Male	No
3	Male	Yes
4	Female	Yes
5	Female	Yes
6	Male	No

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Basic Decision Tree

- Step 2: Which predictor for initial split?
  - Gender



$$Entropy(Male) = -\frac{1}{4} \times \log_2 \frac{1}{4} - \frac{3}{4} \times \log_2 \frac{3}{4} = 0.8113$$

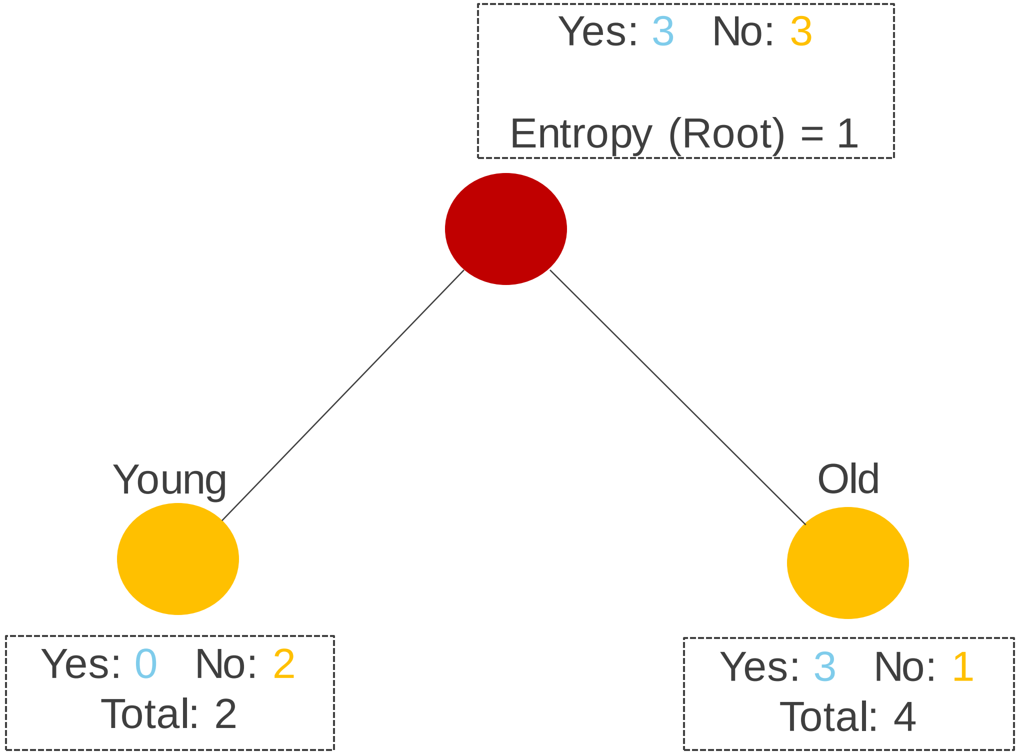
$$Entropy(Female) = -1 \times \log_2 1 - 0 \times \log_2 0 = 0$$

$$Entropy(Gender) = \frac{4}{6} \times 0.8113 + \frac{2}{6} \times 0 = 0.5409$$

DATA ANALYSIS & SOCIAL INQUIRY LAB GRINNELL COLLEGE

Basic Decision Tree

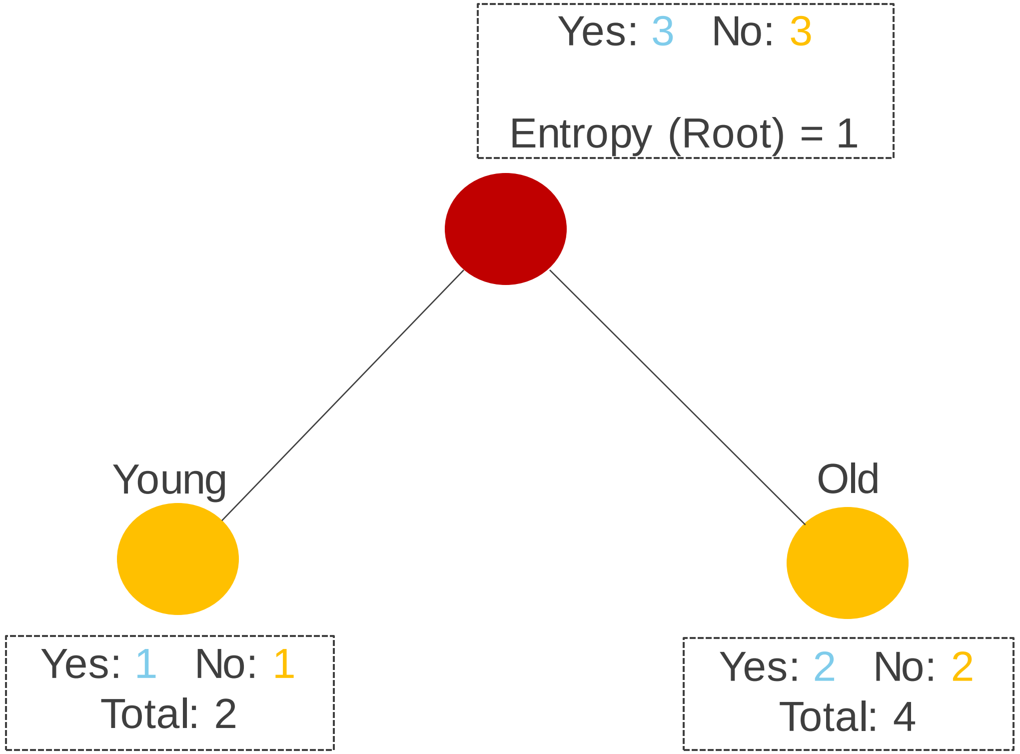
- Step 2: Which predictor for initial split?
  - Age



ID	Age	Leave
1	Old	No
2	Old	No
3	Young	Yes
4	Old	Yes
5	Old	Yes
6	Young	No

Basic Decision Tree

- It is your turn!
  - Age



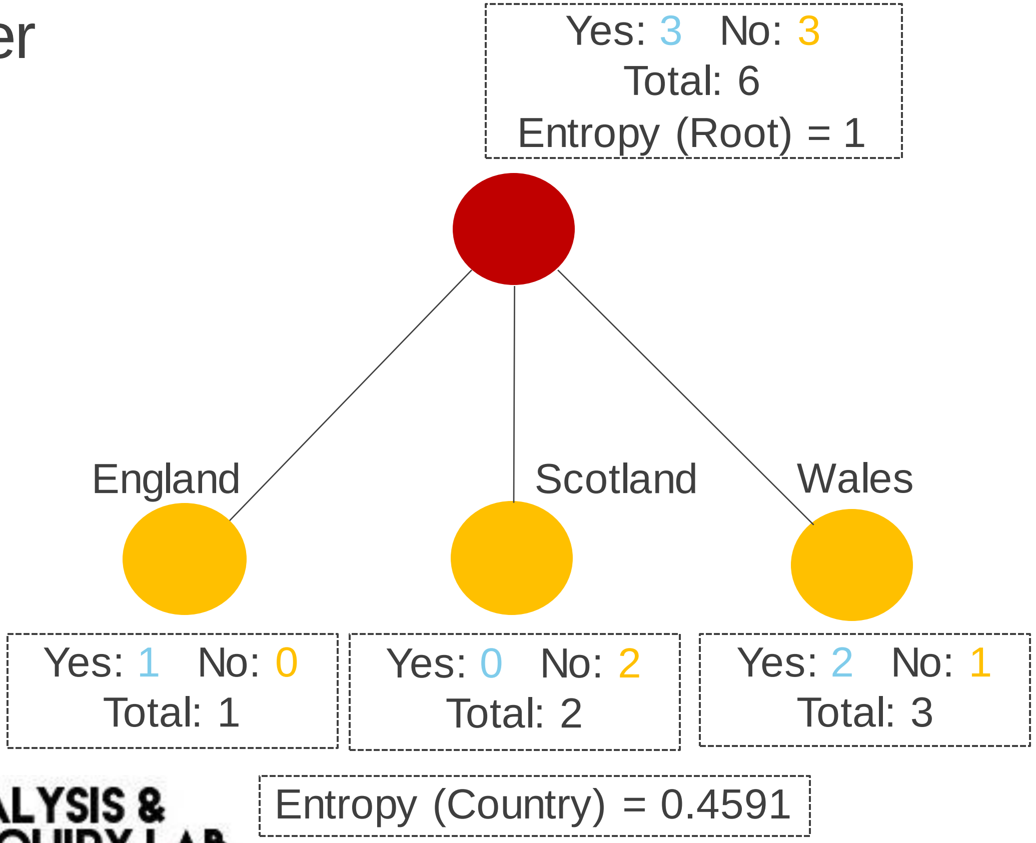
$$Entropy(Young) = ?$$

$$Entropy(Old) = ?$$

$$Entropy(Age) = ?$$

Basic Decision Tree

- Step 3: Now let's compare
  - Country
  - Gender
  - Age



$$Entropy(Root) = 1$$

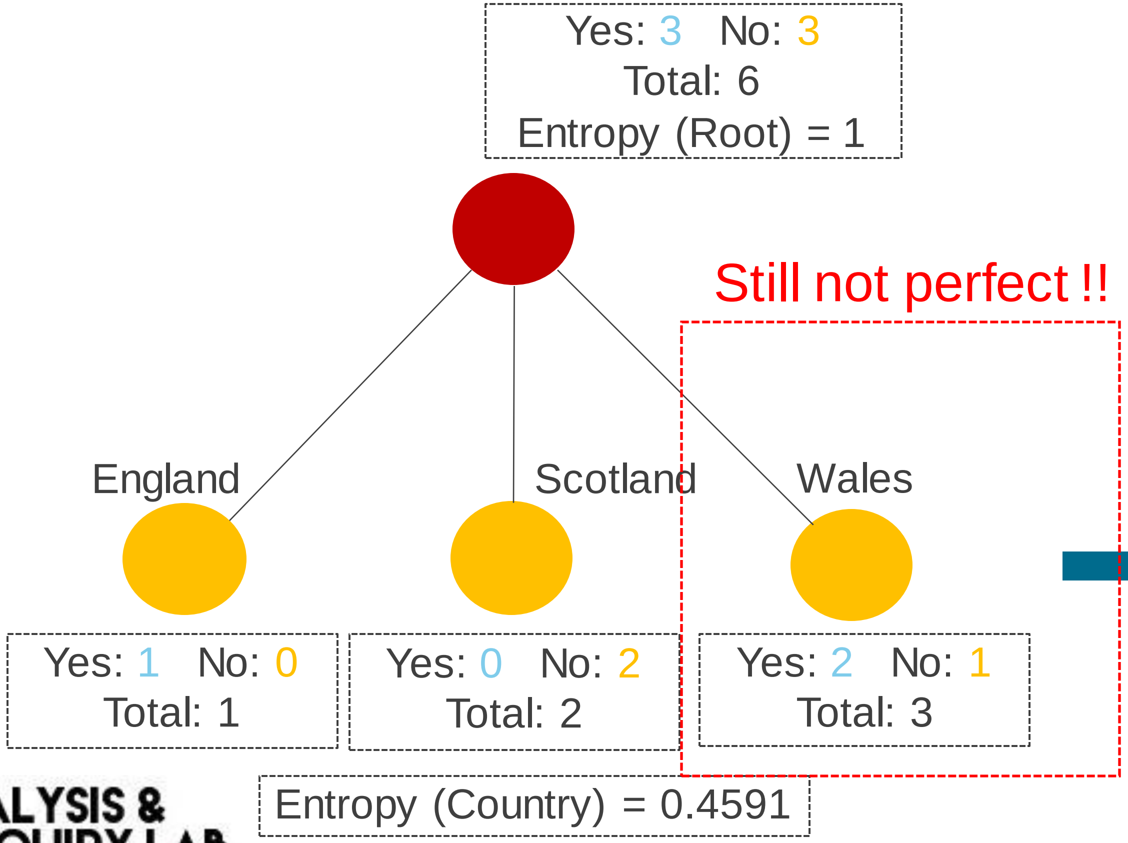
$$Entropy(Country) = 0.4591$$

$$Entropy(Gender) = 0.5409$$

$$Entropy(Age) = 1$$

Basic Decision Tree

- Step 4: Further split



$$Entropy(Root) = 1$$

$$Entropy(Country) = 0.4591$$

ID	Country	Gender	Age	Leave
4	Wales	Male	Old	Yes
5	Wales	Female	Old	Yes
6	Wales	Female	Young	No

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

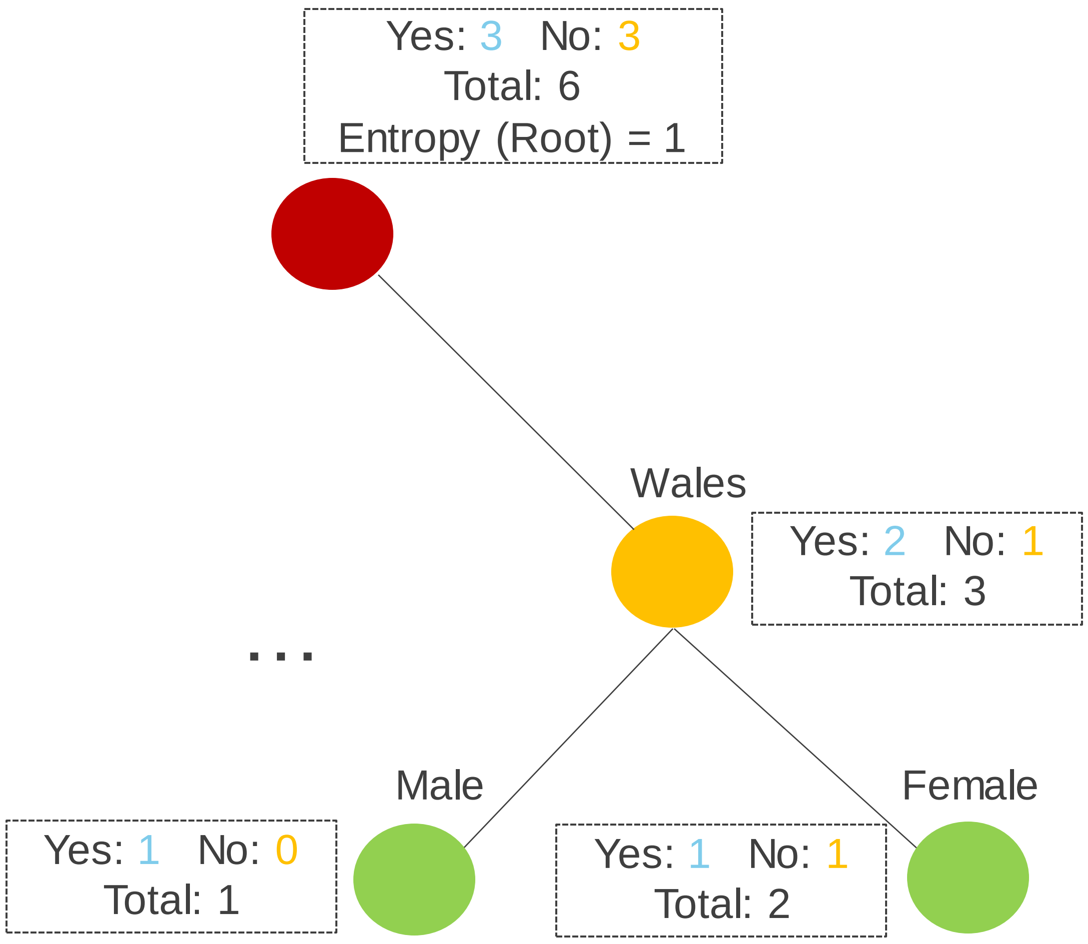
Basic Decision Tree

- Step 4: Further split
  - Gender?

$$Entropy(Male) = -\frac{1}{1} \times \log_2 1 - 0 \times \log_2 0 = 0$$

$$Entropy(Female) = -\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2} = 1$$

$$Entropy(Gen) = \frac{1}{3} \times 0 + \frac{2}{3} \times 1 = 0.667$$



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

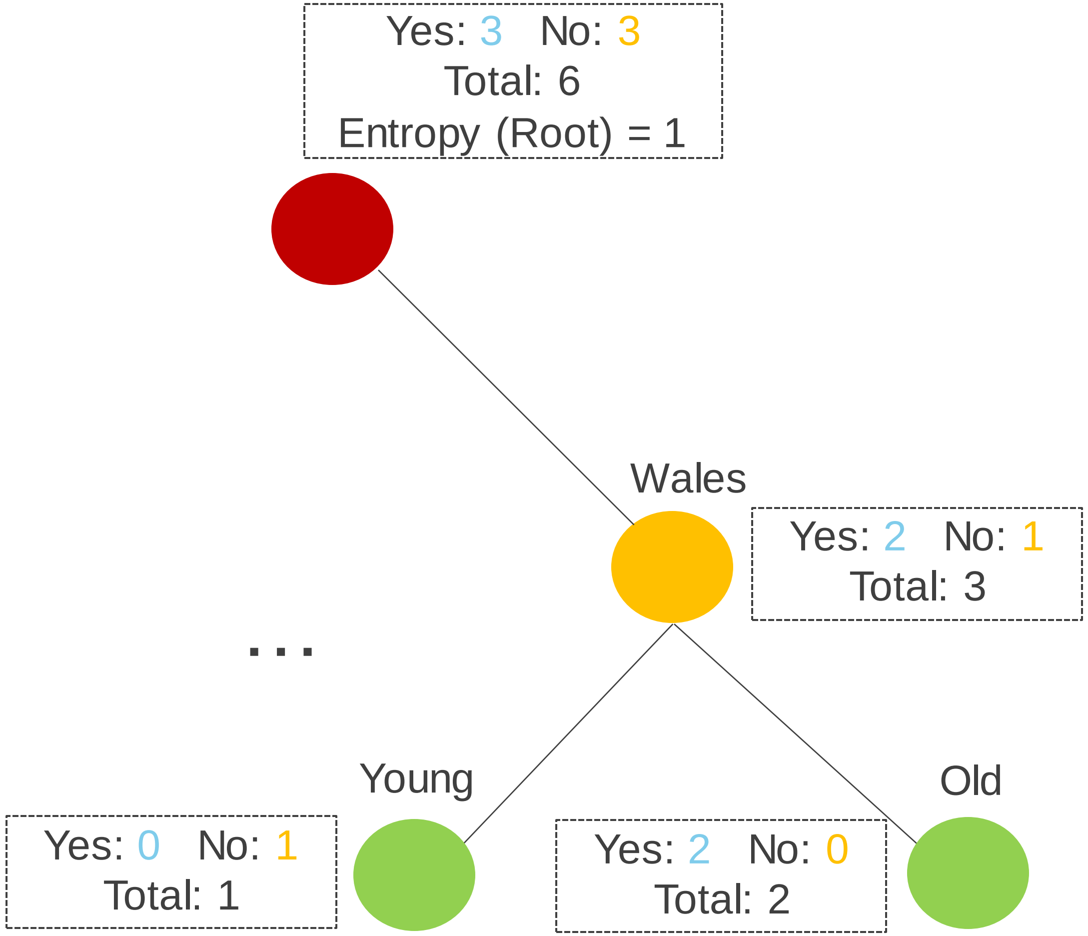
Basic Decision Tree

- Step 4: Further split
  - Age?

$$Entropy(Young) = \color{red}{Your\ turn}$$

$$Entropy(Old) = \color{red}{Your\ turn}$$

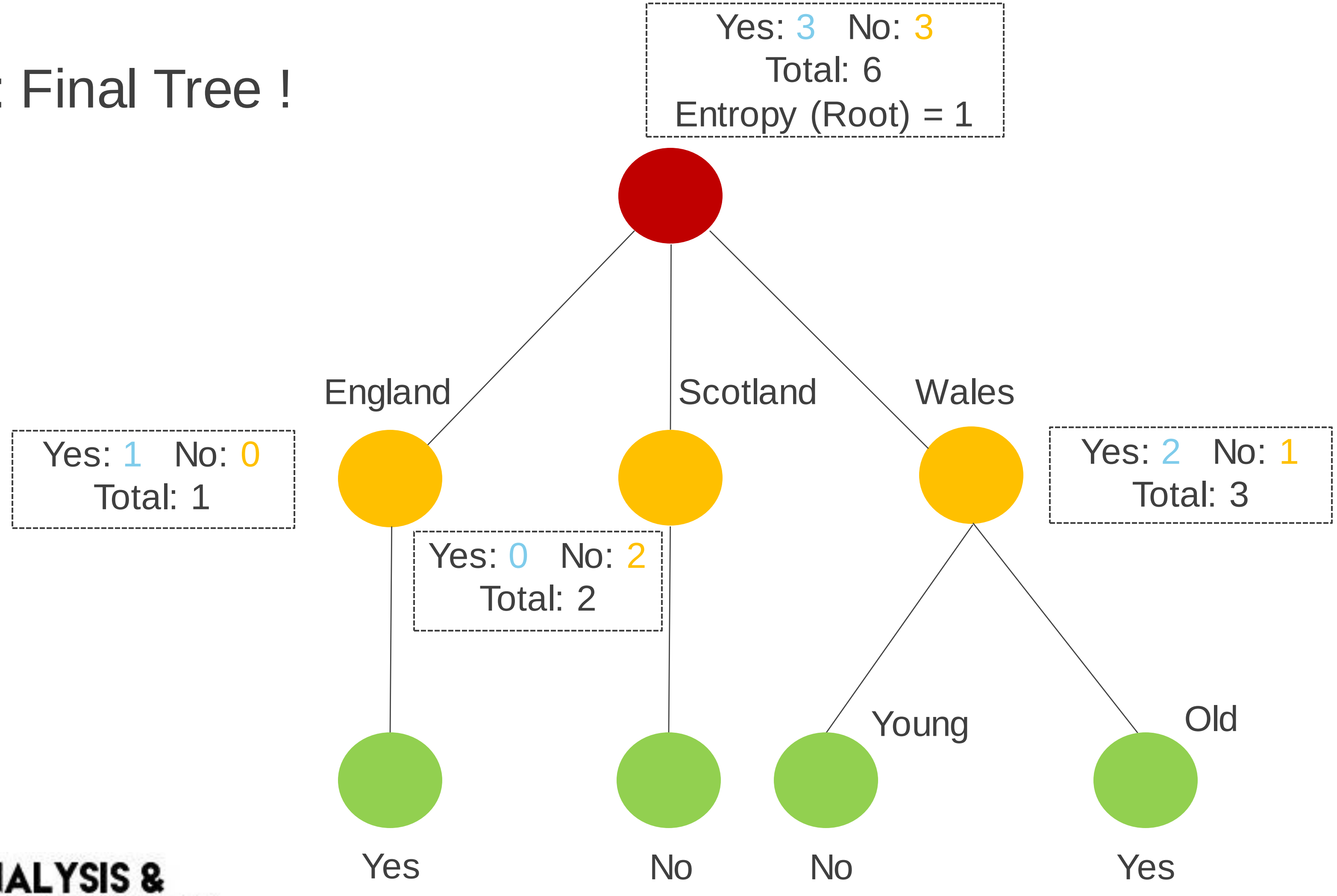
$$Entropy(Age) = \color{red}{Your\ turn}$$



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Basic Decision Tree

- Step 5: Final Tree !



$Entropy = 1$

$Entropy = 0.4591$

$Entropy = 0$!!

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Parameters vs. Hyperparameters

- Parameters are estimated from data
  - e.g. Intercept/slope in linear regression

- Hyperparameters determine what kind of model to fit
- Chosen before you fit the model with data
- Do not depend on the data

Decision Tree Hyperparameters

- Maximum tree depth
- Minimum decrease in entropy needed for a split
- Number of features to consider at each split

Basic Decision Tree

Summary:

- Non-linear model for classification & regression
- Split data into smaller groups to minimize entropy

Q & A