

SSID:
Password:

Register:

<http://bit.ly/wwcodemania>

Github Project:

<https://github.com/wwcodemania/WWCodeManila-ML.AI>

Gitter:

<https://gitter.im/WWCodeManila/Machine-Learning-AI>





WOMEN WHO
CODE
MANILA



Artificial Intelligence Study Group

Twitter: @wwcodemanila
FB: fb.com/wwcodemanila

#WWCodeManila
#AI
#StudyGroup



Issa Tingzon

Research Fellow
PCARI



Our Awesome Mentors

- **Brian Baquiran** – Managing Director for Engineering, Pez AI
- **Marylette Roa** – Researcher at the Philippine Genome Center (PGC)

WOMEN WHO

New Member's Introduction





I am <name>

<your current profession>

<why did you join this study group>

<what's your favorite horror movie/series>

OUR MISSION

Inspiring women to excel in technology careers.



OUR VISION

A world where women are representative as technical executives, founders, VCs, board members and software engineers.



STUDY GROUP

Study groups are events where women can come together and help each other learn and understand a specific programming language, technology, or anything related to coding or engineering.

GUIDELINES

- If you have a question, just **ask**
- If you have an idea, **share it**
- **Make friends** and learn from your study groupmates
- **Do not** recruit or promote your business

TOPIC FOR TODAY

DECISION TREES

Session Resource:

Decision Trees Lecture by Victor Lavrenko (Youtube)

PREREQUISITES

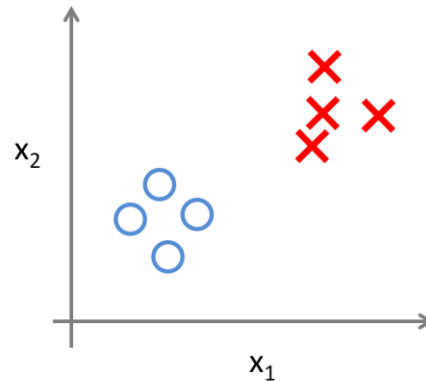
- Knowledge of Python basics
- Accomplished Introduction To Machine Learning
- Understanding of Basic Math Notations

REVIEW

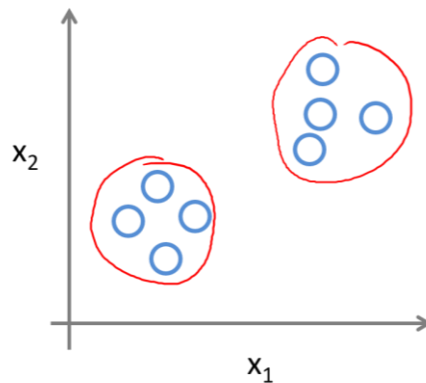
Two types of ML Algorithms:

- **Supervised**
 - Data is labelled
 - Goal: Predict or classify data
- **Unsupervised**
 - Data is unlabelled
 - Goal: Uncover patterns or structure in data

Supervised Learning



Unsupervised Learning



IRIS PLANT CLASSIFICATION

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)

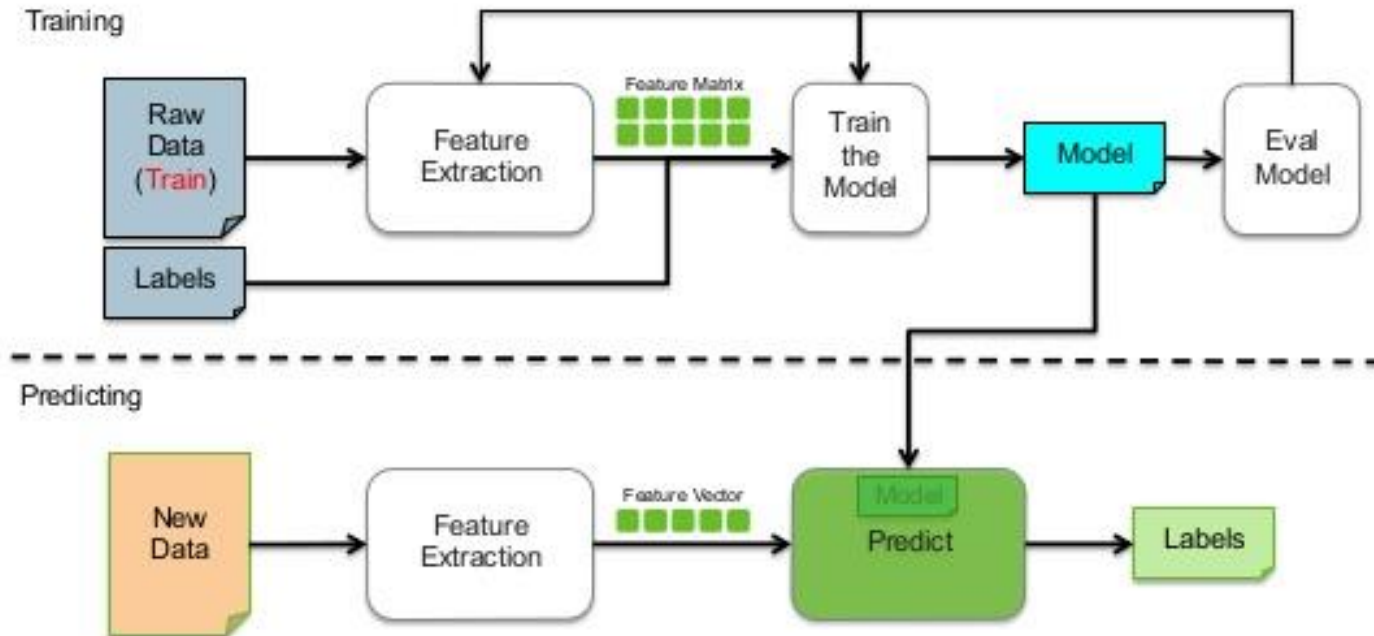
Petal

Sepal



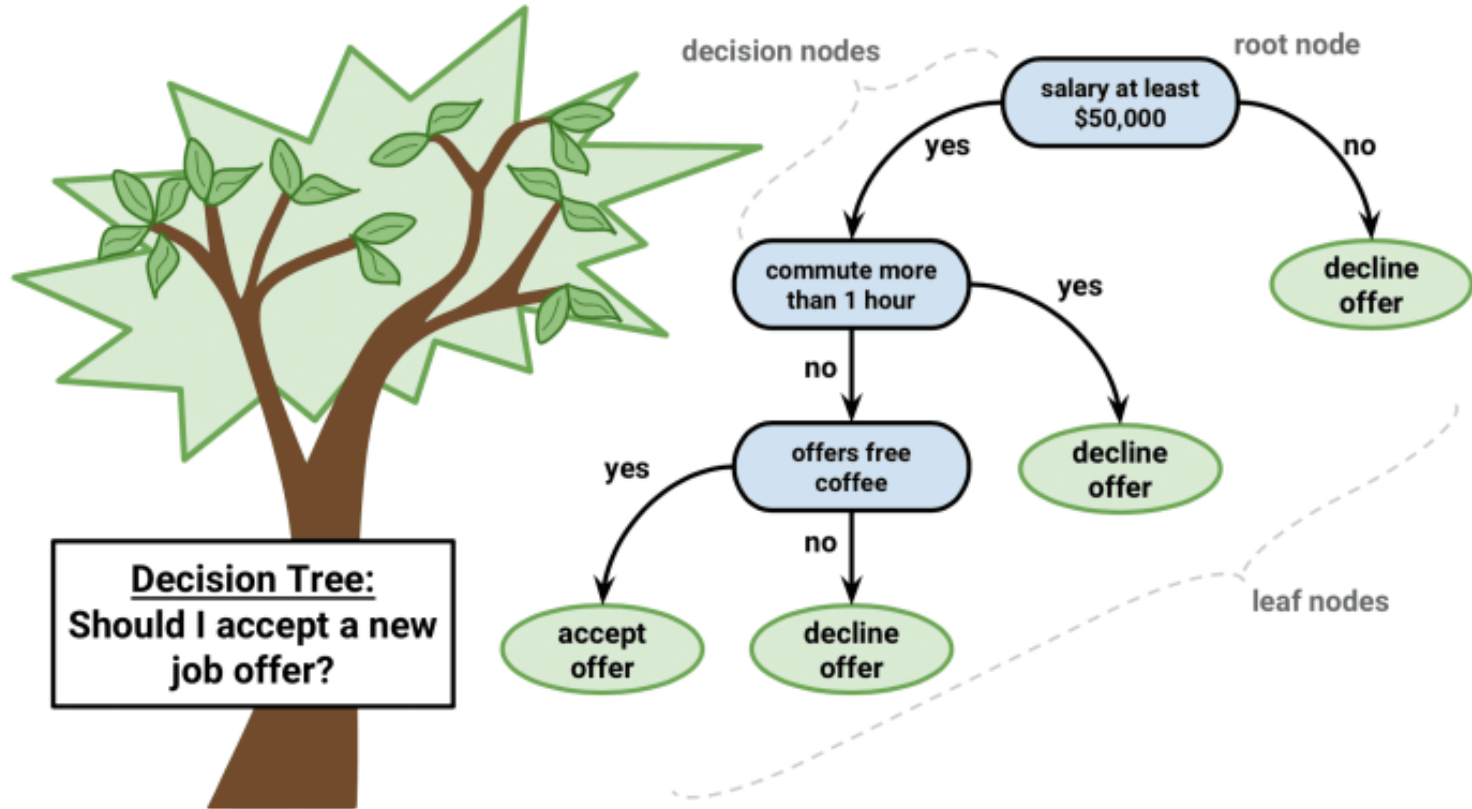
Supervised Learning Workflow

Supervised Learning Workflow



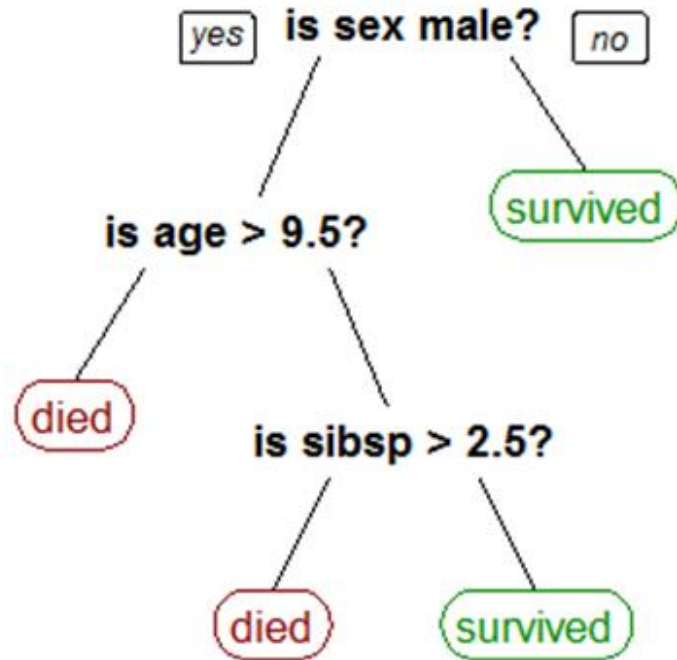
DECISION TREES

- A type of supervised learning algorithm
- Interpretable; mimics human decision making



If Salary > \$50,000 **and** commute is not more than 1 hour **and** offers free coffee, **then** accept offer!

Would you survive the sinking of the Titanic?



TERMINOLOGY

- **Root Node:** represents the entire training set
- **Splitting:** process of dividing a node into two or more subsets/nodes
- **Internal/Decision Node:** corresponds to an attribute
- **Leaf/Terminal Node:** corresponds to a class label

Predict if John will play tennis

Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Predict if John will play tennis

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

New data:

D15	Rain	High	Weak	?
-----	------	------	------	---

Predict if John will play tennis

Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

New data:

D15	Rain	High	Weak	?
-----	------	------	------	---

- Hard to guess
- Try to understand *when* John plays tennis

Predict if John will play tennis

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

New data:

D15	Rain	High	Weak	?
-----	------	------	------	---

- **Divide and Conquer**
 - Split into subsets
 - Are they all “pure”?
 - If yes: stop
 - If not: repeat
- See which subset new data falls into

9 yes / 5 no

Outlook

9 yes / 5 no

Outlook

Sunny

```
graph TD; Outlook([Outlook]) --> Sunny([Sunny]);
```

A decision tree diagram on a light gray background. At the top is a blue oval node labeled 'Outlook'. Above this node is the text '9 yes / 5 no', with '9 yes' in green and '5 no' in red. A blue arrow points from the 'Outlook' node down to a white oval node labeled 'Sunny'.

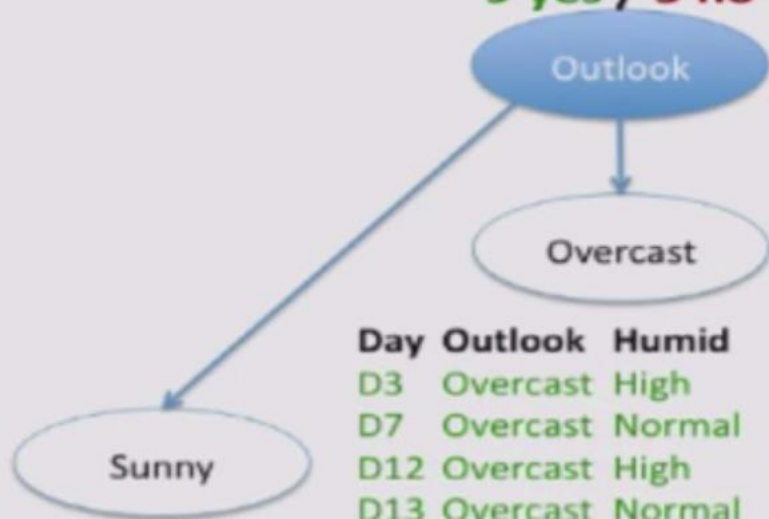
9 yes / 5 no

Outlook

Sunny

Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

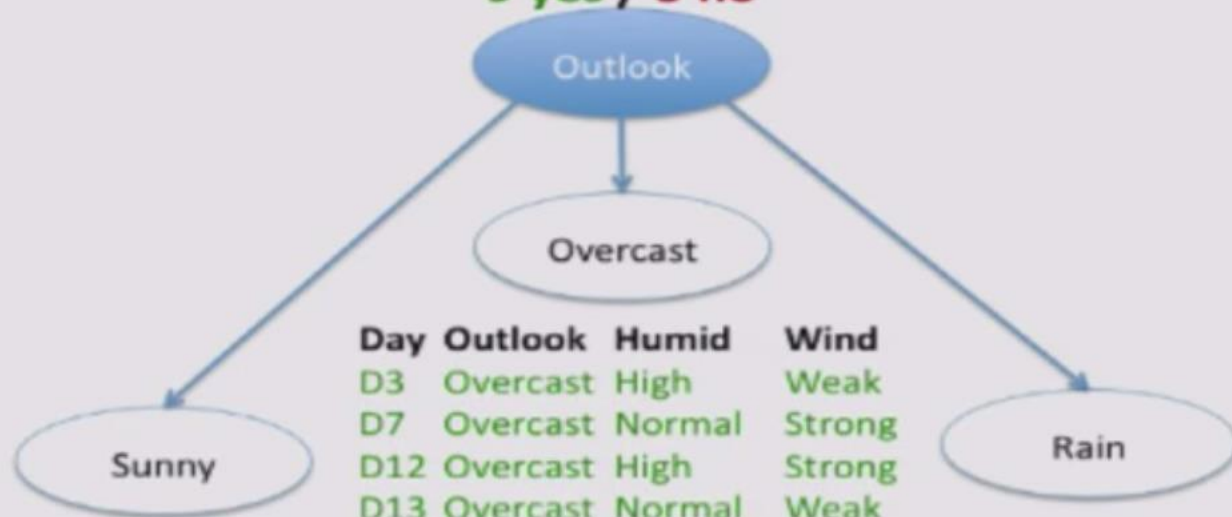
9 yes / 5 no



Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

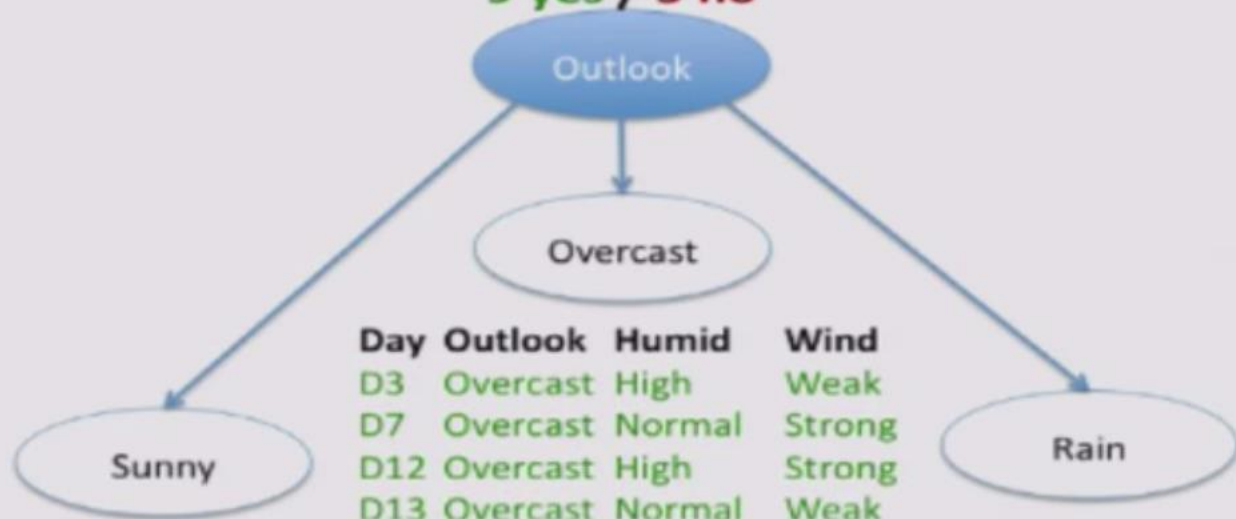
9 yes / 5 no



Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

9 yes / 5 no



Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

4 yes / 0 no
pure subset

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

2 yes / 3 no
split further

3 yes / 2 no
split further

9 yes / 5 no

Outlook

Overcast

Sunny

Rain

Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

2 yes / 3 no
split further

4 yes / 0 no
pure subset

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3 yes / 2 no
split further

9 yes / 5 no

Outlook

Overcast

Sunny

Humidity

High

Normal

Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Rain

4 yes / 0 no
pure subset

Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

Day	Humid	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Day	Humid	Wind
D9	Normal	Weak
D11	Normal	Strong

3 yes / 2 no
split further

9 yes / 5 no

Outlook

Overcast

Sunny

Humidity

High

Normal

Rain

Wind

Weak

Strong

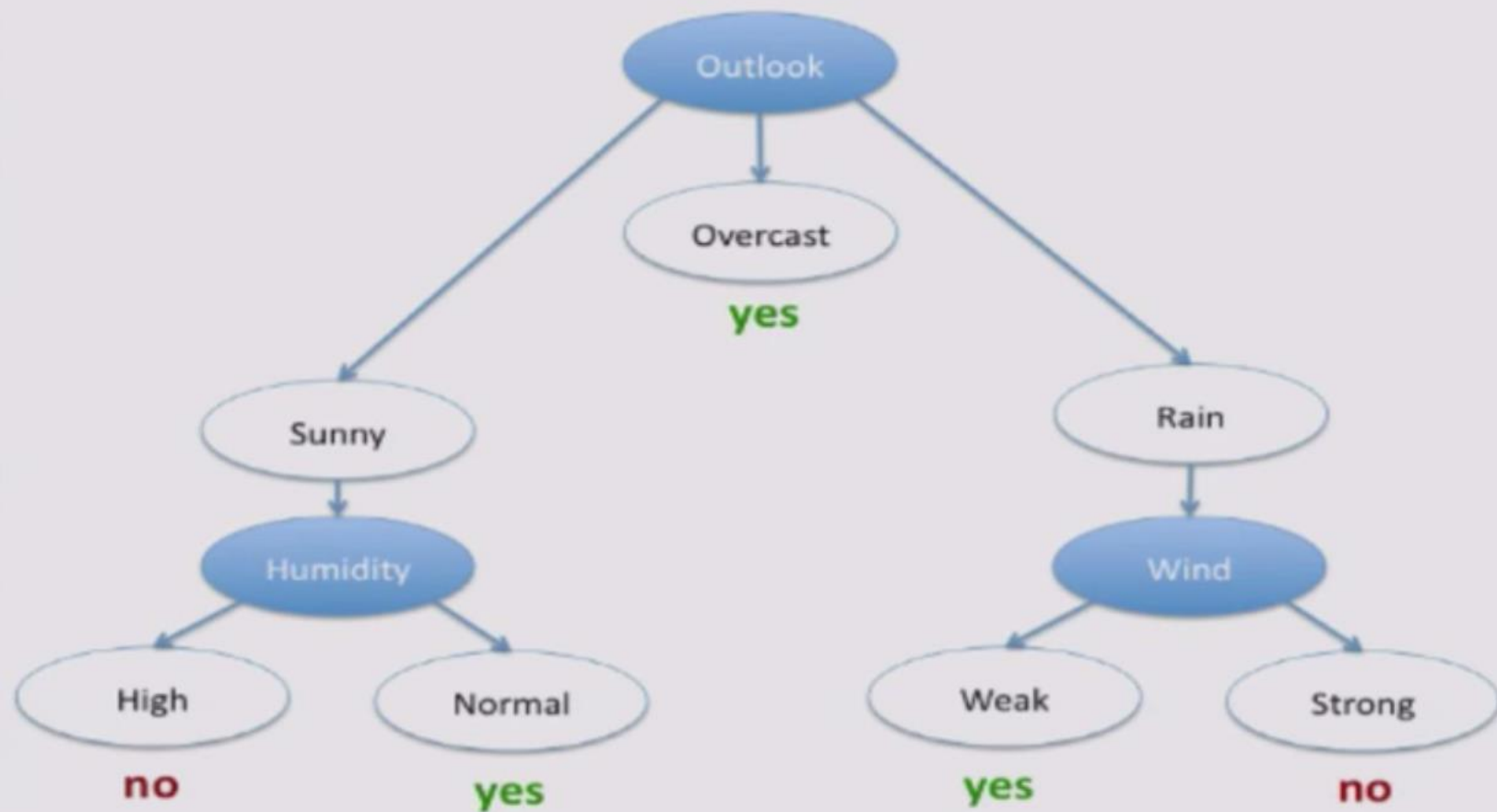
Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

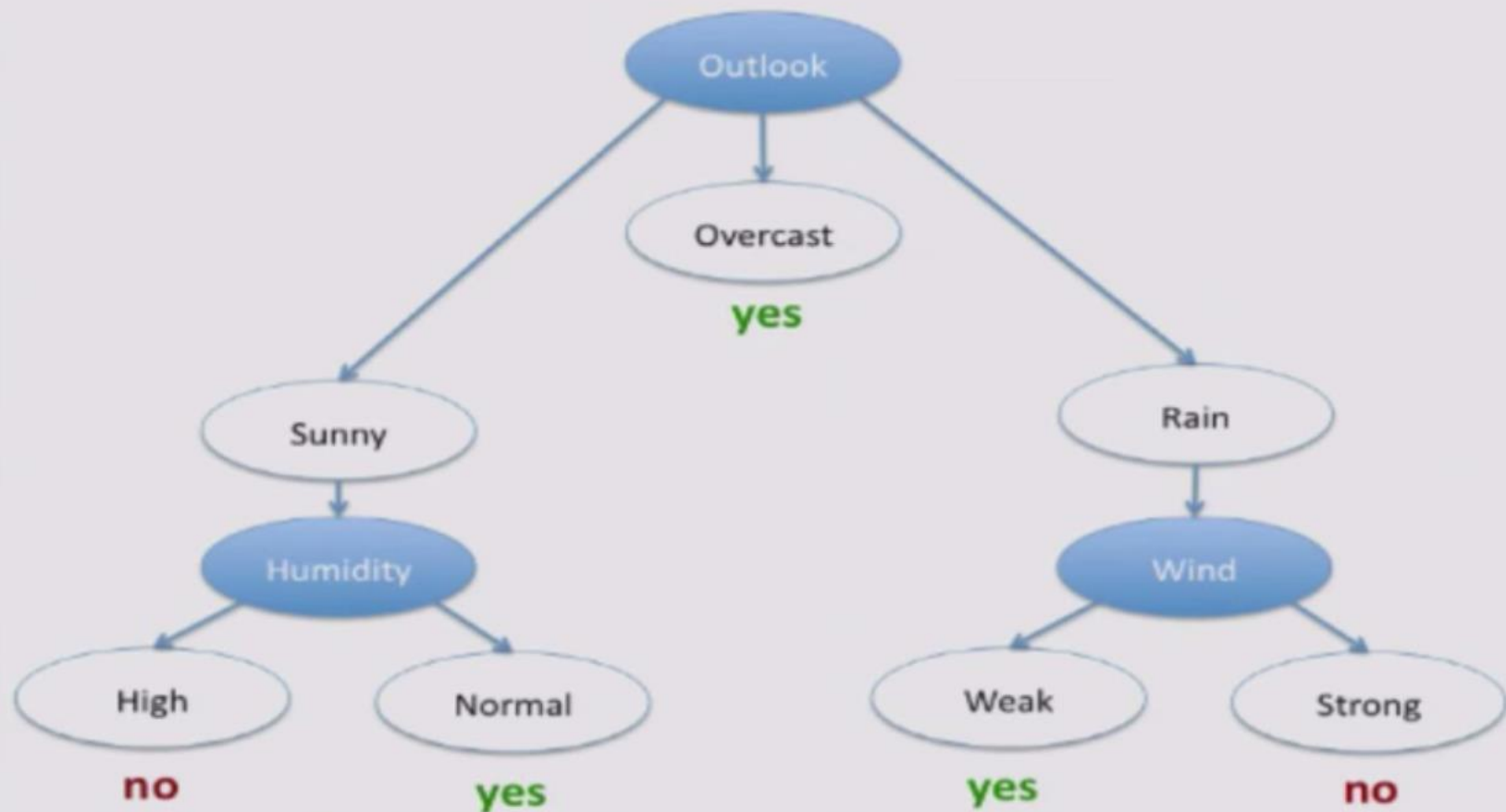
Day	Humid	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Day	Humid	Wind
D9	Normal	Weak
D11	Normal	Strong

Day	Humid	Wind
D4	High	Weak
D5	Normal	Weak
D10	Normal	Weak

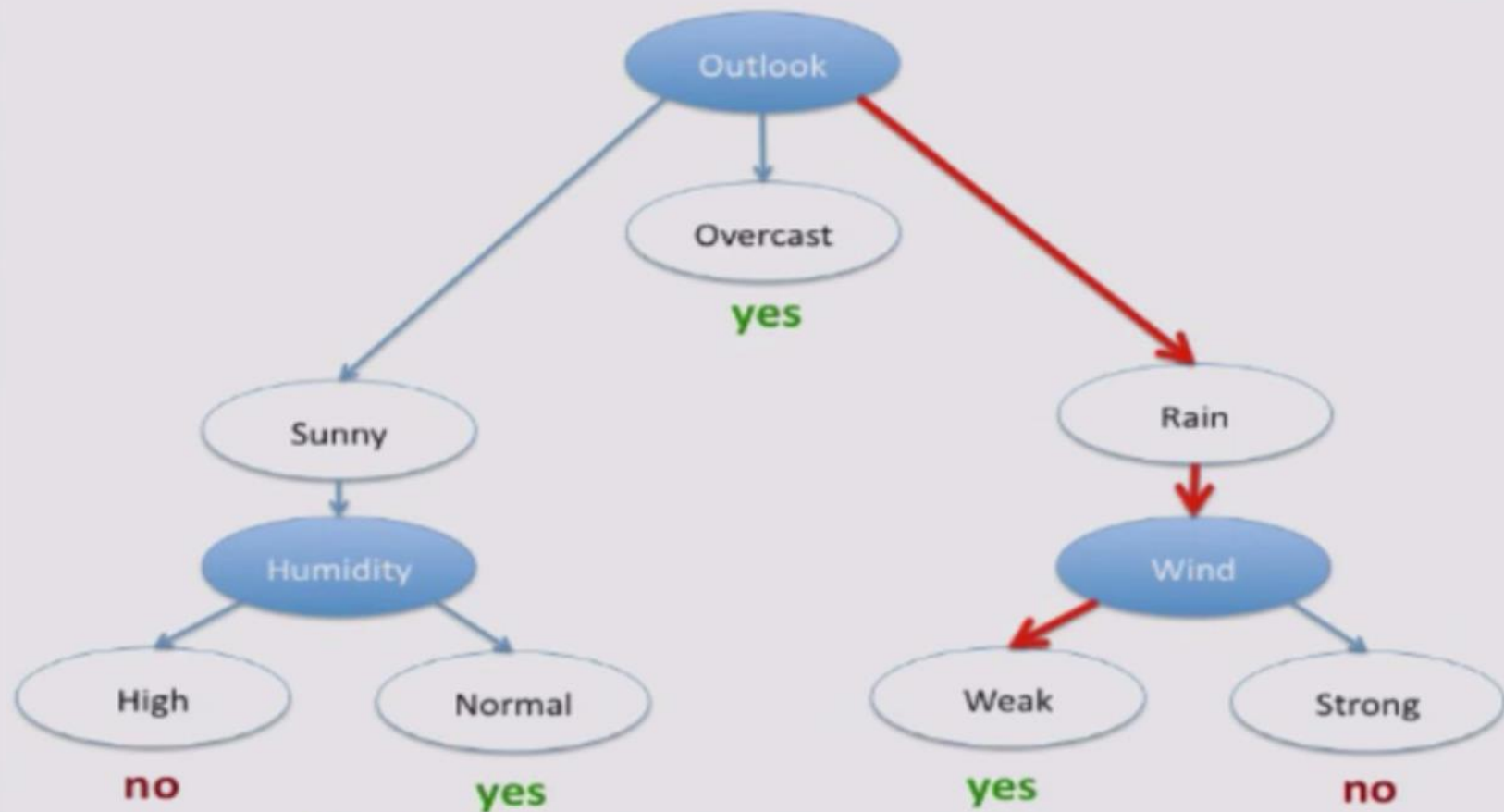
Day	Humid	Wind
D6	Normal	Strong
D14	High	Strong





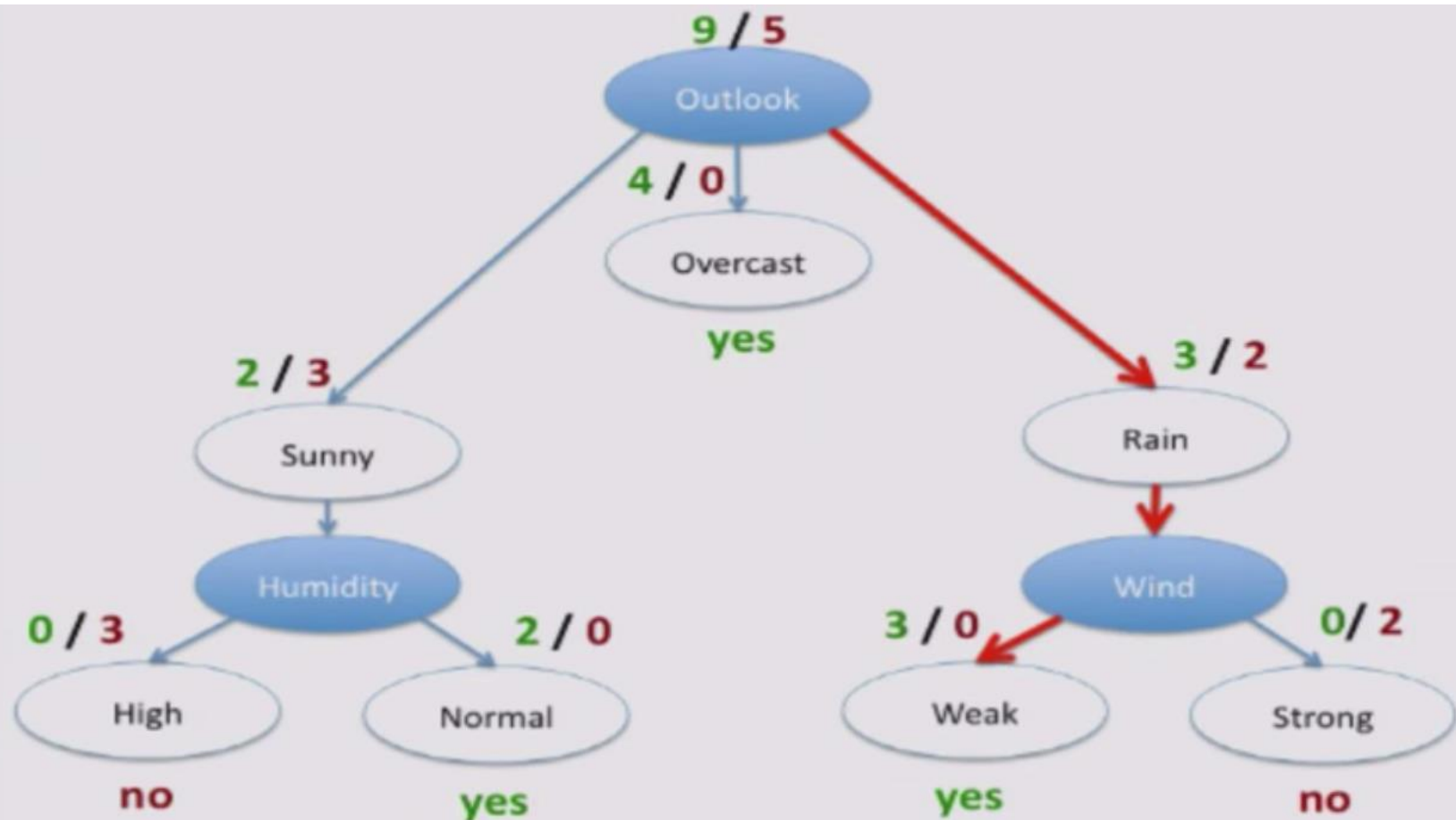
New data:

Day	Outlook	Humid	Wind
D15	Rain	High	Weak



New data:

Day	Outlook	Humid	Wind	
D15	Rain	High	Weak	→ Yes



New data:

Day	Outlook	Humid	Wind	
D15	Rain	High	Weak	→ Yes

ID3 Algorithm

- ID3 - Ross Quinlan, 1986
- Suppose feature A is the **best attribute** to split on.
 - Split entire training set on attribute A
 - For each subset/ child node:
 - If subset is pure: stop
 - Else: split subset

Which attribute to split on?

- Want to measure the “**purity**” of the split
- More certain about Yes/No after the split
 - **Pure set** → (4 Yes / 0 No) 100% certain
 - **Impure Set** → (3 Yes / 3 No) 50 % certain
- **Entropy** – a way to measure certainty
 - Higher entropy → more uncertain

Entropy

$$E(S) = -p_{yes} \log_2 p_{yes} - p_{no} \log_2 p_{no}$$

S – subset of training examples

p_{yes} – proportion of positive (yes) examples

p_{no} – proportion of negative (no) examples

Entropy, more generally

$$E(S) = - \sum_{c \in \text{Classes}} p_c \log_2 p_c$$

S – subset of examples

p_c – proportion of examples in S belonging to class c

Entropy example

e.g. (3 yes / 3 no)

Entropy example

e.g. (3 yes / 3 no)

$$\text{Entropy} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

Entropy example

e.g. (3 yes / 3 no)

$$\text{Entropy} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

e.g. (4 yes / 0 no)

Entropy example

e.g. (3 yes / 3 no)

$$\text{Entropy} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

e.g. (4 yes / 0 no)

$$\text{Entropy} = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

Entropy example

e.g. (3 yes / 3 no)

$$\text{Entropy} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

e.g. (4 yes / 0 no)

$$\text{Entropy} = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

Entropy example

exercise: (9 yes / 5 no)

Entropy example

exercise: (9 yes / 5 no)

Answer: 0.940

Entropy example

exercise: (9 yes / 5 no)

Answer: 0.940

exercise: (3 yes / 4 no)

Answer: 0.985

Entropy

- You can think of **entropy as the measure of disorder or chaos**
- Imagine asking a group of 100 people whether they will vote for Donald Trump or Michelle Obama in the 2020 elections (supposing they run).
 - If half of the people vote for Trump and the other half vote for Obama (50 – Trump / 50 - Obama), then your **entropy is high**. The people don't agree on who should lead them!
 - If everyone votes for Obama (0 – Trump / 100 Obama), then you have stability; thus **entropy is low**. (Same goes if everyone votes for Trump (100 Trump / 0 Obama)). You want stability over disorder.

Entropy

- Entropy tells us how pure **one subset** is.
- But we want a measure of the **effectiveness of an attribute** in classifying the training data.
- We actually want to combine info on different subsets.
- Simply averaging the entropies don't work. (Why not?)

Information Gain

- **Information Gain** – expected reduction in entropy after a split on an attribute

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

A – Attribute

$Values(A)$ – possible values of A

S – subset of training examples

S_v – subset of S for which attribute A have value v

Information Gain

e.g. Find $Gain(S, Wind)$.

Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Information Gain

e.g. Find $Gain(S, Wind)$.

Answer: 0.048

ex. Find $Gain(S, Humidity)$

Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Information Gain

e.g. Find $\text{Gain}(S, \text{Wind})$.

Answer: 0.048

ex. Find $\text{Gain}(S, \text{Humidity})$

Answer: 0.151

ex. Find $\text{Gain}(S, \text{Outlook})$.

Training examples: 9 yes / 5 no				
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Information Gain

e.g. Find $\text{Gain}(S, \text{Wind})$.

Answer: 0.048

ex. Find $\text{Gain}(S, \text{Humidity})$

Answer: 0.151

ex. Find $\text{Gain}(S, \text{Outlook})$.

Answer: 0.246

Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

EXERCISE OPTIONS

1. Solve by hand / calculator (takes a lot of time!)
2. Use python to write function(s) that compute the entropy, gini index, and information gain.
3. Advanced: Write a program that builds a decision tree!
 - Requires some knowledge on recursion, ADTs



Partner/Group/Individual Exercise

Predicting virus infection in files:

	WRITABLE	UPDATED	SIZE	CLASS
1	yes	no	small	infected
2	yes	yes	large	infected
3	no	yes	med	infected
4	no	no	med	clean
5	yes	no	large	clean
6	no	no	large	clean



Challenge

Classify the type of vegetation that is likely to grow in areas of land based on descriptive feature.

ID	Stream	Slope	Elevation	Vegetation
1	False	Steep	High	Chapparal
2	True	Moderate	Low	Riparian
3	True	Steep	Medium	Riparian
4	False	Steep	Medium	Chapparal
5	False	Flat	High	Conifer
6	True	Steep	Highest	Conifer
7	True	Steep	High	Chapparal

WOMEN WHO
CODE

What type of vegetation would likely grow on terrain with a steep slope and medium elevation and is near a stream?

How about on terrain near a stream with high elevation, moderate slope?



WOMEN WHO

Partner/Group/Individual Presentation

CODE



Next time

- More on decision trees - pros and cons
- **Random Forests!**

References

Decision Tree Lecture by Victor Lavrenko (Youtube)

T.I.L.

SHARE IT!
In front!

On Twitter: @wwcodemanila
Or FB: fb.com/wwcodemanila

Don't forget to tag WWCodeManila so we can retweet or share it.

Feedback Form

<https://goo.gl/YzSqcS>

Please don't rate the event on meetup.

Not helpful. It is best to just tell your concerns via the feedback form. We are building a community not a Yelp restaurant.

WOMEN WHO

THANK YOU :)

CODE®