Edge Hill University

</TECHUP_WOMEN>

Introduction to Machine Learning

Topic 5, Module 1, Part 1

Duration: 1 Hour

Durham University

Edge Hill University

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

University of York

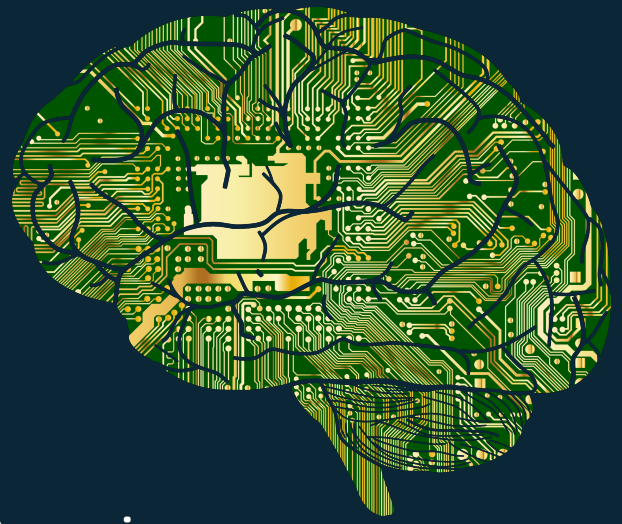
institute of CODING

Office for Students

Edge Hill University

1. Artificial Intelligence

</TECHUP_WOMEN>

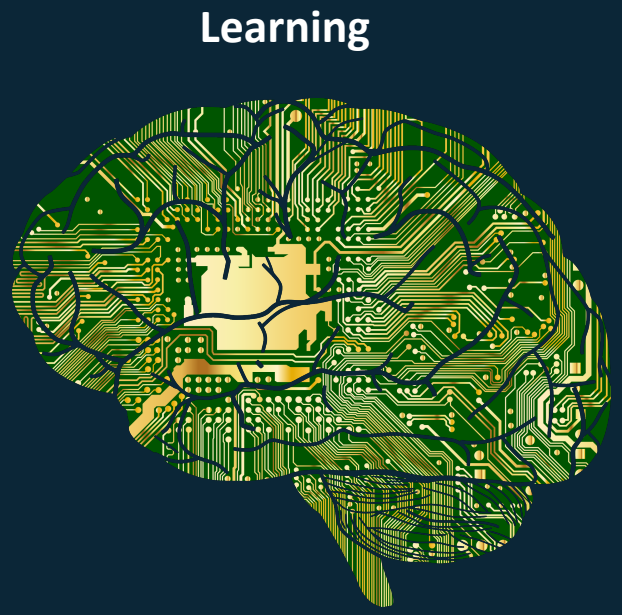


- Artificial Intelligence (A.I.) is a field of study concerned with reproducing/replicating human intelligence.
- A.I. is an umbrella term – it is comprised of many sub-fields of study.
- Here we'll talk about replicating the human capacity for learning.
- This field is known more generally as "Machine Learning".

Edge Hill University

2. Machine Learning

</TECHUP_WOMEN>



- Machine learning (ML) is a fascinating field. It combines insights from statistics, logic, psychology and neuroscience to build automated systems capable of “learning” by themselves.
- ML isn’t concerned with replicating exactly how humans learn, after all, we’re flawed!
- Instead it focuses on developing ways to make optimal decisions/predictions using available information.
- During this module you’ll acquire an understanding of ML that will allow you to apply it moving forward.

Edge Hill University

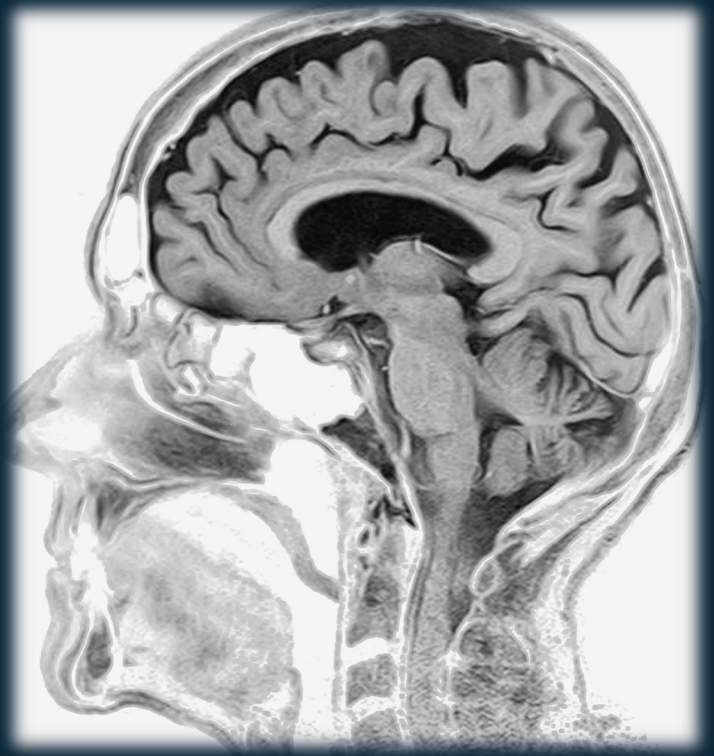
3. Example Applications

</TECHUP_WOMEN>

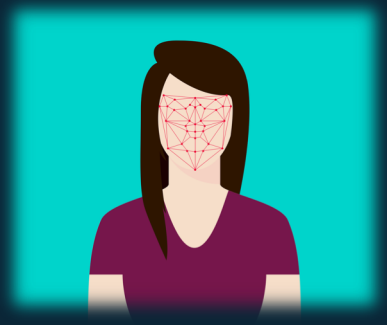
Searching for rare stars in astronomy data

Image Credit: SARAO

Identifying regions of disease in medical images



Climate Modelling



Facial Recognition

Voice Recognition

Uncredited images obtained from https://pixabay.com (Creative Commons License, no attribution required).

Durham University · Edge Hill University · The University of Nottingham UNITED KINGDOM · CHINA · MALAYSIA · University of York · institute of CODING · Office for Students

Edge Hill University

4. What we'll cover

</TECHUP_WOMEN>

This module will introduce…

- Useful terminology
- Key concepts

} Part 1

- Some basic mathematical background
- Our first learning system

} Part 2

- A number of machine leaning algorithms from first principles
- Examples you can try for yourself

} Part 3

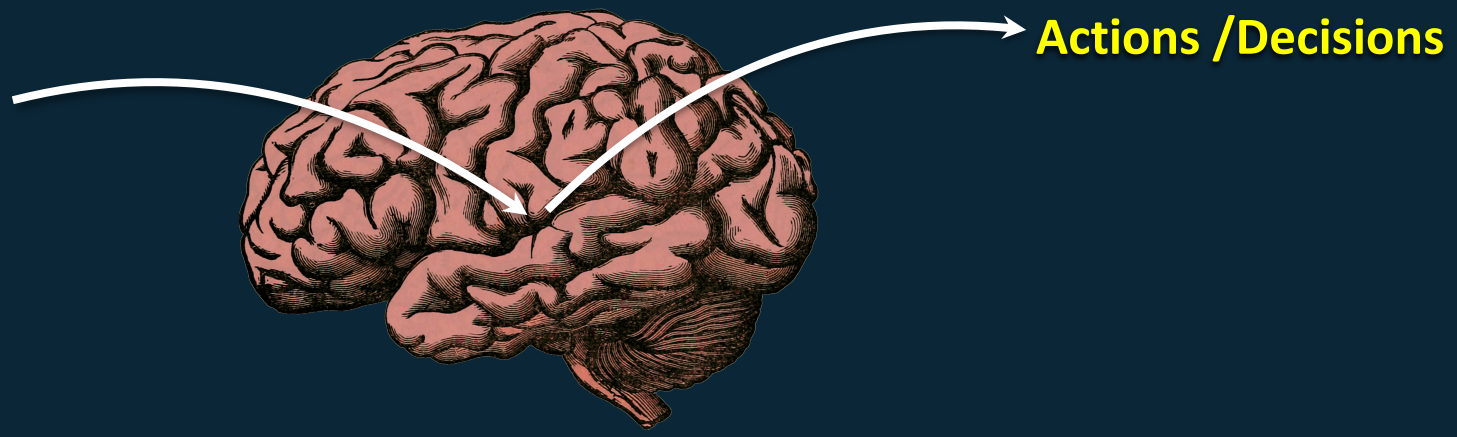
Aim: to help you acquire the foundational knowledge required to apply machine learning in practice.

Lets begin by first considering how human decision making works.

5. Human Decision Making

- **Using available information / past experience we try to make optimal decisions.**

Data (Sight, sound, taste, touch, smell)



Actions /Decisions

6. Quantifying Experience

- **Experience is describable and quantifiable as data.**
- **You already intuitively understand what data is, but to understand ML, we must explain the terminology used in the field.**

Variables / Attributes / Features | **Ground Truth**

Mass (Kg)	Height (cm)	Legs	Colour	Class / Label
4.2	25.4	4	Black	Cat
18	43.5	4	Brown	Dog
0.1	5.1	4	Grey/Black	Mouse
…	…	…	…	…

examples

Table 1. Characteristics of animals we've seen

As the ground truth is known, this is called a labelled dataset.

7. Quantifying Experience

Example notation x_i

	Mass (Kg)	Height (cm)	Legs	Colour
x_1	4.2	25.4	4	Black
x_2	18	43.5	4	Brown
x_3	0.1	5.1	4	Grey/Black

Class / Label y	
Cat	y_1
Dog	y_2
Mouse	y_3

Example notation y_i

Table 1. Characteristics of animals we've seen

Variables / Attributes / Features

Ground Truth

Array

$$x_1 = \{4.2, 25.4, 4, Black\} \qquad y_1 = \{Cat\}$$

$$x_2 = \{18, 43.5, 4, Brown\} \qquad y_2 = \{Dog\}$$

$$x_3 = \{0.1, 5.1, 4, Grey/Black\} \qquad y_3 = \{Mouse\}$$

Edge Hill University

8. Quantifying Experience

</TECHUP_WOMEN>

Variables / Attributes / Features | **Ground Truth**

$$x_1 = \{4.2, 25.4, 4, Black\} \quad y_1 = \{Cat\}$$

$$x_2 = \{18, 43.5, 4, Brown\} \quad y_2 = \{Dog\}$$

$$x_3 = \{0.1, 5.1, 4, Grey/Black\} \quad y_3 = \{Mouse\}$$

$$E = \{(x_1, y_1), (x_2, y_1), \dots, (x_n, y_1)\}$$

- **Experience is comprised of pairs of feature values and ground truth labels.**
- **The pairs are actually called “tuples”.**

Edge Hill University

9. Sometimes Knowledge is Incomplete

</TECHUP_WOMEN>

- **Sometimes we don't know the ground truth. In such cases, we have to collect information, and provide it for ourselves.**
- **This can be costly and time consuming.**

	Variables / Attributes / Features				Ground Truth	
	Ext. Temp (℃)	Int. Temp (℃)	Time (24hr)	Consumption (kWh)	Class / Label y	
x_1	4.2	17.1	15:23	7500	?	$(x_1, ?)$
x_2	33.4	27.8	18:01	7000	?	$(x_2, ?)$
x_3	10.8	19.1	20:36	450	?	$(x_3, ?)$
	…	…	…	…	…	

Table 2. Status of a heating control system

- **If the ground truth is unknown, this is called an "unlabelled" dataset.**

9. Sometimes Knowledge is Incomplete

- **Sometimes we don't know the ground truth. In such cases, we have to collect information, and provide it for ourselves.**
- **This can be costly and time consuming.**

	Variables / Attributes / Features				Ground Truth	
	Ext. Temp (℃)	Int. Temp (℃)	Time (24hr)	Consumption (kWh)	Class / Label y	
x_1	4.2	17.1	15:23	7500	?	$(x_1, ?)$
x_2	33.4	27.8	18:01	7000	?	$(x_2, ?)$
x_3	10.8	19.1	20:36	450	Deactivated	$(x_3, Deactivated)$
	…	…	…	…	…	

Table 2. Status of a heating control system

- **If the ground truth is unknown, this is called an "unlabelled" dataset.**
- **Datasets may also be "partially-labelled".**

10. Feature Values

- **Features can be categorical or numerical.**

Mass (Kg)	Height (cm)	Legs	Colour
4.2	25.4	4	Black
18	43.5	4	Brown
0.1	5.1	4	Grey/Black
…	…	…	…

Class / Label y
Cat
Dog
Mouse
…

Table 1. Characteristics of animals we've seen

Edge Hill University

10. Feature Values

</TECHUP_WOMEN>

- **Features can be categorical or numerical.**
- **Usually we "discretise" categorical features as this makes them easier for many algorithms to work with. This is either done automatically using tools, or manually.**

Mass (Kg)	Height (cm)	Legs	Colour
4.2	25.4	4	1
18	43.5	4	2
0.1	5.1	4	4
…	…	…	…

Class / Label y
1
2
3
…

Table 1. Characteristics of animals we've seen

Edge Hill University

10. Feature Values

</TECHUP_WOMEN>

- Features can be categorical or numerical.
- Usually we "discretise" categorical features as this makes them easier for many algorithms to work with. This is either done automatically using tools, or manually.

Mass (Kg)	Height (cm)	Legs	Colour
0.229	25.4	4	1
1.0	43.5	4	2
0	5.1	4	4
…	…	…	…

Class / Label y
1
2
3
…

Table 1. Characteristics of animals we've seen

- We may also "normalise" our data from time to time.

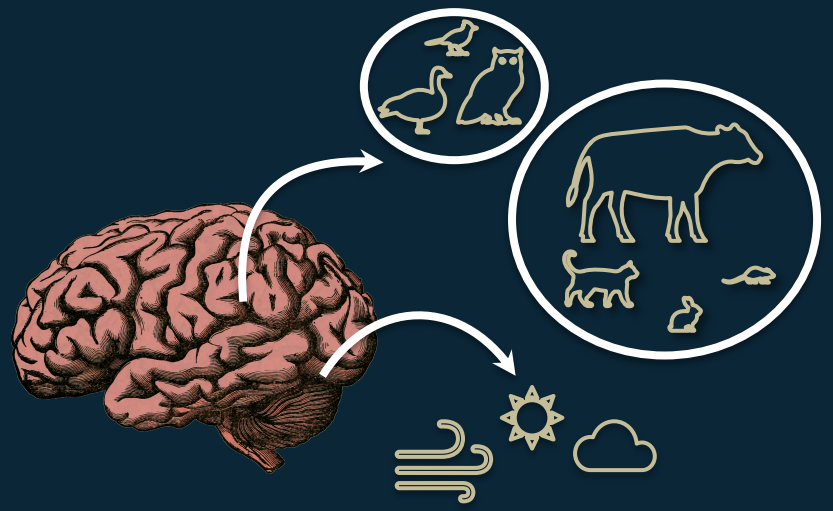
11. Human Decision Making

Features (Variables)

$$x_i = \{ \text{Height, Mass, } \ldots, \text{Age} \}$$

Labels (Feedback)

$$y_i = \{Cat, Dog, \ldots, Bird\}$$

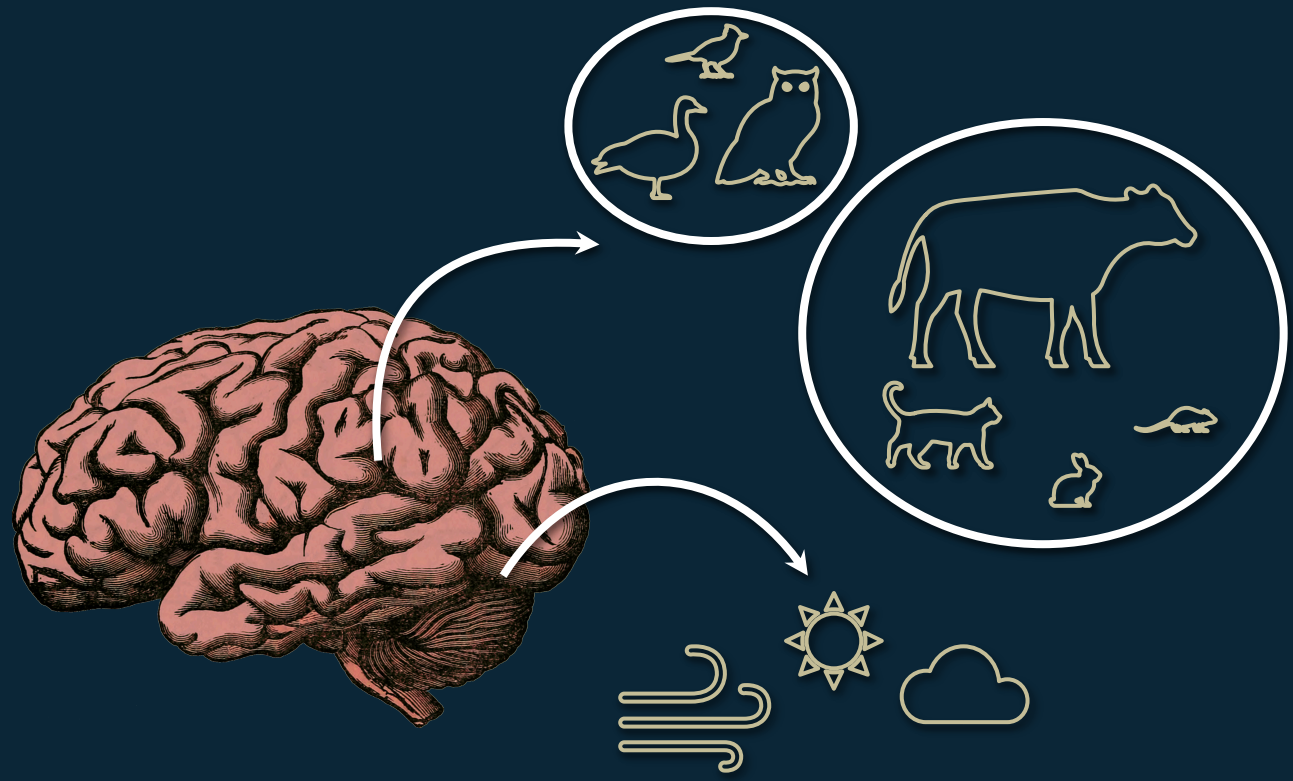


Edge Hill University

12. Human Decision Making

</TECHUP_WOMEN>

- Humans capable of accurately,
  - Clustering
  - Classifying
  - Predicting
- Made possible via a feedback cycle.



$$E = \{(x_1, y_1), (x_2, y_1), \ldots, (x_n, y_1)\}$$

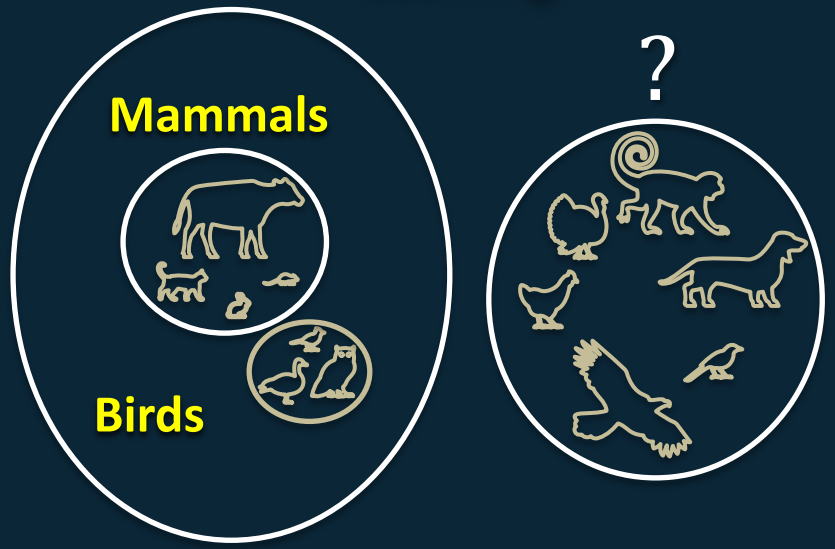
Feedback

Edge Hill University

13. Approaches to Learning

</TECHUP_WOMEN>

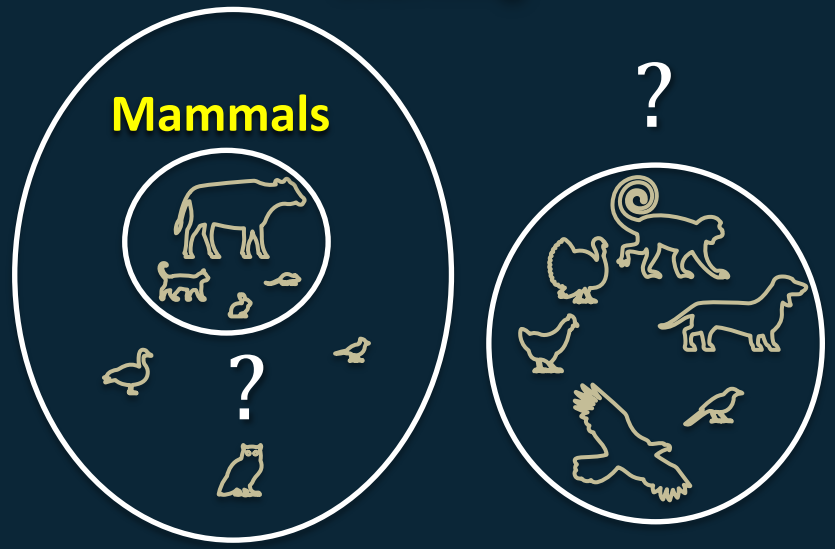
Supervised Learning



E Contains labels for mammals & birds, goal is to correctly group unseen animals into mammal & bird classes.

Labeled data

Semi-supervised Learning



E Contains examples labels for mammals only, goal is to correctly group unseen animals into mammal & bird classes.

Partially-Labeled data

Unsupervised Learning

E Contains no labels. Instead we use the data to self describe class separation.

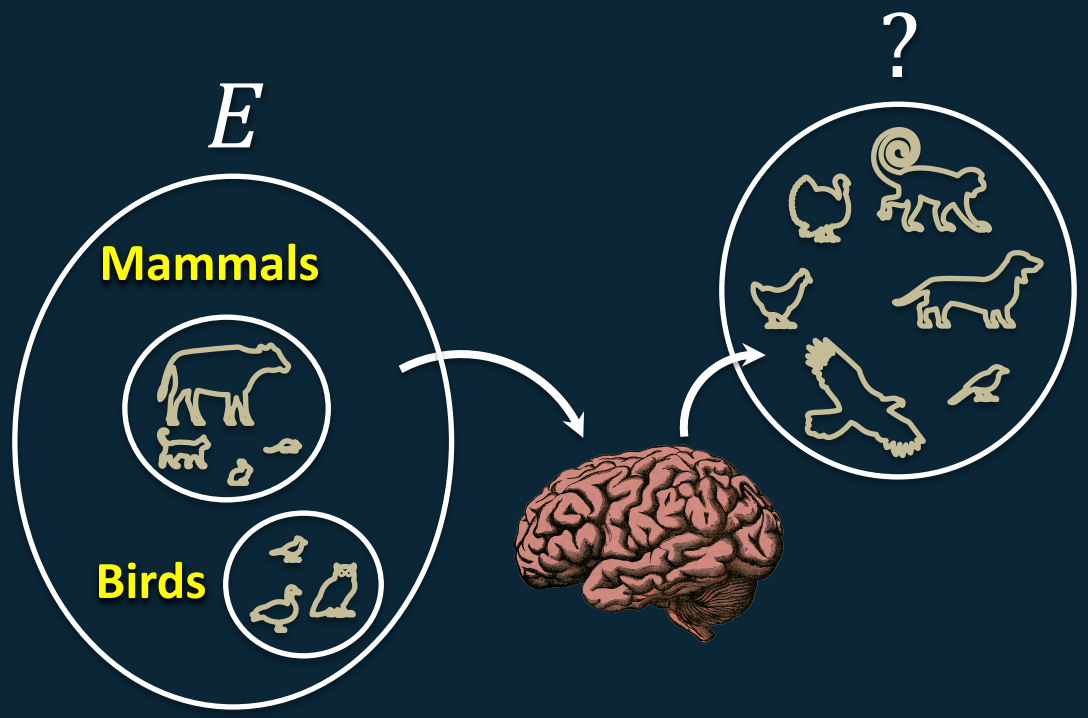
Unlabeled data

Durham University · Edge Hill University · The University of Nottingham (UNITED KINGDOM · CHINA · MALAYSIA) · University of York · institute of CODING · Office for Students

Edge Hill University

14. Classification

</TECHUP_WOMEN>



- These problems are “binary” classification tasks.
- Called binary, as there are two options (mammal or bird).
- Such classification tasks are everywhere, and we do them every day.
- To complete them we make predictions.
- Classification problems in the real-world are usually far more complex, i.e. more than 2 potential classes to predict.
- These are known as multi-class problems. Let’s try one.

15. Example

Feline

Lutrinae
(e.g. Otter)

Canine

Mongoose

Viverridae
(e.g Civet)

16. Could you generalise?

- In the last slide, we saw an animal you may not be familiar with.
- Even if you have extensive knowledge of the animal kingdom, you may not have known the animal was a Fossa.
- Perhaps this animal was not described in the dataset, E .
- Even if you did not know it was a Fossa, you could still tell it was a type of mammal.
- This application of your knowledge, from past experience, illustrates your ability to generalise beyond known facts.
- Humans are gifted generalisers - but they are susceptible to two problems:

Overfitting **Underfitting**

Edge Hill University

17. Notation Recap

</TECHUP_WOMEN>

x_i represents an example (row) in a dataset (replace i with row number).

x_i is an array containing many values (columns).

$$x_1 = \{0.229, 25.4\}$$

y_i is a ground truth label associated with x_i

$$y_1 = \{1\}$$

	Features: Mass (Kg)	Features: Height (cm)	Label: Class / Label y	
x_1	0.229	25.4	1	y_1
x_2	1.0	43.5	2	y_2
x_3	0	5.1	3	y_3

A tuple is a finite ordered list: $(x_1, y_1) = (\{0.229, 25.4\}, \{1\})$

Experience, E, is comprised of many tuples.

$$E = \{(x_1, y_1), (x_2, y_1), \ldots, (x_n, y_1)\}$$

Durham University · Edge Hill University · The University of Nottingham UNITED KINGDOM · CHINA · MALAYSIA · University of York · institute of CODING · Office for Students

Edge Hill University

18. Checkpoint

</TECHUP_WOMEN>

So far we've introduced,

- Data sets, features, and class labels.
- Labelled and unlabelled datasets.
- Different types of learning that we're capable of.
- Bias.
- The concept of classification.
- Generalisation, and under/over fitting.

Next we develop these ideas further to study automated machine learning. We'll find that ML borrows a great deal from how we learn, and is just as susceptible as we are to error!