Edge Hill University

A.I. for Astrophysics: Masterclass

Dr. Rob Lyon

Edge Hill University

robert.lyon@edgehill.ac.uk

@scienceguyrob

Edge Hill University

Welcome

In the next few hours we'll learn how to apply ML to a real-world astrophysics problem. If you've never done machine learning before don't worry. I'll cover everything you need to know including:

- Machine learning basics.
- Python basics.
- The astronomy background you'll need to understand the challenge.

If you have questions at any point – just ask. We'll move through things slowly. We want this masterclass to be enjoyable!

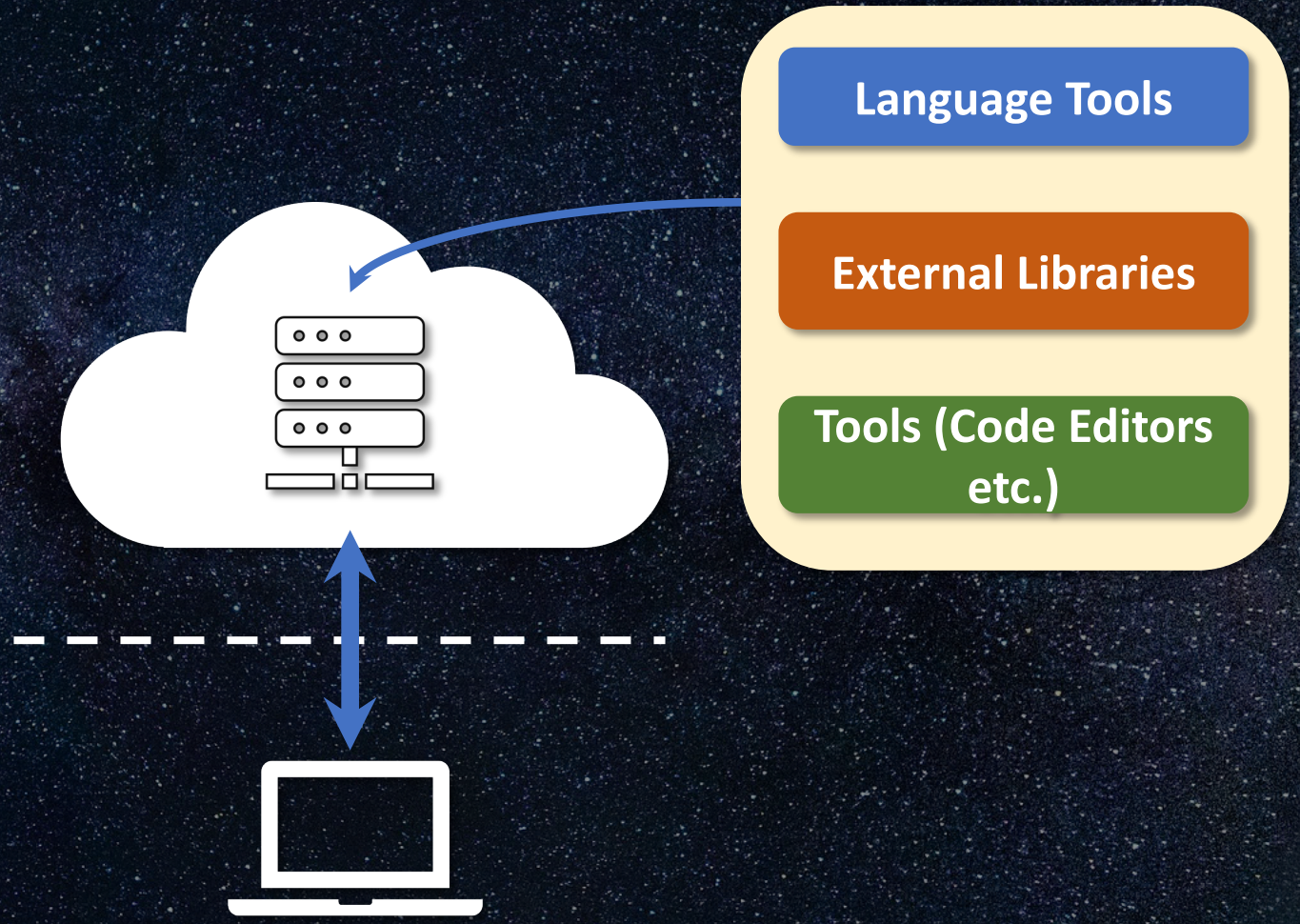
Edge Hill University

Masterclass Environment

Edge Hill University

Google Colab

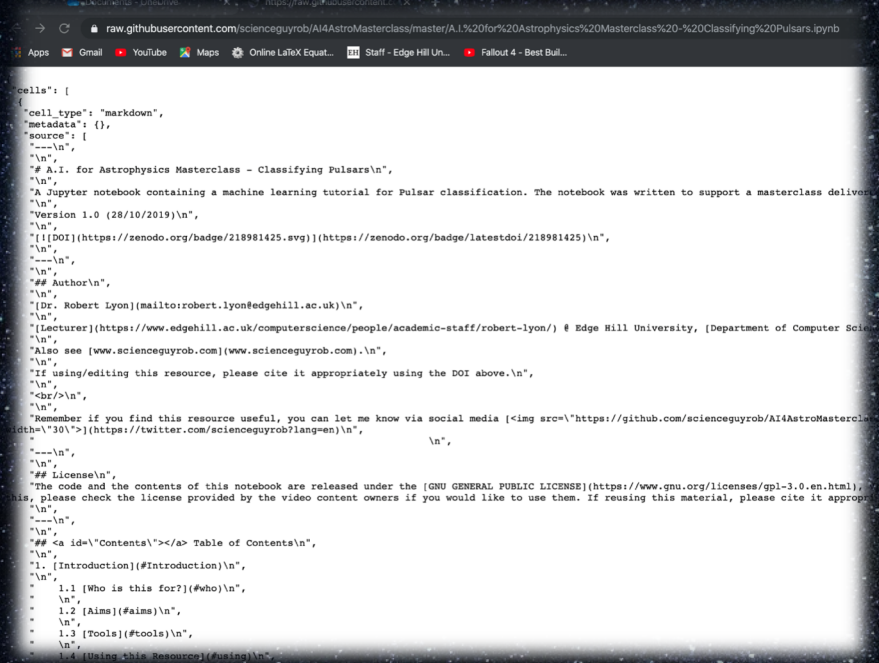
- A cloud environment containing the tools need to write & execute Python programs.
- You'll need is a google account to use, e.g. a Gmail Account and the *Chrome* Web Browser.
- When you login to the Collaboratory, it creates a computer just for you.
- Inside you can create & execute Python code.
- Before proceeding, please create a Google account if you don't already have one.



Edge Hill University

Accessing the Masterclass Resource

1. Open Google Chrome.
2. Navigate to: https://github.com/scienceguyrob/AI4AstroMasterclass
3. Click the file: "`A.I. for Astrophysics Masterclass - Classifying Pulsars.ipynb`"
4. There is a button with the text "Raw" on it – click it. This will load the resource into your browser.
5. Save the content to your own machine – left click and choose the "Save as" option that appears on the popup menu. Keep the `.ipynb` file extension.
6. Open Google Colab: https://colab.research.google.com/
7. Click the File menu, click "Upload Notebook". Choose the file you saved on your machine.



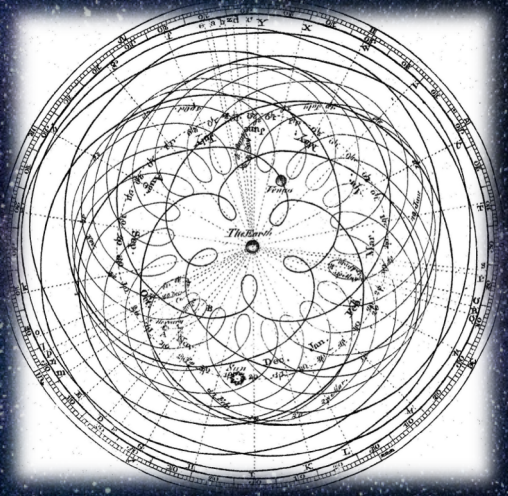
Edge Hill University

Astrophysics Background

Edge Hill
University

Edge Hill University

Some History

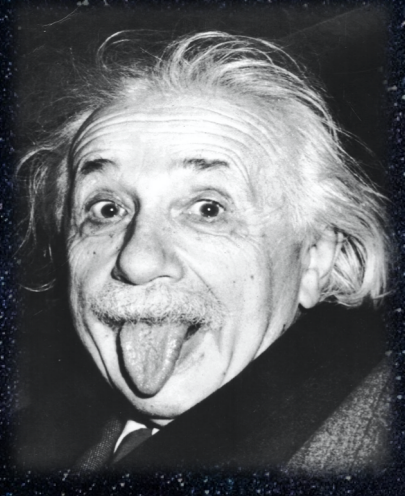


Geocentric Model

Heliocentric Model

Laws of Motion

Gravity

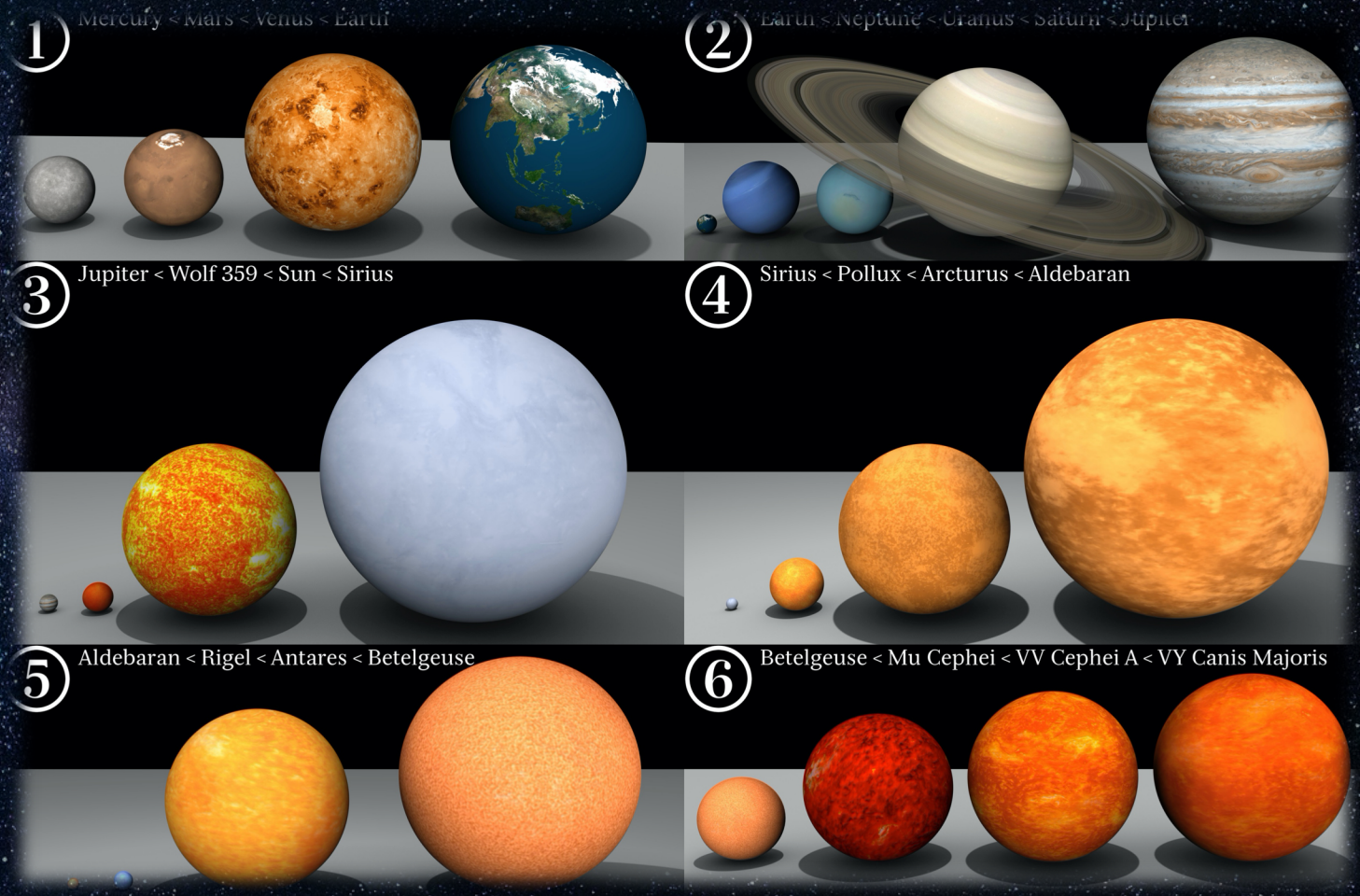


Relativity

Time

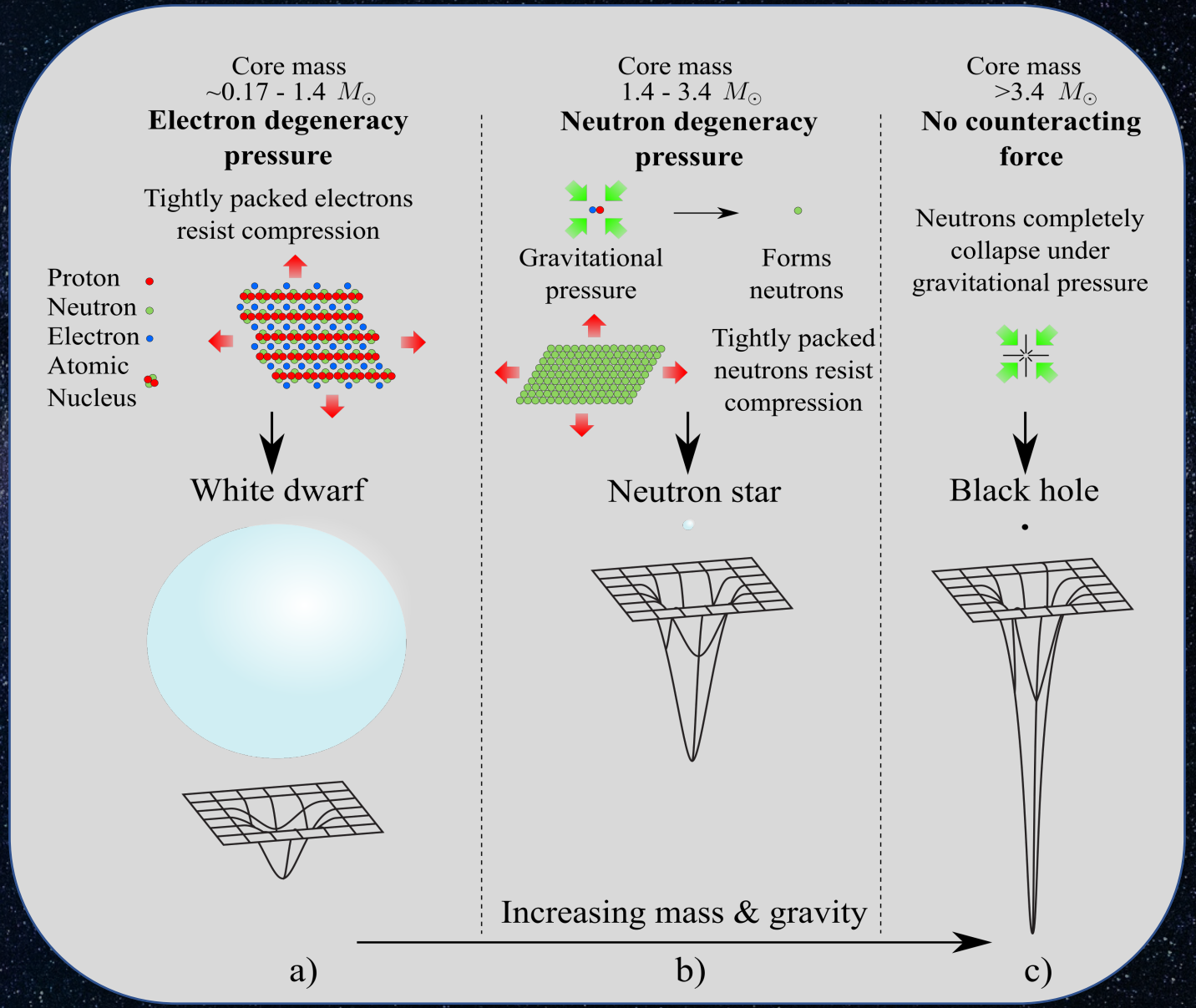
Edge Hill University

Natural Labs?



Edge Hill University

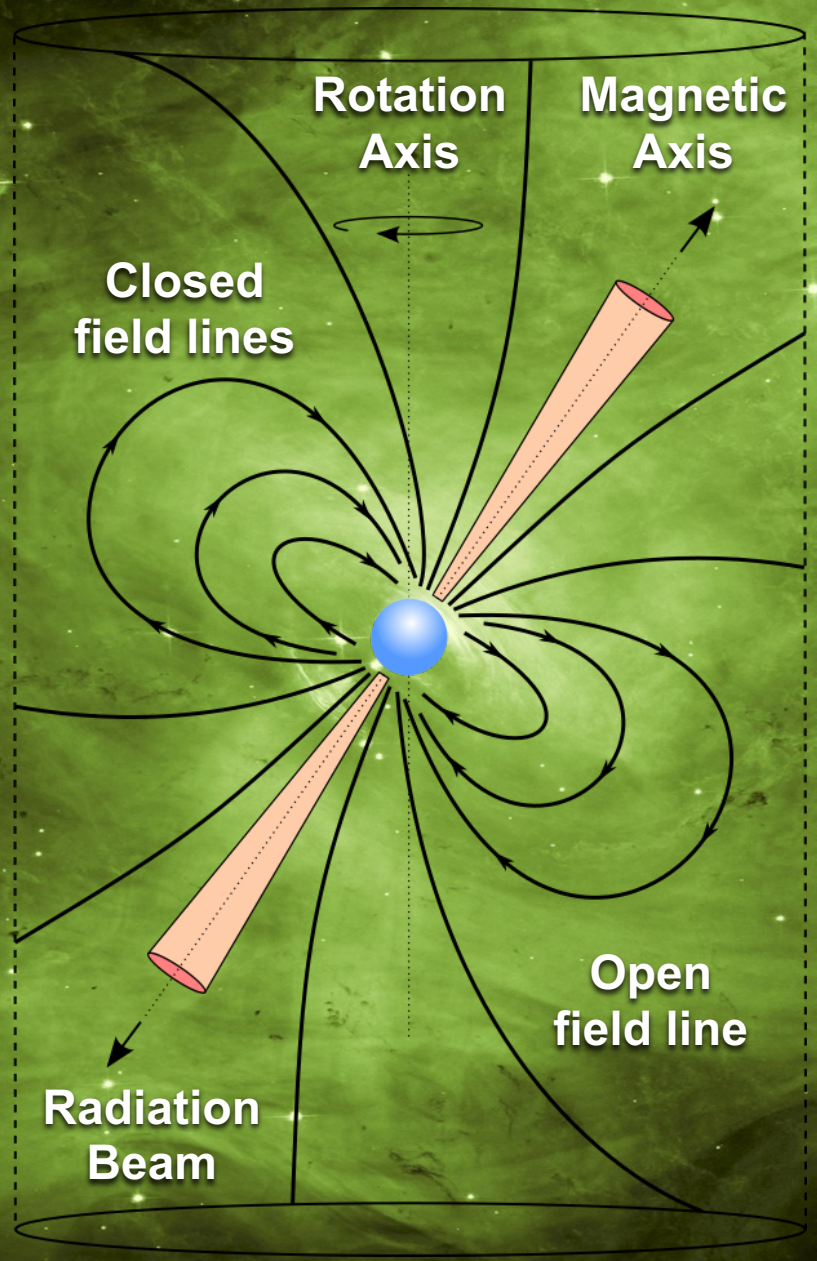
Natural Labs?



Edge Hill University

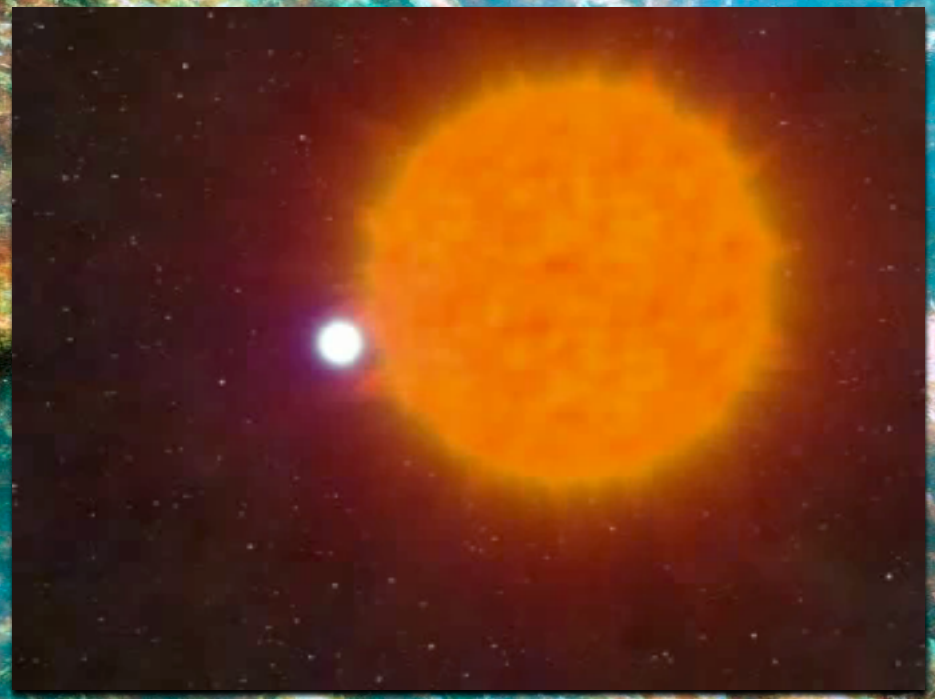
Radio Pulsars

- Stellar remnant
- Very dense
- ~20 km diameter
- Produce radio emission
- Very useful for science



Edge Hill University

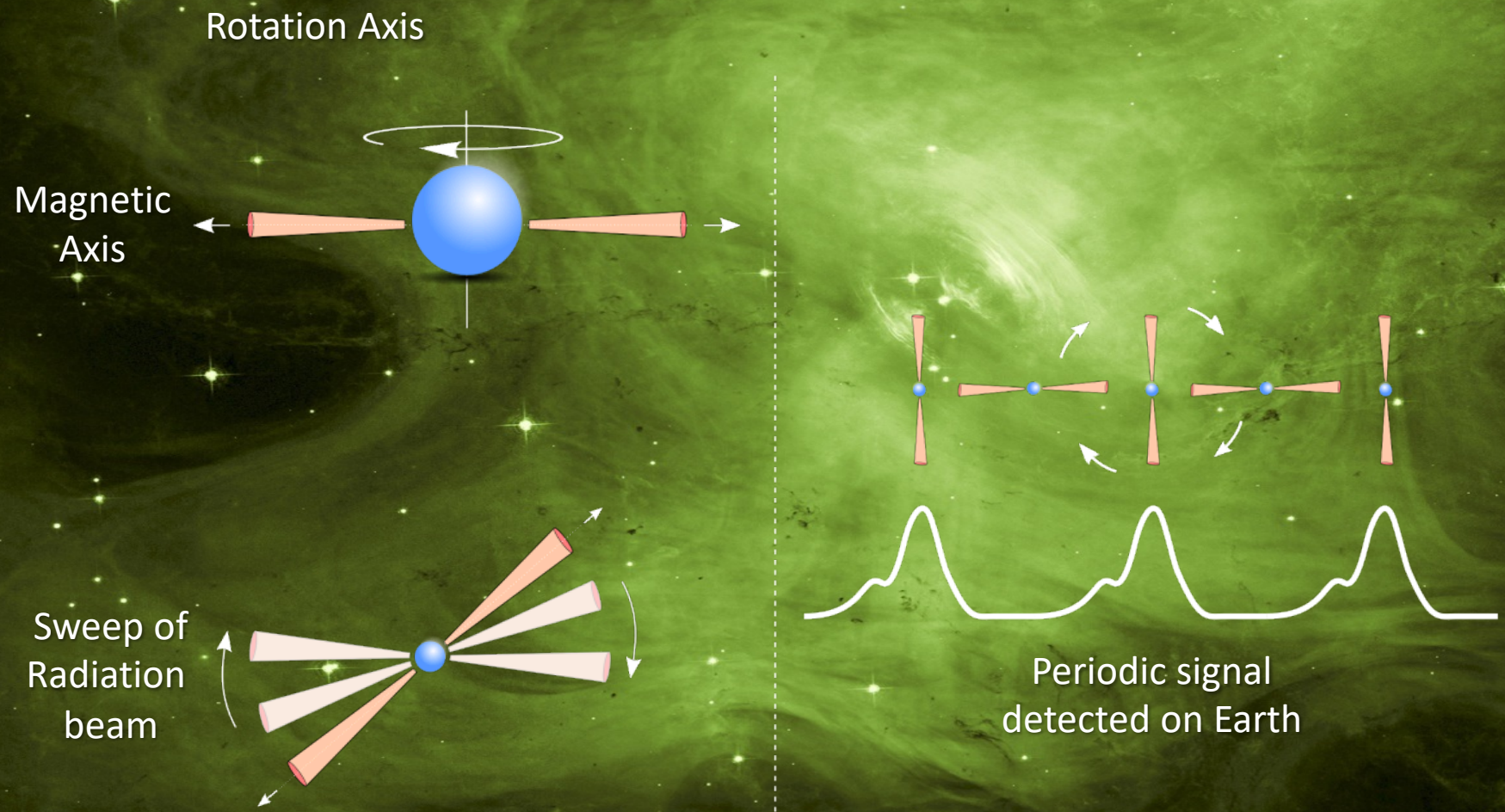
Pulsar Birth



Credit: John Rowe Animations

Edge Hill University

Pulsar Signals



Edge Hill University

Pulse Profiles

Complex Pulse Profiles

Edge Hill University

Search Tools

Credit: Shaun Amy

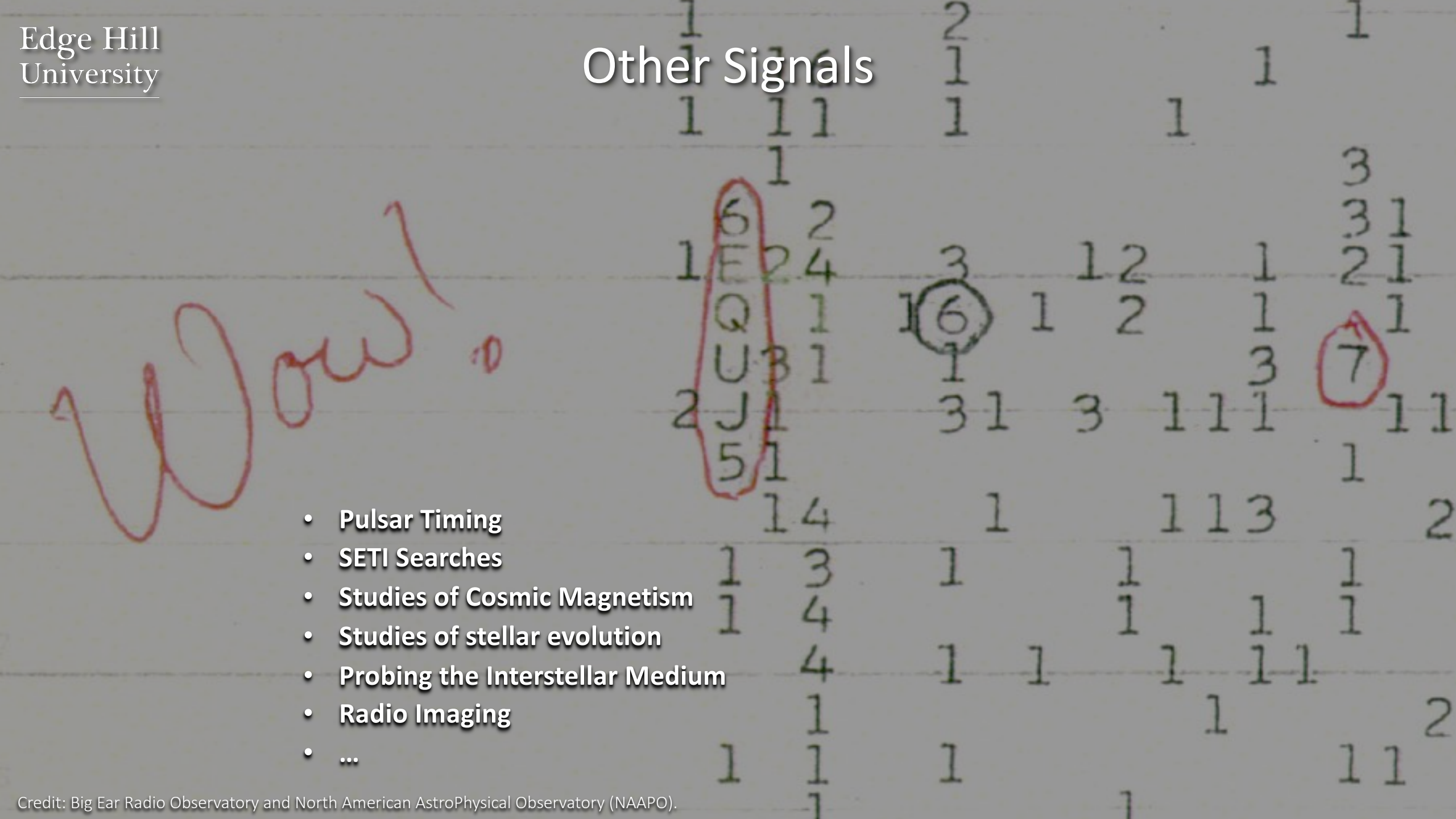






Edge Hill University

Other Signals



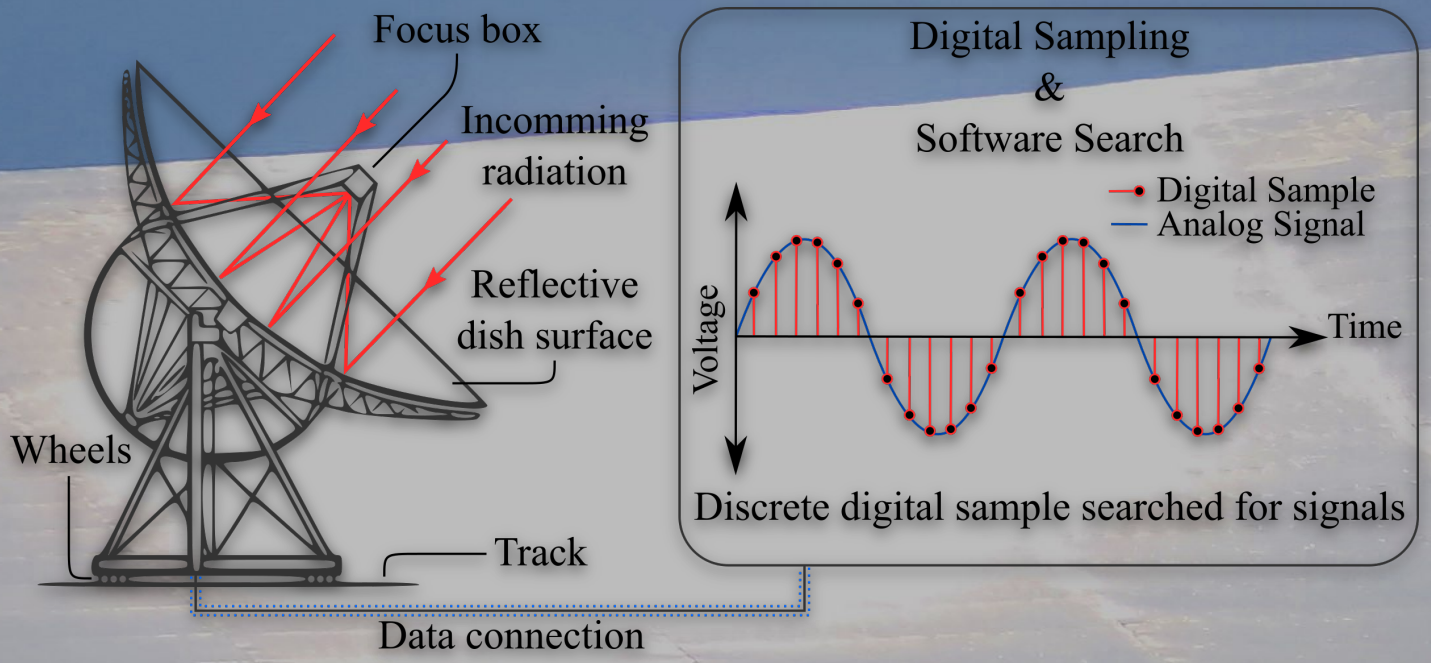
- Pulsar Timing
- SETI Searches
- Studies of Cosmic Magnetism
- Studies of stellar evolution
- Probing the Interstellar Medium
- Radio Imaging
- …

Credit: Big Ear Radio Observatory and North American AstroPhysical Observatory (NAAPO).

Edge Hill University

Data Capture

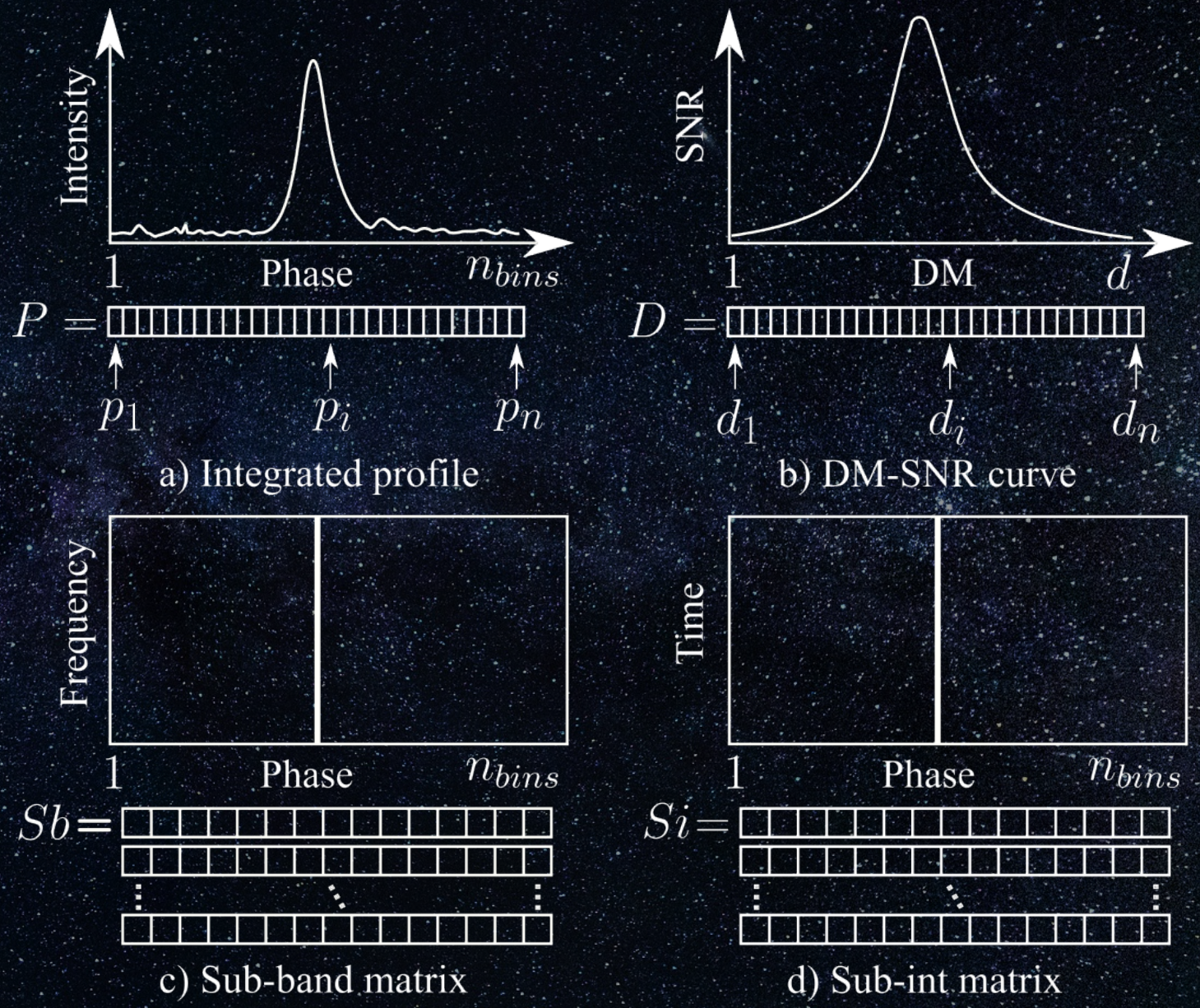
- Analog to digital samples
- Complex search pipeline applied
- Steps:
  - RFI mitigation
  - Dedispersion
  - FFT
  - Harmonic summing
  - Detection
  - Sifting



Edge Hill University

Data Products

- A candidate file
- Describes a signal detection
- Covers time and frequency space
- At least one per detection (duplicates!)
- Classification appears simple - things do get fuzzy!



Edge Hill University

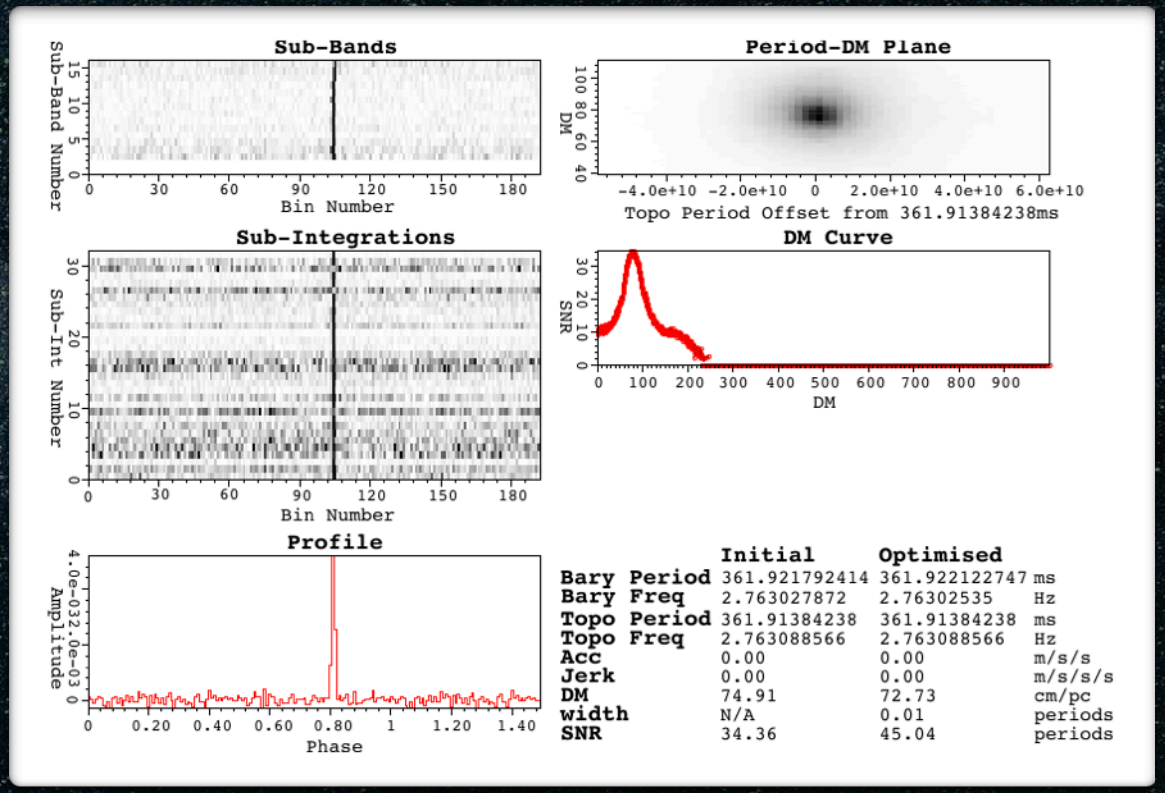
What to look for?

- Clear pulse with a defined peak.
- Evidence that the signal persists in time, i.e. an indication of a periodic source.
- Evidence that the signal persists in frequency - pulsars are broadband emitters.
- A DM value greater than zero.
- Evidence of other effects, such as scintillation?

Edge Hill University

Candidate Examples (1)

- Defined peak
- Clear DM-SNR peak
- Consistent in time
- Persistent in frequency
- Obvious pulsar!

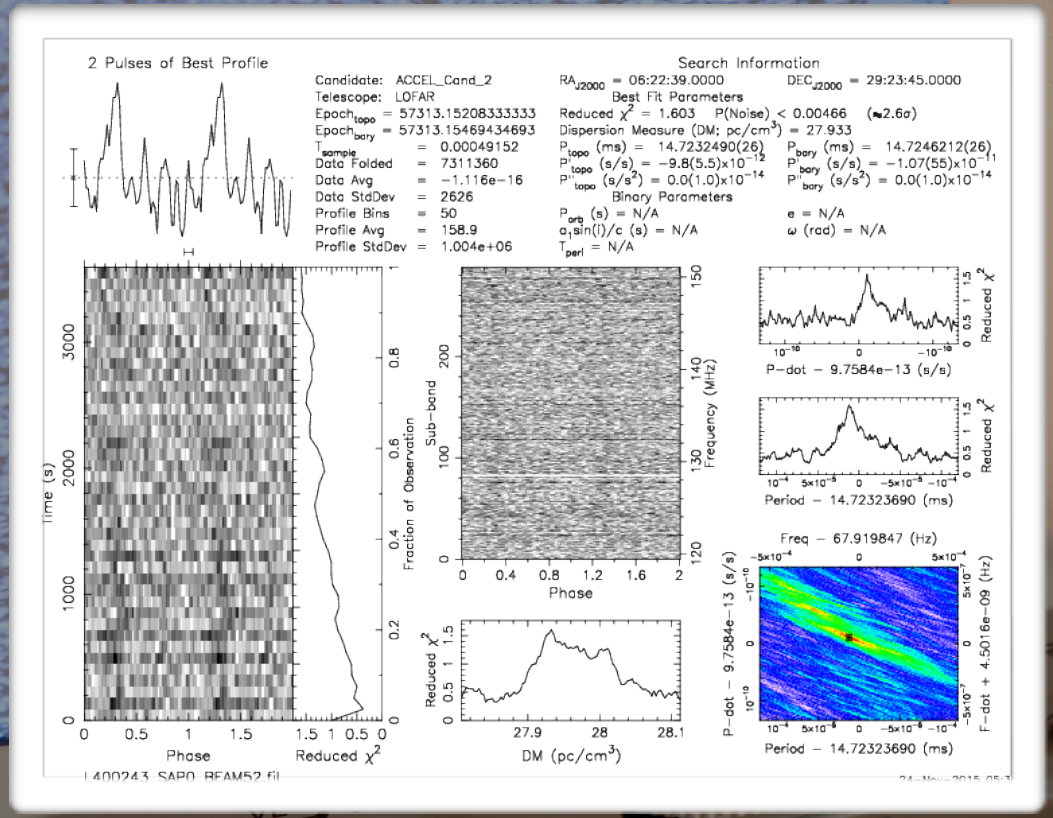


Credit: HTRU Collaboration.

Edge Hill University

Candidate Examples (2)

- Clearly defined peak?
- Some persistence in time?
- A DM value > zero
- Labelled as noise

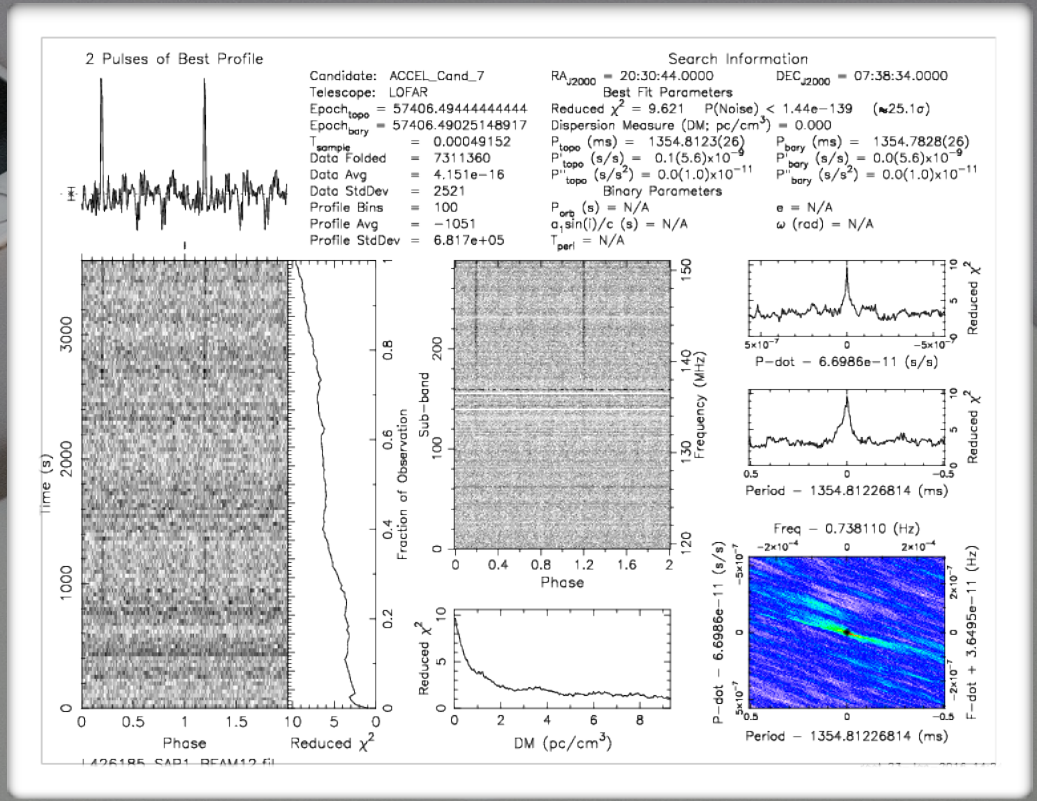


Credit: LOTAAS Collaboration (Chia Min Tan et. al.).

Edge Hill University

Candidate Examples (3)

- Clearly defined peak?
- Some persistence in time?
- Some persistence in frequency?
- RFI



Credit: LOTAAS Collaboration (Chia Min Tan et. al.).

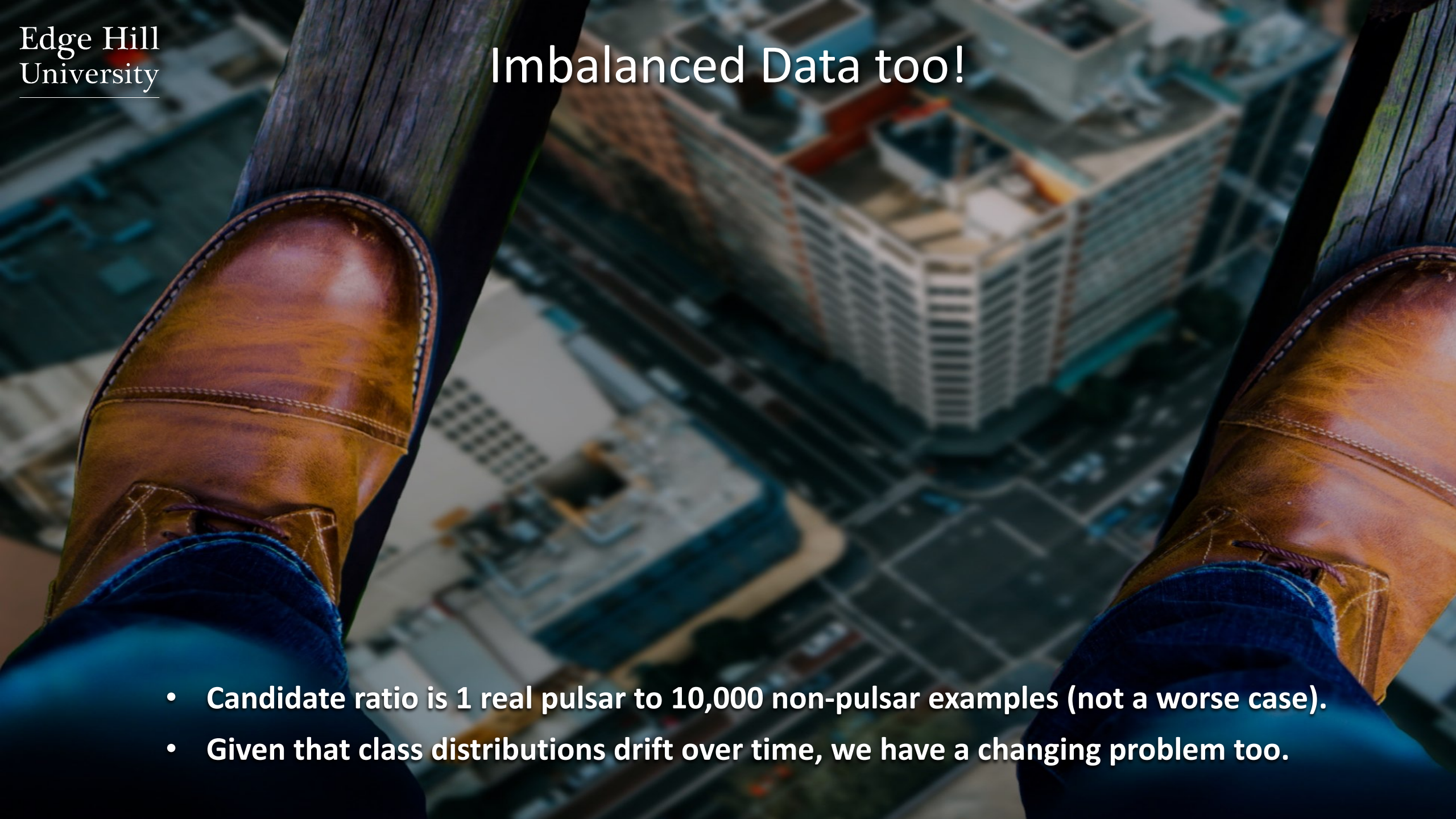
Edge Hill University

Things go wrong in practice

- Not all candidates so easily categorised.
- Huge volume of candidates usually returned by telescopes.
- This makes the problem particularly hard to solve.

Edge Hill University

Imbalanced Data too!



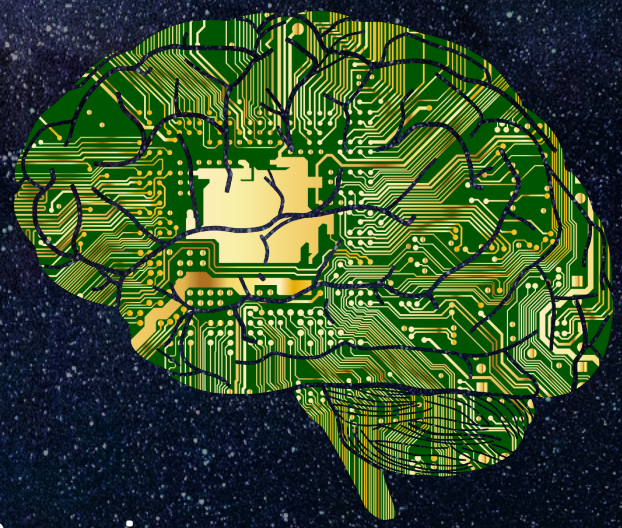
- **Candidate ratio is 1 real pulsar to 10,000 non-pulsar examples (not a worse case).**
- **Given that class distributions drift over time, we have a changing problem too.**

Edge Hill University

Tackling this with A.I.

Edge Hill University

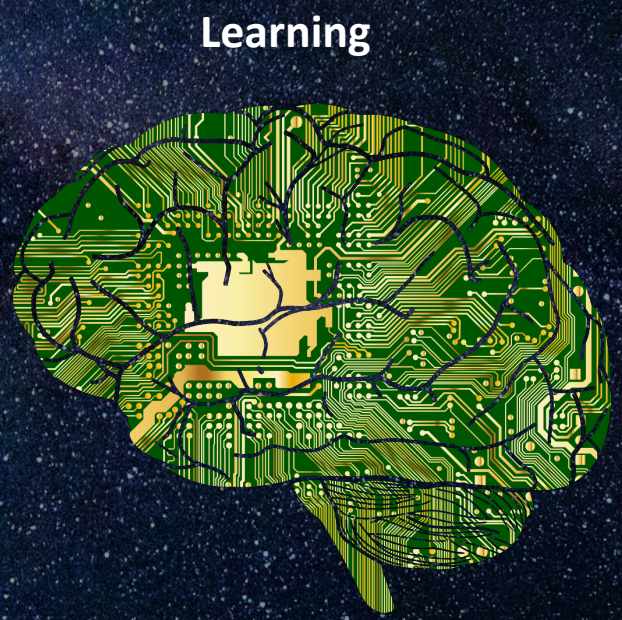
Artificial Intelligence



- 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

Machine Learning



- 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

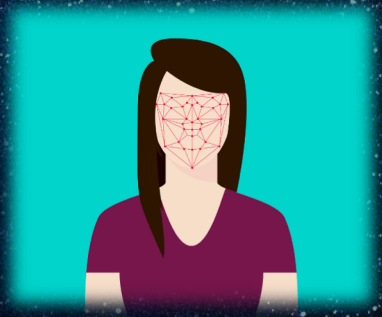
Example Applications

Searching for rare stars in astronomy data

Image Credit: SARAO

Identifying regions of disease in medical images

Climate Modelling



Facial Recognition

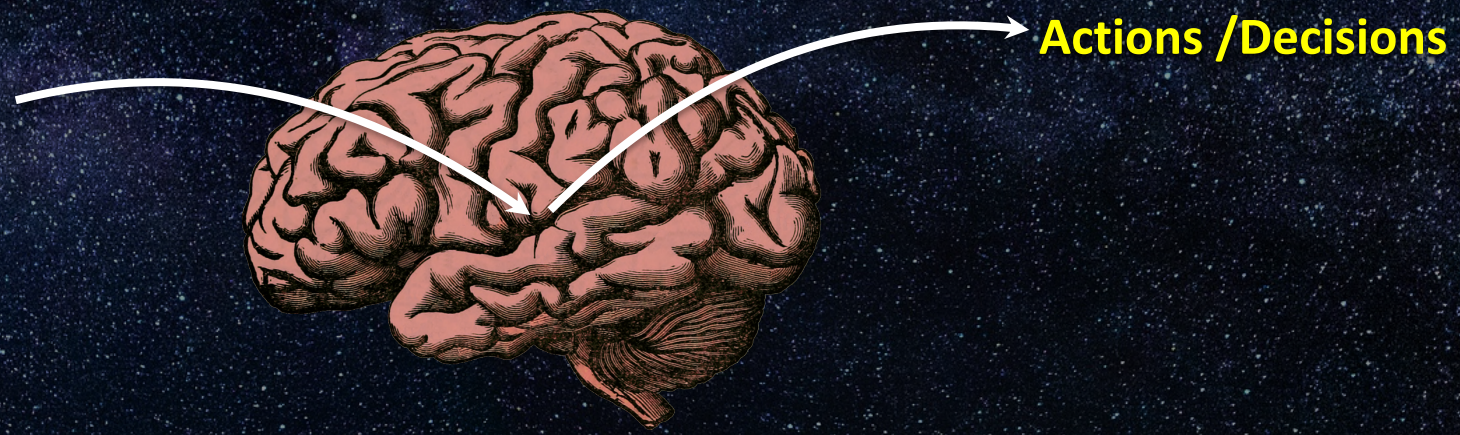
Voice Recognition

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

Edge Hill University

Human Decision Making

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



Edge Hill University

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.

Edge Hill University

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

Quantifying Experience

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

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

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

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

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

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

Edge Hill University

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	$(x_3, ?)$ $(x_3, Deactivated)$
	…	…	…	…	…	

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".**

Edge Hill University

Feature Values

- 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.

Edge Hill University

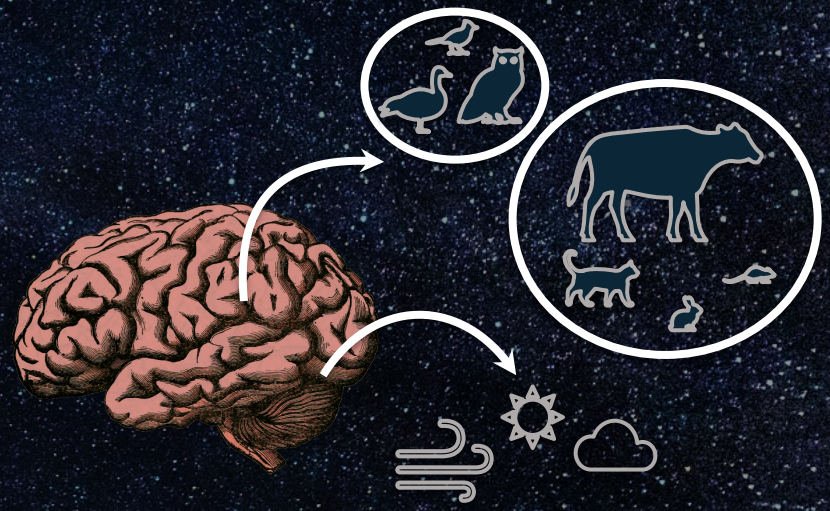
Human Decision Making

Features (Variables)

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

Labels (Feedback)

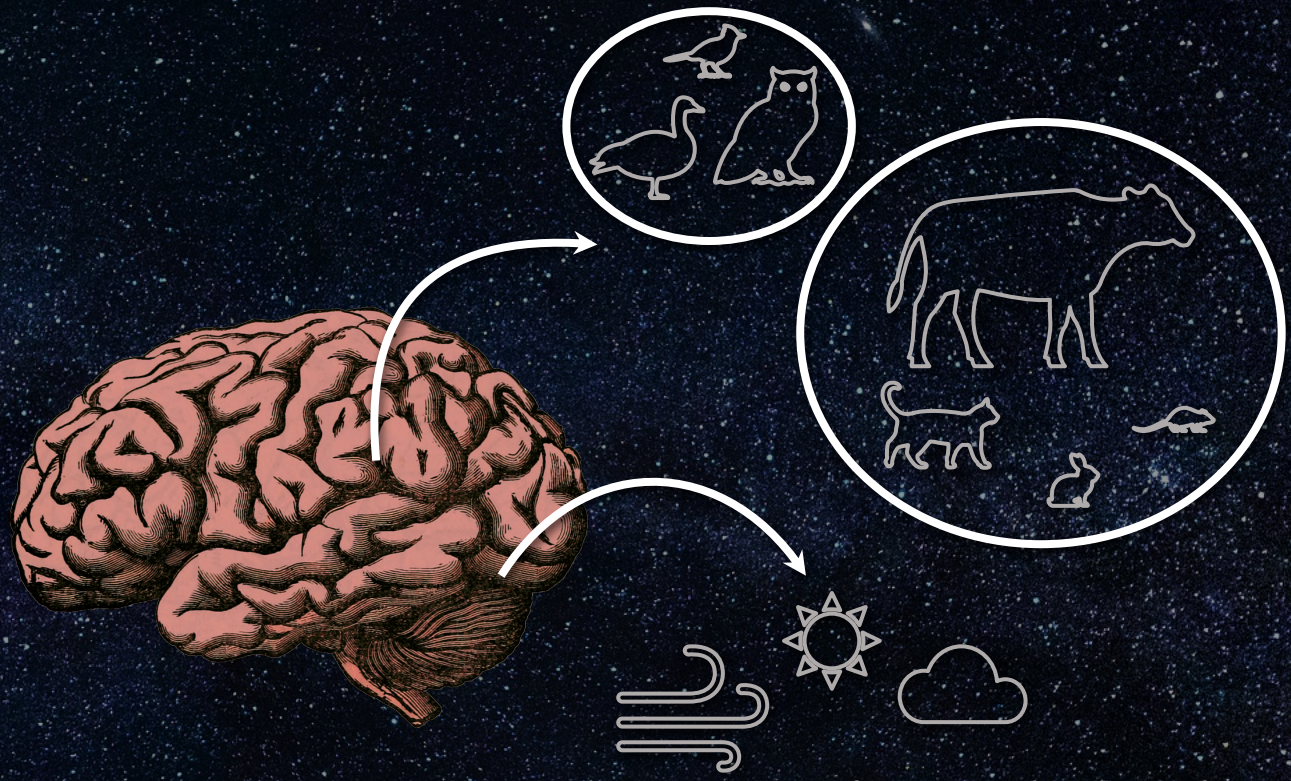
$$y_i = \{Cat, Dog, \ldots, Bird\}$$



Edge Hill University

Human Decision Making

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



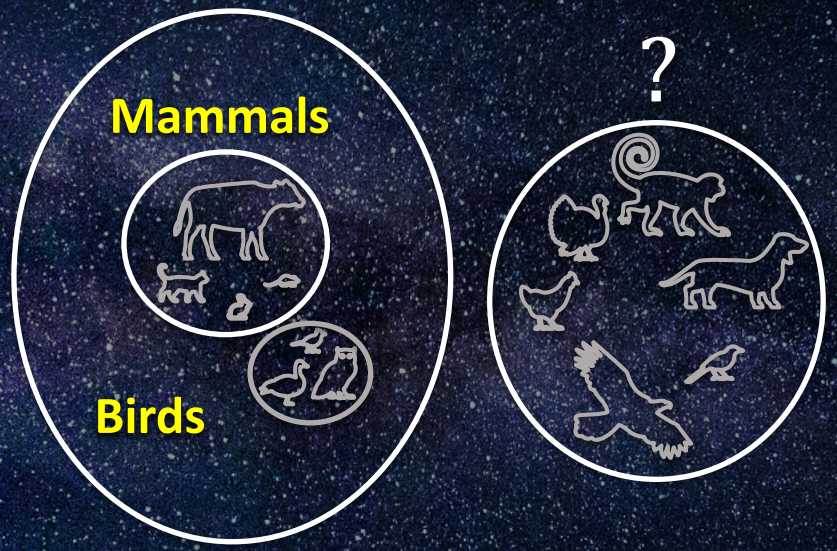
$$E = \{(x_1, y_1), (x_2, y_1), \ldots, (x_n, y_1)\}$$

Feedback

Edge Hill University

Approaches to Learning

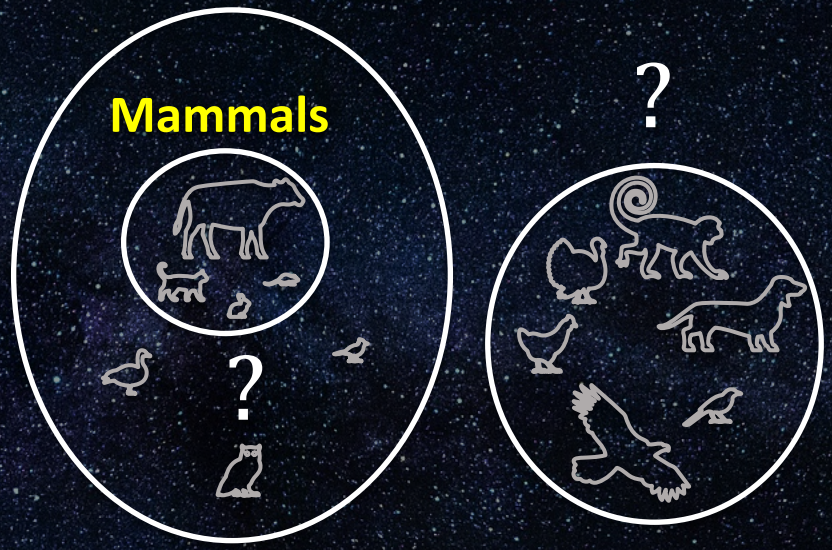
Supervised Learning



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

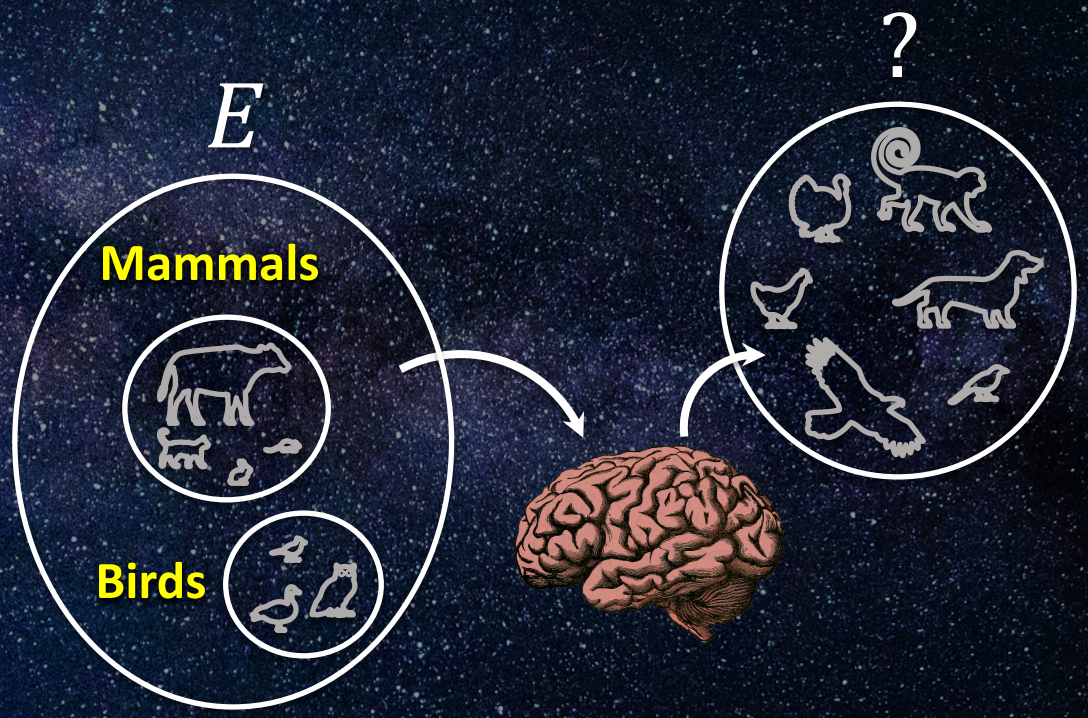
Unsupervised Learning

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

Unlabeled data

Edge Hill University

Classification



- 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.

Edge Hill University

Example

Feline

Lutrinae (e.g. Otter)

Canine

Mongoose

Viverridae (e.g Civet)

Edge Hill University

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

Making Optimal Decisions

- Many have tried to understand how to make optimal decisions.
- We know we should use available evidence at all times.
- Humans are not always so thorough - we do make bad decisions.
- We are biased decision makers - we often use instinct and personal experience to decide.