

School of Electronic Engineering and Computer Science  
Queen Mary University of London

## CBU5201 - Machine Learning – 2023/24

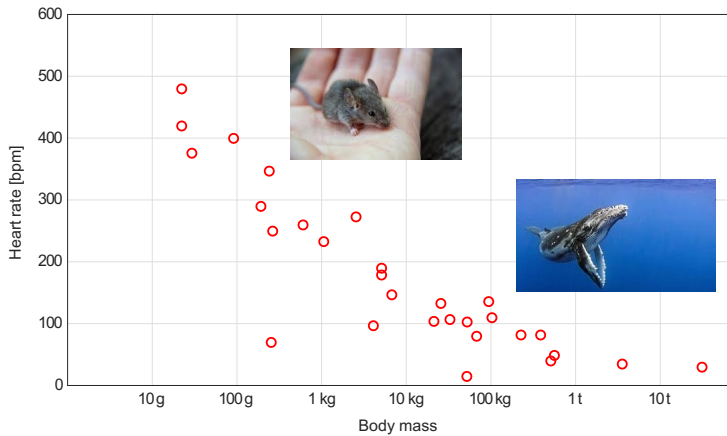
Dr Chao Liu

01 Sep 2023

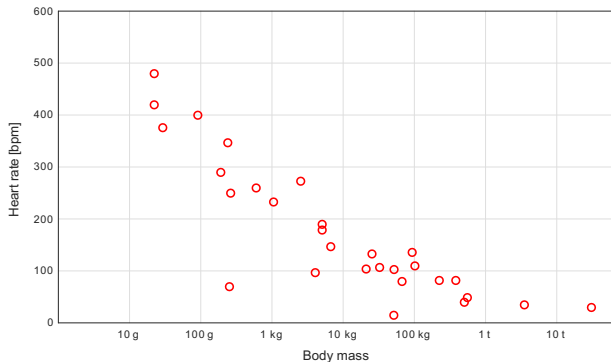
Credit to Dr Jesús Requena Cañón



# From mouse to whale



# From mouse to whale, through rabbit



A rabbit's resting heart beats at

(a)  $\leq 100$  bpm

(b)  $\geq 300$  bpm

(c)  $\geq 100$  bpm and  $\leq 300$  bpm

# Agenda

[What is machine learning?](#)

[The value of knowledge](#)

[The machine learning taxonomy](#)

[About CBU5201](#)

# Machine learning or statistical learning?

**Machine** or **statistical learning** is usually defined as

*The ability to acquire **knowledge**, by extracting patterns from raw **data**.*  
(Goodfellow, Bengio, Courville)

*A set of tools for **modeling** and **understanding** complex **datasets**.*  
(James, Witten, Hastie, Tibshirani)

# What is data?

- **Data** is the materialisation of an **observation** or a **measurement**.
- **Datasets** are data formatted as collections of **items** described by a set of pre-defined **attributes**.

Animal (ID)	Body mass [g]	Heart rate [bpm]
Wild mouse	22	480
Rabbit	$2.5 \times 10^3$	250
Humpback whale	$30 \times 10^6$	30
...	...	...

In machine learning, **our data is always represented as a dataset**.

*Note: Unfortunately, authors use different terms for the same concept. Item, sample, example, instance and point have the same meaning, and so do feature, variable and attribute. You should get used to all of them.*

# What is knowledge?

**Knowledge** can be represented as a

- **Proposition** (statement, law)

*Smaller animals have a faster heartrate.*

- **Narrative** (description, story)

*The size of an animal seems to be related to its heart rate. In general, larger animals tend to have a slow heart rate. For instance, the humpback whale...*

- **Model** (mathematical or computer)

*$r = 235 \times m^{-1/4}$ , where  $r$  is the heart rate and  $m$  is the body mass.*

We will mostly use models to represent knowledge.

# Knowledge as a model

Models describe **relationships** between attributes. **Mathematical** and **computer** models are equivalent:

- Mathematical models can be implemented as computer programs.
- Every computer model has a corresponding mathematical expression.

The mathematical expression  $y = x + 3x^2$  is equivalent to the following Matlab line of code:

```
y = x + 3*x^2
```

or this Python line of code:

```
y = x + 3*x**2
```



# This is about [data] science...

Science is not about using sophisticated instrumentation, maths or techniques, science is about **evaluating** our knowledge.

We use **data** together with **accepted knowledge** in this evaluation.

Which of the two following proposition would you describe as *scientific*

- Proposition 1: *The earth is flat*
- Proposition 2: *The earth is roughly spherical*

Propositions are not scientific or unscientific, but the **way we evaluate them**.



# Data science is much **more than data**

If there is no data, there is no machine learning. This doesn't mean machine learning is all about data.

Specifically, there is no such thing as neutral or objective data that speaks the truth. We need to follow a **rigorous methodology**.

Craniometry (19th century): *The size of a brain is related to its degree of intelligence, e.g. big heads are smarter than small ones, or elongated heads are smarter than short ones.*

# Our notion of machine learning

Many machine learning professionals adopt a **dataset-first** view: we start with a dataset, then (formulate a problem and finally) produce a model.

In CBU5201 we use a **deployment-first** (problem-first) approach: we start with a problem, then secure a dataset and finally produce a model.

Accordingly, we define machine learning as:

*A set of **tools** together with a **methodology**...  
for solving scientific, engineering and business **problems**...  
using **data**.*

# What about AI?

Definitions of AI include creating machines that act like humans, think like humans, act rationally or think rationally.

Some AI solutions use machine learning algorithms, some others do not. In addition, machine learning can be used outside the AI remit.

Most of the time, when media and companies talk about AI, they mean machine learning, basic statistics or even a responsive website.

**This module is about machine learning, not AI.**

# Agenda

[What is machine learning?](#)

[The value of knowledge](#)

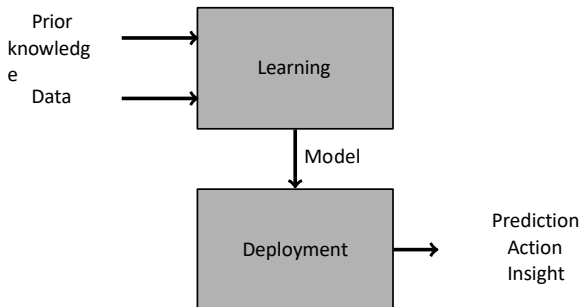
[The machine learning taxonomy](#)

[About CBU5201](#)

# The two stages of Machine Learning

Models can be built, sold and deployed to deliver **value**. During the life of a model, we can distinguish two stages:

1. **Learning** stage: The model is built.
2. **Deployment** stage: The model is used.



# Deployment: eCommerce

Inspired by your shopping trends






Recommendations for you in Grocery





# Data science competitions

The screenshot shows the Kaggle website's 'Competitions' section. The header includes the Kaggle logo, a search bar, and navigation links for Competitions, Datasets, Kernels, Discussion, and Learn. Below the header, there's a blue banner with the word 'Competitions' and buttons for 'Documentation' and 'InClass'. The main content area features a filter bar with 'General' and 'InClass' tabs, a 'Sort by' dropdown set to 'Grouped', and a search bar for competitions. A blue bar indicates '11 Active Competitions'. Three competitions are listed:

Logo	Challenge Name	Prize	Teams
	<b>TGS Salt Identification Challenge</b> Segment salt deposits beneath the Earth's surface <i>Featured</i> · 25 days to go · geology, image data	\$100,000	2,781 teams
	<b>Airbus Ship Detection Challenge</b> Find ships on satellite images as quickly as possible <i>Featured</i> · 10 days to go · image data, object detection, object segmentation	\$60,000	867 teams
	<b>Google Analytics Customer Revenue Prediction</b> Predict how much GStore customers will spend <i>Featured</i> · 2 months to go · regression, tabular data	\$45,000	1,159 teams

# Machine learning basic methodology

In machine learning we are interested in finding models that work **during deployment**. Hence, in addition to building a model, we need to check it works.

Basic machine learning methodologies include two separate tasks:

- **Training:** A model is created using data and a quality metric. We also say that we **fit a model** to a dataset.
- **Testing:** The performance of the model during deployment is assessed using new, **unseen data**.

Without rigorous methodologies, models are very likely to be of little use.

# Agenda

[What is machine learning?](#)

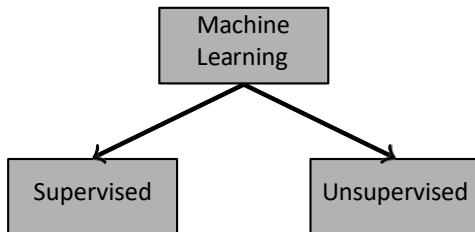
[The value of knowledge](#)

[The machine learning taxonomy](#)

[About CBU5201](#)

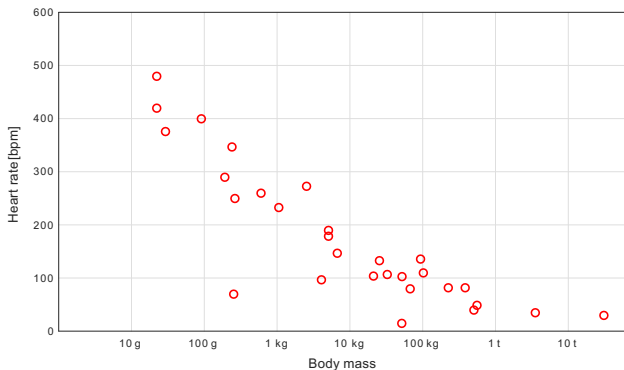
# Problem formulation

What **types of problems** can we formulate in machine learning?



# Supervised learning: Heart rate in the zoo

Can I guess the heart rate of an animal whose body mass I know, by looking at the heart rate and body mass of other animals?

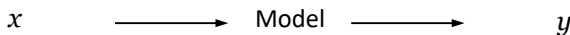


# Supervised learning

In supervised learning, we are given a **new item** (*rabbit*) such that the value of one of its attributes is **unknown** to us (*rabbit's heartbeat*).

Our goal is to **estimate** (*guess*) the **missing value** by learning from a **collection of known items** (*weight and heart rate of other animals*).

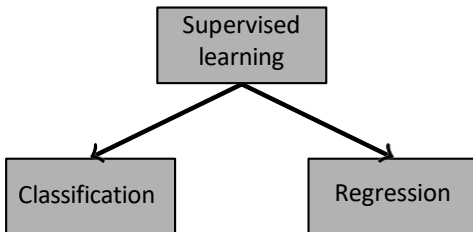
The challenge is then to build a model that maps one attribute  $x$ , known as the **predictor**, to another attribute  $y$ , which we call the **label**, using a dataset of **labelled examples**.



# Supervised learning: Classification and regression

Supervised learning is further divided into two categories depending on the type of label:

- **Classification:** The label is a **discrete** variable.  
*In a spam detector, 0 could mean email is spam, 1 it isn't*
- **Regression:** The label is a **continuous** variable.  
*The heart rate of an animal is a continuous label*



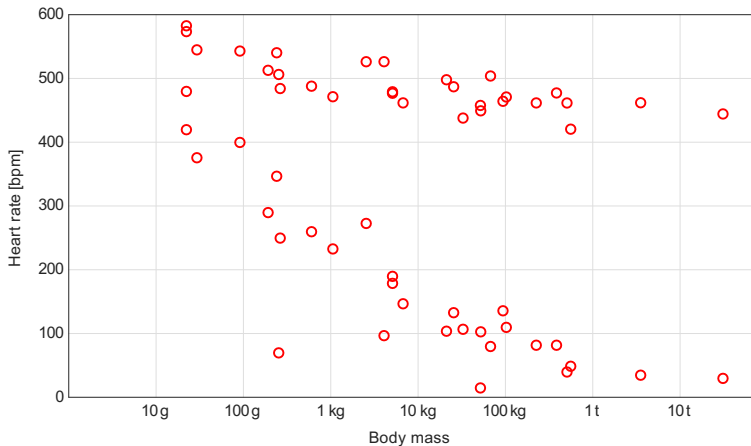
## Unsupervised learning: Heart rate in the galactic zoo





# Unsupervised learning: Heart rate in the galactic zoo

What can you conclude from this distribution of data points?



# Unsupervised learning

In unsupervised learning, we set out to **find the underlying structure** of our dataset. This can be useful to gain understanding, identify anomalies, compress our data and reduce processing time.

Applications of unsupervised learning include:

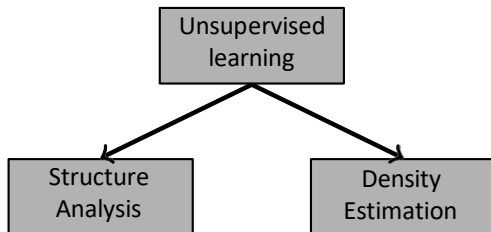
- Customer segmentation.
- Social community detection.
- Recommendation systems.
- Evolutionary analysis.

# Unsupervised learning

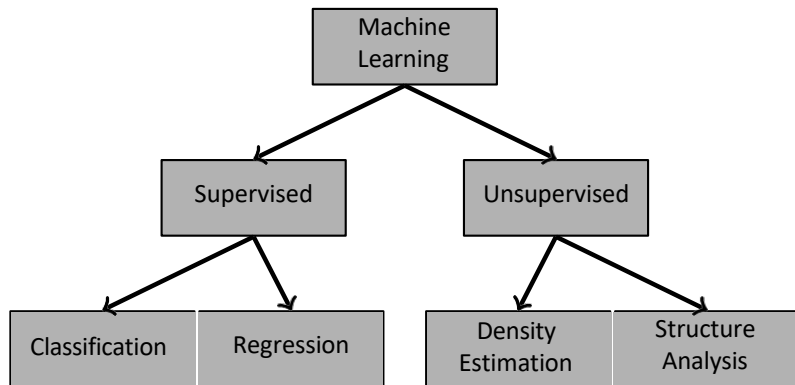
The underlying structure of a dataset can be studied using **structure analysis**, which includes:

- **Cluster analysis:** Focuses on groups of data points.
- **Component analysis:** Identifies directions of interest.

**Density estimation** techniques provide statistical models that describe the distribution of samples in the attribute space.



# Machine Learning taxonomy



# Agenda

[What is machine learning?](#)

[The value of knowledge](#)

[The machine learning taxonomy](#)

[About CBU5201](#)

# Learning goals

1. Understand the principles of ML, its scope and applications.
2. Be able to use the ML taxonomy to formulate meaningful questions and identify suitable techniques to answer them.
3. Discuss the relative merits of different ML techniques.
4. Be able to apply the methodology needed to build and evaluate ML solutions.
5. Be able to independently learn and confidently apply new ML techniques.
6. Critically analyse reports on ML applications and advances.
7. Understand model deployment and its main challenges.

# Module contents

- Week 1.1: Introduction
- Week 1.2: Regression
- Week 1.3: Methodology I
- Week 2.1: Classification I
- Week 2.2: Classification II
- Week 3.1: Methodology II
- Week 3.2: Neural networks & deep learning
- Week 4.1: Structure analysis
- Week 4.2: Density estimation
- Week 4.3: Intro to deployment

# Study outline

This module consists of 150 study hours (lectures, labs, assessment preparation, study time, etc). Its duration is 4 block teaching weeks:

- **8 Lecture sessions each week** (4 block teaching weeks in total).
  - **1 Lab each week** (1h/week) (for each block teaching week).
- Office hour/tutorial

**Check QM+ for more details.**



# Communication

- During our **lectures**.
- **Forum on QM+**: Primary means, questions might have been answered already and answers might be useful to others.
- **Email to c.liu@qmul.ac.uk**: Please make sure its subject is formatted as follows: "[ECS7020P] <DESCRIPTIVE SUBJECT HERE>"
- **Face to face**: On campus in the office hour or after each lecture.

# Assessment and labs

CBU5201 is assessed as follows:

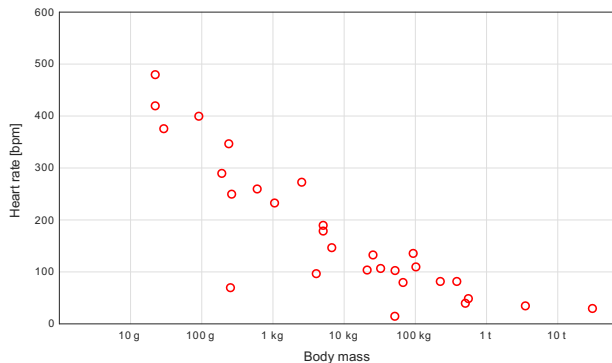
- Final exam: 70%
- CW activities: 30%

CW activities use Python (Jupyter Notebook) and consist of:

- Lab-based quizzes (10 %)
- Mini-project (20 %): Mini-project description will be released in BUPT Teaching Week 14, and the submission deadline will be in BUPT Teaching Week 16.

Check QM+ for deadlines. Note that **deadlines cannot be extended**: it is your responsibility to organise your work so that you can meet them.

# The strange case of the flatworm



The heart rate of a flatworm weighting less than 10 g

(a) Can't be guessed from this dataset

(b) Is  $\geq 300$  bpm

(c) None of the above

Know thy domain!