Queen Mary School Hainan
Queen Mary University of London

QHM5703 Principles of Machine Learning
Introduction

Nikesh Bajaj

Week 9 - 3/4 Nov 2025

Our team

Module Lecturers

- Nikesh Bajaj
- Jesús Requena Carrión

Teaching Fellow

- Kaushal Yadav

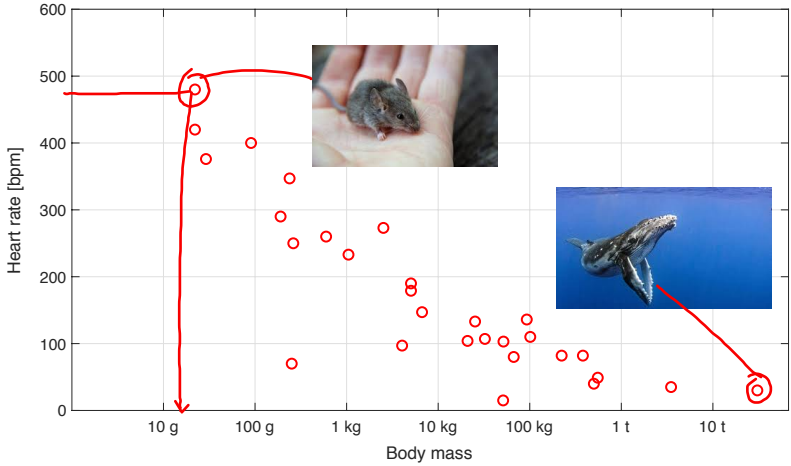
Queen Mary School Hainan
Queen Mary University of London

QHM5703 Principles of Machine Learning
Introduction

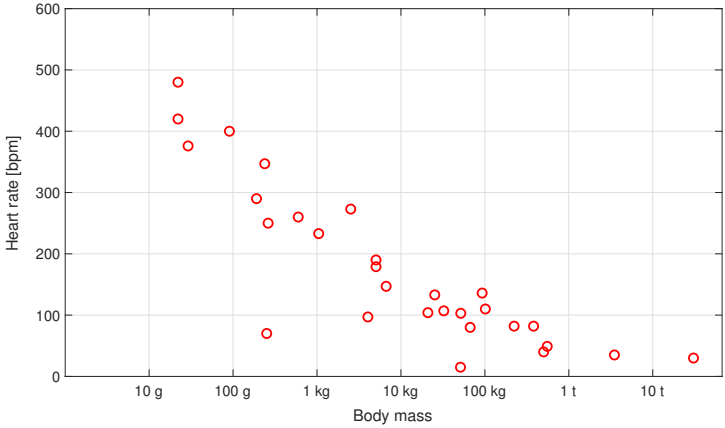
Nikesh Bajaj

Week 9 - 3/4 Nov 2025

From mouse to whale



From mouse to whale, through rabbit



A rabbit's resting heart beats at

(a) ≤ 100 bpm

(b) ≥ 300 bpm

(c) ≥ 100 bpm and ≤ 300 bpm

Go to www.menti.com and use code **6307 3349**

Agenda

What is machine learning?

The value of knowledge

The machine learning taxonomy

About QHM5703

Machine learning a.k.a. 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	2.5×10^3	250
Humpback whale	30×10^6	30
...	...	...

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

Note: Unfortunately, authors use different words 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. Large 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.*

In machine learning, we will 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 line of code (in Python):

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

Machine learning is a branch of **data [science]** ...

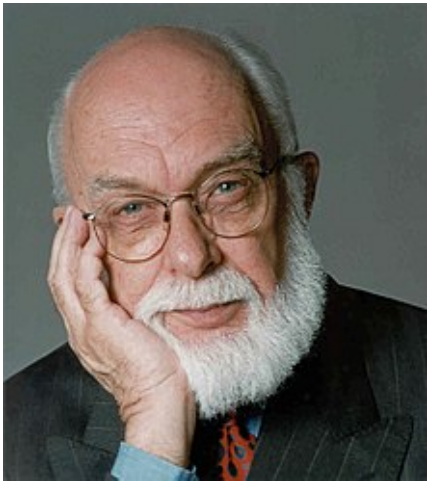
Science is not about sophisticated instrumentation, maths or theories: it is about **evaluating our knowledge**.

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

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

Neither proposition is scientific nor unscientific: they can be true or false. It is the **way we evaluate them** that can be scientific or unscientific.

Pseudoscience



Dowsers claim that using a divining rod or a pendulum they can locate ground water and that current scientific theories are unable to explain it.

Watch *James Randi exposes dowsing*:

- www.youtube.com/watch?v=cqoYrSd94kA
- https://www.bilibili.com/video/BV1oJ1wBjEzX/

Randi: *My concern is not* **how** *they do it, but* **if** *they do it.*

Our notion of machine learning

If there is no data, there is no machine learning. This doesn't mean that machine learning is all about data. **Many machine learning experts use a dataset-first framework**: we start with a dataset, then [formulate a problem and finally] produce a model.

In QHM5703 **we use a deployment-first** (problem-first) **perspective**: we start with a problem, then secure a dataset, finally produce a model.

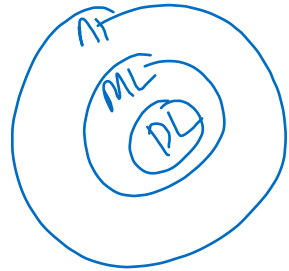
Accordingly, we define machine learning as:

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	think rationally



Warning: When media, companies and even universities talk about AI, they can refer to anything involving computers, including machine learning or a spreadsheet. Be critical!

Some AI solutions use machine learning, some others do not. Also, machine learning can be used to solve non-AI problems.

This module is about machine learning, not AI.

Agenda

What is machine learning?

The value of knowledge

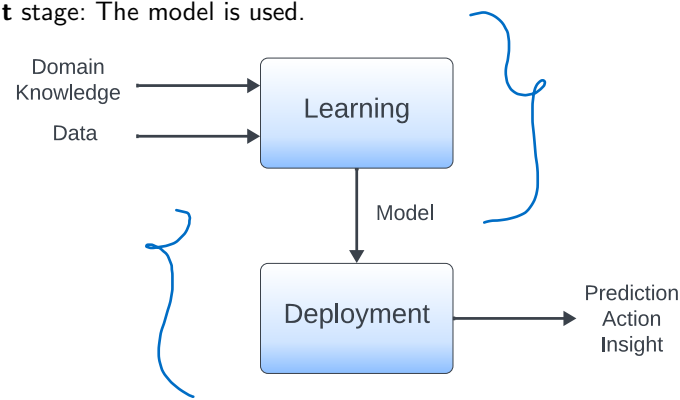
The machine learning taxonomy

About QHM5703

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 basic stages:

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



Deployment: eCommerce

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 a dataset. We 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

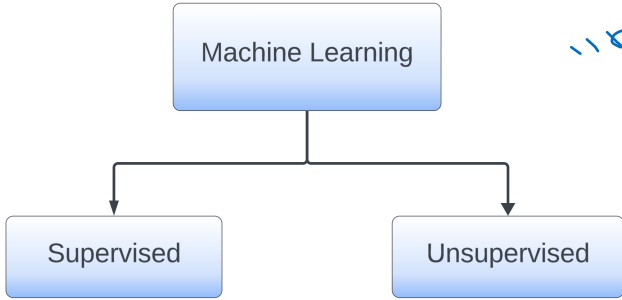
The machine learning taxonomy

About QHM5703

Problem formulation

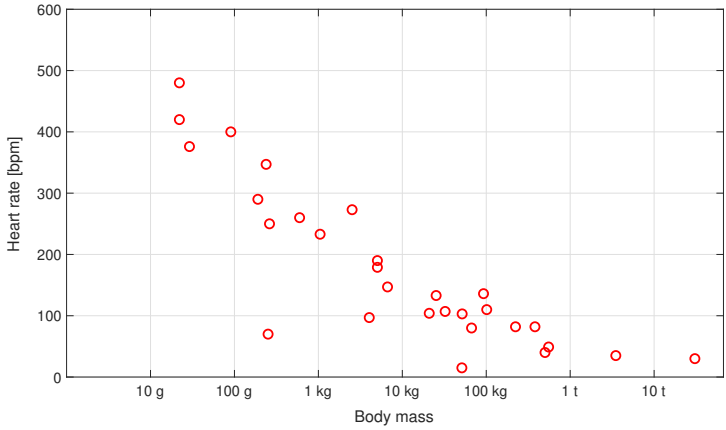
What **types of problems** can we formulate in machine learning?

"problems"



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

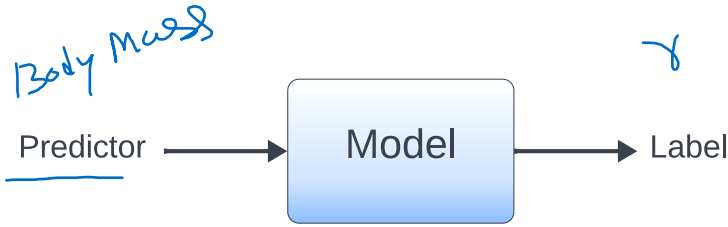
In supervised learning, we are given a **new item** (*rabbit*) such that the value of one of its attributes is **unknown** to us (*rabbit's heartbeat*).

Our goal is to **estimate** (*guess*) the **missing value** based on **what we know** (*body mass*).

The challenge is then to build a suitable model using a **collection of known items** (*weight and heart rate of other animals*).

Supervised learning

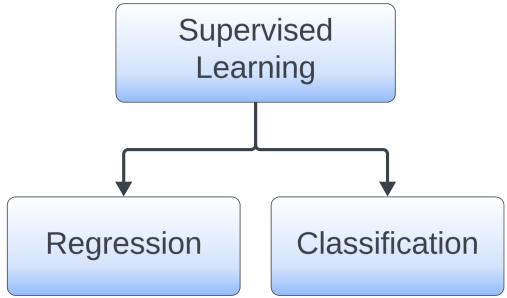
Models in supervised learning map one attribute x, known as the **predictor**, to another attribute y, which we call the **label** and are built using a dataset of **labelled examples**.



Supervised learning: Classification and regression

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

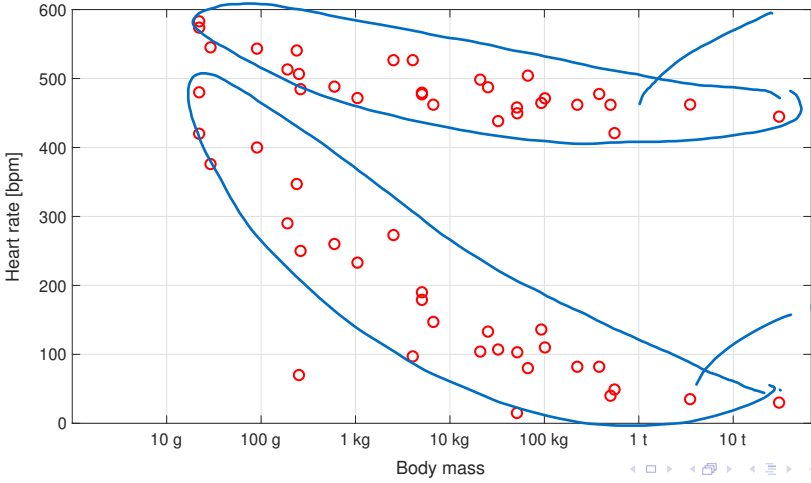
- **Classification**: The label is a **discrete** variable.
  In a spam detector, 0 could mean email is spam, 1 it isn't
- **Regression**: The label is a **continuous** variable.
  The heart rate of an animal is a continuous label



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:

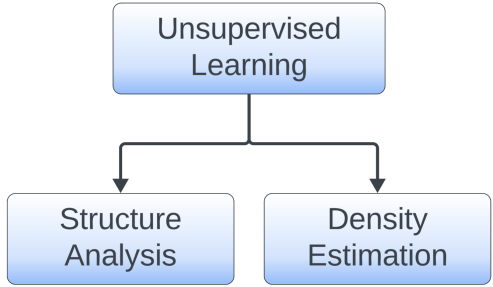
- Customer segmentation.
- Social community detection.
- Recommendation systems.
- Evolutionary analysis.

Unsupervised learning

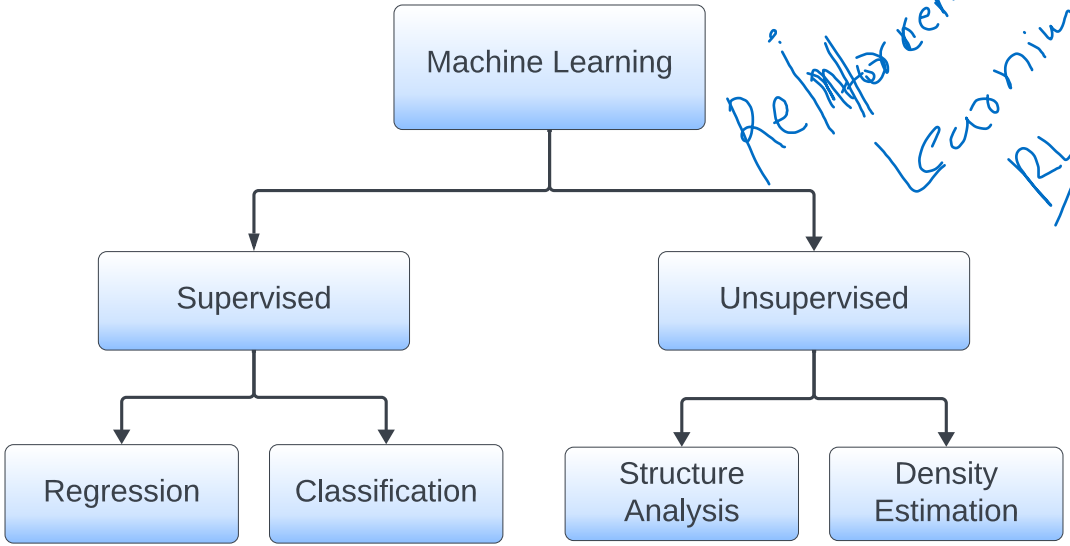
The underlying structure of a dataset can be studied using **structure discovery**, which includes:

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

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



Machine Learning taxonomy



Reinforcement Learning RL

Agenda

What is machine learning?

The value of knowledge

The machine learning taxonomy

About QHM5703

Learning outcomes

1. Demonstrate the understand of principles of machine learning, its **scope and applications**. (T)
2. Able to use **scientific rigour** to formulate a meaningful question and use the machine learning taxonomy to identify suitable techniques to answer them.(T)
3. Discuss the relative merits of different machine learning techniques. (T)
4. Able to **apply the methodology** needed to build and **evaluate** machine learning solutions. (P)
5. Able to independently learn and confidently apply new machine learning techniques. (P)
6. **Critically analyse** the machine learning techniques and their applications.(T+P)
7. Examine the main challenges of model **deployment and pipelines**. (T+P)
8. Able to develop the **critical skills** to formulate the problems. (T)
9. Demonstrate the **professional** dimension of machine learning. (T)

.

High thinking-to-doing ratio

Module contents

- Week 09: Introduction, Regression, (Lab + tutorials)
- Week 10: Methodology I, Classification I, (Lab + tutorials)
- Week 11: Lab: Assessment Support Sessions
- Week 12: Lab + Quiz + Mini Project
- Week 13: Lab + Mini Project
- Week 14: Structure analysis, Density estimation , (Lab + tutorials)
- Week 15: Classification II, Methodology II, Neural Network (Lab + tutorials)

Study outline

This module consists of 150 study hours (lectures, labs, assessment preparation, study time, etc). Its duration is 7 weeks (4 weeks for lectures and 3 weeks for labs):

- **Lectures** (9h/week).
- **Labs** (4h/week).

Check QM+ for more details.

Communication and feedback

- During our **lectures** and **lab sessions**.
- **Forum on QM+**: Primary means, questions might have been answered already and answers might be useful to others.
- **Email**: Please make sure its subject is formatted as follows:
  "[QHM5703] <DESCRIPTIVE SUBJECT HERE>"
- **Face to face**: On campus or remotely (via MS teams).

Assessment and labs

QHM5703 is assessed as follows:

- Final exam: 60 %
- Coursework (CW): 40 %

CW activities use Python (Google Colab environment/**Anaconda Cloud**) and consist of:

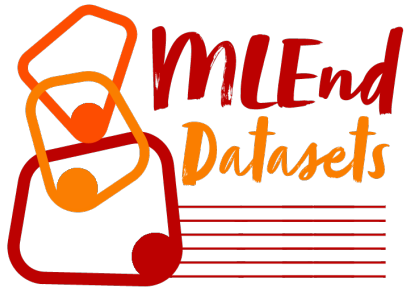
- Lab-based quizzes/activities (16 %).
- Data Collection (4 %) We will create a dataset.
- Mini-project (20 %) We will build a machine learning solution *using collected dataset*.

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

Anaconda Cloud

- Create an account on `https://anaconda.cloud/`

The MLEnd Datasets



https://MLEndDatasets.github.io/

MLEnd Spoken Numerals Dataset (20/21)
MLEnd Hums and Whistles Dataset (21/22)
MLEnd London Sounds Dataset (22/23)
MLEnd Yummy Dataset (23/24)
~~**MLEnd Deception Dataset** (24/25)~~
this year - **MLEnd Hums and Whistles II Dataset** (25/26)

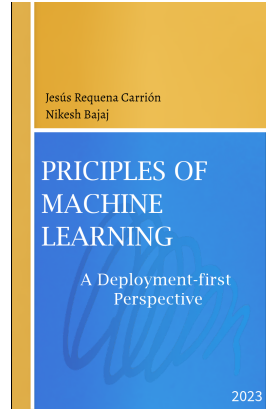
Learning resources

In the Module Content tab on QM+ you will find:

- Lecture slides
- Exercises and solutions
- Lab Jupyter notebooks and other material

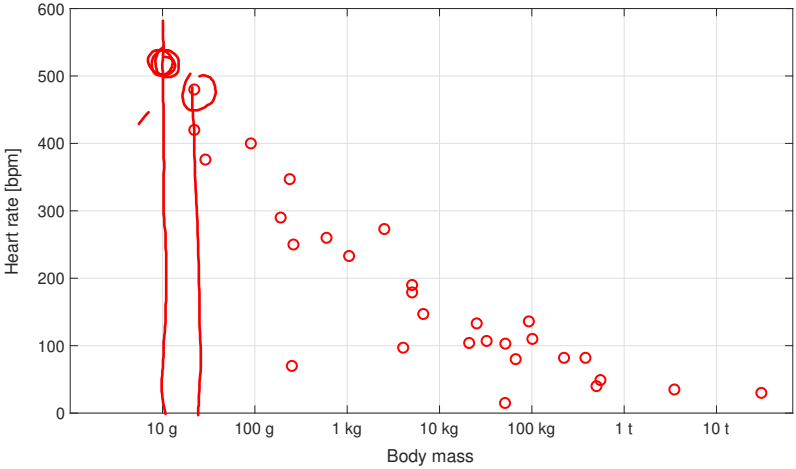
And in the Learning Resources tab you will find:

- A list of recommended textbooks and links to electronic versions.
- Individual chapters of a textbook we are writing for you.



https://PMLBook.github.io
first four chapters are online

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 ≥ 300 bpm

(c) None of the above

Know thy domain!

Queen Mary
University of London