

SCHOOL OF ELECTRONIC ENGINEERING AND COMPUTER SCIENCE
QUEEN MARY UNIVERSITY OF LONDON

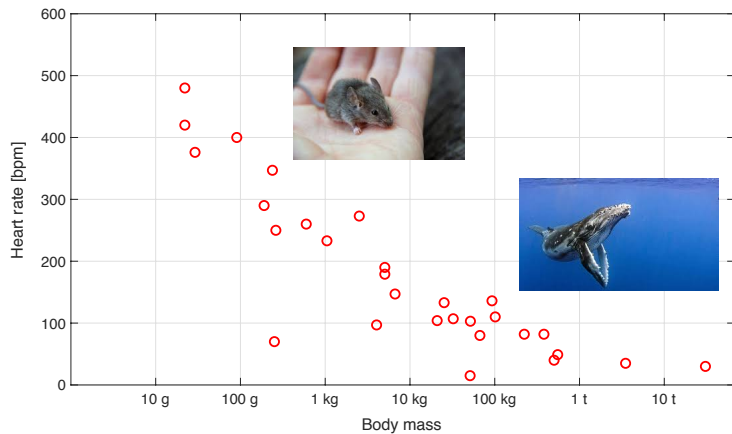
ECS766 Data Mining

Week 1: Introduction to Data Mining

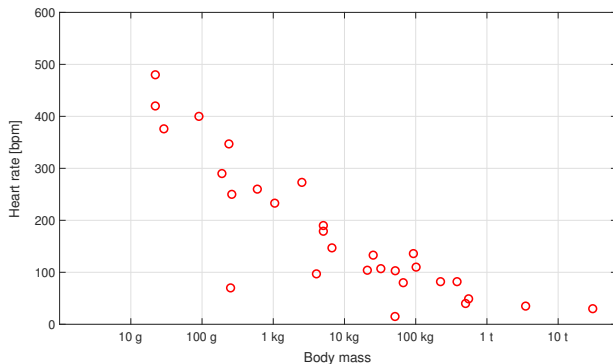
Dr Jesús Requena Carrión

25 Sept 2019

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

Agenda

What is Data Mining?

The value of knowledge

Models in Data Science

A Taxonomy of problems in Data Science

Data Mining

Data Mining is the human activity consisting in extracting **knowledge** from **data**

Data Mining: Data

Data Mining is the human activity consisting in extracting knowledge from **data**.

- **Data** is anything that has been **recorded**.
- **Datasets** are collections of **items** (samples, examples, instances or data points) that are described by a set of **attributes**. One attribute can be seen as one **dimension**.

| Animal | Body mass [g] | Heart rate [bpm] |
|----------------|-------------------|------------------|
| Wild mouse | 22 | 480 |
| Rabbit | 2.5×10^3 | 250 |
| Humpback whale | 30×10^6 | 30 |
| ... | ... | ... |

Data Mining: Knowledge

Data Mining is the human activity consisting in extracting **knowledge** from data.

- **Knowledge** can be represented as a
 - **Proposition** (statement, law)
Example: *Smaller animals have a faster heartbeat*
 - **Narrative** (description, storytelling)
Example: *The size of an animal seems to be related to its heartbeat. In general, larger animals tend to have a slow heartbeat. For instance, the humpback whale...*
 - **Model** (mathematical or computer)
Example: $r = 235 \times m^{-1/4}$

Data Mining uses **Data Science** tools and techniques.

Data and Science

Science is not about using sophisticated instrumentation, models or techniques, science is about **evaluating** propositions, narratives and models.

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

Proposition 1: *The earth is flat*

Proposition 2: *The earth is roughly spherical*

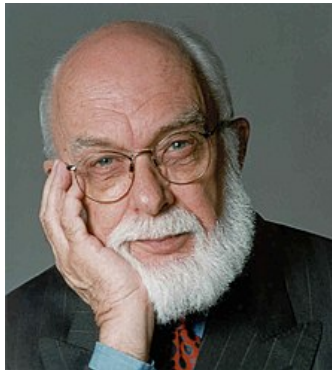
Data and Science

There is no such thing as neutral data and raw data won't prove a proposition is true.

Data need to be **analysed scientifically**.

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

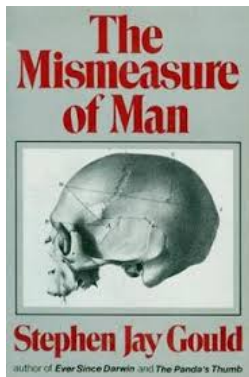
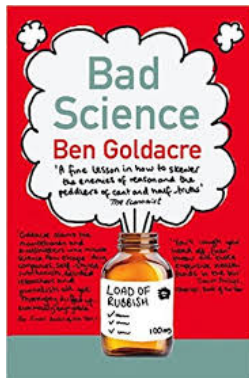
Pseudoscience



James Randi exposes dowsing:

www.youtube.com/watch?v=cqoYrSd94kA

Recommended popular science books



Agenda

What is Data Mining?

The value of knowledge

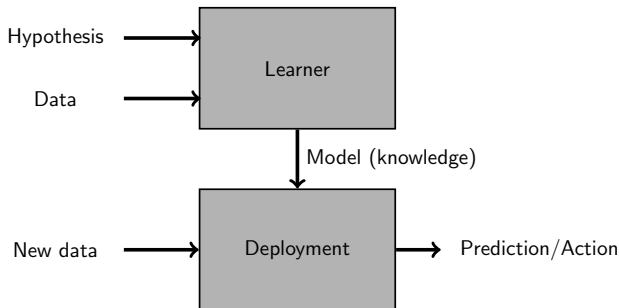
Models in Data Science

A Taxonomy of problems in Data Science

The two stages of Data Mining

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

1. **Learning** stage: The model is built.
2. **Deployment** (inference, production) stage: The model is used.



Data Science example: eCommerce

Inspired by your shopping trends



Recommendations for you in Grocery



Data Science example: Banking



BARCLAYS How much can I borrow?

Get a rough idea of how much you could borrow for a residential mortgage based on your personal circumstances.

How many applicants? [?](#)

☒ 1 ☐ 2 ☐ 3 ☐ 4

My income [?](#)

| | | |
|---|--------|---|
| £ | Yearly | ▼ |
|---|--------|---|

Regular spending

The amount you spend to repay credit and store cards, catalogue purchases, loans, overdrafts, maintenance and your pension. You don't need to tell us about general household spending, such as groceries, travel and utility bills.

| | | |
|---|---------|---|
| £ | Monthly | ▼ |
|---|---------|---|

Reason for mortgage

| | |
|---------------|---|
| Please select | ▼ |
|---------------|---|

Calculate

Data Science example: Face recognition



Data Science example: Spam filter

From zumgala_bouda@aol.fr ☆
Subject **PLEASE VERY URGENT**
To Undisclosed recipients; ☆

Hello,
I know this means of communication may not be morally right to you as a person but I also have had a great thought about it and I have come to this conclusion which I am about to share with you.

INTRODUCTION: I am the Credit Manager U. B. A Bank of Burkina Faso Ouagadougou and in one way or the other was hoping you will cooperate with me as a partner in a project of transferring an abandoned fund of a late customer of the bank worth of \$18,000,000 (Eighteen Million Dollars US).

This will be disbursed or shared between the both of us in these percentages, 55% to me and 35% to you while 10% will be for expenses both parties might have incurred during the process of transferring.

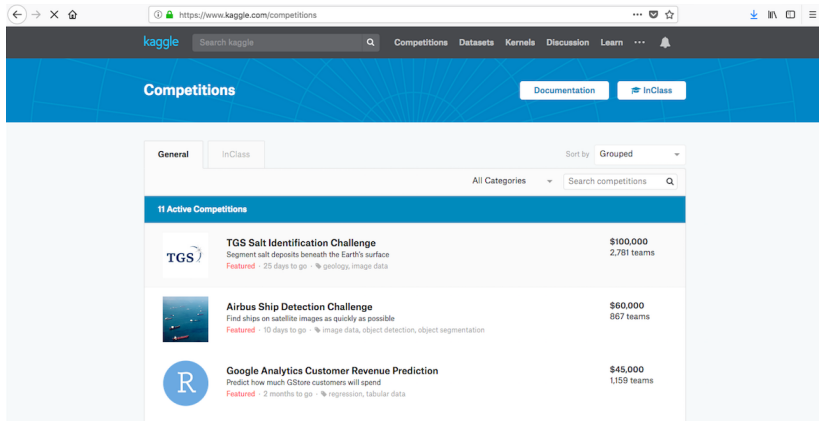
I

await for your response so that we can commence on this project as soon as possible.




PLEASE REPLY ME THOUGH THIS MY PRIVATE EMAIL ADDRSS {sumbala.bouda@yandex.com}

Regards,
MR. BOUDA
Credit Manager U. B. A Bank of
Burkina Faso Ouagadougou

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, Learn, and a bell icon. Below the header, there's a blue banner with the word 'Competitions' and buttons for 'Documentation' and 'InClass'. The main content area has tabs for 'General' and 'InClass', with 'General' selected. A 'Sort by' dropdown is set to 'Grouped'. Below this is a search bar for competitions. A blue bar indicates '11 Active Competitions'. Three competitions are listed:

| Competition Logo | Competition Name | Prize Pool | Teams | Details |
|---|--|------------|-------------|---------|
|  | TGS Salt Identification Challenge Segment salt deposits beneath the Earth's surface <i>Featured</i> · 25 days to go · geology, image data | \$100,000 | 2,781 teams | |
|  | Airbus Ship Detection Challenge 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 | |
|  | Google Analytics Customer Revenue Prediction Predict how much GStore customers will spend <i>Featured</i> · 2 months to go · regression, tabular data | \$45,000 | 1,159 teams | |

Agenda

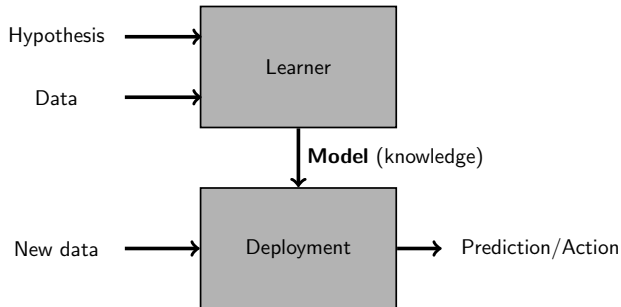
What is Data Mining?

The value of knowledge

Models in Data Science

A Taxonomy of problems in Data Science

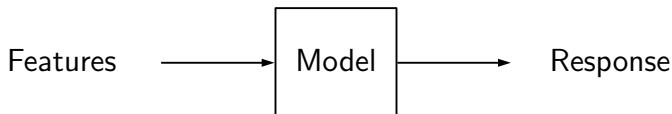
The two stages of Data Mining



What is a model?

Models relate two sets of variables:

- **Features** (*independent variables, input variables, predictors*).
- **Response** (*dependent variables, output variable*).



Types of variables

The basic types of variables are:

- Numeric/continuous
 - Real numbers (temperature, voltage, pixel intensity value)
 - Ordering and distance are defined
- Categorical/discrete
 - Equality is defined
 - Neither ordering nor distance are defined
- Ordinal
 - Categories with ordering (low/medium/high)
 - Ordering and distance are defined

These basic types can be represented by scalar value and lead to vector/arrays.

Mathematical and computer models

Mathematical and computer models are **equivalent**: mathematical models can be implemented numerically and for every computer model there is a mathematical formulation.

- Mathematical models express the relationship between features and responses by using **mathematical expressions**.

$$y = x + 3x^2$$

- Computer or numerical models are **computer programs** that implement the operations necessary to calculate a response for a given set of features.

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

Note on reading maths

γνωθι σεαυτον

$$E_{MSE} = \frac{1}{N} \sum_{i=1}^N e_i^2$$

gnothi seauton

E_{MSE} is equal to 1 over N times the summation from i equals 1 to N of e sub i squared

know thyself

E_{MSE} is the average squared error

Parametric and non-parametric models

Models are grouped into different families. The most basic classification of models distinguishes between:

- **Parametric models** have a pre-defined shape that can be adjusted by tuning a set of **parameters**.
- **Non-parametric models** make no assumptions about the shape and have no parameters that need adjusting.

Non-parametric models are more flexible than parametric ones. However, they need more data and are harder to interpret.

Hyperparameters

Hyperparameters allow us to distinguish specific models within a family of models and shouldn't be confused with conventional parameters.

For instance, take the family of polynomial models. The degree of the polynomial is a hyperparameter and the set of coefficients of a concrete model constitute its parameters.

$$\text{Degree 1: } y = a_1x + a_0$$

$$\text{Degree 2: } y = a_2x^2 + a_1x + a_0$$

$$\text{Degree 3: } y = a_3x^3 + a_2x^2 + a_1x + a_0$$

...

Model training and testing

In Data Mining we are interested in building **the best model**. Hence, in order to identify the best model we need to have a notion of **model quality**. However, our goal is to build models that work well during *deployment*, i.e. when presented with new data.

Data Mining approaches incorporate the following two stages:

- **Training**: Given some notion of quality and data, a model is created.
- **Testing**: Using unseen data, the quality of the model is reassessed.

Therefore, models need to be able to **generalise**. The situation where a model performs well for training data, but poorly for test data, is known as **overfitting**.

Model validation

What if we **train several models**, how can we choose the best one? A **flawed approach** would be to compare the quality of each model during testing and selecting the best (that's peeking!).

We only use model testing on unseen data to assess the quality of a model after training and selecting.

Model **validation** provides different techniques to select our final model. For instance, if we consider a polynomial family of models, validation allows us to set the hyperparameter (degree) and training would adjust the parameters (coefficient).

Agenda

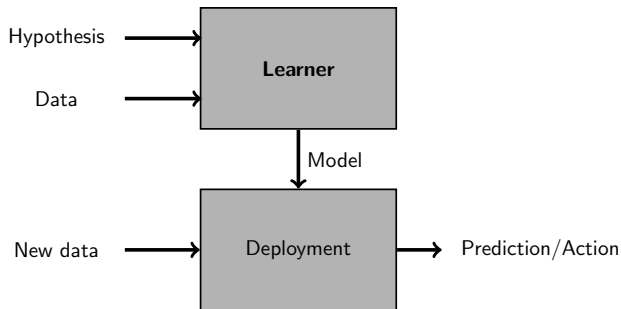
What is Data Mining?

The value of knowledge

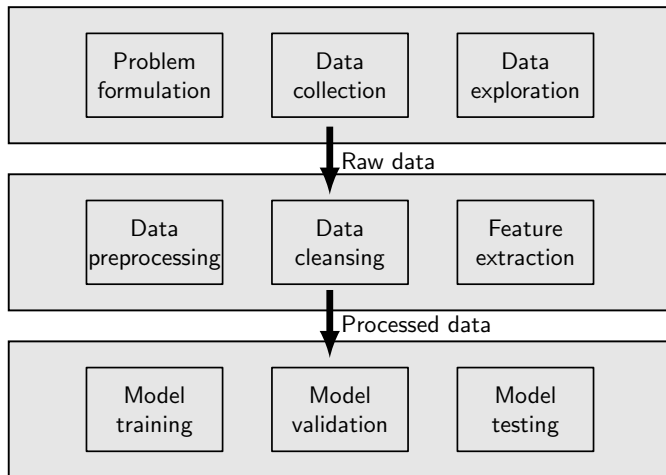
Models in Data Science

A Taxonomy of problems in Data Science

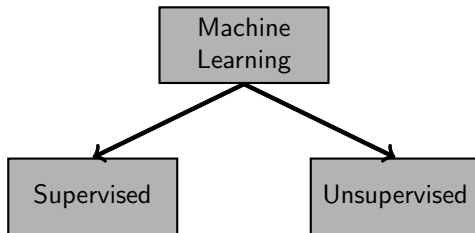
The two stages of Data Mining



Data Science pipeline

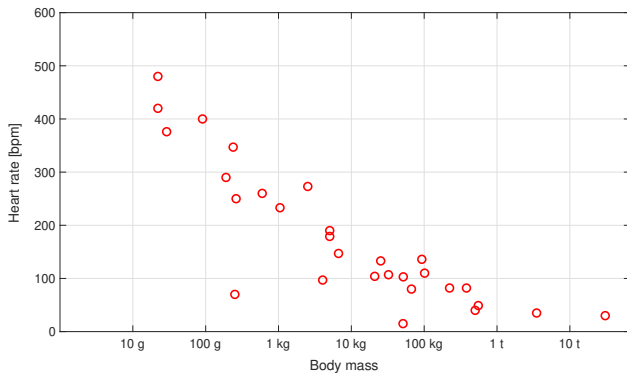


Problem formulation



Supervised learning: Heart rate in the zoo

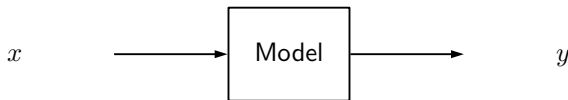
Can I guess the heart rate of an animal whose body mass I know, by looking at the heart rate and body mass of other animals?



Supervised learning

In supervised learning, we are given a **new item** (*rabbit*) and 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 the values of a **collection of previous items** (*weight and heart rate of other animals*).

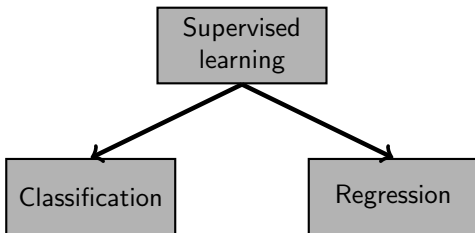
Mathematically, our challenge is to build a model that maps one attribute x to another attribute y which we call the **label**, by learning from a dataset of **labelled examples** (x_i, y_i) .



Supervised learning: Classification and regression

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

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

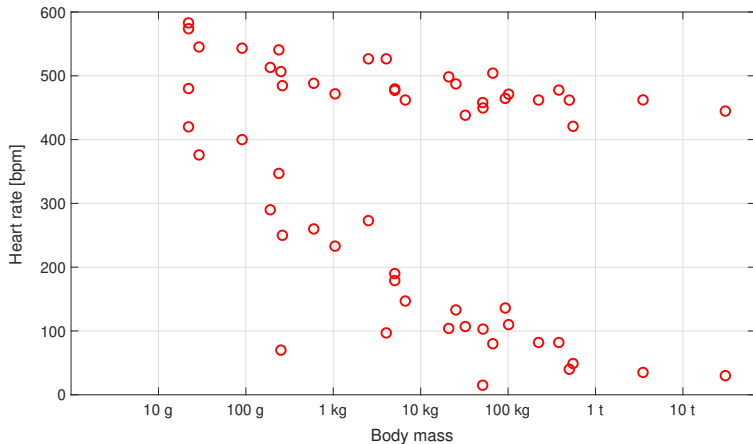


Unsupervised learning: Heart rate in the galactic zoo



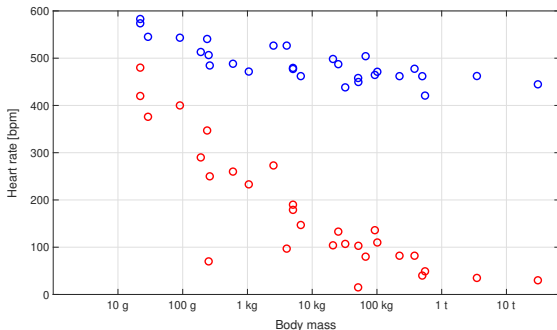
Unsupervised learning: Heart rate in the galactic zoo

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. Among other uses, this can be useful to gain understanding, identify anomalies, compress our data and reduce processing time.

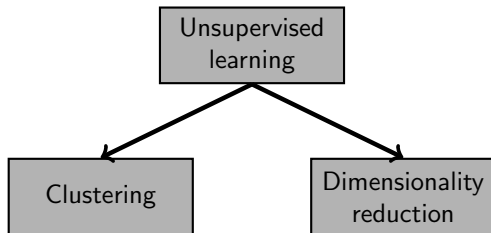


Unsupervised learning: Clustering and dimensionality reduction

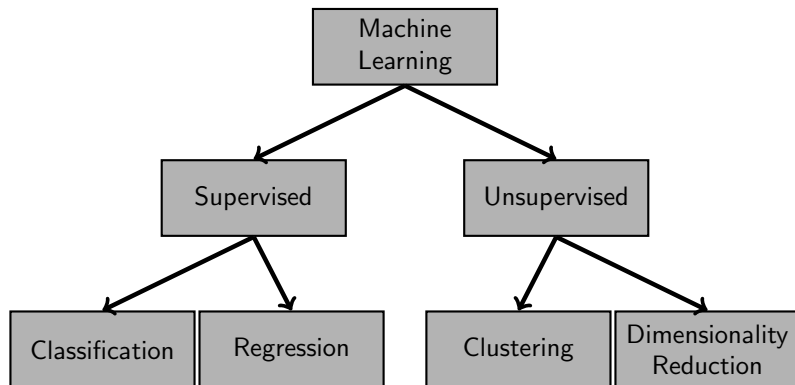
Two main unsupervised learning techniques are:

- **Clustering**: Clusters of data points of similar nature are identified.
- **Dimensionality reduction**: Reduced set of attributes is generated.

Other techniques include **outlier detection** and **quantile estimation**.



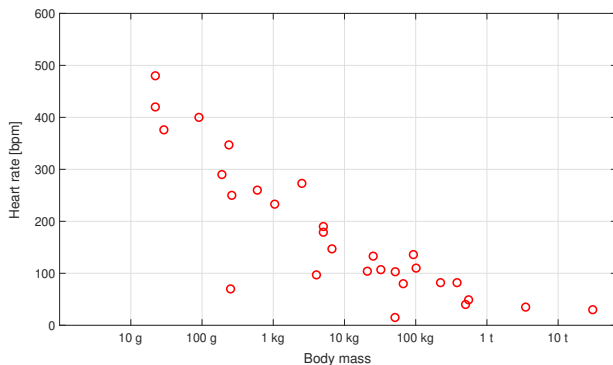
Data science taxonomy



Craniometry sought to justify 19th century society and its inequalities by using "scientific evidence". Is this a problem of the past?

- Our models essentially perform computations, are they immune to the type of unconscious bias that affects humans?
- Can we use anonymised data to build our models to eliminate bias?
- Is it OK for a third-party to store and analyse individuals' data to build a model, for companies to use targeted advertising and for political parties to use targeted campaigning?

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

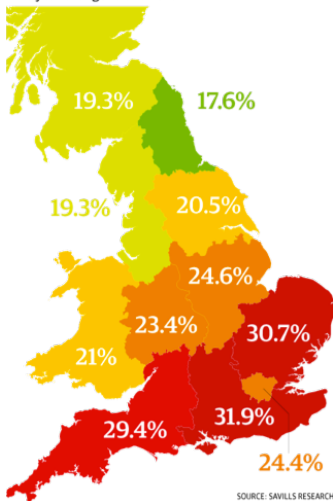
(b) is ≥ 300 bpm

(c) = NaN

House prices

House price growth forecast

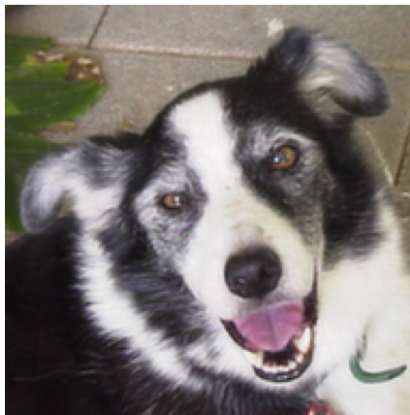
Five year change to end 2018



Predicting the growth of house prices belongs to the following category of Machine Learning problems:

- (a) Classification
- (b) Regression
- (c) Clustering

Is it a dog?



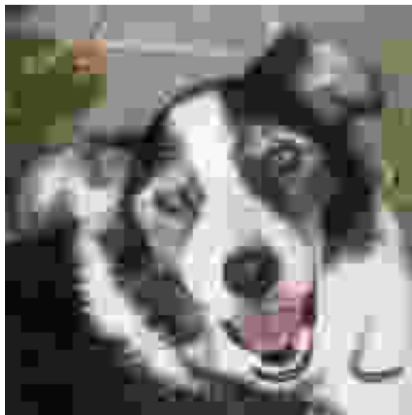
An algorithm that decides whether there is a dog in a picture belongs to the following category of Machine Learning problems:

(a) Classification

(b) Regression

(c) Clustering

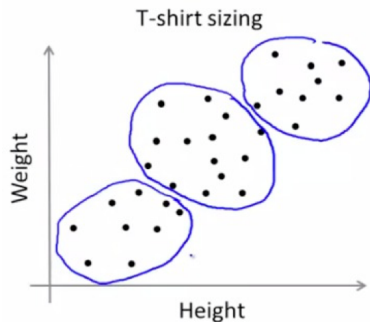
The same dog



Based on the previous picture, we have obtained a new one. Its quality is poorer, but represents the same dog. This is a case of:

- (a) Dimensionality reduction
- (b) Regression
- (c) Clustering

Which size?



Different brands produce T-shirts with different weights and sizes. Based on the weight and size, we define 3 discrete groups, namely S, M and L. We just carried out:

- (a) Dimensionality reduction
- (b) Classification
- (c) Clustering