Queen Mary
University of London

ECS766P Data Mining

Week 1: Introduction

Emmanouil Benetos
emmanouil.benetos@qmul.ac.uk

September 2021

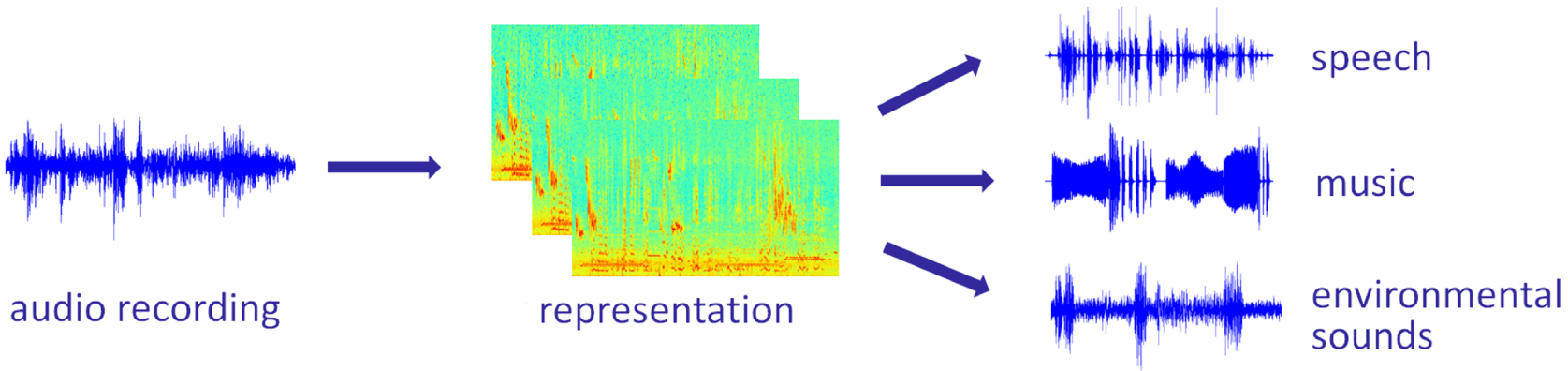
School of EECS, Queen Mary University of London

Preliminaries

Introducing myself

Based at Centre for Digital Music & Centre for Intelligent Sensing

Working on audio analysis - also called machine listening



Communication

- Forum on QM+: primary means, questions might have been answered already and answers might be useful to others
- In the lecture and lab sessions
- Email: will aim to respond within 48 hours. Please include "[ECS766]" in the subject line

Module Contents

- Introduction to data mining (week 1)
- Data (week 2)
- Data preprocessing (week 3)
- Data analysis and visualisation (week 4)
- Storage and data management (week 5)
- Classification and clustering (week 6)
- *Reading week (week 7)*
- Association analysis (week 8)
- Outlier detection (week 9)
- Web mining (week 10)
- Data mining applications & data ethics (week 11)
- *Revision (week 12)*

Assessment and labs

Assessment:

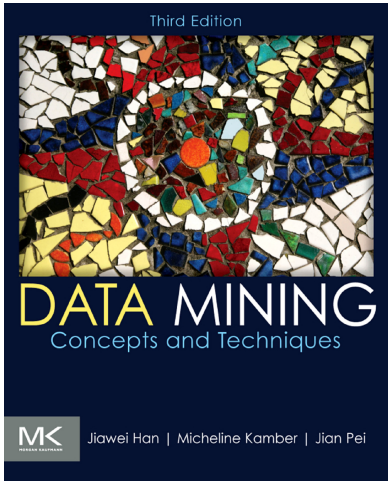
- Final exam: 60%
- 4 lab assignments: 40% (10% each assignment)

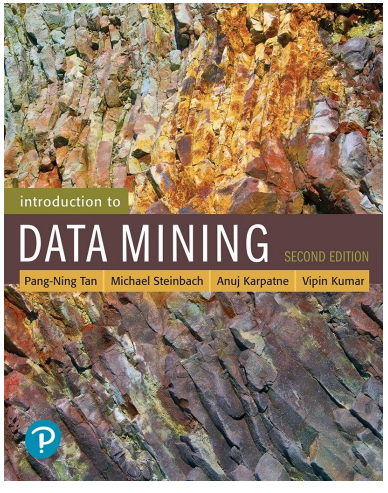
Lab sessions:

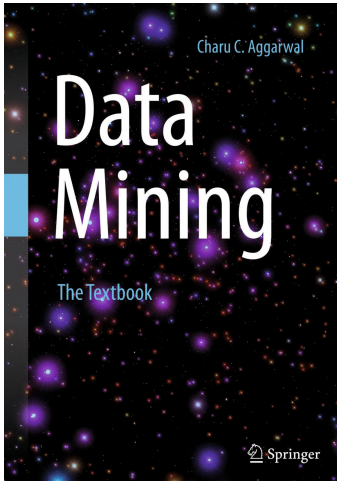
- 9 sessions: 1 introduction + 8 graded labs (2 labs/assignment)
- You have been allocated to one lab slot - please check your personal timetables/calendars for your allocated slot
- You can choose whether to attend online or on-campus
- Lab slot 1: Fridays 11:00-13:00 (on-campus at Queen's Building W206, or online in MS Teams)
- Lab slot 2: Fridays 15:00-17:00 (on-campus at ITL Building 2nd floor, or online in MS Teams)

Reading

- Material uploaded onto QM+ (not just the lecture slides!)
- J. Han, M. Kamber, J. Pei, “Data Mining: Concepts and Techniques”, 3rd edition, Elsevier/Morgan Kaufmann, 2012 [primary]
- P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, “Introduction to Data Mining”, 2nd edition, Pearson, 2019
- C. C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015







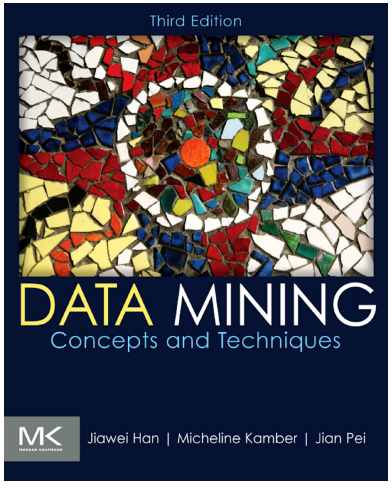
Basic Rules

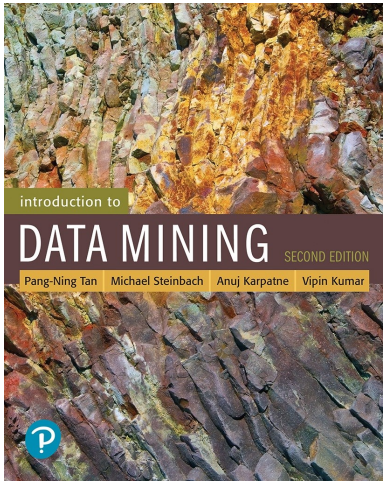
- Slides will be uploaded at least 1 day before the lecture
- All lecture videos will be recorded
- You can use the chat in Collaborate or your microphone to ask questions during the lectures
- Please remember to mute yourself if you are not asking a question
- During the on-campus labs, please ensure you follow the Queen Mary Covid Code

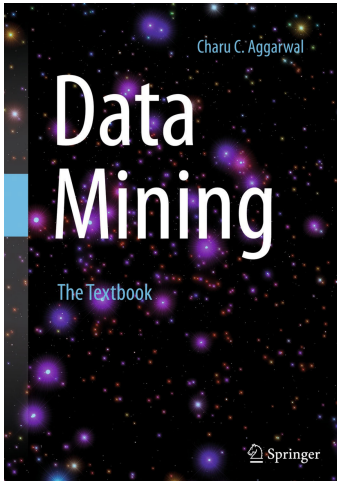
Table of contents

This week's reading

- **Chapter 1 of J. Han, M. Kamber, J. Pei, "Data Mining: Concepts and Techniques", 3rd edition, Elsevier/Morgan Kaufmann, 2012**
- Chapter 1 of P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, "Introduction to Data Mining", 2nd edition, Pearson, 2019
- Chapter 1 of C. C. Aggarwal, "Data Mining: The Textbook", Springer, 2015



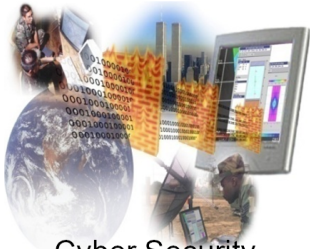




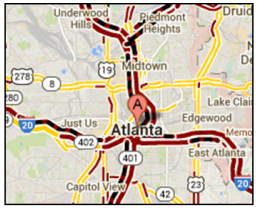
Why data mining?

Data is everywhere!

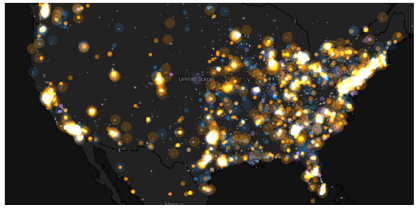
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies.
- New mantra: gather whatever data you can whenever and wherever possible.
- Expectations: Gathered data will have value either for the purpose collected or for a purpose not envisioned.


Cyber Security

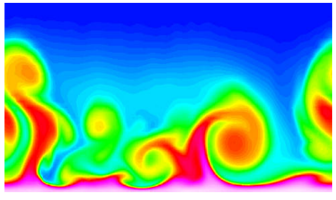
E-Commerce



Traffic Patterns


Social Networking: Twitter


Sensor Networks


Computational Simulations

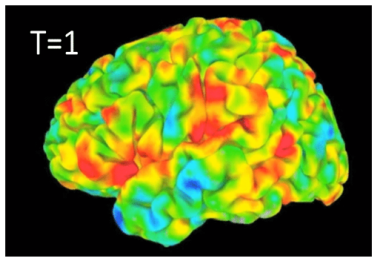
Why data mining? Scientific viewpoint

Data mining helps scientists

- In automated analysis of massive datasets
- In hypothesis formation

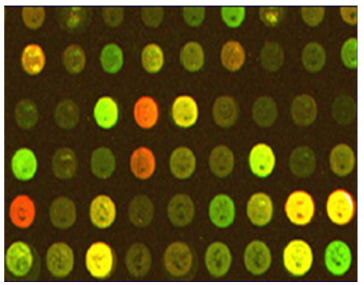
Data is collected and stored at enormous speeds

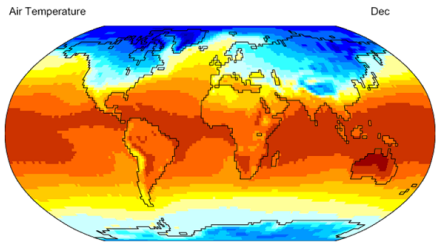
- Remote sensors on satellites
- Telescopes scanning the skies
- Gene expression data
- FMRI data



fMRI Data from Brain


Sky Survey Data


Gene Expression Data



Surface Temperature of Earth

Why data mining? Commercial viewpoint

Lots of data is being collected and warehoused

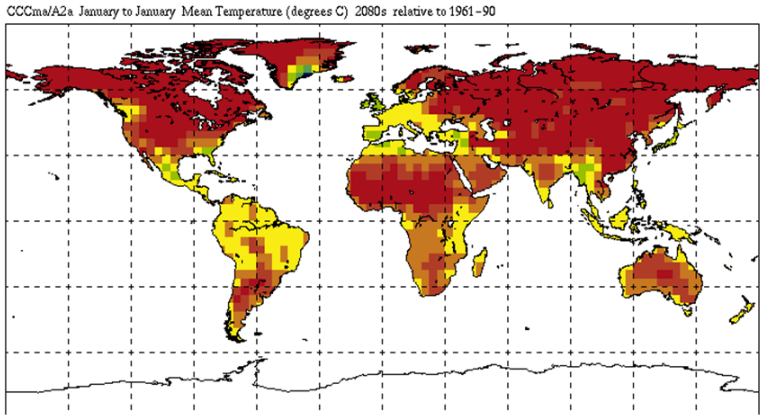
- Web data (e.g. Facebook has billions of active users)
- E-commerce (e.g. Amazon handles millions of deliveries per day)
- Bank/credit card transactions

Computers have become cheaper and more powerful

Competitive pressure is strong: companies strive to provide better customized services

Data mining for solving global challenges

Improving healthcare

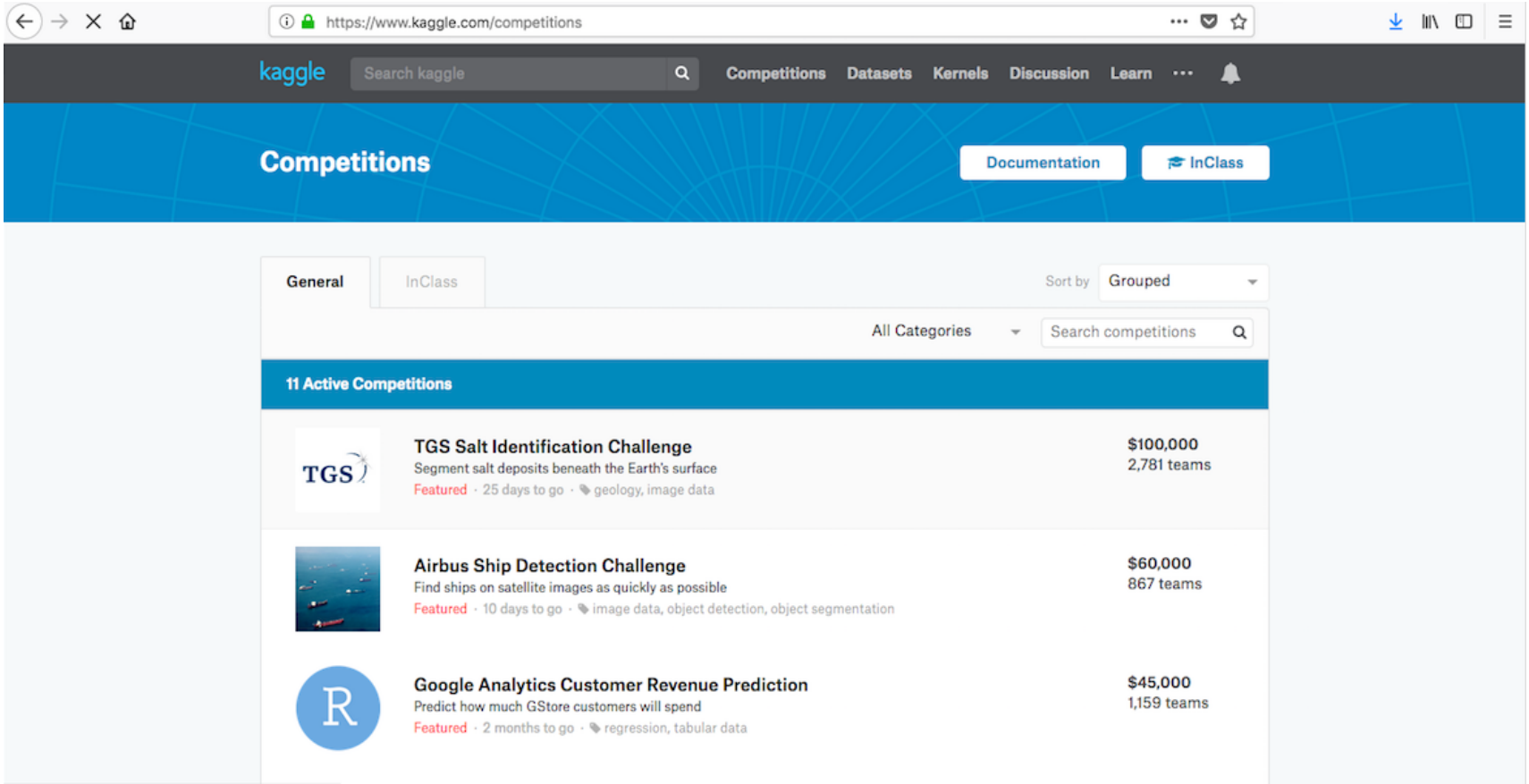


Predicting the impact of climate change

Finding alternative/ green energy sources

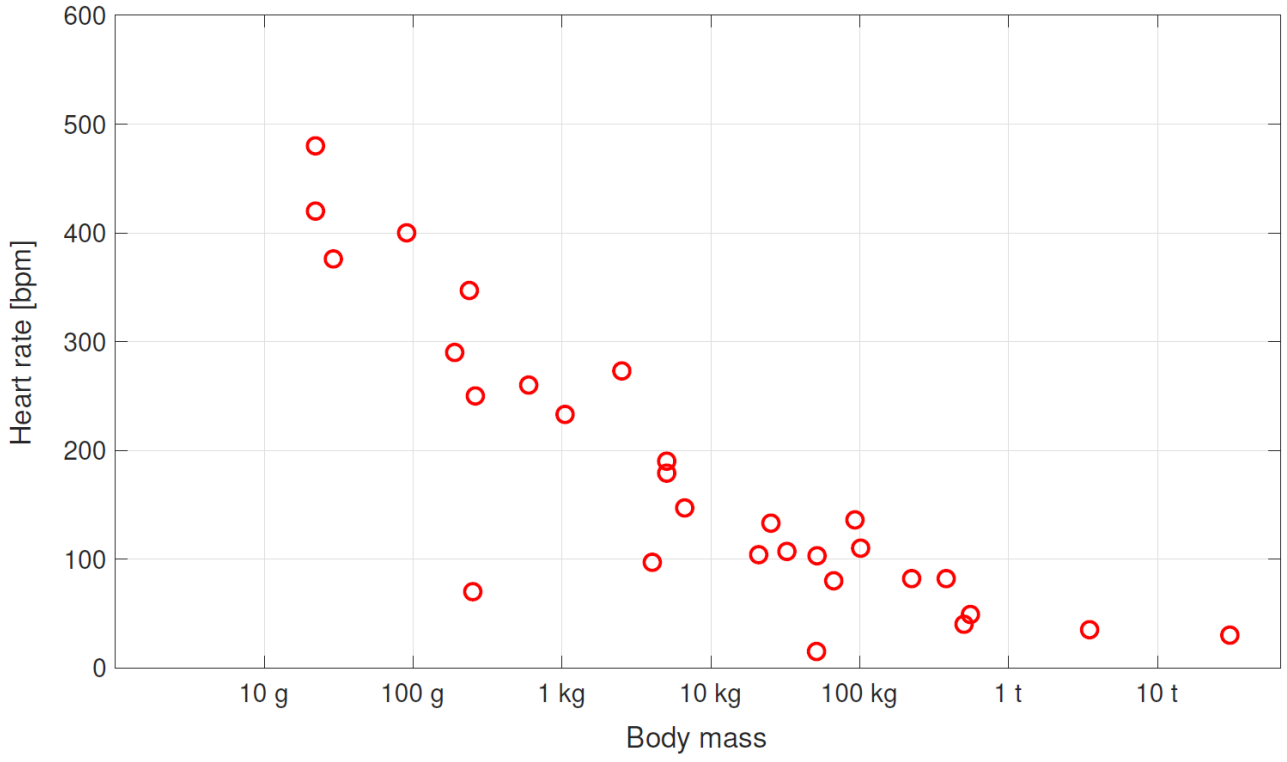
Reducing hunger and poverty by increasing agriculture production

Data mining competitions



What is data mining?

Example: animal body mass vs. heart rate



A rabbit's resting heart beats at:

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

Data

Data refers to characteristics, numerical or categorical, that are collected through observation.

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
...	...	...

Knowledge

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 and Science

Data Mining uses Data Science tools and techniques.

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.

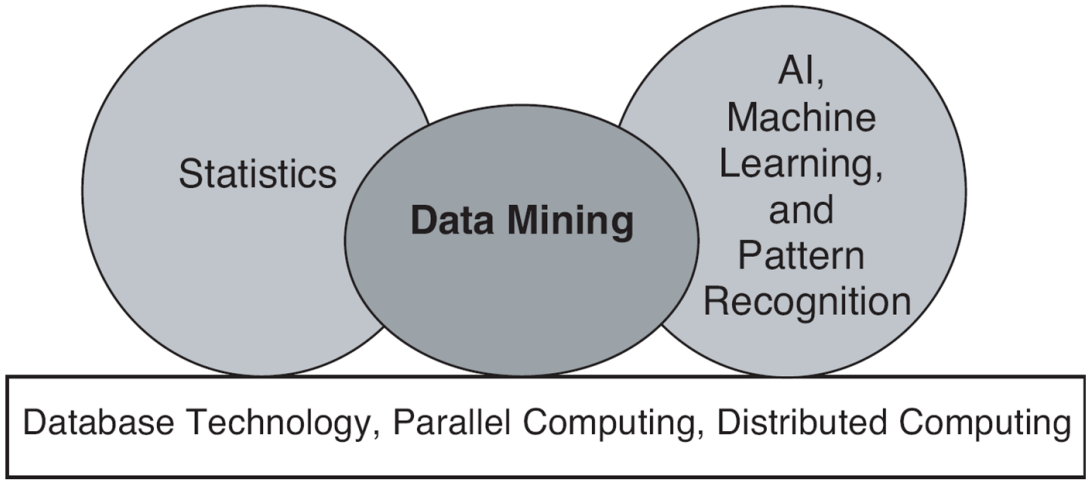
There is no such thing as neutral data and raw data will not prove a proposition is true.

Origins of data mining

Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

A key component of the emerging field of data science and data-driven discovery

Traditional techniques may be unsuitable due to data that is large-scale, high dimensional, heterogeneous, complex, or distributed



Data mining definitions

Definition 1

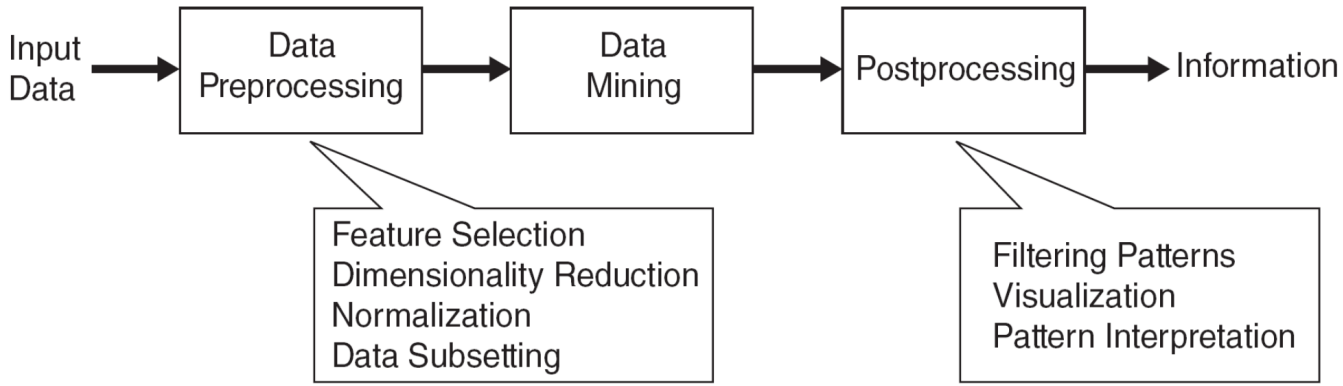
Non-trivial extraction of implicit, previously unknown and potentially useful information from data.

Definition 2

The human activity consisting in extracting knowledge from data.

Definition 3

Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.



What is not data mining?

What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

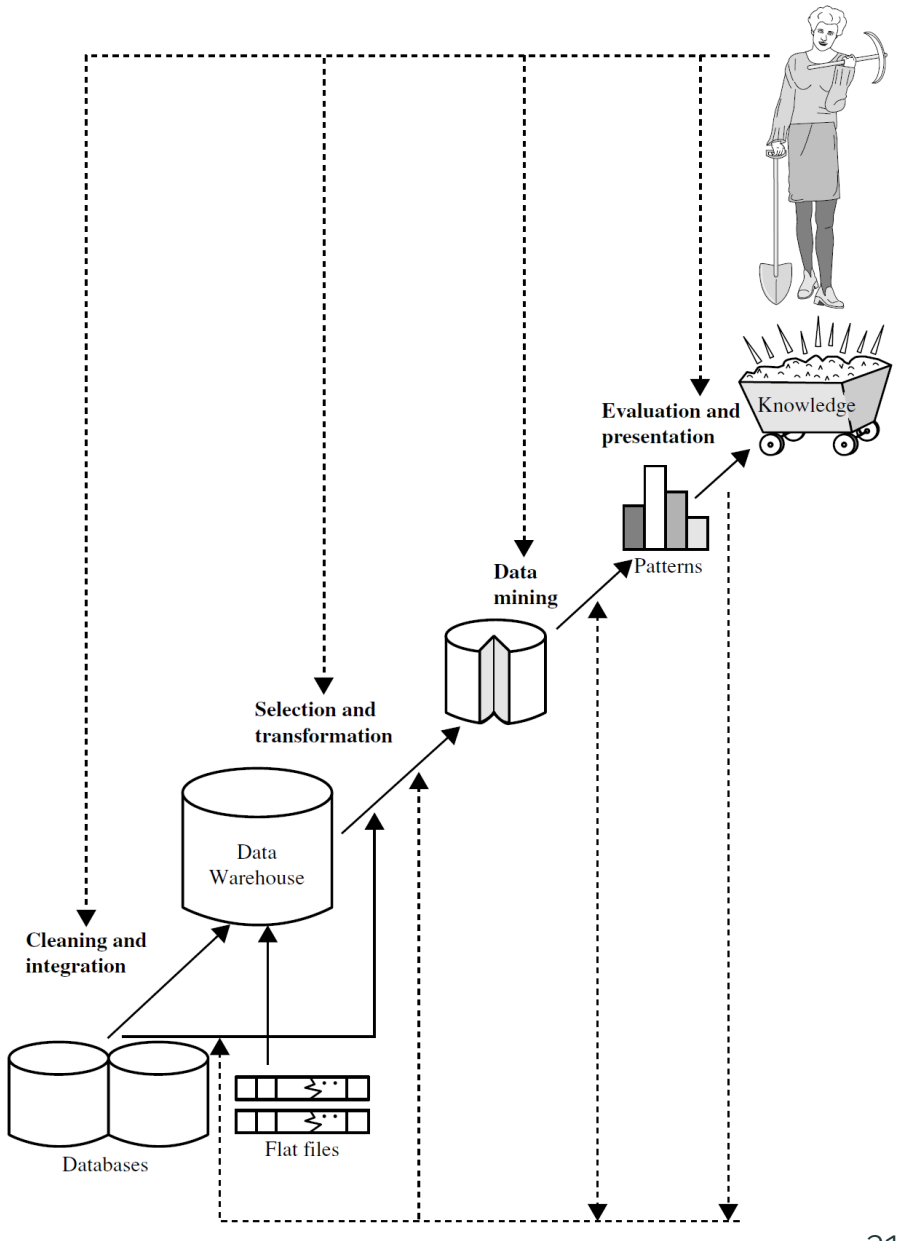
What is Data Mining?

- Certain names are more prevalent in certain UK locations (e.g. Jones, Williams, Taylor)
- Group together documents returned by search engine according to their context (e.g. Amazon rainforest vs. Amazon.com)

Knowledge discovery from data (KDD)

Many people treat data mining as part of knowledge discovery from data (KDD), which includes the following steps:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

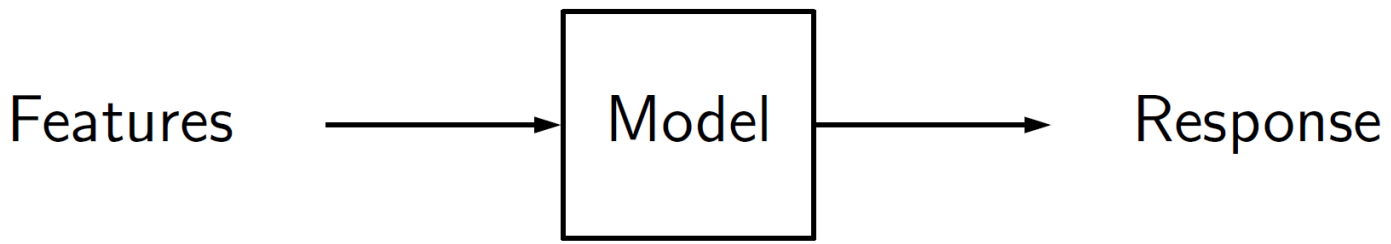


Models in data science

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 values and lead to vectors/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)
```

Mathematical conventions

- Scalar values – x
- Vectors – $\mathbf{x}$
- Matrices – $\mathbf{X}$
- Sets – $\mathcal{X}$

$\rightarrow$ also see FAQ Crib Sheet on Math Notation in QM+

Parametric and non-parametric models

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

- Parametric models have pre-defined assumptions on the shape of the data; models can be adjusted by tuning a set of parameters.
- Non-parametric models make no assumptions about the data 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 should not 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.

$$\begin{aligned} &\text{Degree 1:} && y = a_1 x + a_0 \\ &\text{Degree 2:} && y = a_2 x^2 + a_1 x + a_0 \\ &\text{Degree 3:} && y = a_3 x^3 + a_2 x^2 + a_1 x + a_0 \end{aligned}$$

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.

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 (coefficients).

Data mining tasks

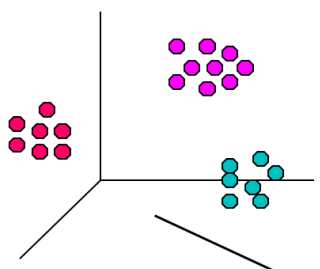
Data mining tasks

In general, data mining tasks can be classified into two categories: descriptive and predictive.

Descriptive mining tasks characterise properties of the data in a target data set.

Predictive mining tasks perform induction on the current data in order to make predictions.

Data mining tasks

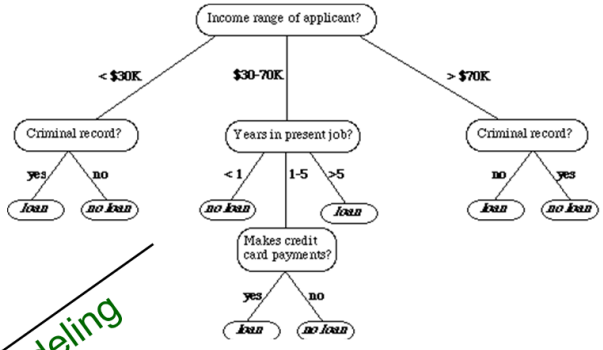


Clustering

Data

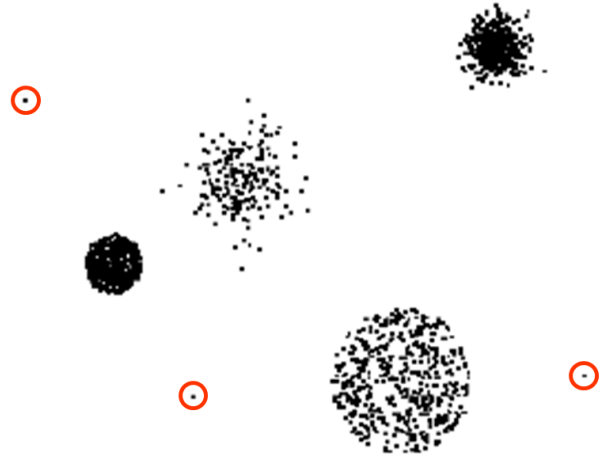
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling



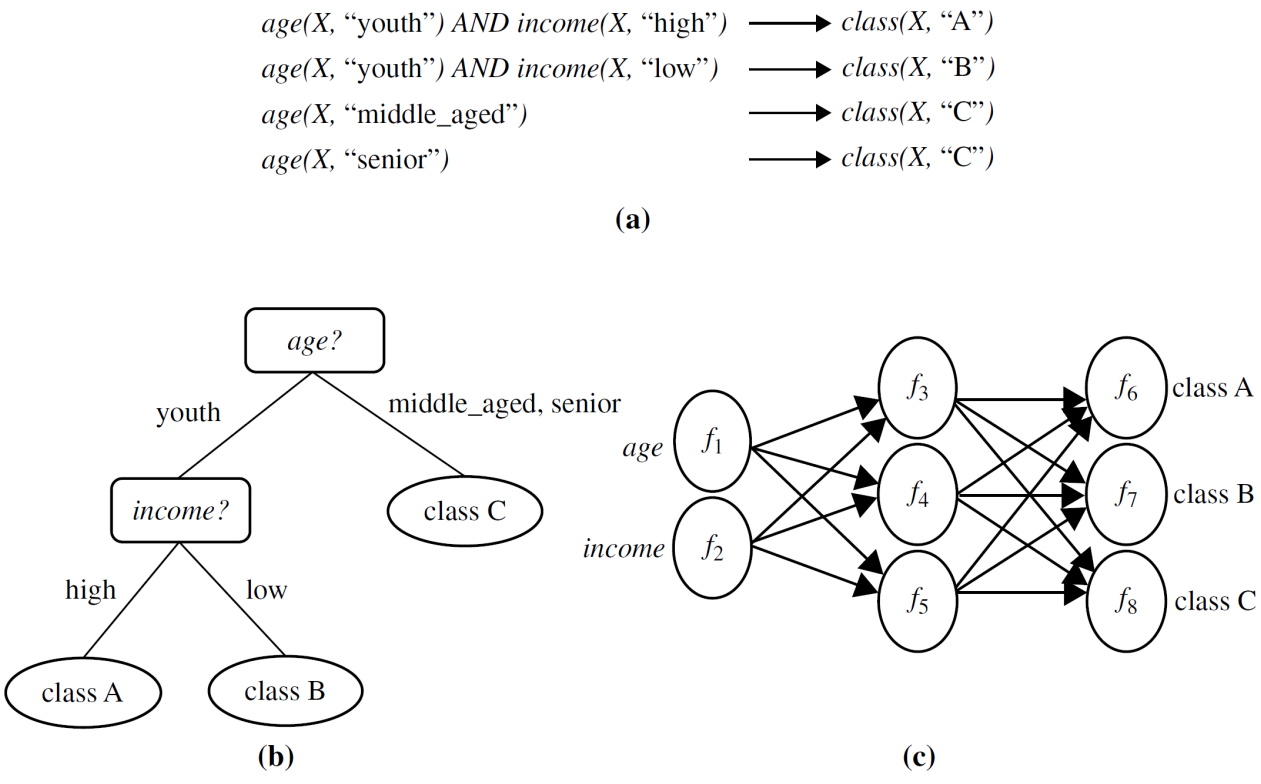
Association Rules

Anomaly Detection



Classification

The process of finding a model (or function) that describes and distinguishes data classes or concepts.



Classification model examples: (a) IF-THEN rules, (b) decision tree, (c) neural network.

Classification Application 1

Fraud Detection

Goal: Predict fraudulent cases in credit card transactions.

Approach:

- Use credit card transactions and the information on its account-holder as attributes (e.g. when does a customer buy, what do they buy...).
- Label past transactions as fraud or fair transactions. This forms the class attribute.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

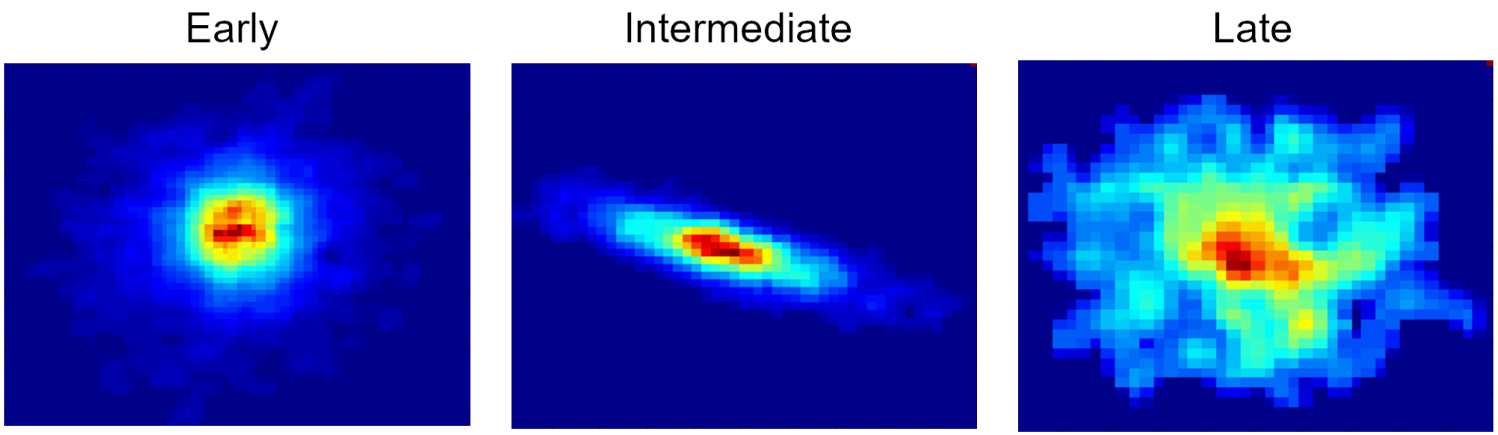
Classification Application 2

Galaxy classification

Goal: To classify galaxies in terms of formation stages

Approach:

- Segment the image.
- Attributes: image features, galaxy characteristics.
- Model galaxy classes based on the above features.



Regression

Whereas classification predicts categorical (discrete, unordered) labels, regression predicts missing or unavailable numerical data values.

Extensively studied in statistics, neural network fields.

Examples:

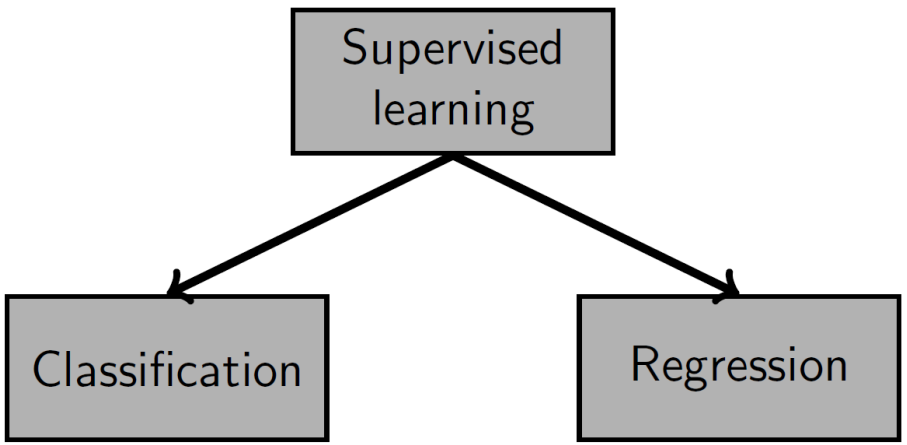
- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices.

Supervised learning

In the field of machine learning, both classification and regression fall under the category of supervised learning.

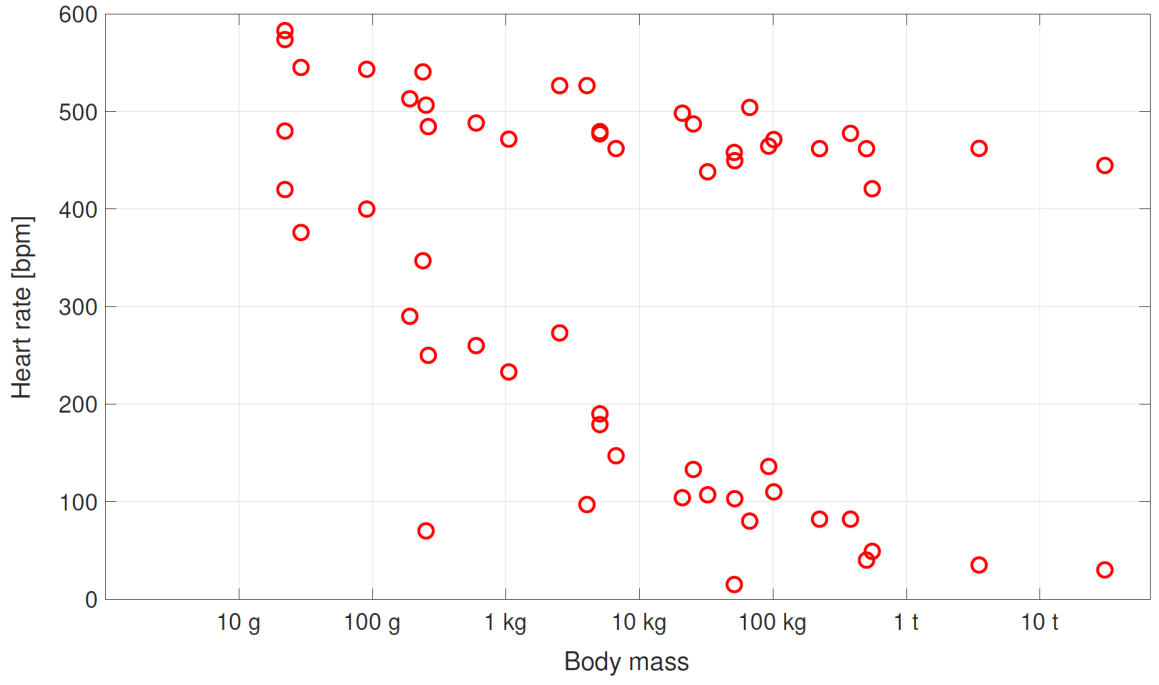
Supervised learning

In supervised learning, we are given a new item and the value of one of its attributes is unknown to us. Our goal is to estimate the missing value by learning from the values of a collection of previous items.



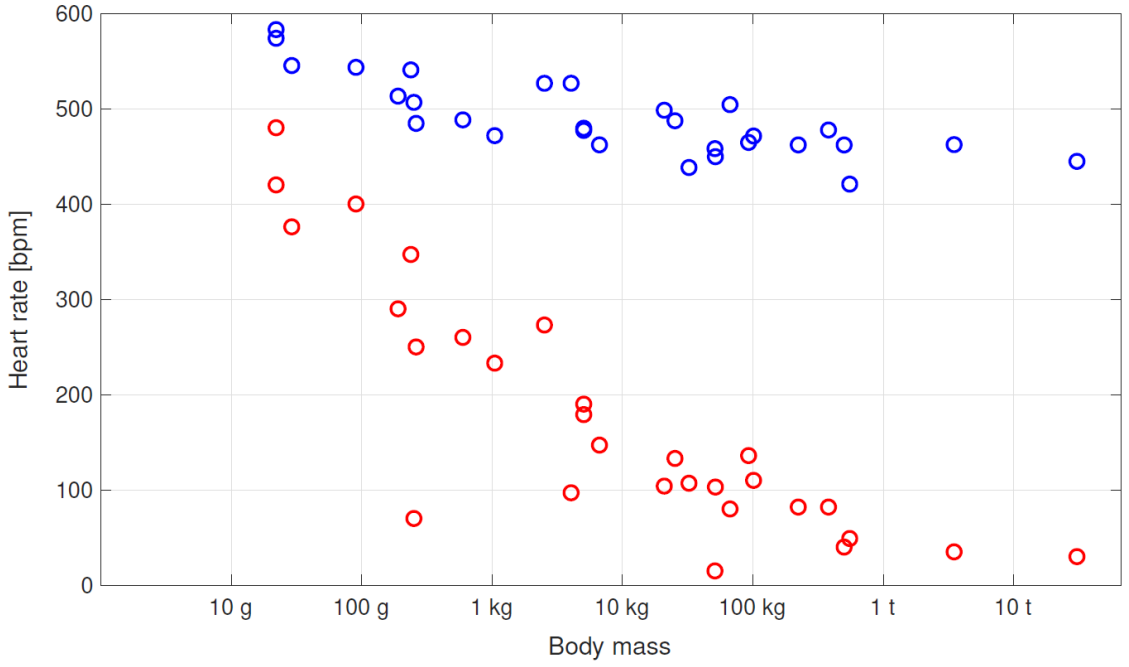
Clustering

Finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups.



Clustering

In clustering, 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.



Clustering Application 1

Document clustering

Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.

Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster the documents.

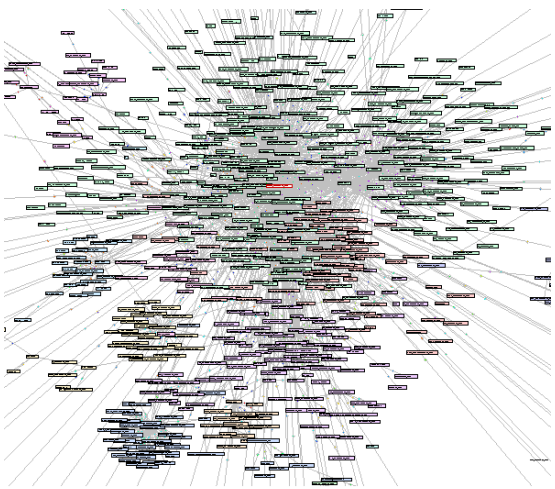


Figure: Enron email dataset (https://www.cs.cmu.edu/∼./enron/)

Clustering Application 2

Market segmentation

Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

Approach:

- Collect different attributes of customers based on their geographical and lifestyle related information.
- Find clusters of similar customers.
- Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

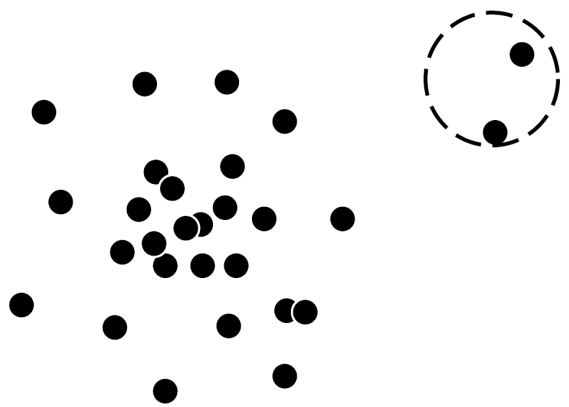
Outlier Analysis / Anomaly Detection

A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers.

The analysis of outlier data is referred to as outlier analysis or anomaly detection.

Applications:

- Credit card fraud detection
- Network intrusion detection
- Monitoring and surveillance in sensor networks
- Detecting changes in the global forest cover



Association Analysis

Given a set of records each of which contain some number of items from a given collection:

- Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, eggs, milk
2	Juice, bread
3	Juice, eggs, butter, milk
4	Juice, bread, butter, milk
5	Eggs, butter, milk

Rules discovered:
{Milk} $\Rightarrow$ {Eggs}
{butter,milk} $\Rightarrow$ {Juice}

Association Analysis Applications

Market-basket analysis:
Rules are used for sales promotion, shelf management, and inventory management.

Telecommunication alarm diagnosis:
Rules are used to find combination of alarms that occur together frequently in the same time period.

Medical Informatics:
Rules are used to find combination of patient symptoms and test results associated with certain diseases.

Are all patterns interesting?

A data mining system has the potential to generate thousands or even millions of patterns or rules.

Are all of the patterns interesting?

- Typically, the answer is no – only a small fraction of the patterns potentially generated would actually be of interest to a given user.

A pattern is interesting if it is:

1. Easily understood by humans
2. Valid on new or test data with some degree of certainty
3. Potentially useful
4. Novel

Are all patterns interesting?

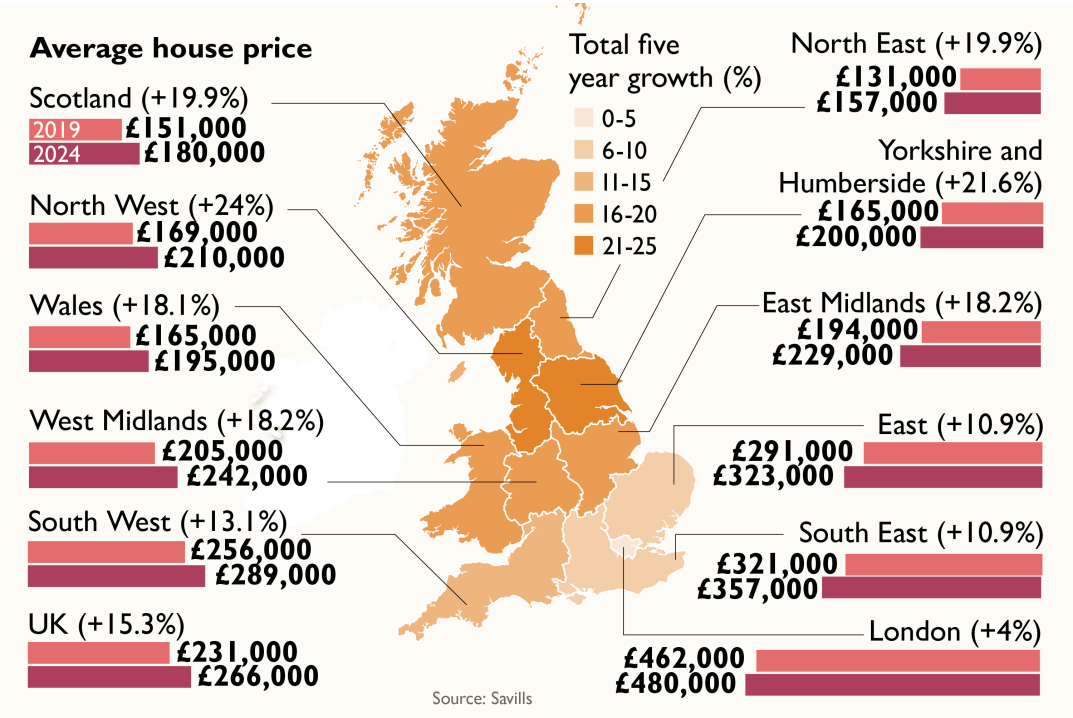
A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge.

Several objective measures of pattern interestingness exist, for example:

- Rule support, representing the percentage of records from a database that the given rule satisfies.
- Confidence, which assesses the degree of certainty of the detected association.

There are also subjective interestingness measures, based on user beliefs in the data.

House prices



Predicting the growth of house prices belongs to the following problem category:

(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 problem category:

(a) Classification

(b) Regression

(c) Clustering

Challenges in data mining

Challenges in data mining

Mining methodology

- Researchers have been vigorously developing new data mining methodologies.
- Current topics: investigation of new kinds of knowledge, mining in multidimensional space, integrating methods from other disciplines...
- Mining methodologies should consider issues such as data uncertainty, noise, and incompleteness.

User Interaction

- How to interact with a data mining system
- How to incorporate a user's background knowledge in mining
- How to visualize and comprehend data mining results

Challenges in data mining

Efficiency and Scalability

- Algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data.
- The wide distribution of data, and the computational complexity of some data mining methods motivate the development of parallel and distributed algorithms.

Diversity of Database Types

- Handling complex types of data
- Mining dynamic, networked, and global data repositories

Challenges in data mining

Data Mining and Society

- Social impacts of data mining: How can we use data mining technology to benefit society? How can we guard against its misuse?
- Privacy-preserving data mining
- Invisible data mining

Data Ethics

An emerging branch of applied ethics which describes the value judgments and approaches we make when generating, analysing and disseminating data.

Questions?
also please use the forum on QM+