MSc Data Mining

Topic 01 : Module Overview

Part 03 : Module Introduction

Dr Bernard Butler and Dr Kieran Murphy

Department of Computing and Mathematics, SETU Waterford.
(bernard.butler@setu.ie; kmurphy@wit.ie)

Spring Semester, 2023

Outline

- Introduction, definitions and context
- Roles, expertise and ethics
- Workflow and process models
- Overview of Machine Learning Algorithms
- Delivery and Assessment
- Resources

What is the AIM of the module?

Aim, as per Module Descriptor*...

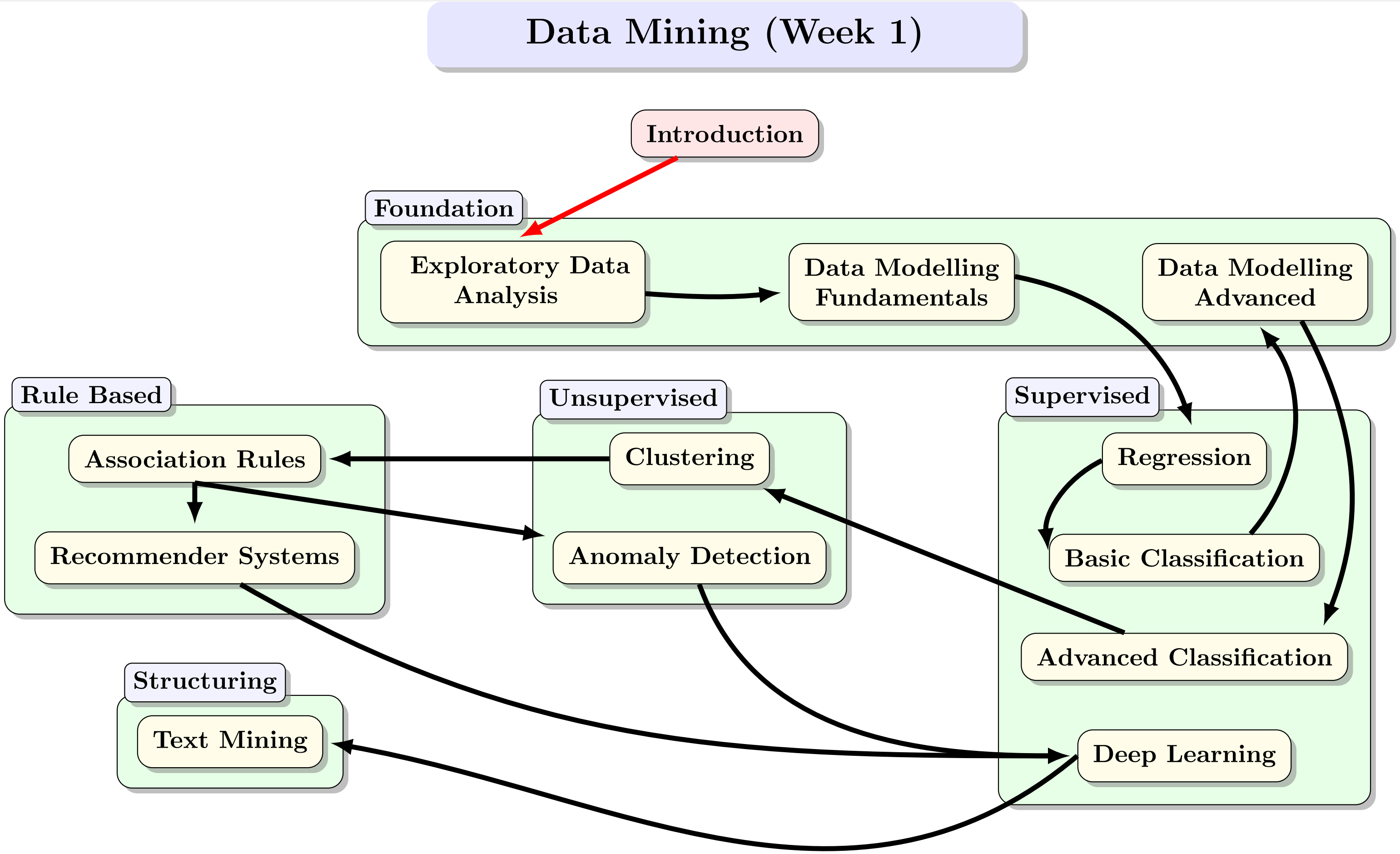
The student will be introduced to the fundamental concepts and techniques of Data Mining. The student will learn the data mining process and experience the steps involved; including data pre-processing, modelling, optimisation, result interpretation and validation...

Translation (Informal Aims)

1. Collect observations from a variety of processes, yielding large amounts of data.
2. Preprocess this data, selecting relevant features only.
3. Use data-intensive analysis techniques to obtain insights.
4. Postprocess analysis results, validate, visualise and refine the process.

*Also, see the module descriptor for the learning outcomes for a more formal description of this module.

What topics does it contain?

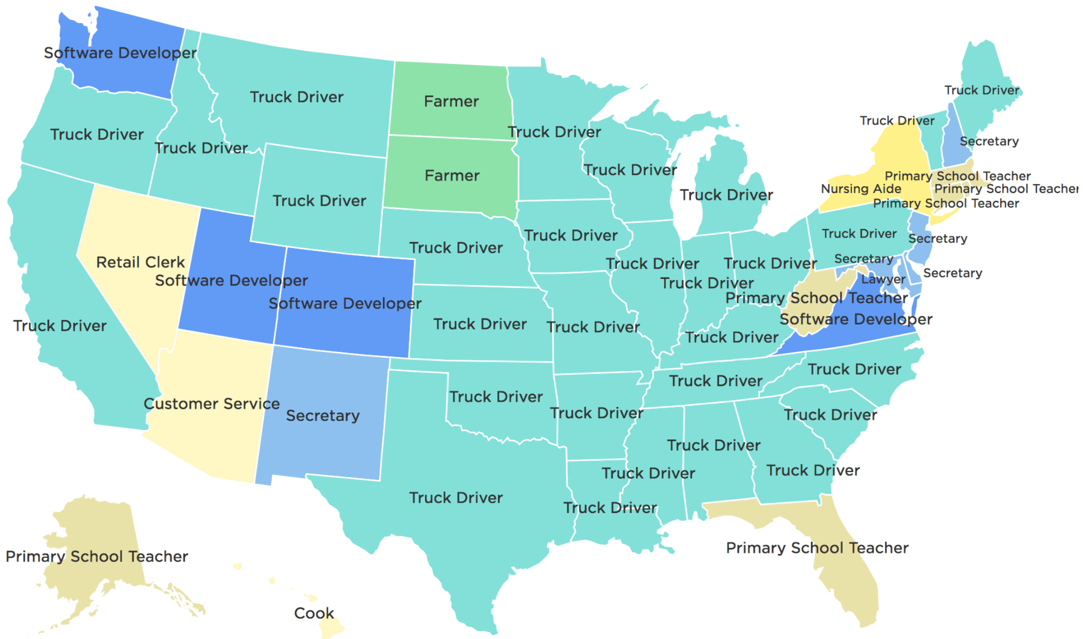


Why study Data Mining?

Machine Learning, Machine Learning, Machine Learning,
Artificial Intelligence, Artificial Intelligence, Artificial Intelligence,
Artificial Intelligence, Artificial Intelligence, Artificial Intelligence,
Artificial Intelligence, Deep Learning, Deep Learning, Deep Learning, Deep
Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Attractive
Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career,
Attractive Career, Attractive Career, Machine Learning, Machine Learning, Machine Learning, Machine Learning,
Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine
Learning, Machine Learning, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence,
Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial
Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning,
Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Attractive Career,
Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive Career, Attractive
Career, Attractive Career, Attractive Career, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Machine
Learning, Machine Learning, Machine Learning, Machine Learning, Machine Learning, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary, Big Salary,
...

Why are Data Mining (and automation) so important?

- Most common job by state in USA (2014)[†] …



- By 2035 autonomous end-to-end delivery can be achieved.

- Current situation:

> Any cognitive task that requires less than 2 seconds to perform can be automated in the short term.

See: Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines

[†] `https://www.npr.org/sections/money/2015/02/05/382664837/map-the-most-common-job-in-every-state`

Selected definitions

IoT: A network of pervasive connected objects able to collect and exchange data from embedded sensors, with the infrastructure and services to support them. (Various)

Big Data: High volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. (Gartner 2012)

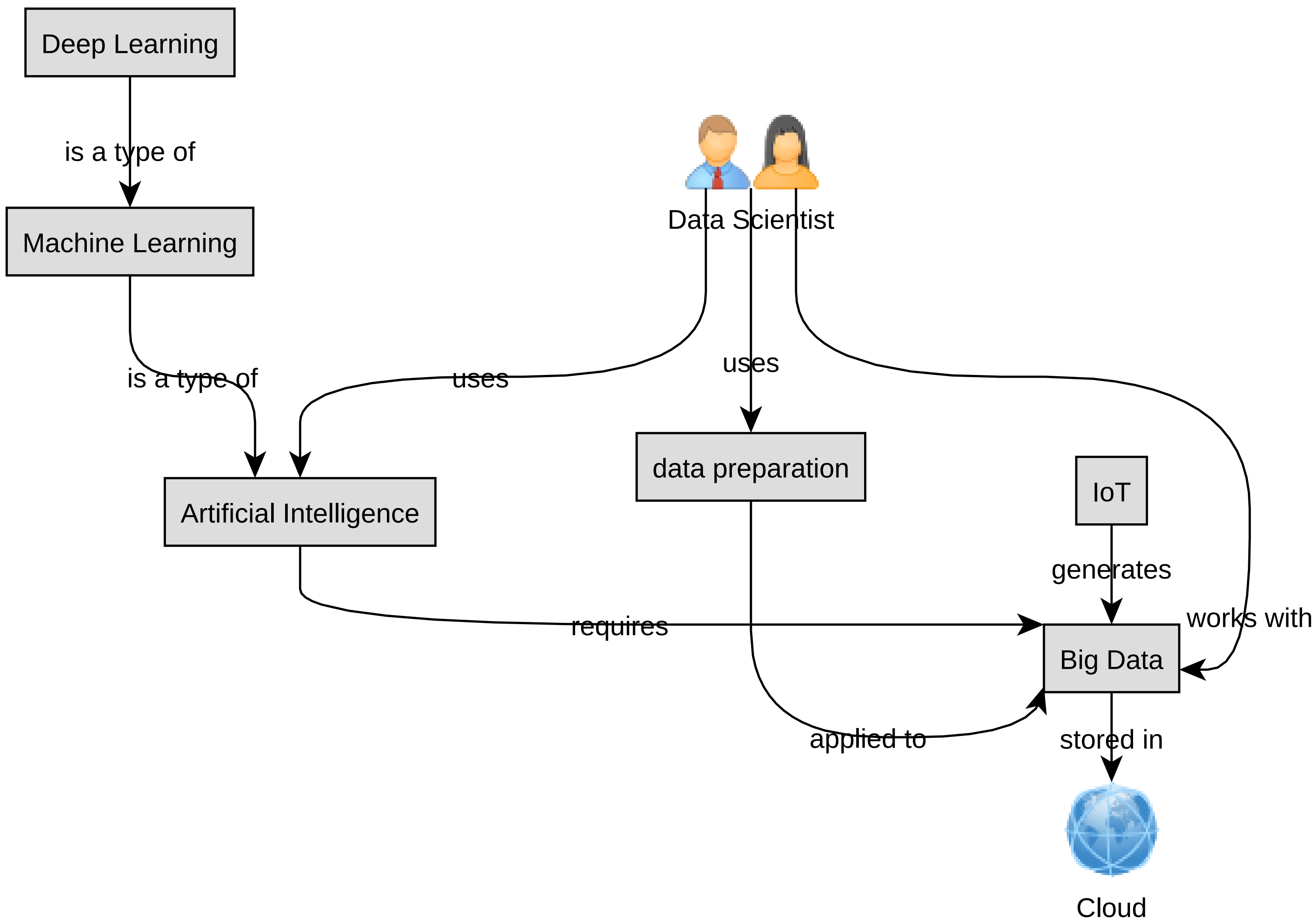
Data Scientist: can ask the right questions, {generate} and consume the results of analysis of Big Data effectively. (McKinsey 2011)

Artificial Intelligence: the capability of a machine to imitate intelligent human behavior (Webster 2017)

Machine Learning: Branch of computer science {and related fields} that gives computers the ability to learn without being explicitly programmed. (Samuel 1959)

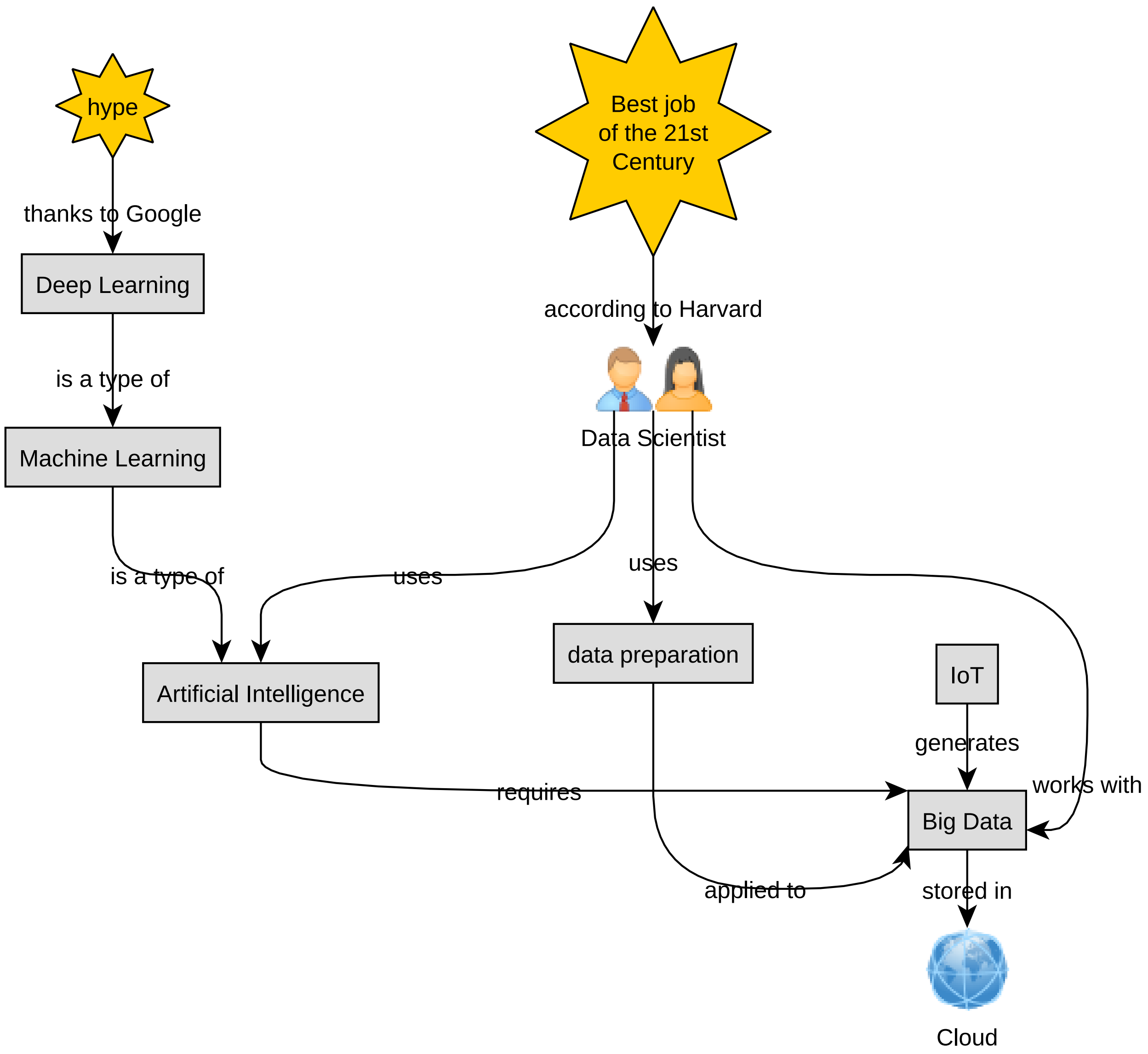
Deep Learning: Use of very large neural networks with many layers of "neurons" that can be trained to generate robust models of their input, whose classification performance scales with the amount of data supplied. (Various)

Relationships between terms



Note that IoT is included because it is a common source of Big Data and hence of scenarios and data for data mining.

Annotated Relationships between terms!



What is data mining and how does it relate to similar terms?

Operational Definitions

- deriving knowledge from large and/or complex datasets, with *guidance* from the data scientist
- “Data mining is the study of efficiently finding structures and patterns in large data sets. It draws from and influences the disciples of programming, mathematics/statistics, database management and machine learning.”

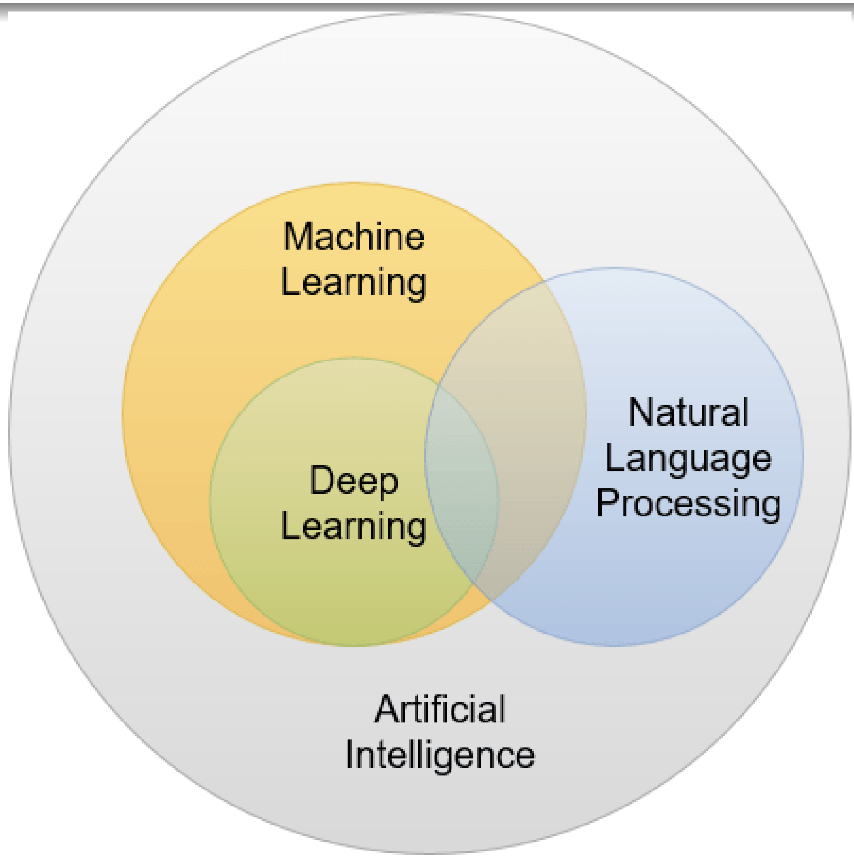
Primary goals

- From messy and noisy raw data, deriving structure and context
- Applying scalable learning algorithms to these higher value data sets

Secondary goals

- Modelling and understanding the error and other consequences of the modelling process.
- Building data-driven processes, architectures & frameworks: *Big Data*

AI vs ML vs DL



- Note the AI $>$ ML $>$ DL hierarchy
- Natural Language Processing (cf., ChatGPT) is very active at the moment, and has some elements that do not fit in ML or DL.

Interlude: Examples of Big Data

Exercise

please consider (real world) processes generating *Big Data*.
Can you come up with 3 examples in 3 minutes?

Prehistory, or before 2007…

Data Generation

- Transactions (bank, retail)
- Activity, e.g., texts
- Basic e-commerce

Data Processing

- Databases, SQL, stored procedures
- Consultants, system integrators
- Proprietary statistical software

Data Analysis

- Reporting: looking back
- Descriptive statistics
- Simple plots

The first (batch) wave: 2007–2011

Data Generation

- As before...
- Web activity: comments, etc.
- 360degree view

Data Processing

- As before...
- NoSQL
- hadoop ecosystem (batch analytics)

Data Analysis

- As before...
- Personalisation and recommendation
- Predictive Analytics

The second (streaming) wave: 2012–2015

Data Generation

- As before...
- Social Media!
- IoT (early adopters)

Data Processing

- As before...
- Apache Spark
- R vs. python

Data Analysis

- As before...
- Data understanding
- Weak AI: assistants, etc.

The machine wave: 2016–2020

Data Generation

- As before...
- Machine-generated (e.g., fake news)
- IoT (mainstream)

Data Processing

- As before...
- Microservices: move function to data
- Decoupled databases with schema-on-read

Data Analysis

- As before...
- Deep learning inflection point
- Visualisation

The connected wave: 2021–date

Data Generation

- As before…
- Aggregation services
- Cyberphysical Systems

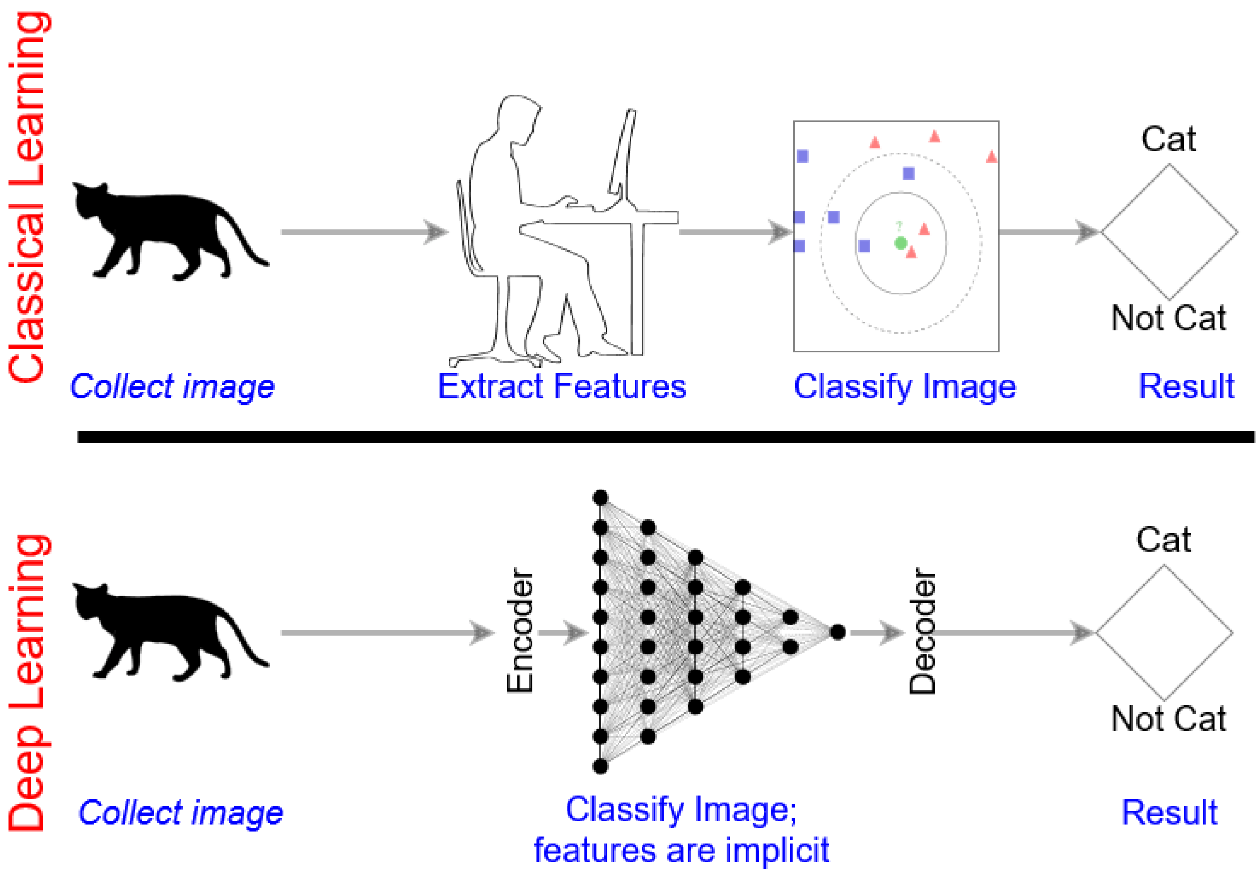
Data Processing

- As before…
- Serverless computing
- Multi-model databases (mix and match, with a common data store)

Data Analysis

- As before…
- Deep learning is dominant
- Interpretability vs fragility

Classical versus Deep Learning - overview



Classical learning requires extensive setup before training.

Classical versus Deep Learning - pros and cons

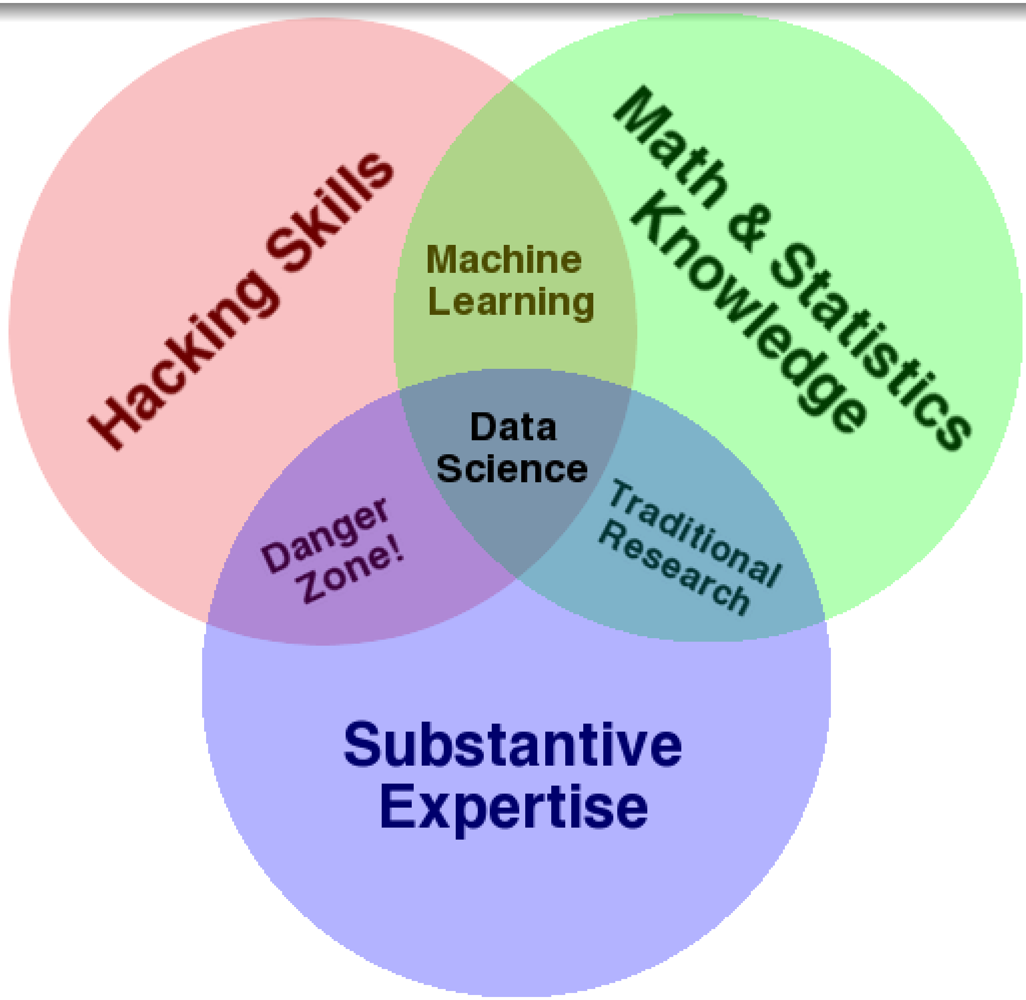
Classical Learning	Deep Learning
Can work well with less data (<10^6 rows, say)	More training data gives more accuracy
Easier to interpret/explain models	Model is opaque and can be fragile
Training is relatively fast	Training can require many epochs
Training requires fewer resources	Training requires massive resources
Accuracy improvement falls off	Accuracy can improve with more training data
Requires feature engineering (by human)	Features are encoded implictly in layers
Complex prediction requires complex model	With enough nodes can represent any function

In conditions where one type of learning is weak, the other is often strong.

Emphasis of this module

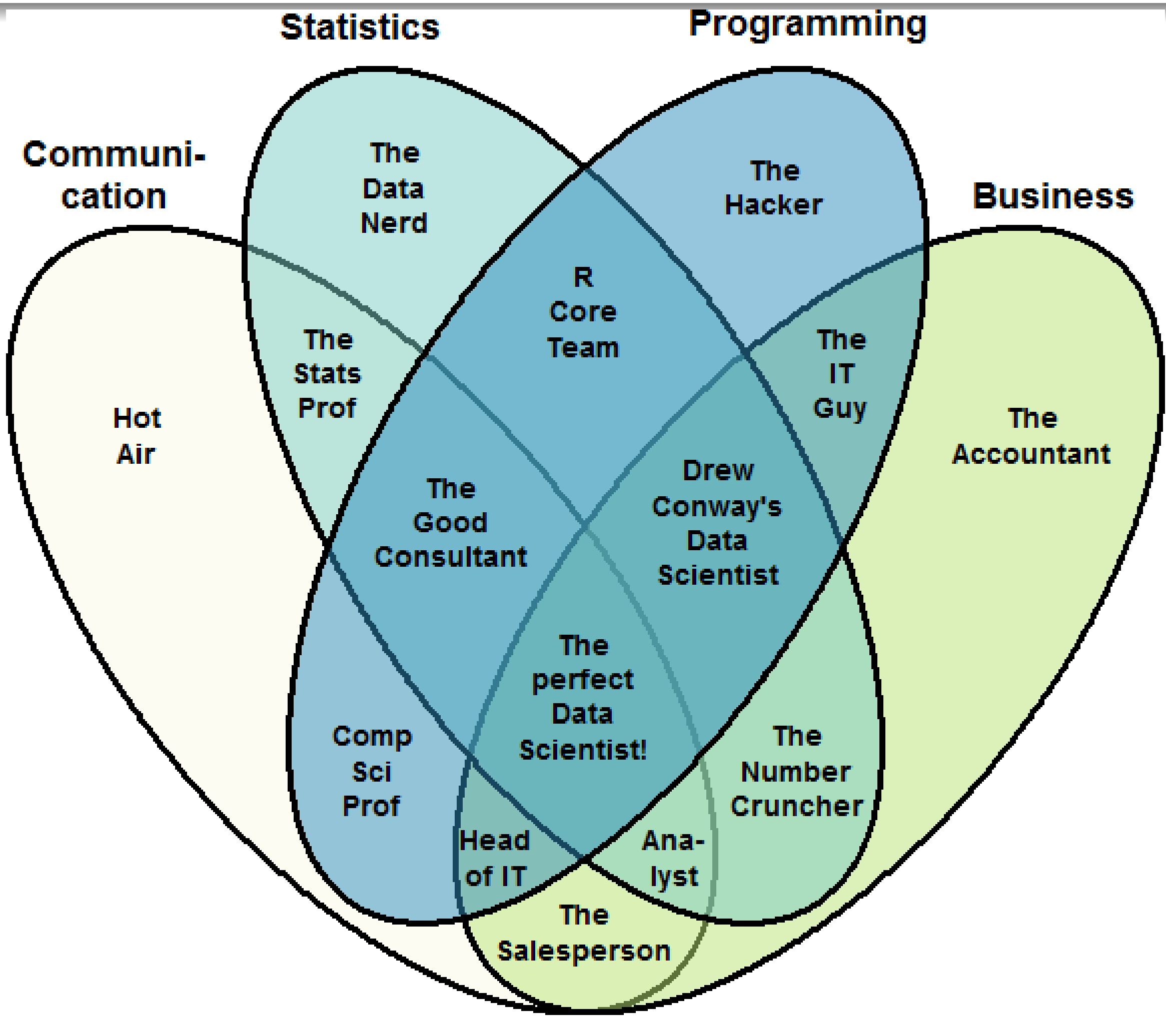
- This module covers foundations, classical models, some deep learning
- Foundations include EDA, notions of error and variance, training vs test data, ...
- Classical modelling includes feature selection and classical approaches to learning from data
- Neural network and Deep learning go straight to predicting based on (encoded) data
- Foundations are shared by both classical and deep learning
- Deep learning is ideal for learning from **labeled, web-scale** big data.
- Day-to-day, classical machine learning is more suitable.

Drew Conway's 3-set Venn Diagram of Data Science Expertise



Source: `http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram`

Stephan Kolassa’s 4-set Venn Diagram of Data Science Expertise

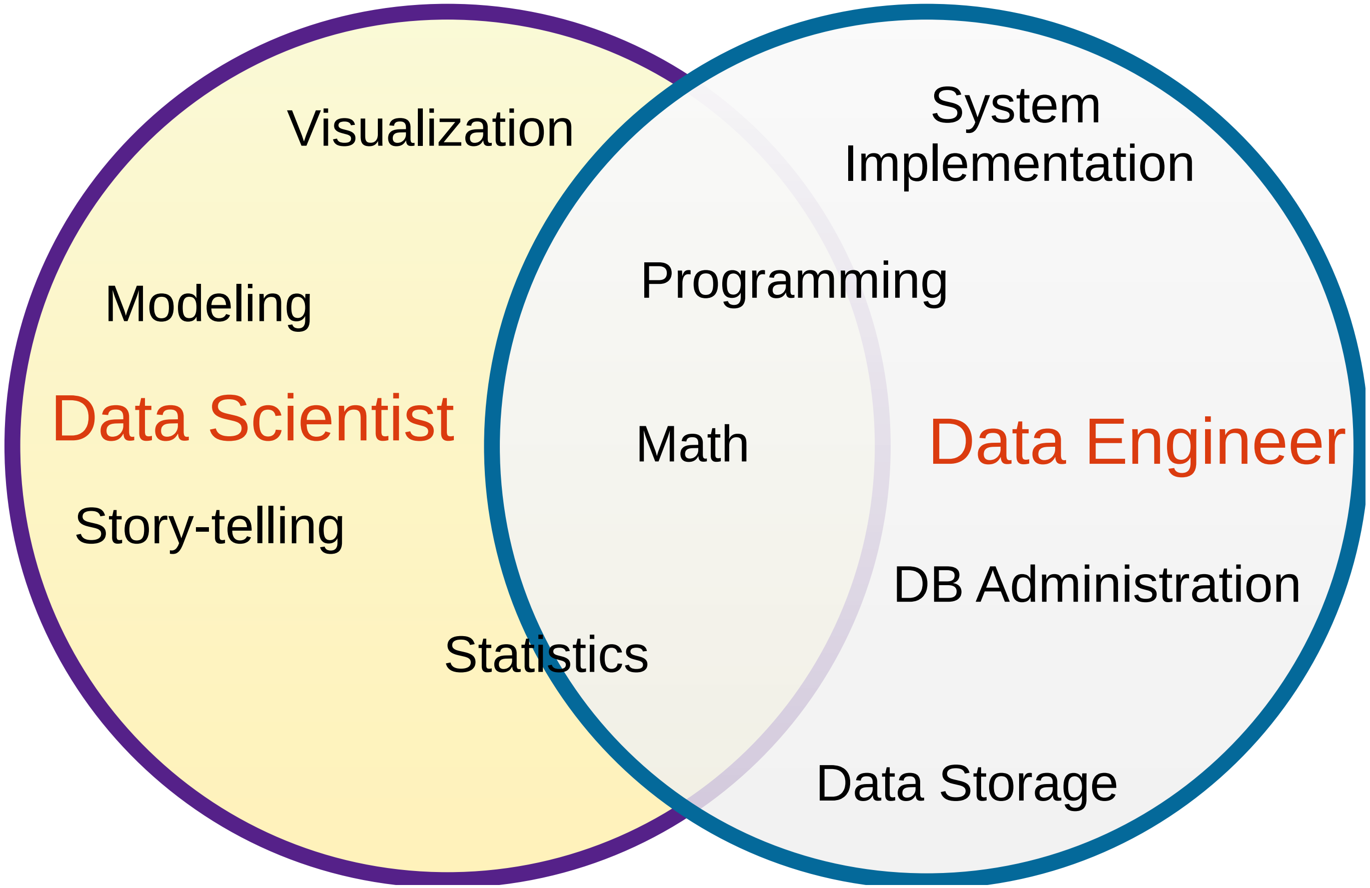


Source: `https://datascience.stackexchange.com/a/2406`

Gartner suggests the need for a *Citizen Data Scientist*

Source: `http://www.kdnuggets.com/2016/03/cartoon-citizen-data-scientist.html`

Data Scientist vs Data Engineer



Source: `https://ryanswanstrom.com/2014/07/08/data-scientist-vs-data-engineer/`

Also the traditional roles of *Data Analyst* and *Software Engineer*...

Complete the following disadvantages of IoT and Big Data

m___ s__v__l_____

i___t__y _h_f_

d___c_ b__n___

d____l _f __r__c_

b__s

l___ o_ t___s__r___y

And those disadvantages are. . .

mass surveillance

identity theft

device botnets

denial of service

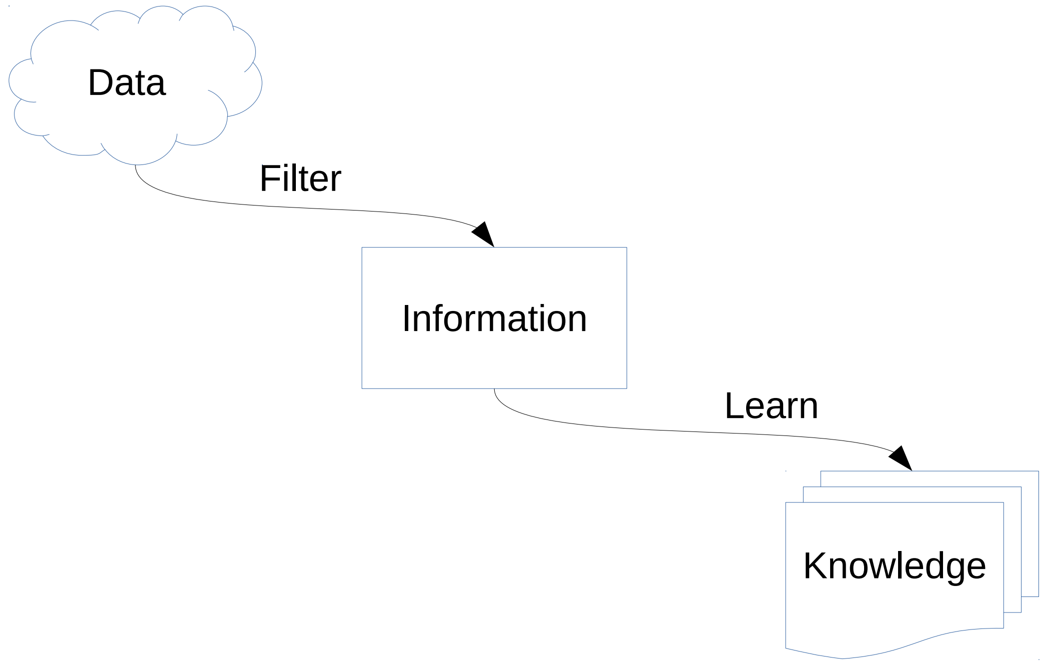
bias

lack of transparency

Ethical Concerns

- protecting privacy (informed consent; undoing pseudonymisation)
- ensuring transparency of decisions (how was the decision made?)
- breaking cycles of bias (biased data leads to biased results)
- enabling validation (ensuring correct usage of techniques)
- enabling decisions to be challenged (openness and due process)

The Data to Knowledge Pipeline



Data Filtering

- Clean (drop unwanted observations)
- Summarise (remove observation detail)
- Reduce (remove/transform variables)

Learning

- Derive models
- Validate models
- Analyse discordance

Data - Information - Knowledge - Wisdom

Example of the DIKW chain

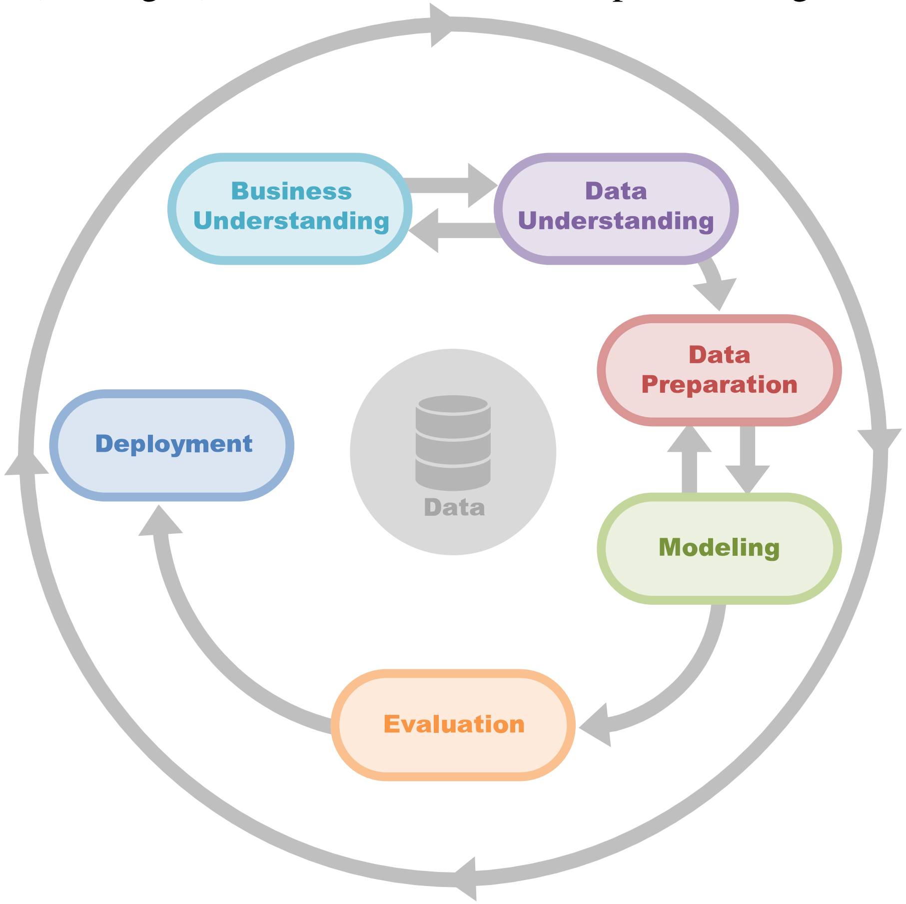
- Servers and applications log events in files and/or databases [DATA]
- *Collector* agents select specific events, in context [INFORMATION]
- Machine learning *classifiers* learn system behaviour and identify anomalies [KNOWLEDGE]
- Humans and software use this knowledge to prevent future problems [WISDOM]

Note that the DIKW chain is often represented as a pyramid.

CRoss Industry Standard Process (for) Data Mining

CRISP-DM

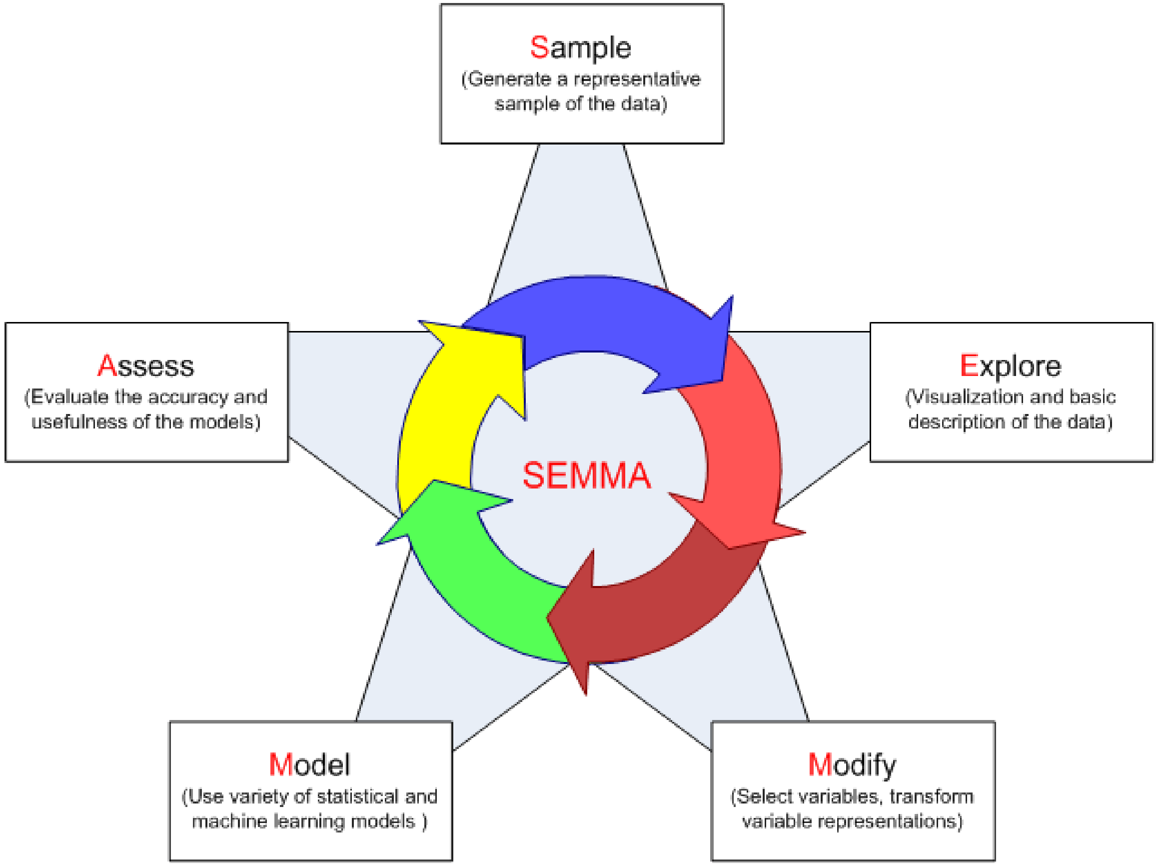
CRISP-DM is a high-level iterative process model. It gives much weight to data understanding and preprocessing and involves the data and problem owners from the start (c.f., Agile). It is the best known but predates Big Data, etc.



Sample, Explore, Model, Modify, Assess

SEMMA

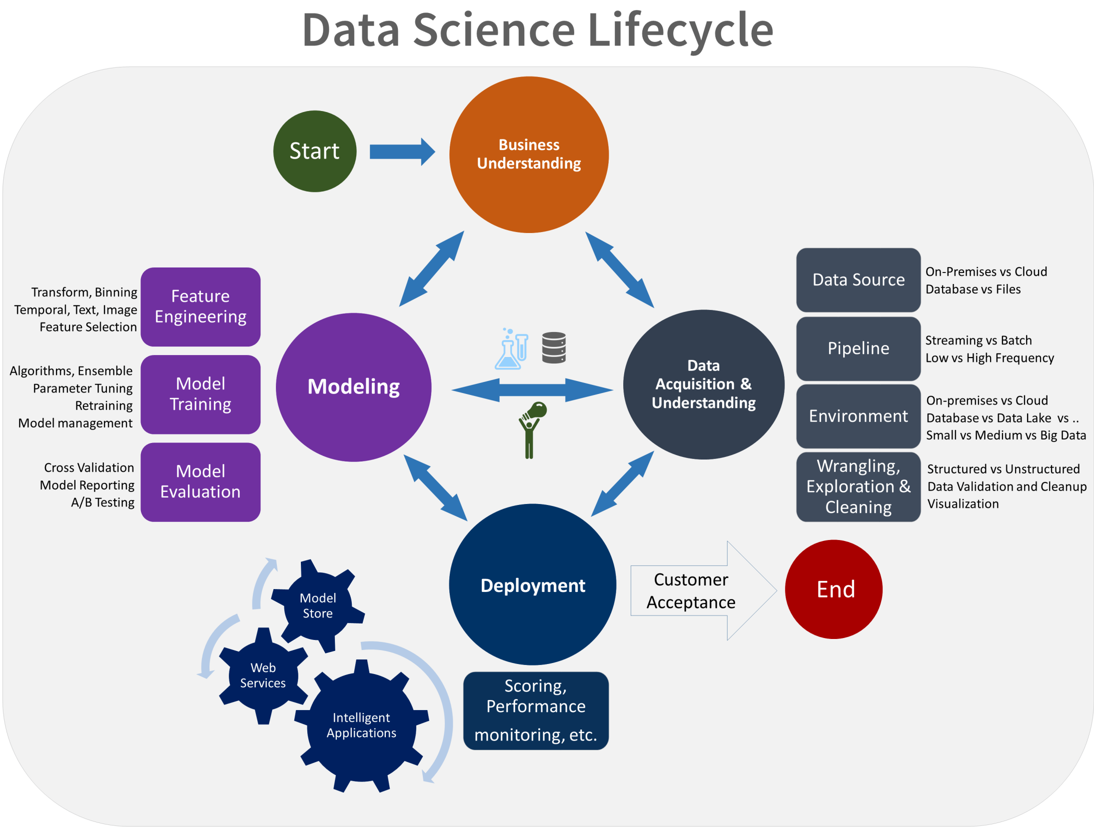
SEMMA is promoted by SAS and takes a more operational view of data mining, using a (statistical) *model-building* metaphor. Business input is essential but largely implicit. It is more concrete than CRISP-DM so it tends to map well to DM tool workflows.



Microsoft Team Data Science Process

TDSP

TDSP is the most detailed process model of the 3. It is much more recent. It is cloud-aware and directly references Azure and other Microsoft technologies. Typically there are two main cycles, one involving the Business, the other involving Deployment. Interestingly, there is a Start and End, so it is more project-focused.



The “5 Tribes of Machine Learning”

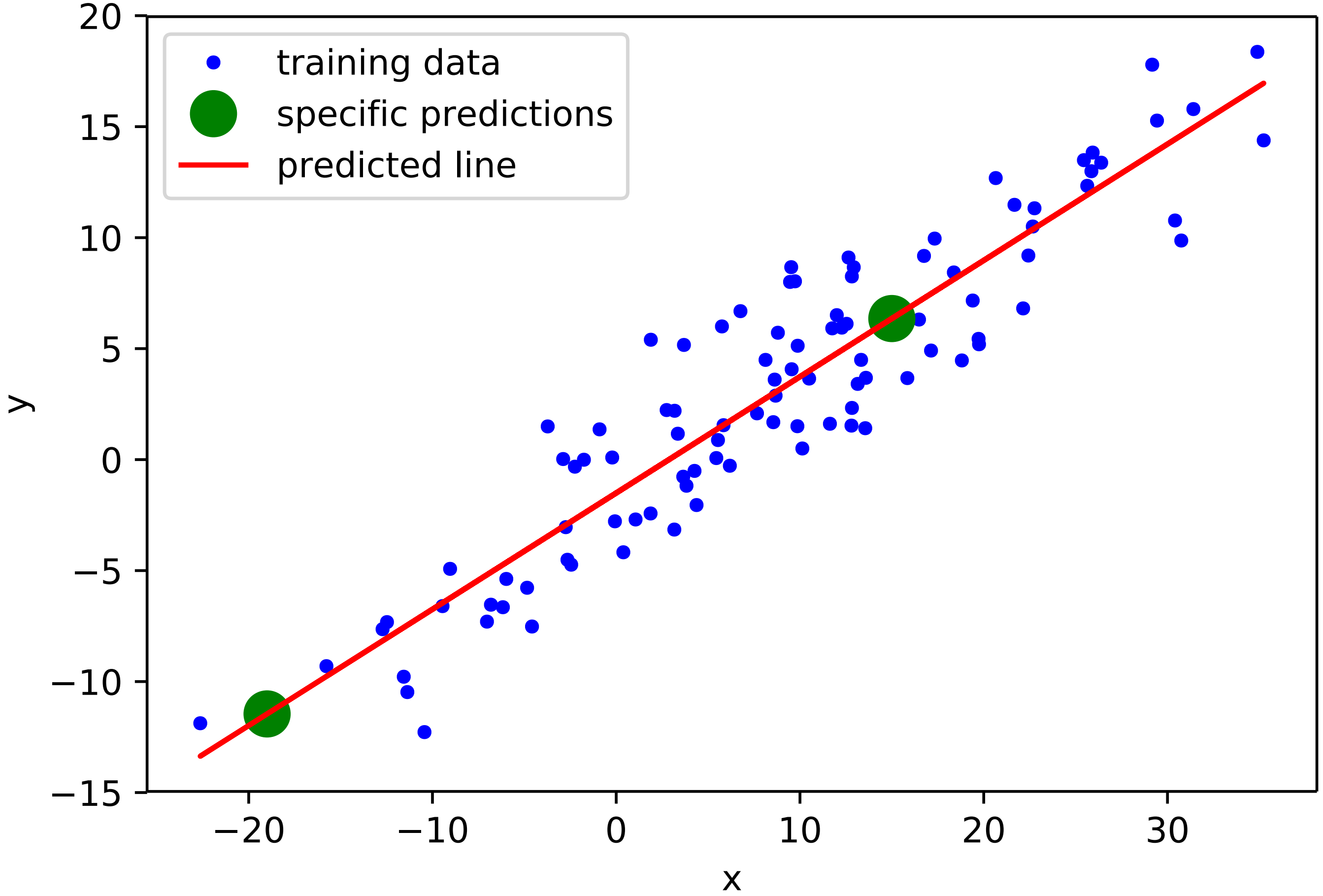
Tribe	Origins	Learning Algorithm
Symbolists	Logic, Philosophy	Inverse Deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Mathematical Biology	Genetic Programming
Bayesians	Statistics	Probabilistic Inference
Analogizers	Psychology	Kernel Machines

Summarised from Domingos (2015) “The Master Algorithm”

Regression

Definition

Given data comprising a set of independent variables (of any type $\boldsymbol{x}$) with a set of dependent variables (numeric only $\boldsymbol{y}$), find the relationship $\boldsymbol{y} = f(\boldsymbol{x})$ having the maximum likelihood given the available observations $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$.



Classification

Definition

Given data comprising a set of independent variables (of any type $\boldsymbol{x}$) with a set of dependent variables (categorical $\boldsymbol{y}$ (labels)), find the relationship $\boldsymbol{y} = f(\boldsymbol{x})$ having the maximum likelihood given the available observations $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$.

There are many ways of representing f: a classification tree is shown here.

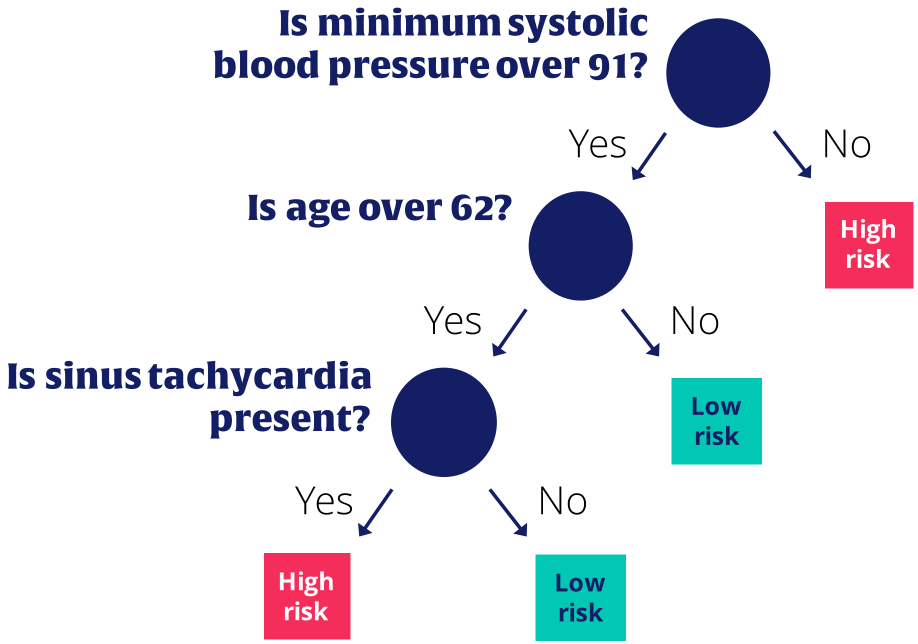
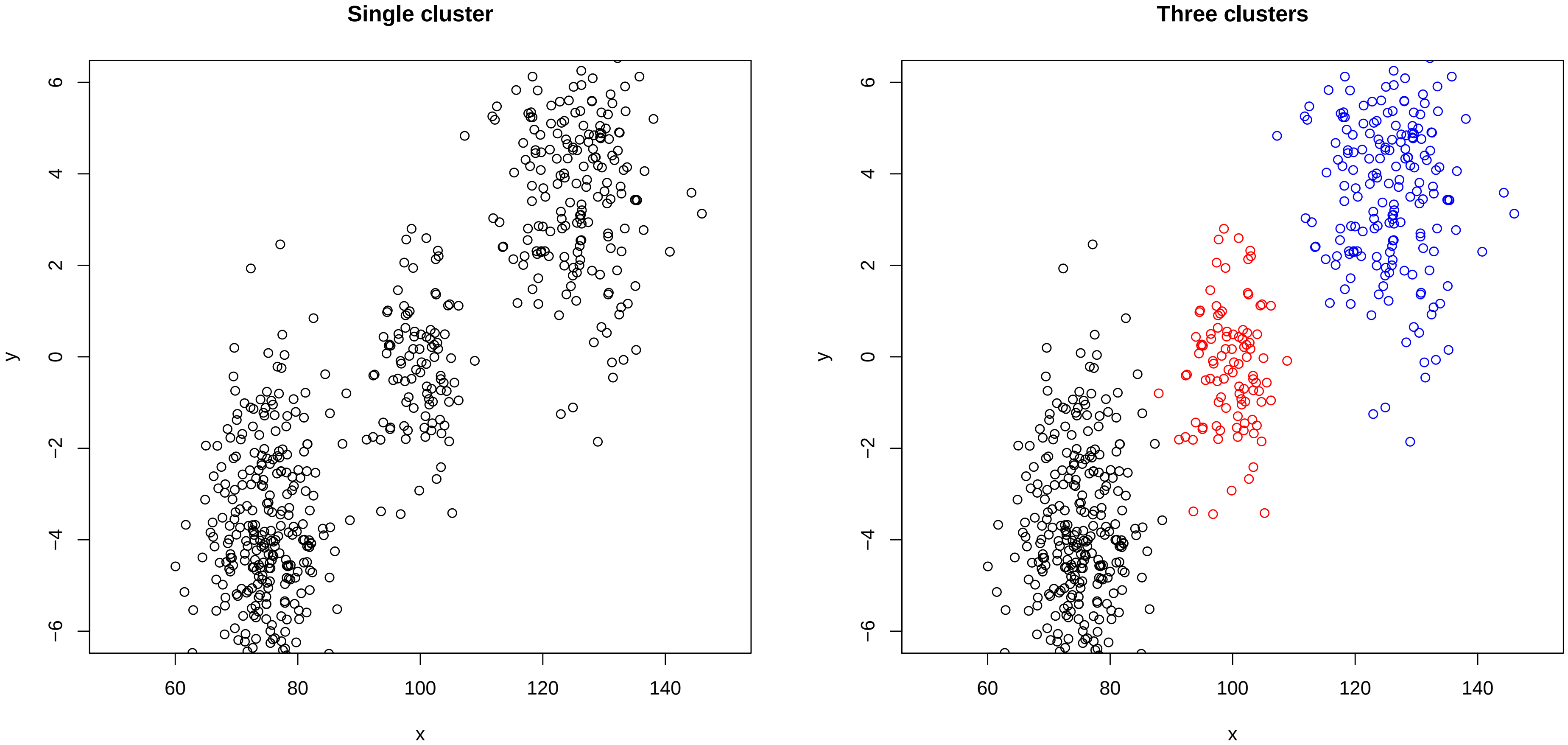


Diagram taken from the book Classification and Regression Trees, by Breiman L., Friedman J., Stone C. and Olshen R. 1984.

Clustering

Definition

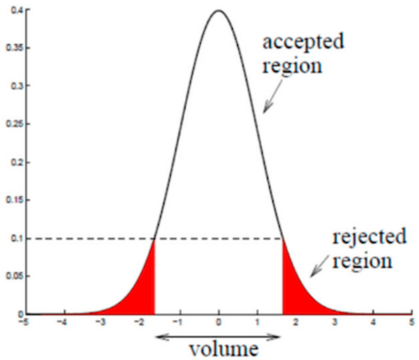
Clustering is the process of grouping data into classes or clusters, so that objects within a cluster have high similarity with each other but are dissimilar to objects in other clusters. Different similarity measures and/or algorithms result in different cluster arrangements.



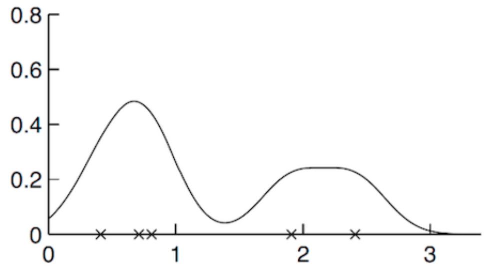
Anomaly Detection

Definition

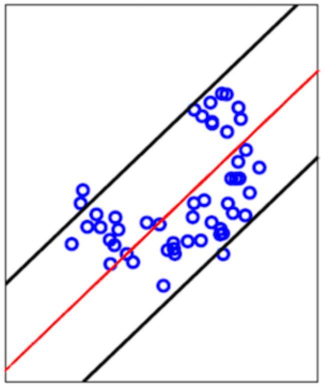
Anomaly detection identifies data points, events, and/or observations that depart from a dataset's normal behavior. Anomalous data can indicate problems, such as fraud, or opportunities, like a surge in demand for a product.



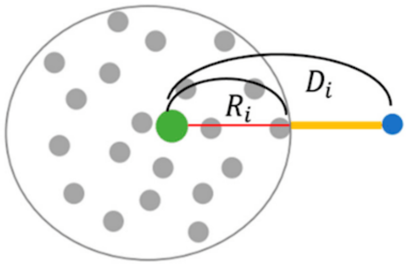
(a)



(b)



(c)



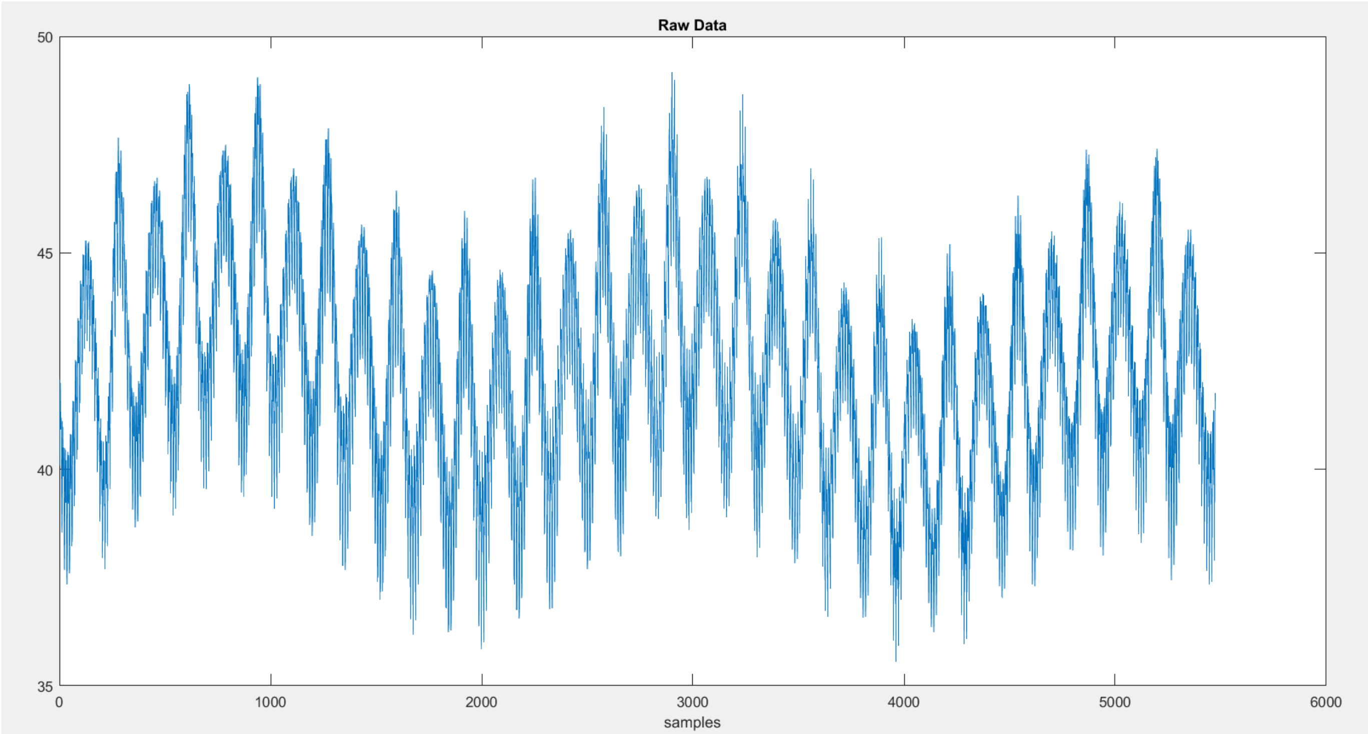
(d)

Source: Appl. Sci. 2019, 9, 4018; doi:10.3390/app9194018

Time Series Analysis

Definition

Time series data is a sequence of observations on the values that a variable taken at regularly-spaced time intervals. This data is sequentially correlated and techniques are needed to determine seasonality, trends, anomalies etc.



Source: https://stats.stackexchange.com/q/458491

Association Rules Mining

Definition

Frequent itemset mining looks for associations and correlations among items in large data sets. Associations are expressed as rules and quantified in terms of their *support* and *confidence*. The classical example is market basket analysis and the famous rule about buying diapers and beer together. See example transaction data below

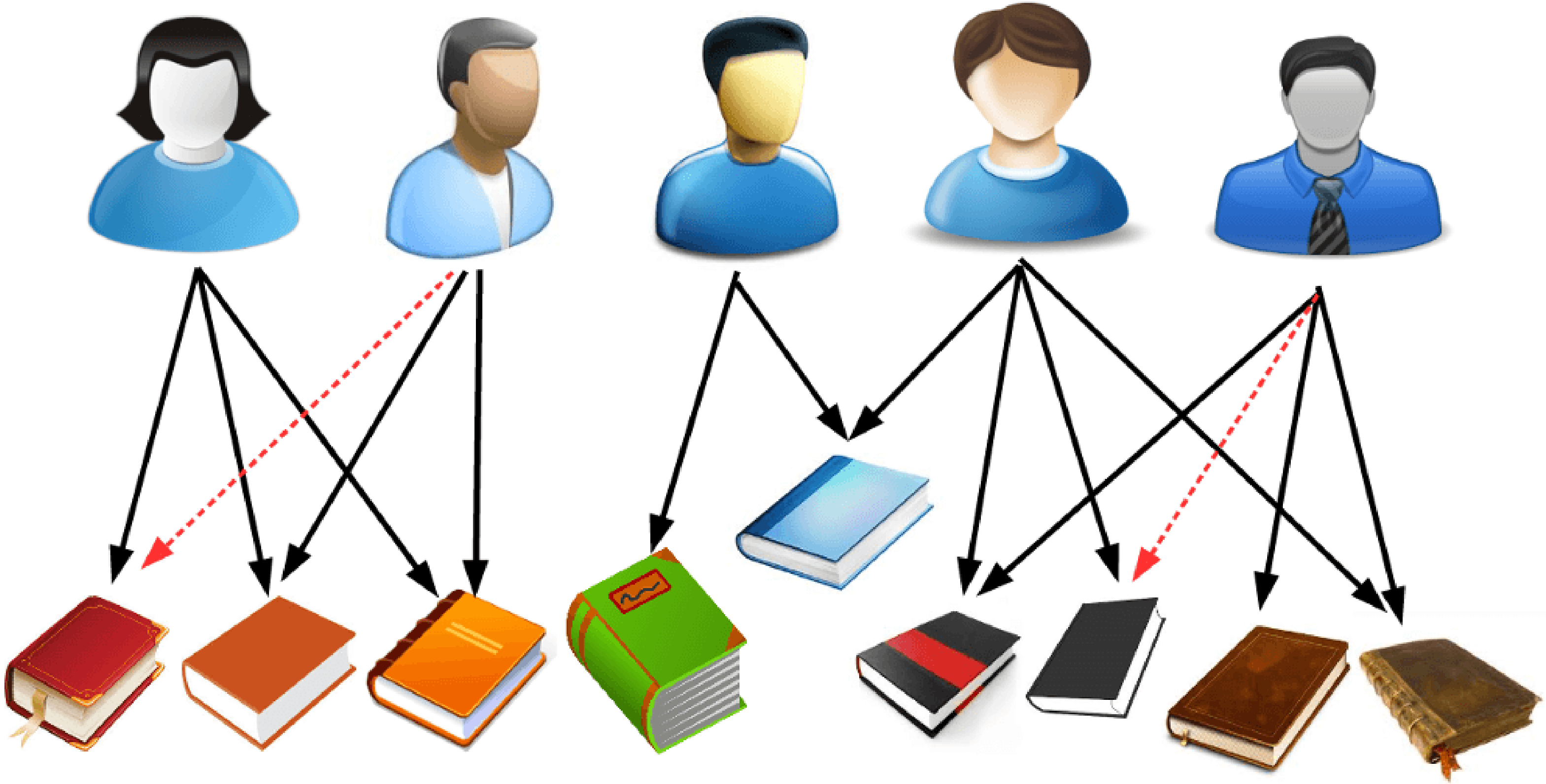
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Recommender Systems

Definition

Collaborative filtering is an extension of item-based association rules mining to consider relationships between users and items. Generally, if two users have similar behaviour and/or preferences, items favoured by one user will also be favoured by the other. This can be used to recommend “new” items to users. Used very commonly on ecommerce websites for cross-selling.



How? — Delivery

Contact hours

- One 2-hour lecture per week.

  —Friday 14:15 (GMT/IST; online over zoom)

  - Presented by Bernard and Kieran, generally on separate weeks.
  - We cover concepts, definitions, examples, etc.
  - Lecture objective is to improve understanding of the topic.
  - Lab is used to develop practical experience of an integrated set of topics
  - Feel free to stop us and ask questions (raise your hand and/or put message in chat) at any point.
- One 2-hour online support session per week on Monday afternoons from 15:15 onwards.
  - Labs use python workbooks to define and implement data mining *workflows*.

(Hardware) Requirements

- Use of own (moderately powerful: multi-core CPU, min 8GB RAM) laptop is recommended!

How? — Assessment Structure

100% Continuous Assessment

1. Issued Week 2, submitted end Week 4 (Data Mining Proposal, 20%)
2. Issued Week 4, submitted end Week 14 (Private Kaggle Competition, 20%)
3. Issued Week 5, submitted end Week 9 (Initial Data Investigation, 30%)
4. Issued Week 10, submitted end Week 14 (End-to-end Data Investigation, 30%)

These are indicative and might change (in terms of when assignments are issued and their relative weightings)...

Resources

moodle

- URL: Moodle: Data Mining-91634-[2022-2023]
- Used for all notices, assignment briefs and practical work submissions.

- URL: Data Mining 2022-2023 pages on github.io
- Used for content delivery (lecture notes and labs).

Software

All software used during this module is open source or freely available for non-commercial use (full details given in notes). Primarily

- Anaconda (**Python 3.9, 64 bit**) www.anaconda.com
- scikit-learn scikit-learn.org
- pandas pandas.pydata.org

Further Reading

Please note that the notes and labs we provide should be sufficient to pass to pass this module, so the books below are intended as *further reading*, not *recommended reading*.

Data Mining, Concepts and Techniques
by *Jiawei Han, Michelline Kamber and Jian Pei*
Broad selection of topics, looks at the entire data mining process including how to collect and preprocess data, discusses selected algorithms in depth.

Mining of Massive Data Sets
by *Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman*
Good mix of mathematical rigour and a treatment of the *Big Data* aspects for data mining at scale.

Python Data Science Handbook
by *Jake vanderPlas*, is both a textbook and a set of freely-available Jupyter notebooks that go into more detail on implementing some of the material in this module.

Neural Networks and Deep Learning
by *Michael Nielsen*, is a free online textbook that demystifies deep learning and is the basis of our treatment of deep clearing later in this module.