(MSc) Data Mining

Topic 01 : Module Overview

Part 03 : Module Introduction

Dr Bernard Butler and Dr Kieran Murphy

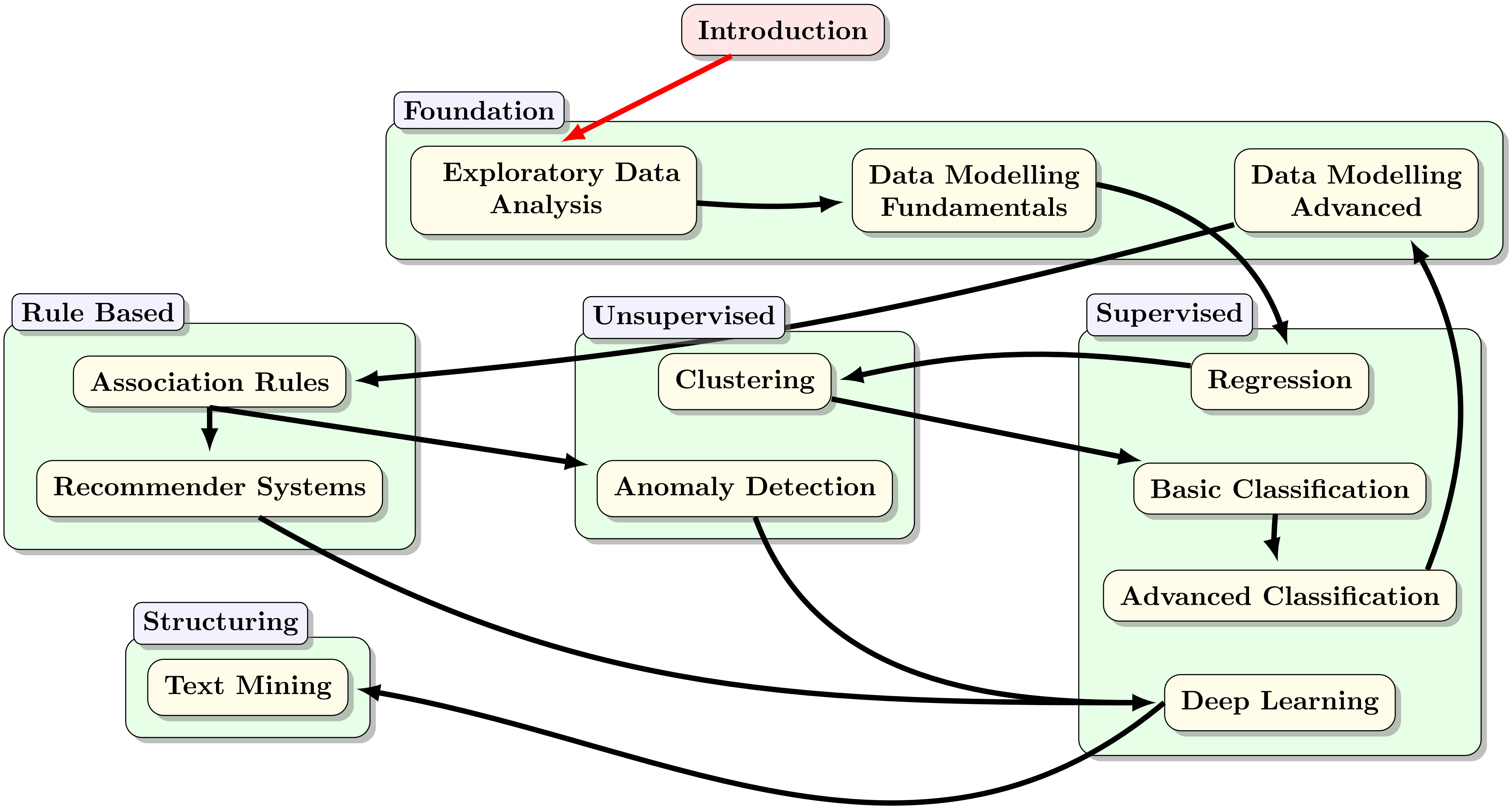
Department of Computing and Mathematics, SETU Waterford.
(bernard.butler@setu.ie; kmurphy@wit.ie)

Spring Semester, 2025

Outline

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

Data Mining (Week 1)



Introduction

Foundation

Exploratory Data Analysis

Data Modelling Fundamentals

Data Modelling Advanced

Rule Based

Association Rules

Recommender Systems

Unsupervised

Clustering

Anomaly Detection

Supervised

Regression

Basic Classification

Advanced Classification

Deep Learning

Structuring

Text Mining

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 alternate 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 practical 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 1, 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%)

What is the AIM of the module?

Aim, as per Module Descriptor[*]...

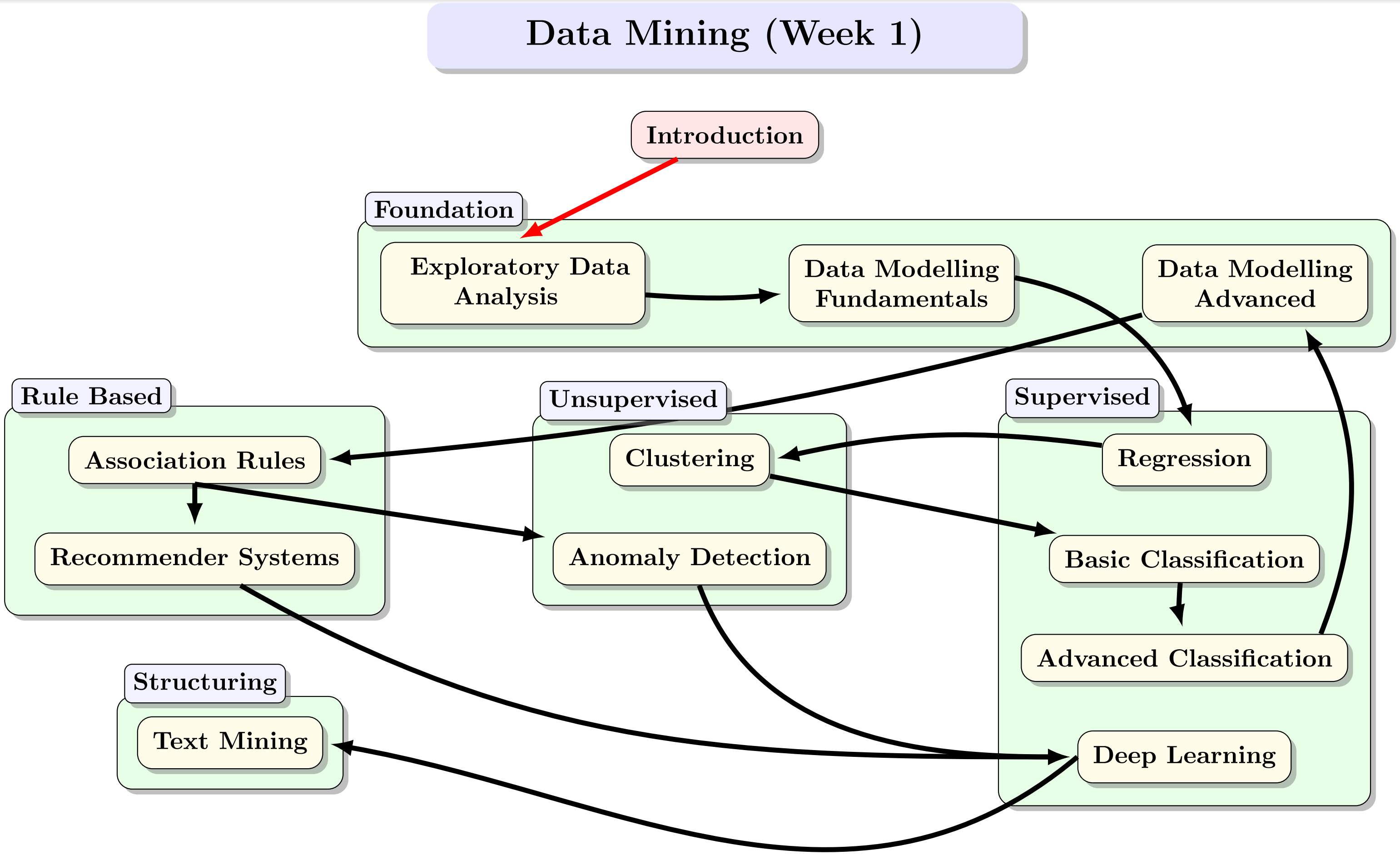
The student will be introduced to the fundamental concepts and techniques of Data Mining. The student will learn the data mining process and experience the steps involved; including data pre-processing, modelling, optimisation, result interpretation and validation...

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?

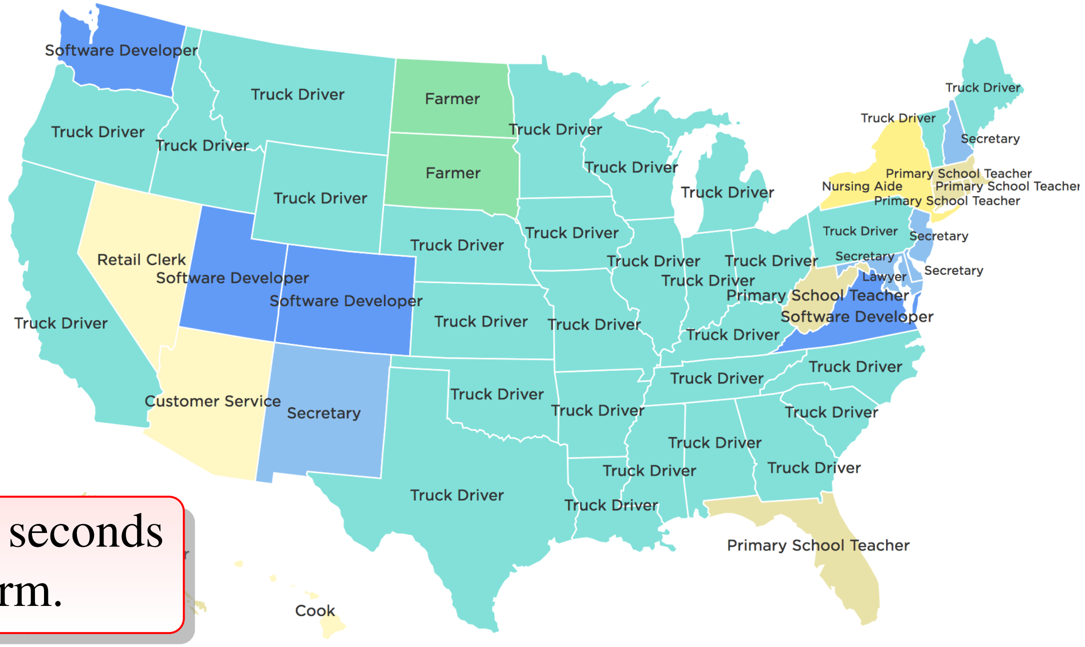
Machine Learning, Machine Learning, Machine Learning, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Artificial Intelligence, Deep Learning, Deep Learning, Deep Learning, Deep Learning, Deep Learning, 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 (10 years from now) 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

Indeed, with generative AI, you could argue that it is happening already, for text, audio, image and video production. . .

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

Selected definitions

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

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)

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

Predictive AI: AI that aims to forecast outcomes (typically, numbers and labels), given training data. After learning, it can predict *outcomes* for new data (Restack 2025).

Discriminative AI: AI that predicts outcomes that are categorical, either labels or boundaries between classes (Restack 2025).

Generative AI: AI that creates *instances of* new data that resemble or extend the training data (Restack 2025).

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.”

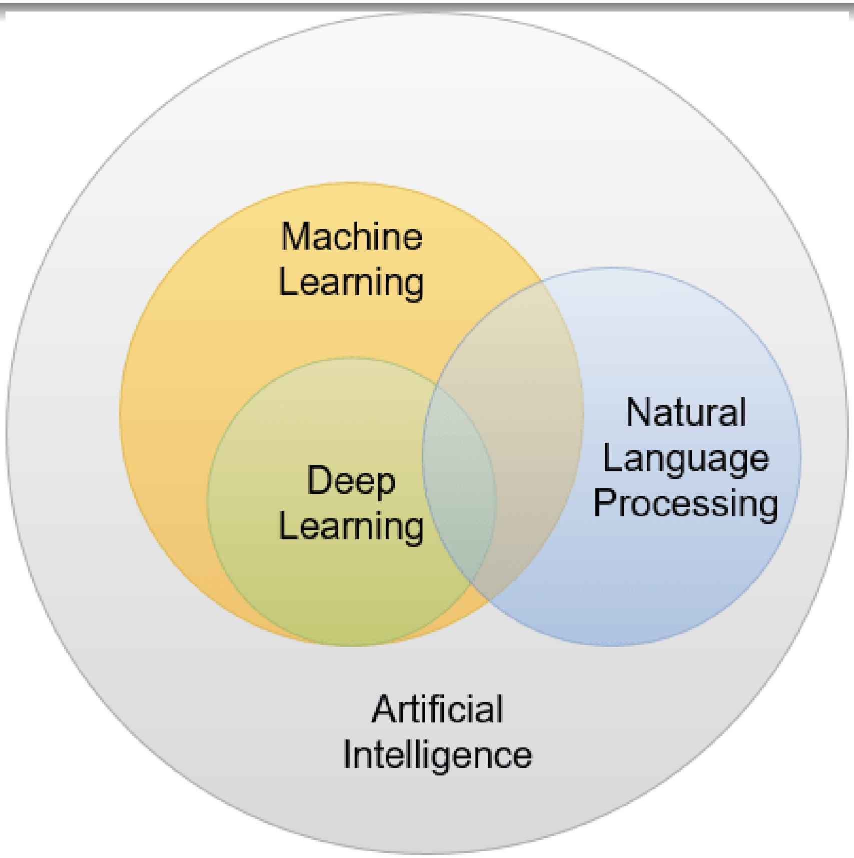
Primary goals

- 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 goes beyond ML and DL.

Interlude: Examples of Data Mining

Exercise

Please consider (real world) challenges where *Data Mining* might help..
Can you come up with 3 examples in 3 minutes?

Analytics and Classical ML

Data Generation

- Financial transactions (bank, retail, ...)
- Social engagement (texts, comments, likes, shares, ...)
- Machine-to-machine (IoT, cyberphysical systems, ...)
- Content curation and recommendation

Data Processing

- From Relational to NoSQL, to Multi-model databases
- From stored procedures to microservices and serverless computing
- From in-database to Apache Hadoop and Spark
- From proprietary statistical software to R and to python
- From consultants, to data scientists to machine learning engineers

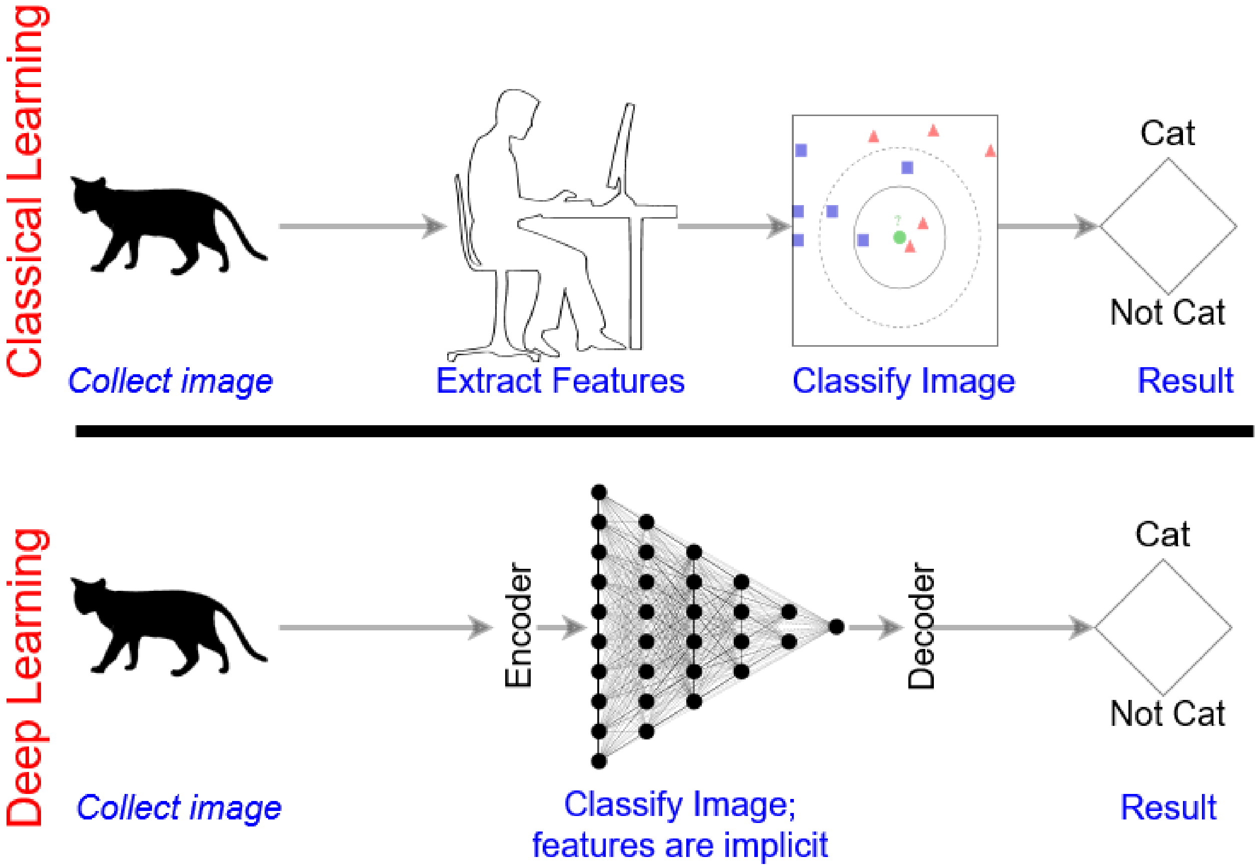
Data Analysis

- Reporting with descriptive statistics and simple plots, to classical ML with predictive analytics
- From predictive accuracy, to robustness vs fragility, to data understanding and visualisation
- Traditional chatbots are limited to specific contexts

The generative wave: 2022–date

- Traditional ML (big data processing for predicting labels or numbers) becomes mainstream
  - ML skills become expected
  - ML is operationalised by incorporating it in other practices (ML + Devops = MLops)
- *Generative AI* becomes practical
  - Generative Adversarial Networks (use generated models when training the desired model)
  - Large Language Models for generating textual responses to prompts, e.g., ChatGPT
  - Generate images, e.g., DALL-E and video, e.g., Sora given textual descriptions.
  - Rise of the Prompt Engineer
- AI is beginning to deliver on the promise identified by Alan Turing and others in the 1950s.

Classical versus Deep Learning - overview



Classical learning requires extensive setup before training.

Classical versus Deep Learning - pros and cons

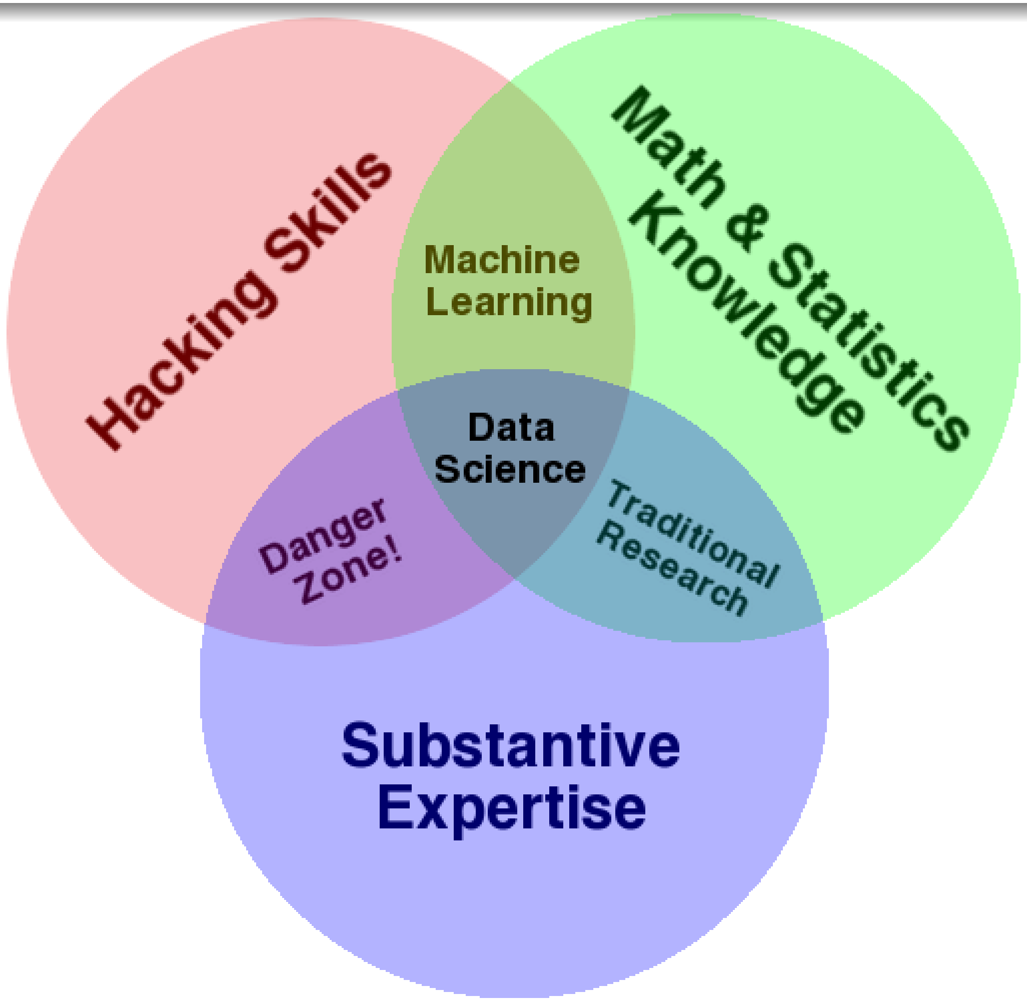
Classical Learning	Deep Learning
Can work well with less data (<10^6 rows, say)	More training data gives more accuracy (within reason!)
Easier to interpret/explain models	Model is opaque and can be fragile
Training is relatively fast	Training can require many epochs
Training requires fewer resources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

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

Emphasis of this module

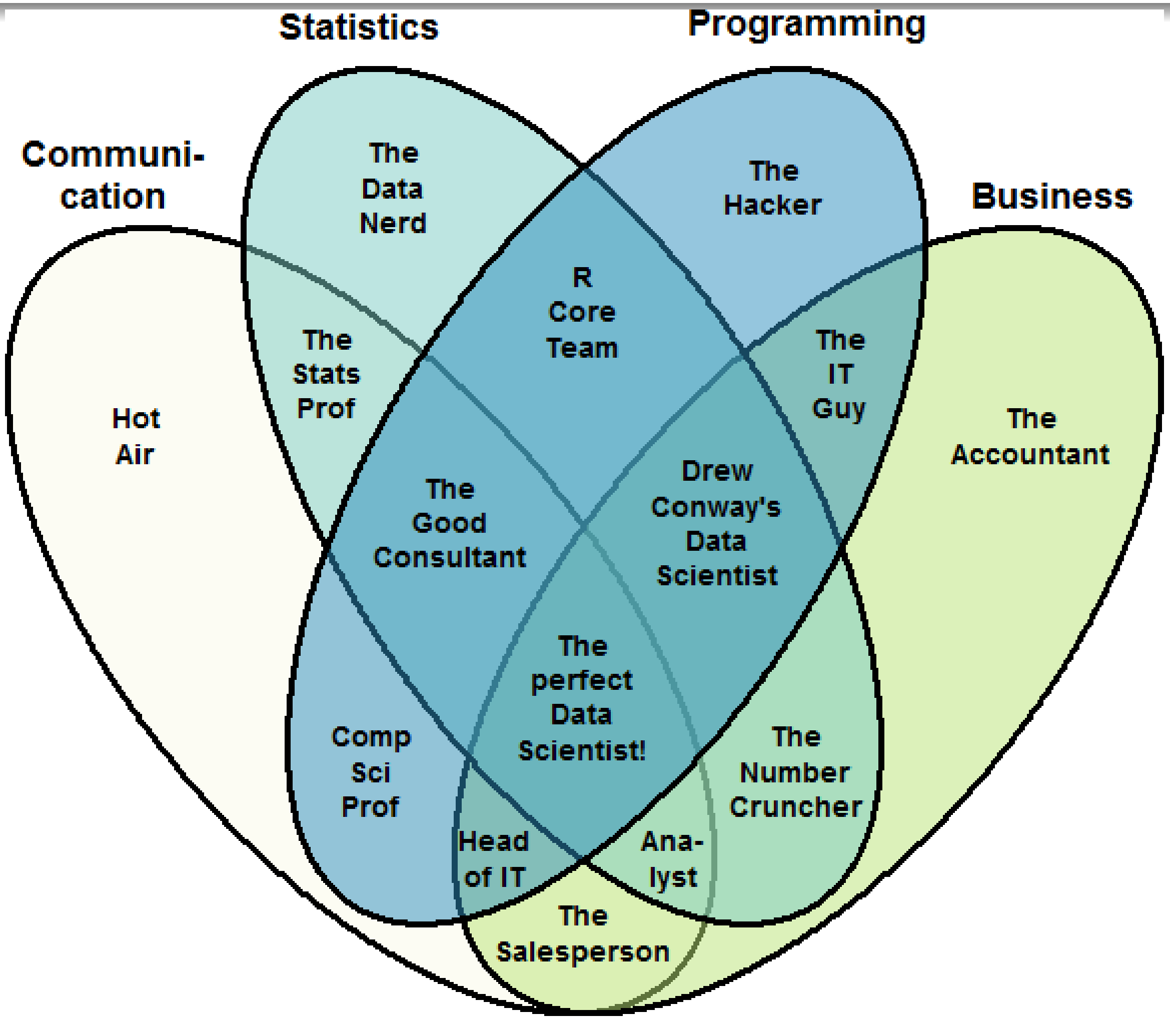
- This module covers foundations, classical models, some deep learning
- Foundations include EDA, notions of error and variance, training vs test data, . . .
- Classical modelling includes feature selection and classical approaches to learning from data
- Neural network and Deep learning go straight 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 often 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



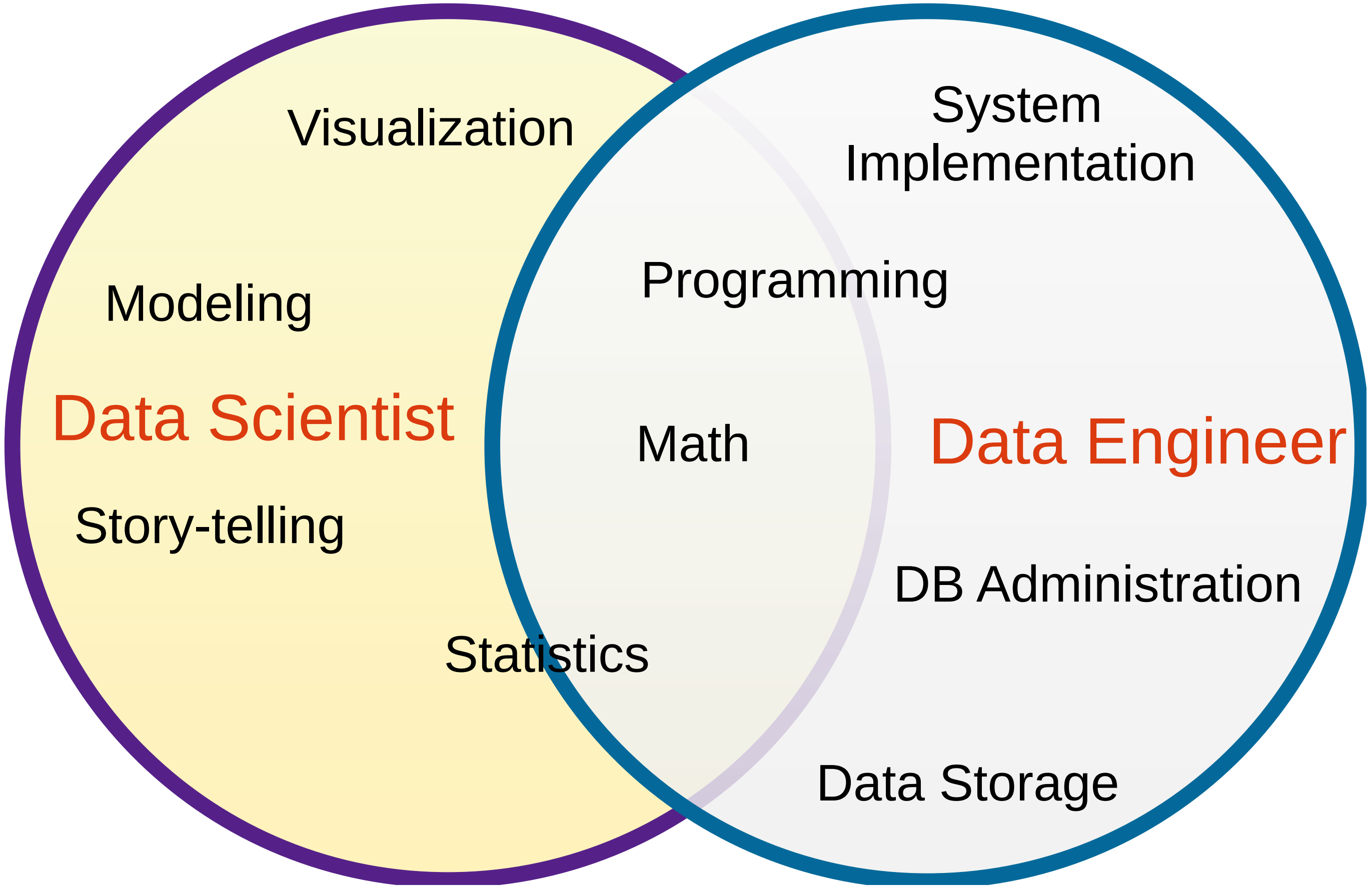
Source: `https://datascience.stackexchange.com/a/2406`

Gartner suggests the need for a *Citizen Data Scientist*

THIS IS BOB, OUR NEW CITIZEN DATA SCIENTIST. HE PREVIOUSLY WORKED AS A CITIZEN DENTIST AND A CITIZEN PILOT.
KDNUGGETS.COM

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

Data Scientist vs Data Engineer



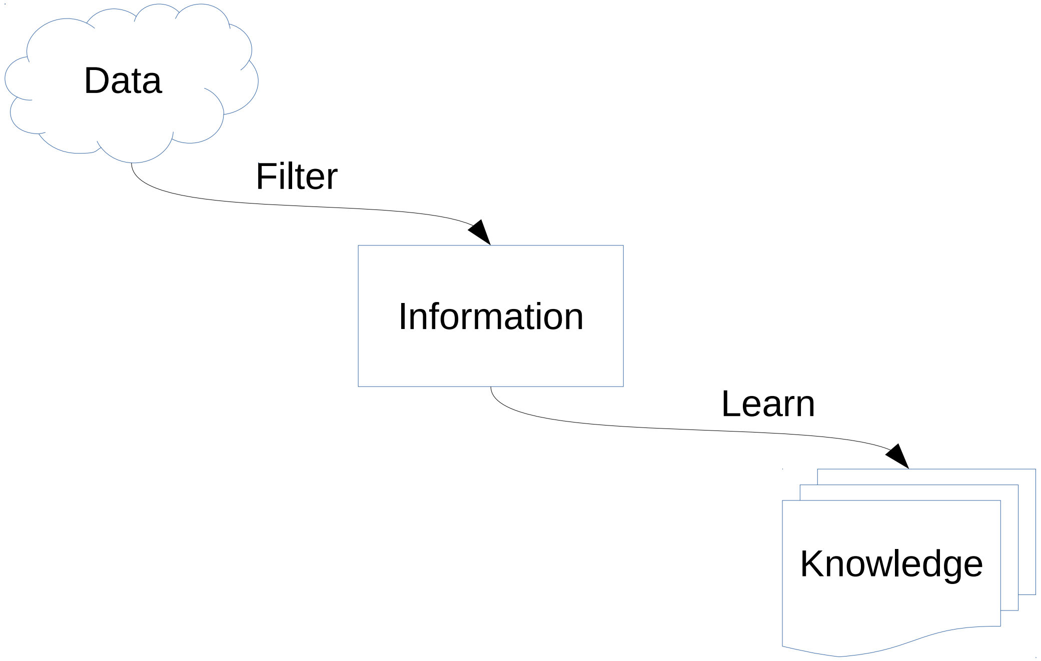
Source: `https://ryanswanstrom.com/2014/07/08/data-scientist-vs-data-engineer/`

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

Ethical Concerns

- protecting privacy
  - obtain informed consent for training data
  - pseudonymisation can be insufficient
- ensuring transparency of decisions
  - opaque models hide how predictions/decisions were made
  - interpretable/explainable models should be used where possible
- breaking cycles of bias
  - Biased training data leads to biased outputs from models
  - Diverse and representative training data gives more reliable predictions
- enabling validation
  - Poor choice of techniques, settings and model parameters can harm prediction quality
  - Careful validation can assess whether predictions are as good as expected
- enabling decisions to be challenged
  - Related to interpretability, predictions without understanding can be misleadinng
  - Need to support an audit trail, validation procedure 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

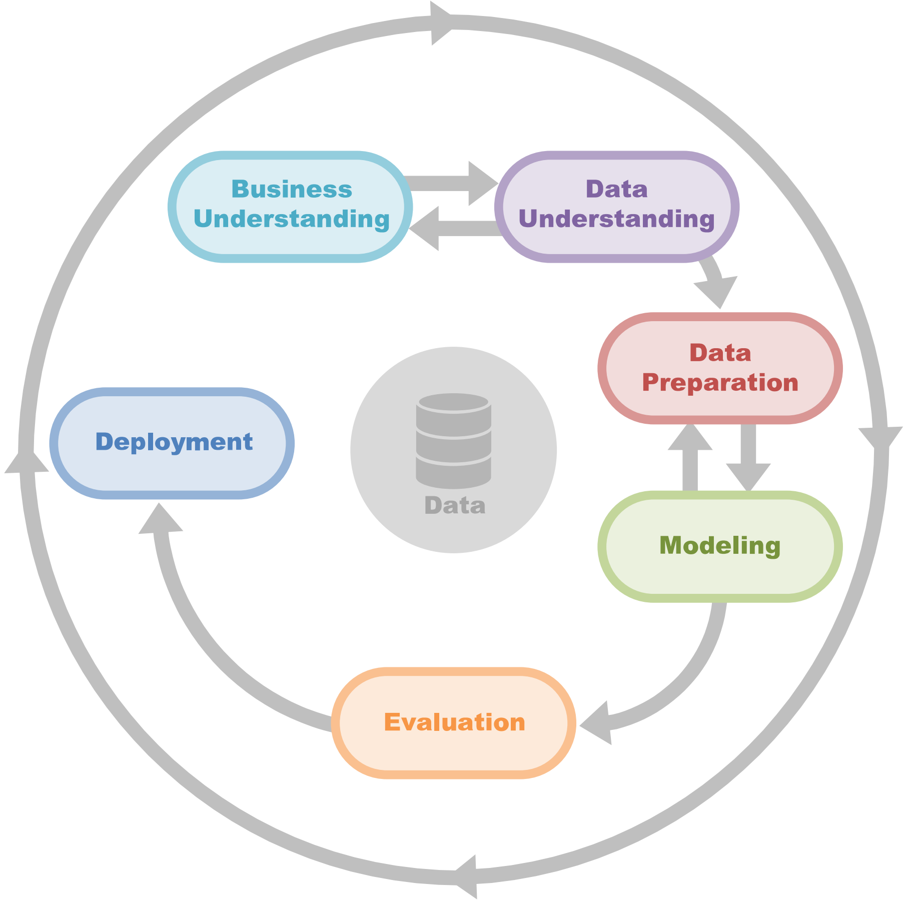
- Servers and applications log events in files and/or databases [DATA]
- *Collector* agents select specific events, in context [INFORMATION]
- Machine learning *classifiers* learn system behaviour and identify anomalies [KNOWLEDGE]
- Humans and software use this knowledge to prevent future problems [WISDOM]

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

CRoss Industry Standard Process (for) Data Mining

CRISP-DM

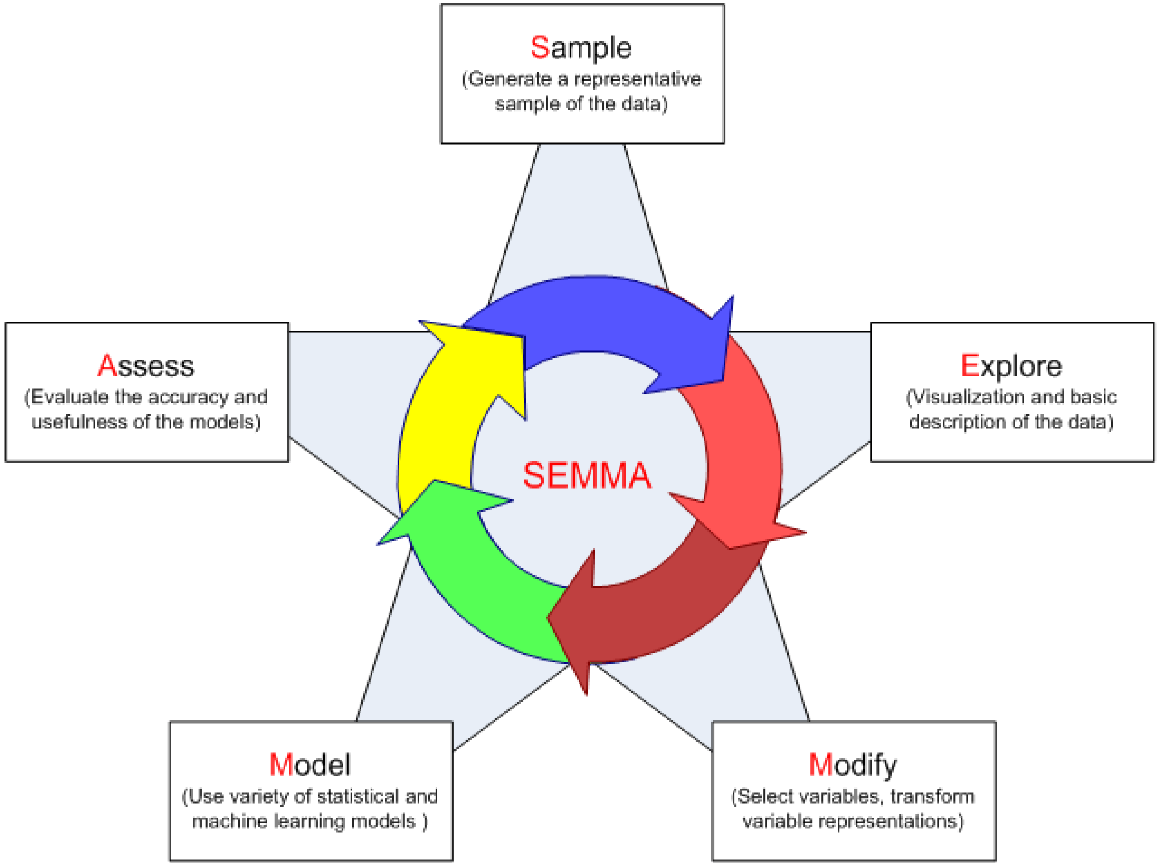
CRISP-DM is a high-level iterative process model. It gives much weight to data understanding and preprocessing and involves the data and problem owners from the start (c.f., Agile). It is the best known but predates Big Data, etc.



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.

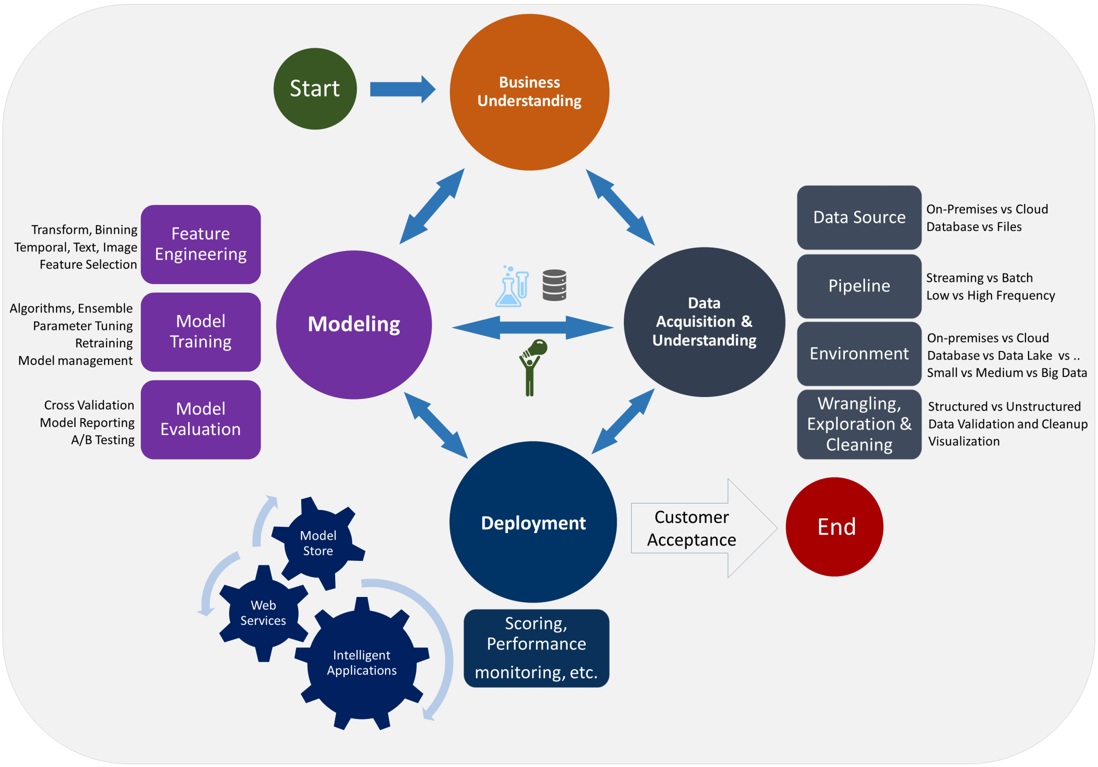

Sample
(Generate a representative sample of the data)
Explore
(Visualization and basic description of the data)
Modify
(Select variables, transform variable representations)
Model
(Use variety of statistical and machine learning models)
Assess
(Evaluate the accuracy and usefulness of the models)
SEMMA

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.

Data Science Lifecycle



Start

Business Understanding

Data Acquisition & Understanding

Data Source — On-Premises vs Cloud; Database vs Files

Pipeline — Streaming vs Batch; Low vs High Frequency

Environment — On-premises vs Cloud; Database vs Data Lake vs ..; Small vs Medium vs Big Data

Wrangling, Exploration & Cleaning — Structured vs Unstructured; Data Validation and Cleanup; Visualization

Modeling

Feature Engineering — Transform, Binning; Temporal, Text, Image; Feature Selection

Model Training — Algorithms, Ensemble; Parameter Tuning; Retraining; Model management

Model Evaluation — Cross Validation; Model Reporting; A/B Testing

Deployment

Scoring, Performance monitoring, etc.

Customer Acceptance

End

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

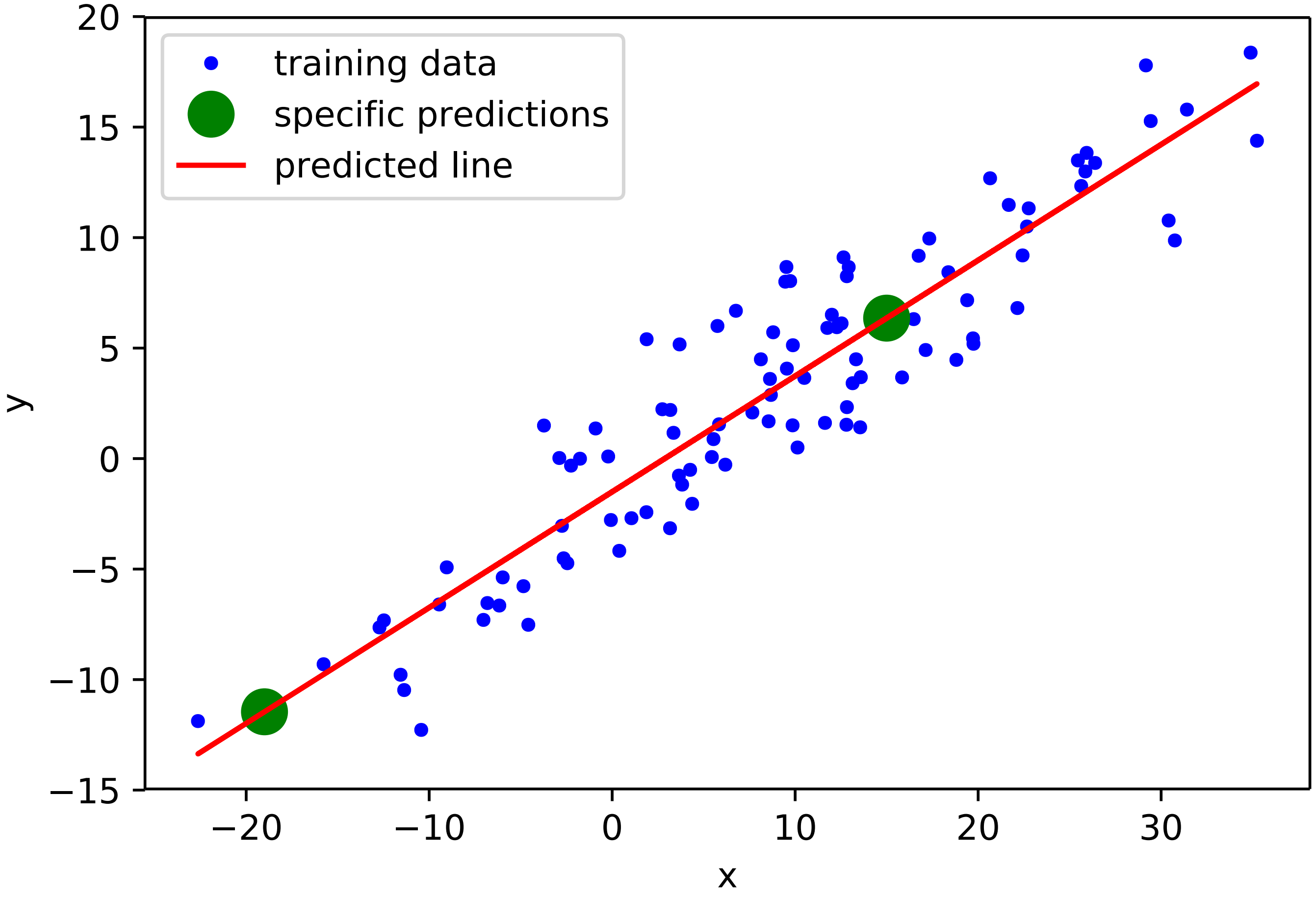
Summarised from Domingos (2015) “The Master Algorithm”

Each approach works well in specific contexts - there is no “master algorithm” yet. . .

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.

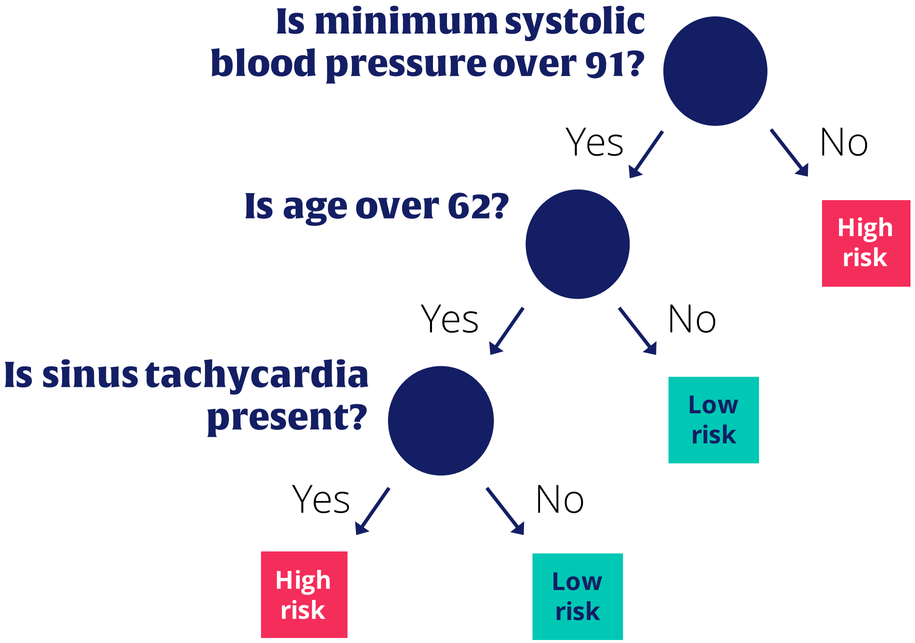
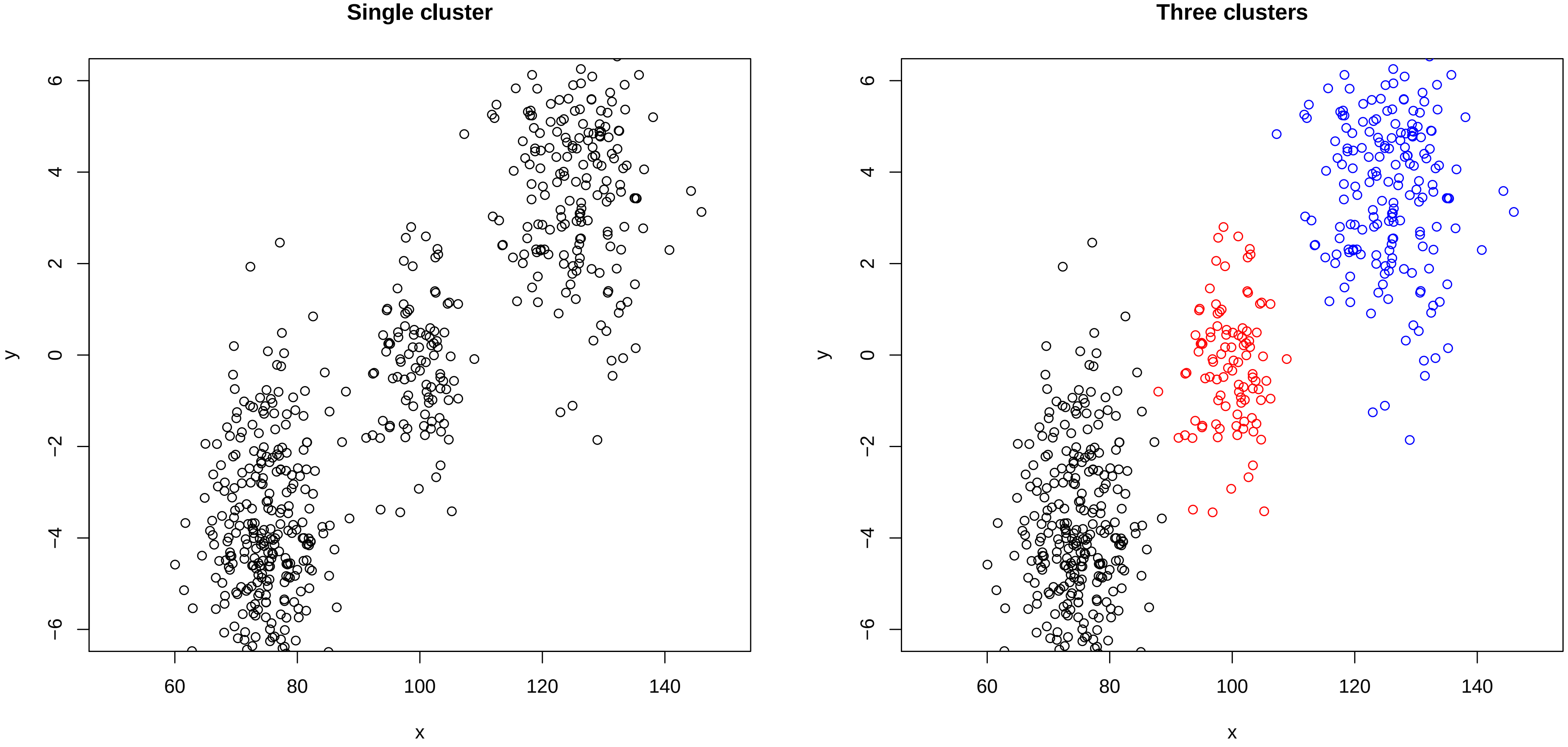

Is minimum systolic blood pressure over 91?
Yes
No
High risk
Is age over 62?
Yes
No
Low risk
Is sinus tachycardia present?
Yes
No
High risk
Low risk

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.


Single cluster

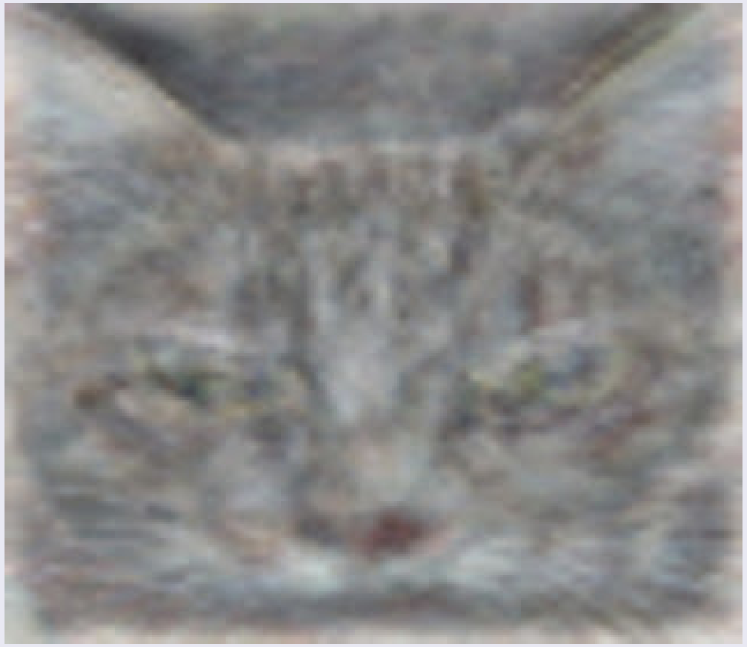
Three clusters

Clustering Example 2: Image Analysis

Le et al (2012)

- Google Brain simulator crawled the web, looking for patterns in photographs on the web.
- It was *not* looking for anything in particular!
- One pattern came up strongly…

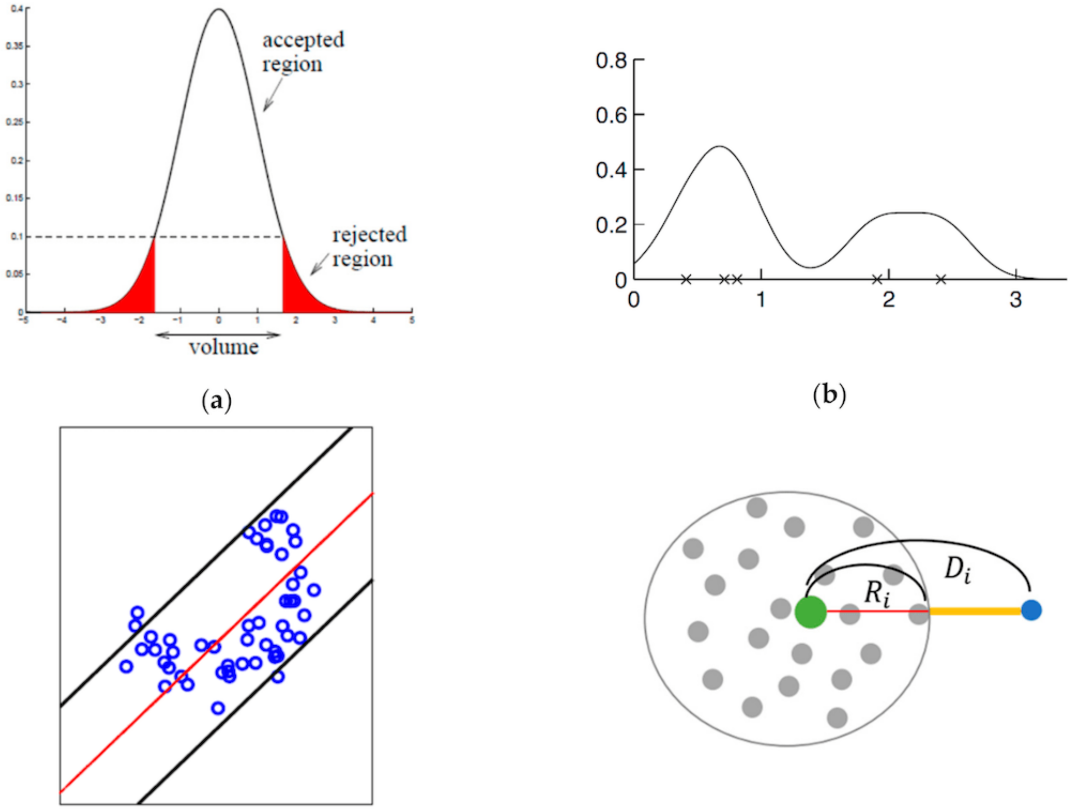
Supervised learning question: what does this represent?



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)

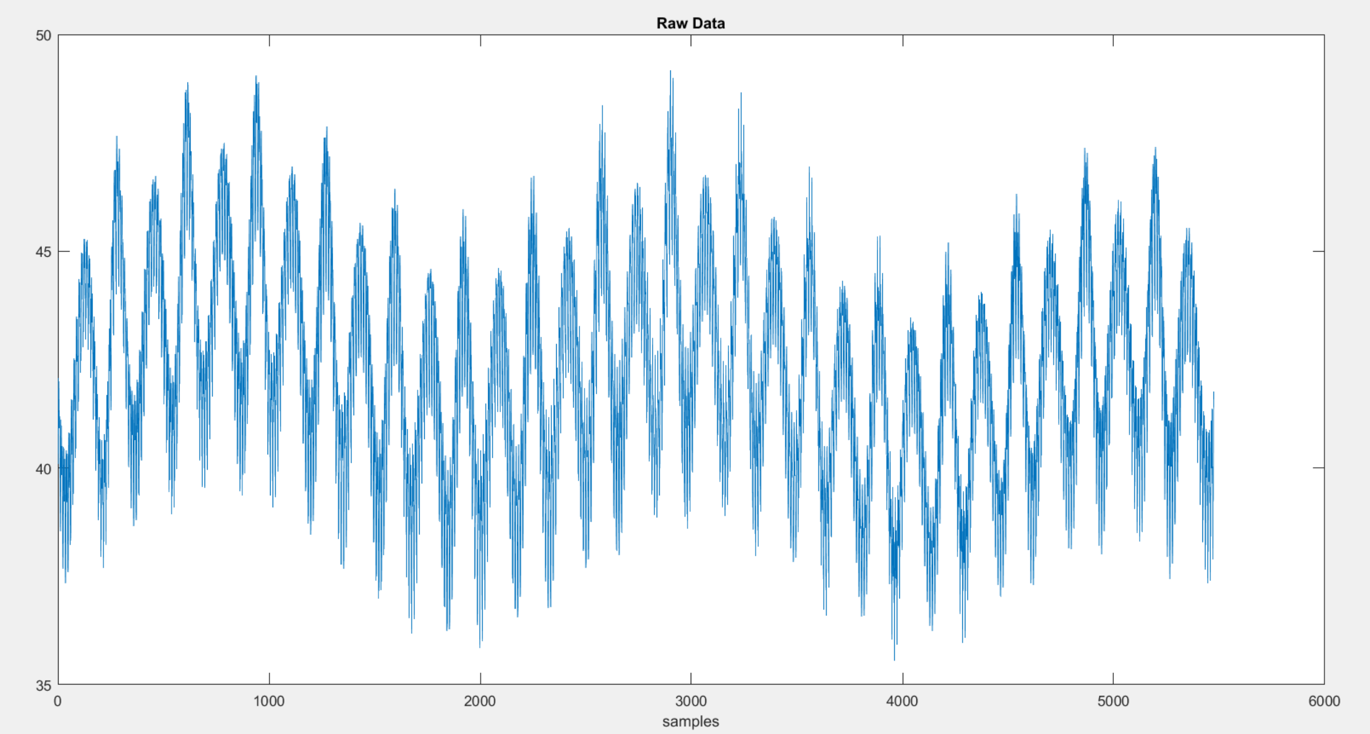
(c) (d)

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.

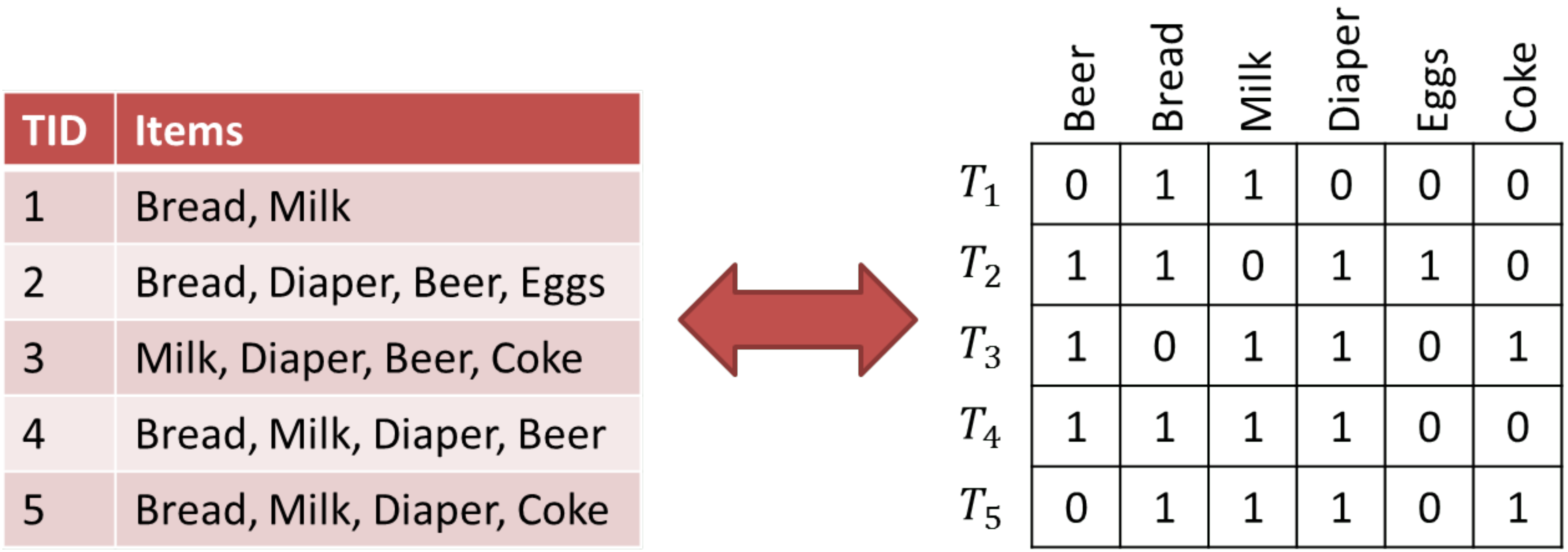


Source: https://stats.stackexchange.com/q/458491

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

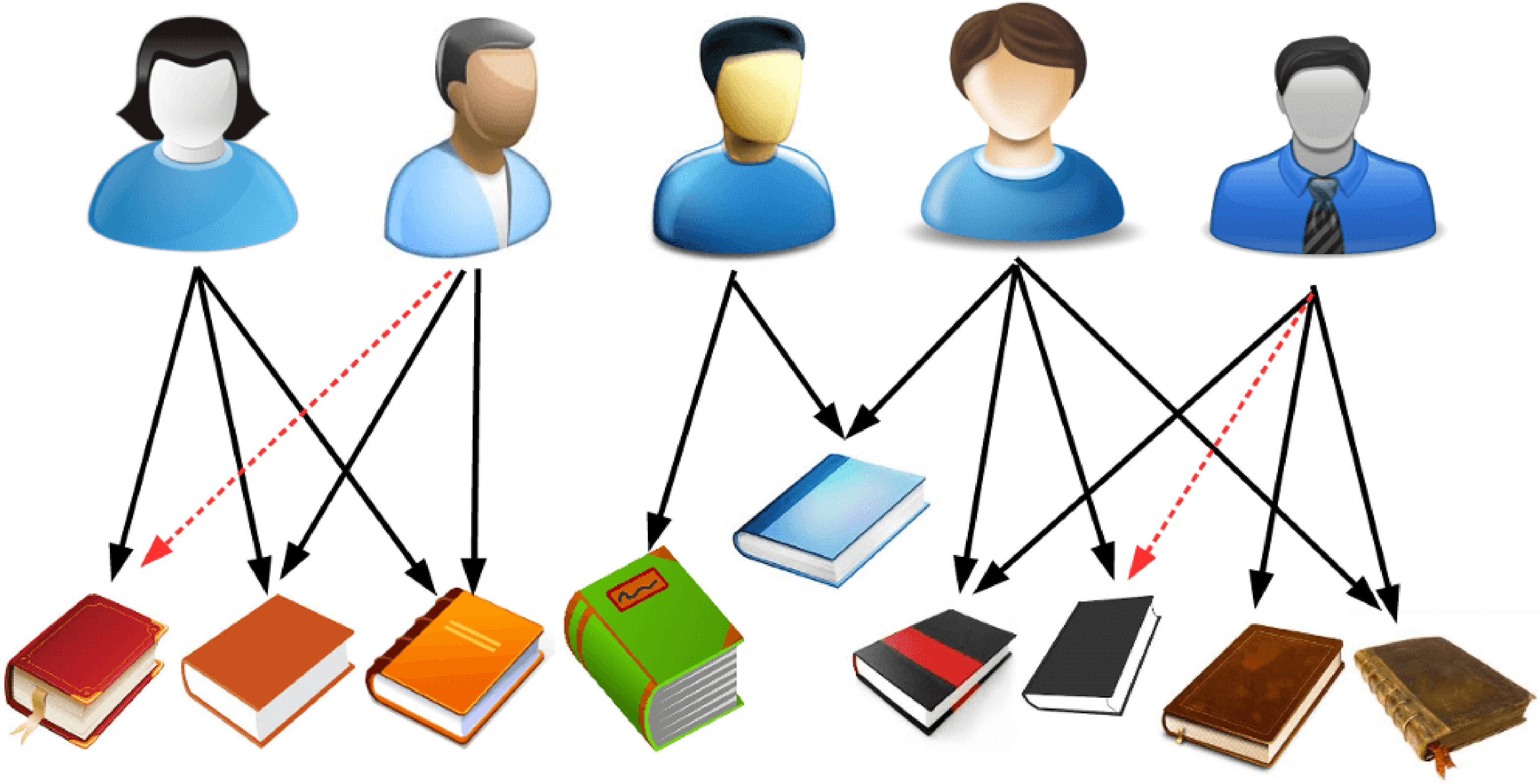
ARM is a special case of structure mining, which includes graph mining, molecular and DNA nucleotide analysis

.

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.



Sample rows from auto-mpg dataset

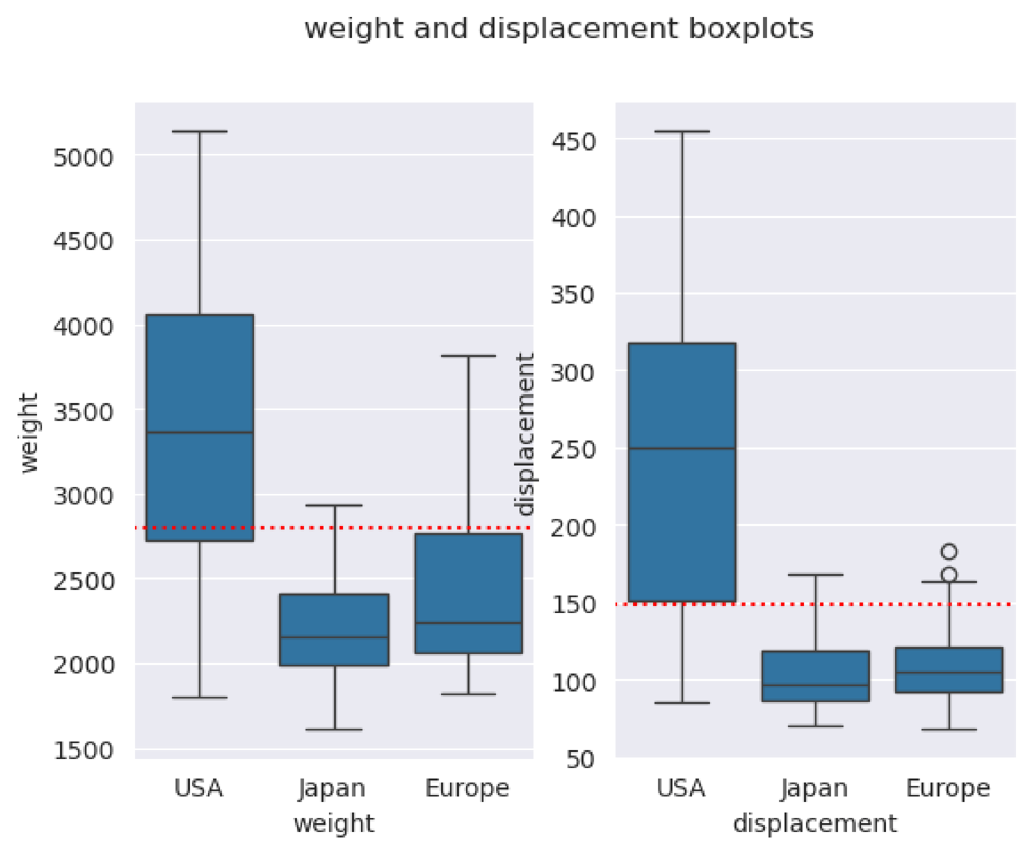
Pandas dataframe: sample rows used for training data

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
32	25.0	4	98.0	?	2046	19.0	71	1	ford pinto
112	19.0	4	122.0	85	2310	18.5	73	1	ford pinto
130	26.0	4	122.0	80	2451	16.5	74	1	ford pinto
168	23.0	4	140.0	83	2639	17.0	75	1	ford pinto
174	18.0	6	171.0	97	2984	14.5	75	1	ford pinto
206	26.5	4	140.0	72	2565	13.6	76	1	ford pinto

- The first (unnamed) “column” is the *dataframe index*; the values are like rowids in a database table.
- The rows contain data about automobiles in the US market in the early 1970s.
- The data is used to train models to predict “mpg” (*target*) given the other columns (features).

Example plot to develop an understanding of the data

Boxplot showing how manufacturer country affects weight and engine size



- Boxplot is one of many plot types available in plotting libraries
- Generated by a few lines of python code
- Side-by-side placement can help comparison
- What do the plots tell us?

Resources

moodle

- URL: Moodle: Data Mining-28115-[2024-2025]
- Used for all notices, assignment briefs and practical work submissions.

- URL: Data Mining 2024-2025 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 3.12, 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.