Intro to Big Data Analytics

PF1&2

Contents

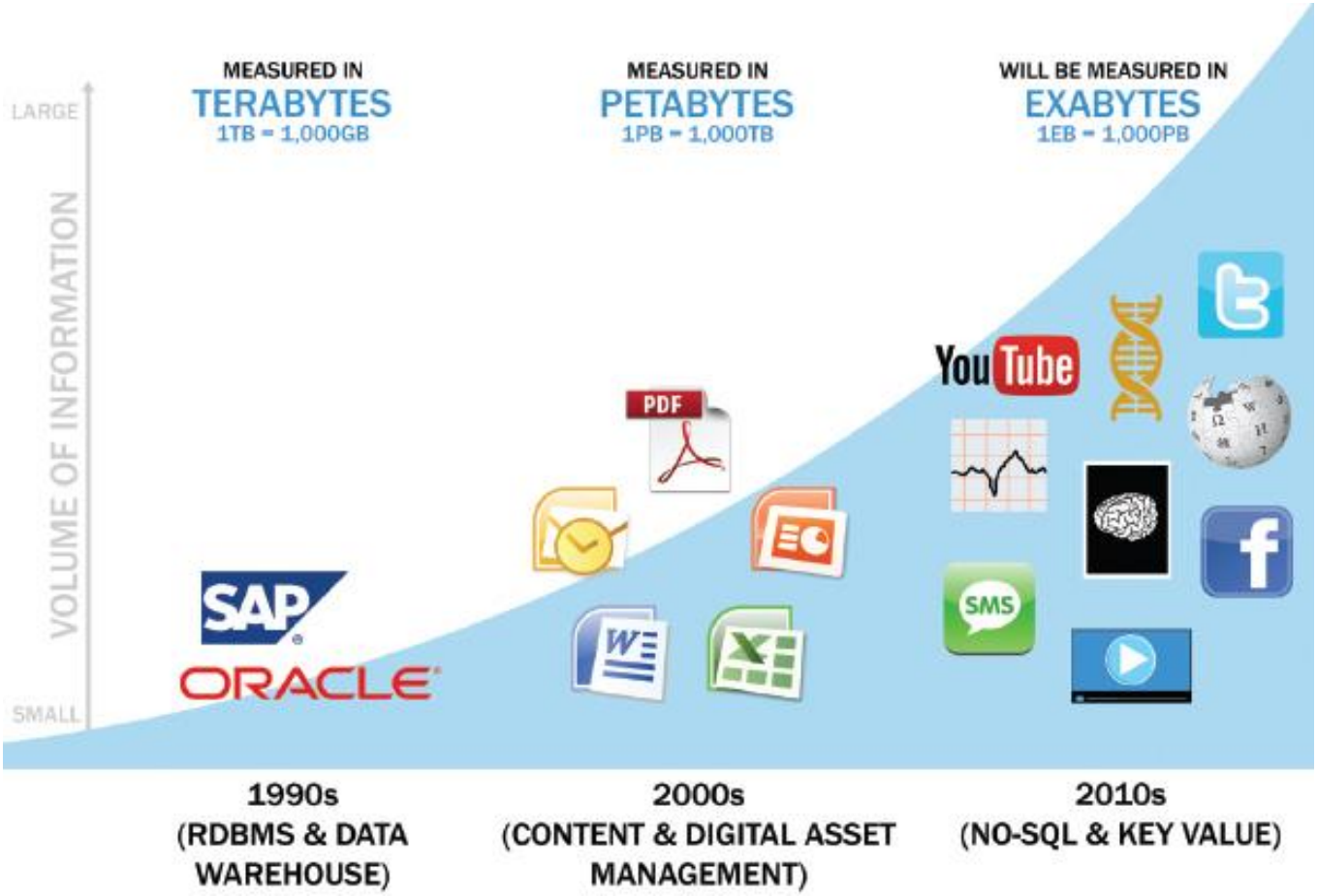
- Big Data: Characteristics and Challenges
- Data-analytic Thinking and Lifecycle
- From Business Problems to Data Mining Tasks
- Data Representation

The Rise of Big Data

- Nowadays, everyone and everything is leaving a digital footprint.

The Rise of Big Data

- The scale and growth rate of data in the past three decades.



Characteristics of Big Data

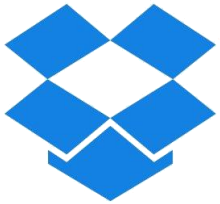
"*Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*"

– Doug Laney of Gartner, Inc.

https://educationalresearchtechniques.com/2016/05/02/characteristics-of-big-data/

Volume

- mid-1970s – mid-2000s
  - Floppy disk (~1.2MB)
- mid-90s – mid-2000s
  - CD-ROM (~700MB)
- Late 2000
  - USB Flash Drive (16MB, 32MB, …, 512MB, 1TB)
- 2007+
  - Web-based storage
    - Dropbox, Google Drive
- 2006+
  - Amazon Web Services
    - Cloud-based services



Velocity

- How fast the data is created, stored and analyzed?
- Velocity refers to the **speed** at which data is being generated and the pace at which data moves from one point to the next.

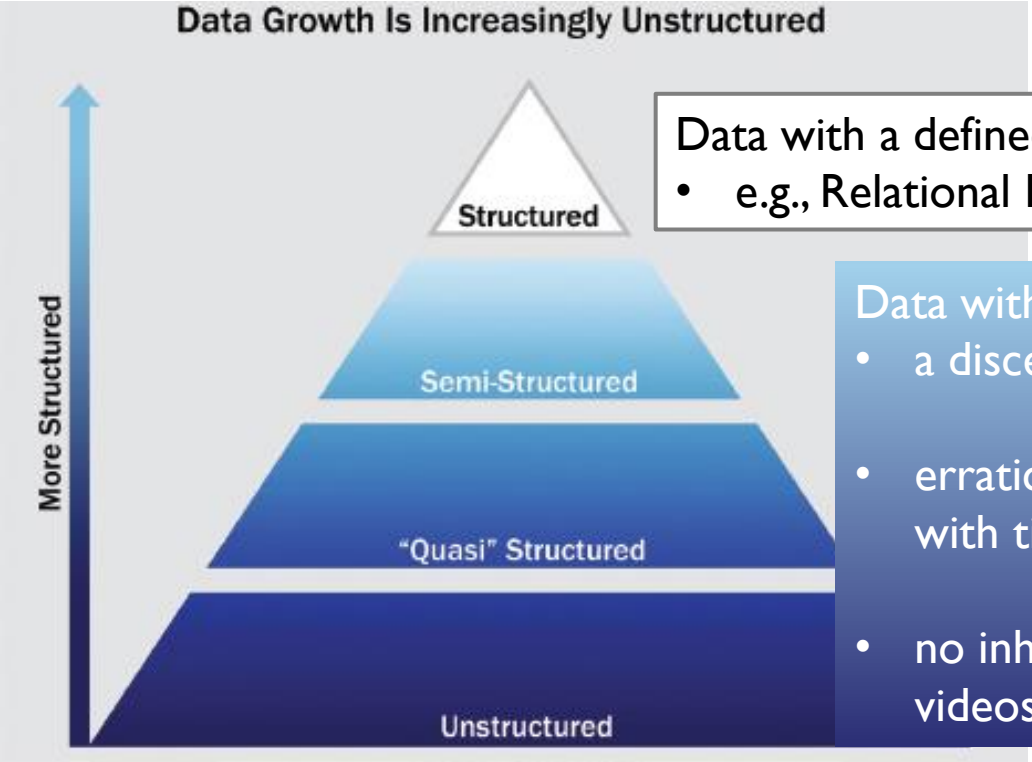
Data Streams: "History can be too long to be stored. "

Variety

Big data draws from text, images, audio, video, sensor data, email etc.

Complexity of Data Structures

- Contrary to traditional data analysis, most of the big data is unstructured or semi-structured.



Data with a defined data type and format
- e.g., Relational Database System, csv files

Data with
- a discernible pattern (e.g., xml, json files);
- erratic patterns that can be formatted with time and effort (e.g., clickstream);
- no inherent structure (e.g., text doc, videos, images).

More Data, Better Decision?

No necessarily, but a different set of skills is needed!

Predictability

More advanced: Social network data

Facebook, Twitter, Phone book

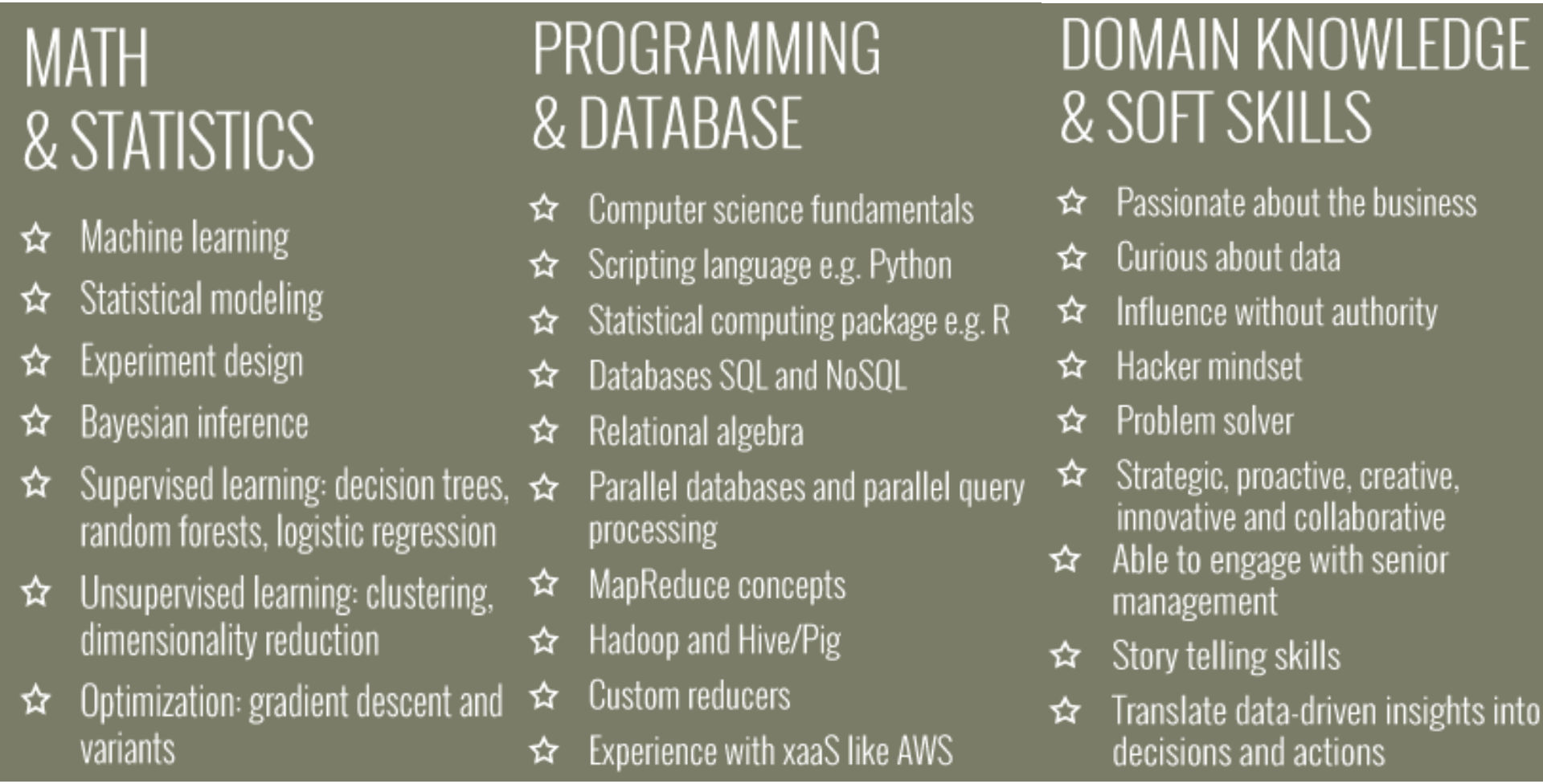
Advanced: Individual transactions data

Credit card records, Apply Pay, Alipay

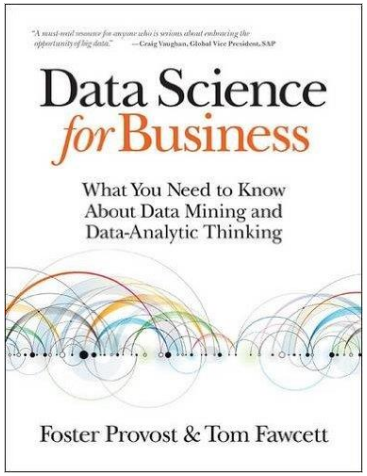
Basic: Sociodemographic data

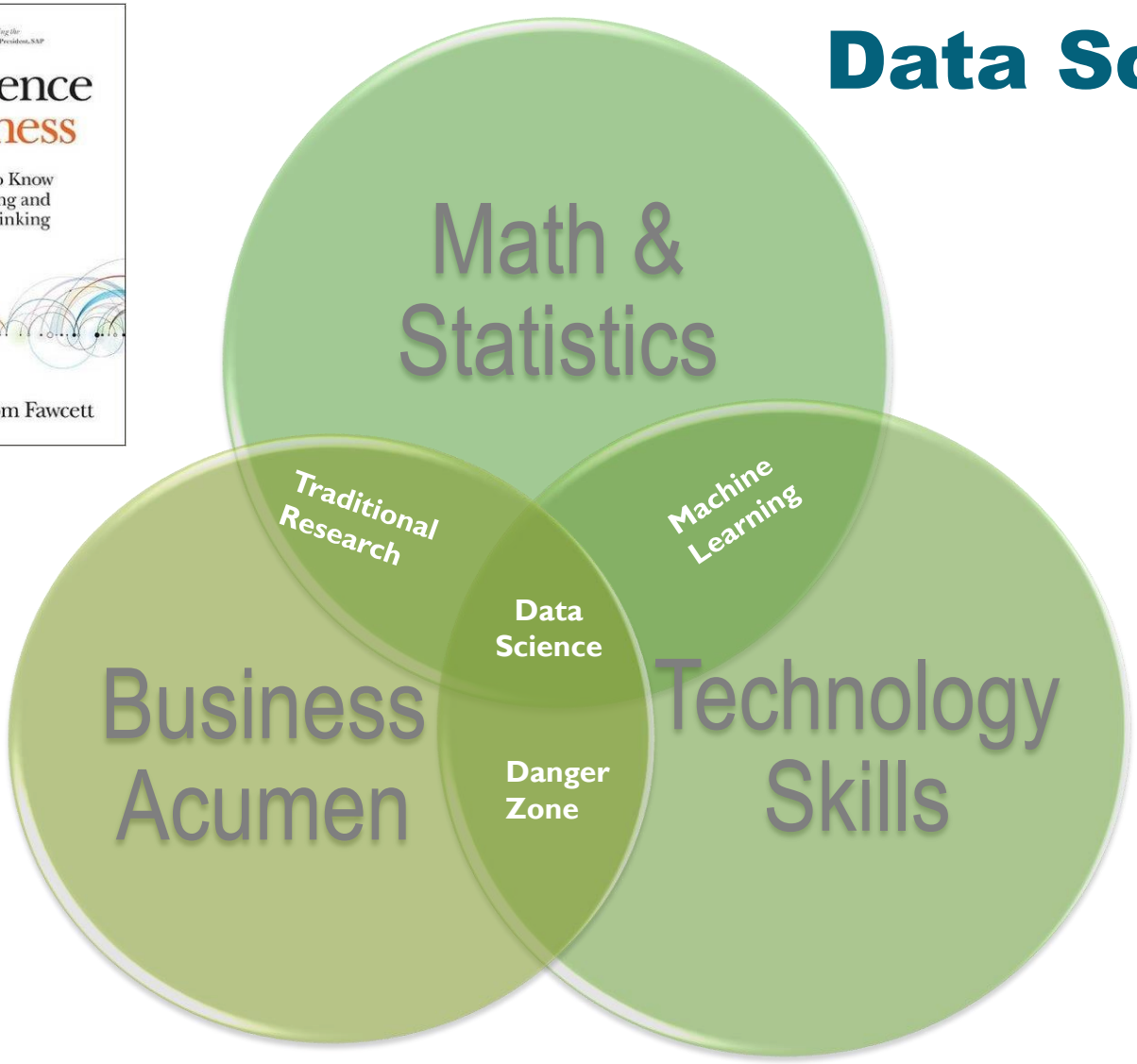
Age, gender, race, income, martial status, birth rate, death rate, household size, education, medical history

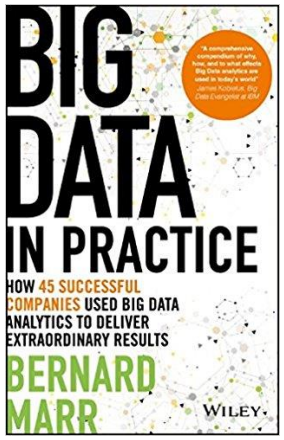
Modern Data Scientist

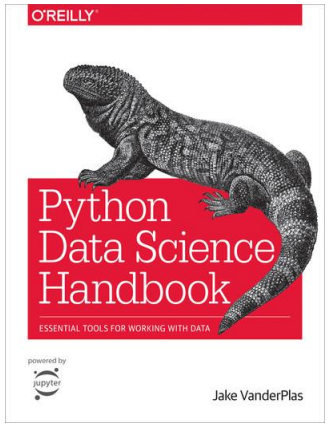


Being a Data Scientist









Data Science vs Data Mining

- **Data science** is a set of *fundamental principles* that guide the extraction of knowledge from data.

- **Data mining** is the extraction of knowledge from data, via *technologies* that incorporate these principles.

Data Science vs. the Work of Data Scientist

"Chemistry is not about testing tubes! "

Data-Analytic Thinking

Data Analysis is Essential for Business Strategies

Data collection within business

- Operations/ Manufacturing
- Supply-chain management
- Customer behavior

Data collection outside business

- Market trends
- Industry news
- Competitors' movements

Data Science/ Data-mining techniques

- **Principles**
- **Algorithms**

Data-driven decisions

- Targeted marketing
- Loan decisions
- Medical treatment

Example: Target

- **Objective**: Sell more baby-related products to pregnant customers before their competitors do.
- **Modeling and Prediction**:
  - Extract relevant data from historical shoppers (needs *domain knowledge)*
    - Pregnant women often change their diets, wardrobes, vitamin regimens…
  - Predict the probability of pregnancy for a female customer.
- **Business Action**:
  - Targeted marketing based on the estimated probability of pregnancy.
    - ☐ Give her the coupon
    - ☐ Do not give her the coupon

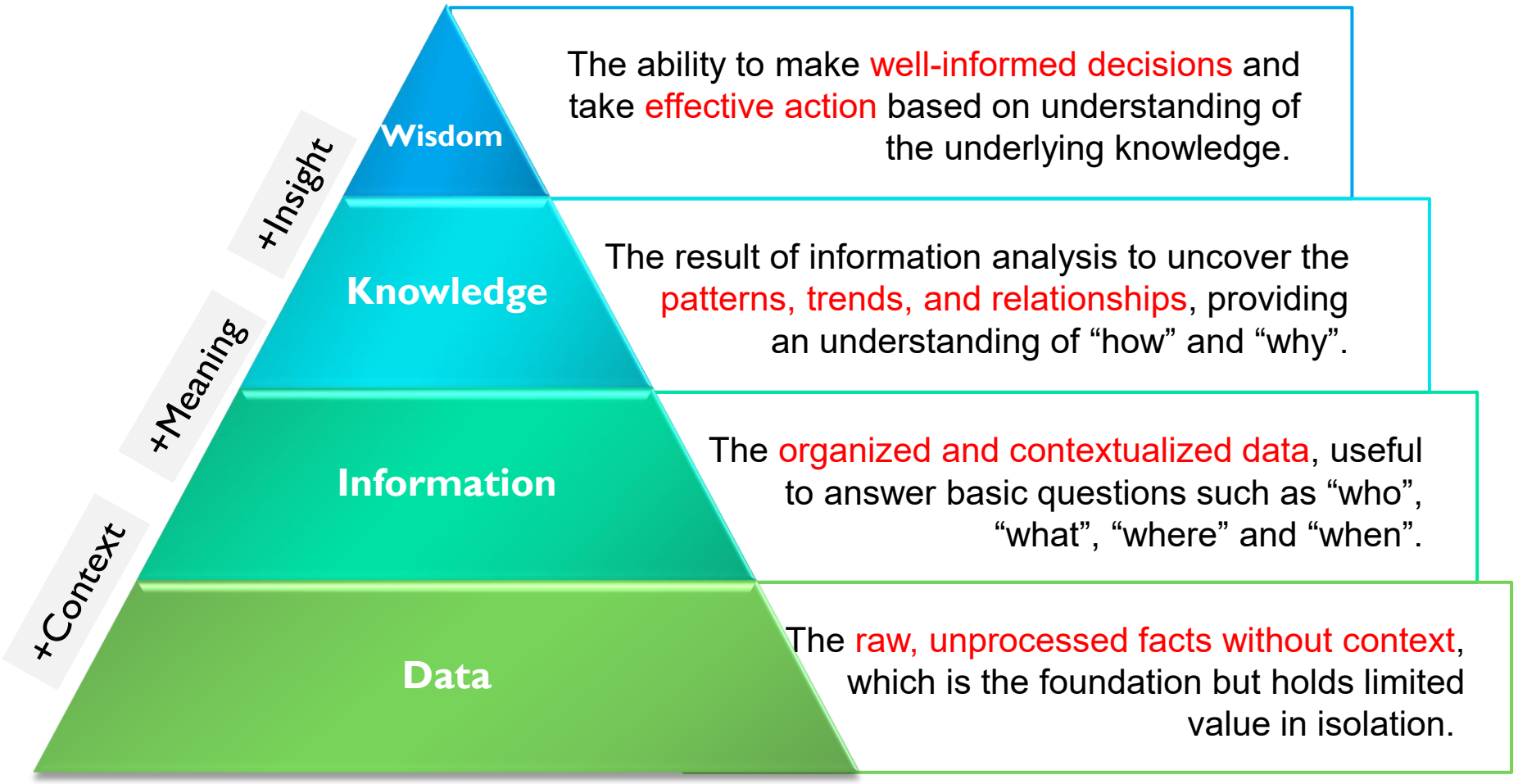
Example: Customer Churn

- **Objective**: retain existing customers due to market saturation.
  - **High churn rate**: 20% of cell phone customers leave for another company when contract expires

- **Modeling and Prediction**:
  - Extract relevant data from historical customers
    - Average charges per month, satisfaction level, usage level, etc.
  - Predict the probability of churning for an existing customer.

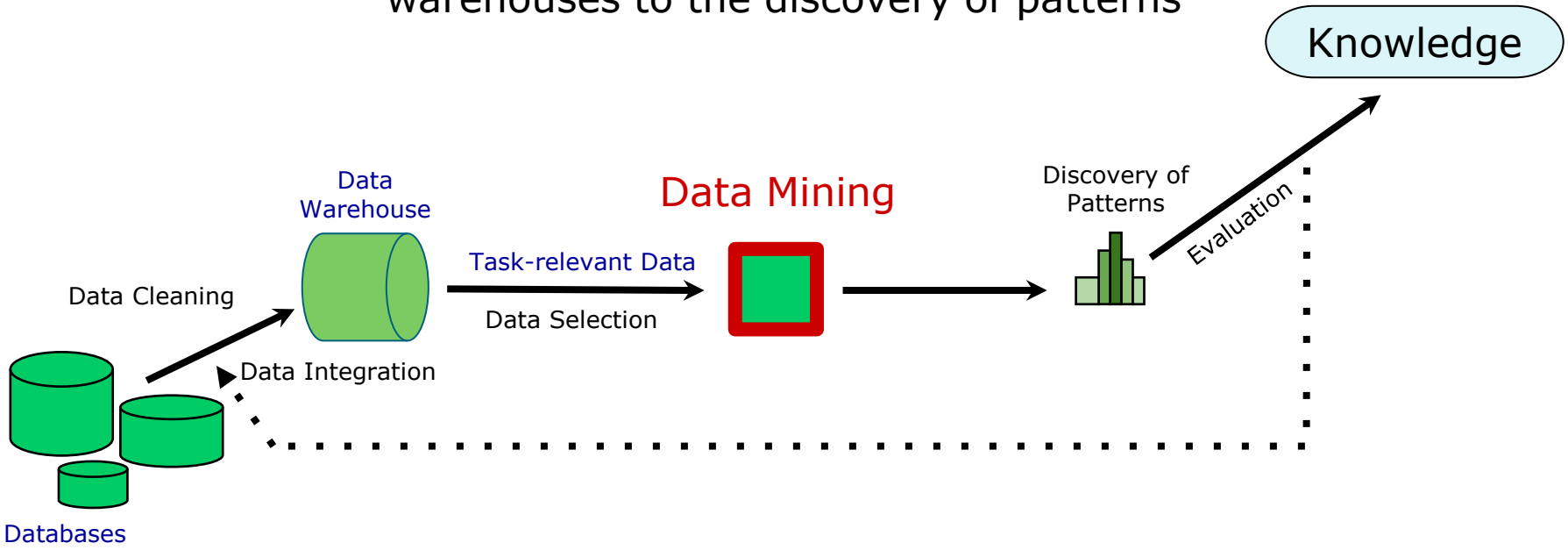
- **Business Action**:
  - Offer a special deal based on a customer's probability to churn.
    - ☐ Give him/her the special offer
    - ☐ Do not give him/ her the offer

DIKW Model: the Data Analysis Process



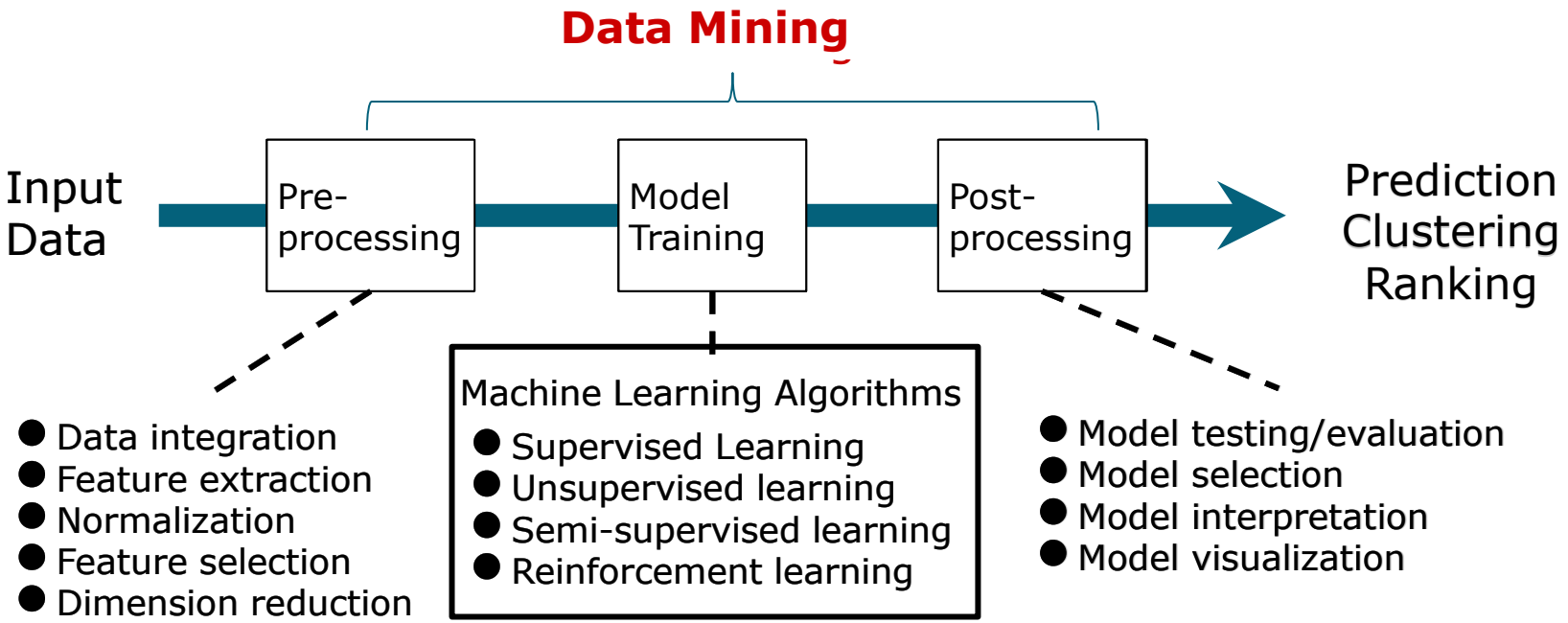
A Database View of Data Mining

Process and techniques that connects data warehouses to the discovery of patterns



Modified figure from Jiawei Han - https://hanj.cs.illinois.edu/bk3/

A Machine Learning View of Data Mining



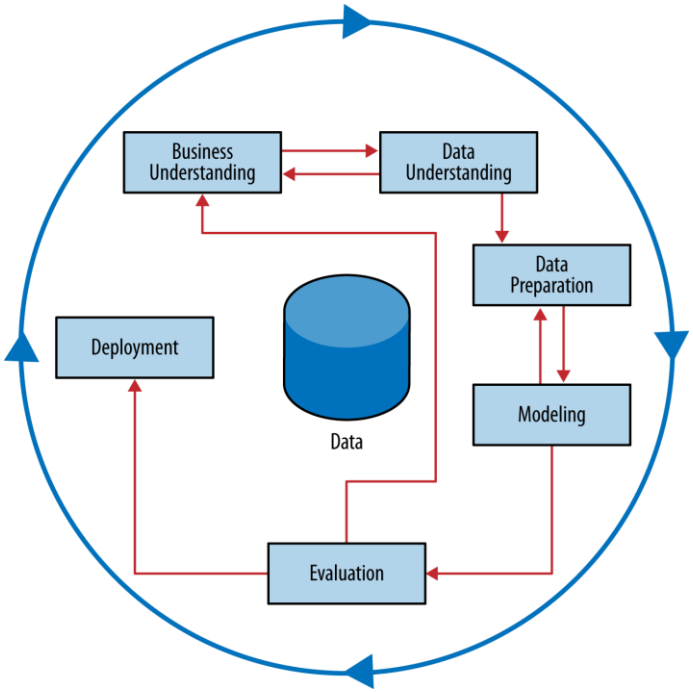
Data Analytics Lifecycle

- How to extract useful knowledge from data to solve business problems?

Cross Industry Standard Process for Data Mining

(CRISP-DM)

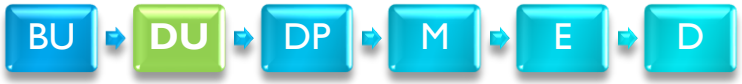
1. **B**usiness **U**nderstanding
2. **D**ata **U**nderstanding
3. **D**ata **P**reparation
4. **M**odeling
5. **E**valuation
6. **D**eployment



Business Understanding

- Analysts' creative problem formulation that reflect actual business need
  - What's the business problem and what is our objective?
  - How would we achieve this objective?
  - What parts of the use scenario constitute possible data mining models?

Data Understanding



- Understand the strength and limitation of the data
  - Good proxy
  - Measurement with errors
- Estimate the costs and benefits of each data source
  - Customer database, transaction database, marketing response database
- Decide whether investment is merited
  - Data is a strategic asset

Data Preparation

- How can we prepare messy data in the real world ready for meaningful analysis?

- Data cleansing and transformation may come to use :
  - Text/images to numerical values
  - Deal with missing values
  - Normalizing variables
  - Data dimension reduction
    - Averaging similar variables
    - Principal component analysis

Modeling

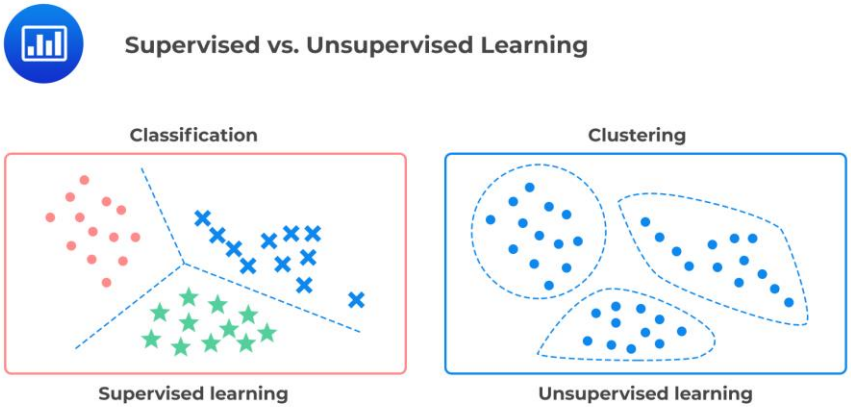
BU → DU → DP → **M** → E → D

Supervised Method

- With specific target
  - Whether a customer will cancel their service?
- **Label**: value of target variable

Unsupervised Method

- No specific target
  - Do the customers fall into different groups?
  - How does each group differ from each other?



Evaluation

- Assess the data mining results rigorously.
  - Are the models both accurate and reliable?
    - Can the model predict the data in-hand, as well as the data we get in the future?
- Evaluate the cost of different prediction errors.
  - **False alarm**: predict an unpregnant customer as pregnant –> waste the coupon and invite complaints.
  - **Missed alarm**: fail to identify a pregnant customer –> miss a business opportunity.
- Ensure that the model satisfies the original business goals.
  - Is the plan justifiable for the cost?

Deployment

- **Decision-making**: models are put into real use, in order to realize some return on investment.
  - Integrate predictive analytics for customer retention.
  - Merge fraud detection model with workforce management information system.

- **Automation**: apply a data mining model into a production or operation system.
  - Automatically build and test models in production.

Problem Formulation Taxonomy

- **Business problems** can be categorized into a set of **common data mining tasks**, so that we can assign appropriate algorithms and prepare data for each task.
  - Classification
  - Regression
  - Clustering
  - Similarity Matching
  - Association Rules
  - Link Prediction
  - Dimensionality Reduction
  - Causal Modelling

  …

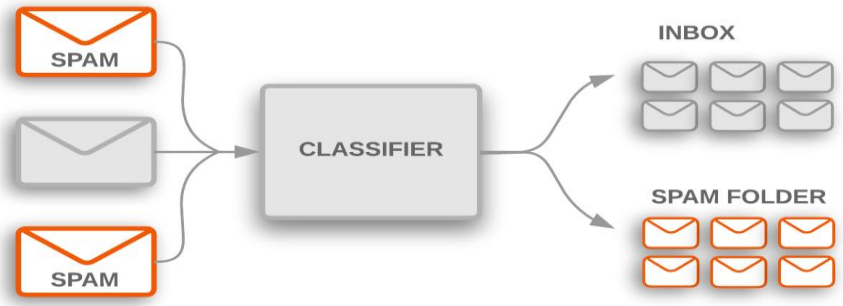
Common Tasks

- ★ **Classification**
  - Will this consumer respond to our campaign given his/her certain features?
    - **Target variable** – Yes or No is categorical
- ★ **Class probability estimation** (Scoring)
  - How likely this consumer will respond to our campaign?
    - Chance that it is “Yes” is 85%
- ★ **Regression**
  - How much will this consumer spend on our service given his/her age, gender and income?
    - **Target variable** ($398 per month) is numerical/continuous

Classification

attributes (X) → categorical outcome (Y)

The attributes of email spam are (X) → Is (Y) email spam?

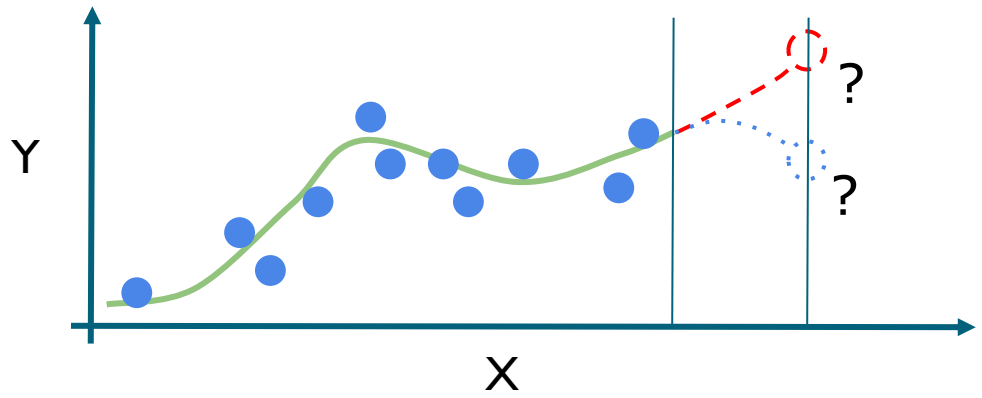


https://developers.google.com/machine-learning/guides/text-classification/

Regression

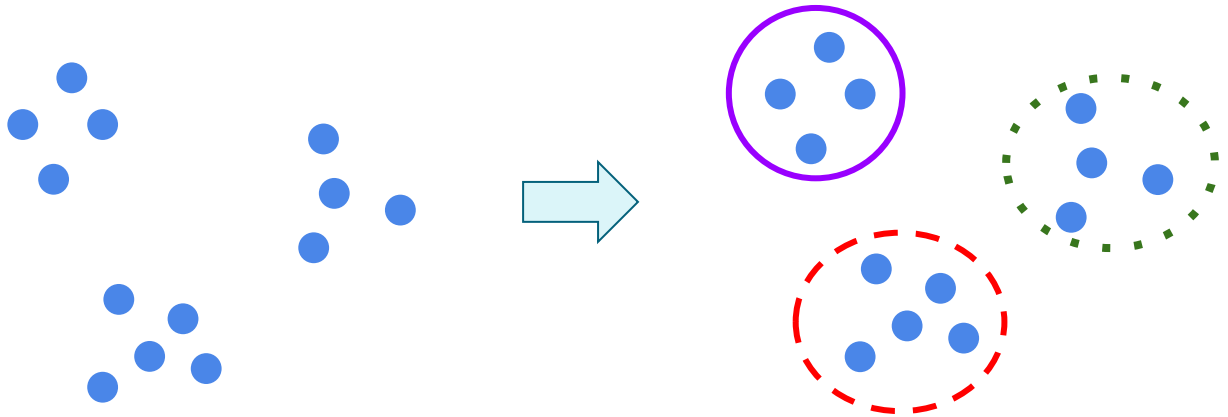
attributes (X) → numerical outcome (Y)

Family annual income (X) → Annual spending (Y) on our service



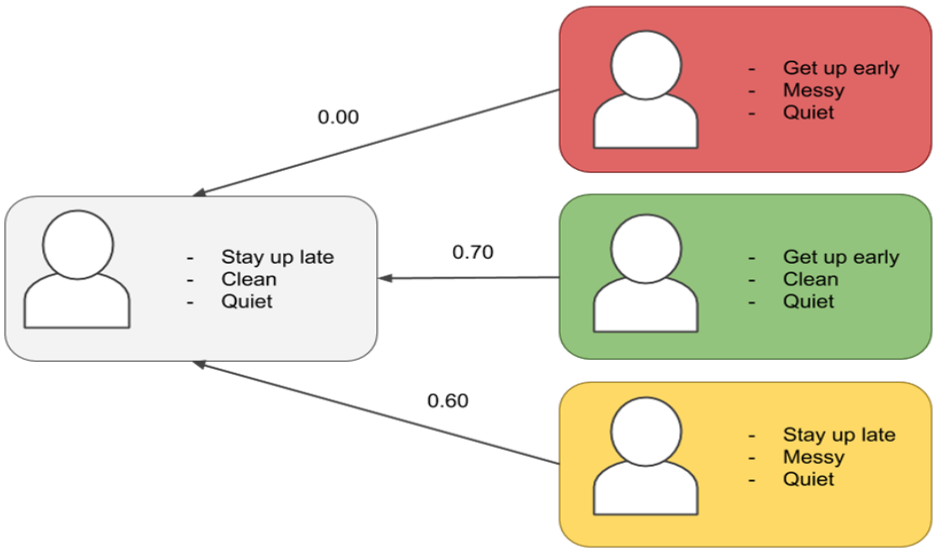
Common Tasks

- ★ **Clustering**
  - Group individual objects in a population by similarity of their own attributes (features)
    - Do my customers form several natural groups?
    - What products or service should we offer or develop for each group?



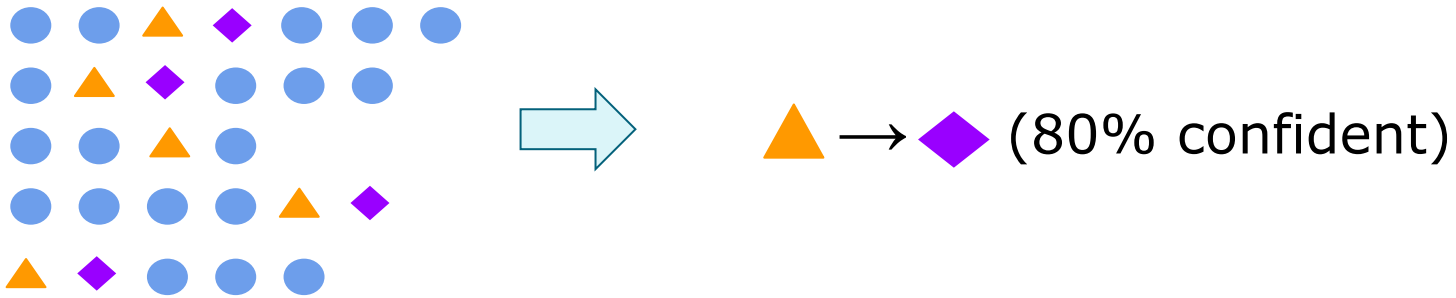
Common Tasks

- ★ **Similarity Matching**
  - Identify similar individuals based on their features.
  - Can we find consumers similar to my best customers?
    - IBM: ***Firmographic*** leads to suitable product recommendations



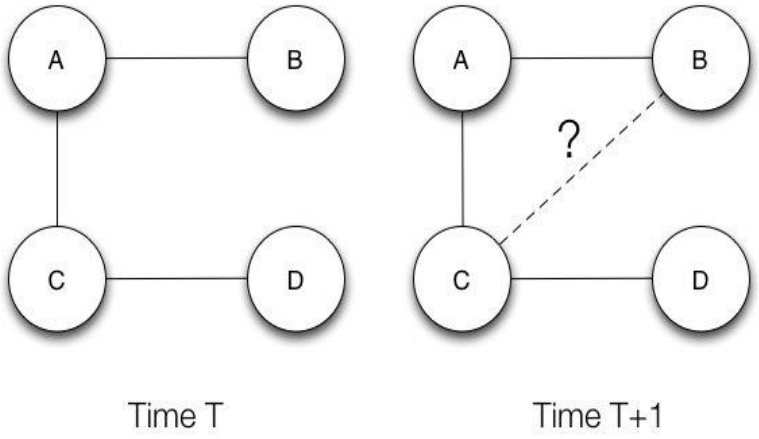
Common Tasks

- **Association Rules**
  - Find association between items based on co-occurrence, usually applied to transaction data
    - What items are commonly purchased together?
    - Applications: Recommendation algorithm



Common Tasks

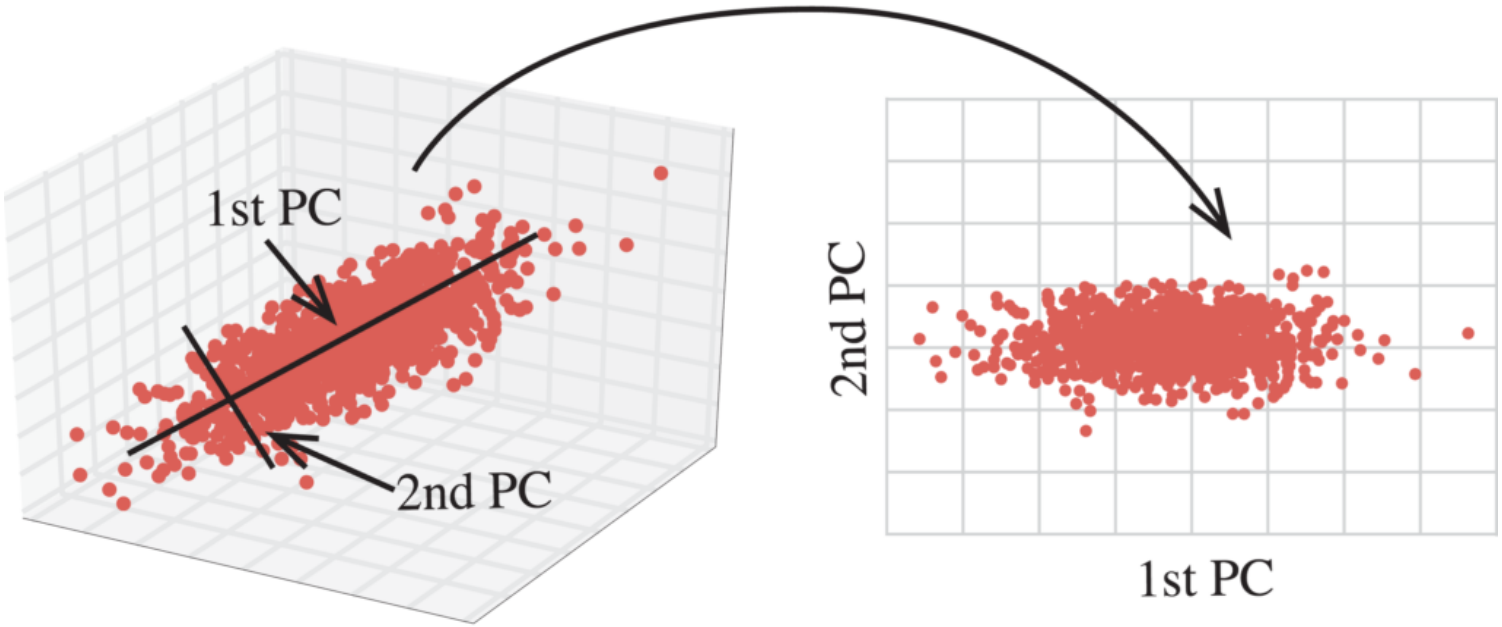
- **Link Prediction**
  - Predict and estimate strength of connections between data items
    - Your friends rated this movie highly, but you haven't watched it. The system suggests a strong link between you and the movie.
    - Since George and Mary share 2 friends, should George become Mary's friend?



https://www.researchgate.net/figure/Link-Prediction-an-example_fig7_281319964

Common Tasks

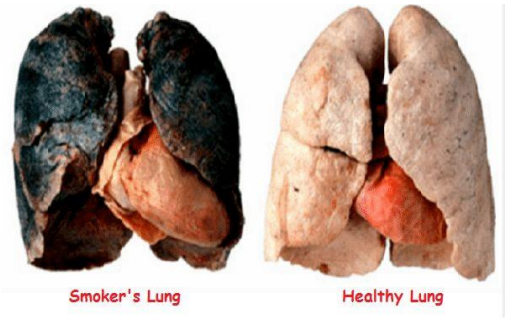
- **Dimensionality Reduction**
  - Replace a large set of features with a smaller one
    - Cost: Loss of information
    - Benefit: Gain insight (e.g., **latent topics**)



https://medium.com/@TheDataGyan/dimensionality-reduction-with-pca-and-t-sne-in-r-2715683819

Common Tasks

- **Causal Modelling**
  - Understand what events or actions actually affect others (**Counterfactual analysis**)
    - Ask “what if" questions to explore alternative realities and assess how different actions could have led to different outcomes.
  - Require substantial investment in data
    - Randomized controlled experiments (A/B tests)

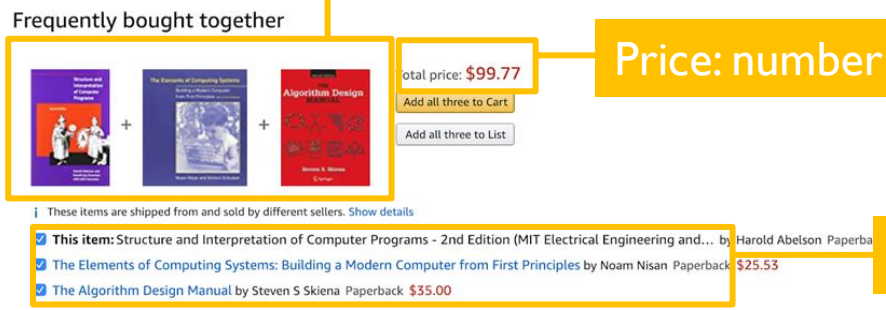


Towards Real World Data

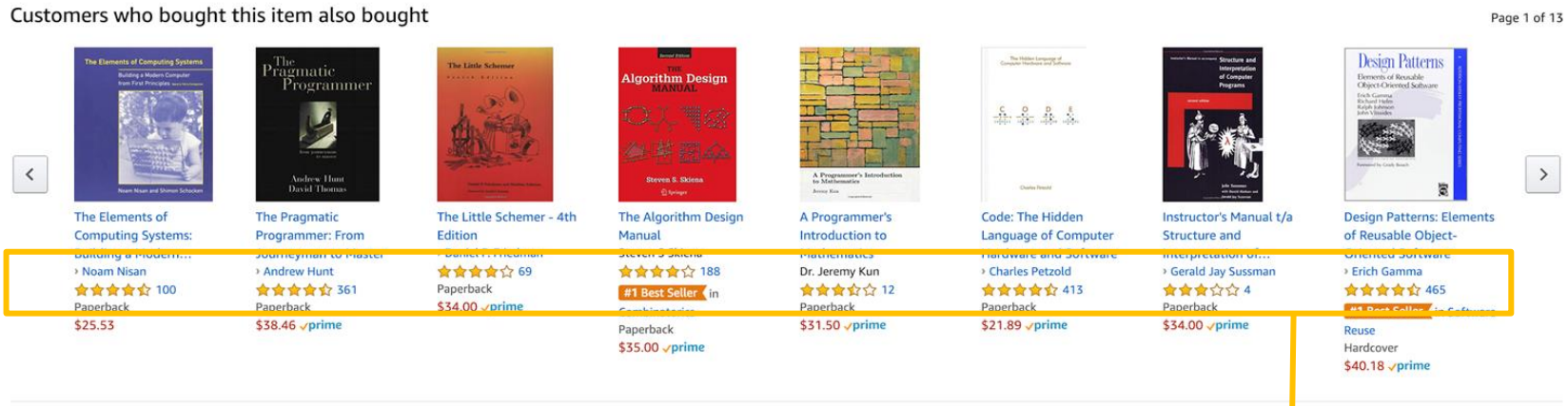
Book covers: image

Price: number

How do we process and analyze such complex and messy information?



Book names: text



Reviews: numbers and texts

Challenge of Real Data

What we are used to:

Binary Code - Christiaan Colen - https://www.flickr.com/photos/christiaancolen/20607150556 - CC BY SA 2.0

What the reality is:

overflowing - zoetnet - https://www.flickr.com/photos/zoetnet/7929093836 - CC-BY-2.0

Stata Center MIT - King of Hearts - https://commons.wikimedia.org/wiki/File:Stata_Center_MIT_October_2014.jpg - CC-BY-SA-3.0

Analyze Real-World Data

- There is a big gap between real data and analytics.
  - Data representation bridges this gap.

- **Data representation** is a **mathematical way** to describe data.
  - What is a basic object of information?
  - What are the attributes/properties of the data object?
  - How are the attributes structured?
  - How to assign values to the attributes?
  - How are different data objects related?

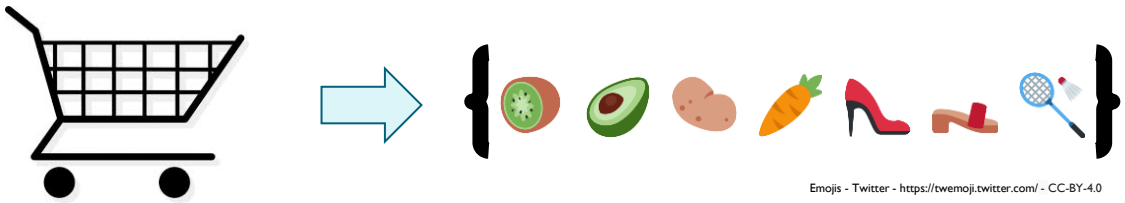
	age	gender	income
1	47.31613	Male	49482.8104
2	31.38684	Male	35546.2883
3	43.20034	Male	44169.1864
4	37.31700	Female	81041.9864
5	40.95439	Female	79353.0144
6	43.03387	Male	58143.3633
7	37.55696	Male	19282.2306
8	28.45129	Male	47245.2385
9	44.20268	Female	48332.5198
10	35.15167	Female	52567.8903

Data Representations

- Item Set
- Vector / Matrix
- Sequence
- Graph / Network
- Time Series
- Spatial/Spatiotemporal
- Stream

Itemset Data

Data Object: a shopping basket (transaction), a piece of text, a board of directors …



Attribute: appearance of a categorical item in the object.

- The item can be a product, a word, a person, etc., depending on the object.

The Itemset Representation

Each data object is represented as a set of items (itemset):

$$X = \{x_1, x_2, \ldots, x_k\}$$

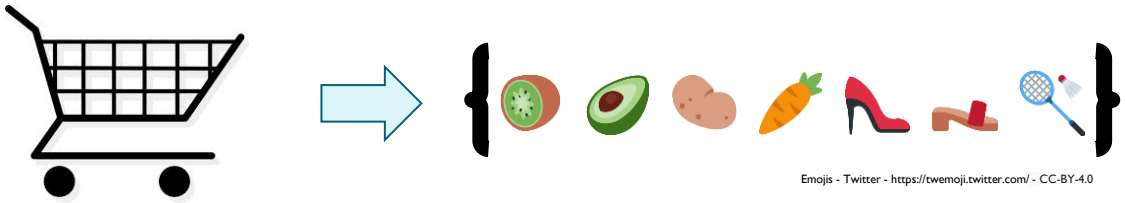
- The categorical item x_i belongs to itemset X if and only if it appears in the itemset (i.e., object).
- Order or counts of the items don't matter.

- Attribute values can also be True/False (or 1/0), depending on whether the item appear in the object.

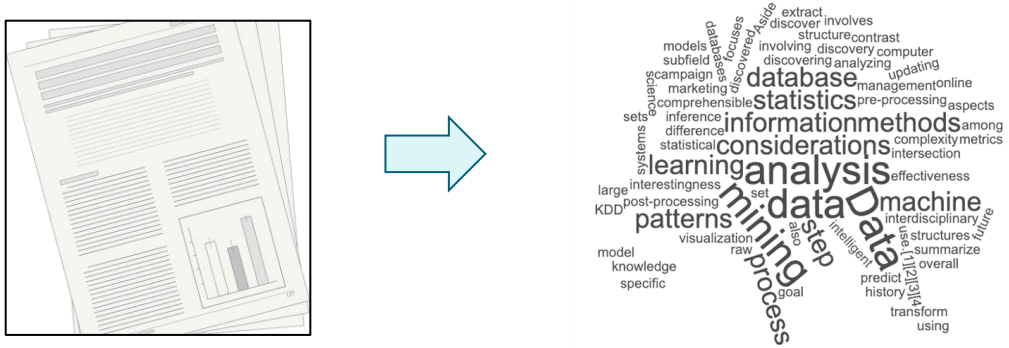
										
0	False	False	False	True	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False

Example of Itemsets

Shopping Baskets:



Text (as bag-of-words):



Vector Data

Data Object: e.g., a user's ratings of various products, course grades of a student

Attribute: a numerical property/feature of the object.

- e.g., Kimono=5; Shoe=4; Piano=3, etc.

The Vector Representation

- Each data object is represented as n-dimensional vector:
  - Each dimension records an attribute (e.g., age, edu). Each attribute is unique.

$$\vec{X} = \langle x_1, x_2, \ldots, x_n \rangle$$

  - X_i is numerical value of object X at the i^{th} attribute.

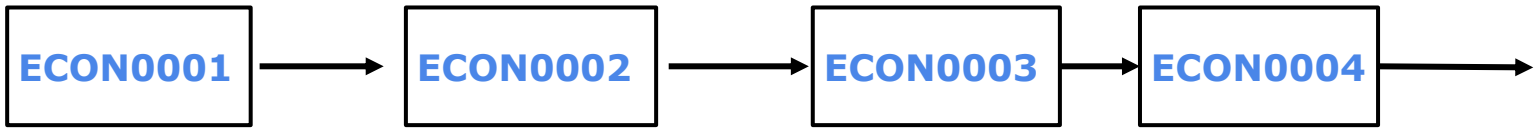
- Multiple objects → A matrix (a collection of vectors).

	👚	👠	🎹	⌚	☂️	🔭	🏀	🎃
👩:	5	4	3	1	4	4	2	5
👱‍♀️:	3	1	3	5	5	5	2	1
👨🏽:	2	4	4	2	5	4	5	3

Sequence Data

- **Data object**: a curriculum path, a DNA sequence, a session of search queries, a sentence (of words), a trace of user actions

- **Attributes**: pairs of positions and categorical item, in a sequential order



(For a degree program, each course and its position are set in a sequential order)

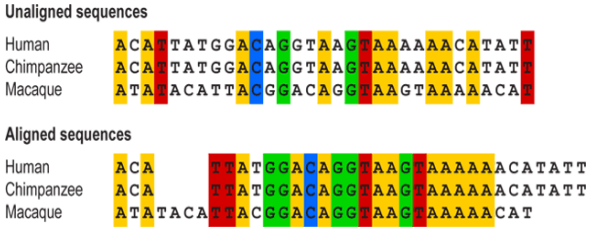
The Sequence Representation

- Each data object is represented as a sequence of items:

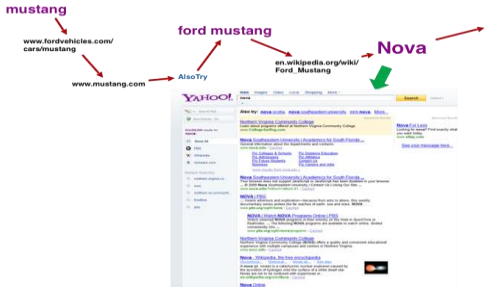
$$X = \{(x_1, 1), (x_2, 2), ..., (x_k, k)\}$$

  - x_i is the categorical item appeared at the i^{th} position of X.

- Other examples of sequence data
  - DNA sequences



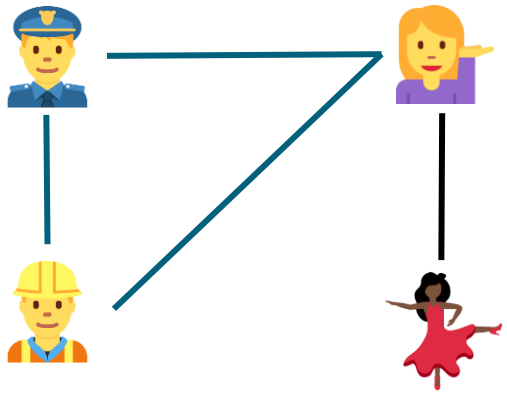
  - A search sequence



Graph (Network) Data

Data objects: an online social network, the Internet, the Web

Attribute: nodes and links



Emojis - Twitter - https://twemoji.twitter.com/ - CC-BY-4.0

The Graph (Network) Representation

- Data representation: $G = (V, E)$

  vertices edges

- **V** is a set of nodes (vertices, entities):

$$V = \{v_1, v_2, ..., v_n\}$$

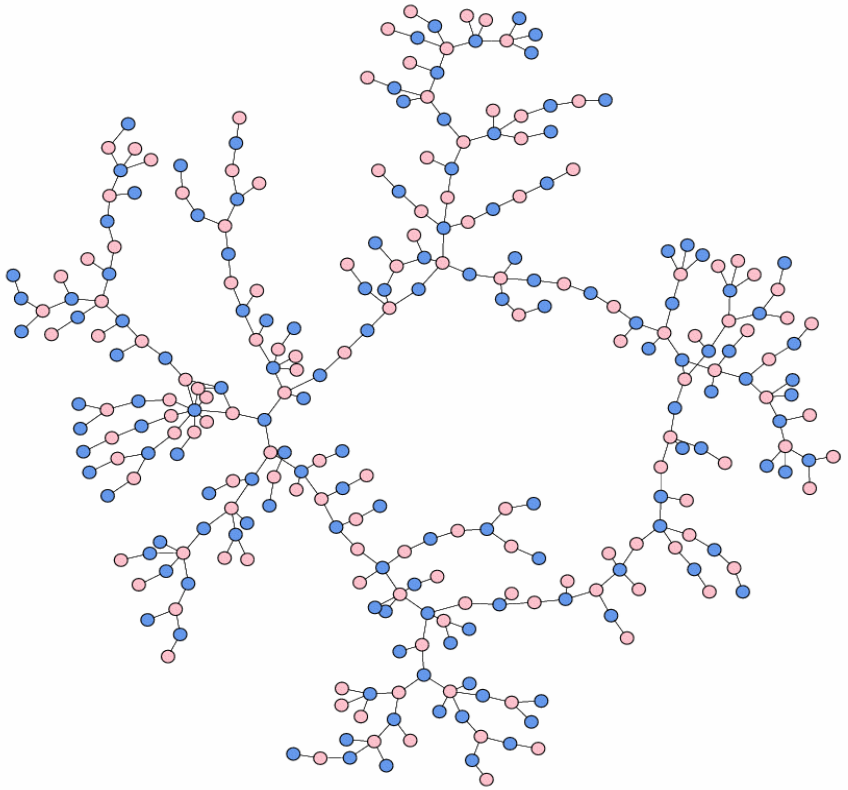
  - A node can be a categorical item or a complex data object.

- **E** is a set of links (edges, relations) between two nodes:

$$E = \{(v_i, v_j), ...\}$$

Examples of Networks

High school dating network



Data Source: Bearman, Moody, and Stovel 2004
Image by Mark N. Newman.

Case Study: How to Represent Text Data?

“To be or not to be”

S2&4

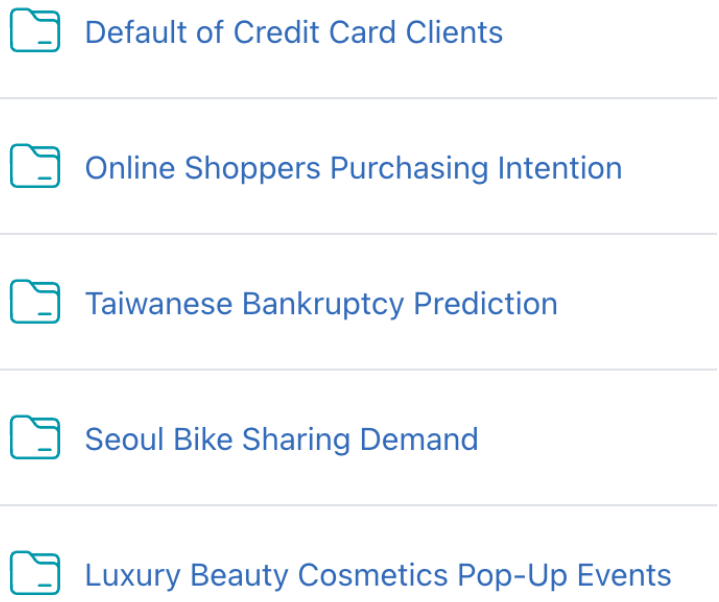
Week	Date	Topic	References
1	6 Sep 2025	Introduction to Big Data Analytics	[PF1-2] [HKP1-2][BM]
2	13 Sep 2025	Python Programming Basics	[WM] [JV2-3]
3	20 Sep 2025	Predictive Modelling: Decision Tree Python Programming – Decision Tree **Assignment 1**	[PF3]
4	27 Oct 2025	Predictive Modelling: Linear Regression Python Programming – Linear Regression **In-class quiz 1**	[MG2.3] [JG14-15] [WM12]
5	4 Oct 2025	Fitting a Model to Data: Objective Functions Python Programming – Logistic Regression **Assignment 2**	[PF4]
6	11 Oct 2025	Overfitting and Its Avoidance Python Programming – SVM & Regularization	[PF5] [MG5.1- 5.2]
7	18 Oct 2025	Midterm Test	
8	25 Oct 2025	Similarity, Neighbors and Clusters Python Programming – KNN & KMeans	[PF6]
9	1 Nov 2025	Data-analytic Thinking: Model Evaluation Python Programming – Metrics **Assignment 3**	[PF7] [MG5.3]
10	8 Nov 2025	Visualizing Model Performance Python Programming – Curves **In-class quiz 2**	[PF8] [MG5.3]
11	15 Nov 2025	Evidence and Probabilities: Naive Bayes Python Programming – Naïve Bayes	[PF9]
12	22 Nov 2025	Association Rules and Itemset Mining Python Programming - Apriori algorithm **In-class quiz 3**	[PF12] [EMC5]
13	29 Nov 2025	Group Project Presentation	

Course Assessment

Class Participation	15%	3 in-class quizzes
Written Assignments	30%	3 assignments
Midterm Test	20%	3-hour test
Group Project & Presentation	15%	18 minutes each group
Final Examination	20%	3-hour exam

Group Project

- **Week 3**: form a project team.
- **Week 5**: choose a dataset from the listed datasets. You are always welcome to use your own dataset.
- **Week 13**: Each group will have 18 minutes (incl. Q &A) to present your analysis and results.



A List of Open-data Providers

Open data site	Description
https://data.gov.hk/	Open data from Hong Kong government
https://data.gov/	The home of the US Government's open data
https://data.europa.eu/	The home of the European Commission's open data
https://data.un.org/	Various open data from the United Data
https://data.worldbank.org/	Open data initiative from the World Bank