1814ict/2814ict/7003ict:
Data Management/
Database Design

Topic 5.1: Data analysis and visualization

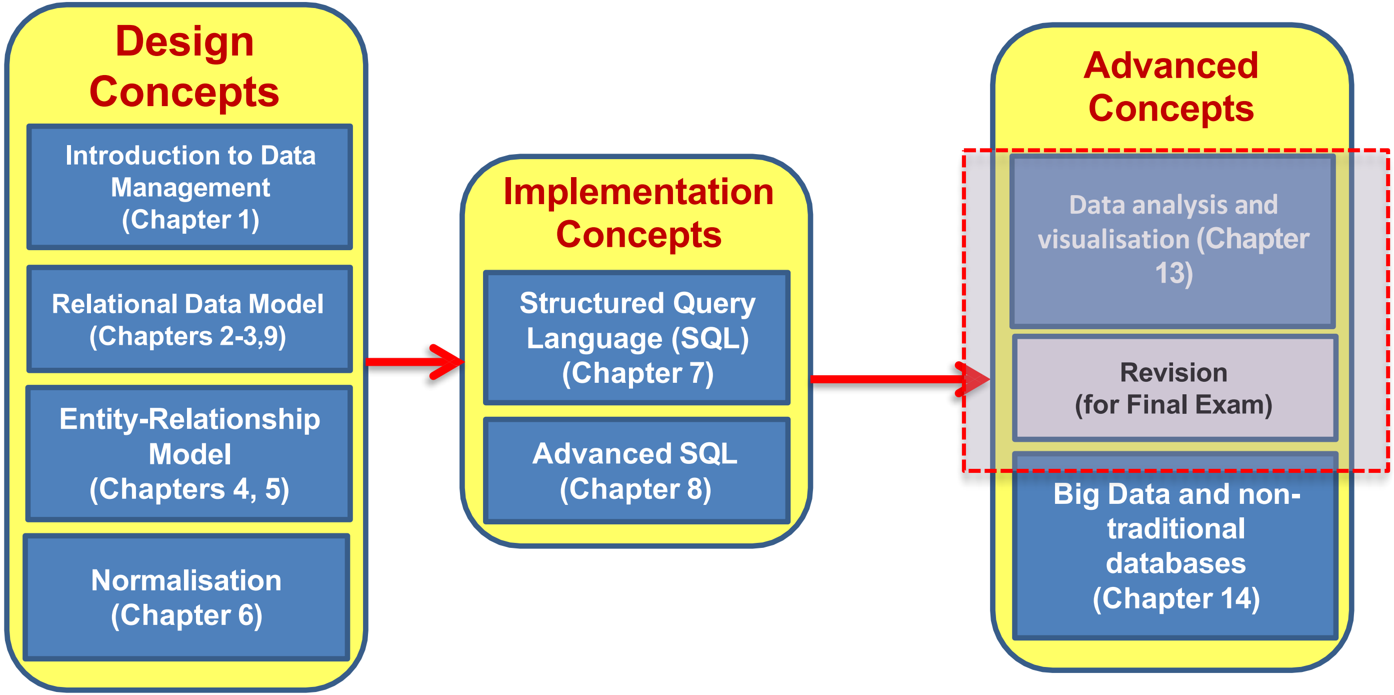
Course convenor: AProf. Henry Nguyen

School of Information and Communication Technology

Course developed by: Dr Mohammad Awrangjeb; AProf John Wang; Dr Zhe Wang

Course bigger picture

- Chapter references are to textbook *Database Systems: Design, Implementation, & Management - By Carlos Coronel and Steven Morris*


Design Concepts
Introduction to Data Management (Chapter 1)
Relational Data Model (Chapters 2-3,9)
Entity-Relationship Model (Chapters 4, 5)
Normalisation (Chapter 6)
Implementation Concepts
Structured Query Language (SQL) (Chapter 7)
Advanced SQL (Chapter 8)
Advanced Concepts
Data analysis and visualisation (Chapter 13)
Revision (for Final Exam)
Big Data and non-traditional databases (Chapter 14)

Learning Outcomes

At the end of this lecture students will be able to:

- Understand the importance of data analysis in business intelligence.
- Know the basics of business intelligence, incl. data warehouse, pre-processing, analytics, and visualisation.

Content

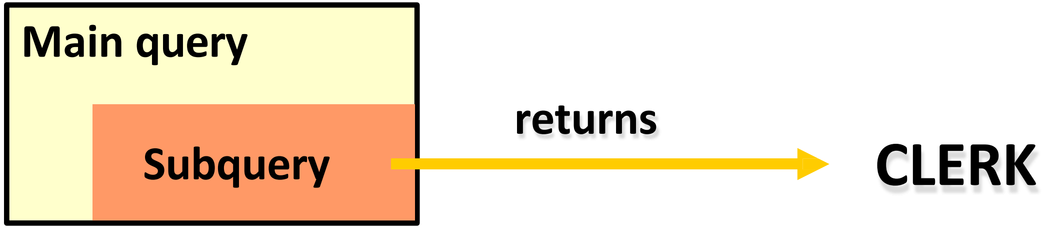
- Data analysis and business intelligence
- Data warehouse
- Data pre-processing: reduction, sampling and clustering
- Data analytics
- Data mining
- Data visualisation

Outcomes 1 to 2

Recap from Topic 4.4

Subquery types

- **Single-row subquery (a single value)**



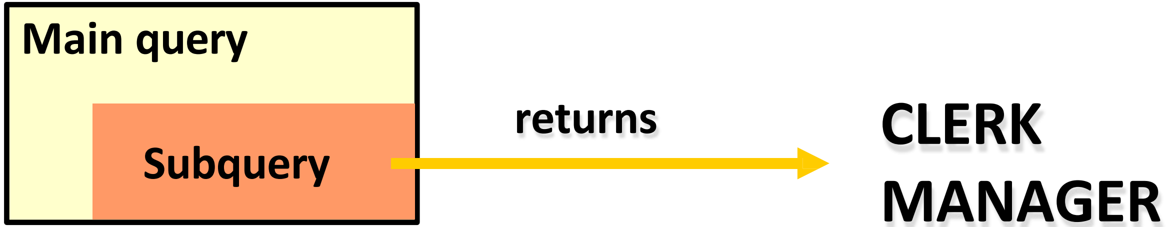
Main query

Subquery

returns

CLERK

- **Multiple-row subquery (a list of values – many rows, one column)**



Main query

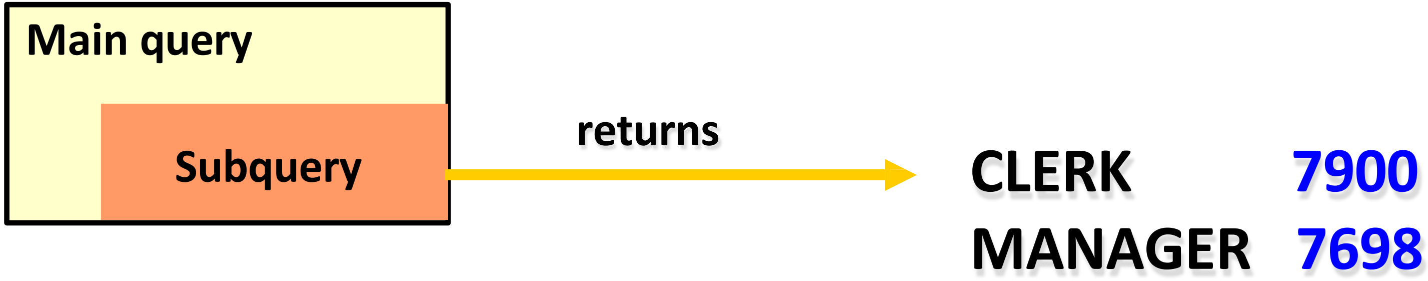
Subquery

returns

CLERK

MANAGER

- **Multiple-column subquery (a virtual table – many rows, many columns)**



Main query

Subquery

returns

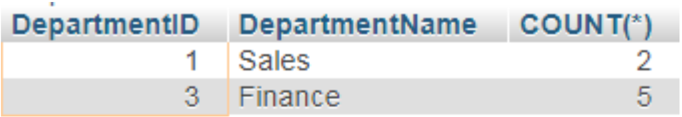
CLERK 7900

MANAGER 7698

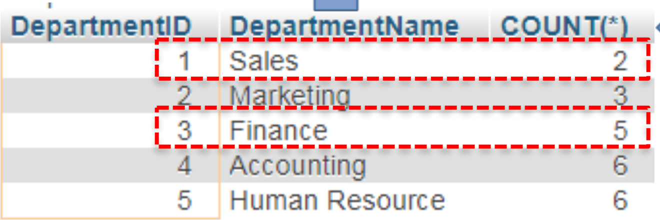
Multiple-Row subquery

- Find the number of staff working in Sales or Finance department.

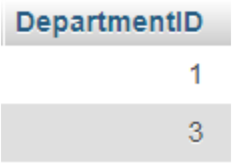
```
SELECT Dp.DepartmentID, Dp.DepartmentName, COUNT(*)
FROM workallocation AS WA, department AS Dp
WHERE WA.DepartmentID = Dp.DepartmentID
GROUP BY WA.DepartmentID
HAVING WA.DepartmentID = ANY(SELECT D.DepartmentID
                            FROM department AS D
                            WHERE D.DepartmentName IN ('Sales','Finance'));
```



DepartmentID	DepartmentName	COUNT(*)
1	Sales	2
3	Finance	5

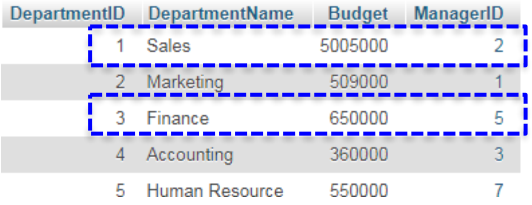


DepartmentID	DepartmentName	COUNT(*)
1	Sales	2
2	Marketing	3
3	Finance	5
4	Accounting	6
5	Human Resource	6



DepartmentID
1
3

Department table



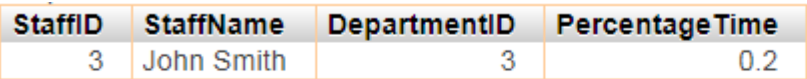
DepartmentID	DepartmentName	Budget	ManagerID
1	Sales	5005000	2
2	Marketing	509000	1
3	Finance	650000	5
4	Accounting	360000	3
5	Human Resource	550000	7

Multiple-Column Subquery

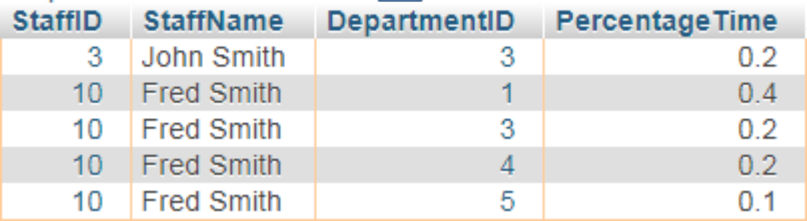
- The number of columns in the main query must match the number of columns returned from the inner query

- Find the staff who work in the same department as Fred Smith and work the same fraction.

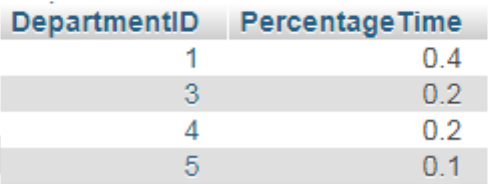
```
SELECT St.StaffID, St.StaffName, Wa.DepartmentID, Wa.PercentageTime
FROM Staff AS St, workallocation AS Wa
WHERE St.StaffID = Wa.StaffID
        AND St.StaffName <> 'Fred Smith'
        AND (Wa.DepartmentID, WA.PercentageTime)
                                            = ANY (SELECT W.DepartmentID, W.PercentageTime
                                                  FROM Staff AS S, workallocation AS W
                                                  WHERE S.StaffID = W.StaffID
                                                  AND S.StaffName = 'Fred Smith');
```



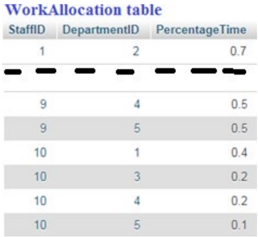
StaffID	StaffName	DepartmentID	PercentageTime
3	John Smith	3	0.2



StaffID	StaffName	DepartmentID	PercentageTime
3	John Smith	3	0.2
10	Fred Smith	1	0.4
10	Fred Smith	3	0.2
10	Fred Smith	4	0.2
10	Fred Smith	5	0.1



DepartmentID	PercentageTime
1	0.4
3	0.2
4	0.2
5	0.1



WorkAllocation table

StaffID	DepartmentID	PercentageTime
1	2	0.7
9	4	0.5
9	5	0.5
10	1	0.4
10	3	0.2
10	4	0.2
10	5	0.1

Business Intelligence

Why Data Analysis?

- **Business decision making:** To have the right data at the right time to support the business decision-making process.

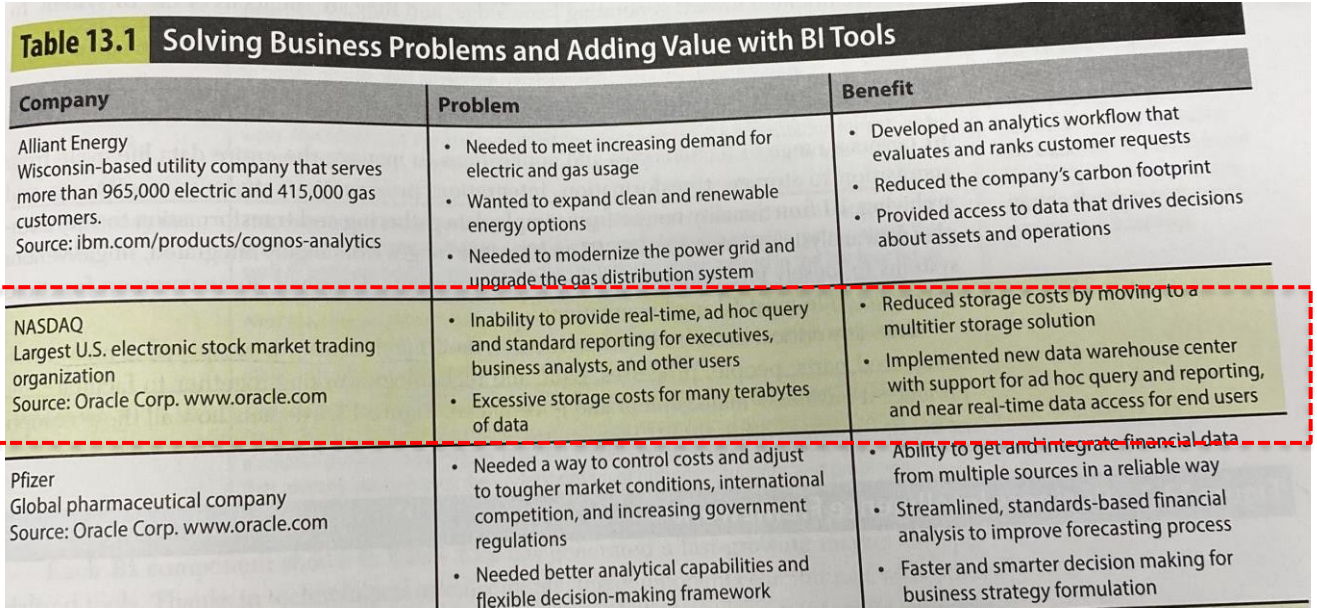
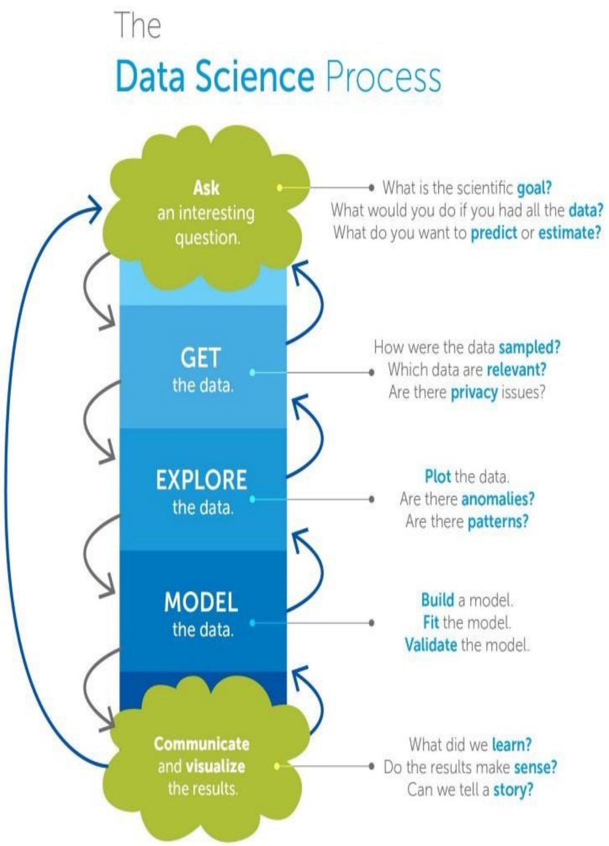


Table 13.1 Solving Business Problems and Adding Value with BI Tools

Company	Problem	Benefit
Alliant Energy Wisconsin-based utility company that serves more than 965,000 electric and 415,000 gas customers. Source: ibm.com/products/cognos-analytics	• Needed to meet increasing demand for electric and gas usage • Wanted to expand clean and renewable energy options • Needed to modernize the power grid and upgrade the gas distribution system	• Developed an analytics workflow that evaluates and ranks customer requests • Reduced the company's carbon footprint • Provided access to data that drives decisions about assets and operations
NASDAQ Largest U.S. electronic stock market trading organization Source: Oracle Corp. www.oracle.com	• Inability to provide real-time, ad hoc query and standard reporting for executives, business analysts, and other users • Excessive storage costs for many terabytes of data	• Reduced storage costs by moving to a multitier storage solution • Implemented new data warehouse center with support for ad hoc query and reporting, and near real-time data access for end users
Pfizer Global pharmaceutical company Source: Oracle Corp. www.oracle.com	• Needed a way to control costs and adjust to tougher market conditions, international competition, and increasing government regulations • Needed better analytical capabilities and flexible decision-making framework	• Ability to get and integrate financial data from multiple sources in a reliable way • Streamlined, standards-based financial analysis to improve forecasting process • Faster and smarter decision making for business strategy formulation

Data Science

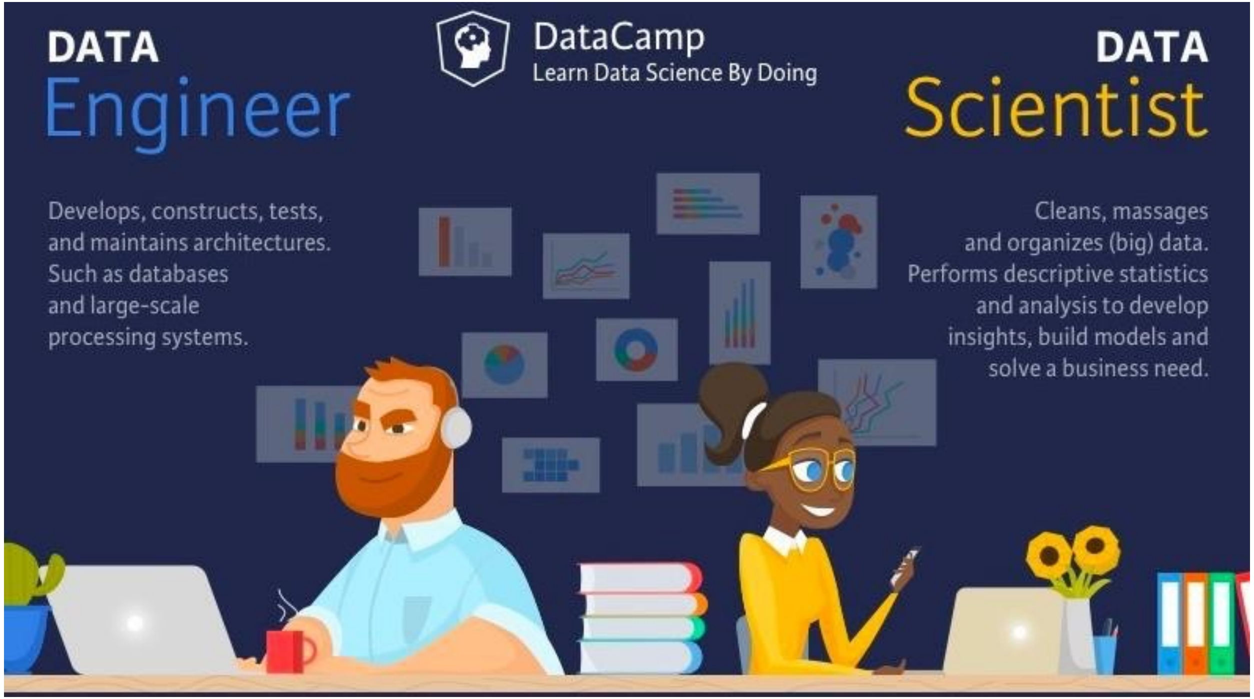
- **Data Science:** The study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of
  - Mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.
  - This analysis helps data scientists to ask and answer questions like
    - What happened,
    - Why it happened,
    - What will happen, and
    - What can be done with the results. (Source: Amazon)


The
Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
GET the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
EXPLORE the data.
Plot the data.
Are there anomalies?
Are there patterns?
MODEL the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course http://cs109.org/.

Data Engineering

- **Data Engineering:** Data engineering is the practice of designing and building systems for collecting, storing, and analyzing data at scale. (Source: coursera.org)


DATA
Engineer
DataCamp
Learn Data Science By Doing
DATA
Scientist
Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.
Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

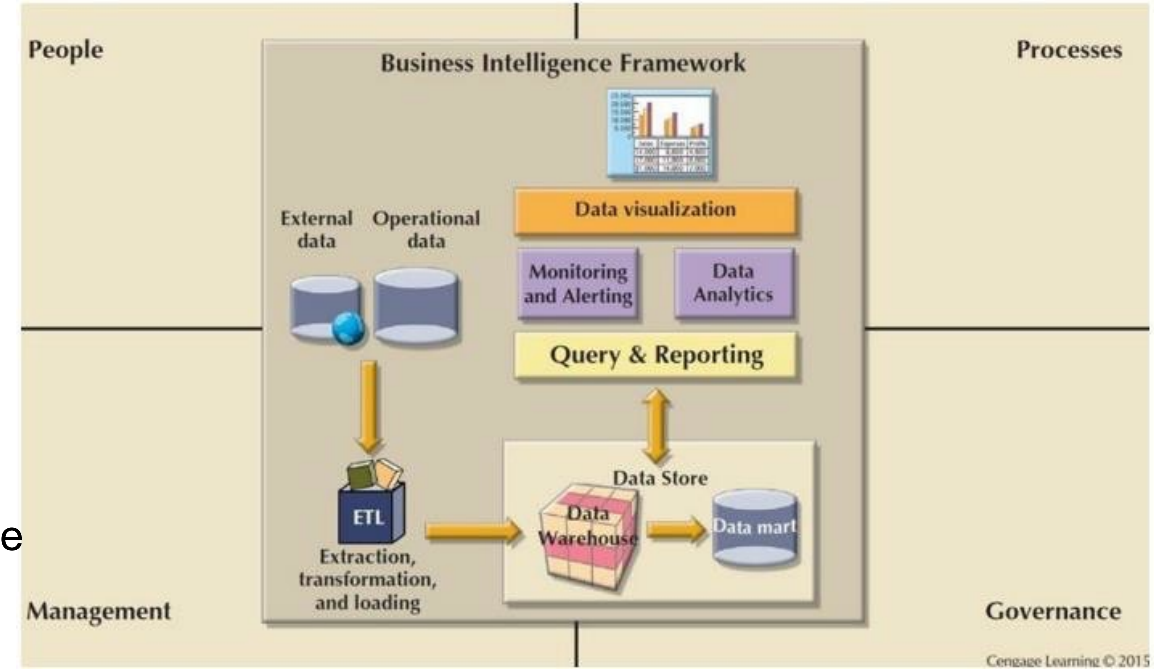
Sources: https://www.datacamp.com/blog/data-scientist-vs-data-engineer

Business Intelligence (BI)

- **Business intelligence:** A comprehensive, cohesive and integrated set of tools and processes used to capture, collect, integrate, store and analyse data with the purpose of generating and processing information to support business decision making.

- **BI framework:**

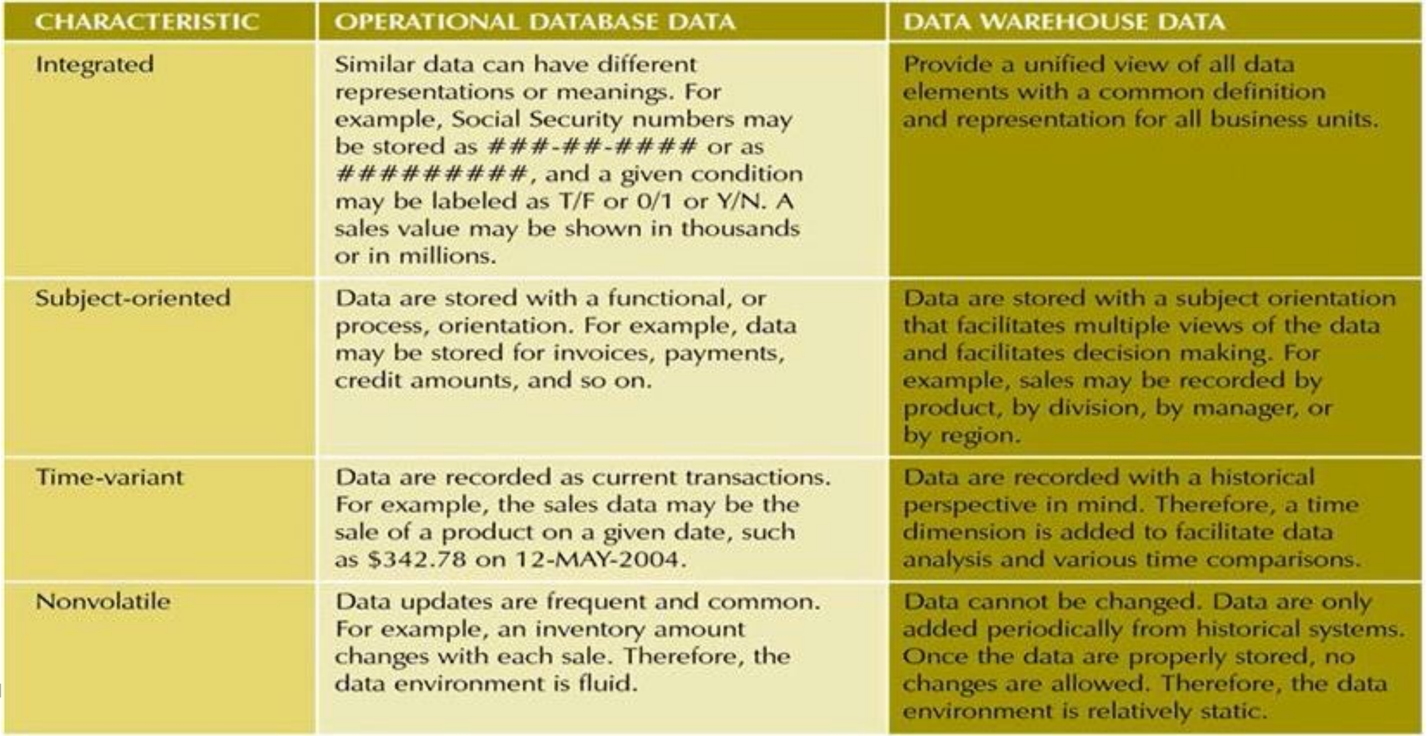
- **BI benefits:**
  - Integrated architecture
  - Common user interface
  - Common data repository
  - Improved business performance



Cengage Learning © 2015

Data Warehouse

- **Data warehouse:** An integrated, subject oriented, time-variant, and non-volatile collection of data that provide support to business decision marking.
- **Table 13.8:** Characteristics of data warehouse data and operational database data



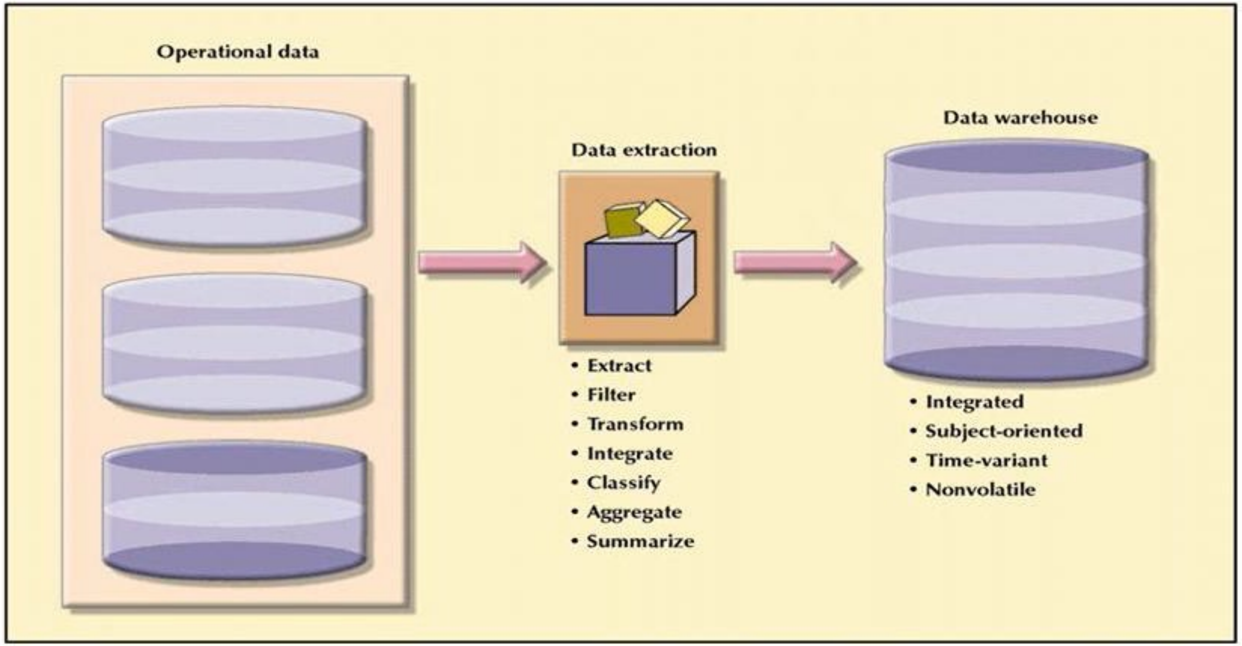
CHARACTERISTIC	OPERATIONAL DATABASE DATA	DATA WAREHOUSE DATA
Integrated	Similar data can have different representations or meanings. For example, Social Security numbers may be stored as ###-##-#### or as #########, and a given condition may be labeled as T/F or 0/1 or Y/N. A sales value may be shown in thousands or in millions.	Provide a unified view of all data elements with a common definition and representation for all business units.
Subject-oriented	Data are stored with a functional, or process, orientation. For example, data may be stored for invoices, payments, credit amounts, and so on.	Data are stored with a subject orientation that facilitates multiple views of the data and facilitates decision making. For example, sales may be recorded by product, by division, by manager, or by region.
Time-variant	Data are recorded as current transactions. For example, the sales data may be the sale of a product on a given date, such as $342.78 on 12-MAY-2004.	Data are recorded with a historical perspective in mind. Therefore, a time dimension is added to facilitate data analysis and various time comparisons.
Nonvolatile	Data updates are frequent and common. For example, an inventory amount changes with each sale. Therefore, the data environment is fluid.	Data cannot be changed. Data are only added periodically from historical systems. Once the data are properly stored, no changes are allowed. Therefore, the data environment is relatively static.

Data Warehouse

- **The ETL process creating a data warehouse:**

Creating a Data Warehouse

FIGURE 12.3 CREATING A DATA WAREHOUSE



Data Pre-processing – major tasks:

- **Data extraction:** Get data from multiple, heterogeneous, and external sources
- **Data cleaning:** Detect errors in the data and rectify them when possible

Data reduction:

- Dimensionality reduction
- Data compression, sampling, and clustering

- **Data transformation:** Convert data from legacy or host format to warehouse format
- **Load:** Sort, summarize, consolidate, compute views, check integrity, and build index and partitions

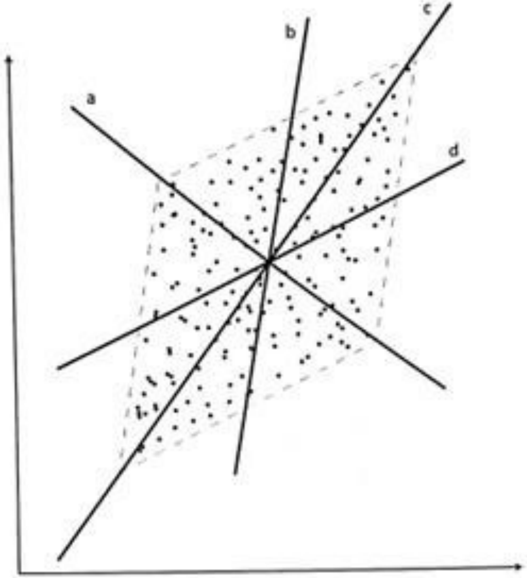
Refresh: propagate the updates from the data sources to the warehouse

Data Reduction - Example

- **Why data reduction?** —A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- **Principal Component Analysis (PCA):**
  - A popular linear dimensionality reduction technique
  - Data in real world is very high dimensional
  - We use PCA reduces the data to 2 dimensions
  - PCA finds orthogonal projections which are independent
  - The first PC is in the direction of maximum variance in the data and so on.

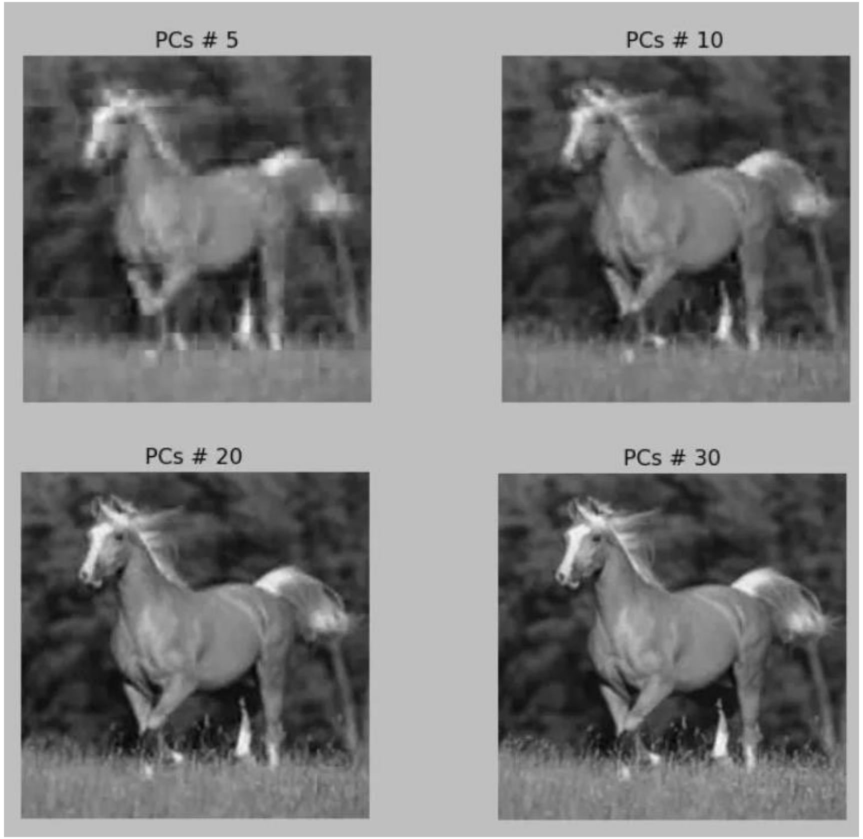
PCA

Project data to 2D dimensional space



Dimensionality Reduction - PCA

- **Principal Component Analysis (PCA):**



To reconstruct the image (high dimensional data from PCs):

- Take PCs 1st, 2nd, 3rd and transform back (reconstruct) the image.
- The more PCs you take, the higher quality, but have more redundancy.
- So, PCA can be used for data reduction (compression)

Source: https://medium.com/analytics-vidhya/principal-component-analysis-of-an-image-7e62105b2fa2

Data Reduction - Sampling

- **What is sampling?** Data sampling is a statistical analysis technique used to select, process, and analyse a representative subset of a population.

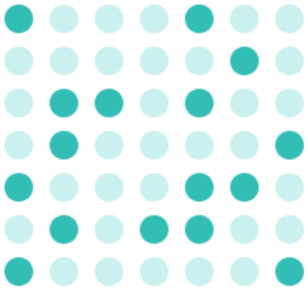
- **Population vs Sample:** A population is a complete set of elements, while a sample is a subset of a population.

Population
VS
Sample

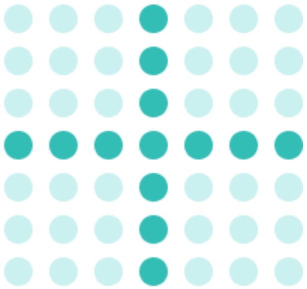
Source: https://www.egnyte.com/guides/life-sciences/data-sampling

Data Reduction - Sampling

- **Probability (random) vs. Non-Probability Sampling Methods:**

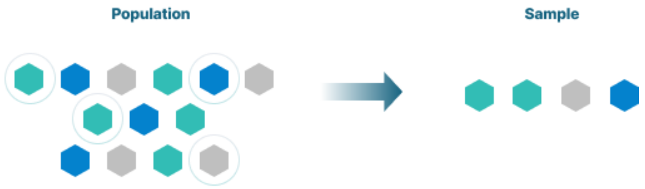


Probability Data Sampling



Non-Probability Data Sampling

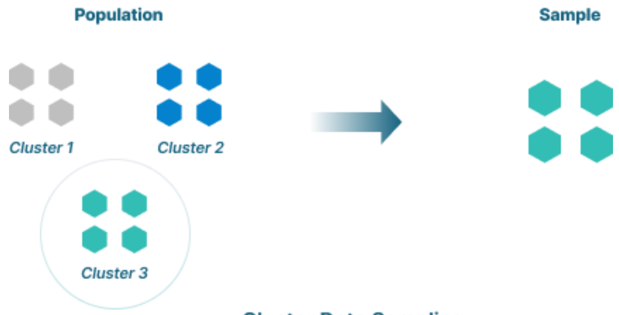
- **Simple Random Sampling**



Simple Random Data Sampling

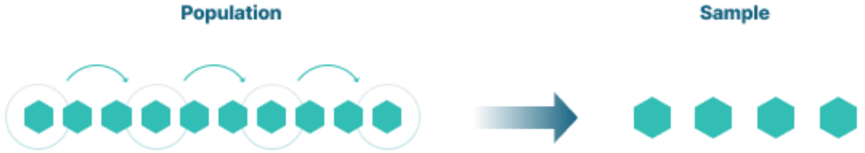
Data Reduction - Sampling

- **Cluster Sampling:**



Cluster Data Sampling

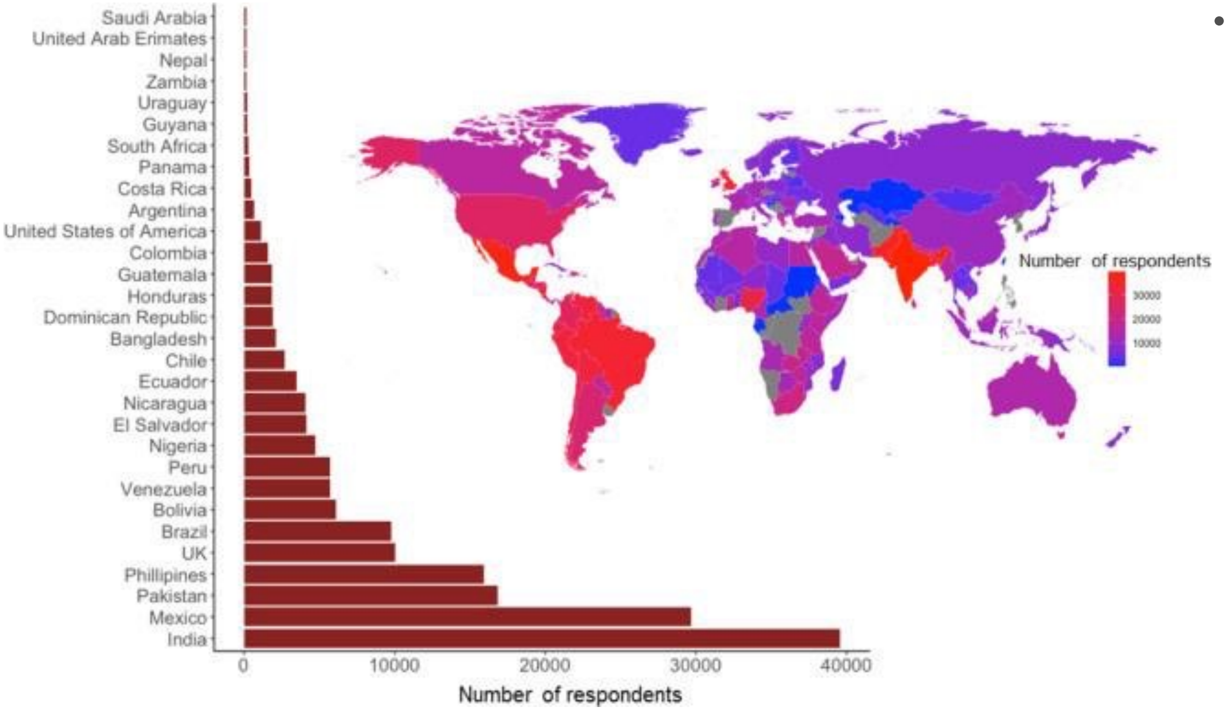
- **Systematic Sampling or Systematic Clustering:**



Systematic Data Sampling

Sampling - Example

- **Global response to COVID-19 symptom survey:**



- 175,566 individuals from 190 different countries responded to the survey.

Figure 1: The bar chart shows the respondent counts for the top-30 countries with highest number of respondents. The map shows the heatmap of the respondent counts for all countries.

Data Clustering

- **What is data clustering?** Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). [Source: Wikipedia]

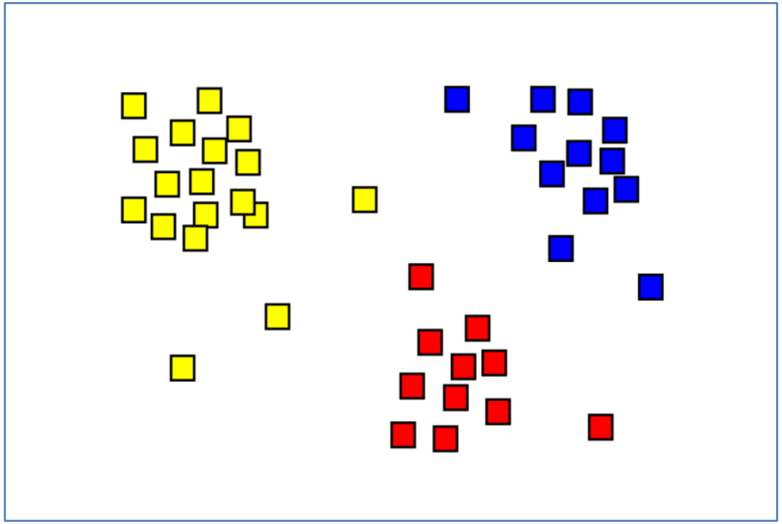


Figure: The result of a cluster analysis shown as the colouring of the squares into three clusters.

Source: https://en.wikipedia.org/wiki/Cluster_analysis

Clustering algorithm: *k*-means

- ***k*-means clustering:** When the number of clusters is fixed to *k*, *k*-means clustering gives a formal definition as an optimisation problem: find the *k* cluster centres and assign the objects to the nearest cluster centre, such that the squared distances from the cluster are minimized. [Source: Wikipedia]

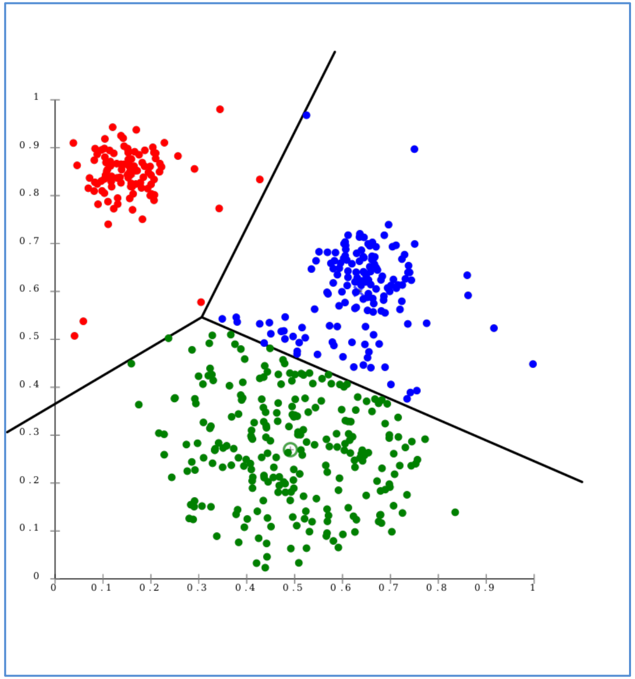


Figure: *k*-means separates data into Voronoi cells.

Source: Source: https://en.wikipedia.org/wiki/Cluster_analysis

Data Analytics

- **Data analytics:** A subset of BI functionality that encompasses a wide range of mathematical, statistical and modelling techniques with the purpose of extracting knowledge from data.
  - The business managers what really want from BI is “*the ability to extract actionable business inside from current events and foresee future problems or opportunities.*”
  - Continuous knowledge acquisition that goes from *discovery → explanation → prediction*

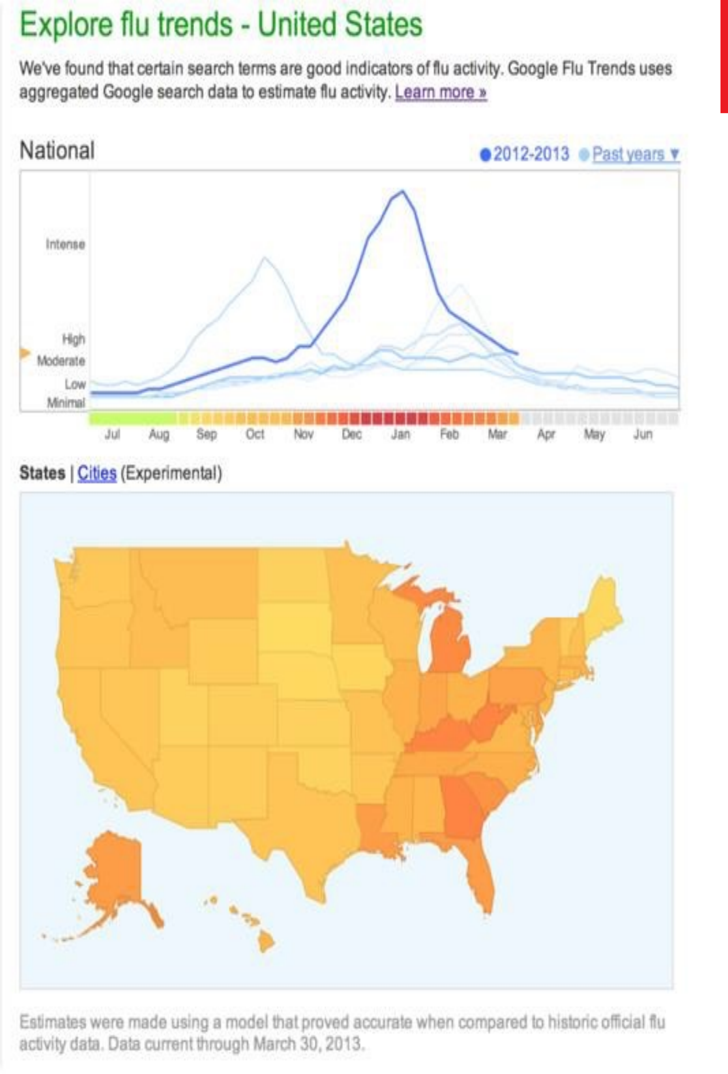
- Two separate areas of data analytics, often overlapping though
  - **Explanatory analytics:**
    - Provides ways to discover relationships, trends and patterns among data
  - **Prediction analytics:**
    - Uses advanced statistical and modelling techniques to predict future business outcome with great accuracy

Application: Flu Monitoring

Google Flu Trend: provide estimates of influenza activity for more than 25 countries

- ➢ Use the search queries related to flu on Google
- ➢ Users tend to search information for potential flu outbreaks
- ➢ Predict flu at a location based on the number of queries and their IP location

Source: https://en.wikipedia.org/wiki/Google_Flu_Trends



Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

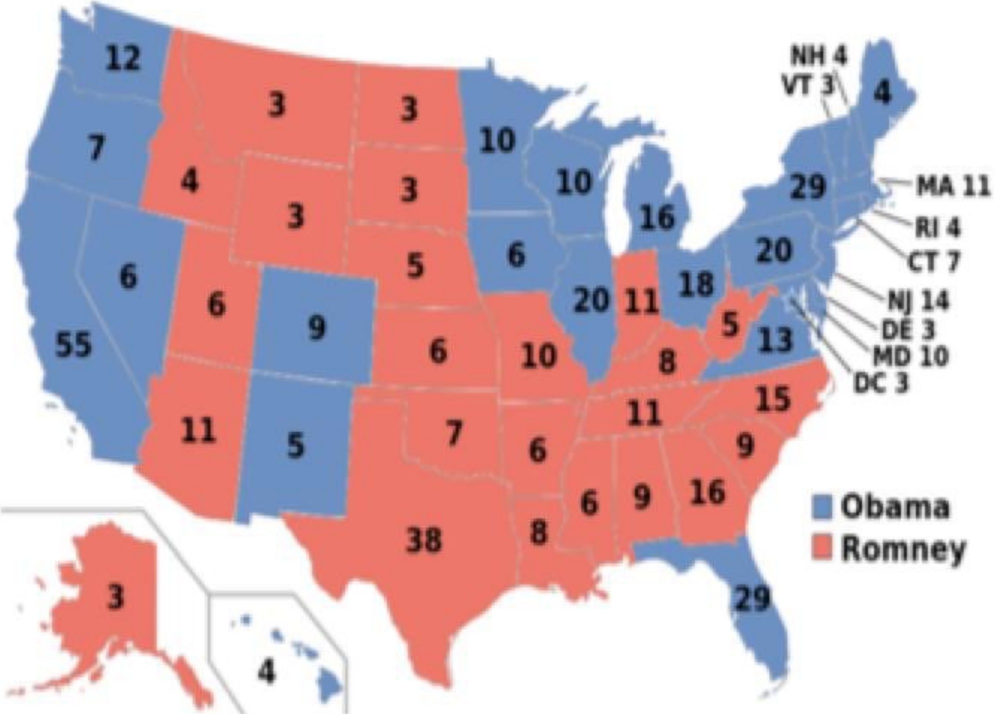
National

States | Cities (Experimental)

Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 30, 2013.

Application: Election Prediction

- Nate Silver made his name by using cold hard math (historical data) to predict elections correctly in 49 out of 50 states in the 2008 and all 50 states in 2012



Data Mining

- **Data mining:** A process that employs automated tools
  - To analyse data in a data warehouse or other sources and
  - To proactively identify possible relationships and anomalies.

- Knowledge discovery from hidden patterns

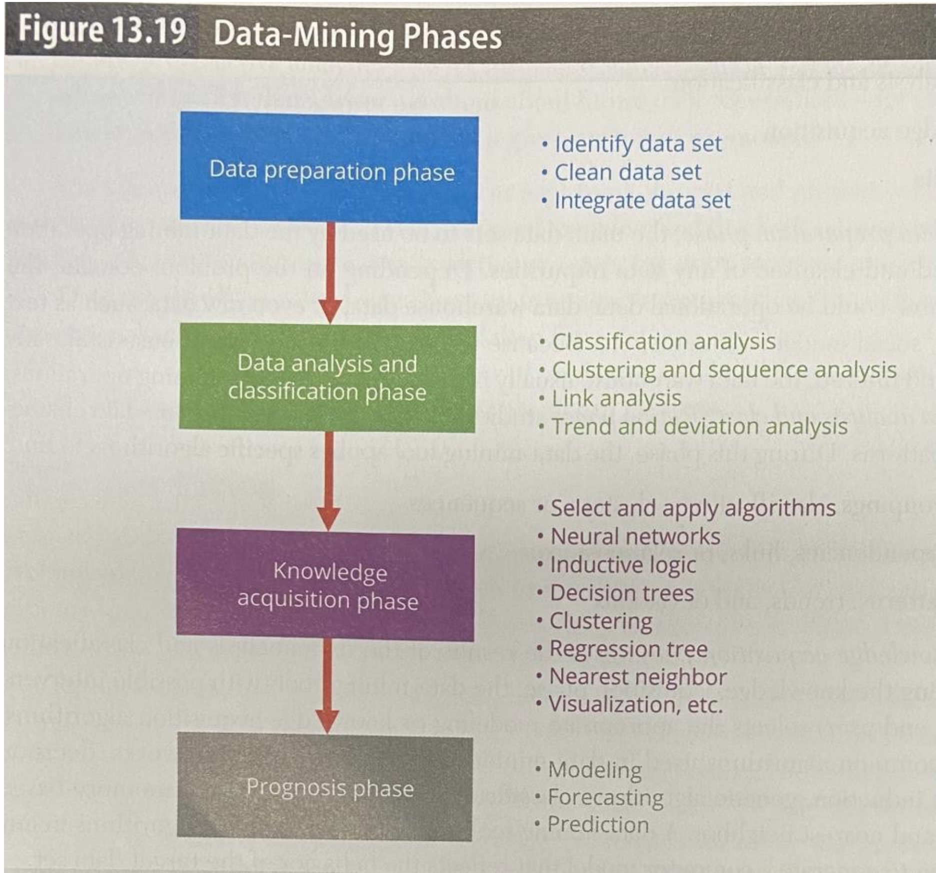


Figure 13.19 Data-Mining Phases

Data preparation phase
- Identify data set
- Clean data set
- Integrate data set

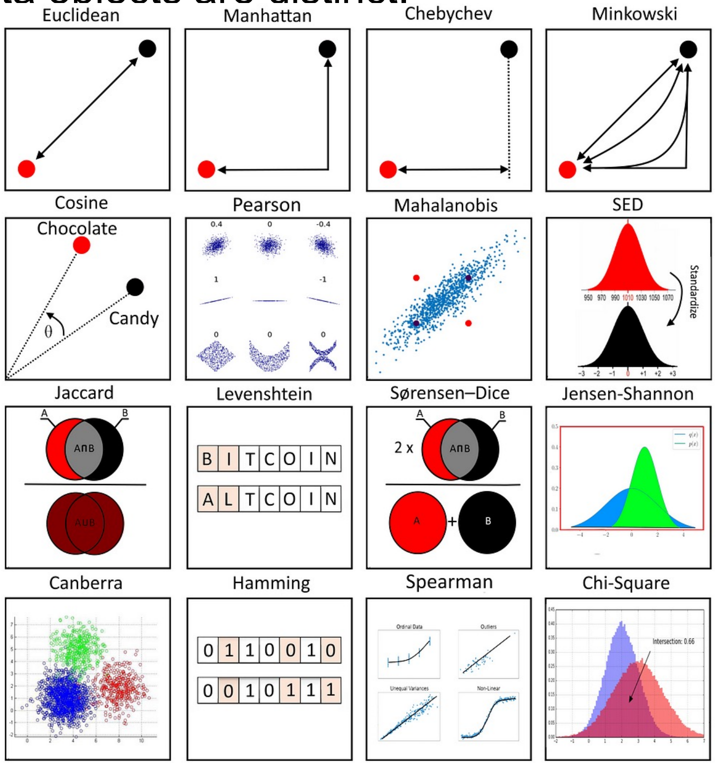
Data analysis and classification phase
- Classification analysis
- Clustering and sequence analysis
- Link analysis
- Trend and deviation analysis

Knowledge acquisition phase
- Select and apply algorithms
- Neural networks
- Inductive logic
- Decision trees
- Clustering
- Regression tree
- Nearest neighbor
- Visualization, etc.

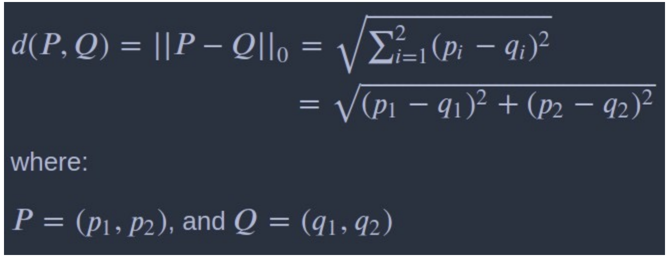
Prognosis phase
- Modeling
- Forecasting
- Prediction

Data Mining

- Data similarity vs dissimilarity:
  - Similarity measure is to measure how data samples are related or closed to each other.
  - Dissimilarity measure is to tell how much the data objects are distinct.



- The Euclidean distance:
  - Shortest distance between two data points.
  - i.e., The hypotenuse of a right-angle triangle.
  - In 2-dimensional data space



$$d(P,Q) = ||P - Q||_0 = \sqrt{\sum_{i=1}^{2} (p_i - q_i)^2}$$
$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

where:

$P = (p_1, p_2)$, and $Q = (q_1, q_2)$

Data Mining

- The Euclidean distance in machine learning:
  - In a dataset, we have 3 flower types: Iris-Setosa, Iris-Versicolor, and Iris-Virginica
  - And 4 features for each flower: sepal length, sepal width, petal length, petal width, so we have a 4-dimensional space
  - Let’s train two flower types in a 2-diminesional space.

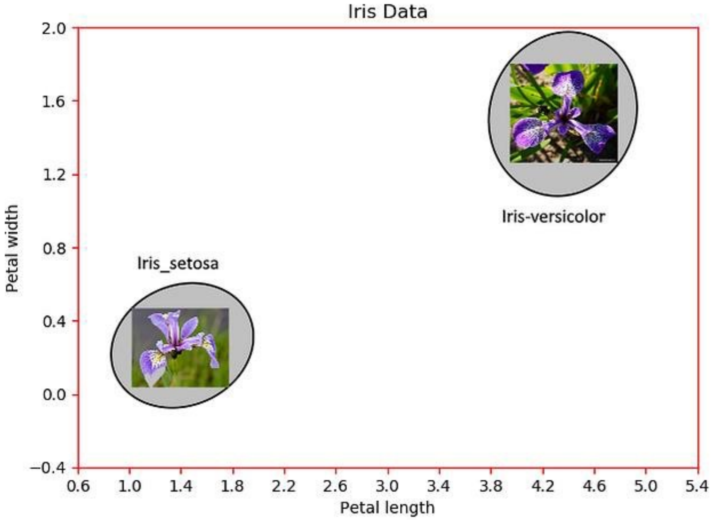


Figure: Iris dataset for two types of flowers in two features’ space.

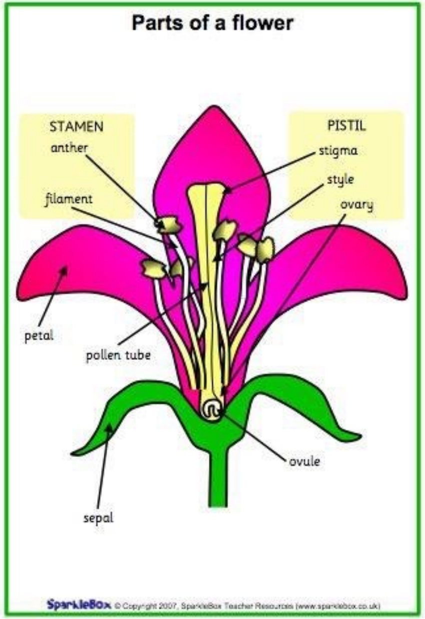


Figure: Indicating sepal and petal in a flower.

Data Mining

- The Euclidean distance in machine learning:

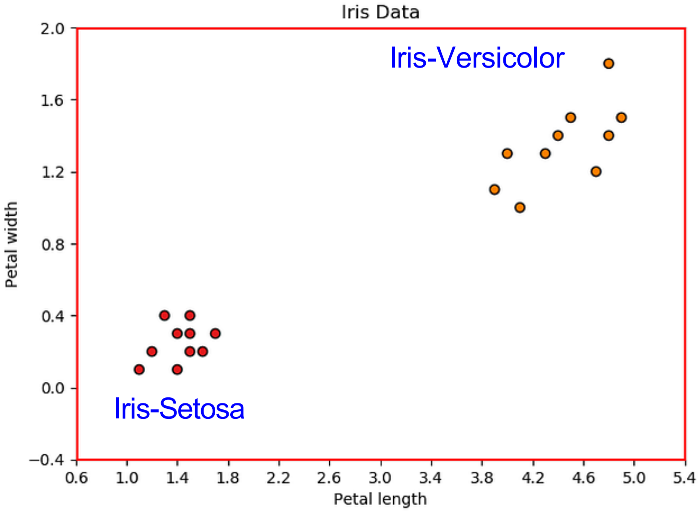


Figure: Training dataset.

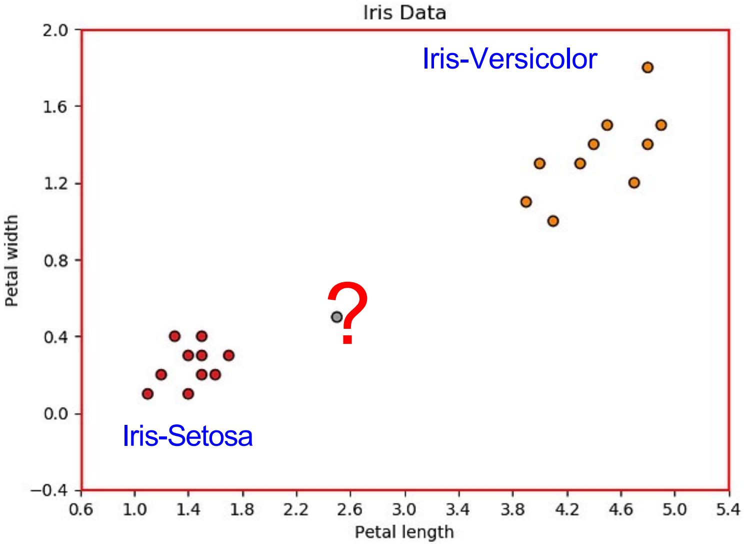
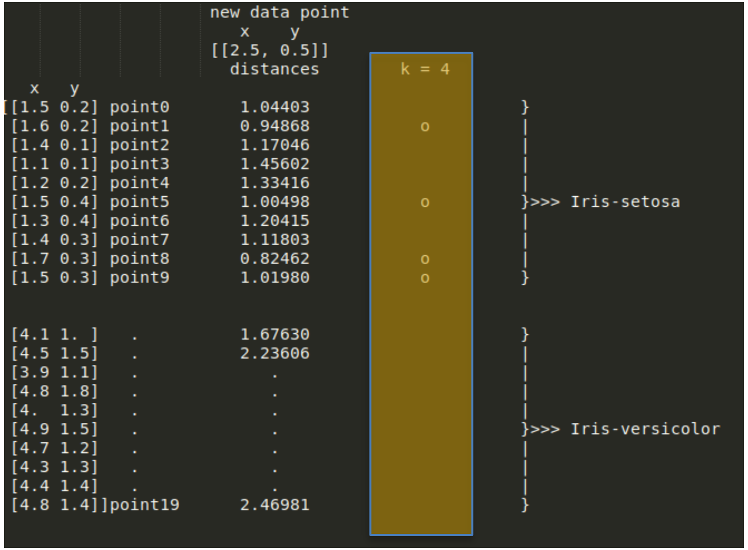


Figure: Predict the label for a new data point.

Data Mining

- The Euclidean distance in machine learning:



```
                          new data point
                              x     y
                          [[2.5, 0.5]]
                            distances           k = 4
  x    y
[[1.5 0.2] point0          1.04403                              }
 [1.6 0.2] point1          0.94868                o             |
 [1.4 0.1] point2          1.17046                              |
 [1.1 0.1] point3          1.45602                              |
 [1.2 0.2] point4          1.33416                              |
 [1.5 0.4] point5          1.00498                o             }>>> Iris-setosa
 [1.3 0.4] point6          1.20415                              |
 [1.4 0.3] point7          1.11803                              |
 [1.7 0.3] point8          0.82462                o             |
 [1.5 0.3] point9          1.01980                o             }


 [4.1 1. ]    .            1.67630                              }
 [4.5 1.5]    .            2.23606                              |
 [3.9 1.1]    .               .                                 |
 [4.8 1.8]    .               .                                 |
 [4.   1.3]   .               .                                 |
 [4.9 1.5]    .               .                                 }>>> Iris-versicolor
 [4.7 1.2]    .               .                                 |
 [4.3 1.3]    .               .                                 |
 [4.4 1.4]    .               .                                 |
 [4.8 1.4]]point19         2.46981                              }
```

Figure: Calculation.

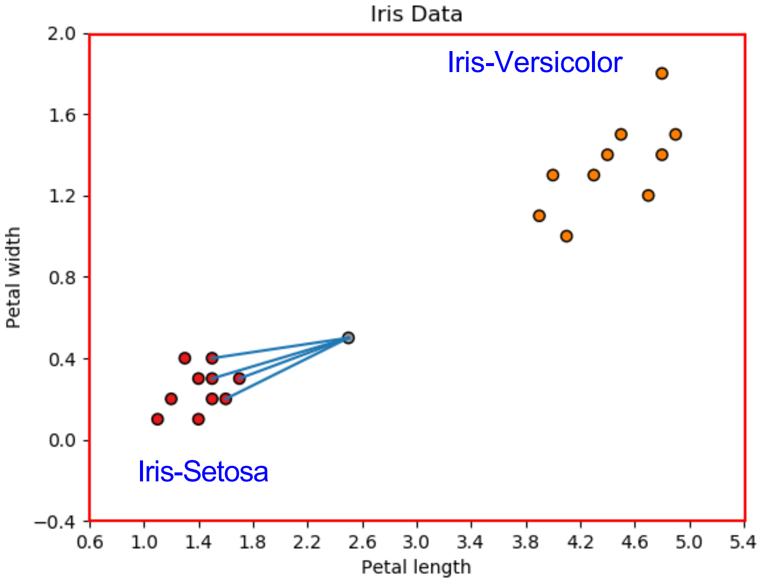


Figure: Four neighbours voted for Iris-Setosa.

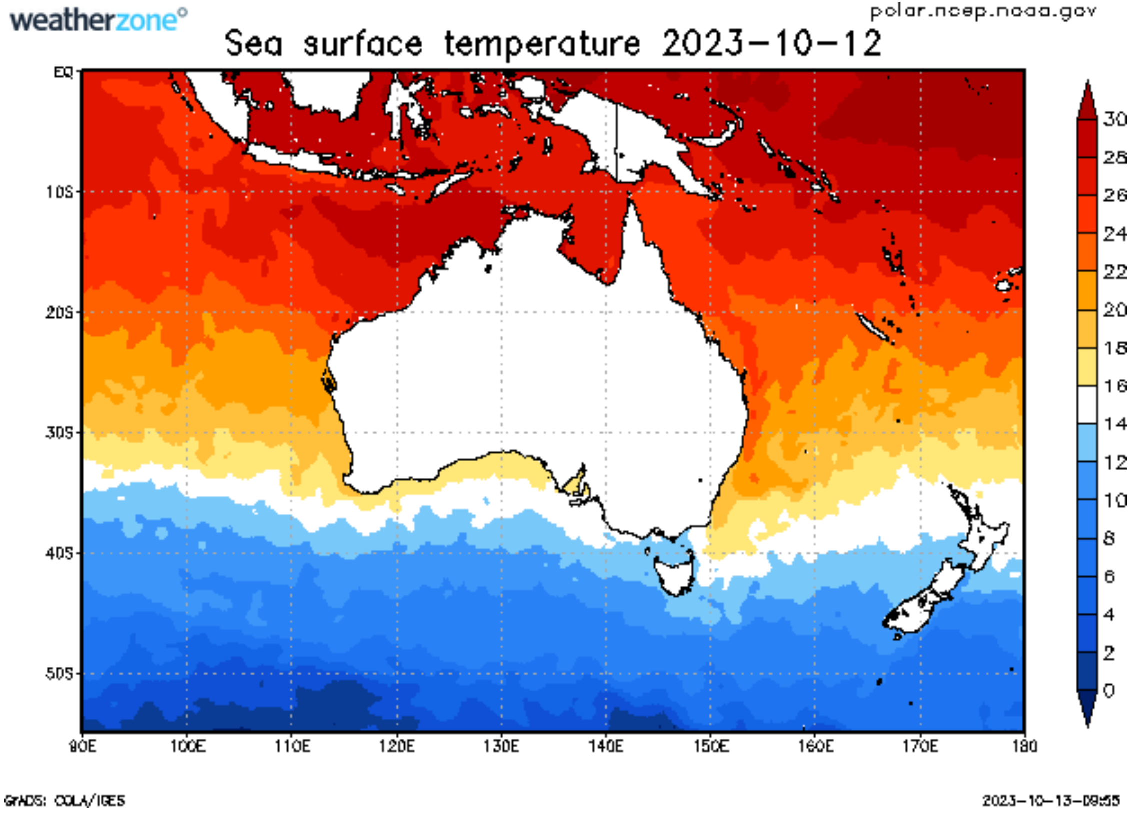
- Calculate the Euclidean distance from new flower to all flowers in two classes.
- Find the k-nearest neighbours based on the Euclidean distance to see to which class the new flower closer to.

Data Visualisation

- **Data visualisation:** Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationship among data items or attributes can be analyzed or reported.
- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well-developed ability to analyze a large amount of information that is presented visually
    - Can detect general patterns and trends
    - Can detect outliers and unusual patterns

Data Visualisation - Example

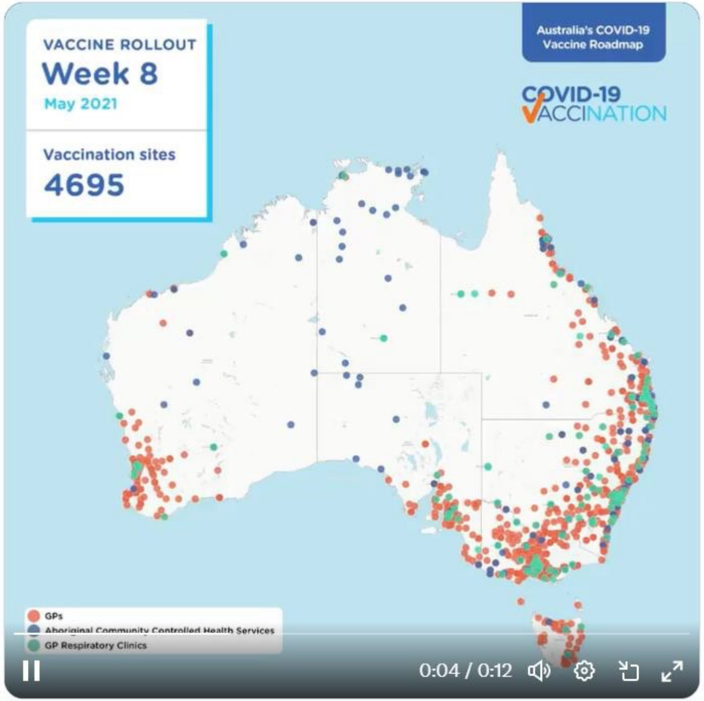
- **Contour plot:** Sea Surface Temperature around Australia



Source: https://www.farmonlineweather.com.au/climate/indicator_sst.jsp?c=sst

Data Visualisation - Example

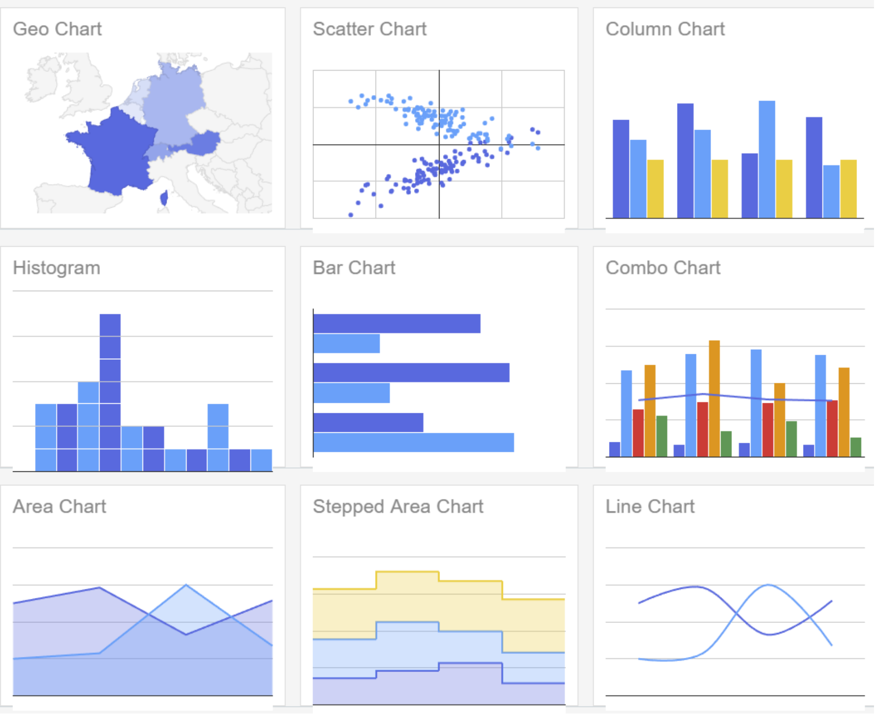
- Covid-19 vaccine rollout in 2021

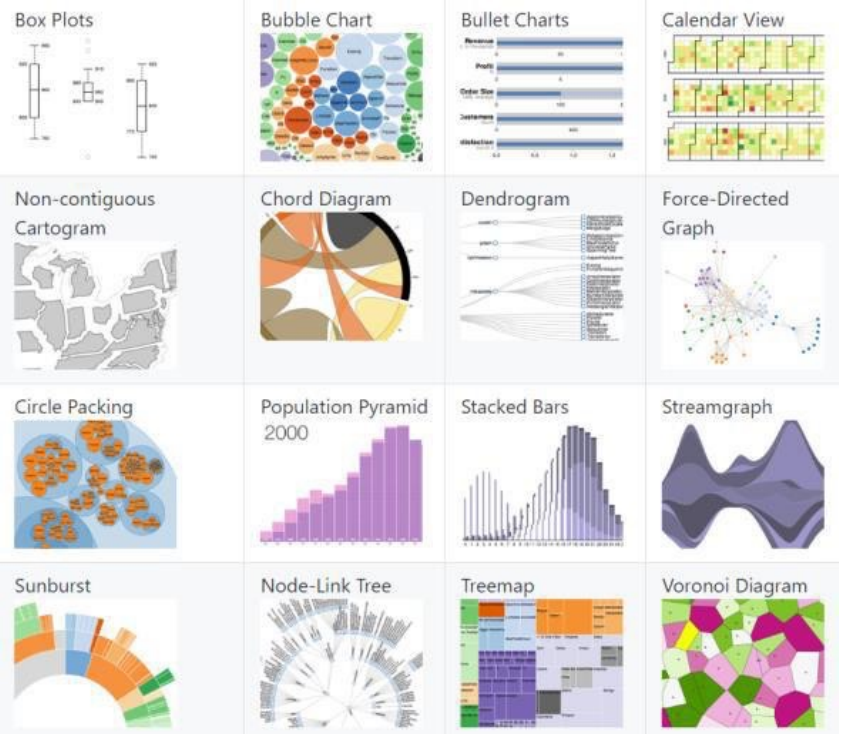


- Click to watch online: https://twitter.com/healthgovau/status/1428615312950300672

Data Visualisation - Example

- Other visualisation options:





Thank you