R：Statistical Programming Methods
R：程式、機率與統計

Exploratory Data Analysis (1)

Data Science

- Data and technology enabled platform drawing on knowledge of **computer science**, **engineering**, **business and mathematics**.
- A continuous loop of improving decision-making from business ideas, production, marketing, consumption and further enhancements.
- Data science focuses on capability development and data analytics development.
- GIGO (Garbage-in, garbage-out)
  - quality of input and methods of statistical inference determines the quality of output

Careers in Data Science

Data Analyst
資料分析師

- Domain knowledge is the KEY
- Interpret data to make decision
- Communicating the results
- Data Visualization

Data Scientist
資料科學家

- Statistical Analysis
- Exploring data and identify trends (data-driven)
- Programming, mathematics and statistics

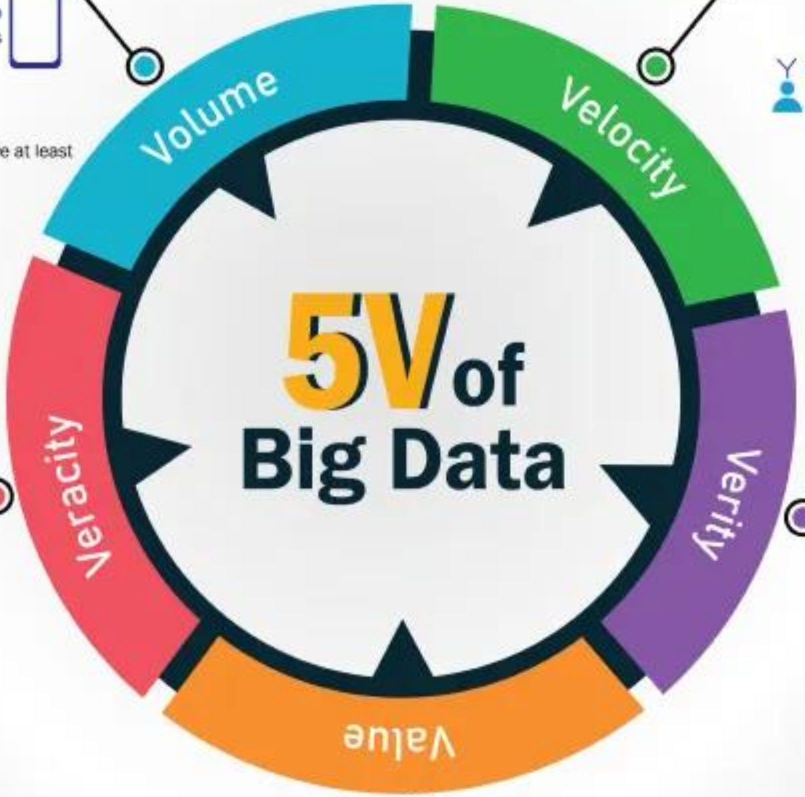
Data Engineer
資料工程師

- Data Acquisition
- Preparation of data for modeling
- Data warehousing

Big Data

- Data as strategic assets
- Big data: very large for traditional data processing systems, and therefore require new processing technologies, e.g., computing power.
- What examples of big data can you use for your dream job?

The Five V's of Big Data



Scale of Data

This refers to the sheer volume of data being generated every second.

6 Billion People have cell phones

40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005

Most companies in the U.S. have at least **100 Terabytes** of data stored.

The New York Stock Exchange capture **1 TB of Trade Information**

Analysis of Streaming Data

Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.

By 2016 it is projected there will be **18.9 Billion** network connections

Modern cars have close to **100 Sensors**

4 Billion+ hours of video are watched on You Tube each month

30 Billion pieces of content are shared on facebook every month

400 Million tweets are sent per day by about 200 million monthly active users

Diffrent forms of data

As more and more data is being digitized.

1 in 3 Business leaders don't trust the information they use to make decisions

Uncertainty Of Data

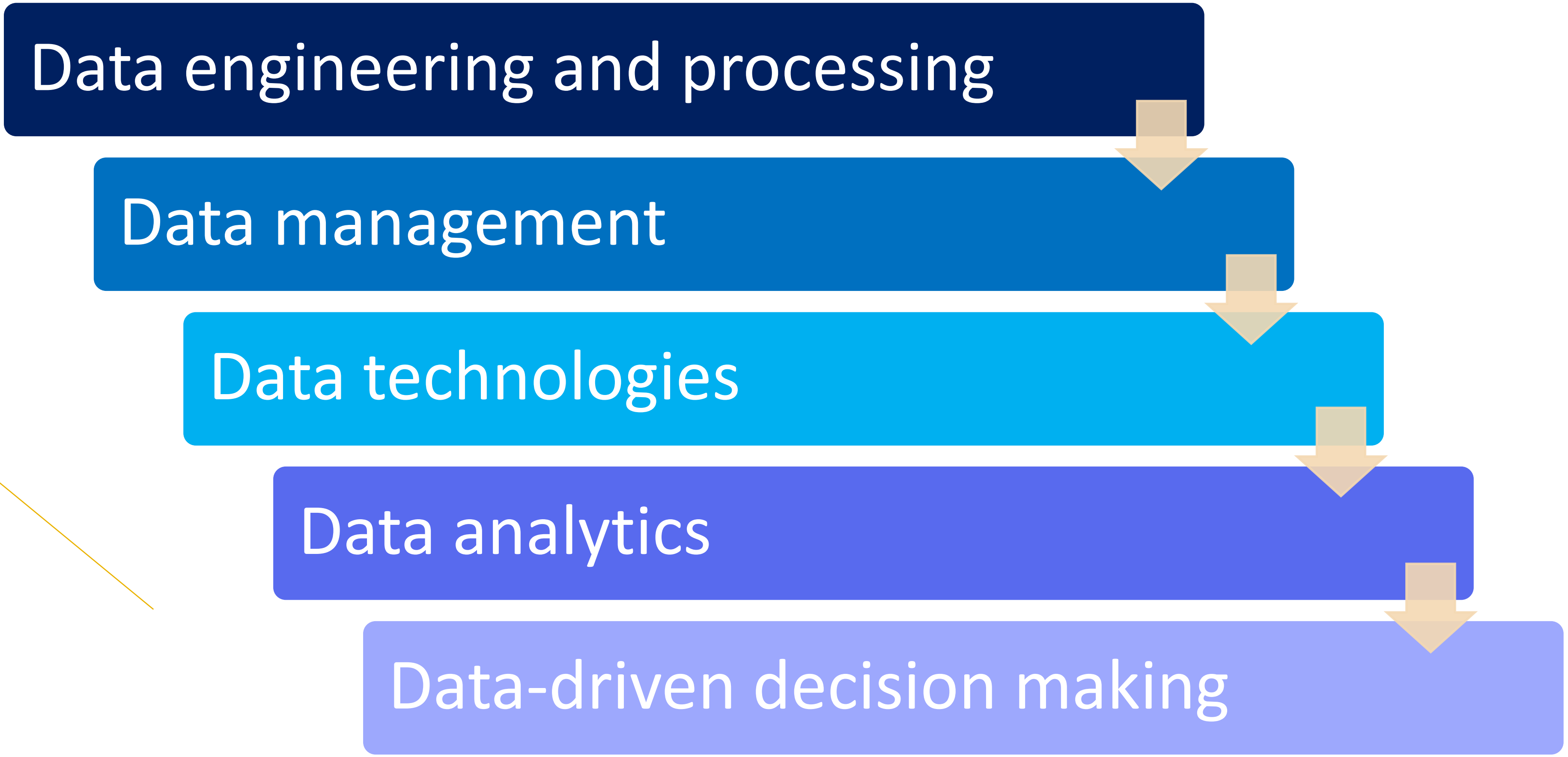
This refers to the discrepancies found in the data.

Poor data quality costa the US economy around **$ 3.1 Trillion a year**

Value Of Data

Having access to big data is all well and good but that's only useful if we can turn it into a value.

Making Data-driven decision



Data engineering and processing

Data management

Data technologies

Data analytics

Data-driven decision making

Data Analytics

Descriptive Analytics

- What’s happening right now?

Diagnostic Analytics

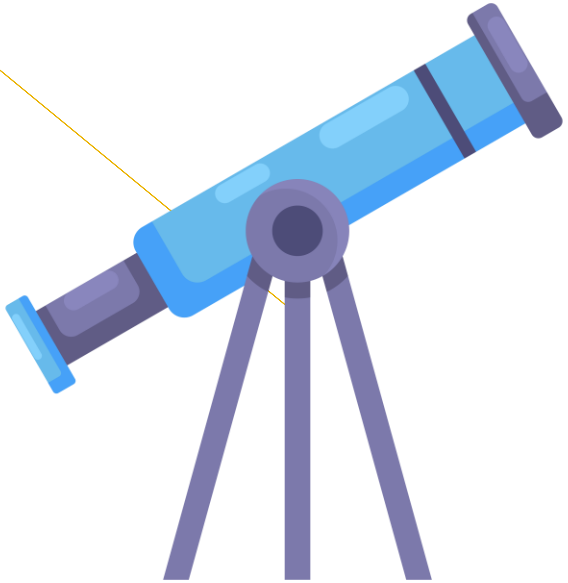
- Why did that happen?

Predictive Analytics

- What will happen next?

Prescriptive Analytics

- What should we do?



Structured Data 結構性資料

- Data comes from many sources: sensor measurements, events, text, images, and videos
- Unstructured raw data must be processed and manipulated into a structured form

  → Table with rows and columns

- Structured Data
  - Numeric: Continuous 連續 v.s. Discrete 離散
  - Categorical: Fixed set of values (e.g., Binary Data, Ordinal Data 順序)

- Data typing in software acts as a signal to the software on how to process the data.

Estimation of Location/Variability

- A basic step in exploring your data is getting a “typical value” for each feature (variable): an estimate of where most of the data is located (i.e., its central tendency).
    - Mean 平均值
    - Median 中位數
    - Percentile 百分比

- Variability, or dispersion measures whether the data values are tightly clustered or spread out.
    - Variance 變異數
    - Standard deviation 標準差