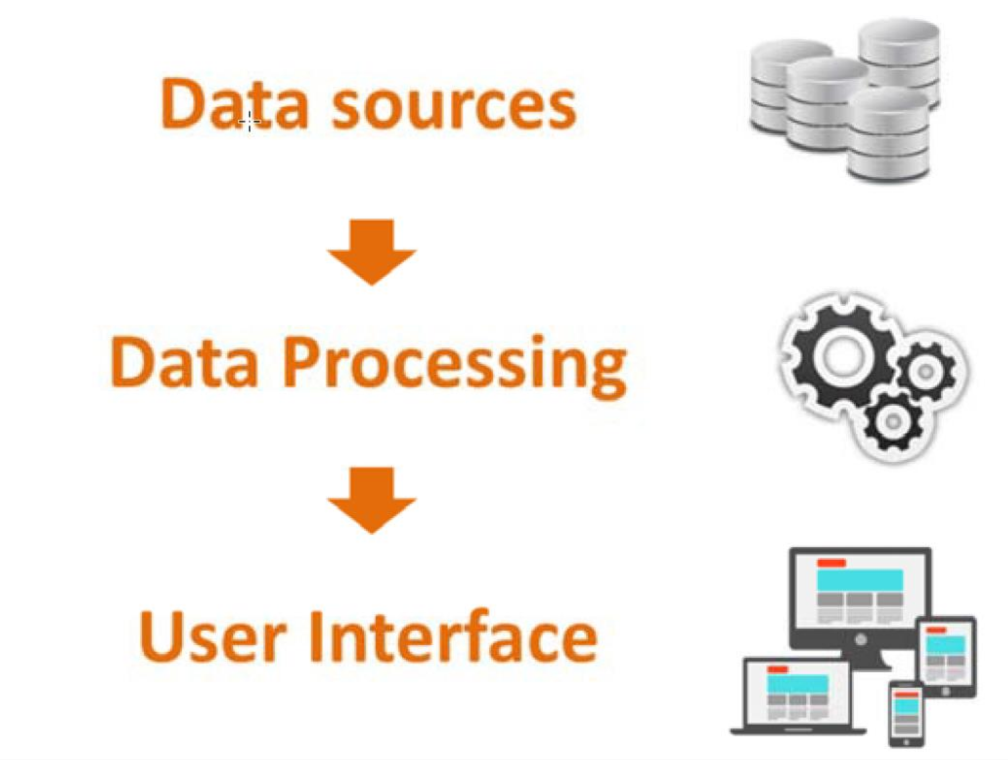
Chapter 1: What is big data?

HO
GENT

Contents

1. Information Management
2. What are big data?
3. The origin of big data
4. The 4 V’s of big data
5. Some other V’s
6. Bottlenecks
7. Examples of how big data can be used

1. Information Management

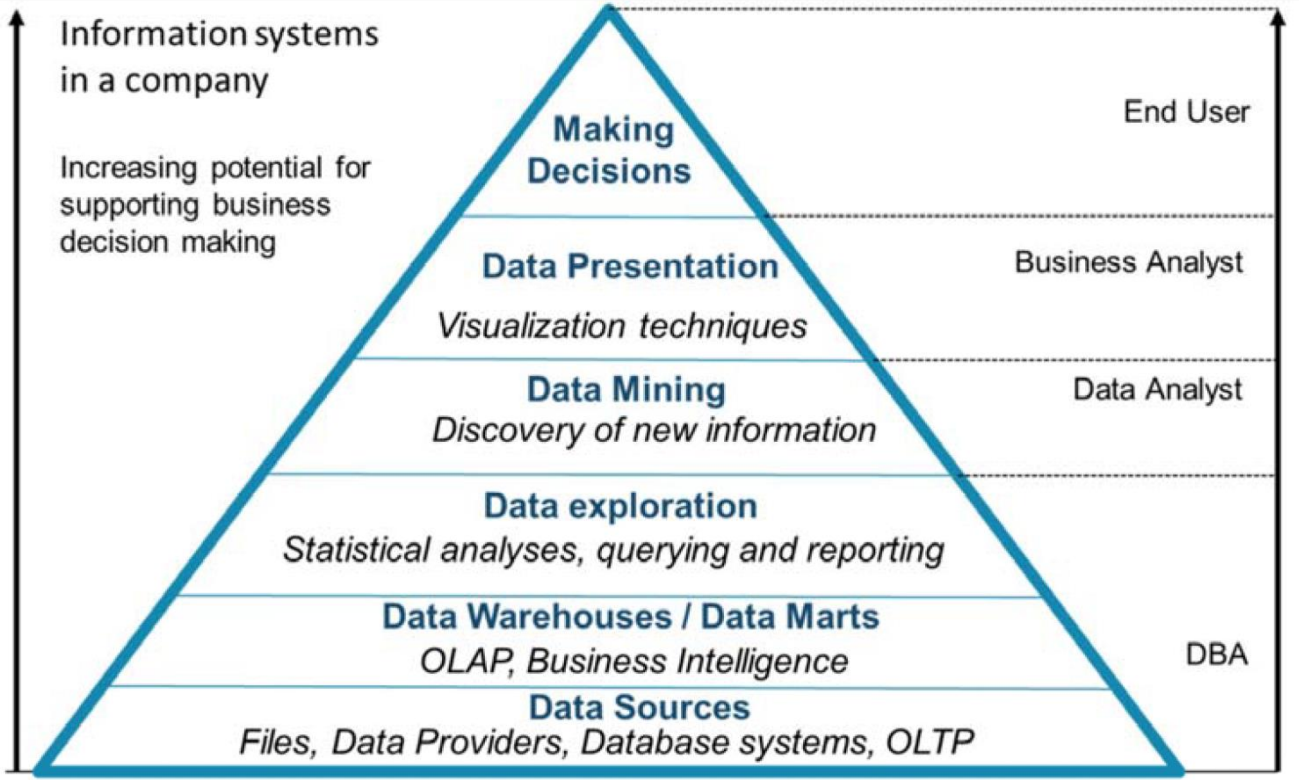


Data sources

Data Processing

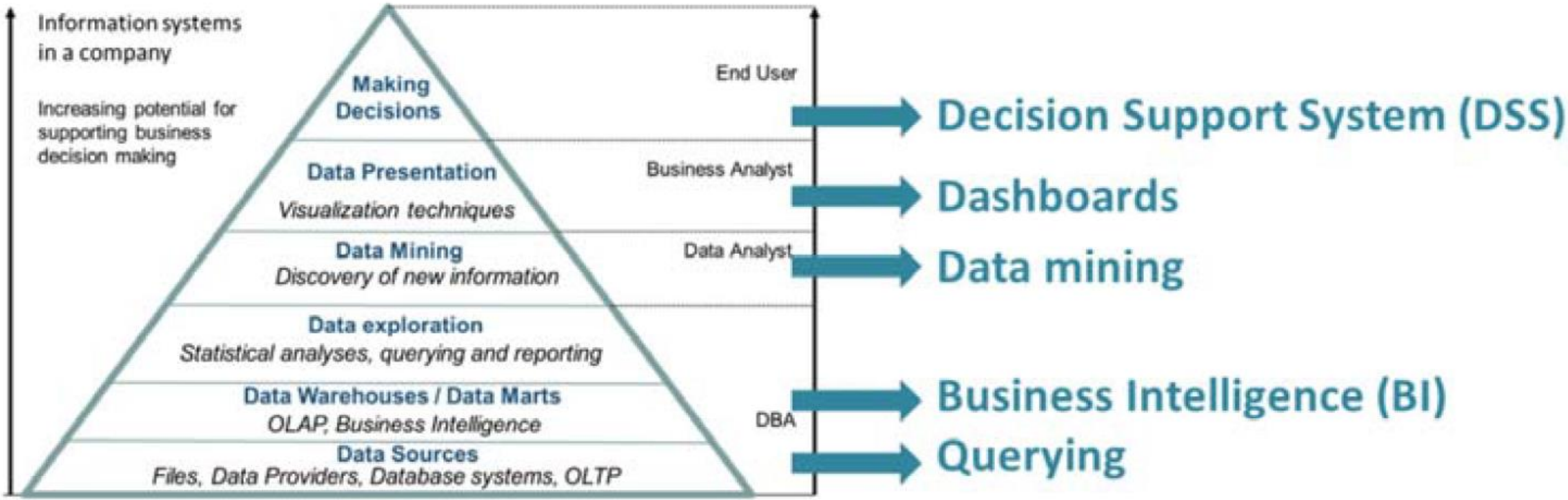
User Interface

Decision Support Systems


Information systems in a company
Increasing potential for supporting business decision making
Making Decisions
End User
Data Presentation
Visualization techniques
Business Analyst
Data Mining
Discovery of new information
Data Analyst
Data exploration
Statistical analyses, querying and reporting
Data Warehouses / Data Marts
OLAP, Business Intelligence
DBA
Data Sources
Files, Data Providers, Database systems, OLTP

A decision support system (DSS) is a computer program application used to improve a company's decision-making capabilities. It analyzes large amounts of data and presents an organization with the best possible options available.

Traditional systems


Information systems
in a company
Increasing potential for supporting business decision making
Making Decisions
End User
Data Presentation
Visualization techniques
Business Analyst
Data Mining
Discovery of new information
Data Analyst
Data exploration
Statistical analyses, querying and reporting
Data Warehouses / Data Marts
OLAP, Business Intelligence
DBA
Data Sources
Files, Data Providers, Database systems, OLTP
Decision Support System (DSS)
Dashboards
Data mining
Business Intelligence (BI)
Querying

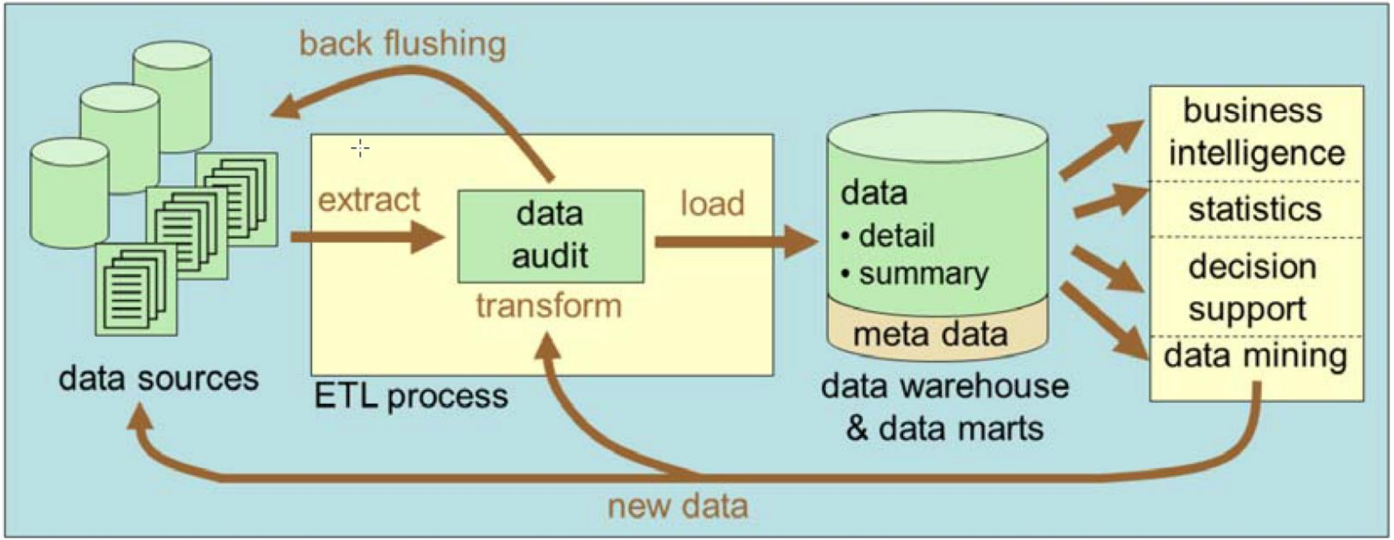
OLTP: Online Transaction Systems, day-to-day operations

OLAP: Online Analytical Processing, midterm and longterm decision support

What do users want?

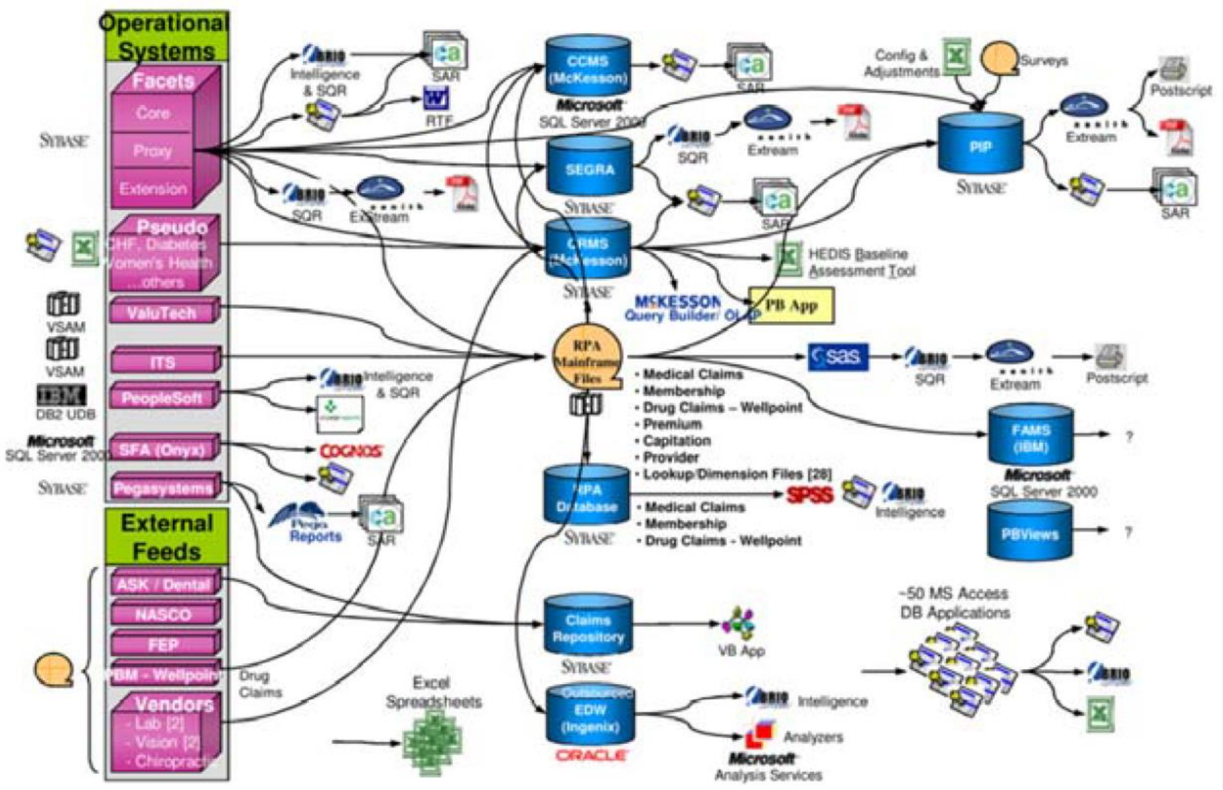
Users want a single logic database for their analyses: it looks like all distributed, heterogeneous data is available from one single database.

What do we usually have? … in theory

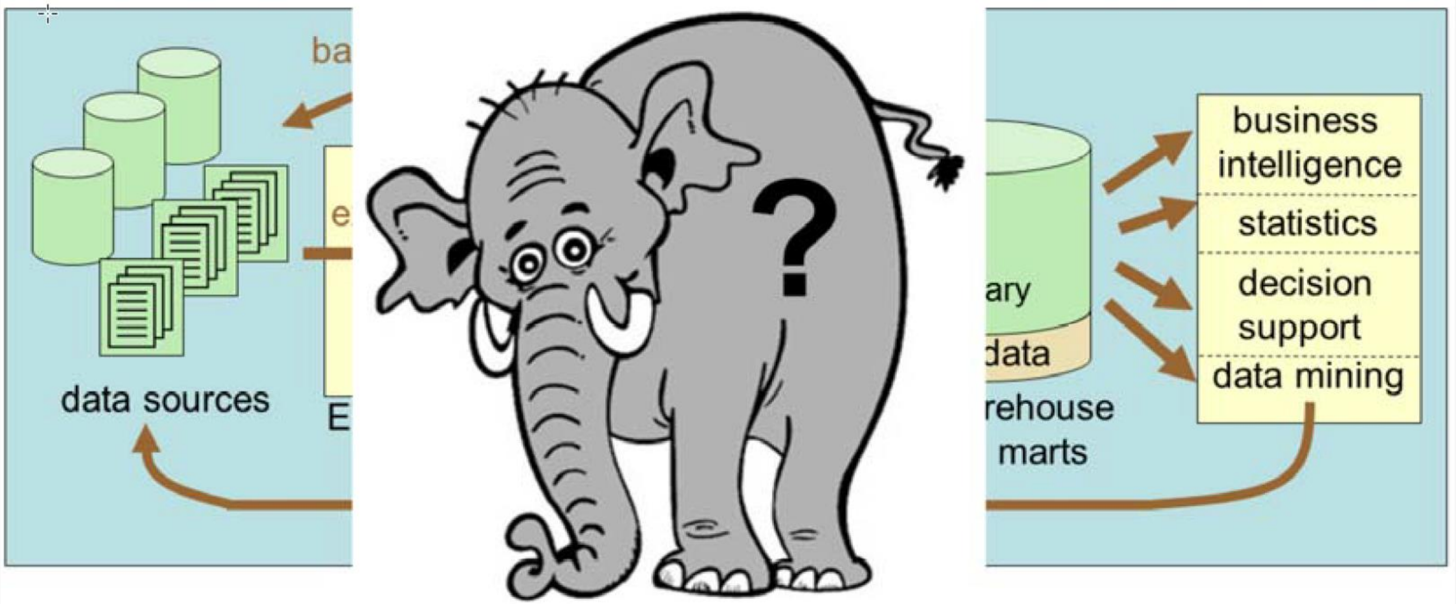


This is the theoretical process of **E**xtracting data from different source systems, **T**ransforming and cleaning it to a suitable format and **L**oading it into a Datawarehouse for easy Business Intelligence.

What do we usually have? … in practice

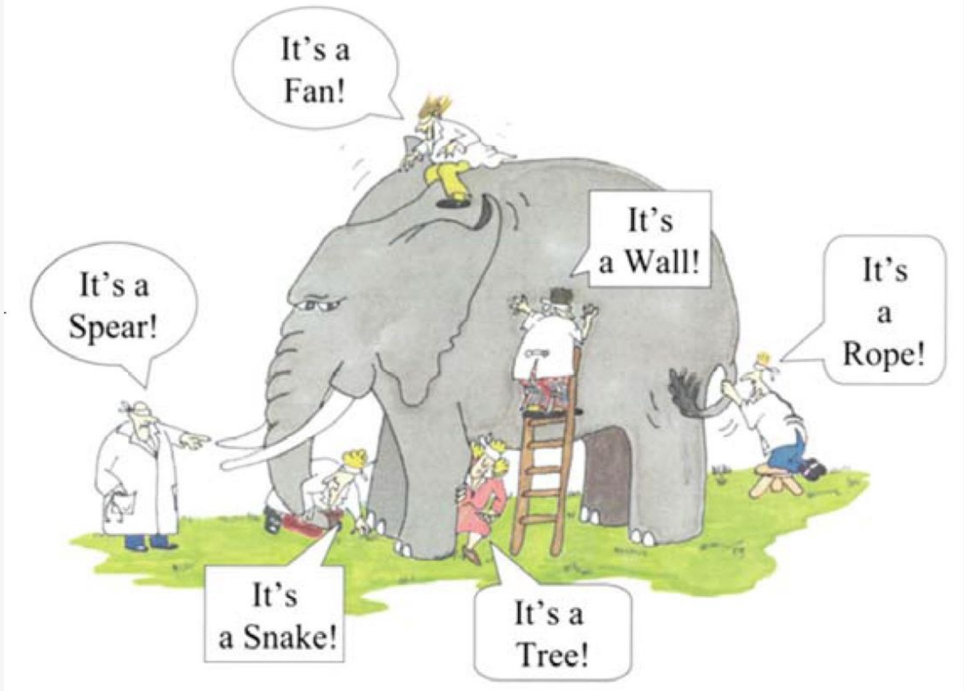


How does Big Data fit in these pictures?



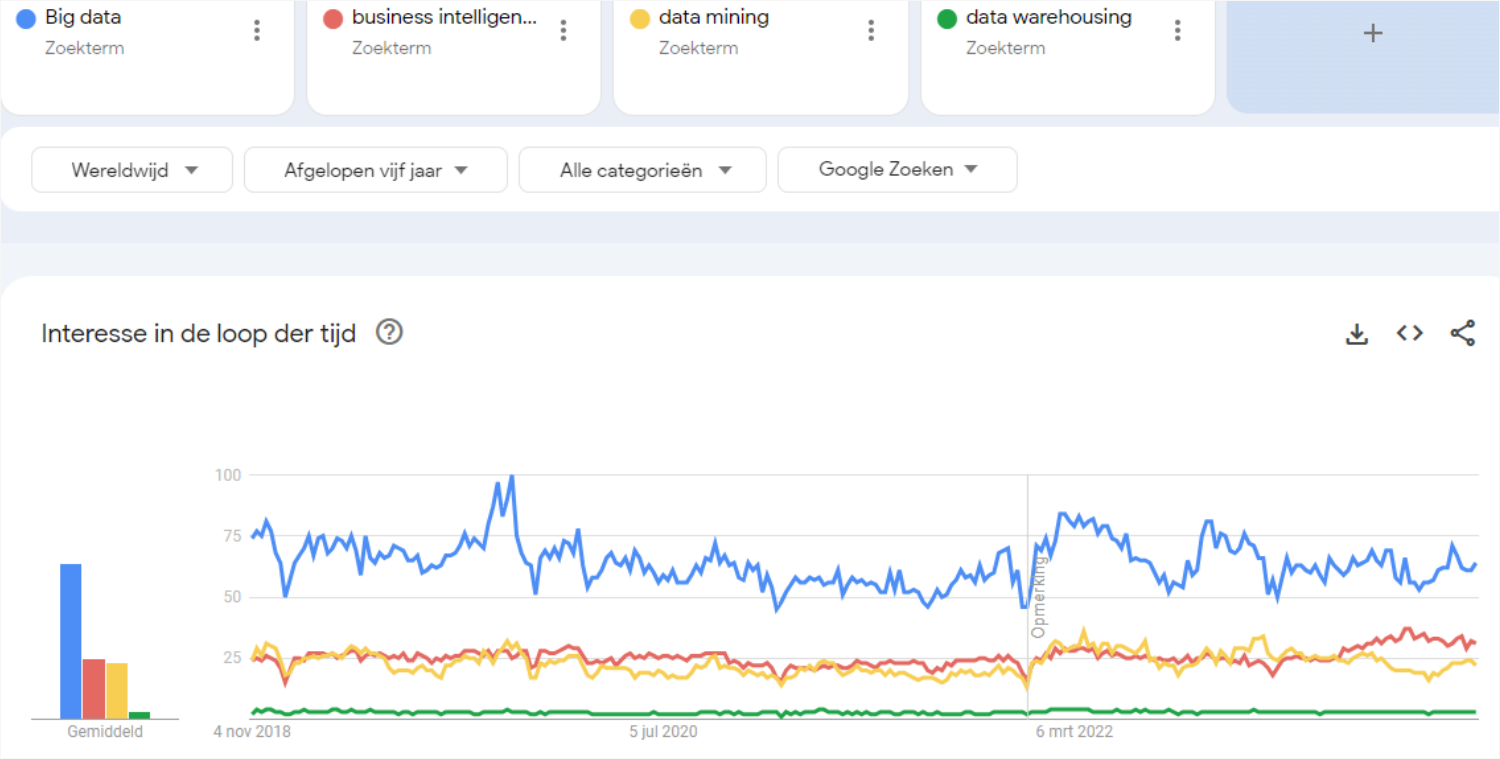
Furthermore, traditional databases, ETL processes, datawarehouses, data mining techniques etc. are not suited for processing data with “specific” characterics. Such data are called “big” data.

2. What are big data?



Many definitions of big data exist, some are leading to confusion and unjustified criticism …

Big Data: Just the next hype or reality?



● Big data
Zoekterm

● business intelligen...
Zoekterm

● data mining
Zoekterm

● data warehousing
Zoekterm

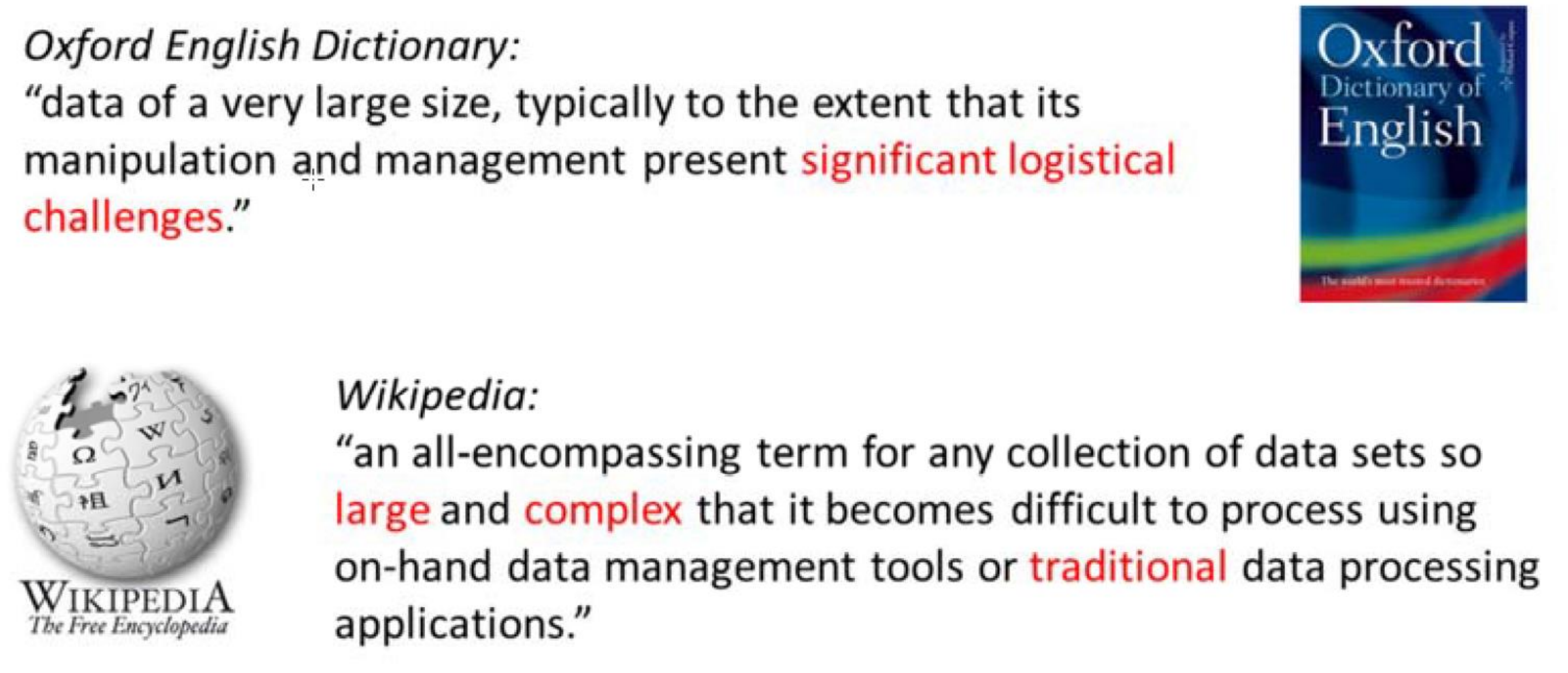
Wereldwijd ▾ | Afgelopen vijf jaar ▾ | Alle categorieën ▾ | Google Zoeken ▾

Interesse in de loop der tijd

Source: Google Trends

What are ‘big’ data?

Definitions of big data



Oxford English Dictionary:

“data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.”

Wikipedia:

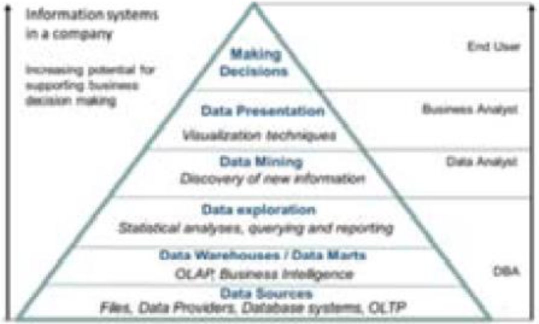
“an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.”

Definitions of big data

My personal definition (Johan Decorte):

“Data that is not (consciously) entered by a user but that arise, often spontaneously, as a by-product of other processes and that are (usually) used for purposes for which they were not originally intended”.

3. The origin of ‘big’ data



Information systems in a company

Increasing potential for supporting business decision making

Making Decisions — End User

Data Presentation — *Visualization techniques* — Business Analyst

Data Mining — *Discovery of new information* — Data Analyst

Data exploration — *Statistical analyses, querying and reporting*

Data Warehouses / Data Marts — *OLAP, Business Intelligence* — DBA

Data Sources — *Files, Data Providers, Database systems, OLTP*

➡ New data characteristics

➡ New data challenges

+

+

+

Internet Of Things

VIDEO

What is the Internet of Things (IoT)?

LinkedIn Learning · May 2019 · From the course: Tech Sense

2m 21s

https://www.linkedin.com/learning/tech-sense/what-is-the-internet-of-things-iot?autoplay=true&u=77666441

(watch for free with your HOGENT-account)

4. The 4 V’s of big data

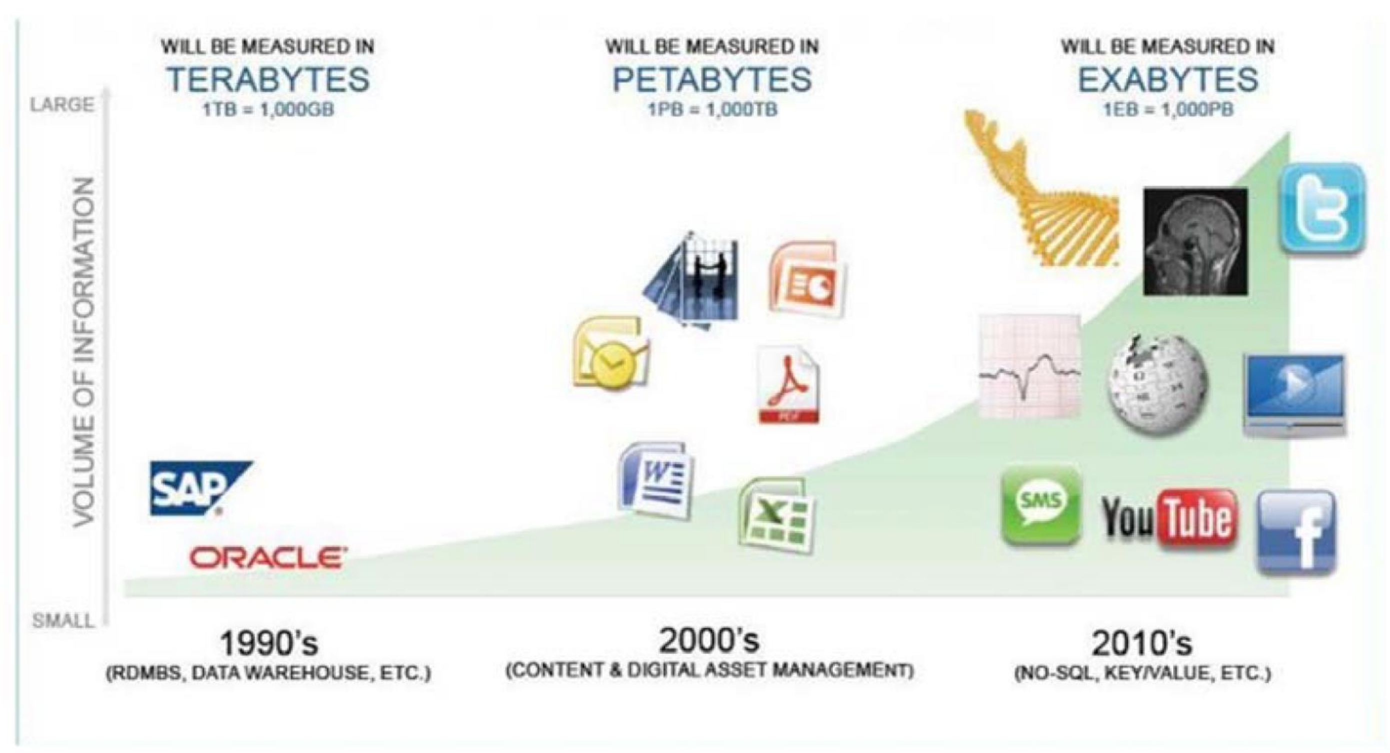
- **Volume**: the amount of data, also referred to the data “at rest”
- **Variety**: the range of data types and sources that are used, data in its “many forms”
- **Velocity**: the speed at which data comes in and goes out, data “in motion”, streaming data
- **Veracity**: the uncertainty of the data, data “in doubt”

Big Data Challenges

- **Volume**: BIG data → horizontal scaling/distributed data
- **Variety**: varied data → schemaless databases
- **Velocity**: fast data → NoACID databases
- **Veracity**: bad data → Data veracity handling

The 4 V’s of big data

BIG Data



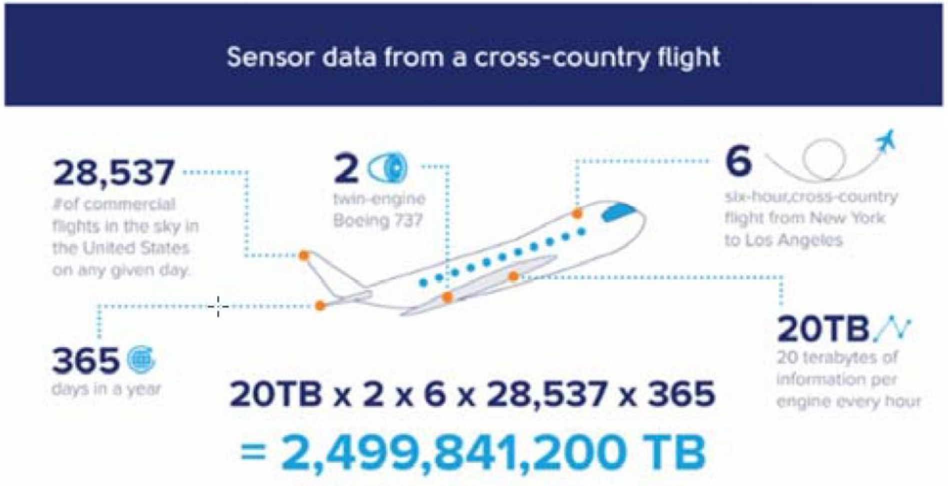
- Don’t decide too fast you have BIG data: tradition SQL databases can easily store several terabytes of data. Relational databases and SQL are still the preferred technology for most data related applications
- Automated and extremely fast generation and (near) real-time data processing of unstructured data require new technologies
  - NoSQL databases
  - Big data analysis via Python and R

Varied Data

To store business cards in a traditional database system you have to create a uniform database structure with fixed database schema which requires expensive data transformations

→ Use case for NoSQL databases

Fast Data (IoT)



Sensor data from a cross-country flight

28,537 #of commercial flights in the sky in the United States on any given day.

2 twin-engine Boeing 737

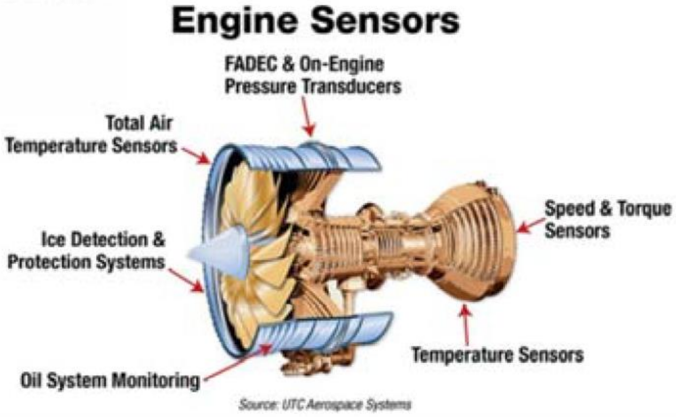
6 six-hour,cross-country flight from New York to Los Angeles

20TB 20 terabytes of information per engine every hour

365 days in a year

20TB x 2 x 6 x 28,537 x 365

= 2,499,841,200 TB

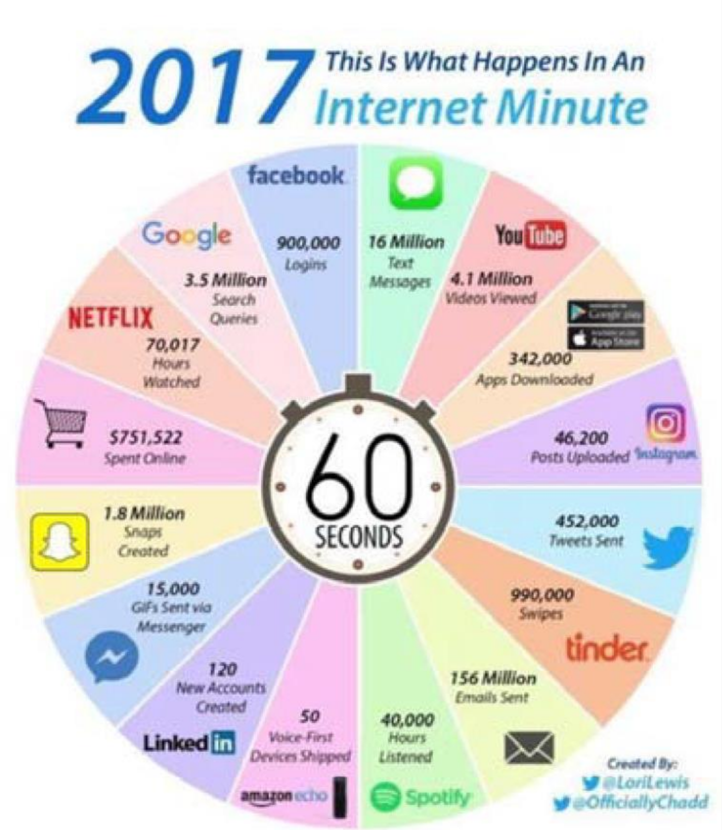


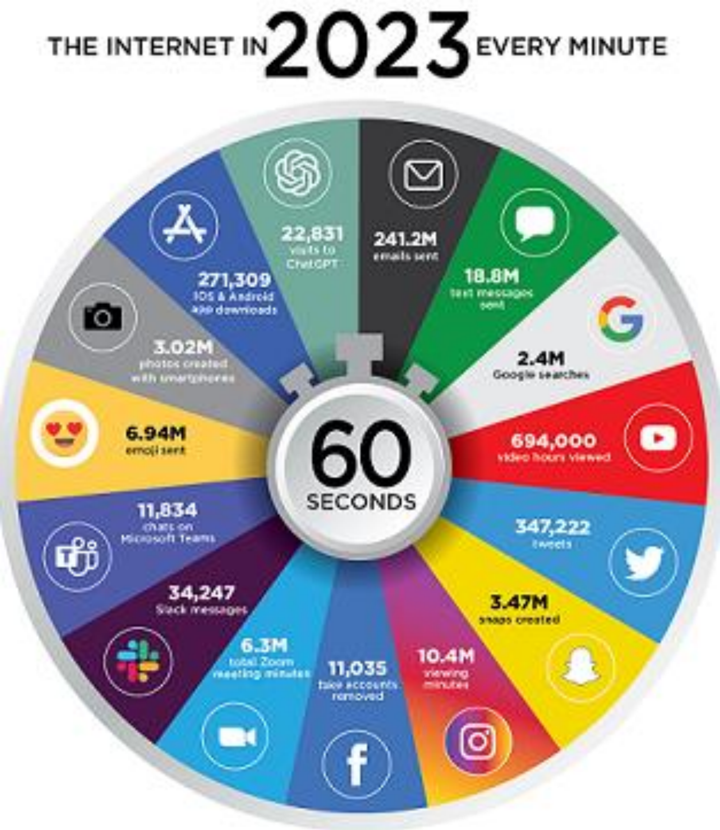
Engine Sensors

FADEC & On-Engine Pressure Transducers
Total Air Temperature Sensors
Ice Detection & Protection Systems
Speed & Torque Sensors
Temperature Sensors
Oil System Monitoring

Source: UTC Aerospace Systems

Voluminous, Fast and Varied Data (social media)





The 4 V's of big data

Bad Data (bad veracity)

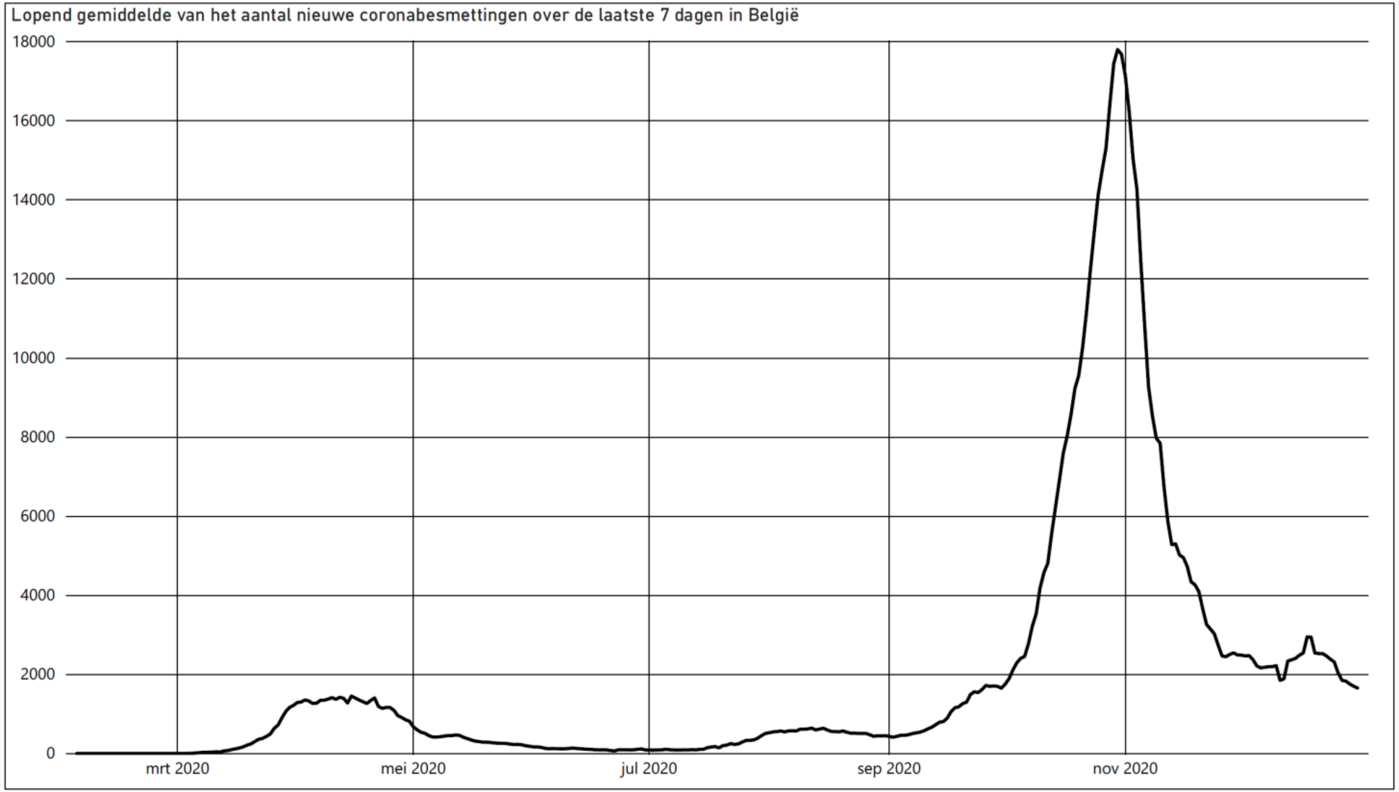
- Imprecise data
- Vague data
- Uncertain data
- Incomplete (missing) data
- Inconsistent data

→ Large volumes of varied data that need to be processed very fast are susceptive to bad data quality (failing sensors, networks, …)

→ Handling bad or unknown quality is a big challenge.

→ Such data is also called big data

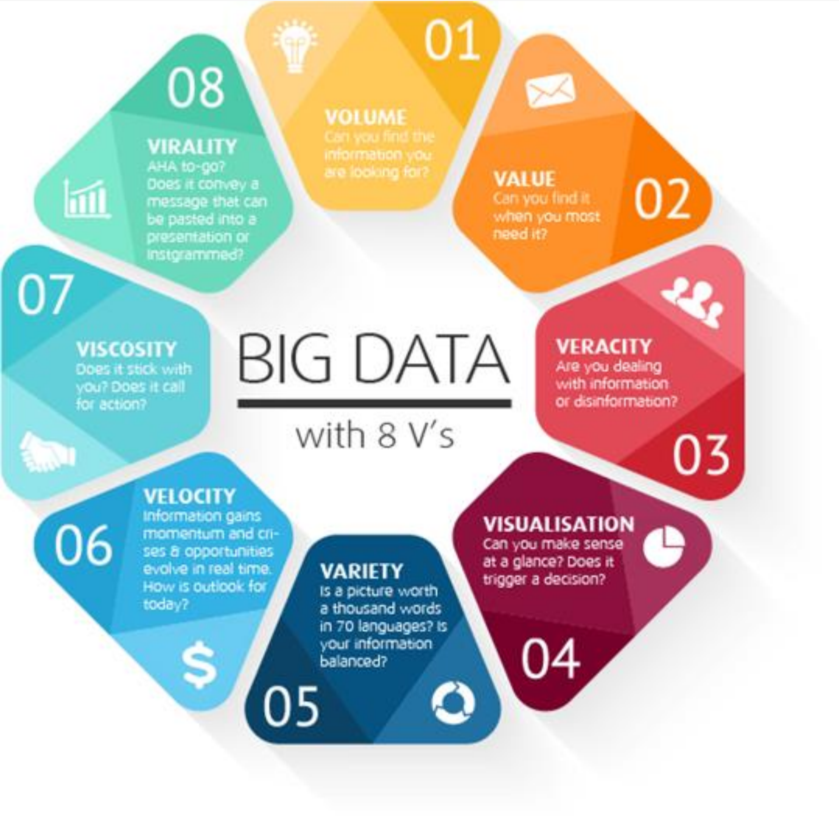
Example of inconsistent data



In the first COVID wave in Belgium, the laboratories detected an estimated one-thirtieth of the actual corona infections, in the second wave an estimated one-third. This gives the wrong impression that the second wave was much worse than the first in terms of number of infections.

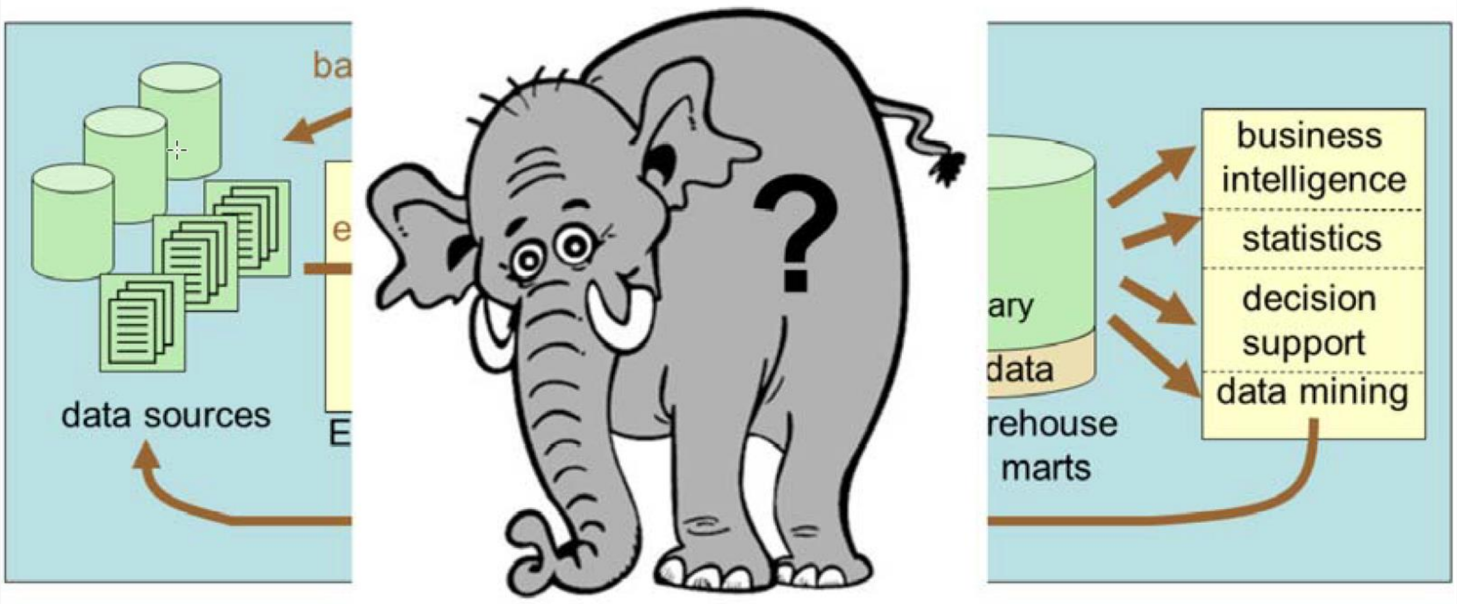
Data source: https://ourworldindata.org/coronavirus-source-data. Visualisation software used: Microsoft Power BI.

5. Some other V’s


01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and cri-ses & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?
BIG DATA
with 8 V’s

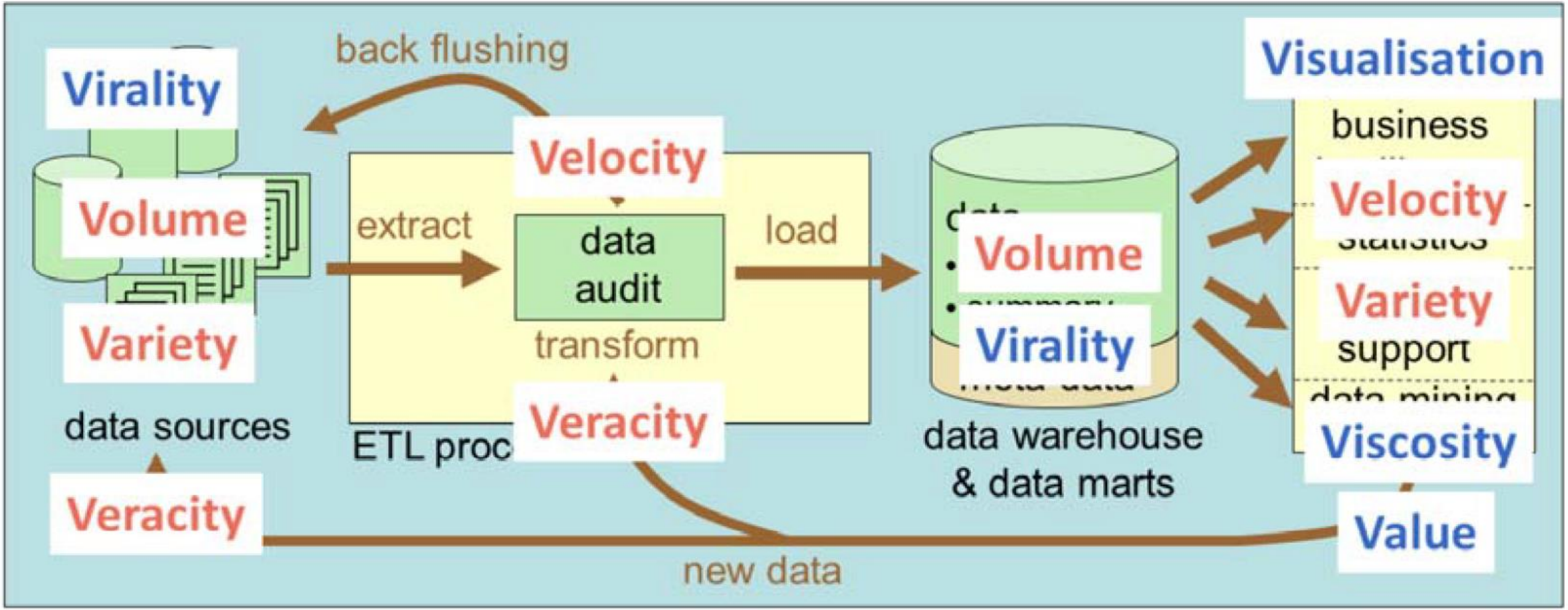
- **Virality**: How long do we need to keep data, when does it become outdated?
- **Viscosity**: Do we have enough data to perform (statistically) relevant analyses?
- **Visualisation**: can the results easily be presented?
- **Value**: what the value of our data? “Data is the new gold”.

6. Bottlenecks



Where do we find these challenges in the (theoretical) information processing?

Bottlenecks



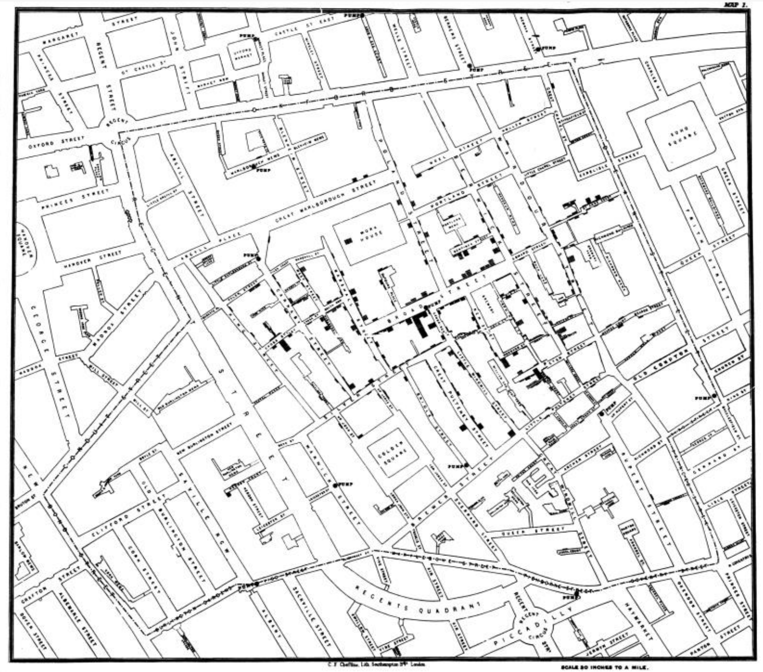
7. Examples of how big data can be used

- London (Soho) cholera outbreak in 1854 → (deaths per household, centralized around infected water pump)

- Counting the number of visitors of the 'Gentse Feesten': https://www.nieuwsblad.be/cnt/dmf20180722_03626134

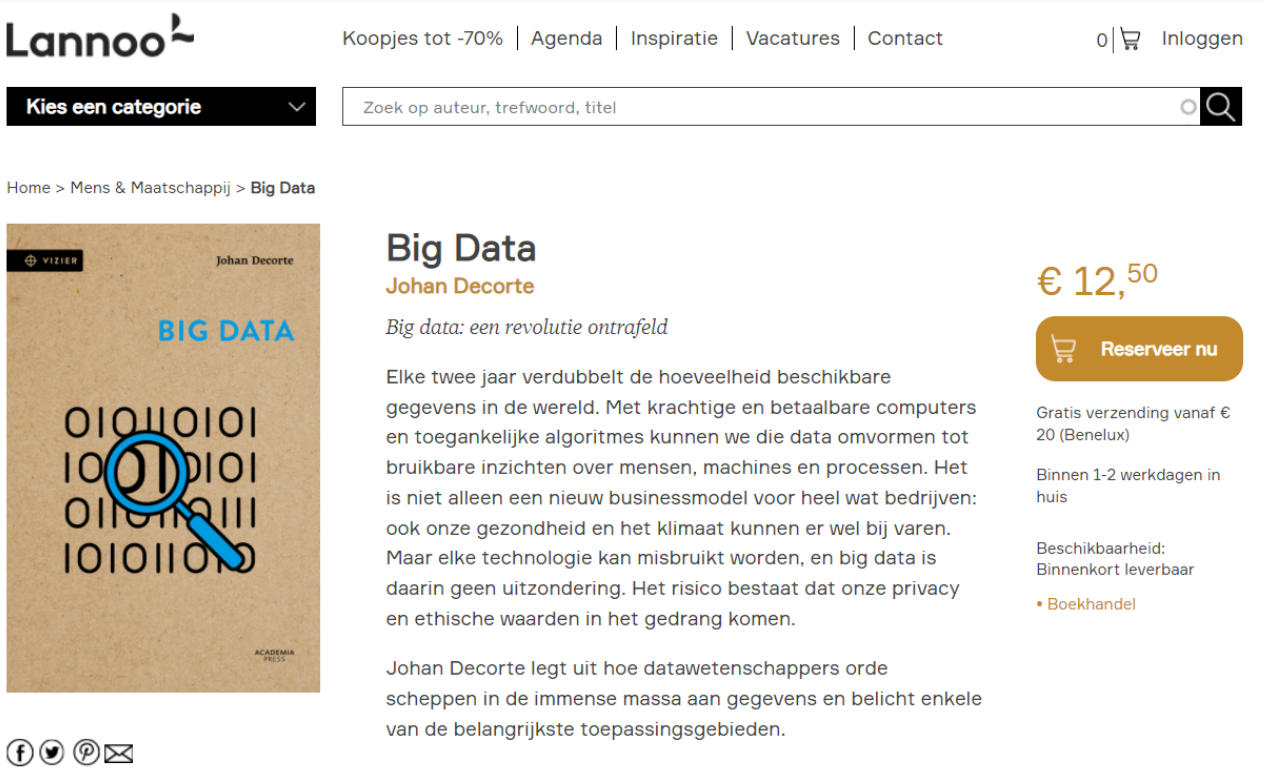
- Better predictions using Big Data: see next video (dutch):

  https://www.youtube.com/watch?v=m9YcrXIc2IM



IN CE 260E

If you want to now more …



Big Data

Johan Decorte

Big data: een revolutie ontrafeld

Elke twee jaar verdubbelt de hoeveelheid beschikbare gegevens in de wereld. Met krachtige en betaalbare computers en toegankelijke algoritmes kunnen we die data omvormen tot bruikbare inzichten over mensen, machines en processen. Het is niet alleen een nieuw businessmodel voor heel wat bedrijven: ook onze gezondheid en het klimaat kunnen er wel bij varen. Maar elke technologie kan misbruikt worden, en big data is daarin geen uitzondering. Het risico bestaat dat onze privacy en ethische waarden in het gedrang komen.

Johan Decorte legt uit hoe datawetenschappers orde scheppen in de immense massa aan gegevens en belicht enkele van de belangrijkste toepassingsgebieden.

€ 12,50

Gratis verzending vanaf € 20 (Benelux)

Binnen 1-2 werkdagen in huis

Beschikbaarheid: Binnenkort leverbaar

• Boekhandel