

Odissee
DE CO-HOGESCHOOL

Big Data - Introductie



Jens Baetens



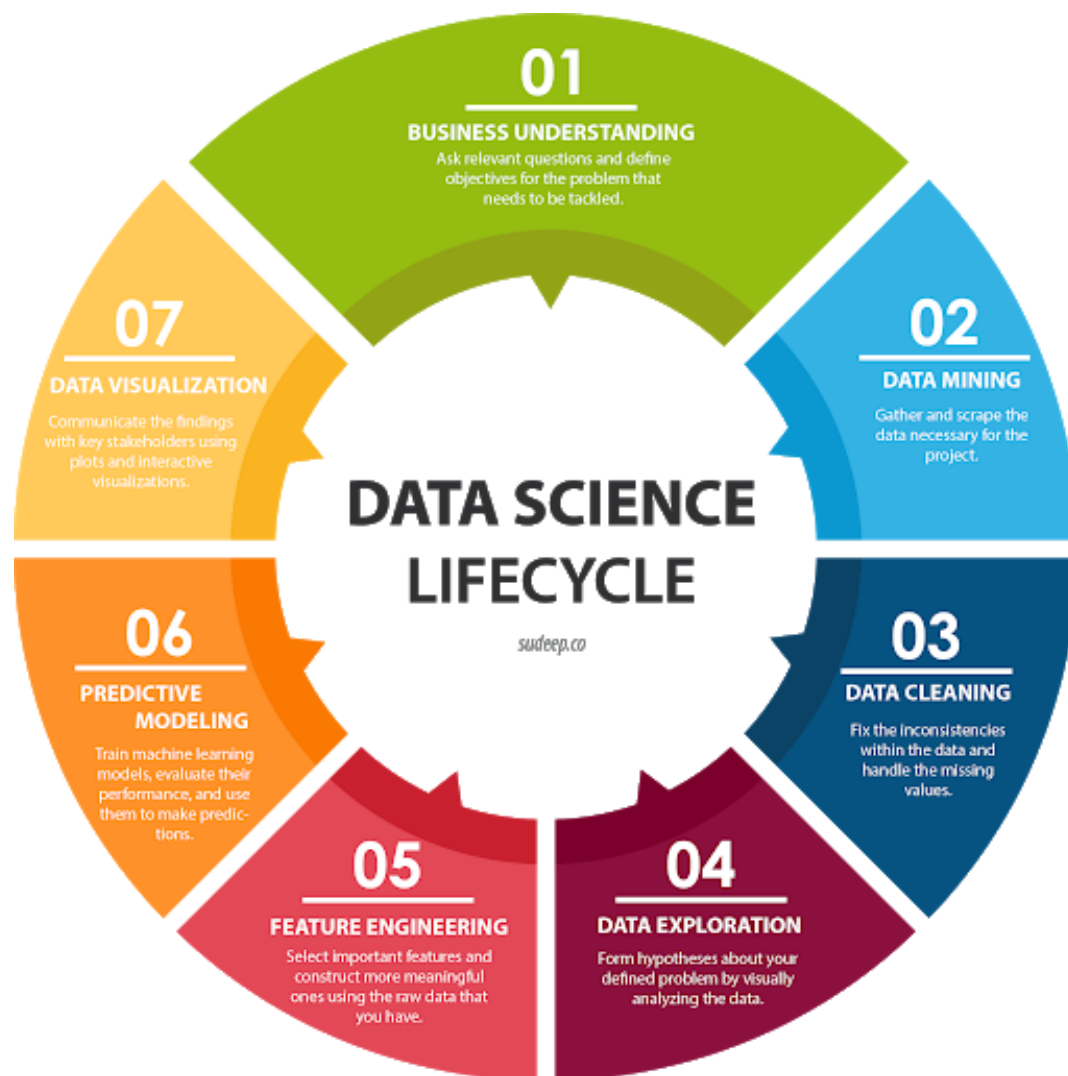
Structuur

- ▣ Recap – Data Science
- ▣ Kenmerken van Big Data
- ▣ Distributed Storage
- ▣ Distributed Computing
- ▣ Tools



Recap – Data Science

Data Lifecycle

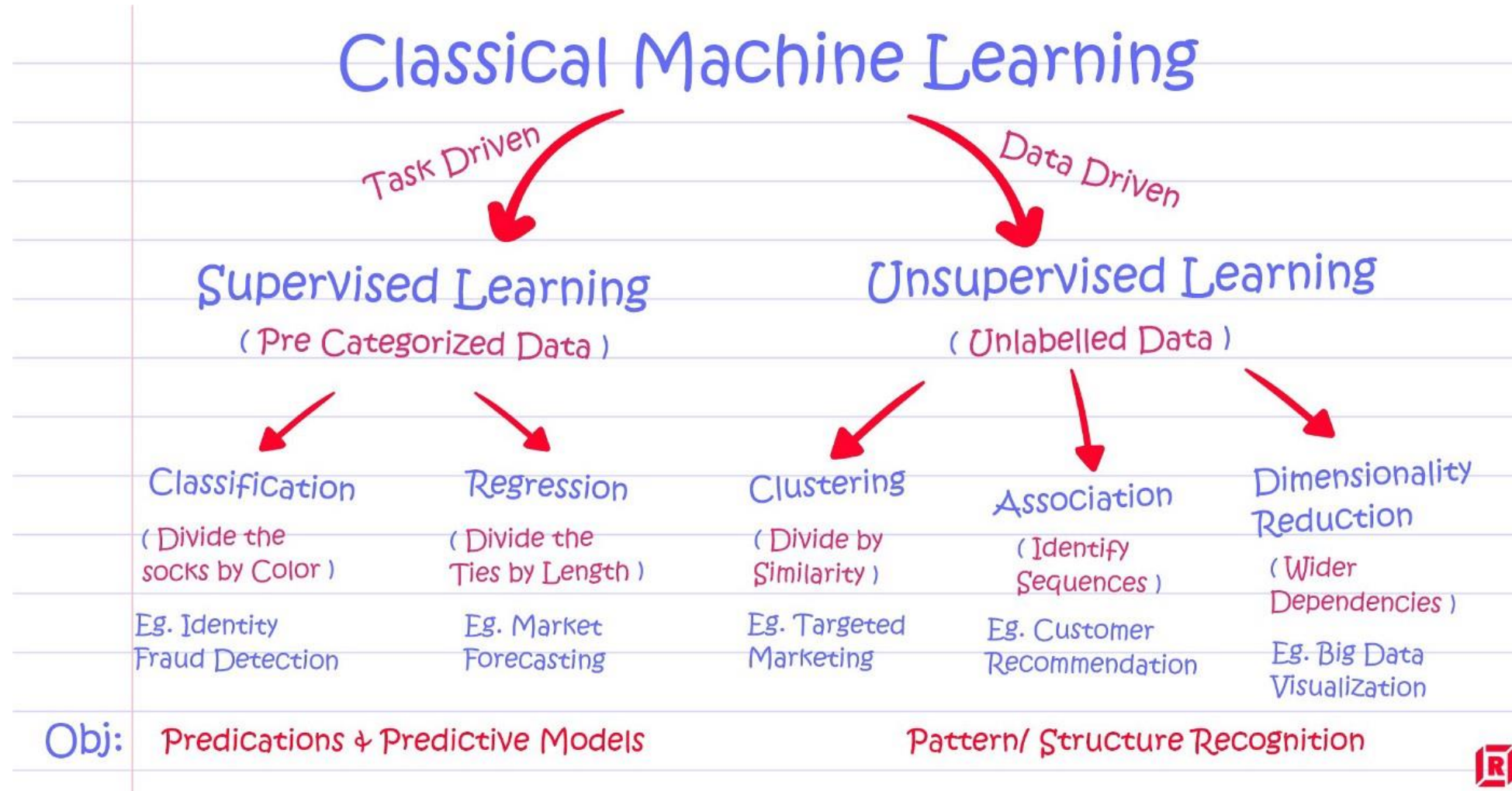




Data Cleaning & Exploration

- ▣ Bestuderen beschikbare datasets
 - ▬ Vinden van correlaties en verbanden
 - ▬ Informatie over de beschikbare data en hoe bruikbaar ze is
- ▣ Opschonen en bewerken van beschikbare data
 - ▬ Omzetten dataformaten (datums, bag of words, scaling ...)
 - ▬ Privacy van personen
 - ▬ Oplossen problemen in de data (typo's, vertalingen, ontbrekende data, ...)

Data Modelling



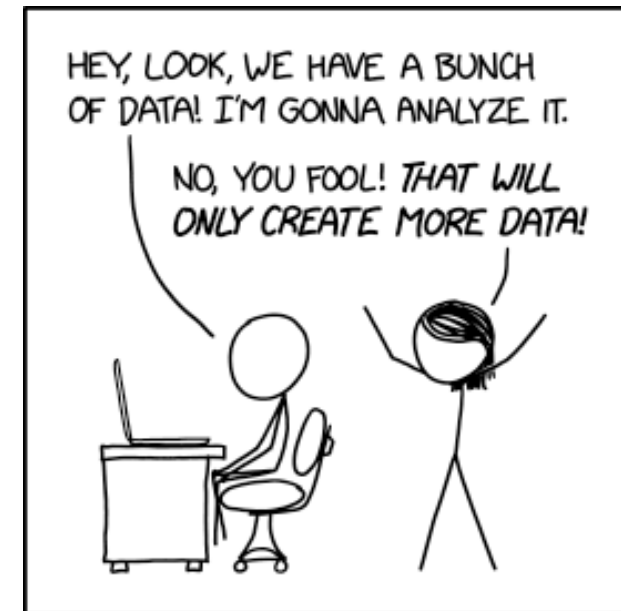


Gebruikte datasets

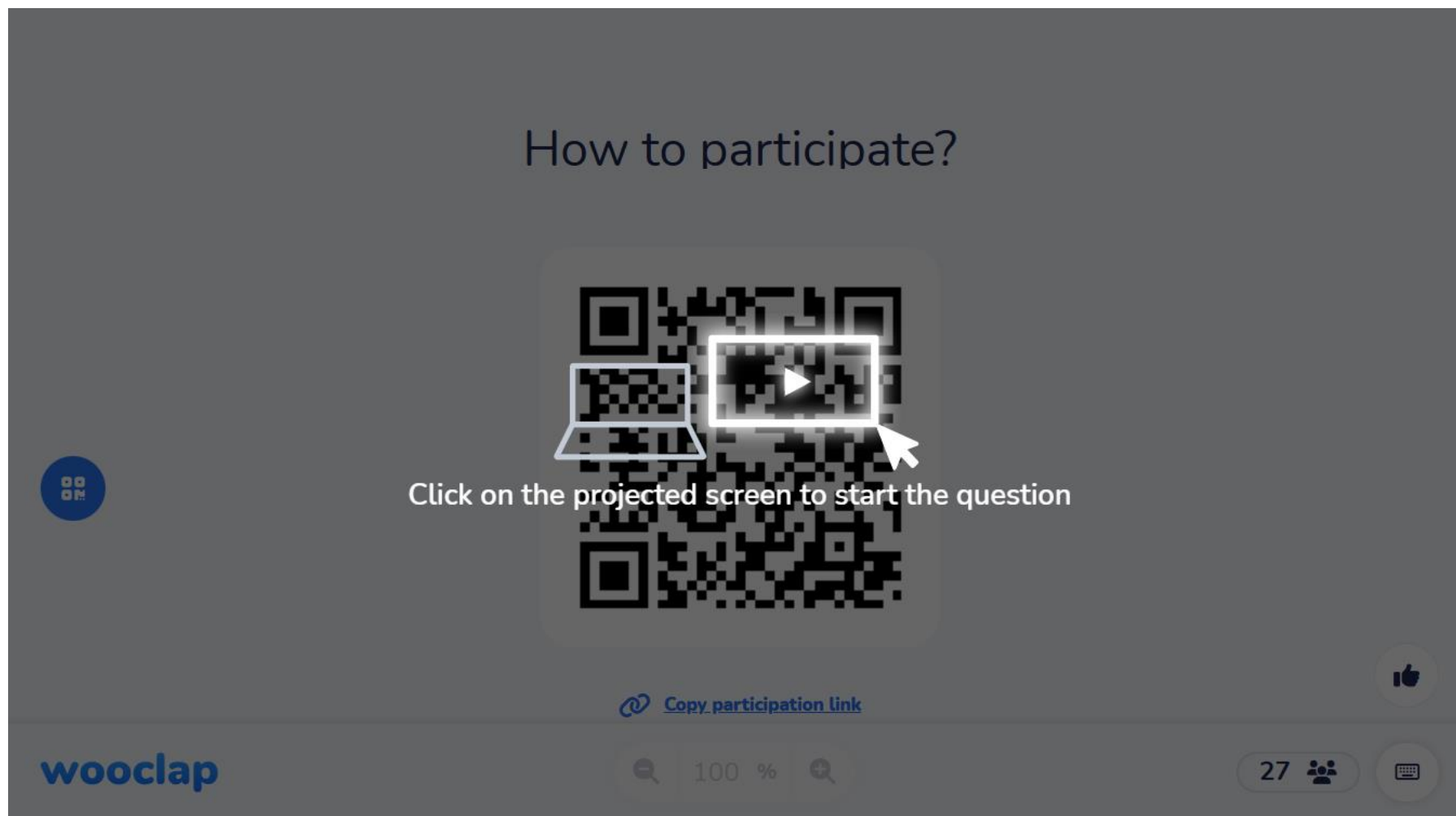
- ▣ Aantal honderden MB
- ▣ Csv of jpegs
- ▣ Gedownload naar harde schijf
- ▣ Volledig ingeladen in memory voor verwerking

2.

Is dit altijd mogelijk?



Zien jullie problemen?



Zien jullie problemen?

Go to **wooclap.com** and use the code **DBEIAG**

Zien jullie problemen met de daarnet besproken methode?

Ja

Ik denk van niet

Ja, sommige datasets zullen te groot zijn.

dat er niet genoeg of niet goeie soort data is

Ja, de data is te groot

Click on the projected screen to start the question

wooclap

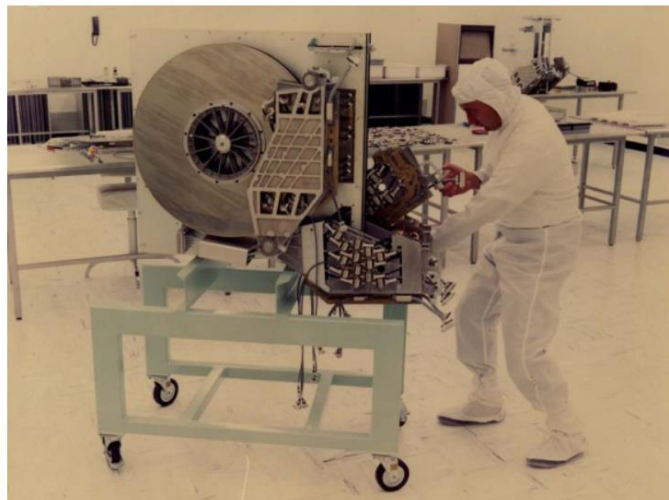
100 %

9

Grootte harde schijven?



1956: 5 MB



1975: 250 MB



1988: 1 GB



2019: 1 TB

$\times 1000$

$\times 1000$

Prijs?

Backblaze Average Cost per Drive Size

By Quarter: Q1 2009 - Q2 2017



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter



4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

DEMYSTIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
b bit	0 or 1	1/8 of a byte
B byte	8 bits	1 byte
KB kilobyte	1,000 bytes	1,000 bytes
MB megabyte	1,000 ² bytes	1,000,000 bytes
GB gigabyte	1,000 ³ bytes	1,000,000,000 bytes
TB terabyte	1,000 ⁴ bytes	1,000,000,000,000 bytes
PB petabyte	1,000 ⁵ bytes	1,000,000,000,000,000 bytes
EB exabyte	1,000 ⁶ bytes	1,000,000,000,000,000,000 bytes
ZB zettabyte	1,000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB yottabyte	1,000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

3.9bn

people use emails

4TB

of data produced by a connected car

Intel



Searches made a day

5bn

Searches made a day from Google

3.5bn

Smart Insights



ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

44ZB

PwC

2013

2020

463EB

of data will be created every day by 2025

IOC

95m

photos and videos are shared on Instagram

Instagram Business



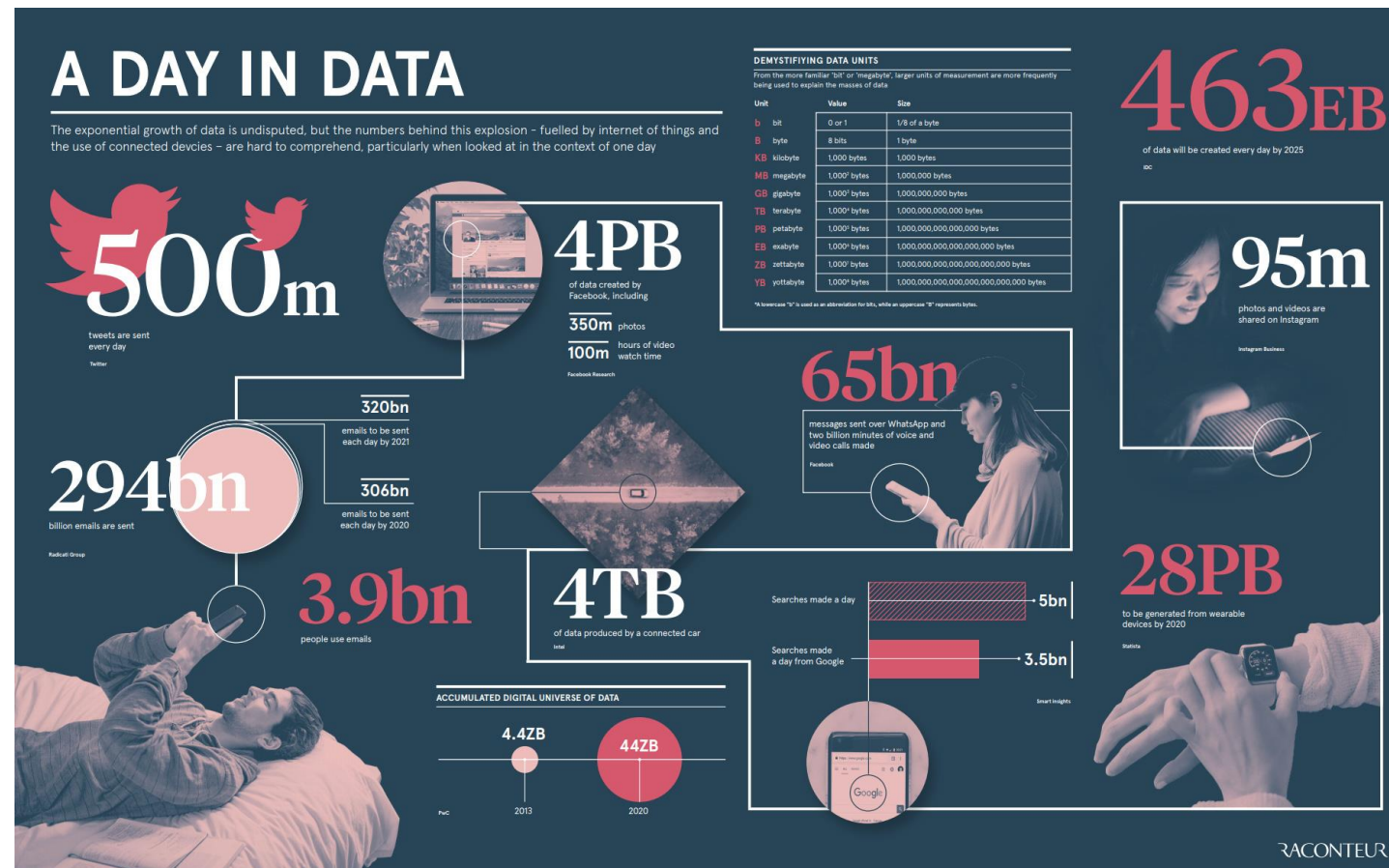
28PB

to be generated from wearable devices by 2020

Statista



- ▣ 1 PB = 125 8TB HDD's
- ▣ 1 EB = 125000 8TB HDD's
- ▣ 1 ZB = 125 000 000 8TB HDD's



Wat zijn je opties voor dit soort data bij te houden?

Go to **wooclap.com** and use the code **DBEIAG**

Welke opties zijn er voor dit volume van data bij te houden?

- Data warehouse
- cloud opslag
- Edge computing
- Sql database
- Magnetic tape

wooclap

100 %

10

Is het mogelijk om alles lokaal bij te houden om te verwerken?

- Onmogelijk om computers te kopen die deze hoeveelheid data bijhoudt.
- RAM-geheugen nodig om data in te laden (Ook niet mogelijk)

-> Distributed Computing

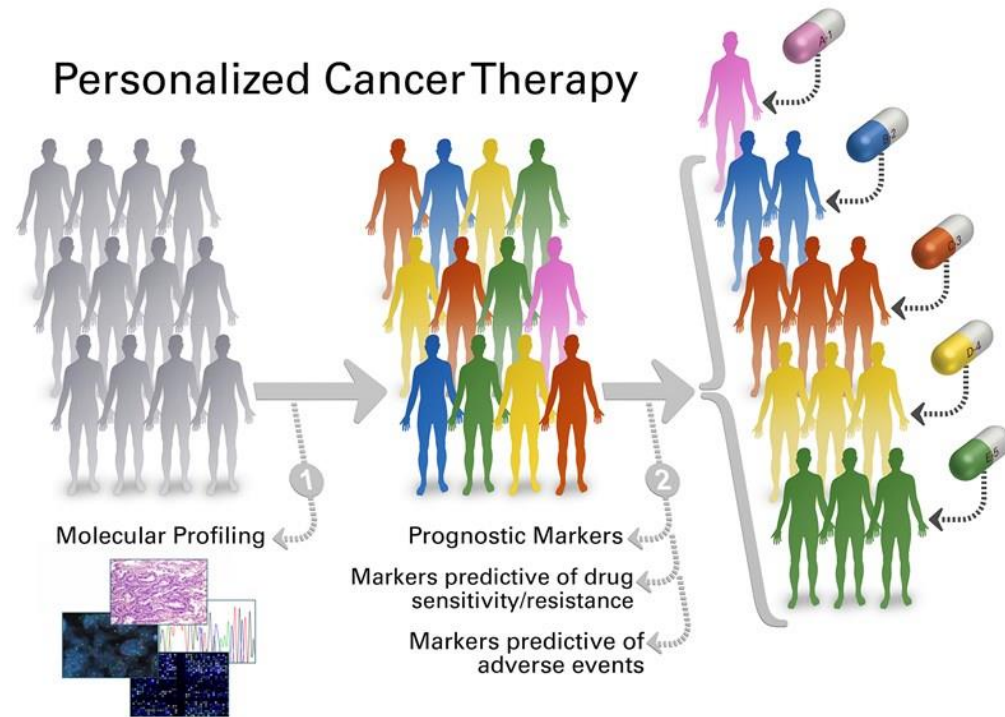
-> Cloud Computing

| *horizontaal schalen*

1 sterke PC = vert schalen

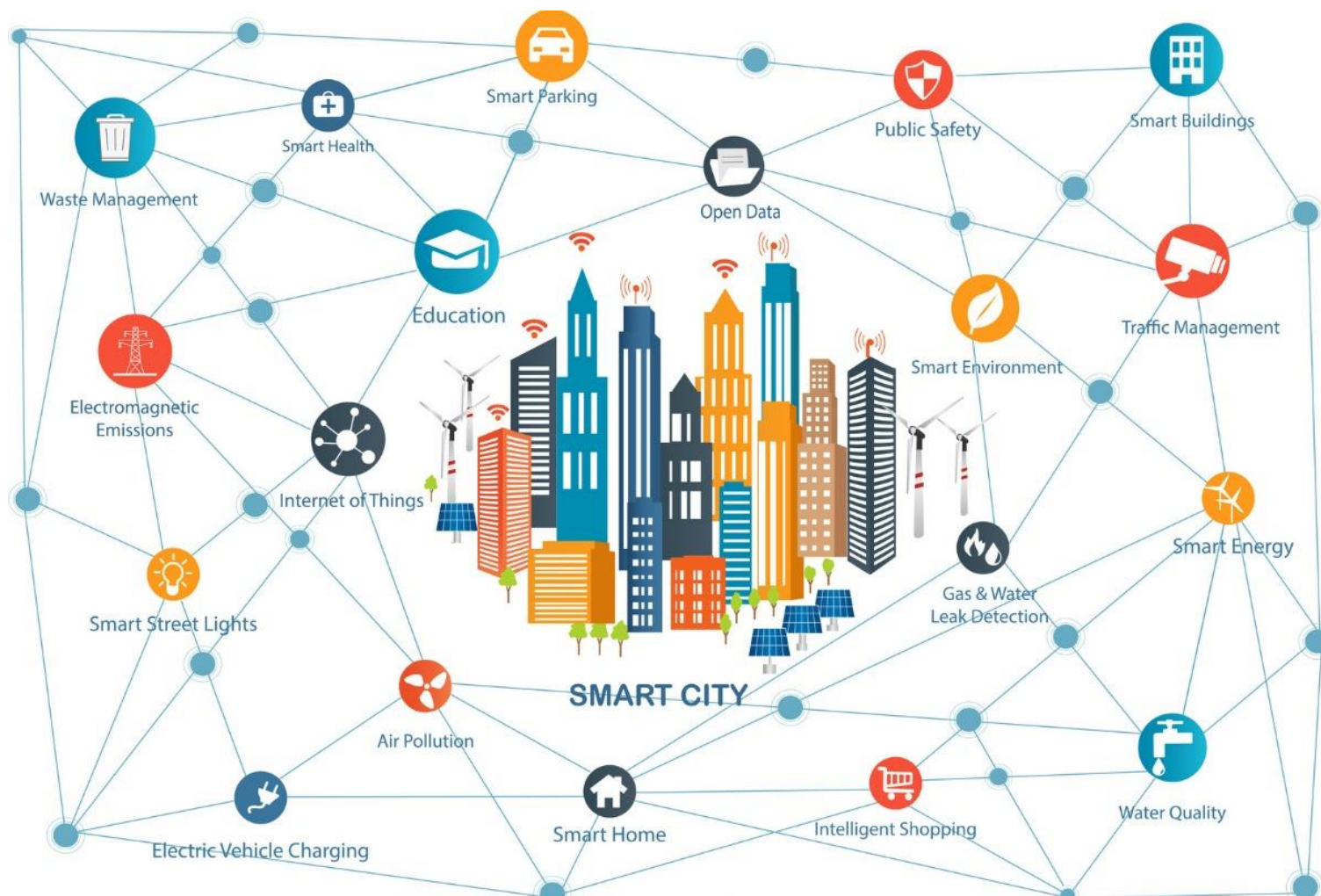
Waarom zoveel data nodig?

- ▣ Meer data -> betere modellen -> betere voorspellingen / verder vooruit voorspellen
- ▣ Menselijk DNA = 100 GB




Waarom zoveel data nodig?

Smart Cities





<https://www.youtube.com/watch?v=i3zx3gF9AUU>

- 
- ▣ Large Hadron Collider : 90 PBs per jaar
 - ▣ Boeing 737: 20 TB per uur per motor



3.

Big Data

Definitie - Wikipedia

- **Big data** of **massadata**^[1] zijn gegevensverzamelingen (datasets) die te groot en te weinig gestructureerd zijn om met reguliere databasemanagementsystemen te worden onderhouden. De gegevens hebben een direct of indirect verband met privégegevens van personen. ^[2] Big data spelen een steeds grotere rol. De hoeveelheid data die opgeslagen wordt, groeit exponentieel. Dit komt doordat consumenten bij sociale media in toenemende mate data opslaan in de vorm van bestanden, foto's en films (bijvoorbeeld op Facebook of YouTube, waar Facebook ook de door de gebruikers gewiste data bewaart) en organisaties, overheden en bedrijven steeds meer data over burgers produceren en opslaan, en doordat apparaten zelf data verzamelen, opslaan en uitwisselen (het zogenaamde internet der dingen). Hierdoor is er steeds meer sensordata beschikbaar. Niet alleen de opslag van deze hoeveelheden is een uitdaging, maar ook het analyseren ervan. Deze data bevatten namelijk informatie voor doeleinden zoals marketing, wetenschappelijk onderzoek, of preventief onderhoud.



Definitie - Gartner

- ▣ **Big data** is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.



Kenmerken

Polling – Wat bepaalt wanneer een data science project een big-data project is?

Go to **wooclap.com** and use the code **DBEIAG**

Wat bepaalt of een data science project een big data project is?

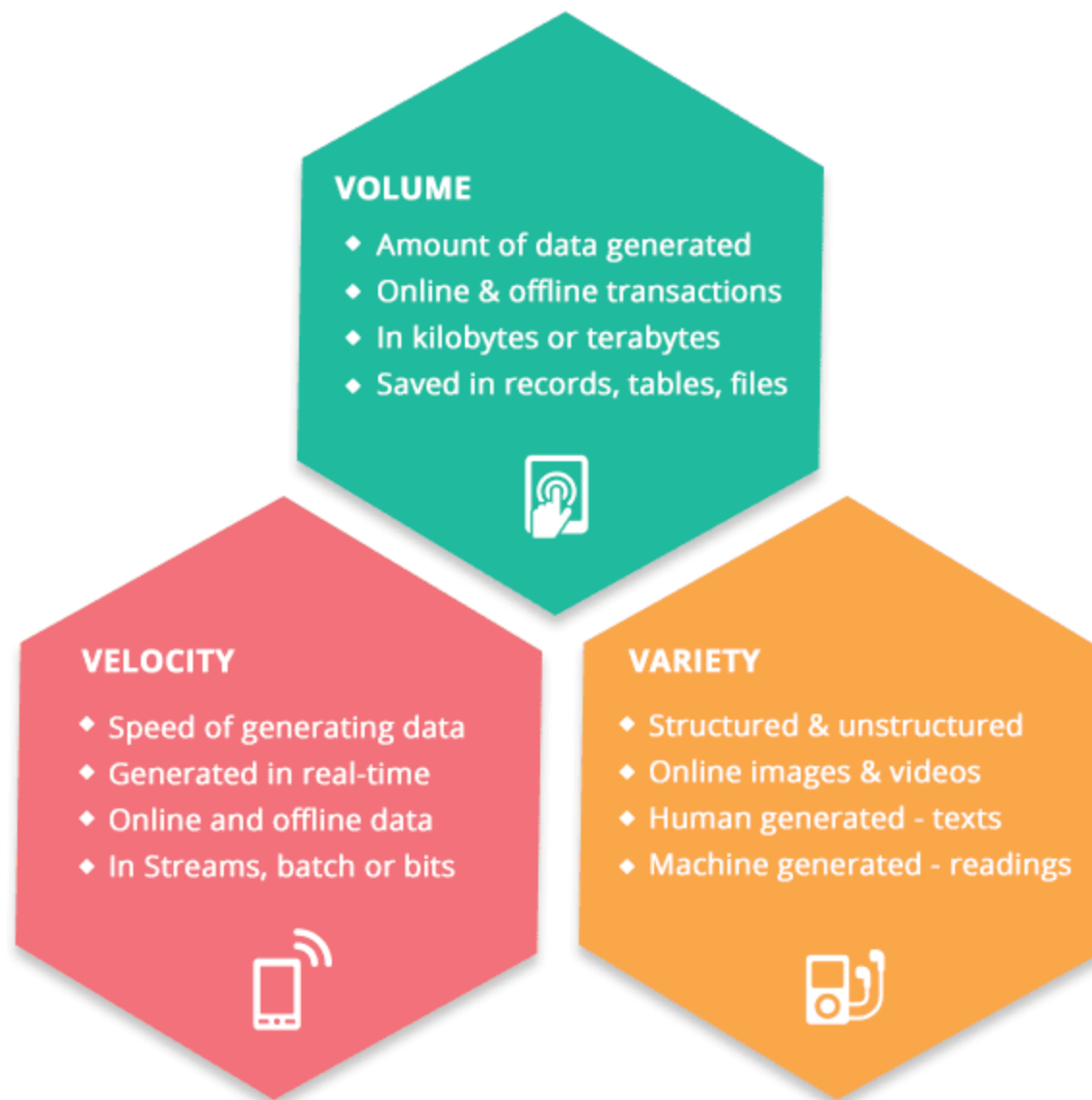
- variatie in de data
- Het hoeveelheid data
- de hoeveelheid data en hoe snel het verwerkt wordt
- hoe snel de data binnenkomt
- Wanneer de data gemaklijker is te transformeren horizontaal dan verticaal

Click on the projected screen to start the question

wooclap

7

De drie hoofd V's



Extra V: Veracity



THE 4 V'S OF BIG DATA

40 ZETTABYTES
of data will be created by
2020, an increase of 300
times from 2005



6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION



Volume

SCALE OF DATA

2.5 QUINTILLION BYTES
of data are created
each day



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



As of 2011, the global size of
data in healthcare was
estimated to be
150 EXABYTES



**30 BILLION
PIECES OF CONTENT**
are shared on facebook
every month



Variety

DIFFERENT
FORMS OF DATA

**4 BILLION +
HOURS OF VIDEO**
are watched on
You Tube each month



4 MILLION TWEETS
are sent per day by about
200 million monthly active
users



The New York Stock
Exchange captures
**1TB OF TRADE
INFORMATION**
during each trading
session



Velocity

ANALYSIS OF
STREAMING DATA

Modern cars have
close to
100 SENSORS
that monitor items such as
fuel level and tire pressure



**1 IN 3 BUSINESS
LEADERS**
don't trust the information
they use to make
decisions



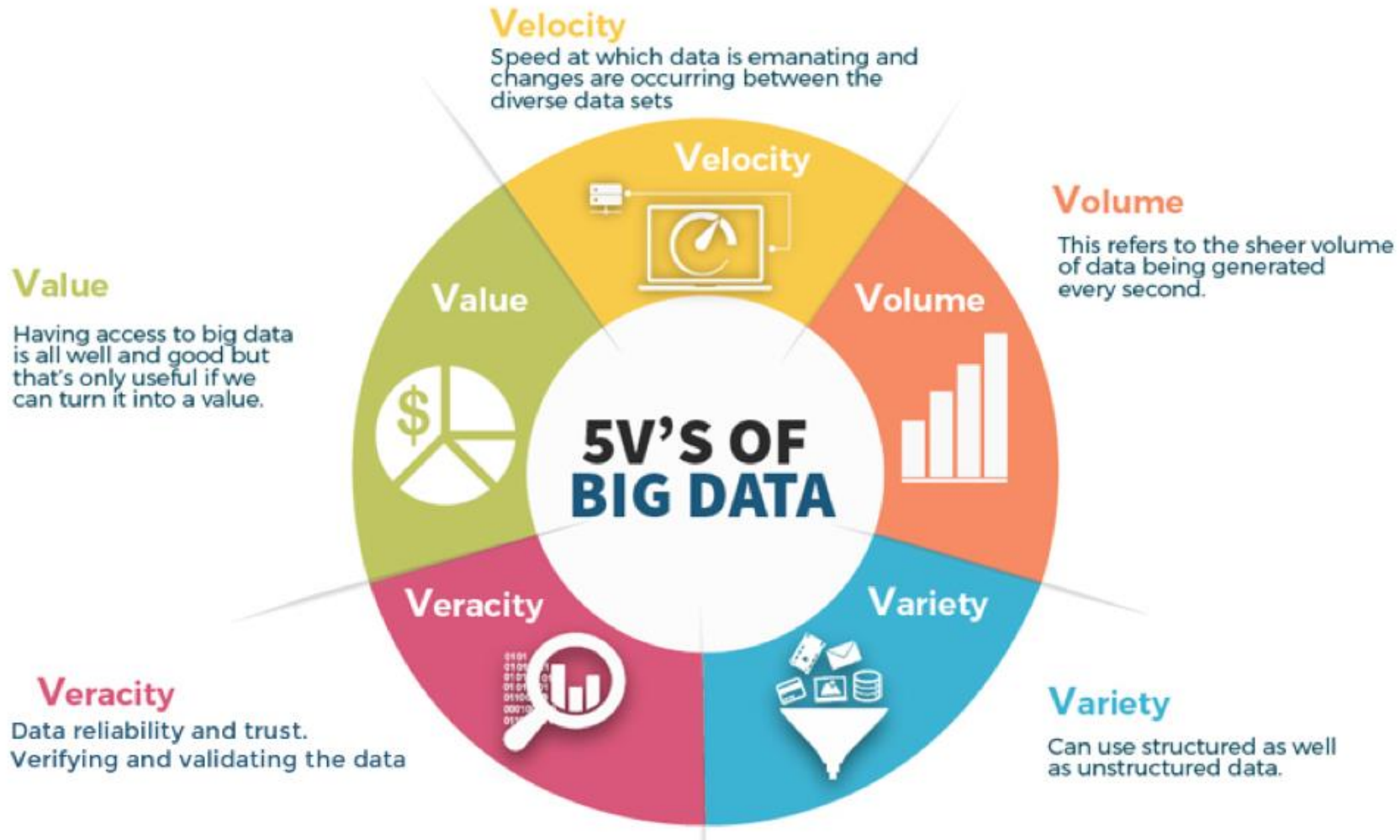
Veracity

UNCERTAINTY
OF DATA

27% OF RESPONDENTS
in one survey were unsure
of how much of data
was inaccurate









Of 5 V's: Value

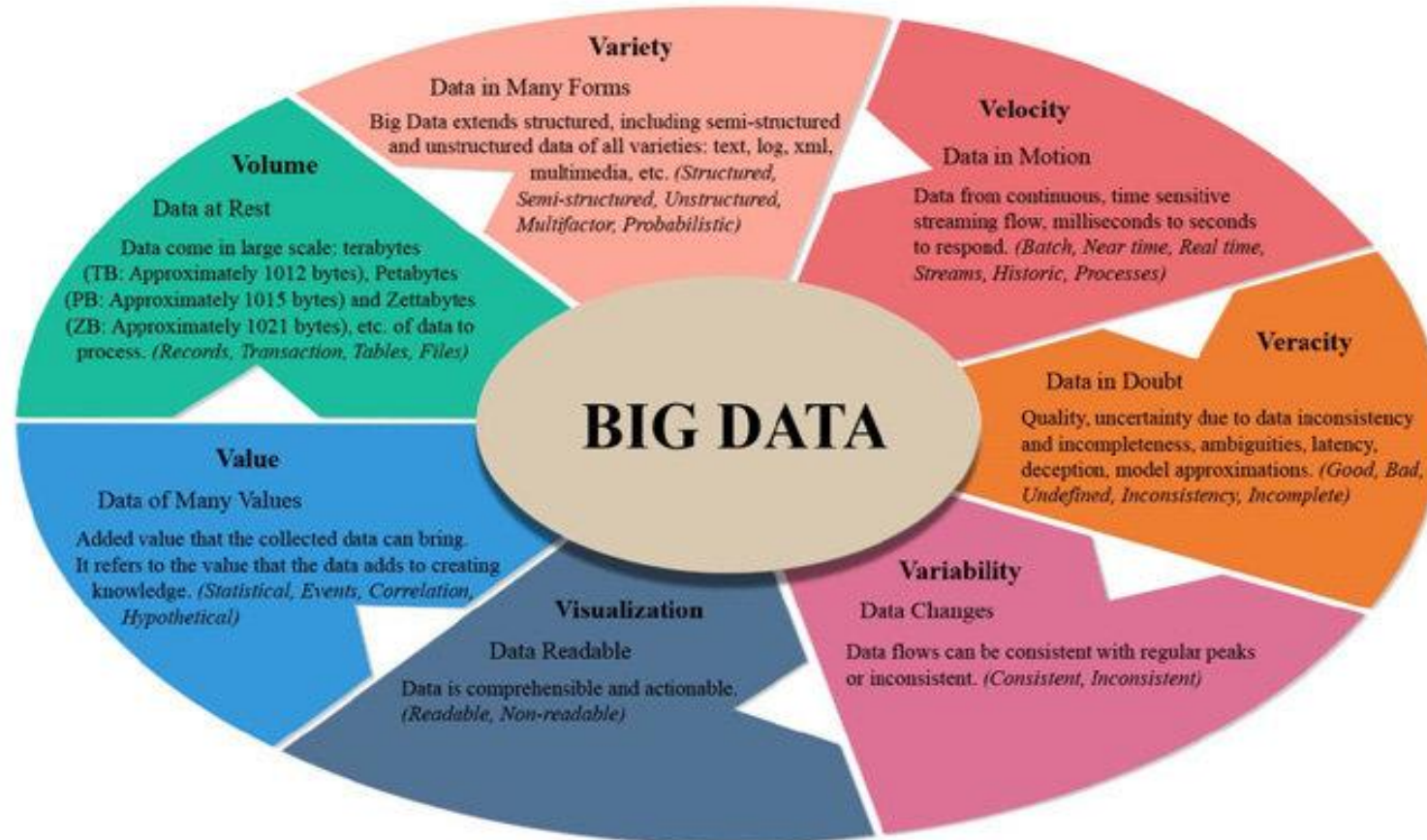


Of 6? Variability

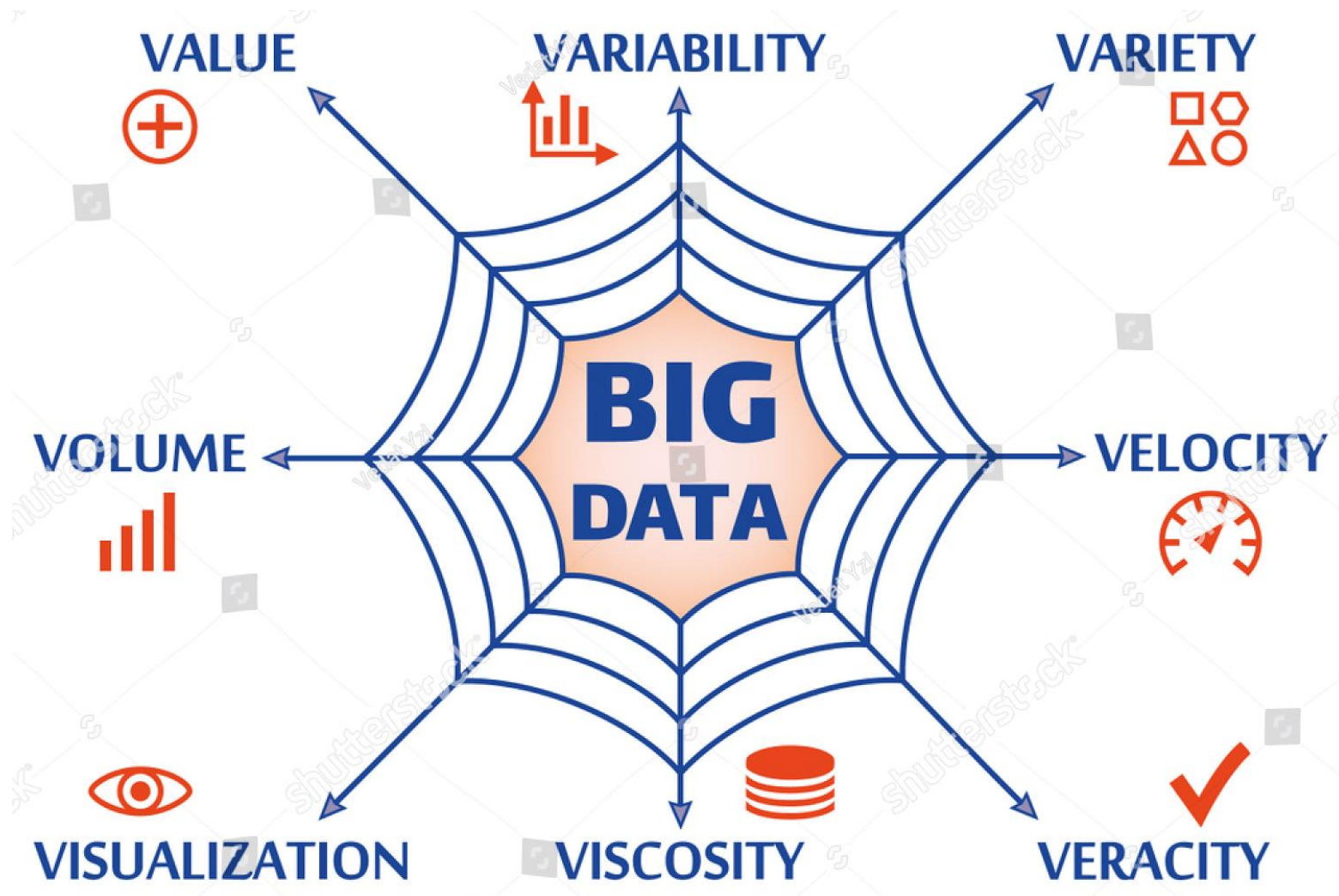
Vanaf hier niet te kennen

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					

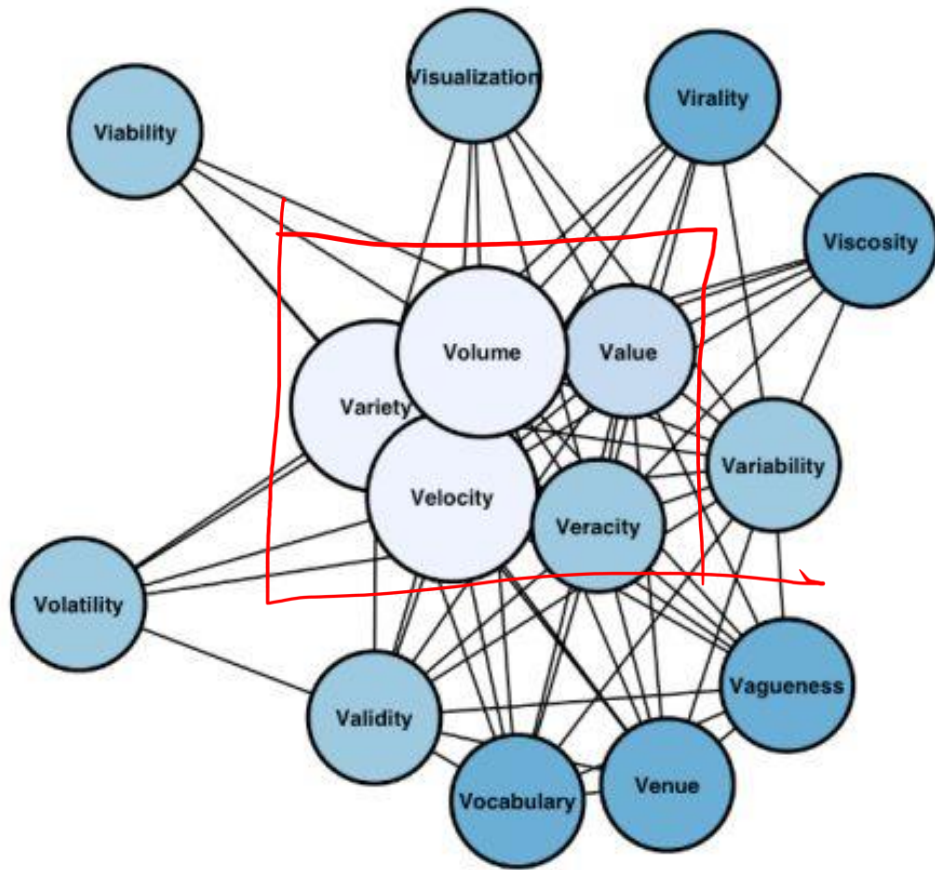
Of 7? Visibility



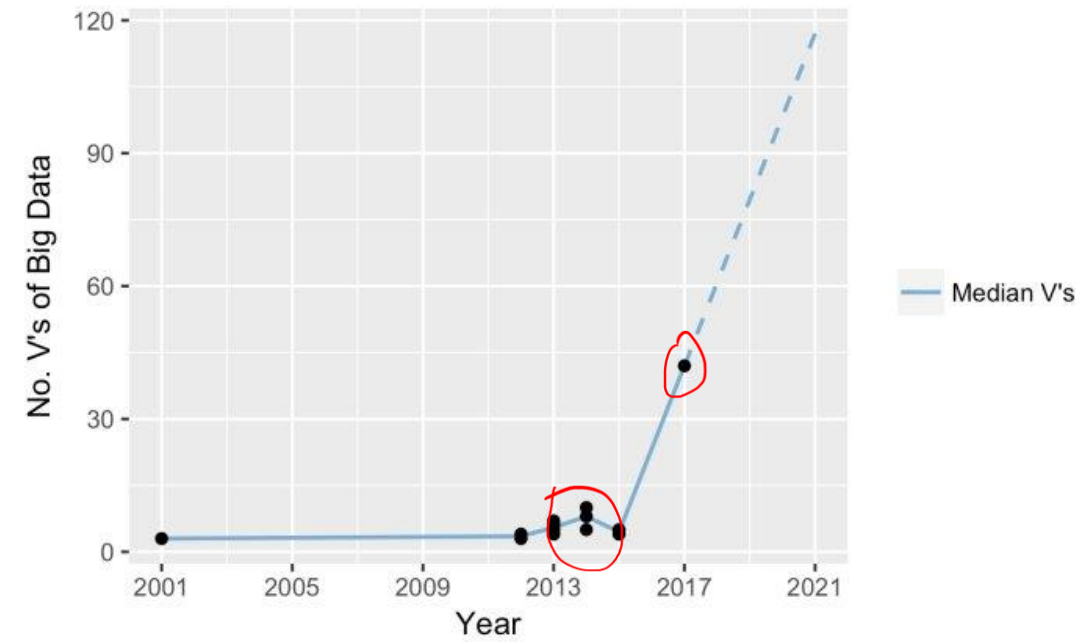
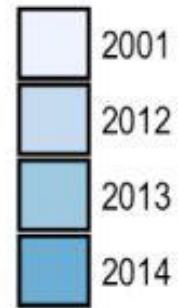
Of 8? Viscosity



10 V's of meer?



First Occurrence





3.

Soorten data

Structured data

- ▣ Vast data formaat in tabel vorm met rijen en kolommen
- ▣ Alle formaten vooraf vastgelegd
- ▣ Excel files, Sql-database, csv, ...



Unstructured data

- ▣ Geen vaste structuur in de data
- ▣ Moeilijk om in te zoeken
- ▣ Foto's, video's, audio, tekst ...

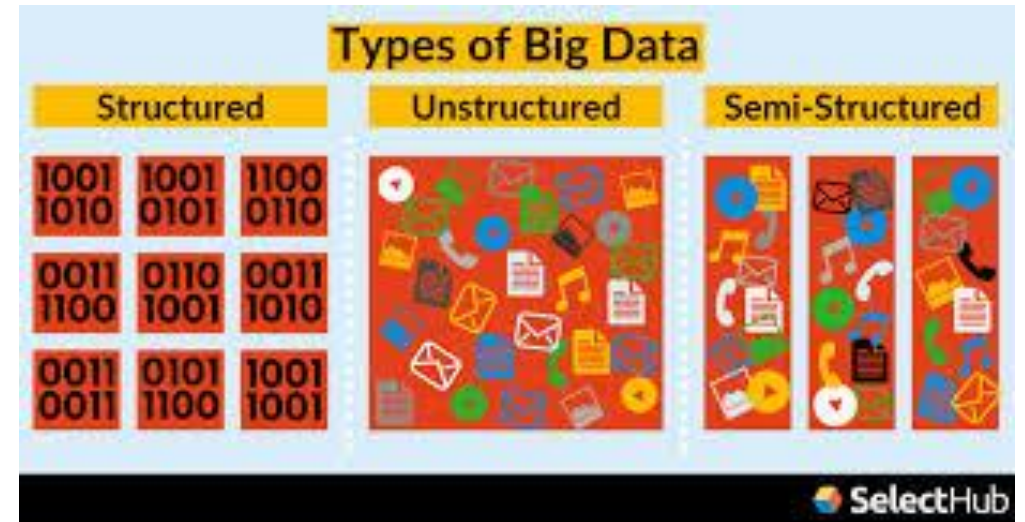
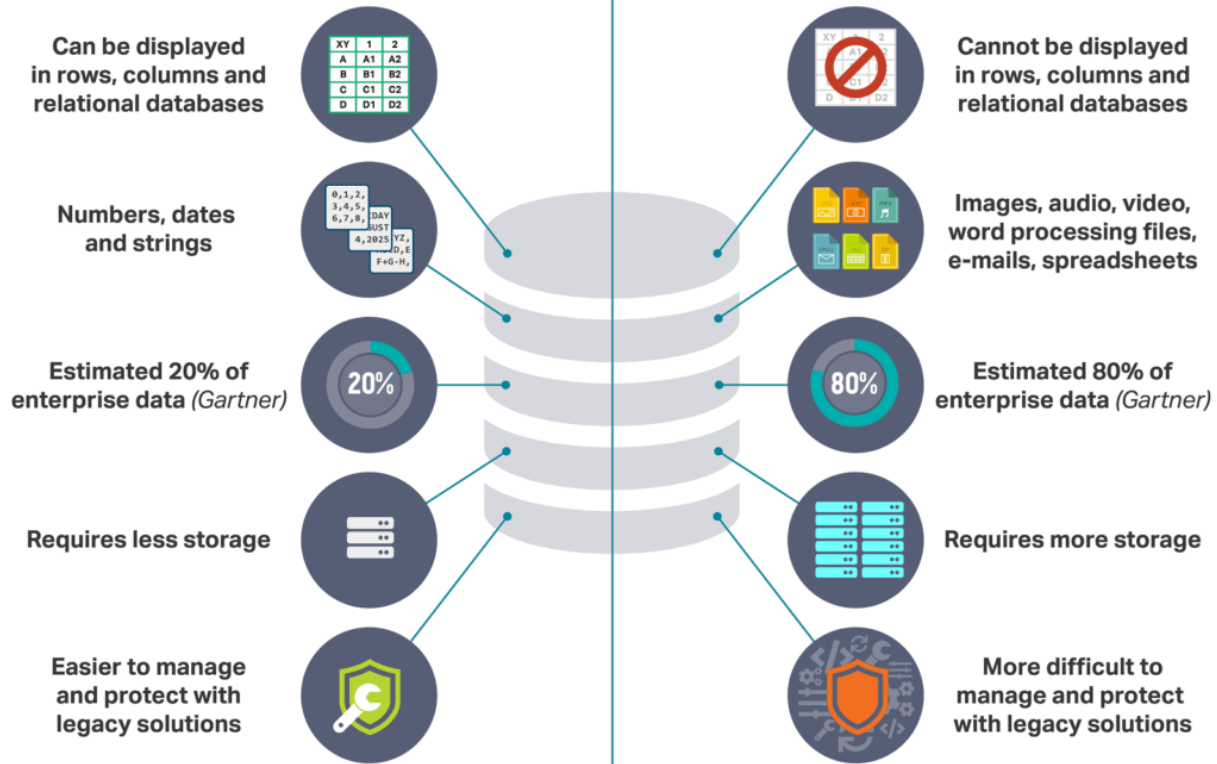


Semi-Structured data

- ▣ Licht-georganiseerde data
- ▣ Tags/metadata verzorgt de structuur
- ▣ Html, xml, json, ...



Structured Data vs Unstructured Data



Door wie is de data geproduceerd?

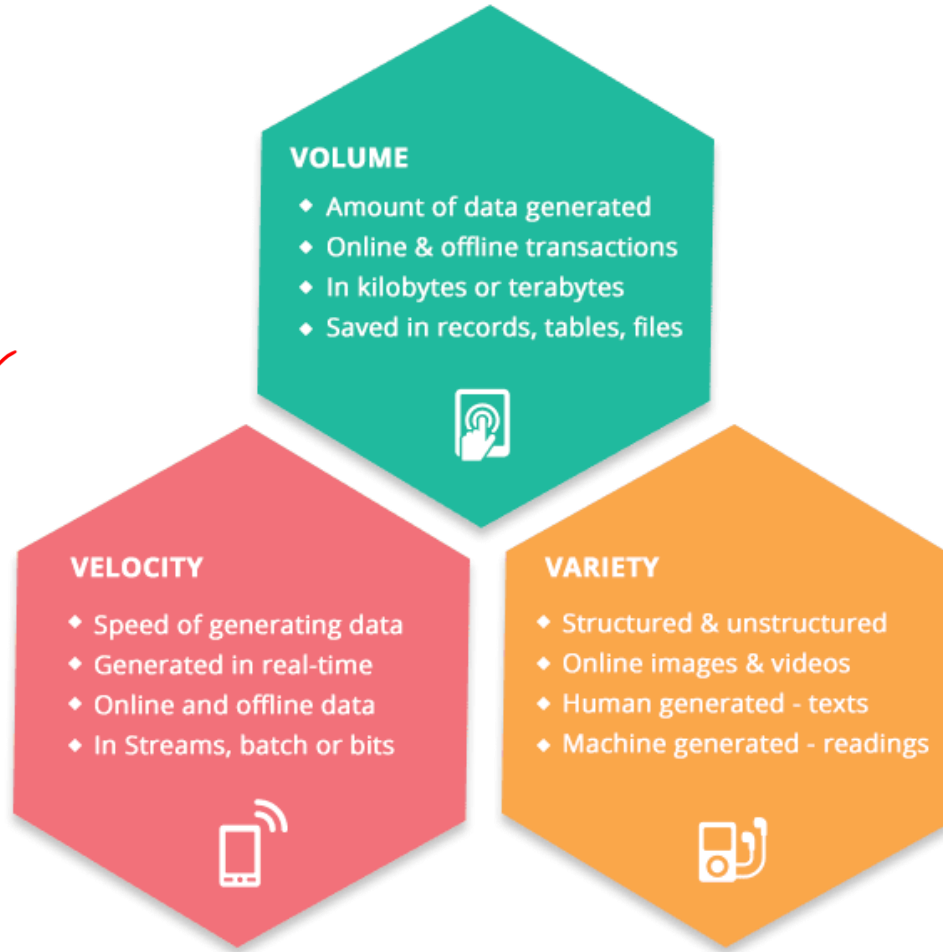
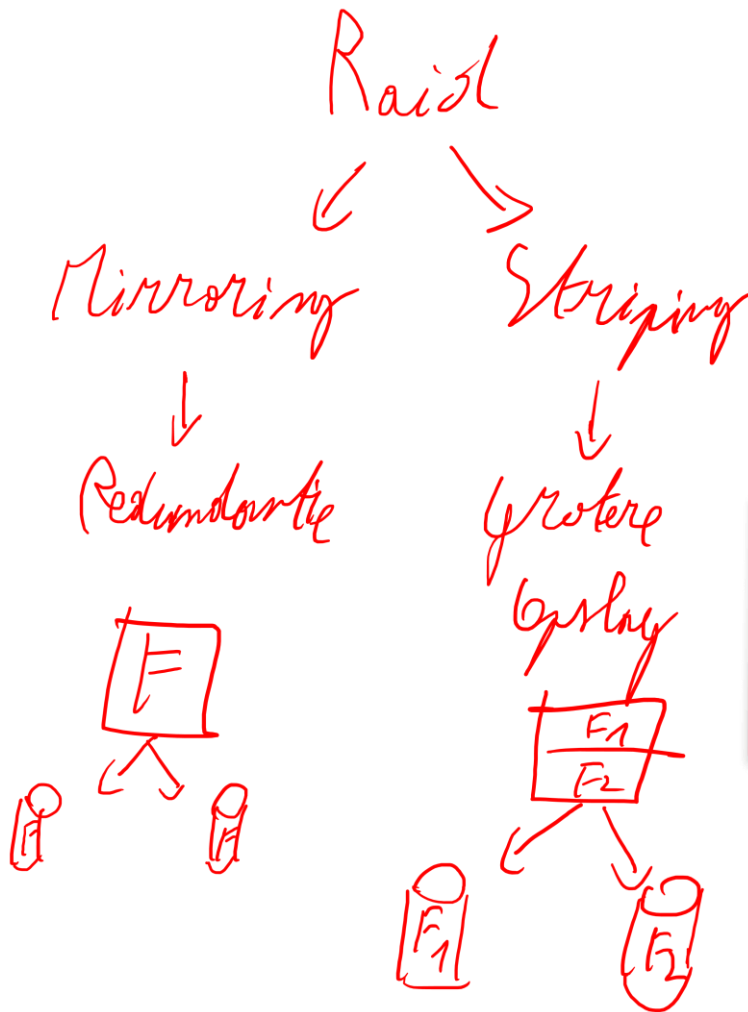




Kritiek op Big Data

- ▣ Op de onderliggende theorie:
 - ▢ Toekomst gelijkaardig aan het verleden
 - ▢ Context afhankelijk
- ▣ V-model focust op schaalbaarheid en rekenkracht, niet op verklaarbaarheid
- ▣ Grote datasets en analyses bestaan reeds decennia, niet zo nieuw als veel denken
- ▣ Buzzword om aandacht te trekken naar je product
- ▣ Privacyschendingen, datalekken, controles, ...

Hoe kan je omgaan met deze problemen?



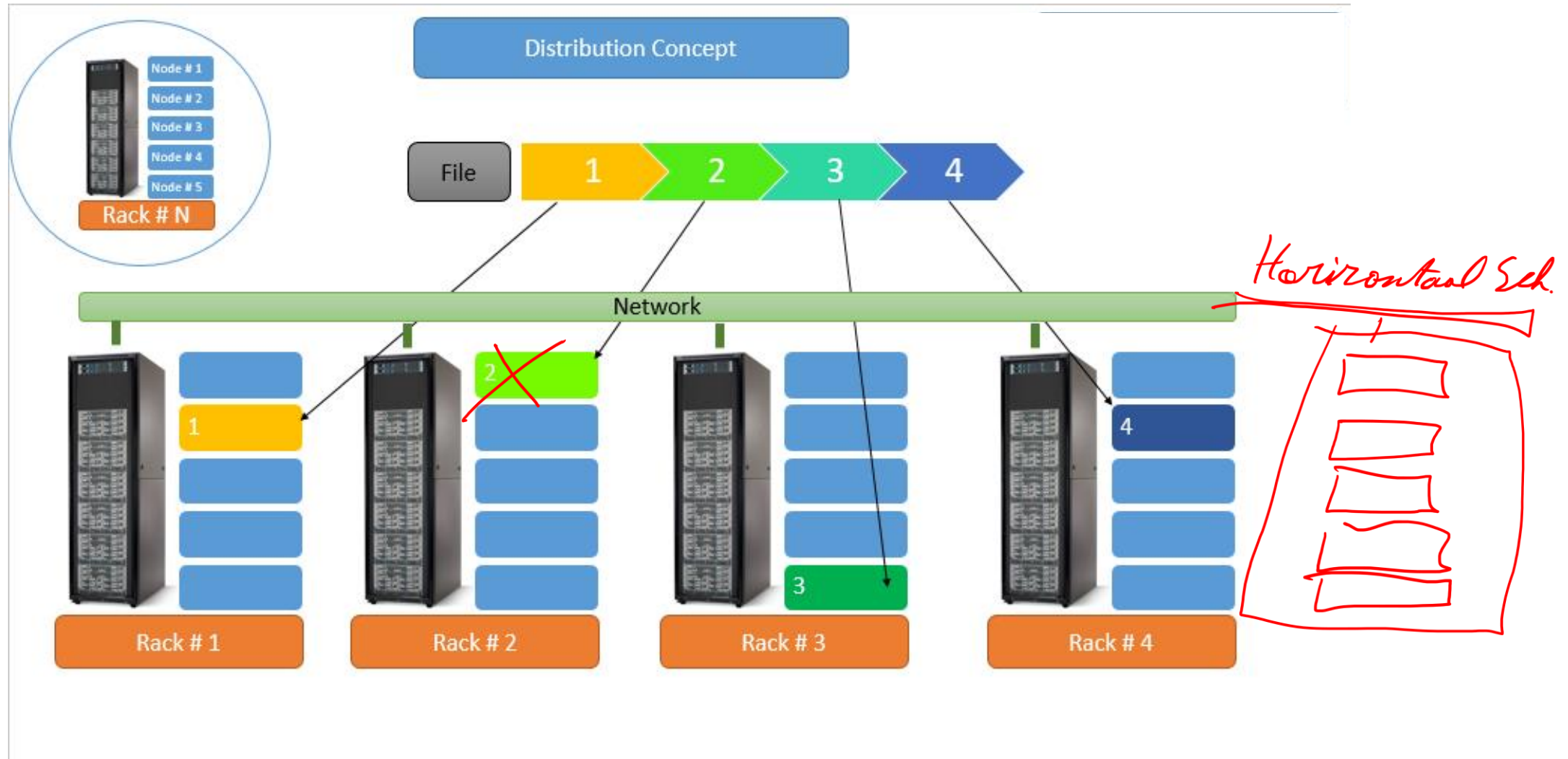


Distributed storage

Van Pc -> Rack -> Datacenter -> Cloud

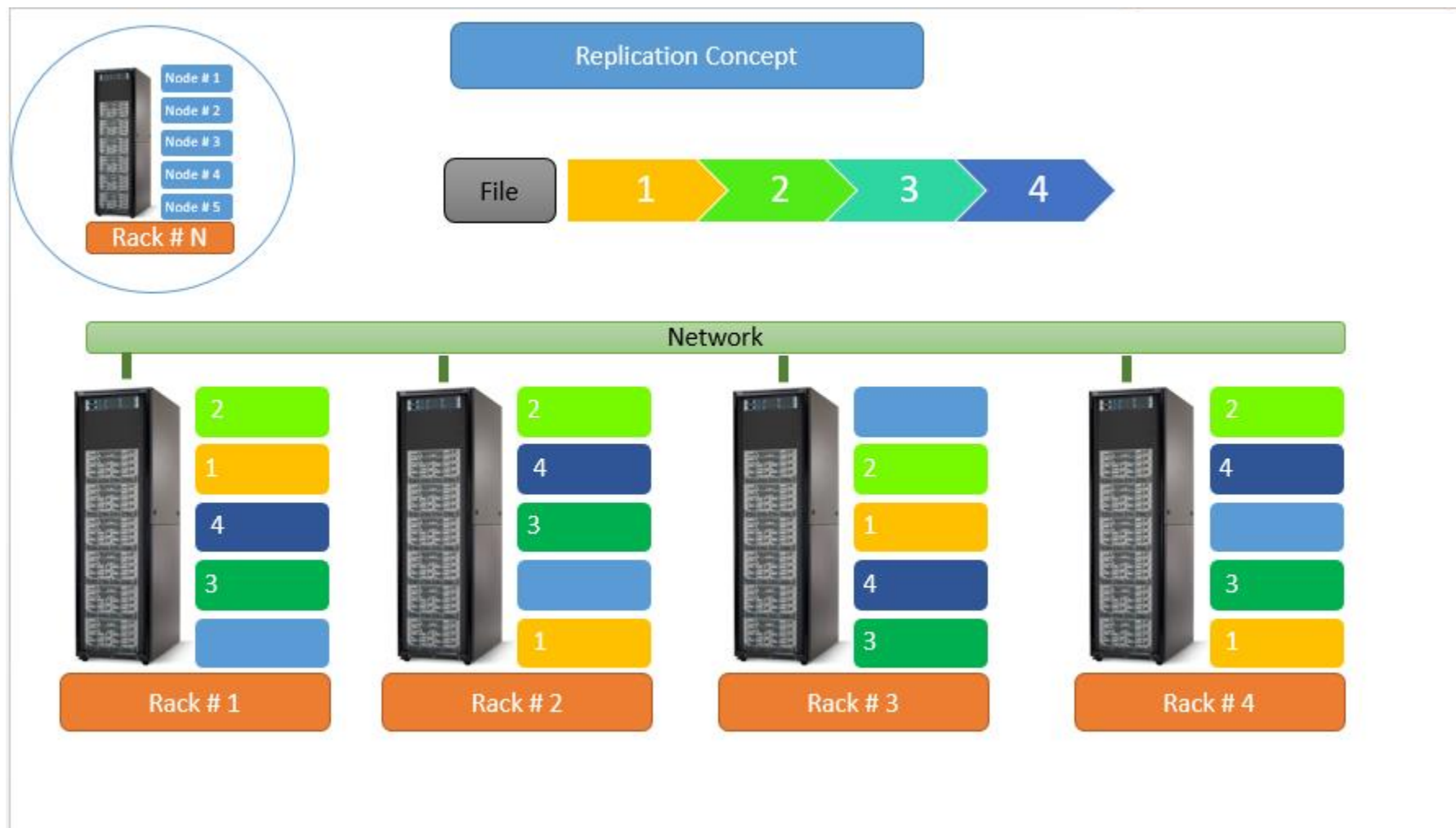


Vert Schaalbaar = 1 PC sterken maken



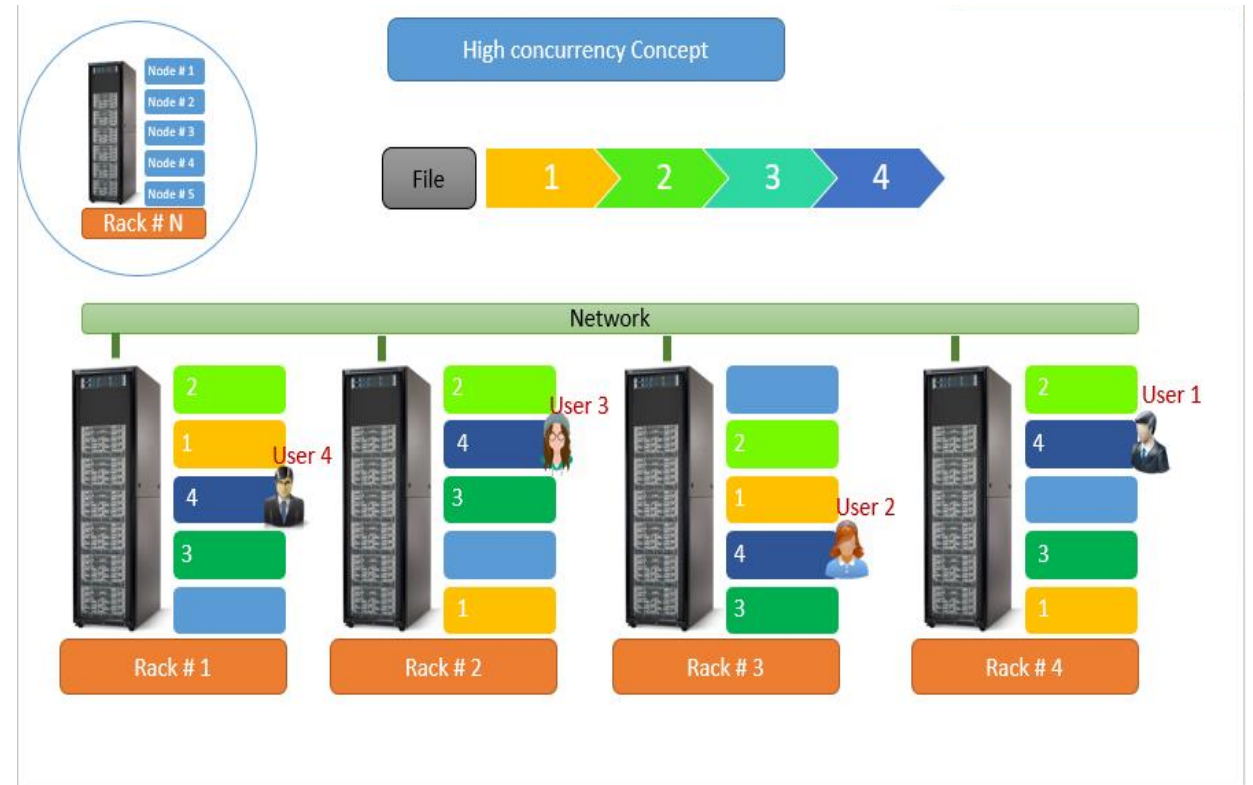
Replication

▣ Fault tolerance



Voordelen

- ▣ Schaalbaar
- ▣ Fout tolerant
- ▣ Nodige rekenkracht ook verdeeld
 - Concurrency
- ▣ Goedkoper
 - Minder gespecialiseerde computers
 - Commodity clusters





Nadelen

- ▣ Meer management van welke data op welke server zit nodig
- ▣ Replication of data maakt het nodig om synchronisatie te doen
 - ▬ Wat bij geografisch verspreide data?
 - ▬ Wat bij uitvallen van server/ datacenter / ...?

Doet dit jullie aan een andere techniek denken?





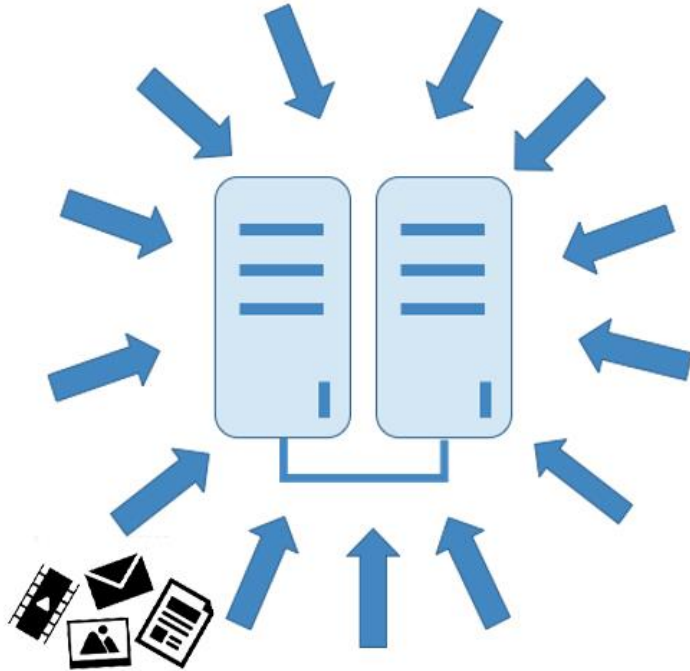
Belangrijke termen voor distributed storage

- ▣ Horizontale schaalbaarheid
- ▣ Duplicatie/replica's
- ▣ Fault-Tolerance



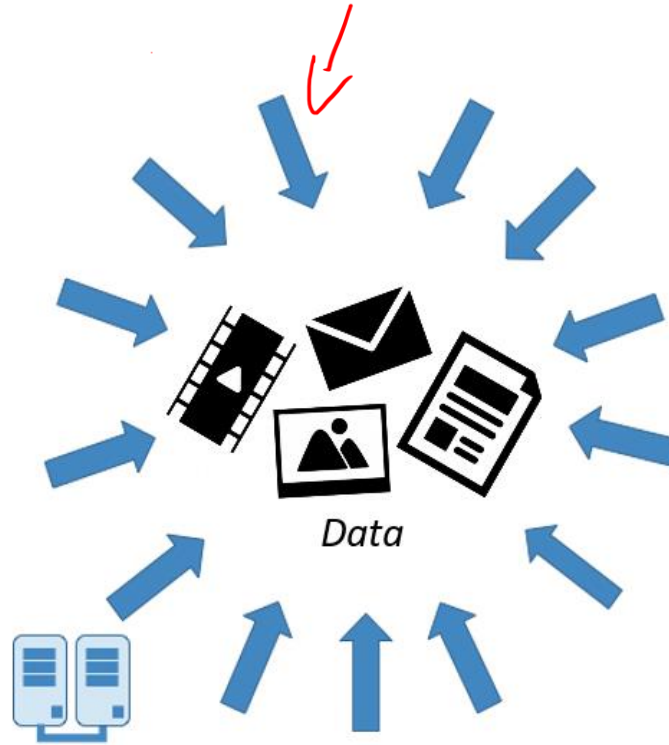
Bring computing to data

Computing to data



Code / Programma niet gemakkelijk te migreren
Rekeneenheid moet krachtig zijn

Data → computing



Code / Programma gemakkelijk te migreren
Volume van data is groot en schaalbaar
Rekenkracht is verspreid
Berekeningen gebeuren asynchroon en verspreid

compute → data

Computing to data

▣ Sneller

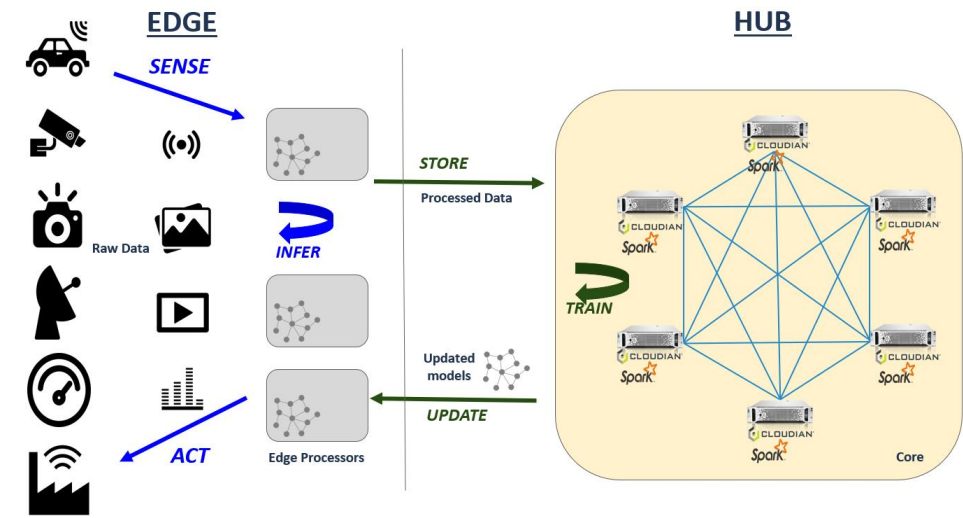
- ▬ Geen transmissietijd voor real-time beslissingen

▣ Goedkoper

- ▬ Verplaatsen data is kostelijk, schaalbaarheid ook belangrijk voor kosten te beperken

▣ Veiliger

- ▬ Data is gemakkelijker te onderscheppen bij verplaatsen



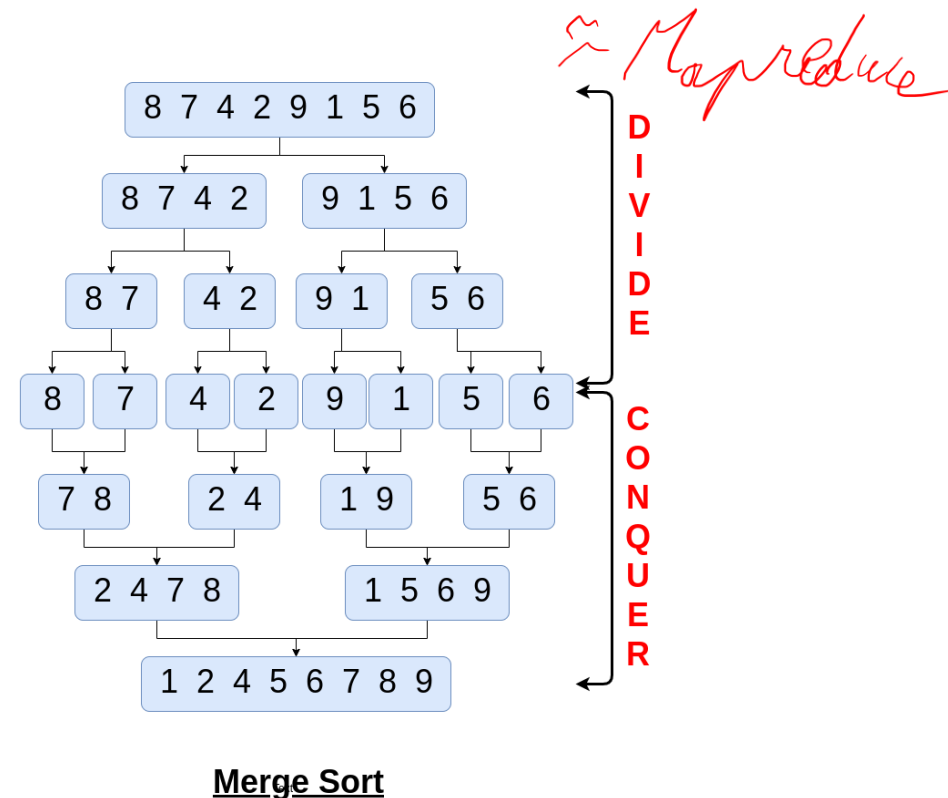
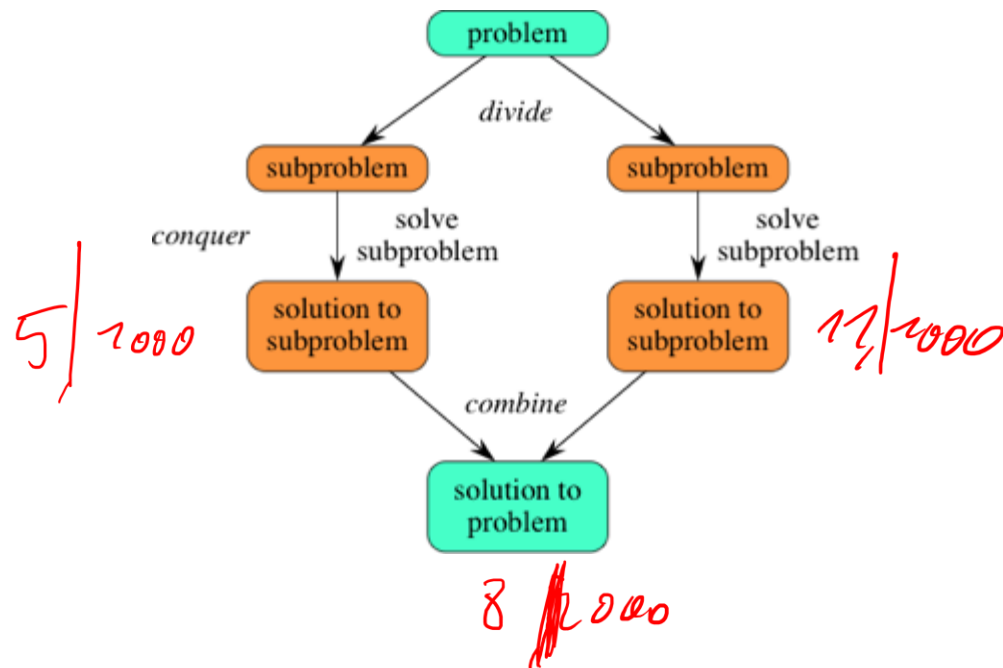


Veel gebruikte concepten

Veel gebruikte concepten bij distributed computing

■ Divide-And-Conquer

- Elke node in de cluster voert maar berekeningen uit op een klein stukje van de data



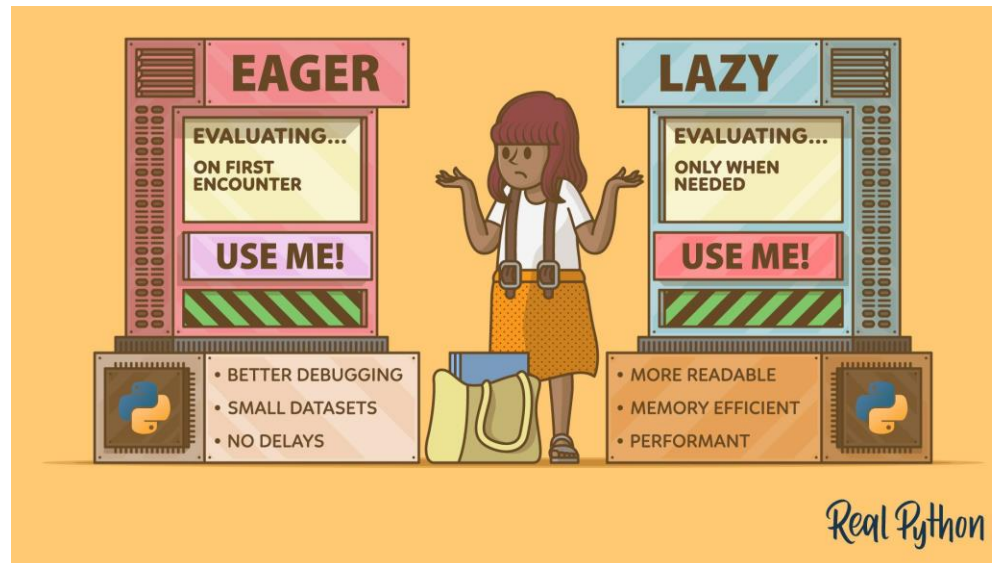
Veel gebruikte concepten bij distributed computing

■ Divide-And-Conquer

■ Lazy-Evaluation

- Je code bouwt een stappenplan op. Enkel als er een eind-resultaat verwacht wordt, worden er berekeningen uitgevoerd
 - Onnodige operaties worden vermeden, optimalisatie van de beschikbare rekencapaciteit

```
a = [  
b = mean(a)  
c = sum(a) ✓  
print(c) ✓
```



Veel gebruikte concepten bij distributed computing

- ▣ Divide-and-conquer
- ▣ Lazy evaluation
- ▣ Eventually consistent
 - Updates van data zijn non-blocking, replica's kunnen tijdelijk nog de oude data bevatten

