

# Big data



Nog niet Gecertificeerde NLT module voor VWO

# Colofon

Deze module is ontwikkeld door leraren in opleiding als eindopdracht voor het vak OCE (Ontwerpen van Communicatieve en Educatieve producten) in de masteropleiding Science Education & Communication van de TU-Delft.

- Tekstschrijvers

Larbi el Yaakoubi  
Julian de Groot  
Bert Lobbezoo

- Inhoudelijk expert

Martin Bruggink (SEC, TU Delft)

- Vakcoach en hoofdredacteur

Wim Sonneveld (SEC, TU Delft)

© mei 2017. Versie 1.1

Aangepaste versies van deze module mogen alleen verspreid worden, indien in dit colofon vermeld wordt dat het een aangepaste versie betreft, onder vermelding van de naam van de auteur van de wijzigingen.

Het auteursrecht op de module berust bij Stichting Leerplan Ontwikkeling (SLO). SLO is derhalve de rechthebbende zoals bedoeld in de hieronder vermelde creative commons licentie.

De auteurs hebben bij de ontwikkeling van de module gebruik gemaakt van materiaal van derden en daarvoor toestemming verkregen. Bij het achterhalen en voldoen van de rechten op teksten, illustraties, enz. is de grootst mogelijke zorgvuldigheid betracht.

Mochten er desondanks personen of instanties zijn die rechten menen te kunnen doen gelden op tekstgedeeltes, illustraties, enz. van deze module, dan worden zij verzocht zich in verbinding te stellen met SLO.

De module is met zorg samengesteld en getest. Landelijk Ontwikkelpunt NLT, Stuurgroep NLT, SLO en auteurs aanvaarden geen enkele aansprakelijkheid voor onjuistheden en/of onvolledigheden in de module. Ook aanvaarden Landelijk Ontwikkelpunt NLT, Stuurgroep NLT, SLO en auteurs geen enkele aansprakelijkheid voor enige schade, voortkomend uit (het gebruik van) deze module.

Voor deze module geldt een

Creative Commons Naamsvermelding-Niet-commercieel-Gelijk delen 3.0 Nederland Licentie

► <http://creativecommons.org/licenses/by-nc-sa/3.0/nl>



Afbeelding voorpagina: [www.sanwen8.cn](http://www.sanwen8.cn)

# Inhoudsopgave

1	De definitie van Big Data .....	4
	Leerdoelen .....	4
	1.1 Van media naar nullen en enen.....	4
	1.2 Wanneer is data groot?.....	13
	1.3 Data uit verschillende bronnen .....	20
2	Toepassingen.....	24
	Leerdoelen .....	24
	2.1 Sneller meten, weten en van feit naar beleid .....	25
	2.2 Duurzaamheid en privacy .....	29
	2.3 De noodzaak en de mensen .....	34
3	Technieken .....	40
	Leerdoelen .....	40
	3.1 Associatieanalyse.....	40
	3.2 Associatieanalyse met Geogabra .....	48
	3.3 Clusteranalyse .....	52

# 1 De definitie van Big Data

Het opslaan van informatie heeft afgelopen tientallen jaren een enorme vlucht genomen. Door de digitalisering van de samenleving zijn enorm veel mogelijkheden ontstaan om betere producten te maken. Tegenwoordig is er zoveel data, dat er nieuwe technieken ontwikkeld moeten worden om de data op een slimme manier te verwerken in iets dat waarde heeft voor een bedrijf of voor de overheid.

Dit hoofdstuk bevat drie paragrafen. De eerste paragraaf gaat in hoe je diverse typen informatie kunt opslaan. De tweede paragraaf gaat dieper in op de definitie van Big Data, en behandelt de vraag wanneer je data echt Big Data kunt noemen. De laatste paragraaf gaat in op de voordelen en uitdagingen van het combineren van verschillende bronnen met informatie.

## Leerdoelen

Na dit hoofdstuk:

- weet je welke waarden een bit kan hebben
- kun je een tekst kunt omzetten in bits
- kun je muziek omzetten in bits
- kun je een afbeelding omzetten in bits
- kun je een video omzetten in bits
- kun je berekenen hoeveel opslag je nodig hebt voor een bepaalde hoeveelheid informatie
- kun je berekenen hoe lang het versturen dan wel ontvangen duurt van een bepaalde hoeveelheid informatie
- kun je berekenen hoe lang het duurt om een bepaalde berekening uit te voeren op de data
- kun je bepalen of data uit verschillende bronnen consistent is
- kun je bepalen welke bron minder betrouwbaar is
- kun je bepalen in hoeverre data van verschillende bronnen zijn te combineren

## 1.1 Van media naar nullen en enen

### Inleiding

Het bewaren van informatie is ontzettend belangrijk. Denk bijvoorbeeld aan het onthouden van een wachtwoord. Je kunt proberen het wachtwoord uit je hoofd te leren, maar als je vijf verschillende wachtwoorden hebt is het al moeilijk om het uit je hoofd te leren. Daarbij komt dat herinneringen door onze hersenen in de loop van de tijd vervormd worden. Bij een wachtwoord heeft dit vervelende gevolgen, want als je één teken verkeerd herinnert klopt het wachtwoord niet meer. Ook het onthouden van financiële gegevens is belangrijk. Mensen hebben veel moeite om financiële gegevens te onthouden, daarom schrijven ze deze gegevens meestal op.

### Voorbeeld A

Zet de volgende tekens om in bits. Gebruik hiervoor de bovenstaande tabel.

- a. 'e'
- b. 'n'
- c. 'en'
- d. 'en dan'

a	00000001
b	00000010
c	00000011
d	00000100
e	00000101
n	00001110
SPATIE	10000000

Tabel 1.1

### Antwoorden

- a. 00000101
- b. 00001110
- c. 00000101 00001110
- d. 00000101 00001110 10000000  
00000100 00000001 00001110

### Voorbeeld B

Een audio-CD slaat geluidsgolven op met 44.100 metingen per seconde. Elke meting heeft 256 verschillende hoogtes.

- a. Hoeveel bits zijn nodig voor één meting?
- b. Hoeveel bits zijn nodig voor een seconde geluid?

### Antwoorden

- a. Er zijn acht bits nodig voor één meting
- b.  $44.100 \text{ metingen} \times 8 \text{ bits per meting} = 352.800 \text{ bits}$ .

Sinds tientallen jaren hebben we computers om informatie op te slaan. Na het ontstaan van de computer is veel nagedacht op welke manier je informatie zo slim mogelijk kunt opslaan. Een 'aan/uit' of 'een/nul' is erg gemakkelijk om op te slaan. Dit kan op veel verschillende manieren. Ook kun je gemakkelijk rekenen met aan en uit. Misschien heb je al gewerkt met elektrische schakelingen met het vak natuurkunde. Door twee schakelingen 'aan' te zetten kun je een vervolg schakeling ook 'aan' zetten. Hierdoor kun je berekeningen uitvoeren.

## Waarom nullen en enen

Een 'aan/uit' of 'een/nul' wordt een bit genoemd. Het woord is een Engelstalig woord. Het is een combinatie van de woorden 'binary' en 'digit'. Het is belangrijk voor de rest van de module dat je de betekenis weet van een bit. Vanaf nu wordt het woord 'bit' gebruikt in plaats van 'nullen en enen'.

De reden om zo'n bit te gebruiken is dat je er gemakkelijk berekeningen mee kunt uitvoeren via elektronische circuits. Een andere reden is dat het gemakkelijk is om extra bits bij te plaatsen. Je hebt vast weleens meegemaakt dat er te weinig vrije ruimte op je computer was. Door bits te gebruiken wordt het gemakkelijker om de ruimte groter te maken.

Ten slotte is de dataopslag met behulp van bits veel gemakkelijker. Denk aan een lamp, deze kan aan staan, of uit. Of denk aan een ouderwetse conducteur, die kan een gaatje in je ticket knippen, zodat de ticket 'uit' wordt geschakeld. Een gat staat dan voor 'uit', en geen gat staat voor 'aan'. Als je de getallen nul tot negen zou moeten opslaan, dan is dit onmogelijk met een lamp. De conducteur zou een tienrittenkaart hiervoor inzetten, maar bij een tienrittenkaart moet je minstens acht of negen keer reizen voordat je de prijs eruit hebt gehaald. Een tienrittenkaart is niet handig als je maar één of twee keer hoeft te reizen. Zo zijn er meerdere mogelijkheden te bedenken om materiaal te gebruiken waarmee je bits kunt opslaan.

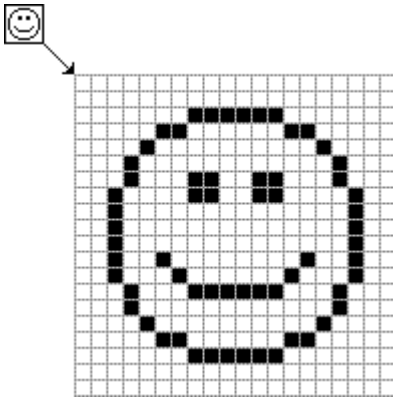
## Tekst representeren met bits

Een tekst bestaat uit meerdere zinnen. Een zin bestaat uit meerdere letters en tekens. Het Nederlandse alfabet heeft 26 letters. Om een tekst te schrijven gebruiken wij verder nog de 'spatie' en 'enter' knop. Verder gebruiken we leestekens, zoals een komma of punt. Een tekst heeft ook hoofdletters. Uitgaand van de Nederlandse taal kunnen er dus 26 verschillende hoofdletters gebruikt worden.

Als we alle mogelijke tekens optellen, komen we aan maximaal 100 tekens die gebruikt kunnen worden in een tekst. Met acht bits is het mogelijk om 256 verschillende cijfers te representeren. Deze cijfers op hun beurt kunnen weer een teken representeren. Om een teken te representeren, hebben we dus acht bits nodig.

## Muziek representeren met bits

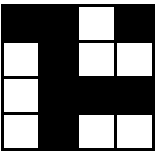
Geluid bestaat uit trillingen. Deze trillingen gedragen zich als een golf. Je kunt geluid vergelijken met een rivier of golvende zee. Hoe hoger



Figuur 1.1 Pixels in een afbeelding

**Voorbeeld C**

Zet de volgende zwart-wit afbeelding om in bits.



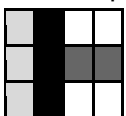
Figuur 1.2

**Antwoord**

0010 1011 1000 1011

**Voorbeeld D**

Zet de volgende grijswaarden afbeelding om in bits. Gebruik twee bits per punt.



Figuur 1.3

**Antwoord**

0100111110100101001001111

de golf, hoe harder het geluid. De natuurkundige eigenschappen van geluidsgolven zijn heel anders dan bij golven van water. Zo komen er op een strand maximaal twee golven per seconde aan. Met geluid is het mogelijk dat er wel tweeduizend geluidsgolven per seconde aankomen bij onze oren.

De hoogte van golven kun je meten. Een golf op het strand is bijvoorbeeld tussen de 25 cm en 200 cm hoog. We weten ondertussen dat acht bits maximaal 255 tekens kan representeren. Op dezelfde manier kunnen acht bits maximaal 255 verschillende golfhoogtes representeren. Bij golven noemen we zo'n meting van de hoogte van een golf of dat een **sample**.

Op een strand kunnen er meerdere golven en dalen per seconde aankomen. Als we een mooie grafiek willen krijgen kunnen we ervoor kiezen per seconde vier keer zo'n meting te verrichten. De frequentie van deze meting is dan vier keer per seconde, oftewel 4 Hz.

Een geluidsgolf heeft ook een hoogte. De manier om deze hoogte te representeren kan precies op dezelfde manier als golven die aankomen op het strand. Daarnaast heeft een geluidsgolf een snelheid. Deze ligt veel hoger dan de snelheid waarmee golven aankomen op het strand. Voor mooi geluid zonder haperingen hebben we minstens 25.000 metingen per seconde nodig. Dit komt overeen met 25.000 hertz (Hz) of 25 kilohertz (kHz).

### Zwart-wit afbeelding representeren met bits

Een zwart-wit afbeelding omzetten naar bits kan als volgt. Allereerst bepalen we hoeveel punten er nodig zijn zodat de afbeelding de juiste kwaliteit heeft. We beginnen links bovenin met deze procedure. We schrijven een nul als een punt 'uit' en is dan 'zwart'. We schrijven één als een punt 'aan' en is dan 'wit'. Daarna gaan we een punt naar rechts, en bepalen weer of de punt een nul of één is. We gaan door naar rechts totdat we de rand van de afbeelding hebben bereikt. Als we niet verder naar rechts kunnen gaan we een rij naar beneden weer helemaal naar links. Dit wordt herhaald totdat we alle punten hebben gehad.

### Grijswaarden afbeelding representeren met bits

Tot nu toe hebben we alleen zwart en wit gebruikt voor de kleuren van een afbeelding. Een 'nul' staat voor zwart, en een 'één' staat voor wit.

0	1

Tabel 1.2

We kunnen grijze kleuren krijgen door zwart en wit met elkaar te mengen. Hierdoor wordt het mogelijk meerdere grijstinten te maken. We kunnen nu de combinatie 00 gebruiken voor zwart, de combinatie 01 gebruiken voor donkergrijs, de combinatie 10 gebruiken voor lichtgrijs en de combinatie 11 gebruiken voor wit.





### Voorbeeld E

Hoeveel bits zijn nodig voor een grijswaarden afbeelding?

- a. met in totaal 1000 punten?
- b. met 1000 punten breed en 500 punten hoog?









### Antwoorden

- a.  $1000 \text{ punten} \times 8 \text{ bits per punt} = 8.000 \text{ bits}$
- b.  $1000 \text{ punten} \times 500 \text{ punten} \times 8 \text{ bits per punt} = 4.000.000 \text{ bits.}$

			
00	01	10	11

Tabel 1.3

Deze stap kunnen we weer herhalen.

							
000	001	010	011	100	101	110	111

Tabel 1.4

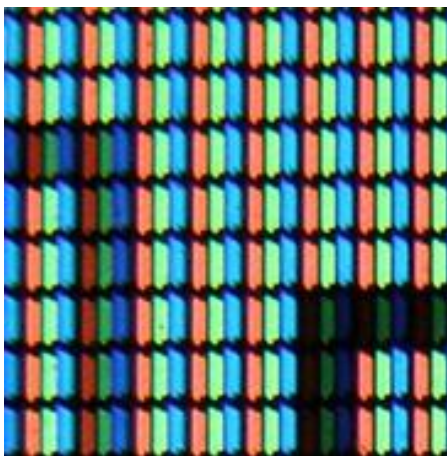
Op den duur zal de overgang tussen de grijs tinten niet meer zichtbaar zijn. Dit gebeurt bij ongeveer acht bits, dus als er 256 verschillende grijs tinten zijn.

## Kleurenafbeelding representeren met bits

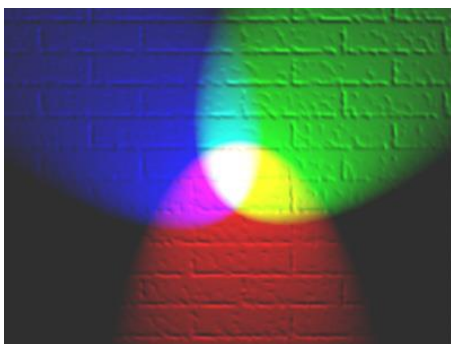
Als je met een vergrootglas naar een beeldscherm kijkt, zie je een patroon van vierkantjes. Elk vierkantje is een pixel. Als je met een sterker vergrootglas kijkt, zie je dat elke pixel op zijn beurt weer bestaat uit drie kleinere pixels: subpixels. Deze subpixels zijn rood, groen en blauw.

Door de kleuren rood, groen en blauw te mengen kun je alle mogelijke kleuren verkrijgen. Vergelijk het met het mengen van verfkleuren. Een beeldscherm mengt ook kleuren, alleen het mengen gebeurt niet met verfkleuren maar met lichtkleuren.













Bij de grijswaarden afbeelding zijn er meerdere bits nodig per pixel om de juiste grijswaarde te representeren. Ditzelfde gaan we nu ook doen voor de kleurenafbeelding.



Figuur 1.4 Een close-up van een beeldscherm



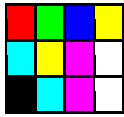
Figuur 1.5 Het mengen van kleuren door middel van licht

												
R	00	01	10	11	00	00	00	00	00	00	00	00
G	00	00	00	00	00	01	10	11	00	00	00	00
B	00	00	00	00	00	00	00	00	00	01	10	11

Tabel 1.5

### Voorbeeld F

Zet de volgende kleuren afbeelding om in bits. Gebruik drie bits per punt.



Figuur 1.6

### Antwoord

100010001110011110

Als we nu twee kleuren gaan mengen, komen we uit bij de volgende representaties:

R	00	01	10	11	00	00	00	00	00	01	10	11
G	00	01	10	11	00	01	10	11	00	00	00	00
B	00	00	00	00	00	01	10	11	00	01	10	11

Tabel 1.6

Als we alle drie de kleuren mengen, komen we uit op de volgende kleuren:

R	00	01	10	11
G	00	01	10	11
B	00	01	10	11

Tabel 1.7

### Voorbeeld G

Jan wil uitrekenen hoeveel opslag hij nodig heeft voor een video van 3 minuten. De video bevat 30 frames per seconde. De video bevat alleen grijswaarden en heeft een resolutie van 640x480 pixels.

- Hoeveel bits zijn er nodig om deze video op te slaan?
- Hoeveel bytes zijn er nodig om deze video op te slaan?

### Antwoorden

- 180 seconden x 30 frames per seconde = 5.400 frames  
5.400 frames x 640 pixels x 480 pixels x 8 bits per pixel = 13.271.040.000 bits
- 13.271.040.000 bits / 8 bits per byte = 1.658.880.000 bytes

## Video representeren met bits

In de vorige paragraaf is uitgelegd hoe je een kleurenafbeelding kunt representeren in bits. Een video bestaat uit meerdere afbeeldingen achter elkaar. Als je deze afbeeldingen snel genoeg afwisselt, ontstaat er een video.

Het menselijk oog kan maximaal 24 verschillen waarnemen per seconde. Als ons oog een video met minder dan 24 verschillende afbeeldingen per seconde ziet, dan ziet dat er naar ons gevoel schokkerig uit. Daarom moet je voor een vloeiende video minstens 24 afbeeldingen per seconde hebben. Deze afbeeldingen heten dan **frames**. Het aantal **frames per seconde** van een video vertelt eigenlijk hoeveel afbeeldingen de video per seconde bevat.



## Vragen en opdrachten

De vragen en opdrachten zijn te maken op drie verschillende niveaus. Kies het niveau waarvan je denkt dat je dat momenteel hebt. Als blijkt dat het niveau te moeilijk of juist te gemakkelijk is kun je altijd beginnen aan een ander niveau.

### Niveau 1

1.

Hoeveel waardes

- Kan één bit bevatten?
- Kunnen twee bits bevatten?

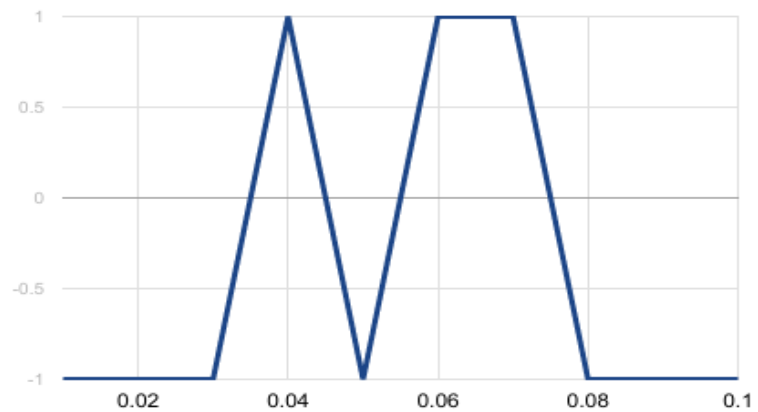
2.

Zet de volgende tekst om in bits. Gebruik hiervoor Tabel 1.1.

- 'a'
- 'b'

3.

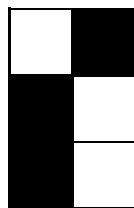
Zet de volgende muziek om in bits.



*Figuur 1.7*

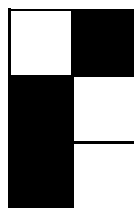
4.

Zet de volgende afbeelding om in bits.

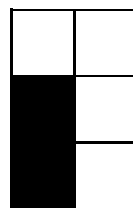


5.

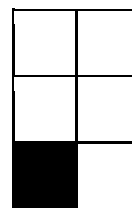
Zet de volgende video om in bits.



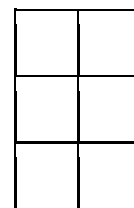
*Sample 1*



*Sample 2*



*Sample 3*



*Sample 4*

## Niveau 2

1.

Hoeveel waardes kunnen acht bits bevatten?

2.

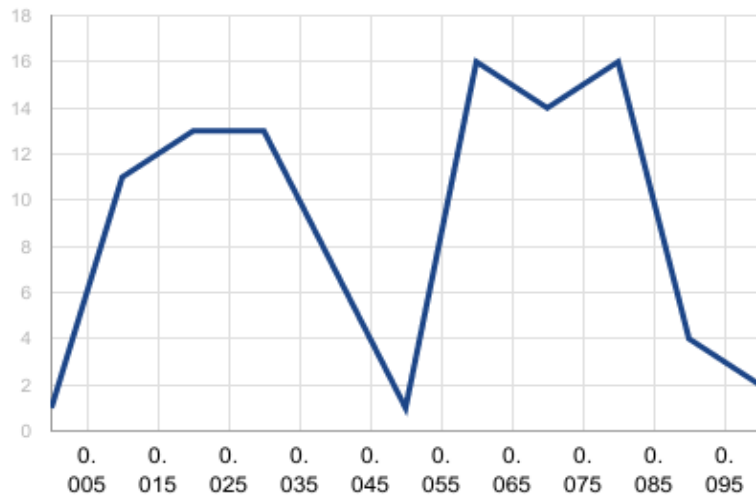
Zet de volgende tekst om in bits. Gebruik hiervoor Tabel 1.1.

a. 'abc'

b. 'a b c'

3.

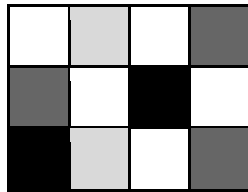
Zet de volgende muziek om in bits. Gebruik twee bits per pixel.



*Figuur 1.8*

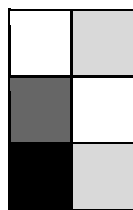
4.

Zet de volgende afbeelding om in bits.

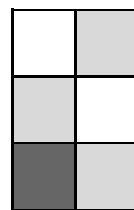


5.

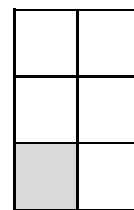
Zet de volgende video om in bits. Gebruik twee bits per pixel.



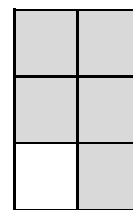
*Sample 5*



*Sample 6*



*Sample 7*



*Sample 8*

### Niveau 3

1.

Hoeveel waardes kunnen 16 bits bevatten?

2.

Zet de volgende tekst om in bits. Gebruik hiervoor de ASCII codering, zie Tabel 1.8.

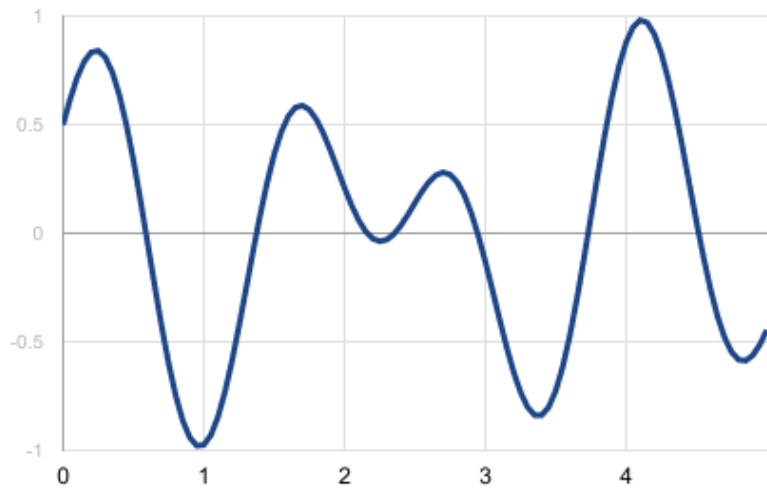
a. 'U-bocht'

b. '1+1=2'

c. 'LëêßTékèñs'

3.

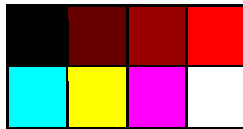
Zet de volgende muziek om in bits.



Figuur 1.9

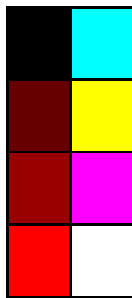
4.

Zet de volgende afbeelding om in bits. Gebruik drie bits per pixel.

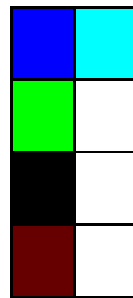


5.

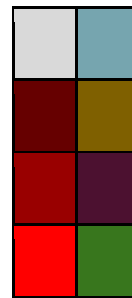
Zet de volgende video om in bits. Gebruik drie bits per pixel.



Sample 9



Sample 10



Sample 11

## Table of ASCII characters 0-127 (x0000-x007f)

000	--0	--1	--2	--3	--4	--5	--6	--7	010	--0	--1	--2	--3	--4	--5	--6	--7
00-	<u>N</u> <sub>U</sub> <sub>L</sub>	<u>S</u> <sub>O</sub> <sub>H</sub>	<u>S</u> <sub>T</sub> <sub>X</sub>	<u>E</u> <sub>T</sub> <sub>X</sub>	<u>E</u> <sub>O</sub> <sub>T</sub>	<u>E</u> <sub>N</sub> <sub>Q</sub>	<u>A</u> <sub>C</sub> <sub>K</sub>	<u>B</u> <sub>E</sub> <sub>L</sub>	10-	@	A	B	C	D	E	F	G
01-	<u>B</u> <sub>S</sub>	<u>H</u> <sub>T</sub>	<u>L</u> <sub>F</sub>	<u>V</u> <sub>T</sub>	<u>F</u> <sub>F</sub>	<u>C</u> <sub>R</sub>	<u>S</u> <sub>O</sub>	<u>S</u> <sub>I</sub>	11-	H	I	J	K	L	M	N	O
02-	<u>D</u> <sub>L</sub> <sub>E</sub>	<u>D</u> <sub>C</sub> <sub>1</sub>	<u>D</u> <sub>C</sub> <sub>2</sub>	<u>D</u> <sub>C</sub> <sub>3</sub>	<u>D</u> <sub>C</sub> <sub>4</sub>	<u>N</u> <sub>A</sub> <sub>K</sub>	<u>S</u> <sub>Y</sub> <sub>N</sub>	<u>E</u> <sub>T</sub> <sub>B</sub>	12-	P	Q	R	S	T	U	V	W
03-	<u>C</u> <sub>A</sub> <sub>N</sub>	<u>E</u> <sub>M</sub>	<u>S</u> <sub>U</sub> <sub>B</sub>	esc	<u>F</u> <sub>S</sub>	<u>G</u> <sub>S</sub>	<u>R</u> <sub>S</sub>	<u>U</u> <sub>S</sub>	13-	X	Y	Z	[	\	]	^	_
04-		!	"	#	\$	%	&	'	14-	`	a	b	c	d	e	f	g
05-	(	)	*	+	,	-	.	/	15-	h	i	j	k	l	m	n	o
06-	0	1	2	3	4	5	6	7	16-	p	q	r	s	t	u	v	w
07-	8	9	:	;	<	=	>	?	17-	x	y	z	{		}	~	<u>D</u> <sub>E</sub> <sub>L</sub>

Tabel 1.8 De ASCII tekens

## 1.2 Wanneer is data groot?

### Inleiding

Big Data betekent letterlijk 'grote data'. Maar wanneer is groot nu eigenlijk groot? Voor een mier is een takje groot, maar een olifant kan met bomen spelen. Wat vroeger als veel informatie werd beschouwd, is een lachertje voor moderne computers.

Data is pas werkelijk Big Data als er nieuwe manieren en technieken nodig zijn om geld te verdienen met behulp van deze data. Deze strikte definitie van Big Data wordt over het algemeen niet gebruikt als men over Big Data spreekt.

Data pas echt groot als het voldoet aan drie kenmerken. Deze kenmerken zijn **hoeveelheid**, **snelheid**, en **diversiteit**. Deze drie kenmerken worden uitgelegd in deze paragraaf.

### De hoeveelheid data

Het eerste kenmerk van Big Data is de **hoeveelheid** data. Bij natuurkunde heb je vast al kennis gemaakt met de voorvoegsels 'kilo' en 'mega'. Deze worden bij computers ook gebruikt om duizenden of miljoenen bytes op een gemakkelijk manier op te schrijven.

Er is alleen een probleem bij het gebruiken van bytes om deze om te zetten naar duizendtallen. Als je namelijk gebruik maakt van binair rekenen, kom je niet uit op 1.000 bytes = 1 kilobyte, maar 1.024 bytes = 1 kilobytes. Daarom is afgesproken dat bij binair rekenen je niet 'kilo' en 'mega' gebruikt, maar 'kibi' en 'mebi'. In de praktijk wordt nog steeds gebruik gemaakt van 'kilo' en 'mega', maar dit is dus niet de exacte hoeveel opslag die er daadwerkelijk is. Bekijk Tabel 1.9 maar eens.

Naam	Waarde	Binair	Waarde
Byte	$1 = 10^0$	Byte	$1 = 2^0$
Kilobyte	$1.000 = 10^3$	Kibibyte	$1.024 = 2^{10}$
Megabyte	$1.000.000 = 10^6$	Mebibyte	$1.048.576 = 2^{20}$
Gigabyte	$1.000.000.000 = 10^9$	Gibibyte	$1.073.741.824 = 2^{30}$
Terabyte	$1.000.000.000.000 = 10^{12}$	Tebibyte	$1.099.511.627.776 = 2^{40}$
Petabyte	$1.000.000.000.000.000 = 10^{15}$	Pebibyte	$1.125.899.906.842.624 = 2^{50}$
Exabyte	$1.000.000.000.000.000.000 = 10^{18}$	Exbibyte	$1.152.921.504.606.846.976 = 2^{60}$

Tabel 1.9 De naam van verschillende grote hoeveelheden data

Media	Bytes nodig
Letter	1
Zwart/wit afbeelding	1/8
Grijswaarden afbeelding	1 byte
Kleuren afbeelding	3 per pixel
Video	3 per pixel per frame

Tabel 1.10 Benodigde hoeveelheid pixels per media

Je ziet dat er weinig verschil is. Daarentegen bij hogere machten, bij 'peta' en 'exa' loopt het verschil op in meer dan 10%! Als je een harde schijf koopt van 1 terabyte, bevat deze in werkelijkheid geen  $1.099.511.627.776 = 2^{40}$  maar  $1.000.000.000.000 = 10^{12}$  bytes. Dit is een significant verschil. Omdat in het in het taalgebruik is ingeburgerd en omdat je op deze manier minder kunt verkopen voor dezelfde prijs wordt nu nog steeds over terabytes gesproken in plaats van de wiskundige tebibytes.

Onderstaande tabel bevat een overzicht hoeveel byte je ongeveer nodig hebt om verschillende type data te representeren in bytes.

#### Voorbeeld A

Sanne wil berekenen hoeveel dataopslag hij nodig heeft om alle namen de gebruikers van Facebook op te slaan. Ga ervanuit dat Facebook 1.86 miljard gebruikers heeft. Ga er verder vanuit dat de lengte van de naam van gebruiker gemiddeld 16 tekens bevat.

- Geef het antwoord in gigabytes
- Geef het antwoord in gibibytes

#### Antwoorden

- $1.860.000.000 \times 16 = 29.760.000.000$  byte = 29.760.000 kilobyte = 29.760 megabyte = 29.8 gigabyte
- $b. 1.860.000.000 \times 16 = 29.760.000.000$  byte  
 $29.760.000.000 \text{ byte} / 1.024 = 29.062.500$  kibibyte  
 $29.062.500 \text{ kibibyte} / 1.024 = 28.382$  mebibyte

### Voorbeeld B

Henk heeft een camera opgehangen om zijn buurman in de gaten te houden. Hij wil de beelden graag 200 dagen lang bewaren.

De camera slaat 4 beelden per seconde op met een resolutie van 1024 x 768 pixels.

- Bereken hoeveel frames de opslag bevat.
- Bereken de dataopslag in terabytes als de camera zwart-wit beelden zou opnemen
- Bereken de dataopslag in terabytes als de camera grijswaarden beelden zou opnemen
- Bereken de dataopslag in tebibytes als de camera kleurenbeelden zou opnemen

### Antwoorden

- $200 \text{ dagen} \times 24 \text{ uur} \times 60 \text{ minuten} \times 60 \text{ seconden} = 17.280.000$  seconden  
 $17.280.000 \text{ seconden} \times 4 \text{ frames per seconde} = 69.120.000$  frames.
- Een frame bevat  $1.024 \times 768 = 786.432$  pixels.  
 $786.432 \text{ pixels} \times 0.125 \text{ byte per pixel} = 98.304 \text{ bytes}$ .  
 $98.304 \text{ bytes} \times 69.120.000 \text{ frames} = 6.794.772.480.000 \text{ bytes}$   
 $6.794.772.480.000 \text{ bytes} = 6.794.772.480 \text{ kilobytes} = 6.794.773$  megabytes = 6.795 gigabyte = 6.8 terabytes.
- $6.794.772.480.000 \text{ pixels} \times 8 \text{ byte per pixel} = 54.358.179.840.000$  bytes  
 $54.358.179.840.000 \text{ bytes} = 54.358.179.840 \text{ kilobytes} =$   
 $54.358.180 \text{ megabytes} = 54.358 \text{ gigabytes} = 54.4 \text{ terabytes}$
- $6.794.772.480.000 \text{ pixels} \times 24 \text{ bytes per pixel} =$   
 $163.074.539.520.000 \text{ bytes}$   
 $163.074.539.520.000 \text{ bytes} / 1.024 = 159.252.480.000 \text{ kibibytes}$   
 $159.252.480.000 \text{ kibibytes} / 1.024 = 155.520.000 \text{ mebibytes}$   
 $155.520.000 \text{ mebibytes} / 1.024 = 151.875 \text{ gibibytes}$   
 $151.875 \text{ gibibytes} / 1.024 = 148,3 \text{ tebibytes}$

## De snelheid om data op te vragen

Het tweede kenmerk van Big Data is de **snelheid** om de data op te vragen. Bekijk het onderstaande voorbeeld.

### Voorbeeld C

Sofie wil de gemiddelde leeftijd berekenen van alle Facebook gebruikers. Ga ervanuit dat Facebook 1.86 miljard gebruikers heeft. Ga er verder vanuit dat er 8 bits nodig zijn per leeftijd.

- a. Bereken hoeveel bytes er nodig zijn voor de dataopslag van de leeftijden

De harde schijf waarop de data is opgeslagen heeft een leessnelheid van 10 megabytes per seconde.

- b. Hoe lang duurt het voordat alle leeftijden zijn ingelezen?

### Antwoorden

- a.  $1.86 \text{ miljard bytes} = 1.86 \text{ miljoen kilobytes} = 1.860 \text{ megabytes} = 1,86 \text{ gigabytes.}$
- b.  $1.860 \text{ megabytes} / 10 \text{ megabytes per seconde} = 186 \text{ seconden.}$

Een smartphone of computer heeft maar een bepaalde ruimte om data op te slaan. Uit een eerder besproken voorbeeld blijkt dat er 150 tebibytes nodig zijn om voor 200 dagen kleurenvideo op te slaan. Op het moment dat deze module is geschreven kan één harde schijf maximaal 10 tebibytes bevatten. Er zijn dus 15 harde schijven nodig om zoveel data op te slaan. Deze passen niet in een enkele PC. Je kunt wel een netwerk bouwen met meerdere PC's, waarbij elke PC een gedeelte van de opslag voor zijn rekening neemt. Bekijk nu het volgende voorbeeld.

### Voorbeeld D

Chris wil de gemiddelde kleur weten van alle pixels van 200 dagen kleurenvideo. Deze kleurenvideo heeft een opslag nodig van 150 tebibytes. Hiervoor heeft hij 15 PC's ingezet, elk met een opslagcapaciteit van 10 tebibytes.

De PC waarop Chris de berekening uitvoert, heeft een verbinding met deze 15 PC's, en het is mogelijk om 500 mebibytes per seconde te transporteren over deze verbinding.

- a. Hoelang duurt het voordat alle data ontvangen is op de PC waarop Chris de berekening uitvoert? Ga ervanuit dat deze PC slechts één verbinding tegelijk kan maken.
- b. Hoelang duurt het voordat alle data ontvangen is op de PC waarop Chris de berekening uitvoert? Ga er nu vanuit dat deze PC vijftien verbindingen tegelijk kan maken.

### Antwoorden

- a. In totaal moet er 150 tebibytes getransporteerd worden. Dit gaat met een snelheid van 0,0005 tebibytes per seconde. Er zijn dus  $150 / 0,0005 = 300.000$  seconden nodig om alle data te transporteren. Dit komt overeen met  $300.000 \text{ seconden} / 3.600 \text{ seconden per uur} = 83,3$  uur.
- b. In totaal moet er per verbinding 10 tebibytes getransporteerd worden. Dit gaat met een snelheid van 0,0005 tebibytes per seconde. Er zijn dus  $10 / 0,0005 = 20.000$  seconden nodig om alle data te transporteren naar de PC waarop de berekening kan plaatsvinden. Dit komt overeen met  $20.000 \text{ seconden} / 3.600 \text{ seconden per uur} = 5,6$  uur.



Chris heeft bedacht dat het veel slimmer is om de gemiddelde kleur te berekenen op de PC waarop de videobeelden zelf staan. Vervolgens wordt de waarde gestuurd via de verbinding naar de PC waarop Chris de berekening heeft gestart.

De reden dat dit veel sneller is komt onder er geen gebruik meer gemaakt hoeft te worden van een trage netwerkverbinding Een PC Dit komt omdat ze niet gebruik hoeven te maken van een trage netwerkverbinding. De data kan direct in de PC getransporteerd worden naar de centrale berekeningseenheid in een PC. Een ander woord voor dit onderdeel is de processor van de computer, of de CPU (Central Processing Unit).

**Voorbeeld E**

Stel dat je dus 15 PC's hebt. De snelheid van de processor om het gemiddelde te berekenen van bits is 250 gebibytes per seconde.

- a. Hoeveel tijd kost het een PC om het gemiddelde te berekenen van 15 tebibytes?
- b. Hoeveel tijd kost het vijftien PC's om het gemiddelde te berekenen van 15 tebibytes?
- c. Hoeveel tijd kost het om vijftien getallen ter grootte van 1 kibibyte te sturen over een verbinding met een maximale snelheid van 500 mebibytes per seconde?

**Antwoorden**

- a.  $15 \text{ tebibytes} / 0,250 \text{ tebibytes per seconde} = 60 \text{ seconden}$
- b. Deze berekeningen worden tegelijkertijd uitgevoerd, dus het antwoord is gelijk aan het vorige antwoord, namelijk 60 seconden.
- c. In totaal moet er dus 15 kibibytes ontvangen worden op de PC waar vandaan de berekening is gestart. Er is dus  $0,015 \text{ mebibytes} / 500 \text{ mebibytes per seconde} = 0,00003 \text{ seconde} = 0,003 \text{ milliseconde}$  voor nodig om de uitkomsten berekent op de 15 PC's te ontvangen

#	Kleur
1	Geel
2	Rood
3	Blauw
4	Rood

Tabel 1.11 Data in de normale representatie

#	Kleur
1	Geel
2, 4	Rood
3	Blauw

Tabel 1.12 Een andere representatie

Deze manier van berekenen wordt ook wel **distributed computing** genoemd. Deze techniek is een goed voorbeeld van techniek die ontwikkeld is en verder ontwikkeld moet worden om Big Data snel genoeg te kunnen verwerken.

**De diversiteit van data**

Het derde kenmerk van Big Data heeft alles te maken met de **diversiteit** van de data. Als data weinig divers is, is het gemakkelijk om de data te analyseren en om de gegevens uit de data te halen die je wilt. Kijk maar naar Tabel 1.11.

Je kunt zien dat de data weinig divers is. Er zijn vier objecten, en elke object heeft een kleur. Als je het objectnummer weet, kun je gemakkelijk nagaan wat de bijbehorende kleur is. Andersom is iets lastiger, maar nog steeds doenlijk. De kleur 'geel' hoort bij object 1, de kleur 'blauw' hoort bij object 3, en de kleur rood hoort bij de objecten 2 en 4.

Bekijk nu ook Tabel 1.12. Als je beide tabellen met elkaar vergelijkt, zie je dat de tweede tabel meer diversiteit aan data bevat. De tweede

#	Kleur	Vorm
1	Geel	Driehoek
2	Rood	Vierkant
3	Blauw	Vierhoek
4	Rood	Ster

Tabel 1.13 Diverse data

kolom van de tweede bevat namelijk complexere data, namelijk een lijst met objecten in plaats van een enkel object. In de eerste tabel is de data minder divers, omdat elk object een eigen kleur heeft. Overigens is de data van beide tabellen even divers, alleen de **representatie** van de data is anders. Hierdoor komt de diversiteit van de data beter tot haar recht. De diversiteit van data is niet afhankelijk van hoe de data is gerepresenteerd. Daarom loont het altijd om eerst de data op de juiste manier te representeren voordat je gaat onderzoeken hoe divers de data is.

Je kunt je voorstellen dat er meer data bekend is van elk object. Bekijk Tabel 1.13.

Elk object heeft ook een vorm. **Data fusion** maakt het mogelijk om de gegevens van de objecten te combineren. Het woord ‘fusion’ komt van fuseren, wat samenvoegen, samensmelten betekent. Bekijk de volgende tabel maar, waar de data is samengevoegd.

Een mooi voorbeeld van de diversiteit in data kun je vinden in Tabel 1.14. In de laatste kolom zie je dat er een videofragment is toegevoegd aan elk object.

#	Kleur	Vorm	Video
1	Geel	Driehoek	24 frames
2	Rood	Vierkant	48 frames
3	Blauw	Vierhoek	24 frames
4	Rood	Ster	24 frames

Tabel 1.14 Meer diverse data

Stel dat we nu alleen de video hebben, en we gaan op zoek bij welk object de video hoort. Als de video 48 frames heeft, is het een gemakkelijk taak in dit geval, dan hoort de video bij object 3. Als de video 24 frames heeft, wordt het erg ingewikkeld. We moeten dan de bits in de video op de een of andere manier gaan analyseren om erachter te komen bij welk object de video hoort. Een mogelijke aanpak is om de waardes van de rode, groene en blauwe kleur van alle pixels in de video op te tellen. Hopelijk heeft object 1 minder rode pixels dan object 2 en 4. Als dat niet het geval is, moeten we nog een complexere oplossing bedenken. Dit is een goed voorbeeld waaruit blijkt dat de diversiteit van data een rol speelt in de definitie van Big Data. Data is pas echt groot als het moeilijk is te analyseren, of als alleen via een complexe manier het mogelijk is de juiste gegevens uit de data te halen.

### Voorbeeld F

Gegeven onderstaande tabel.

- a. Hoeveel frames moet je met elkaar vergelijken om de twee frames te vinden die het meest op elkaar lijken?
- b. Om een vergelijking tussen bits uit te voeren gebruik je twee wiskundige berekeningen: aftrekken en de absolute waarde. Hoeveel berekeningen zijn er nodig om twee frames met elkaar te vergelijken? Ga er vanuit dat elke frame 640x480 pixels bevat, en dat de video in kleurenbeelden bevat.
- c. Hoeveel wiskundige berekeningen zijn er nodig om de twee frames te vinden die het meest op elkaar lijken?

### Antwoorden

- a. Object 1 heeft 24 frames. Die moeten met de frames van de objecten 2, 3 en 4 vergeleken worden, dus met  $48 + 24 + 24 = 96$  frames.

$24 \times 96 = 2.304$  vergelijkingen.

Object 2 heeft 48 frames. Die moeten met de frames van de objecten 3 en 4 vergeleken worden, dus met  $24 + 24 = 48$  frames.

$48 \times 48 = 2.304$  vergelijkingen.

Object 3 heeft 48 frames. Die moeten met de frames van object 4 vergeleken worden, dus met 24 frames

$24 \times 24 = 576$  vergelijkingen.

In totaal zijn er dus  $2.304 + 2.304 + 576 = 5.184$  vergelijkingen nodig.

- b. Per kleur moeten er  $640 \times 480 = 307.200$  vergelijkingen uitgevoerd worden. In totaal zijn er drie kleuren, dus er zijn  $307.200 \times 3 = 921.600$  vergelijkingen nodig.

Elke vergelijking heeft twee wiskundige berekeningen nodig, dus er zijn  $921.600 \times 2 = 1.843.200$  wiskundige berekeningen nodig.

- c.  $1.843.200 \times 5.184 = 9.555.148.800$  wiskundige berekeningen zijn er nodig.

## Vragen en opdrachten

De docent zal aan het einde van deze les een toets geven van ongeveer een kwartier. Hieronder staan een aantal voorbeeldopgaven.

1. Hoeveel kilobytes heb je nodig voor het opslaan van een document van 10.000 letters?
2. Hoeveel kleurenfoto's van 4096x2048 pixels kun je maken met een fototoestel die een opslagcapaciteit heeft van 4 gibibytes?
3. Hoe lang duurt het om 10 gigabyte aan informatie te sturen over een netwerkverbinding met een snelheid van 200 kilobytes per seconde?
4. Eva heeft 20 grijswaarden foto's gemaakt van 1024x768 pixels. Ze gaat op zoek naar de donkerste foto. Hoeveel berekeningen moet ze uitvoeren voordat ze deze foto heeft gevonden?

## 1.3 Data uit verschillende bronnen

### Inleiding

Eén van de grootste overnames binnen de sociale media vond plaats in 2014. Facebook nam toentertijd WhatsApp over. Hierdoor heeft Facebook ineens de beschikking over veel meer data.

Deze paragraaf gaat over het combineren van bronnen met data. Dit levert namelijk vaak problemen op. Hoe kan Facebook zijn data combineren met de data van WhatsApp? Een mogelijkheid is te zoeken naar gebruikers met dezelfde naam op Facebook en WhatsApp. Maar wat als er meerdere gebruikers zijn met dezelfde naam?

Een andere mogelijkheid is de gebruikers te combineren op basis van hun telefoonnummer. Maar ook dan zijn er problemen. Wat als iemand geen telefoonnummer heeft ingevoerd bij Facebook? Wat als iemand twee telefoonnummers heeft? En wat als iemand het telefoonnummer van zijn ouders of vrienden heeft gebruikt?

Big Data bestaat vaak uit meerdere bronnen. Eén van de uitdagingen van Big Data is daarom om deze bronnen op een slimme manier te combineren. In het kort kun je zeggen dat er drie uitdagingen zijn. Allereerst kan er variatie zijn in de data. Ook de kwaliteit van de data kan een rol spelen. Tot slot kan de complexiteit van de data verschillen.

### Variatie in de data

Data uit verschillende bronnen kan variëren. Dit betekent dat de data uit een bron niet overeenkomt met data uit een andere bron. Een simpel voorbeeld: voor Facebook gebruikt een gebruiker een andere profielfoto dan voor zijn WhatsApp.

Een ander voorbeeld zijn weersvoorspellingen. De ene weerman geeft vaak andere voorspelling dan een andere weerman. En hoe extremer de voorspelling, hoe groter de kans dat deze weerman het landelijk nieuws haalt. Als dit gebeurt, leest iedereen deze extreme voorspelling. Hoe kun je als nieuwsstation nu deze slechte voorspellingen weg filteren? En hoe kun je de verschillende weersvoorspellingen het slimst combineren? Een oplossing is het gemiddelde nemen van deze weersvoorspellingen.

#### **Voorbeeld A**

Voor de tweede kamer verkiezingen zijn vier peilers actief. Jan is redacteur bij een krant, en wil alle vier de peilingen op een zo slim mogelijke manier combineren. Hierdoor vangt Jan de extremen af waardoor zijn krant een eerlijk beeld geeft van de huidige peilingen. Hoe kan hij dit het beste doen?

#### **Antwoord**

Het gemiddelde nemen van de peilingen.

## Kwaliteit van de data

De betrouwbaarheid van de data uit verschillende bronnen kan verschillen. Deze betrouwbaarheid of kwaliteit van de data kan op twee manieren onderscheiden worden.

De eerste manier is hoe **precies** de data is. Vergelijk de kwaliteit of 'preciesheid' van de leeftijd van Sofie. Uit bron A blijkt dat Sofie 15 jaar oud is. Uit bron B blijkt dat Sofie 15 jaar, 3 maanden en 6 dagen oud is.

Een ander voorbeeld is de kwaliteit van een afbeelding. Twee bronnen hebben dezelfde afbeelding, maar bron A heeft de afbeelding in 1600x1200 pixels en bron B heeft de afbeelding in 1024x768 pixels, De kwaliteit van de afbeelding in bron A is beter dan de kwaliteit van de afbeelding in bron B.

De tweede manier is de **correctheid** van de data. Vergelijk de kwaliteit of 'correctheid', 'juistheid' van de leeftijd van Sofie. Volgens haar officiële paspoort is ze 15 jaar, 3 maanden en 6 dagen oud. Maar op haar Facebook staat dat je 18 jaar, 3 maanden en 6 dagen oud is. Welke bron is meer betrouwbaar? Er is een logische verklaring waarom Sofie op haar Facebook heeft ingevuld dat ze 18 jaar is. Hierdoor heeft ze toegang tot informatie die alleen toegankelijk is voor volwassenen. Verder bevat haar paspoort officiële gegevens, als ze een vals paspoort heeft kan dit grote problemen veroorzaken als Sofie op vakantie gaat naar andere landen. Daarom is de betrouwbaarheid van de informatie op haar paspoort groter dan de betrouwbaarheid van de informatie van Facebook.

Een tweede voorbeeld zijn de weersvoorspellingen. Het is bekend dat als je als weerman een extreme voorspelling doet, je het landelijk nieuws haalt. Maar als je dit te vaak doet, neemt jouw betrouwbaarheid af als weerman en zullen de landelijke media jouw voorspellingen negeren, hoe correct ze ook zijn. Door verschillende weersvoorspellingen op de juiste manier te combineren neemt de kans toe dat de voorspelling correct is. Een veelgebruikte manier is het gemiddelde nemen, hierdoor worden extremen afgevlakt.

### **Voorbeeld B**

Is een kleurenafbeelding meer of minder betrouwbaar dan een grijswaarden afbeelding?

### **Antwoord**

Een kleurenafbeelding is betrouwbaarder omdat deze meer informatie bevat.

**Voorbeeld C**

Heeft een foto genomen in de nacht meer of minder kwaliteit dan een foto genomen in daglicht? Ga ervanuit dat het fototoestel exact dezelfde instellingen gebruikt voor nacht- en dagfoto's.

**Antwoord**

Een foto genomen in de nacht heeft minder kwaliteit, omdat de correctheid van een nachtfoto minder is.

## Complexiteit van data

Niet altijd is het mogelijk om data uit verschillende bronnen te combineren. Er zijn twee manieren waarop je de complexiteit van data kunt uitleggen.

De eerste manier is dat het combineren van data wel, niet, of in bepaalde mate **mogelijk** is. Denk aan het combineren van een Facebookgebruiker en een Twitter gebruiker. Dit is perfect mogelijk, aangezien het één en dezelfde persoon betreft.

Een ander voorbeeld is het combineren van alle gegevens van auto's met alle gegevens van dieren. Qua informatie voegen de bronnen niets toe, maar je kunt wel een vergelijking maken tussen de snelheid van een auto en de snelheid waarmee een dier kan rennen. Dan is het combineren van de data in beperkte mate mogelijk.

Het derde voorbeeld is het combineren van een bron met de snelheden van alle auto's met een bron met de namen van alle dieren. Het combineren van deze informatie is op geen enkele manier mogelijk.

De tweede manier is dat het combineren van data uitsluitend op een manier mogelijk is waardoor de **uitkomst** van deze combinatie erg complexe gegevens oplevert.

Een voorbeeld waarbij het wel mogelijk is om data te combineren maar waarbij de data erg complex is, is het combineren van weersvoorspellingen met de voorspellingen van een Tweede Kamer verkiezing. Als er regen wordt voorspeld, zal de opkomst *waarschijnlijk* lager zijn dan wanneer er mooi weer wordt voorspeld. Ook *zouden* misschien mensen die een bepaalde partij stemmen niet gaan stemmen omdat ze het dan te veel moeite vinden. Een partij met trouwe kiezers *zou* hierdoor winst *kunnen* behalen. De schuingedrukte woorden duiden aan dat het niet zeker is of de data te combineren is. Er is wel enige mogelijkheid om de bronnen te combineren, maar op basis van de weersvoorspelling alleen is het onmogelijk om te beredeneren wat de uitslag van de Tweede Kamer verkiezingen zal worden. Andersom, op basis van de peiling van de verkiezingen is het onmogelijk om te beredeneren wat voor weer het de komende dagen zal worden.

**Voorbeeld D**

Wat kun je zeggen over de complexiteit van het combineren van de telefoonnummers van een gebruiker op Facebook en een gebruiker op WhatsApp.

- a. Is het mogelijk deze gegevens te combineren? Leg uit waarom.
- b. Is de uitkomst van het combineren interessant? Leg uit waarom.

**Antwoorden**

- a. Ja. De accounts met hetzelfde telefoonnummer verwijzen vermoedelijk naar dezelfde persoon.
- b. Ja, zo kun je nagaan of iemand zijn echte telefoonnummer gebruikt, of dat hij twee telefoonnummers gebruikt

**Voorbeeld E**

Wat kun je zeggen over de complexiteit van het combineren van videobeelden van een beveiligingscamera in een metro met geluidsopnames van een chimpansee?

- a. Is het mogelijk deze gegevens te combineren?
- b. Is de uitkomst van het combineren interessant?

**Antwoorden**

- a. Nee
- b. Deze vraag is niet meer relevant want het combineren van de data is niet mogelijk.

## Vragen en opdrachten

Voor deze les heeft de docent gevraagd of jullie de tekst van deze paragraaf van tevoren kunnen doorlezen. Als je dat gedaan hebt, kun je beginnen met de drie werkvormen.

**Werkvorm 1: mixed piles**

Maak drie stapels met de categorieën 'variatie', 'kwaliteit' en 'complexiteit'. Plaats de kaarten die je van de docent hebt gekregen op stapel met de juiste categorie bij welke de kaart hoort.

**Werkvorm 2: domino**

Leg de kaarten die je van de docent op de juiste manier achter elkaar

**Werkvorm 3: memory**

Je mag samen met je buurman/buurvrouw het spel memory spelen om de lesstof te leren. De kaarten hiervoor krijg je van de docent.

# 2 Toepassingen

Na een uitgebreide definitie van Big Data is het nu tijd om meer in detail te gaan kijken naar de toepassingen. Je zult zien dat Big Data op heel veel verschillende manieren in ons leven een belangrijke rol speelt.

Uit onderzoek van Gartner blijkt bij succesvolle integratie van Big Data dat een bedrijf 20% beter presteert dan concurrerende bedrijven. Die potentie is er ook voor de publieke sector ondanks dat men daar niet een specifiek winsttoegmerk heeft. Onderzoeker McKinsey stelt zelfs dat Big Data, wanneer juist toegepast, de publieke sector in Europa 150 tot 300 miljard per jaar kan opleveren.

Wat zijn dan die unieke toepassingen binnen de private sector en wat kan de publieke sector daar mee?

In de eerste paragraaf gaan we hier dieper op in. Hoe onze maatschappij verder vorm krijgt en de potentie heeft tot verduurzaming met behulp van Big Data komt samen met de valkuilen in de tweede paragraaf aan bod. Uiteindelijk ronden we in de laatste paragraaf af met de steeds belangrijkere rol van Big Data voor organisaties, het soort mensen die hierbij nodig zijn en de opleidingen die daarbij horen. Misschien heb je na deze module wel zoveel interesse gekregen dat je business analist of data scientist wil worden!

De **publieke sector** is de verzamelnaam voor alle overheidsorganisaties en semioverheidsorganisaties. De publieke sector is de tegenhanger van de private sector.  
Bron: Wikipedia

## Leerdoelen

Na dit hoofdstuk kun je:

- minstens drie voorbeelden geven van manieren waarop Big Data in de praktijk wordt toegepast
- opnoemen binnen welke gebieden van de publieke en private sector Big Data een grote invloed heeft of zal gaan hebben
- de rol van het Internet of Things (IoT) in combinatie met Big Data benoemen
- de stappen uit de feedback cyclus benoemen binnen de context van Big Data
- voorbeelden geven op wat voor manier Big Data bijdraagt aan de verduurzaming van onze maatschappij
- de randvoorwaarden benoemen als het gaat om de inzet van persoonsgegevens bij Big Data
- de ethische kanten van Big Data benoemen
- de primaire noodzaak om te veranderen benoemen voor organisaties in de publieke- en private sector.
- vijf organisatieprincipes benoemen die een steeds grotere rol zullen gaan spelen in de toekomst.
- benoemen welke vier rollen nodig zijn voor het model van de voorspellende analyse
- voorbeelden geven van beroepen en opleidingen die horen bij Big Data
- via online cursusmateriaal de basisconcepten van een data scientist jezelf eigen maken



## 2.1 Sneller meten, weten en van feit naar beleid

### De private sector

Met bijna 100 miljoen betalende abonnees heeft Netflix behoorlijk wat informatie van klanten te verwerken om het aanbod te blijven verbeteren. Netflix wil natuurlijk graag films en series aan het assortiment toe blijven voegen waar behoefte naar is. Om die behoeftes in kaart te brengen verzameld Netflix gegevens over ons kijkgedrag: wat kijk je? welke genres? hoe lang? pauzeer je vaak? wanneer precies? hoe lang moest je zoeken? Op basis van deze **data mining** kan Netflix persoonlijke aanbevelingen doen, de klant tevreden houden en nog belangrijker: nieuwe klanten werven!

Alleen data verzamelen is niet genoeg om te blijven verbeteren. Hoe weet Netflix op basis van al deze data of een nieuwe film gewild zal zijn? Hiervoor richt Netflix zich op **Big Data analytics**. Met behulp van Big Data technieken proberen organisaties voorspellingen te doen. Dat is precies wat Netflix ook doet. Door middel van uitgebreide analysetechnieken wordt waardevol inzicht verkregen. Het ingenieuze deel aan de oplossing van Netflix is het gegeven dat het bedrijf klanten uitbetaald om films en series te voorzien van tags. Vervolgens geeft Netflix op basis van jouw streamgedrag en veel voorkomende tags suggesties voor soortgelijke films en series. Zo kan Netflix het aanbod steeds verder verfijnen en blijven leren van haar klanten.

### De publieke sector

Netflix is een van de vele voorbeelden uit de private sector waar bedrijven met gerichte inzet van Big Data beter draaien, maar waar liggen precies mogelijkheden binnen de publieke sector? Binnen onze maatschappij zijn er bepaalde zaken die wij als burger belangrijker vinden dan anderen. De kunst is om een probleem op een zodanige manier op te lossen dat het de meeste waarde oplevert. Big Data kan hier voor het nodige inzicht zorgen. Organisaties in de publieke sector voelen echter veel minder de noodzaak om te veranderen.

Neem als voorbeeld 'de zorg' waarbij in de meest ideale situatie alle gegevens van elke operatie en elke patiënt verzameld, verwerkt en geanalyseerd kunnen worden om

- te gebruiken voor het maken van betere beslissingen
- het optimaliseren van alle stappen binnen het zorgproces
- het beter omgaan met beschikbaar geld en middelen
- meer betekenisvolle patiëntgegevens bij te houden
- voorspellingen te doen omtrent ziektebeelden

Uiteindelijk kan dat tot een betere zorg leiden. Dat hier nog wat haken en ogen aan zitten bespreken we in paragraaf 2.2 en 2.3.

## Internet of Things (IoT)

Belangrijk is dat Big Data ons in staat stelt om zogenaamde ‘real time’ ontwikkelingen vanuit verschillende bronnen te volgen en er voortdurend aan te **meten**. Om zoveel mogelijk nuttige ‘real time’ data te verzamelen worden vaak allerlei sensoren en aan het internet-gekoppelde apparaten gebruikt. Dit noemen we ook wel het **internet of things (IoT)**. Je zou IoT kunnen zien als een manier om bij Big Data daadwerkelijk voor de grote hoeveelheid data te zorgen. Al die metingen maken het vervolgens mogelijk om in een vroeg stadium op basis van de beschikbare data voorspellingen te doen zodat problemen kunnen worden voorkomen.

Een uitgebreide analyse op basis van de data en ondernomen acties zorgt dan voor een beter **inzicht** en mogelijkheden voor besluitvorming waarbij behoeften en **kansen** voor verbetering sneller in beeld komen. In plaats van voor iedereen dezelfde acties te ondernemen is het dan mogelijk om maatwerk te leveren en dat op te nemen in het beleid. Uitgebreide analyse op de verzamelde data maakt het zelfs mogelijk om voorspellingen te doen over de haalbaarheid van de te ondernemen **actie**.

## De feedback cyclus

Deze feedback cyclus van **meten - inzicht - kansen - actie** kan met Big Data snel worden doorlopen en de effecten van een bepaald beleid direct inzichtelijk maken die voorheen niet of nauwelijks afgeleid konden worden door de grote hoeveelheid aan data. Hierdoor is het mogelijk om sneller af te leiden welke delen van het beleid wel of niet effectief zijn, waar nog geoptimaliseerd kan worden zodat er ruimte is voor **innovatie**. Uiteindelijk kunnen hierdoor weer nieuwe producten en diensten ontstaan. Waar bedrijven voorheen jaren nodig hadden om beleid te evalueren is het met Big Data mogelijk direct te evalueren en aan te passen waardoor innovatieprocessen sterk versneld worden. Een mooi voorbeeld van de uitwerking van deze stappen uit de cyclus is het ‘London Smart City initiative’:

### **meten**

- sensoren die verkeersdichtheid meten van het wegennet
- Wi-Fi systemen die de bewegingen van mensen in het metronetwerk volgen

### **inzicht**

- op basis van de metingen achterhaalt men waar de knelpunten zitten

### **kansen**

- door eerder genomen acties te analyseren kan voor nieuwe verkeerssituaties en op basis van het verkregen inzicht de mate van succes worden bepaald en gebruikt worden

### **actie**

- op specifieke plekken verkeersomleidingen realiseren om het wegennet te ontlasten
- of een extra buslijn omdat dit op basis van de data de meest succesvolle oplossing blijkt te zijn
- of door een nieuwe weg aan te leggen of
- te investeren in een ander ticketsysteem waardoor men sneller kan reizen zodat vertragingen verminderd worden

Terwijl je dit leest maakt jouw school hoogstwaarschijnlijk ook gebruik van Big Data. Zo zijn er tegenwoordig digitale leeromgevingen en online methodes waar jij als leerling in kunt werken en docenten jou in kunnen volgen en assisteren. Gegevens over jouw voortgang zoals welke oefeningen nog niet lukken, hoe lang je ergens over hebt gedaan kunnen door middel van **Learning Analytics** zowel de leerling als de docent een beter beeld geven over de voortgang zodat jij beter kan leren en de docent jou daar gerichter bij kan helpen!

## Vragen en opdrachten

Behalve het gebruiken van de theorie is het aan te raden om bij onderstaande vragen gebruik te maken van het internet om tot volledige antwoorden te komen.

### 1. Technieken en sectoren

- Hoe heet de techniek waarmee grote hoeveelheden gegevens worden vergaard?
- Hoe heet de techniek waarmee de grote hoeveelheid vergaarde gegevens nauwkeurig bekeken en verwerkt worden?
- Leg uit waarom volgens jou organisaties in de publieke sector over het algemeen nog steeds minder met Big Data werken t.o.v. organisaties in de private sector

### 2. Van waarnemen tot innoveren

- De regering zou graag meer gebruik willen maken van zonnepanelen en wil o.a. grootschalig in kaart brengen over het land waar de meeste zonuren zijn om zoveel mogelijk energie op te wekken. Wat voor soort apparatuur zou jij hiervoor gebruiken en hoe verzamel jij die data?
- Een waterzuiveringsbedrijf laat door een effectieve inzet van Big Data en concepten van IoT de inzet van chemicaliën verminderen. Door de data uitgebreid te analyseren kan de installatie op de juiste momenten de juiste stappen ondernemen. Waarom kan het verzamelen van de data niet offline?
- Geef voor Netflix een uitwerking van elke stap binnen de feedbackcyclus

### 3. Leren met Big Data

- Benoem een aantal voordelen van het gebruik van Big Data in het onderwijs.

- b. Benoem mogelijke nadelen van het gebruik van Big Data in het onderwijs

#### 4. Onze toekomst in beeld

- a. Bekijk het volgende Engelstalige YouTube filmpje: <https://www.youtube.com/watch?v=eVSfJhssXUA&t=3s>. Het filmpje schetst een toekomst waarin de computer via automatisering en Big Data op veel punten werk van de mens overneemt. Verschillende stappen uit de feedbackcyclus komen indirect voorbij. Leg uit bij welke van de stappen in de feedbackcyclus volledige automatisering minder snel zal plaatsvinden.
- b. Lees het Engelse artikel op de volgende website: <https://www.weforum.org/agenda/2017/02/big-data-how-we-can-manage-the-risks>. Leg aan de hand van het artikel uit waarom het voor elk mens van belang is dat hij/zij over enige basiskennis beschikt van computer algoritmen en de toepassing ervan bij Big Data?

#### 5. Samen pitchen door de feedback loop

Vorm een tweetal en kies een fictieve organisatie/bedrijf waar jij verbeteringen met behulp van Big Data op loslaat. Let op! dit mag niet een voorbeeld uit de theorie of voorgaande opgaven zijn! Kies een aantal aspecten (zoals klanttevredenheid bv) binnen de organisatie dat je uitlicht om verbetering op toe te passen.

Teken nu op een A4 jullie cyclus en plaats bij elk gedeelte binnen de cyclus een stuk context hoe dat precies in jullie organisatie/bedrijf van toepassing is.

Zodra de schets af is laat je het geheel controleren bij de docent en ga je na goedkeuring zonder gebruik te maken van de schets en/of andere bronnen om de beurt een pitch geven voor jouw partner over jouw goede idee om Big Data toe te passen.

## 2.2 Duurzaamheid en privacy

### Inleiding

Uit de voorgaande paragraaf werd al duidelijk dat de inzet van Big Data niet alleen maar voordelen kent en dat er grote verschillen zijn in aanpak tussen de publieke en private sector. Toch zijn er voor beide sectoren veel mogelijkheden op het gebied van duurzaamheid. In alle gevallen moet men veel data verzamelen wat soms te herleiden is tot een bepaald individu. In deze paragraaf gaan we van de vele duurzame mogelijkheden naar de grootste valkuilen van Big Data.

### Toepassingen

Het toepassen van procesoptimalisatie om energie en grondstoffen te besparen noemen we verduurzaming door middel van Big Data. Die besparingen kunnen oplopen tot miljoenen euro's. In de vorige paragraaf zagen we voorbeelden van manieren hoe een succesvolle toepassing van Big Data tot grote besparingen kan leiden.

Behalve de verduurzaming in de zorg (zie paragraaf 2) zijn er nog veel meer mogelijkheden. Denk aan de volgende zaken:

#### **Meer energie opwekken met Big Data**

Hierbij kun je denken aan windmolens in de zee die worden voorzien van sensoren om de effecten van het weer en het water te meten.

Door deze gegevens te analyseren is het mogelijk om vooraf te kunnen bepalen hoe een nieuwe windmolen hierop zal reageren. Bedrijven die de windmolens plaatsen kunnen hiermee rekening houden en er uiteindelijk voor zorgen dat een windmolen langer blijft werken en het windmolenpark dus meer energie op zal leveren.

Men voorspelt hiermee dat de windmolenparken 20% meer energie op kunnen wekken

#### **Meer voedsel met Big Data (de moderne boer)**

Hierbij moet je vooral denken aan het begrip precisielandbouw dat zorgt voor een grote revolutie in de agrarische sector. Met behulp van allerlei sensoren die pH-waarden meten in de grond kan uiterst nauwkeurig bepaald worden welk deel van het land meer of minder kalk nodig meer of waar extra bemesting verstandig is.

Dit leidt onder andere tot een vermindering in de hoeveelheid kunstmest en bestrijdingsstoffen dat zowel goed voor het milieu als de portemonnee is!

#### **Beter water zuiveren met Big Data**

Met behulp van aan het internet gekoppelde sensoren (IoT) en computersimulaties is het mogelijk om beter inzicht te krijgen in de werking van een waterzuiveringsinstallatie en in welke situaties er meer zuurstof, energie en chemicaliën nodig zijn. Door middel van uitgebreide analyse kan de installatie op de juiste momenten de juiste stappen ondernemen.

In Spanje is een dergelijk systeem al in werking gegaan waarbij na de toepassing van Big Data zorgde voor 13% minder

elektriciteitsconsumptie, 14% minder gebruik van chemicaliën en 17% minder afval.

### **Slimme huizen met Big Data**

Naast energiezuinige woningen met besparende kranen, zonnepanelen en slimme meters, zit de echte kracht in de aaneenschakeling en onderlinge afstemming van verschillende apparaten in het huis. In 2017 worden verschillende proeven gedraaid met zogenaamde pilotwoningen die zijn voorzien met allerlei sensoren. Het hoofdsysteem van de woning, genaamd Iris, leert van het gedrag van de bewoners, stemt hier de aansturing van bijvoorbeeld de verwarming op af om uiteindelijk zoveel mogelijk energie te kunnen besparen.

### **Beter werken met Big Data**

Het is niet ondenkbaar dat jij binnenkort in een 'slim' kantoorgebouw komt te werken waarbij met behulp van allerlei sensoren (IoT) data verzameld wordt van beweging, koolstofdioxide waarden, temperatuur en luchtvochtigheid. Aansturen van systemen als verlichting en ventilatie zorgen dan niet alleen voor energiebesparing, maar ook voor 'gezondere' en betere werkplekken die weer voor een verbeterde productiviteit kunnen zorgen!

### **Betere zorg met Big Data**

In de vorige paragraaf werden de mogelijkheden besproken voor verbeteringen in de zorg. Onderzoekers van de Vrije Universiteit in Amsterdam hebben Big Data, in de vorm van geanonimiseerde elektronische patiëntendossiers, daadwerkelijk gebruikt om darmkanker te detecteren. Door een computersysteem de grote hoeveelheden gegevens te laten scannen zijn patronen ontdekt en verbanden gelegd die voorheen nog niet in beeld waren gekomen. Door Big Data is het mogelijk om in de toekomst in een vroeger stadium kanker te constateren.

Bovenstaande lijst is een greep uit de vele mogelijkheden van verduurzaming door middel van Big Data. Door het aanbrengen en inwinnen van informatie afkomstig van sensoren in allerlei situaties is het mogelijk om veel meer inzicht te krijgen over de manier waarop wij omgaan met de grondstoffen die de Aarde ons te bieden heeft en hoe wij daar als mensheid mee om kunnen gaan.

Dit klinkt fantastisch en veel hiervan wordt al gerealiseerd, maar als er overall informatie over ons gedrag wordt ingewonnen, hoe zit het dan met onze privacy?

## **Van webpagina naar online profiel**

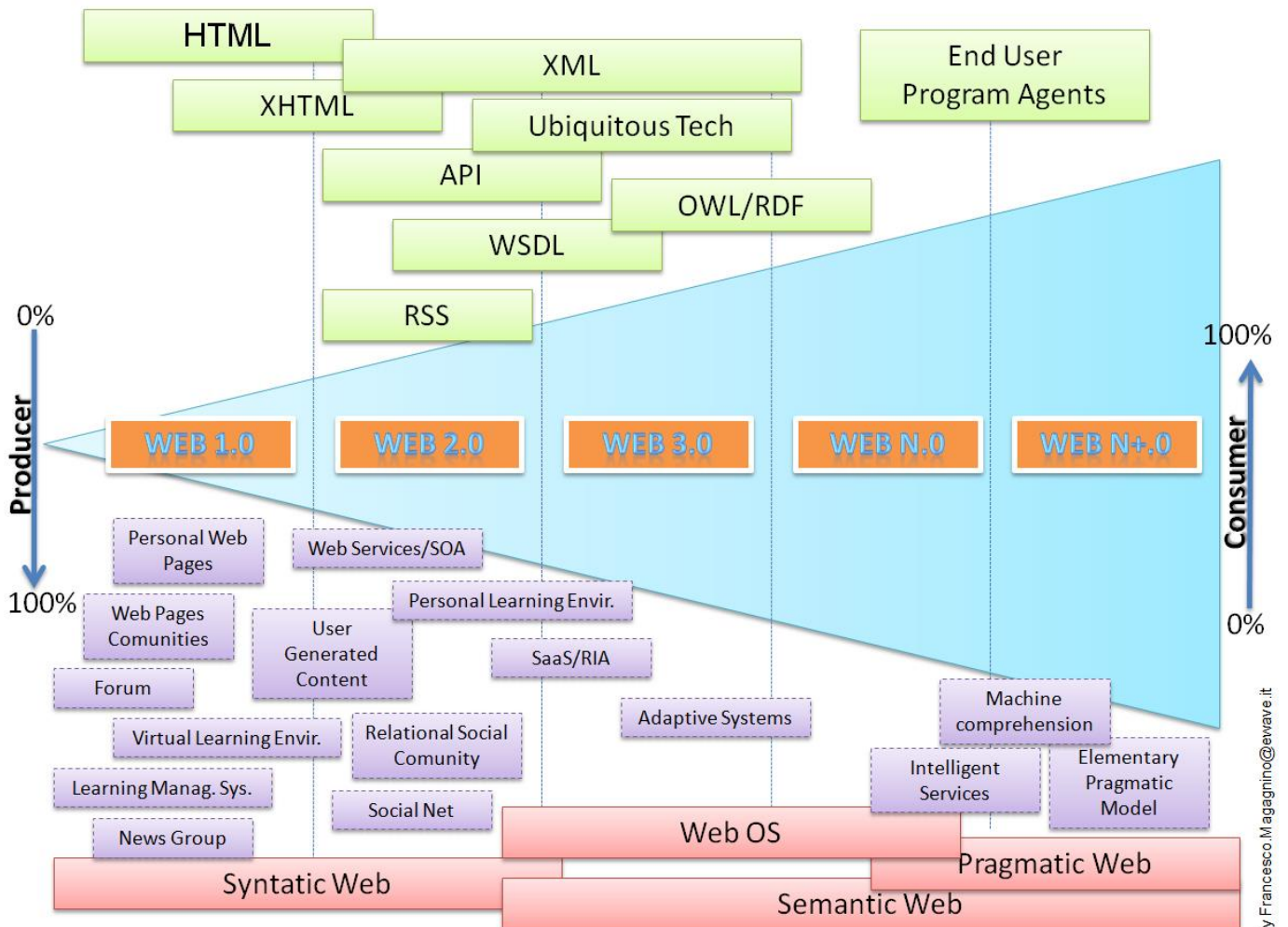
Het internet heeft de afgelopen jaren sterke ontwikkelingen meegemaakt die gekenmerkt kunnen worden door generatienamen als Web 1.0, Web 2.0, Web 3.0 enz. (zie Figuur 2.1)

Door deze generaties heen heeft de mens steeds meer data zelf gemaakt. In het begin, tijdens Web 1.0, waren vooral hobbyisten actief in het maken van webpagina's, maar tegenwoordig kan iedereen zo beginnen met het bouwen van een eigen website en ligt de nadruk op online gemeenschappen en socialiseren; mensen (Web 2.0).

Ondertussen zijn we aangekomen in Web 3.0 waar Big Data de grootste rol speelt. Via het Internet of Things (IoT) zijn steeds meer apparaten aan het internet gekoppeld en verzamelen we overal grote hoeveelheden informatie over onze omgeving, producten die we gebruiken, maar vooral ook onszelf.

Op dit moment zien we vooral inzet van die gegevens om allerlei producten te verbeteren, procesoptimalisatie en grotere efficiëntie door te voeren. De vraag is dan: hoe ziet een toekomst eruit waarin vrijwel al onze privé-acties op een bepaalde manier worden opgeslagen?

Nu al vraagt men zich af hoe bedrijven als Facebook omgaan met de grote hoeveelheden persoonlijke data, ook wel Big Social Data genoemd. Kunnen wij elkaar daarin nog vertrouwen? Krijgen we bijvoorbeeld inzicht wat er over ons is opgeslagen en waar die data allemaal naartoe gaat? Hoe zit het precies met onze privacy?



By Francesco.Magagnino@evrve.it

Figuur 2.1 De ontwikkeling van internet

## Randvoorwaarden

Iedereen is blij met energiebesparing, kostendalingen, verbeterde dienstverlening, zorg en misdaadpreventie door middel van Big Data, maar als hierbij onze rechten als burger geschonden worden gaat dat te ver. Een belangrijke voorwaarde hierbij is de mogelijkheid om data te kunnen herleiden naar een individu. Het College Bescherming Persoonsgegevens (CBP) houdt zich in Nederland onder andere hier mee bezig. Zij geven bijvoorbeeld aan dat bedrijven niet in alle gevallen toestemming moeten vragen voor het gebruik van naam, adres e.d. Dit soort data noemen we persoonsgegevens omdat op basis daarvan iemand geïdentificeerd kan worden. Als een bedrijf die gegevens nodig heeft voor de normale bedrijfsvoering dan ziet het CBP dat als een 'gerechtvaardigd belang'.

Wanneer persoonsgegevens **geanonimiseerd** zijn, zijn zij voor Big Data doeleinden bruikbaar. Als een organisatie gebruik wil maken van herleidbare gegevens moet er zijn voldaan aan de eisen van de Wet bescherming persoonsgegevens (Wbp). Herleidbare gegevens zijn data die het mogelijk maken om jouw identiteit vast te stellen. Deze mogen alleen gebruikt worden voor het **doel** waarvoor de gegevens zijn verzameld. Het is mogelijk om daar toestemming voor te vragen, maar dan moet iemand zeer goed geïnformeerd zijn en op de hoogte zijn van de manier waarop de gegevens zullen worden gebruikt. De overheid die dergelijke gegevens van ons verzameld heeft naast verantwoordelijkheid voor de beveiliging en juistheid van de gegevens ook de verantwoordelijkheid om de gegevens te anonimiseren zodat er geen sprake meer is van herleidbaarheid.

De wet Wbp, die opgesteld is in 2001; voor het tijdperk van Big Data, is ondanks zijn heldere criteria toe aan vernieuwing. Zo stijgt het aantal rechterlijke zaken waarbij overheden op de vingers worden getikt wegens verkeerde toepassing van Big Data op hun burgers. Een overzicht hiervan is op deze site te vinden: <http://www.privacyfirst.nl/rechtszaken-1.html>

Burgers uiten steeds vaker zorgen over elektronische dataopslag door de overheid. Een bekend voorbeeld is het elektronische patiëntendossier, waarin al de medische gegevens staan van een burger. Je moet maar hopen dat die gegevens niet benaderbaar zijn door de verkeerde personen. Er is veel mogelijk binnen de kaders van de wet, maar niet alles mag zomaar over ons worden bijgehouden en opgeslagen worden. Bedrijven en overheden moeten daarom bewust omgaan met het privacy vraagstuk als het gaat om Big Data.

## Vragen en opdrachten

### 1. Duurzaamheid

- a. Leg uit op welke manier Big Data bij kan dragen aan een beter milieu.
- b. Leg uit op welke manier Big Data bij kan dragen aan een betere zorg.

### 2. Big Social Data



Bekijk de eerste 10 minuten van de volgende documentaire over Big Data:

URL:

<https://www.vpro.nl/programmas/tegenlicht/kijk/afleveringen/2013-2014/persoonlijke-data.html>

Aan het einde van de video komt Michael Kosinski aan het woord van Cambridge University en legt uit hoe je aan de hand van iemand zijn Facebook profiel hun persoonlijkheid kunt afleiden. Dit gaan we uitproberen. Weest gerust, jouw resultaten en de demo zijn anoniem. Log in op Facebook en sta de app toe om jouw profiel in te lezen. Als je geen Facebook profiel hebt vraag je of je mee mag kijken bij een klasgenoot.

Ga vervolgens naar de volgende URL van de Cambridge University:

<https://applymagicsauce.com/demo.html>

Doe de test en kijk naar de resultaten. Vergeet niet naar beneden te scrollen!

- a. Leg uit welke 'Big Five personality traits' deze app af kan leiden
- b. Leg uit waarom jij denkt dat dit soort informatie geld waard is
- c. Hoe denk je dat bedrijven waar jij later gaat solliciteren hier mee om (zouden) gaan?

### **3. Het Big Data debat**

Onder begeleiding van de docent gaan jullie nu een debat voeren. Het idee is dat je aan de hand van de theorie en informatie uit deze paragraaf leert na te denken en een discussie te voeren over de verschillende standpunten aangaande privacy en Big Data. Wat kan er wel en wat kan er niet? We gaan dus kijken naar de ethische kant van Big Data.

Een van de belangrijkste uitdagingen bij Big Data is de balans tussen publieke en persoonlijke belangen.

Hierbij kunnen we dan denken aan publieke belangen als veiligheid, terrorismebestrijding, cybercrime, zorg, onderwijs en milieu en persoonlijke belangen als privacy.

Jullie gaan aan de slag met de volgende stelling:

Het is met Big Data mogelijk om deze balans tussen publieke en persoonlijke belangen halen. De docent zal de debatvorm met jullie bespreken.

## 2.3 De noodzaak en de mensen

### Inleiding

Het is duidelijk dat een goede inzet van Big Data voor grote verbeteringen in onze maatschappij kan zorgen. Toch zijn er zoals je in de vorige paragraaf kon lezen de nodige valkuilen en ziet niet elk bedrijf een noodzaak om te veranderen. Waarom die noodzaak er is gaan we in deze paragraaf bespreken en kijken welke mensen en opleidingen daarbij nodig zijn om dat in goede banen te leiden.

### De noodzaak om te veranderen

Uit de vorige paragrafen bleek al dat organisaties goed moeten nadenken over de manier waarop zij data verzamelen en dat bij een correcte inzet ook sprake is van een verbetering in de beleids- en innovatieprocessen. Voor veel bestaande grote organisaties is dat erg lastig: afdelingen hebben onderlinge afhankelijkheden, weten vaak niet van elkaar wie welke informatie heeft en de informatiebronnen zijn vaak niet of lastig op elkaar af te stemmen wat voor een succesvolle toepassing van Big Data noodzakelijk is.

In de private sector zien we daarom dat nieuwe bedrijven, de zogenaamde startups, de 'grote jongens' vaak inhalen doordat klanten tegenwoordig andere diensten en meer maatwerk vragen. Omdat de startups vanaf het begin hun organisatiestructuur hebben opgebouwd aan de hand van de huidige technologische mogelijkheden zijn zij vaak beter in staat om op korte termijn veranderingen door te voeren tegen lagere kosten. Klanten zijn hier vaak erg van gecharmeerd omdat het bedrijf daarmee als het ware meebuigt, snel aanpassingen maakt en dat ook nog tegen een redelijke prijs.

Gezien de afwezigheid van het winst oogmerk is te begrijpen dat van dergelijke ontwikkelingen in de publieke sector minder sprake is. Toch is er ook in de publieke sector een noodzaak om te veranderen. Zo moet men wegens economische druk zoeken naar manieren om kosten te drukken zodat er enerzijds financiële ruimte is voor vernieuwing maar ook lastenverlichting in de vorm van belasting aan de kant van de burger.

Zo stelt Prof. dr. Theo de Vries van de universiteit Twente dat een nationaal anti-fraudebureau ons land miljarden euro's kan besparen en dat door een correcte inzet van Big Data soortgelijke besparingen elders ook mogelijk zijn.

### Nieuwe organisatieprincipes

Het is nu duidelijk geworden dat organisaties moeten vernieuwen, maar waar liggen dan precies de pijnpunten?

Hieronder een opsomming:

- langzaam verloop in de feedbackcyclus; in paragraaf 2.1 werd duidelijk dat dit noodzakelijk is om snel tot verbetering te komen. In de praktijk zorgt het omgekeerde voor veel frustratie.
- het lerend vermogen van de organisatie; er ligt teveel de nadruk op het maken van lange termijn (meerjaren) plannen en

vergaderen waardoor de problemen in het 'hier en nu' te weinig aandacht krijgen, de organisatie kansen mist en daardoor constant achter de feiten aanloopt

- verdraaien en verkeerd interpreteren van de werkelijkheid; op basis van beperkte of incomplete gegevens worden problemen gereduceerd tot acties die niet direct leiden tot de oplossing van het probleem

Organisaties moeten veranderen op volgende wijze op deze pijnpunten uit de weg te gaan:

- Beslissen op basis van (kennis uit) de data; **de data is leidend**
- Gebruikmaking van zoveel mogelijk databronnen om meer inzicht te verkrijgen
- Een korte feedbackcyclus waarbij constant feedback is vanuit de omgeving en waar direct op gereageerd kan worden
- Meer nadruk op korte lijntjes, minder lagen, afdelingen en minder hiërarchie
- Meer opdrachtgestuurd werken en vernieuwen aan de hand van inzichten vanuit de data

Het is duidelijk dat bij dergelijke veranderingen de rol van managers en bestuurders in het algemeen steeds verder zal afnemen. De nadruk komt te liggen op professionals die de data van het bedrijf kunnen analyseren, inzichten kunnen verschaffen en aan de hand hiervan snel een nieuwe strategie kunnen uitzetten op basis van de data. Dit vereist een feit (op basis van data) gestuurde organisatie die kortcyclisch bottom-up werkt waar de beleidsmedewerkers zich richting uitvoering begeven. Door inzet van nieuwe technologie en Big Data zal de voorheen gescheiden wereld van beleid en uitvoering verdwijnen en meer integratie van verschillende organisatiedelen plaatsvinden.

## Nieuwe organisatieprincipes

Het is nu duidelijk geworden dat organisaties moeten vernieuwen, maar waar liggen dan precies de pijnpunten?

Stel je wil als bedrijf graag uit de grote hoeveelheid beschikbare data inzicht verkrijgen en op basis daarvan voorspellingen doen.

Hoe pak je dat dan aan en wie heb je daarbij nodig? We maken hierbij onderscheid tussen de volgende vier rollen:

- Business manager
- Business analist
- IT system manager (ook vaak big data manager genoemd)
- Data scientist (ook vaak data analist genoemd)

Het is natuurlijk ook mogelijk voor iemand om meerdere rollen tegelijk op zich te nemen.

Een verdere uitwerking van de verschillende stappen in dit model zijn hieronder te vinden:

### Probleemformulering

In het begin probeert de **business manager** grip te krijgen op het geheel door alles concreet te maken en het proces af te bakenen.

### **Data voorbereiding en verkenning**

Zodra het 'probleem' goed in beeld is begint men samen met de **IT system manager**, **data scientists** en **business analyst** mogelijke databronnen in kaart te brengen om vervolgens het geheel gereed te maken voor verzameling en opschoning zodat er sprake is van optimale bruikbaarheid van de data. Omdat hier een combinatie van veel professionals noodzakelijk is heeft deze stap veel tijd nodig.

### **Data transformatie en selectie**

Bij deze stap zijn de **data scientists** bezig om met behulp van speciale software de data te doorzoeken op patronen, onderliggende relaties en eventuele trends waar de organisatie het beste op in kan spelen.

### **Modelontwikkelfase**

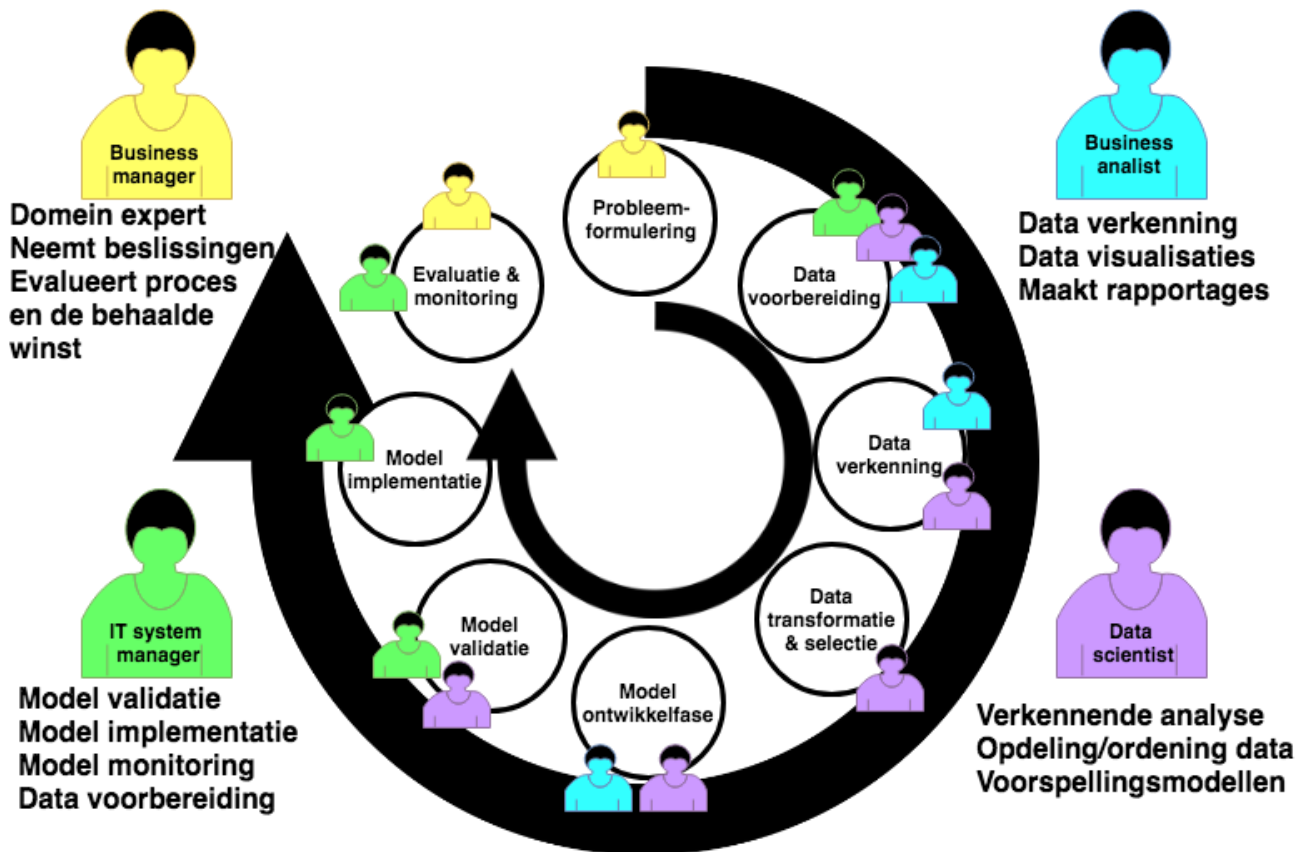
Vervolgens bouwen de **data scientists** via statistische en data-mining software een algoritmisch model. Hierbij maakt men in veel gevallen gebruik van zelflerende modellen die aan de hand van de oude data met bijbehorende resultaten voor nieuwe data uitkomsten met een hogere voorspellende waarde kunnen genereren. Deze resultaten moeten constant tegen het licht worden gehouden door de **business analyst** om te zien op welke manier het geheel bijdraagt aan het doel dat tijdens de voorbereiding was opgesteld.

### **Model validatie en implementatie**

Tijdens deze stap voeren de **data scientists** samen met de **IT system manager** en achterliggende ICT-ers de laatste checks uit om eventuele fouten in het model op te sporen en de werking vooraf te testen. De daadwerkelijke implementatie en toepassing van het model ligt volledig bij de **IT system manager** en de bijbehorende IT-ers.

### **Evaluatie en monitoring**

Na de implementatie is het noodzakelijk voor de ontwikkeling van de organisatie om alles te monitoren en op basis van de resultaten de genomen stappen te evalueren en waar nodig te zorgen voor aanpassing. Door nauw overleg tussen de **IT system manager** en de **business manager** worden nieuwe doelen gesteld en/of problemen snel geïdentificeerd om vervolgens weer opnieuw de cyclus te doorlopen aan de hand van een bijgewerkte strategie. Het geheel is als cyclus van voorspellende analyse in een figuur als volgt weer te geven:



Figuur 2.2 Het model van de voorspellende analyse

## Nieuwe organisaties, nieuwe banen

### Data scientists

Data scientists hebben vaak een vooropleiding gehad binnen de volgende velden: (toegepaste) wiskunde, statistiek, technische informatica en econometrie.

In Delft bestaat hier bijvoorbeeld de studie Technische Informatica en Technische Wiskunde voor en in Rotterdam de studie Econometrie. Er bestaan tegenwoordig echter ook al zeer gerichte opleidingen op het gebied van data science. Zo biedt de universiteit in Eindhoven de opleiding Data Science aan.

### Business analisten & business managers

Business analisten en business managers hebben vaak een vooropleiding gehad in de vorm van business management met een combinatie van IT. In Rotterdam bestaat hier bijvoorbeeld de opleiding Business Information Management voor.

Er zijn ook vaak aansluitende masters of minors binnen de opleiding voor Big Data te vinden waardoor iemand zich vanuit het bedrijfs perspectief al kan richten op het werken met Big Data.

### IT system manager

IT system managers hebben vaak een vooropleiding gehad in de vorm van informatica of bedrijfskundige informatica. Dergelijke opleidingen

worden tegenwoordig door het hele land aangeboden en geven de mogelijkheid om heel breed uit te stromen.

### **Zelfstudie**

Door het grote aanbod van opleidingen die in meer of mindere mate te maken hebben met Big Data is de afgelopen jaren de nadruk komen liggen op zelfstudie.

Werkgevers vragen vaak om aantoonbare kennis in het gebruik van bepaalde softwarepakketten en/of resultaten van online sites zoals <https://bigdatauniversity.com/> , <https://www.datacamp.com> en <https://www.kaggle.com/>

Het is daarom belangrijk om naast een opleiding ook goed te blijven kijken naar het online cursusaanbod dat binnen jouw vakgebied en interesses ligt. In de laatste opdracht van deze paragraaf gaan we daar dieper op in.

## **Vragen en opdrachten**

### **1. Veranderen**

- a. Leg uit wat de primaire noodzaak om te veranderen is in de private sector
- b. Leg uit wat de noodzaak om te veranderen is in de publieke sector

### **2. Kansen laten liggen**

Binnen het onderwijs is het effect van Big Data te zien bij de grote uitgeverijen. Zij worden ingehaald door nieuw bedrijven die via online leeromgevingen/methoden hun eigen materiaal constant evalueren en bijwerken op basis van de wensen van de actuele wensen van scholen. Waar traditioneel gezien een uitgeverij elke paar jaar een nieuw boek uitbracht als vernieuwing moet er nu om de paar weken het product en de effectiviteit ervan onder de loep worden genomen.

- a. Leg uit waar het grootste pijnpunt ligt bij deze uitgeverijen aan de hand van de genoemde punten in de theorie.
- b. Leg uit welke organisatieprincipes van toepassing moeten zijn om dit weer in goede banen te leiden.

### **3. De mensen**

- a. Benoem de typen experts dat nodig is bij het opzetten van een Big Data probleem waarbij voorspellingen gemaakt moeten kunnen worden.
- b. Welk type expert zet de daadwerkelijke strategie uit en neemt daarop beslissingen?
- c. Welk type expert is verantwoordelijk voor de diepgaande analyses en ontwikkeling van wiskunde 'voorspellende' modellen op basis van de data?
- d. Welk type expert zit met de ene voet in de programmeer- en wiskundige kant van data analyses en met de andere voet in het management?

#### 4. Big Data University - MOOC

In de theorie heb je kunnen lezen dat er tegenwoordig steeds meer de nadruk is komen te liggen op zelfstudie via online platformen. Binnen de IT en aansluitende vakgebieden gaan de ontwikkelingen zo hard dat het voor veel werkgevers zelfs een vereiste is om aan te tonen dat jouw vaardigheden en kennis 'op peil' zijn.

Sommige van deze cursussen zijn betaald maar er zijn ook veel gratis MOOC's (Massive Open Online Courses). Binnen Big Data nemen data scientists de meest prominente rol in. Zonder hun uitgebreide kennis van statistiek en computerwetenschappen is het niet mogelijk om de voorspellende modellen te ontwikkelen waar een organisatie op zit te wachten. Er is veel vraag naar hun expertise en veel websites waar de vaardigheden als data scientist aangeleerd kunnen worden laten als stimulans zelfs het gemiddelde salaris zien om mensen aan te trekken. Neem maar eens een kijkje op de volgende site:

URL: <https://www.datacamp.com/tracks/data-scientist-with-python>

Om je kennis te laten maken met de hoofdconcepten van 'data science' gaan we online een beginnerscursus volgen op de site van bigdatauniversity.com. De cursus is in het Engels en levert bij succesvolle afronding een badge/certificaat op. Hoe hoger jouw score binnen de cursus, hoe beter jouw deelcijfer.

De URL: <https://bigdatauniversity.com/courses/data-science-101/>

Maak een account aan of log in met jouw Google of Facebook account. Deze online cursus is een afsluiting van dit hoofdstuk waarvan de genoemde technieken in het volgende hoofdstuk met meer wiskundige achtergrond naar voren zullen komen.

# 3 Technieken

Elektrische energie is in het dagelijks leven onmisbaar. Het is om twee redenen aantrekkelijk. Elektrische energie is gemakkelijk te vervoeren en met weinig verlies om te zetten in andere vormen van energie. Maar hoe komt die elektrische energie eigenlijk tot stand? En hoe komt die elektrische energie naar je toe?

## Leerdoelen

Na dit hoofdstuk:

- kun je in eigen woorden het doel en de taak van associatieanalyse bij datamining toelichten.
- Kun je de begrippen verzamelingenleer, associatieanalyse associatieregel uitleggen.
- weet je de toepassingen van associatieanalyse, bijvoorbeeld Market basket analyse
- ben je in staat praktische voorbeelden van associaties opnoemen.
- kun je Apriori algoritme gebruiken bij het vinden van de associatieregel in een eenvoudig probleem.
- gebruiken bij het bepalen van Doorsnede en de vereniging van twee of meer verzamelingen.
- gebruiken bij het berekenen van Support van een deelverzameling
- gebruiken bij het berekenen van Confidence van een associatieregel van een deelverzameling
- als rekenmiddel voor de uitvoering van de benodigde berekeningen bij elke iteratie van Apriori algoritme.
- weet je het doel en de taak van clusteranalyse bij datamining.
- kun je een paar toepassingen van clustering noemen;
- kun je de begrippen datapunt, object, cluster, centrum, centroïde, Euclidische afstand, Manhattan afstand; centrum uitleggen .
- kun je de afstand tussen zowel twee objecten als twee clusters met behulp van afstandsmaten zoals Euclidische afstand en Manhattan afstand berekenen.
- Kun je de centrum of centroïde tussen zowel twee objecten als twee clusters met behulp van van afstandsmaten zoals Euclidische afstand en Manhattan afstand berekenen.
- Ben je in staat de clusteringsmethoden, zoals de  $k$ -means methode op passende problemen toepassen.

Notatie	Uitleg
$\in$	'is een element van', 'behoort tot'
{ , }	accolades worden gebruikt om een verzameling aan te duiden,
$\subset$	'is een deelverzameling van'
$\cap$	'doorsnede', het gemeenschappelijke deel van de twee verzamelingen
$\cup$	'unie', 'vereniging', de elementen van de eerste en tweede verzameling worden samengenomen

Tabel 3.1 Symbolenschema

## 3.1 Associatieanalyse

### Verzamelingenleer

In plaats van een 'groep' spreken we van een *verzameling* ( $V$ ). In een verzameling bevindt zich individuen, beter bekend als *elementen*  $\{v_1, v_2, v_3, \dots v_n\}$ .

Een verzameling representeert een aantal elementen met een bepaalde gezamenlijke *eigenschap* ( $x$ ), zoals de eigenschap 'mens' of



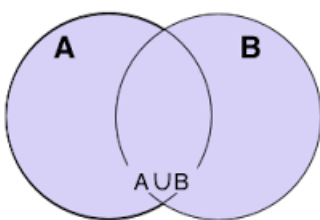
‘sterfelijk’. Een verzameling bestaat uit elementen, en als  $x$  een element van een verzameling  $X$  noteren we dit als  $x \in X$ . We gebruiken veelal accolades  $\{$  en  $\}$  om verzamelingen te noteren. We gebruiken binnen de accolades vaak het symbool  $|$  om elementen met bepaalde eigenschappen aan te geven, zoals in  $\{x \in N \mid x \text{ is deelbaar door } 2\}$

Zo is bijvoorbeeld  $\{Melk, Brood, Kaas\}$  de verzameling met de elementen Melk, Brood, en Kaas.  $Melk \in \{Melk, Brood, Kaas\}$

Als voor twee verzamelingen  $X$  en  $Y$  geldt  $(\forall x: x \in X \rightarrow x \in Y)$ , ofwel dat elk element van  $X$  ook een element is van  $Y$ , dan zeggen we dat  $X$  een deelverzameling is van  $Y$ .

Zo is bijvoorbeeld  $\{Melk, Brood\}$  een deelverzameling van  $\{Melk, Brood, Kaas\}$

De doorsnede  $X \cap Y$  bestaat uit de gemeenschappelijke elementen van  $X$  en  $Y$ .



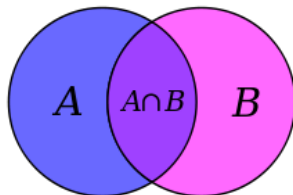
Figuur 3.1 De doorsnede van  $X$  en  $Y$

$X \cap Y = \{x \mid x \in X \wedge x \in Y\}$ . Deze verzameling heet de doorsnede van  $X$  en  $Y$ .

Zo is bijvoorbeeld  $\{Kaas, Melk, Brood\} \cap \{Brood, Fles, Melk\}$  de verzameling met de elementen *Melk* en *Brood*.

$$\{Kaas, Melk, Brood\} \cap \{Brood, Fles, Melk\} = \{Melk, Brood\}$$

De vereniging  $X \cup Y$  bestaat uit de elementen die tot  $X$ , tot  $Y$  of tot beide horen.



Figuur 3.2 De vereniging van  $X$  en  $Y$

$X \cup Y = \{x \mid x \in X \vee x \in Y\}$ . Deze verzameling heet de vereniging van  $X$  en  $Y$ .

Zo is bijvoorbeeld  $\{Kaas, Melk, Brood\} \cup \{Brood, Fles, Melk\}$  de verzameling met de elementen *Kaas*, *Melk*, *Brood* en *Fles*.

$$\{Kaas, Melk, Brood\} \cup \{Brood, Fles, Melk\}$$

$$= \{Kaas, Melk, Brood, Fles\}$$

Voor eindige verzamelingen  $A$  definieerden we simpelweg  $\#(A)$  als het aantal elementen in  $A$ . Zo is bijvoorbeeld

$$\#\{\{Chips, zeep, appels\}\} = 3.$$

Een element van een verzameling kan zelf ook een verzameling zijn.

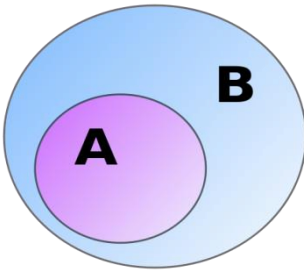
Zo bestaat de verzameling

$$C = \{\{Chips, zeep, appels\}, \{Chips, zeep, appels\}, \{Zeep, bananen\}, \{Chips, zeep, bananen\}\}$$

uit vier elementen. Een klasse verzamelingen is een collectie die bestaat uit verzamelingen.

## Associatieregels

De associatieregels  $X \Rightarrow Y$  betekent: als de deelverzameling  $X$  voorkomt dan komt ook de deelverzameling  $Y$  voor.



Figuur 3.3 A is een deelverzameling van B

**Voorbeeld A**

De onderstaande tabel bevat een collectie.

$T_1$	{Chips, zeep, appels}
$T_2$	{Chips, zeep}
$T_3$	{Zeep, bananen}
$T_4$	{Chips, zeep, bananen}

**Voorbeeld B**

{Melk, Brood, Suiker} is een itemset.

{Melk, Brood, Suiker} is een 3-itemset.

{Melk, Brood}, {Brood, Suiker} en {Suiker, Melk} zijn een 2-itemset.

{Melk}, {Brood} en {Suiker} zijn een 1-itemset.

De **support** van een deelverzameling  $X$  is het aantal deelverzamelingen waar de elementen van  $X$  in voorkomen gedeeld door het aantal elementen van de klasse  $C$  verzamelingen.

Notatie:  $support(X) = \frac{freq(X)}{\#(C)}$

Waarbij:

$freq(X)$  het aantal deelverzamelingen is waarin de elementen van  $X$  voorkomen.

Bijvoorbeeld is de collectie:

$$C = \{\{banaan\}, \{chips\}, \{melk\}, \{frum\}, \{banaan, rum\}, \{chips, rum\}, \{banaan, chips, melk\}, \{banaan, rum, melk, frum\}\}$$

Er geldt:

- $freq(\{banaan\}) = 4$ , want er zijn 5 deelverzamelingen waartoe element banaan behoort.
- $support(\{melk\}) = \frac{3}{8}$ , want er zijn 3 van de 8 deelverzamelingen van  $C$  waar het element *melk* in voorkomt.
- $support(\{banaan, rum\}) = \frac{2}{8}$ , want er zijn 2 van de 8 deelverzamelingen waarin elementen banaan en rum voorkomen.

Bij **marktonderzoek** is de support van een stel artikelen het aantal klanten dat al die artikelen koopt, meestal als percentage van het totaal.

Een stel artikelen met hoge support (boven een zekere drempel) heet frequent

Itemset is een deelverzameling van elementen van een stel artikelen.  $k$ -itemset is een deelverzameling die  $k$  elementen bevatten.

Er moet voldoende vertrouwen zijn in de associatieregel. Om dit te kunnen bepalen, berekent men de confidence van de associatieregel.

De confidence van  $X \Rightarrow Y$  is per definitie de support van de vereniging van  $X$  en  $Y$  gedeeld door de support van  $X$ .

In notatievorm:  $confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)}$

Omdat  $\frac{support(X \cup Y)}{support(X)} = \frac{freq(X \cup Y)}{freq(X)}$ , er geldt ook:

$$confidence(X \Rightarrow Y) = \frac{freq(X, Y)}{freq(X)}$$

Bij marktonderzoek geeft dit een maat voor de kans dat iemand  $Y$  koopt, gegeven dat hij  $X$  ook in zijn boodschappenmandje heeft.

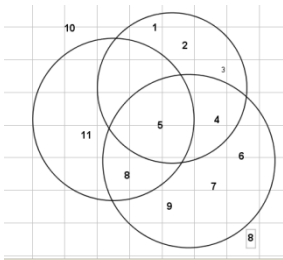
### Voorbeeld C

Associatieregel:  $\{banaan\} \Rightarrow \{rum\}$

$$C = \left\{ \begin{array}{l} \{banaan\}, \{chips\}, \{melk\}, \{rum\}, \{banaan, rum\}, \{chips, rum\} \\ \{banaan, chips, melk\}, \{banaan, rum, melk, rum\} \end{array} \right\}$$

$$\begin{aligned} \text{confidence}(\{banaan\} \Rightarrow \{rum\}) &= \frac{\text{support}(\{banaan, rum\})}{\text{support}(\{banaan\})} \\ &= \frac{2/8}{4/8} = 0.5 \text{ (50\%)} \end{aligned}$$

Want:  $\text{support}(\{banaan, rum\}) = \frac{2}{8}$  en  $\text{support}(\{banaan\}) = \frac{4}{8}$



	Items
$T_1$	$\{Chips, zeep, appels\}$
$T_2$	$\{Chips, zeep\}$
$T_3$	$\{Zeep, bananen\}$
$T_4$	$\{Chips, zeep, bananen\}$

Tabel 3.2

	Items
1	$\{Pasta, Kaas, Tomaat\}$
2	$\{Pasta, Room\}$
3	$\{Pasta, Tomaat\}$
4	$\{Kaas, Room\}$
5	$\{Kaas, Tomaat\}$
6	$\{Pasta, Tomaat, Room\}$
7	$\{Tomaat, Room\}$
8	$\{Kaas, Tomaat\}$
9	$\{Pasta, Kaas, Tomaat, Room\}$
10	$\{Pasta\}$

Tabel 3.4

### Vragen en opdrachten

1.

Gegeven:  $A = \{1,2,3,4,5\}$ ,  $B = \{4,5,6,7\}$  en  $C = \{5,8,11\}$ .

- Schrijf de verzamelingen  $A \cup B$ ,  $A \cup C$  en  $B \cup C$  op.
- Schrijf de verzamelingen  $A \cap B$ ,  $A \cap C$  en  $B \cap C$  op.
- Schrijf de verzameling  $A \cup B \cup C$  op.

2.

Gegeven de volgende deelverzamelingen:

$$T_1 = \{Chips, zeep, appels\}, T_2 = \{Chips, zeep\},$$

$$T_3 = \{Zeep, Bananen\}, T_4 = \{Chips, zeep, bananen\}$$

- Schrijf de verzamelingen  $T_1 \cap T_2$ ,  $T_1 \cap T_4$  en  $T_3 \cap T_4$  op.
- Schrijf de verzamelingen  $T_1 \cup T_2$ ,  $T_1 \cup T_4$  en  $T_3 \cup T_4$  op.
- Schrijf de verzameling  $T_1 \cup T_2 \cup T_3 \cup T_4$  op.

3.

Gegeven de klasse verzamelingen A:

$$A = \{\{5,8\}, \{2,6,8\}, \{5,6,8\}, \{5,4,7,10\}, \{2,5,8\}\}$$

Bereken:

- $\text{Support}\{1\}$
- $\text{Support}\{4,5\}$
- $\text{Confidence}(\{1\} \Rightarrow \{5\})$
- $\text{Confidence}(\{2\} \Rightarrow \{8\})$

4.

Gegeven de klasse verzamelingen T:

$$T = \{\{P, K, T\}, \{P, R\}, \{P, T\}, \{K, R\}\}$$

$$P = \text{Pasta}, K = \text{Kaas}, T = \text{Tomaat}, R = \text{Room}$$

Bereken:

- $\text{Support}\{R\}$
- $\text{Support}\{P, T\}$
- $\text{Confidence}(\{P\} \Rightarrow \{T\})$

5.

Gegeven de klasse verzamelingen  $A$ :

$$A = \{\{a, b, c\}, \{b, c, d, e\}, \{c, d\}, \{a, b, d\}, \{a, b, e\}\}$$

Bereken:

- Support $\{b\}$
- Support  $\{a, b\}$
- Confidence( $\{a, b\} \Rightarrow \{c\}$ )

6.

Stel:  $S = Spaghetti$ ,  $T = Tomatensaus$ ,  $B = Brood$ .  $S_1$  is een collectie van de deelverzamelingen met grootte 1.  $S_1 = \{\{S\}, \{T\}, \{B\}\}$ . Verder is  $S_2$  is een collectie van de deelverzamelingen met grootte 2. Dus  $S_2 = \{\{S, T\}, \{T, B\}, \{S, B\}\}$ . En  $S_3$  is een collectie van de deelverzamelingen met grootte 3, dus  $S_3 = \{\{S, T, B\}\}$ .

- Bereken de support van alle deelverzamelingen van  $S_1$
- Bereken de support van de alle deelverzamelingen van  $S_2$
- Bereken de support van de alle deelverzamelingen van  $S_3$

ID	Items
1	Spaghetti, tomatensaus
2	Spaghetti, brood
3	Spaghetti, tomatensaus, brood
4	Brood, tomatensaus

Tabel 3.5

7.

Een kok experimenteert met gerechten met verschillende ingrediënten.

Bekijk de volgende deelverzameling:  $\{Pasta, Kaas, Tomaat\}$ .

Laat zien dat  $Confidence(\{Pasta, Kaas\} \Rightarrow \{Tomaat\})$  hoger is dan die van  $\{Pasta\} \Rightarrow \{Kaas, Tomaat\}$ .

8.

Gegeven de volgende dataset hiernaast:

$$C_1 = \{\{Spaghetti\}, \{Tomatensaus\}, \{Brood\}, \{Boter\}\}$$

Bereken de support van alle deelverzamelingen van  $C_1$  en vul de volgende tabel in:

$\{Spaghetti\},$	
$\{Tomatensaus\}$	
$\{Brood\}$	
$\{Boter\}$	

$$C_2 = \{\{Spaghetti, tomatensaus\}, \{Spaghetti, brood\},$$

$$\{tomatensaus, brood\}\}$$

Bereken de support van alle deelverzamelingen van  $C_2$  en vul de volgende tabel in:

$\{Spaghetti, Tomatensaus\}$	
$\{Tomatensaus, Brood\}$	
$\{Brood, Spaghetti\}$	

$$C_3 = \{\{Spaghetti, tomatensaus, brood\}\}$$

Bereken de support van alle deelverzamelingen van  $C_3$  en vul de volgende tabel in:

TID	Items
1	$\{Pasta, Kaas, Tomaat\}$
2	$\{Pasta, Room\}$
3	$\{Pasta, Tomaat\}$
4	$\{Kaas, Room\}$
5	$\{Kaas, Tomaat\}$
6	$\{Pasta, Tomaat, Room\}$
7	$\{Tomaat, Room\}$
8	$\{Kaas, Tomaat\}$
9	$\{Pasta, Kaas, Tomaat, Room\}$
10	$\{Pasta\}$

Tabel 3.6

3-items	Support
{ <i>Spaghetti, tomatensaus, brood</i> }	

Bereken de confidence van de volgende associatieregels en vul de tabel in:

Associatieregels	Confidence
{ <i>Spaghetti</i> } ⇒ { <i>Spaghetti, Tomatensaus</i> }	
{ <i>Spaghetti</i> } ⇒ { <i>Spaghetti, Brood</i> }	

## Apriori algoritme

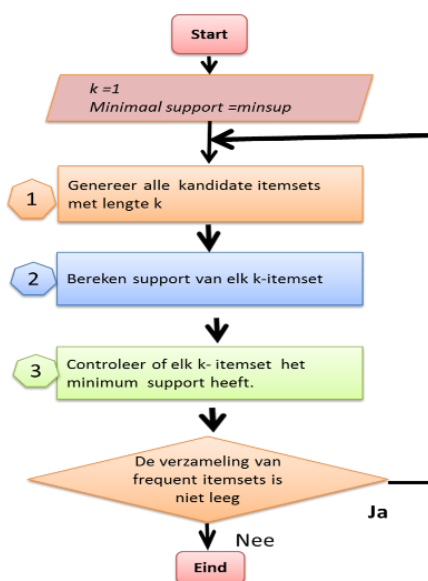
Er bestaan veel algoritmes om associatieregels te ontdekken. De oudste en meest bekende methode is het Apriori principe. Het is in de jaren negentig bedacht door Agrawal en anderen. Het is gebaseerd op de volgende observatie. Iedere deelverzameling van een frequente itemset is zelf ook weer een frequente itemset.

Het Apriori algoritme bestaat uit twee fasen:

- 1 Het genereren van verzamelingen of itemsets die mogelijk de minimale support hebben. Dit worden kandidaten genoemd.
- 2 Het controleren van de kandidaten die daarvan de minimale support hebben.

Deze fasen worden per niveau uitgevoerd, dus voor verzamelingen of items met grootte 1, met grootte 2, enzovoort.

Het algoritme stopt wanneer er geen kandidaten meer gevonden kunnen worden.



### Methode

**Start:**  $k = 1$

- ① Genereer alle deelverzamelingen (itemsets) kandidaten met lengte  $k$ .
- ② Bereken support van alle deze deelverzamelingen ( $k$ -itemsets)
- ③ Controleer met behulp van de minimaal support of deze  $k$ -itemsets kandidaten zijn. Alle itemsets die afgekeurde  $k$ -itemset kandidaten bevatten, zijn verwijderd en worden niet meer als kandidaten beschouwd.
- ④ Als er minstens één frequente  $k$ -itemsets bestaat, herhaal de stappen ①, ② en ③ voor alle  $k + 1$  itemsets 'kandidaten'. Anders ga naar je Einde.

**Einde:** Genereer alle associatieregels die aan minimaal confidence Voldoen.

TID	Itemset
1	<i>Spaghetti, Tomatensaus</i>
2	<i>Spaghetti, Brood</i>
3	<i>Spaghetti, Tomatensaus, Brood</i>
4	<i>Brood, Boter</i>
5	<i>Brood, Tomatensaus</i>

Tabel 3.7

### Voorbeeld C

Stel je hebt als winkel een groot aantal gegevens, zie Tabel 3.10.

Vraag: Genereer de k-itemsets met behulp van het Apriori-algoritme.

Zorg voor een minimale support van 40%.

### Stappenplan (Algoritme)

Start:  $k = 1$

#### Iteratie 1

①

In de eerste stap genereert het Apriori-algoritme alle deelverzamelingen (sets) van grootte 1:

$$C_1 = \{\{Spaghetti\}, \{Tomatensaus\}, \{Brood\}, \{Boter\}\}$$

②

Itemset	Support(%)
<i>{Spaghetti}</i>	60%
<i>{Tomatensaus}</i>	60%
<i>{Brood}</i>	80%
<i>{Boter}</i>	20%

③

*{Boter}* is geen kandidaat (minder frequent)

De frequente 1-itemsets zijn dus:

$$F_1 = \{\{Spaghetti\}, \{Tomatensaus\}, \{Brood\}\}$$

④

$F_1$  is niet leeg → ga naar **Iteratie 2**

#### Iteratie 2

①

Enkel die verzamelingen van lengte 2 waarvoor geldt dat al hun deelverzamelingen van lengte 1 frequent zijn, worden als kandidaat beschouwd:

$$C_2 = \{\{Spaghetti, tomatensaus\}, \{Spaghetti, brood\}, \{tomatensaus, brood\}\}$$

②

Itemset	Support(%)
<i>{Spaghetti, tomatensaus}</i>	40%
<i>{Spaghetti, brood}</i>	40%
<i>{tomatensaus, brood}</i>	40%

③

De frequente 2-itemsets zijn dus:

$$F_2 = \{\{Spaghetti, brood\}, \{Spaghetti, tomatensaus\}, \{tomatensaus, brood\}\}$$

④

$F_2$  is niet leeg → ga naar **Iteratie 3**

### Iteratie 3

①

$C_3 = \{\{Spaghetti, tomatensaus, brood\}\}$

②

3-Itemset	Support(%)
$\{Spaghetti, tomatensaus, brood\}$	20%

③

$\{Spaghetti, Tomatensaus, Brood\}$  is geen kandidaat.

$F_3 = \{\}$

④

$F_3$  is leeg → ga naar **Einde**

### Einde

De uiteindelijke set van frequente verzameling itemsets is:

$F = \{\{\}, \{Spaghetti\}, \{Tomatensaus\}, \{Brood\}, \{Spaghetti, Brood\}, \{Spaghetti, Tomatensaus\}, \{Tomatensaus, Brood\}\}$


Associatieregels	Confidence(%)
$\{Spaghetti\} \Rightarrow \{Spaghetti, Tomatensaus\}$	67%

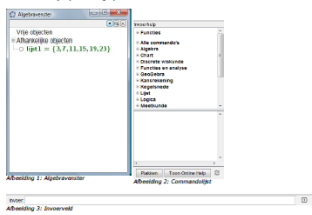
## 3.2 Associatieanalyse met Geogebra

### Invoerveld

In deze cursus wordt meermaals gevraagd een code in te voeren. Tenzij anders aangegeven, is het de bedoeling dat de code wordt ingevuld in het invoerveld (Afbeelding 3).

### Commandolijst

Open de commandolijst door te klikken op  (Afbeelding 3). Aan de rechterkant verschijnt nu de commandolijst (Afbeelding 2).



In deze paragraaf ga je met behulp van het computerprogramma Geogebra berekeningen doen met betrekking tot associatieanalyse.

## Verzamelingen

In de invoerregel van Geogebra kun je bijvoorbeeld verzamelingen als lijsten invoeren.

### Voorbeeld A

Gegeven de volgende dataset.

TID	Items
1	<i>Appel, Banaan, Kiwi, Citroen</i>
2	<i>Appel, Peer, Kiwi, Citroen</i>

Voer deze data in bij Geogebra.

$$D_1 = \{Appel, Banaan, Kiwi, Citroen\} \rightarrow$$

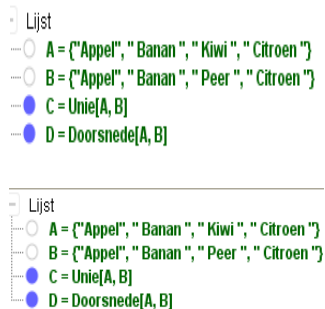
$$Lijst A = \{ "Appel", " Banaan ", " Kiwi ", " Citroen " \}$$


$$D_2 = \{Appel, Banaan, Kiwi, Citroen\} \rightarrow$$

$$Lijst B = \{ "Appel", " Banaan ", " Peer ", " Citroen " \}$$

TID	Items
1	<i>Appel, Banaan, Kiwi, Citroen</i>
2	<i>Appel, Peer, Kiwi, Citroen</i>

Tabel 3.8



Met de optie commandolijst kun je verzamelingen creëren. Open de commandolijst door te klikken op  (Afbeelding 3). Aan de rechterkant verschijnt de commandolijst (Afbeelding 2).

Om de verenigingen van twee verzamelingen  $A$  en  $B$  te bepalen voer je in de invoerregel in: **Unie** $[A, B]$ .

$$A \cup B = \text{Unie}[A, B] = \{ "Appel", " Banaan ", " Kiwi ", " Citroen ", " Peer " \}.$$

De optie **Unie** $[A, B]$  geeft de verenigingen tussen twee verzamelingen  $A$  en  $B$ .

Om de doorsnede van twee verzamelingen  $A$  en  $B$  te bepalen voer je in de invoerregel in: **Doorsnede** $[A, B]$ .

$$A \cap B = \text{Doorsnede}[A, B] = \{ "Appel", " Banaan ", " Citroen " \}.$$

De optie **Doorsnede** $[A, B]$  geeft de verenigingen tussen twee verzamelingen  $A$  en  $B$ .

Op dezelfde manier kun je ook een klasse van verzamelingen als lijst van deelverzamelingen invoeren.



### Voorbeeld B

Gegeven de volgende dataset:

TID	Items
1	<i>Appel</i>
2	<i>Appel, Banaan, Kiwi</i>
3	<i>Banaan, Citroen</i>
4	<i>Banaan, Kiwi</i>

Voer deze data in bij Geogabra.

$$D_1 = \{Appel\} \rightarrow L_1 = \{\"Appel\"\}$$

$$D_2 = \{Appel, Banaan, Kiwi\} \rightarrow L_2 = \{\"Appel\", \"Banaan\", \"Kiwi\"\}$$

$$D_3 = \{Banaan, Citroen\} \rightarrow L_3 = \{\"Appel\", \"Banaan\", \"Kiwi\"\}$$

$$D_4 = \{Banaan, Kiwi\} \rightarrow L_4 = \{\"Banaan\", \"Kiwi\"\}$$

$D$

$$= \{\{Appel\}, \{Appel, Banaan, Kiwi\}, \{Banaan, Citroen\}, \{Banaan, Kiwi\}\}$$

$$\rightarrow L = \{L_1, L_2, L_3, L_4\}$$

ID	Items
1	<i>Spaghetti, tomatensaus</i>
2	<i>Spaghetti, brood</i>
3	<i>Spaghetti, tomatensaus, brood</i>
4	<i>Brood, tomatensaus</i>

Tabel 3.9

Om de support van een  $k$ -item  $X$  te berekenen, moet je eerst  $\text{frq}(X)$  berekenen.

$\text{frq}(X)$  is het aantal deelverzamelingen  $D_i$  van de dataset  $D$  waarin  $X$  een deelverzameling is.

Als  $X$  een deelverzameling van  $D_i$  is, dan  $f_i = 1$ , anders 0.

$$\rightarrow f_i = \text{Als}[\text{Doorsnede}[X, D_i] == X], 1, 0]$$

Om de lengte van de dataset te berekenen gebruik je optie  $\text{Lengte}[\text{lijst}] \rightarrow n = \text{Lengte}[D]$

$$\text{support}(X) = \frac{\text{frq}(X)}{\#(C)} \rightarrow \text{support} = \text{Gem}[f_1 : f_n]$$

TID	Items
1	<i>\{Pasta, Kaas, Tomaat\}</i>
2	<i>\{Pasta, Room\}</i>
3	<i>\{Pasta, Tomaat\}</i>
4	<i>\{Kaas, Room\}</i>
5	<i>\{Kaas, Tomaat\}</i>
6	<i>\{Pasta, Tomaat, Room\}</i>
7	<i>\{Tomaat, Room\}</i>
8	<i>\{Kaas, Tomaat\}</i>
9	<i>\{Pasta, Kaas, Tomaat, Room\}</i>
10	<i>\{Pasta\}</i>

Tabel 3.10

# Opdrachten



Bij uitvoering van de opdrachten van 1 tot en met 6 maak je gebruik maken van de Geogebra commando's.

1.

Gegeven:  $A = \{1,2,3,4,5\}$ ,  $B = \{4,5,6,7\}$  en  $C = \{5,8,11\}$

Bepaal de volgende verzamelingen:

- $A \cup B$ ,  $A \cup C$  en  $B \cup C$ .
- $A \cap B$ ,  $A \cap C$  en  $B \cap C$ .
- $A \cup B \cup C$ .

2.

Gegeven de volgende deelverzamelingen:

$T_1 = \{Chips, zeep, appels\}$ ,  $T_2 = \{Chips, zeep\}$

$T_3 = \{Zeep, bananen\}$  en  $T_4 = \{Chips, zeep, bananen\}$

- Schrijf de verzamelingen  $T_1 \cap T_2$ ,  $T_1 \cap T_4$  en  $T_4 \cap T_3$  op.
- Schrijf de verzamelingen  $T_1 \cup T_2$ ,  $T_1 \cup T_4$  en  $T_4 \cup T_3$  op.
- Schrijf de verzameling  $T_1 \cup T_2 \cup T_3 \cup T_4$  op.

3.

Gegeven de klasse verzamelingen  $T$ :

$T = \{\{P, K, T\}, \{P, R\}, \{P, T\}, \{K, R\}\}$

$P = Pasta$ ,  $K = Kaas$ ,  $T = Tomaat$ ,  $R = Room$

Bereken:

- Support  $\{R\}$
- Support  $\{P, T\}$
- Confidence( $\{P\} \Rightarrow \{T\}$ )

4.

Stel:  $S = Spaghetti$ ,  $T = Tomatensaus$ ,  $B = Brood$ .  $S_1$  is een collectie van de deelverzamelingen met grootte 1 dus  $S_1 = \{\{S\}, \{T\}, \{B\}\}$ . Verder is  $S_2$  een collectie van de deelverzamelingen met grootte 2, dus  $S_2 = \{\{S, T\}, \{T, B\}, \{S, B\}\}$ . En  $S_3$  is een collectie van de deelverzamelingen met grootte 3 dus  $S_3 = \{\{S, T, B\}\}$ .

- Bereken de support van alle deelverzamelingen van  $S_1$ .
- Bereken de support van de alle deelverzamelingen van  $S_2$ .
- Bereken de support van de alle deelverzamelingen van  $S_3$ .

5.

Een kok experimenteert met gerechten met verschillende ingrediënten. Bekijk de volgende deelverzameling:

$\{Pasta, Kaas, Tomaat\}$

- Bereken de support van  $\{Pasta, Kaas, Tomaat\}$ .
- Laat zien dat Confidence( $\{Pasta, Kaas\} \Rightarrow \{Tomaat\}$ ) hoger is dan Confidence( $\{Pasta\} \Rightarrow \{Kaas, Tomaat\}$ ).

TID	Items
1	<i>a, b, c</i>
2	<i>b, c, d, e</i>
3	<i>c, d</i>
4	<i>a, b, d</i>
5	<i>a, b, c</i>

Tabel 3.11

TID	Items
1	<i>Brood, Melk</i>
2	<i>Brood, Luier, Bier, Eieren</i>
3	<i>Melk, Luier, Bier, Cola</i>
4	<i>Brood, Melk, Luier, Bier</i>
5	<i>Brood, Melk, Luier, Cola</i>

Tabel 3.12

TID	Items
1	<i>a, b, c, d</i>
2	<i>a, b, d</i>
3	<i>b, c, d, e</i>
4	<i>b, c, e</i>
5	<i>a, d, e</i>
6	<i>f, g</i>
7	<i>g</i>
8	<i>a, d, g</i>
9	<i>b, d, f</i>
10	<i>g, f</i>

Tabel 3.13



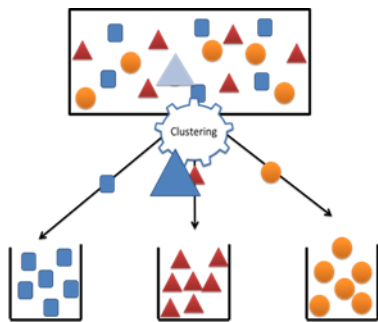
Bij uitvoering van de opdrachten van 6 tot en met 9 maak je gebruik maken van de Geogebra applet **Associatie.ggb**

6.
  - a. Illustreer het Apriori-algoritme door te tonen hoe dit algoritme alle itemsets met een minimale support van 50% in deze dataset vindt.
7.
  - a. Pas het Apriori-algoritme toe de dataset gegeven in Tabel 3.11.
  - b. Geef alle associatieregels met een support van minimaal 40% en Confidence van minimaal 70% in de volgende transactiedatabase.
8.
 

Pas het Apriori-algoritme toe op de dataset in Tabel 3.12. Toon in jouw oplossing de verschillende tussenstappen.

Leg uit hoe deze deelverzamelingen gebruikt kunnen worden om alle associatieregels te vinden met een support van 50% en een confidence van 60%.
9.
 

Gebruik het Apriori-algoritme toe op de dataset gegeven in Tabel 3.13 om alle associatieregels te vinden met een confidence van 80% en een minimale support van 40% in deze dataset.



Figuur 3.4 Het classificeren van objecten

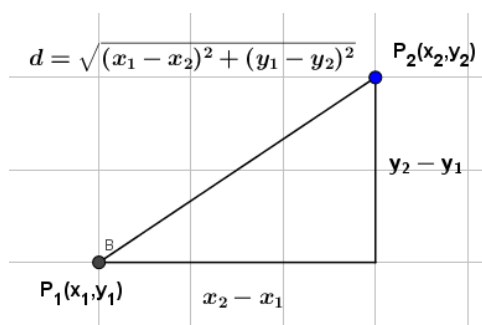
## 3.3 Clusteranalyse

### Inleiding

Clusteranalyse is het classificeren of het groeperen in 'clusters' of 'klassen' van objecten op grond van hun kenmerken. Het doel van clusteranalyse is het vormen van deelverzamelingen die elk hun eigen gedeelde kenmerken bevatten.

Het doel van clusteranalyse is het verdelen van een dataset in groepen. Deze groepen en het aantal groepen zijn vooraf niet bekend. Het streven is zoveel mogelijk gelijkheid binnen een groep en zoveel mogelijk verschil tussen de groepen te krijgen. In tegenstelling tot classificatie: daar weten we de indeling in groepen al, en willen we een nieuw object in de juiste groep krijgen. Clusteranalyse is vooral bekend uit marktonderzoek en wordt veel ingezet om het koopgedrag van klanten te onderzoeken. Het doel van de analyse is niet het voorspellen van dit koopgedrag, maar het zoeken naar een beperkt aantal groepen klanten met hetzelfde koopgedrag. Online kan deze techniek heel goed worden ingezet om websitebezoek te onderzoeken. Hiermee kunnen verschillende doelgroepen op een effectieve wijze benaderd worden. Een bedrijf krijgt zo zicht op welk product of dienst een groep klanten het beste aansluit, en welke product eventueel niet haalbaar of minder bereikbaar zijn voor de klantengroep.

In de biologie zijn er meerdere gebieden waar clusteranalyse wordt toegepast. Denk bijvoorbeeld aan de classificatie van verschillende organismen. Elk organisme hoort bij een soort. Soorten kunnen op hun beurt weer worden onderverdeeld in lagere taxa, zoals ondersoort en variëteit. Soorten zelf worden samengevoegd in geslachten en deze weer in families en in taxa van nog hogere rang. Een ander voorbeeld van het gebruiken clustertechnieken in de biologie is het maken van groepen met genen die zie een bepaalde erfelijke ziekte kunnen bevatten. Door het gebruik van clustermethodes kunnen de groepen met genen gevonden worden. Als deze groepen bekend zijn, wordt het gemakkelijk een medicijn te ontwikkelen dat de erfelijke ziekte kan voorkomen of genezen.

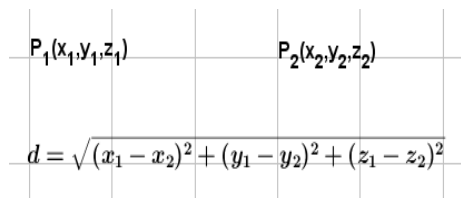


Figuur 3.5 Stelling van Pythagoras

### Afstandsmaten

#### Afstand tussen twee punten

In de wiskunde kan de afstand worden berekend als de wortel uit de som van de kwadraten van de verschillen tussen de coördinaten, volgens de stelling van Pythagoras, zie Figuur 3.5. In drie dimensies geldt analoog hiervoor de Euclidische afstand, zie Figuur 3.6



Figuur 3.6 Euclidische afstand bij drie dimensies

#### Afstand tussen datapunten (objecten):

Dit is een maat die aangeeft hoe groot de 'overeenkomst' of het 'verschil' is tussen twee kenmerk datapunten.

Er zijn twee bekende methoden om de afstand tussen waarnemingen:

De definitie van de **Euclidische afstand** is:

$$d(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

$$= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2}$$

Waarbij:

<b>A</b>	<b>B</b>	<b>A - B</b>	<b>(A - B)<sup>2</sup></b>
<b>a<sub>1</sub></b>	<b>b<sub>1</sub></b>	<b>a<sub>1</sub> - b<sub>1</sub></b>	<b>(a<sub>1</sub> - b<sub>1</sub>)<sup>2</sup></b>
<b>a<sub>2</sub></b>	<b>b<sub>2</sub></b>	<b>a<sub>2</sub> - b<sub>2</sub></b>	<b>(a<sub>2</sub> - b<sub>2</sub>)<sup>2</sup></b>
<b>a<sub>i</sub></b>	<b>b<sub>i</sub></b>	<b>a<sub>i</sub> - b<sub>i</sub></b>	<b>(a<sub>i</sub> - b<sub>i</sub>)<sup>2</sup></b>
<b>a<sub>n</sub></b>	<b>b<sub>n</sub></b>	<b>a<sub>n</sub> - b<sub>n</sub></b>	<b>(a<sub>n</sub> - b<sub>n</sub>)<sup>2</sup></b>

De definitie van de **Manhattan of City-block afstand** is

$$d(A, B) = \sum_{i=1}^n |a_i - b_i|$$

$$= |a_1 - b_1| + |a_2 - b_2| + \dots + |a_n - b_n|$$

Waarbij:

<b>A</b>	<b>B</b>	<b>A - B</b>	<b> A - B </b>
<b>a<sub>1</sub></b>	<b>b<sub>1</sub></b>	<b>a<sub>1</sub> - b<sub>1</sub></b>	<b> a<sub>1</sub> - b<sub>1</sub> </b>
<b>a<sub>2</sub></b>	<b>b<sub>2</sub></b>	<b>a<sub>2</sub> - b<sub>2</sub></b>	<b> a<sub>2</sub> - b<sub>2</sub> </b>
<b>a<sub>i</sub></b>	<b>b<sub>i</sub></b>	<b>a<sub>i</sub> - b<sub>i</sub></b>	<b> a<sub>i</sub> - b<sub>i</sub> </b>
<b>a<sub>n</sub></b>	<b>b<sub>n</sub></b>	<b>a<sub>n</sub> - b<sub>n</sub></b>	<b> a<sub>n</sub> - b<sub>n</sub> </b>

### Voorbeeld A

Geef de afstand tussen de volgende objecten:

Object A	10	12	15	13	9
Object B	18	23	13	15	17

De Manhattan of City-block afstand is:

$$d(A, B) =$$

$$= |10 - 18| + |12 - 23| + |15 - 13| + |13 - 15| + |9 - 17| = 25$$

De Euclidische afstand is:

$$d(A, B) =$$

$$= \sqrt{(10 - 18)^2 + (12 - 23)^2 + (15 - 13)^2 + (13 - 15)^2 + (9 - 17)^2}$$

$$= 16,16$$

### Afstandsmatrix

Een matrix van afstanden is een matrix waarvan de elementen de afstanden tussen de punten aangeven.

Twee dimensionale Euclidische afstand:

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

#### Voorbeeld B

Gegeven een dataset met vier objecten. Geef de afstandsmatrix.

	$x$	$y$
A	2	1
B	4	2
C	6	1
D	7	2

De afstandsmatrix ziet eruit als volgt:

Afstand	A	B	C	D
A	0	2.24	4	5.1
B	2.24	0	2.24	3
C	4	2.24	0	1.41
D	5.1	3	1.41	0

De afstandsmaat die hier wordt gebruikt is de gewone Euclidische afstand.

### Afstand tussen clusters

Als je de afstand tussen elk paar van objecten weet, wat is dan de afstand tussen twee clusters  $C_1$  en  $C_2$ ?

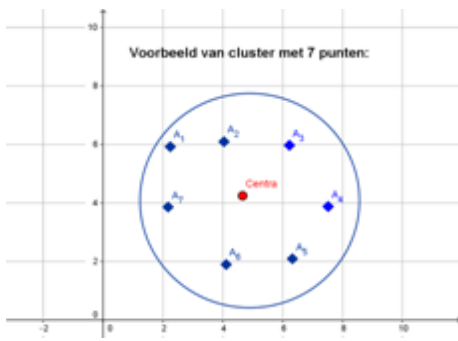
#### Single Linkage

De afstand tussen twee clusters  $C_1$  en  $C_2$  is:

$$d(C_1, C_2) = \min\{d(A, B) \mid A \in C_1 \text{ en } B \in C_2\}$$

#### Euclidische afstand

$$d(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$
$$= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2}$$



Figuur 3.7 Het centrum van een cluster met zeven objecten

### Voorbeeld C

Gegeven een dataset met zes punten.

	x	y
A <sub>1</sub>	1	3
A <sub>2</sub>	1	4
A <sub>3</sub>	2	2
B <sub>1</sub>	5	1
B <sub>2</sub>	5	2
B <sub>3</sub>	7	2

De clusters zijn  $C_1 = \{A_1, A_2, A_3\}$  en  $C_2 = \{B_1, B_2, B_3\}$ .

De afstandsmatrix is

	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
A <sub>1</sub>	4,47	4,12	6,03
A <sub>2</sub>	5	4,47	6,32
A <sub>3</sub>	3,16	3	5

Geef de minimale afstand tussen  $C_1$  en  $C_2$ .

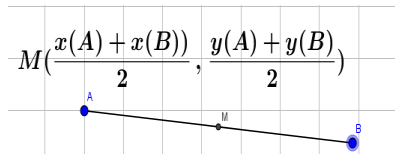
$$d(C_1, C_2) = d(C_1, C_2) = \min\{d(A, B) | A \in C_1, B \in C_2\}$$

$$d(A_3, B_2) = \sqrt{(5-2)^2 + (2-2)^2} = 3$$

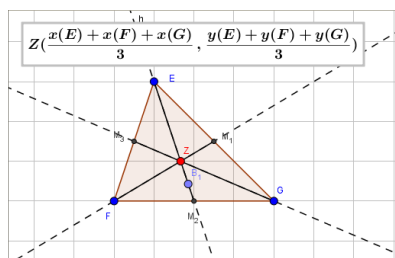
Met de groene kleur wordt de minimale afstand aangegeven.

### Het centrum van een cluster

In de wiskunde is het centrum van een cluster met twee punten  $A$  en  $B$  het **middelpunt**  $M$  van het lijnstuk  $[A, B]$ .



Heb je te maken met drie punten, dan gebruik je het zwaartepunt. Hieronder zie je de drie zwaartelijnen in een driehoek die door één punt gaan. Dat punt heet het zwaartepunt  $Z$  van de driehoek.



Het centrum van een cluster met drie punten  $E, F$  en  $G$  moet gezien worden als het **zwaartepunt** van een driehoek  $\Delta EFG$

Het centrum  $C(x, y)$  van een cluster bepaal je door gemiddelde van alle punten  $A_i(x_i, y_i, \dots, z_i)$  in die cluster te

Om het **centrum** van een cluster te berekenen gebruik je de volgende formule

$$M(x, y, \dots, z) = \left( \frac{\sum_1^n x_i}{n}, \frac{\sum_1^n y_i}{n}, \dots, \frac{\sum_1^n z_i}{n} \right)$$

Ofwel

$$M(x, y, \dots, z) = \left( \frac{x_1+x_2+\dots+x_n}{n}, \frac{y_1+y_2+\dots+y_n}{n}, \dots, \frac{z_1+z_2+\dots+z_n}{n} \right)$$

### Voorbeeld D

Gegeven een cluster  $C$  die de volgende objecten bevat:

Datapunten	x	y
A <sub>1</sub>	0	0
A <sub>2</sub>	7	8
A <sub>3</sub>	4	8
A <sub>4</sub>	3	0

Bereken het centrum van de cluster  $C$ .

$$C = \{A_1, A_2, A_3, A_4\}$$

$$M((0+7+4+3)/4, (0+8+8+0)/4)$$

Het centrum van de cluster is  $M(3.5, 4)$ .

# Clustermethoden

Er zijn twee methodes om tot clusters te komen: hiërarchische of partitiemethode. In deze paragraaf gaan we ons beperken tot niet-hiërarchisch clustering.

Niet-hiërarchisch of partitioneren betekent het stap voor stap verbeteren van een bestaande clustering.

Het basialgoritme is hier:

- a. Begin met een willekeurige clustering in  $k$  clusters. (nadeel:  $k$  ligt vooraf vast)
- b. Herhaal (totdat je meent klaar te zijn): stop een object in een andere cluster, zodanig dat de kwaliteit van de clusters verbetert

## K-means-clustering

Het oudste en meest bekende clusteralgoritme is de K-means methode. Later zijn er veel varianten ontwikkeld die gebaseerd zijn op deze methode. K-means is een eenvoudige, iteratieve manier van clusteren. Vooraf wordt bepaald hoeveel clusters je wilt hebben. Om de optimale verbetering van een bestaande clustering ga je als volgt te werk:

**Start:** Kies de centra van de clusters eerste keer gewoon willekeurig.

① Daarna bepaal je van elk datapunt de afstand tot ieder centrum, en wijs het datapunt toe aan de cluster waarvan het centrum het dichtstbij is.

② Nadat alle datapunten aan een cluster zijn toegevoegd, bereken je de centums van de clusters opnieuw. Hiervoor gebruik je de volgende formule:

$$C(x, y, z) = \left( \frac{\sum_1^n x_i}{n}, \frac{\sum_1^n y_i}{n}, \dots, \frac{\sum_1^n z_i}{n} \right)$$

③ Als de centums niet veranderen, herhaal dan de stappen ① en ②. Dit gaat door de centums niet meer veranderen, of totdat er een vooraf gekozen aantal stappen is geweest.

## Voorbeeld E

Gegeven de volgende dataset:

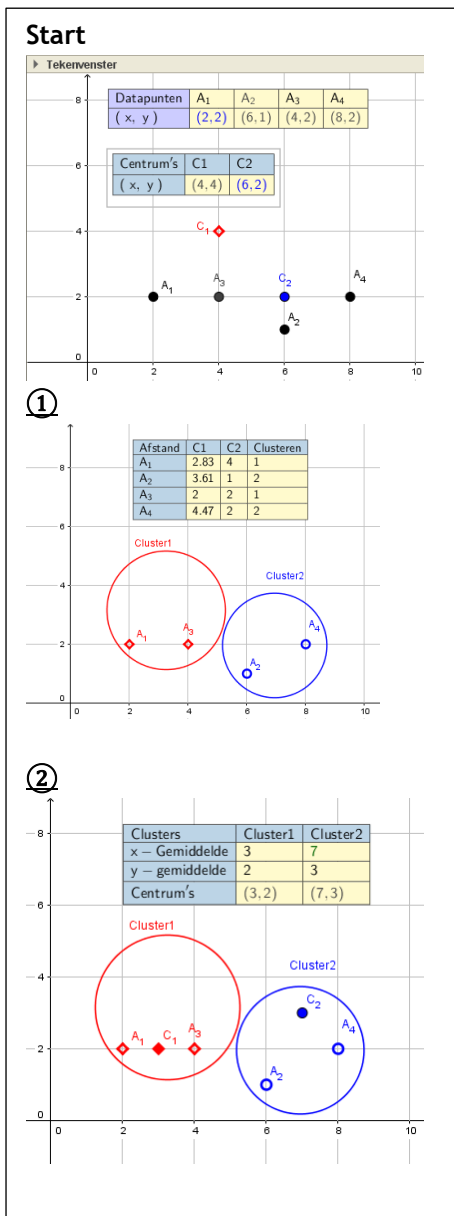
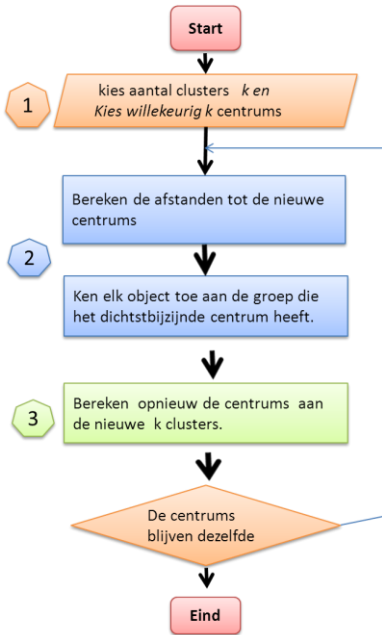
Datapunten	$x$	$y$
$A_1$	2	2
$A_2$	4	2
$A_3$	6	1
$A_4$	8	2

We bepalen vooraf dat we twee clusters willen, dus  $k = 2$ . Geef de uitwerking van de eerste iteratie.

**Start**

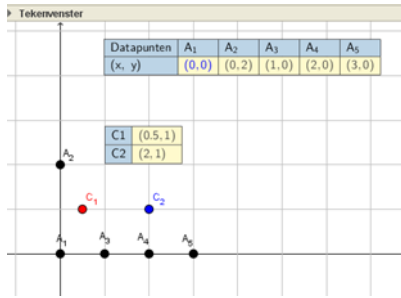
Kies twee willekeurig centums, bijvoorbeeld:

$$M_1 = (0.5, 1)$$





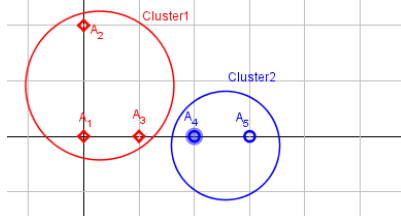
## Start



## Iteratie 1

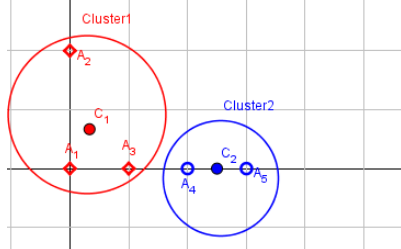
①

Afstand	C1	C2	Clusteren
A <sub>1</sub>	1,12	2,24	1
A <sub>2</sub>	1,12	2,24	1
A <sub>3</sub>	1,12	1,41	1
A <sub>4</sub>	1,8	1	2
A <sub>5</sub>	2,69	1,41	2



②

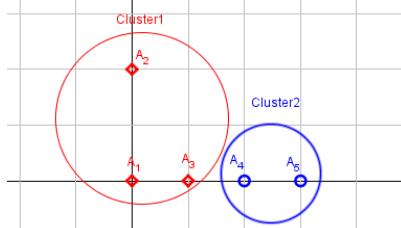
Clusters	Cluster1	Cluster2
x – Gemiddelde	0,33	2,5
y – gemiddelde	0,67	0
Centrum	(0,33, 0,67)	(2, 0)



## Iteratie 2

①

Afstand	C1	C2	Clusteren
A <sub>1</sub>	0,75	2	1
A <sub>2</sub>	1,37	2,83	1
A <sub>3</sub>	0,94	1	1
A <sub>4</sub>	1,8	0	2
A <sub>5</sub>	2,75	1	2



$$M_2 = (2,1)$$

## Iteratie 1

①

Als afstandsmaat gebruik je de normale (Euclidische) afstand.

Afstand	C <sub>1</sub>	C <sub>2</sub>	Toewijzing
A <sub>1</sub>	2,83	4	1
A <sub>2</sub>	3,61	1	2
A <sub>3</sub>	2	2	1
A <sub>4</sub>	4,47	2	2

De nieuwe clusters zijn:

C <sub>1</sub>	{A <sub>1</sub> , A <sub>3</sub> }
C <sub>2</sub>	{A <sub>2</sub> , A <sub>4</sub> }

②

De nieuwe centrusms zijn:

$$M_1 = (3,2)$$

$$M_2 = (7,3)$$

## Voorbeeld F

Gegeven de volgende dataset:

Datapunten	x	y
A <sub>1</sub>	0	0
A <sub>2</sub>	2	0
A <sub>3</sub>	0	1
A <sub>4</sub>	0	2
A <sub>5</sub>	3	0

## Start

We zijn op zoek naar twee clusters, dus  $k = 2$ . Als afstandsmaat kun je de normale (Euclidische) afstand gebruiken. Geef de uitwerking van de eerste twee iteraties.

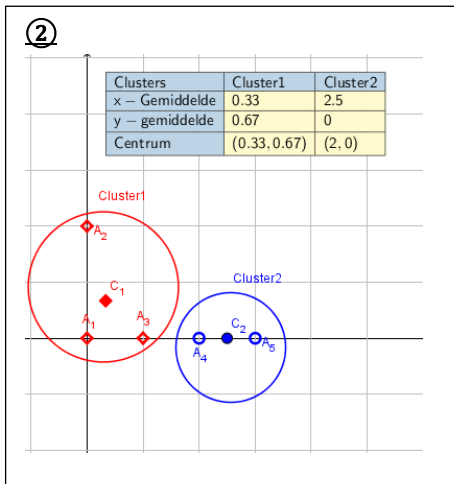
$$M_1 = (0,5,1)$$

$$M_2 = (2,1)$$

## Iteratie 1

①

Afstand	C <sub>1</sub>	C <sub>2</sub>	Clusteren
A <sub>1</sub>	1,12	2,24	1
A <sub>2</sub>	1,12	2,24	1
A <sub>3</sub>	1,12	1,41	1
A <sub>4</sub>	1,8	1	2
A <sub>5</sub>	2,69	1,41	2



Nieuwe clusters:

$C_1$	$\{A_1, A_2, A_3\}$
$C_2$	$\{A_4, A_5\}$

②

De nieuwe centrusms zijn

$M_1 = \left(\frac{1}{3}, \frac{2}{3}\right)$
$M_2 = (2, 0)$

③

Herhaal nu de stappen ① en ② → Iteratie 2

Iteratie 2

①

Afstand	$C_1$	$C_2$	Toekenning
$A_1$	0,75	2	1
$A_2$	1,37	2,83	1
$A_3$	0,94	1	1
$A_4$	1,8	0	2
$A_5$	2,75	1	2

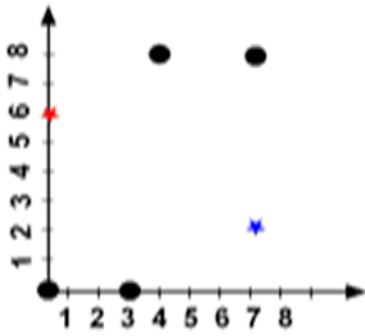
②

Er niets veranderd. De centrusms blijven dezelfde.

$$M_1 = \left(\frac{1}{3}, \frac{2}{3}\right)$$

$$M_2 = (2, 0)$$

In dit geval kunt je zeggen dat  $k$ -means algoritme convergeert.



Figuur 3.10 De objecten van opdracht 4

## Vragen en opdrachten

1.

Gegeven is een dataset met 6 datapunten.

Data	$X_1$	$X_2$	$X_3$	$X_4$
$A_1$	6	3	4	5
$A_2$	2	3	5	4
$A_3$	5	4	6	3
$A_4$	9	1	1	8
$A_5$	8	2	0	9
$A_6$	8	0	1	8

Als afstandsmaat gebruiken we de normale (Euclidische) afstand

Bereken de Euclidische afstand van  $A_1$  en  $A_2$

Vul de volgende tabel in:

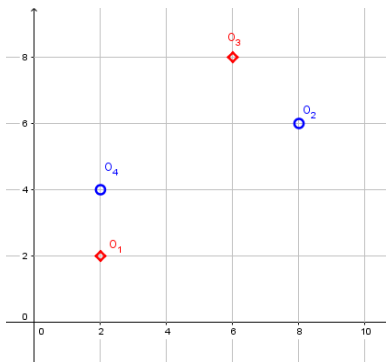
Afstand	$A_4$	$A_5$	$A_6$
$A_1$			
$A_2$			
$A_3$			

Clusteren deze gegevens in twee clusters:

$C_1 = \{A_1, A_2, A_3\}$  en  $C_2 = \{A_4, A_5, A_6\}$ .

Geef de afstand tussen clusters  $C_1$  en  $C_2$ .

Bereken de centrums van clusters  $C_1$  en  $C_2$ .



Figuur 3.9 De objecten in opdracht 2

2.

Gegeven de volgende vier objecten:

Object	$x_1$	$x_2$
$O_1$	2	2
$O_2$	8	6
$O_3$	6	8
$O_4$	2	4

We willen deze gegevens clusteren in twee clusters ( $k = 2$ ), met behulp van het k-means algoritme.

Initialiseren het algoritme: objecten 1 en 3 in één cluster  $C_1$  en objecten 2 en 4 in de andere cluster  $C_2$ .

Noteer  $C_1 = \{O_1, O_3\}$  en  $C_2 = \{O_2, O_4\}$

Als afstandsmaat gebruik je de normale (Euclidische) afstand.

Bereken het centrum  $M_1$  van  $C_1$  en het centrum  $M_2$  van  $C_2$

Vul de onderstaande tabel in en bereken daarmee de afstand tussen de cluster  $C_1$  en  $C_2$ .

Afstand	$O_2$	$O_4$
$O_1$		
$O_3$		

Bereken de afstanden tot de centra's  $M_1$  en  $M_2$  en vul de volgende tabel in:

Afstand	$M_1$	$M_2$
$O_1$		
$O_2$		
$O_3$		
$O_4$		

3.

Gegeven is een dataset met 6 datapunten (zie tabel hieronder):

Data	$x_1$	$x_2$
$D_1$	6	3
$D_2$	2	3
$D_3$	5	4
$D_4$	9	1
$D_5$	8	2
$D_6$	8	0

We willen deze gegevens clusteren in twee clusters ( $k = 3$ ) met behulp van het  $k$ -means algoritme.

We willen objecten  $D_1$  en  $D_2$  in één cluster  $C_1$ , de objecten  $D_3$  en  $D_4$  in de cluster  $C_2$  en  $D_5$  en  $D_6$  in de cluster  $C_3$ .

Dus:  $C_1 = \{D_1, D_2\}$ ,  $C_2 = \{D_3, D_4\}$  en  $C_3 = \{D_5, D_6\}$

Als afstandsmaat gebruik je de normale (euclidische) afstand

Bereken de centra's  $M_1$  voor  $C_1$ ,  $M_2$  voor  $C_2$  en  $M_3$  voor  $C_3$ .

Bereken de afstanden tot de centra's  $M_1$  en  $M_2$  en vul de volgende tabel in:

Afstand	$M_1$	$M_2$	$M_3$
$O_1$			
$O_2$			
$O_3$			
$O_4$			

4.

Gegeven is een dataset met vier objecten.

Objecten	$x$	$y$
$A_1$	0	0
$A_2$	7	8
$A_3$	4	8
$A_4$	3	0

Cluster de objecten door middel van het  $k$ -means algoritme. Geef de uitwerking van de tot twee iteraties.

Start het algoritme met de willekeurige centra's  $(0,6)$  en  $(7,2)$

Noteer voor elke iteratie welke cluster gevormd worden en wat de centra's van de cluster zijn en vul deze resultaten in de onderstaande tabel in:

	Clusters	Centrums
Start		$M_1 = (0,6)$ $M_2 = (7,2)$
Iteratie 1	$C_1 = \{\dots\dots\dots\dots\dots\dots\}$ $C_2 = \{\dots\dots\dots\dots\dots\dots\}$	$M_1 =$ $M_2 =$
Iteratie 2	$C_1 = \{\dots\dots\dots\dots\dots\dots\}$ $C_2 = \{\dots\dots\dots\dots\dots\dots\}$	$M_1 =$ $M_2 =$

5.

Een reisagentschap wil zijn klanten opdelen in twee clusters ( $k = 2$ ), op basis van leeftijd en duur van geboekte vakanties. De informatie van het reisagentschap is gegeven in de onderstaande tabel. De leeftijd van klanten uitgedrukt in aantal jaren, de duur van de vakantie in het aantal dagen.

ID	Leeftijd	Duur vakantie
$TO_1$	19	3
$TO_2$	25	8
$TO_3$	43	14
$TO_4$	61	14
$TO_5$	30	7
$TO_6$	22	10

Cluster de objecten door middel van het  $k$ -means algoritme

Kies  $TO_1$  als centrum voor cluster  $C_1$  en Kies  $TO_5$  als centrum voor cluster  $C_2$ .

Noteer voor elke iteratie welk cluster gevormd wordt en wat de centra's zijn. Vul deze resultaten in de onderstaande tabel in:

Iteratie	Clusters	Centrums
Start		$M_1 = (19,3)$ $M_2 = (30,7)$
Iteratie 1	$C_1 = \{\dots\dots\dots\dots\dots\dots\}$ $C_2 = \{\dots\dots\dots\dots\dots\dots\}$	$M_1 =$ $M_2 =$
Iteratie 2	$C_1 = \{\dots\dots\dots\dots\dots\dots\}$ $C_2 = \{\dots\dots\dots\dots\dots\dots\}$	$M_1 =$ $M_2 =$

6.

Hieronder staan gegevens van 4 (fictieve) studenten van een data mining cursus. We noteren respectievelijk het aantal bijgewoonde lessen, het aantal dagen examenvoorbereiding, en of ze al dan niet voor het examen kwamen opdagen:

$$D = \{S_1(0.0, 0, 1), S_2(7.8, 2, 1), S_3(4.8, 2, 1), S_4(3.0, 0, 1)\}$$

We willen deze objecten clusteren, wat we kunnen doen door middel van de  $k$ -means methode.

Kies  $k = 2$ , en als initiële willekeurige cluster centra's:  $M_1 = (0.6, 1, 0)$  en  $M_2 = (7.2, 1, 0)$ .

Pas de  $k$ -means methode toe op  $D$  tot een maximum van 3 stappen. Noteer voor elke iteratie welke clusters gevormd worden en wat de centra's zijn.

Convergeert de methode? Zo ja, leg uit.

Vul de resultaten in de onderstaande tabel in:

	Clusters	Centrums
Start		$M_1 = (0,6,1,0)$ $M_2 = (7, 2, 1, 0)$
Iteratie 1	$C_1 = \{ \dots \}$ $C_2 = \{ \dots \}$	$M_1 =$ $M_2 =$
Iteratie 2	$C_1 = \{ \dots \}$ $C_2 = \{ \dots \}$	$M_1 =$ $M_2 =$
Iteratie 3	$C_1 = \{ \dots \}$ $C_2 = \{ \dots \}$	$M_1 =$ $M_2 =$
.....	.....	.....