



# Data2Diamonds® - BIG Data TestNet – Summer School

July 9, 2014



# Agenda

- Opening
- BIG Data demonstrator
- Introduction BIG Data, Tim van Soest & Henk van Haaster
- Vitens case, Jan-Willem Lankhaar
- Privacy & BIG Data, Caroline Massart
- Brainstorm Workshop Preparation, Bram Bronneberg
- Break
- Brainstorm Workshop



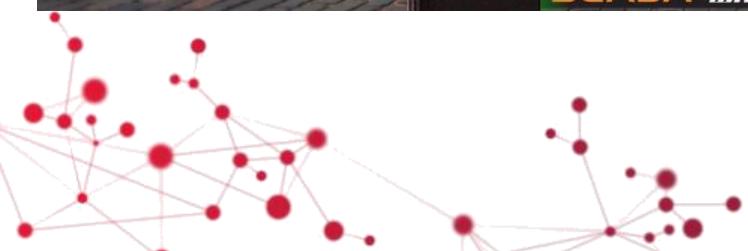


# Data2Diamonds® - BIG Data Inspire TestNet – Summer School

Tim van Soest & Henk van Haaster  
July 9, 2014

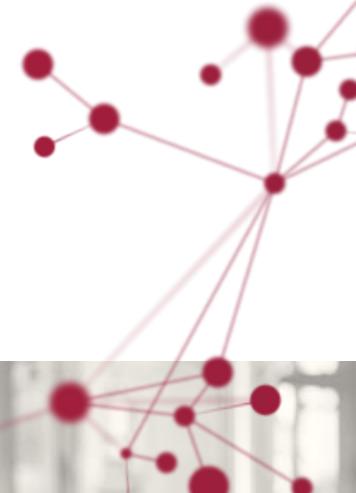


# BIG Data Demonstrator



**CGI**

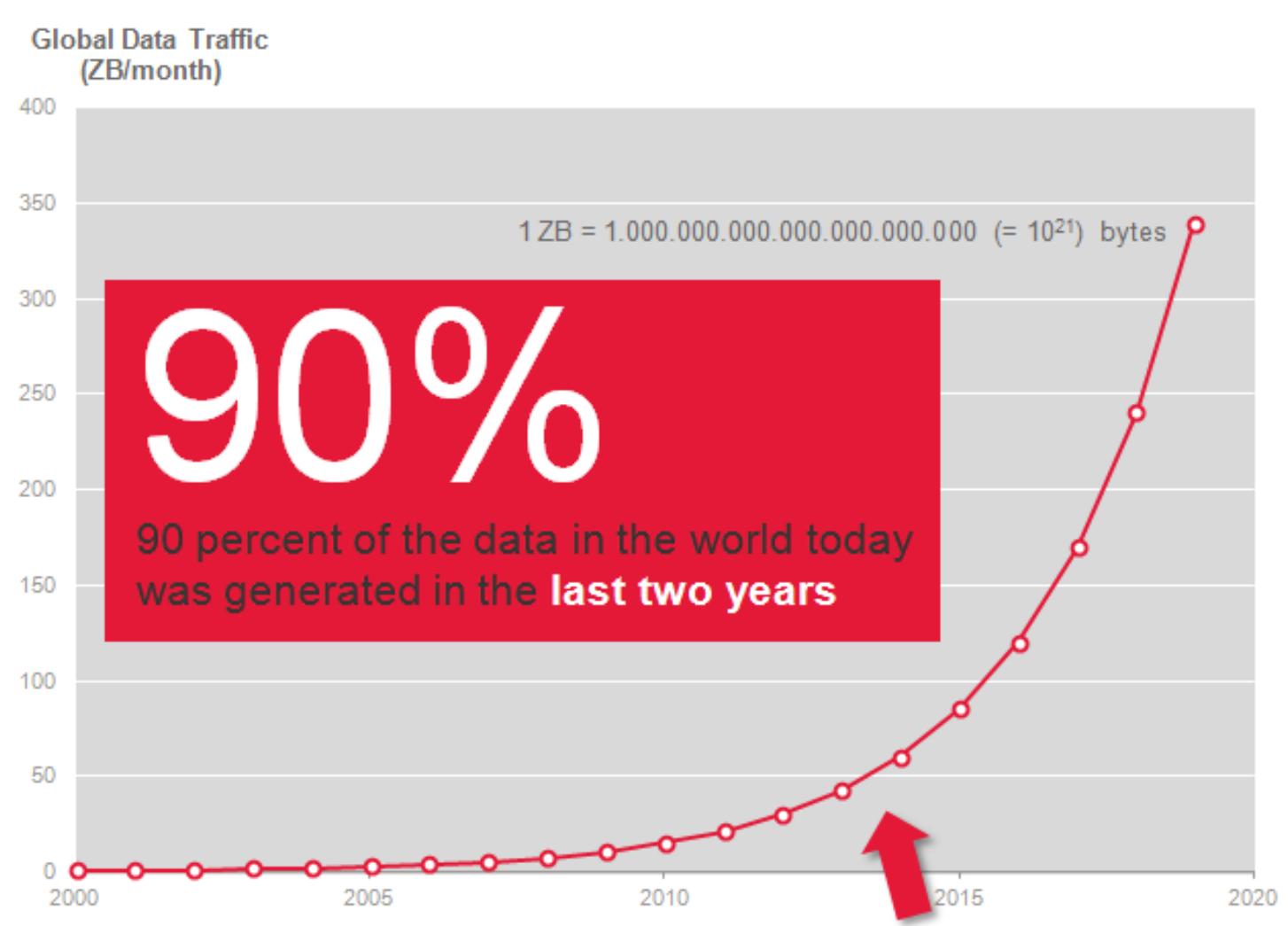
# Introduction BIG Data



**CGI**

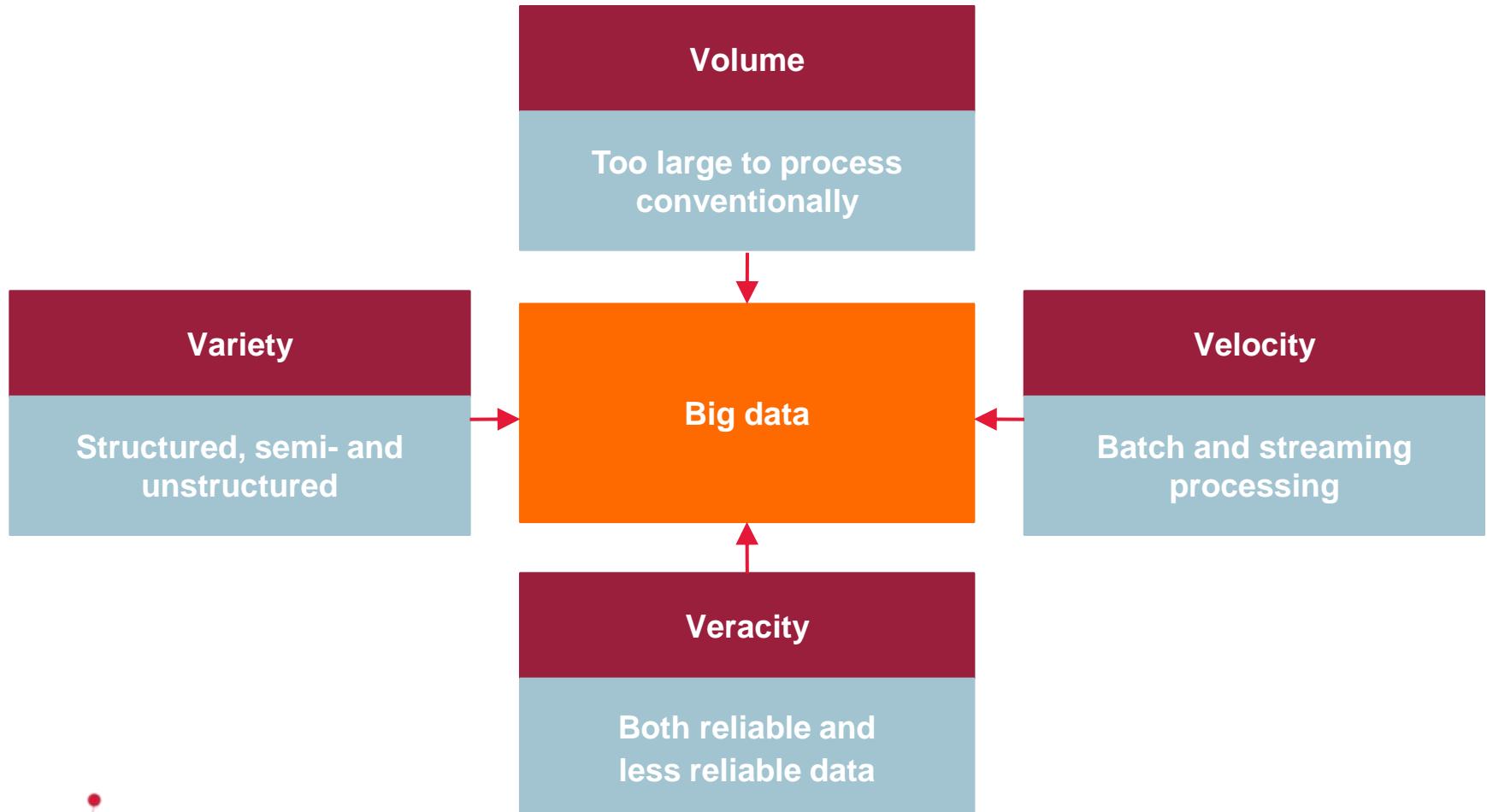
Experience the commitment®

# BIG Data: Data explosion!





# How to define BIG Data?



# Big data! Big deal?



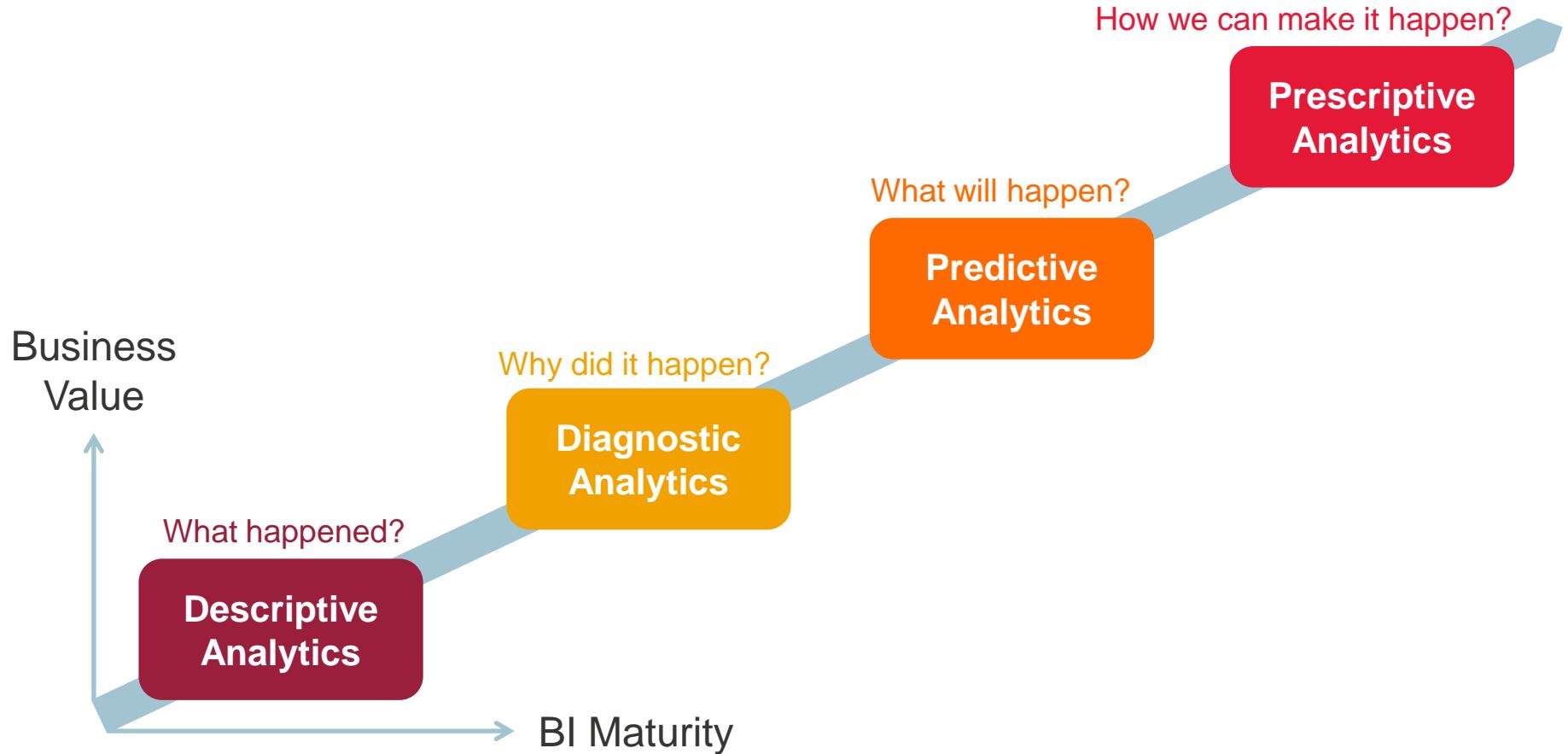
**CGI**

# Additional or new business models: *Shining diamonds*

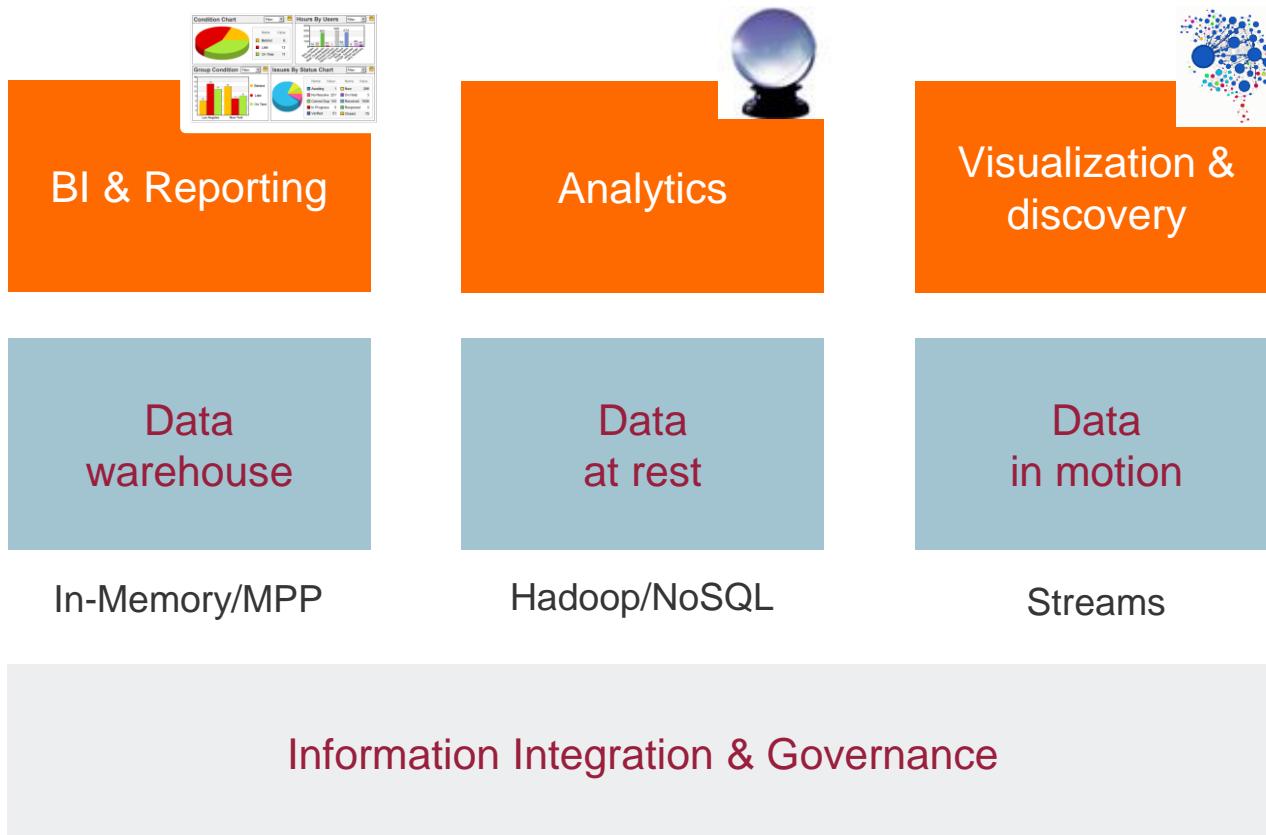


**CGI**

# The evolution of BI and Analytics



# BIG Data solutions and traditional BI



# Important Big Data applications



# Data2Diamonds® - BIG Data-programme

- BIG Data in four steps
- Focused on business value

- Inspire
- Deep dive
- Proof of Value
- Implementation



**CGI**



# Data2Diamonds® - BIG Data @ Vitens TestNet – Summer School

Jan-Willem Lankhaar  
July 9, 2014



# Data2Diamonds® - Big Data Proof of Value Klantcase



Model ontwikkelen voor lokaliseren van lekken in waterleidingnetwerk

- Interne en externe (historische) data in lab
- Data geanalyseerd (40 storingen)

- Heatmaps voor lokalisatie lekken
- 50% van lekken < 2,5 km van hotspot

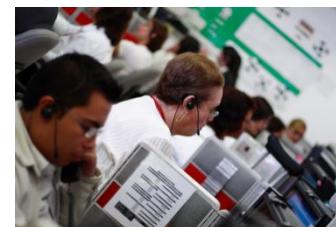


# Lekken in het waterleidingnetwerk

## Lekken

- ... bedreigen leveringszekerheid
- ... leiden tot hoge callcenterbelasting
- ... kunnen tot ernstige gevolgschade leiden
- ... veroorzaken verlies aan waterkwaliteit

- Detecteren van lekken is lastig
- Localiseren van lekken is bewerkelijk



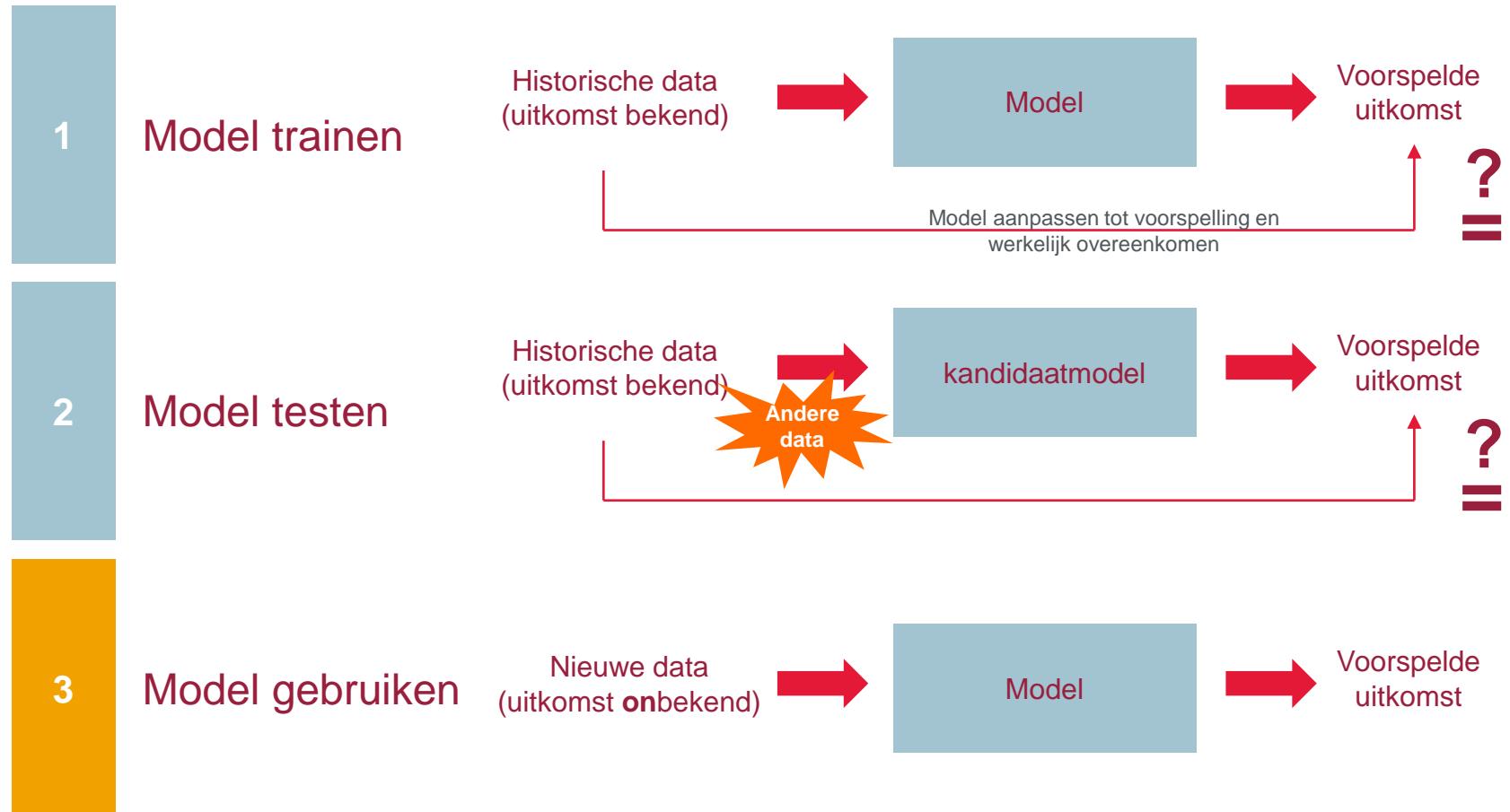
# Data2Diamonds® - Big Data Proof of Value

## De uitdaging

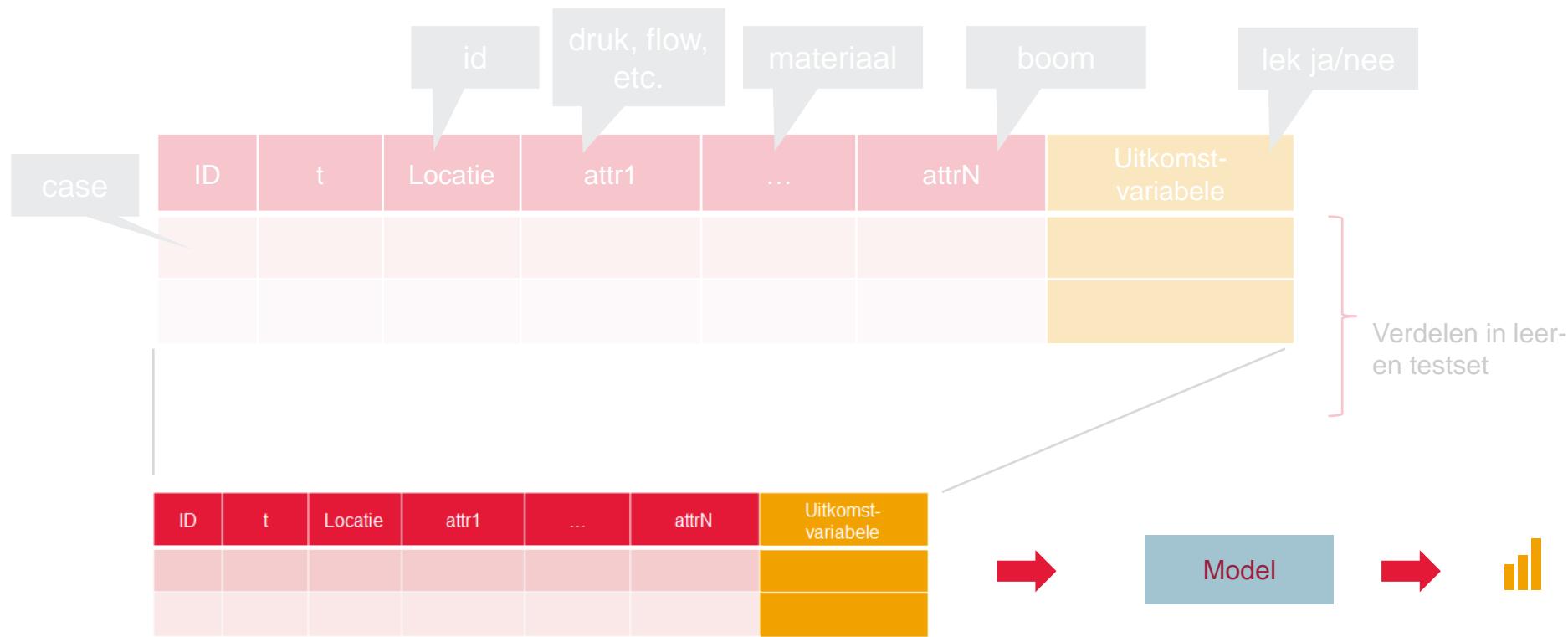
Vind een innovatieve manier om lekken te lokaliseren op basis van interne en externe data



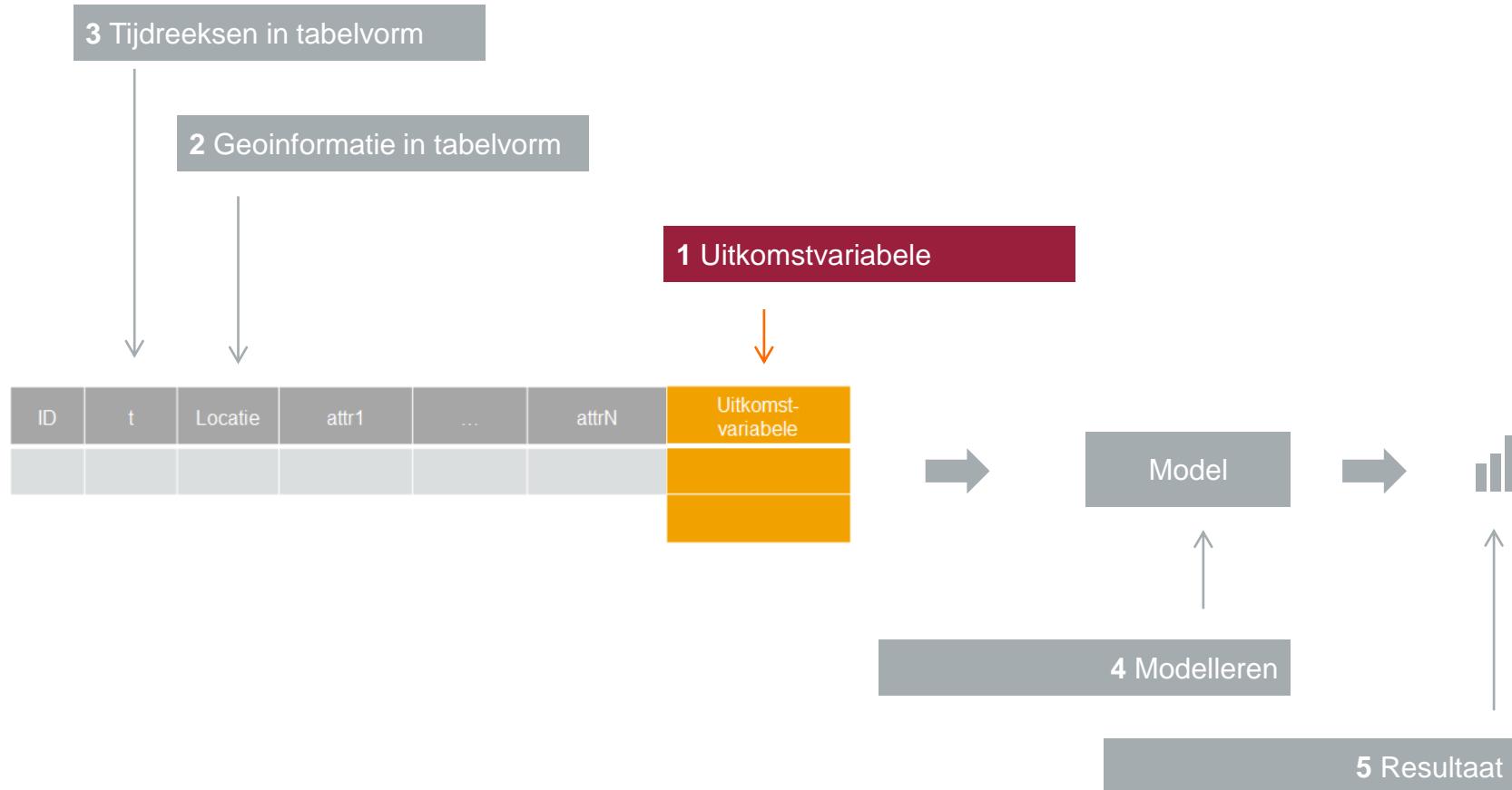
# Detectie/lokalisatie = *machine learning*



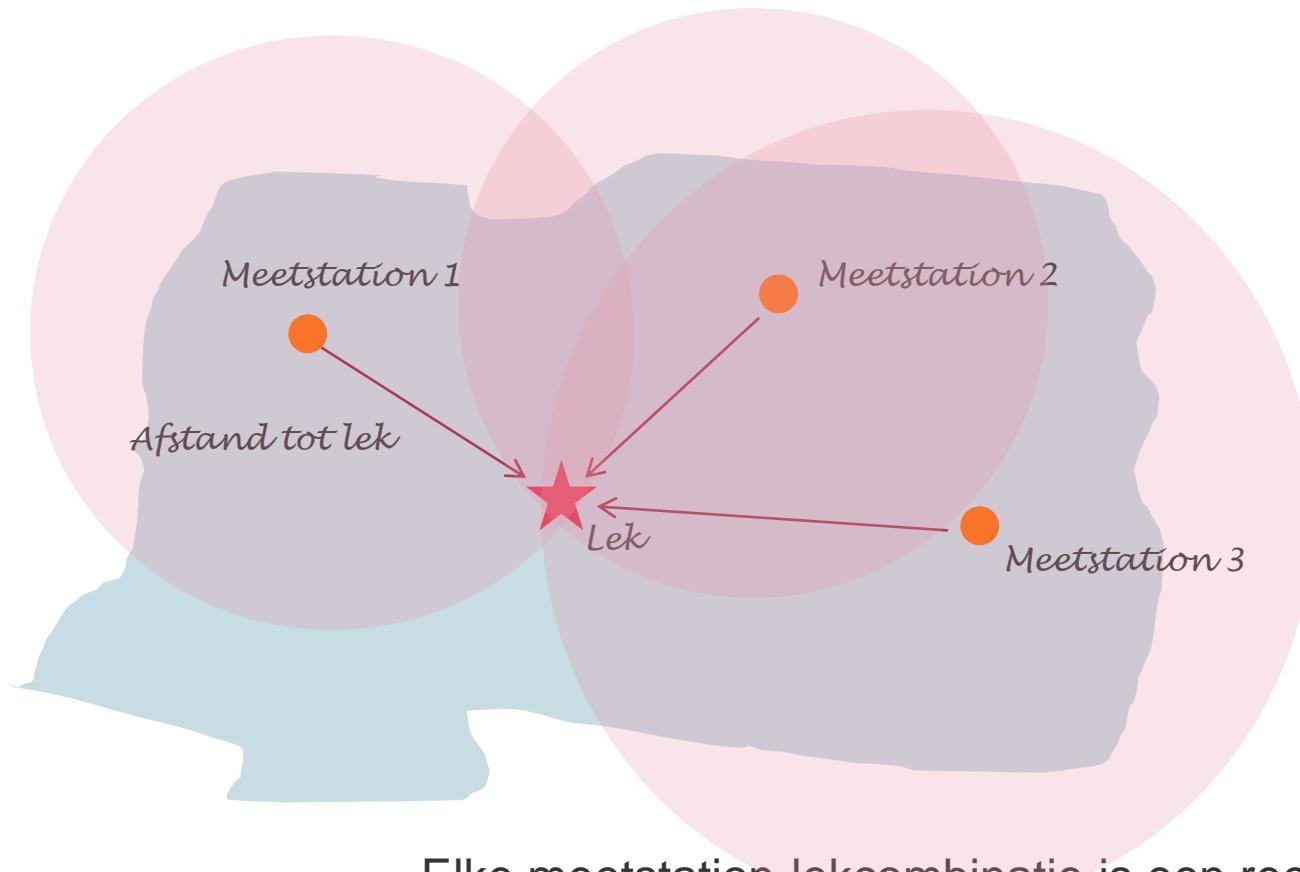
# Aanpak machine learning



# Stap 1: Uitkomstvariabele analyseren



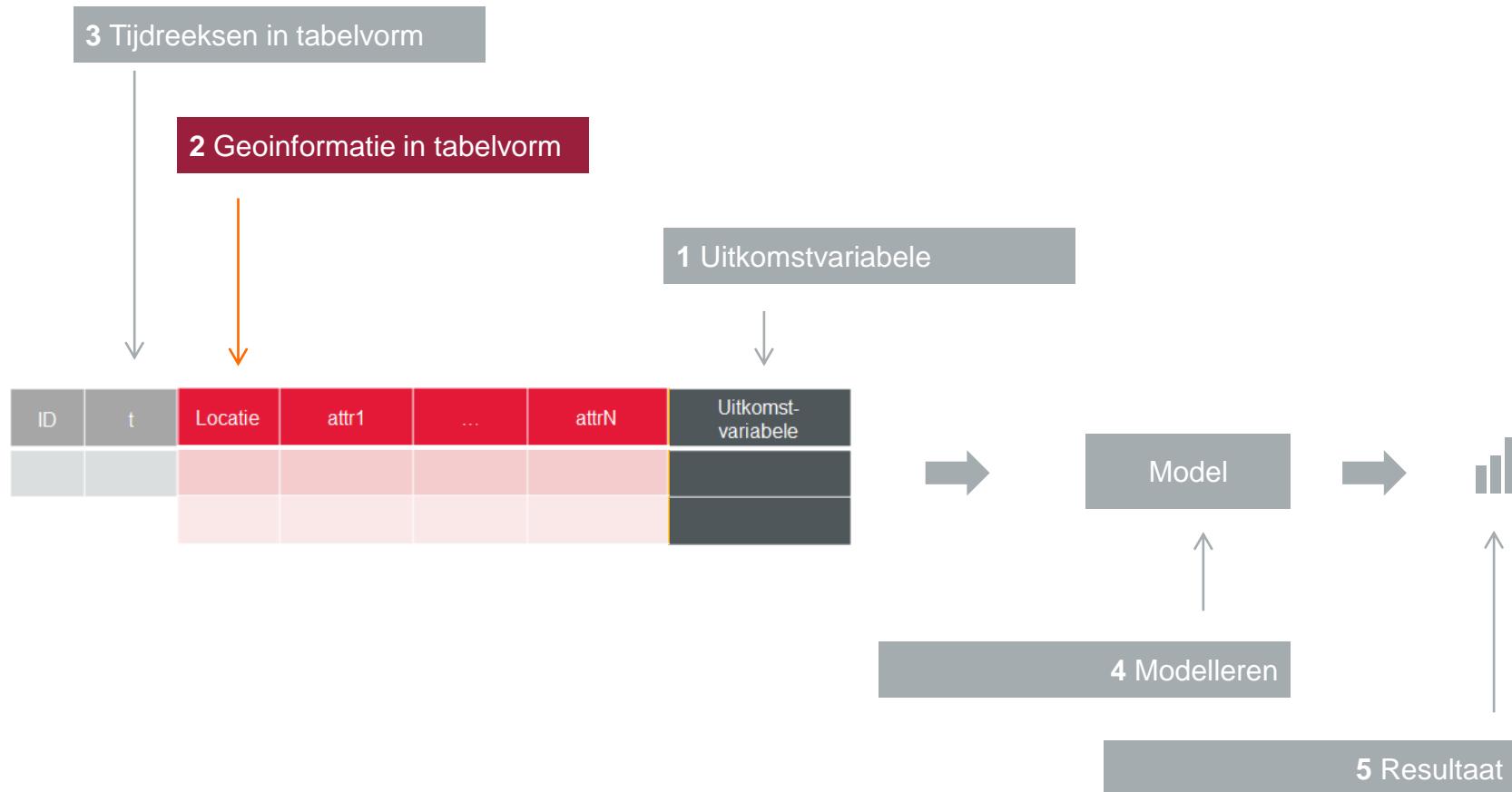
# Afstand tot lek als uitkomstvariable

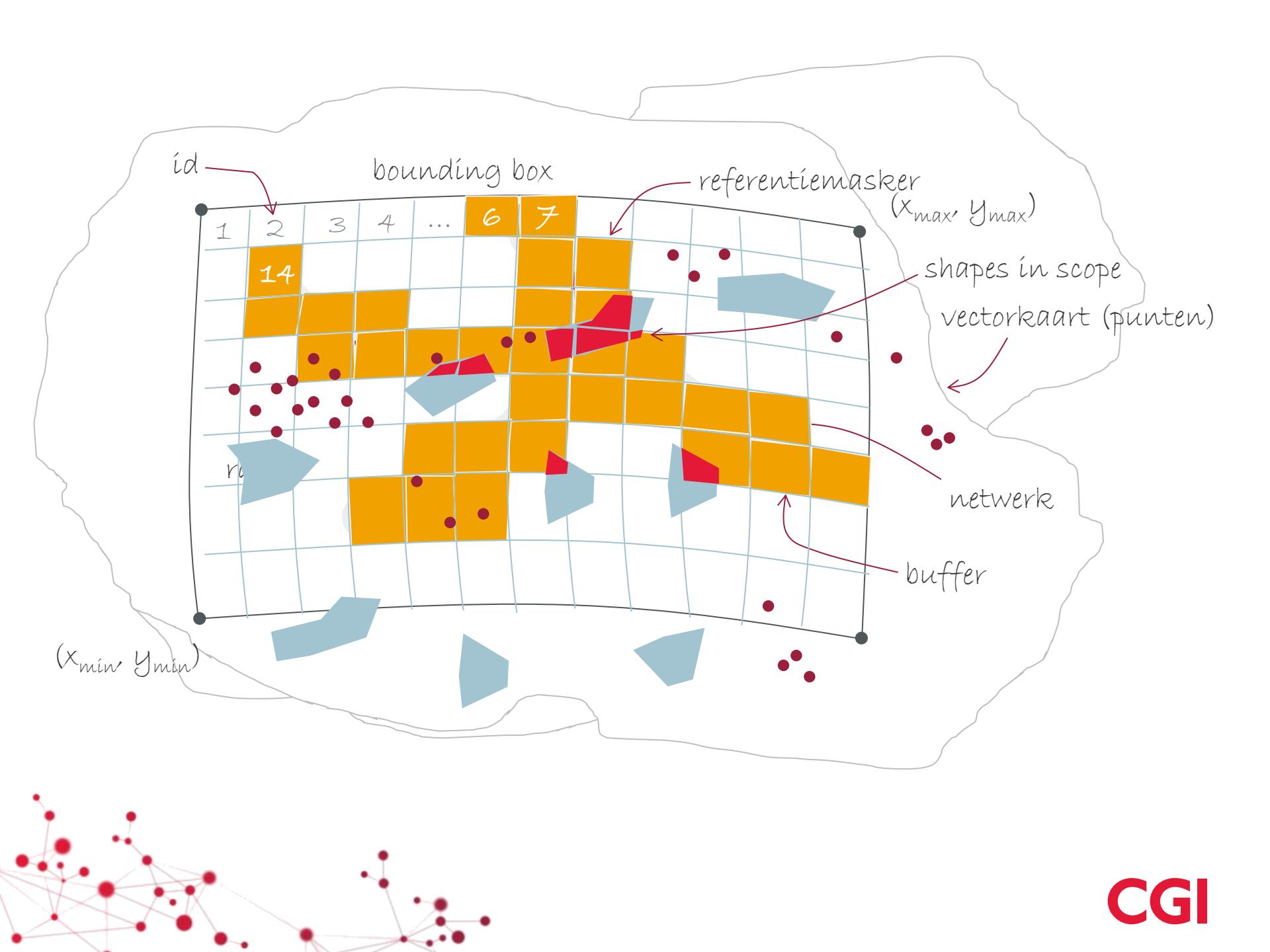


Elke meetstation-lekcombinatie is een record  
Aantal meetpunten groter

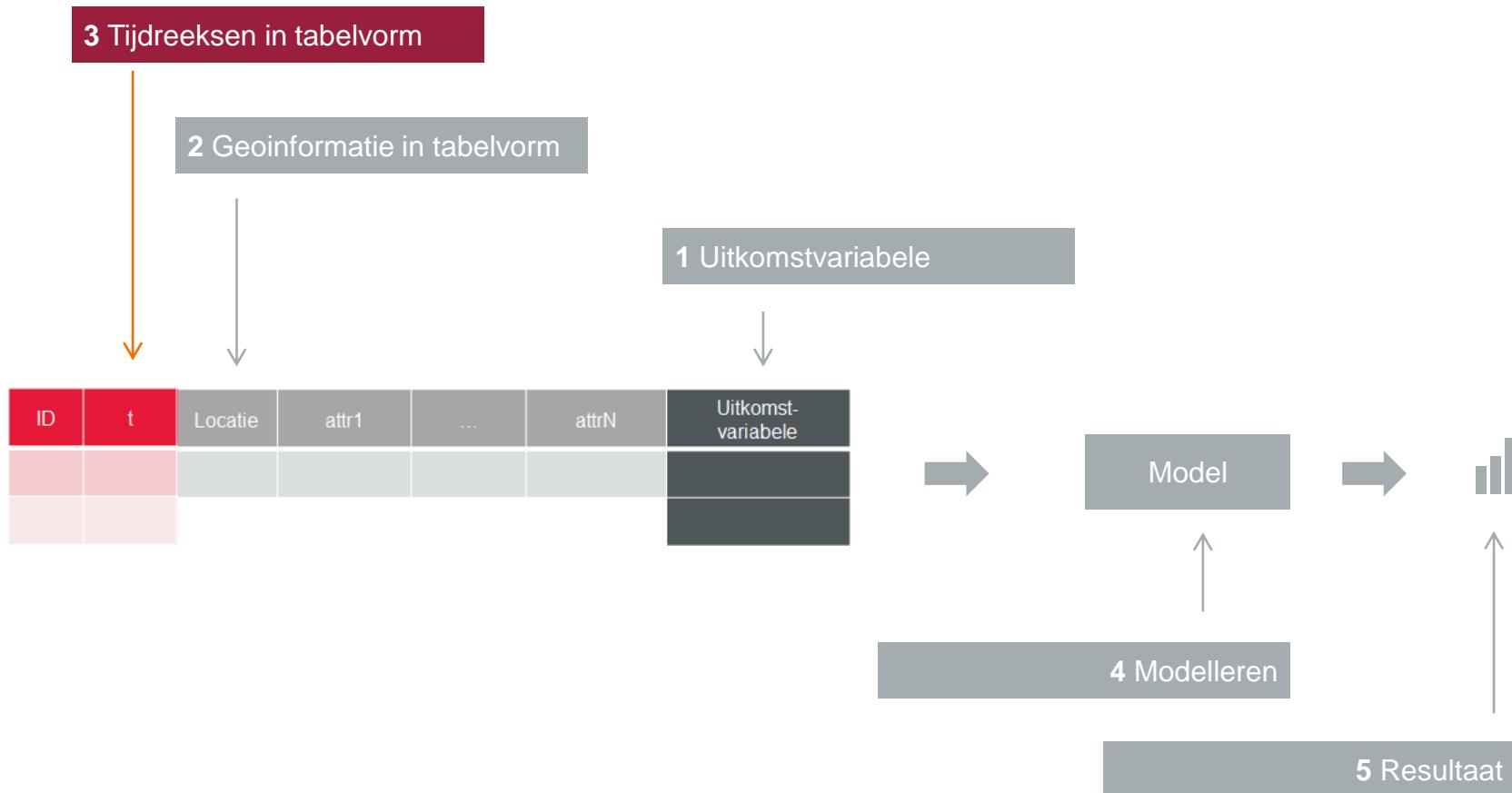


# Stap 2: geoinformatie in tabelvorm





# Stap 3: tijdreeksanalyse



# Tijdreeksanalyse voor het ontdekken van patronen... ...maar wat is een patroon?

The screenshot shows a news article from Volkskrant.nl. The headline reads "Stokoude records sneuvelen bij Brazilië-Duitsland". The text discusses how Germany's victory over Brazil was a repeat of their 1920 defeat to Uruguay. Below the text is a photograph of three German players celebrating. A red box highlights the sentence: "Zo evenaarde Brazilië zijn grootste nederlaag ooit. Tot dinsdag leek de 6-0 verliesbeurt tegen Uruguay in de Copa America van 1920 'onaantastbaar'." To the right of the main article is a sidebar with a small image of a field and some text.

weggegeven. Ik heb een heel goed gevoel. Iedereen zal alles geven, de elf spelers die in de basis staan en de zeven reservespelers."

► Nu

Robben kan ervoor zorgen dat de reeks' wordt voortgezet. Sinds 2002 geen Nederlandse Nederlander in de finale van de WK.

Guardiola

Bayern München-trainer Josep Guardiola

**Volkskrant.nl**

HOME NIEUWS POLITIEK OPINIE BUITENLAND SPORT  
BINNENLAND CULTUUR ECONOMIE REIZEN WETENSCHAP & KUNST

DOSSIER HET WEER

Zonnigste maart ooit: levendige voorjaarslucht

31/03/14, 11:14 - bron: Weerplaza

Stokoude records sneuvelen bij Brazilië-Duitsland

Gepubliceerd: 9 juli 2014 07:08 | Laatst gewijzigd: 9 juli 2014 07:13

Voetbalstatistici maakten dinsdagavond overuren tijdens de historische WK-wedstrijd tussen Brazilië en Duitsland.

Verschillende stokoude records sneuvelden of werden geëvenaard tijdens de halve finale die de 'Mannschaft' met zeldzaam machtsvertoon en met klinkende cijfers (7-1) won.

Zo evenaarde Brazilië zijn grootste nederlaag ooit. Tot dinsdag leek de 6-0 verliesbeurt tegen Uruguay in de Copa America van 1920 'onaantastbaar'.

De laatste keer dat Brazilië op eigen bodem verloor was in 2002. Paraguay triomfeerde toen in een vriendschappelijke ontmoeting.

Duitsland vernedert Brazilië op WK

1 / 11

ÖZIL 8

rij!

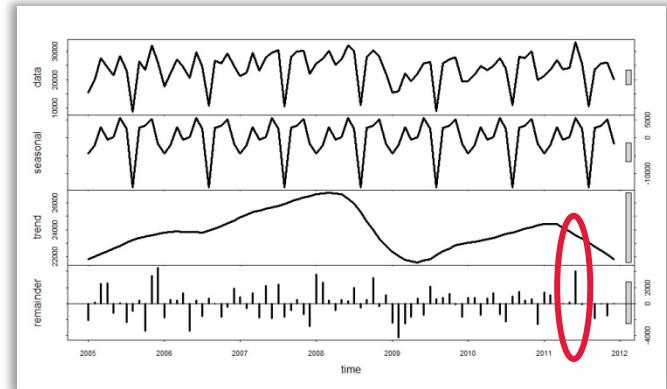
# Tijdreeksanalyse: waarom en hoe?

## Doel tijdreeksanalyse

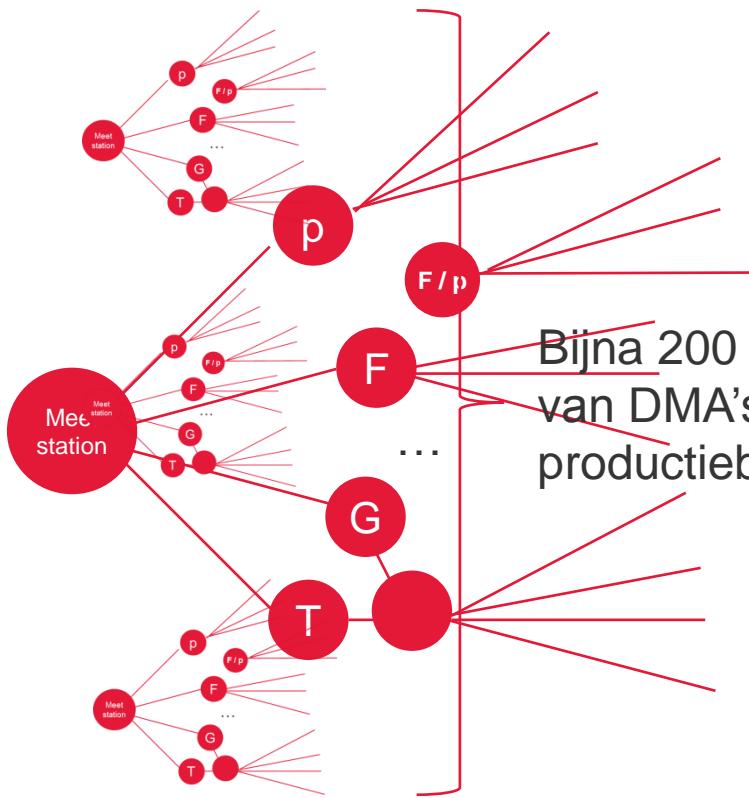
- Uitzonderingen opsporen
- Uitzonderingen versterken

## Technieken

- Middelen, standaarddeviatie (beschrijvend)
- Trendanalyse
- Signalen combineren (domeinkennis)
- Spectraalanalyse
- Modelleren
- Filteren
- Waveletanalyse
- ....

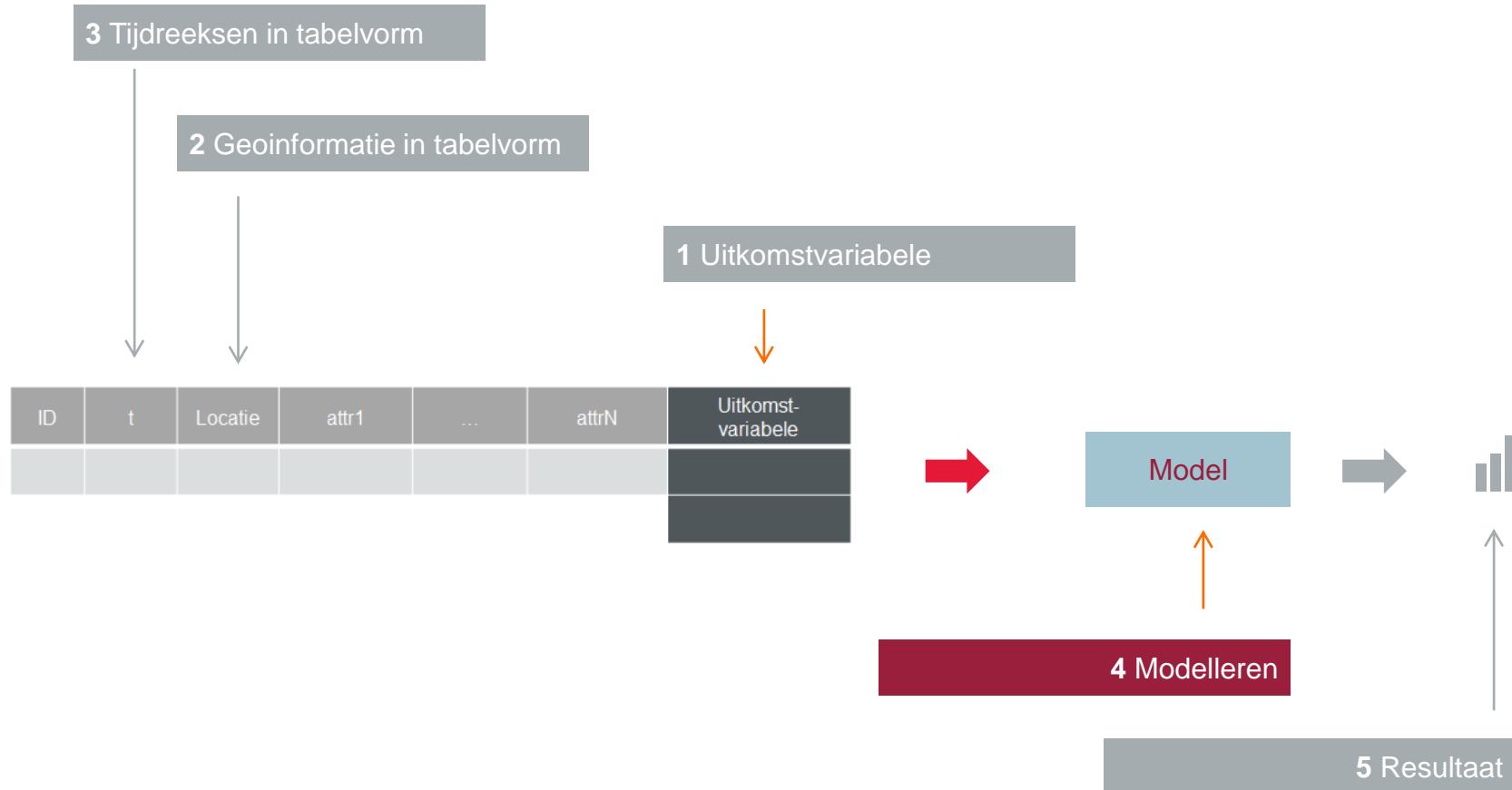


# Afgeleide en afgeleiden van combinaties



Bijna 200 originele en afgeleide signalen afkomstig  
van DMA's, opjager, distributiereservoirs,  
productiebedrijf, deelbalansgebiedvolumestroom

# Stap 4: modelleren



# Aanpak afstand tot lek

- Voorspel afstand van lek tot meetstation op basis van inputsignalen
- Iedere meetstation-lekcombinatie is een datapunt
- Construeer cirkels om tags

alle mogelijke voorspellers



Feature selection

goede voorspellers



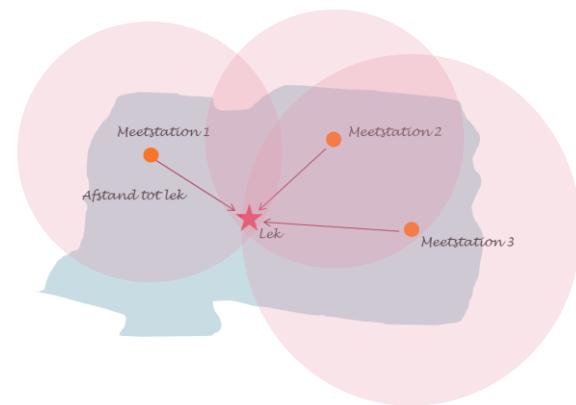
Regressiemodel

Schatting afstand(lek, tag)

Cirkels om tags construeren

Heatmap

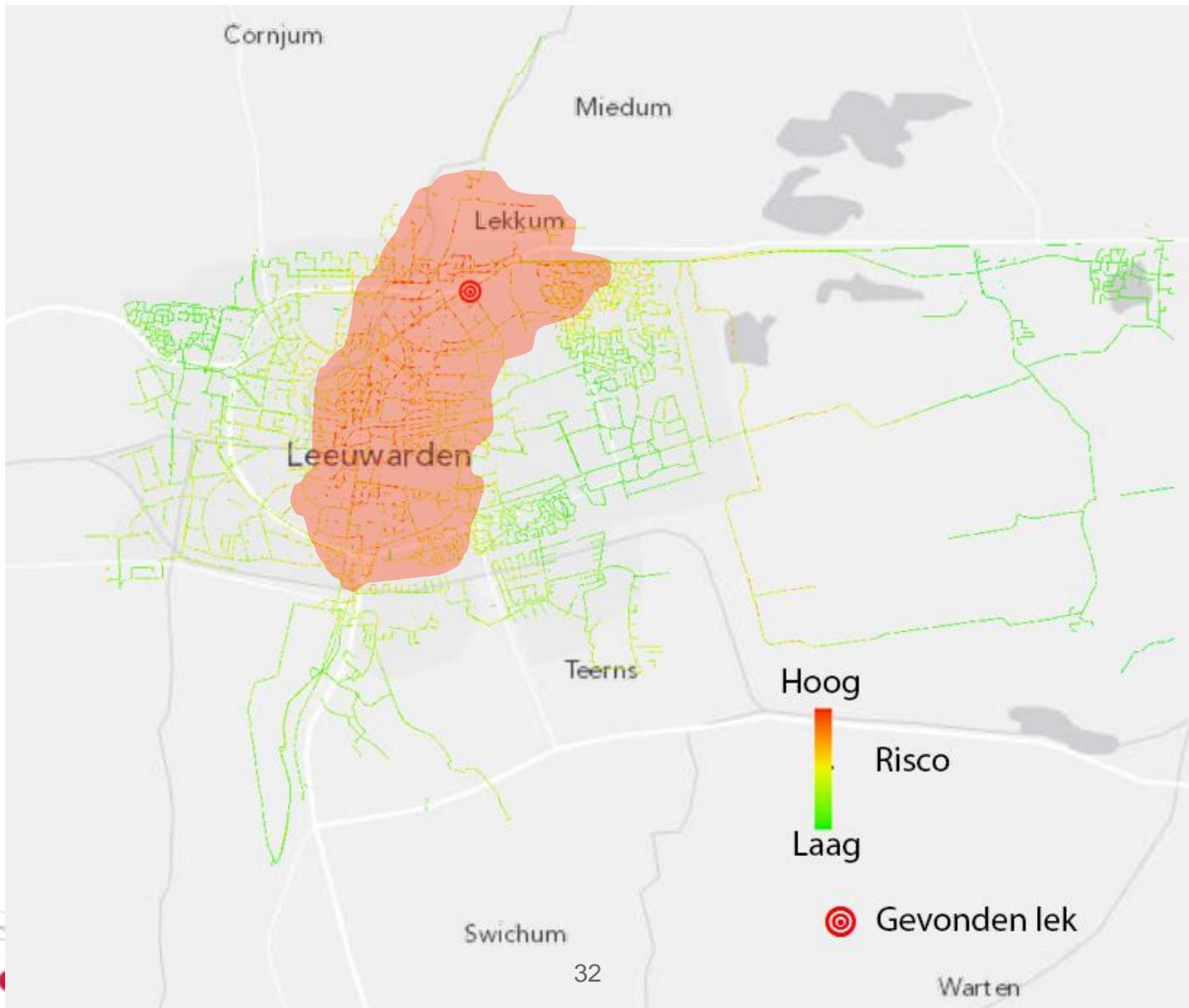
Hotspots duiden op lek



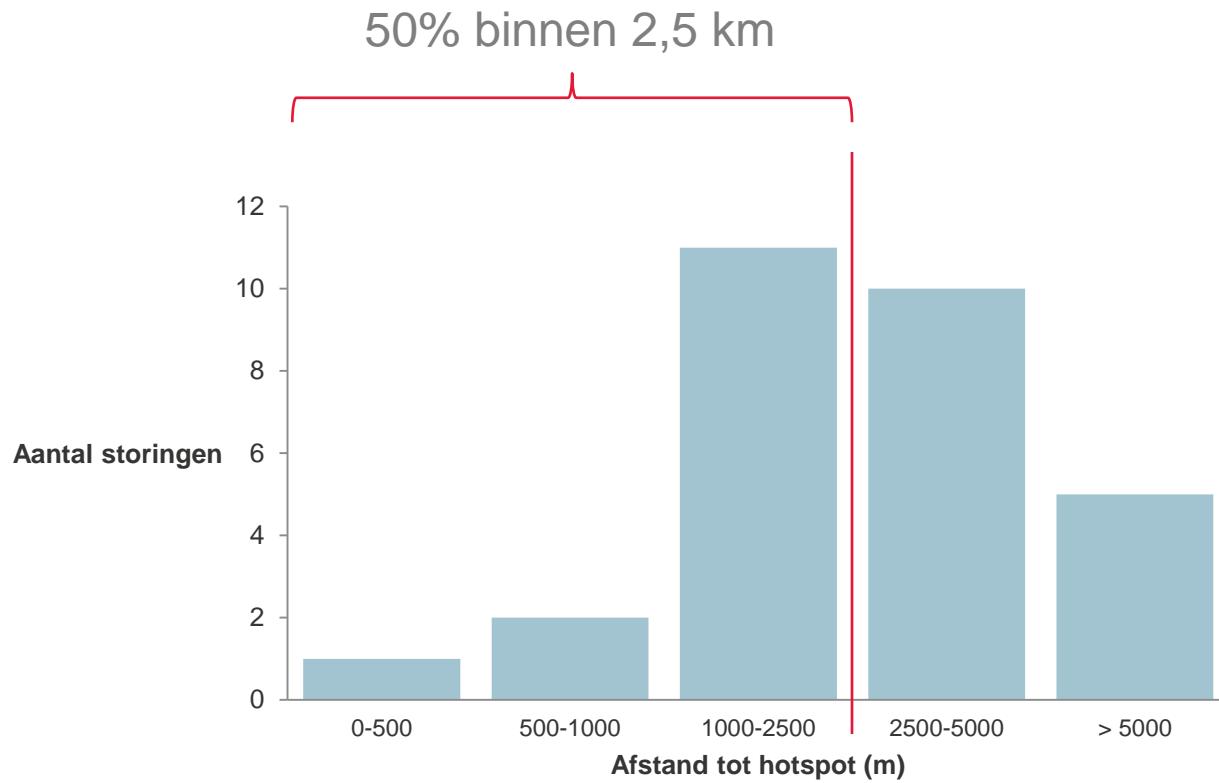
# Stap 5: Resultaat



# Leklokalisatie



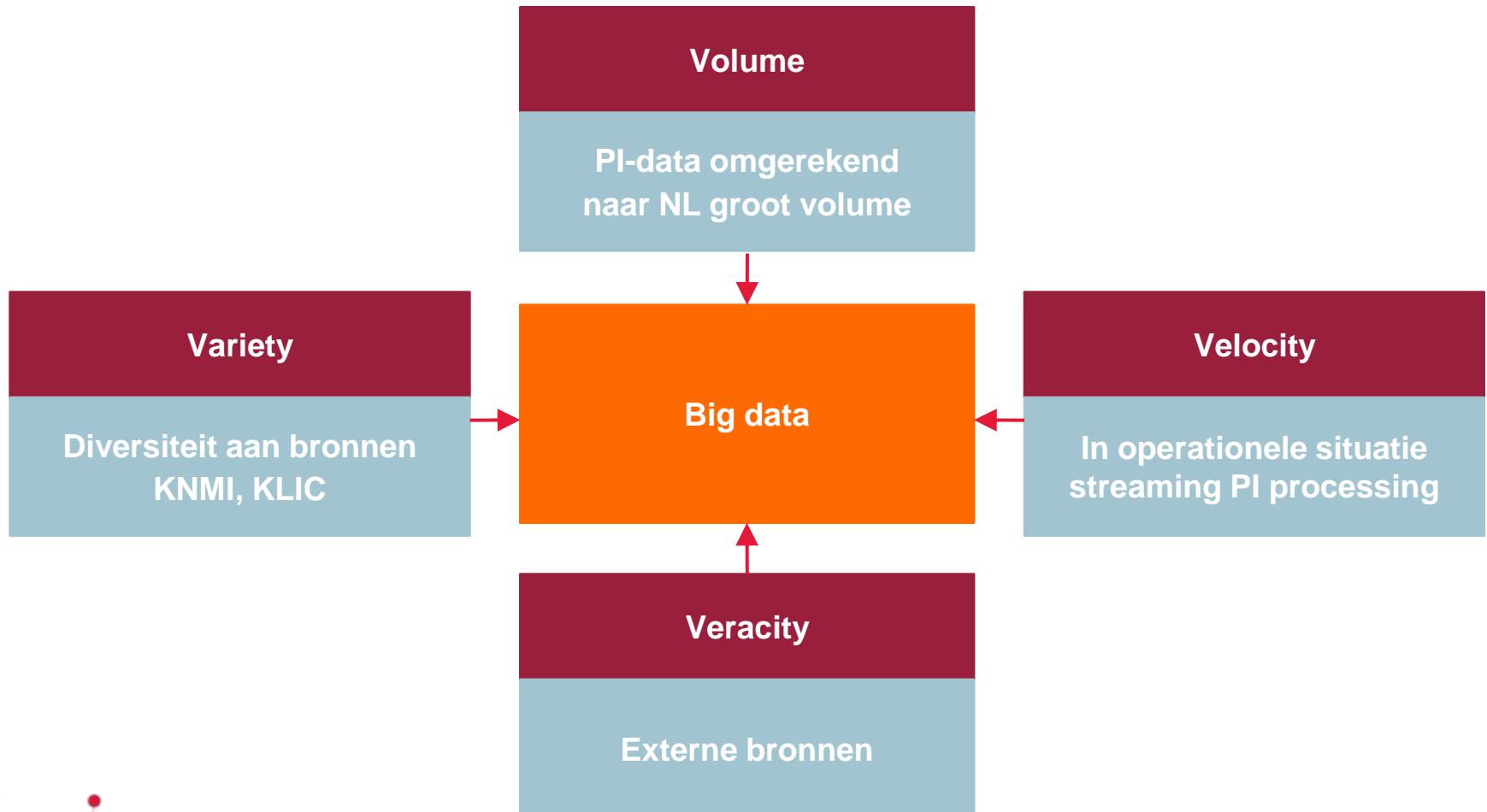
# Resultaat



< 2,5 km



# Proof of Value: 'little' big data



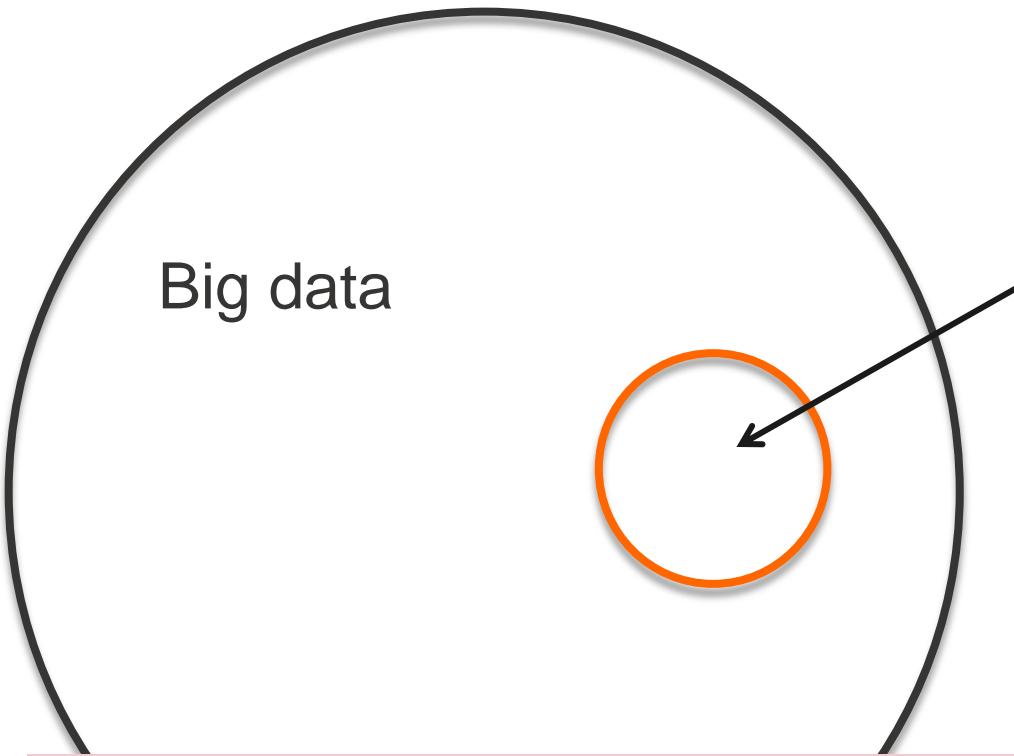


# Data2Diamonds® - Privacy & BIG Data TestNet – Summer School

Caroline Massart  
July 9, 2014



# Personal data is common in big data projects



## Personal data

*“any information relating to an identified or identifiable natural person”*

(European Commission, 1995)

*“Personal data is the new oil of the internet and the new currency of the digital world”*

(Meglena Kuneva, European Consumer Commissioner 2009)

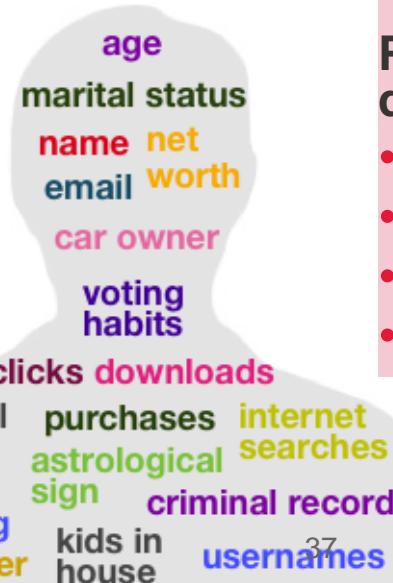


# Personal data is often sensitive

Degree of sensitivity is *contextual and personal*

The use of **sensitive personal data of customers** poses magnified risks to organizations:

- Compliance
- Data breaches
- Brand reputation



## Perceived high sensitive personal data:

- balance of savings account
- balance of checking account
- medical history
- e-mail content
- location tracking

## Perceived low sensitive personal data:

- Marital status
- Full name
- E-mail address
- Educational level

# Customer information privacy

- What is information privacy?

Definition: “*one’s ability to control information about oneself*”

Practical: in the commercial/customer sense it involves:

- 1. Protection and careful use of the personal information of customers**
- 2. Meeting the expectations of customers about the use of their personal information**



# Three important stakeholders involved in the privacy landscape

Legislator



Organisation



Customer



# Legislation lags behind

- Technology moves faster than the law
- Minimal enforcement by CBP (College Bescherming Persoonsgegevens)
- First universal point of reference: the OECD Privacy Principles, developed in 1980. Updated recently to fit online data context.
- Will lead in 2016 to EU Data Directive.

## OECD 8 Privacy Principles

1. Collection Limitation
2. Data Quality
3. Purpose Specification
4. Use Limitation
5. Security Safeguards
6. Openness
7. Individual participation
8. Accountability



# Customer: privacy is hot topic!

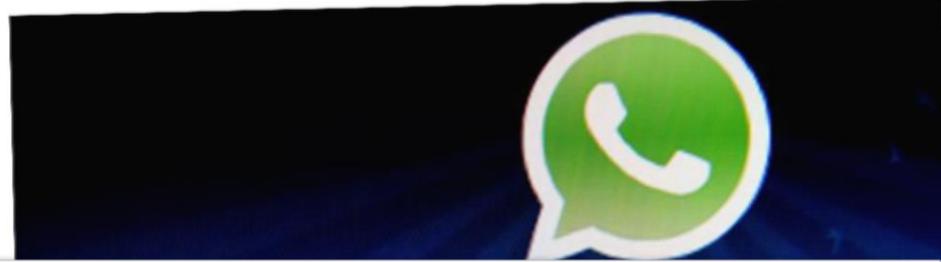
**INTERNET**  
D66 wil aparte eurocommissaris voor privacy  
Gepubliceerd: 12 mei 2014 15:48  
Aangepast: 12 mei 2014 15:48

f t g+

'Google biedt bedrijven snel internet in ruil voor gegevens'  
een groter plan om meer mensen web-diensten van Google te laten gebruiken op een zwaar gesubsidieerd WiFi-netwerk

## Whatsapp reageert op privacyzorgen

dinsdag 18 mrt 2014, 03:43 (Update: 18-03-14, 06:45)



## Google Trends

News headlines  Forecast ?

May 2014 (partial data)

■ privacy big data: 100

## Privacygevoelige proef ING

maandag 10 mrt 2014, 07:36 (Update: 10-03-14, 17:09)

Equens ziet voorlopig af van doorverkopen  
pingegevens klanten  
Nrc.nl, 24 mei 2013

# Organizations need a position on privacy

- For many organizations that depend upon personal data, privacy has become a strategic factor.
- Position should reflect the strategy of the organization and be in line with its business objectives.
- The Information Commissioner's Office (UK) has identified three broad approaches of privacy protection, ranging from very narrow to the very broad.

Approaches of privacy protection

Minimalist

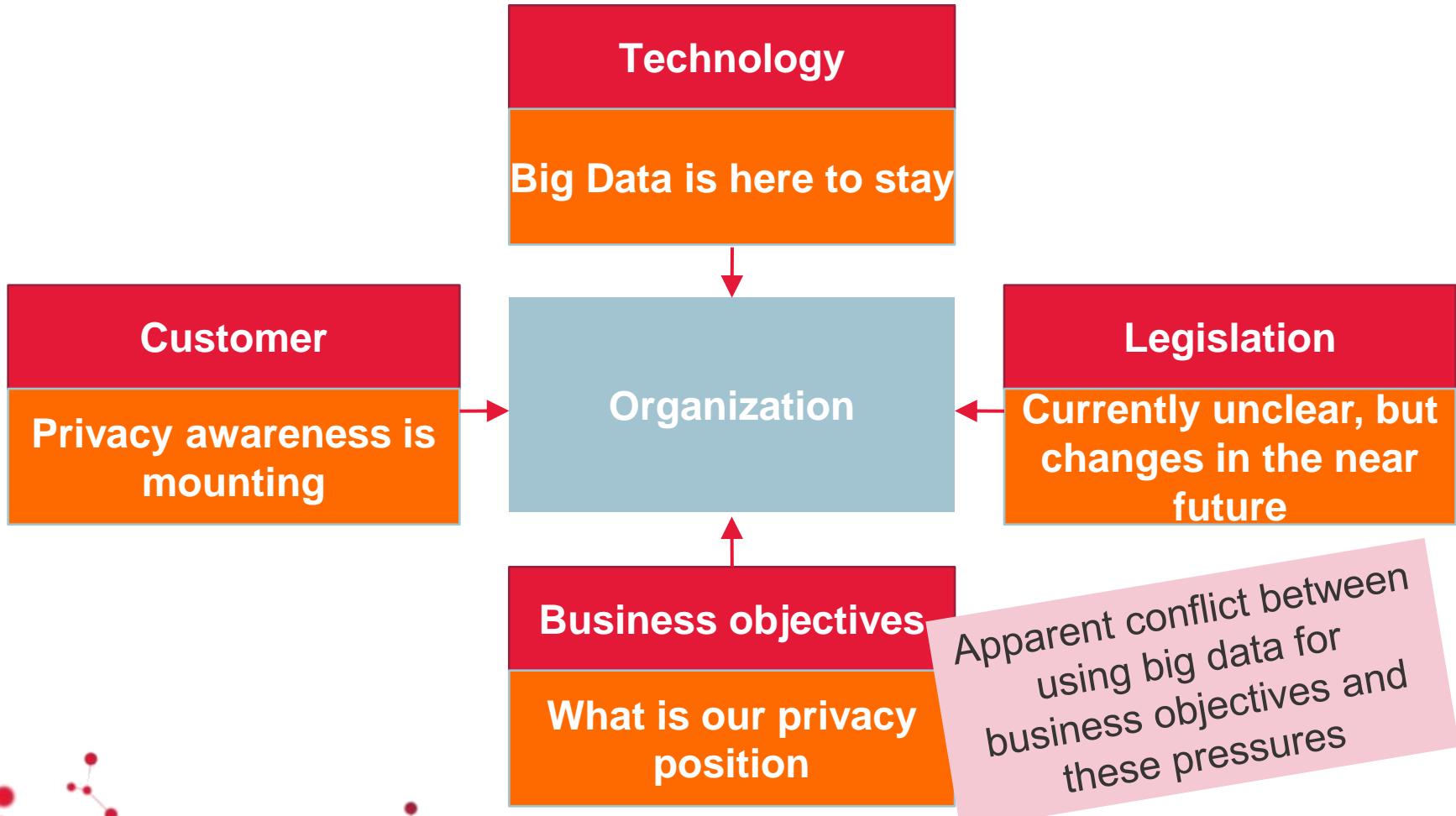
Comprehensive

Social Impacts

- But what is the position of your organization?

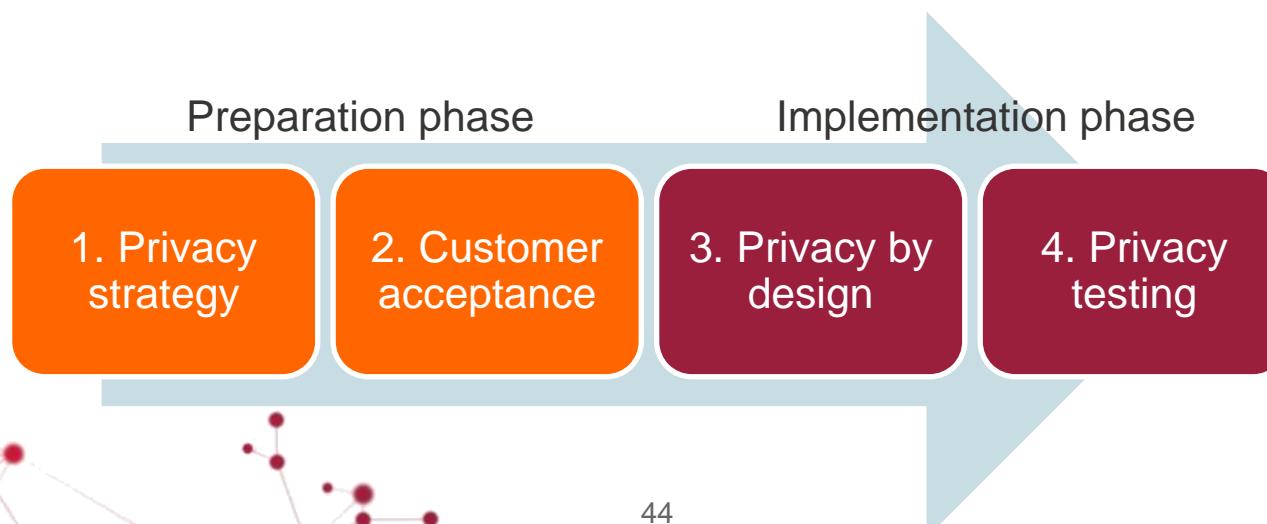
# Problem: privacy landscape under pressure

***Customer privacy is a business problem that must be addressed!***



# Approach

1. Organizations should develop a clear *privacy strategy*
2. Assess '*privacy strategy*' for customer acceptance
3. Privacy can be implemented early on in systems by 'privacy by design' principle
4. Testing whether privacy is implemented in the design correctly



# 1. Determine your ‘privacy strategy’

OECD Privacy Principles	Approaches of privacy protection		
	Minimalist	Comprehensive	Social Impacts
1. Collection Limitation	The organization collects personal data that is irrelevant to the purpose of collecting, or may be inaccurate, incomplete and not-up-to date.		
2. Data Quality			
3. Purpose Specification			
4. Use Limitation			
5. Security Safeguards			
6. Openness		There is a policy of openness to customers	An individual has full insight into the data obtained about him/her and can adjust the data
7. Individual participation			
8. Accountability			

## 2. Assess customer acceptance

### Example: ongoing study

Survey:

Each respondent was described a scenario of a fictitious bank ( privacy strategy + sensitivity of data) and questions

Closing date:  
1 june 2014

520  
respondents

202  
women

318  
men

average age

37  
CGI

#### RETHINKING PRIVACY IN THE ERA OF BIG DATA

The impact of corporate data privacy  
strategies on consumer behavior in the  
financial services industry

CAROLINE MASSART  
15 JUNE 2014

RSM  
*Erasmus*  
ERASMUS  
UNIVERSITY

CGI

## 2. Assess customer acceptance

### Example: ongoing study

Results of survey:

- A “social impacts” privacy strategy results in lower privacy concerns in comparison to broader privacy appetites.
- The results are irrespective of high or low sensitivity of personal data.



In other words; with the right approach, privacy is not a constraint!

### 3. Privacy by Design

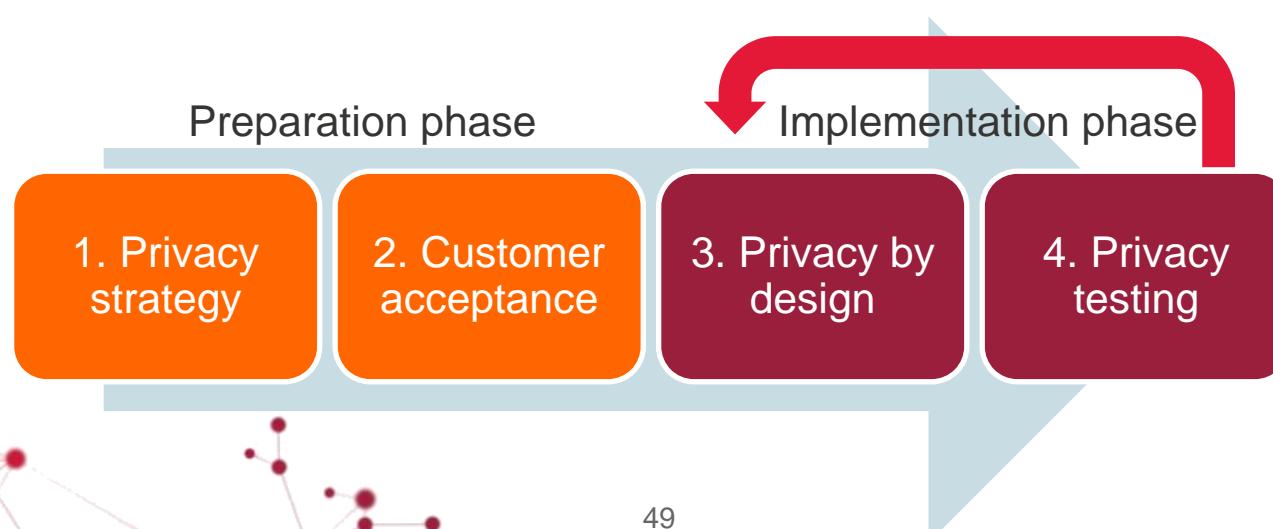
- Implement privacy early on in the design of the system

Privacy Enhancing Technologies			
<i>Protecting</i>	Onion routing	Anonymous credentials	
	Cryptography	Blind signatures	
	Privacy Guaranteeing Execution Container	Pseudonyms	
	Mix networks	Private information retrieval	
	Secret Sharing		Attribute certificates
	Concealment or encryption schemes		
	Steganography	Zero-knowledge proofs	
<i>Enabling</i>	Information expiration date		
	Support for legal protection: sticky policies, log files & watermarking		
	P3P Privacy-Enhancing Intelligent Software Agents		
<i>Transparency</i>	Credibility rating systems		
	Reputation systems		
	Audit logs Transparency tools		

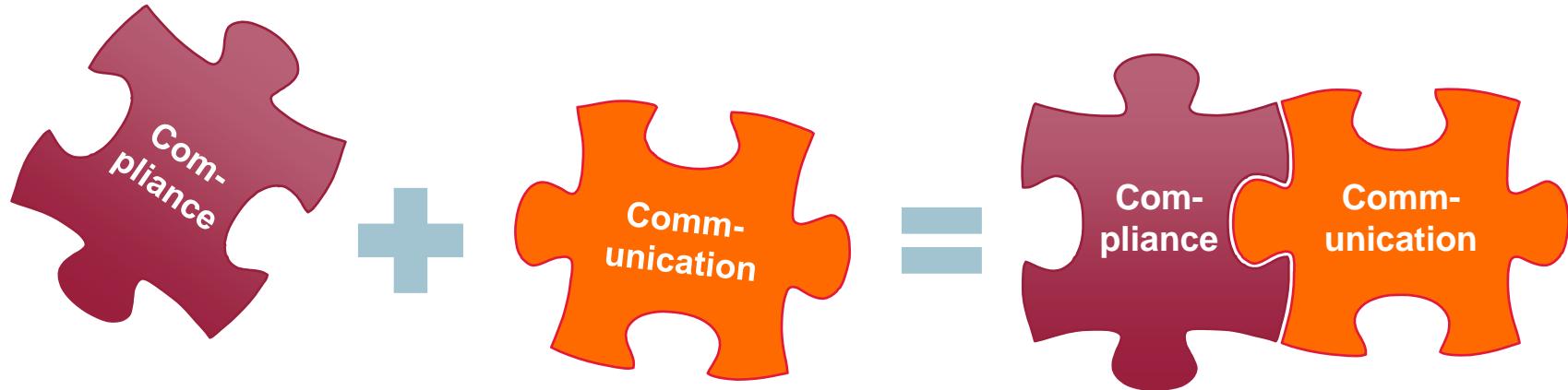
# 4. Testing privacy

***Technological tools are available, but issues arise...***

- Is the collection limitation principle applicable?
- How to interpret ‘vague’ legislation correctly?
- Tradeoff between using production data or fictitious data
- Anonymised data can be re-identified with a particular individual through matching with other data



Think of privacy as a means to building trust  
rather than a matter of compliance



Privacy

Trust





# Data2Diamonds® - Brainstorm Workshop

## TestNet – Summer School

Bram Bronneberg  
July 9, 2014



# Brainstorm workshop

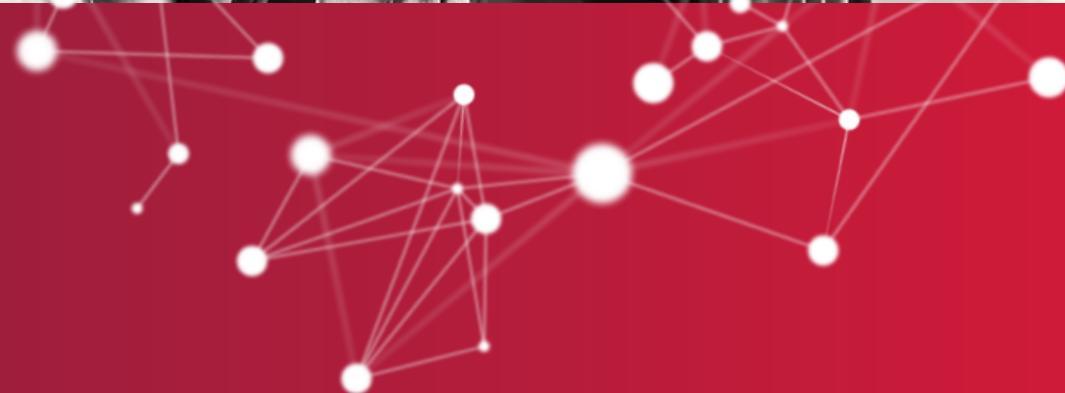
Hoe zou jij **BIG Data** oplossingen testen?

## Werkvorm:

- Vorm 2 groepen
- 3 ronden bestaande uit (3 x 25 min)
  - Centrale uitleg (5 min)
  - Brainstormen in groepjes (15 min)
  - Onderling presenteren (5 min)
- Centrale samenvatting



# Break



**CGI**

Experience the commitment®

# Brainstorm workshop

Hoe zou jij **BIG Data** oplossingen testen?

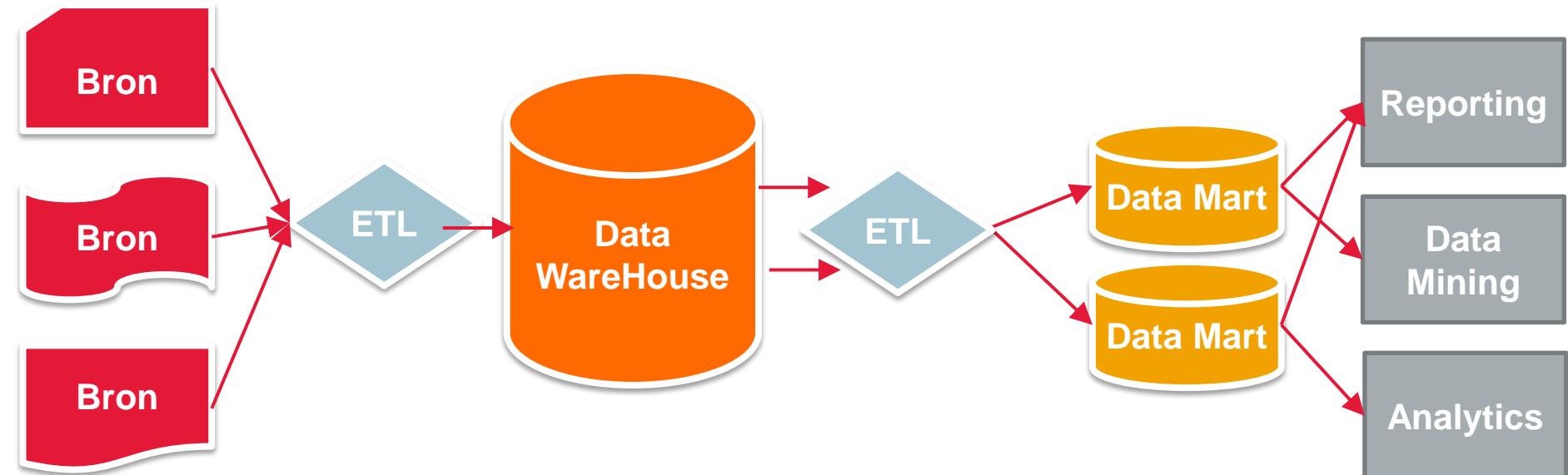
## Werkvorm:

- Vorm 2 groepen
- 3 ronden bestaande uit (3 x 25 min)
  - Centrale uitleg (5 min)
  - Brainstormen in groepjes (15 min)
  - Onderling presenteren (5 min)
- Centrale samenvatting



# 1<sup>e</sup> ronde – Klein beginnen... BI Testen

Hoe zou jij BI testen aanpakken?



# 2<sup>e</sup> ronde - 2 Test V's en 4 BIG Data V's

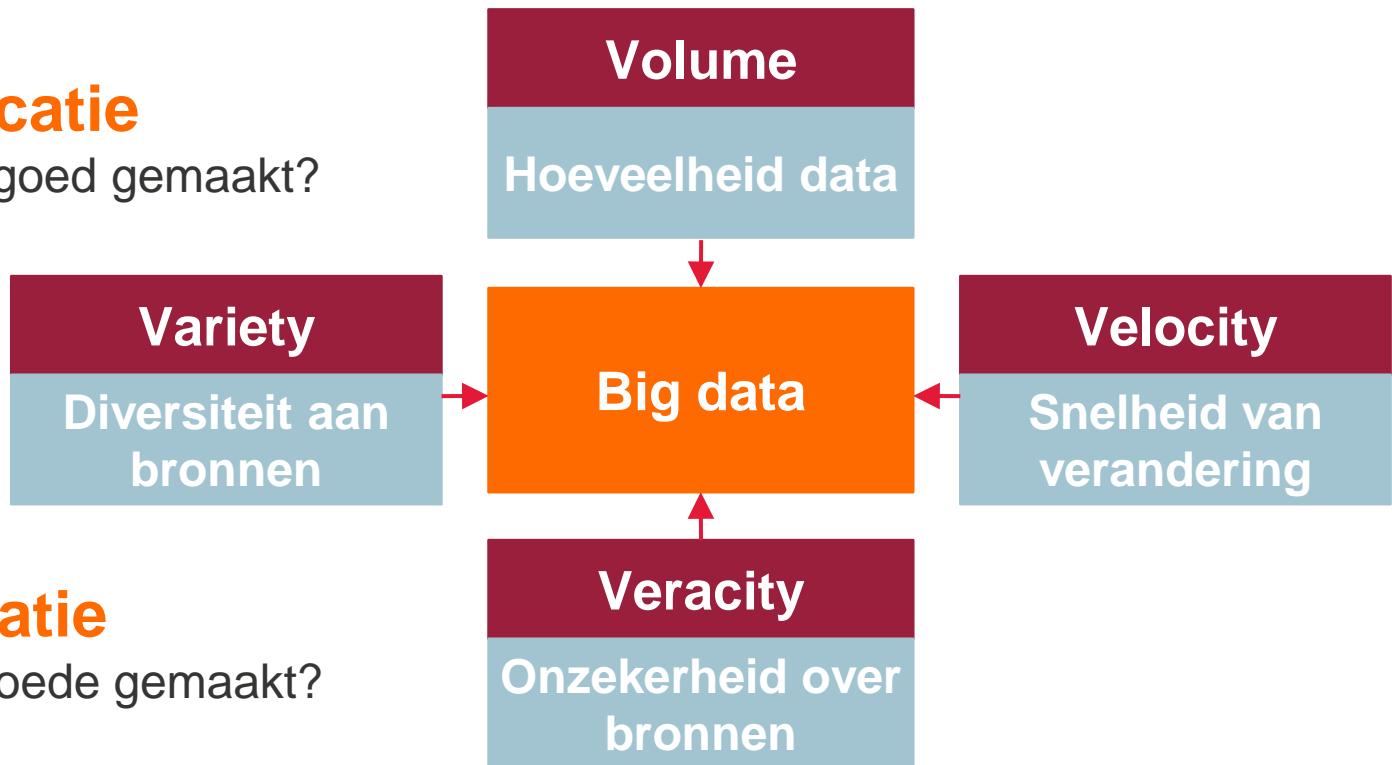
Wat voor testdoelen / kwaliteitstoetsen kun je bedenken voor de 4 verschillende Big Data V's. Denk daarbij aan verificatie maar ook aan validatie.

## Verificatie

Hebben we het goed gemaakt?

## Validatie

Hebben we het goede gemaakt?



# 3<sup>e</sup> ronde – BIG Data = small privacy?

BIG Data heeft grote implicaties m.b.t. privacy.

*Combineren van geanonimiseerde bronnen bijvoorbeeld, kan de anonimisering teniet doen!*

Hoe test je privacy?

Waar hou je rekening mee in je proces?

## OECD 8 Privacy Principles

1. Collection Limitation

2. Data Quality

3. Purpose Specification

4. Use Limitation

5. Security Safeguards

6. Openness

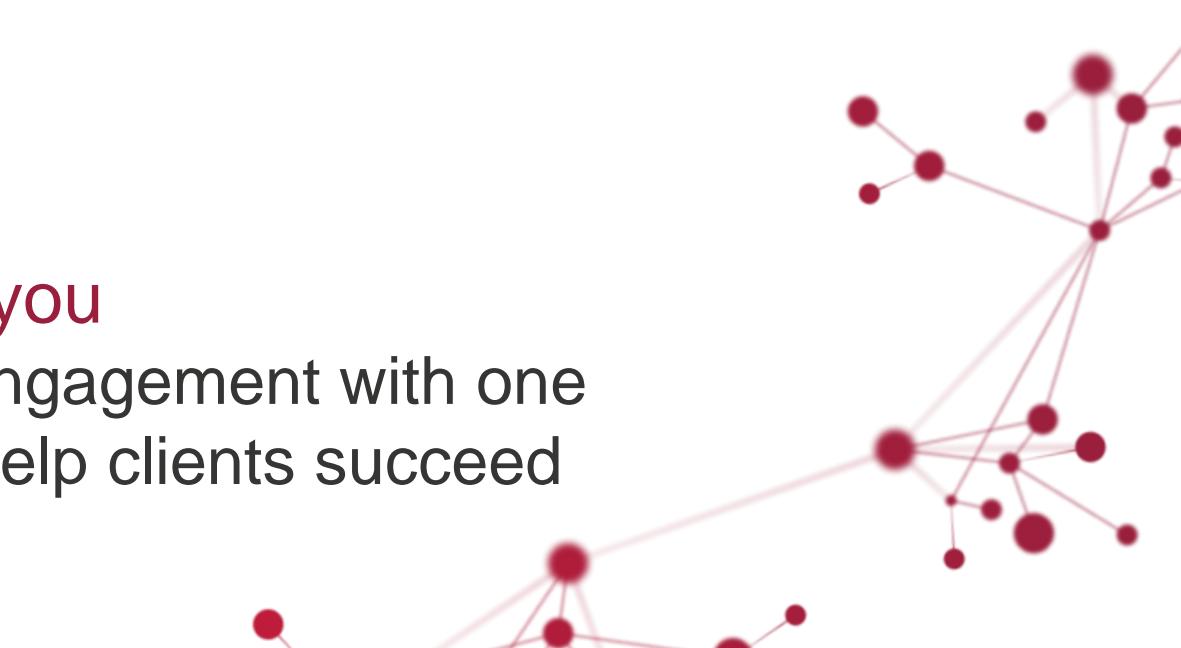
7. Individual participation

8. Accountability



## Our commitment to you

We approach every engagement with one objective in mind: to help clients succeed



**CGI**

Experience the commitment®