Data2Diamonds® - BIG Data

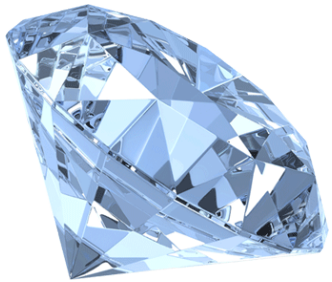
TestNet – Summer School

July 9, 2014

Experience the commitment®

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



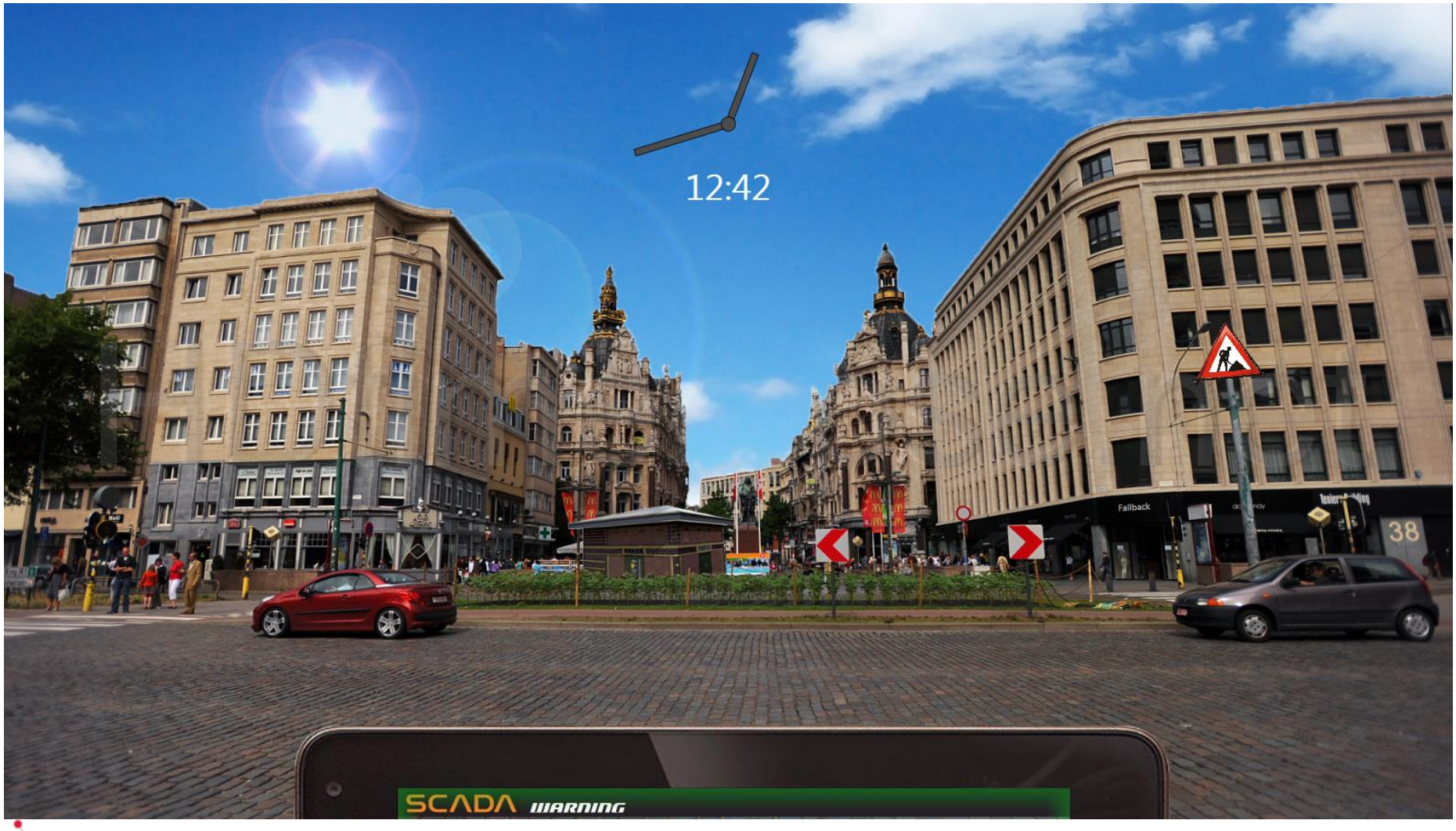
Data2Diamonds® - BIG Data Inspire

TestNet – Summer School

Tim van Soest & Henk van Haaster
July 9, 2014

Experience the commitment®

BIG Data Demonstrator

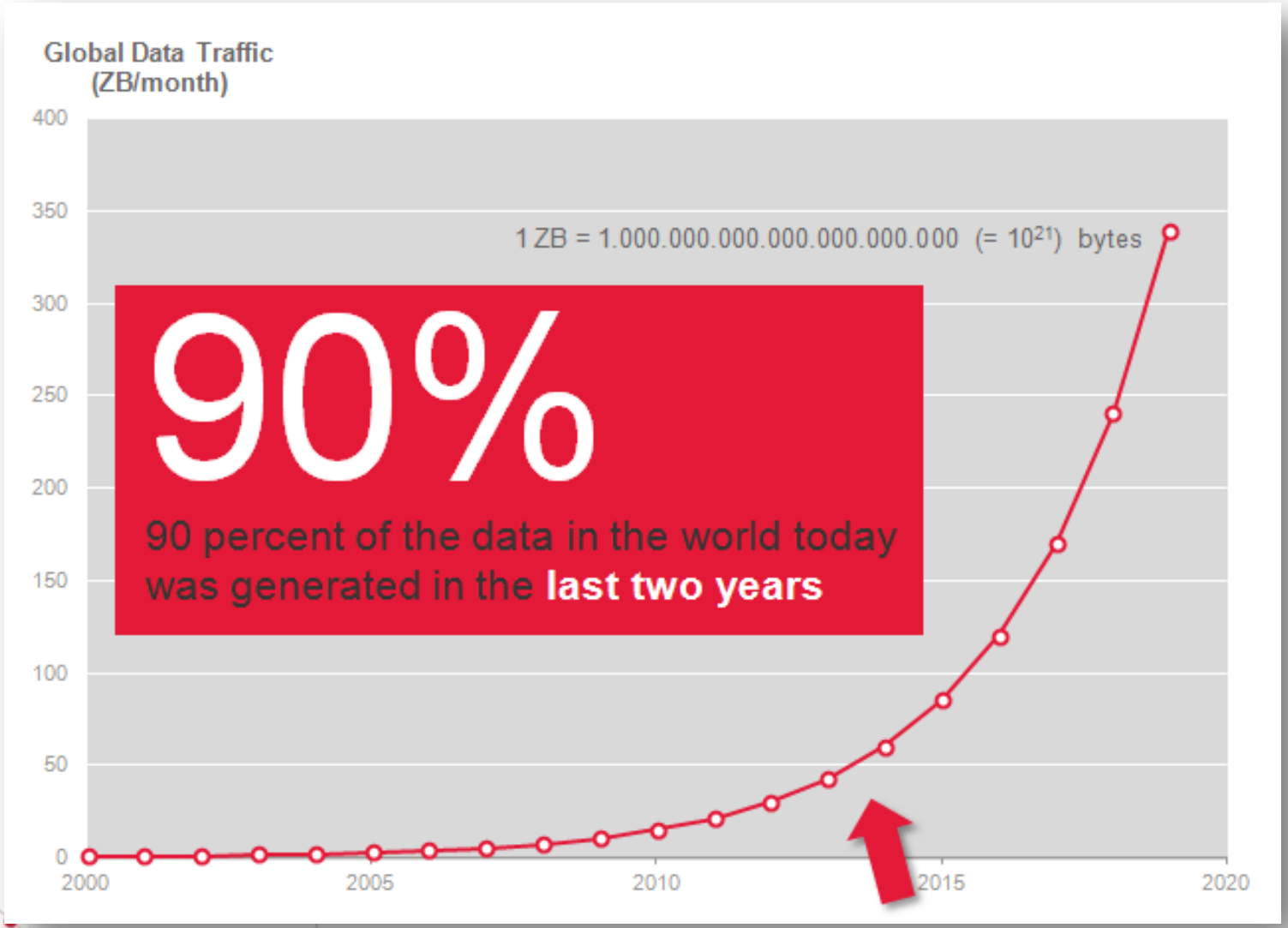


CGI

Introduction BIG Data

Experience the commitment®

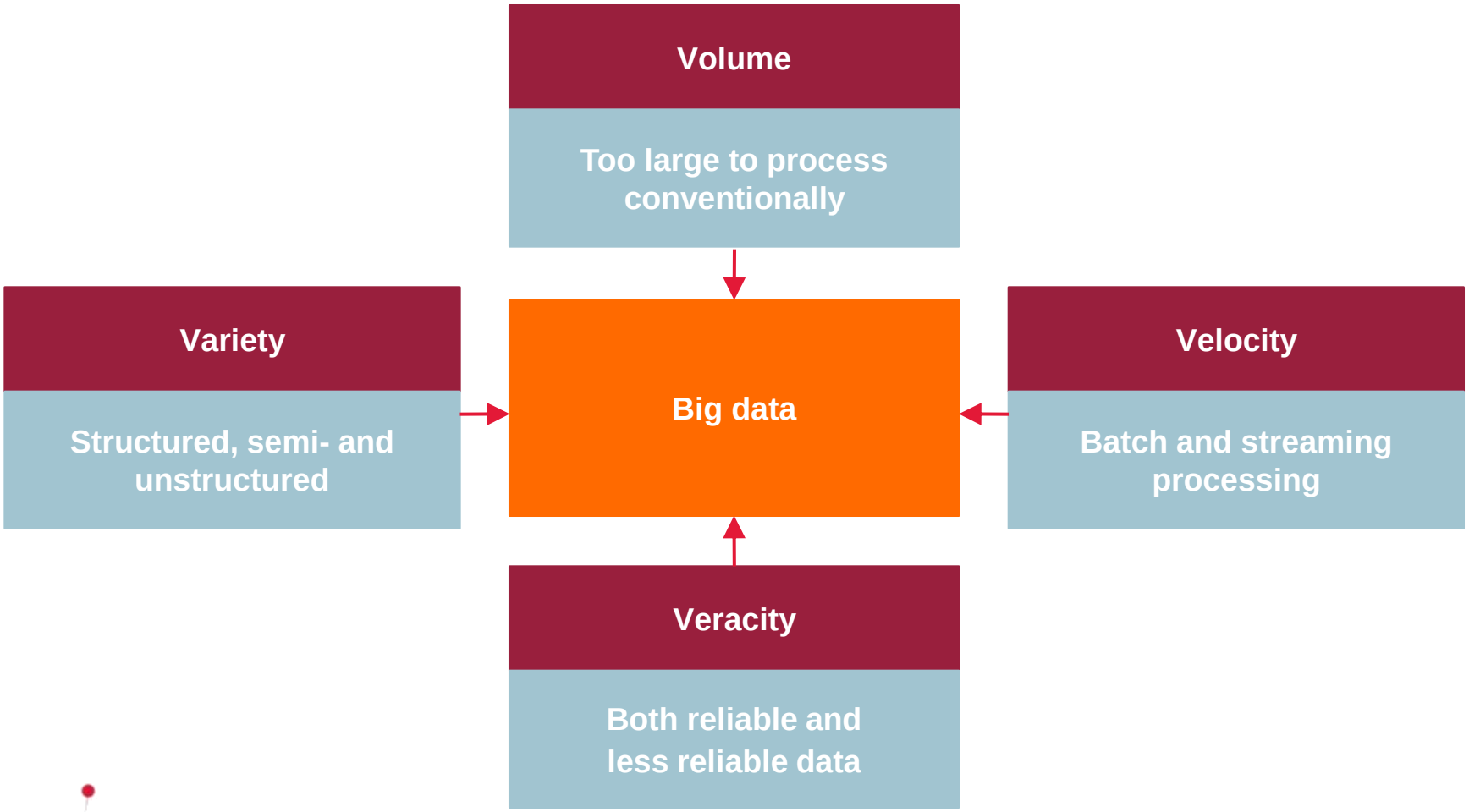
BIG Data: Data explosion!



CGI


Google
Spotify
YouTube
DOMAINS
facebook
2
14
72
HOURS
70
41
15
THOUSAND
skype
1.8
1.4
350GB
347
WordPress
flickr
20
MILLION
571
WEBSITES
278
THOUSAND
twitter
17
THOUSAND
Walmart
104
THOUSAND
20
THOUSAND
tumblr.
snapchat
$83,000
204
MILLION
11
THOUSAND
11
THOUSAND
3,600
EMAIL
amazon.com
LinkedIn
Pinterest
Instagram

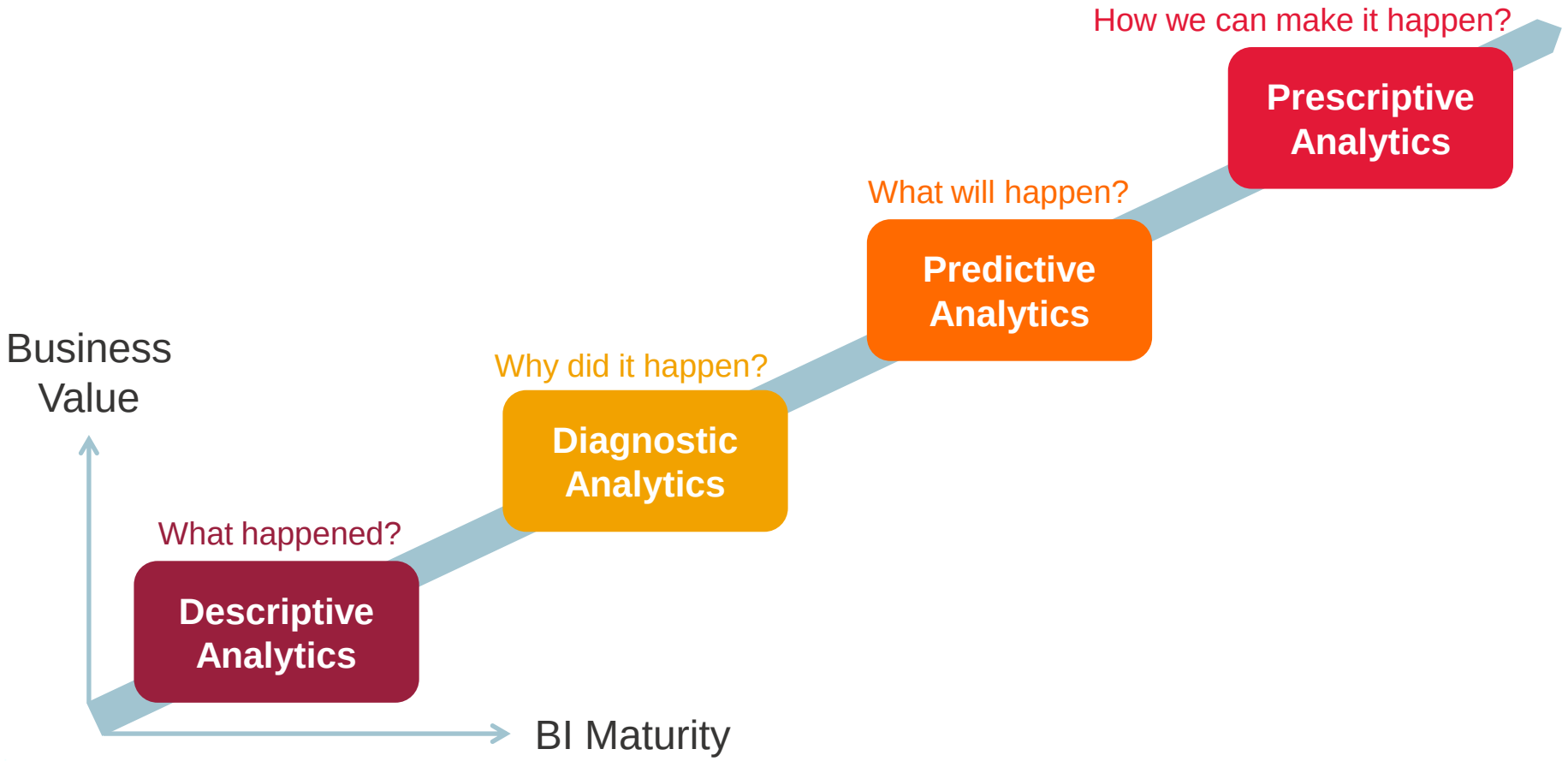
How to define BIG Data?



Big data! Big deal?

Additional or new business models:
Shining diamonds

The evolution of BI and Analytics



BIG Data solutions and traditional BI

BI & Reporting

Analytics

Visualization & discovery

Data warehouse

Data at rest

Data in motion

In-Memory/MPP

Hadoop/NoSQL

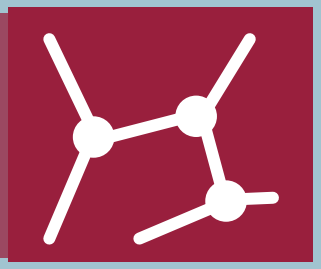
Streams

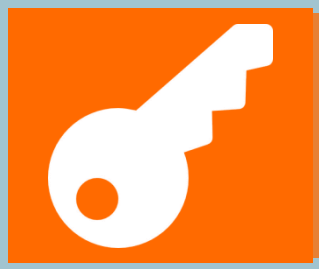
Information Integration & Governance

CGI

Important Big Data applications

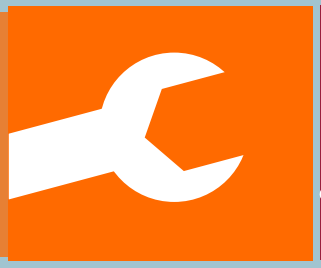
Network analytics
water, defense, social

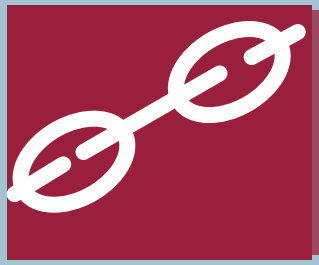




Fraud detection
finance, public

Predictive maintenance
utitlities, petrochemistry





Supply chain optimization
chemistry, retail, manufacturing

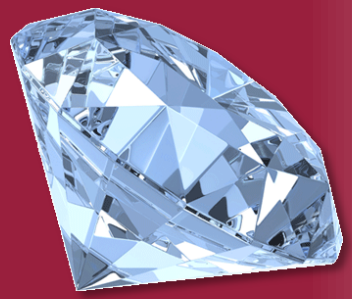
Profiling
customers, patients, citizens

CGI

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

Experience the commitment®

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

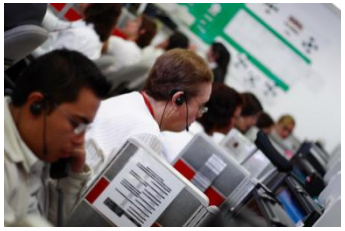
… leiden tot hoge callcenterbelasting

… kunnen tot ernstige gevolgschade leiden

… veroorzaken verlies aan waterkwaliteit

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

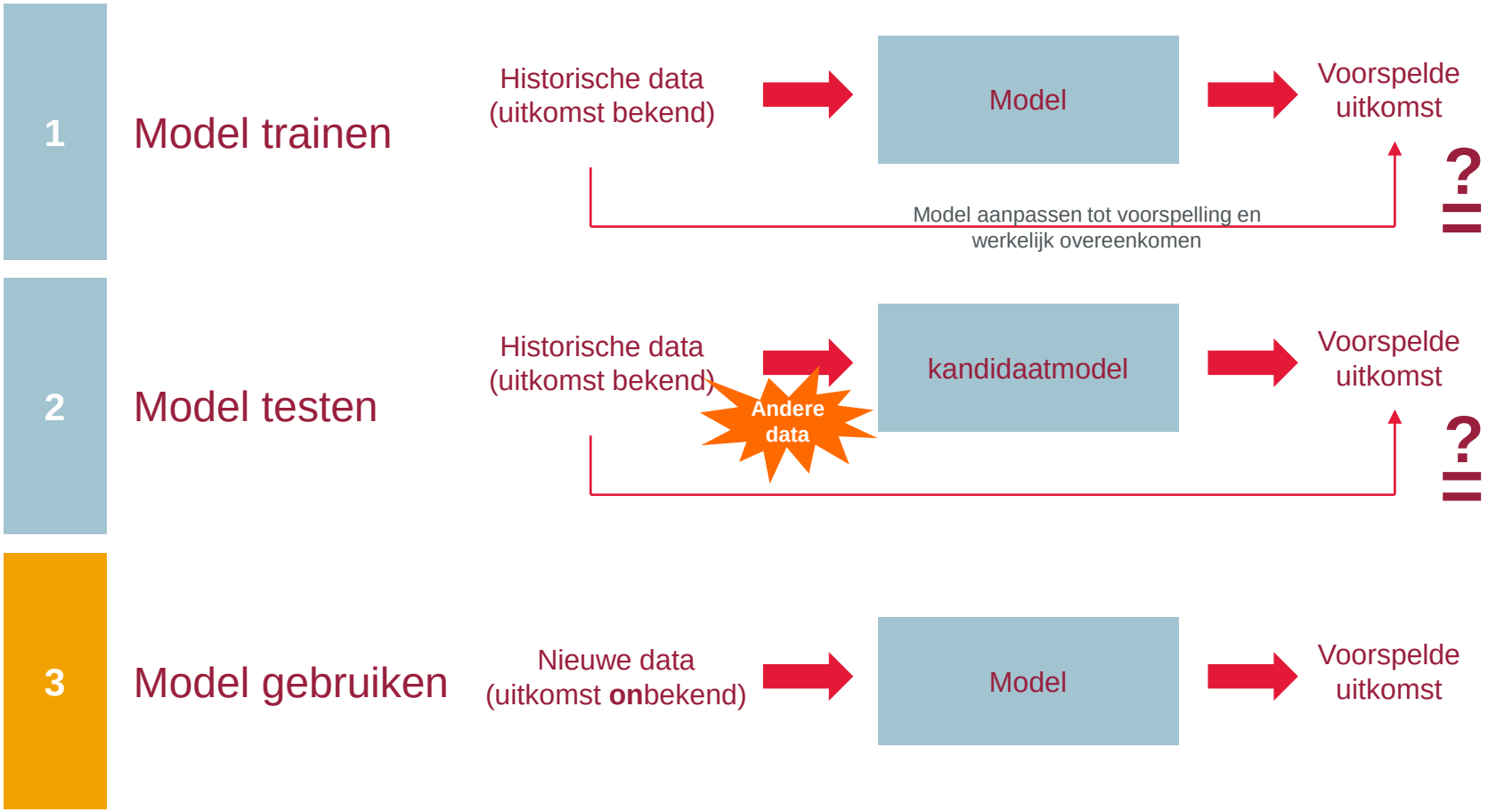


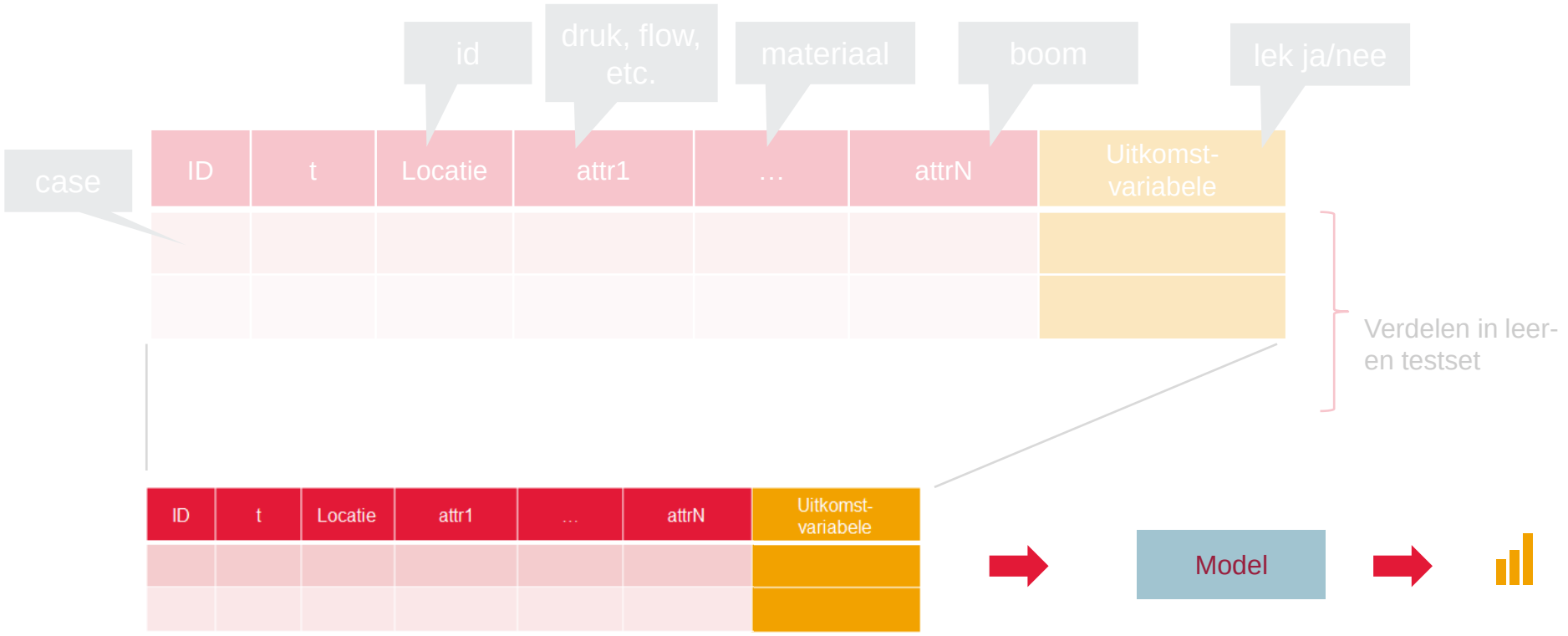

Data2Diamonds® - Big Data Proof of Value
De uitdaging

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

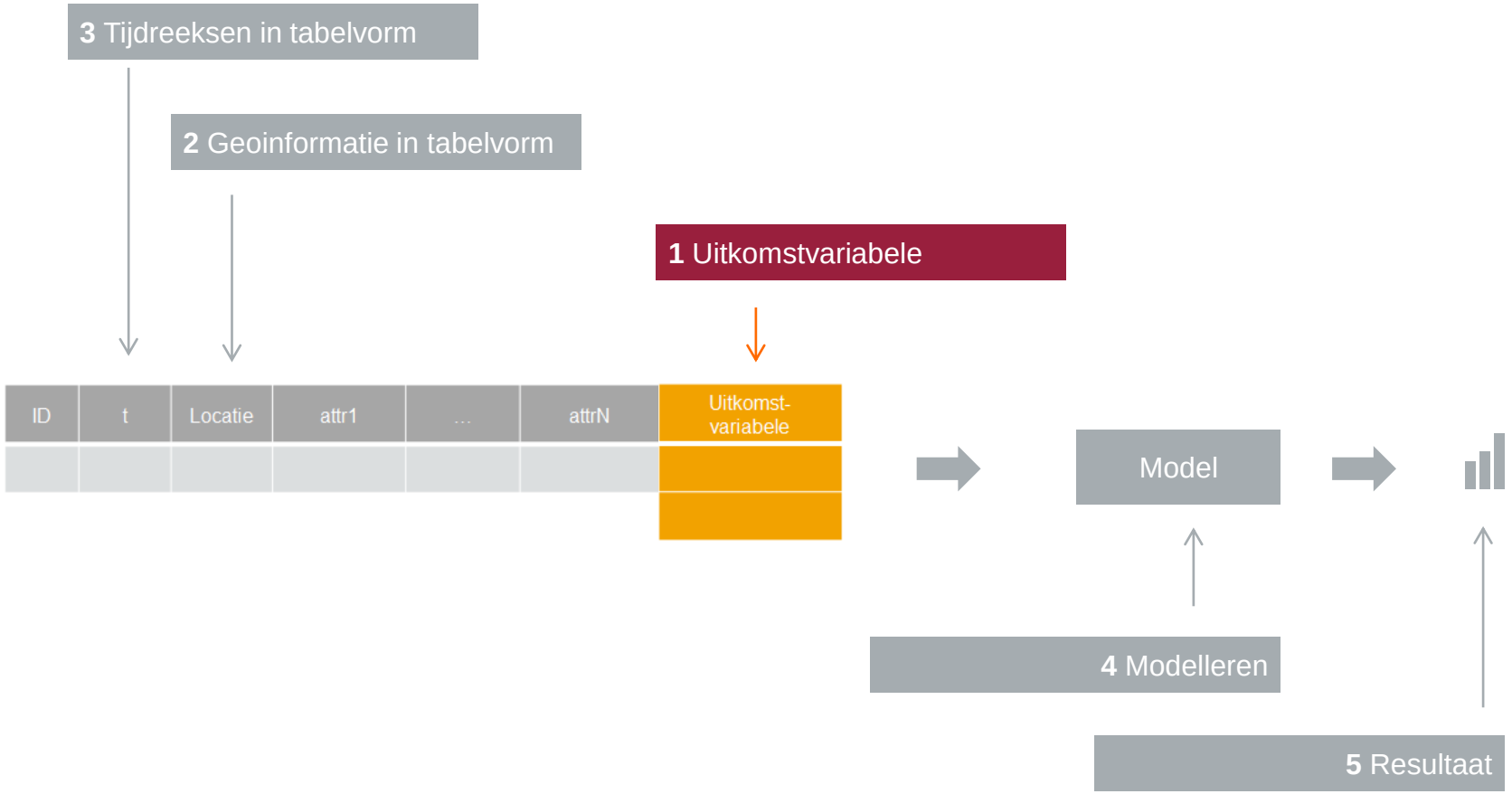
Detectie/lokalisatie = *machine learning*



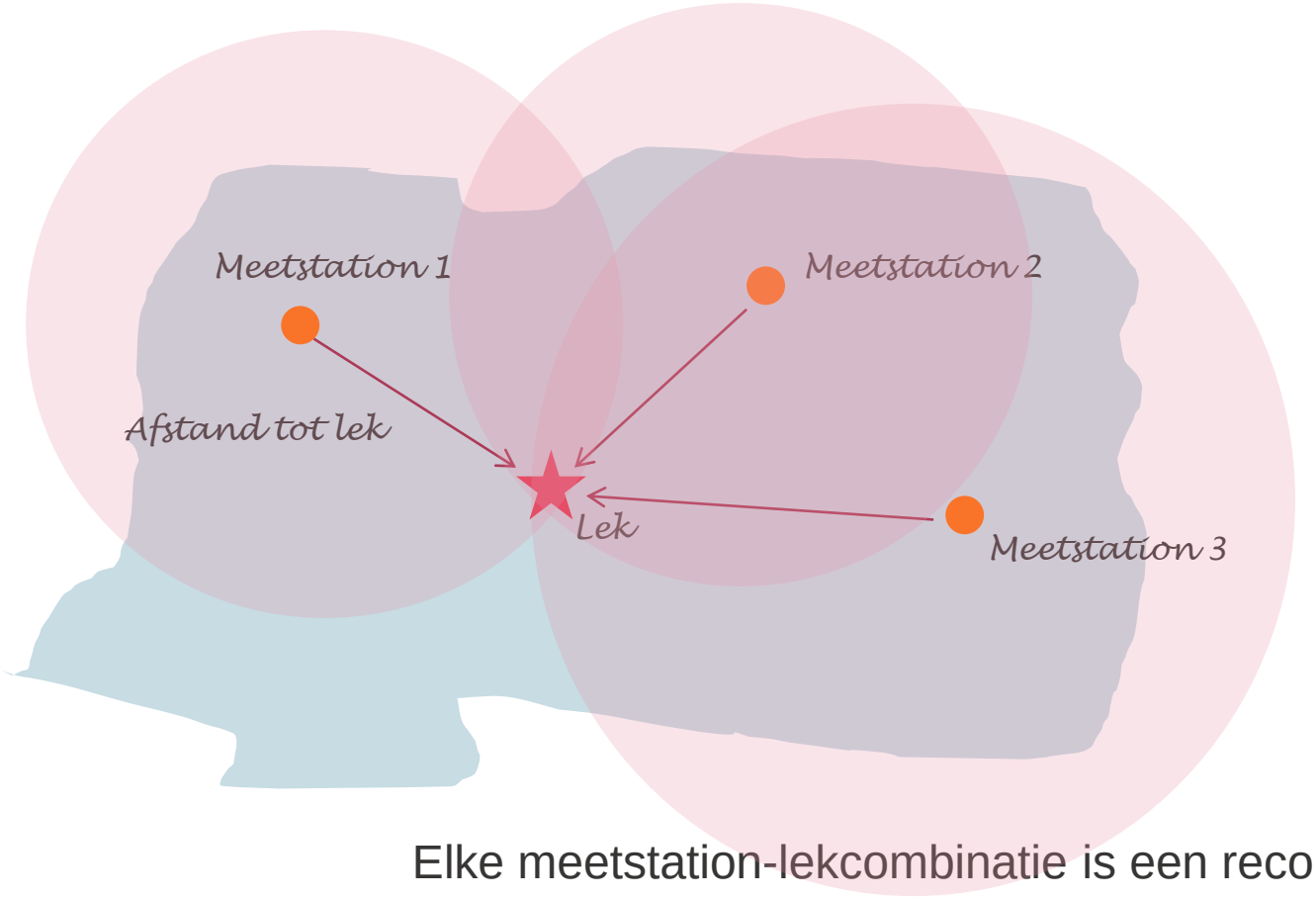
Aanpak machine learning



Stap 1: Uitkomstvariabele analyseren

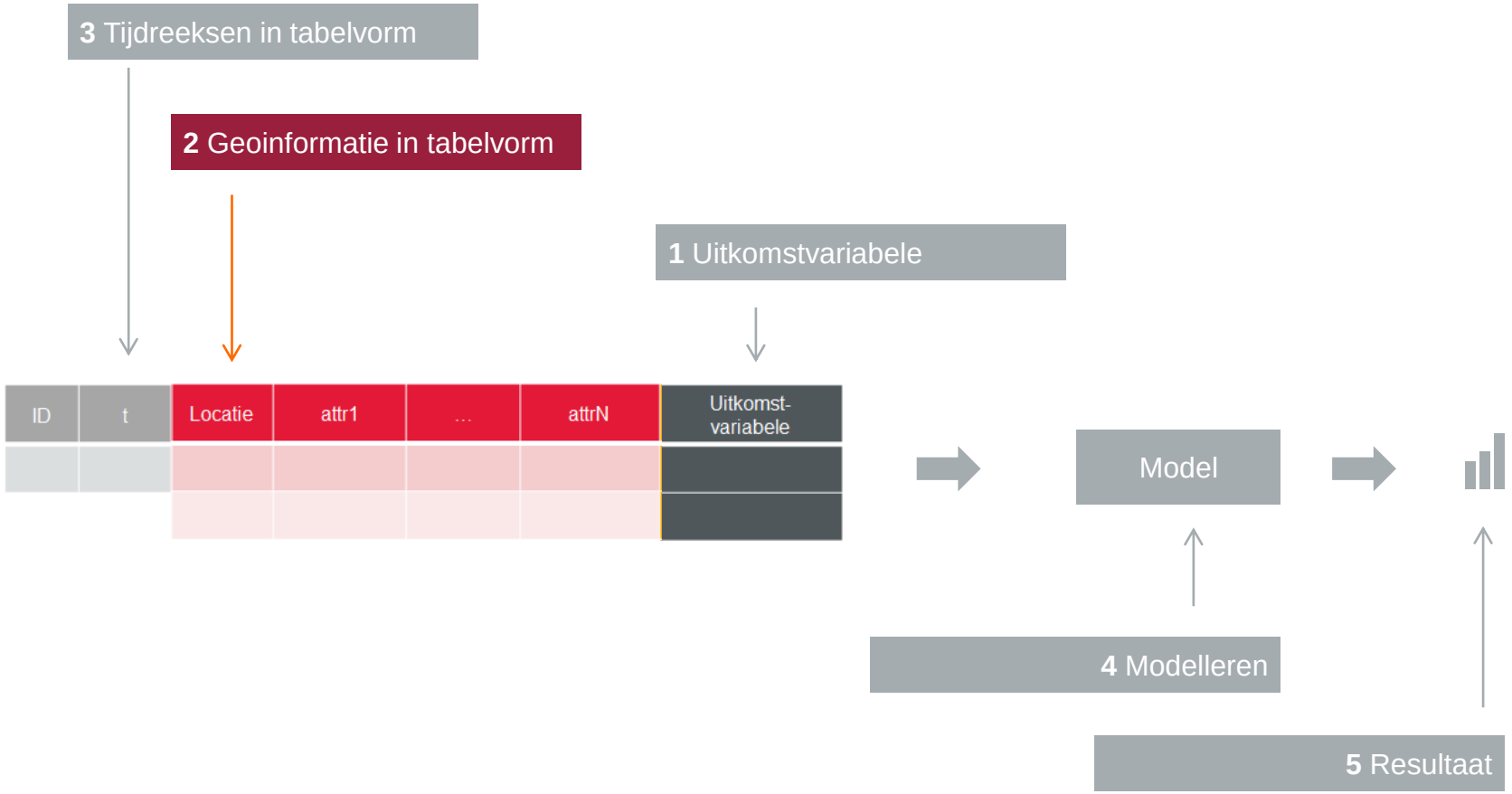


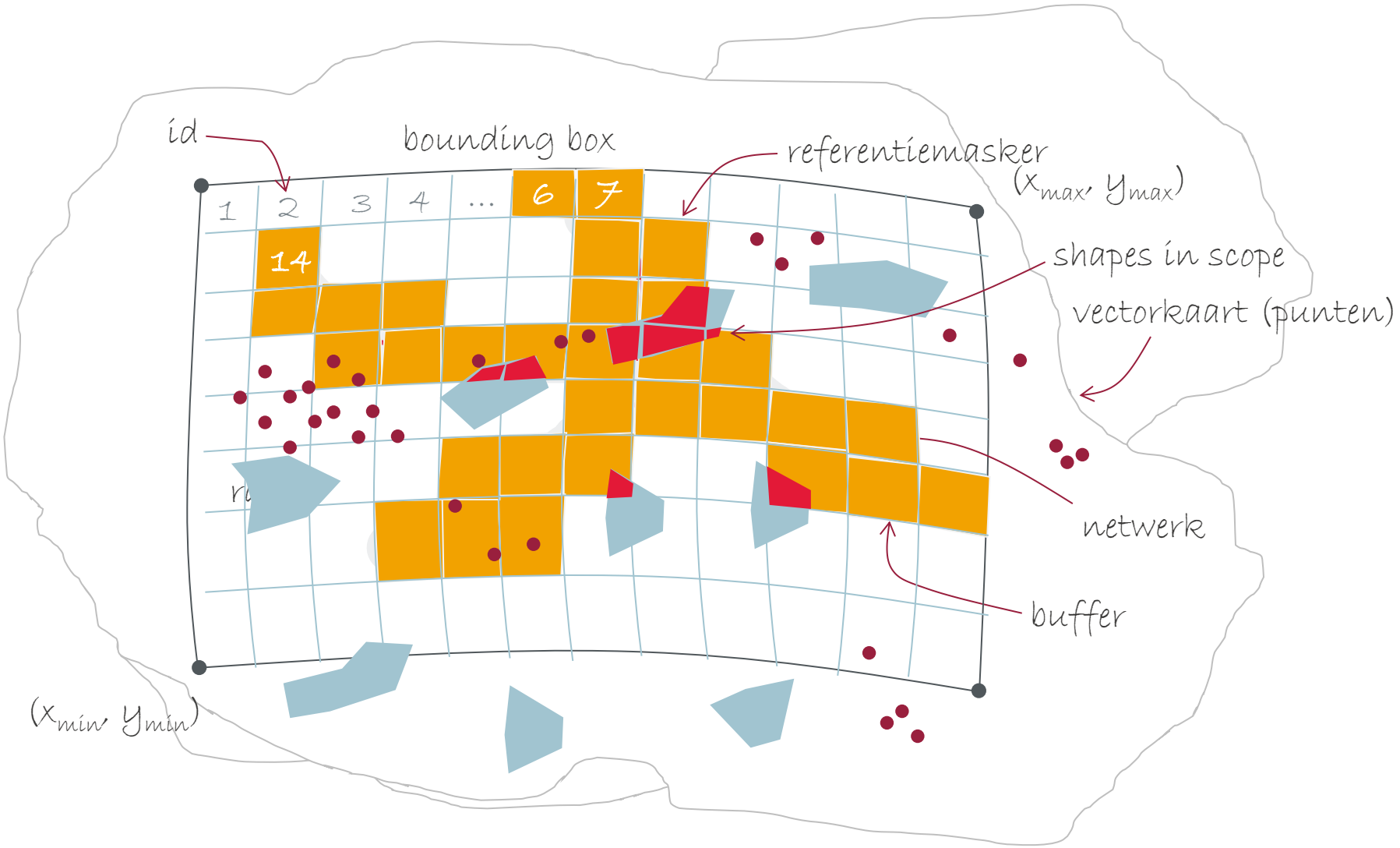
Afstand tot lek als uitkomstvariabele



Elke meetstation-lekcombinatie is een record
Aantal meetpunten groter

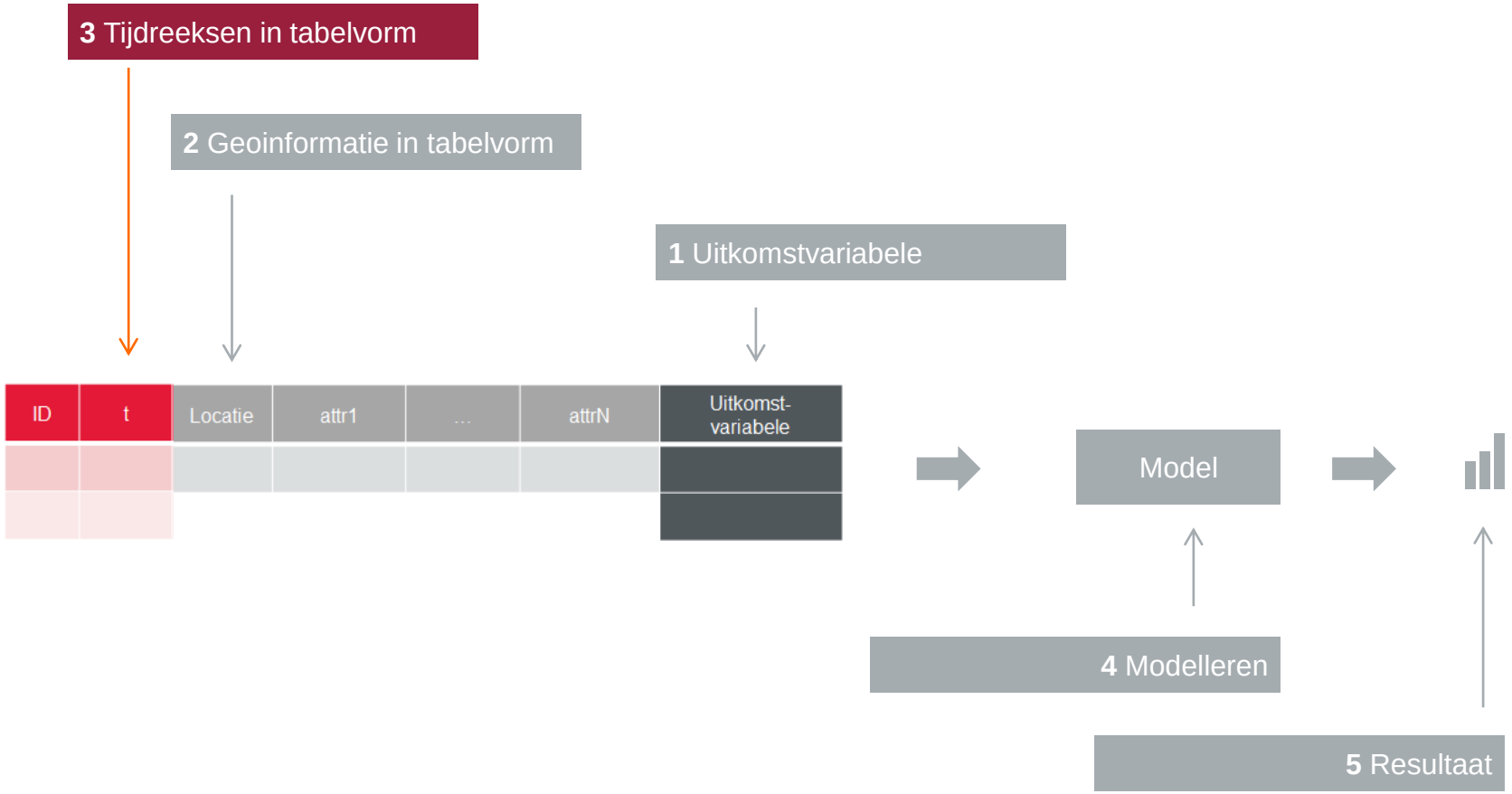
Stap 2: geoinformatie in tabelvorm



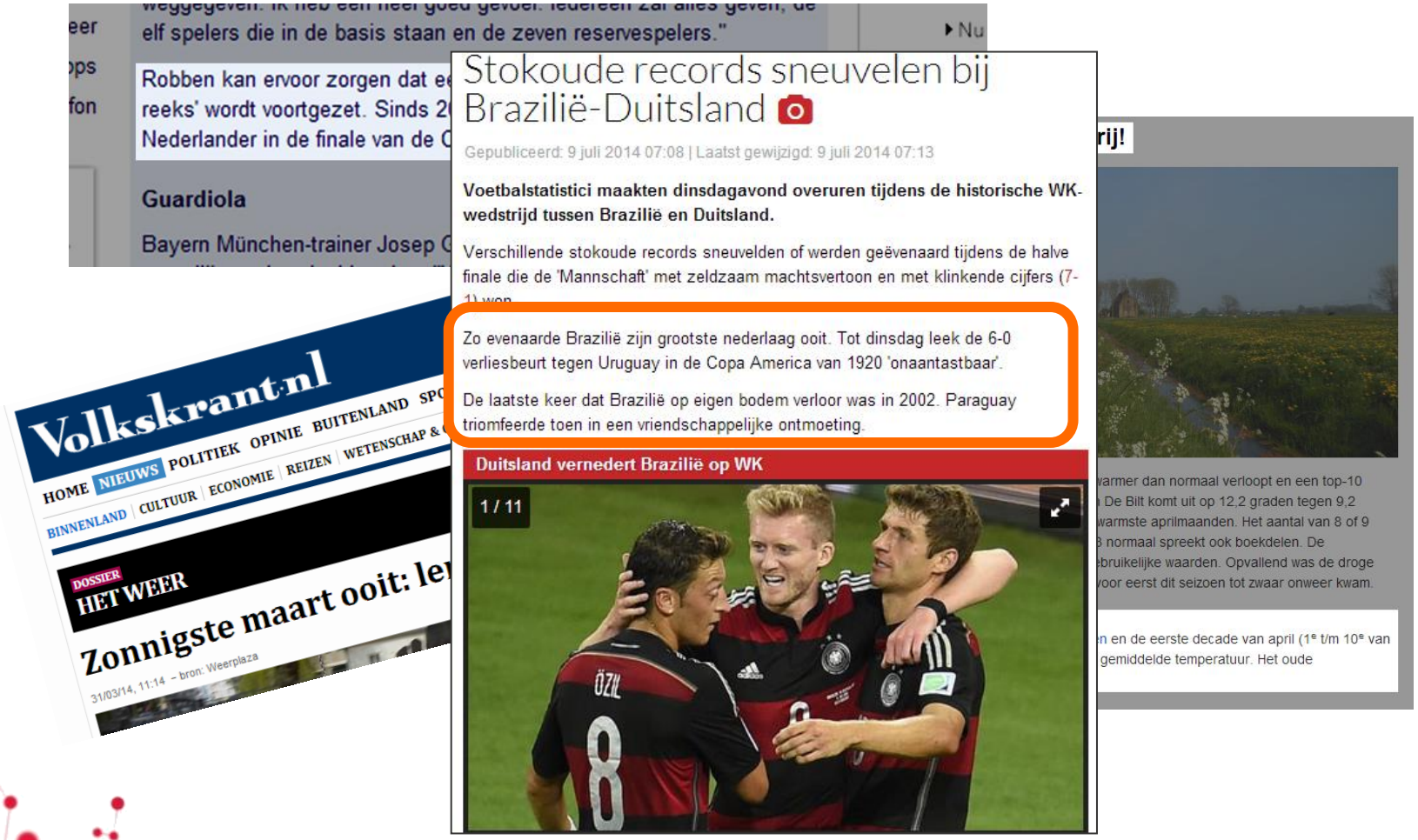

id
bounding box
referentiemasker
(x_{max}, y_{max})
1
2
3
4
...
6
7
14
shapes in scope
vectorkaart (punten)
netwerk
buffer
(x_{min}, y_{min})

CGI

Stap 3: tijdreeksanalyse



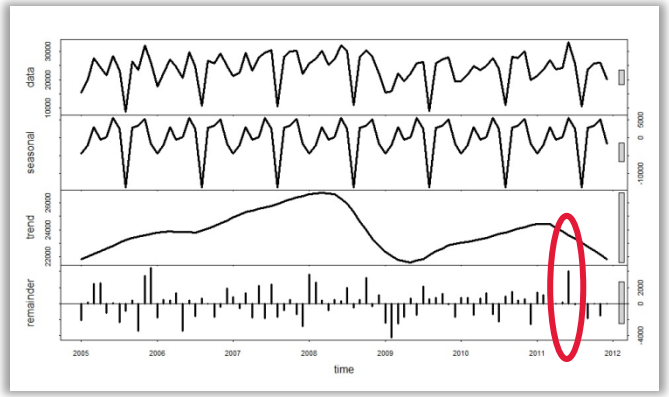
Tijdreeksanalyse voor het ontdekken van patronen…
…maar wat is een patroon?



Tijdreeksanalyse: waarom en hoe?

Doel tijdreeksanalyse

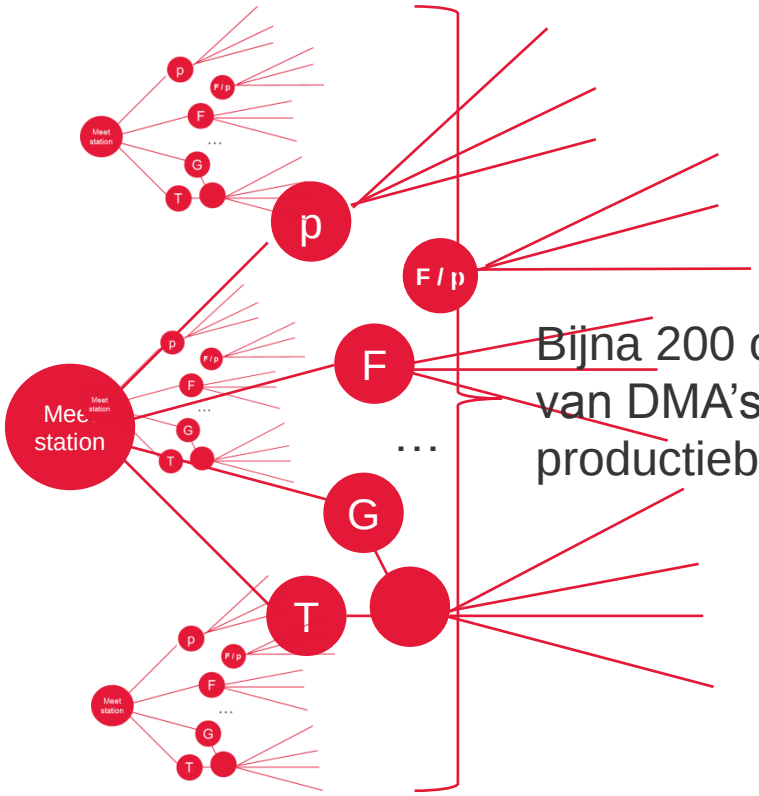
- Uitzonderingen opsporen
- Uitzonderingen versterken



Technieken

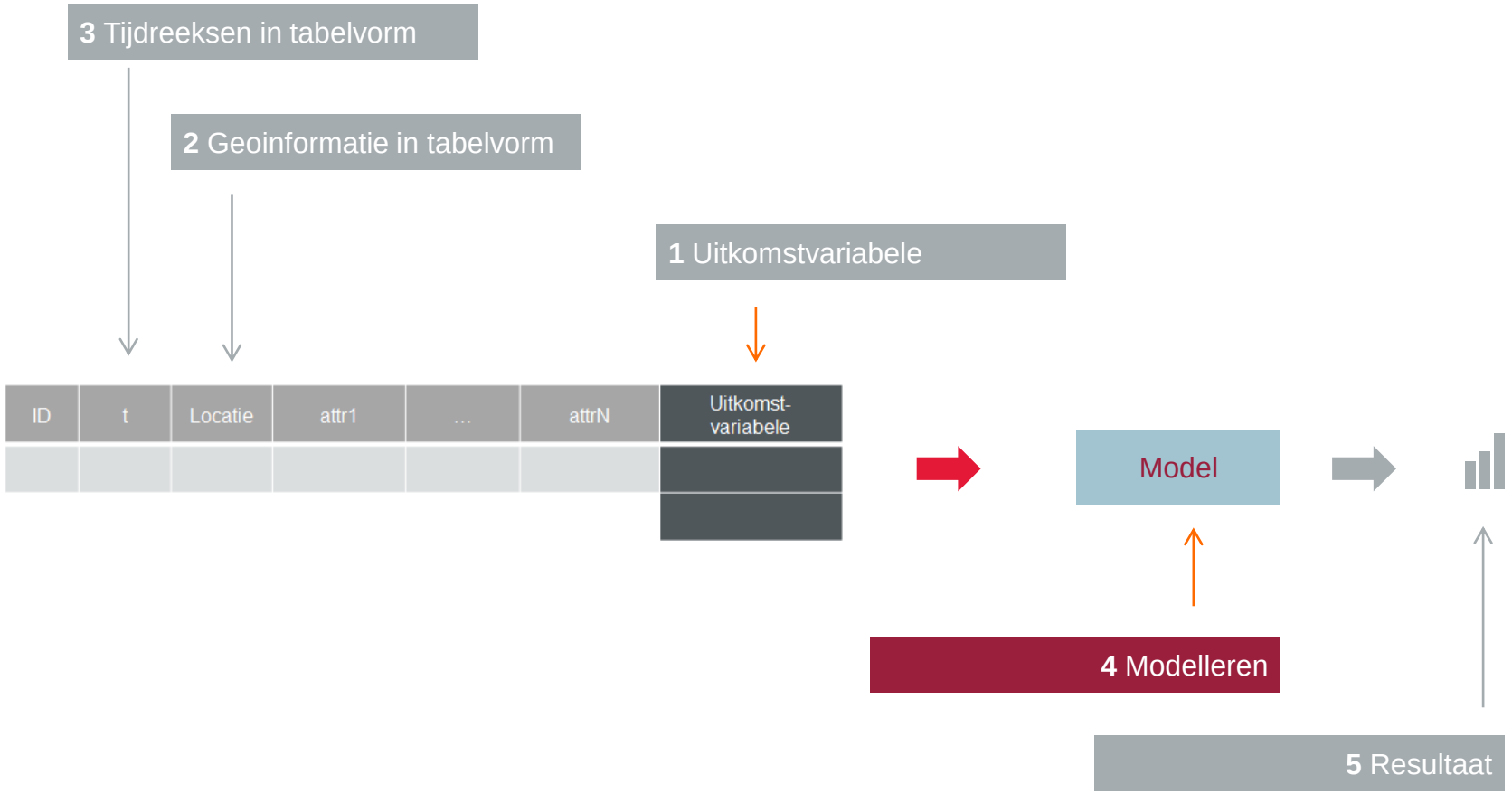
- Middelen, standaarddeviatie (beschrijvend)
- Trendanalyse
- Signalen combineren (domeinkennis)
- Spectraalanalyse
- Modelleren
- Filteren
- Waveletanalyse
- ….
- …

Afgeleide en afgeleiden van combinaties



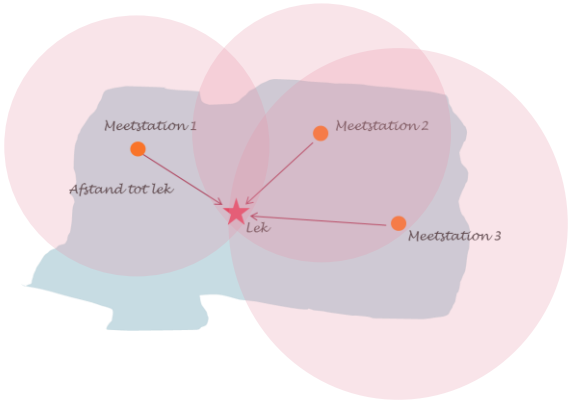
Bijna 200 originele en afgeleide signalen afkomstig van DMA’s, opjager, distributiereservoirs, productiebedrijf, deelbalansgebiedvolumestroom

Stap 4: modelleren

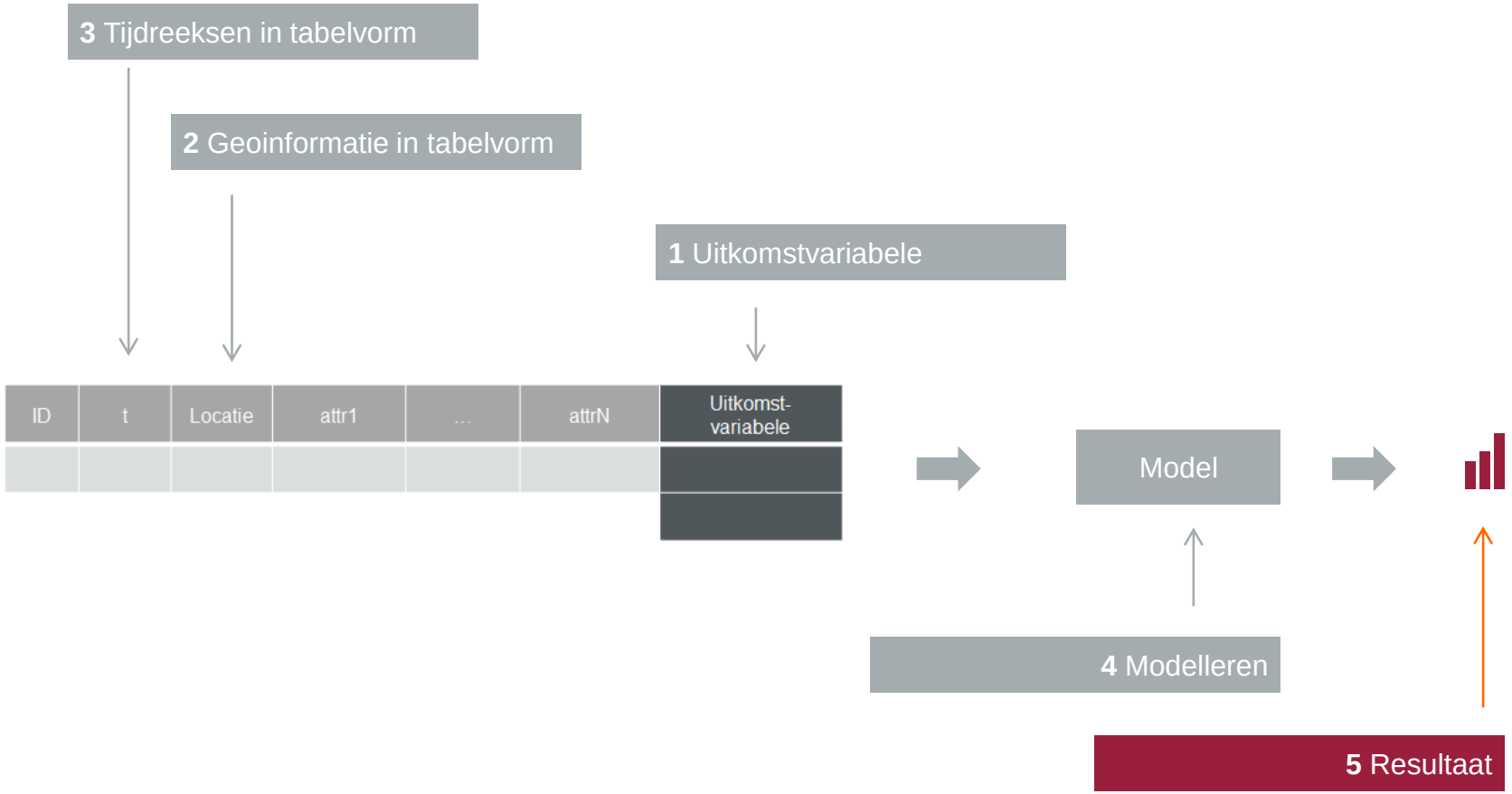


Aanpak afstand tot lek

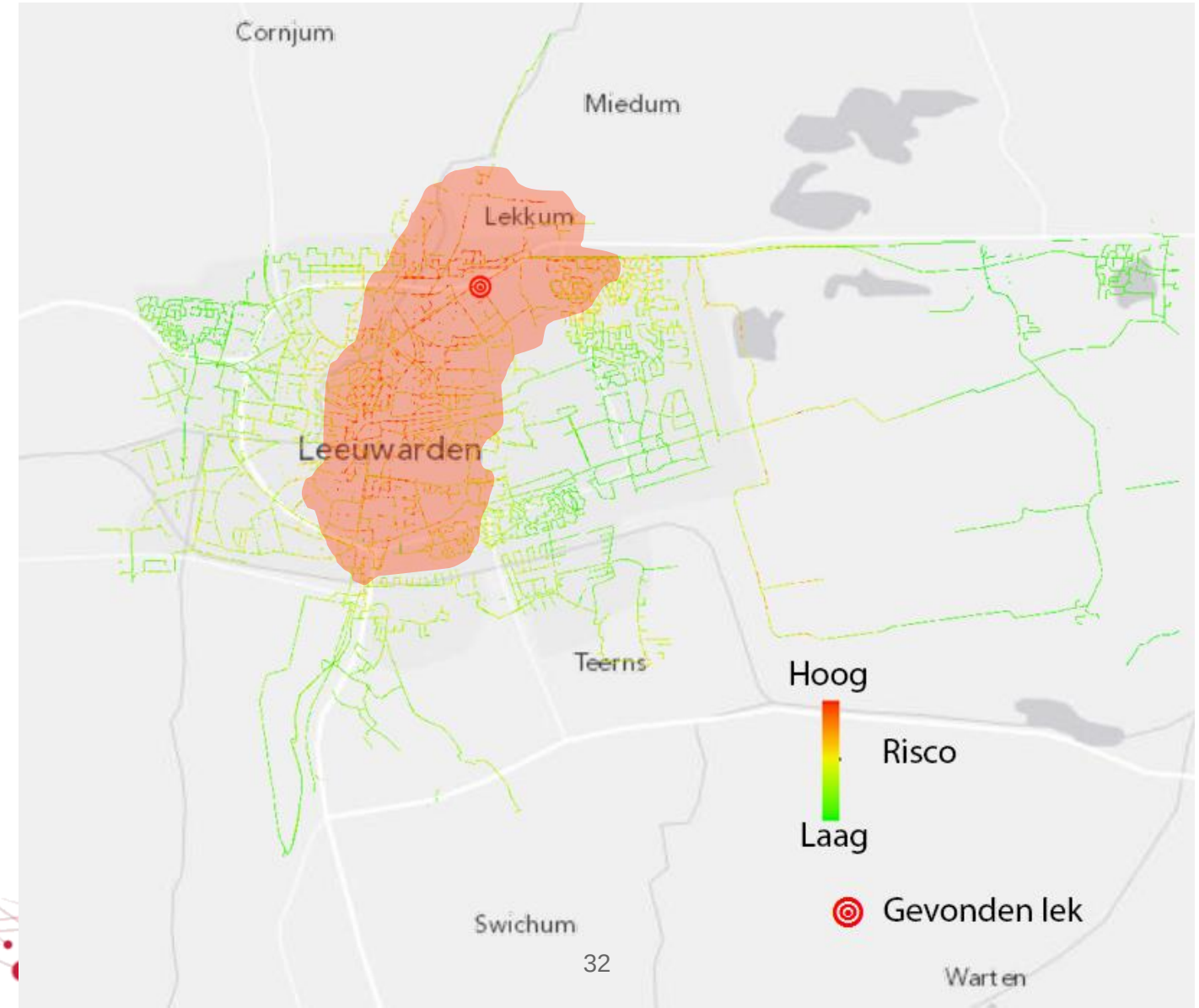
- Voorspel afstand van lek tot meetstation op basis van inputsignalen
- Iedere meetstation-lekcombinatie is een datapunt
- Construeer cirkels om tags



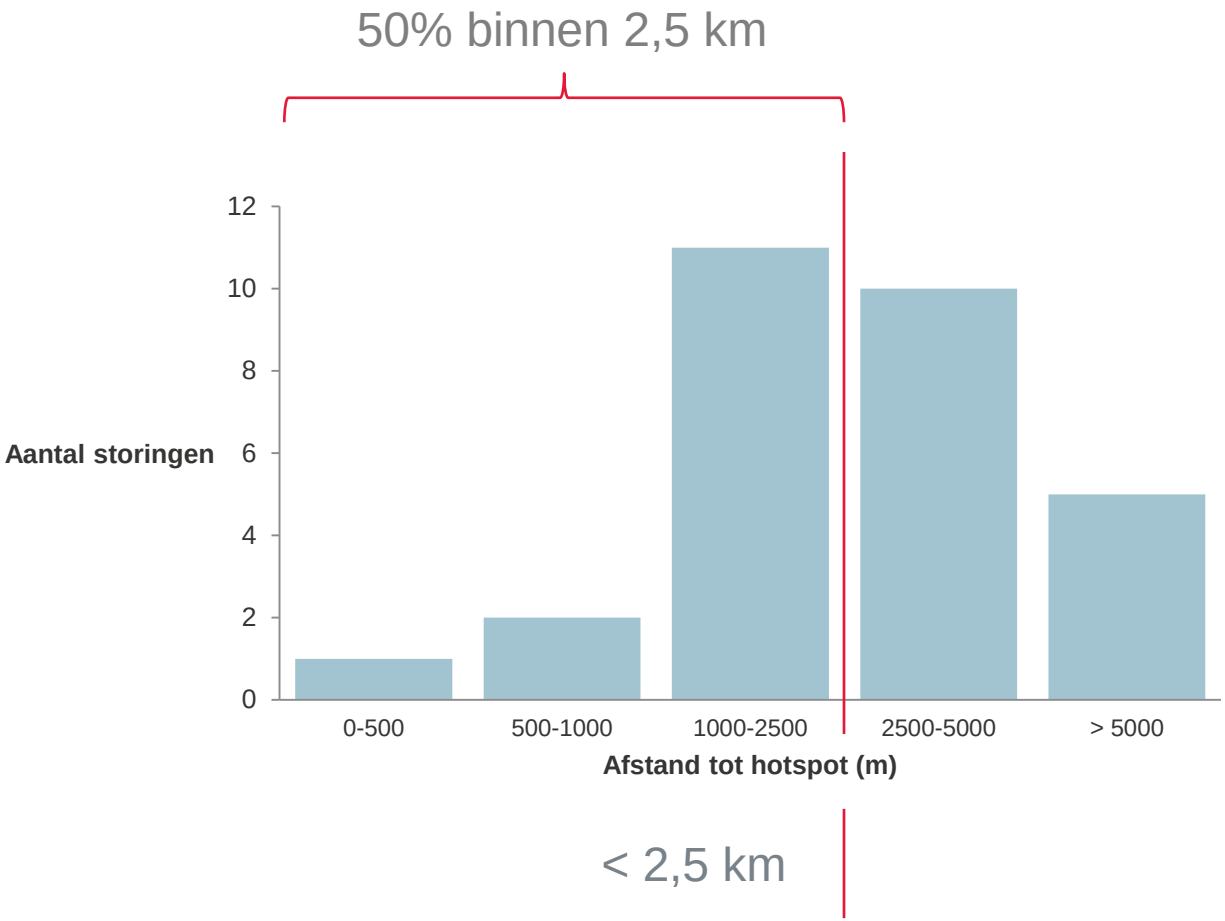
Stap 5: Resultaat



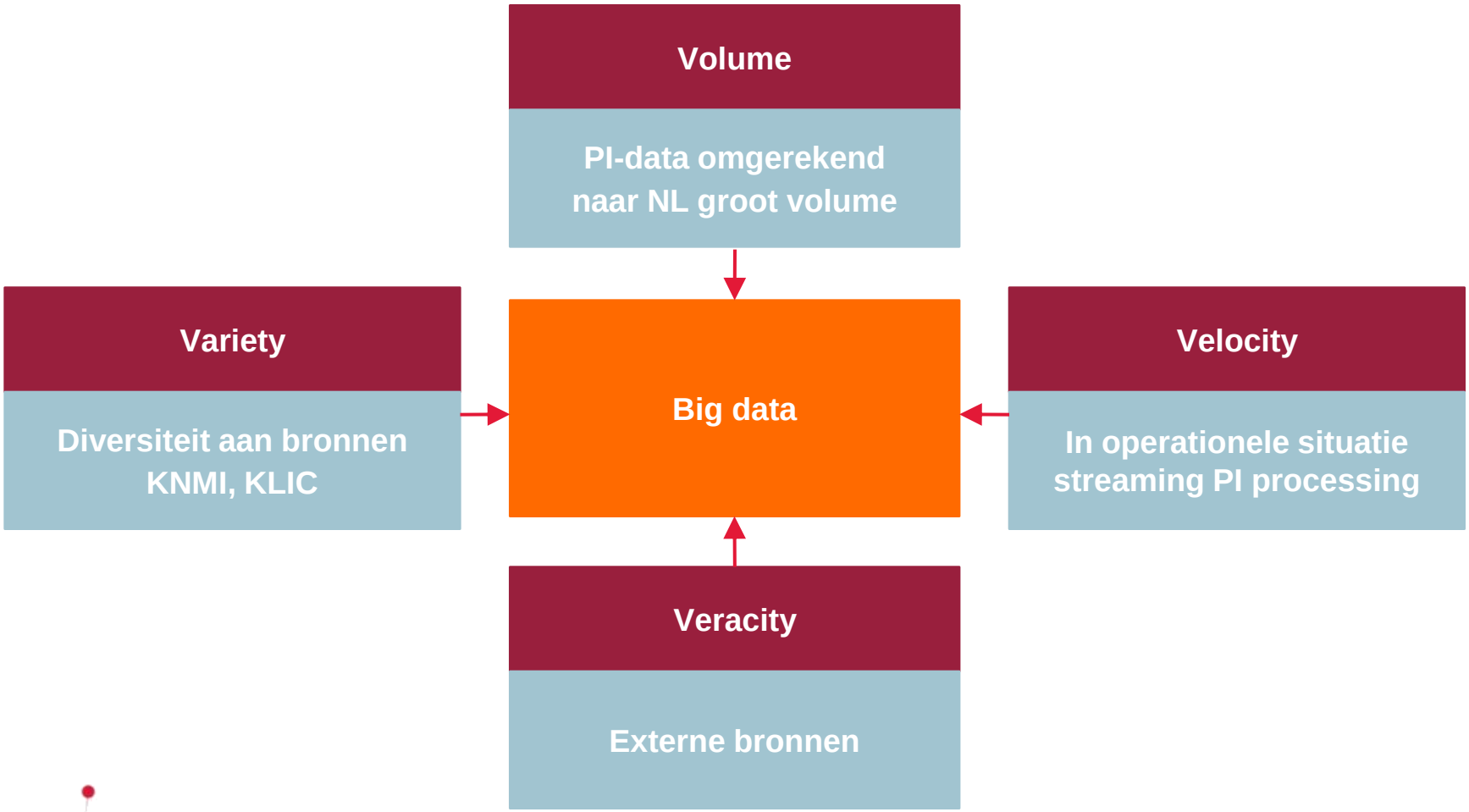
Leklokalisatie



Resultaat


50% binnen 2,5 km
Aantal storingen
12
10
8
6
4
2
0
0-500
500-1000
1000-2500
2500-5000
> 5000
Afstand tot hotspot (m)
< 2,5 km

Proof of Value: ‘little’ big data



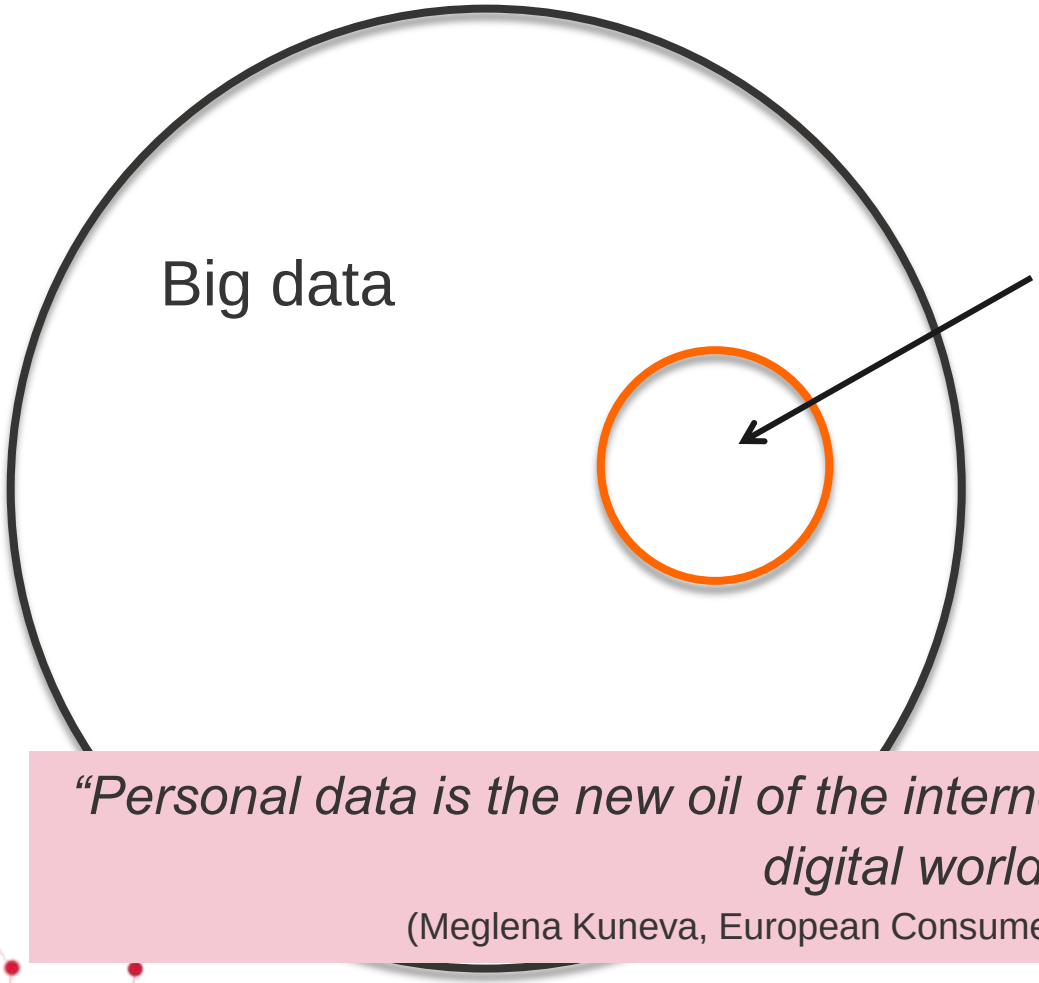
Data2Diamonds® - Privacy & BIG Data

TestNet – Summer School

Caroline Massart
July 9, 2014

Experience the commitment®

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

CGI

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

CGI

Customer: privacy is hot topic!

‘Google biedt bedrijven snel internet in ruil voor gegevens’

groter plan om meer mensen web-diensten van Google te laten
een zwaar gesubsidieerd WiFi-netwerk

Whatsapp reageert op privacyzorgen

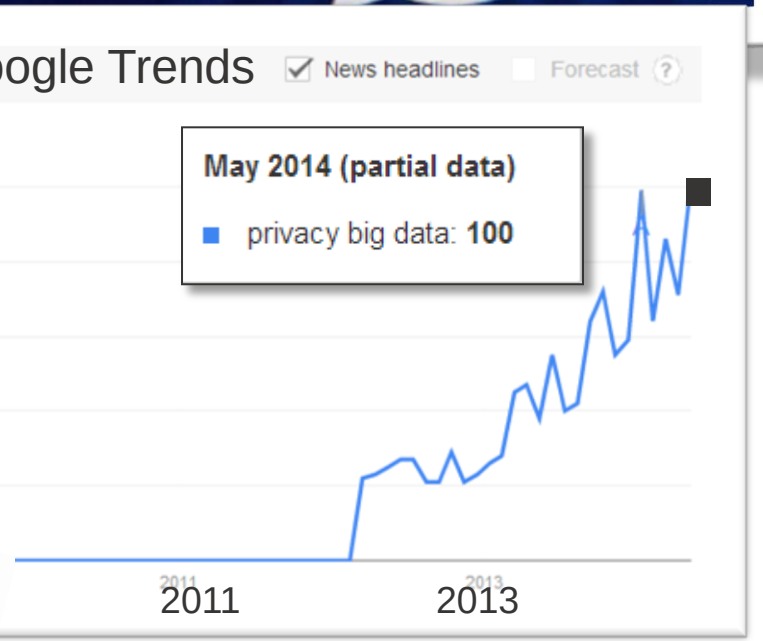
dinsdag 18 mrt 2014, 03:43 (Update: 18-03-14, 06:45)

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

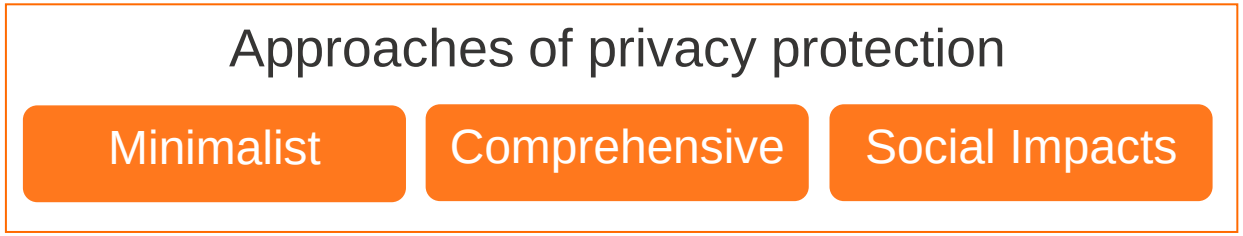


Organizations need a position on privacy

- 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.

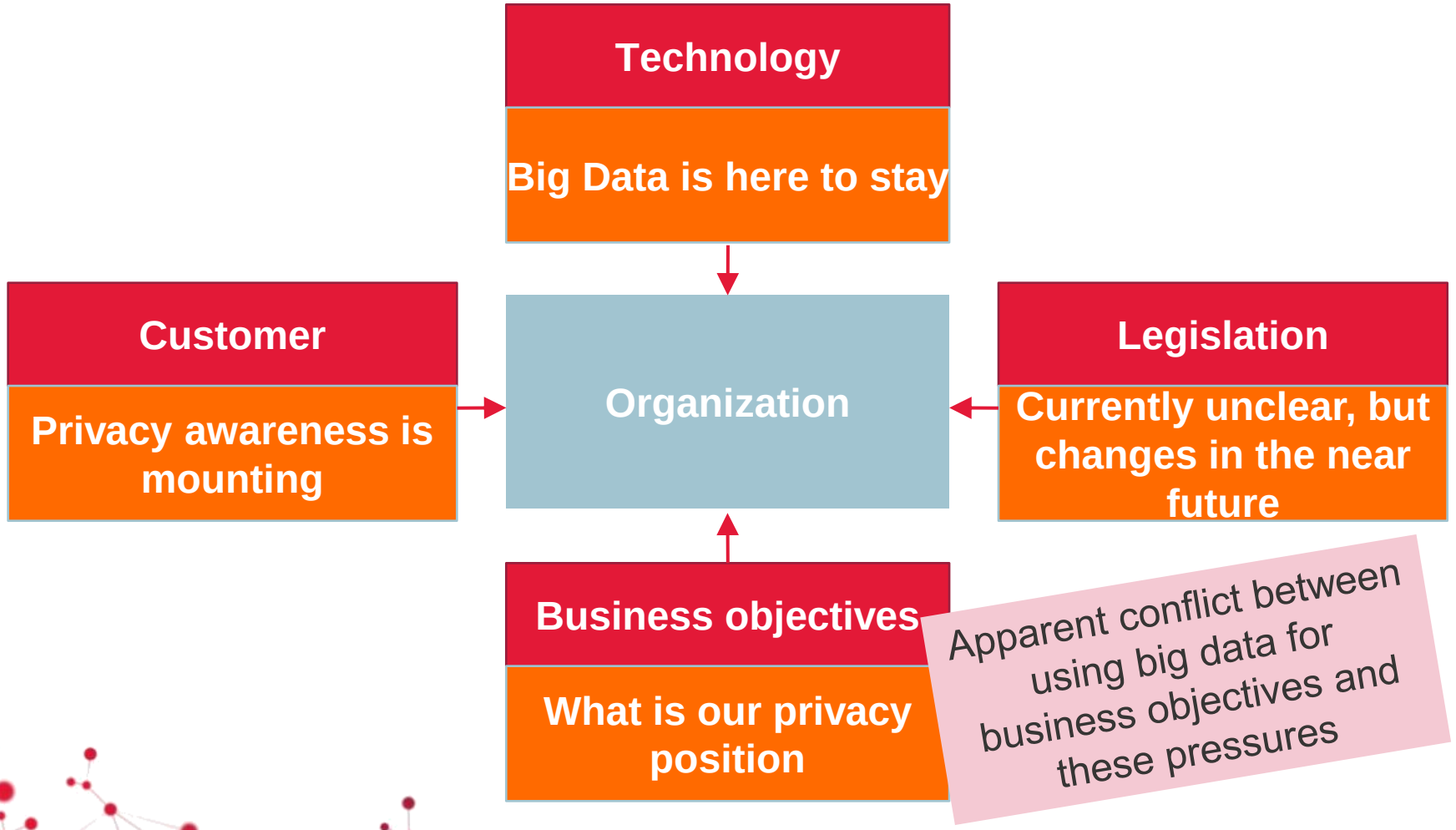
- The Information Commissioner's Office (UK) has identified three broad approaches of privacy protection, ranging from very narrow to the very broad.



- But what is the position of your organization?

Problem: privacy landscape under pressure

Customer privacy is a business problem that must be addressed!



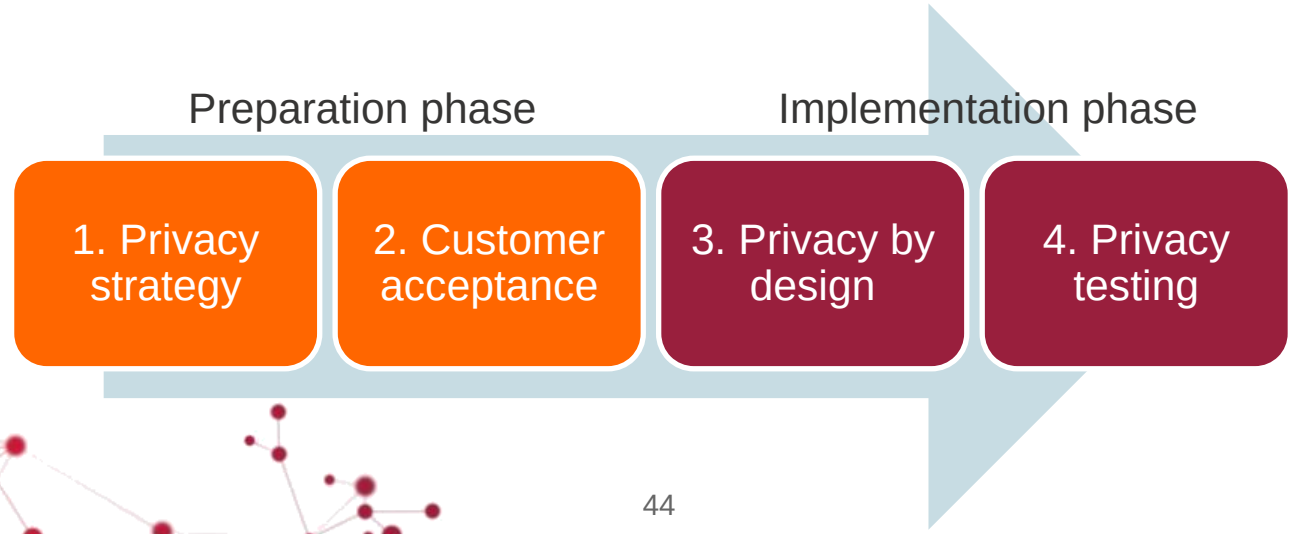
Approach

1. Organizations should develop a clear ***privacy strategy***

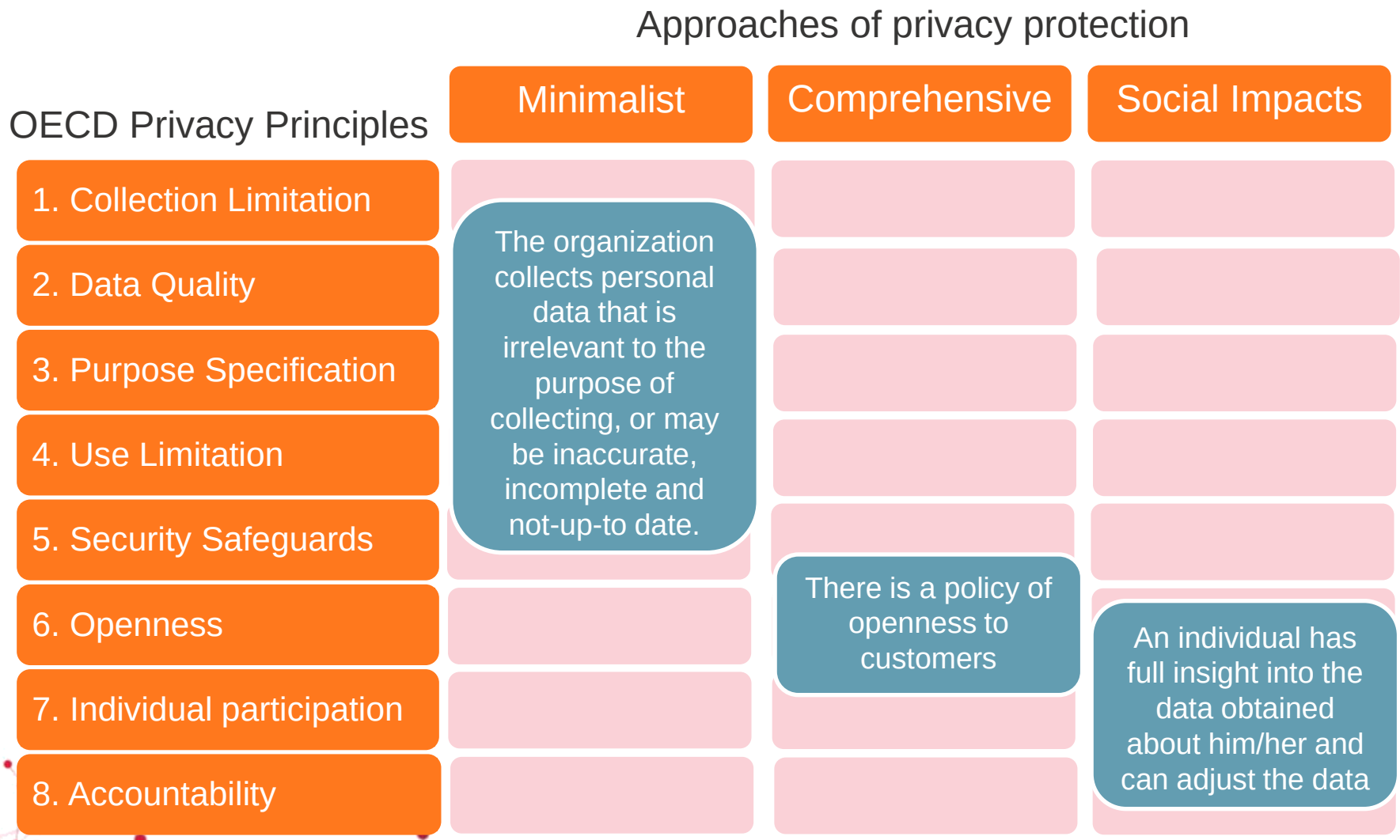
2. Assess '*privacy strategy*' for customer acceptance

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’



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

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!

CGI

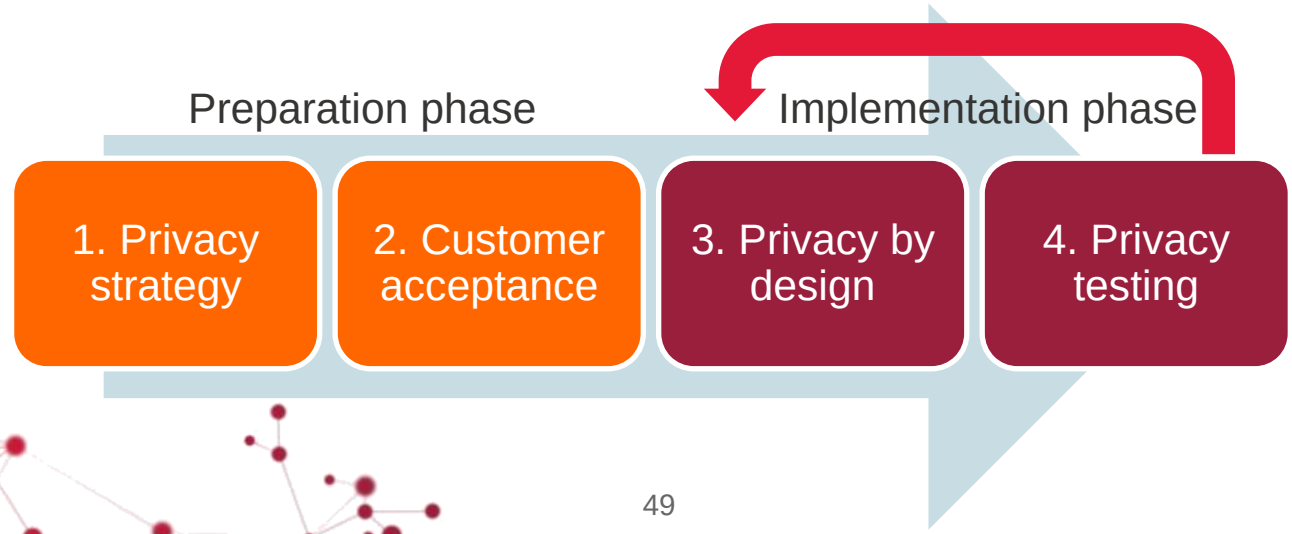
3. Privacy by Design

- Implement privacy early on in the design of the system

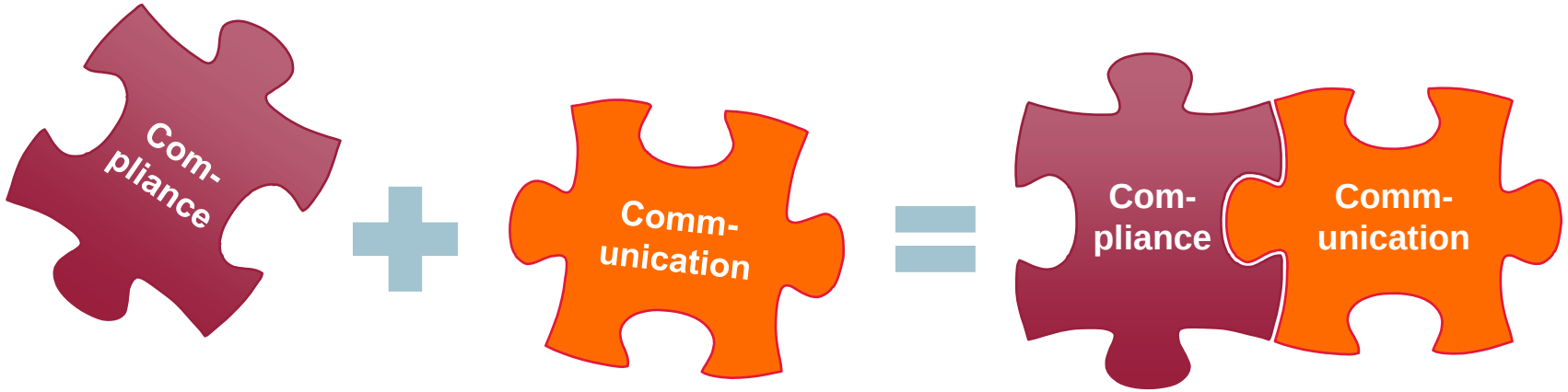
4. Testing privacy

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

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

Break

Experience the commitment®

Brainstorm workshop

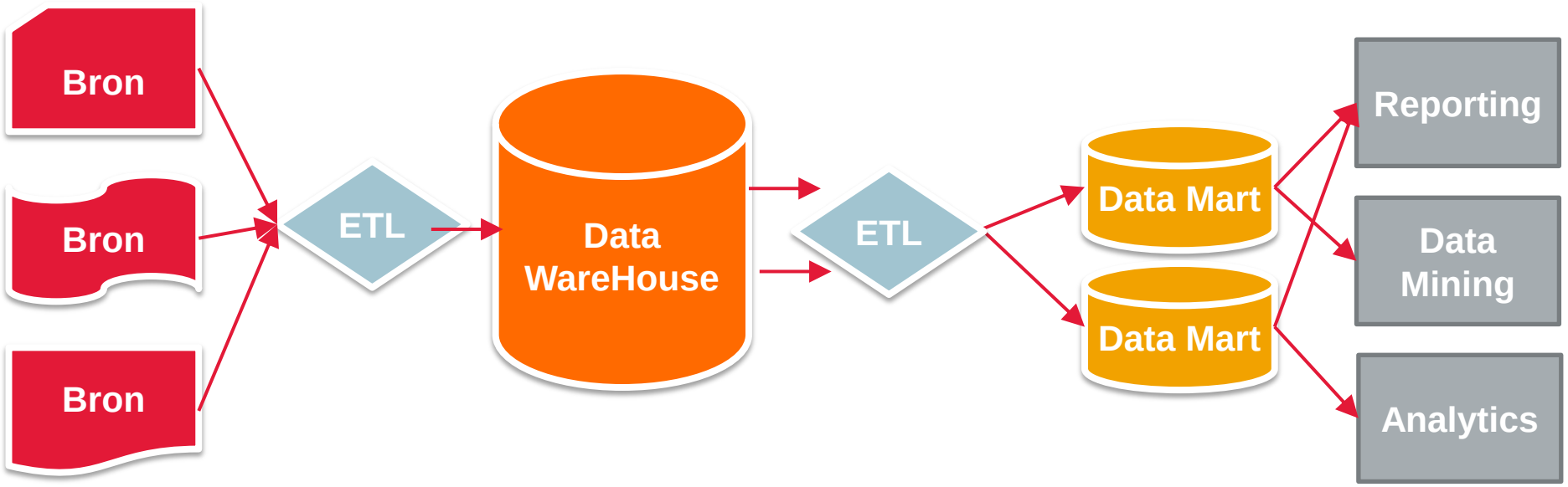
Hoe zou jij BIG Data oplossingen testen?

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

1e ronde – Klein beginnen… BI Testen

Hoe zou jij BI testen aanpakken?



2e ronde - 2 Test V's en 4 BIG Data V's

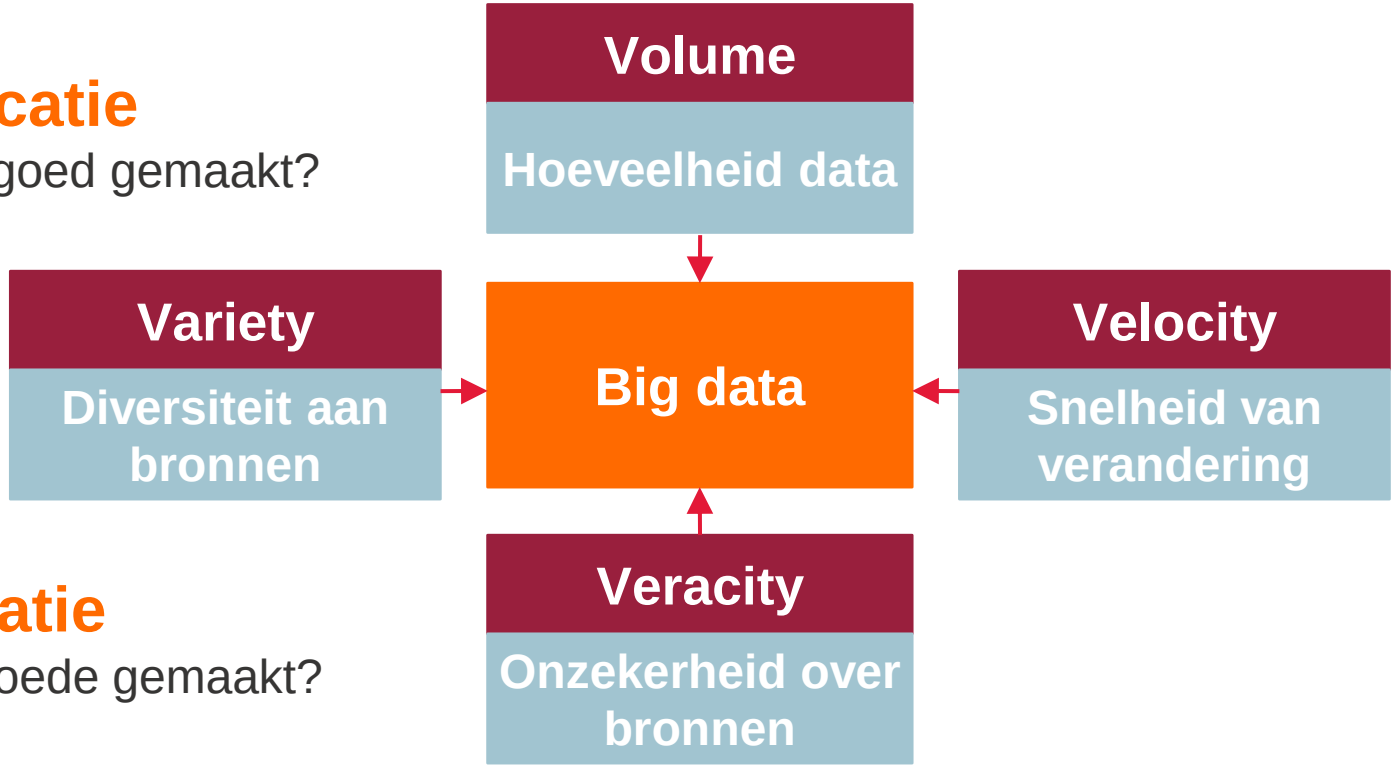
Wat voor testdoelen / kwaliteitstoetsen kun je bedenken voor de 4 verschillende Big Data V's. Denk daarbij aan verificatie maar ook aan validatie.

Verificatie

Hebben we het goed gemaakt?

Validatie

Hebben we het goede gemaakt?



3e 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

CGI

Our commitment to you

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

CGI

Experience the commitment®