

The logo for AethiQs, featuring the company name in a white serif font on a blue rounded rectangular background.

**AethiQs.**

Remain Relevant!



# VSAE congress

## *Break-out interactive session*

*The relationship between Data Quality and Artificial Intelligence  
In the light of the new pension scheme in the Netherlands*

*Including use cases in practice*

**AethiQs – Aron Jeurinck & Joost Opvel**  
**March 5<sup>th</sup>, 2024**





# Introduction

## Backgrounds

- Econometrics, actuarial science and data science
- Business analytics, data optimization and data quality
- In-company Python and Power BI trainings
- In-company prompt engineering training
- Cross-sectoral projects in the field of business analytics



# *We work for social-cultural relevant organisations:*



**Health care**



**Media & culture**



**Education**



**Pension funds**



**Insurance**



**Governments**

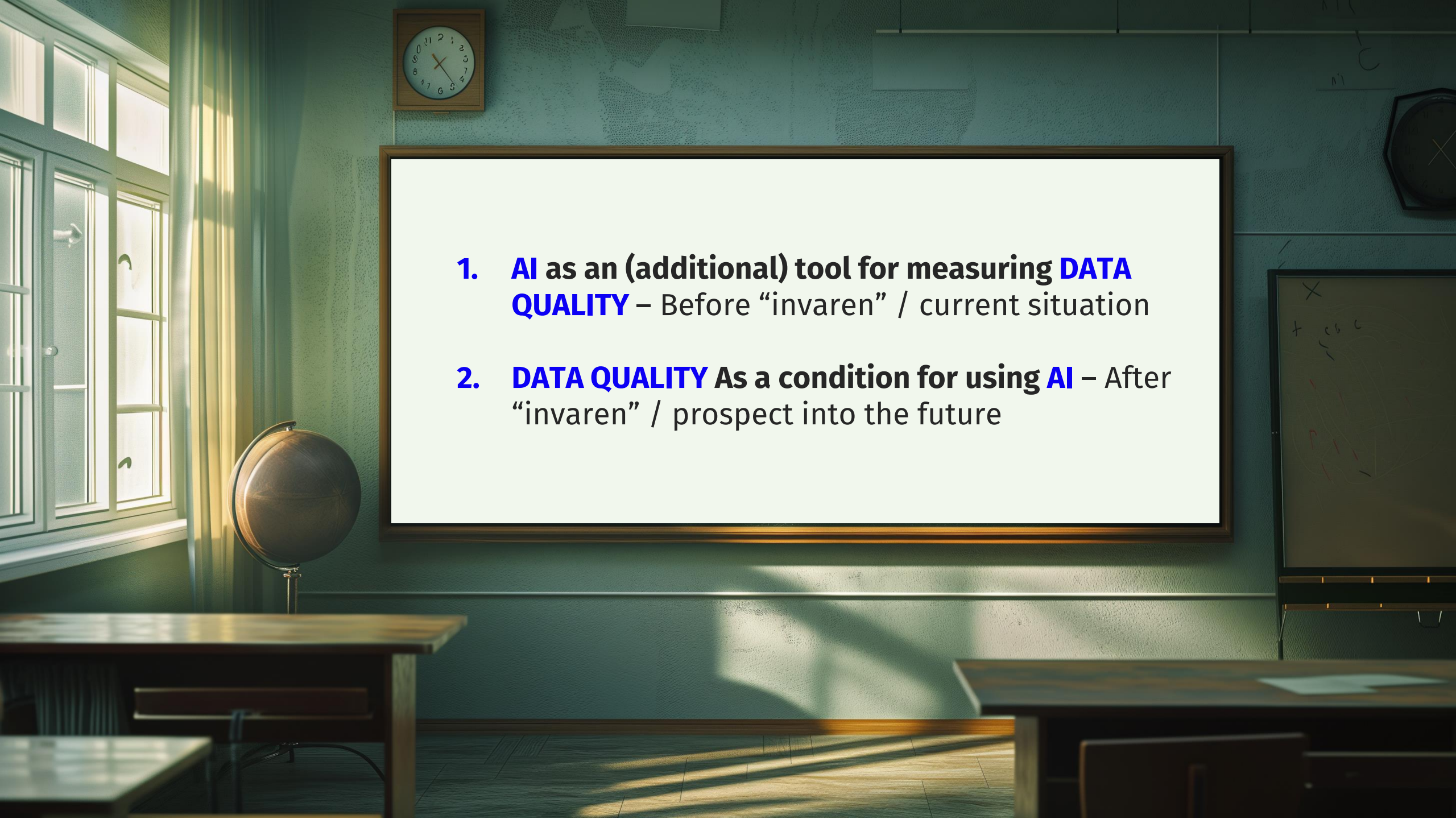
***Data Science at the heart of our services, cross-sectoral!***





**The relationship between Data Quality and Artificial Intelligence**  
*In the light of the new pension scheme in the Netherlands*



- 
- A classroom scene with a large screen displaying text. The screen is the central focus, showing two numbered points. To the left of the screen is a window with sunlight streaming in, a globe on a stand, and a round wall clock. To the right is a whiteboard with some faint markings. The room has a greenish-blue wall and wooden desks in the foreground.
1. **AI as an (additional) tool for measuring DATA QUALITY** – Before “invaren” / current situation
  2. **DATA QUALITY As a condition for using AI** – After “invaren” / prospect into the future



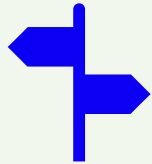
**AI**

**as an (additional) tool for  
measuring**

**DATA  
QUALITY**



# “Kader datakwaliteit” by Pensioenfederatie



Board sets the guidelines  
(risicobereidheid,  
correctie/herzieningenbeleid,  
kritische data-elementen,  
maximale toegestane afwijking)

**Fase 1: Opzet datakwaliteit**



Performing risk analyses  
and data quality  
measures

**Fase 2: Risico-inventarisatie en beoordeling**

**Fase 3: Data-analyses en deelwaarnemingen**

**Fase 4: Rapportage en beoordeling**



Check proces via Agreed Upon  
Procedures  
(by external accountant or IT-  
auditor)

**Fase 5: Overeengekomen specifieke  
werkzaamheden door externe accountant  
of IT-auditor**



The board decides  
about the data quality  
before “invaren”

**Fase 6: Besluit over datakwaliteit vóór  
invaren**

# Intermezzo: Understanding data?

## What is the concept “data” for a pension fund?

**First layer:** Data as concept and discipline. The word data is used as summary and has a lot of meanings.

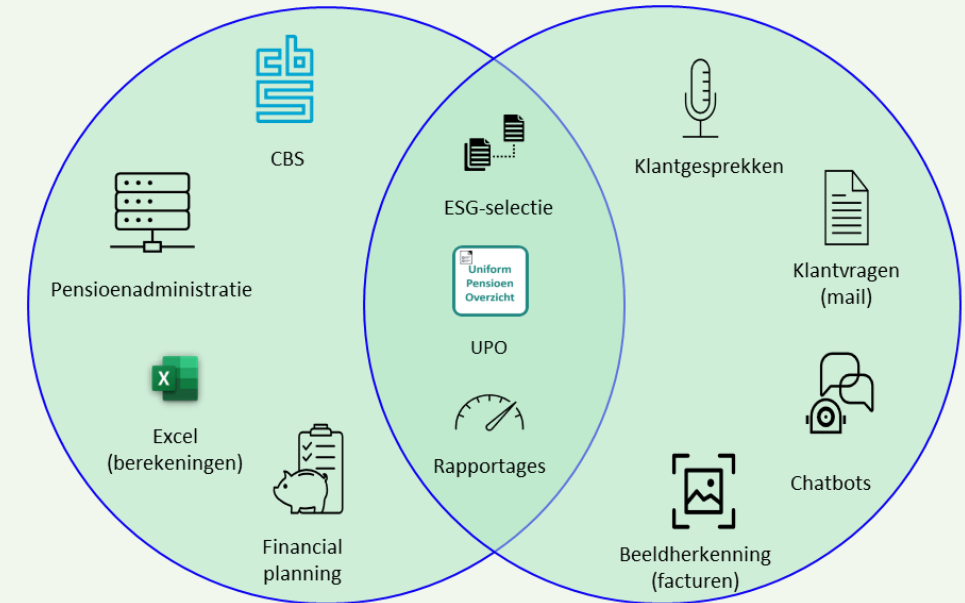
**Second layer:** The discipline data has some underlying themes and domains. As example: data management, data science and data tooling. These are all topics within the discipline data, with each their own depth.

**Third layer:** In this layer, it is all about specific applications within a theme. For example: data quality research and data controls within the data management domain.

**Fourth layer:** The level of the data fields, where data like gender, date of birth and salary are registered.



*“Data refers to factual information or data that is gathered, measured and reported. It can consist out of numbers, text, images, audio or other forms of information that can be used to gather insights, draw conclusions and make decisions. Data is the source that is used to generate information and knowledge.”*





# Intermezzo: Understanding Data Quality



*“Data Quality is important, because it leads to better decisions, trust in data, efficiency, innovation, customer satisfaction and compliance with laws. It improves analyses, saves costs and prevents errors.”*

*“De quality of the data of a pension fund is determined by the completeness, accuracy and suitability of the data.”*

*– DNB Good Practices 2017*

*“Quality is defined as the degree to which dimensions of a data concept meet requirements.”*

*– DAMA Dimensions of Data Quality*

*“Artificial intelligence (AI) is technology that enables computers and digital devices to learn, read, write, talk, see, create, play, analyze, make recommendations, and do other things humans do.” - IBM*



$$Y_i = \beta_0 + \beta_1 X_i$$

Diagram illustrating the components of a linear regression equation:

- $Y_i$  is labeled as the **Dependent Variable**.
- $\beta_0$  is labeled as the **Constant/Intercept**.
- $\beta_1$  is labeled as the **Slope/Coefficient**.
- $X_i$  is labeled as the **Independent Variable**.

Search

Sophia Ciocca

MADE FOR SOPHIA

## Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your favourites!

Made for Sophia Ciocca by Spotify - 30 songs, 2 hr 3 min

PLAY FOLLOWING ...

FOLLOWER 1

Filter Download

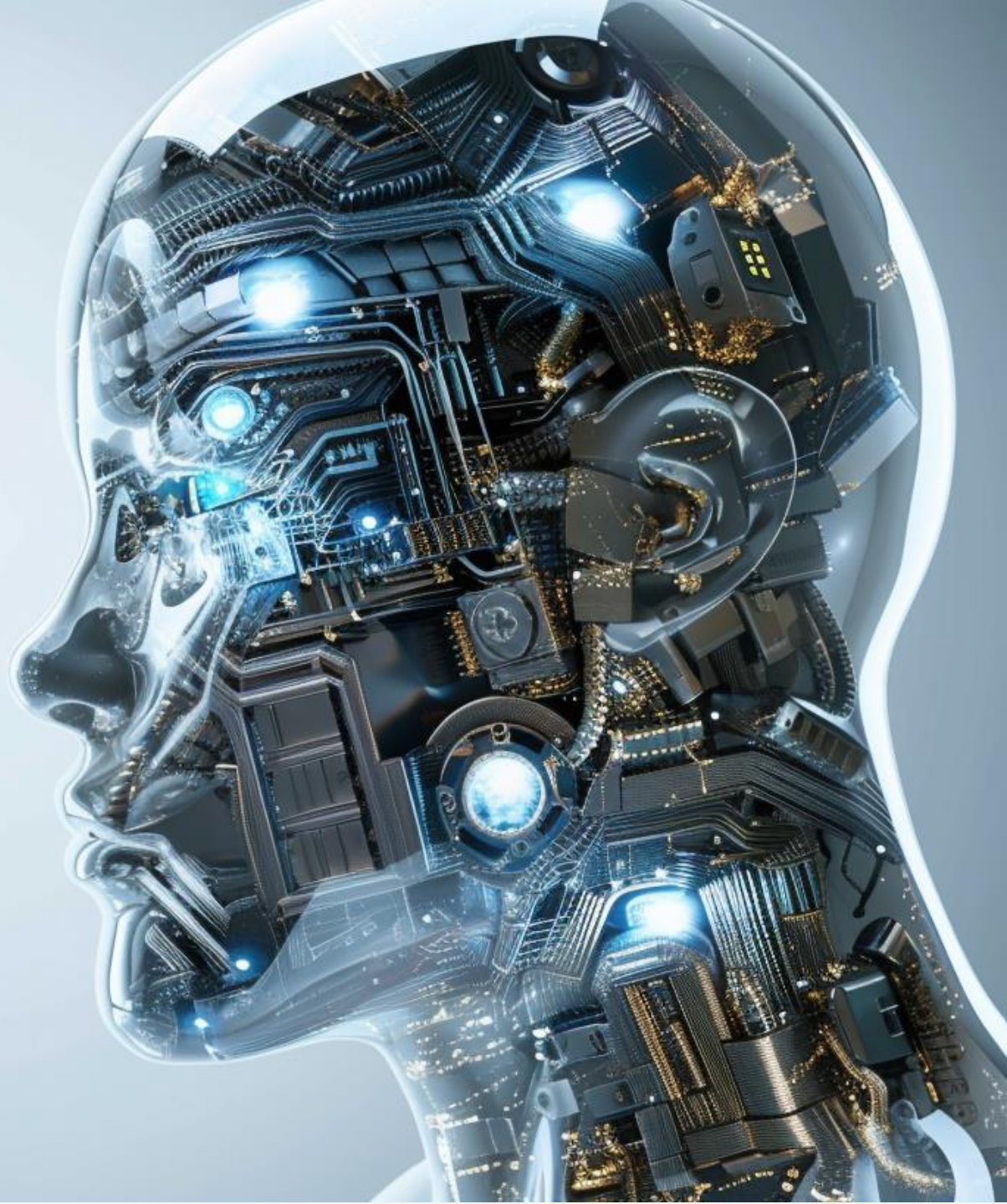
## Suggestion for data science and AI methods in “Kader Datakwaliteit”

### Suggestie inzet van data science technieken

In deze fase kan de pensioenuitvoerder gebruik maken van data science technieken om de analyses op efficiënte wijze uit te voeren. Bijvoorbeeld door via multivariate analyse outliers te detecteren die via business rules of queries minder snel te detecteren zijn. Of om via machine learning technieken te onderzoeken in welke mate de waarde van KDE's zich laat voorspellen uit (combinaties van) andere KDE's. Door de deelnemers waarvoor deze voorspellende waarde (meer dan een gekozen grenswaarde) afwijkt van de overige deelnemers als outliers te kiezen kunnen risicovolle posten worden onderkend die nader onderzocht kunnen worden.

***“In this phase (3), pension funds can use data science techniques to perform efficient analyses.”***





***Human Intelligence and Artificial Intelligence reinforce each other***

## Fase 2 Risico-inventarisatie en -beoordeling

### 2.1 Risico-inventarisatie

#### 2.1.1 Profiel pensioenuitvoerder

- a. Profiel en kenmerken pensioenuitvoerder
- b. Datakwaliteitsbeheersingsraamwerk en datastromen in processen
- c. Events
- d. Incidenten en klachten
- e. Tijdshorizon

#### 2.1.2 Profiel deelnemers

- a. Deelnemersrisico-indicatoren (DRI's)
- b. Combinaties DRI's
- c. Risicogroepen vaststellen

### 2.2 Risicobeoordeling

- a. Risico-beschrijving
- b. Bruto risico
- c. Aanwezige beheersmaatregelen
- d. Netto risico
- e. Maximaal toegestane afwijking (MTA)

### 2.3 Vaststellen aanvullende activiteiten

- a. Aanvullende data-analyses/deelwaarnemingen
- b. Aanvullende beheersmaatregelen (stay clean)

## Fase 3 Data-analyses en deelwaarnemingen

### 3.1 Data profiling

- a. Compleetheid
- b. Juiste typering
- c. Binnen domeinwaarden
- d. Distributiekennmerken

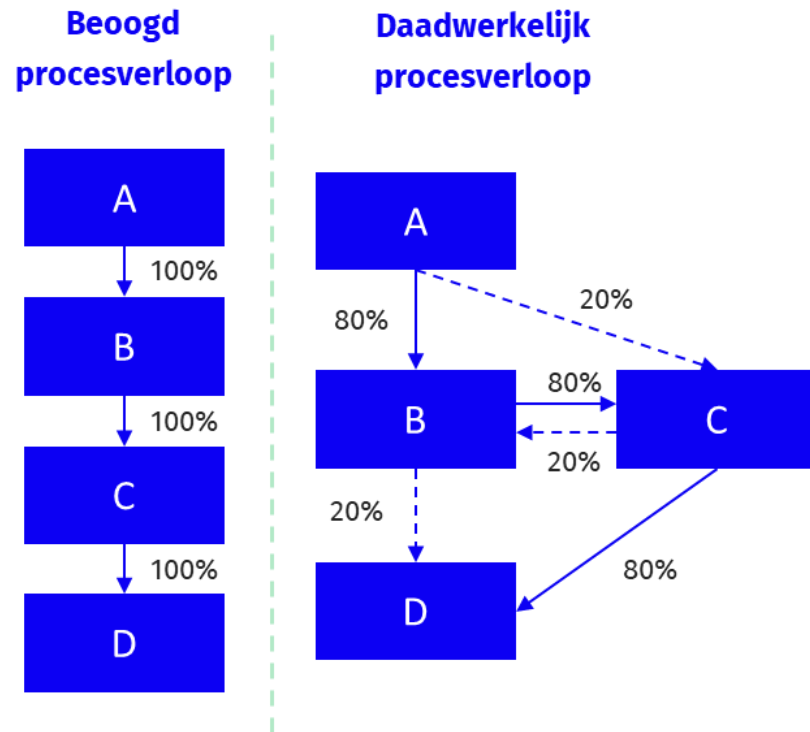
### 3.2 Data-analyse

- a. Generiek
- b. Specifiek

### 3.3 Deelwaarnemingen

- a. Outliers
- b. Risicogroepen

- Identification of potential risks is based on human knowledge about historic events and incidents.
- Risk assessment is subjective.
- Data-analyses are based on best practices and domain knowledge of professionals
- Let's add some AI!



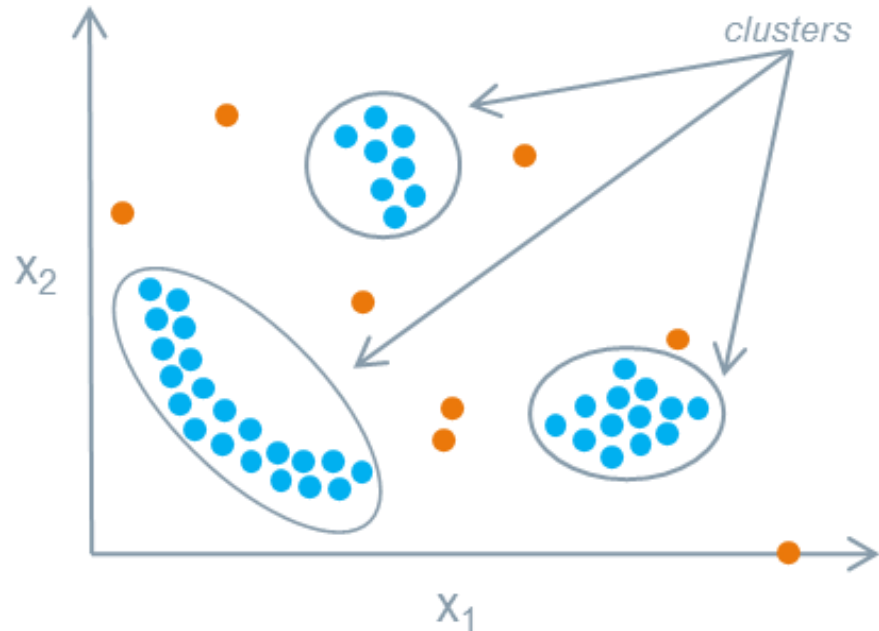
- Expected process flow versus real process flow
- Detection of wrong flows of processes
- Can be a helpful tool to detect potential risks in the administration system
- With regards to phase 2: identification of potential data quality risks, in a data-driven manner.



# Outlier detection (clustering)

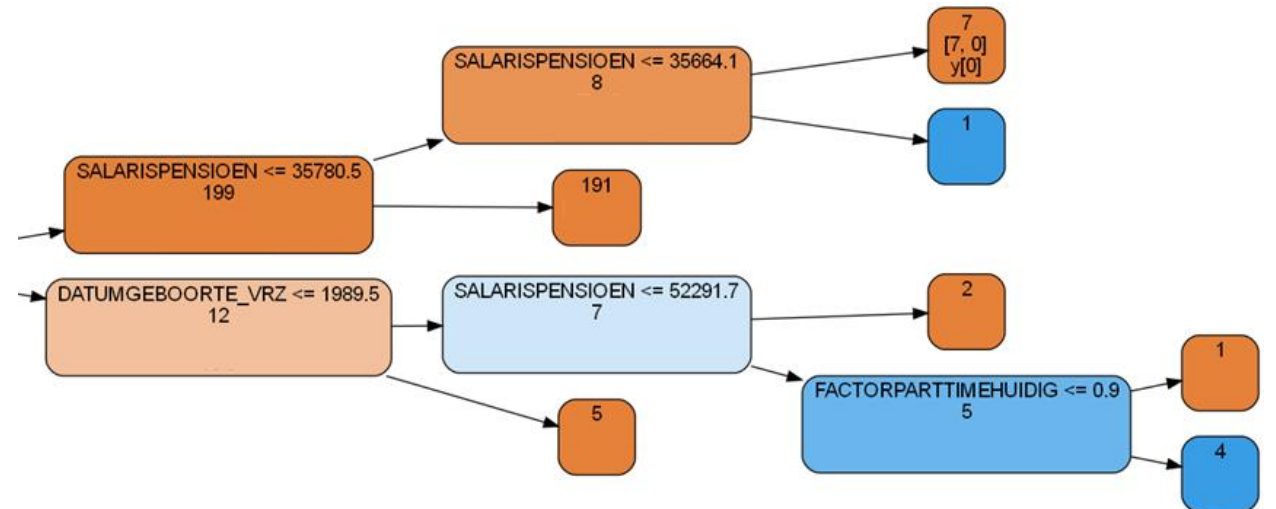
1

Outlier detectie



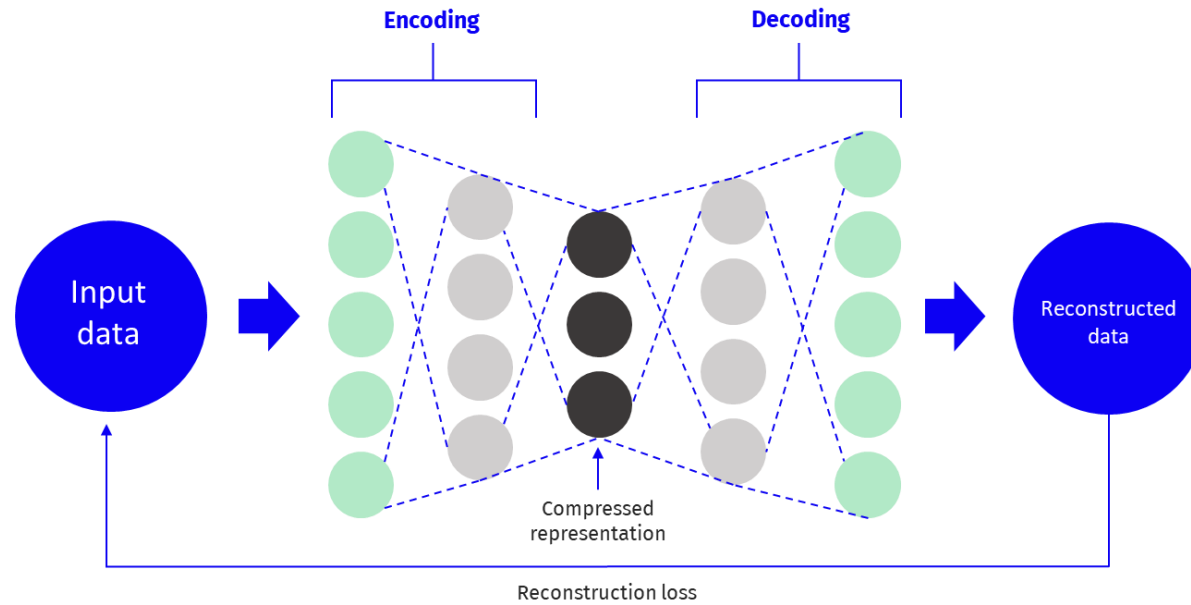
2

Reverse engineering

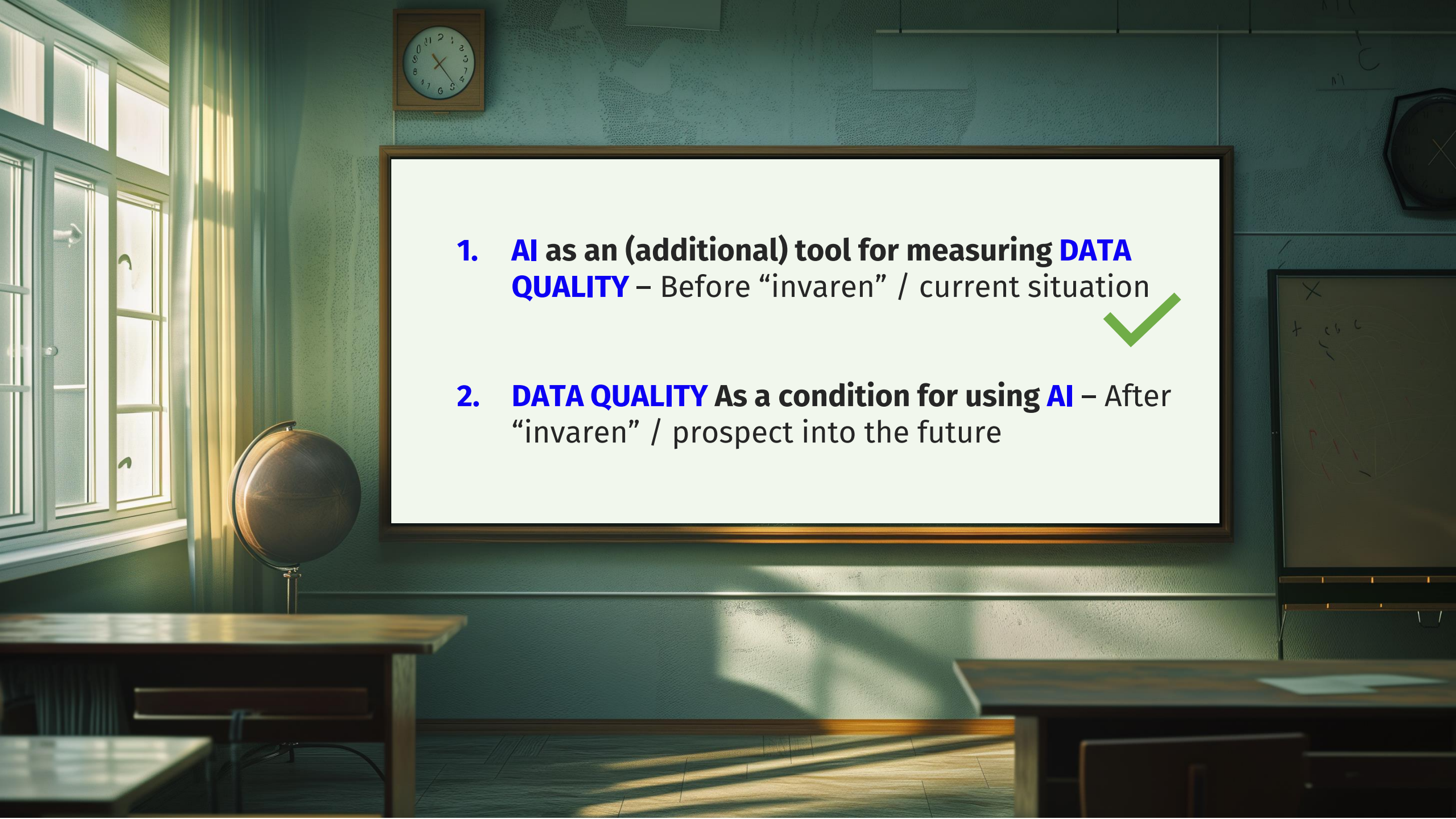


- Detects outliers in data, based on multiple characteristics of the participants in the administration, and explains why the participant is seen as an outlier.
- Follow-up: use anomalies as “risk groups” for further research or use it as data-driven plausible checks.

## Outlier detection (expectation vs. reality)



- Deep learning model to detect outliers in data.
- Follow-up: use anomalies as “risk groups” for further research or use it as data-driven plausible checks.

- 
1. **AI as an (additional) tool for measuring DATA QUALITY** – Before “invaren” / current situation ✓
  2. **DATA QUALITY As a condition for using AI** – After “invaren” / prospect into the future



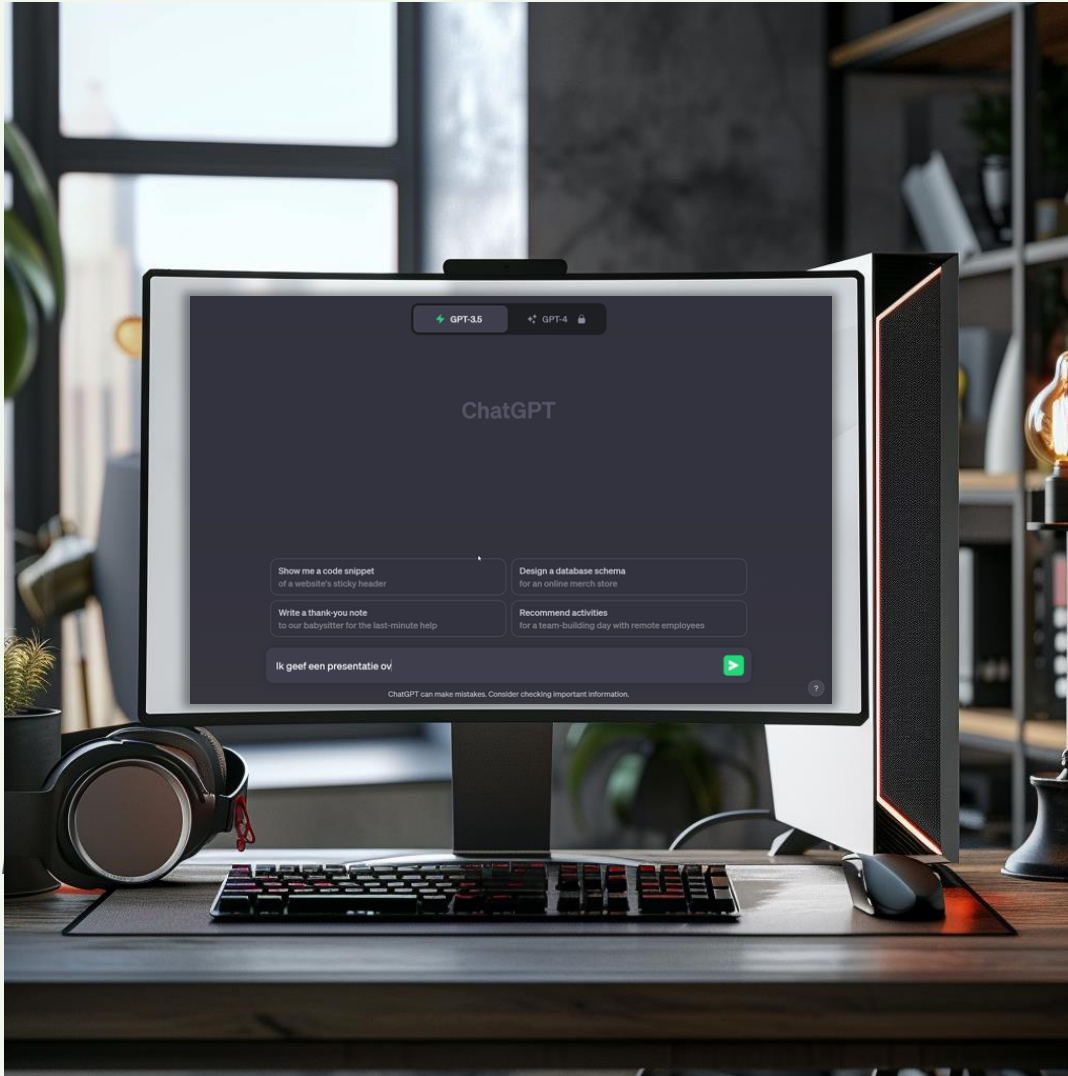


# DATA QUALITY

**As a condition for using**

# AI

# Rise of language models and chatGPT of OpenAI



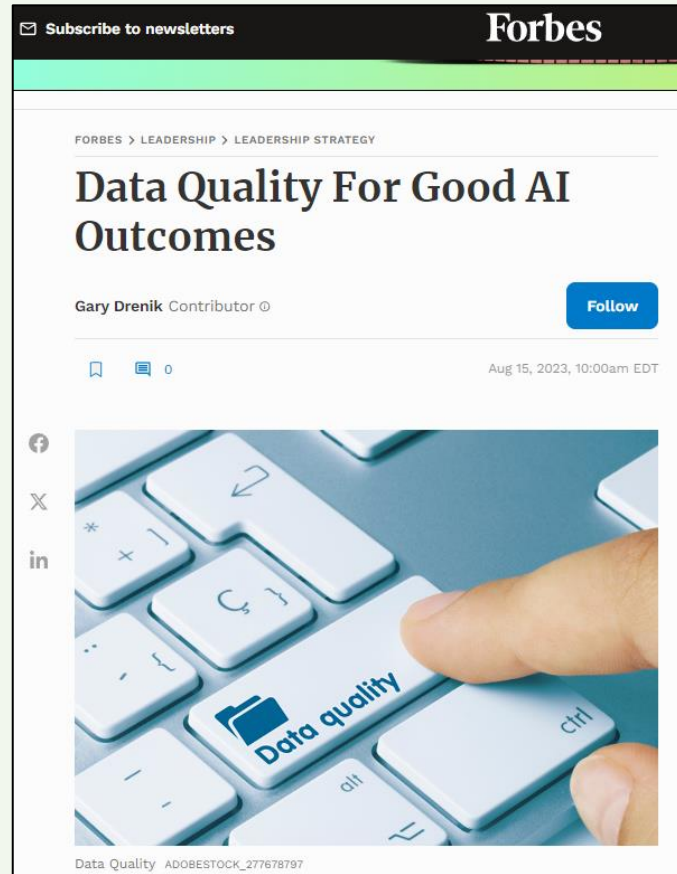
## De technique

- **NLP: Natural Language Processing**
- **LLM: Large Language Models**
- **GPTs: Generative Pre-trained Transformers**

In Italy I prefer to eat... **Pizza (90%)**  
In Italy I prefer to eat ... **French Fries (5%)**  
In Italy I prefer to eat ... ... **(5%)**

- Multimodal learning
- New skills: prompt engineering, both implementation as creativity to ask questions

# Data quality is essential for the correct functioning of AI



Artificial Intelligence

## Slechte data zijn als junkfood voor je AI

## Onderzoek: 'Data kwaliteit een van de grootste uitdagingen voor adoptie AI'

Nieuws april 22, 2020

<https://informatieprofessional.nl/onderzoek-datakwaliteit-een-van-de-grootste-uitdagingen-voor-adoptie-ai/>  
<https://www.salesforce.com/nl/blog/data-centric-ai/>  
<https://www.forbes.com/sites/garydrenik/2023/08/15/data-quality-for-good-ai-outcomes/?sh=6ab87a441184>




# The rapid development of AI



4 sept 12:17

## AI ontketent een enorm gevecht om data

 The Economist

Bouwers van AI-modellen zijn op zoek naar nieuwe, betere bronnen om hun razende honger naar data te stillen. Ondertussen bekijken bedrijven die over enorme hoeveelheden data beschikken hoe ze die het best te gelde kunnen maken.

Bronnen: FD

- No more training of your own models, but using pre-trained large models of large tech companies (especially for unstructured data)
- The opportunities seem endless, computing power is present. But is there a shortage of **high quality data**?

NOS Nieuws • Zaterdag 1 juli, 20:46 • Aangepast zaterdag 1 juli, 22:12



## Twitter beperkt aantal tweets dat je per dag mag lezen

Twitter heeft een limiet gesteld aan het aantal tweets dat gebruikers mogen lezen. Volgens Twitter-baas Elon Musk is dat nodig om het platform goed te laten draaien en gaat het om "een tijdelijke noodmaatregel".

Volgens Musk is het zogenoemde *scraping* een groot probleem voor Twitter.

Hierbij gebruiken bedrijven software om geautomatiseerd data te vergaren op Twitter voor het ontwikkelen van modellen voor kunstmatige intelligentie. Dat kan er volgens de topman toe leiden dat Twitter trager of onbetrouwbaarder wordt.

Bronnen: NOS en FD

- Tech companies that develop language models, are trying to corporate with organisations that have a lot of data available.
- Having **data of high quality** gives you a competitive advantage.

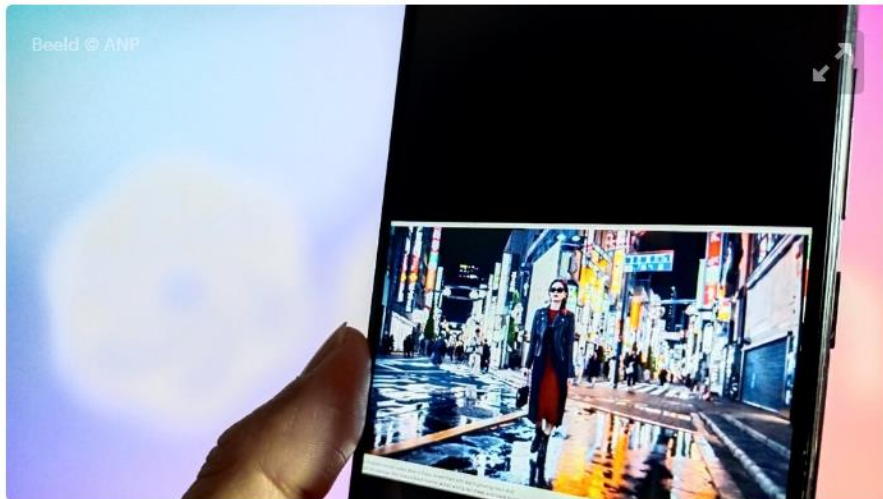


## Zorgen om AI van Sora: 'Golf aan desinformatie komt op ons af'



Door Harm Teunis

20 februari 2024 12:14 • Aangepast 20 februari 2024 12:14



Een stijlvolle vrouw die wandelt door Tokio, terwijl neonlicht weerspiegeld wordt door plassen op de grond. Het is een van de video's die ChatGPT-maker OpenAI online heeft gezet. De video is gemaakt door kunstmatige intelligentie met een nieuw programma dat Sora heet. Knap, zeker. Maar niet zonder gevaar, waarschuwen experts. "Het was eerst best ingewikkeld om te maken. Maar de rem die er was, is er nu af."



### Net bi

13:25 DN  
hoe

13:22 Aja  
ach  
Led

13:16 Bos  
neg

12:59 Mir  
Ord  
doc

12:57 Arr  
sto  
'be

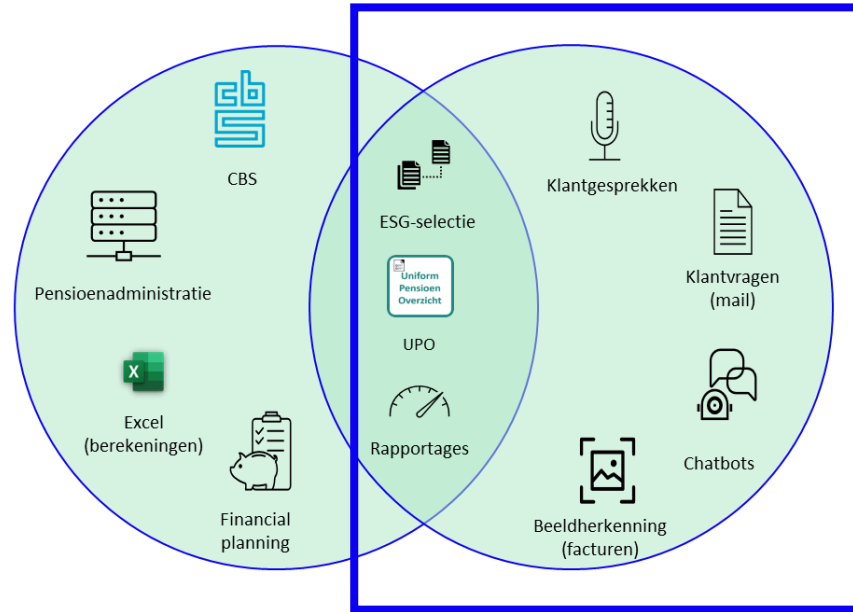
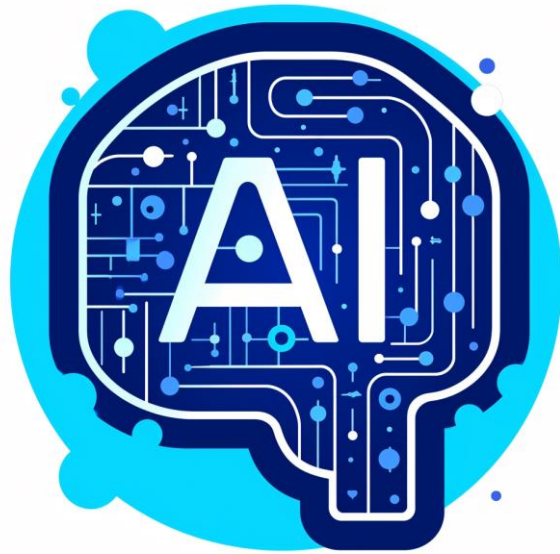
- The rise of AI models leads to potential increase of fake news
- Data Quality of outputs of these models more and more important
- Data Quality definition gets even more dimensions:
  - **Credibility:** The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.
  - **Lineage:** Measures whether factual documentation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration.

<https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf>





# Enormous potential in the pension sector (now and after “invaren”)



Breakthrough in **AI: NLP, LLM and GPT's** and fast improvements in these fields

A lot of **available unstructured data** that was out of scope before

**Data Quality** gets a broader scope, both in types of data as in definition

# Use case: Searching in the huge amounts of data of pension funds

**WERKEN AAN ONZ PENSIOEN**

Home

**Filters**

- Type
- Document
- Geavanceerd

**13 resultaten gevonden**

Selecteer type  Selecteer document

**Algemene toelichting**

**2. Besluit toekomst pensioenen**

Gevonden tekst:  
Datakwaliteit  
Dit besluit stelt hiernaast enkele specifieke eisen met betrekking tot de datakwaliteit indien een pensioenfonds besluit tot een collectieve waardeoverdracht van de opgebouwde pensioenaanspraken en pensioenrechten. In dat geval moet het pensioenfonds verantwoorden dat de datakwaliteit voldoende is op de transitieperiode. Het implementatieplan geeft het pensioenfonds aan welke werkzaamheden in het transitieproces doorgevoerd worden.

**WERKEN AAN ONZ PENSIOEN**

**MAATI**  
Zoeken met AI

Hallo, ik ben MAATI

Ik ben een AI-powered virtuele assistent aan wie u vragen over de transitie naar het nieuwe pensioenstelsel kunt stellen. Mijn antwoorden baseer ik op de informatie van het informatieplatform Werkenaanonspensioen.nl, de Wet toekomst pensioenen en bijbehorende lagere regelgeving. Ik kan u het beste helpen als u uw vragen duidelijk en volledig formuleert. Een generiekere vraag krijgt doorgaans ook een generiek antwoord. Het helpt om daarbij te denken in termen van 'wat', 'hoe', 'waarom', 'wie' vragen.

Weet u niet waar u moet beginnen? Probeer dan een van de onderstaande zoekopdrachten:

- Wat houdt het hoorrecht in?
- Welke vereisten zijn er voor het risicopreferentieonderzoek?
- Welke premieovereenkomsten mogen verzekeraars aanbieden?
- Wat is de rol van het verantwoordingsorgaan bij inwaven?
- Waar kan de solidariteitsreserve voor ingezet worden?

Waar kan de solidariteitsreserve voor ingezet worden?

De solidariteitsreserve kan voor verschillende doelen worden ingezet. Het pensioenfonds en de sociale partners bepalen gezamenlijk hoe de solidariteitsreserve wordt gebruikt. Enkele mogelijke doelen zijn het verbeteren van het collectieve beleggingsresultaat door het vergroten van de beleggingshorizon, het delen van bepaalde risico's tussen generaties en het aanvullen van pensioenvermogens en uitkeringen. Het precieze gebruik van de solidariteitsreserve wordt vastgelegd in de afspraken tussen het pensioenfonds en de sociale partners, zoals de uitvoeringsovereenkomst en het pensioenreglement.

Vond u dit antwoord nuttig?

Waar kan de solidariteitsreserve voor ingezet worden?

Disclaimer

[https://zoeken.werkenaanonspensioen.nl/waop\\_search/](https://zoeken.werkenaanonspensioen.nl/waop_search/)

- A lot of unstructured data is available.
- New pension scheme: more focus on individual participant.
- Participants are getting used to a specific level of service, outside the pension sector.
- Developments of techniques give new opportunities to search in huge amounts of unstructured data.
- And how about chatting with your unstructured, but also with your structured data...?



# Use cases in practice





# The definition of Data Quality evolves

The technique is present and growing, but... People are needed to insure responsible AI, including quality checks:

- Insuring **data quality** of input data and output data
- Asking the **right** question
- Applying the **right** techniques and models for a specific business question
- Preparation of datasets: insure a **balanced** dataset (bias models)
- Set up data quality policies, data management policy or AI policy
- Interpreting models and results, **quality controls** and detecting fake news / wrong outputs.

**These are competences suited for...?**



**Questions?**

# Thank you!



Remain Relevant!

**Tell us your story!**

**Aron Jeurinck**, relevantieadviseur at AethiQs

**Joost Oprel**, relevantieadviseur at AethiQs

E: [Aron.Jeurinck@AethiQs.nl](mailto:Aron.Jeurinck@AethiQs.nl)

T: +31 6 114 32 617

[www.AethiQs.nl](http://www.AethiQs.nl)