Verbeteren data kwaliteit en informatie kwaliteit

Een data science perspectief

SR Dielbandhoesing
5/10/23

Garbage *in* Garbage **OUT**

Source: solutions.ait.ac.th/garbage-in-garbage-out/

“Waste Management Consultant”

Wat is de ‘garbage’

- Missende waarden(dependent variabele)
- Dubbele waarden
- Anomaliteiten
- Inconsistentie
- Kwaliteit (pixels)
- Missende metadata
  - Bron
  - Doel
  - Vertrouwelijkheid
- Interoperabiliteit (character-sets, datum naar rekengetallen)
- Uniciteit (wat definieert een record uniek)

Voorbeelden waar modellen hele andere uitkomsten geven bij elk van de kwaliteit problemen

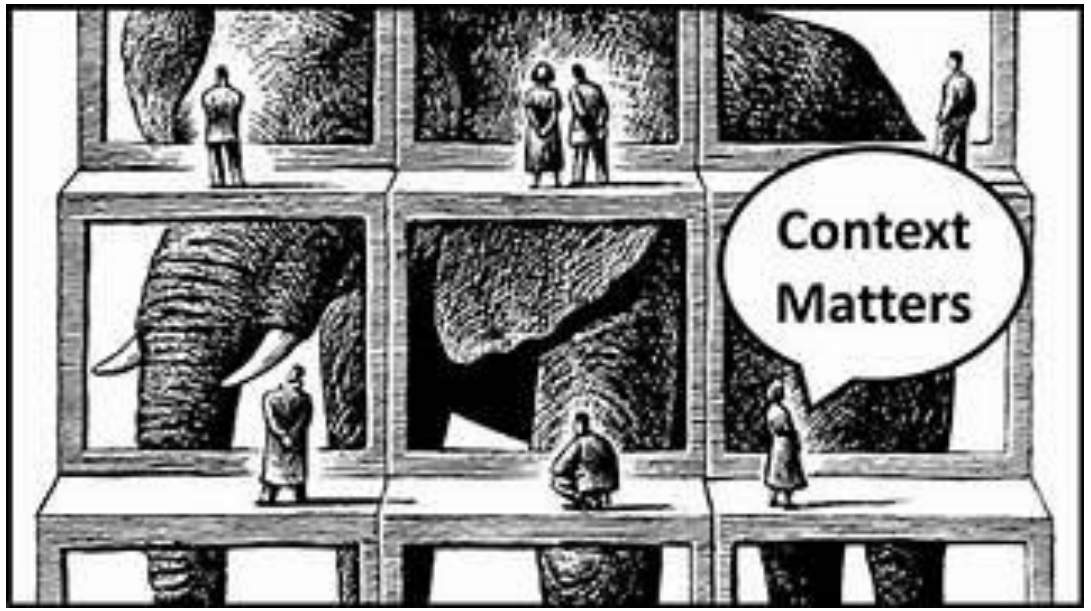
Voorbeelden hoe AI kan helpen

Data kwaliteit en AI

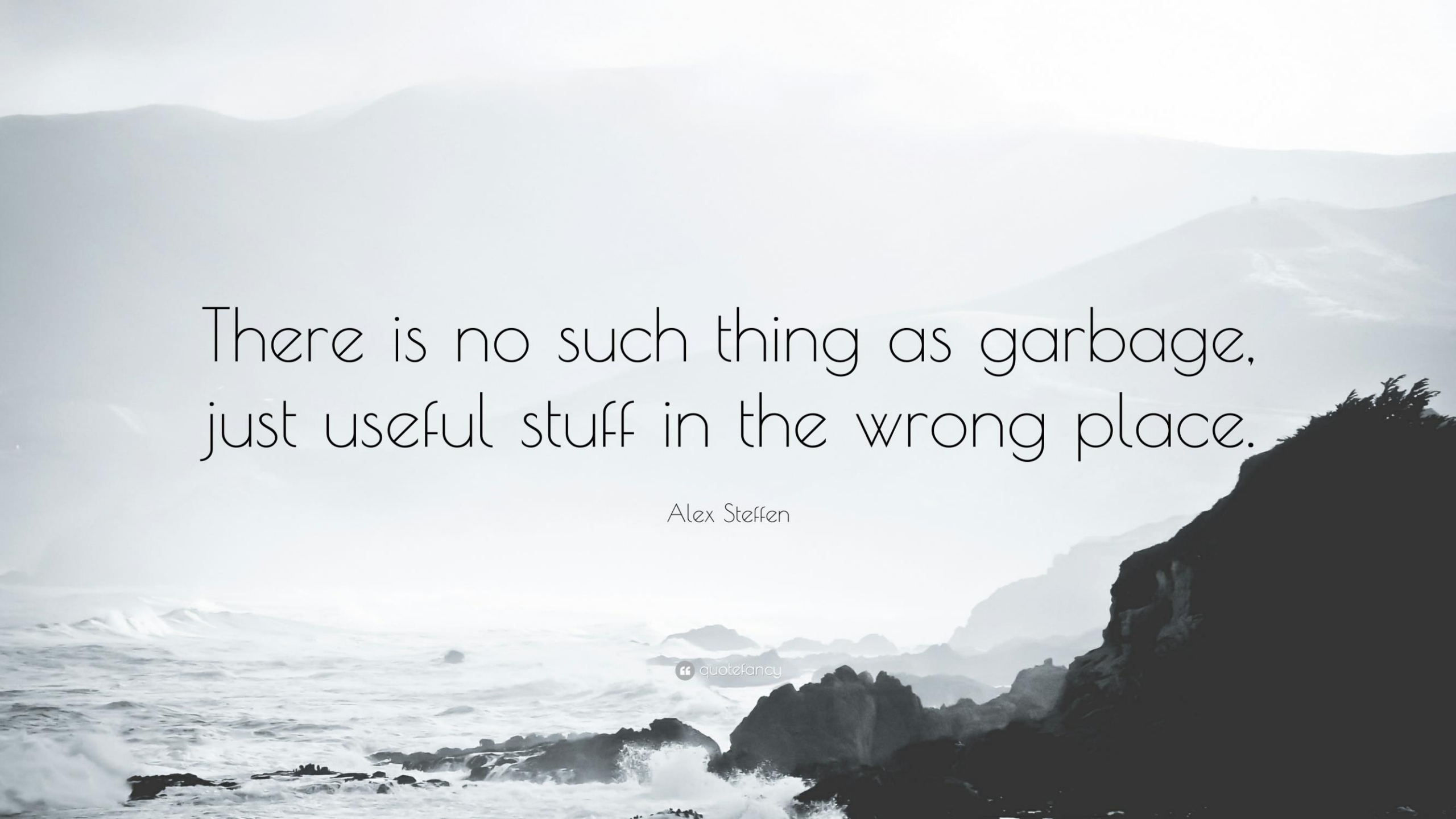
- Metadata
  - Algoritmen: data is de uitkomst van een algoritme (voorspellende waarde, generatieve algoritme)
- Privacy gevoeligheid → Anonimiseren

Data kwaliteit en Ethiek

- Open toegang → Open data beleid
- Bias: voorbeeld omgaan met missing values
  - **Confirmation bias**: zelfde profiel → missing vervangen door gemiddelde
  - **Ingroup bias**: dominante heeft het voor het zeggen → missing vervangen door modus
  - **Selection bias**: → missende geslacht
  - **Cobra effect**: oplossing maakt probleem groter → missende invullen ipv verwijderen


Context
Matters


CONTENT
CONTEXT


There is no such thing as garbage,
just useful stuff in the wrong place.
Alex Steffen
quotefancy