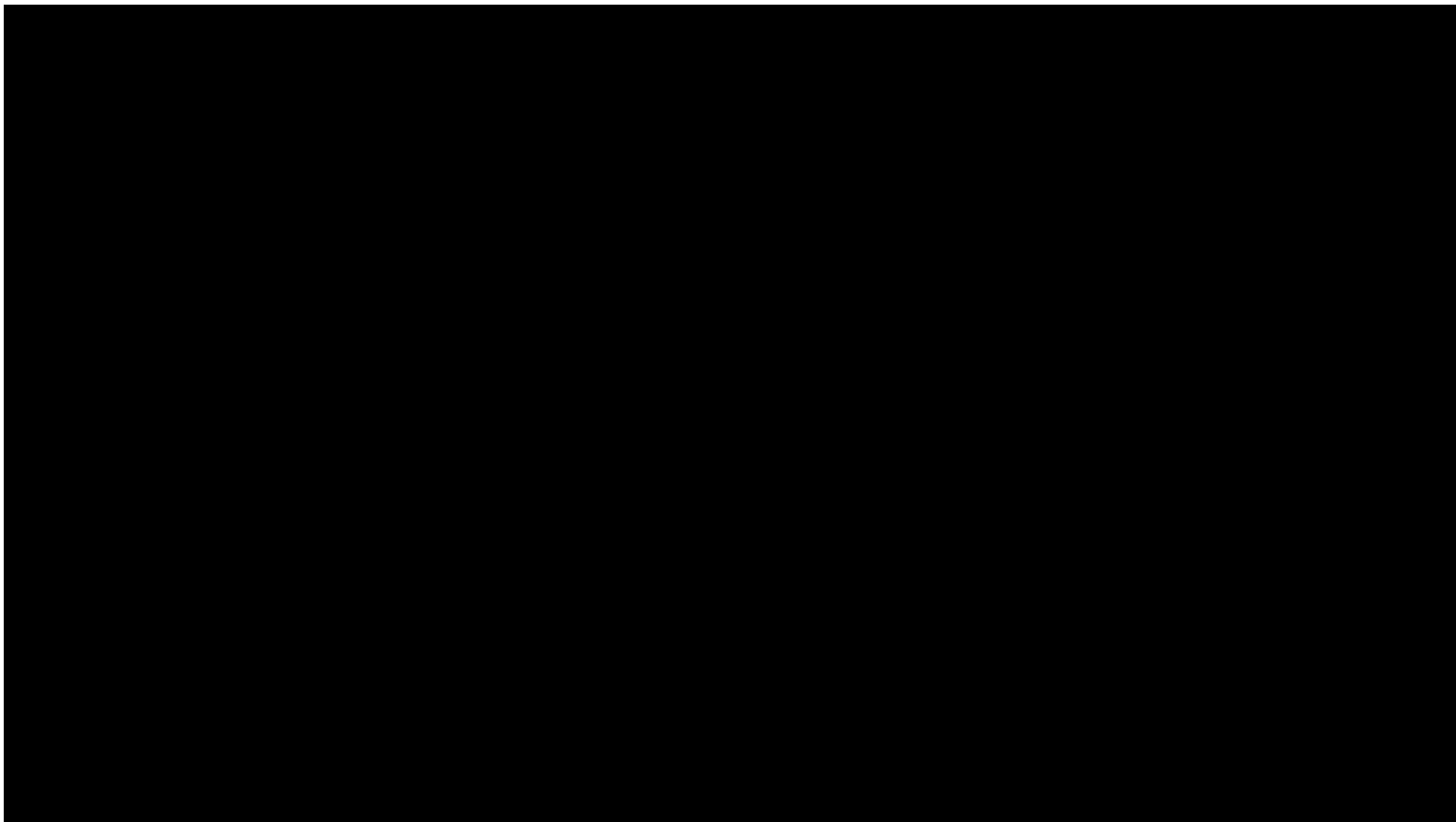




Data Collection

JENS BAETENS





Primary	Secondary
Zelf verzamelde data door enquêtes, logs, ...	Gebruik van reeds bestaande datasets
Specifiek voor je probleem	Meer generieke data
Veronderstellingen over welke data belangrijk is	Data die niet aanwezig is kan niet gebruikt worden
<div> <div></div> <div>kan lang duren en veel geld kosten</div> </div>	<div> <div></div> <div>Zeer snel om mee aan de slag te gaan</div> </div>

Meer Overheadkosten

Geen opstartkosten

Quantitatieve data	Qualitatieve data
Numerieke waarden (leeftijd, gewicht, ...)	Niet-numerieke waarden om eigenschappen, meningen gevoelens te beschrijven. <i>tekst</i>
Statistische evaluatie mogelijk <i>describe()</i>	Clusteren of groeperen van gelijkaardige waarden <i>data</i>
Hoeveel heeft je wafel gekost?	Waarom heb je een wafel gekocht?
70% van de aanwezigen hebben een wafel gekocht	Het was een koude dag en mijn trein had vertraging.

Soorten Secondary data Sources

Bestaande publieke online datasets → *Kaggle / statal*

Gebruik maken van aangeboden API's → *twitter / big data*

Scraping of websites

Bestaande online datasets

<https://www.kaggle.com/>

<https://statbel.fgov.be/en/open-data> of andere overheden

Github bevat ook een reeks datasets

Google-Search

API's van Social media

Bijvoorbeeld voor Twitter kan je een API (Rest) gebruiken om:

- Verzamel alle tweets waarin je bedrijf voorkomt
- Verdeel de tweets in groepen: klachten, positieve review, ...
- Geef de geclassificeerde tweets door aan je klantendienst

Scraping of websites

Indien een bedrijf geen API aanbied om data op te vragen kan je ook rechtstreeks webpagina's gaan 'scrapen' of bestuderen.

Nadelen:

- Afhankelijk van de structuur van de te scrapen website
- Duurt langer om in te stellen
- Gevoelig aan wijzigingen van de website

→ smart scrapers bestaan

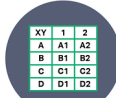
Soorten data

Structured Data

vs

Unstructured Data

Can be displayed
in rows, columns and
relational databases



XY	1	2
A	A1	A2
B	B1	B2
C	C1	C2
D	D1	D2

Numbers, dates
and strings



0, 1, 2,	DAY
3, 4, 5,	AUS
6, 7, 8,	4, 2025
9, 1	Y, 1
F+G-H,	

Estimated 20% of
enterprise data (Gartner)

20%

Requires less storage



Easier to manage
and protect with
legacy solutions



Cannot be displayed
in rows, columns and
relational databases



Images, audio, video,
word processing files,
e-mails, spreadsheets



Estimated 80% of
enterprise data (Gartner)

80%

Requires more storage



More difficult to
manage and protect
with legacy solutions



Data Science

→ machine learning

↳ big data