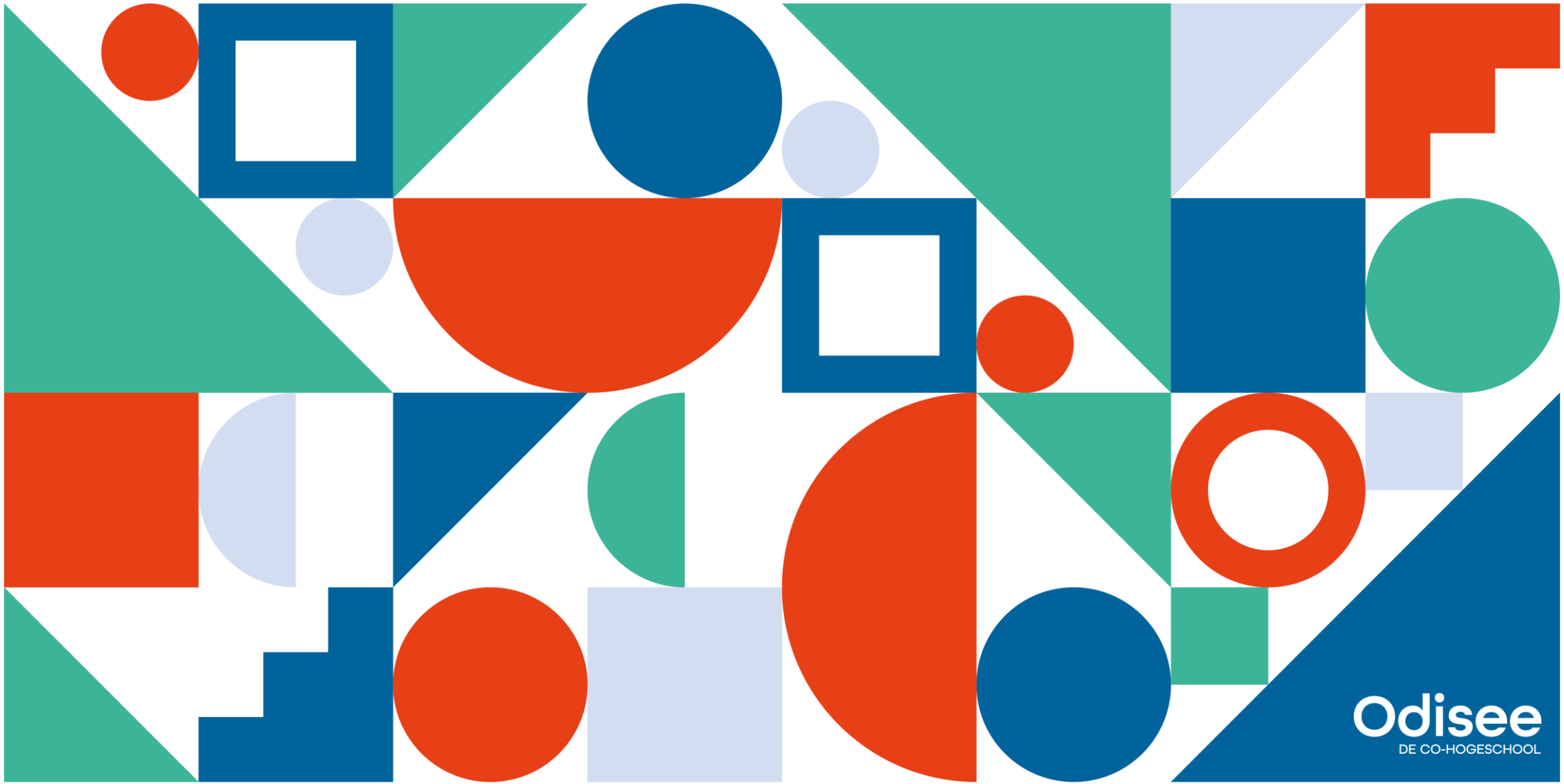
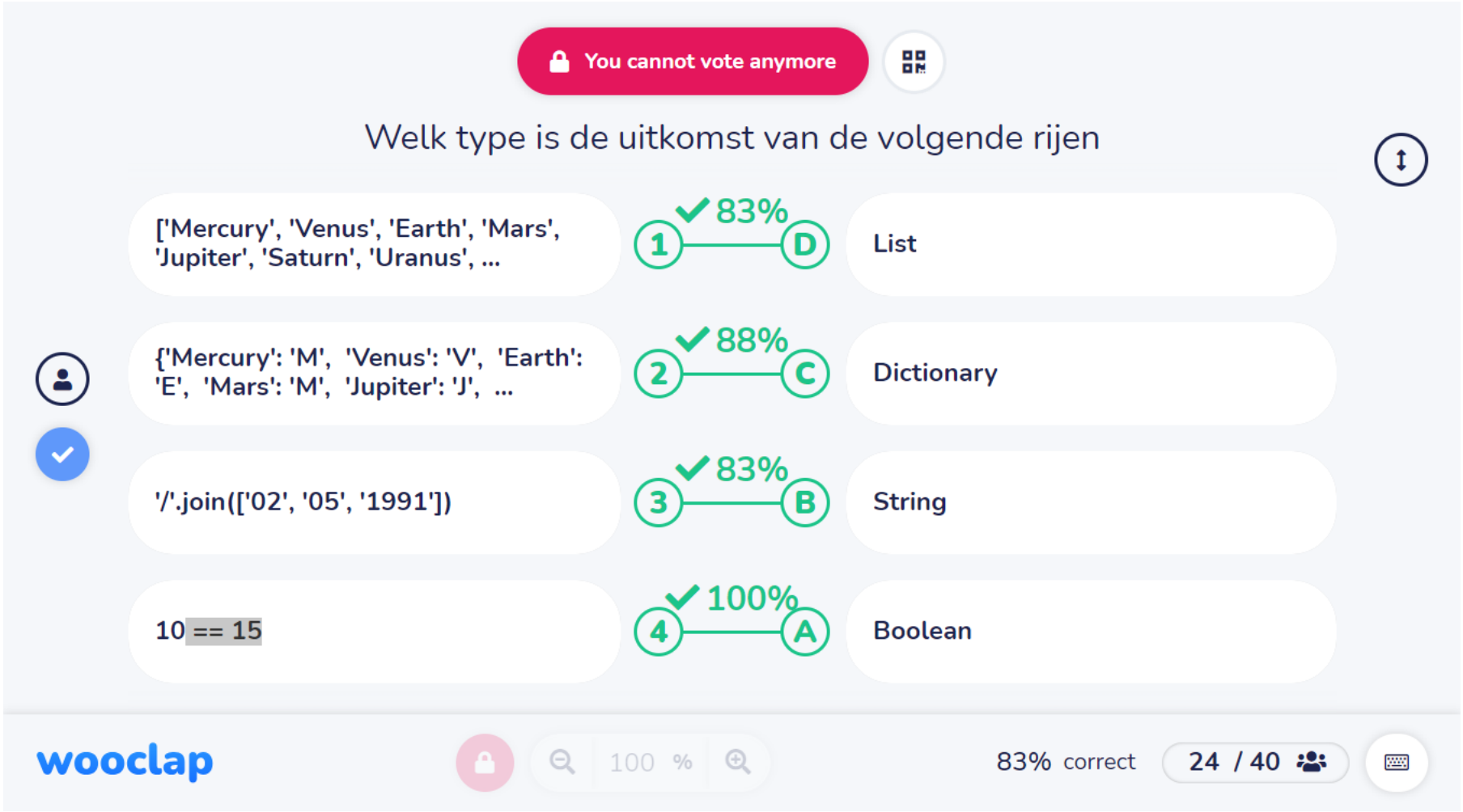

Odisee
DE CO-HOGESCHOOL

Data Science – week 2

Jens Baetens

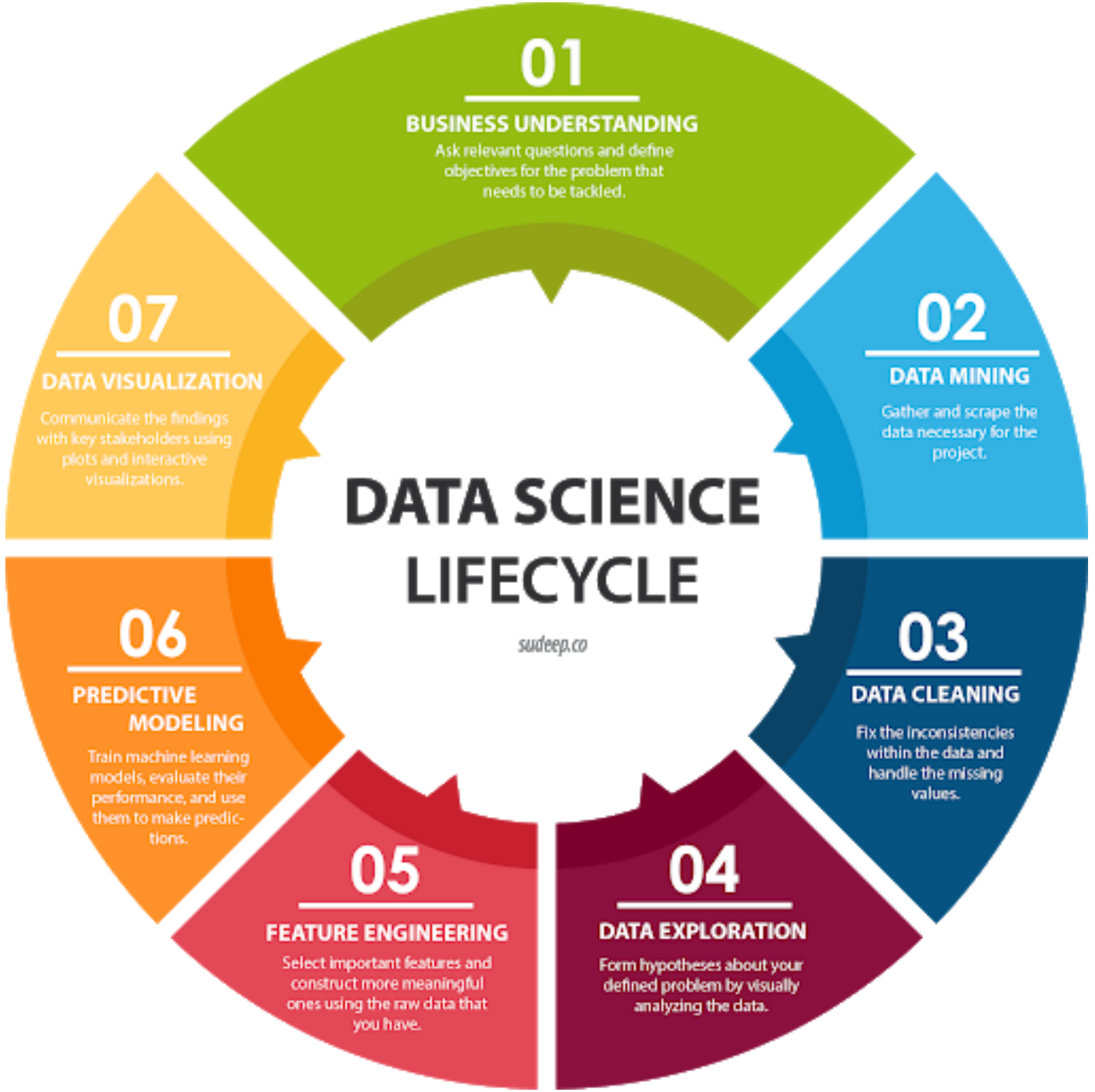

How to participate?
Click on the projected screen to start the question
Copy participation link
wooclap
100 %
40


Go to wooclap.com and use the code ZPKZMR
Wat is het verschil tussen een scripting taal zoals Python en een standaard..
Close this message
python code wordt uitgevoerd door een interpreter, een standaard taal niet
wooclap
100 %
21


You cannot vote anymore
Welk type is de uitkomst van de volgende rijen
['Mercury', 'Venus', 'Earth', 'Mars', 'Jupiter', 'Saturn', 'Uranus', ...
83%
1
D
List
{'Mercury': 'M', 'Venus': 'V', 'Earth': 'E', 'Mars': 'M', 'Jupiter': 'J', ...
88%
2
C
Dictionary
'/'.join(['02', '05', '1991'])
83%
3
B
String
10 == 15
100%
4
A
Boolean
wooclap
100 %
83% correct
24 / 40


Go to wooclap.com and use the code ZPKZMR
Wat is de output van het volgend stukje code:...
1
True
23%
5
2
False
77%
17
wooclap
100 %
22 / 42

Data science lifecycle


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

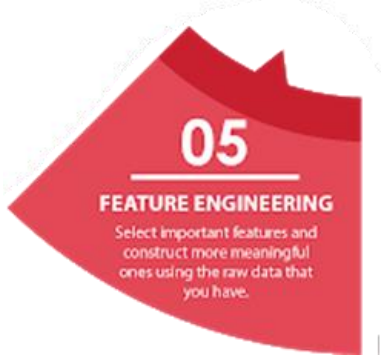
- Wat is de gestelde vraag of het probleem?
- Formuleer de vragen waarop een antwoord moet gevonden worden
- 5 soorten vragen:

Hoeveel?	Regressie
Wat is het?	Classificatie
Is het sterk gelijkend op?	Clustering
Is het vreemd?	Anomaly Detection
Welke optie is het beste?	Recommendation

- Verzamel data van verschillende bronnen
- Welke data is er nodig?
- Hoe geraak ik aan deze data?
  - Lokale databases
  - Scraping van webpaginas
  - Verzamelen van data van sensoren / apps / satellieten …
- Hoe bewaar ik de verzamelde data?

- Belangrijke stap voor betrouwbare resultaten te bekomen:
  - Garbage In -> Garbage Out
- Het doel is om problemen op te lossen in de datasets:
  - Ontbrekende data
  - Verkeerd gelabelde data (0/1 vs true/false)
  - Verschillende dataformaten (male/m/Male or dates)
  - Verbeteren van typos, vertalen van sommige velden, …

- Fase waarin je de verzamelde data bestudeerd
- Zoek naar bestaande patronen en controleer of er een bias aanwezig
- Visualiseer en analyseer deze patronen
- Detecteer outliers
- Stel een aantal hypotheses voor
- Ook exploratory data analysis genoemd: https://en.wikipedia.org/wiki/Exploratory_data_analysis

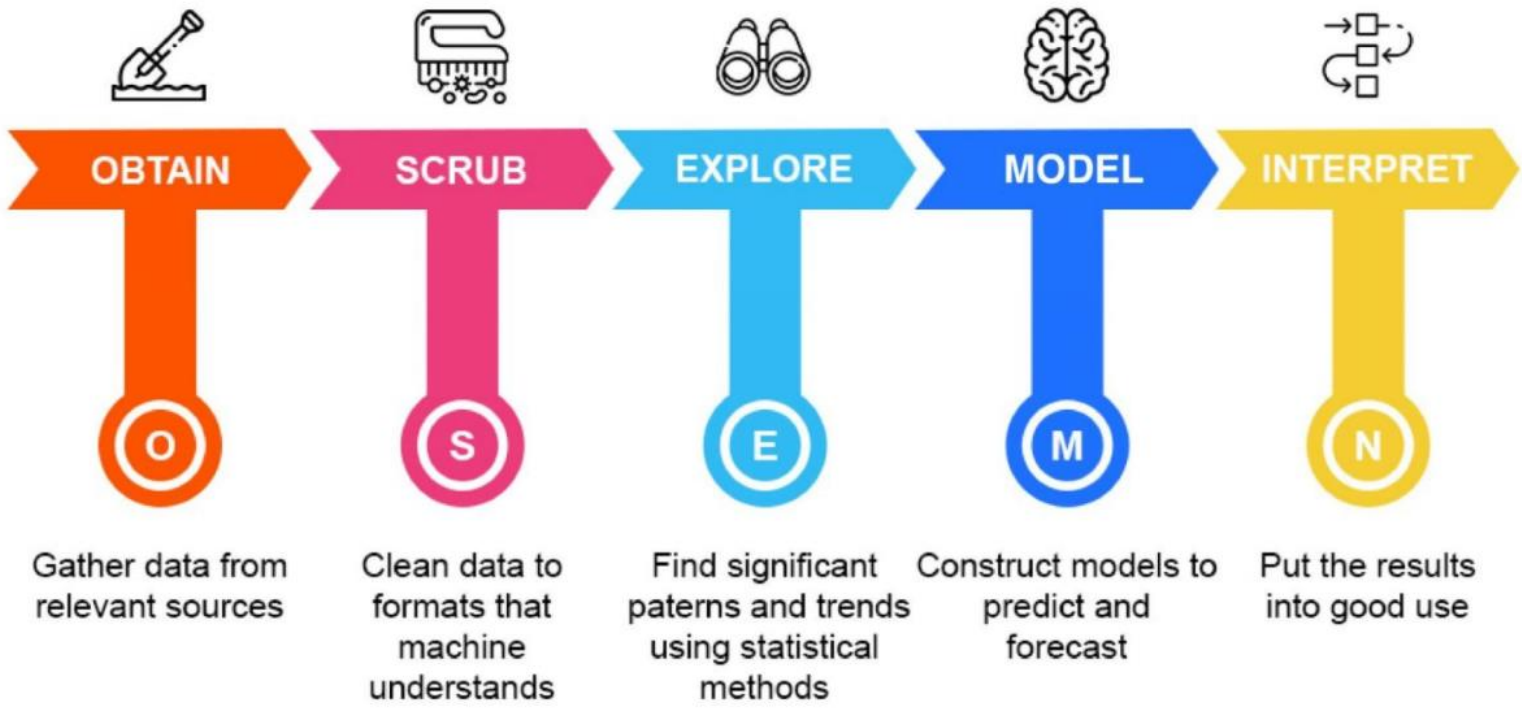


- Feature = Een meetbare eigenschap van een geobserveerd datapunt
- Feature engineering = Zoeken naar de beste features om iets te bereiken
  - Vereist domein kennis
- Feature Selection
  - Verwijder onbruikbare features/datapunten
  - Curse of dimensionality
- Feature Construction
  - Nieuwe features op basis van bestaande
  - Vaak belangrijk in het geval van beelden
  - vb: Enkel geinteresseerd of iemand volwassen is en niet de exacte leeftijd.

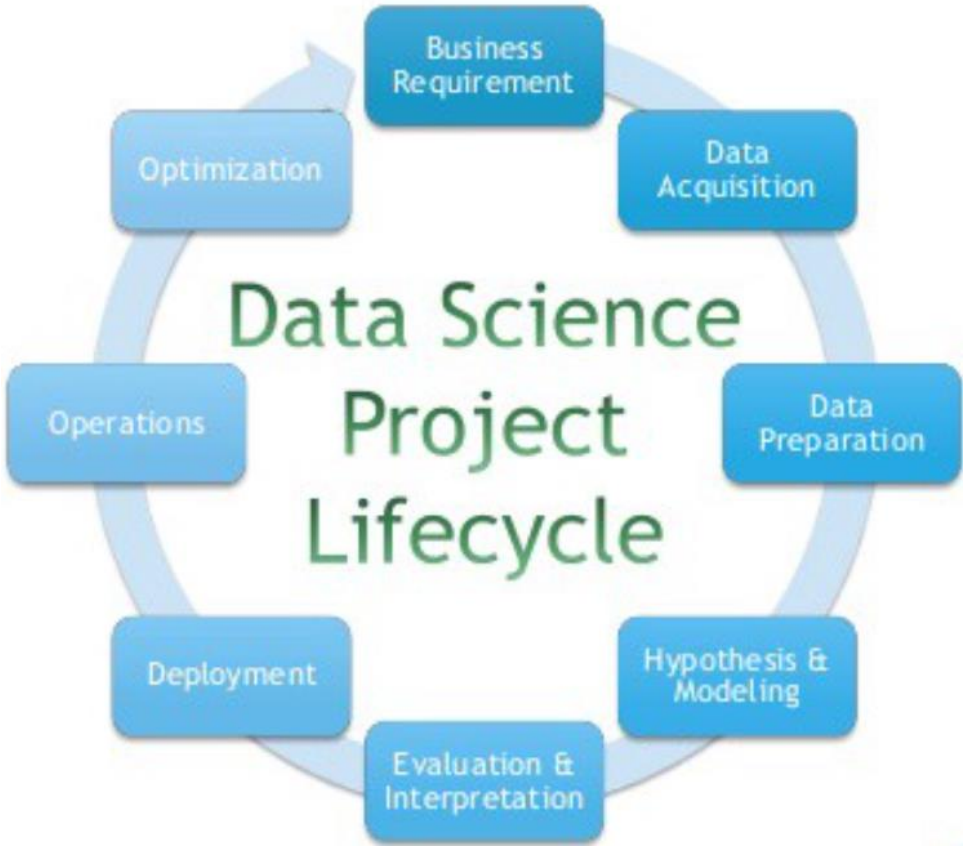
- Machine learning model opbouwen
  - Probeer verschillende varianten en evalueer elk model
  - Zie cheat sheet voor een aantal mogelijkheden
- Beste keuze hang af van:
  - Hoeveelheid, type en kwaliteit van de data
  - Beschikbare computer-capaciteit
  - Gewenste output type

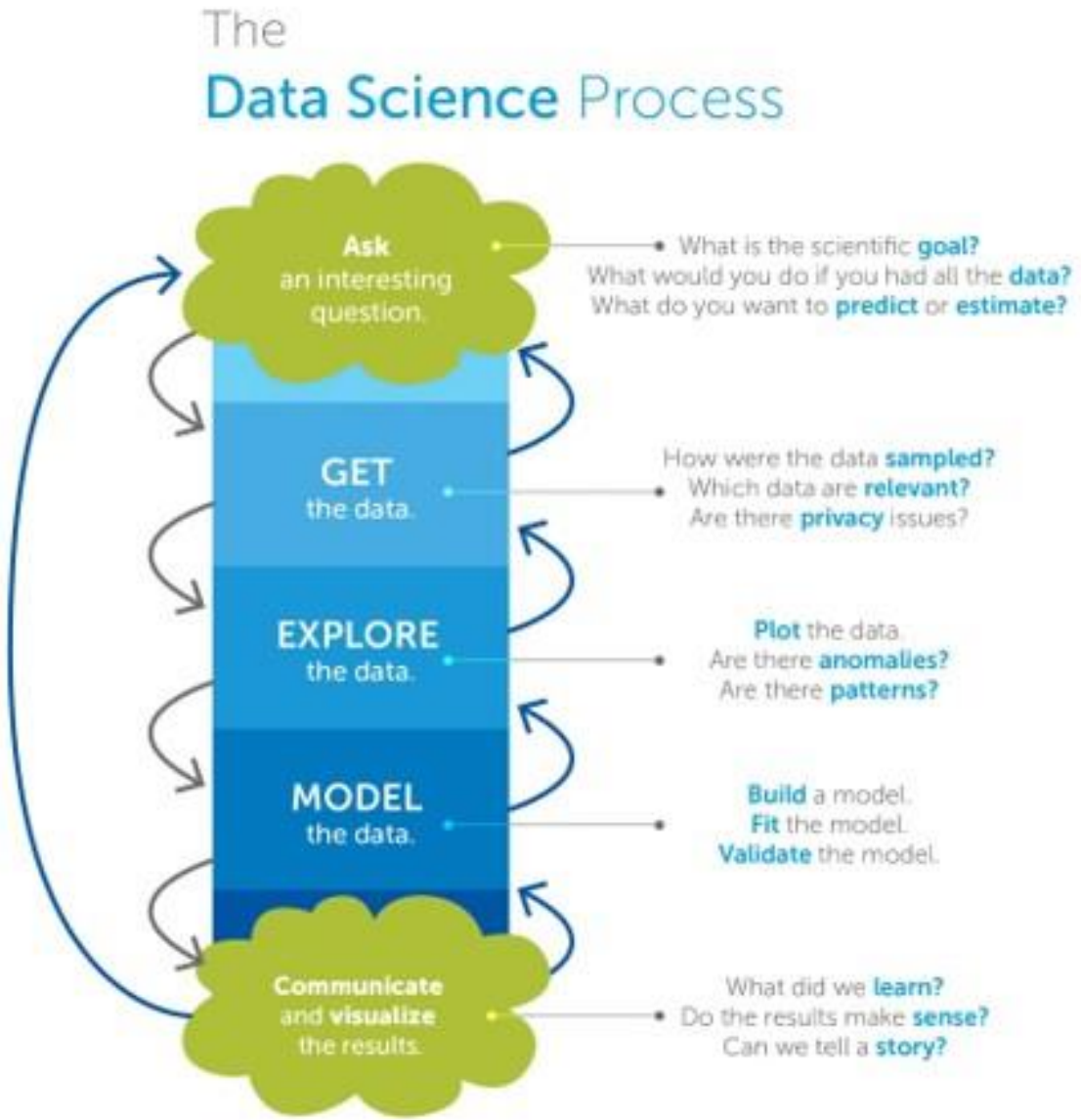
- Visualiseer de resultaten en inzichten
- Communicatie aangepast aan het doelpubliek

Data Science lifecycle – alternatieve modellen



Data Science lifecycle – alternatieve modellen




The
Data Science Process
Ask
an interesting
question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
GET
the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
EXPLORE
the data.
Plot the data.
Are there anomalies?
Are there patterns?
MODEL
the data.
Build a model.
Fit the model.
Validate the model.
Communicate
and visualize
the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Belangrijke termen

- Data Collection
- Data Cleaning
- Exploratory Data Analysis
- Data Modelling
- Training
- Feature
- Feature Engineering
- Modelling
- Training
- Curse of dimensionality

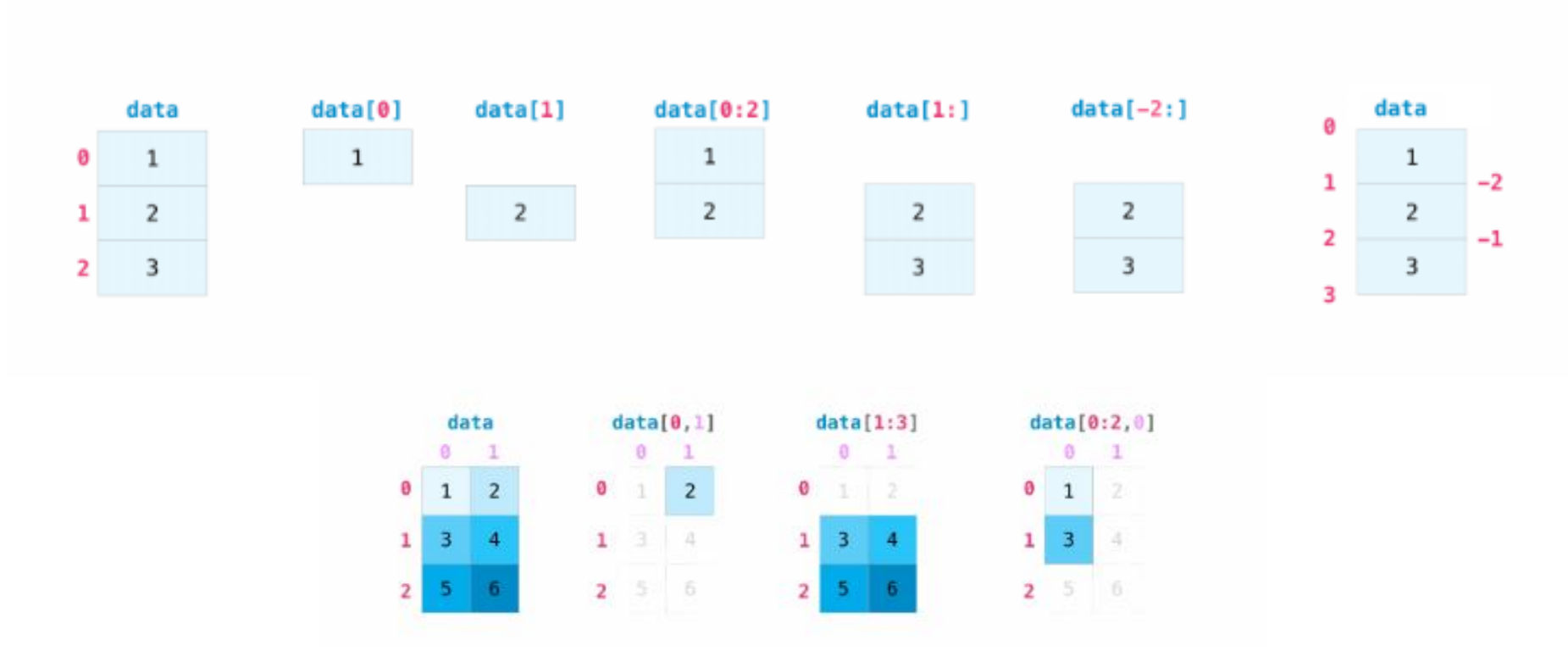
Numpy en Pandas

Numpy

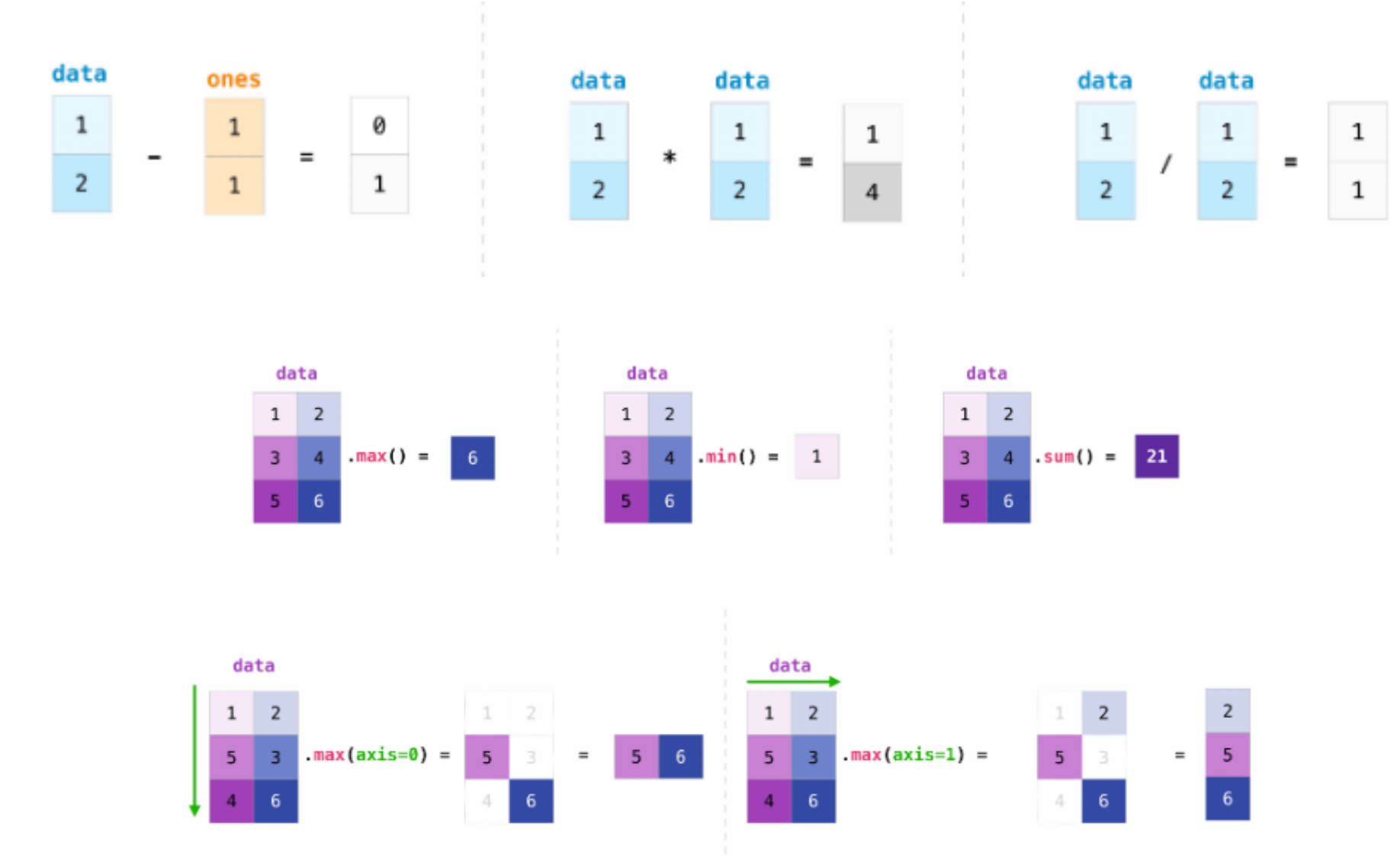
- Fundamentele package voor wetenschappelijke/wiskundige berekeningen
  - Pre-compiled C-code voor snelheid

- Nd-array
  - N-dimensional array
  - Vaste grootte
  - Matrixberekeningen

Numpy - slicing

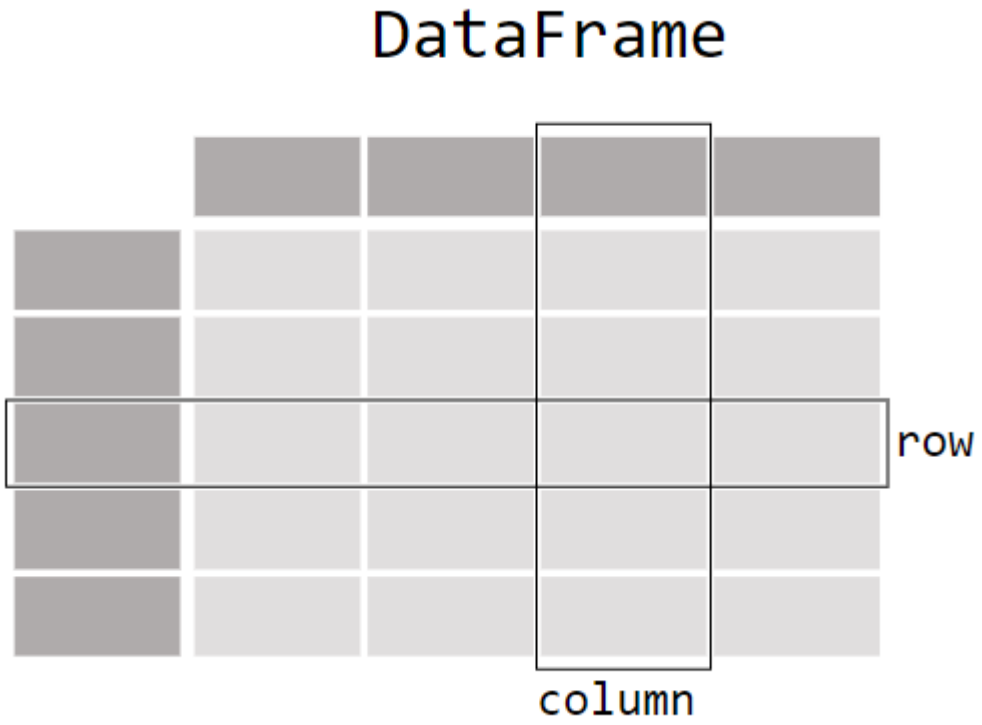


Numpy – matrix bewerkingen



Pandas

- Voor statistische analyse
- Gestructureerde data
- SQL-like in code
- Pre-compiled in C voor snelheid
- Dataframes



Data Collection


Data Collection
Methods

Hoe onderscheid maken in soorten data?

Primary	Secondary
Zelf verzamelde data door enquetes, 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
Kan lang duren en veel geld kosten	Zeer snel om mee aan de slag te gaan

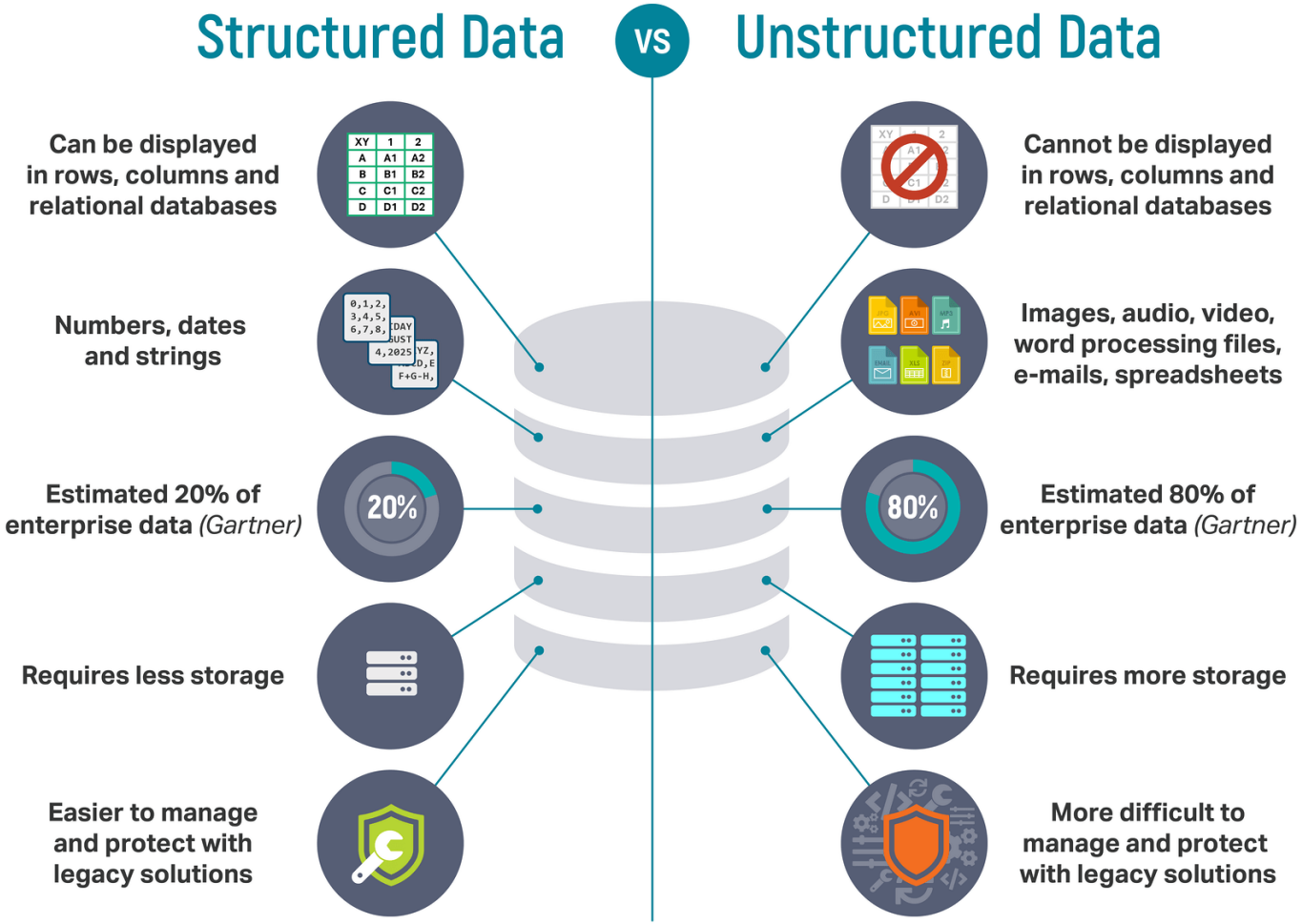
Soorten secundaire databronnen

- Bestaande publieke online datasets
  - Kaggle, statbel.fgov.be, Github, …
- Gebruik maken van aangeboden API’s
  - Facebook API, twitter API, …
- Scraping of websites
  - Vaak zelf te schrijven
  - Afhankelijk van structuur website en gevoelig voor wijzigingen

Hoe onderscheid maken in soorten data?

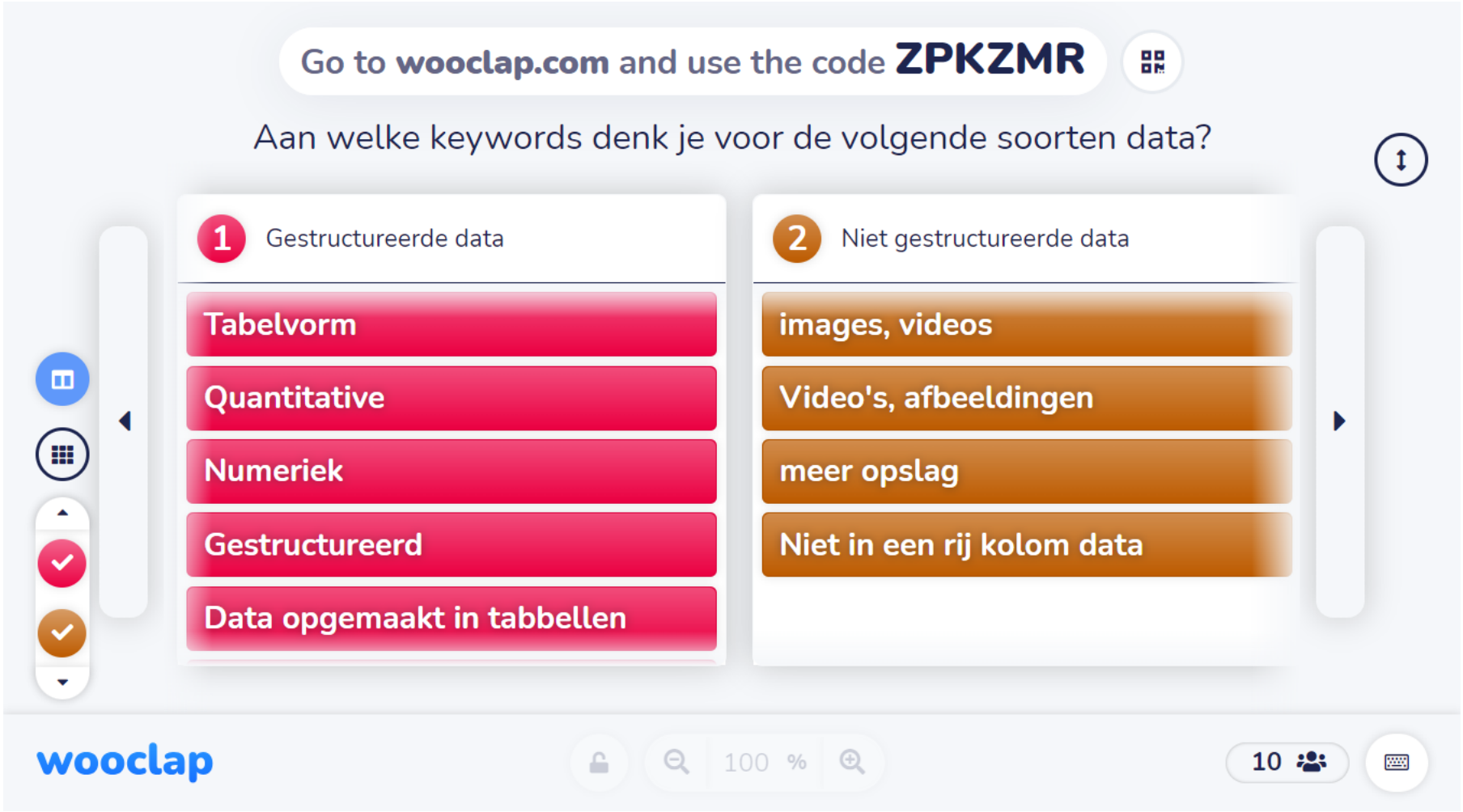
Quantitatieve data	Qualitatieve data
Numerieke waarden (leeftijd, gewicht, …)	Niet-numerieke waarden om eigenschappen, meningen gevoelens te beschrijven.
Statistische evaluatie mogelijk	Clusteren of groeperen van gelijkaardige waarden
Hoeveel heeft je wafel gekost?	Waarom heb je een wafel gekocht?
70%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

Hoe onderscheid maken in soorten data?




Go to wooclap.com and use the code ZPKZMR
Geef een aantal keywords voor de volgende soorten data
1 Primary data
Tijdrovend, duur
2 Secondary data
wooclap
100 %
1


Go to wooclap.com and use the code ZPKZMR
Waar kan je data verzamelen?
api's
github
KAGGLE
Scraping
Github
Veel soorten websites (kaggle?)
spreadsh eeta
github
statbel
API
Kaggle
Fysieke archiven
database s
kaggle
wooclap
100 %
19


Go to wooclap.com and use the code ZPKZMR
Aan welke keywords denk je voor de volgende soorten data?
1 Gestructureerde data
Tabelvorm
Quantitative
Numeriek
Gestructureerd
Data opgemaakt in tabbellen
2 Niet gestructureerde data
images, videos
Video's, afbeeldingen
meer opslag
Niet in een rij kolom data
wooclap
100 %
10

Belangrijke termen

- Primaire data
- Secundaire data
- Quantitatieve data
- Qualitatieve data
- Gestructureerde data
- Niet-gestructureerde data

Numpy en pandas tutorial

- Ga naar:
  - https://www.kaggle.com/learn/pandas
  - Volg de volledige tutorial voor een pandas introductie

  - Deze tutorials zijn deel van de leerstof van dit vak en helpen om vlot te kunnen beginnen
    - Studeer ze dus volledig tegen volgende les

Pandas oefening

- Opgave: https://classroom.github.com/a/UMeI7Tn9
- Maak de oefening individueel tegen volgende week

- Dit is een opwarmer en wordt niet geëvalueerd