

Odissee
DE CO-HOGESCHOOL

Lifecycle overzicht



Jens Baetens



01

BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

- 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, ...

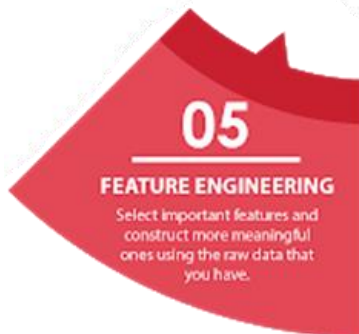


04

DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

- ▣ 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 geïnteresseerd 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



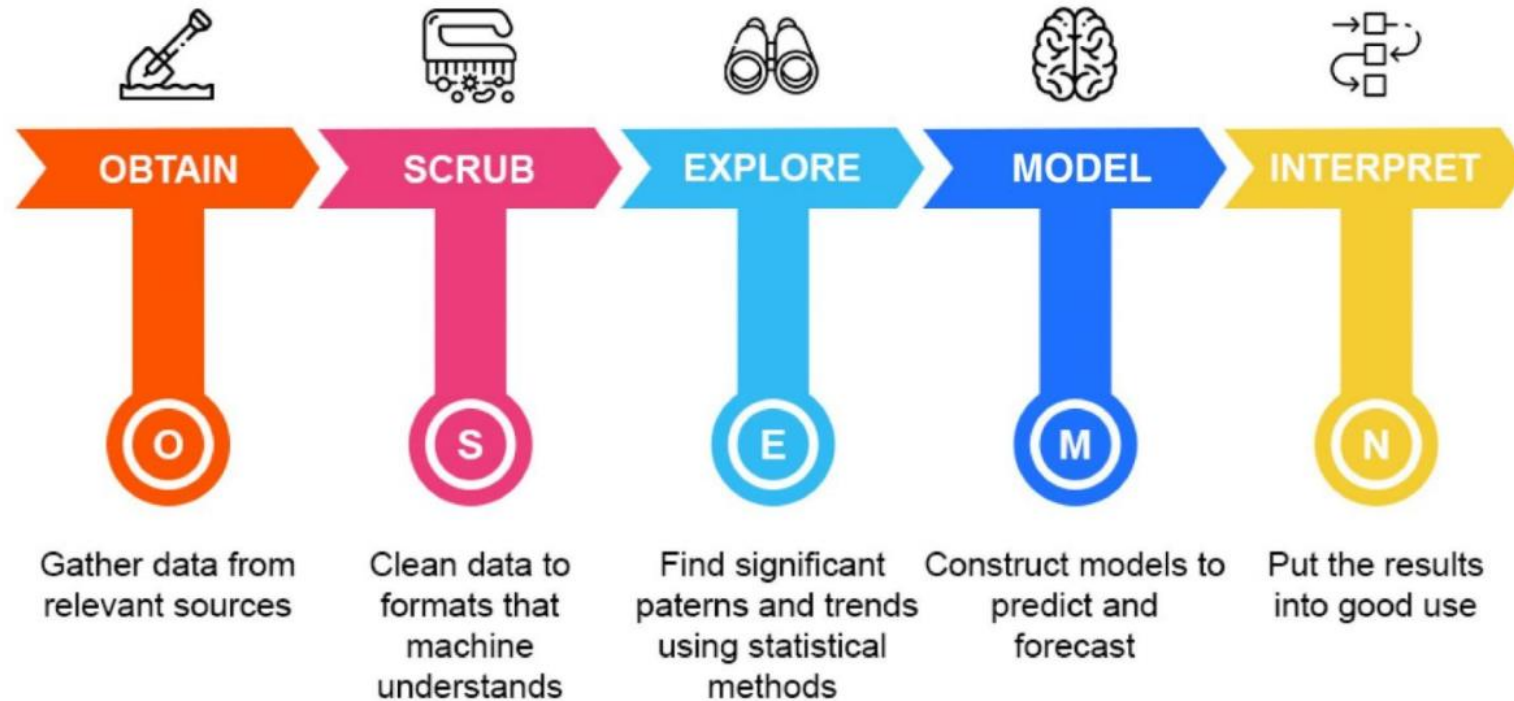
07

DATA VISUALIZATION

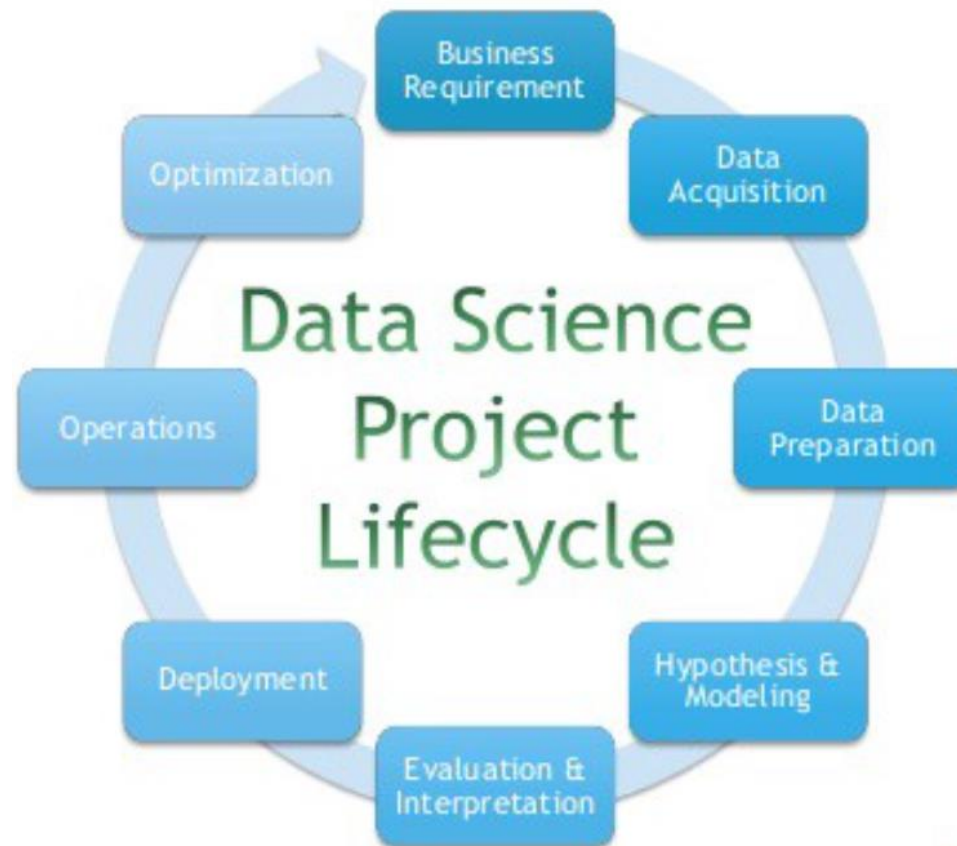
Communicate the findings
with key stakeholders using
plots and interactive
visualizations.

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

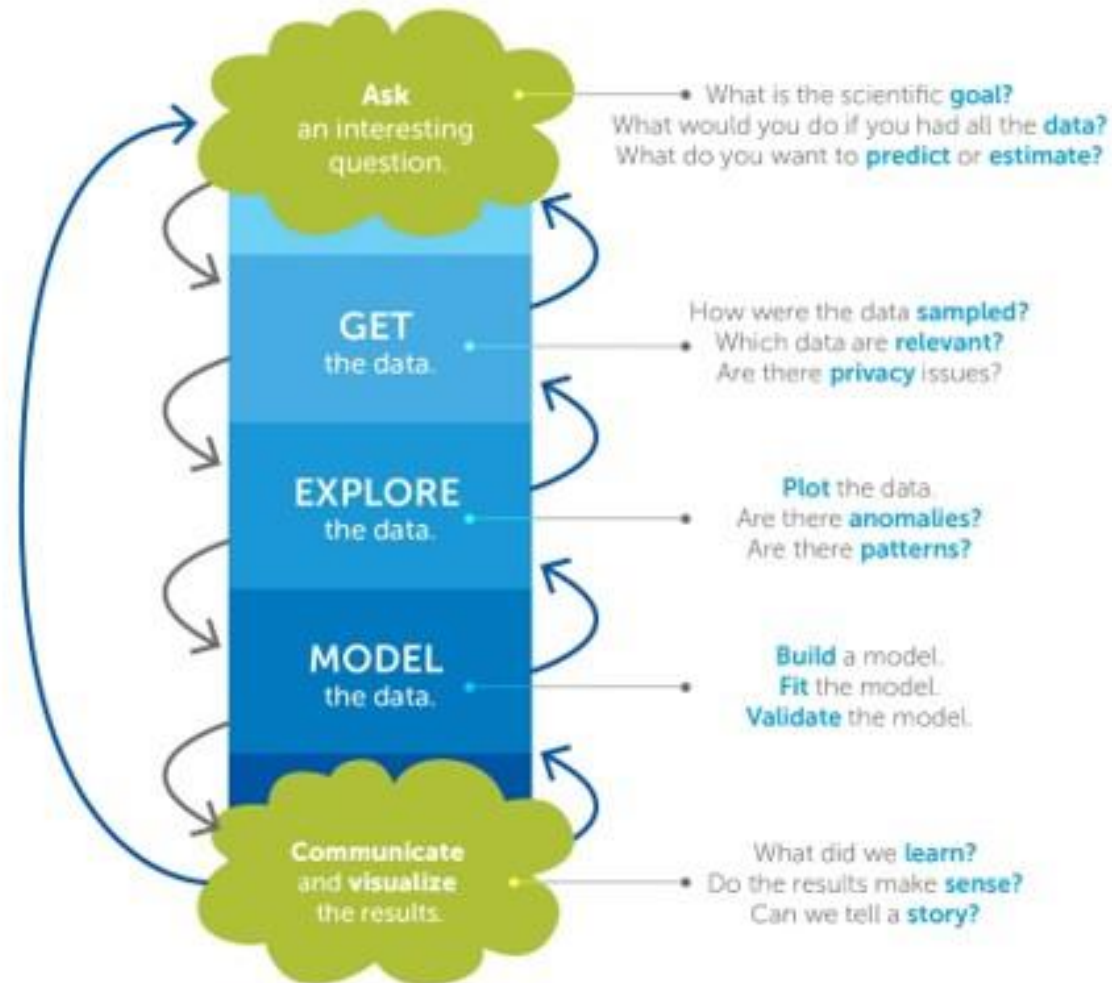
Data Science lifecycle – alternatieve modellen



Data Science lifecycle – alternatieve modellen



The Data Science Process





Belangrijke termen

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

Hoe deelnemen?



Klik op het geprojecteerd scherm om de vraag te activeren

 [Deelnamelink kopiëren](#)

wooclap

 100 % 

0



Go to **wooclap.com** and use the code **PZLPCB**



Wat zijn de te volgen stappen in de data science lifecycle



① Feature engineering

② Data mining

③ Business understanding

④ Data exploration

⑤ Data vizualisation

⑥ Data cleaning

⑦ Predictive modeling



Click on the projected screen to start the question

wooclap



100 %



0 / 0



Go to **wooclap.com** and use the code **PZLPCB**



Wat is het belangrijkste dat er gebeurt in elke stap van de Data Science ...



Business Understanding

1

A

Zoeken naar een op te lossen vraag



Data Mining

2

B

Verzamelen van data



Data Cleaning

3

C

Split de data af die je gaat gebruiken

Data Exploration

4

D

Zoeken naar verbanden in de data

Click on the projected screen to start the question

wooclap



100 %



0 / 0

