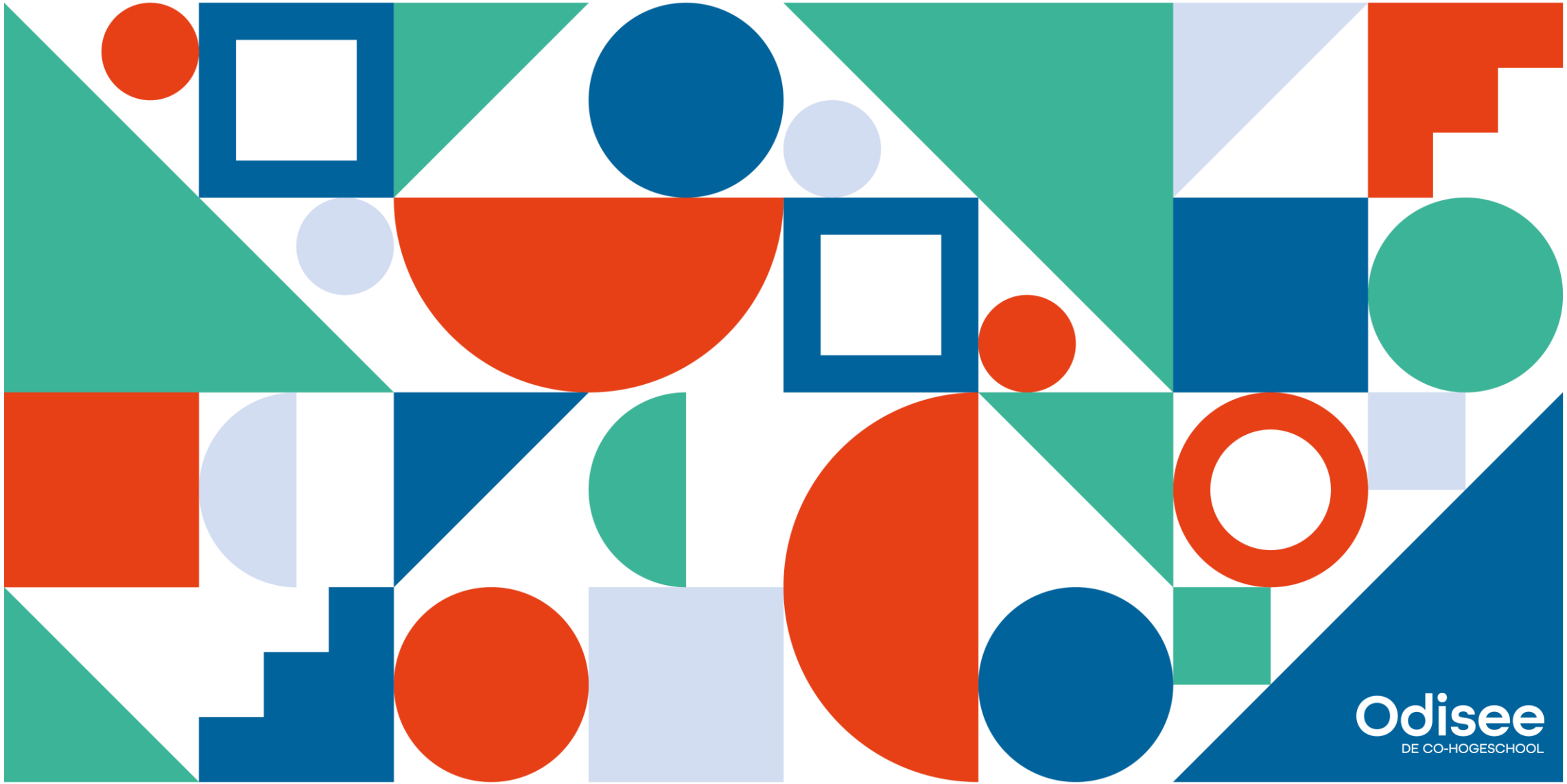
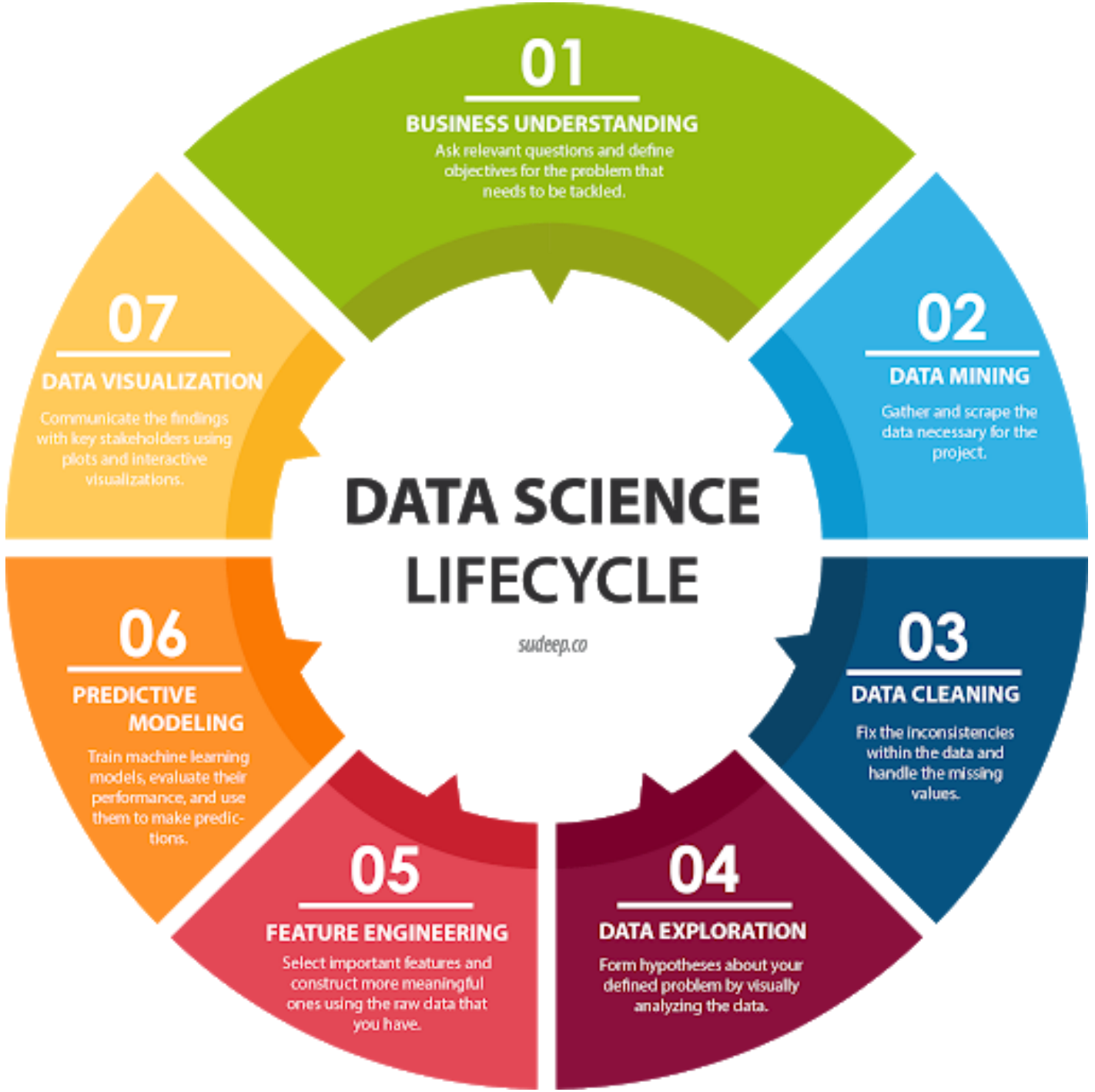

Odisee
DE CO-HOGESCHOOL

Lifecycle overzicht

Jens Baetens


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

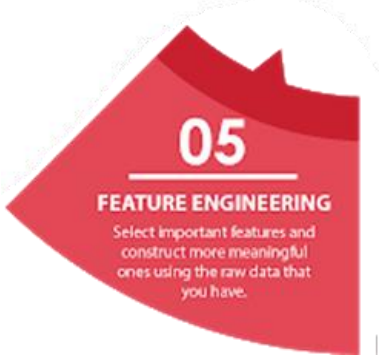
- Wat is de gestelde vraag of het probleem?
- Formuleer de vragen waarop een antwoord moet gevonden worden
- 5 soorten vragen:

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

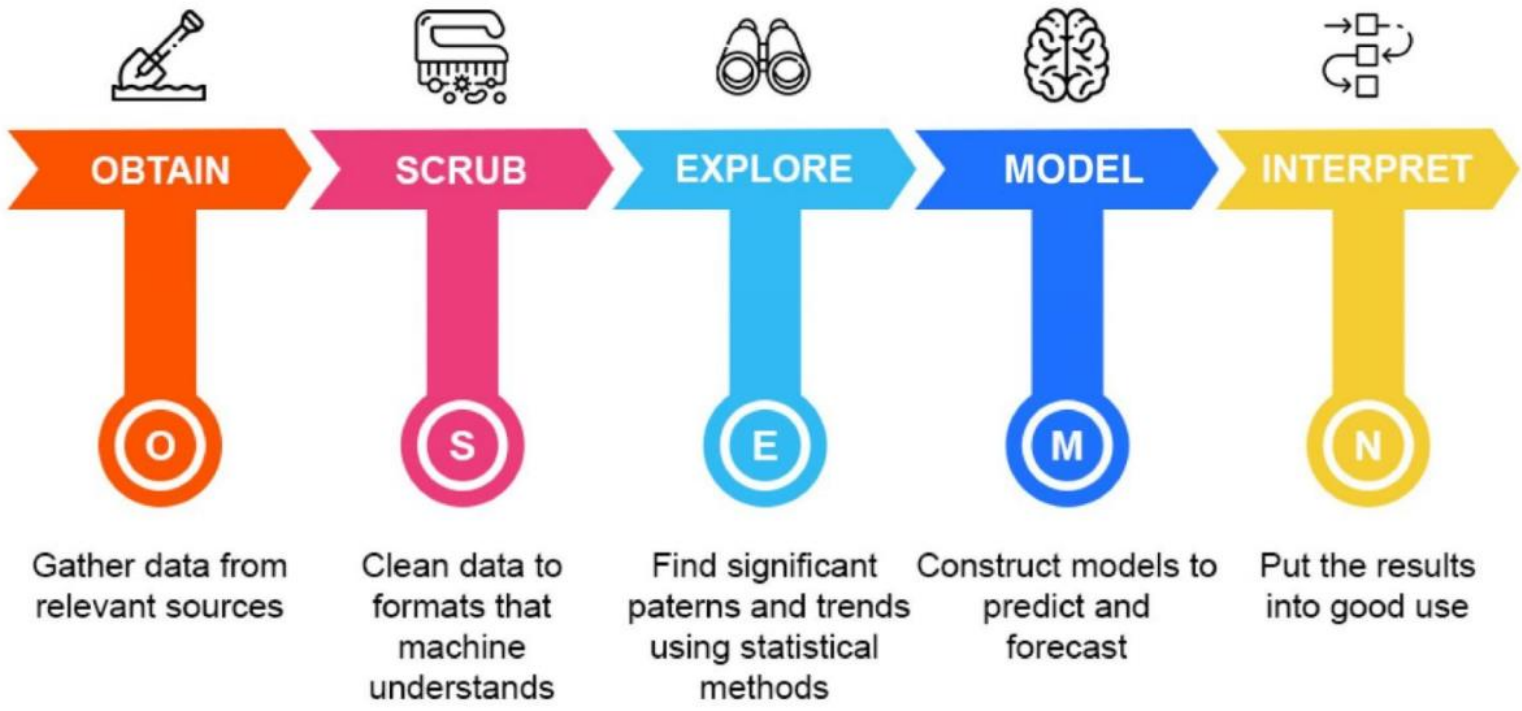


- Feature = Een meetbare eigenschap van een geobserveerd datapunt
- Feature engineering = Zoeken naar de beste features om iets te bereiken
  - Vereist domein kennis
- Feature Selection
  - Verwijder onbruikbare features/datapunten
  - Curse of dimensionality
- Feature Construction
  - Nieuwe features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

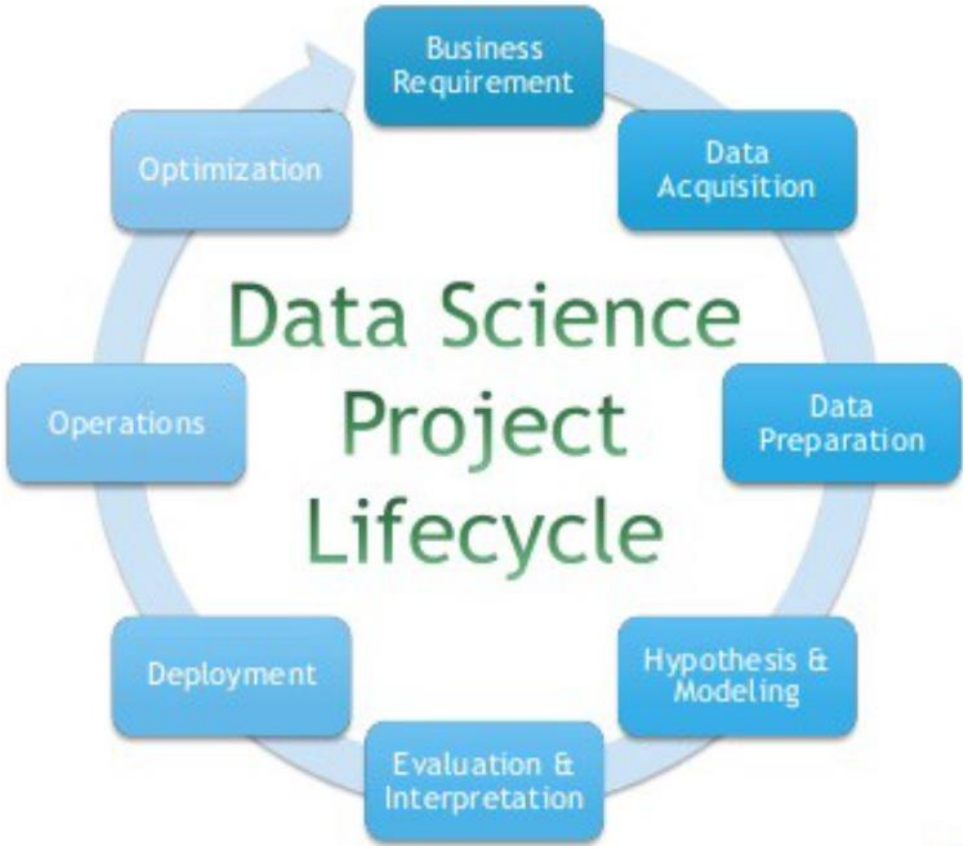
- Machine learning model opbouwen
  - Probeer verschillende varianten en evalueer elk model
  - Zie cheat sheet voor een aantal mogelijkheden
- Beste keuze hang af van:
  - Hoeveelheid, type en kwaliteit van de data
  - Beschikbare computer-capaciteit
  - Gewenste output type

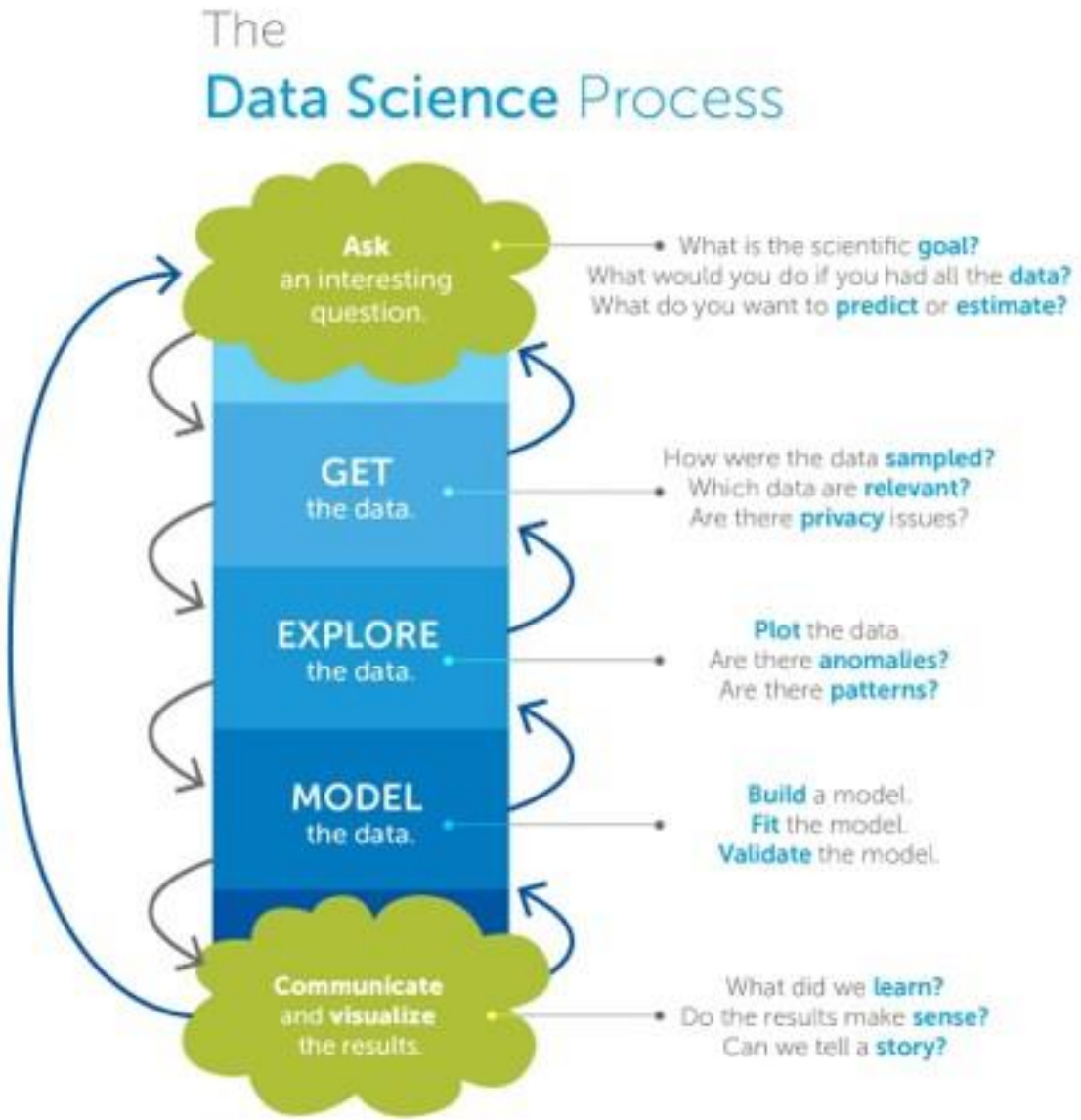
- Visualiseer de resultaten en inzichten
- Communicatie aangepast aan het doelpubliek

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
- Feature
- Feature Engineering
- Modelling
- Training
- Curse of dimensionality