

# Data Foundations

## Hoofdstuk 1

Hassan Haddouchi



# Praktische afspraken

Theorie:

Donderdag om de 2 weken.

Slides komen na de les op Digitap.

Labo:

Elke donderdag.

Indienen opdracht meestal om de 2 á 3 weken.

## **Evaluatie:**

Permanente evaluatie (20%): portfolio over de ingediende labo's.

## **Examen**

Praktisch (40%)

Theorie (40%)

# Inhoud van de cursus (onder voorbehoud)

Introductie

Lab: Python en data

Data-visualisatie

Lab:

Python data-verwerking

Data-visualisatie

Data-analytics

Lab:

Power BI

Data engineering

Leren uit data

Lab:

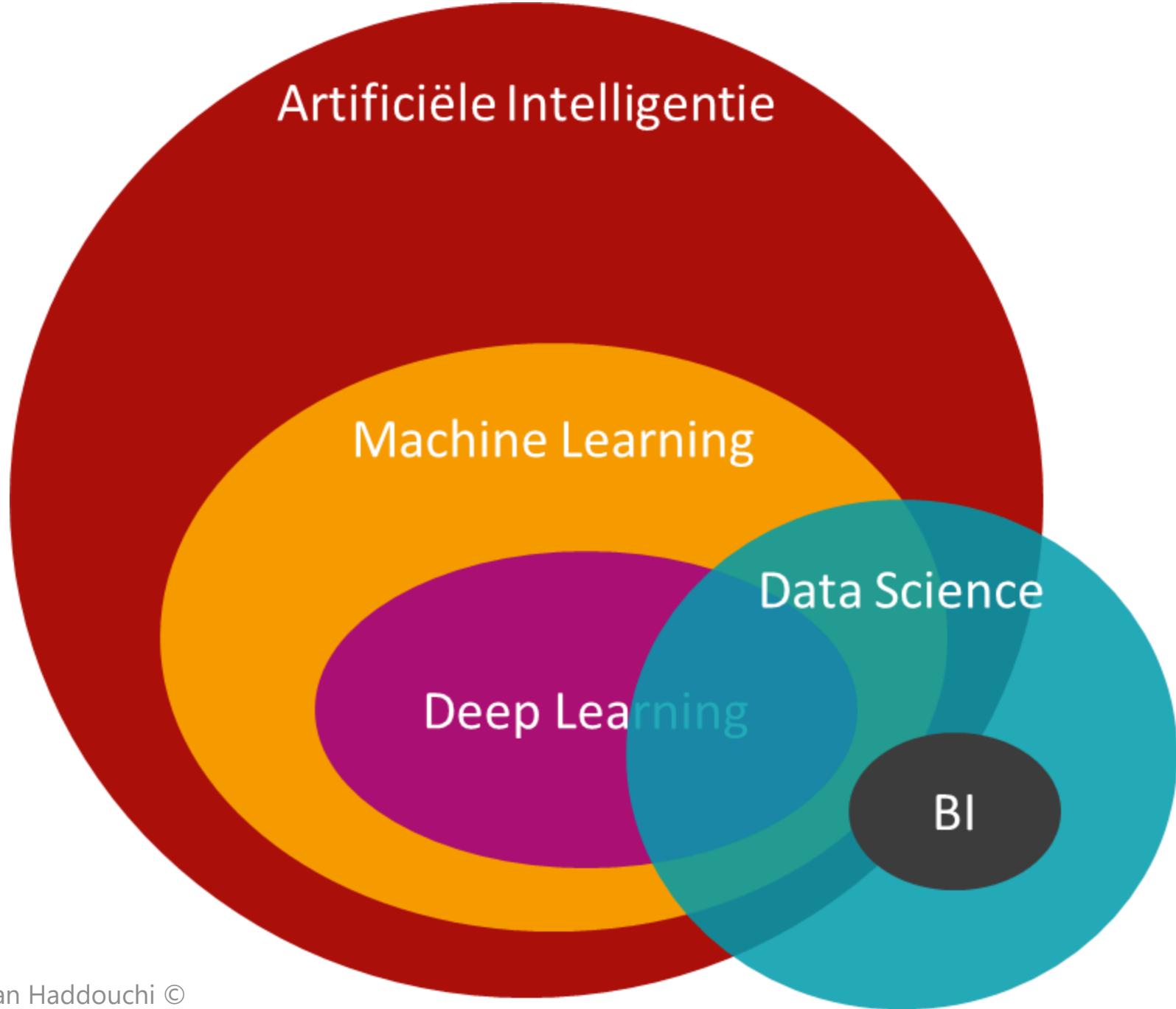
RapidMiner

Data ethics & privacy (gastspreker)

Lab:

Ethisiek en data

**Wat is het eerste waar jullie aan denken bij data?**



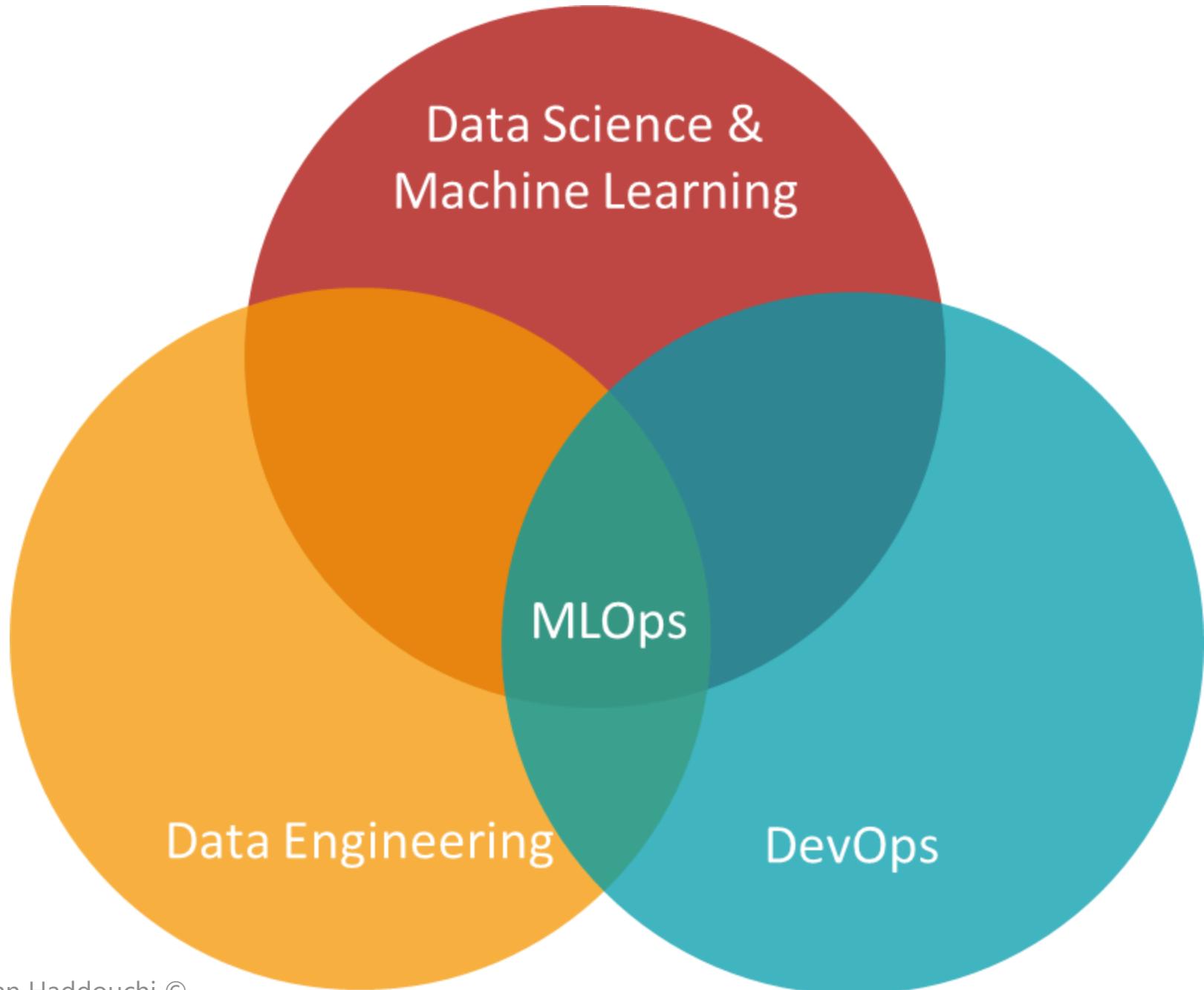
# Enkele termen

## Data Science

Business domein dat grote hoeveelheden data gebruikt om nieuwe inzichten te ontdekken, voorspellingen te doen en beslissingen te ondersteunen.

## Business Intelligence (BI)

Analyseren van data om business te ondersteunen (rapporteren, dashboards, analytics, ...).



## **Data Engineering**

Installeren en onderhouden van infrastructuur die data toegankelijk maakt.

## **DevOps**

Praktijken binnen software ontwikkeling die toelaten om snel wijzigingen door te voeren.

## **ML Operations (MLOps)**

Praktijken die toelaten om ML modellen betrouwbaar en efficiënt in productie te zetten.

**Welke toepassingen van ML en Deep Learning (DL)  
kennen jullie?**

**Om aan ML of DL te doen, heb je data nodig.**

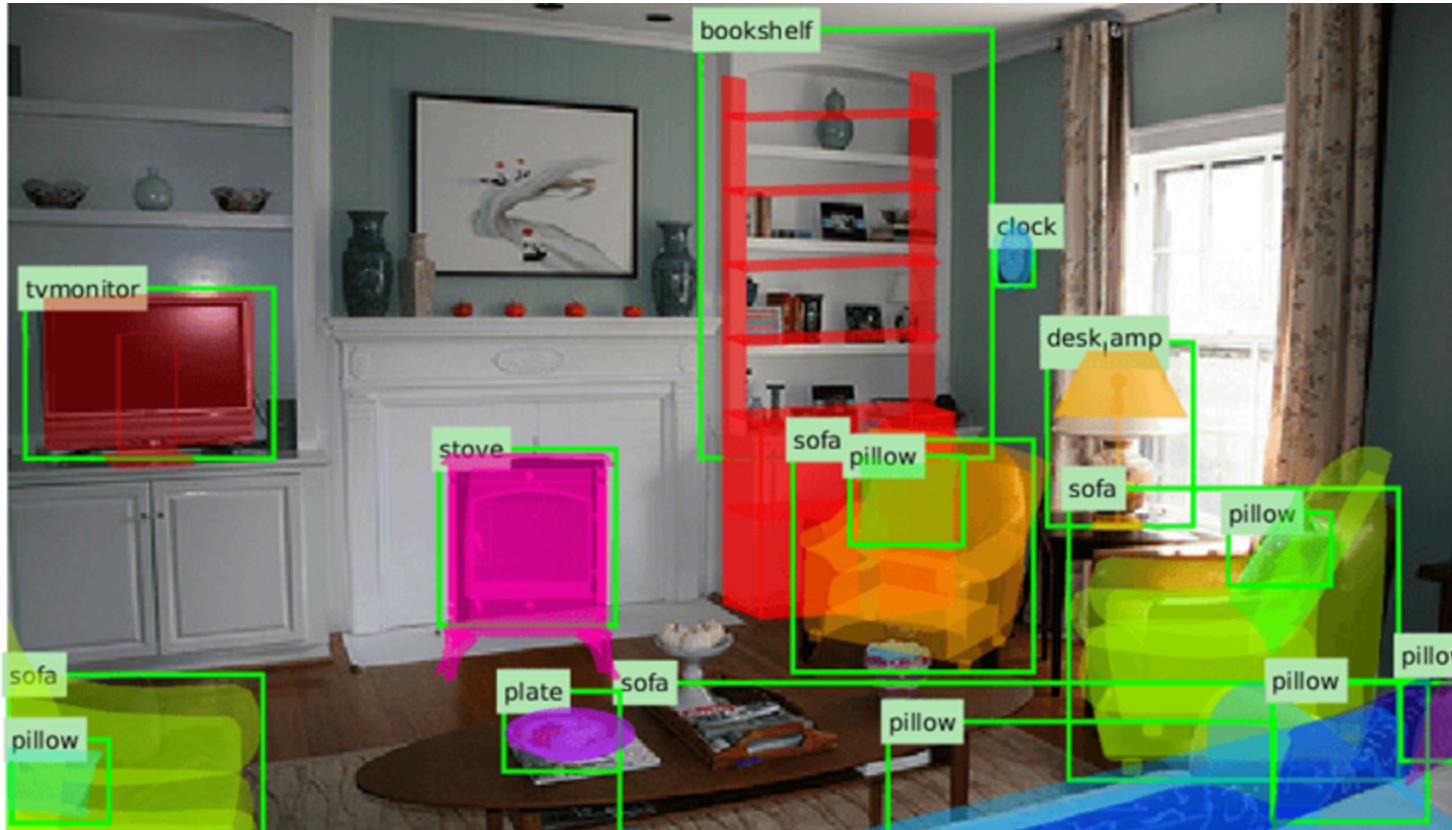
# Voorbeelden van DL: vertalen

The screenshot shows the Google Translate interface. At the top, there's a navigation bar with a menu icon, the "Google Vertalen" logo, and a user profile icon with a 'W'. Below the bar are three tabs: "Tekst" (selected), "Documenten", and "Websites". The main area has two language rows. The first row has "TAAL HERKENNEN" (Language detection), "ENGELS" (selected), "NEDERLANDS", and "FRANS". The second row has "NEDERLANDS" (selected), "ENGELS", "FRANS", and a dropdown arrow. Between these rows is a double-headed arrow icon. Below the rows is a large input field containing a vertical pipe character ("|") and a microphone icon. To its right is a smaller output field containing the word "Vertaling". At the bottom of the input field is a progress bar showing "0 / 5.000" and a settings icon. A "Feedback sturen" link is at the bottom right.

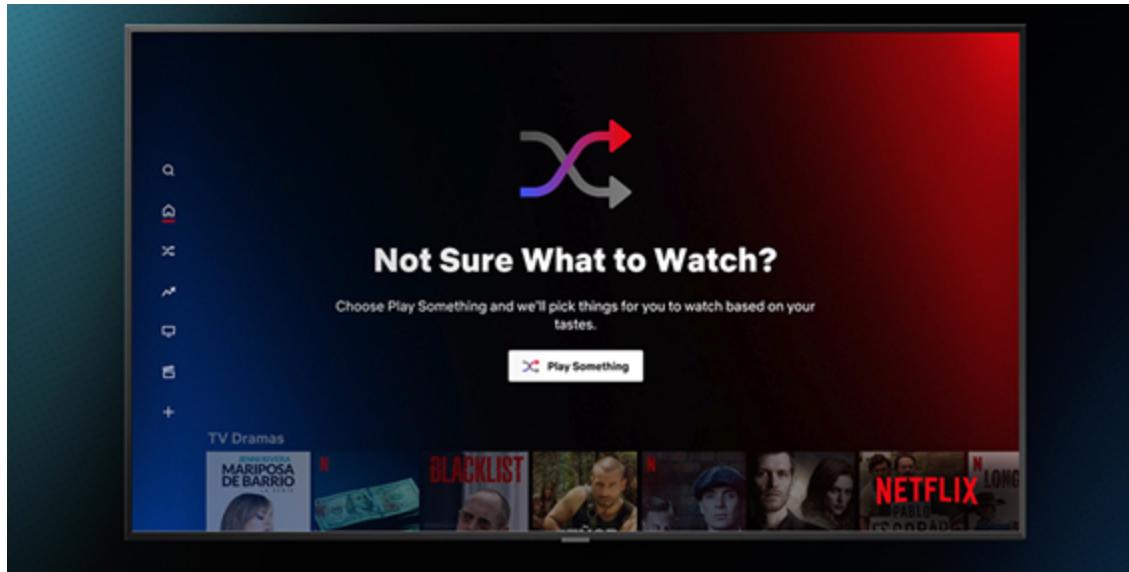
# Voorbeelden van DL: spraakherkenning



# Voorbeelden van DL: beeldherkenning



# Voorbeelden van DL: persoonlijke aanbevelingen



# Voorbeelden van DL: games



## **Nu weten we waar data voor nodig is.**

Namelijk voor databeheer en analyse, inzichten verkrijgen uit uw data en om ML en DL toepassingen te ontwikkelen.

# Soorten data

# Datatypes

Data komen in verschillende vormen voor, elk met zijn unieke kenmerken.

Het begrijpen van deze typen is cruciaal voor effectief databeheer en analyse.

Er bestaan 3 primaire datatypen:

- Gestructureerd
- Semi-gestructureerd
- Ongestructureerd

# Gestructureerde data

Gestructureerde data is erg georganiseerd en volgt een vast schema.

Meestal opgeslagen in relationele databases.

Voorbeelden

tabellen in een database,

spreadsheets en CSV-bestanden.

Goed geschikt voor numerieke en categorische gegevens.

Gemakkelijk te bevragen, analyseren en visualiseren.

## Semi-gestructureerde data

Semi-gestructureerde data is gedeeltelijk georganiseerd en mist een strikt schema.

Het omvat gegevensformaten zoals JSON, XML en NoSQL-databases.

Biedt flexibiliteit in de representatie van gegevens.

Bevat metadata of tags voor elementen.

Gebruikt in scenario's waar gegevensstructuren kunnen evolueren of variëren.

# Ongestructureerde data

Ongestructureerde data is de meest diverse en uitdagende om mee te werken.

Het ontbreekt aan een vooraf gedefinieerde structuur en organisatie.

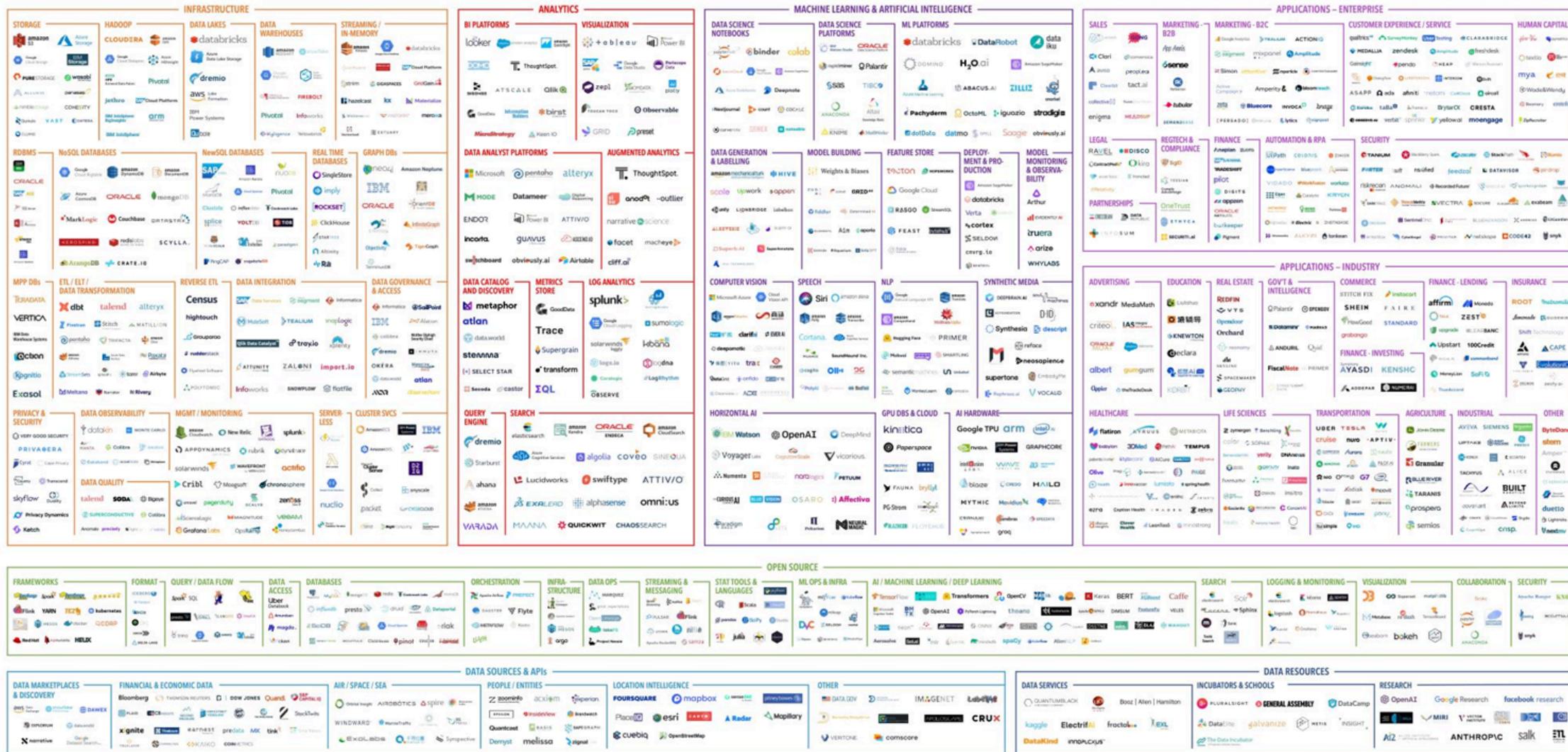
Voorbeelden: tekstbestanden, afbeeldingen, audio- en videobestanden.

Vereist geavanceerde technieken zoals natuurlijke taalverwerking (NLP) en computervisie voor analyse.

Inzichten halen uit ongestructureerde data is complex maar essentieel voor het begrijpen van klantensentiment, beeldherkenning en meer.

# Tooling

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



Version 3.0 - November 2021

© Matt Turck (@mattturck), John Wu (@john\_d\_wu) & FirstMark (@firstmarkcap)

[mattturck.com/data2021](http://mattturck.com/data2021)

FIRSTMARK  
EARLY STAGE VENTURE CAPITAL

Hassan Haddouchi ©

# Tooling

Het is onmogelijk om al deze tools bij te houden.

Het data landschap is nog volop in ontwikkeling.

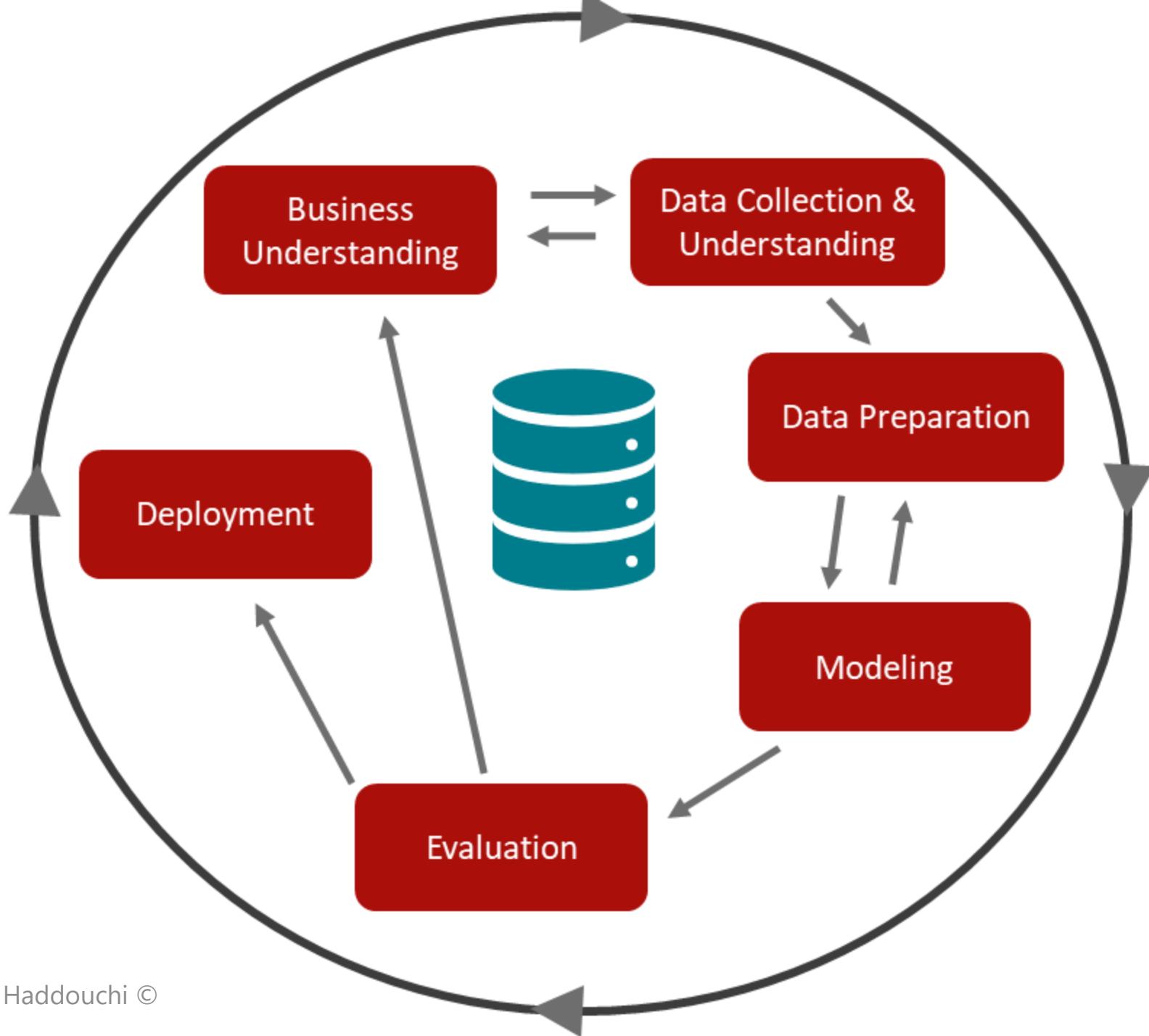
Focus eerder op concepten dan tools.

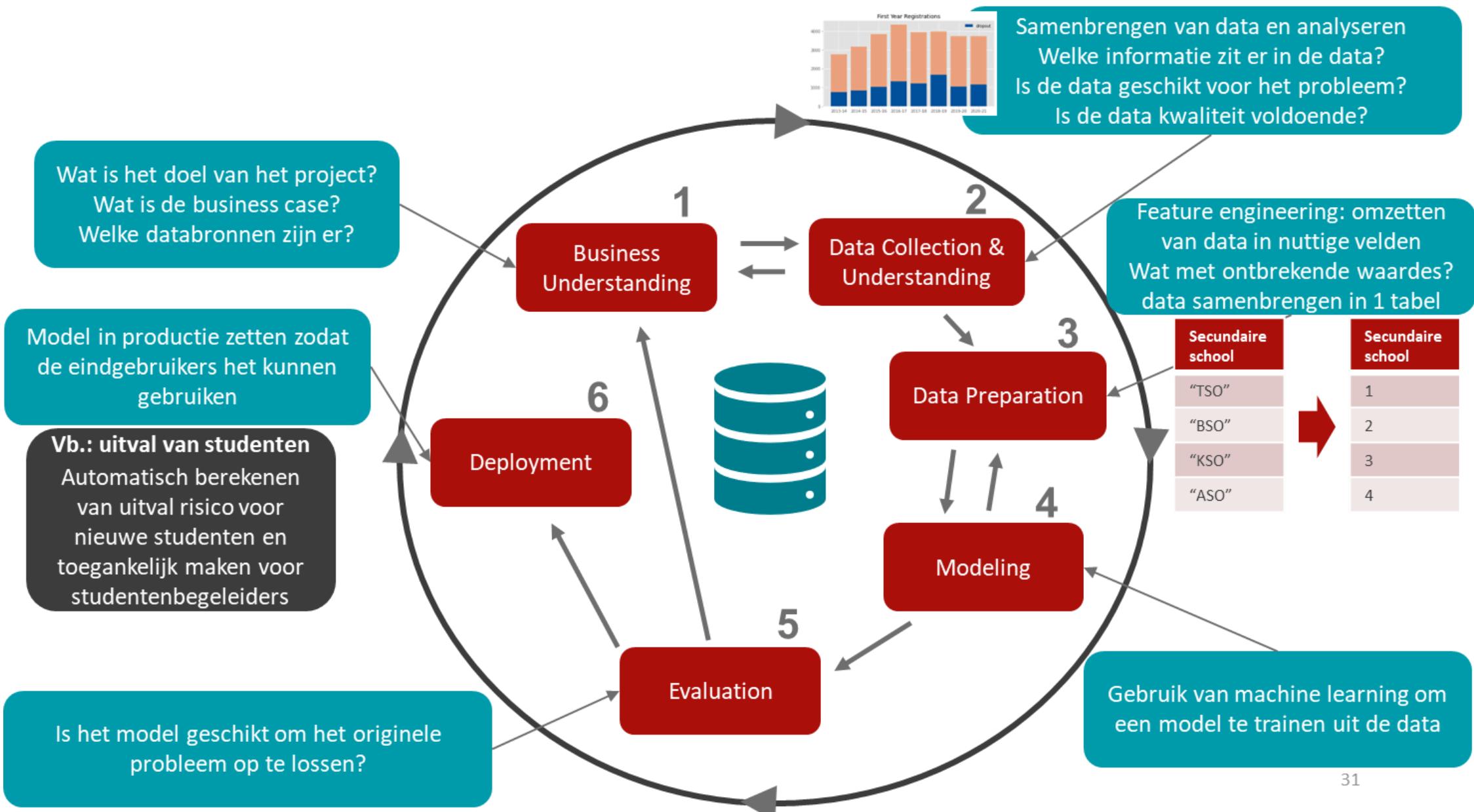
Python is wel al enkele jaren bezig met een opmars rond data.

R is traditioneel nog steeds populair.

# Data Science

## Data Science Lifecycle





# Data Science Lifecycle

Bespreek de verschillende stappen voor volgende projecten:

- Ontwikkeling software voor autonoom voertuig
- Opsporen van fraude met kredietkaarten
- Spam filter samen tijdens de les
- Gepersonaliseerde aanbevelingen op een webshop
- Een automatische kattenluik met gezichtsherkenning
- Detectie van fake news

## Voorbeeld: spam filter

### **Business Understanding:**

Hoe zal het systeem werken? Hoe nauwkeurig moet het spam kunnen detecteren?  
Welke data is er? Database met mails met aanduiding van welke spam zijn?

### **Data Collection & Understanding:**

Samenbrengen historische mails en kijken hoeveel er spam zijn.  
Evolueert spam doorheen de tijd?

## **Data preparation:**

Maken van informatieve datavelden (bevat mail bepaalde woorden zoals "Proficiat!")

## **Modeling:**

Classificatie

## **Evaluatie:**

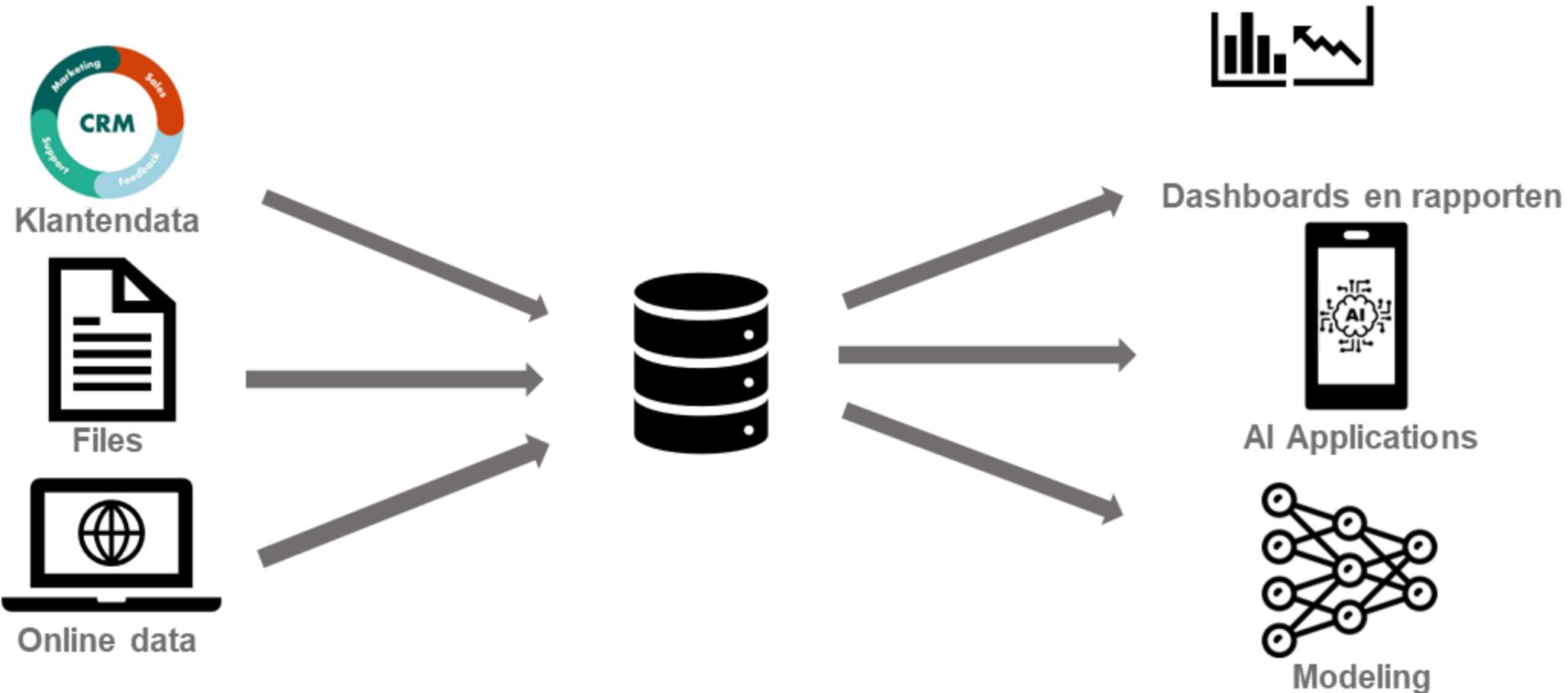
Hoeveel spam wordt er daadwerkelijk gedetecteerd? Is dit voldoende voor een filter?

## **Deployment:**

Gebruik van model in mailbox

# Data Engineering

Installatie en onderhoud van systemen die data samenbrengen en beschikbaar maken.



# Data visualisatie

- Rauwe data (bv. tabellen) is moeilijk om te interpreteren.
- Data visualisaties laat toe om data te presenteren op een intuïtieve manier.
- Data Storytelling: een verhaal vertellen met data.



# Waarom is data nu zo belangrijk??

Het kloppende hart van de moderne wereld.

- Informatiebron
- Besluitvorming
- Innovaties zoals AI
- ...

**Enkele vragen voor we afronden.**

**Wat zijn de populairste programmeertalen voor het werken met data?**

# Wat is Data Engineering?

# Waar dient de Data Science Lifecycle voor?

# Lab

# Dataverwerking met Python

## Pandas

Hassan Haddouchi ©



Pandas is een erg populair Python pakket voor dataverwerking en analyse.

Open source.

Combinatie met Matplotlib voor visualisaties.

## Werkt met Pandas Dataframe:

	name	id	nametype	recclass	mass (g)	fall	year	reclat	reclong	GeoLocation
0	Aachen	1	Valid	L5	21	Fell	01/01/1880 12:00:00 AM	50.77500	6.08333	(50.775, 6.08333)
1	Aarhus	2	Valid	H6	720	Fell	01/01/1951 12:00:00 AM	56.18333	10.23333	(56.18333, 10.23333)
2	Abree	6	Valid	EH4	107000	Fell	01/01/1952 12:00:00 AM	54.21667	-113.00000	(54.21667, -113.0)
3	Acapulco	10	Valid	Acapulcoite	1914	Fell	01/01/1976 12:00:00 AM	16.88333	-99.90000	(16.88333, -99.9)
4	Achiras	370	Valid	L6	780	Fell	01/01/1902 12:00:00 AM	-33.16667	-64.95000	(-33.16667, -64.95)