odisee

Data Science : Intro

katja.verbeeck@odisee.be tim.vermeulen@odisee.be
joris.maervoet@odisee.be

February 13, 2019

Inhoud

Wat geef jij prijs op Facebook?

Cambridge Analytica stopt ermee

02/05/18 om 22:01 - Bijgewerkt om 22:01
Bron: Belga

Het Britse dataverwerkingsbedrijf Cambridge Analytica gaat dicht, zo heeft het bedrijf dat in het middelpunt staat van het privacyschandaal rond Facebook, zelf meegedeeld.

1
Keer gedeeld

Lees later

Cambridge Analytica © Belga

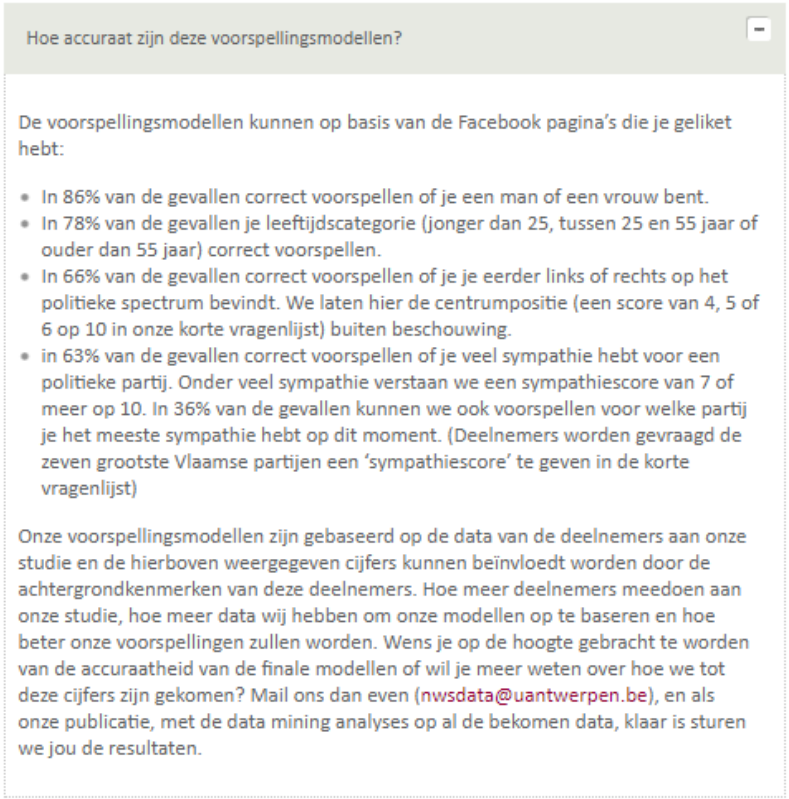
Wat geef jij prijs op Facebook?

https://www.uantwerpen.be/nl/projecten/nws-data/facebookstudie/dataverzameling/

Wat geef jij prijs op Facebook?

https://www.uantwerpen.be/nl/projecten/nws-data/facebookstudie/dataverzameling/

Iedereen vindt Nutella leuk



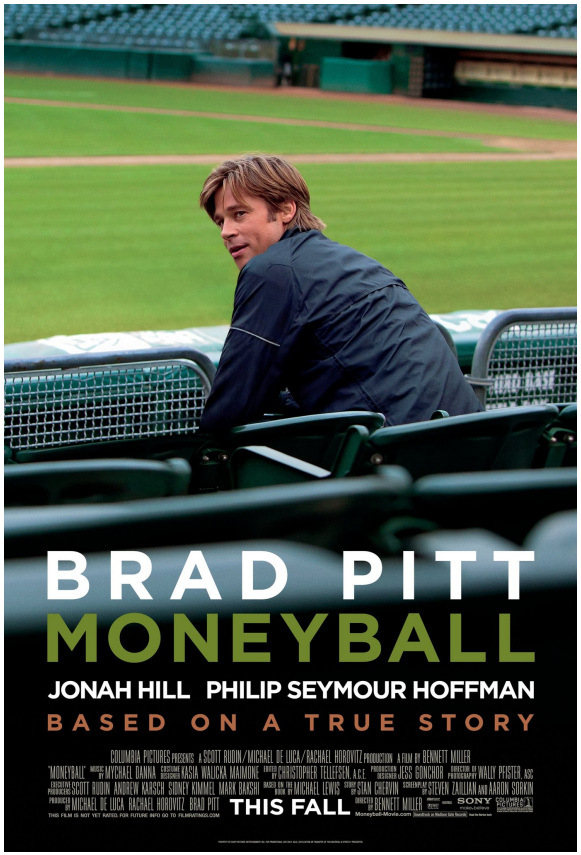
Hoe accuraat zijn deze voorspellingsmodellen?

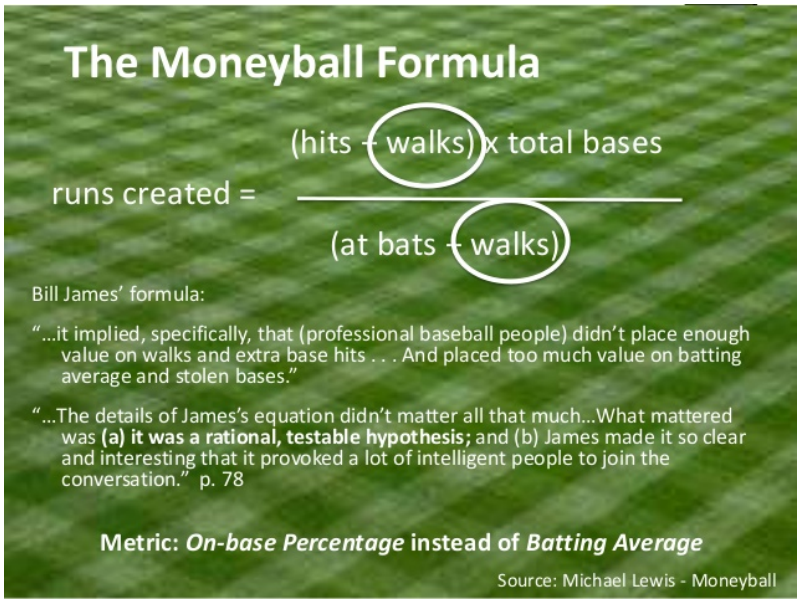
De voorspellingsmodellen kunnen op basis van de Facebook pagina's die je geliket hebt:

- In 86% van de gevallen correct voorspellen of je een man of een vrouw bent.
- In 78% van de gevallen je leeftijdscategorie (jonger dan 25, tussen 25 en 55 jaar of ouder dan 55 jaar) correct voorspellen.
- In 66% van de gevallen correct voorspellen of je je eerder links of rechts op het politieke spectrum bevindt. We laten hier de centrumpositie (een score van 4, 5 of 6 op 10 in onze korte vragenlijst) buiten beschouwing.
- in 63% van de gevallen correct voorspellen of je veel sympathie hebt voor een politieke partij. Onder veel sympathie verstaan we een sympathiescore van 7 of meer op 10. In 36% van de gevallen kunnen we ook voorspellen voor welke partij je het meeste sympathie hebt op dit moment. (Deelnemers worden gevraagd de zeven grootste Vlaamse partijen een 'sympathiescore' te geven in de korte vragenlijst)

Onze voorspellingsmodellen zijn gebaseerd op de data van de deelnemers aan onze studie en de hierboven weergegeven cijfers kunnen beïnvloedt worden door de achtergrondkenmerken van deze deelnemers. Hoe meer deelnemers meedoen aan onze studie, hoe meer data wij hebben om onze modellen op te baseren en hoe beter onze voorspellingen zullen worden. Wens je op de hoogte gebracht te worden van de accuraatheid van de finale modellen of wil je meer weten over hoe we tot deze cijfers zijn gekomen? Mail ons dan even (nwsdata@uantwerpen.be), en als onze publicatie, met de data mining analyses op al de bekomen data, klaar is sturen we jou de resultaten.

Sport + Data Science





Moneyball (2011) gaat over het toepassen van statistiek in de sportwereld. De film vertelt het verhaal van de Oakland Athletics die met een zeer beperkt budget een enorm succes boeken. Billy Beane wil met zijn honkbalteam de Major League winnen, maar zijn budget is minuscuul vergeleken met dat van andere teams. Beane komt met een origineel plan. Waar anderen hoge bedragen neertellen voor spelers met een hoog slaggemiddelde of veel binnengeslagen punten, graaft hij dieper in de statistieken en combineert op het eerste gezicht onlogische spelers tot een winnend team. In navolging van deze film werd statistiek steeds meer toegepast in verschillende sporten.

Sport + Data Science

Visterin Fileert: Voetbal is oorlog, maar vooral data

Waarom Roberto Martinez zo slim is, Kylian Mbappé (en Eden Hazard) zo speciaal en Neymar niet te duur

7 november 2018 08:43 | William Visterin

Topic Digital Innovation

'We beginnen bij voetballers en eindigen bij voetballers. En alles daartussen zijn data.' Giels Brouwer van Scisports, hofleverancier van de Rode Duivels, oogt erg ontspannen tijdens het interview. Zijn bedrijf bestaat vijf jaar en telt vijftig medewerkers, vooral datawetenschappers en ontwikkelaars. En dat allemaal met voetbal.

Voetbal is de belangrijkste bijzaak, zo hoor ik wel eens. Maar dat is het natuurlijk niet waar. Want die zogenaamde bijzaak staat al sinds deze zomer bijna aanhoudelijk op de voorpagina van mijn krant. Eerst tijdens het WK (elke dag), dan met de operatie 'Propere Handen' en nu met de Football Leaks. Voetbal is niet alleen oorlog. Voetbal is gesjoemel, zo blijkt meer en meer. Maar voetbal wordt meer en meer wiskunde.

Sport + Data Science

April 19, 2018

Go Fast and Win: The Big Data Analytics of F1 Racing

Alex Woodie

(Abdul-Razak-Latif/Shutterstock)

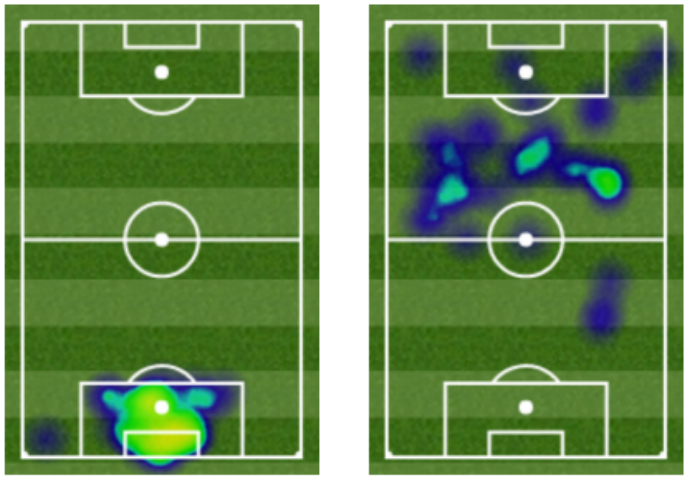
What's the secret to winning in Formula One racing? Simple: Go faster than everybody else. But finding a lasting recipe for success on the racetrack today requires getting hundreds of variables right, so it's no wonder that F1 teams are turning to big data analytics for help.

For a case study in how big data analytics is impacting racing, there's no better example than Mercedes-AMG Petronas Motorsport, which propelled driver Lewis Hamilton to the Formula One driver's championship in November 2017, and which took home the constructor's title to boot. It was Hamilton's fourth F1 title.

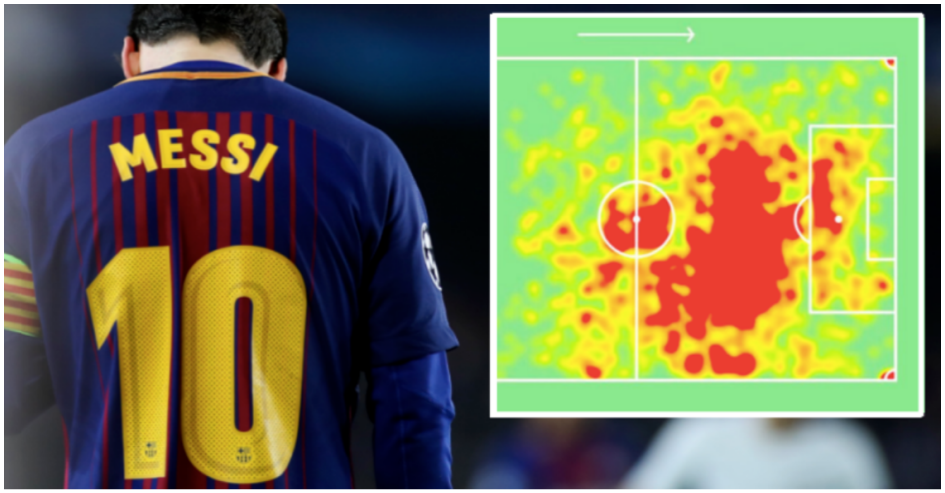
Mercedes-AMG Petronas Motorsport crewmembers at the Malaysian Grand Prix in September 2017 (By Hafiz Johari/Shutterstock)

Sport + Data Science

data visualisatie bron: Pieter Ideler



Figuur 1.1: Heatmaps van de wedstrijd België-Ierland
Wedstrijd op 18 juni 2016 tijdens het Eufa Euro 2016 toernooi.
Links: keeper Thibaut Courtois. Rechts: spits Romelu Lukaku. [1]

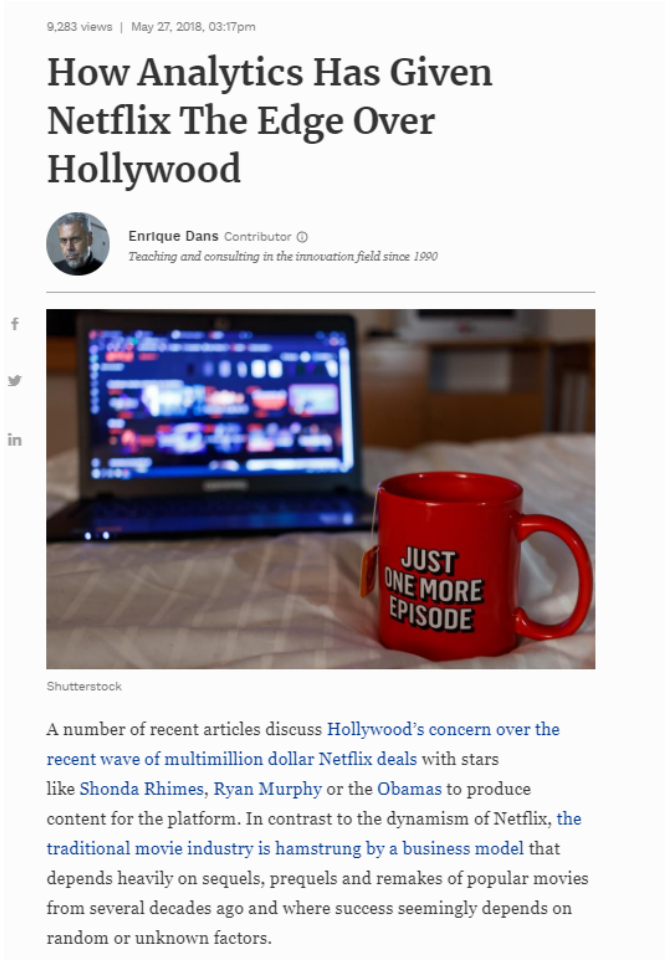


De 'heatmap' van Messi dit seizoen zet zijn genialiteit in een nieuw daglicht

PERFORMANCE

Netflix


NETFLIX
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Congratulations!
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.
FAQ | Forum | Netflix Home
© 1997-2009 Netflix, Inc. All rights reserved.


9,283 views | May 27, 2018, 03:17pm
How Analytics Has Given Netflix The Edge Over Hollywood
Enrique Dans Contributor
Teaching and consulting in the innovation field since 1990
JUST ONE MORE EPISODE
Shutterstock
A number of recent articles discuss Hollywood's concern over the recent wave of multimillion dollar Netflix deals with stars like Shonda Rhimes, Ryan Murphy or the Obamas to produce content for the platform. In contrast to the dynamism of Netflix, the traditional movie industry is hamstrung by a business model that depends heavily on sequels, prequels and remakes of popular movies from several decades ago and where success seemingly depends on random or unknown factors.

Netflix Challenge

https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf

Price Oktober 2006: Netflix beloont het eerste team dat het eigen recommenderalgoritme met 10% kan verbeteren. De beloning bedraagt $1M ! Zolang die 10% verbetering niet gehaald wordt, wordt er jaarlijks een bedrag uitgereikt van $10K.

Training data Volgende data werd ter beschikking gesteld :

- meer dan 100M ratings : ongeveer 500,000 anonieme klanten en hun score van ongeveer 17,770 films. Elke film werd gescoord op een schaal van 1 tot 5 sterren.
- In totaal 6 jaar aan data uit de periode 2000-2005

Test data Deelnemende teams moesten een score voorspellen voor een test set van ongeveer 3M. Gebruik makende van de echte (niet verspreide) scores van de test-set berekende Netflix een error (root-mean-square error RMSE).

Netflix training en test data

user	movie	score	date
1	21	1	2002-01-03
1	213	5	2002-04-04
2	345	4	2002-05-05
2	123	4	2002-05-05
2	768	3	2003-05-03
3	76	5	2003-10-10
4	45	4	2004-10-11
5	568	1	2004-10-11
5	342	2	2004-10-11
5	234	2	2004-12-12
6	76	5	2005-01-02
6	56	4	2005-01-31

user	movie	score	date
1	212	?	2003-01-03
1	1123	?	2002-05-04
2	25	?	2002-07-05
2	8773	?	2002-09-05
2	98	?	2004-05-03
3	16	?	2003-10-10
4	2450	?	2004-10-11
5	2032	?	2004-10-11
5	9098	?	2004-10-11
5	11012	?	2004-12-12
6	664	?	2005-01-02
6	1526	?	2005-01-31

$$\hat{r}_{ui} = q_i^T p_u$$

De geschatte score $\hat{r}_{ui}$ van gebruiker u van film i uitgedrukt als dot product tussen de 2 vectoren in de gedeelde feature ruimte die gebruikers en films karakteriseren. Hoe meer overlap tussen de twee, hoe groter de score.

Methods

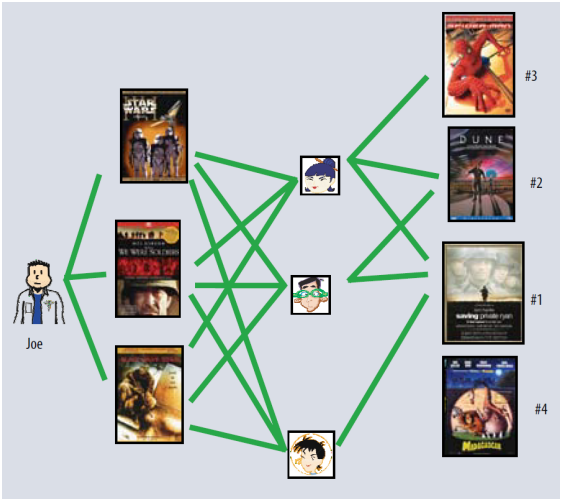


Figure 1. The user-oriented neighborhood method. Joe likes the three movies on the left. To make a prediction for him, the system finds similar users who also liked those movies, and then determines which other movies they liked. In this case, all three liked *Saving Private Ryan*, so that is the first recommendation. Two of them liked *Dune*, so that is next, and so on.

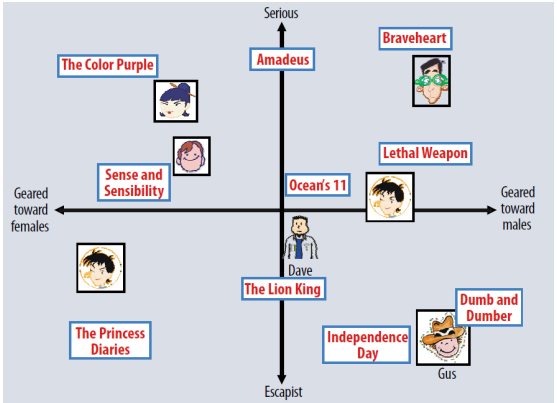


Figure 2. A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist.

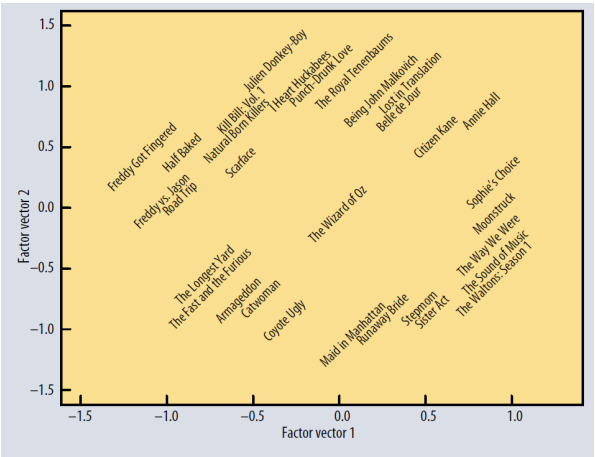


Figure 3. The first two vectors from a matrix decomposition of the Netflix Prize data. Selected movies are placed at the appropriate spot based on their factor vectors in two dimensions. The plot reveals distinct genres, including clusters of movies with strong female leads, fraternity humor, and quirky independent films.

Leertaak : Bepaal de vectoren p_u en u_i zodat de error tussen de geschatte score en werkelijk score minimaal is :

$$min_{q*,p*} \sum_{(u,i)r_{ui}gekend} (r_{ui} - q_i^T p_u)^2 + \lambda(||q_i||^2 + ||p_u||^2)$$

De laatste term is een regularizatie om overfitting te vermijden.

Het gevaar van (big) data

MICROSOFT \ WEB \ TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

60

By James Vincent | Mar 24, 2016, 6:43am EDT

SHARE

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Twitter bot that the company described as an experiment in

Het gevaar van (big) data

Onderzoekers van de universiteit van Houston bestudeerden Google Flu Trends. Deze dienst monitort de griep aan de hand van zoekopdrachten. Wanneer mensen bijvoorbeeld zoeken op symptomen van het griepvirus mag men ervan uitgaan dat de griep in die omgeving heerste. Op basis van die zoekopdrachten kan Google Flu Trends inzicht geven in waar de griep toe heeft geslagen en of de epidemie toeneemt of afneemt. Maar de dienst overschatte de griep herhaaldelijk, hij bleek gewoon winterseizoenen te voorspellen en miste de enorme piek van griep in 2013 helemaal. Googles algoritme was oa. vatbaar voor overfitting door seizoensgebonden termen die niet gerelateerd werden door griep zoals *high school basketball.*

Het gevaar van (big) data

from Alexa

Bezos: Alexa, buy me something from Whole Foods.
Alexa: Buying Whole Foods.
Bezos: Wait, what?

Scottish Voice Recognition Elevator

ELEVEN! : `https://www.youtube.com/watch?v=MNuFcIRlwdc`

Data Science

Data science is een overkoepelende term die het proces aanduidt om structuur te brengen in (big) data , patronen erin te ontdekken en na grondige data-analyse (strategische) beslissingen te ondersteunen en inzichten helpen te verwerken.

Big Data

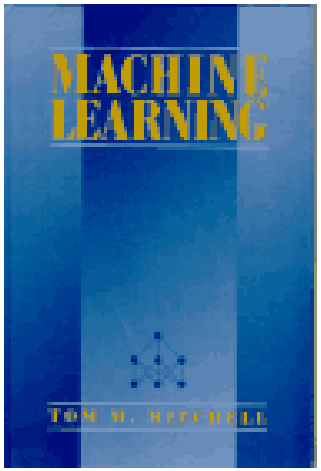
Big Data zijn enorme hoeveelheden aan niet gestructureerde data afkomstig van verschillende bronnen, die typisch niet verwerkt kunnen worden gebruik makend van traditionele applicaties.
Big data storage tools (Hadoop, Greenplum, MapReduce, etc.)

Machine learning : Kan een computer programma echt leren?

Figure: (from Eric Xing lectures on ML)

Definitie van leren

Wanneer een computerprogramma door ervaring **E**, steeds beter wordt in een bepaalde taak **T** volgens een gemeten performantiewaarde **P**, dan zegt men dat dit programma leert uit zijn ervaring E.
(Mitchell, Machine Learning, McGraw-Hill, 1997)



leren is bvb patronen herkennen

Of anders, kan een programma een patroon herkennen zoals mensen dat doen? Voor mensen is het eenvoudig om bvb:

- gezichten te herkennen
- concepten te herkennen (welke voertuigen zijn bussen bv.)
- gesproken taal / woorden te herkennen
- een handschrift te herkennen / lezen
- herkenning of een credit card transactie frauduleus is of niet
- ...

Herkennen van handgeschreven karakters

- Taak T: herken handgeschreven karakters (dit is een **classificatietaak**)

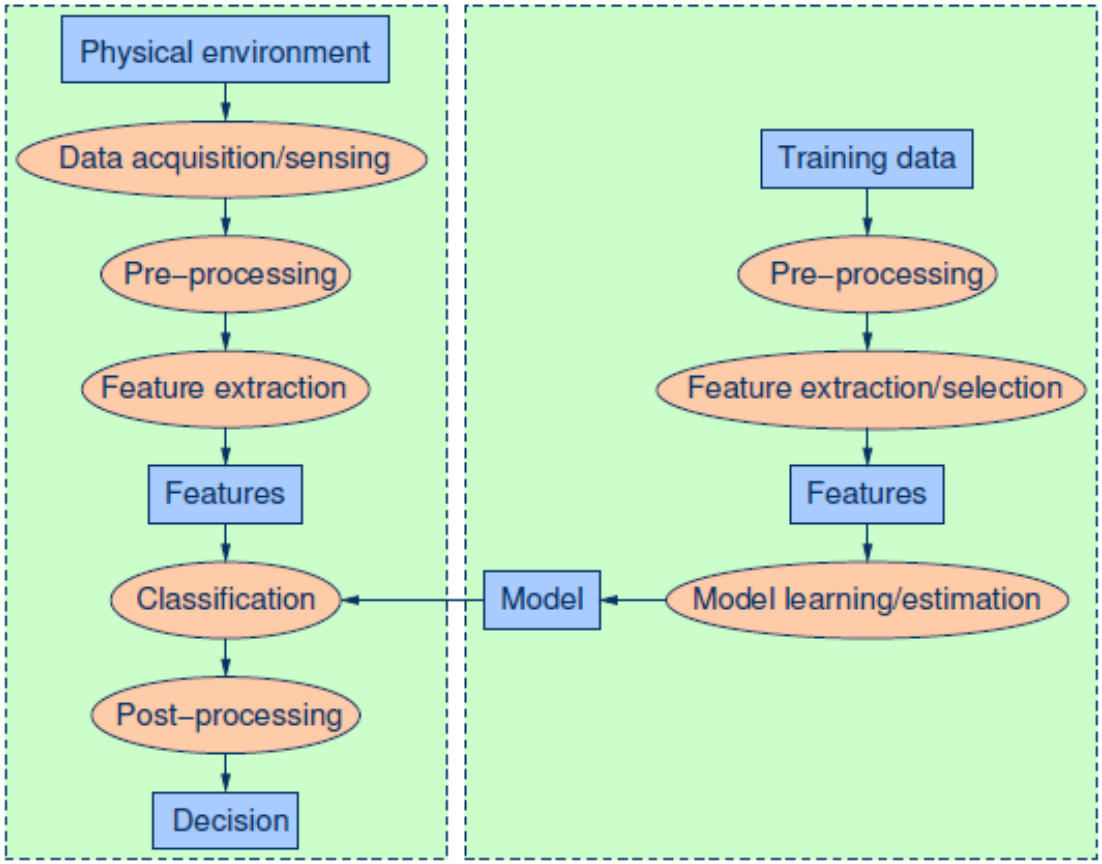
Herkennen van handgeschreven karakters

- Taak T: herken handgeschreven karakters (dit is een **classificatietaak**)
- Metriek P: wat is het percentage van de karakters die correct geclassificeerd werden?

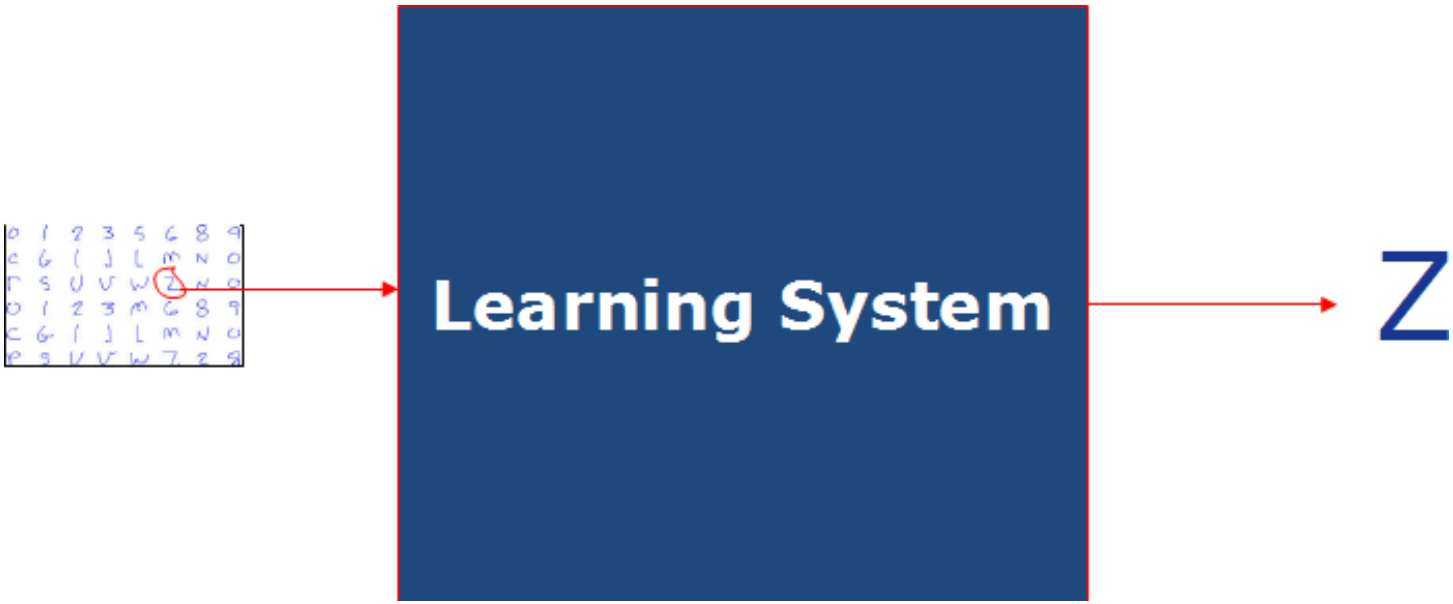
Herkennen van handgeschreven karakters

- Taak T: herken handgeschreven karakters (dit is een **classificatietaak**)
- Metriek P: wat is het percentage van de karakters die correct geclassificeerd werden?
- Training data E: een databank van handgeschreven karakters waarvan de klassificatie gekend is

Onderdelen van een patroonherkenningssysteem

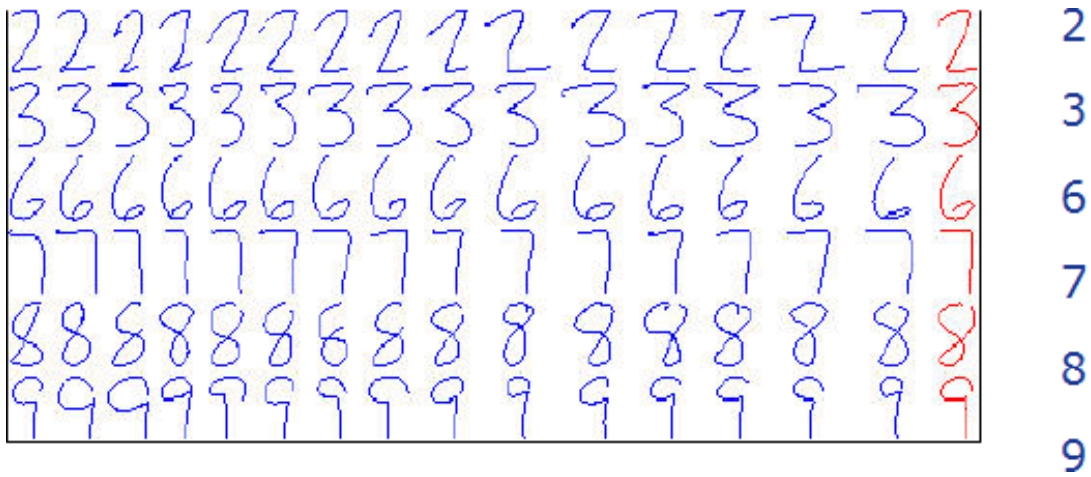


Stap 0



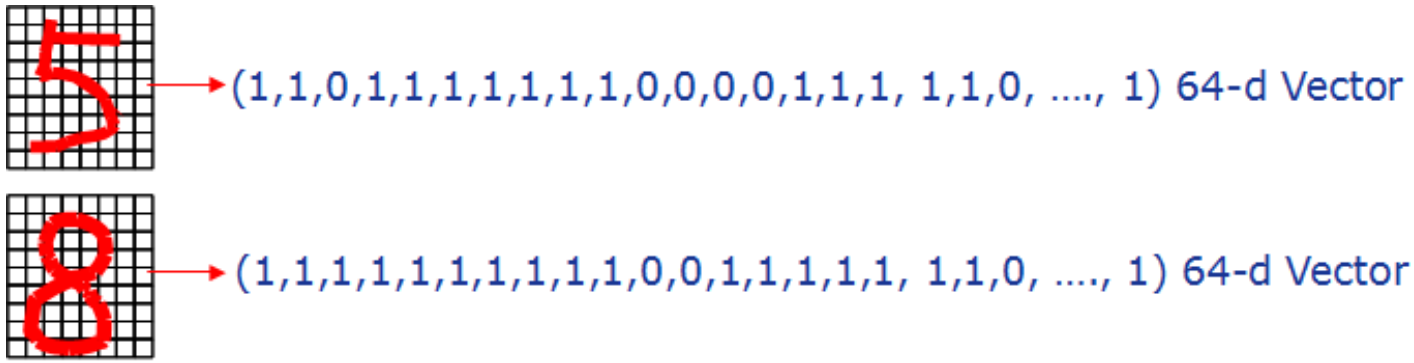
Het leersysteem is een `Black Box` die we zullen gaan benaderen

Stap 1 : verzamelen van trainingsdata



Leren gebeurt hier aan de hand van vele voorbeelden.

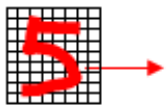
Stap2 : Kies een gepaste representatie



Welke eigenschappen (`attributen` of `features`) zijn van belang om een voorbeeld weer te geven of te beschrijven? De verschillen in deze attributen zullen aanleiding geven tot verschillende patronen en klassificaties.

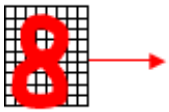
Stap2 : Kies een gepaste representatie

D = (d0, d1, d2, d3, d4, d5, d6, d7, d8, d9)



X = (1,1,0,1,1,1,1,1,1,1,0,0,0,0,1,1,1, 1,1,0, …., 1); 64-d Vector

D= (0,0,0,0,0,1,0,0,0,0)

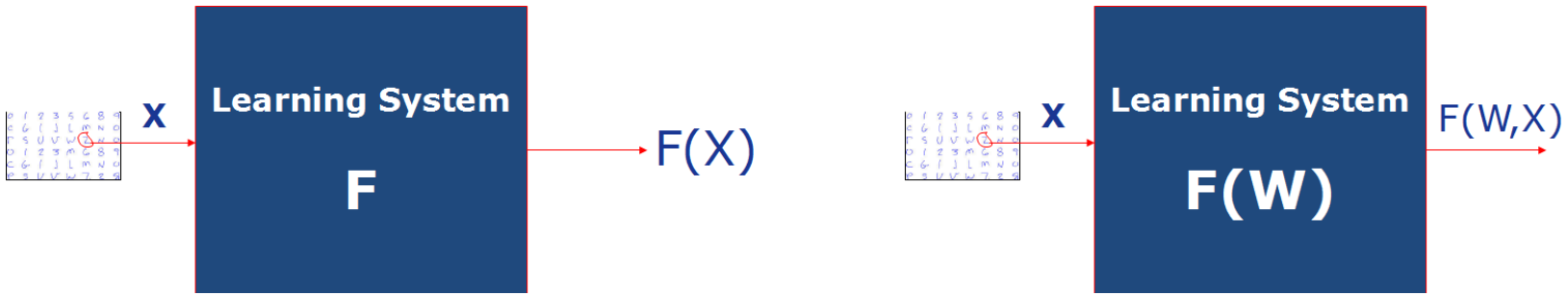


X= (1,1,1,1,1,1,1,1,1,1,0,0,1,1,1,1,1, 1,1,0, …., 1); 64-d Vector

D= (0,0,0,0,0,0,0,0,1,0)

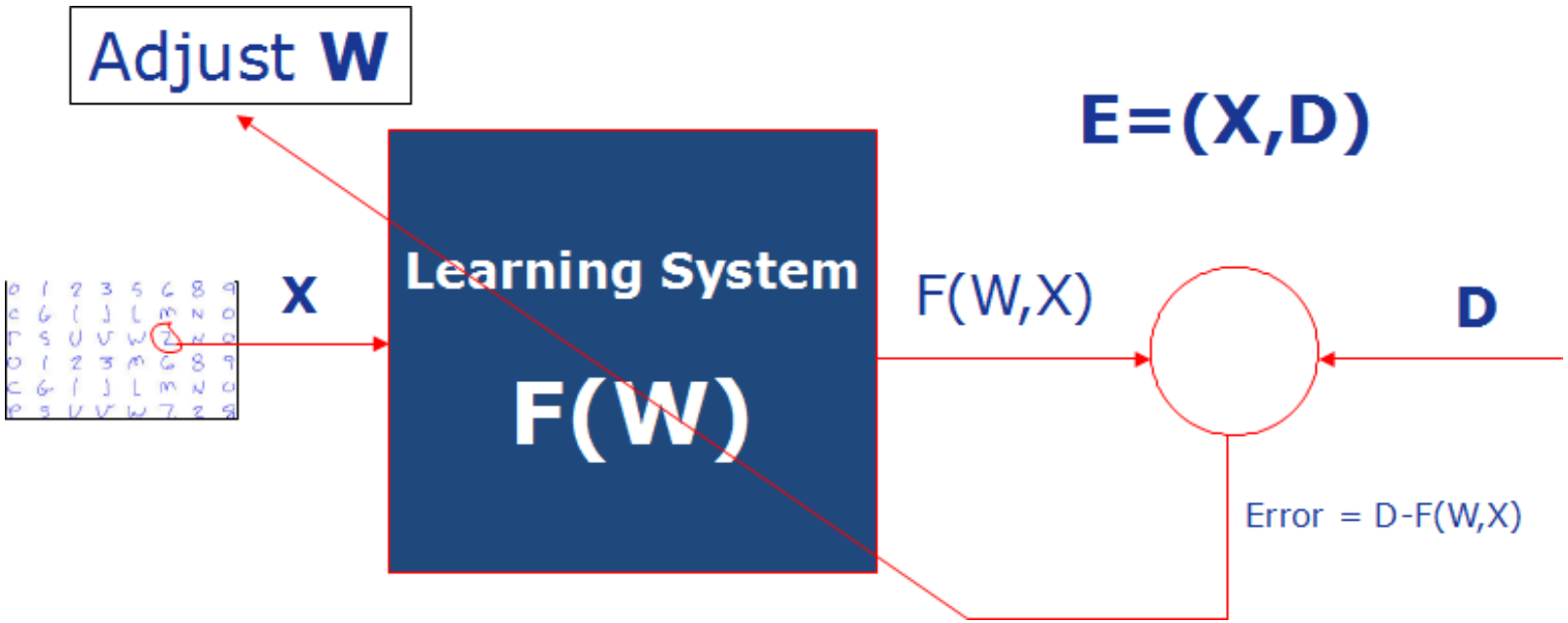
Ook de ouput moet beschreven worden. Van elke ervaring (trainingsvoorbeeld) moeten we kennis hebben over de classificatie.

Stap 3 : Kies een leeralgoritme



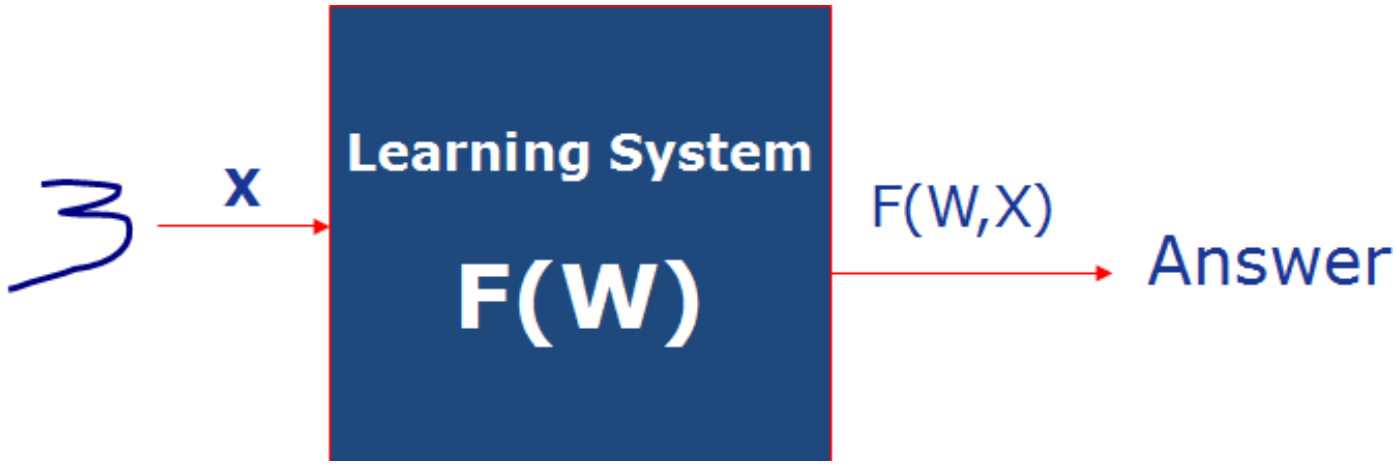
Probeer je blackbox te vervangen door een hypothese. Door een geparametrizeerd leeralgoritme te kiezen kan je de parameter zo tunen dat de classificatie beter wordt.

Stap 3 : Kies een leeralgoritme



Laat het algoritme leren.

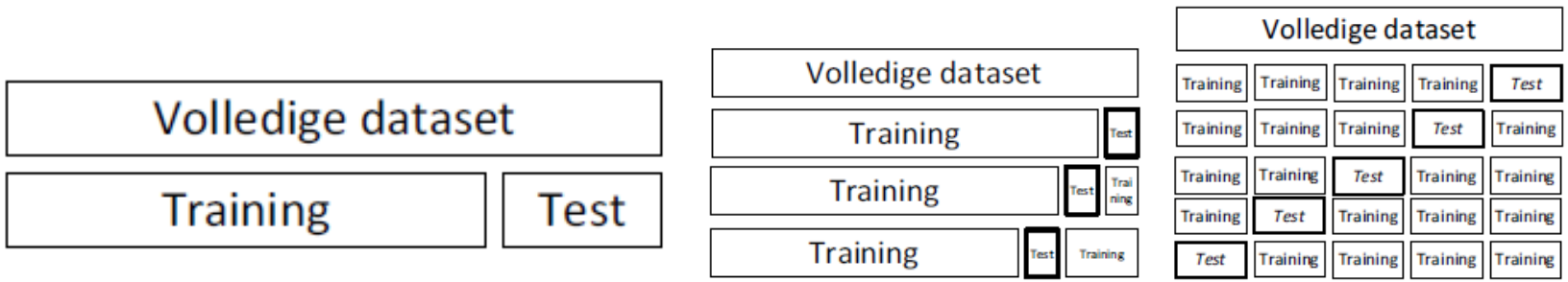
Stap 4 : Test je systeem



Hoe goed doet je systeem het op je trainingsdata?
Hoe goed presteert je systeem op ongeziene data?

Testen en Valideren

Om te kunnen voorspellen hoe goed het geleerde model presteert op ongeziene data wordt typisch slechts een deel van de data gebruikt om te trainen (typisch 2/3de). De rest van de data gebruik je om te testen. Wanneer het model ook nog moet leren welke parameterinstellingen de beste zijn, ga je ook een validatieset uit de trainingsdata halen.



Opdeling van de data : 2/3 als trainingsset, 1/3 testset - leave-one-out : de testset bestaat telkens slechts uit 1 element - cross-validatie : deel je data op in partities, kies telkens 1 partitie als test set

Welke vormen van leren zijn er?

Welke vormen van leren zijn er?

- Supergeviseerd leren : leren aan de hand van een trainer / leraar. De trainingsdata is voorzien van de juiste categorie

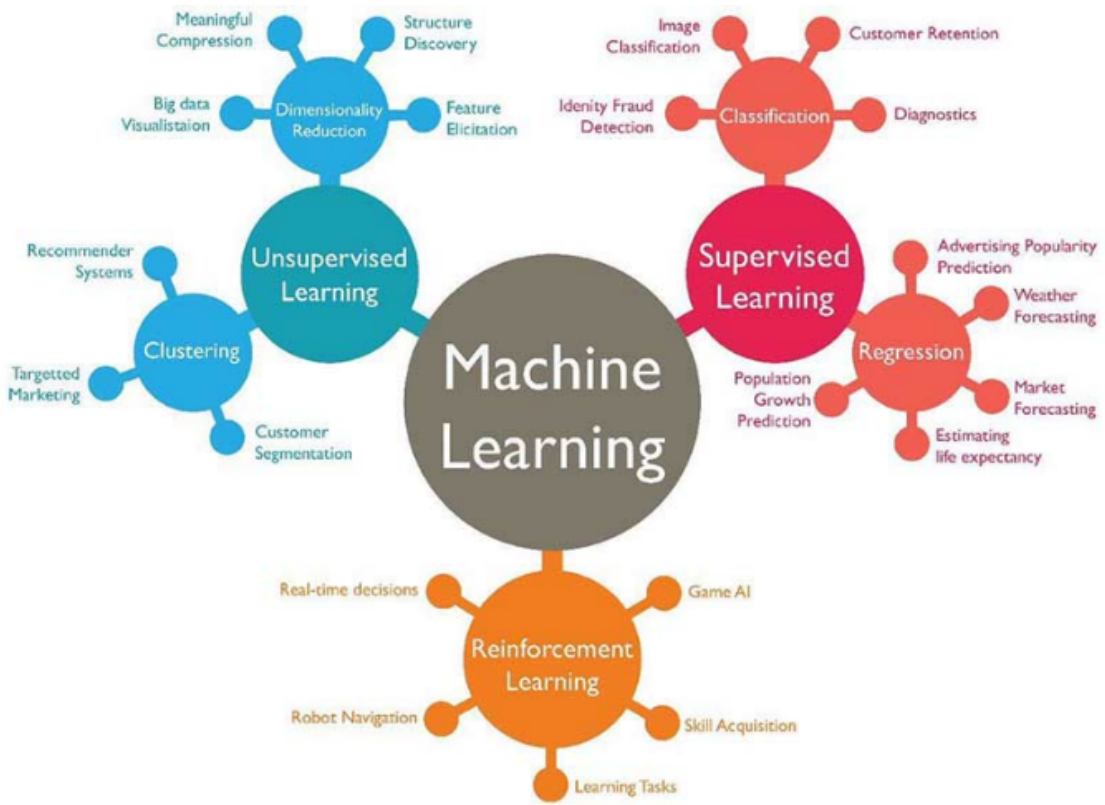
Welke vormen van leren zijn er?

- Supergeviseerd leren : leren aan de hand van een trainer / leraar. De trainingsdata is voorzien van de juiste categorie
- Niet gesuperviseerd leren : er is geen kennis over de trainingsdata. De data kan gegroepeerd of geclusterd worden volgens de gelijkenissen van de input features.

Welke vormen van leren zijn er?

- Supergeviseerd leren : leren aan de hand van een trainer / leraar. De trainingsdata is voorzien van de juiste categorie
- Niet gesuperviseerd leren : er is geen kennis over de trainingsdata. De data kan gegroepeerd of geclusterd worden volgens de gelijkenissen van de input features.
- Reinforcement leren : leren om acties te nemen in een ongekende omgeving aan de hand van een signaal uit de omgeving

Machine Learning



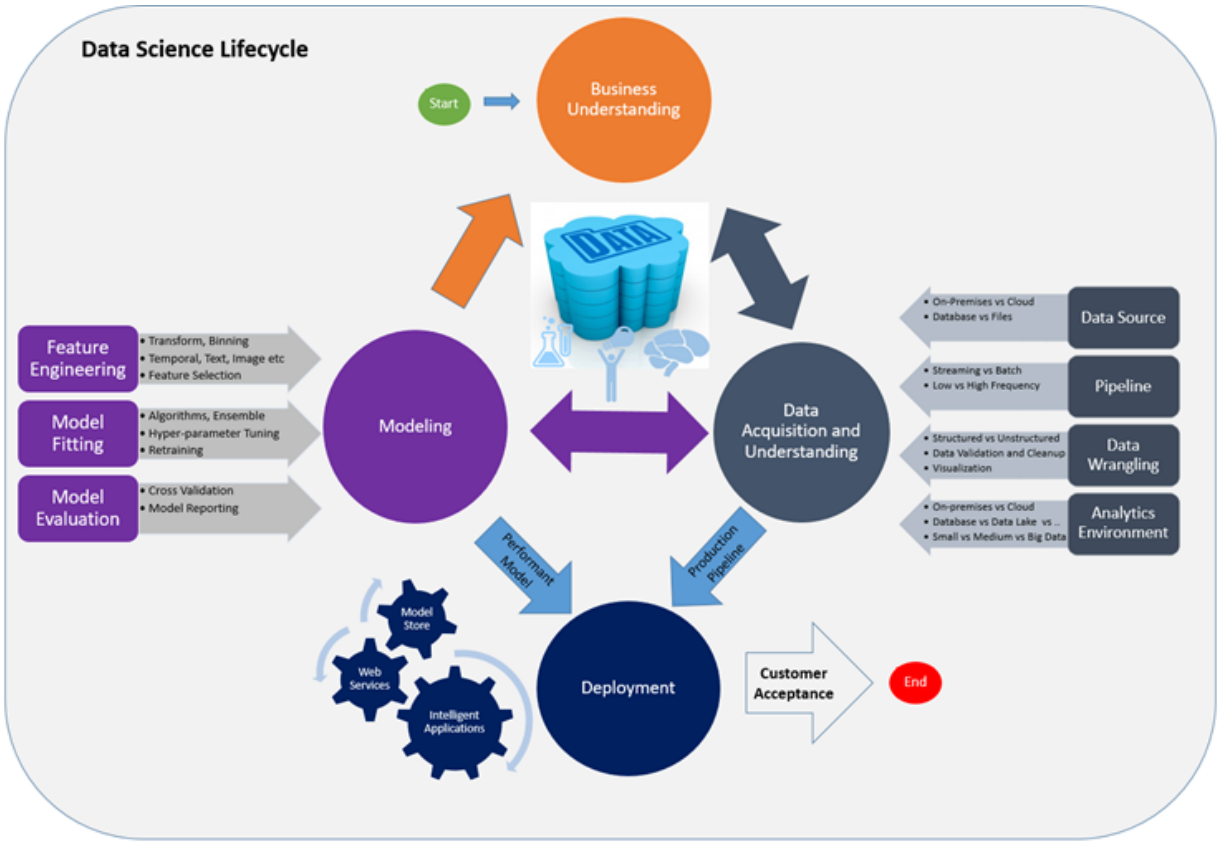
datavibes.nl

Data Mining en analyse

Data Mining is het process waarbij data gecollecteerd wordt en patronen in die data gezocht wordt. Hierbij worden Machine learning algoritmen toegepast op (big) data. Binnen het volledige data science process is dit de eerste stap.

Data Analyse duidt op die fase in het proces waarbij de resultaten van de data mining geanalyzeerd worden. Dankzij deze analyse kan de data scientists ondersteuning bieden naar het nemen van strategische beslissingen en het geven van inzichten in deze beslissingen.

Samengevat



https://blog.revolutionanalytics.com/2016/10/the-team-data-science-process.html

Python: The Meaning of Life in Data Science

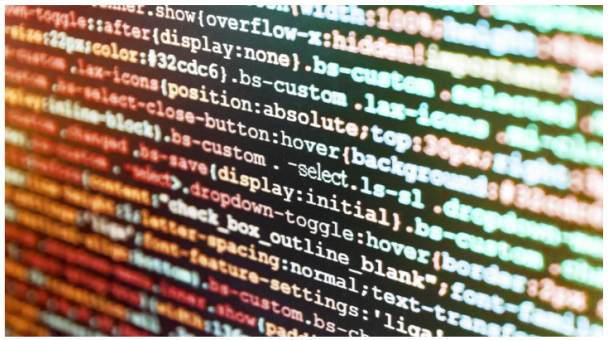
TECHZINE Secties Gerelateerde artikelen Recent in Cloud & Data

Cloud & Data

Populaire programmeertaal Python meest gebruikt voor data science

Geplaatst op 8 Februari 2019 12:57 door Zamire Willems

Home » Nieuws » Cloud & Data » Populaire programmeertaal Python meest gebruikt voor data science



Wat mag je verwachten?

- wekelijkse labo's + theorie
- Toledo tests tijdens het labo
- paper in jupyter notebook met volledige uitwerking van een DS probleem
- schriftelijk Toledo examen