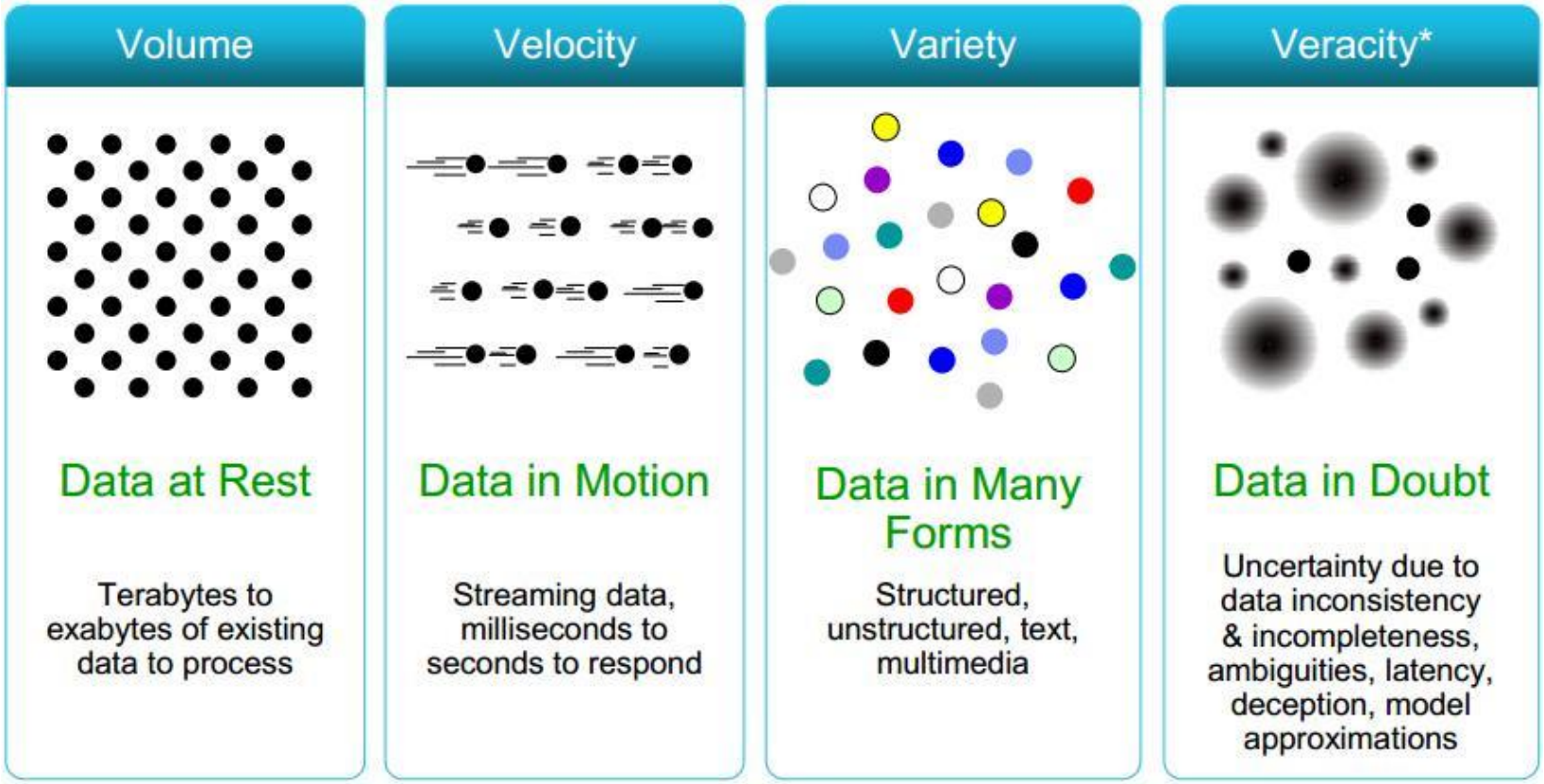
Microsoft Modern Data Platform

The world is changing

Big Data

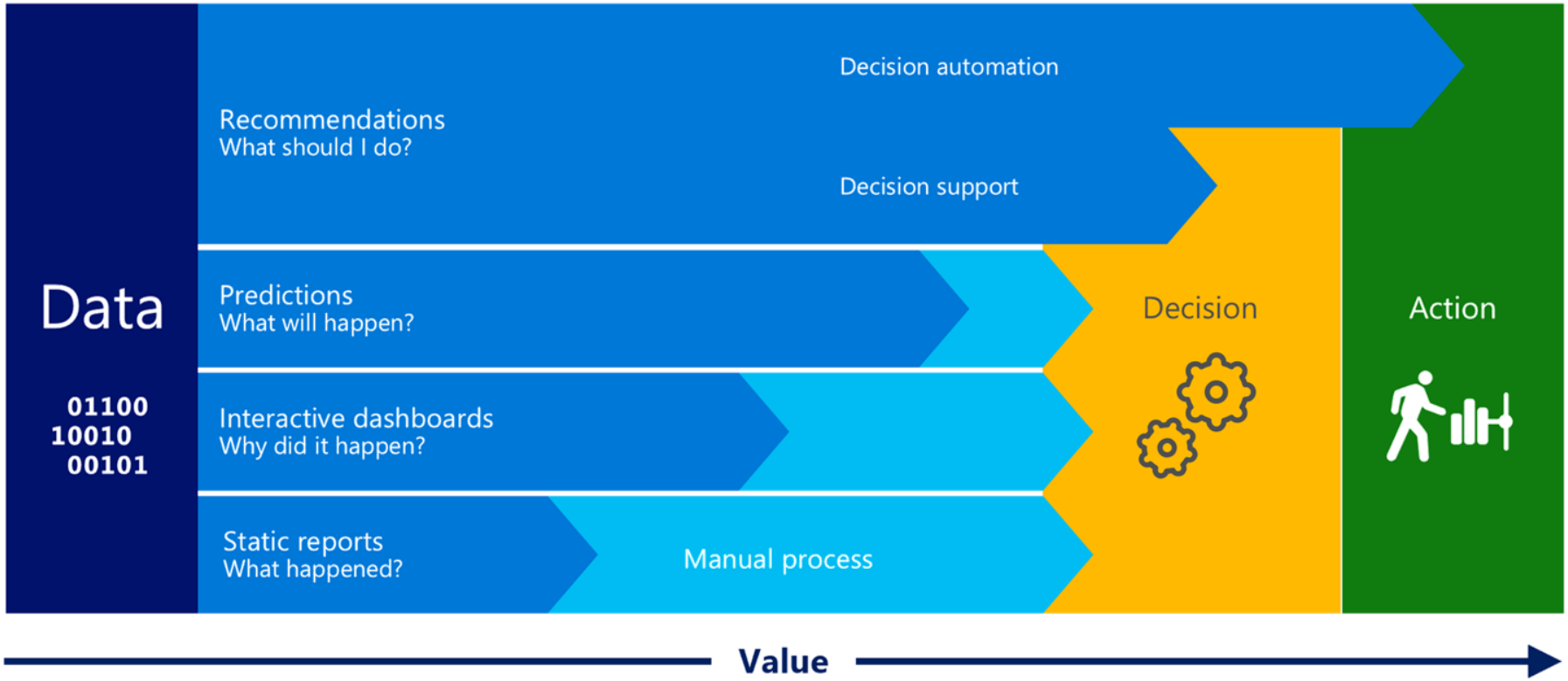
Grote uitdagingen, groot potentieel



Bron: http://www.rosebt.com/blog/data-veracity

Het moderne data platform draagt bij aan

Data gedreven beslissingen en acties



Azure Machine Learning

Powerful predictive analytics in Azure

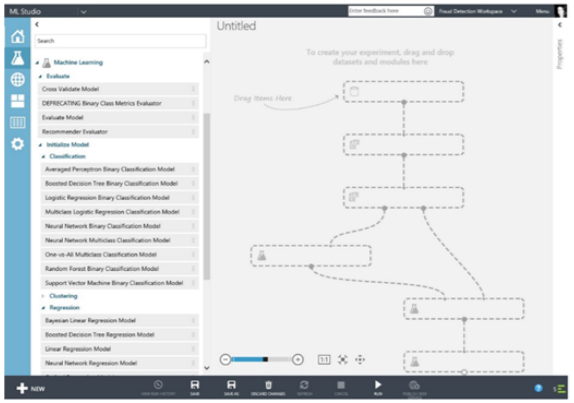
ML Algorithms are best of breed and embrace OSS
- MS + R + Python + BYOA

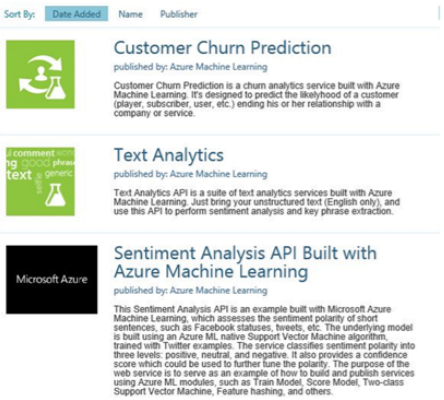
ML Studio for productive development
- Faster experiments results in faster improvements
- Visual Workflows & ML Experiments at Cloud Scale

ML Operationalization to remove deployment friction
- Build entire ML Apps & Deploy as Cloud APIs

ML Applications Marketplace
- Provide ML applications like apps in an 'app store'
- Publish/consume APIs in a 2 sided market

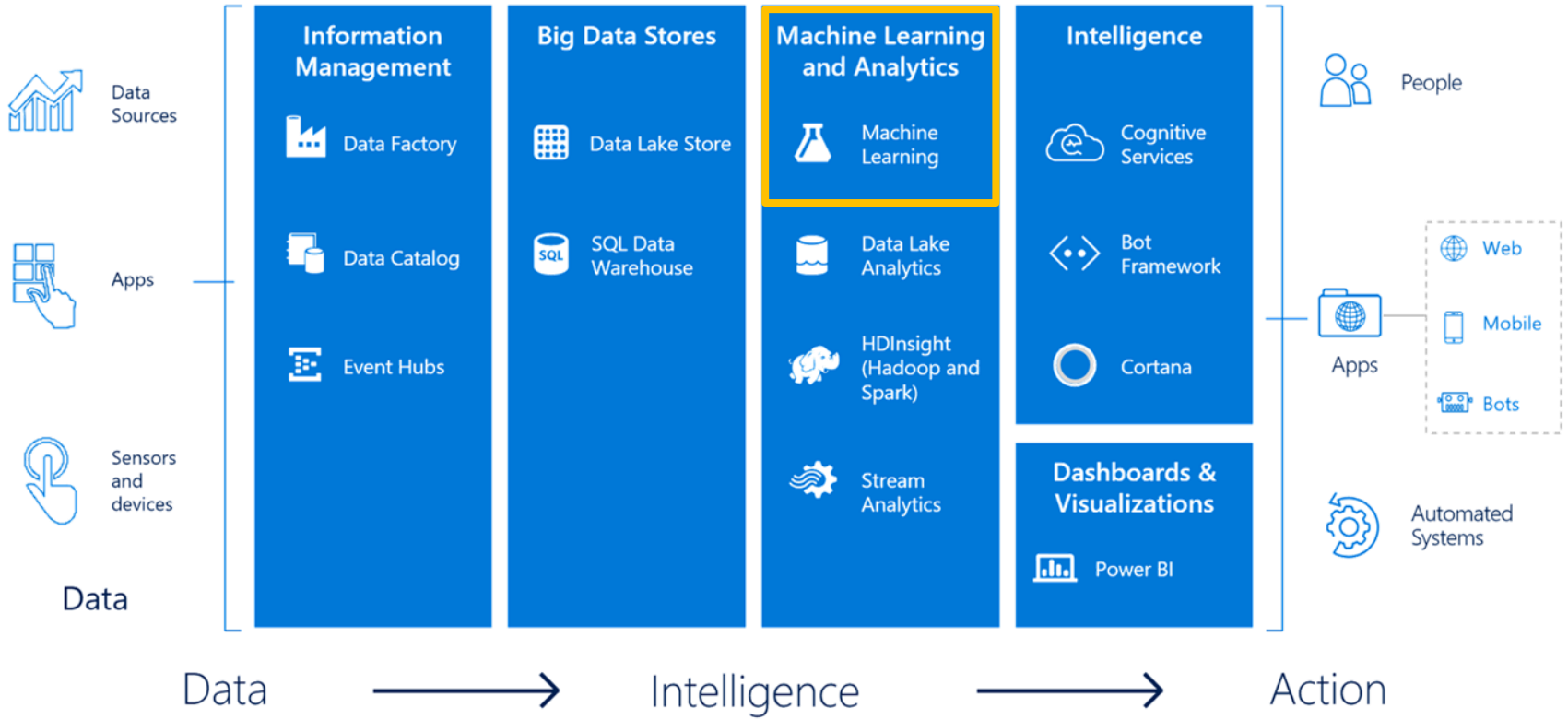
Help organizations eliminate undifferentiated heavy lifting





Cortana Intelligence Suite services

Building blocks of data intelligence

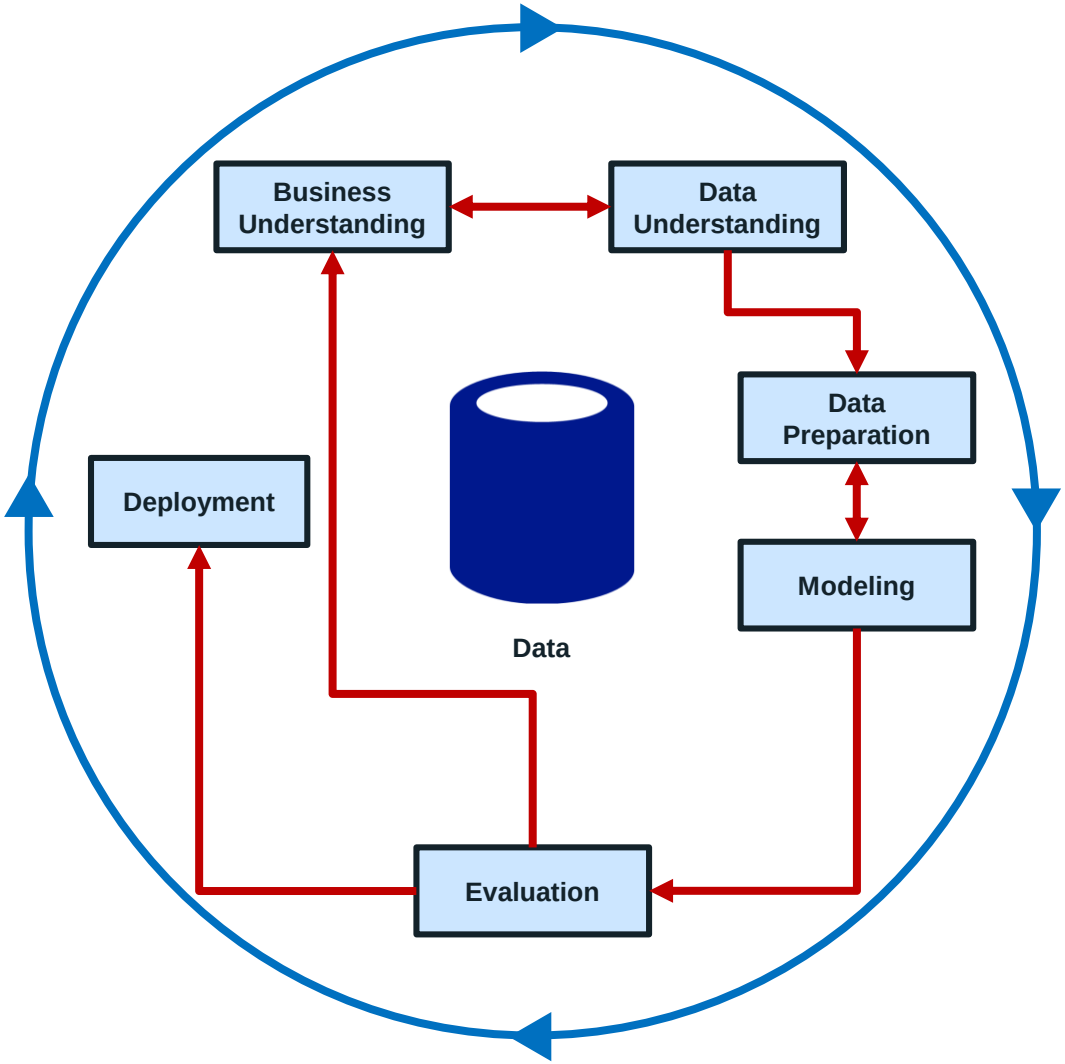


Data Science

Kerncompetenties

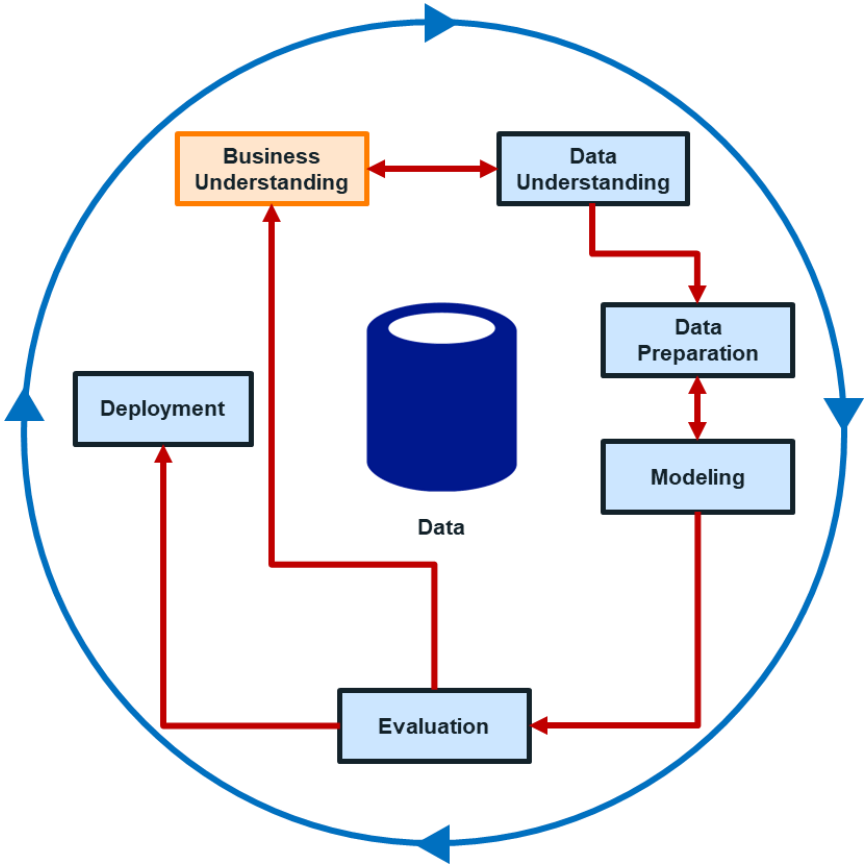
The CRISP data mining process

KDD + Business & Data Understanding



Business Understanding

Think about your problem



Business Understanding

Denk na over het probleem en over use cases!

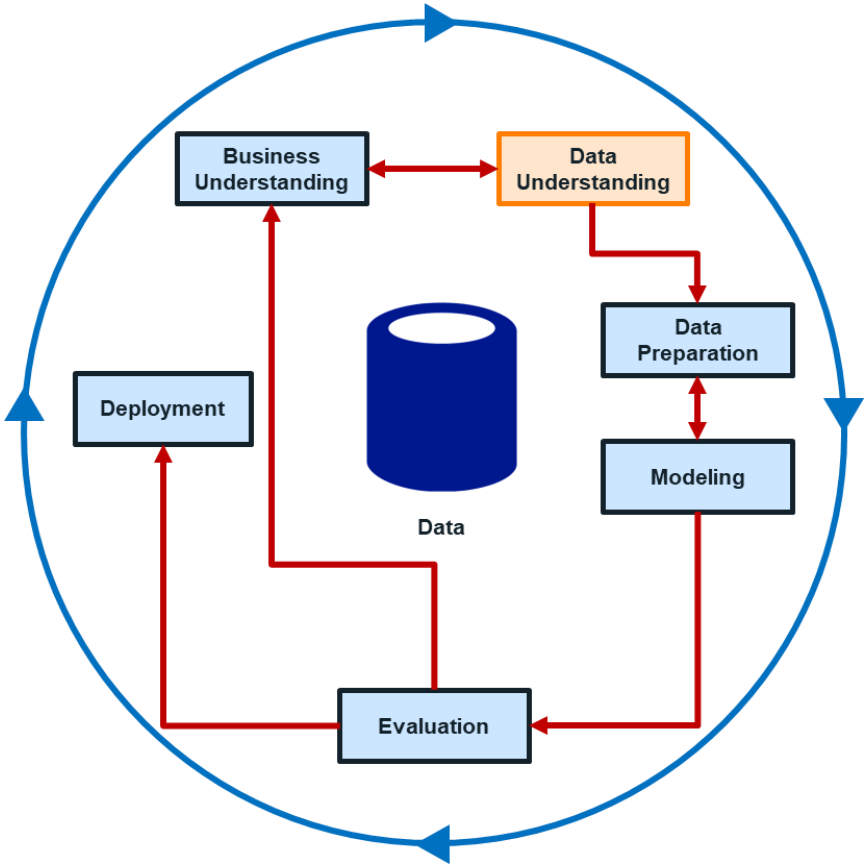
- **Welk probleem wil ik gaan oplossen (de wat)?**

- **Welk algoritme past hierbij (de hoe)?**
  - Classification / class probability estimation?
  - Regression (“value estimation”)?
  - Similarity matching?
  - Clustering?
  - Co-occurence grouping (frequent itemset mining, market-basket analyse)?
  - Profiling?
  - Link prediction?
  - Data reduction?
  - Causal modeling?

- **Welk gedeelte van de use cases dragen mogelijk bij aan het data mining model?**

Data Understanding

Think about your required data



Data Understanding

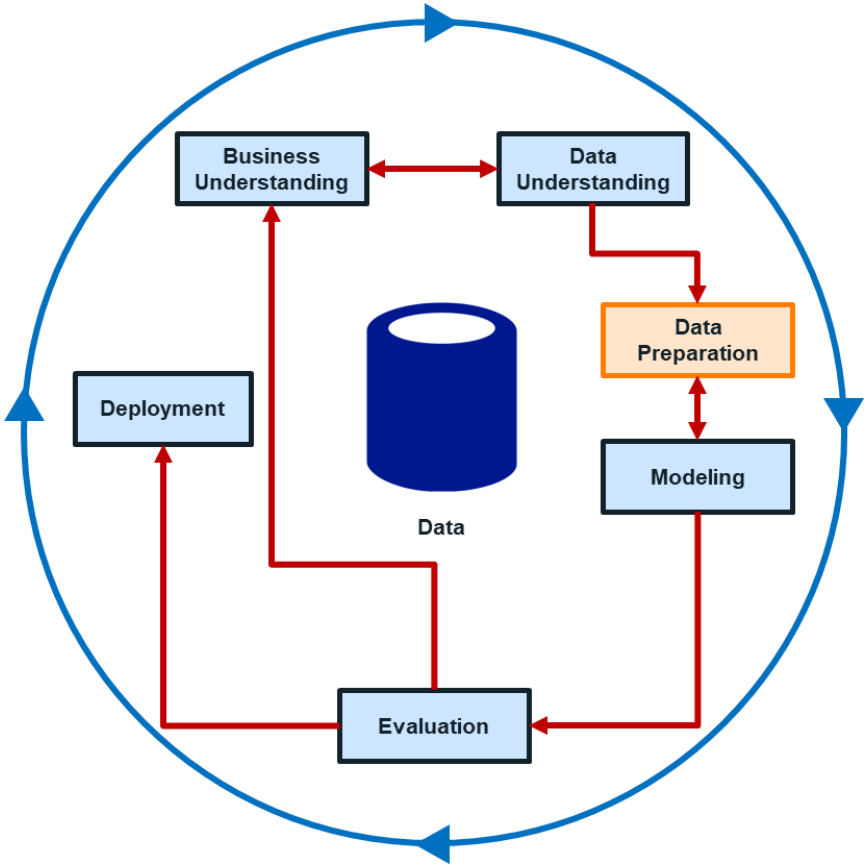
Denk na over de consequenties van de data!

Belangrijke vragen

- Kan ik met mijn huidige data mijn probleemstelling oplossen?
- Is mijn data volledig?
- In welke vorm ontvang ik de data?
- Waarin moet ik mijn data ontsluiten?
- Is mijn data gestructureerd of ongestructureerd?
- In welke hoeveelheden krijg ik mijn data binnen?
- Met welke frequentie ontvang ik mijn data?
- Is mijn data gratis beschikbaar, of moet ik hiervoor betalen?
- Mag ik deze data juridisch gezien wel verzamelen en/of verwerken (privacy/AVG)?
- Verzamel/verwerk ik mijn data voor ethisch verantwoorde redenen?

Data Preparation

Clean and structure your data



Data Preparation

Garbage in – garbage out

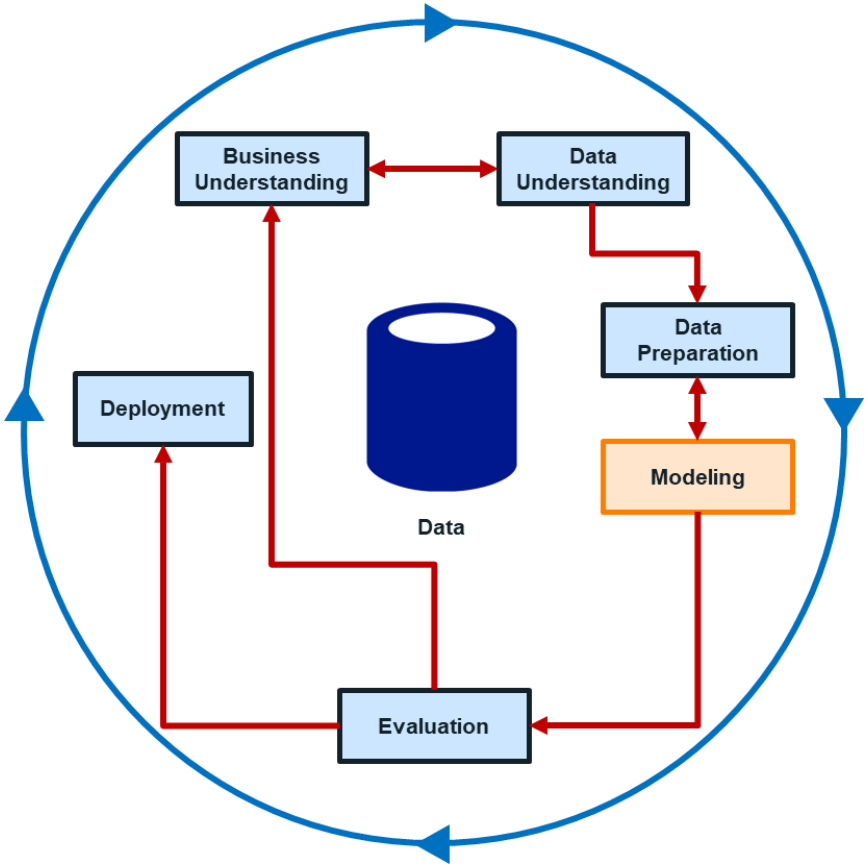
- **90% van de tijd zit hier**

- **Veelvoorkomende handelingen zijn:**
  - Data typen veranderen
  - Data converteren naar een tabelstructuur
  - Datamodellen bouwen vanuit meerdere bronnen
  - Data prepareren om hiervan een vector te genereren
  - Oplossen van data gerelateerde problemen

- **Veelvoorkomende problemen zijn:**
  - Missende en herhalende data (dubbele records)
  - Uitschieters (outliers) en foutmeldingen (#VALUE)
  - Schaalverdeling numerieke data (feature a: 1 t/m 10, feature b: 1000 t/m 5000)
  - Class imbalance (veel voorkomend issue bij classificeren van data)
  - Data lekkage (niet van het type AVG, hiermee wordt bedoeld dat een variabele binnen historische data informatie over de target variabele geeft)

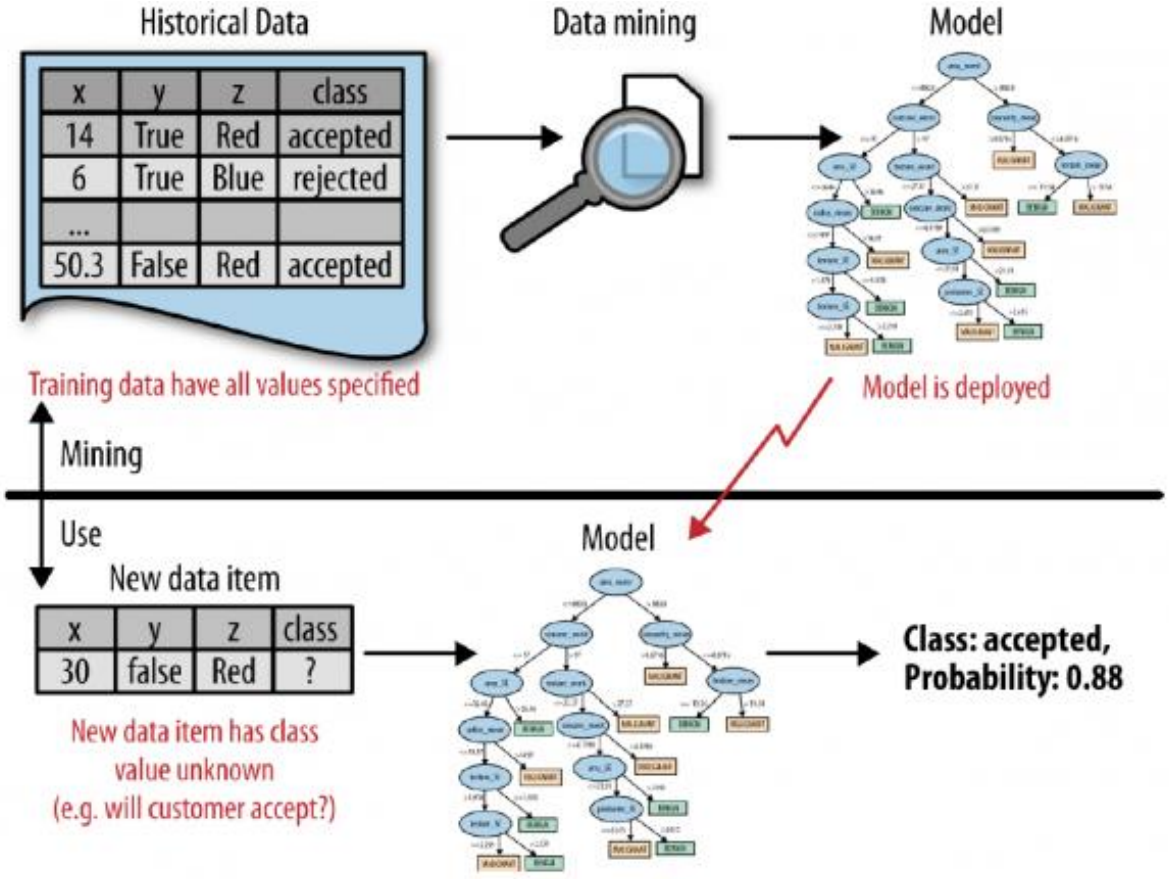
Modeling

Find patterns in your data



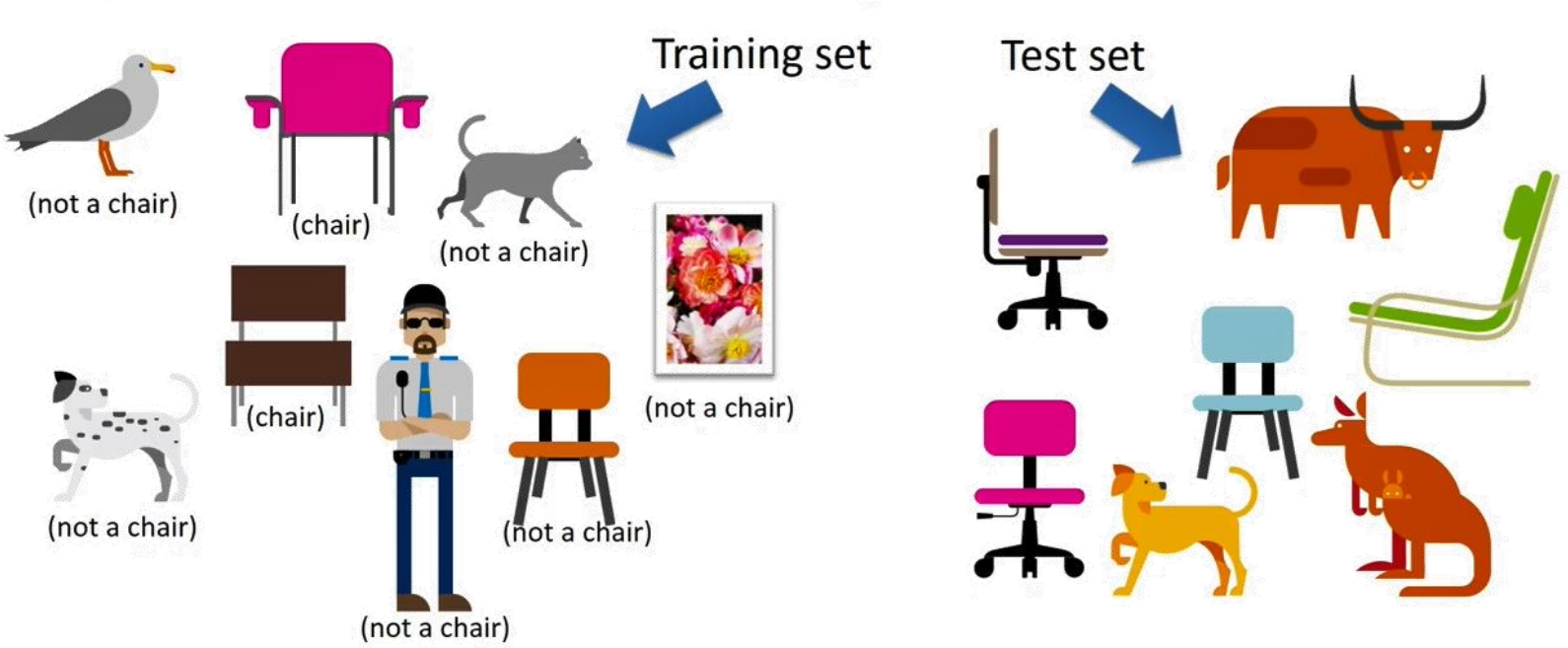
Modeleren

Classificeren, regressie en clusteren



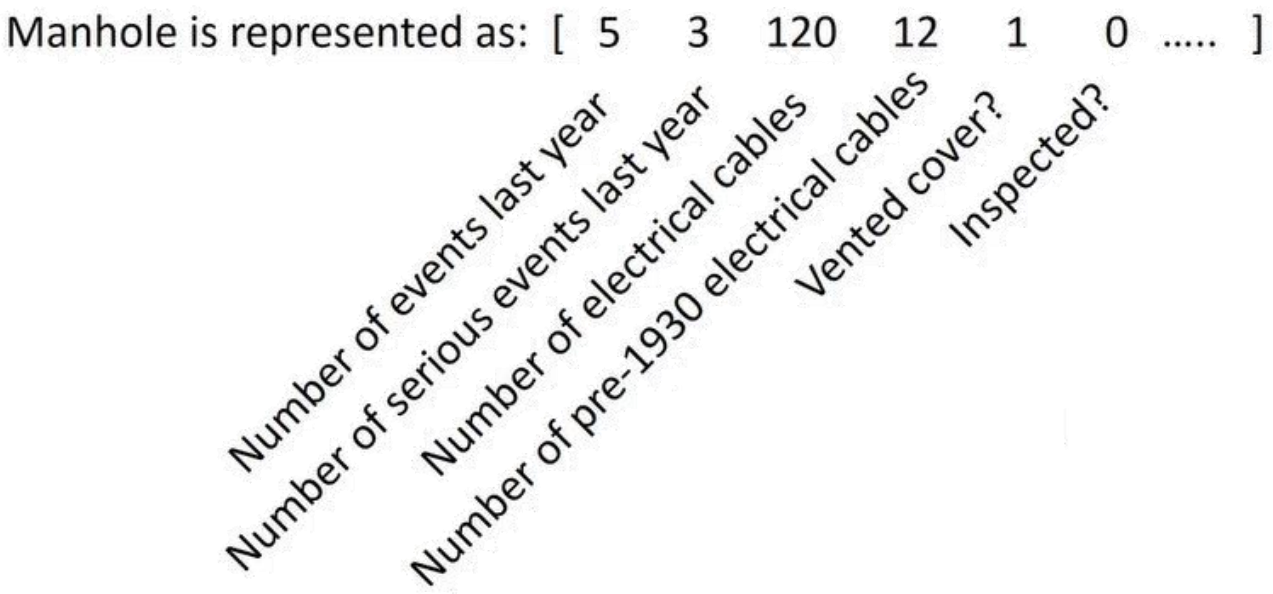
Classificeren

Voorspel antwoorden op ja/nee vragen



Classificeren

Hoe werkt het? 1/3



Classificeren

Hoe werkt het? 2/3

Manhole is represented as:

[	5	3	120	12	1	0	.....	] -1
[	0	0	89	5	1	1	.....	] 1
[	1	0	20	0	0	1	.....	] -1

Features, called X

Labels, called Y

(Predictors, Covariates,
Explanatory Variables,
Independent Variables)

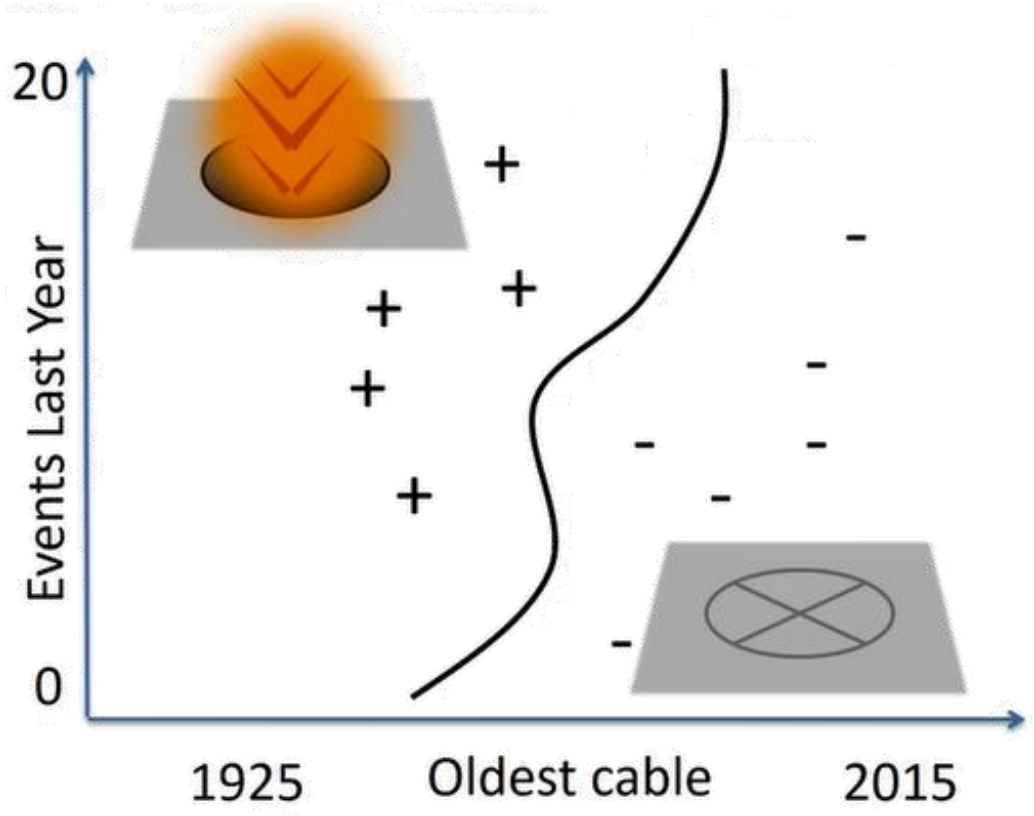
Classificeren

Hoe werkt het? 3/3

f(x) = function(Events Last Year, Oldest Cable)

Manhole is represented as: [1925 15]

Year oldest cable installed
Number of events last year



Regressie

Voorspel (real) numerieke waarden

A person is represented as: [5]
[0]
[1]
:

Single feature, called X

Income

84
32
-10

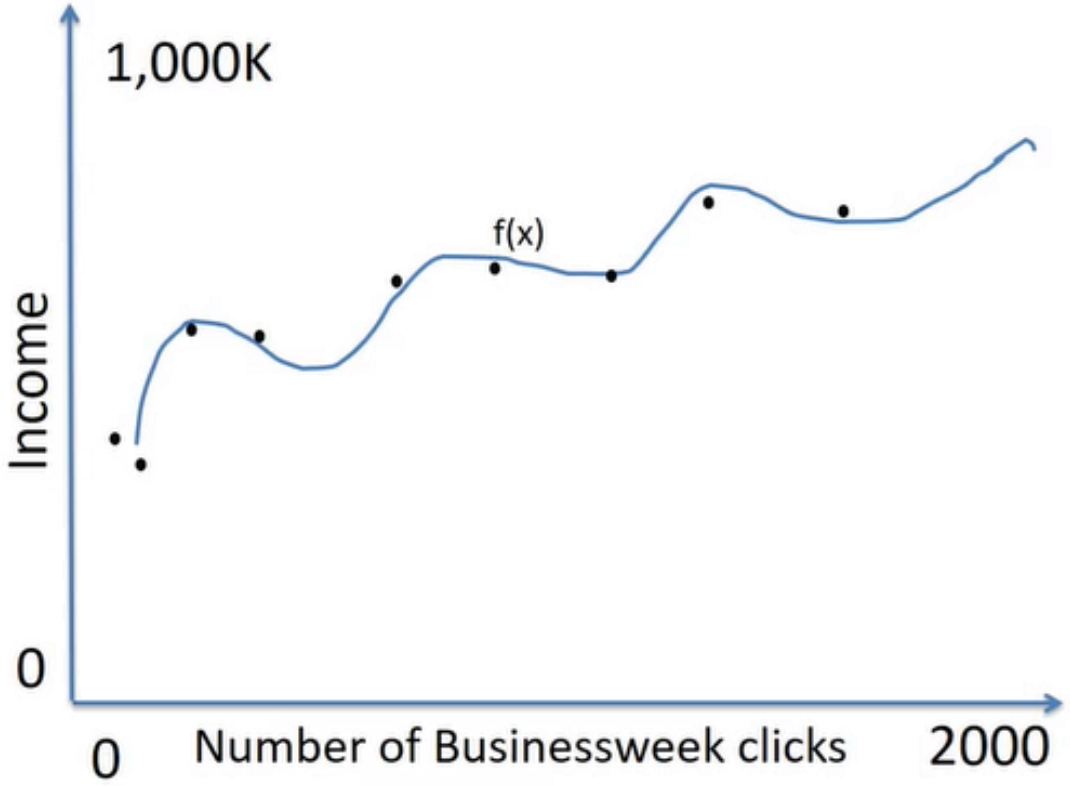
Labels, called Y

Regressie

Hoe werkt het? 1/3

f(x) = function(Number of Businessweek clicks)

(Overfitting?)

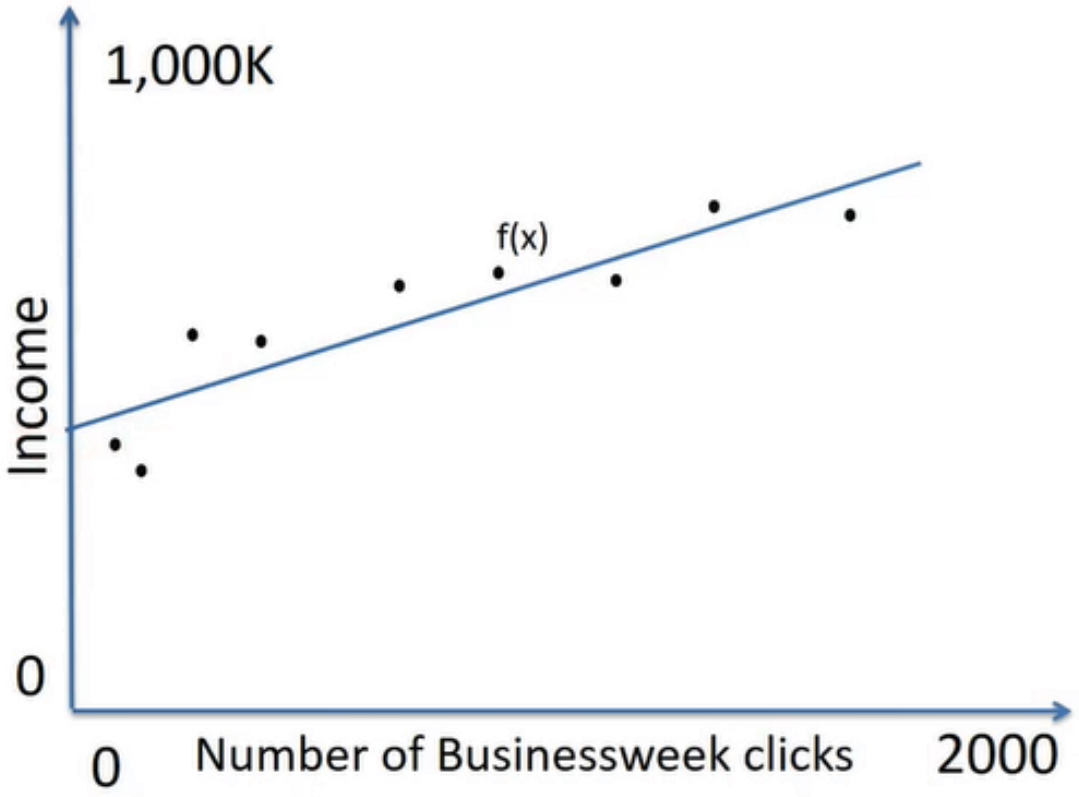


Regressie

Hoe werkt het? 2/3

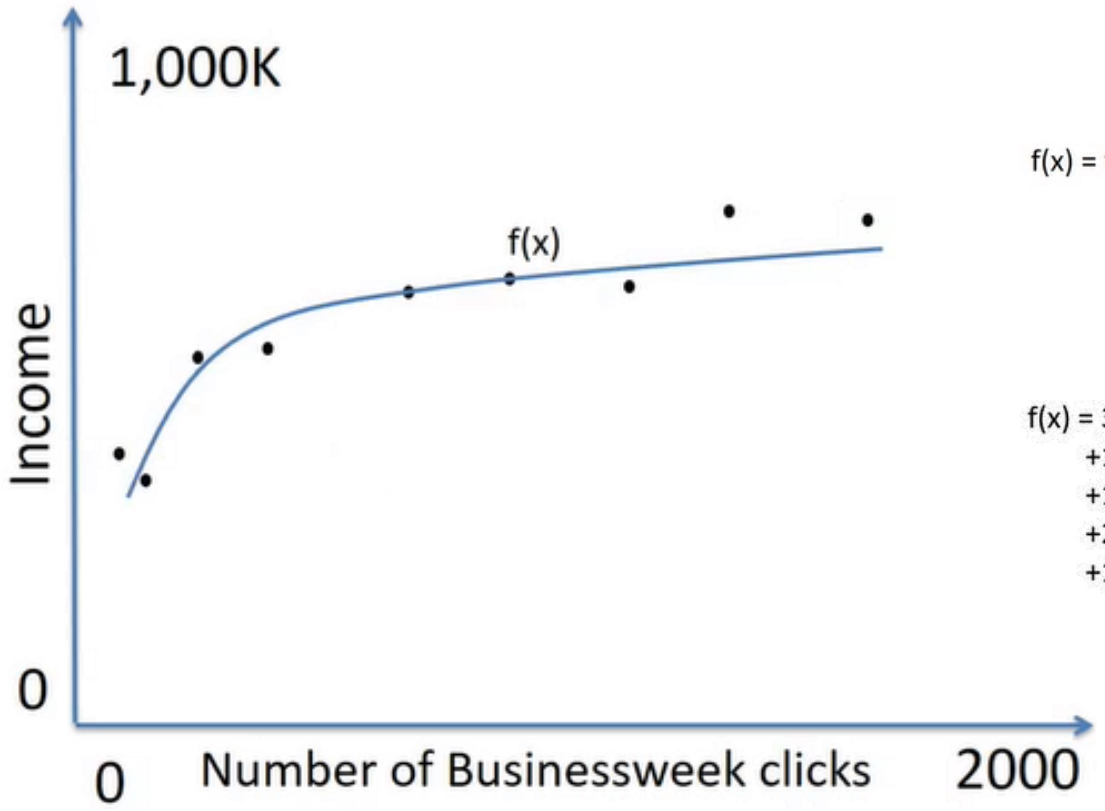
f(x) = function(Number of Businessweek clicks)
= 5K*Number of Businessweek clicks + 100K

(Underfitting?)



Regressie

Hoe werkt het? 3/3



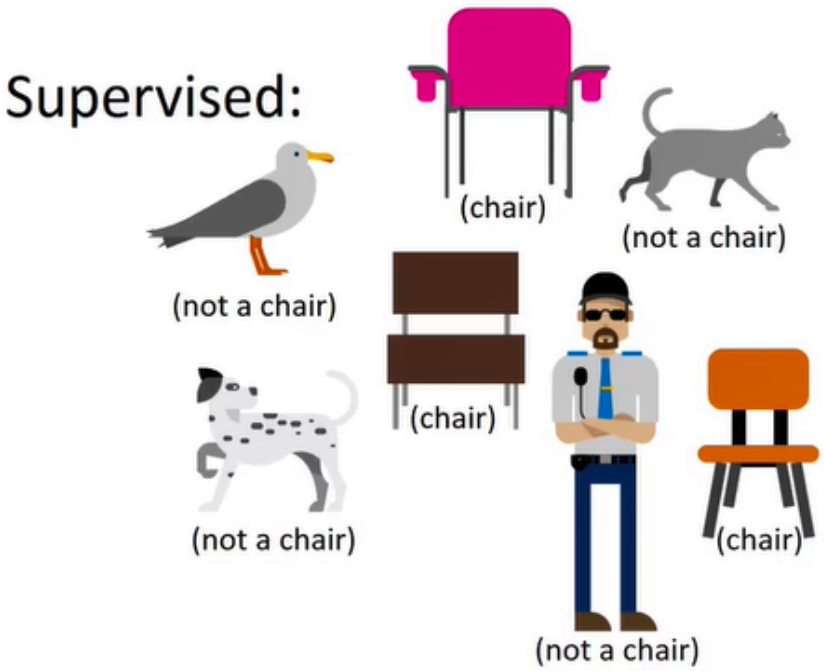
f(x) = function(Number of Businessweek clicks)

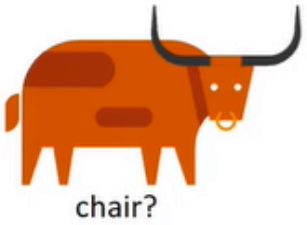
(Just right?)

f(x) = 3*Number of visits to upscale furniture websites
+10*Number of Businessweek clicks
+100*Number of distinct people emailed per day
+2*Number of purchases of over 5K within the last month
+10*Number of visits to airlines

Clusteren

Vindt patronen in soortgelijke objecten



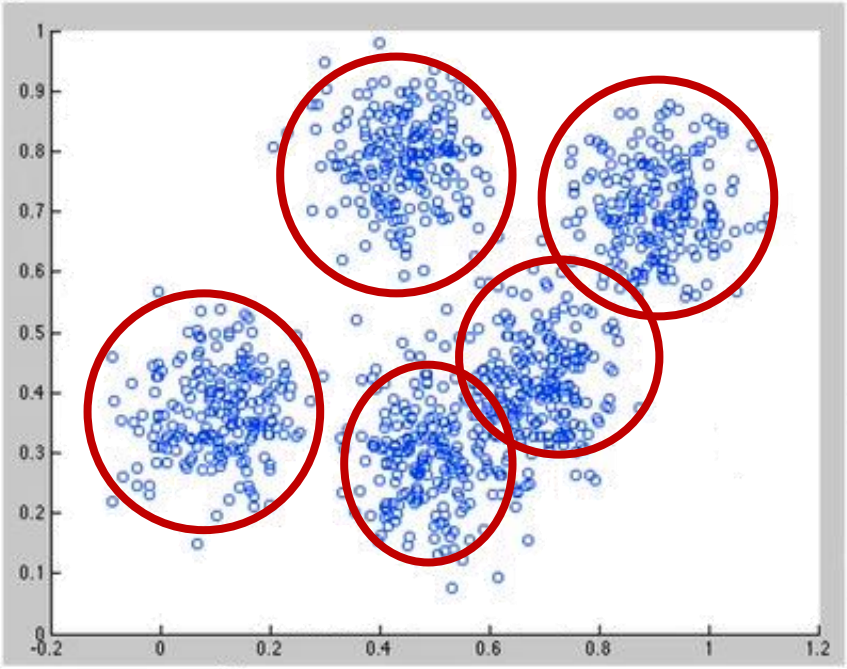


Clusteren

Vindt patronen in soortgelijke objecten

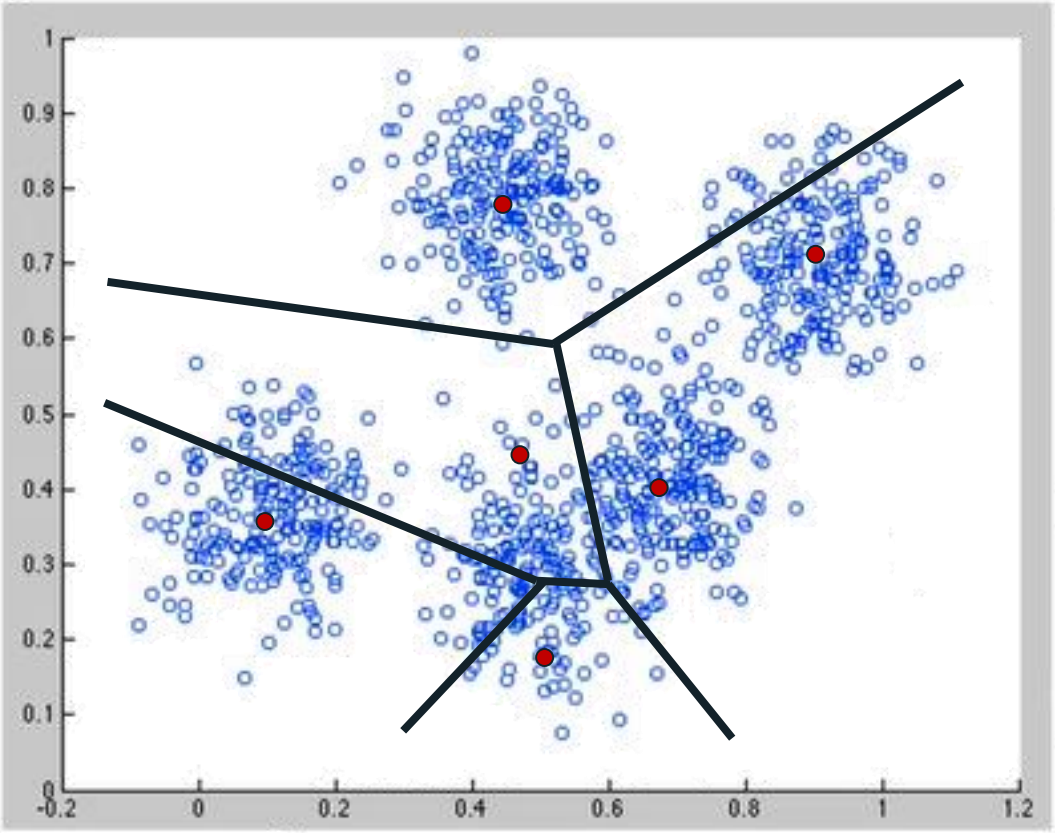
(K-Means) Clusteren

Hoe werkt het? 1/3



(K-Means) Clusteren

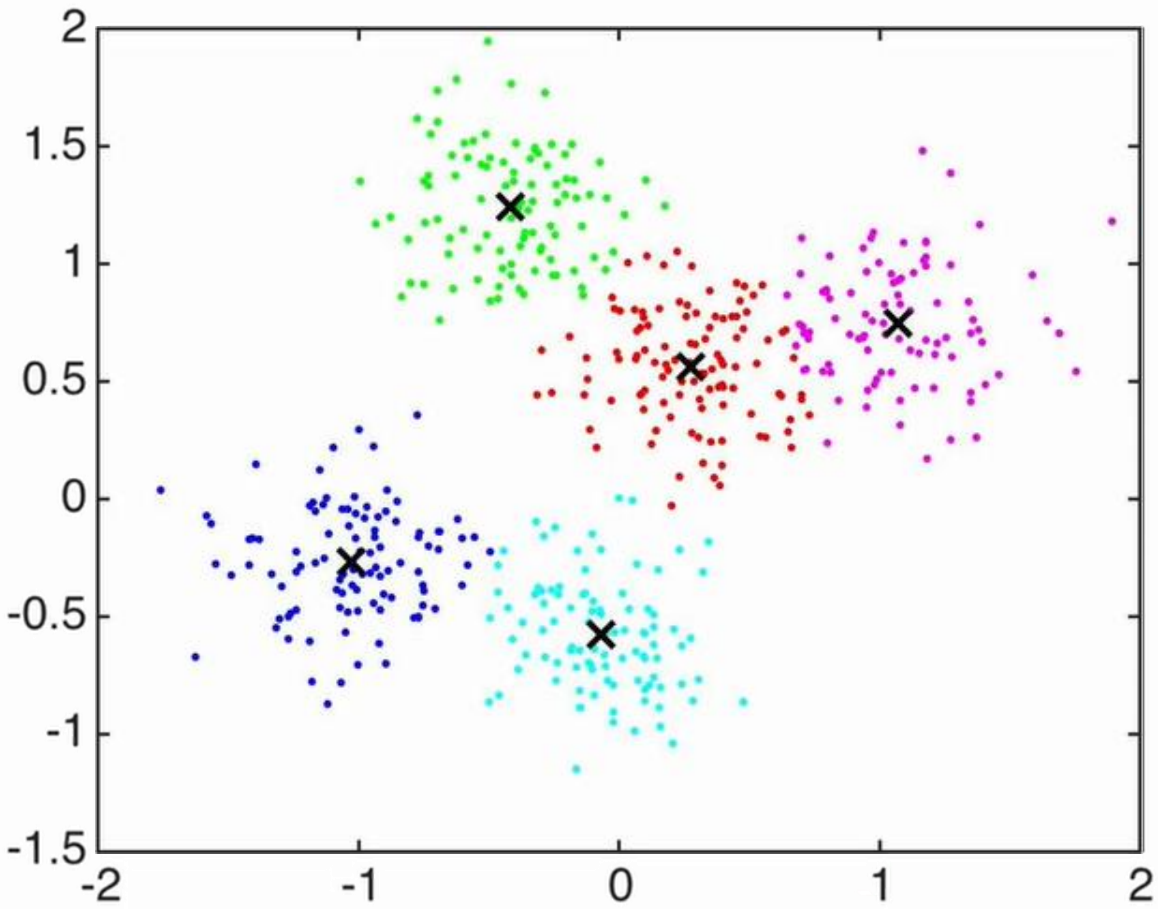
Hoe werkt het? 2/3



1. Input number of clusters, randomly initialize centers
2. Assign all points to the closest cluster center
3. Change cluster centers to be in the middle of its points
4. Repeat until convergence

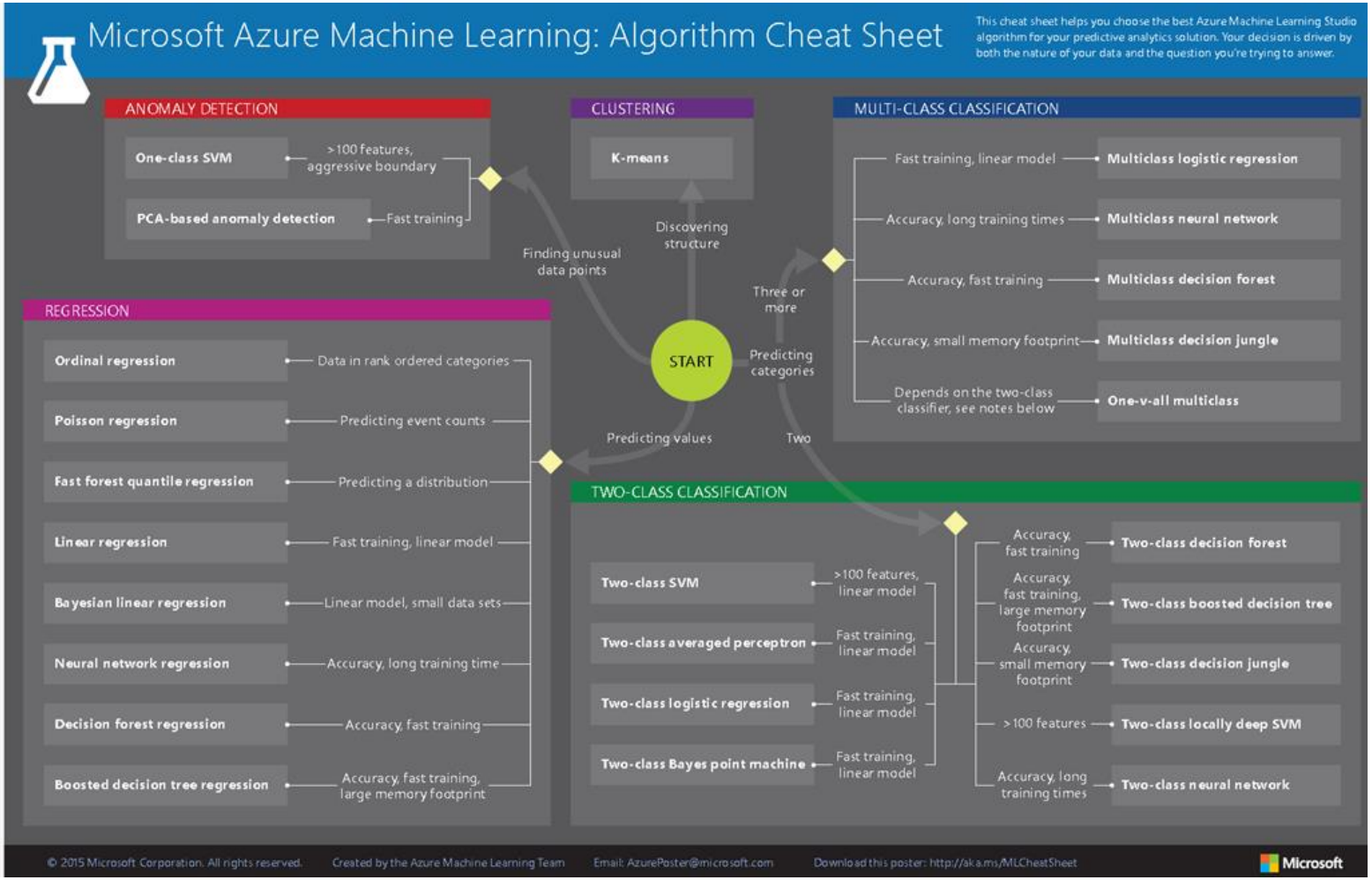
(K-Means) Clusteren

Hoe werkt het? 3/3



Welk algoritme past bij mijn probleem?

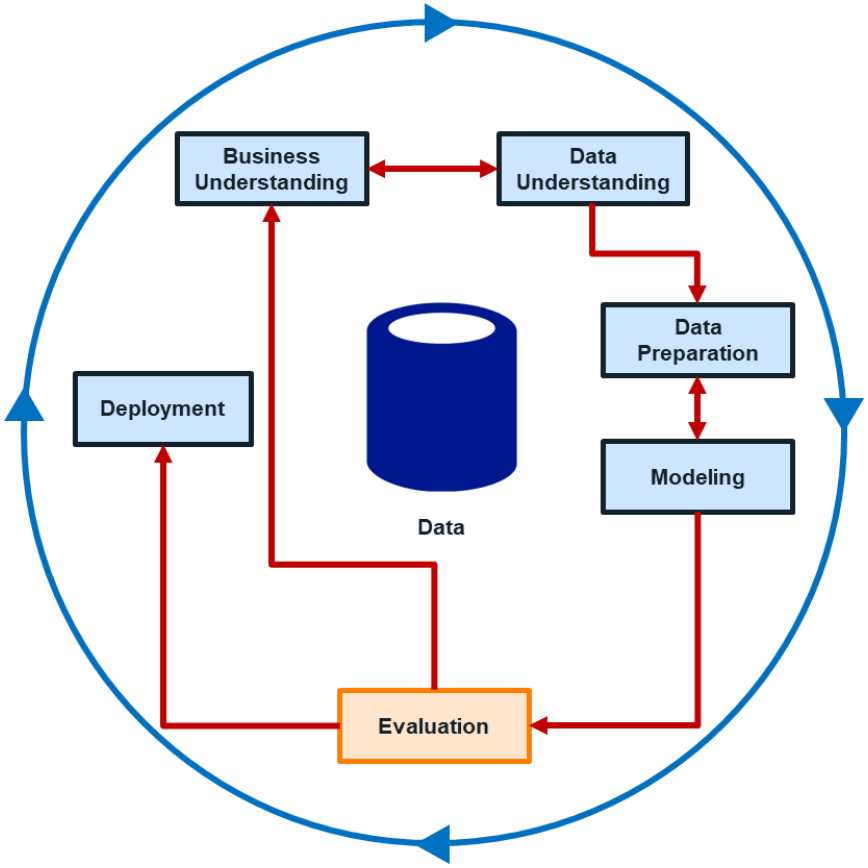
Handig spiekbriefje!



ilionx
Your partner in digital business

Evaluation

Assess your data mining results



Evaluatiemeetwaarden voor Classificatie

Confusion matrix & accuracy

Confusion matrix

	y = p	y = n
f(x) p	True positives	False positives (“Type I”)
f(x) n	False negatives (“Type II”)	True negatives

Misclassification error

Number of false decisions / total number of decisions

Accuracy

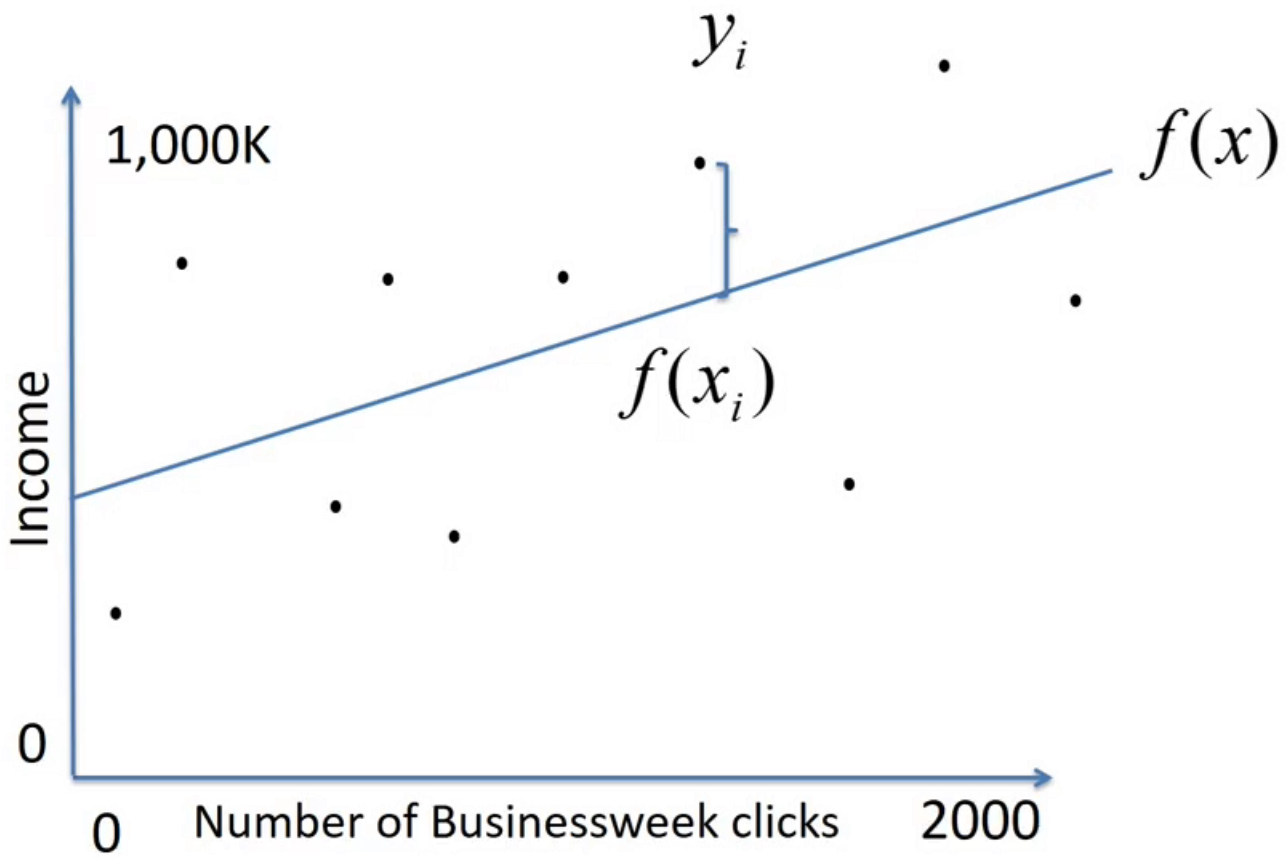
Number of correct decisions / total number of decisions

Evaluatiemeetwaarden voor Regressie

Sum of Squares Error (SSE)

$$\left(y_i - f(x_i)\right)^2$$

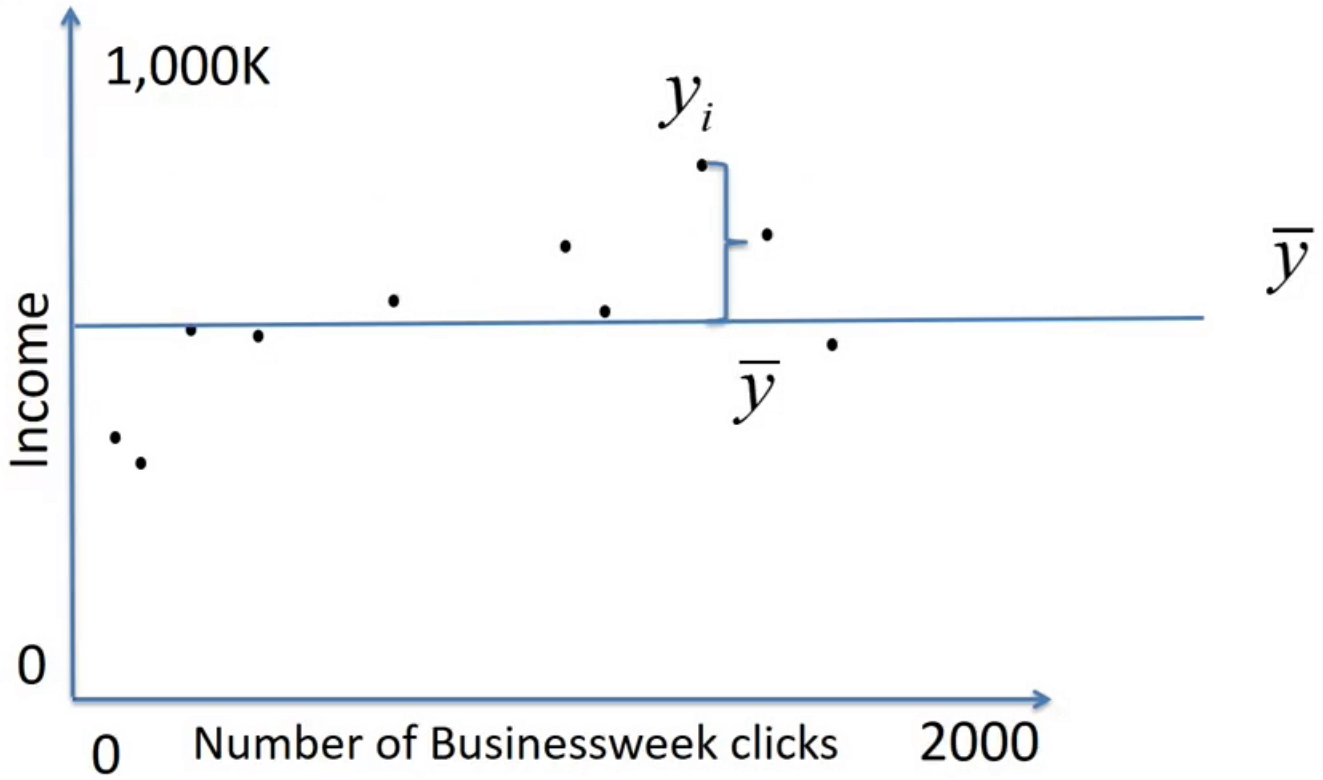
$$SSE = \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2$$



Evaluatiemeetwaarden voor Regressie

Sum of Squares Total (SST) / total variation

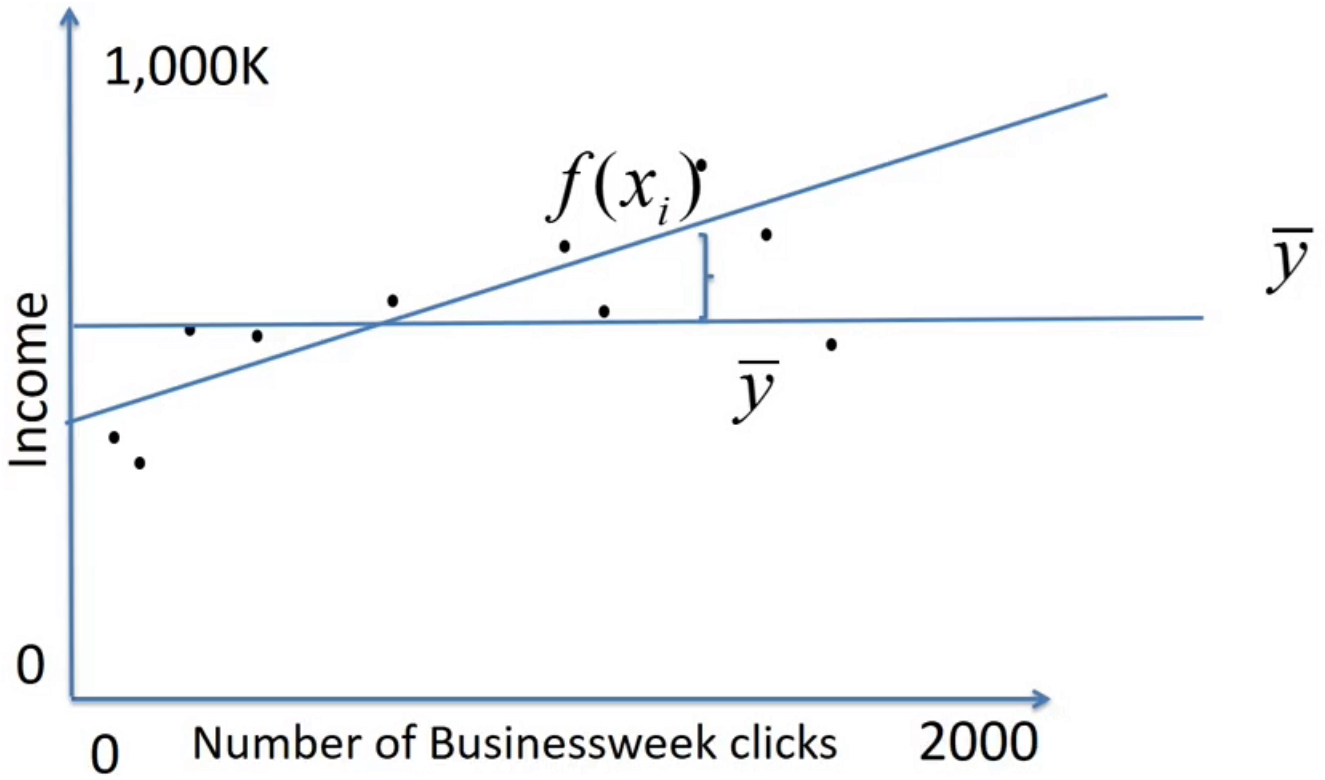
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



Evaluatiemeetwaarden voor Regressie

Sum of Squares Regression (SSR)

$$SSR = \sum_{i=1}^{n} \left(f(x_i) - \bar{y} \right)^2$$



Evaluatiemeetwaarden voor Regressie

Root Mean Square Error (R^2)

Sometimes (e.g. least squares):

SST = SSE + SSR

$$1 - \frac{SSE}{SST}$$

Close to 1

Close to 0

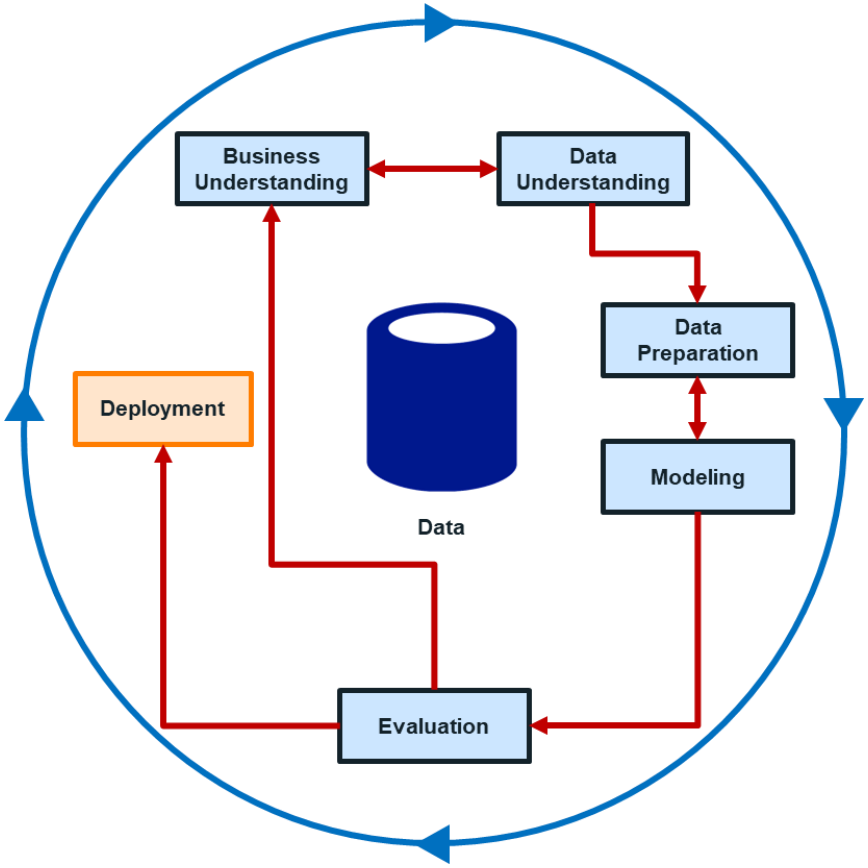
$$\frac{SSE}{SST}$$

Close to 0

Close to 1

Deployment

Implement your predictive model



Uitrollen en implementeren

Het eindstation van CRISP?

- **Tot hier was het alleen maar investeren**

- **Dat een model goed in de ontwikkelomgeving functioneerde wilt niet zeggen dat het goed in productie zal draaien**
  - Vaak moet het algoritme worden aangepast op productie
  - “Your model is not what the data scientists design, it’s what the engineers build”

- **Uitrollen en implementeren van predictive models wordt vereenvoudigd door middel van Microsoft Azure en Cortana Analytics**