

# Microsoft Modern Data Platform

**illionX**

Your partner in digital business

# The world is changing

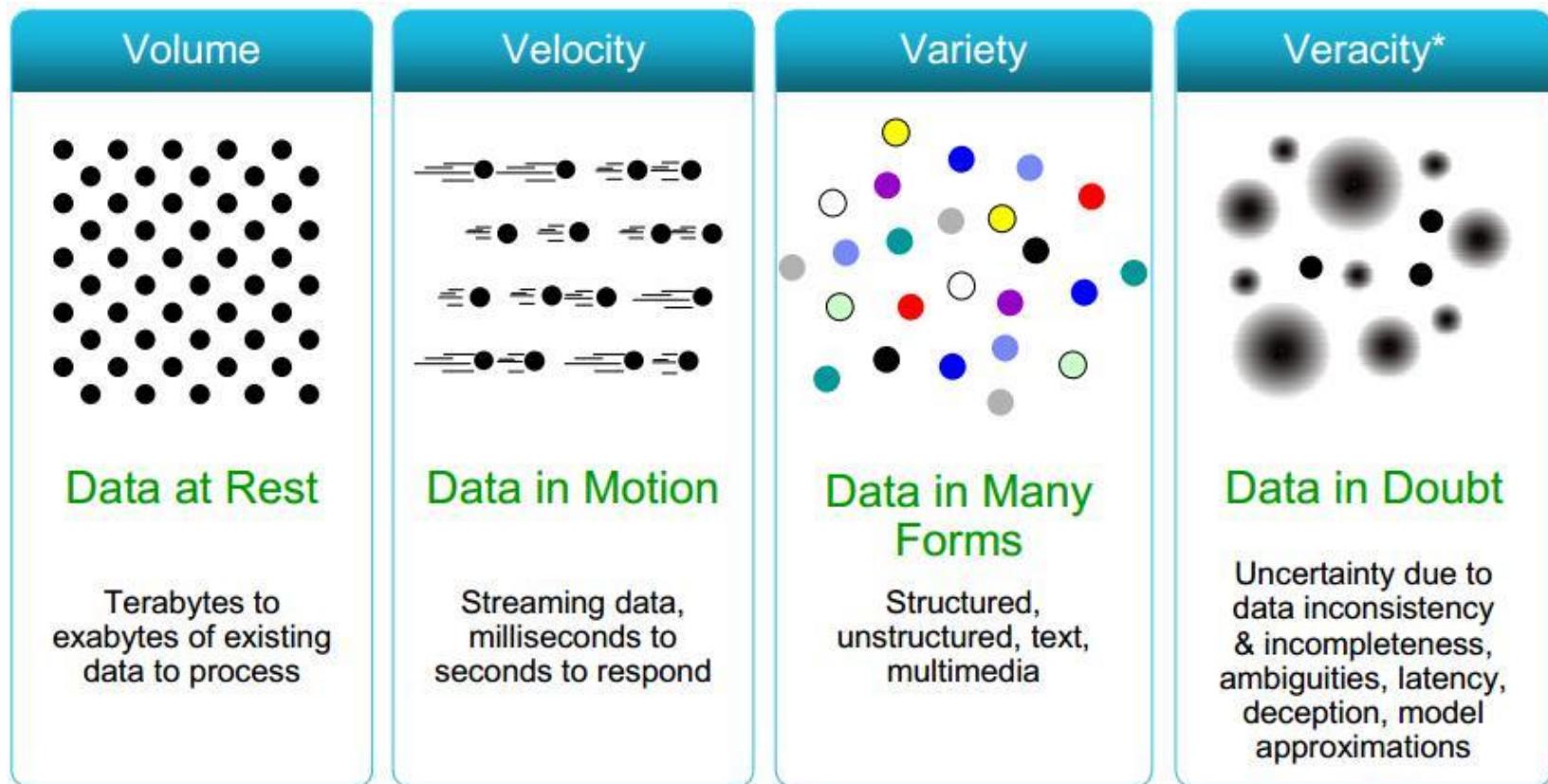
DATA

CLOUD

AI

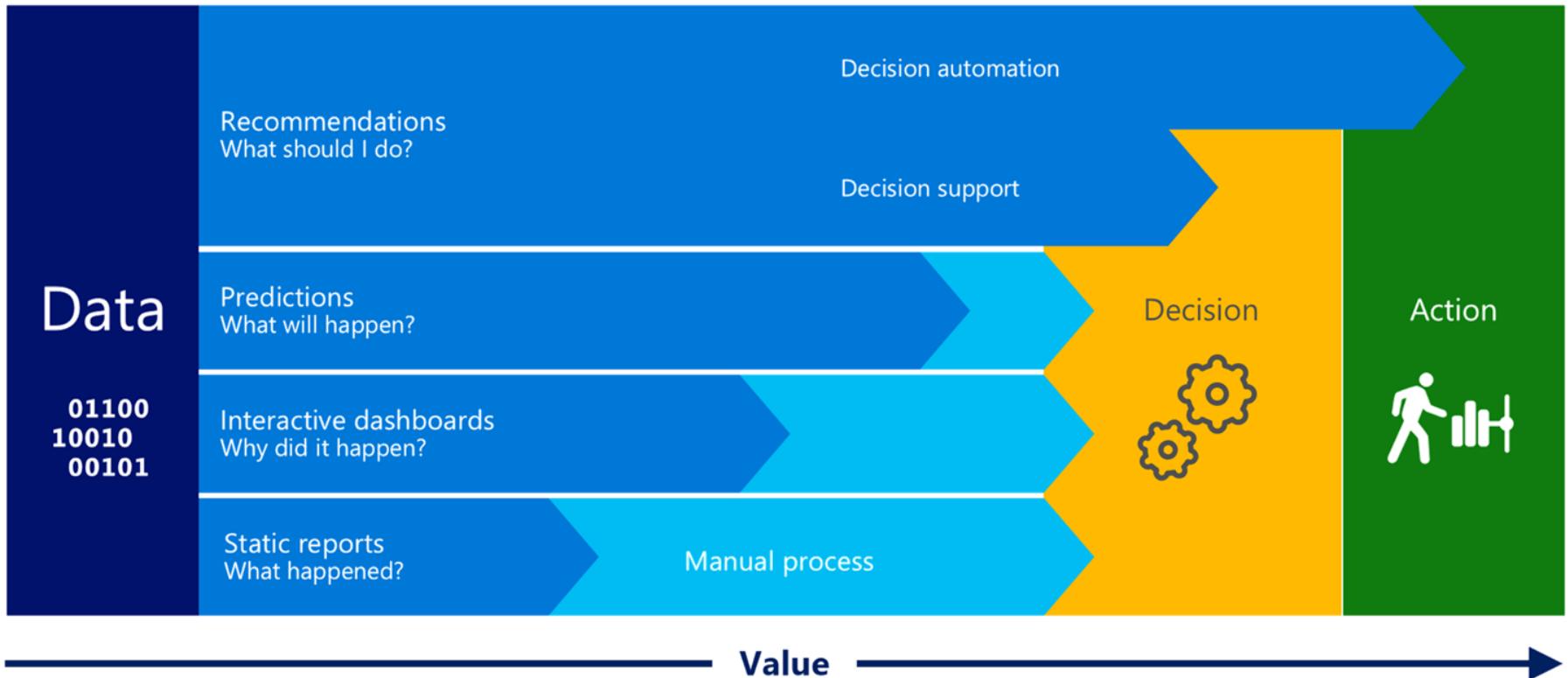
# Big Data

## 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

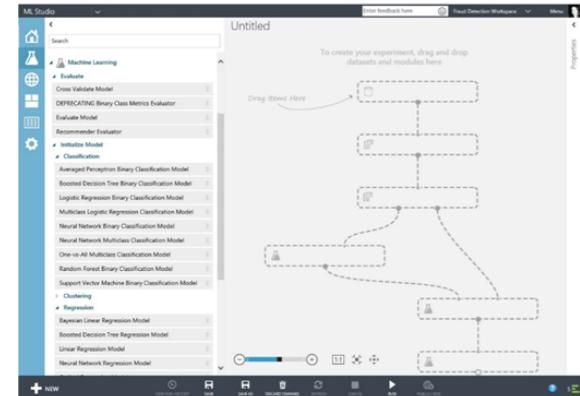
ML Operationalization to remove deployment friction

- 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



The screenshot shows the Azure Machine Learning Studio interface. On the left, there is a sidebar with a search bar and a list of modules categorized under 'Machine Learning'. The categories include 'Evaluate' (with 'Cross Validate Model' and 'DEPRECATING: Binary Class Metrics Evaluator'), 'Recommender Evaluation' (with 'Evaluate Model'), 'Classification' (with 'Averaged Perceptron Binary Classification Model', 'Boosted Decision Tree Binary Classification Model', 'Logistic Regression Binary Classification Model', 'Multinomial Logistic Regression Classification Model', 'Neural Network Binary Classification Model', 'Neural Network Multiclass Classification Model', 'One-vs-All Multiclass Classification Model', 'Random Forest Binary Classification Model', and 'Support Vector Machine Binary Classification Model'), 'Regression' (with 'Bayesian Linear Regression Model', 'Boosted Decision Tree Regression Model', 'Linear Regression Model', and 'Neural Network Regression Model'), 'Clustering' (with 'K-Means Clustering Model'), and 'Text Analytics' (with 'Text Analytics API'). The main workspace is titled 'Untitled' and shows a visual workflow with various dashed boxes connected by arrows, representing data flow between different steps. Below the workspace, there is a toolbar with icons for file operations and a status bar.

Sort By: Date Added Name Publisher

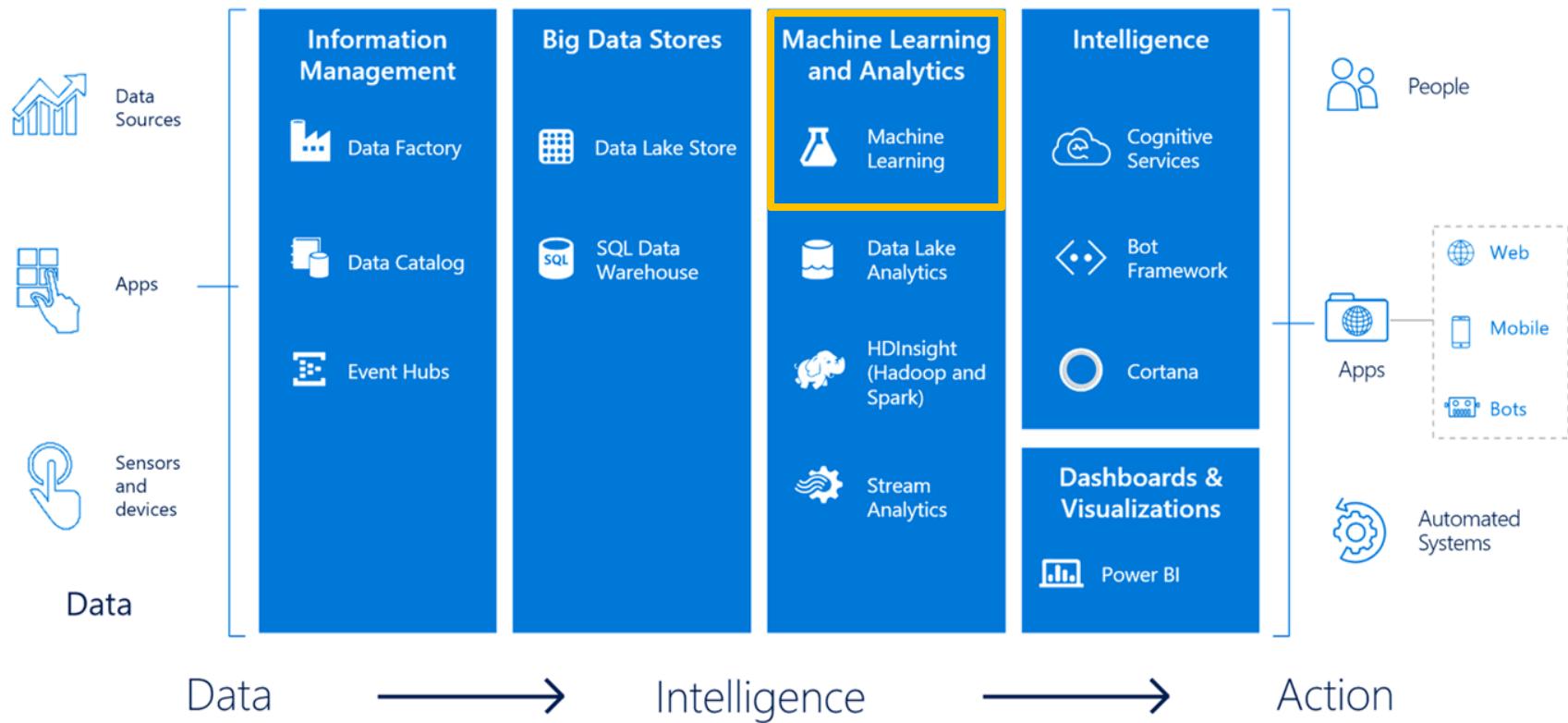
**Customer Churn Prediction**  
published by: Azure Machine Learning  
Customer Churn Prediction is a churn analytics service built with Azure Machine Learning. It's designed to predict the likelihood of a customer (player, subscriber, user, etc.) ending his or her relationship with a company or service.

**Text Analytics**  
published by: Azure Machine Learning  
Text Analytics API is a suite of text analytics services built with Azure Machine Learning. Just bring your unstructured text (English only), and use this API to perform sentiment analysis and key phrase extraction.

**Sentiment Analysis API Built with Azure Machine Learning**  
published by: Azure Machine Learning  
This Sentiment Analysis API is a sample built with Microsoft Azure Machine Learning. It analyzes the sentiment polarity of short sentences, such as Facebook statuses, tweets, etc. The underlying model is built using an Azure ML native Support Vector Machine algorithm. Instead of outputting a raw score, it maps the output into three levels: positive, neutral, and negative. It also provides a confidence score which could be used to further tune the polarity. The purpose of the web API is to demonstrate how to build and publish public services using Azure ML modules, such as Train Model, Score Model, Two-class Support Vector Machine, Feature hashing, and others.

# Cortana Intelligence Suite services

## Building blocks of data intelligence



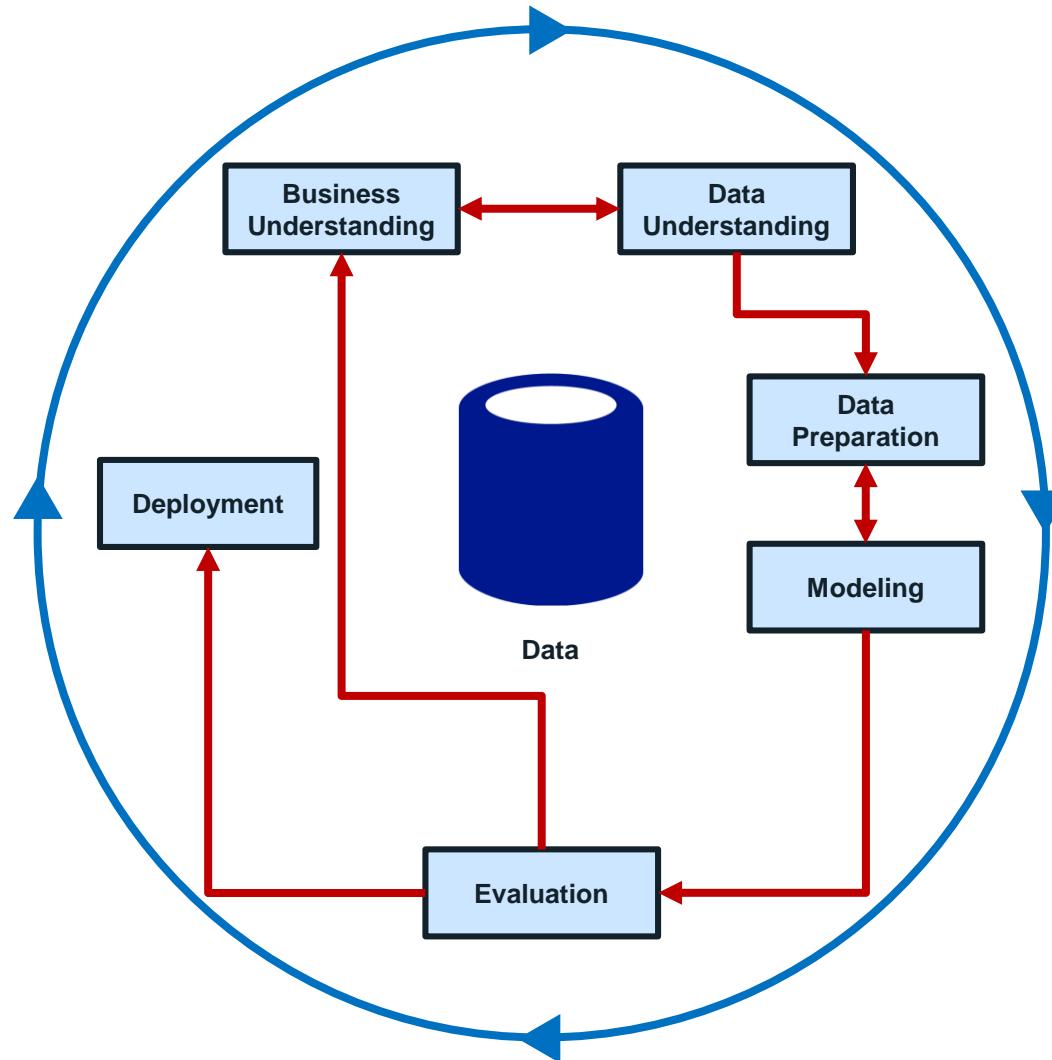
# Data Science

## Kerncompetenties



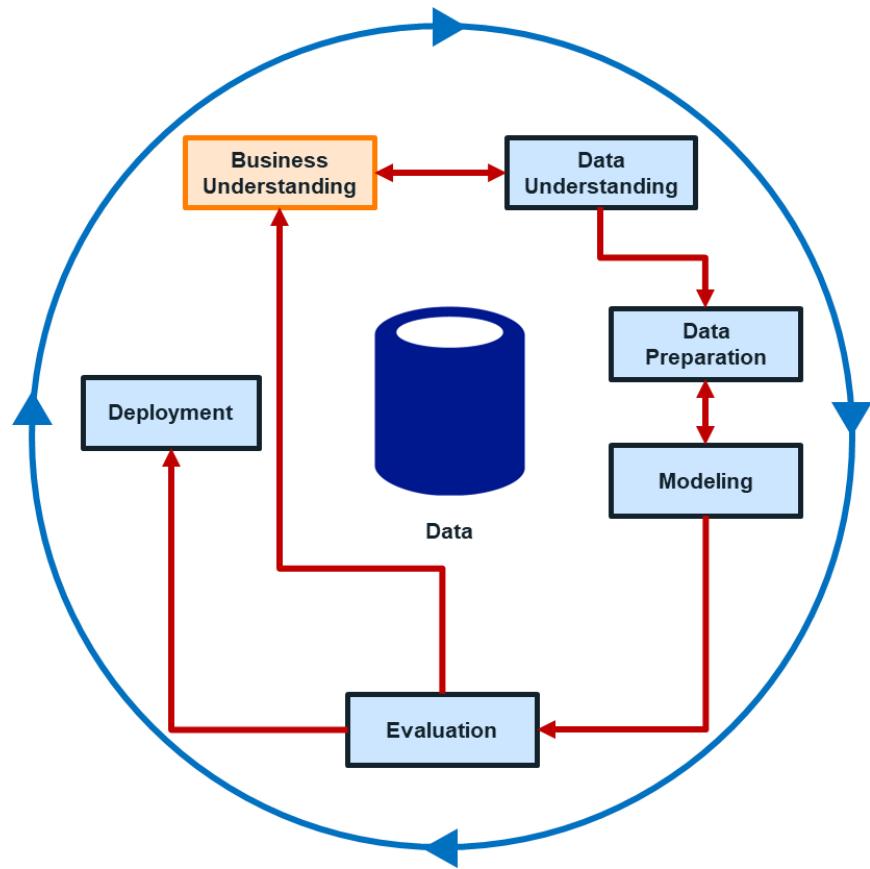
# The CRISP data mining process

## KDD + Business & Data Understanding



# Business Understanding

Think about your problem



# Business Understanding

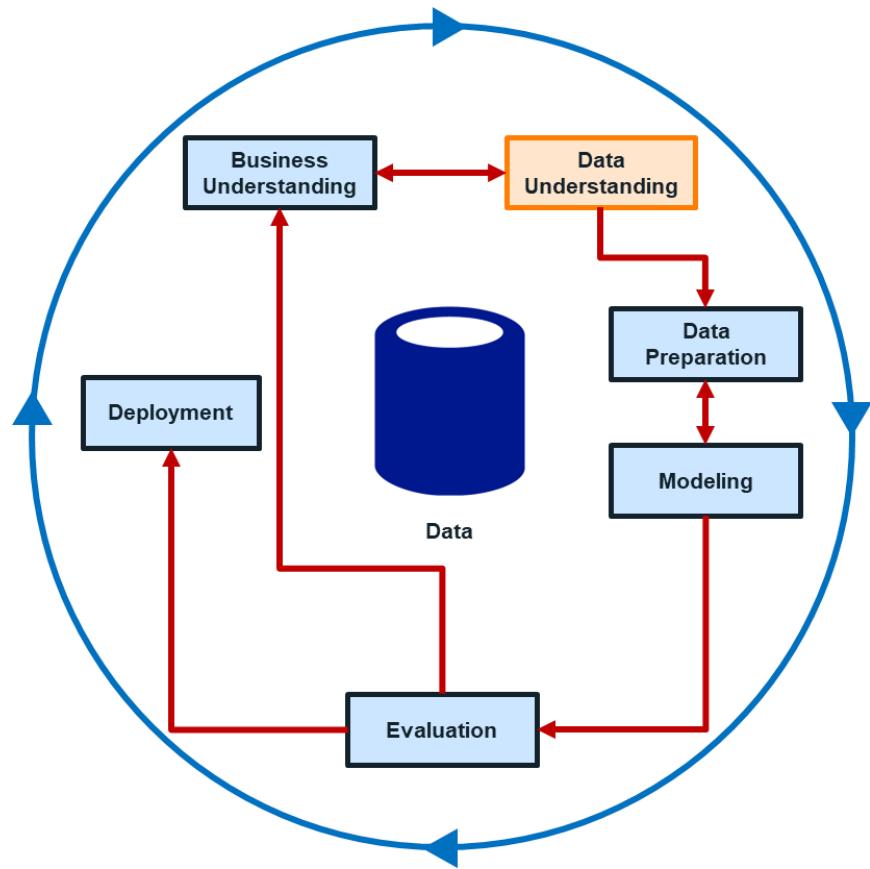
Denk na over het probleem en over use cases!

---

- **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-occurrence 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!

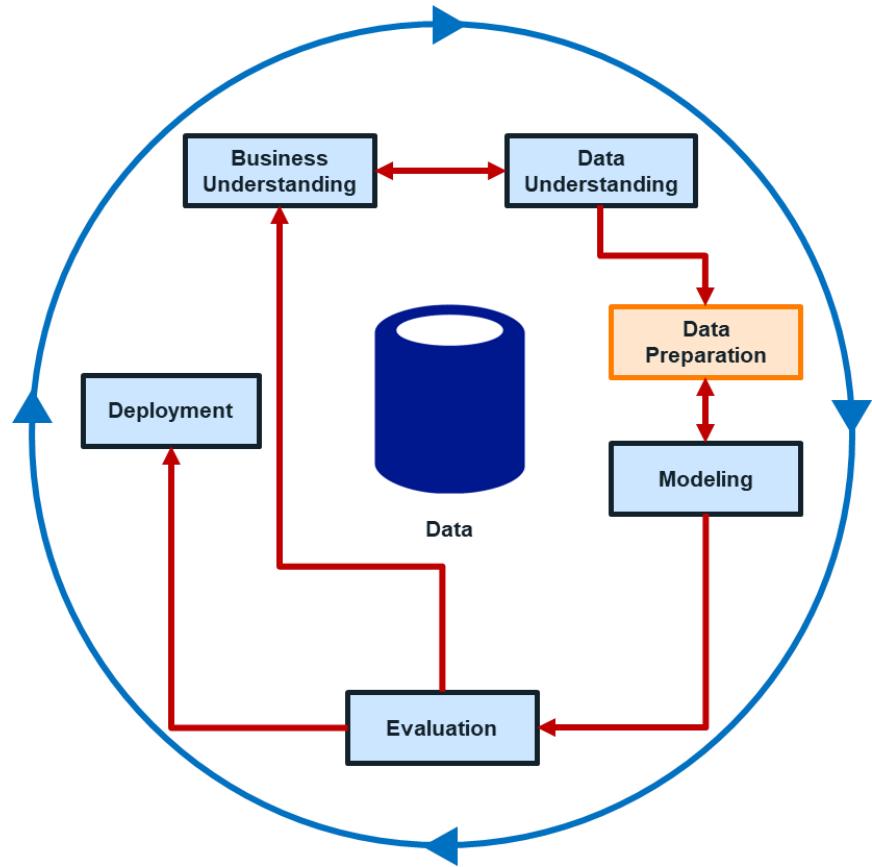
---

### 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

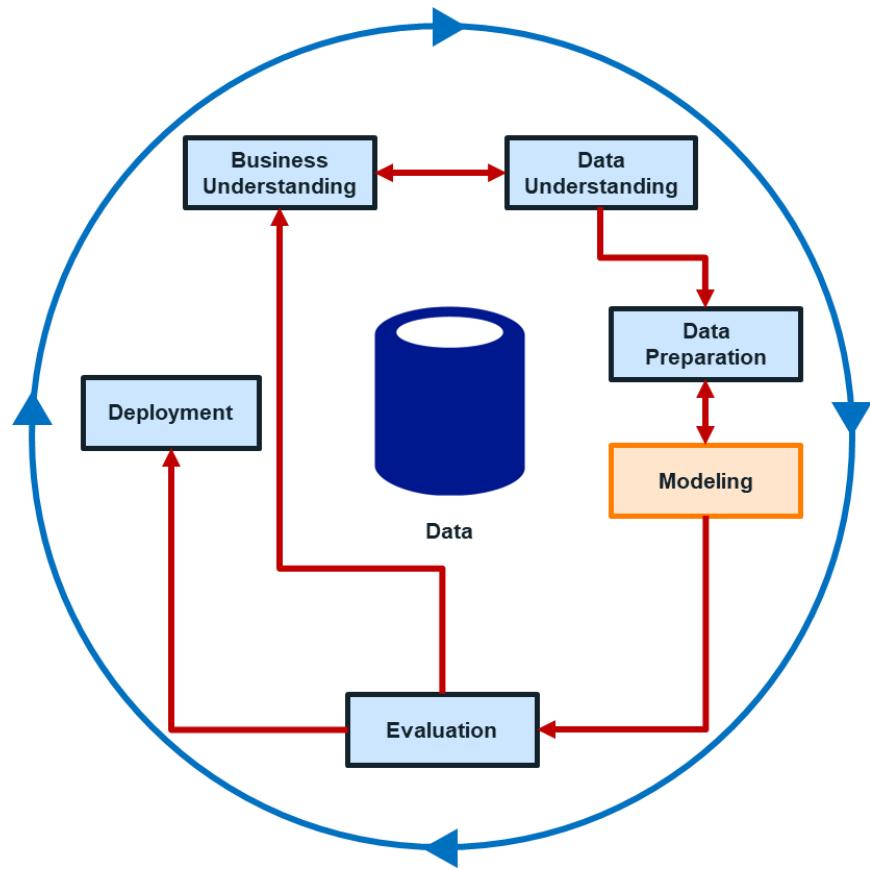
## Garbage in – garbage out

---

- **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)

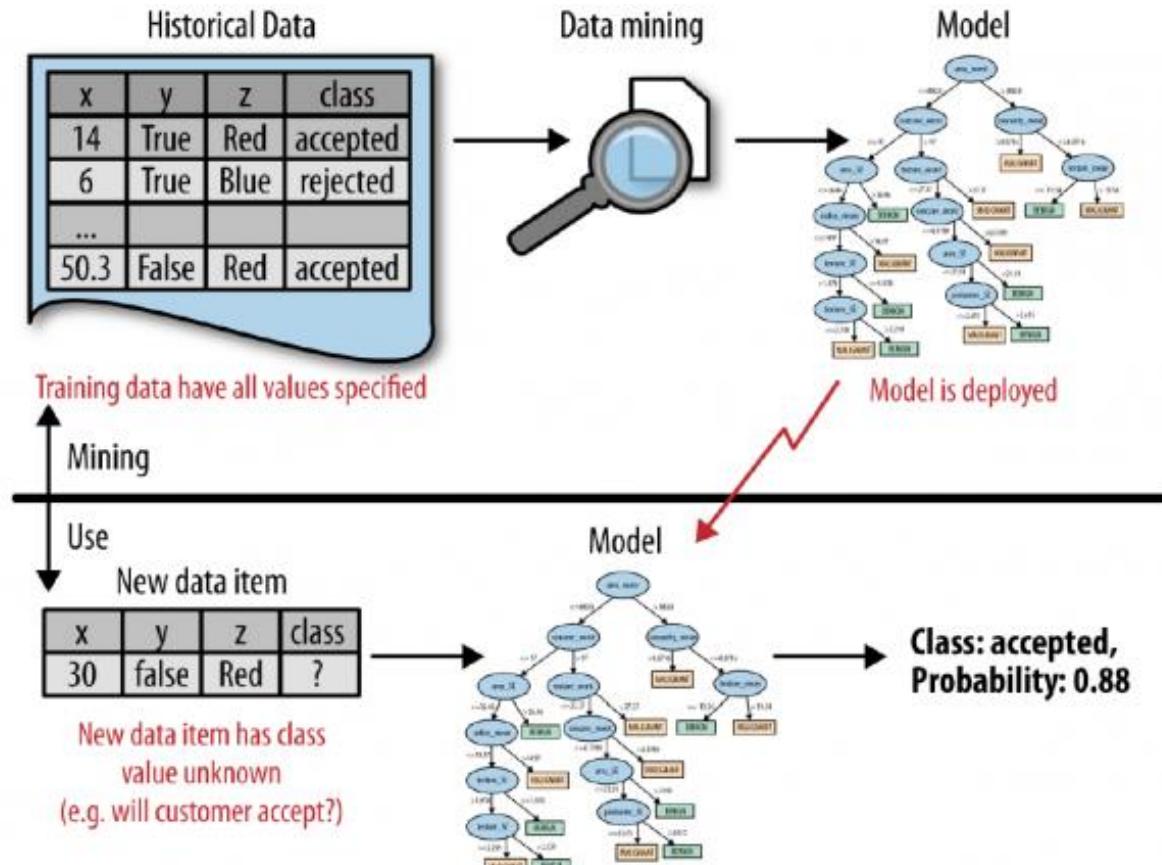
# Modeling

Find patterns in your data



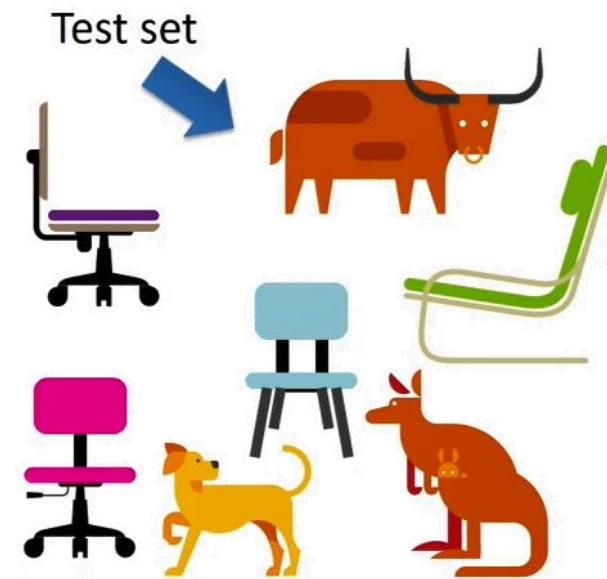
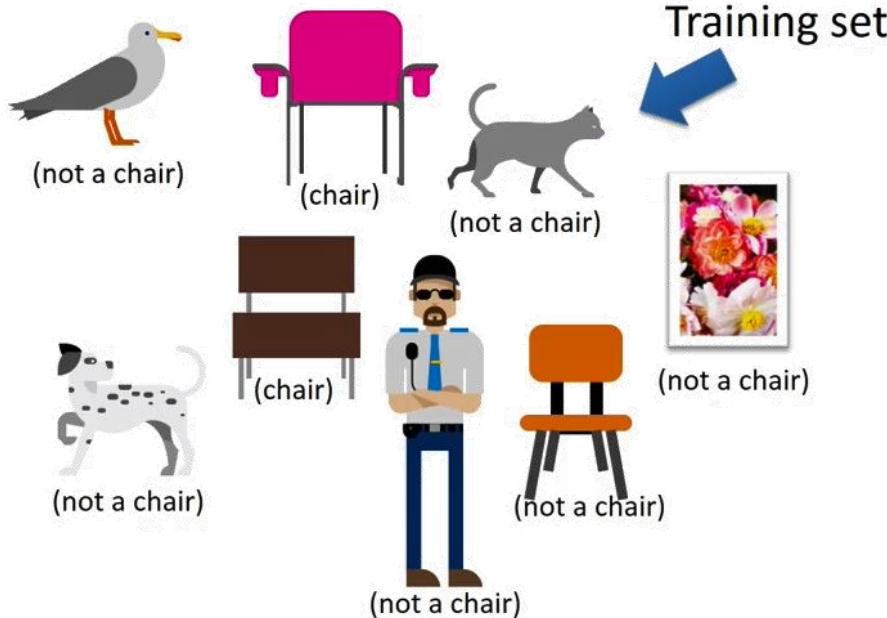
# Modeleren

## Classificeren, regressie en clusteren



# Classificeren

## Voorspel antwoorden op ja/nee vragen



# Classificeren

## Hoe werkt het? 1/3

Manhole is represented as: [ 5 3 120 12 1 0 ..... ]

Number of events last year  
Number of serious events last year  
Number of electrical cables  
Number of pre-1930 electrical cables  
Vented cover?  
Inspected?

# 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)

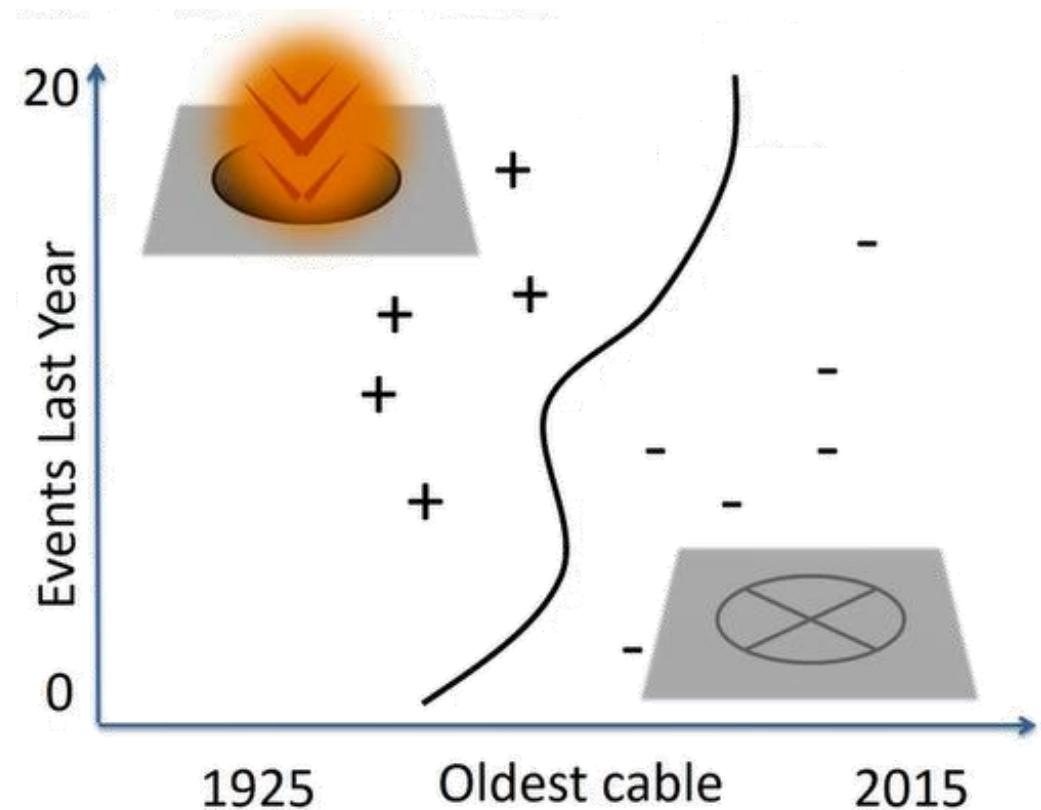
# Classificeren

## Hoe werkt het? 3/3

$f(x) = \text{function}(\text{Events Last Year}, \text{Oldest Cable})$

Manhole is represented as: [ 1925 15]

Year oldest cable installed  
Number of events last year



# Regressie

## Voor spel (real) numerieke waarden

A person is represented as:

[	5	]
[	0	]
[	1	]
:		



Single feature, called X

Income
84
32
-10



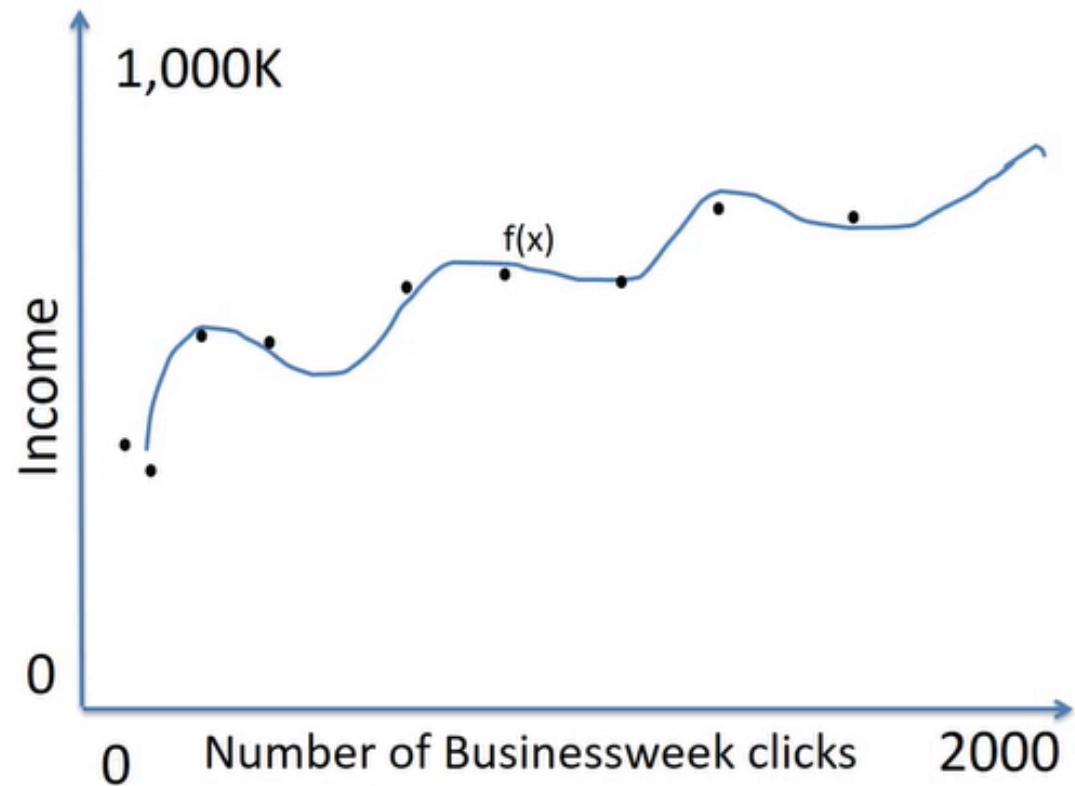
Labels, called Y

# Regressie

## Hoe werkt het? 1/3

$f(x) = \text{function}(\text{Number of Businessweek clicks})$

(Overfitting?)

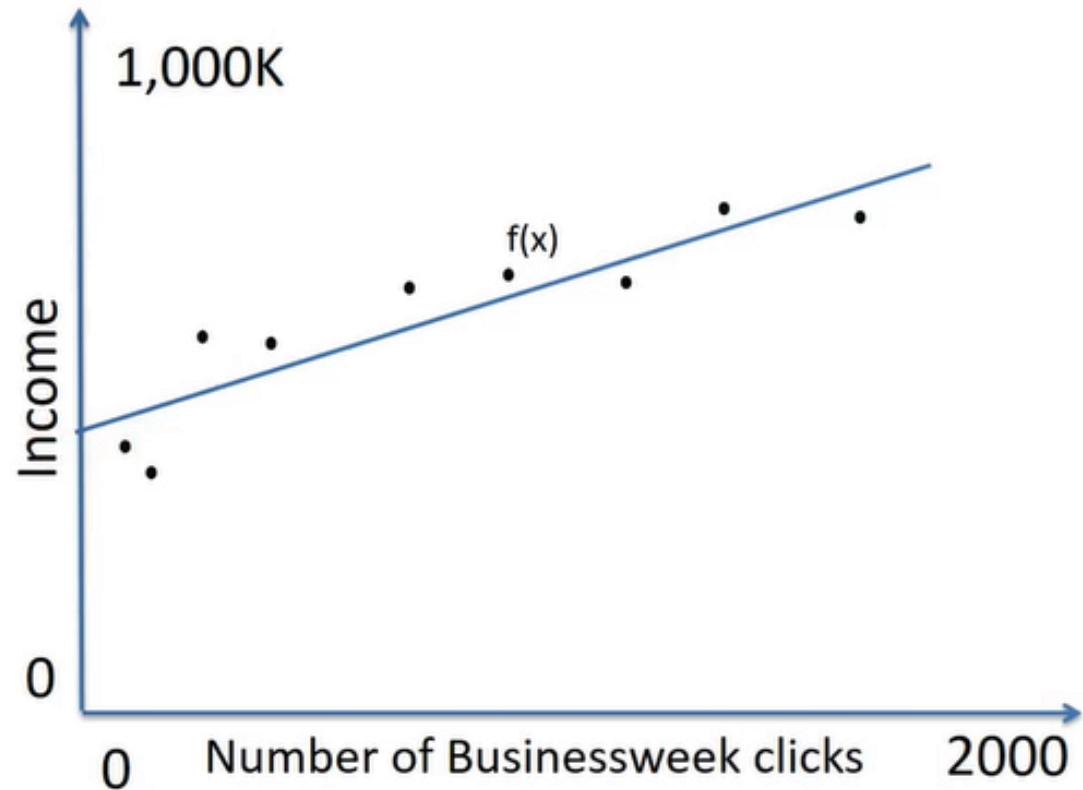


# Regressie

## Hoe werkt het? 2/3

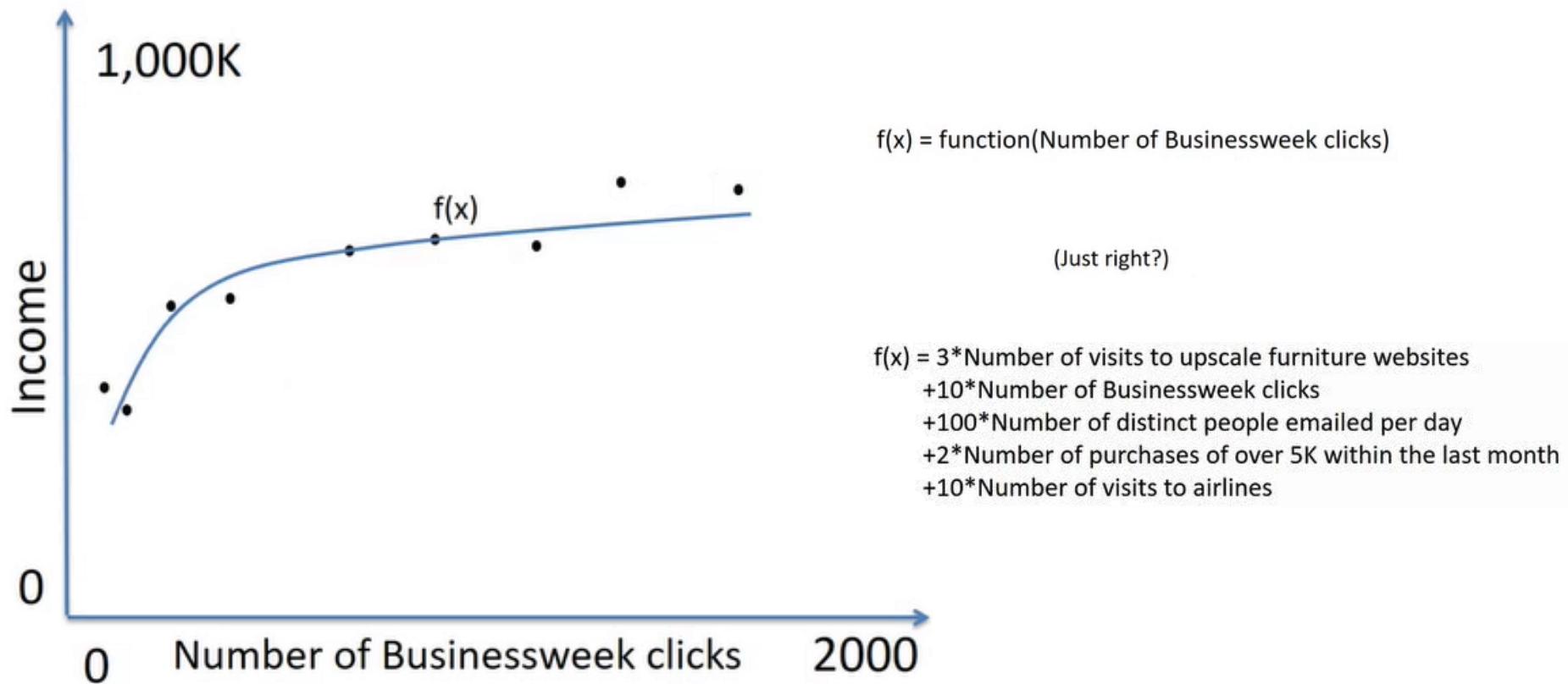
$f(x) = \text{function}(\text{Number of Businessweek clicks})$   
 $= 5K * \text{Number of Businessweek clicks} + 100K$

(Underfitting?)



# Regressie

## Hoe werkt het? 3/3



# Clusteren

Vindt patronen in soortgelijke objecten

Supervised:



(not a chair)



(not a chair)



(chair)



(not a chair)



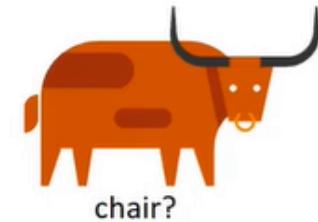
(chair)



(not a chair)



(chair)

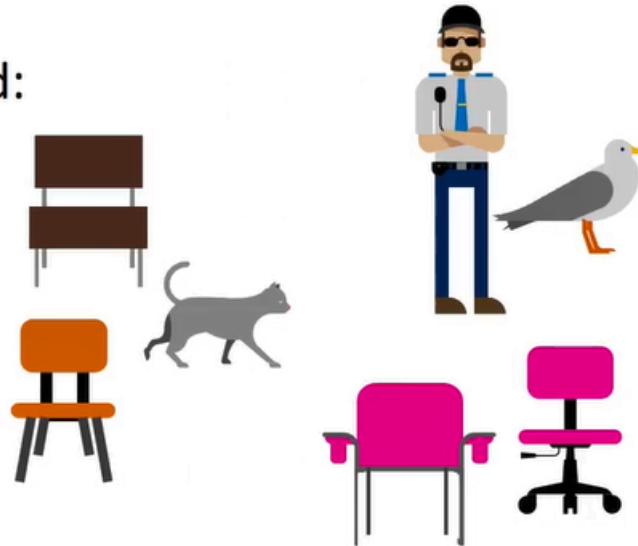


chair?

# Clusteren

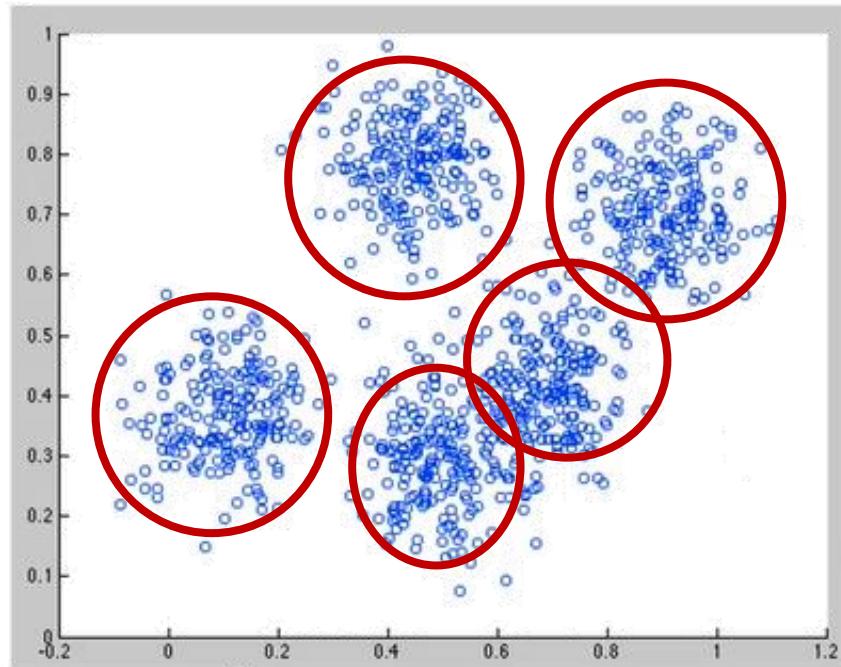
Vindt patronen in soortgelijke objecten

Unsupervised:



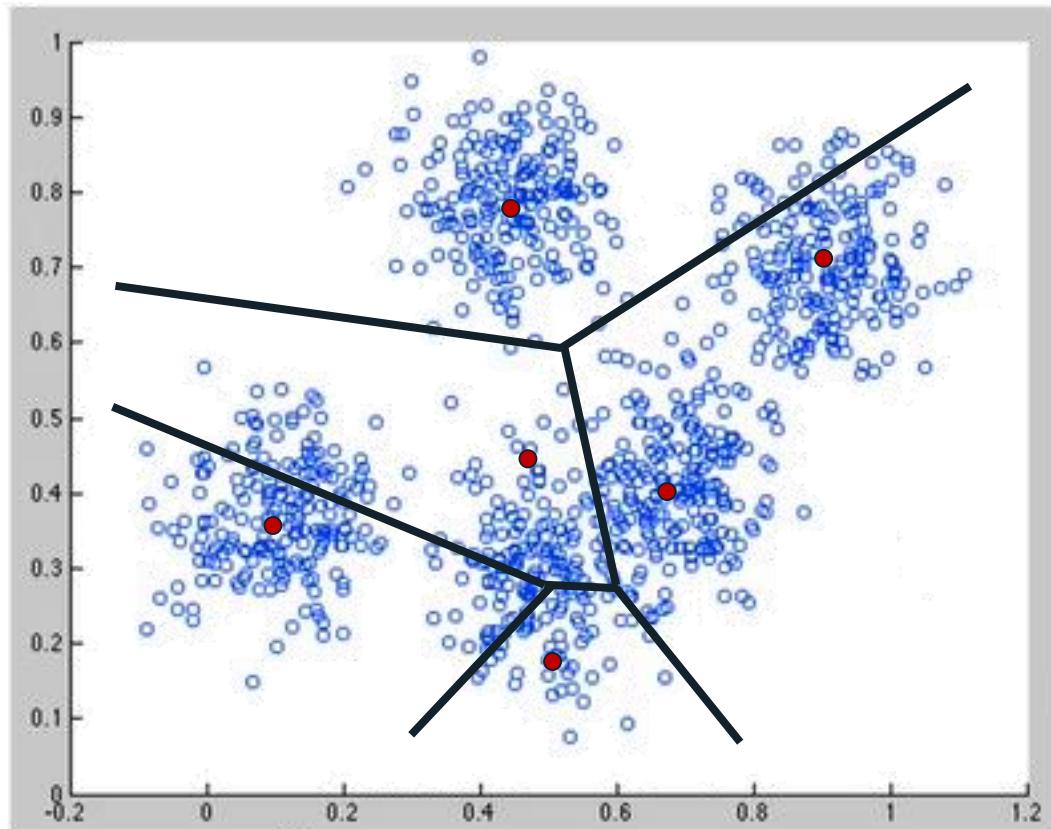
# (K-Means) Clusteren

## Hoe werkt het? 1/3



# (K-Means) Clusteren

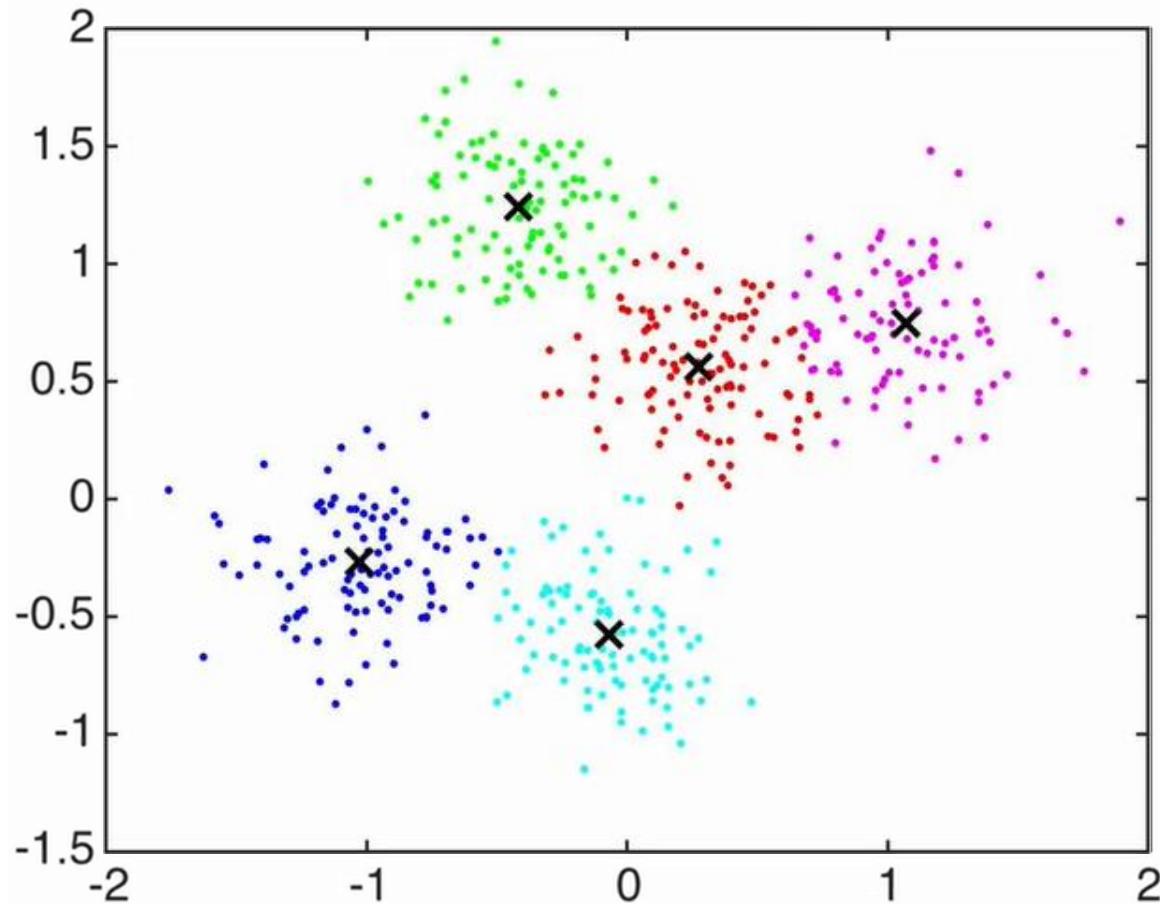
## Hoe werkt het? 2/3



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

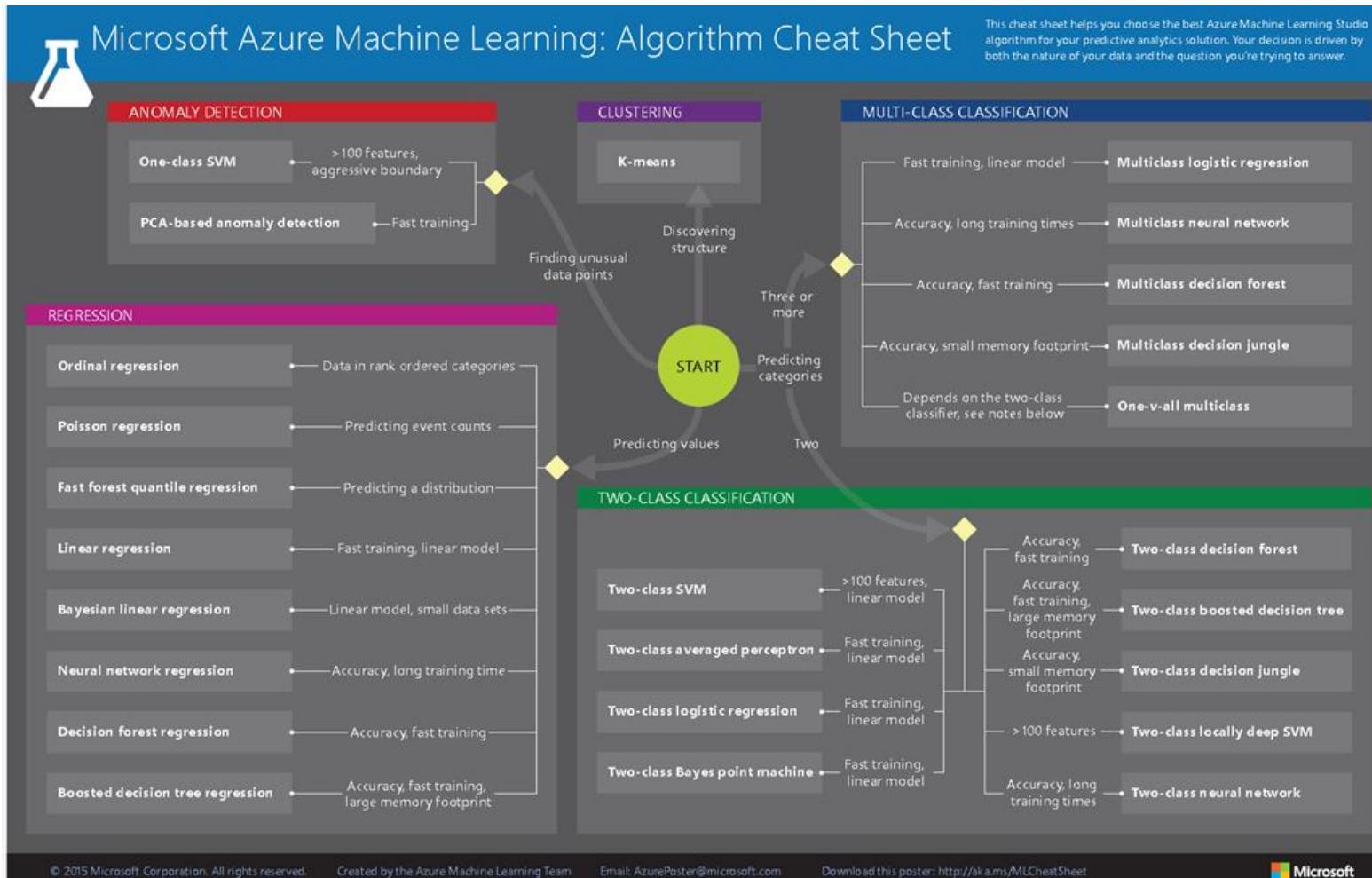
# (K-Means) Clusteren

Hoe werkt het? 3/3



# Welk algoritme past bij mijn probleem?

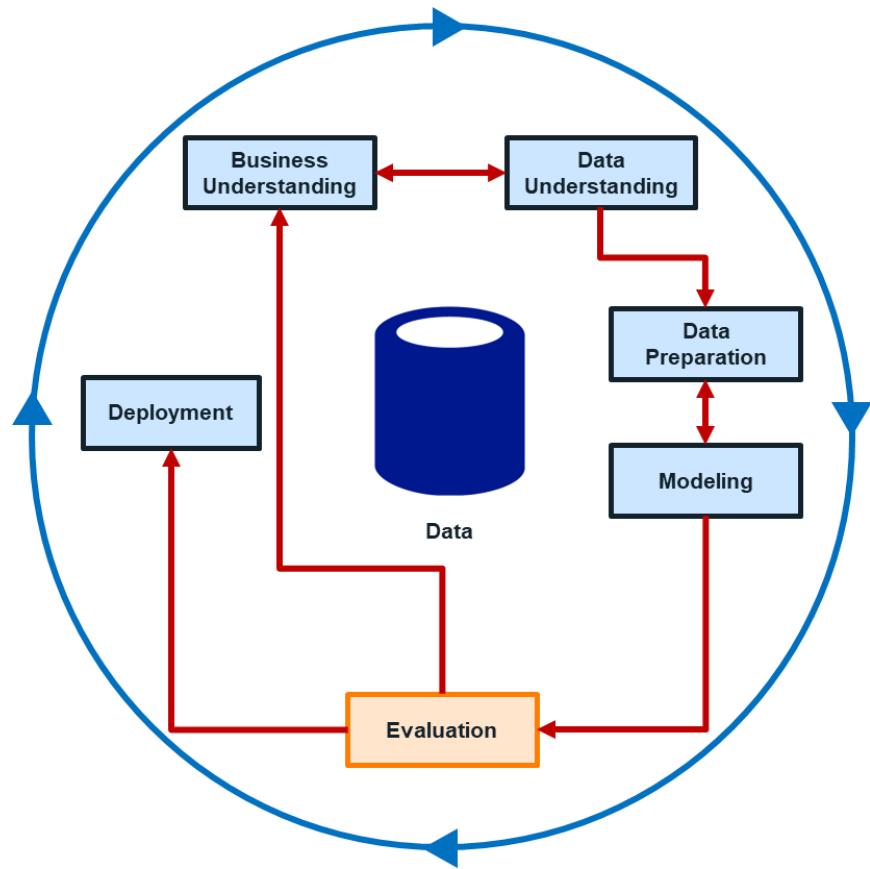
## Handig spiekbriefje!



# 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

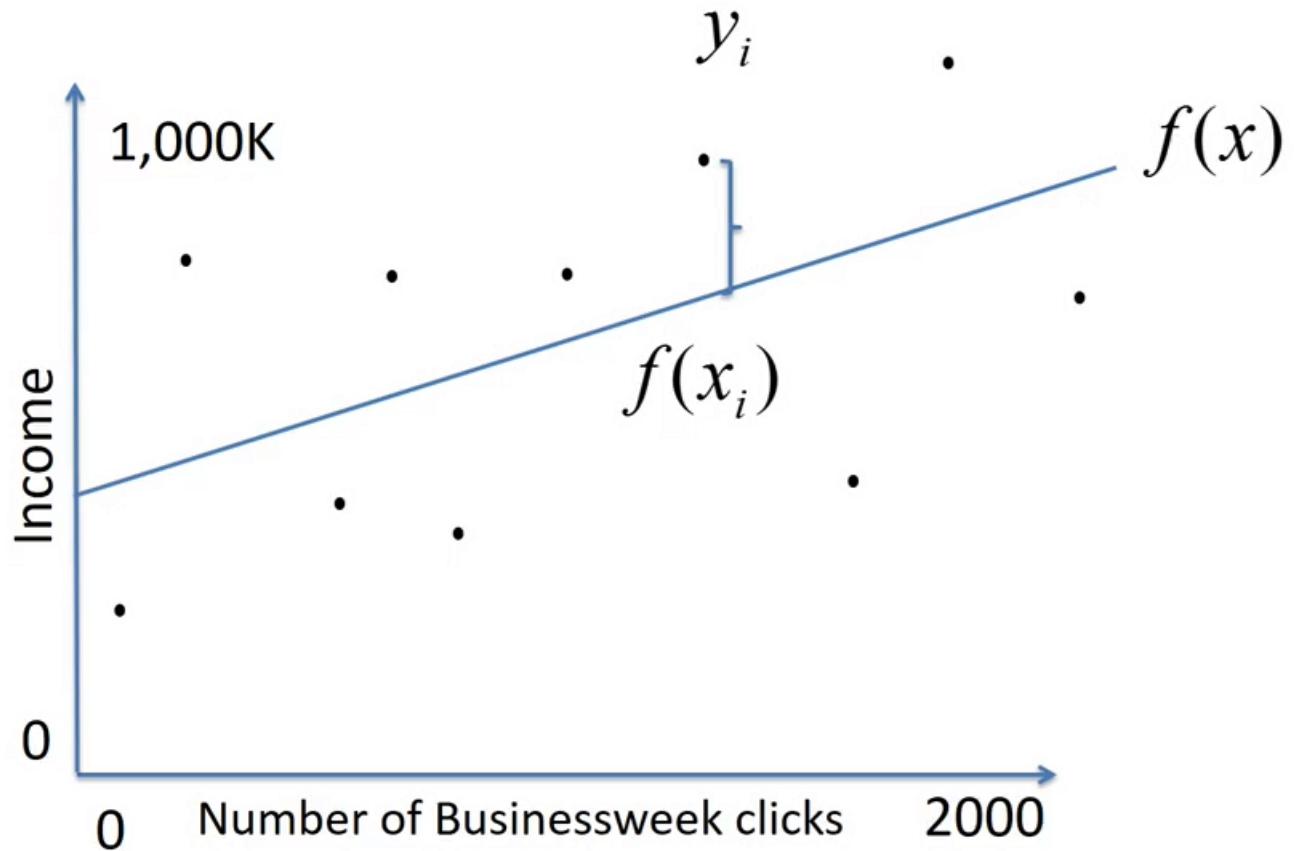
### Accuracy

Number of correct decisions / total number of decisions

# Evaluatiemeetwaarden voor Regressie

## Sum of Squares Error (SSE)

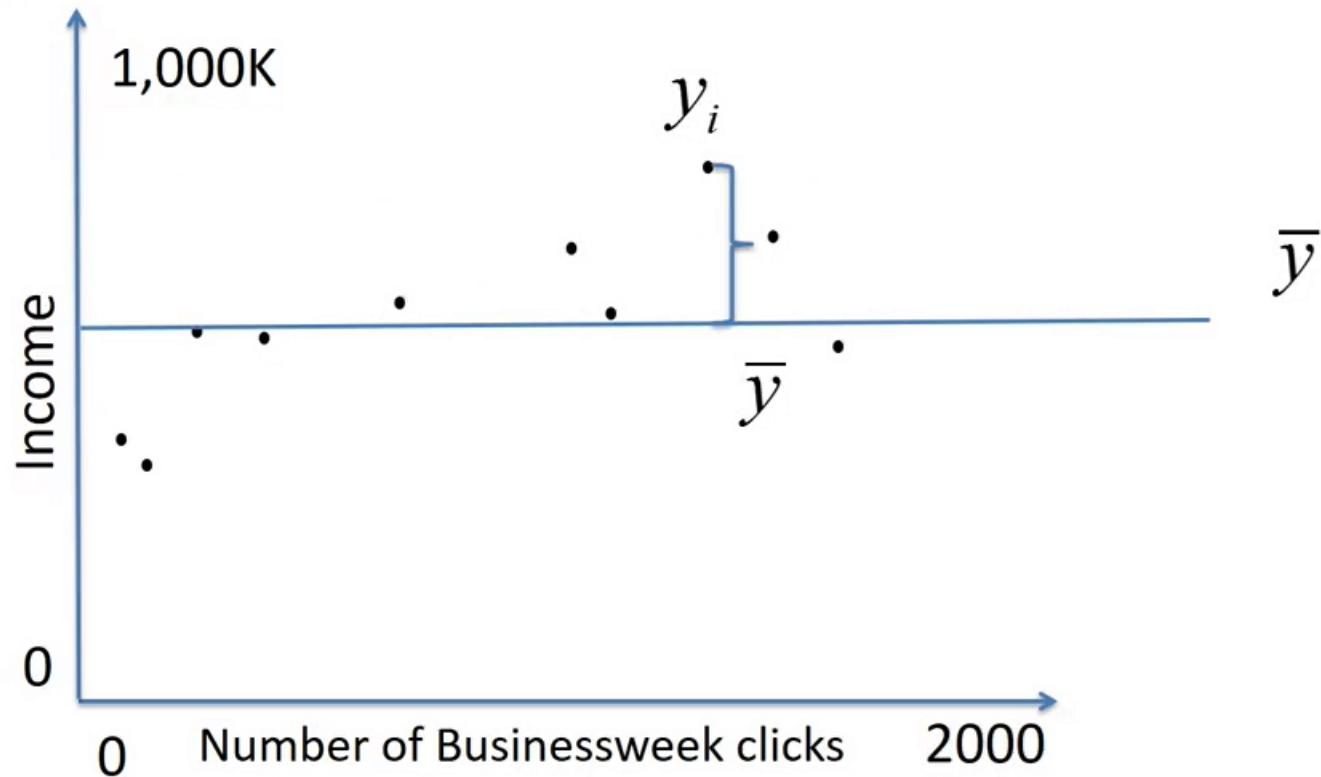
$$(y_i - f(x_i))^2$$
$$SSE = \sum_{i=1}^n (f(x_i) - y_i)^2$$



# Evaluatiemeetwaarden voor Regressie

## Sum of Squares Total (SST) / total variation

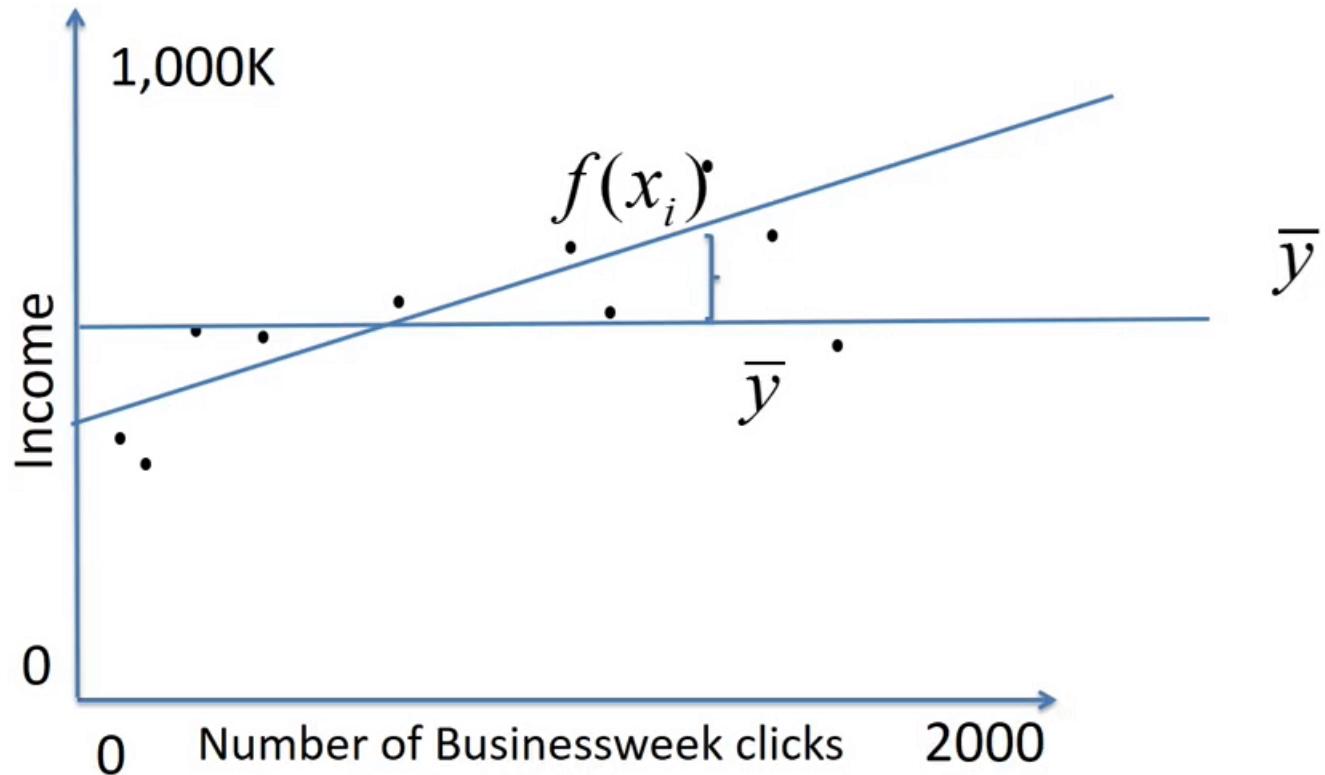
$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$



# Evaluatiemeetwaarden voor Regressie

## Sum of Squares Regression (SSR)

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



# Evaluatiemeetwaarden voor Regressie

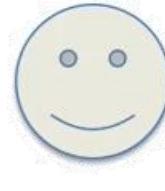
## Root Mean Square Error ( $R^2$ )

Sometimes (e.g. least squares):

$$SST = SSE + SSR$$

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

Close to 1      Close to 0



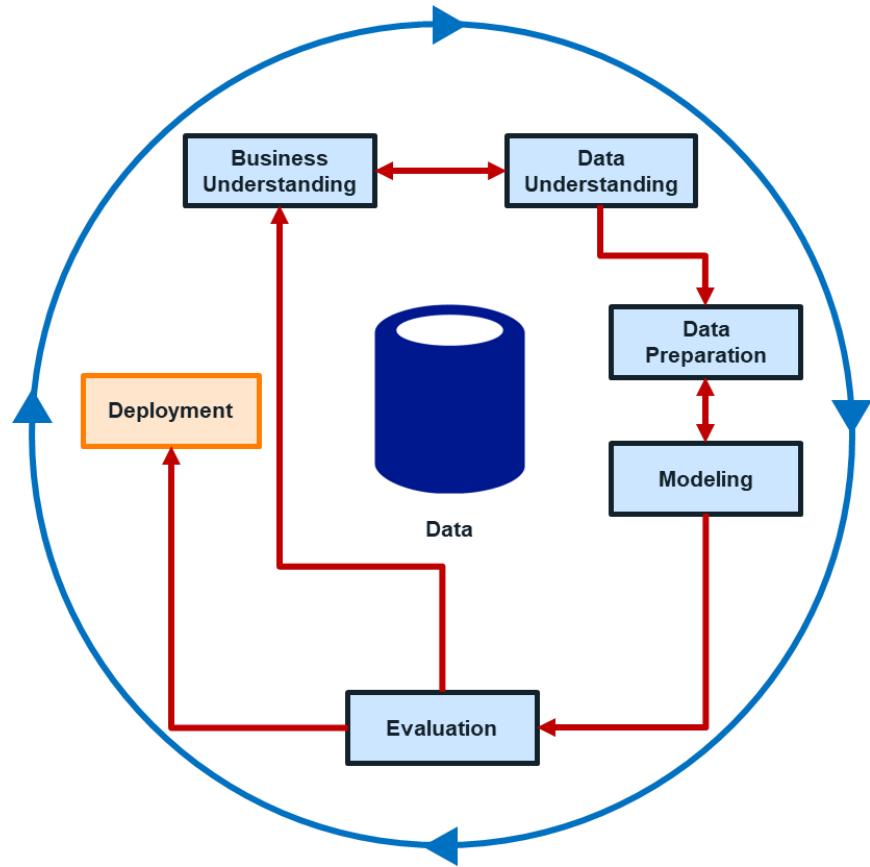
$$\frac{SSE}{SST}$$

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**