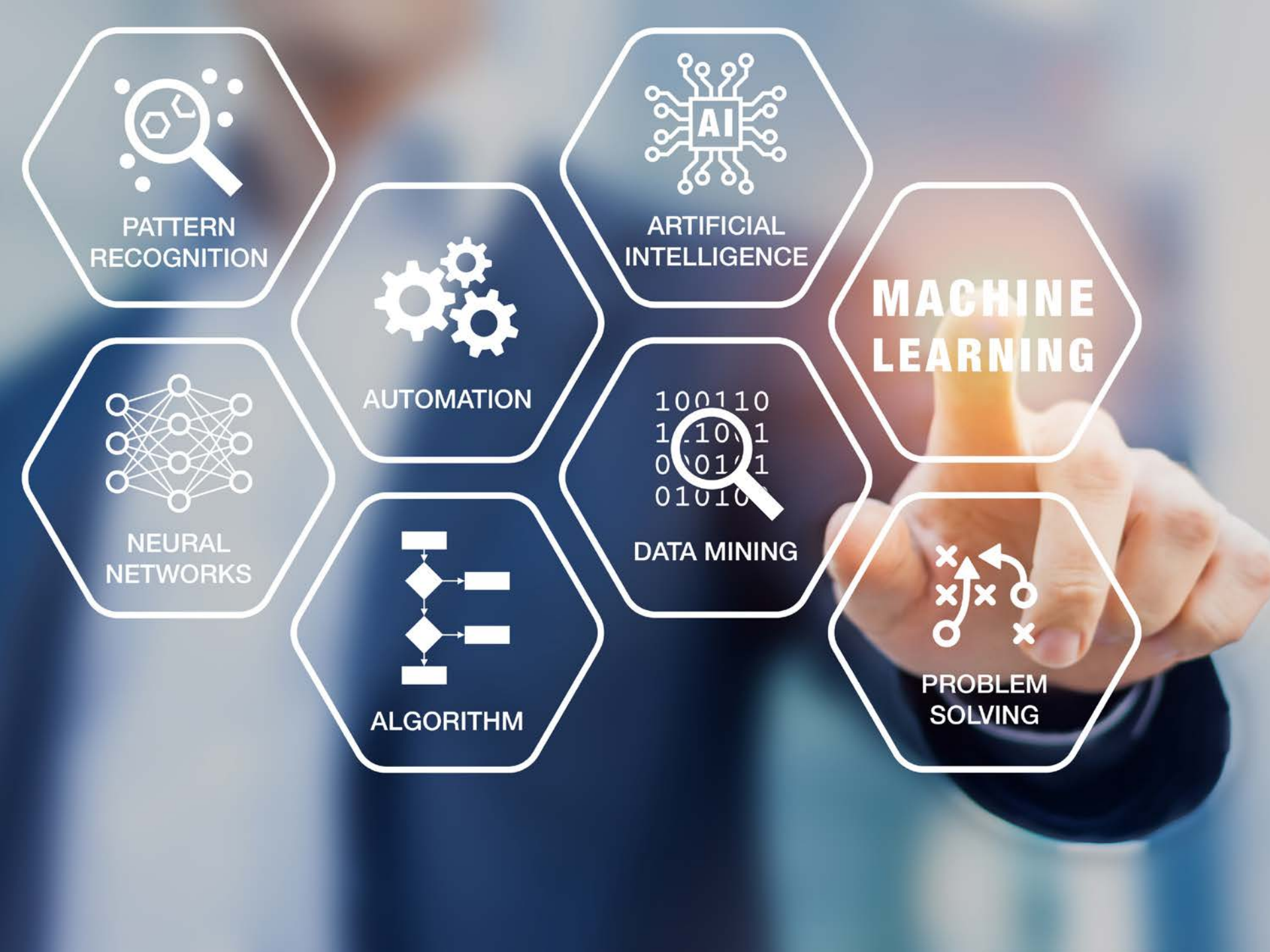
Introduction to Data Mining

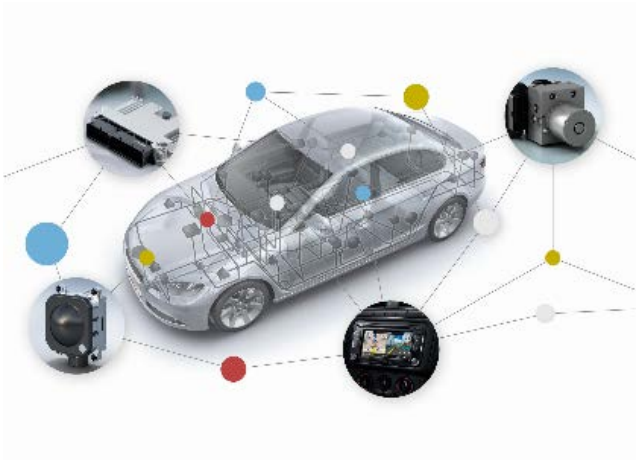
Prof. Dr. Stephan Trahasch
Offenburg University of Applied Sciences

Outline

- Introduction
- Applications
- Summary


PATTERN RECOGNITION
AI
ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
AUTOMATION
NEURAL NETWORKS
100110
DATA MINING
ALGORITHM
PROBLEM SOLVING

Digitalization in technology, production, living etc. is steadily increasing



Flexible Assignement

(Predictive) Analytics

Energy

Industrial Production

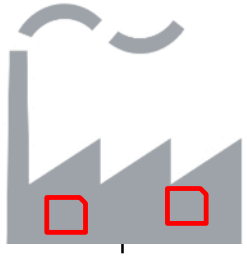
Sensor Actor

Working Environment

Machine

Human

Product



Local Communication

Middleware

Cloud

Internet

More and more data is generated, collected and stored

Banking, telecommunication, commerce

150 PB on 50k+ servers
running 15k apps (6/2011)

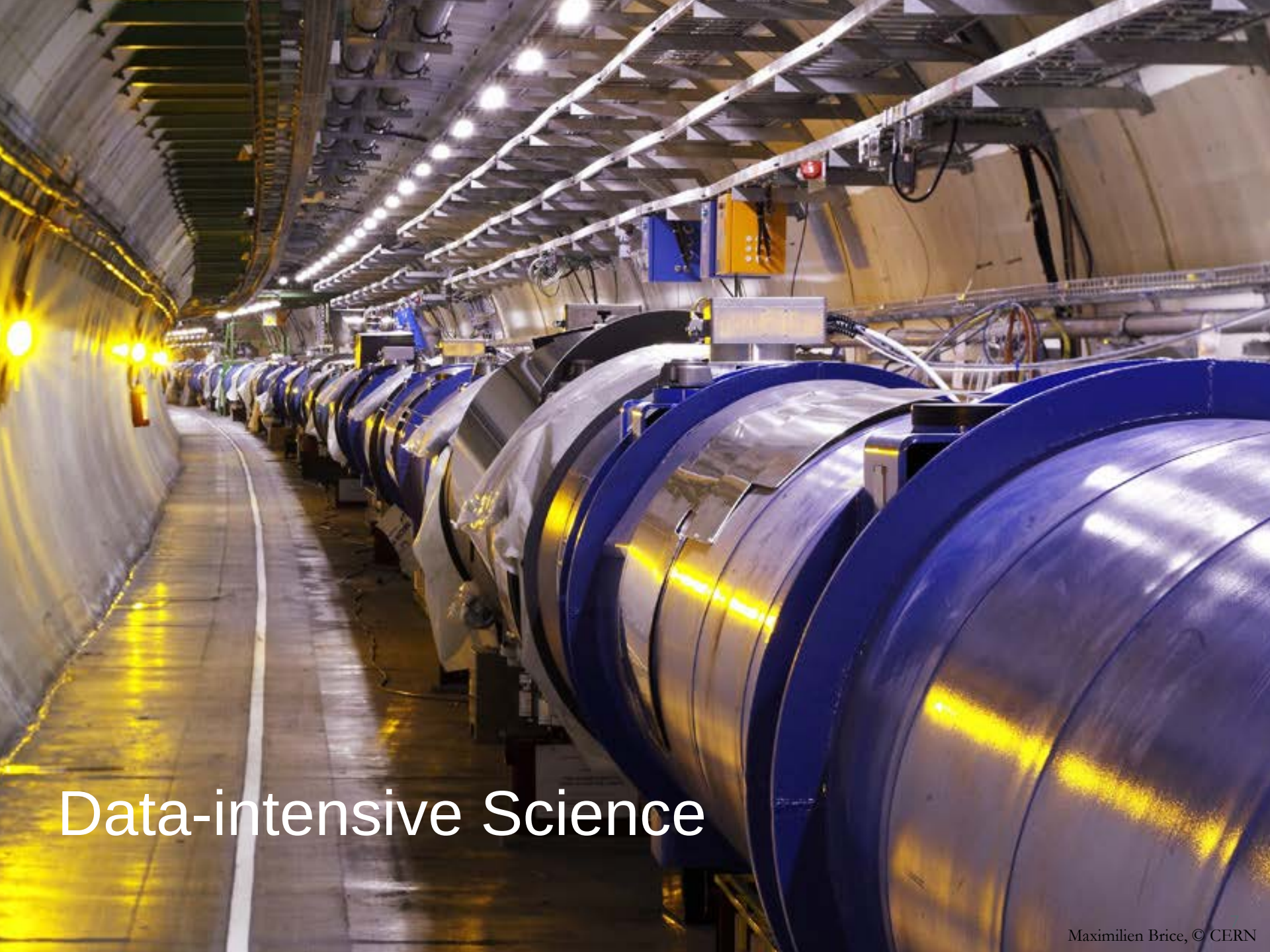
Web, mobile, social media, TV …

Crawls 20B web pages a day (2012)
Search index is 100+ PB (5/2014)
Bigtable serves 2+ EB, 600M QPS (5/2014)

Scientific data and engineering

LHC: ~15 PB a year

NSA …



Maximilien Brice, © CERN

Data is a valuable resource and a competitive advantage

However, raw data is useless.

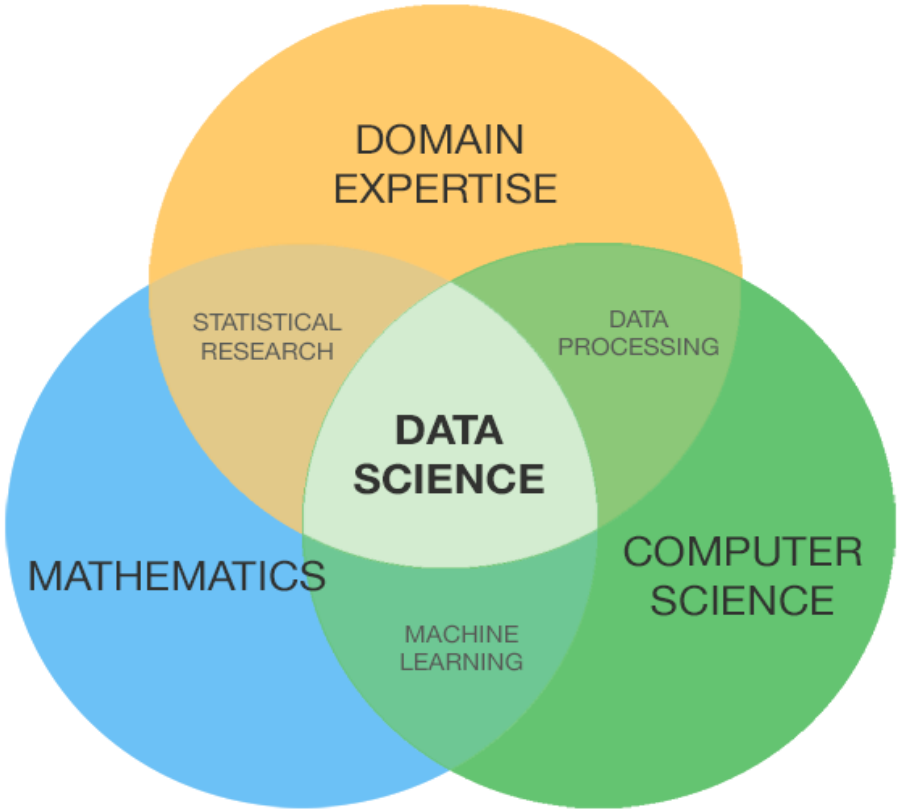
Techniques for the automatic extraction of information are required.

How do I extract knowledge from data?

This is one of the most exciting questions of the information age

→ Data Mining / Machine Learning

Data Mining is an important part of Data Science



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Data Mining

Quelle: http://www.eintagdeutschland.de/sites/default/files/arbeiten/Tjaden_Oliver_B_02.jpg

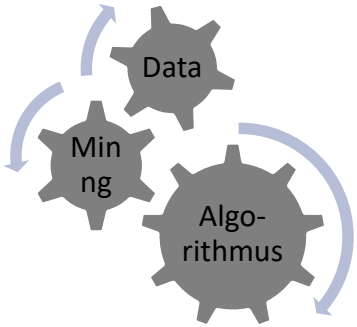
Definition of Data Mining

Data Mining is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. (Fayyad et al. 1996)

Extraction of

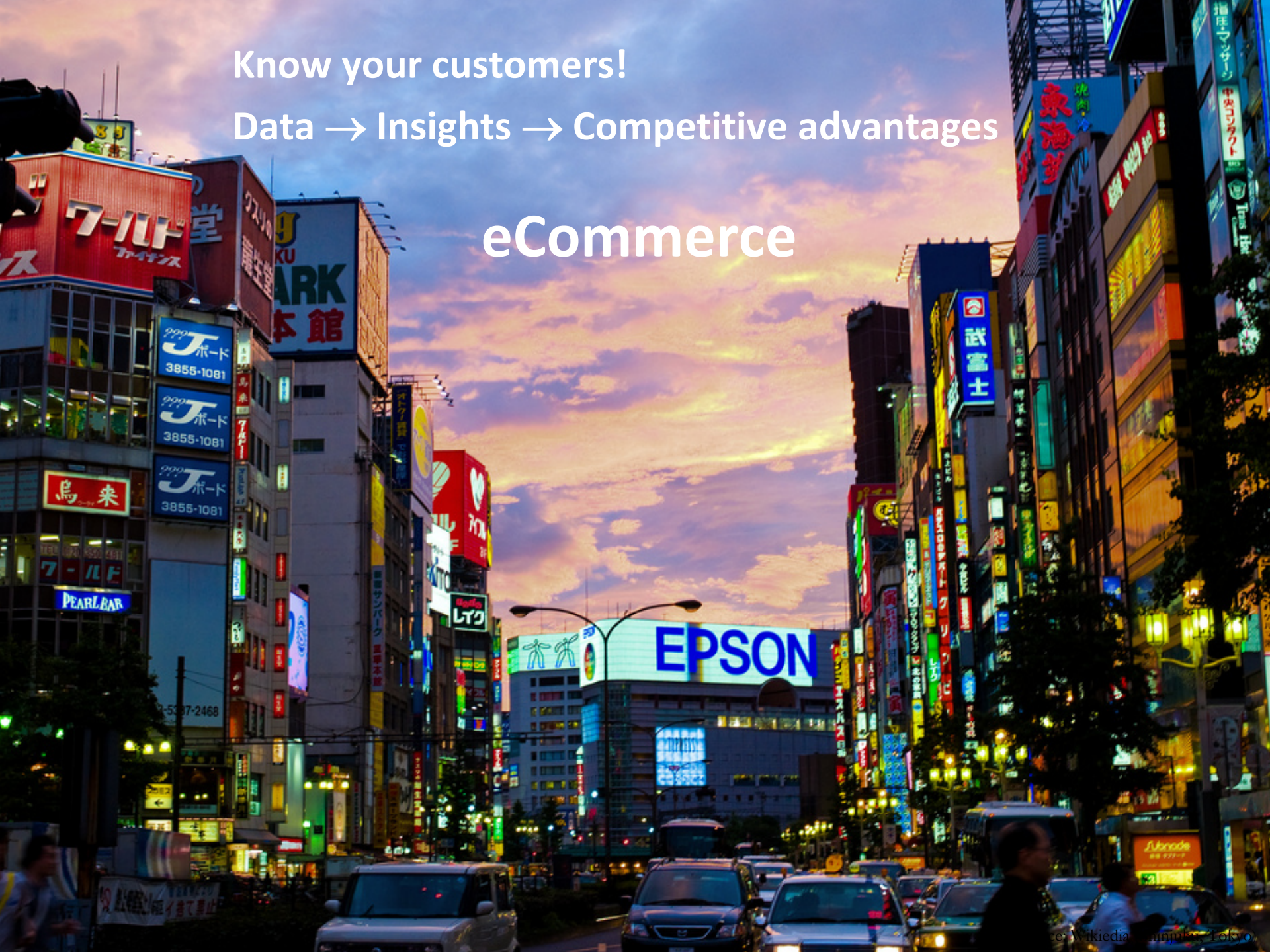
Implicit, new and (hopefully) useful

patterns.



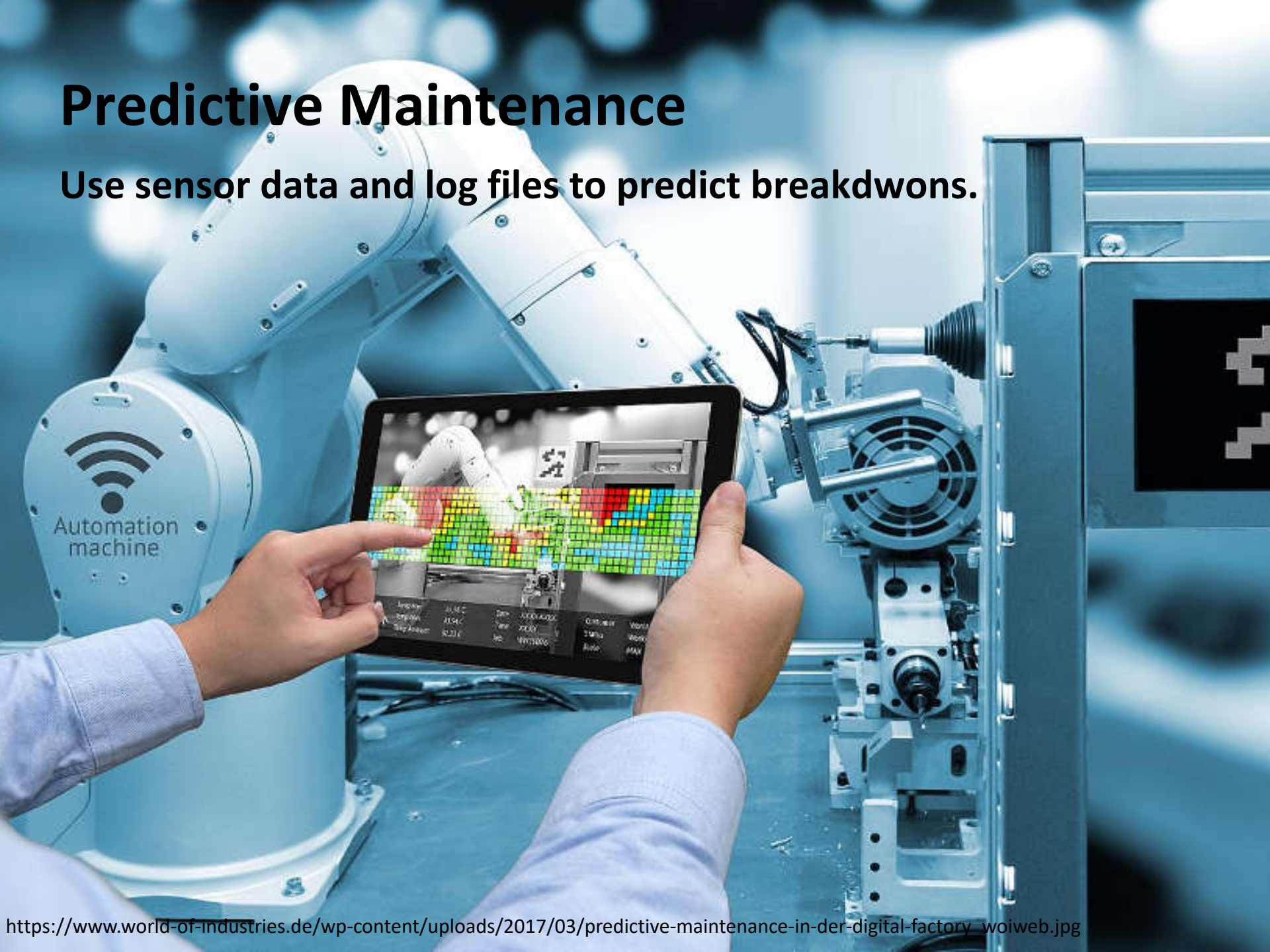
Fundamental Questions

1. How do I extract knowledge from data?
2. How to automatically measure the quality of the prediction (aka knowledge)?
3. How to decide, if a pattern is useful or not?
4. Should I try to predict everything?
   Should I use all attributes for prediction?


Know your customers!
Data → Insights → Competitive advantages
eCommerce
EPSON

Predictive Maintenance

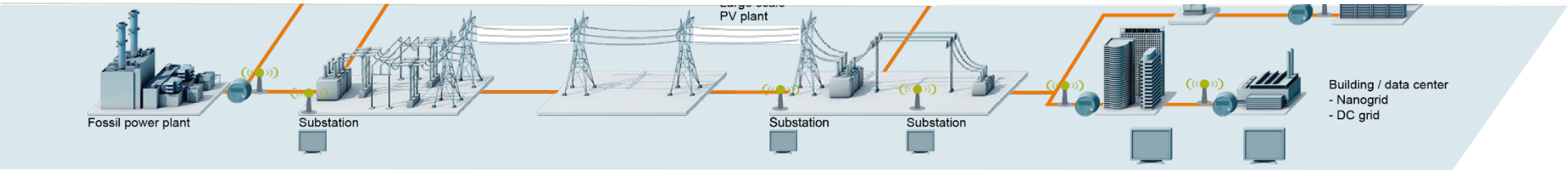
Use sensor data and log files to predict breakdwons.



https://www.world-of-industries.de/wp-content/uploads/2017/03/predictive-maintenance-in-der-digital-factory_woiweb.jpg


CYBER
CYBER CRIME

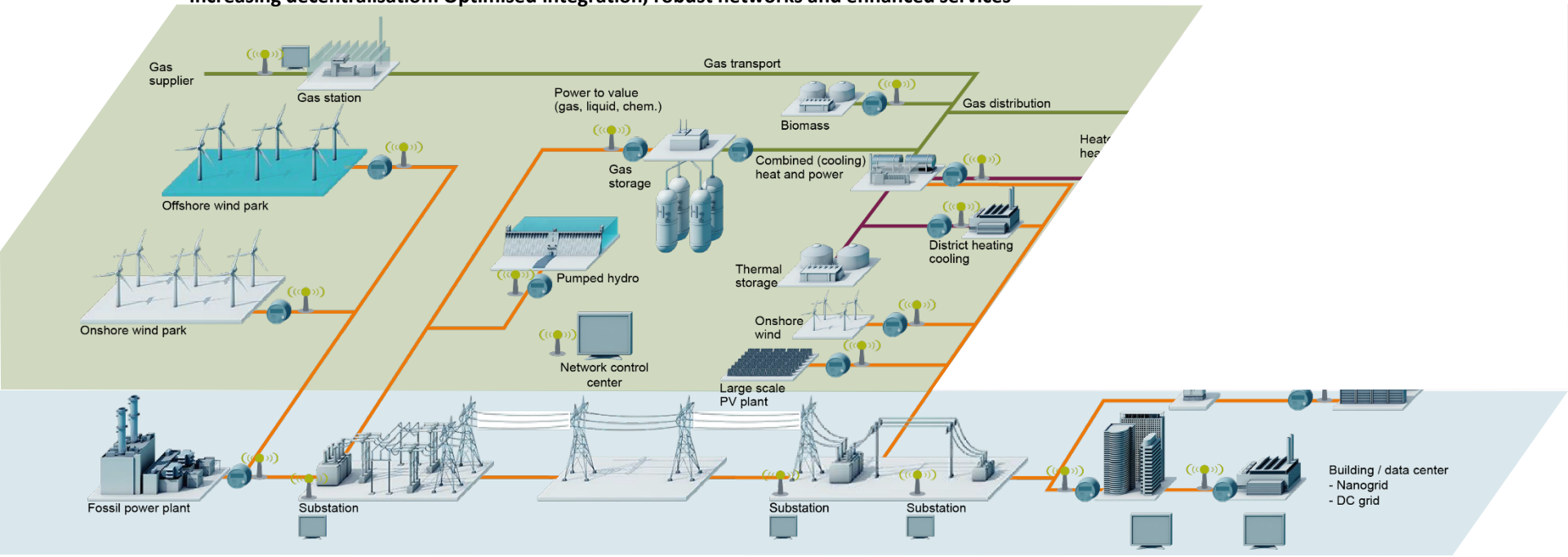
Our energy system is changing rapidly!



Classic model: supplier to consumer

Our energy system is changing rapidly!

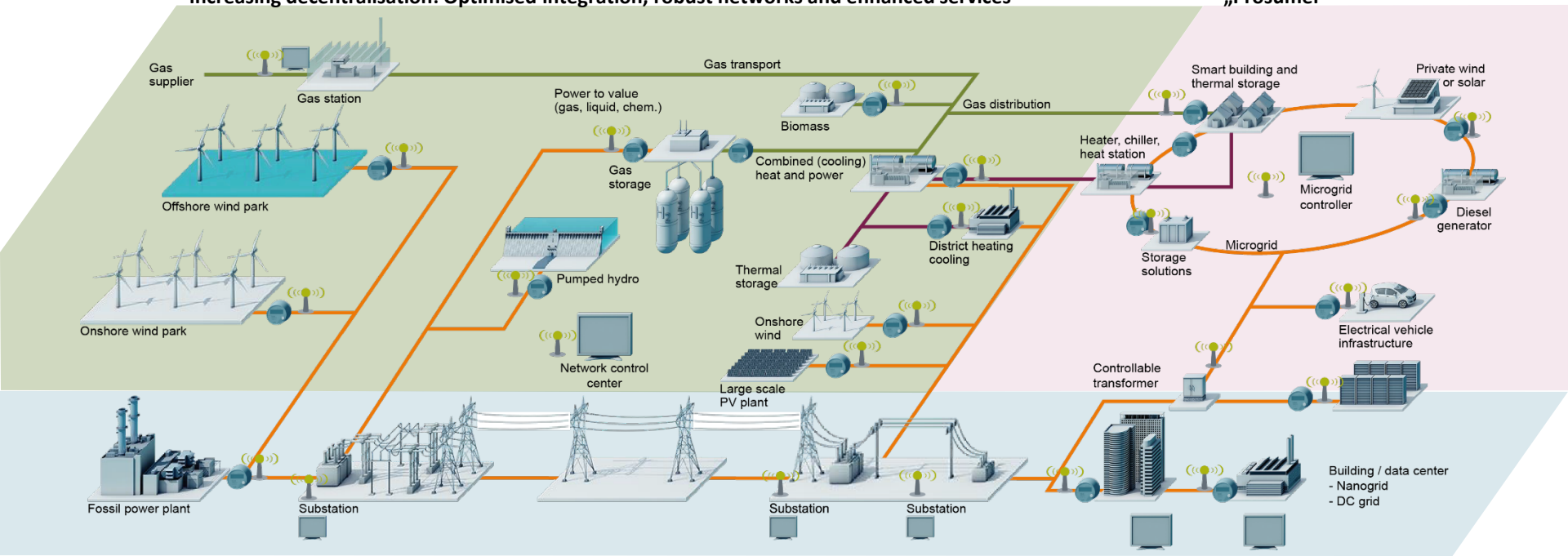
Increasing decentralisation: Optimised integration, robust networks and enhanced services



Classic model: supplier to consumer

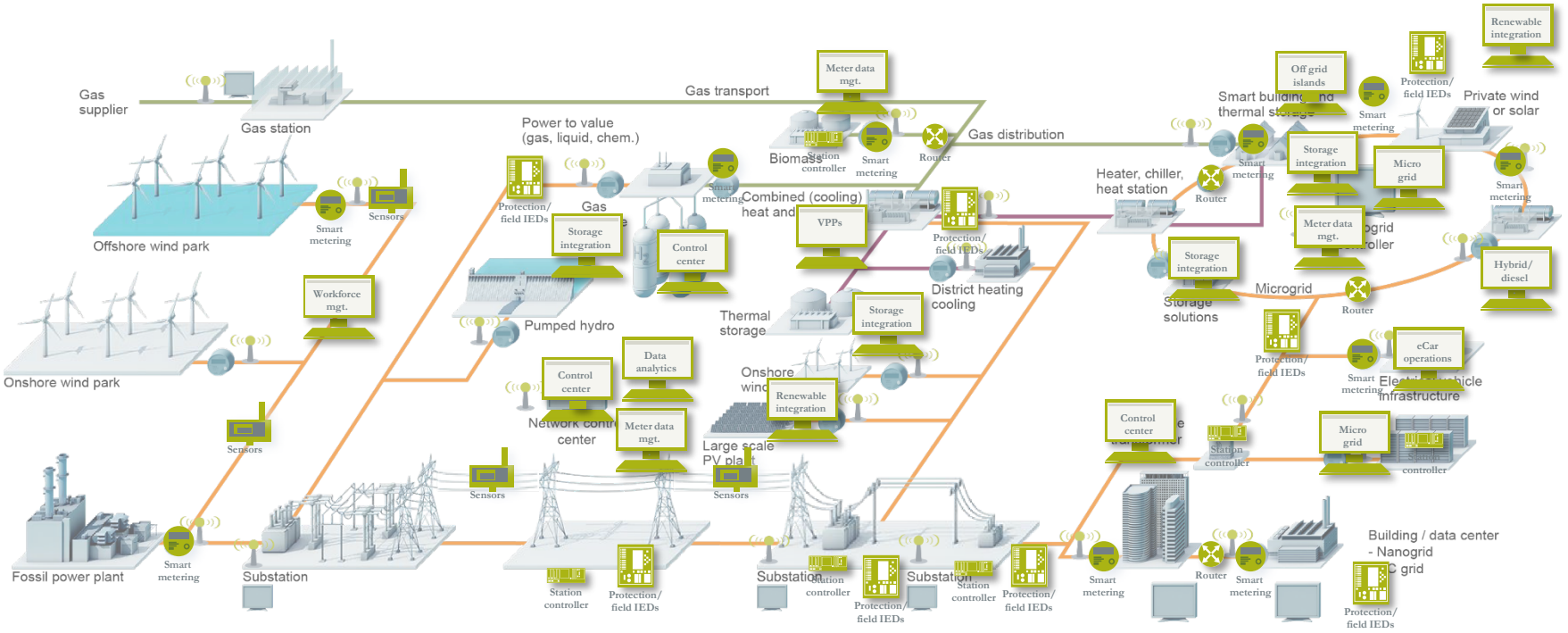
Our energy system is changing rapidly!

Increasing decentralisation: Optimised integration, robust networks and enhanced services **„Prosumer“**



Classic model: supplier to consumer

Energy Industry is in the Digitalization Transformation!

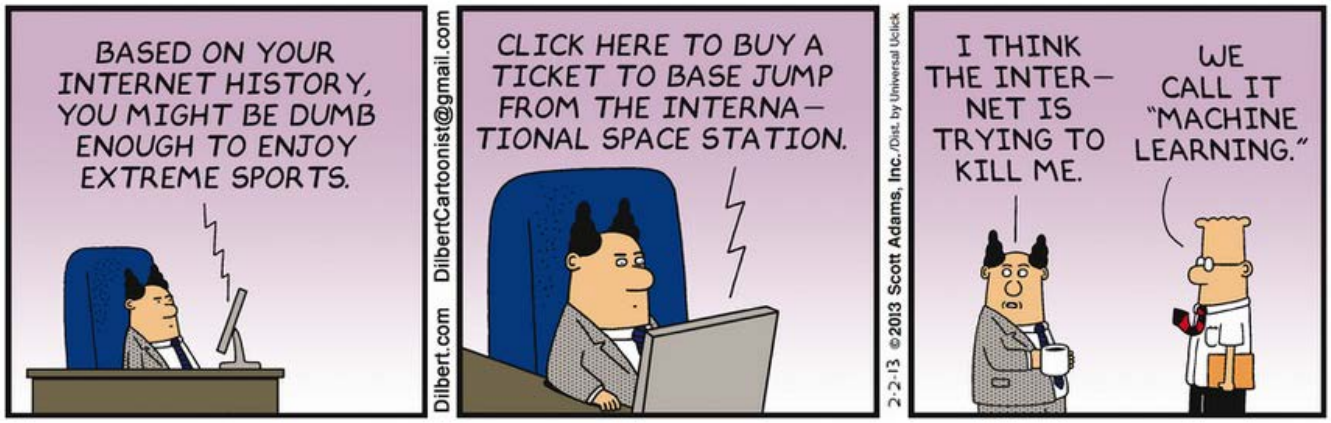


Machine Learning

Machine Learning is a research field that deals with the computer-aided modelling and realization of learning.

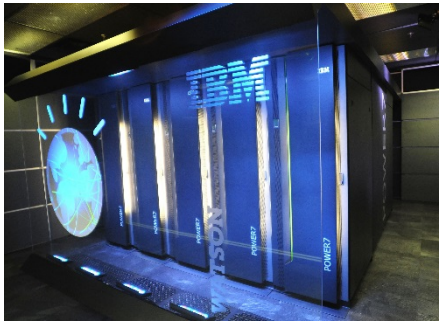
Machine Learning uses learning techniques to achieve adaptive behavior.

Applications: Games, Robotics, autonomous vehicles etc.

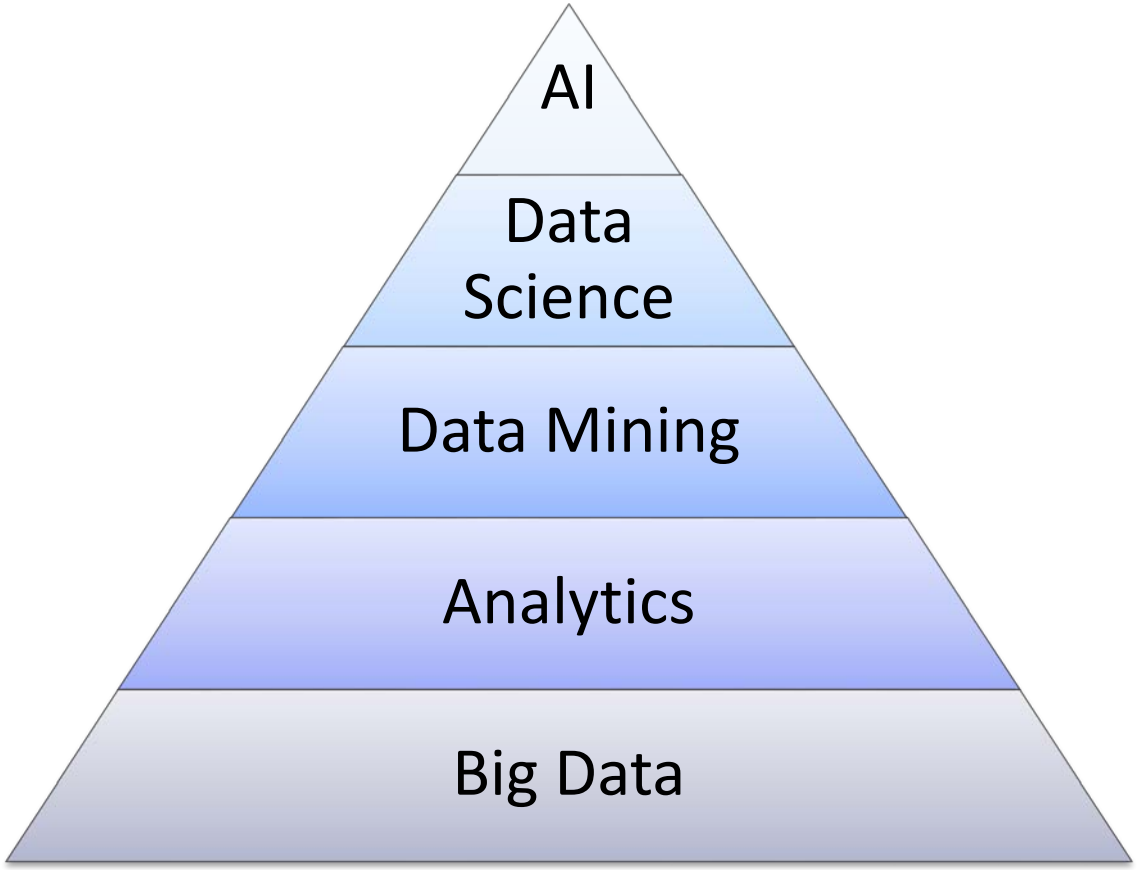


Artificial Intelligence (AI)

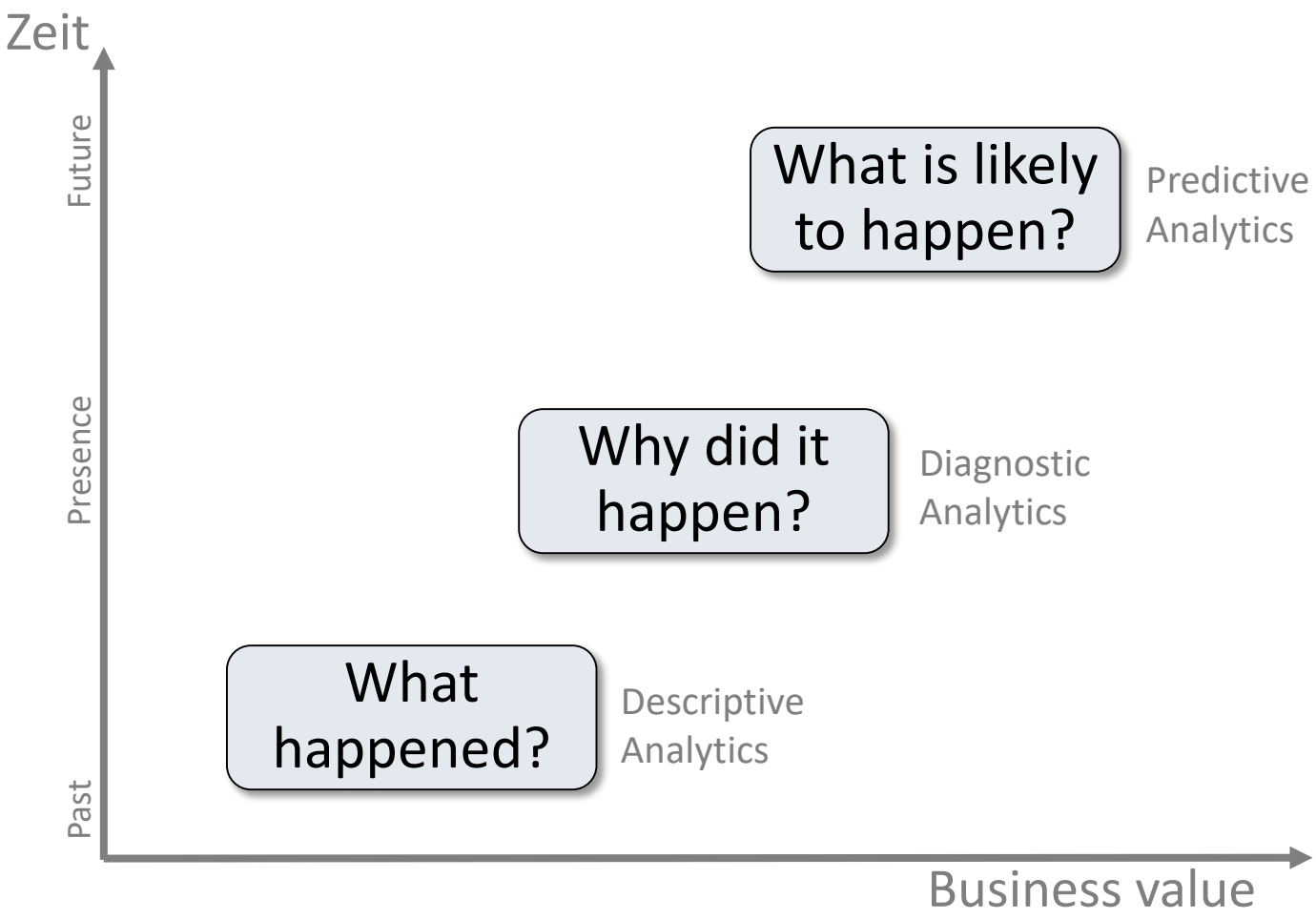
Building “intelligent systems”. Lots of parts to intelligent behavior



AI, Data Science, Data Mining … What's the difference?



The Evolution of Data Analysis

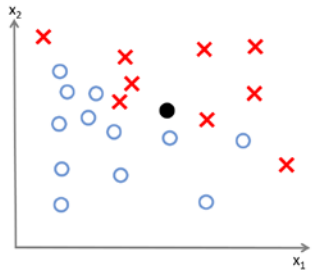


The Evolution of Data Analysis

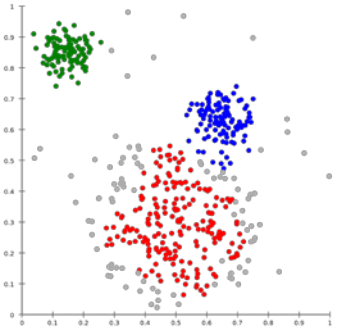
Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), SQL, ODBC	Oracle, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Ware-housing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	OLAP, Multi-dimensional databases, data warehouses	SAP, Oracle, IBM …	Retrospective dynamic data delivery at multiple Levels
Data Mining	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive Databases	Weka, R, Python, RapidMiner, …	Prospective, Proactive information delivery

Data Mining Tasks

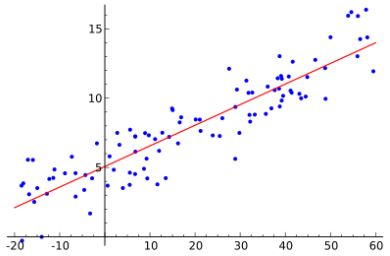
Classification



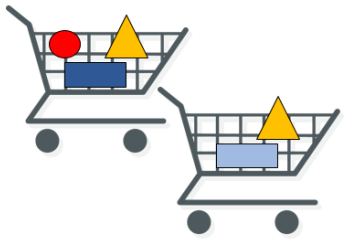
Clustering



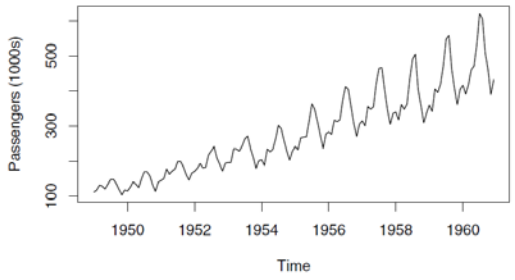
Regression



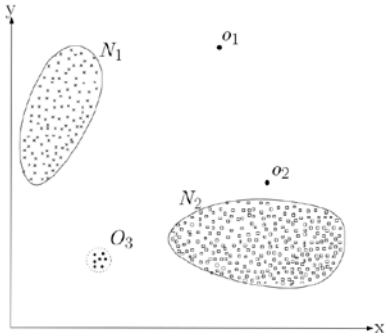
Association Pattern



Time Series



Outlier Analysis



Data Mining

Requirements: We have data

Task: Find patterns and underlying rules in data.

How can we solve the problem automatically?
→ Data Mining Algorithms

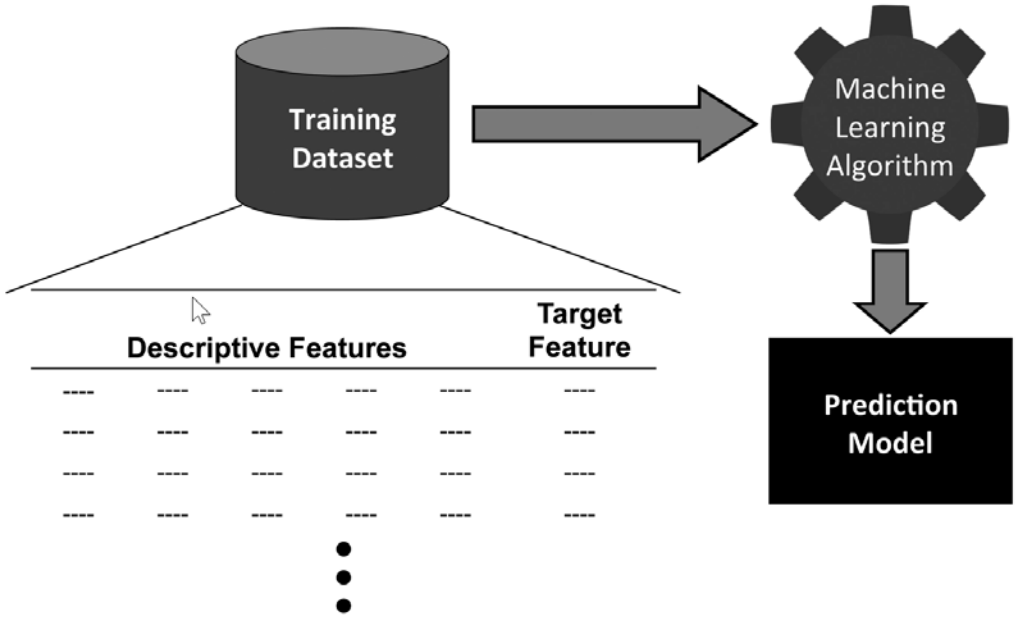
Challenges:

- How to automatically measure the quality of the result?
- How to decide, if a pattern is useful or not?

Learned patterns and rules can be used for prediction.

Keep in mind that there are not always underlying patterns and rules that can be learned, even if we have a lot of data!

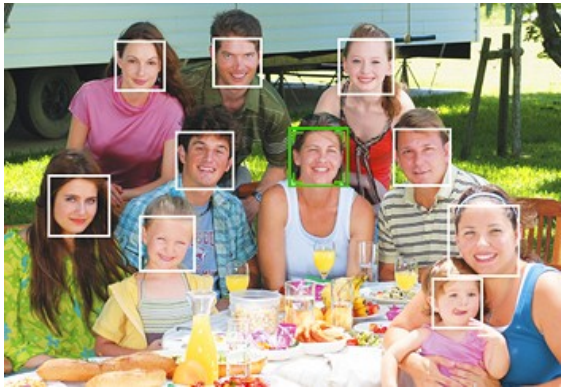
Data Mining techniques automatically learn a model of the relationship between a set of descriptive features and a target feature from a set of historical examples.



Source: Fundamentals of Machine Learning for Predictive Data Analytics, 2015

Types of prediction problems

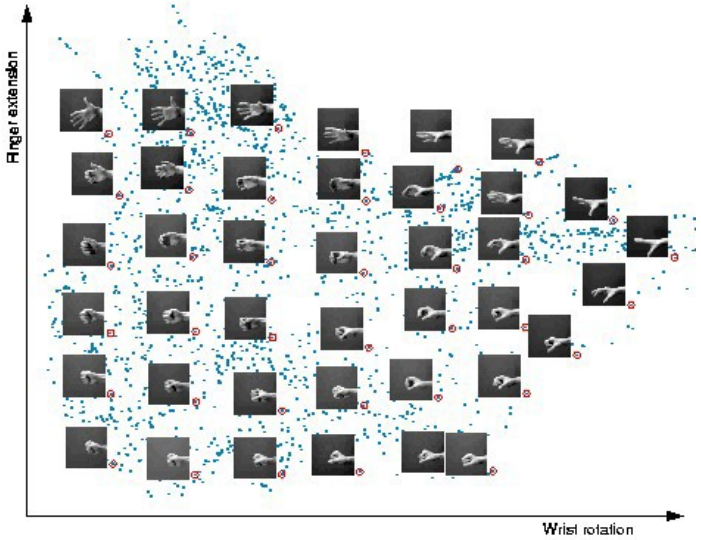
- **Supervised learning**
  - "Labeled" training data
  - Every example has a desired target value, a "best answer"
  - Reward prediction being close to target



57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0	1
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0	0
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0	1
18,M,165,0,110,80,41,30,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0	0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	1
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	1

Types of prediction problems

- Supervised learning
- **Unsupervised learning**
  - No known target values
  - No targets = nothing to predict?
  - Reward “patterns” or “explaining features”
  - Often, data mining



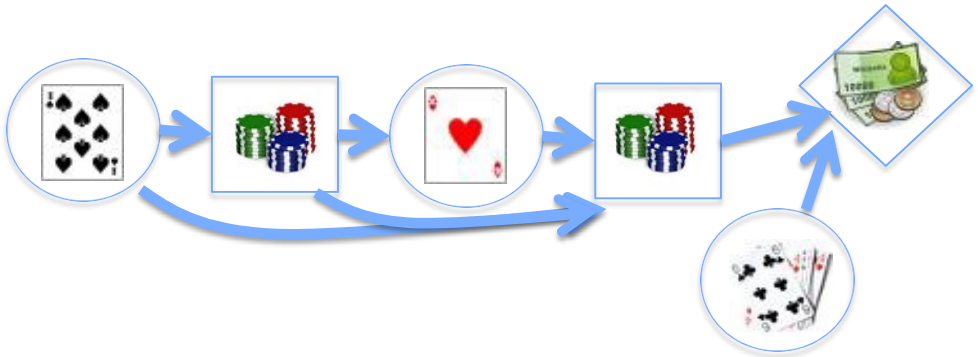

```
57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1
```

Types of prediction problems

- Supervised learning
- Unsupervised learning
- **Semi-supervised learning**
  - Similar to supervised
  - some data have unknown target values
  - Example: medical data
    Lots of patient data, few known outcomes
  - Example: image tagging
    Lots of images on Flickr, but only some of them tagged

Types of prediction problems

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- **Reinforcement learning**
  - "Indirect" feedback on quality
  - No answers, just "better" or "worse"
  - Feedback may be delayed



Terminology

Object/concept	Objects of interest like customer, product, machine
Features or attributes	A set of characteristic features or attributes that describe the object (quantitativ and/or qualitative)
Label or target	A specific attribute to be predicted
Sample data	Data of observations. One observation is called an instance or example.
Model	A algorithm or function that assigns an output value of the target attribute to objects. This function minimizes an error function or optimizes a quality function.

Example: Predict price of a used car

Object/concept	Used car
Features or attributes	Type, kilometer, fuelType, condition, color ,engine power …
Label or target	Price
Sample data	Historical sales data
Model	A function that predicts the price of a used car. The model determines how the features are combined to determine the price. This is what we want to learn.

Example: Predict price of a used car

dateCrawled	name	seller	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage	dateCreated	nrOfPictures	postalCode	lastSeen
24.03.2016 11:52	Golf_3_1.6	privat	Angebot	480	test		1993	manuell	0	golf	150000	0	benzin	volkswagen		24.03.2016 00:00	0	70435	07.04.2016 03:16
24.03.2016 10:58	A5_Sportback_2.7_Tdi	privat	Angebot	18300	test	coupe	2011	manuell	190		125000	5	diesel	audi	ja	24.03.2016 00:00	0	66954	07.04.2016 01:46
14.03.2016 12:52	Jeep_Grand_Cherokee_"Overland"	privat	Angebot	9800	test	suv	2004	automatik	163	grand	125000	8	diesel	jeep		14.03.2016 00:00	0	90480	05.04.2016 12:47
17.03.2016 16:54	GOLF_4_1_4__3TÜR	privat	Angebot	1500	test	kleinwagen	2001	manuell	75	golf	150000	6	benzin	volkswagen	nein	17.03.2016 00:00	0	91074	17.03.2016 17:40
31.03.2016 17:25	Skoda_Fabia_1.4_TDI_PD_Classic	privat	Angebot	3600	test	kleinwagen	2008	manuell	69	fabia	90000	7	diesel	skoda	nein	31.03.2016 00:00	0	60437	06.04.2016 10:17
04.04.2016 17:36	BMW_316i___e36_Limousine___Bastlerfahrzeug__Export	privat	Angebot	650	test	limousine	1995	manuell	102	3er	150000	10	benzin	bmw	ja	04.04.2016 00:00	0	33775	06.04.2016 19:17
01.04.2016 20:48	Peugeot_206_CC_110_Platinum	privat	Angebot	2200	test	cabrio	2004	manuell	109	2_reihe	150000	8	benzin	peugeot	nein	01.04.2016 00:00	0	67112	05.04.2016 18:18
21.03.2016 18:54	VW_Derby_Bj_80__Scheunenfund	privat	Angebot	0	test	limousine	1980	manuell	50	andere	40000	7	benzin	volkswagen	nein	21.03.2016 00:00	0	19348	25.03.2016 16:47
04.04.2016 23:42	Ford_C___Max_Titanium_1_0_L_EcoBoost	privat	Angebot	14500	control	bus	2014	manuell	125	c_max	30000	8	benzin	ford		04.04.2016 00:00	0	94505	04.04.2016 23:42
17.03.2016 10:53	VW_Golf_4_5_tuerig_zu_verkaufen_mit_Anhaengerkupplung	privat	Angebot	999	test	kleinwagen	1998	manuell	101	golf	150000	0		volkswagen		17.03.2016 00:00	0	27472	31.03.2016 17:17
26.03.2016 19:54	Mazda_3_1.6_Sport	privat	Angebot	2000	control	limousine	2004	manuell	105	3_reihe	150000	12	benzin	mazda	nein	26.03.2016 00:00	0	96224	06.04.2016 10:45
07.04.2016 10:06	Volkswagen_Passat_Variant_2.0_TDI_Comfortline	privat	Angebot	2799	control	kombi	2005	manuell	140	passat	150000	12	diesel	volkswagen	ja	07.04.2016 00:00	0	57290	07.04.2016 10:25
15.03.2016 22:49	VW_Passat_Facelift_35i__"7Sitzer"	privat	Angebot	999	control	kombi	1995	manuell	115	passat	150000	11	benzin	volkswagen		15.03.2016 00:00	0	37269	01.04.2016 13:16
21.03.2016 21:37	VW_PASSAT_1.9_TDI_131_PS_LEDER	privat	Angebot	2500	control	kombi	2004	manuell	131	passat	150000	2		volkswagen	nein	21.03.2016 00:00	0	90762	23.03.2016 02:50
21.03.2016 12:57	Nissan_Navara_2.5DPF_SE4x4_Klima_Sitzheizg_Bluetooth.Doppelkabine	privat	Angebot	17999	control	suv	2011	manuell	190	navara	70000	3	diesel	nissan	nein	21.03.2016 00:00	0	4177	06.04.2016 07:45
11.03.2016 21:39	KA_Lufthansa_Edition_450€_VB	privat	Angebot	450	test	kleinwagen	1910		0	ka	5000	0	benzin	ford		11.03.2016 00:00	0	24148	19.03.2016 08:46
01.04.2016 12:46	Polo_6n_1_4	privat	Angebot	300	test		2016		60	polo	150000	0	benzin	volkswagen		01.04.2016 00:00	0	38871	01.04.2016 12:46
20.03.2016 10:25	Renault_Twingo_1.2_16V_Aut.	privat	Angebot	1750	control	kleinwagen	2004	automatik	75	twingo	150000	2	benzin	renault	nein	20.03.2016 00:00	0	65599	06.04.2016 13:16
23.03.2016 15:48	Ford_C_MAX_2.0_TDCi_DPF_Titanium	privat	Angebot	7550	test	bus	2007	manuell	136	c_max	150000	6	diesel	ford	nein	23.03.2016 00:00	0	88361	05.04.2016 18:45

Used cars database

Over 370,000 used cars scraped from Ebay Kleinanzeigen

https://www.kaggle.com/orgesleka/used-cars-database

Example: Possible questions

- Predict price (Regression or Classification)
- Will the car be sold? (Classification)
- How long will it take until the car is sold? (Classification or Time)
- Identifiy common features of sold cars (Clustering)

Example: Classification

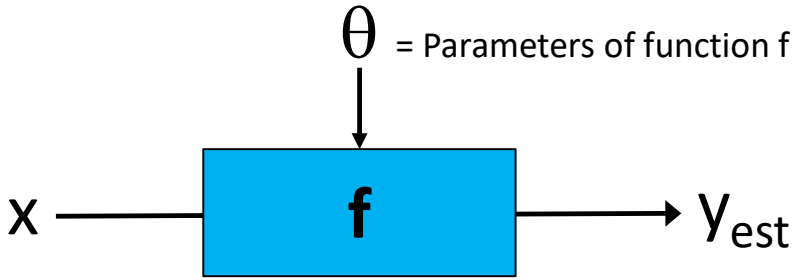
Given Class Y, $y \in \{0, 1\}$,

Set $T = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset X \times Y$ of sample data,

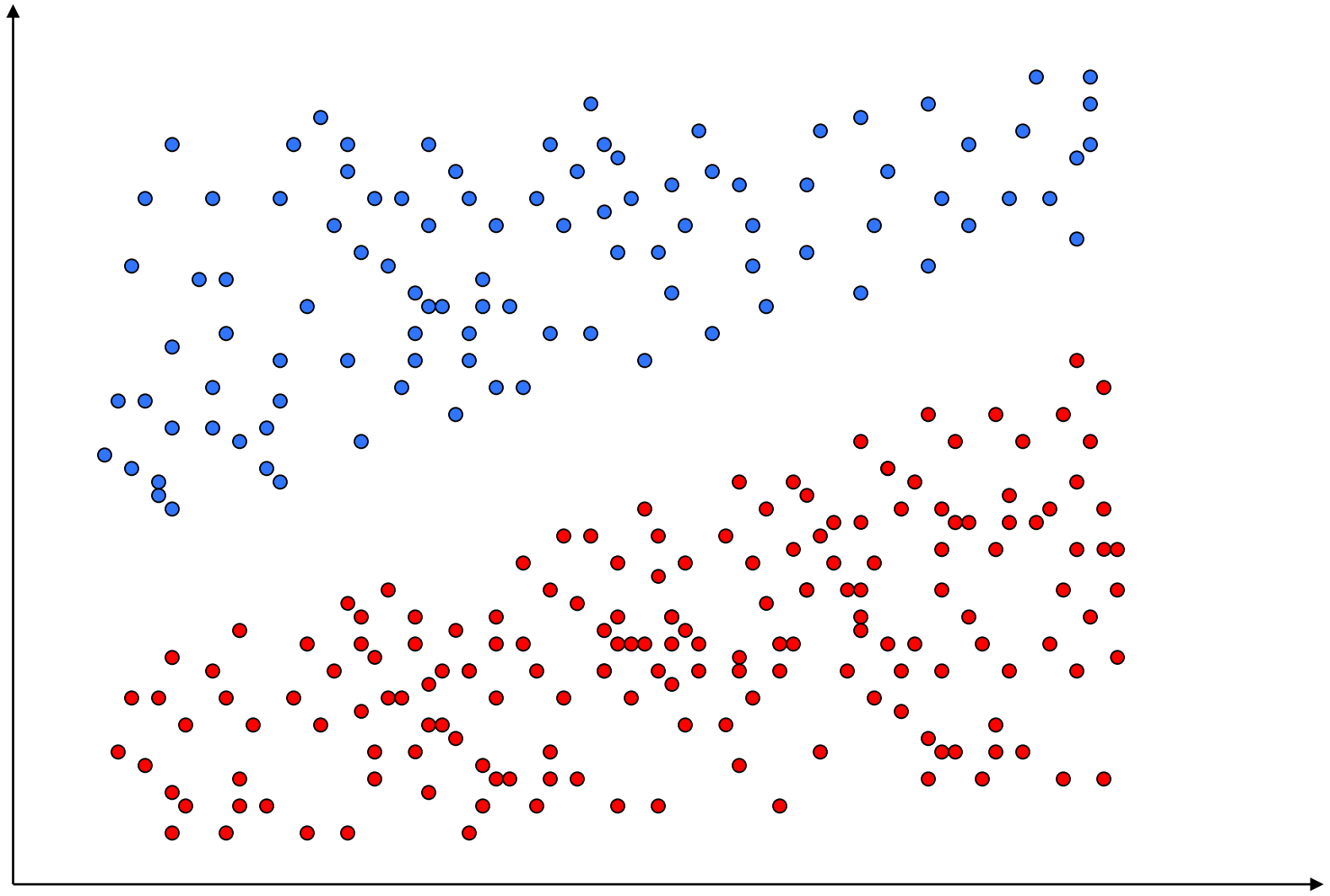
Quality function q

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	1
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	0
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	1
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1	1

Find a function $f : X \rightarrow Y$, which optimizes the quality function.

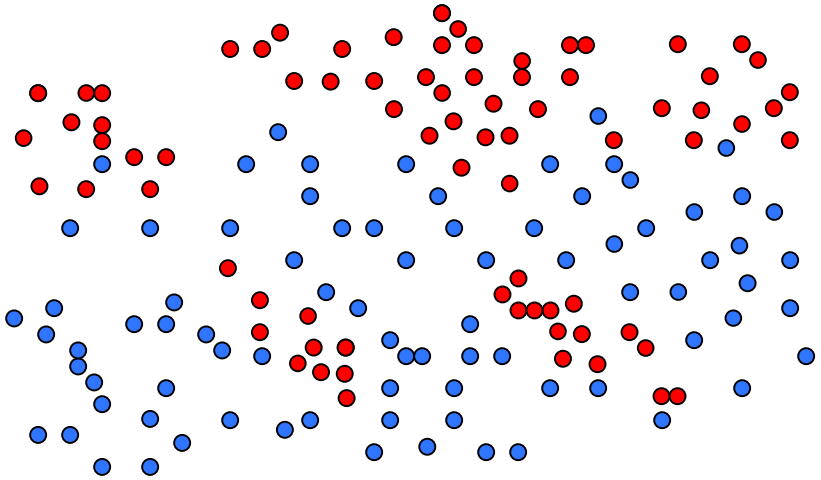


Which function should be selected?

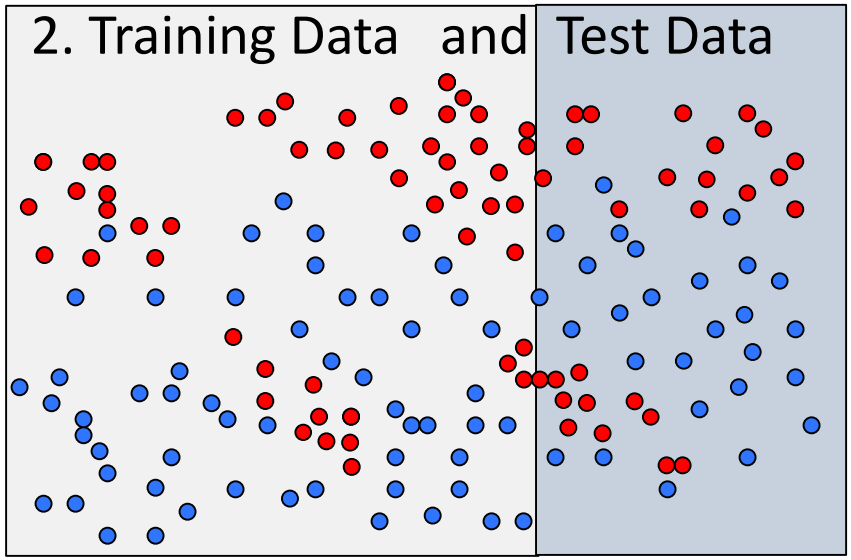


Training and test set

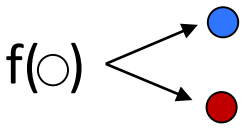
1. Sample Data



2. Training Data and Test Data



3. Find function f



$f(x) = y_{est}$

4. Evaluation of f with test data
   Minimize prediction error

Training and test set

We divide the data we have into a

Training set

Data of the training set is input to our learning algorithm.

From this he learns the function

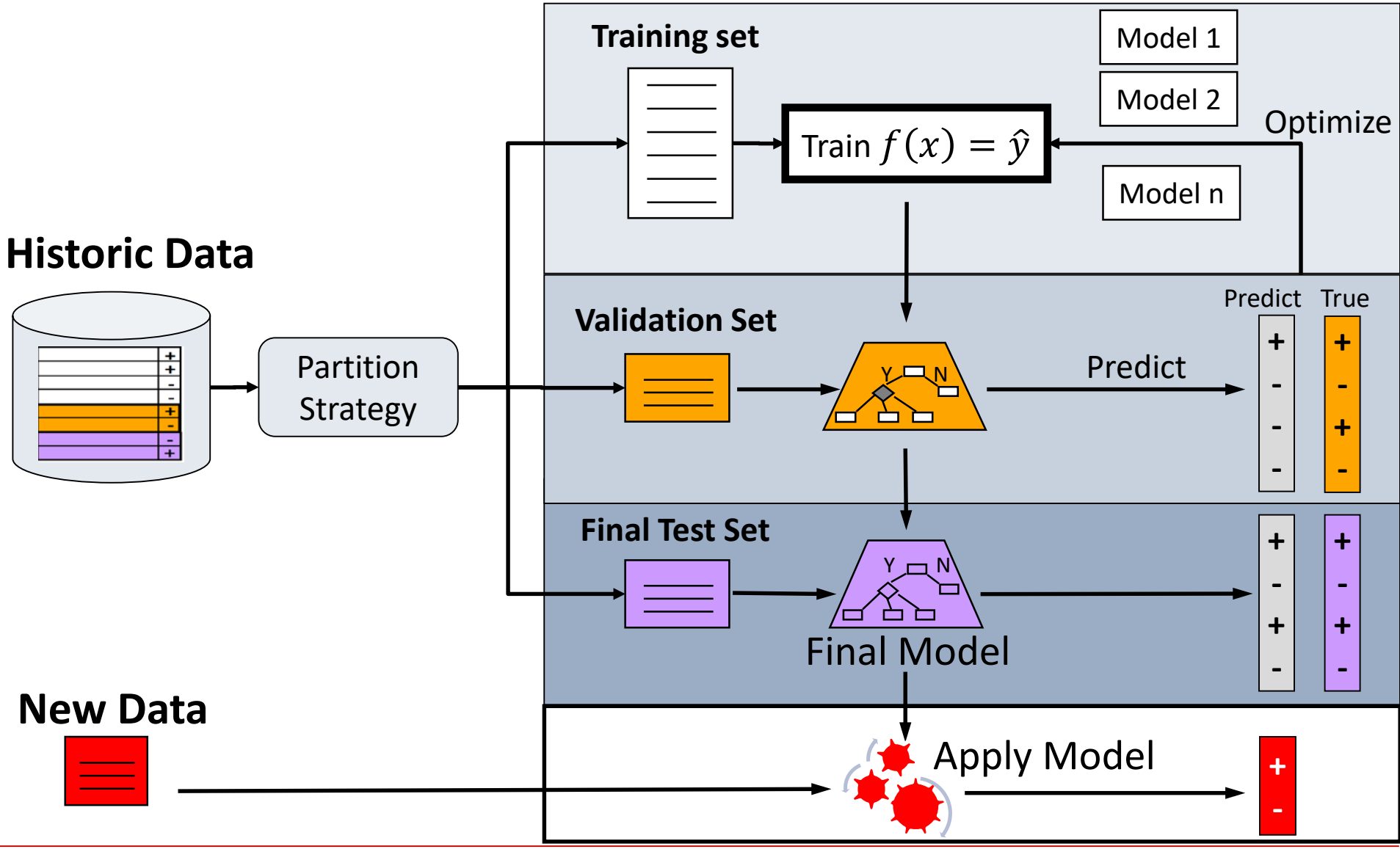
$$f(x) = \hat{y}$$

Test set

To compute the error we compare y (true target value) with $\hat{y}$ (predicted value) -> test of residulas.

Data of the test set is NEVER used for learning/training!

Training, validation and test set



Hyperparameter

Hyperparameters are, for example, weightings on penalty straps, number of decision trees in Random Forets, number of hidden nodes of a neural network,. ....

Division of data into training, validation and test set.

- The **model is trained on the training data** with different values of the hyperparameter.
- We optimize and select the model with the appropriate hyperparameters, which showed the **best performance on the validation data**. → Training Error
- The error of this optimized model is calculated with the **test data**. → Test Error

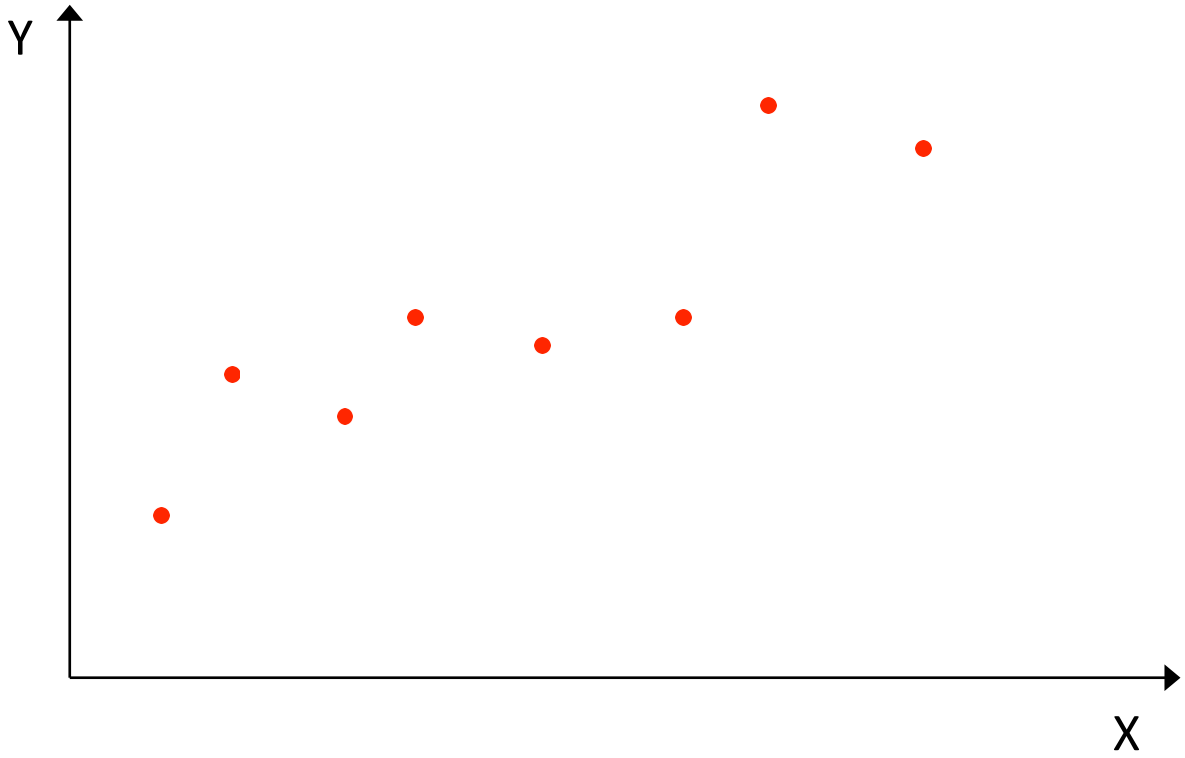
What can go wrong?

Data Mining algorithms work by searching through sets of potential models.

There are two sources of information that guide this search:

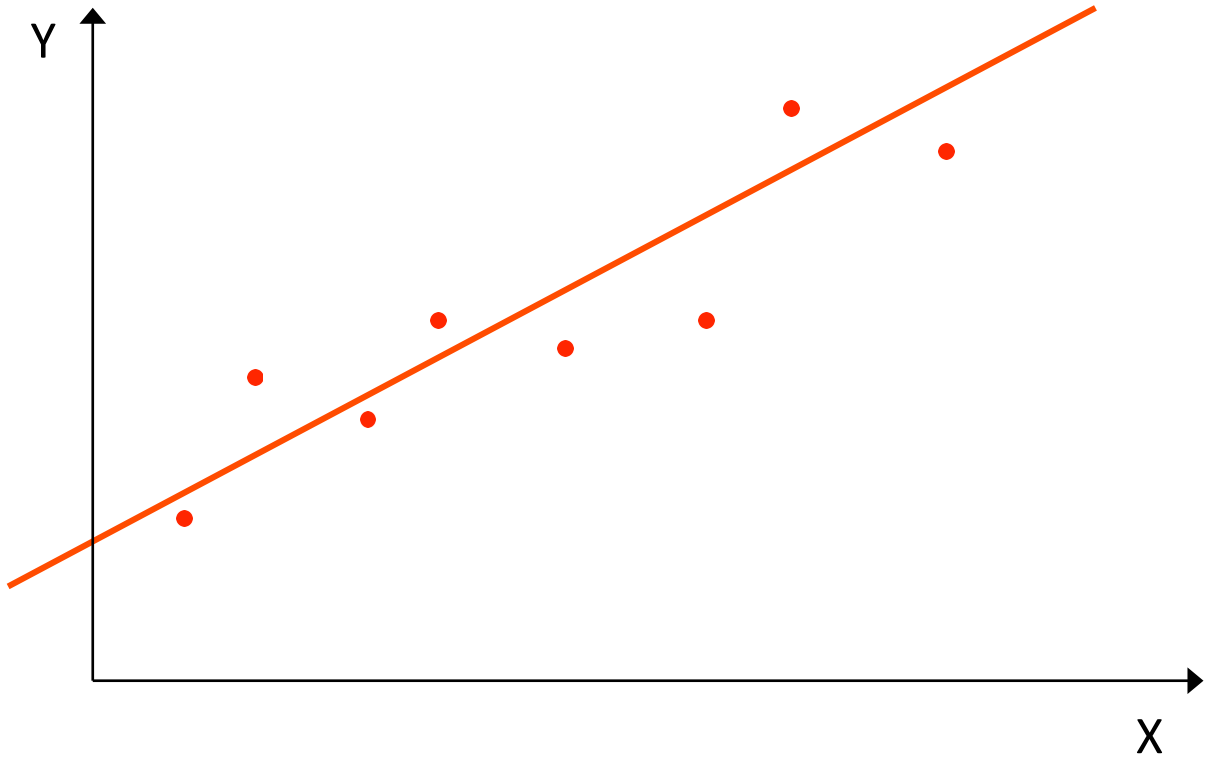
1. the training data,
2. the inductive bias of the algorithm

Overfitting and complexity

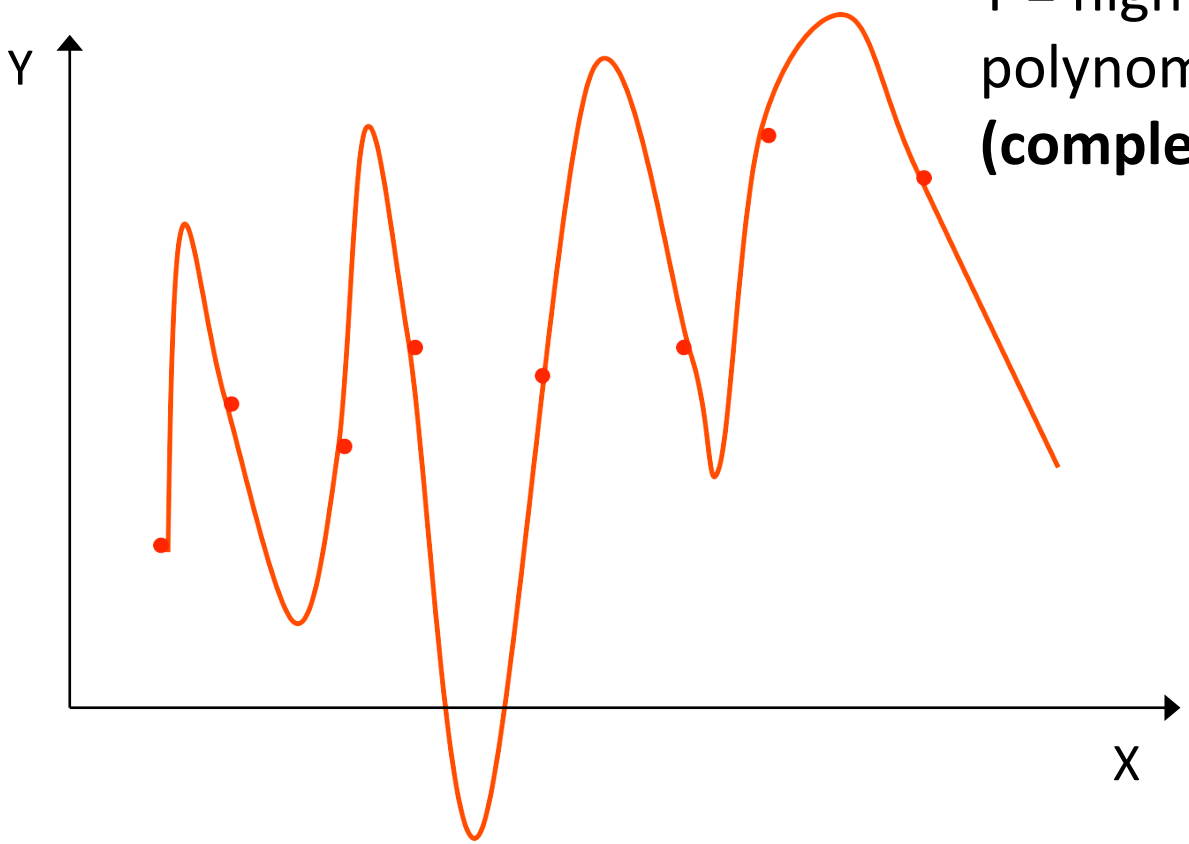


Overfitting and complexity

Simple model: $Y = aX + b + e$

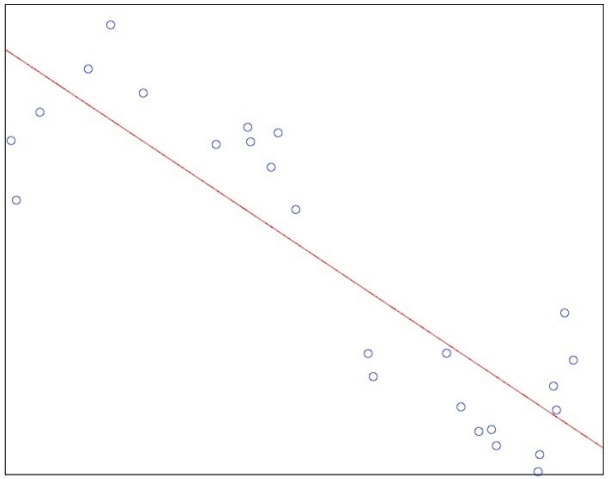


Overfitting and complexity

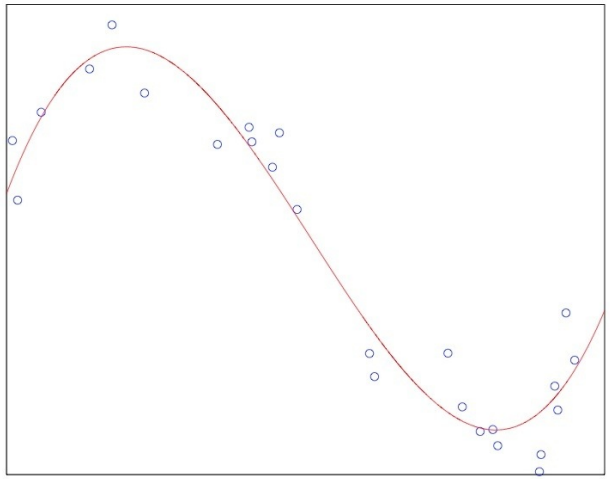


Y = high-order polynomial in X **(complex model)**

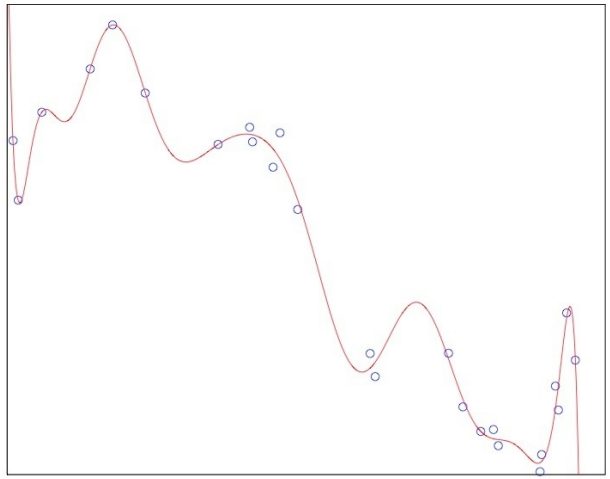
Example of Under- and Overfitting



underfit
degree = 1



Ideal fit
degree = 3



overfit
degree = 20

error(X) = noise(X) + bias(X) + variance(X)

Under- and Overfitting

Underfitting:

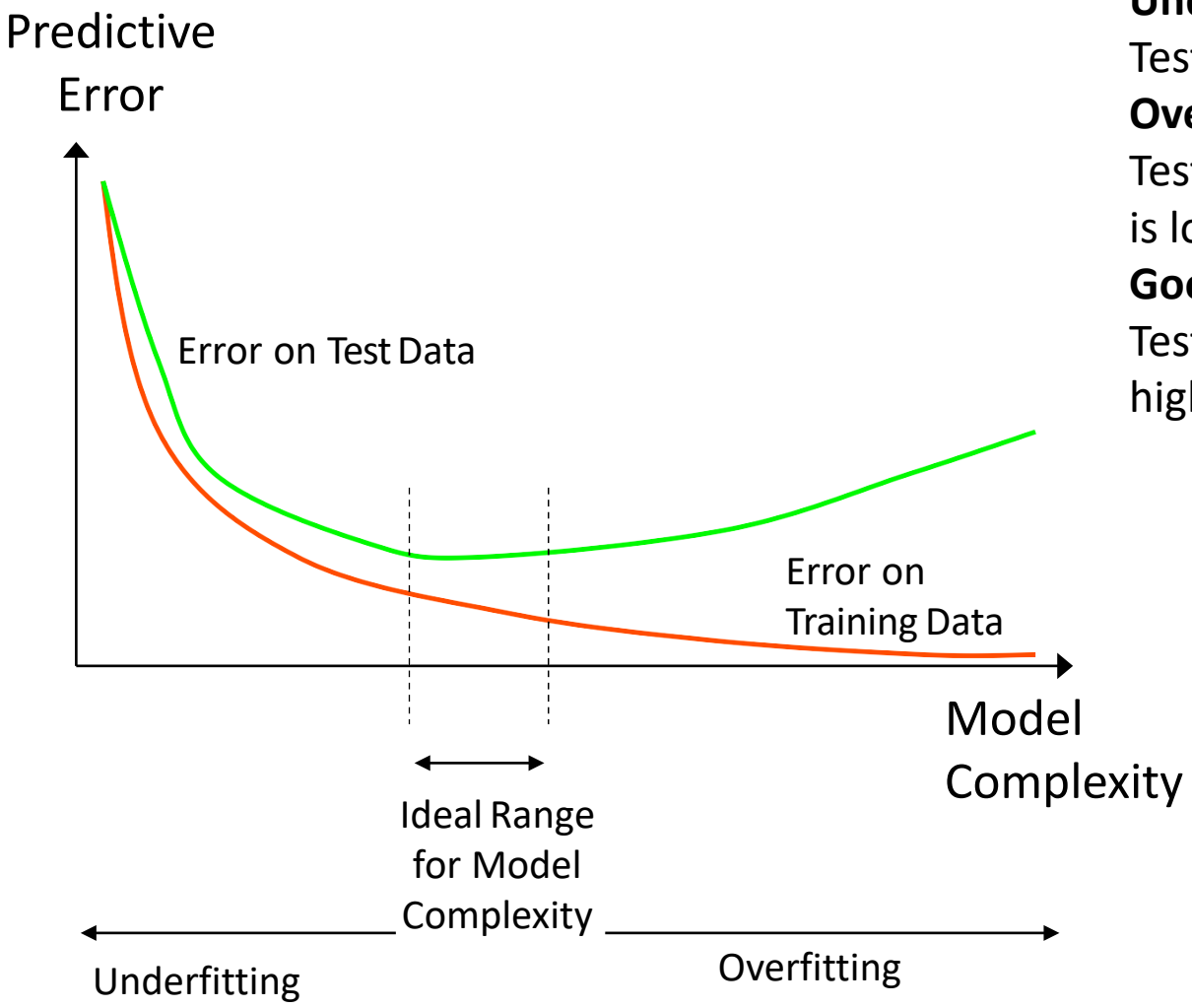
Noisy data and too few examples lead to false results.

Overfitting:

We have learned a model with a small error on the training set, but the error on the test set is bigger. The model (hypothesis) found is over-fitted to the data.

There is a better model or hypothesis with a bigger error on the training data than the one found, but the total error in the test data is smaller.

How Overfitting affects Prediction



Underfitting
Test and training error are both high

Overfitting
Test error is high while training error is low

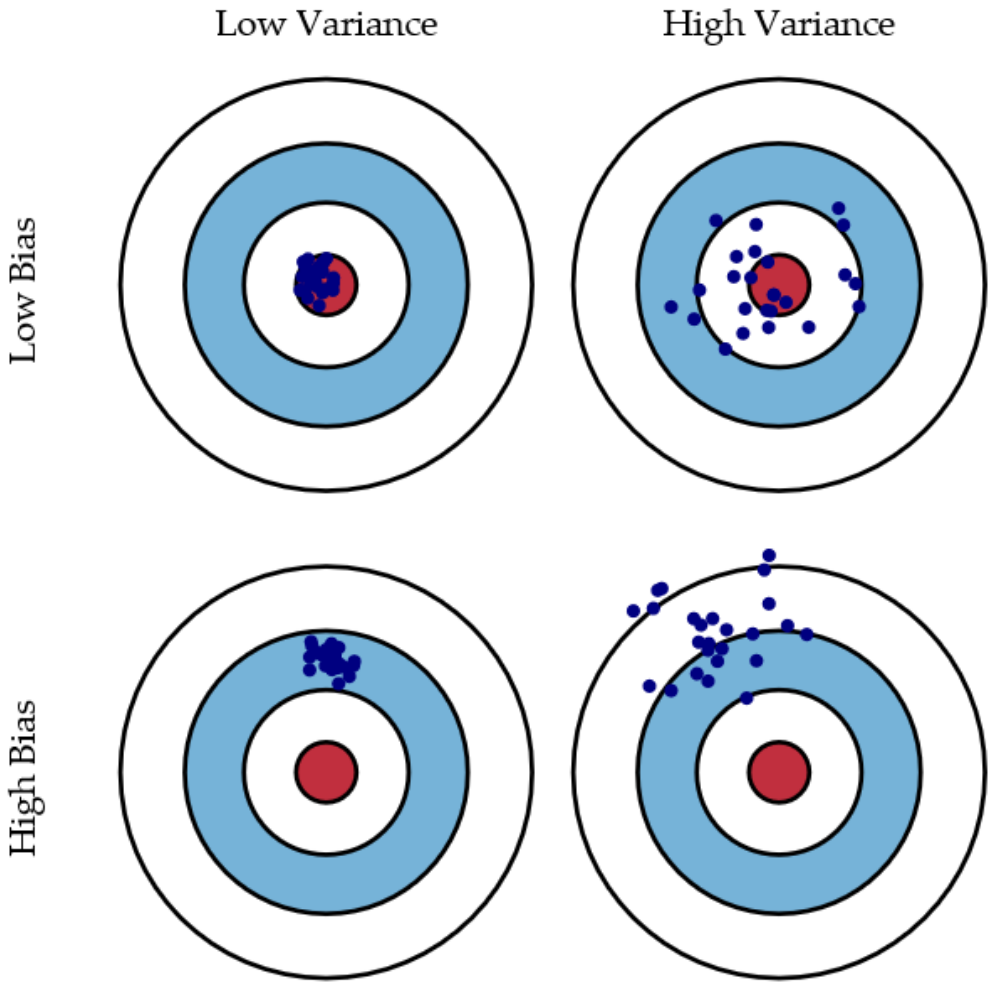
Good fit
Test error is low, and only slightly higher than the training error

Bias-Variance Tradeoff

- **Bias:** difference between what you expect to learn and truth
  - Measures how well you expect to represent true solution
  - Decreases with more complex model

- **Variance:** difference between what you expect to learn and what you learn from a from a particular dataset
  - Measures how sensitive learner is to specific dataset
  - Increases with more complex model

Bias-Variance Tradeoff

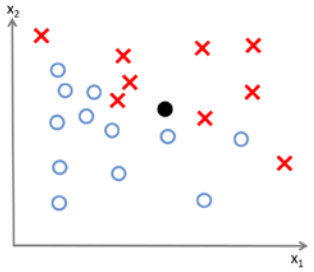


Outline

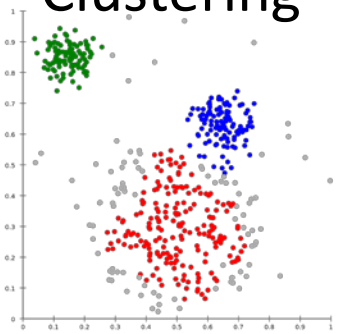
- Introduction
- Applications
- Summary

Data Mining Tasks

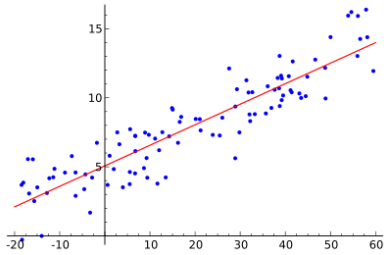
Classification



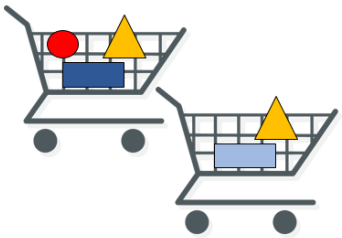
Clustering



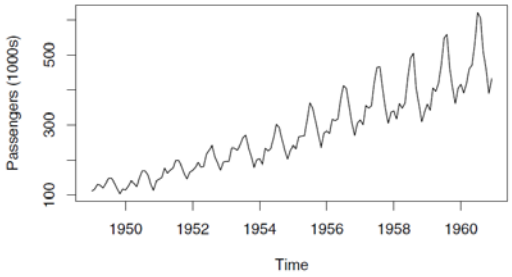
Regression



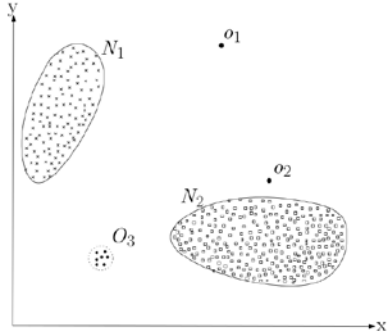
Association Pattern



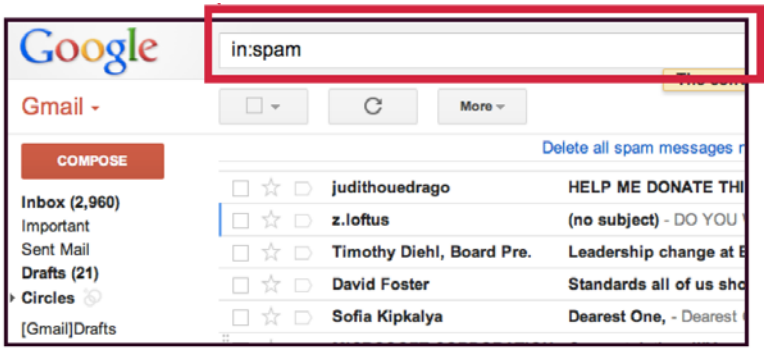
Time Series



Outlier Analysis

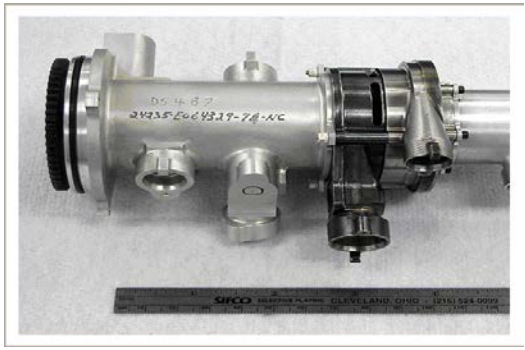


Classification Examples



Examples for classification applications

- Will it rain tomorrow?
- Will the applicant cause liability damage next year?
- Is there a case of insurance fraud?
- Is the customer able to repay the loan?
- Will the device fail soon?
- Is there a traffic jam?



Clustering

- Identification of a finite set of categories, classes or groups (clusters) in the data
- Objects in the same cluster should be as similar as possible.
- Objects from different clusters should be as dissimilar as possible to each other

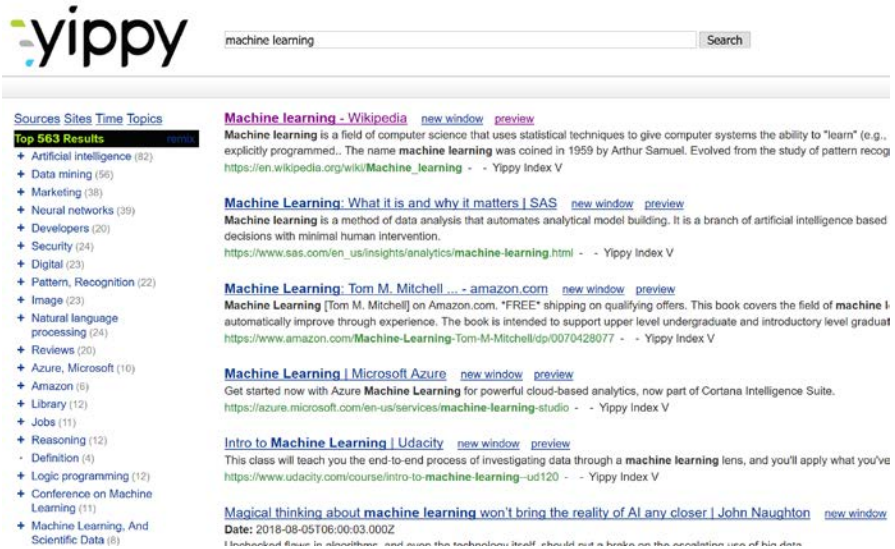
Application examples?

Clustering

Clustering of customers according to their consumption patterns

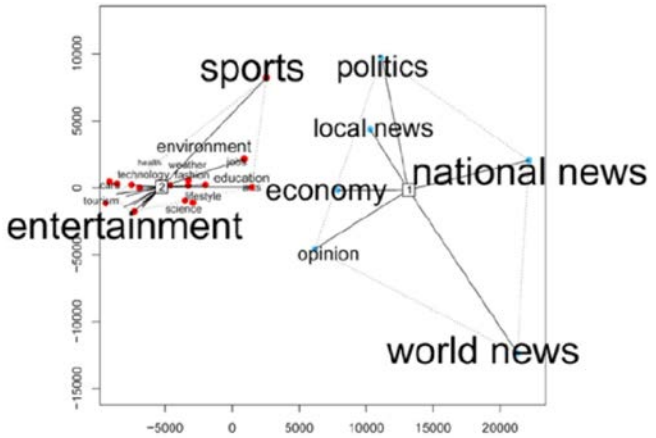
Source: https://mapr.com/blog/apache-spark-machine-learning-tutorial/

Clustering

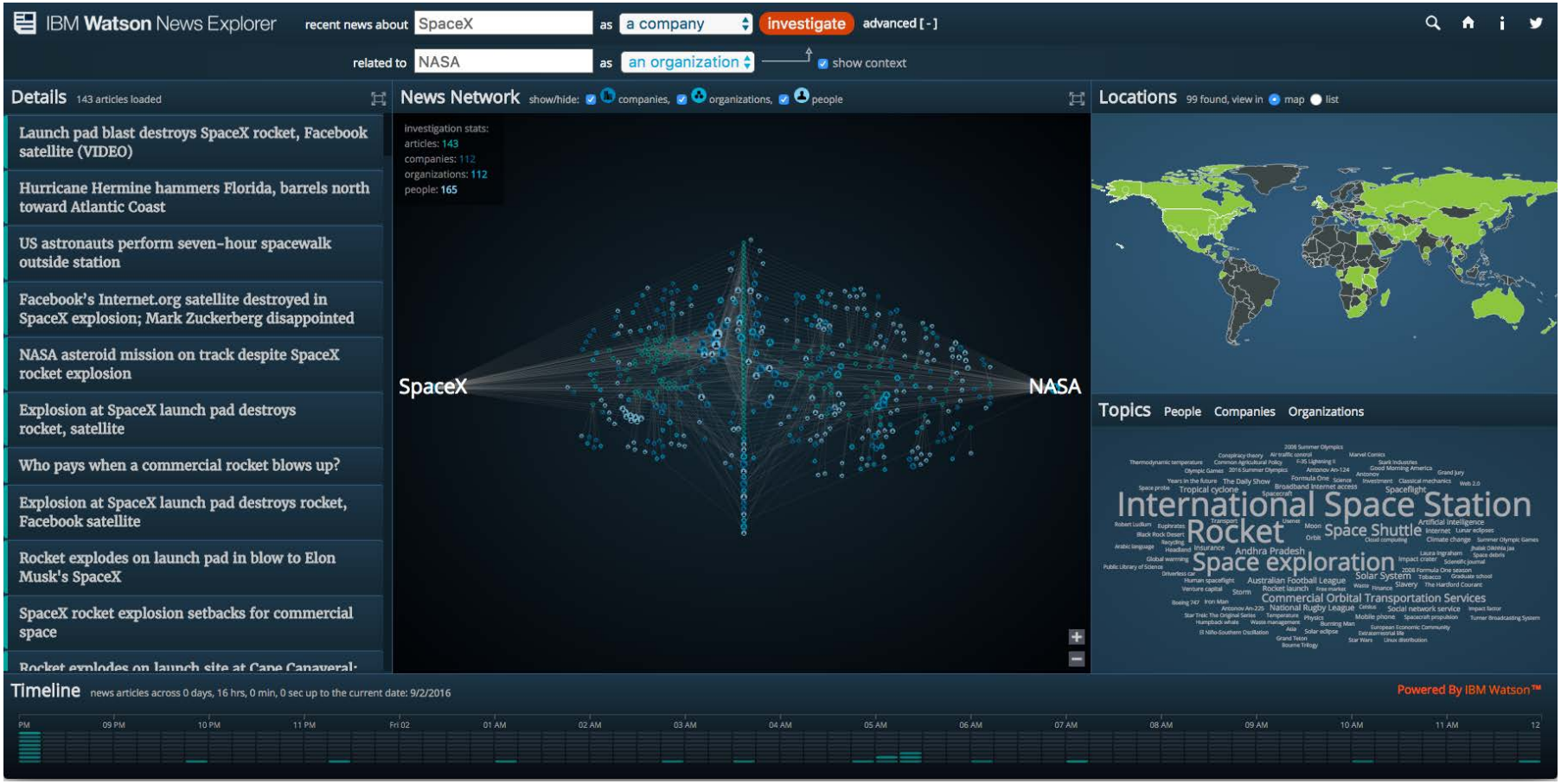


Document clustering according to content similarity

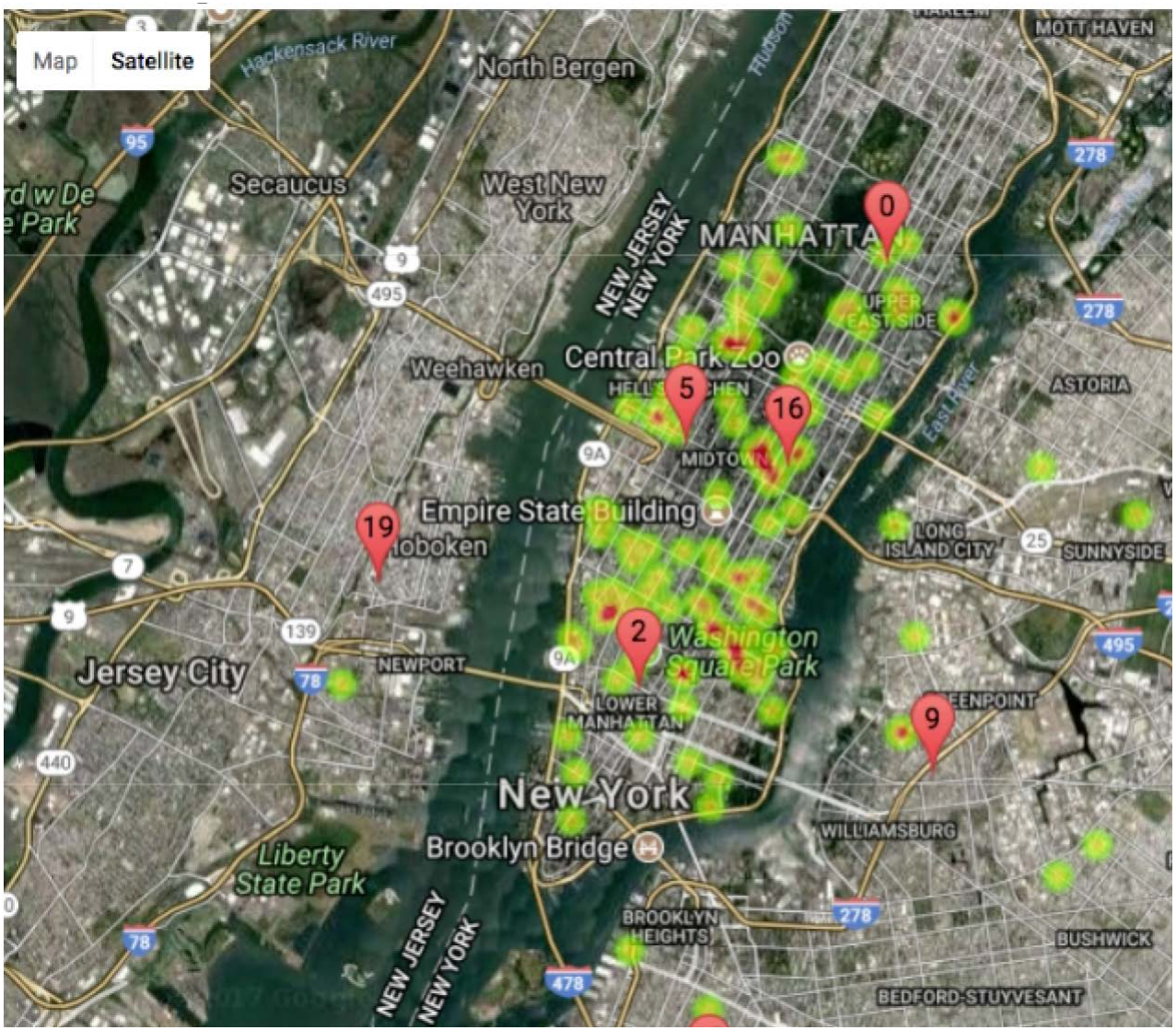
Group articles in different categories



Clustering


IBM Watson News Explorer
recent news about SpaceX as a company investigate advanced [-]
related to NASA as an organization show context
Details 143 articles loaded
Launch pad blast destroys SpaceX rocket, Facebook satellite (VIDEO)
Hurricane Hermine hammers Florida, barrels north toward Atlantic Coast
US astronauts perform seven-hour spacewalk outside station
Facebook's Internet.org satellite destroyed in SpaceX explosion; Mark Zuckerberg disappointed
NASA asteroid mission on track despite SpaceX rocket explosion
Explosion at SpaceX launch pad destroys rocket, satellite
Who pays when a commercial rocket blows up?
Explosion at SpaceX launch pad destroys rocket, Facebook satellite
Rocket explodes on launch pad in blow to Elon Musk's SpaceX
SpaceX rocket explosion setbacks for commercial space
News Network show/hide: companies, organizations, people
investigation stats:
articles: 143
companies: 112
organizations: 112
people: 165
SpaceX
NASA
Locations 99 found, view in map list
Topics People Companies Organizations
International Space Station
Rocket
Space Shuttle
Space exploration
Commercial Orbital Transportation Services
Timeline news articles across 0 days, 16 hrs, 0 min, 0 sec up to the current date: 9/2/2016
Powered By IBM Watson™
PM 09 PM 10 PM 11 PM Fri 02 01 AM 02 AM 03 AM 04 AM 05 AM 06 AM 07 AM 08 AM 09 AM 10 AM 11 AM 12

Uber Trip Clustering



Association analysis and Recommender Systems

Find association rules that predict the occurrence of an item depending on the occurrence of other items or users.

Energy

- Demand Side Management (DSM):

  How would the consumption reduce if certain operational changes are made, such as lowering thermostat settings, ventilation rates or indoor lighting levels?

- Operation and maintenance (O&M):

  How much energy could be saved by retrofits to building shell, changes to air handler operation from constant air volume to variable air volume operation, or due to changes in the various control settings, or due to replacing the old chiller with a new and more energy efficient one?

- Monitoring and verification (M&V):

  If the retrofits are implemented to the system, can one verify that the savings are due to the retrofit, and not to other causes, e.g. the weather or changes in building occupancy?

Energy

- Automated fault detection, diagnosis and evaluation (AFDDE):
  How can one automatically detect faults in heating, ventilating, air-conditioning and refrigerating?
- Optimal operation:
  How can one characterize HVAC&R equipment (such as chillers, boilers, fans, pumps,…) in their installed state and optimize the control and operation of the entire system?

Example of Spatial Data Mining

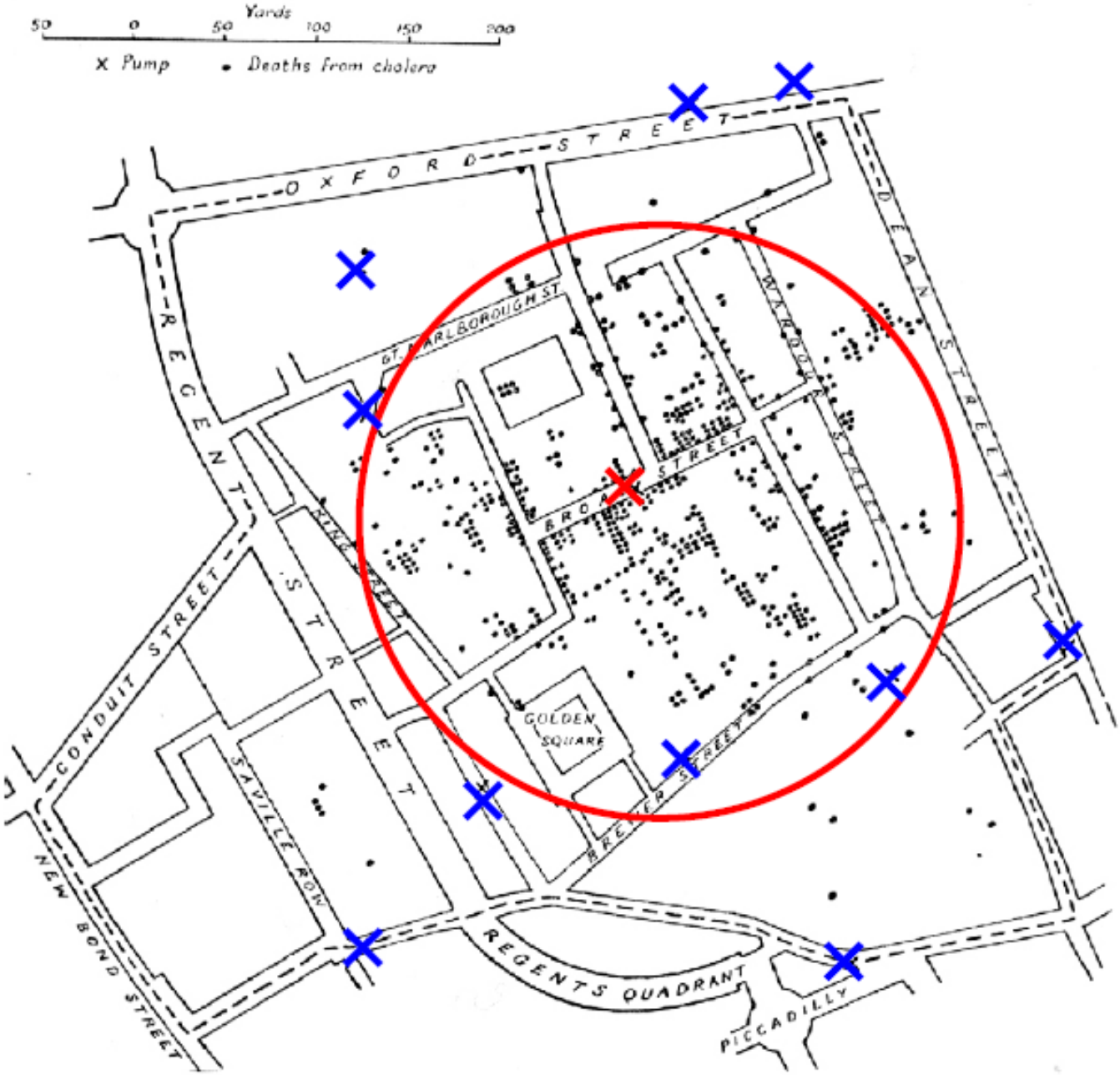
In 1854 a cholera epidemic occurred in London.

John Snow, an English physician, found the cause of this disease using spatial data mining methods.

Myth?


50 0 50 100 150 200 Yards
X Pump
• Deaths from cholera
OXFORD STREET
REGENT STREET
GT. MARLBOROUGH ST.
BROAD STREET
CARNABY STREET
DEAN STREET
KING STREET
CONDUIT STREET
SAVILLE ROW
NEW BOND STREET
GOLDEN SQUARE
BREWER STREET
REGENTS QUADRANT
PICCADILLY


50 0 50 100 150 200 Yards
X Pump • Deaths from cholera
OXFORD STREET
REGENT STREET
CONDUIT STREET
ST. MARLBOROUGH ST
BROAD STREET
GOLDEN SQUARE
SAVILLE ROW
NEW BOND STREET
REGENTS QUADRANT
PICCADILLY
DEAN STREET

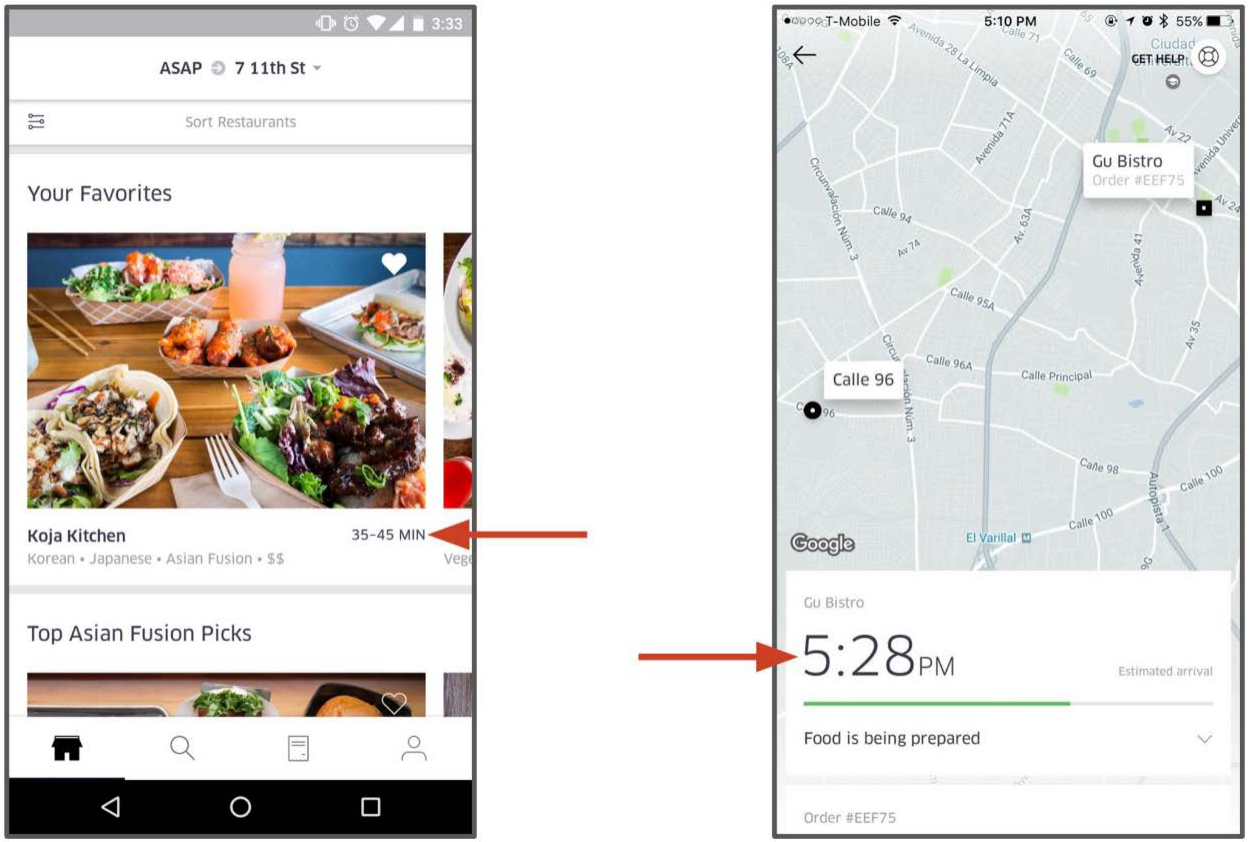

50 0 50 100 150 200 Yards
X Pump
• Deaths from cholera
OXFORD STREET
REGENT STREET
CONDUIT STREET
NEW BOND STREET
SAVILLE ROW
GOLDEN SQUARE
REGENTS QUADRANT
PICCADILLY
DEAN STREET
BROAD STREET

Spatial Data Mining

Drinking water well in the Broad Street, London

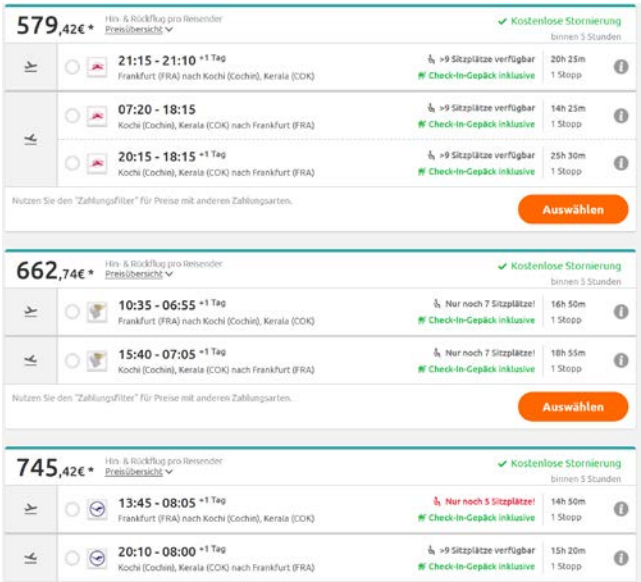
Quelle: http://spatialgeek.com/blog/?p=139

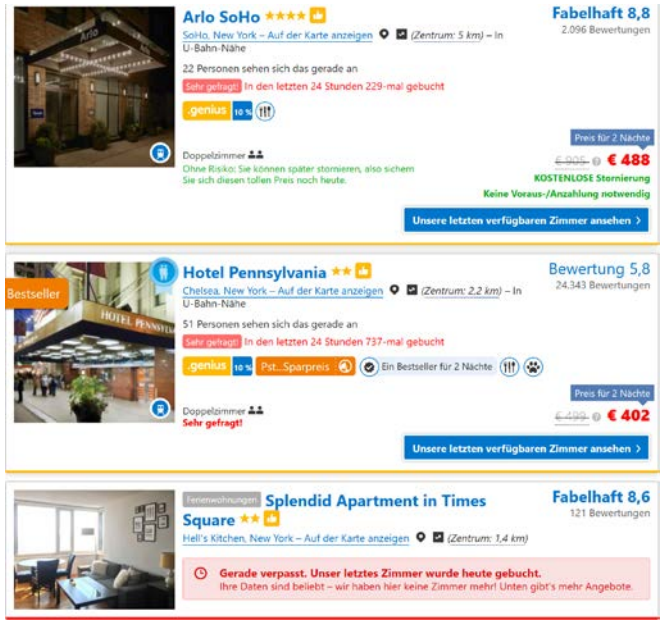
Uber Eats: Meal Delivery Time



Forecast: Numerical Prediction

- How many tourists and business travellers need on xx. xx. xx. a hotel in our city?
- How many people want to be on xx. xx. from A to B?
- How many items are sold tomorrow?
- Best price?





Automatic decision of credit applications

Given:
Questionnaire with information about the person and their financial circumstances

Problem:
Should the loan be granted?
Simple statistical method covers 90% of all cases
But: 50% of all borderline cases lead to credit defaults

Solution (?): reject all borderline cases
No! Borderline cases are among the customers with the highest turnover
→ credit scoring

Credit scoring

- 1000 training examples for borderline cases
- 20 attributes:
  age, length of service with the current employer, Duration of residence at the current address, duration of the client relationship with the bank, other loans granted,...

- Learned control quantity correctly predicts 2/3 of the borderline cases!
- In addition, the company likes the rules because they can be used to explain the credit decision to customers.

Prediction of power consumption

- Electric power stations need forecasts
  - future energy cvonsumption
  - at certain points in time
- Accurate predictions of the minimum and maximum load within each hour result in significant savings.
- Given:
  manually constructed static model, which requires "normal" weather conditions
- Problem: Adaptation to specific weather conditions
- Parameters of the static model:
  Base load for the current year, seasonal load trends, influence of holidays

Improvements with Data Mining

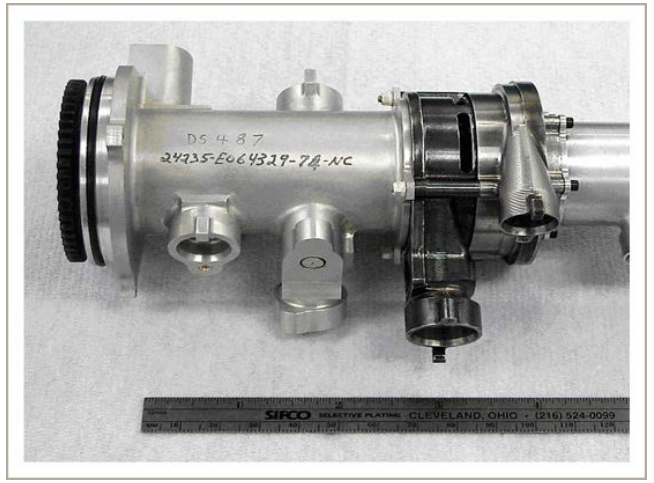
- Improved forecasting by searching for "most similar days"
- Attributes like temperature, humidity, wind speed, cloudiness, and additional difference between actual and predicted load
- The mean difference of the three most similar days is added to the static model.
- Coefficients of the linear regression function represent attribute weights in the similarity function.

Predictive Maintenance

classic field of application for expert systems

Given:

Fourier analysis of vibrations at different points of the case



Task:

Preventive maintenance of electromechanical motors and generators.

The data is very noisy.

Previous: Diagnosis by expert/manually created rules.

With data mining better prediction could be created.

Marketing and sales I

Companies collect large amounts of sales and marketing data

Possible applications:

- Customer loyalty: Identification of potential customers who "jump off" soon by recognising changes in their behaviour (e. g. banks, telephone companies). Churn Management.
- Special offers: Identifying profitable customers (e. g. reliable customers of credit card companies who need a higher holiday time limit).

Marketing and sales II

Shopping Cart Analysis

- Association techniques to find groups of goods that are often bought together
- Analysis of purchasing patterns in the past
- Identification of good customers
- Focusing of advertising mail (advertising campaigns are cheaper than mass mailings)

Career opportunities

At present, graduates who are proficient in data analysis are in high demand in many industries!

Examples of companies and institutions:

- Psiori, Freiburg
- BlueYonder, Karlsruhe
- RapidMiner, Dortmund, Boston, mehr als 40 Mitarbeiter, wachsend
- Fraunhofer, St. Augustin, Intelligente Analyse- und Informationssysteme, mehr als 200 Wissenschaftler
- Amazon, Berlin, Machine Learning HQ mit mehr als 50 Mitarbeiter
- Zalando, Berlin, sucht Data Scientists
- Banken, Versicherungen, Hedgefonds, und viele, viele mehr

Outline

- Introduction
- Applications
- Summary

Summary

Data Mining techniques automatically learn the relationship between a set of descriptive features and a target feature from a set of historical data.

Challenges of Data Mining algorithms

1. generalize,
2. underfitting and overfitting,
3. bias and variance,
4. productive.

Striking the right balance between model complexity and simplicity (between underfitting and overfitting) is the hardest part.