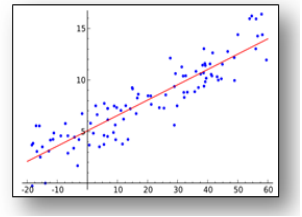
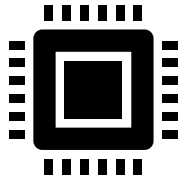
RAAK Top-up AIoTValley

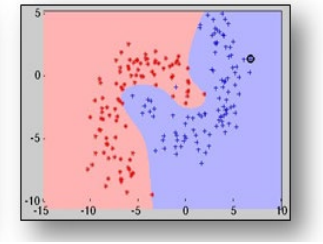
Artificial Intelligence Training session 1
Methodology & data mining

Jeroen Linssen

2021-05-21







SAXION UNIVERSITY OF APPLIED SCIENCES

TVALLEY

BOOST smart industry

Who am I?

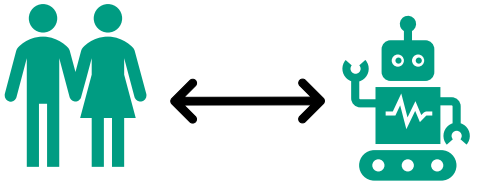
Associate Lector Ambient Intelligence @Saxion

Jeroen Linssen

Artificial Intelligence

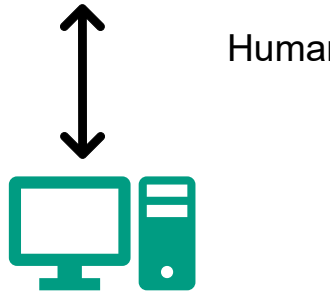
Predictive Maintenance

Health monitoring



Human-Robot Interaction

Teaching



Human-Media Interaction

SAXION UNIVERSITY OF APPLIED SCIENCES

Modus operandi

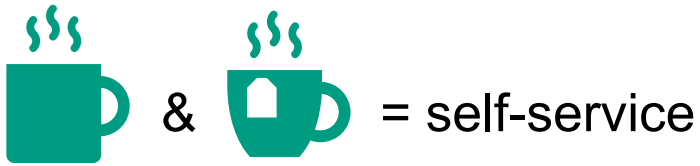
Webinar, but interactive

Ask questions in chat

Assistance by Linda Maalderink

Max. 3 hours with 2 breaks

Session is being recorded



Introduction square

Motivation

RAAK projects + BOOST

Involved partners

Learning goals

RAAK-mkb Focus op Vision & Data in Smart Industry

Focus op Vision (2019 – 2021)

- Enabling companies in ‘maakindustrie’ to use computer vision
- Disseminating knowledge on (AI for) vision

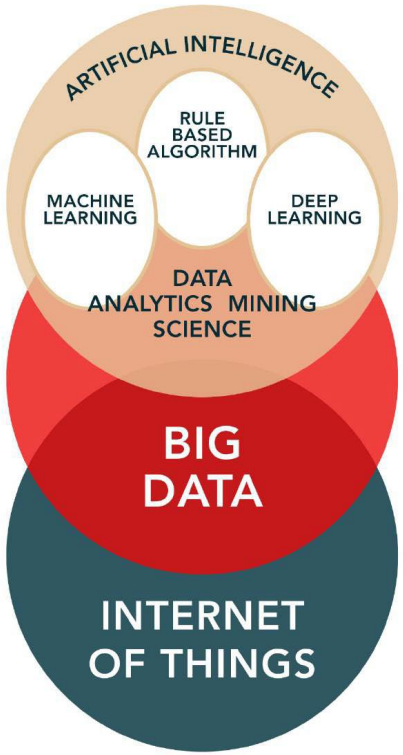
Data in Smart Industry (2017 – 2020)

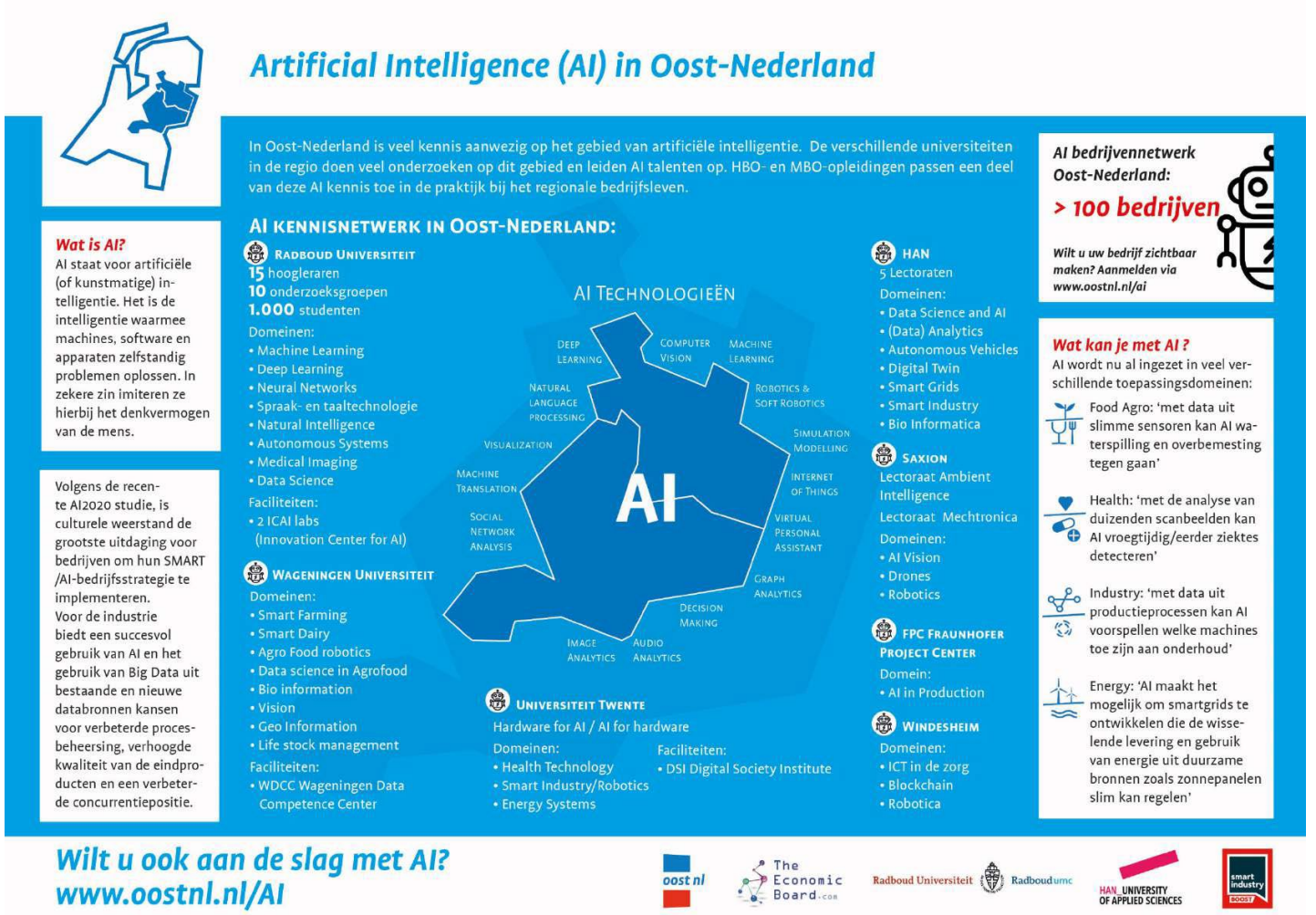
- Enabling companies in ‘maakindustrie’ to use data for process optimization
- Disseminating knowledge on IoT and AI for data acquisition and analysis

AIoTValley (2020 – 2021)

- ‘Top-up’ of DSI to create open access educational material
- Including AI and IoT in TValley (robotics & mechatronics fieldlab)

Bai: ‘BOOST artificial intelligence’





SAXION UNIVERSITY OF APPLIED SCIENCES

boostsmartindustry.nl

TValley

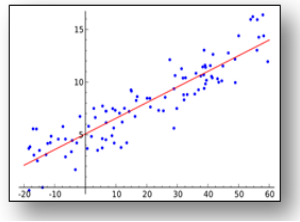
tvalley.nl

SAXION UNIVERSITY OF APPLIED SCIENCES

Ambient Intelligence – ‘Enabling IT for a smart world’

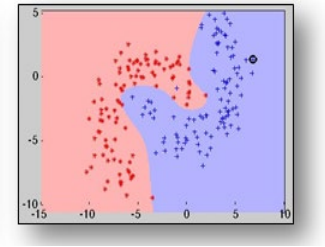
- Research group at Saxion
- 22 members, 11 nationalities
- Applied research to adapt innovative technology to real-world usage
- Technical industry & healthcare

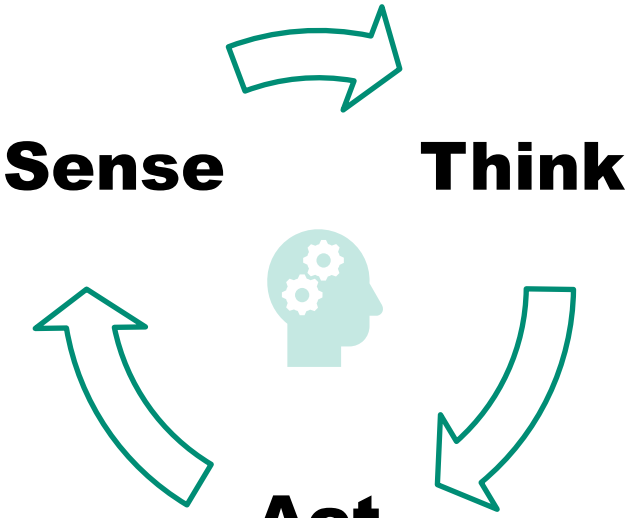
saxion.nl/ami



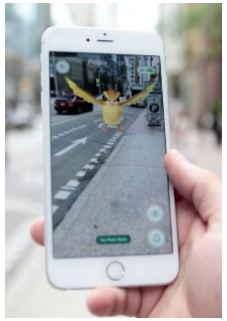
Connected Embedded Systems

Applied Data Science





Augmented Interaction



SAXION UNIVERSITY OF APPLIED SCIENCES

Overview

Training structure

AI, broadly speaking

CRISP-DM, a methodology for data mining

Data mining techniques

Recap & outlook

Training structure

Three sessions
1. Methodology & data mining
2. Machine learning
3. Deep learning

Overview, but in-depth

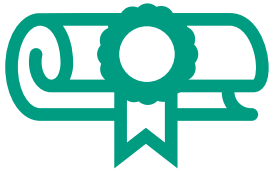
Some hands-on practice

Learning goals

Methodology & data mining
- Employing CRISP-DM
- Data mining: exploration and visualization

Machine learning
- Fundamentals of learning algorithms
- Training and tweaking models
- Evaluating models

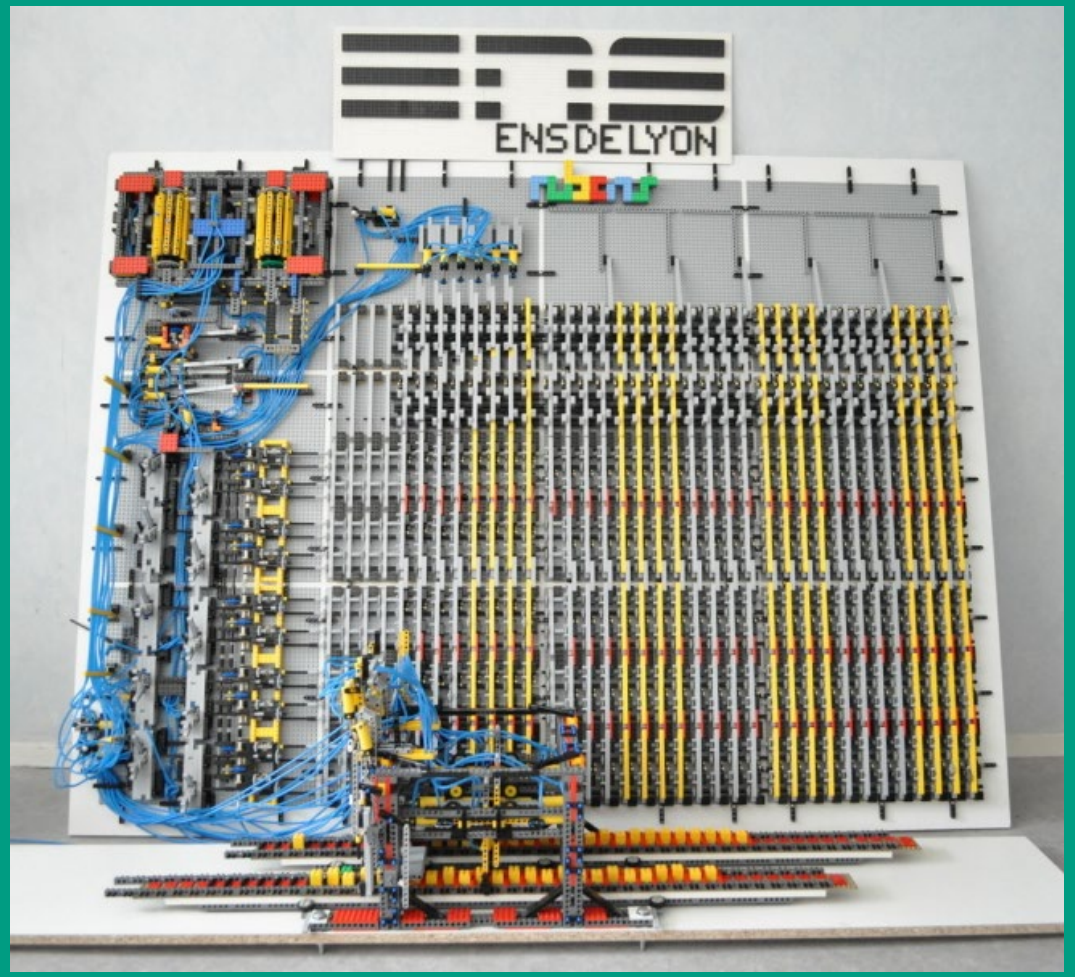
Deep learning
- Fundamentals of neural networks
- Training and tweaking models
- Evaluating models



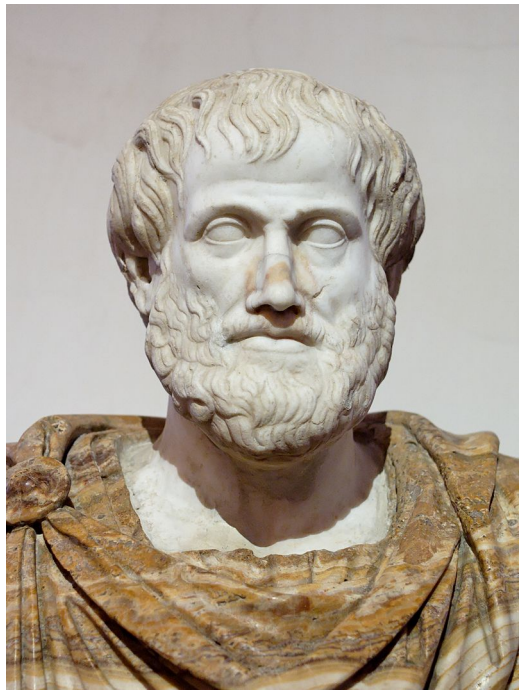
AI, broadly speaking

A short history of AI

Data science, AI, machine learning, deep learning, etc.



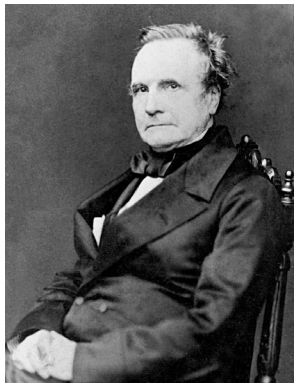
A short history of AI



Aristoteles

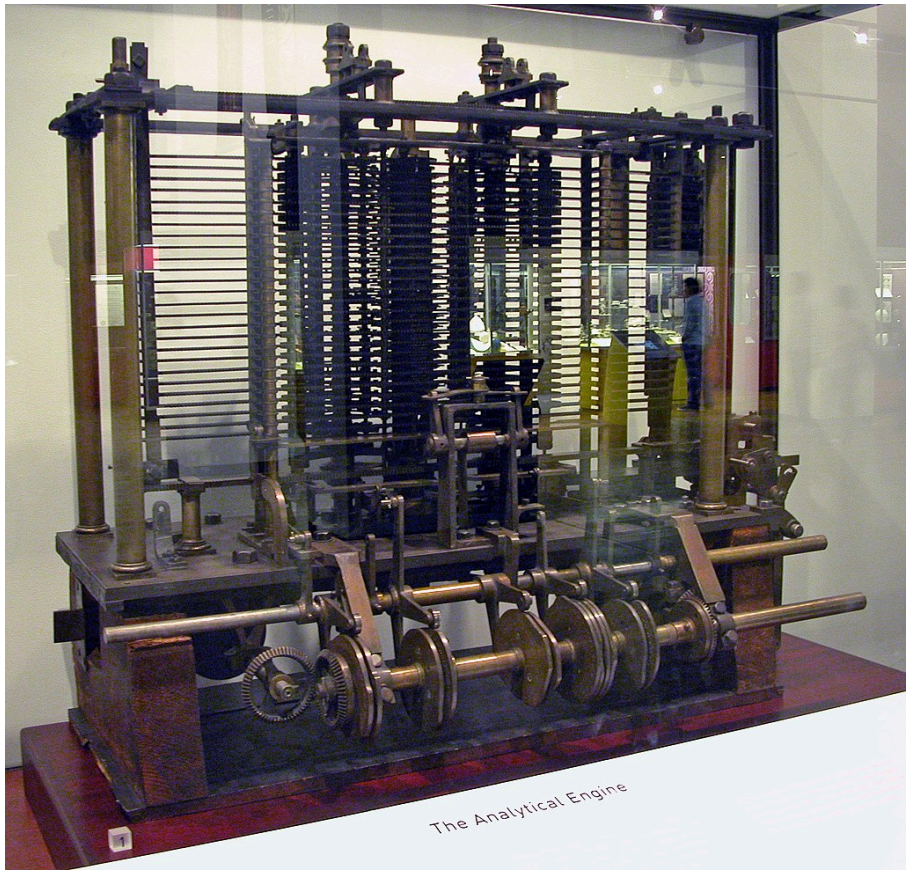
Syllogism

All men are mortal.
Socrates is a man.
Therefore, Socrates is mortal.

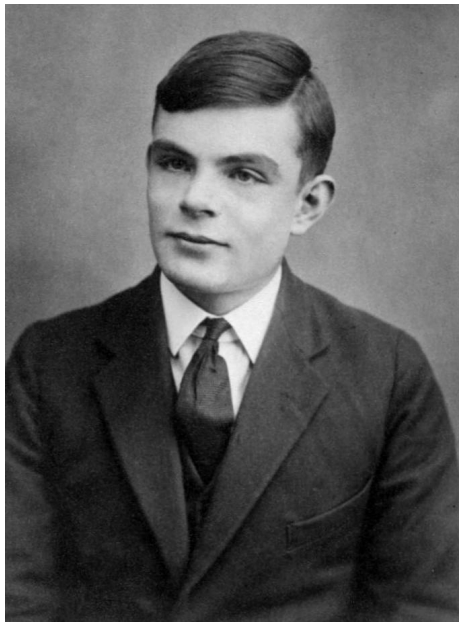


Babbage

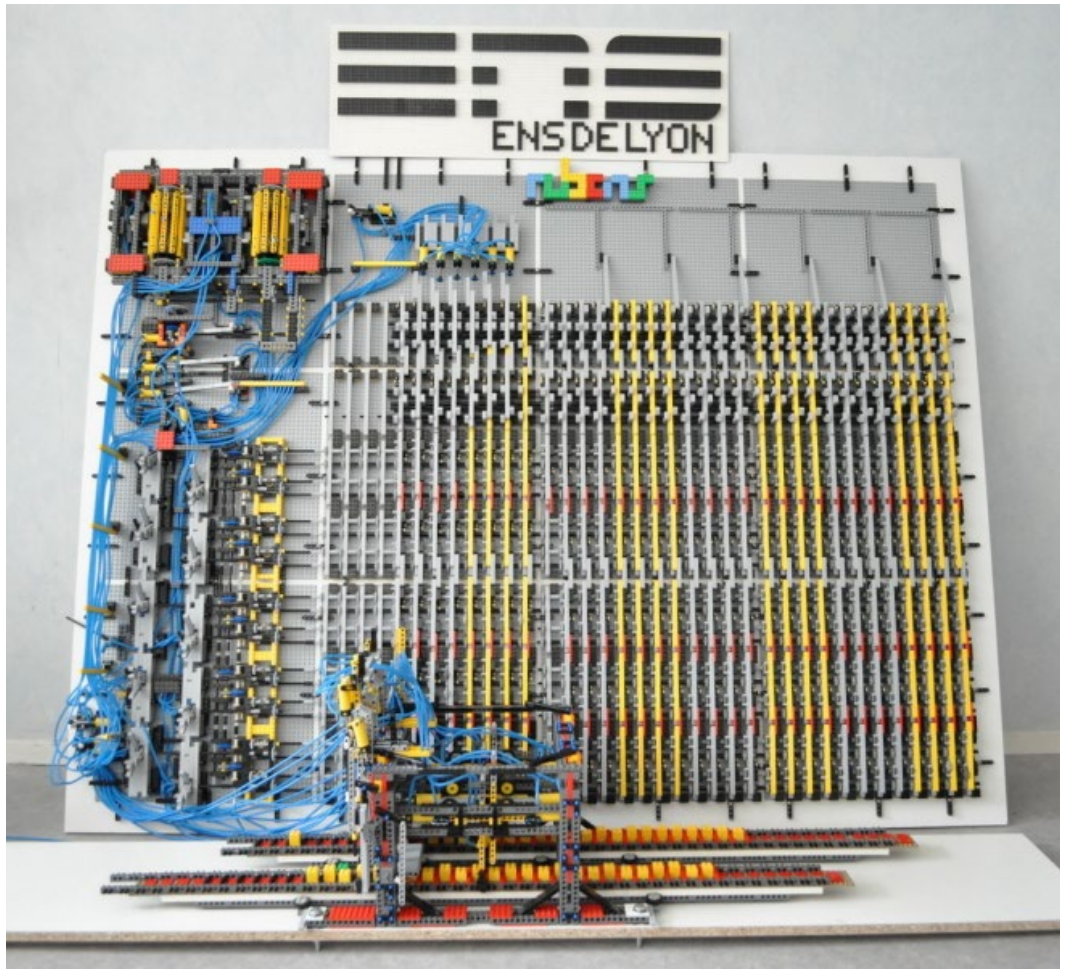
Lovelace

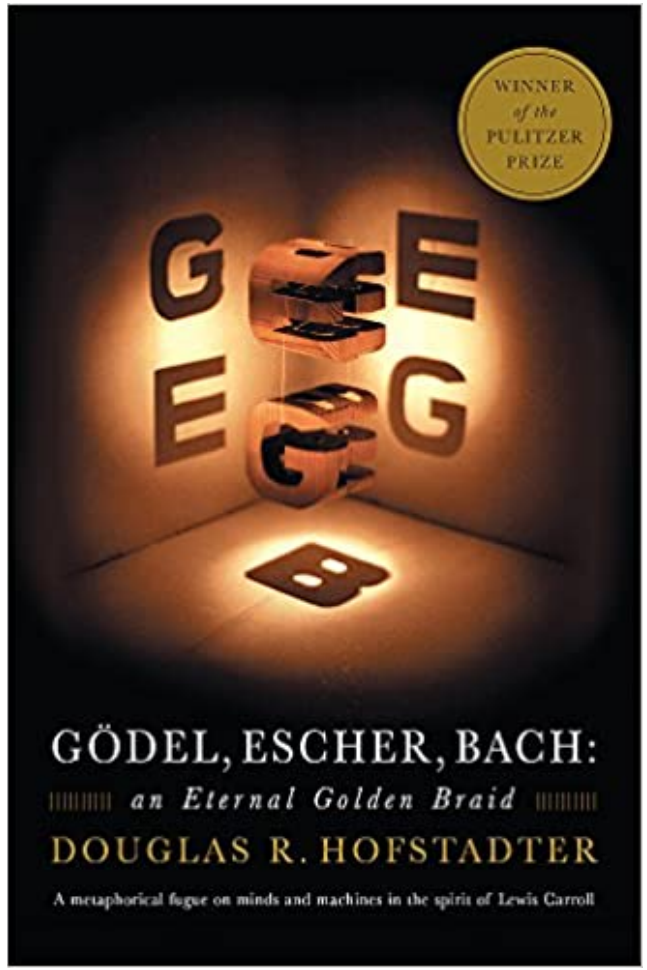


A short history of AI

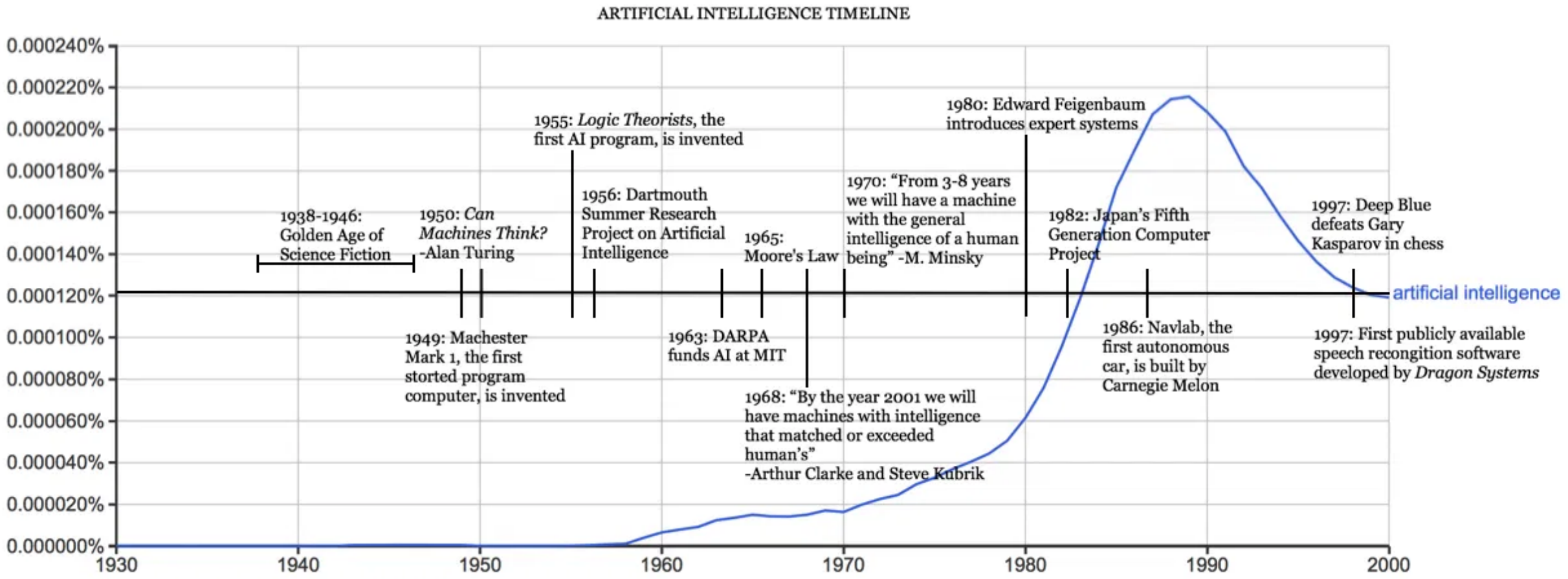


Turing





A short history of AI



ARTIFICIAL INTELLIGENCE
≠
MACHINE LEARNING

ARTIFICIAL INTELLIGENCE

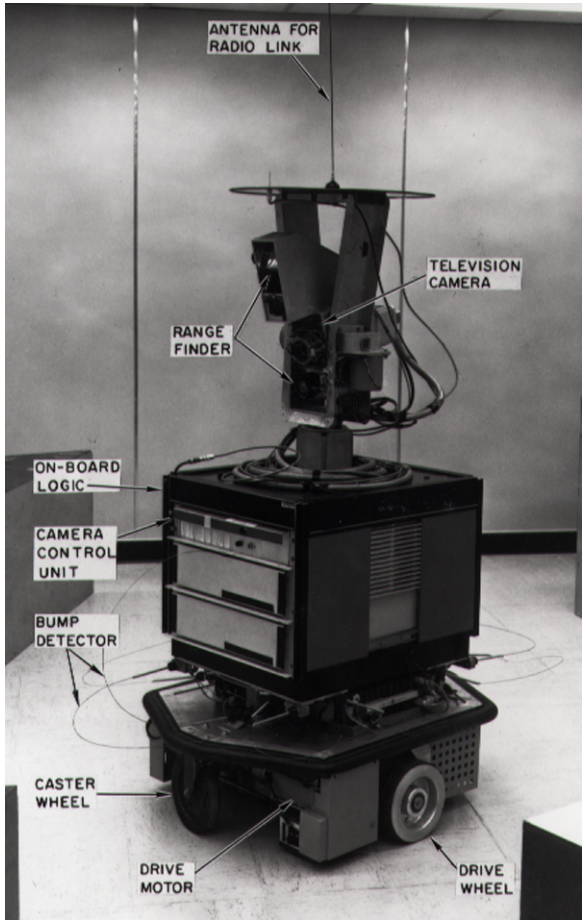
=

(AUTOMATED) PATTERN RECOGNITION + PLANNING FOR ACTION

AI in use

Spam filters: keywords, Bayesian classification

Route planning: A*



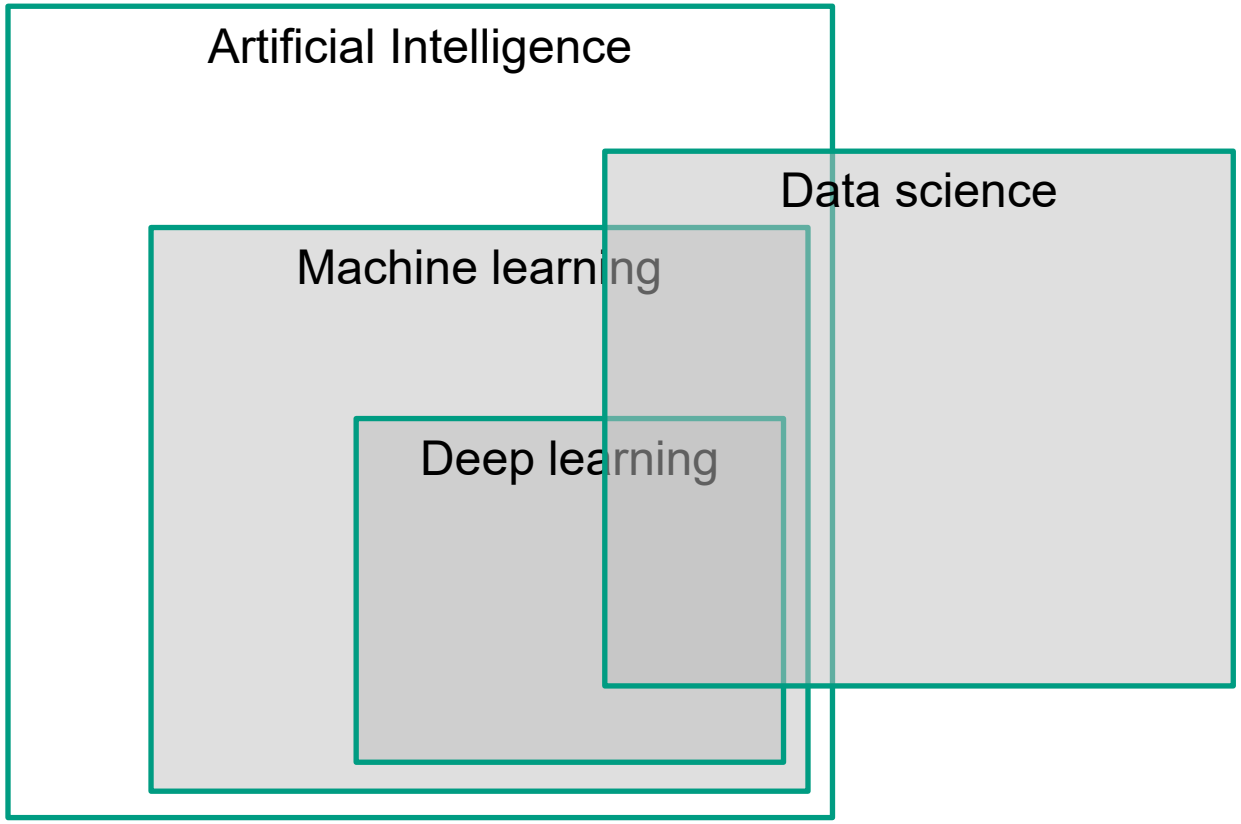
Shakey, the robot

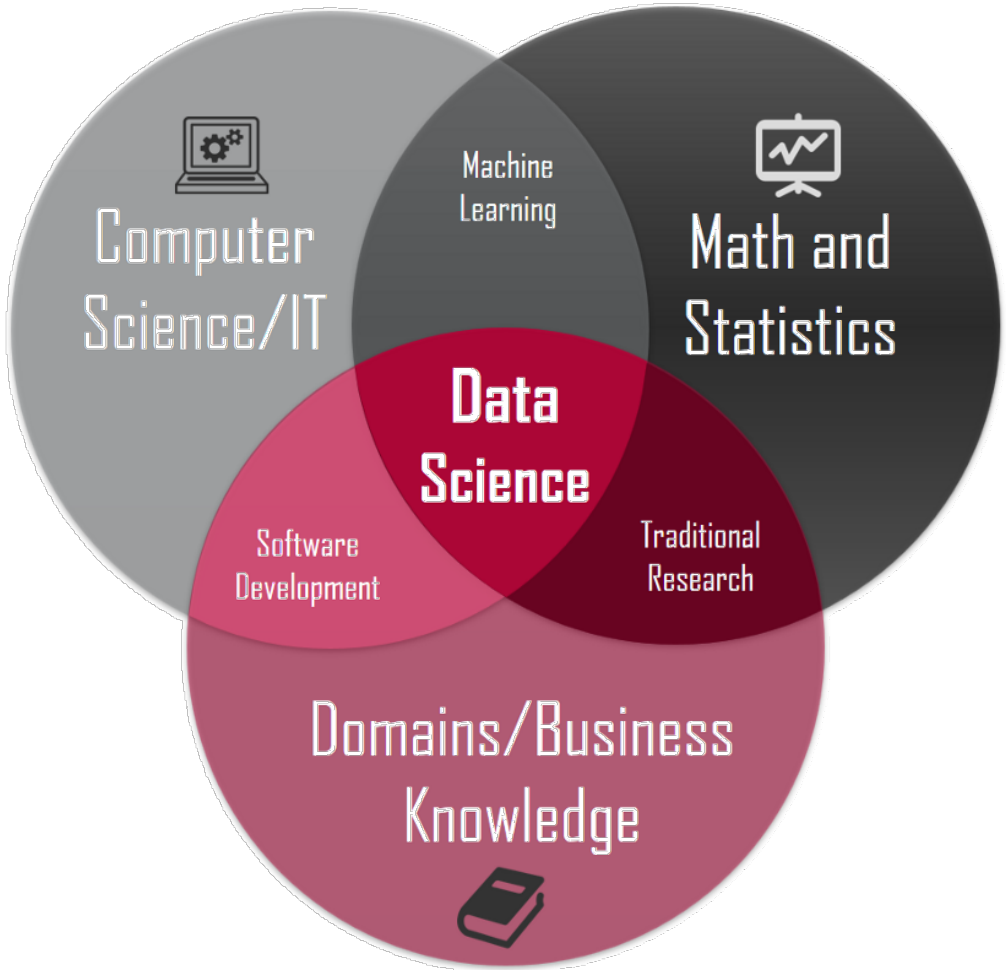
AI in use

Virtual characters

Human-robot interaction

Data science, AI, machine learning





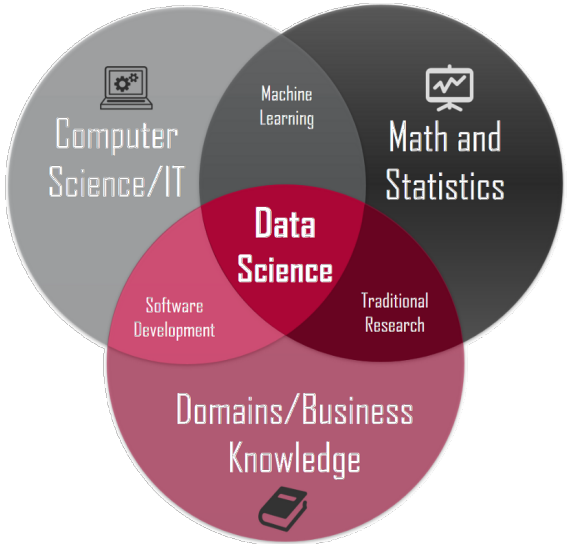
https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Data science

Data science is an interdisciplinary field aiming to turn data into real value.

Data may be:
- Structured or unstructured
- Big or small
- Static or streaming

Value may be provided in the form of:
- Predictions
- Automated decisions
- Models learned from data
- Any type of data visualization delivering insights



Data science includes:
- Data extraction
- Data preparation
- Data exploration
- Data transformation
- Storage and retrieval
- Computing infrastructures
- Various types of mining and learning
- Presentation of explanations and predictions
- Exploitation of results taking into account ethical, social, legal, and business aspects

Van Der Aalst, W. (2016). Data science in action. In *Process mining* (pp. 3-23). Springer, Berlin, Heidelberg.

AI AD

Data science

Embedded systems

Machine learning

Smart industry

Companies

Education

Innovation

Artificial Intelligence

Healthcare

Morals & ethics

Virtual & augmented reality

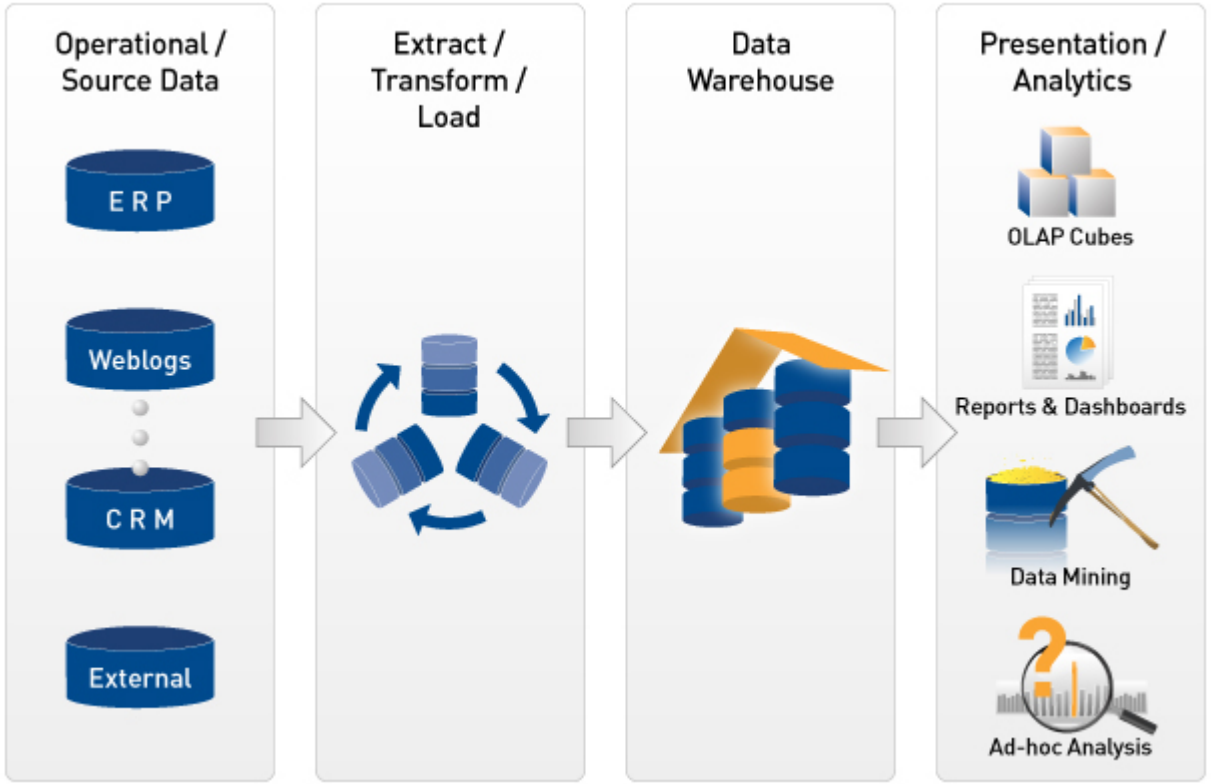
SAXION HOGESCHOOL

Things we will not cover in these sessions

Internet of Things (IoT)

Databases, data warehousing:
Extract, Transform, Load (ETL)

Data governance, privacy, ethics

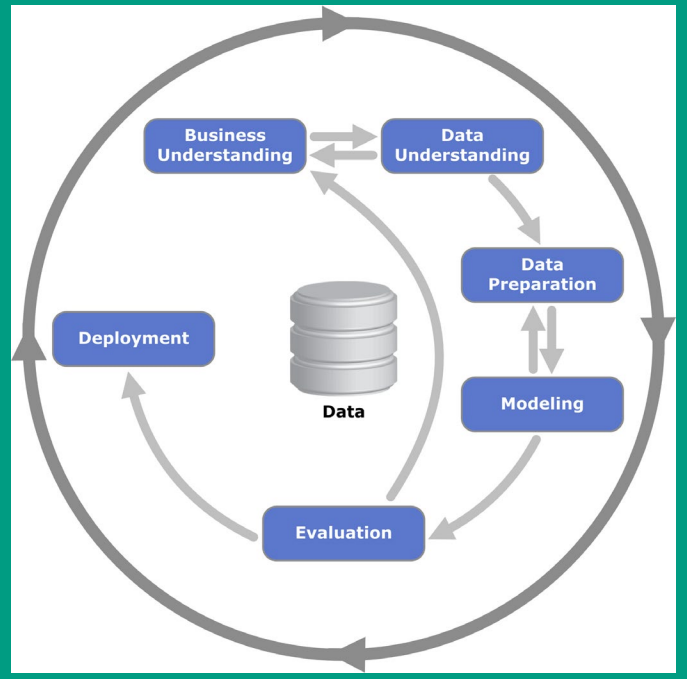


CRISP-DM, a methodology for data mining

Methodology

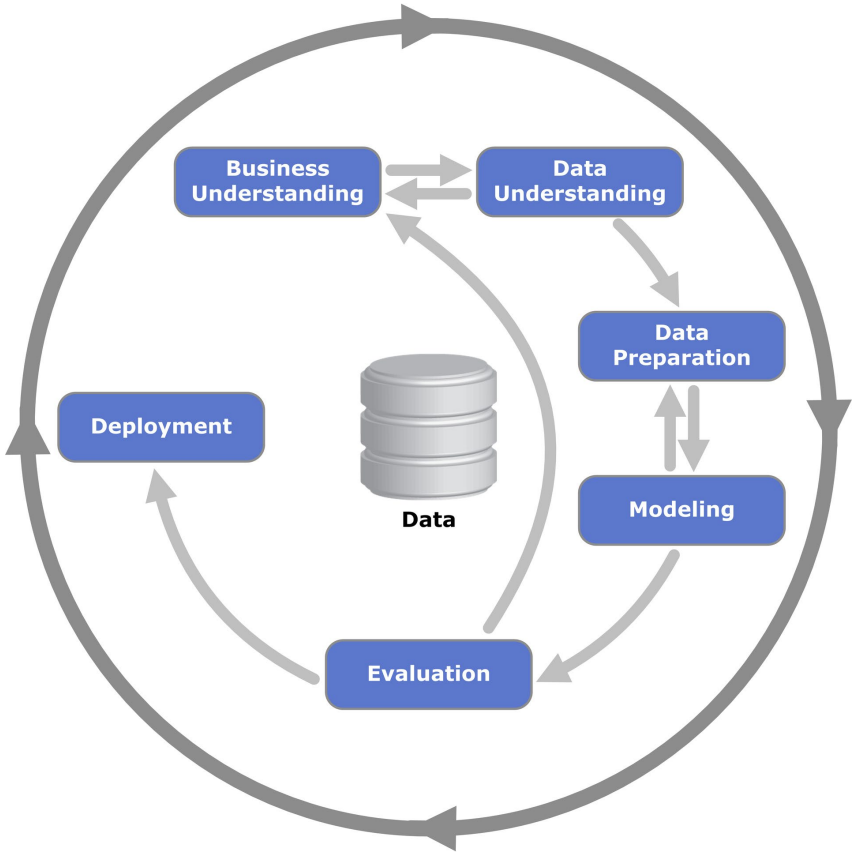
Example use case

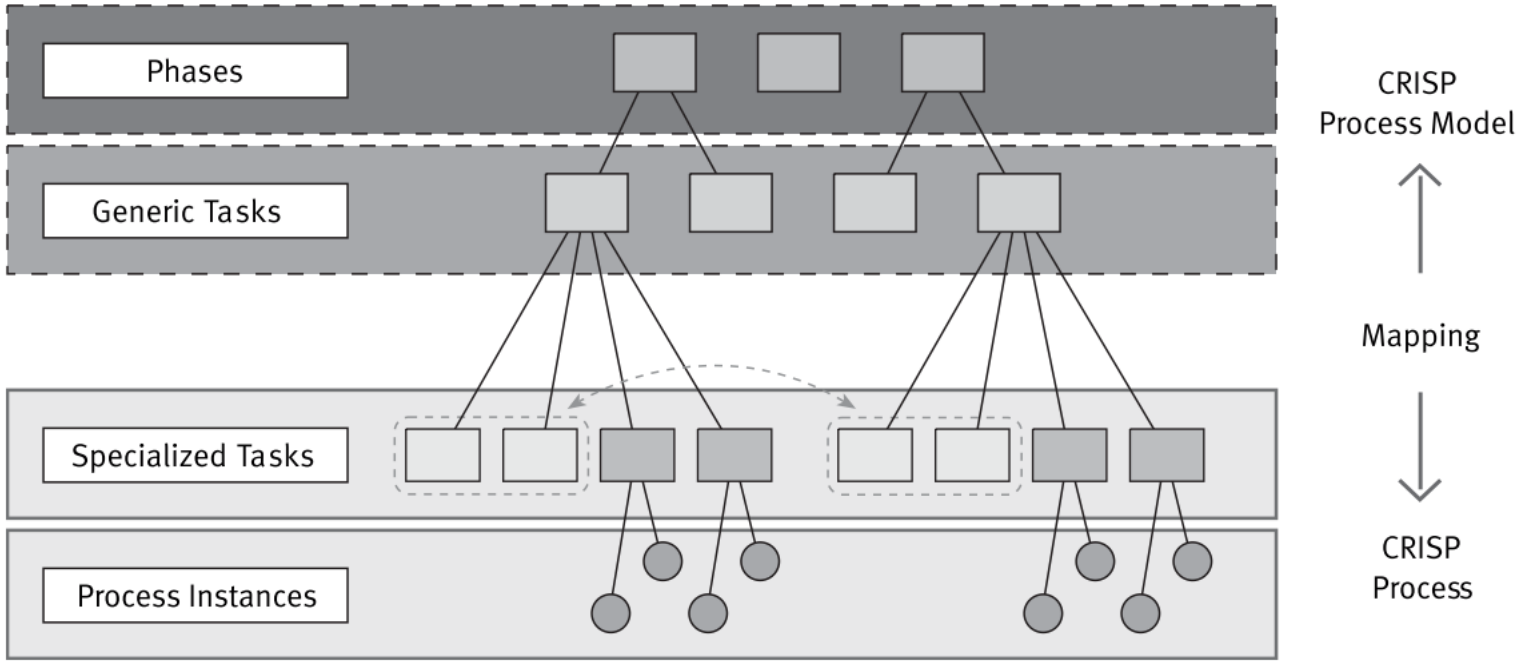
Hands-on practice



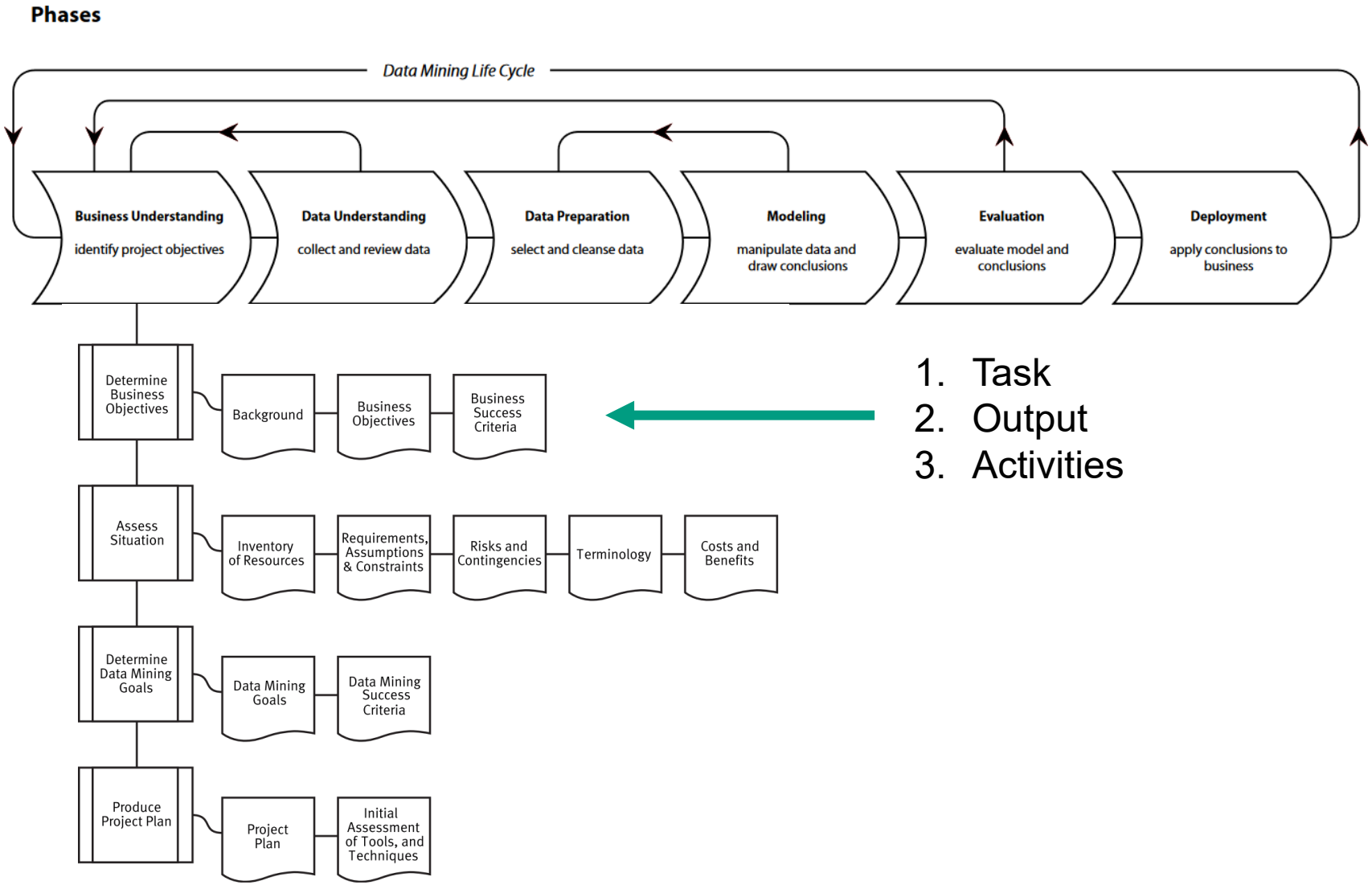
CRISP-DM

CRoss-Industry Standard Process for Data Mining





The CRISP-DM cycle



1. Task
2. Output
3. Activities

CRISP-DM Phase 1

Business objectives
For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor.

Assess situation
Resources, availability of data, requirements/constraints.

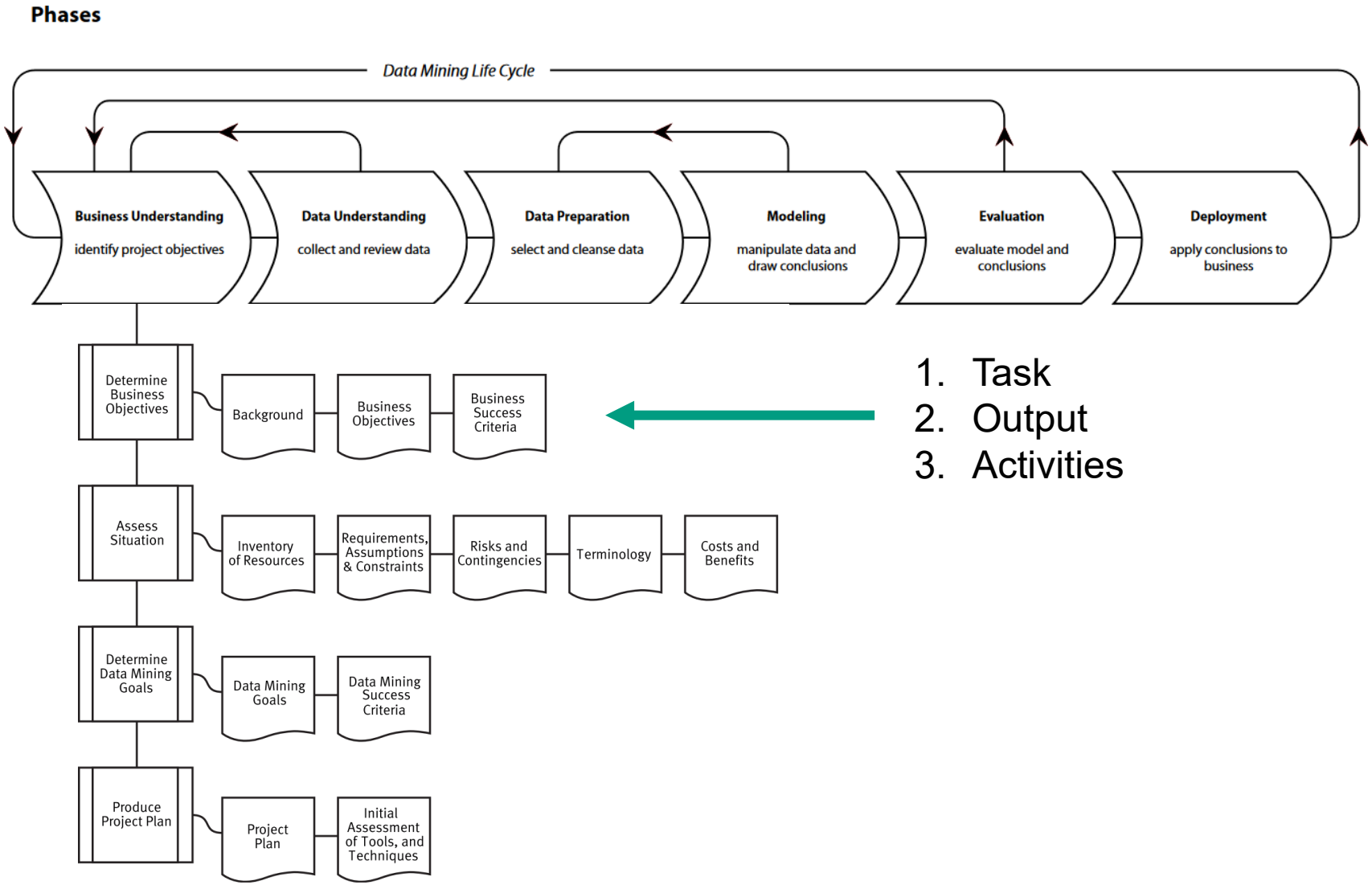
Data mining goals
For example, the business goal might be “Increase catalog sales to existing customers.” A data mining goal might be “Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city, etc.), and the price of the item.”

Project plan
List the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. It is also important to analyze dependencies between time schedule and risks.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0*. IEEE.

INTERMISSION 1

The CRISP-DM cycle



1. Task
2. Output
3. Activities

Example use case: carriers at Scania

Scania Production Zwolle

- 2,000 employees
- 180 trucks per day, 60 % of Europe
- Lean, on-demand production

Business goals:

- Reducing downtime
- Possible causes of unplanned (corrective) maintenance on carriers
- Move towards predictive maintenance

Example use case: carriers at Scania

Assess situation

- Resources
  - Personnel: data/IT experts, domain experts
  - Data warehouse
  - Data: registration of notifications and alarms
- Requirements/constraints
  - Documentation
  - Sensor placement
  - Source alignment possible
  - Available working hours
- Risks
  - Access to sources
  - Data sources unrelatable

Example use case: carriers at Scania

Data mining goals

- Finding correlations in notifications
- Finding possible causations for a specific alarm for the carriers

Project plan

- Plan according to CRISP-DM
- Iterations incorporated
- Risks and contingencies incorporated
- Implicit agile development with this type of plan

Let’s do business!

Walkthrough of CRISP-DM Phase 1

hypersay

Go to live.hypersay.com/HYDEHE (link in chat)

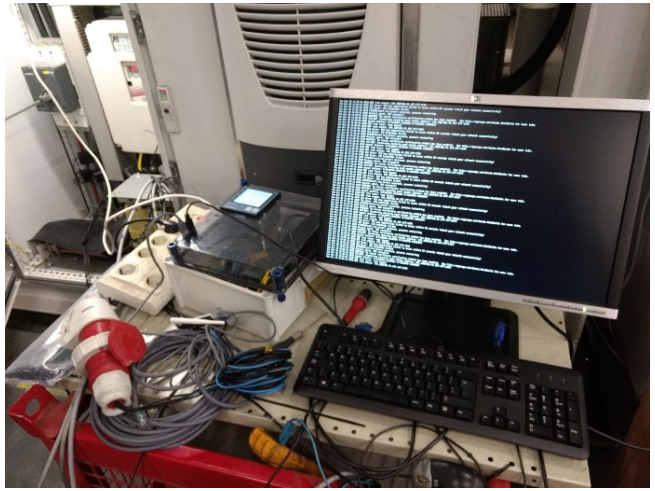
From plan to data: data acquisition

Data availability

Formats

Sensor data

- Selection
- Monitoring installation
- Frequency
- Acquisition
- Storage
- Internet of Things (IoT)



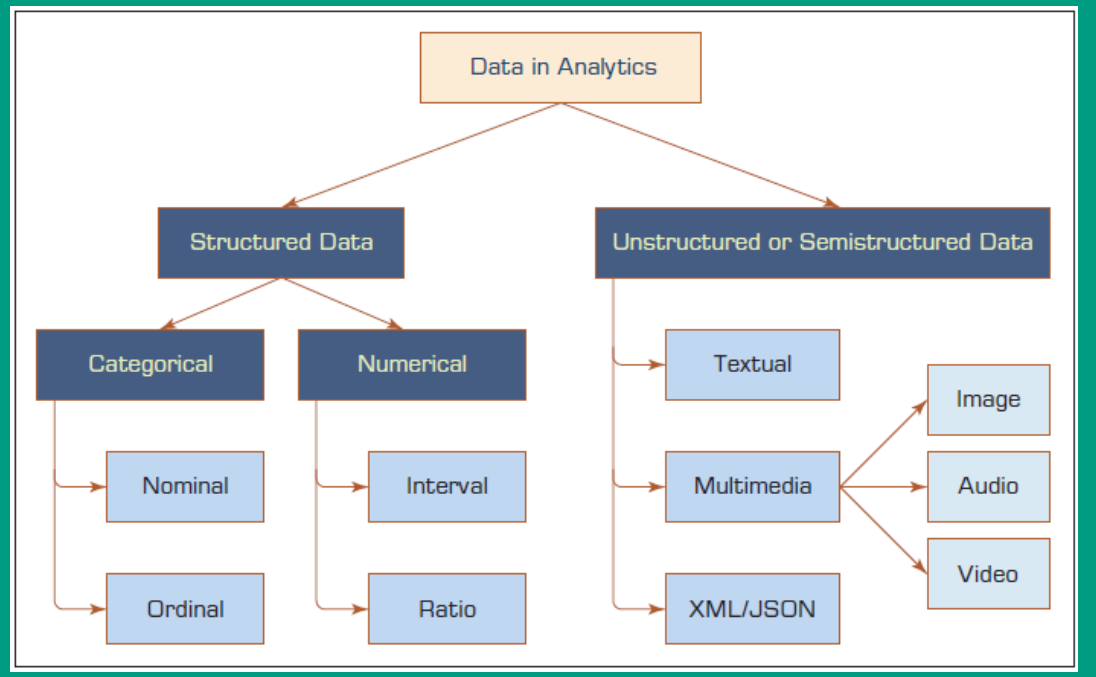
INTERMISSION 2

Data mining techniques

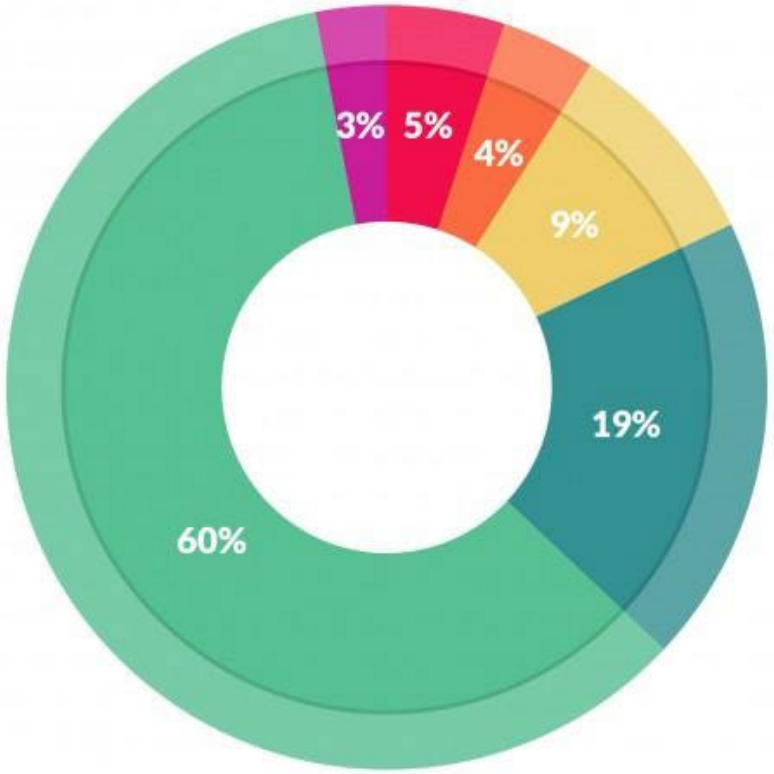
Data quality

Data pre-processing

Data exploration



What does a data scientist do?



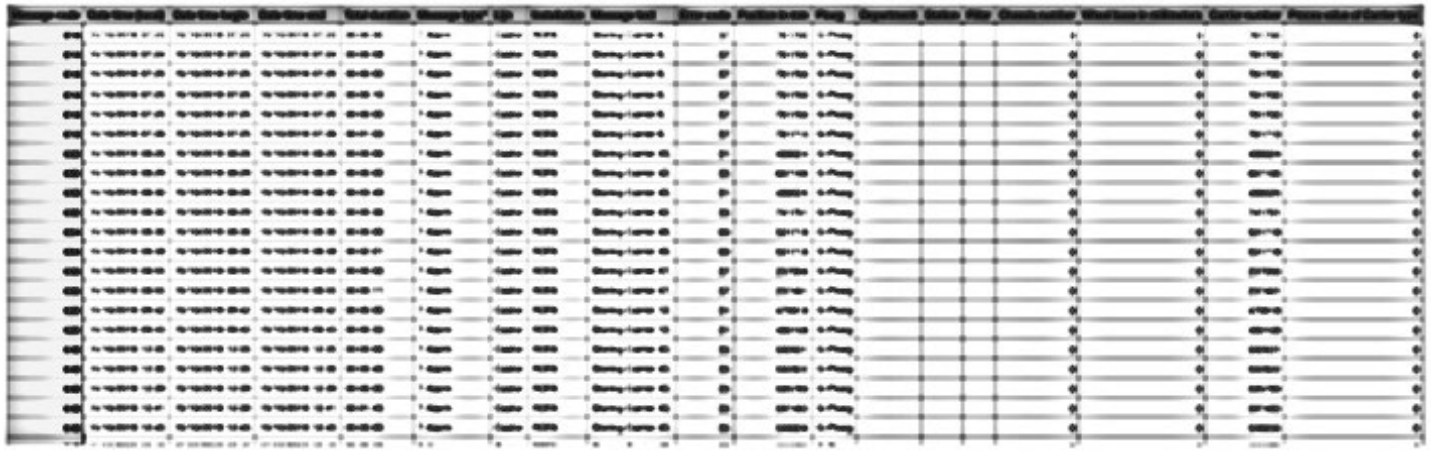
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Example use case: carriers at Scania

CRISP-DM Phase 2: data understanding

- Collect
- Describe
- Explore
- Verify quality



Attribuut	Type	Waarden	Beschrijving
Message code	Int		
Date time (local)	Date/time		
Date time begin	Date/time		
Date time end	Date/time		
Total duration	Time		
Message type	String		
Lijn	String		
Installation	String		
Message text	String		
Error code	Int		
Position in mm	Int		
Ploeg	String		
Department	String		
Station	String		
Pillar	String		
Chassis number	Int		
Wheel base in millimeters	Int		
Carrier number	Int		
Process value of Carrier type	Int		

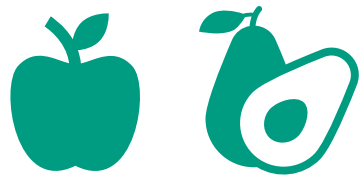
Types of variables

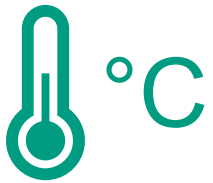
Nominal
- Labels
- Categories

Ordinal
- Order is important
- Differences between values unknown

Interval
- Ordered
- Numeric scale

Ratio
- Ordered, numeric scale
- Absolute zero





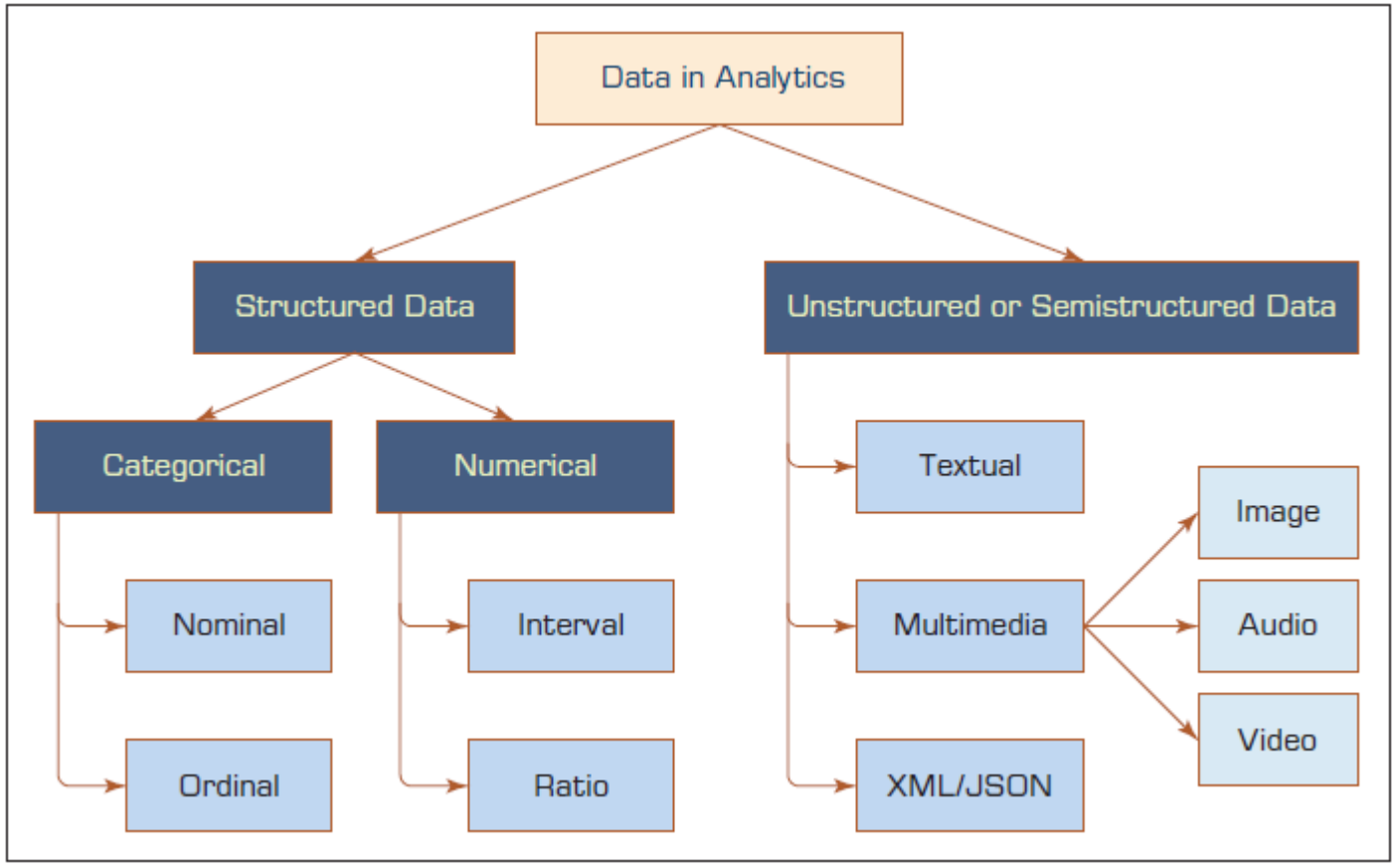
Data types

Data types depend on domain.

Methods of analysis depend on data types.

Data is:

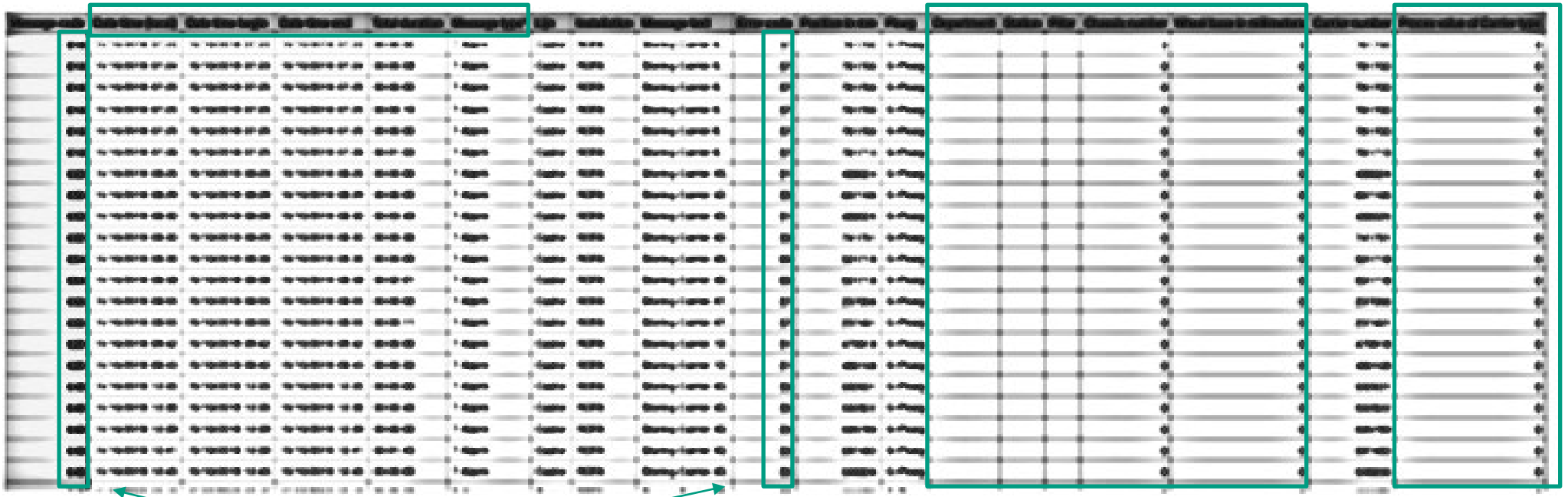
- Messy
- Unaligned
- Complex
- Inaccurate



Example use case: carriers at Scania

Meaning of variables?

Data quality issues



Redundant?

Back to business understanding phase to refine data mining goals.

Example use case: carriers at Scania

Data preparation

- Discard redundant information
- Impute missing values
- Transform values
  - Simpler labels
  - Normalize, standardize

Pandas Profiling Report on Carrier Errors

Overview Variables Correlations Missing values Sample

Overview

Dataset info

Number of variables	19
Number of observations	5889
Missing cells	17667 (15.8%)
Duplicate rows	59 (1.0%)
Total size in memory	3.9 MiB
Average record size in memory	691.9 B

Variables types

CAT	9
NUM	6
UNSUPPORTED	3
BOOL	1

Toggle Reproduction Information
Toggle Warnings

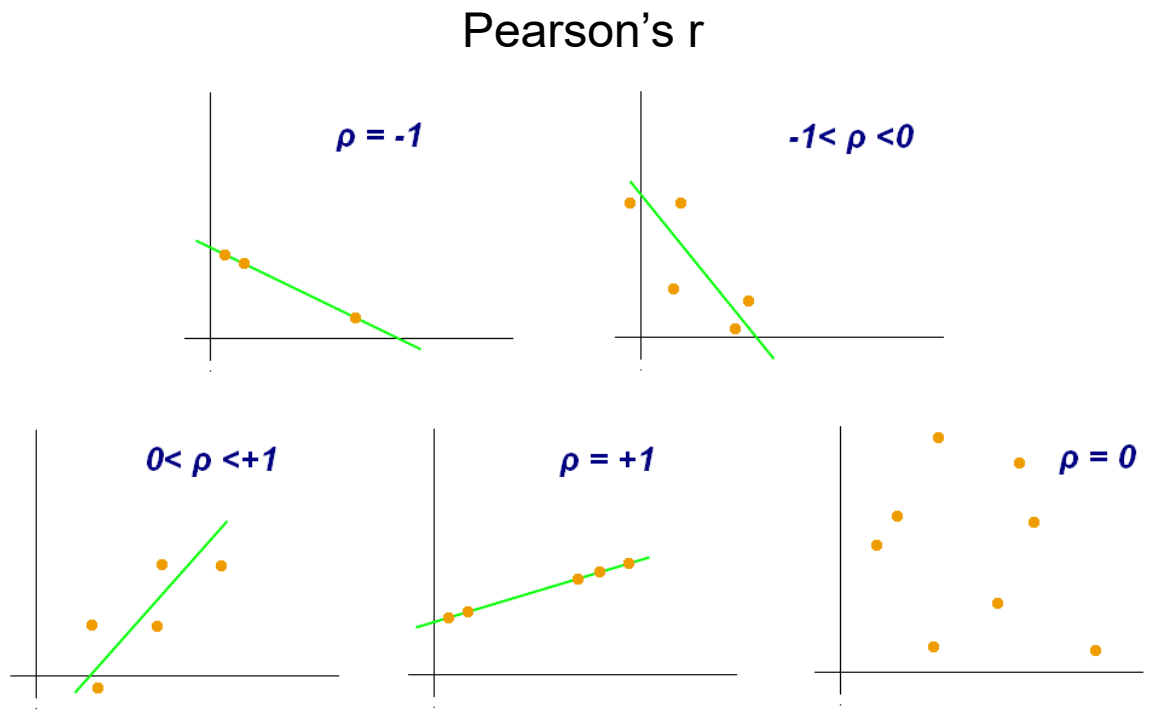
Warnings

Dataset has 59 (1.0%) duplicate rows	Warning
has 392 (6.7%) zeros	Zeros
has 4441 (75.4%) zeros	Zeros
has a high cardinality: 4355 distinct values	Warning
has a high cardinality: 2742 distinct values	Warning
has a high cardinality: 4355 distinct values	Warning
has 5889 (100.0%) missing values	Missing
is an unsupported type, check if it needs cleaning or further analysis	Warning
has constant value	Rejected
has constant value	Rejected
has constant value	Rejected
has 5889 (100.0%) missing values	Missing
is an unsupported type, check if it needs cleaning or further analysis	Warning
has 392 (6.7%) zeros	Zeros
has 4753 (80.7%) zeros	Zeros
has 5889 (100.0%) missing values	Missing

SAXION UNIVERSITY OF APPLIED SCIENCES

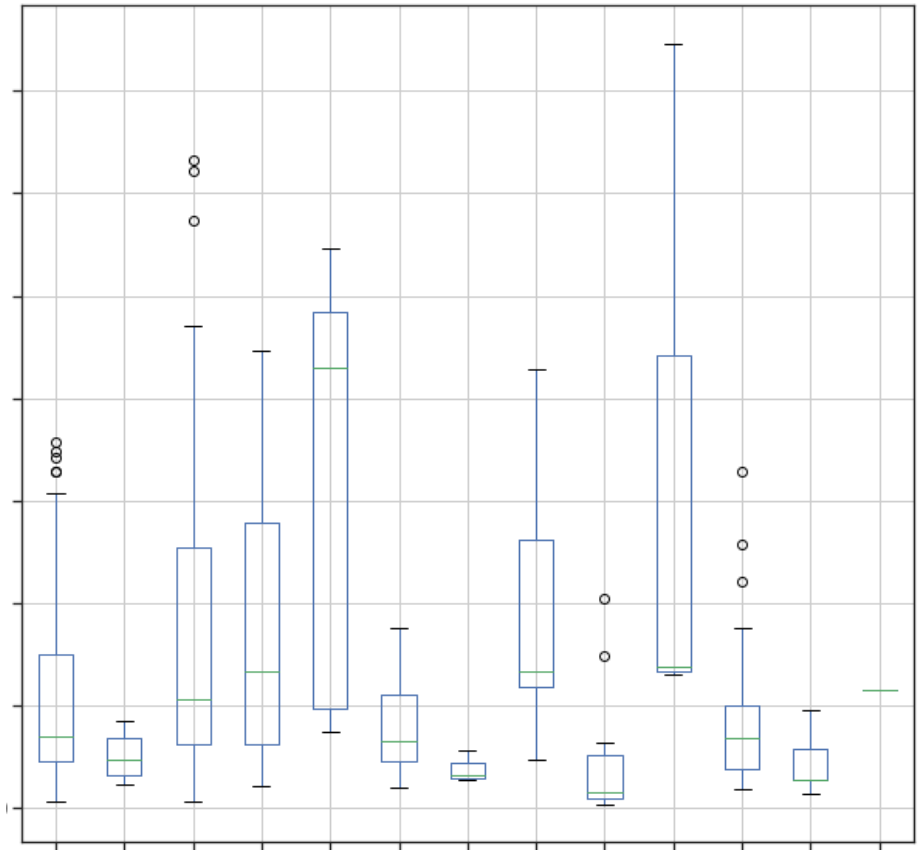
Example use case: carriers at Scania

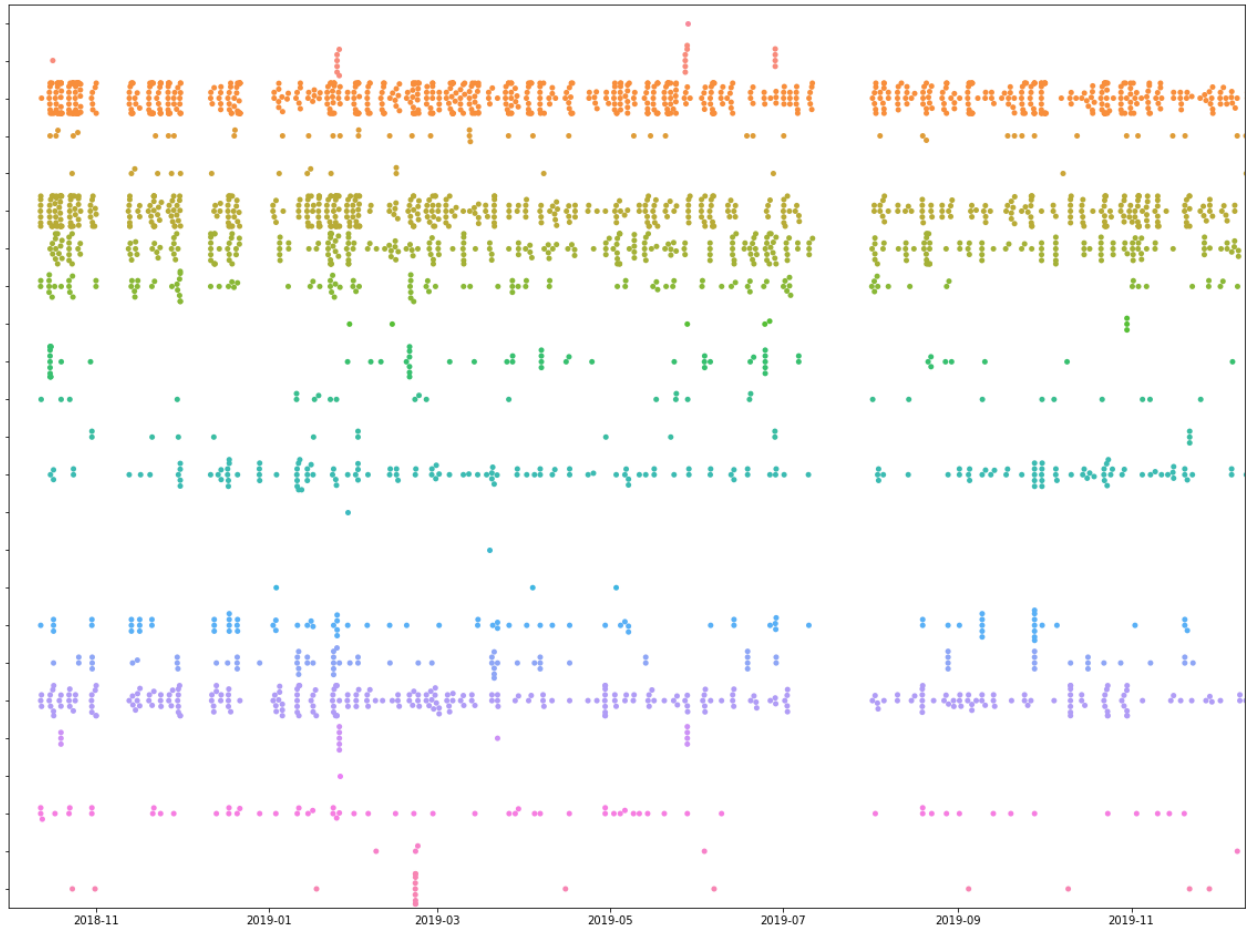
Data exploration ≈ data visualization



Example use case: carriers at Scania

Data exploration ≈ data visualization

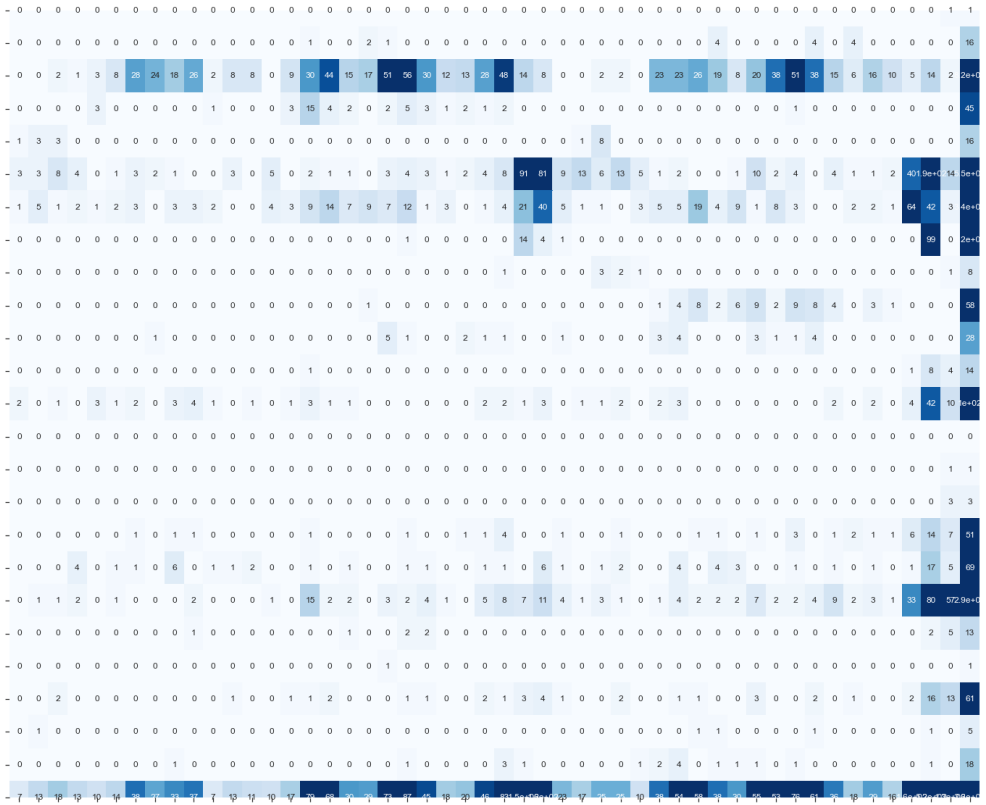


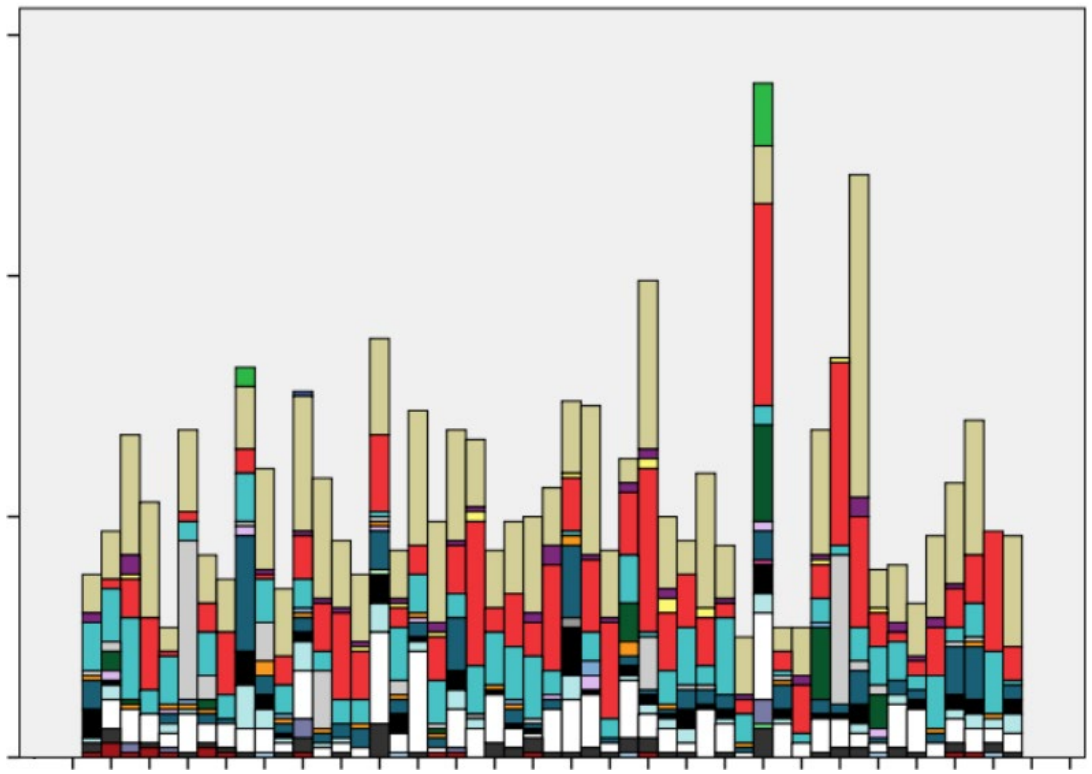


SAXION UNIVERSITY OF APPLIED SCIENCES

Example use case: carriers at Scania

Data exploration ≈ data visualization





SAXION UNIVERSITY OF APPLIED SCIENCES

Example use case: carriers at Scania

Lessons learned (and confirmed)

- Certain locations correlate with notifications
- Certain notifications are main time-consumers
- No direct causal links for urgent notifications
- Necessity of lower-level (system) data
- Never mistrust the gut feeling of a domain expert

- GIGO: Garbage In, Garbage Out

Current steps

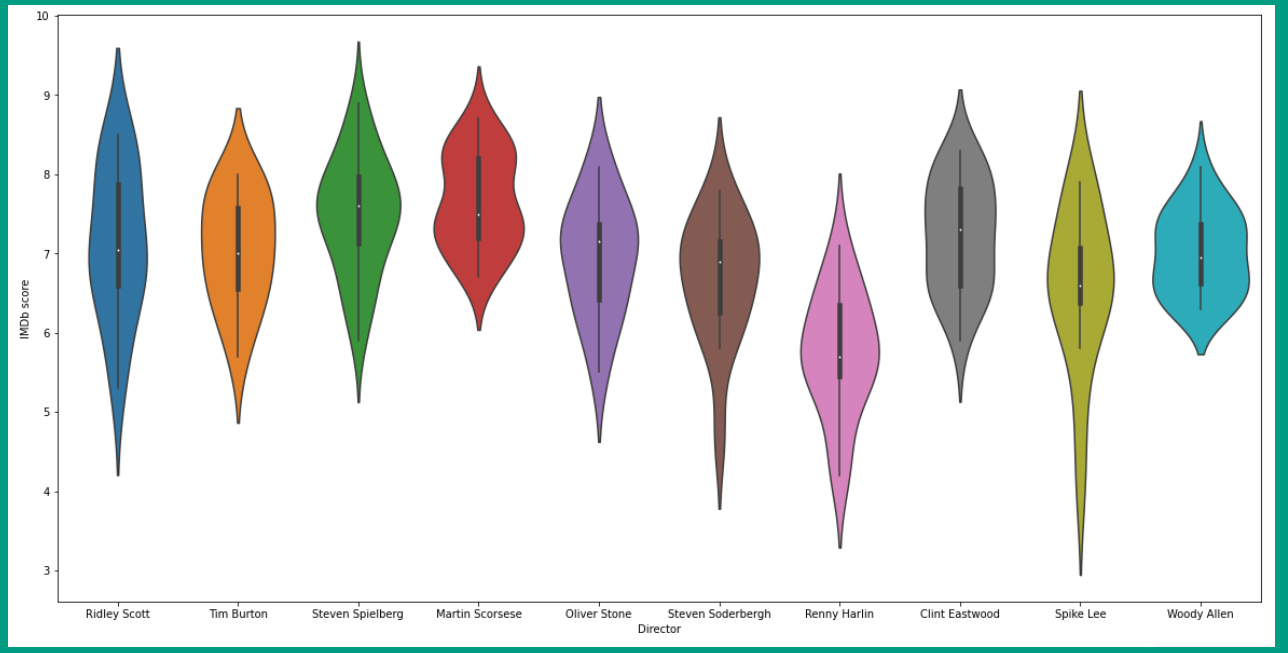
- Collect lower-level (system) data
- Design IoT solution to acquire data
- Align more with in-house domain expertise

Data science with Pandas

Data exploration

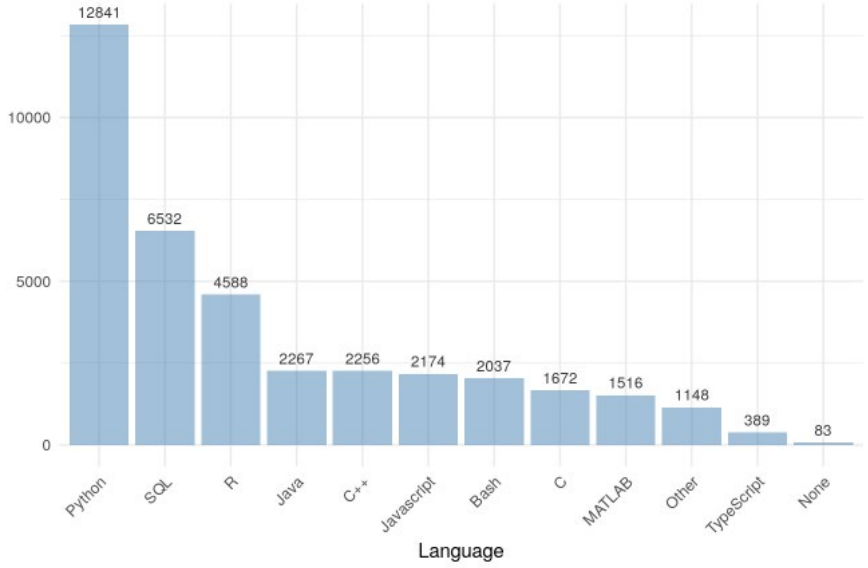
Data visualization


```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

plt.figure(figsize=(20, 10))

movie_score_histogram = sns.distplot(movie_simple['imdb_score'],
                                     bins=20
                                     )
```



Pandas, in short

Python
De facto standard
for data scientists
(next to R)

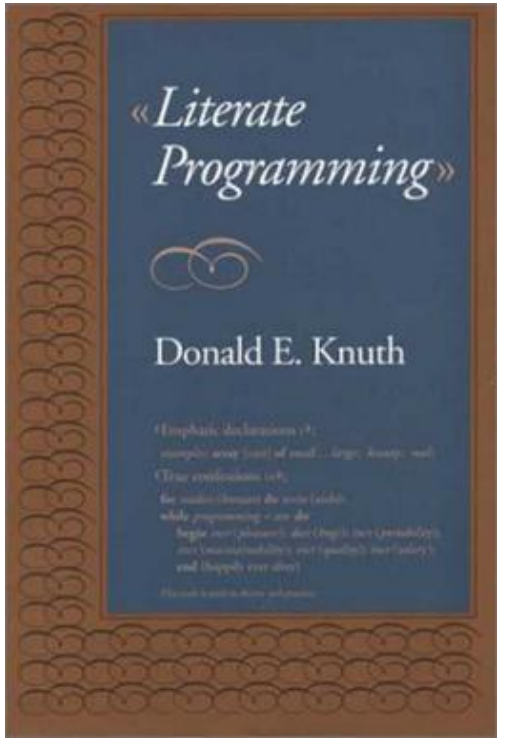


python™ pandas

colab

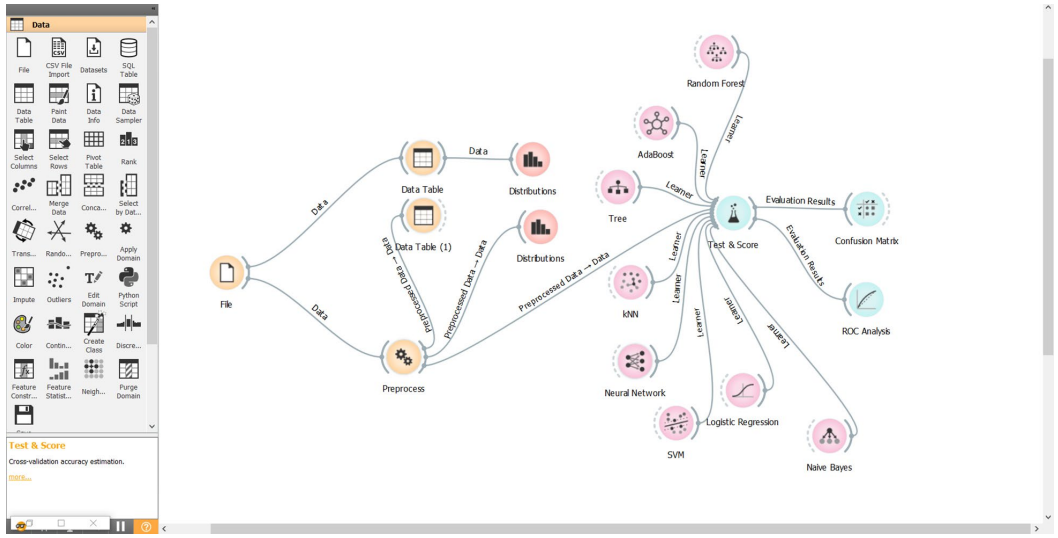
Python notebooks + Google Colab

Literate programming (Donald Knuth):
intersperse source code with explanation in natural language

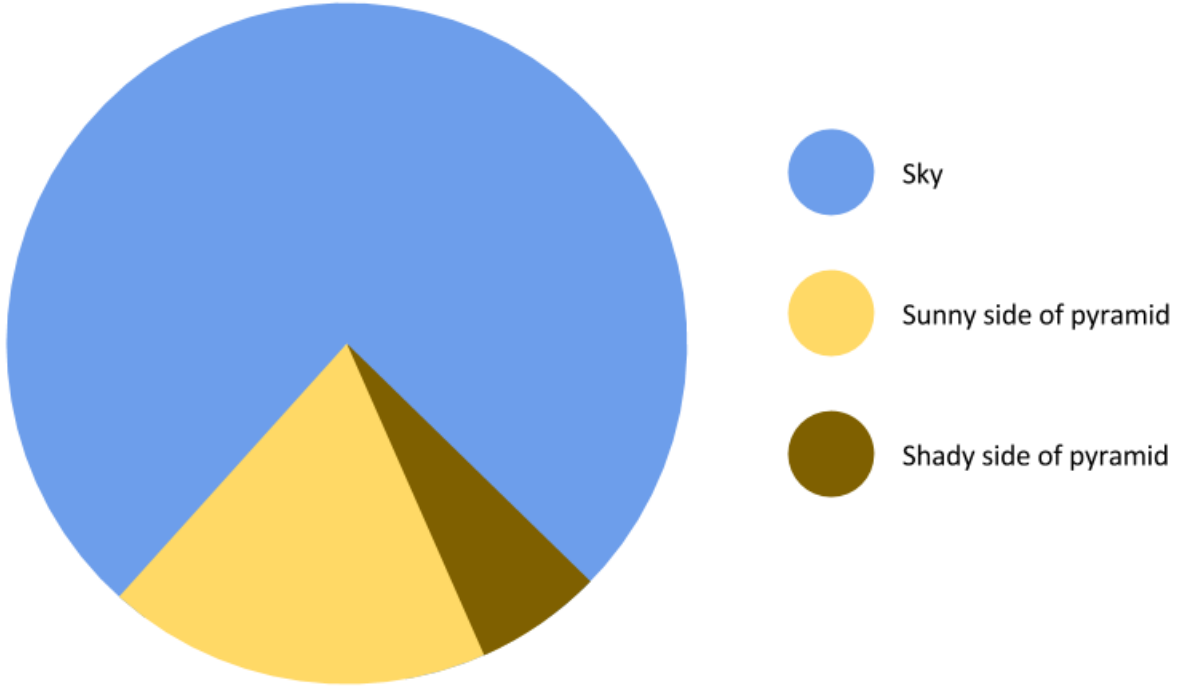


https://colab.research.google.com/drive/1ZR9k-76CFS1CVG3hicifc-314QUQVDGE?usp=sharing

Other tools for data exploration

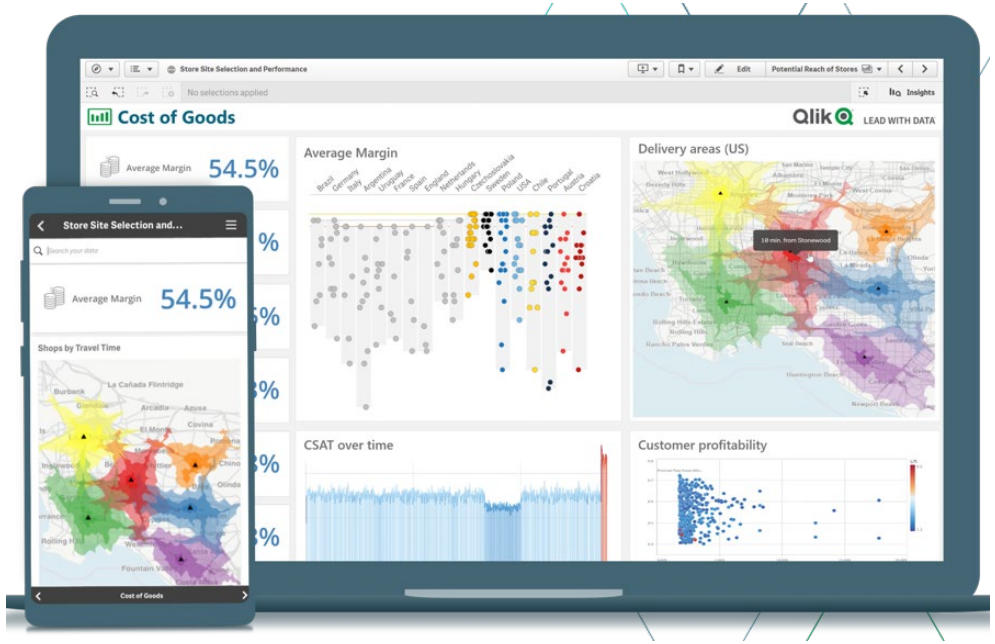


orange.biolab.si

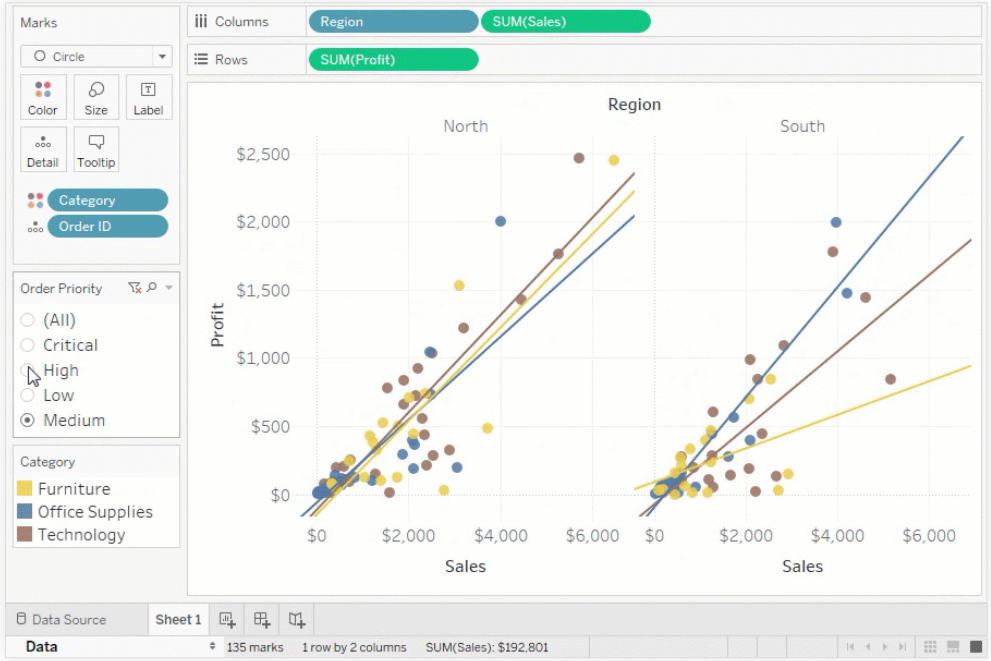


data-to-viz.com

Other tools for data exploration



Qlik Sense



Tableau

More data mining resources

Python Pandas Tutorial: A Complete Introduction for Beginners
https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/

Orange Data Mining Tutorial Videos
https://www.youtube.com/channel/UCIKKWBe2SCAEyv7ZNGhIe4g

Tableau (Official) Free Training Videos
https://www.tableau.com/learn/training/20201

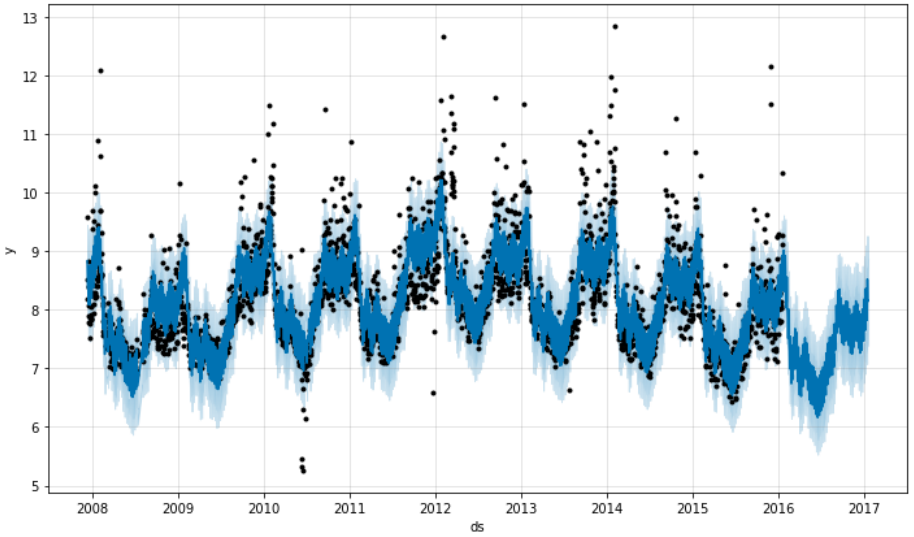
From data mining to machine learning

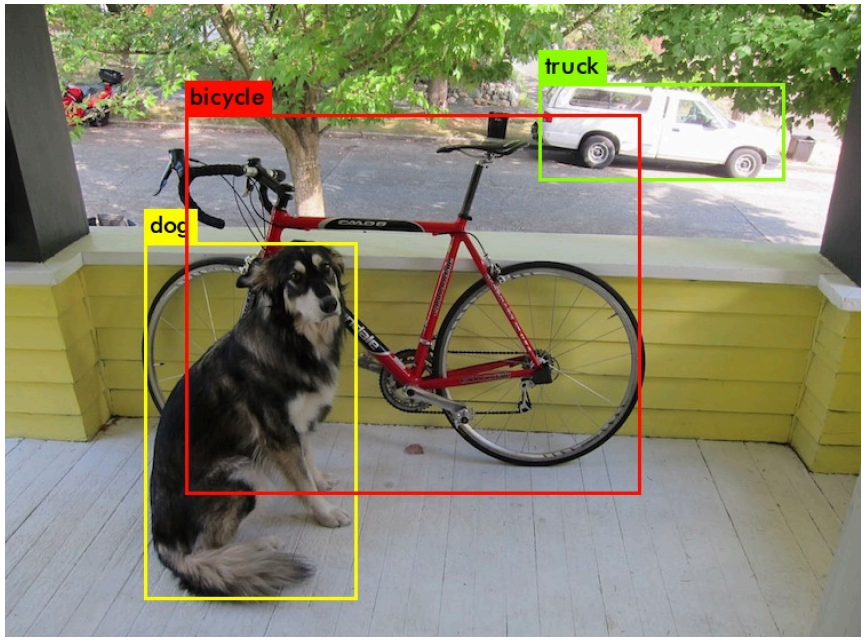
Worthy data
- Business objective clear
- Data mining objective clear
- Clean data

Be prepared to iterate (enter a 2nd, 3rd, 4th… loop)

Certain (niche) cases are well-developed
- Time series forecasting
- Object recognition

Challenges remain
- Anomaly detection
- Unlabelled data





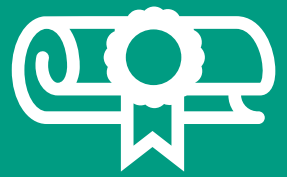
Recap

Learning goals

- Employing CRISP-DM
  - Business objective
  - Data understanding
  - Data exploration

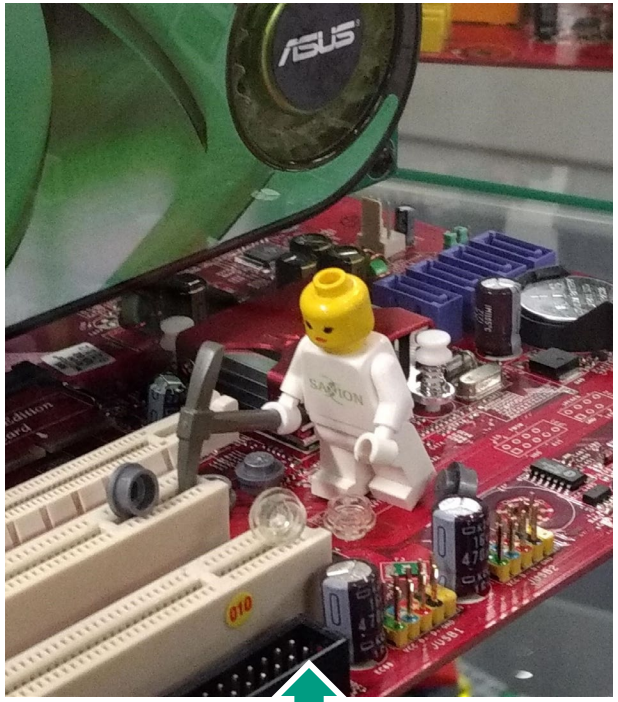
- Data mining
  - Exploration ≈ visualization
  - Use the tool that fits

Open questions?



All learning material will be made available.

Thanks for your attention! Want more?



(Data mining)

Attend our next sessions!

4 June: machine learning
18 June: deep learning

Send an email!

j.m.linssen@saxion.nl

Visit our websites!

saxion.nl/ami

tvalley.nl

boostsmartindustry.nl

SAXION UNIVERSITY OF APPLIED SCIENCES

Media sources

Oost NL AI map: https://oostnl.nl/ai
Lovelace: By Antoine Claudet - File:Ada Byron daguerreotype by Antoine Claudet 1843 or 1850.jpg, Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=82757509
Analytical Engine: By Bruno Barral (ByB), CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=6839854
AI timeline: http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/
Shakey: By SRI International - SRI International, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17294520
A*: https://commons.wikimedia.org/w/index.php?curid=14916867
A* train: By Srossd - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=35900474
Data type taxonomy: Turban, E., Sharda, R., & Delen, D. (2017). Business Intelligence, Analytics, and Data Science: A Managerial Perspective.
Data scientist time spending: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/
CRISP-DM: By Kenneth Jensen - Own work based on: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1), CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24930610
CRISP-DM figures: Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0*. IEEE.
Pearson's r: By Kiatdd - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=37108966
Literate programming: Fair use, https://en.wikipedia.org/w/index.php?curid=31213513
Kaggle survey language usage: https://www.kaggle.com/etsc9287/python-vs-r-the-data-science-rivalry
Pie chart vs bar plot: https://www.data-to-viz.com/caveat/pie.html
Egyptian Pie Chart: https://www.patheos.com/blogs/religionprof/2015/02/egyptian-venn-diagram.html
Qlik Sense: https://www.qlik.com/us/products/qlik-sense
Tableau: https://www.tableau.com/products/desktop
YoloV3 object recognition: https://pjreddie.com/darknet/yolo/
Time series forecasting: https://facebook.github.io/prophet/docs/quick_start.html