Introduction to Data Science

Lecture 1

IDS-L1

Chair of Process and Data Science

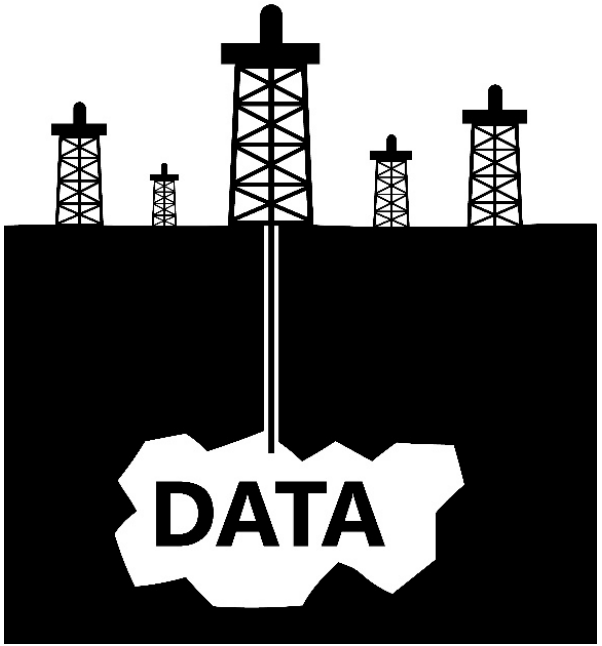
RWTH AACHEN UNIVERSITY

Outline of Today's Lecture

- Outline of course
- Organization and exam regulations
- Motivation (Big Data & Skills)
- DS pipeline
- Types of data
- Terminology
- Data science process
- Challenges

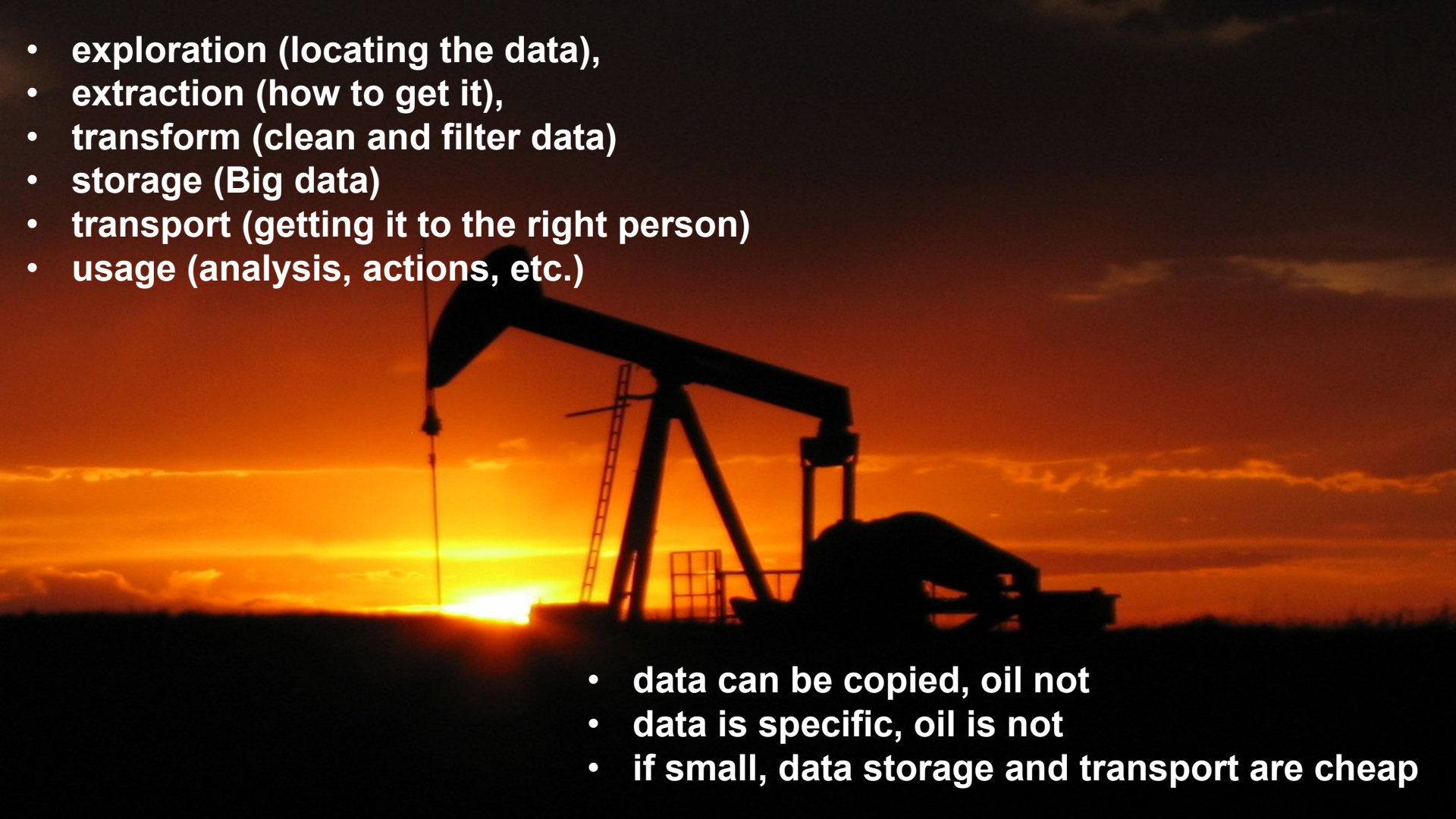
Outline of course

Our society is driven by data, the new oil



- At all levels: personal, device, system, system of systems, organization, nation, world.
- Data volume grows exponentially.
- Allowing for new products and services.

Chair of Process and Data Science



- exploration (locating the data),
- extraction (how to get it),
- transform (clean and filter data)
- storage (Big data)
- transport (getting it to the right person)
- usage (analysis, actions, etc.)

- data can be copied, oil not
- data is specific, oil is not
- if small, data storage and transport are cheap


Four generic
data science
questions

#1

#2

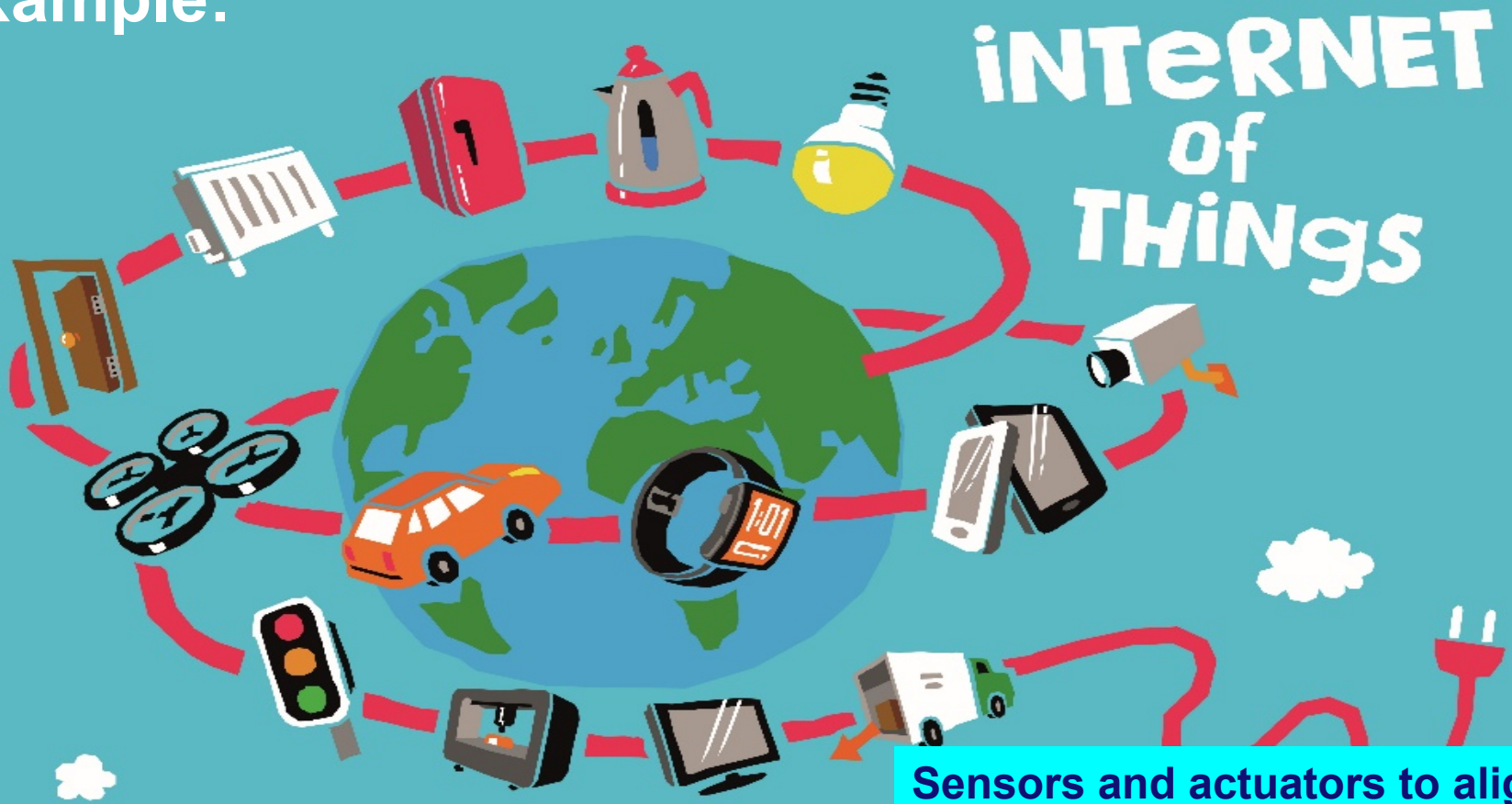
#3


#4
What is
the best that
can happen?

Why now?


10
00
11
10
10
00
0101 0100110001111010100001110001110010011101100010001
110
010
111
100
010
001
111
1100101010101010101011100000101001010010100101100010011110001000010001010001
1000100101001010010110001001010001000110101010001100101001001111100010000100010100
0100101010101111100101001010011100000010100111000000101001011000100111100010000100010100
0000010001001010010100101100110000001010011100000010100101001102101010101010010101010101010100100101100010001
1000100101001010010110001000010000001000101000100110000001010011100000010100101011010100011001010010011111000010
0100101010101111100101001010011100000010100111000000101001011000100111100010000100010100101000101010101010
0000010001001010010100101100000100000010001100000010100111000000101001011000100111100010000100010100101000100101

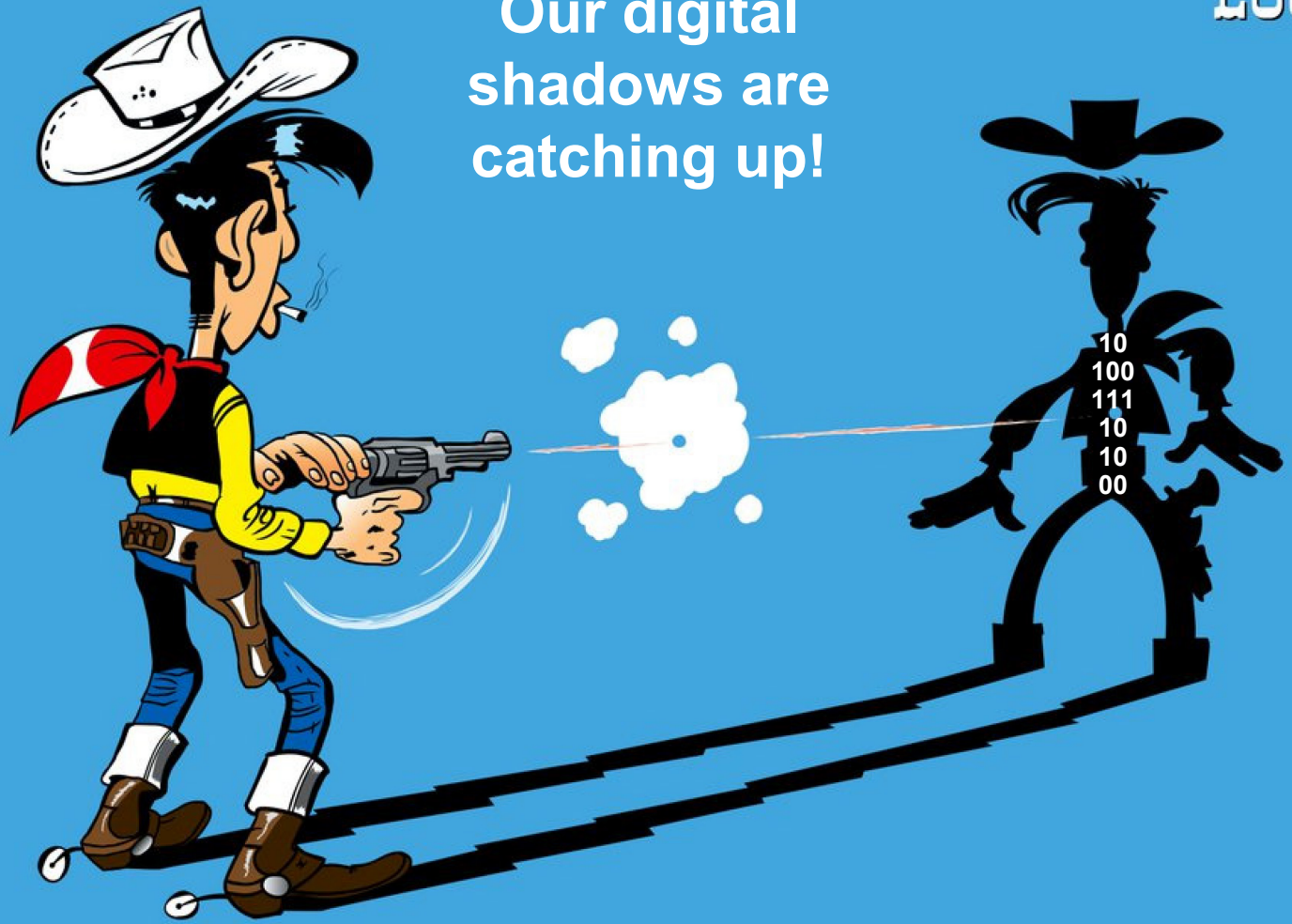
Example:



Sensors and actuators to align the digital and physical world!

Our digital shadows are catching up!

LUCKY LUKE



PAULO FIGUEIREDO
2016

Dimensions

- Different types of data (structured, unstructured, text, images, events, etc.).
- Different types of tasks (supervised, unsupervised).
- Human versus machine (who does what?).
- Algorithm versus visualization.
- Flexibility versus usability (e.g., Python versus Disco).
- Scalability versus quality.
- Responsibility (fairness, privacy, transparency, etc.).

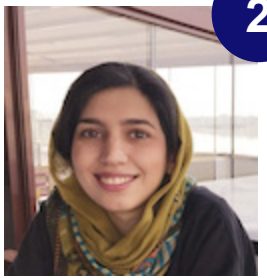
Lectures

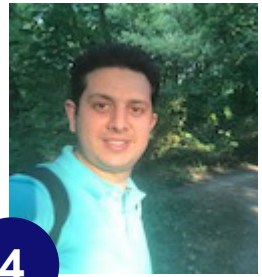
1. Introduction
2. Crash course in Python
3. Basic data visualization/exploration
4. Decision trees (IG)
5. Regression
6. Support vector machines
7. Neural networks (1/2)
8. Neural networks (2/2)
9. Evaluation of supervised learning problems
10. Clustering (K-means)
11. Frequent items sets (a priori and FP growth)
12. Association rules
13. Sequence mining
14. Process mining (unsupervised)
15. Process mining (supervised)
16. Text mining (1/2)
17. Text mining (2/2)
18. Data preprocessing, data quality, binning, etc.
19. Visual analytics & Information visualization
20. Responsible data science: Fairness (discrimination-aware data mining, ethics, etc.)
21. Responsible data science: Confidentiality (various types of encryption and anonym.)
22. Big data (1/2): MapReduce & Distribution
23. Big data (2/2): Systems (Scala, Hadoop, etc.?)
24. Closing

Organization and exam regulations

People involved

1. Wil van der Aalst
2. Mahsa Bafrani
3. Marco Pegoraro
4. Majid Rafiei
5. Yaguang Sun
6. Anja Syring





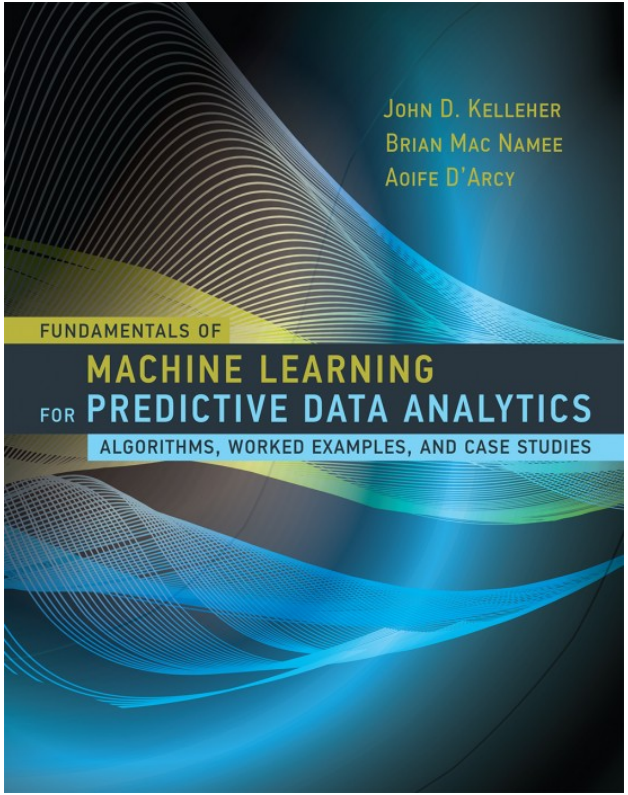
Chair of Process and Data Science

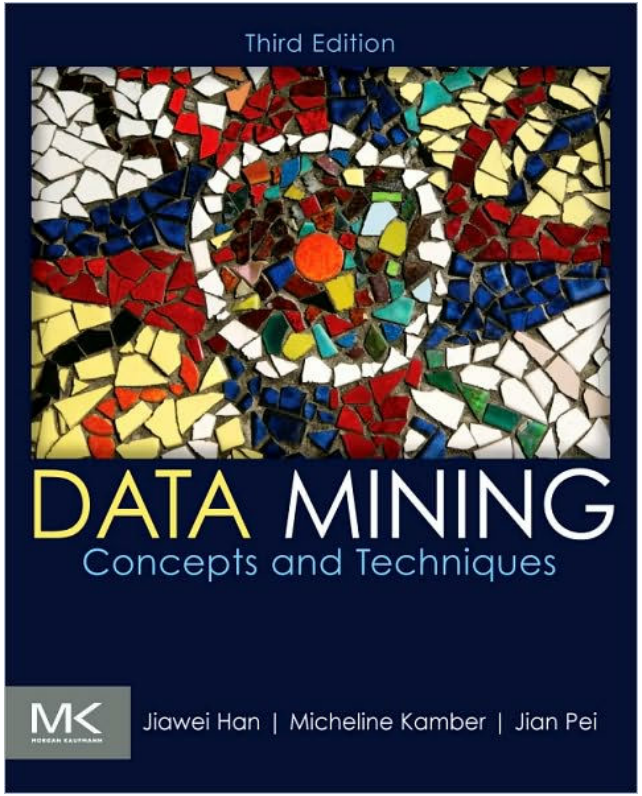
Lectures

- The course starts on 10-10-2018.
- Lectures are on normally Wednesdays and Thursdays from 8.30 to 10.00 in Aula 2 (2352|021).
- Instructions are on Fridays from 8.30 to 10.00 also in Aula 2 (2352|021).
- Note that there are a few exceptions!

#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2	*Decision trees and data visualization/explora*	*19/10/2018*	*Friday*
Lecture 5	Regression	24/10/2018	Wednesday
Lecture 6	Support vector machines	25/10/2018	Thursday
Instruction 3	*Regression and support vector machines*	*26/10/2018*	*Friday*
Lecture 7	Neural networks (1/2)	31/10/2018	Wednesday
Instruction 4	*Neural networks and supervised learning*	*02/11/2018*	*Friday*
Lecture 8	Neural networks (2/2)	07/11/2018	Wednesday
Lecture 9	Evaluation of supervised learning problems	08/11/2018	Thursday
Instruction 5	*Neural networks and supervised learning*	*09/11/2018*	*Friday*
Lecture 10	Clustering	14/11/2018	Wednesday
Lecture 11	Frequent items sets	15/11/2018	Thursday
Lecture 12	Association rules	21/11/2018	Wednesday
Lecture 13	Sequence mining	22/11/2018	Thursday
Instruction 6	*Clustering, frequent items sets, association ru*	*23/11/2018*	*Friday*
Lecture 14	Process mining (unsupervised)	28/11/2018	Wednesday
Lecture 15	Process mining (supervised)	29/11/2018	Thursday
Instruction 7	*Process mining and sequence mining*	*30/11/2018*	*Friday*
Lecture 16	Text mining (1/2)	05/12/2018	Wednesday
Instruction 8	*Text mining and process mining*	*06/12/2018*	*Thursday !!*
Lecture 17	Text mining (2/2)	12/12/2018	Wednesday
Lecture 18	Data preprocessing, data quality, binning, etc	13/12/2018	Thursday
Lecture 19	Visual analytics & information visualization	19/12/2018	Wednesday
backup		20/12/2018	Thursday
Instruction 9	*Text mining, preprocessing and visualization*	*21/12/2018*	*Friday*
Lecture 20	Responsible data science (1/2)	09/01/2019	Wednesday
Lecture 21	Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	*Responsible data science*	*11/01/2019*	*Friday*
Lecture 22	Big data (1/2)	16/01/2019	Wednesday
Lecture 23	Big data (2/2)	17/01/2019	Thursday
Instruction 11	*Big data*	*18/01/2019*	*Friday*
Lecture 24	Closing	23/01/2019	Wednesday
backup		24/01/2019	Thursday
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*
backup		30/01/2019	Wednesday
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*

Material





- **Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies by John D. Kelleher, Brian Mac Namee and Aoife D'Arcy. MIT Press. ISBN: 9780262029445, 624 pages, July 2015 (http://machinelearningbook.com/).**
- **Data Mining: Concepts and Techniques (3rd edition) by Jiawei Han , Micheline Kamber , Jian Pei. The Morgan Kaufmann Series in Data Management Systems, Elsevier. ISBN: 9780123814807, 744 pages, 2011 (http://hanj.cs.illinois.edu/book).**

Material

- Additional background material will be provided as the course progresses.
- Given the nature of the course there is not one textbook (many superficial books or books that focus on one aspect; a rapidly developing area).
- Distribution of slides, assignments, etc. via RWTHmoodle.

Software (see next lecture)

- During the first instruction of the course, we will give detailed installation procedures for all the software.
- You will also find the procedures on RWTHmoodle.
- The study guide describes the required software. For the course, you will need a working installation of Python 3.6.x with some common Data Science software packages like:
  - numpy
  - scipy
  - pandas
  - scikit-learn
  - jupyter
- More packages will need to be installed later in the course.
- Anaconda Python is the advised Python distribution.
  Pip is the advised package manager.
  JetBrains Pycharm Community Edition is the advised IDE.

Examinations and Assignments

- The exam consists of three parts: two assignments (Schriftliche Hausarbeit) each counting for 20% of the final result, and the final written test which counts for remaining 60% of the final result.
- Participation in the assignments is required for participation in the final test.
- Only the final test can be retaken in this semester (there will be one re-exam). Assignments can only be redone in the next academic year.

Examinations and Assignments

- **Final written test (60%) Questions to test the theoretical knowledge of the algorithms and techniques learned:**
  - **First option (PT1): 25-02-2019 09:00 – 11:00 in Aula 2**
  - **Second option (PT2): 25-03-2019 09:00 – 11:00 in Aula 2**
- **Schriftliche Hausarbeit / DS Assignment 1 (20%): Analysis of a real-life and/or synthetic data sets using the techniques and tools provided in the course. This assignment is used to test the understanding of the material in lectures 1-10. Deadline Sunday 09-12-2018 23:59.**
- **Schriftliche Hausarbeit / DS Assignment 2 (20%): Analysis of more complex data sets using various data science techniques. This includes the interpretation of the results and creatively using multiple views on the data. The focus is on the lectures 11-21. Deadline Sunday 20-01-2019 23:59.**

Important: participation in both Schriftliche Hausarbeiten / Assignments is a prerequisite for taking the written exam. The three parts form a whole and it is not possible to retake parts of the course, i.e., the results of the assignments expire after the exam.

Make sure

- to register for the course and final exam, and
- to hand in the assignments in time.

(sounds simple, but there are always sad cases)

Who can take this course?

- As long as capacity allows we welcome all students.
- However, the course is a master course. It is mandatory for students taking the Data Science master (it is listed as a Wahlpflichtfach, but a requirement for doing a master thesis). It is a Wahlpflichtfach for Informatik, Media Informatics, and Software Engineering.
- Other students are free to participate, but it is up to the management and rules of the corresponding programs to decide whether the course "counts". (Ask your Studienberater/Prüfungsamt.)
- If you have problems with RWTHonline, RWTHmoodle, etc., that are not specific for this course, please contact the persons responsible for these systems and not the lecturer.

Warning

- **Slides are intended to be self-contained, but …**
- **Overlap with earlier / later courses is unavoidable.**
  (This is an introduction course and gives a broad overview.)
- **The diversity of the course will make it tricky, so stay synchronized.**

Motivation (Big Data & Skills)

Impact and size of data are obvious

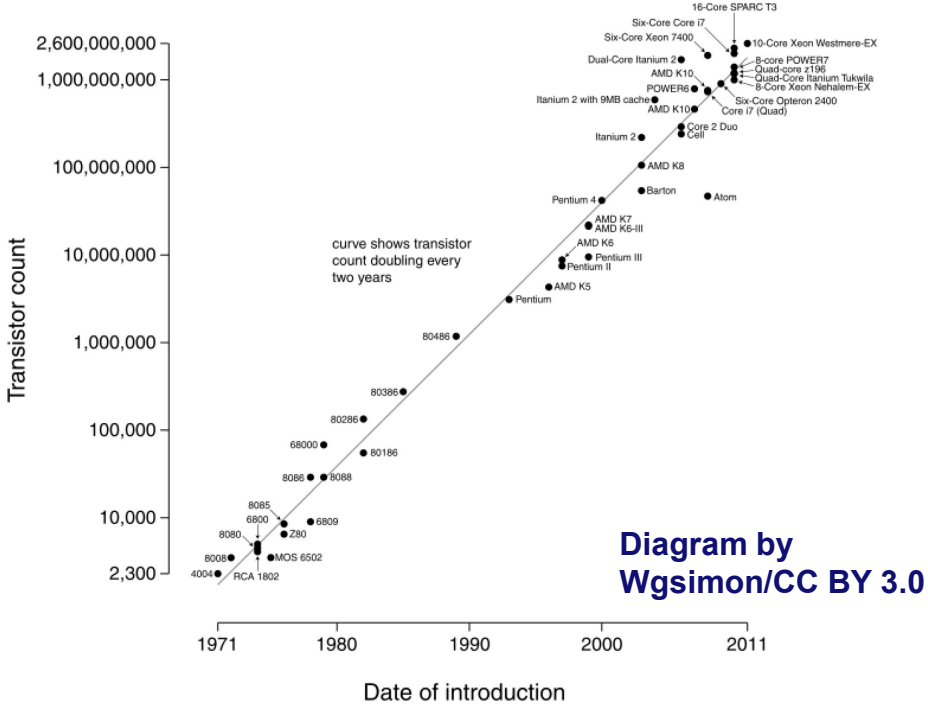
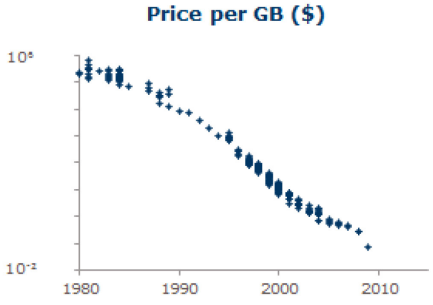
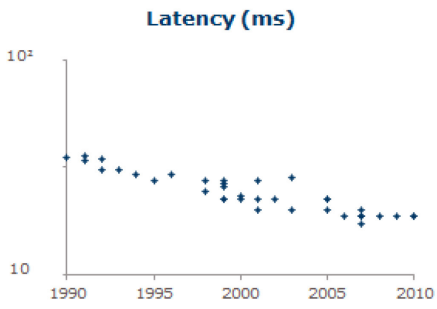
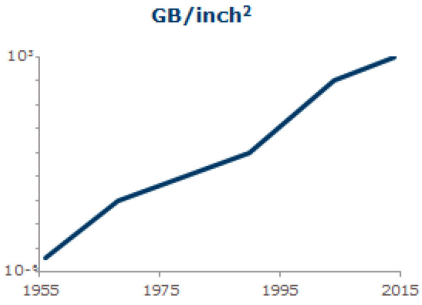
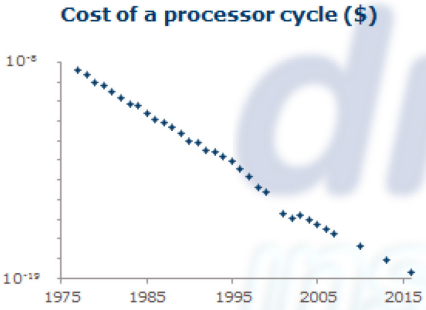
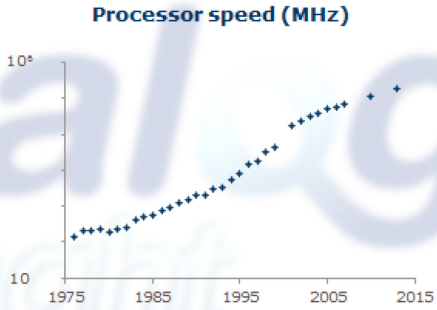
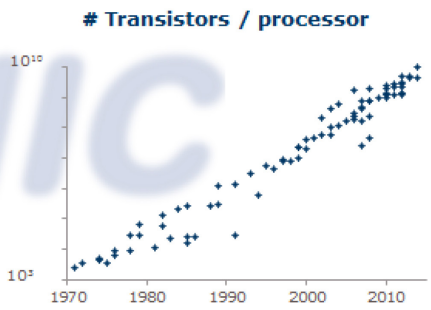
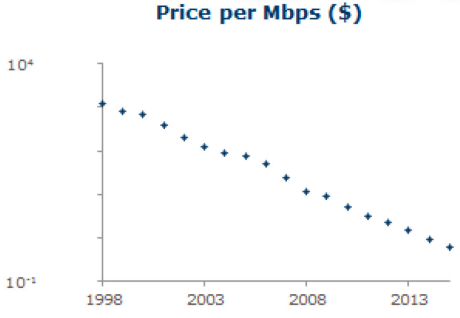
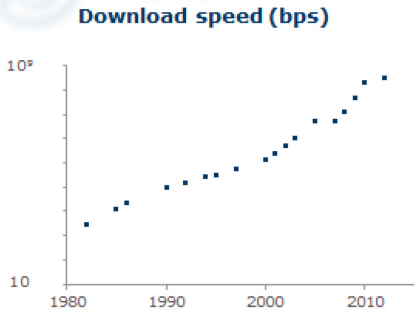
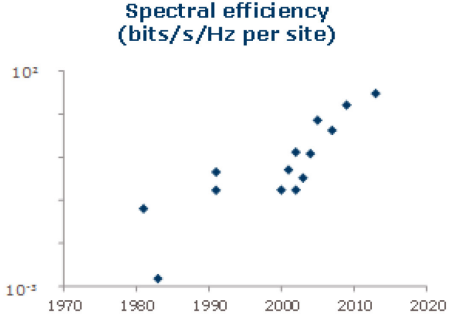


Diagram by Wgsimon/CC BY 3.0

2^{20} = 1.048.576x in 40 years

Chair of Process and Data Science

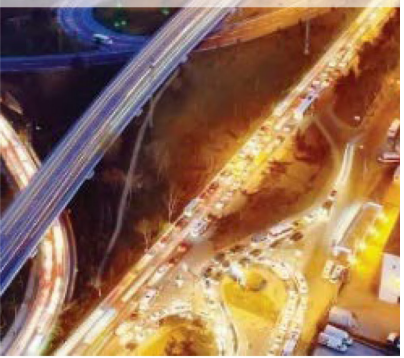
	Cheaper	Faster	More compact
Storage			
Processing			
Transmission			

(Source: Dialogic, 2014. The impact of ICT on the Dutch economy based on *Bronnen: mkomo.com (linksboven), storagenewsletter.com (middenboven), IBM (rechtsboven), singularity.com (linksmidden + midden), Moore's law Wikipedia (rechts midden), drpeering.net (linksonder), Nielsen's law (midden onder), Spectrale efficiëntie mobiele technologie 1G t/m 4G: Wikipedia.org (rechtsonder).*

pwc.co.nz/automation

Will robots really steal our jobs?

An international analysis of the potential long term impact of automation

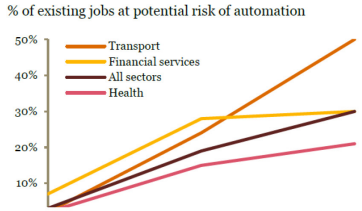


pwc

Key findings: impact of automation

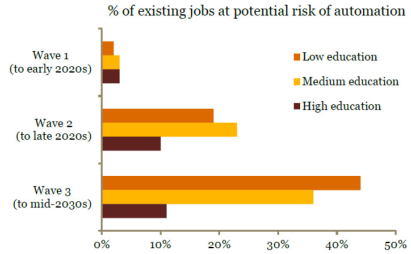
Financial services jobs could be relatively vulnerable to automation in the shorter term, while transport jobs are more vulnerable to automation in the longer term

Figure 1 – Potential job automation rates by industry across waves



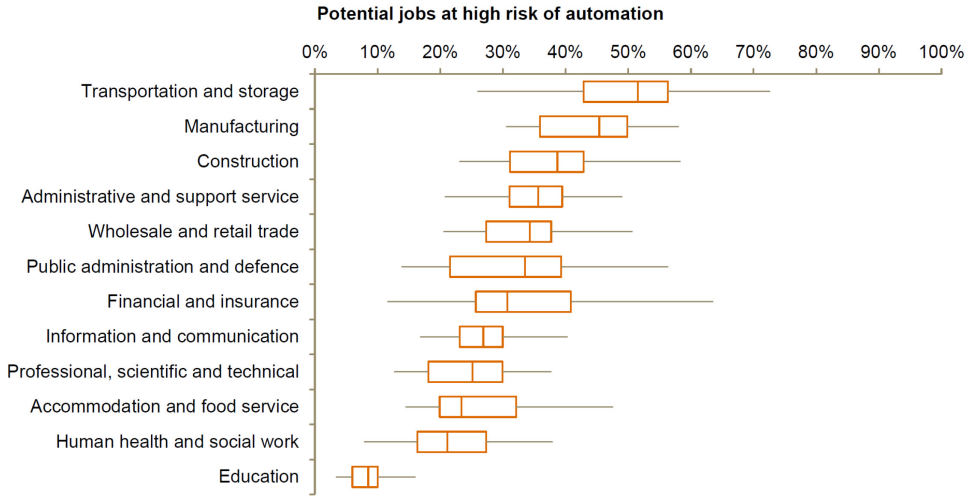
In the long run, less well educated workers could be particularly exposed to automation, emphasising the importance of increased investment in lifelong learning and retraining

Figure 2 – Potential job automation rates by education level across waves



Source: PwC estimates based on OECD PIAAC data (median values for 29 countries)

Figure 4.1 –Share of jobs with potential high automation rates by industry



Waves	Description and impact
Wave 1: **Algorithmic wave (to early 2020s)**	Automation of simple computational tasks and analysis of structured data, affecting data-driven sectors such as financial services.
Wave 2: **Augmentation wave (to late 2020s)**	Dynamic interaction with technology for clerical support and decision making. Also includes robotic tasks in semi-controlled environments such as moving objects in warehouses.
Wave 3: **Autonomous wave (to mid-2030s)**	Automation of physical labour and manual dexterity, and problem solving in dynamic real-world situations that require responsive actions, such as in transport and construction.

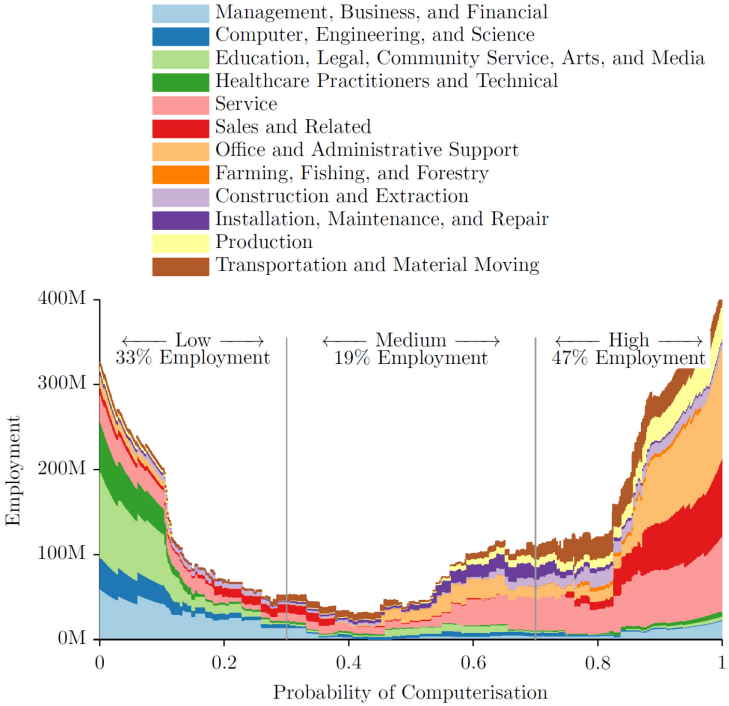
“According to our estimates around 47 percent of total us employment is in the high risk category. We refer to these as jobs at risk – i.e. jobs we expect could be automated relatively soon, perhaps over the next decade or two.”

WORKING PAPER

OXFORD MARTIN SCHOOL

UNIVERSITY OF OXFORD

Published by the Oxford Martin Programme on Technology and Employment



Occupations and probability of computerization (sample from 702 occupations):

- **Healthcare Social Workers 0.35%**
- **Firefighters 17%**
- **Statisticians 22%**
- **Accountants and Auditors 94%**
- **Bookkeeping and Accounting 98%**
- **Tax Preparers 99%**

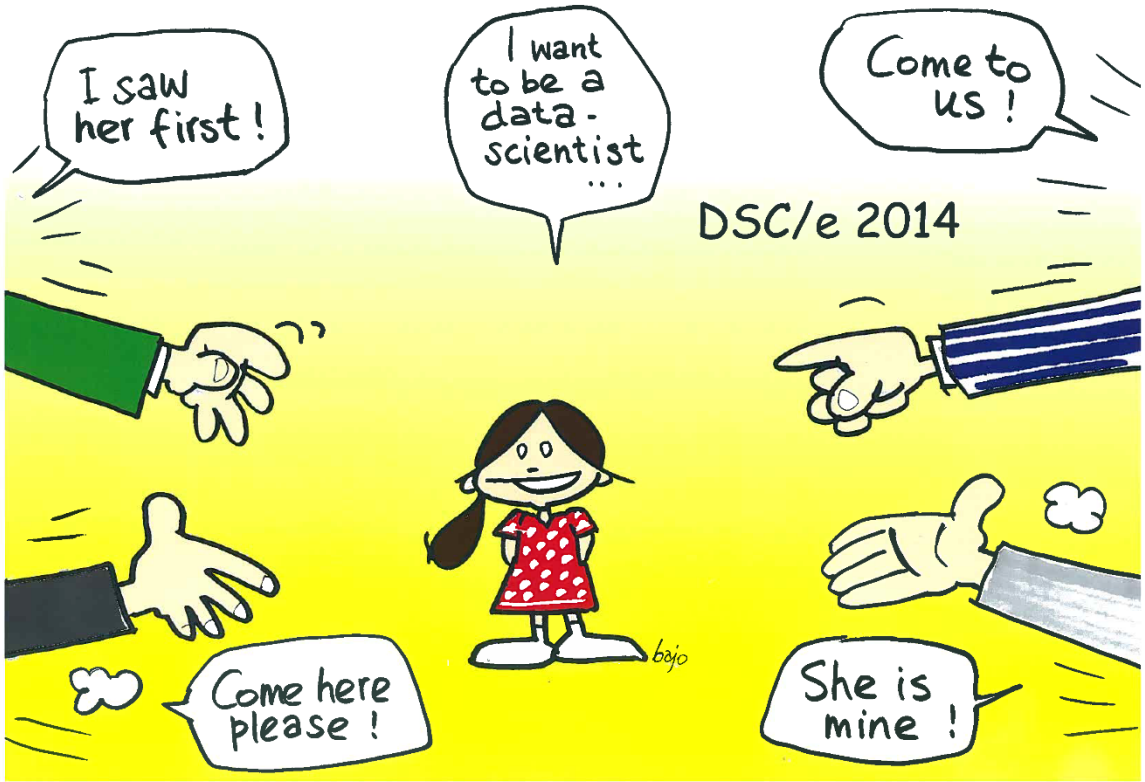
Chair of Process and Data Science

Scope and impact of the growth of data are undeniable

data scientist

Chair of Process and Data Science

Data scientists are valuable

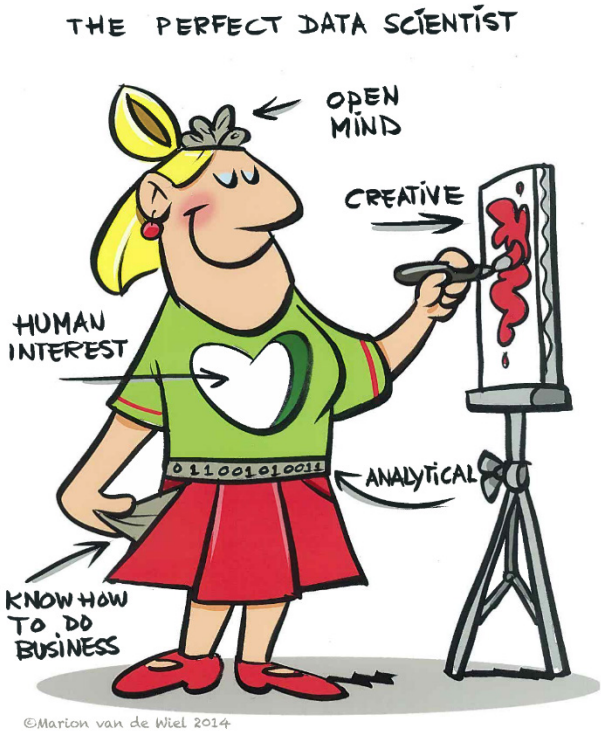


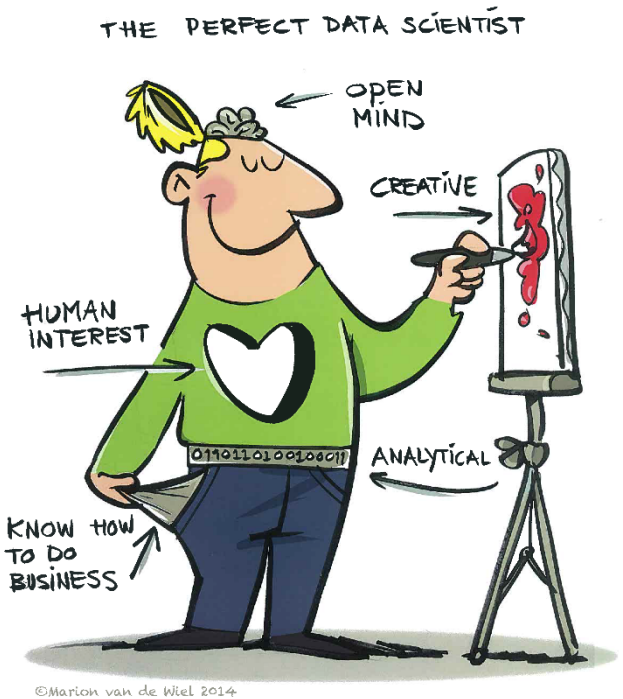
Cartoons where made for DSC/e were I was Scientific Director from 2013-2018

DSC/e

Chair of Process and Data Science

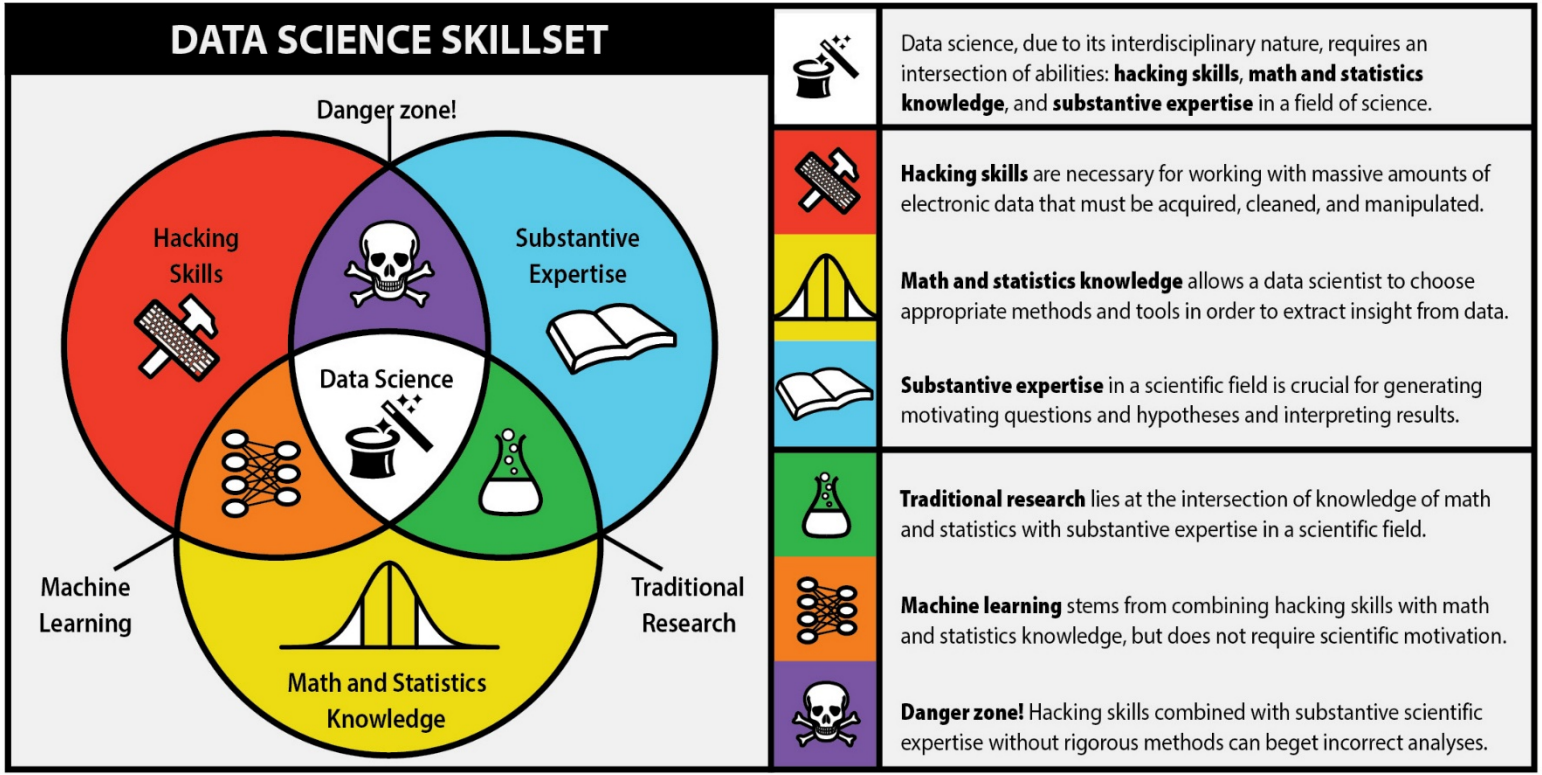
Data scientists need to combine different skills





Chair of Process and Data Science

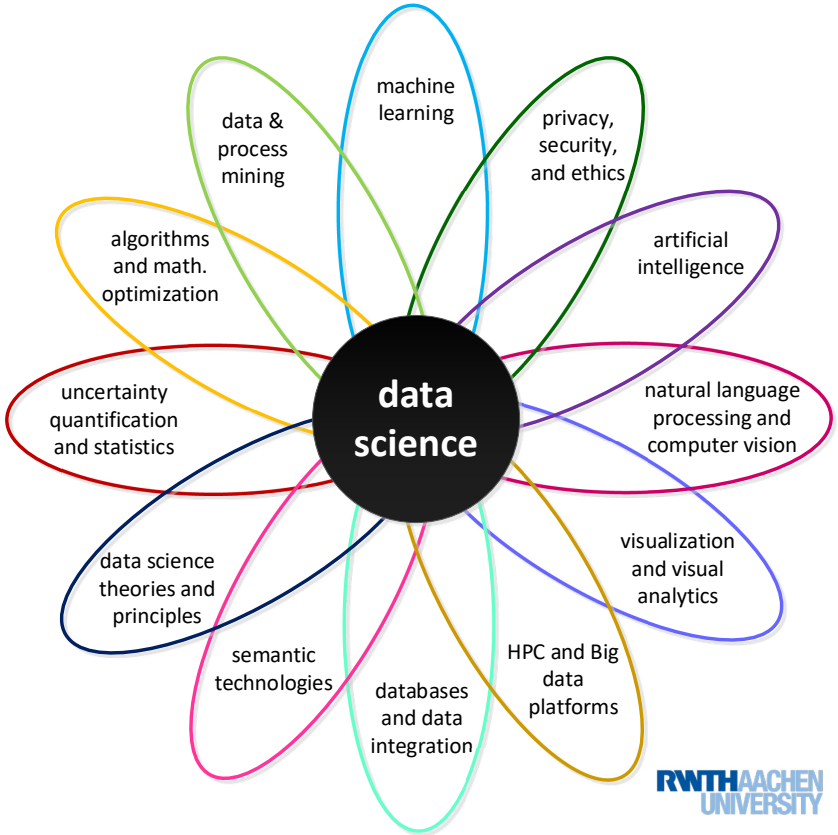
Another characterization



Picture by Natalia Bilenko, Drew Conway, et al.

Chair of Process and Data Science

RWTH Data Science Flower



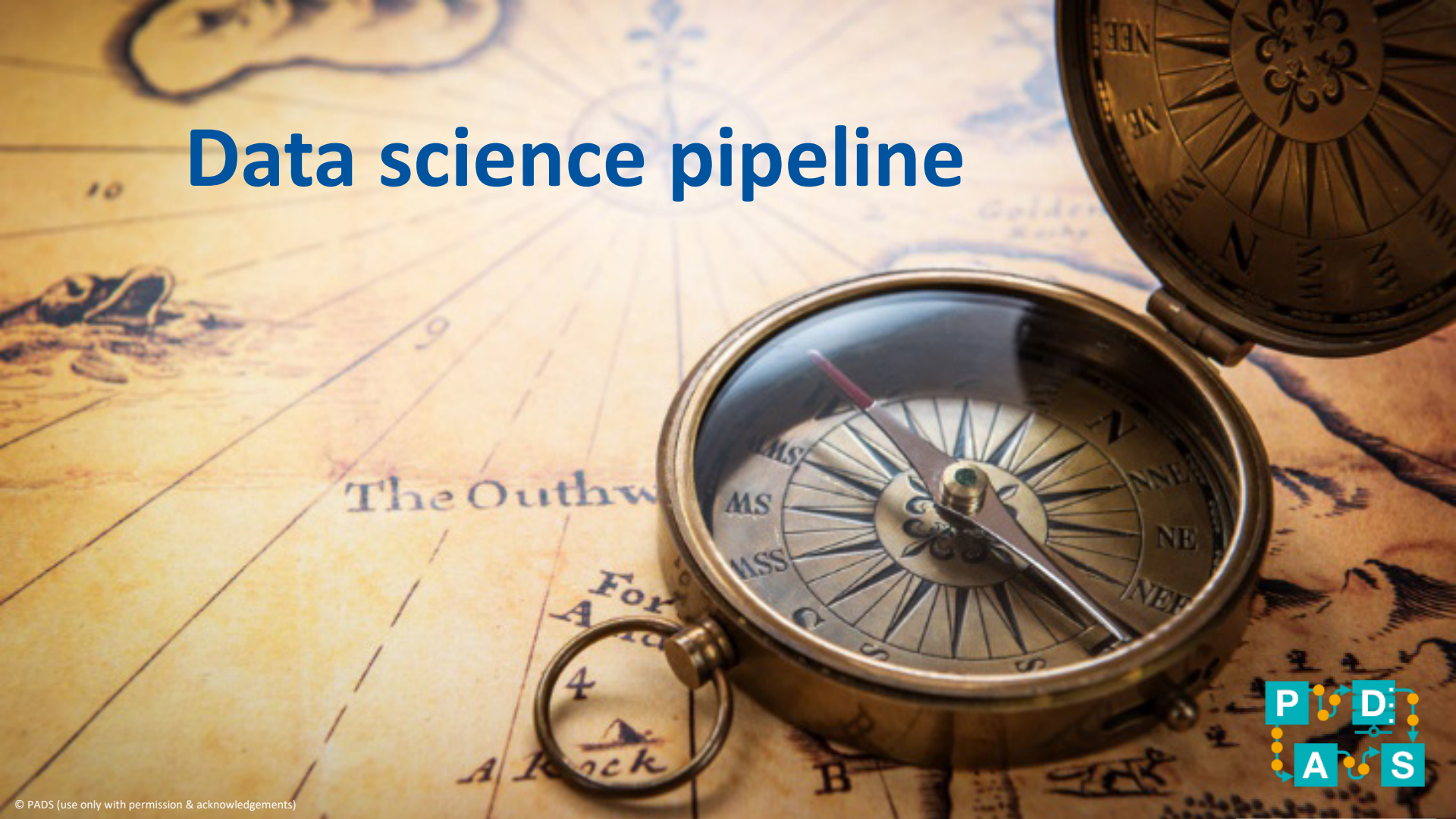
RWTH AACHEN UNIVERSITY

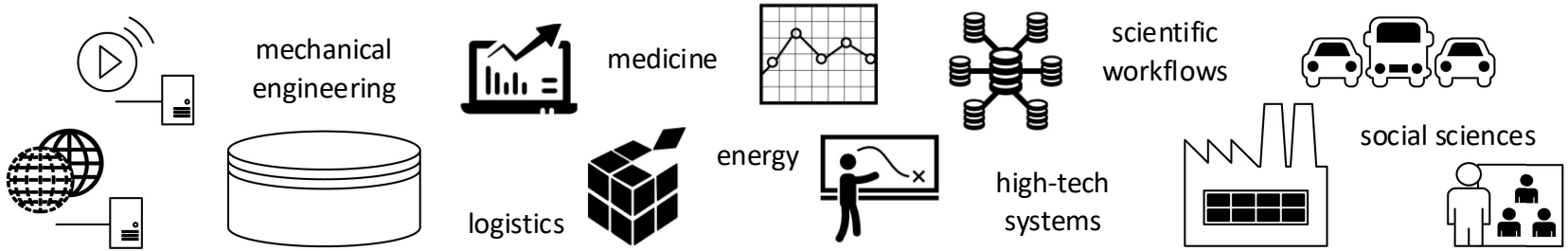
PADS Chair of Process and Data Science

It is not easy to turn data into value

Chair of Process and Data Science

Data science pipeline





infrastructure

"volume and velocity"

- big data infrastructures
- distributed systems
- data engineering
- programming
- security
- ...

analysis

"extracting knowledge"

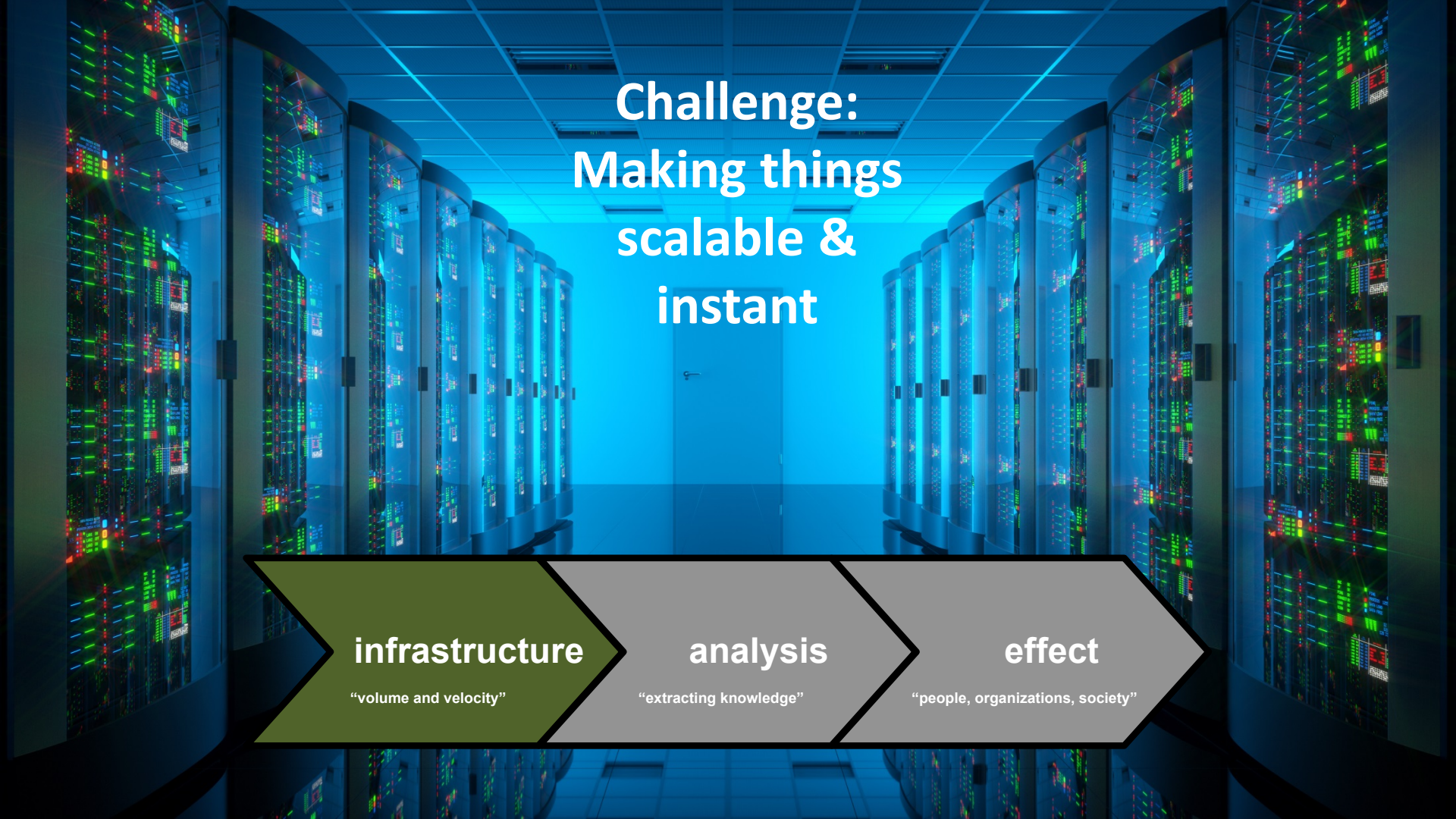
- statistics
- data/process mining
- machine learning
- artificial intelligence
- visualization
- ...

effect

"people, organizations, society"

- ethics & privacy
- IT law
- operations management
- business models
- entrepreneurship
- ...

Chair of Process and Data Science


Challenge:
Making things
scalable &
instant
infrastructure
"volume and velocity"
analysis
"extracting knowledge"
effect
"people, organizations, society"

Challenge: Providing answers to known and unknown unknowns

United States Secretary of Defense
Donald Rumsfeld (12-2-2002)

Challenge: Doing all of this in a responsible manner!

Types of data

Types of data

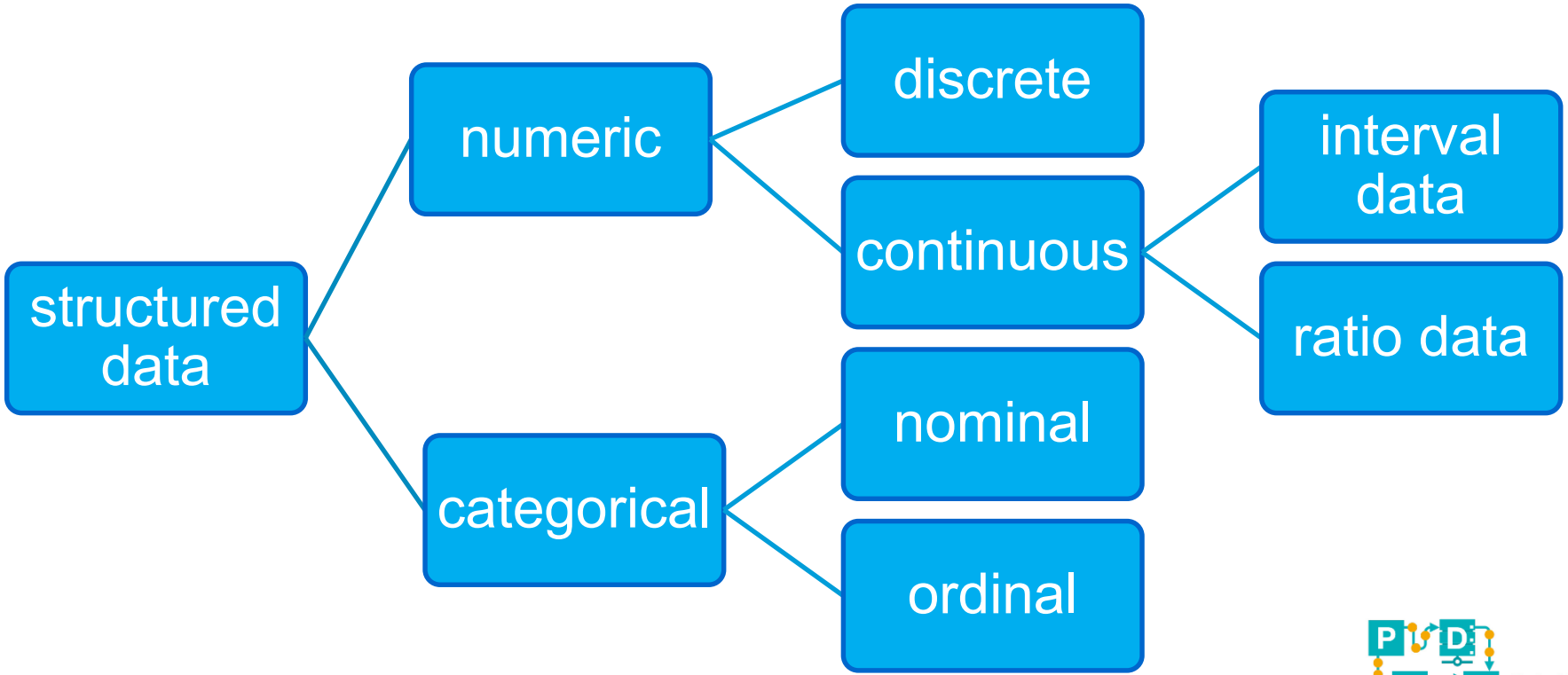
Structured data

- Numerical data (age, time, temperature)
- Categorical data (gender, color, country, class)

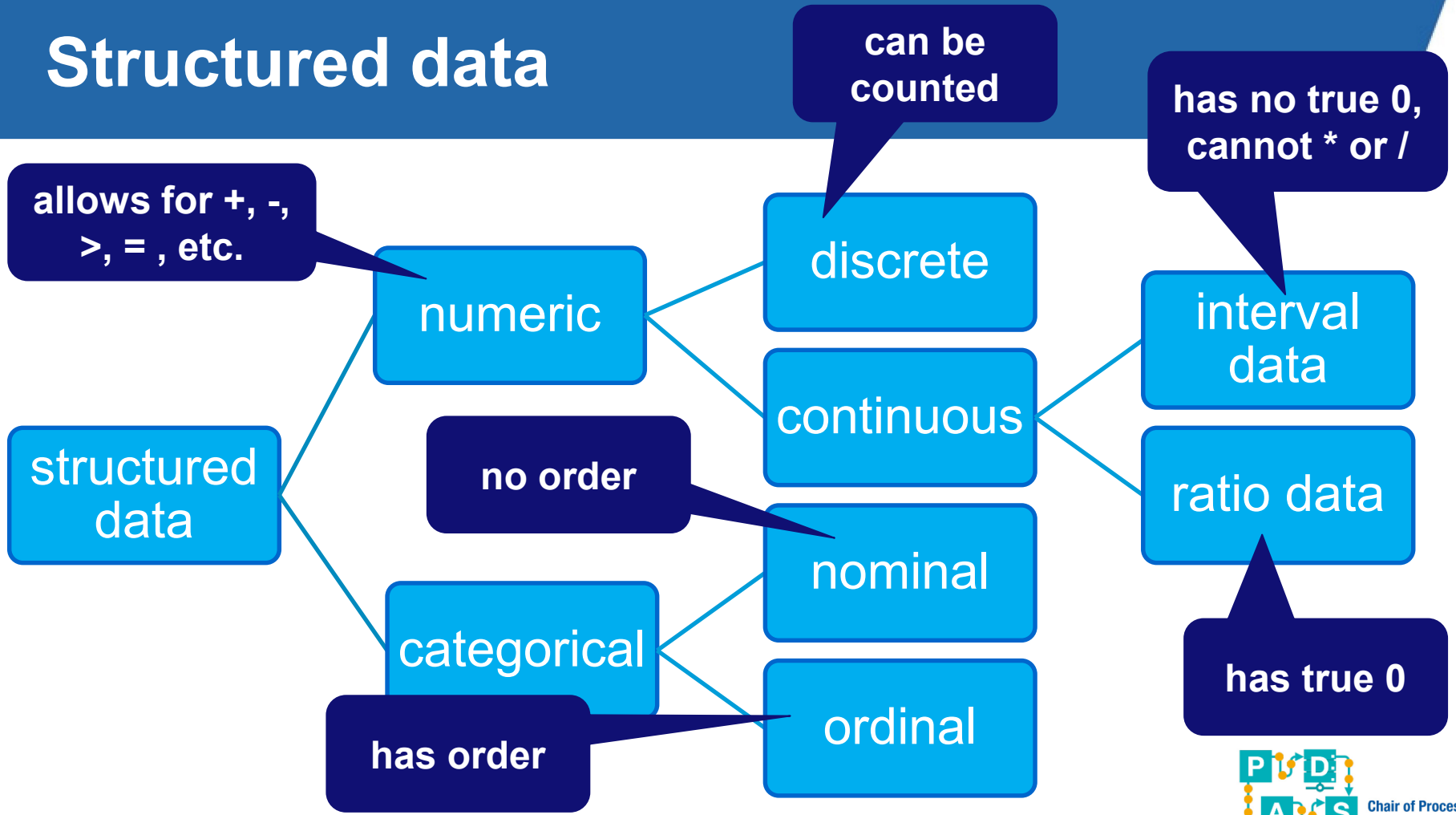
Unstructured data

- Text
- Audio
- Image
- Signal
- Video
- etc.

Structured data

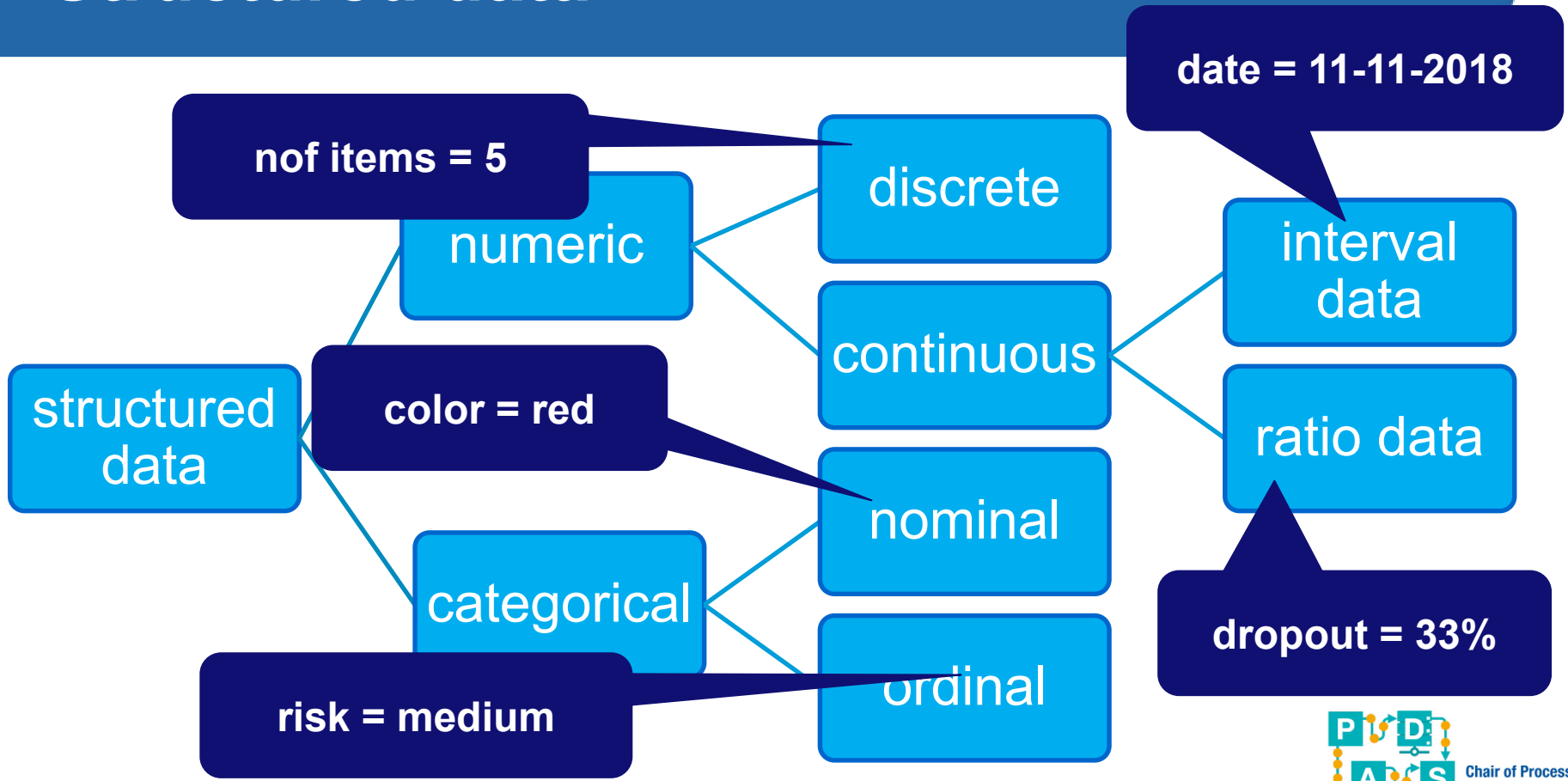


Structured data



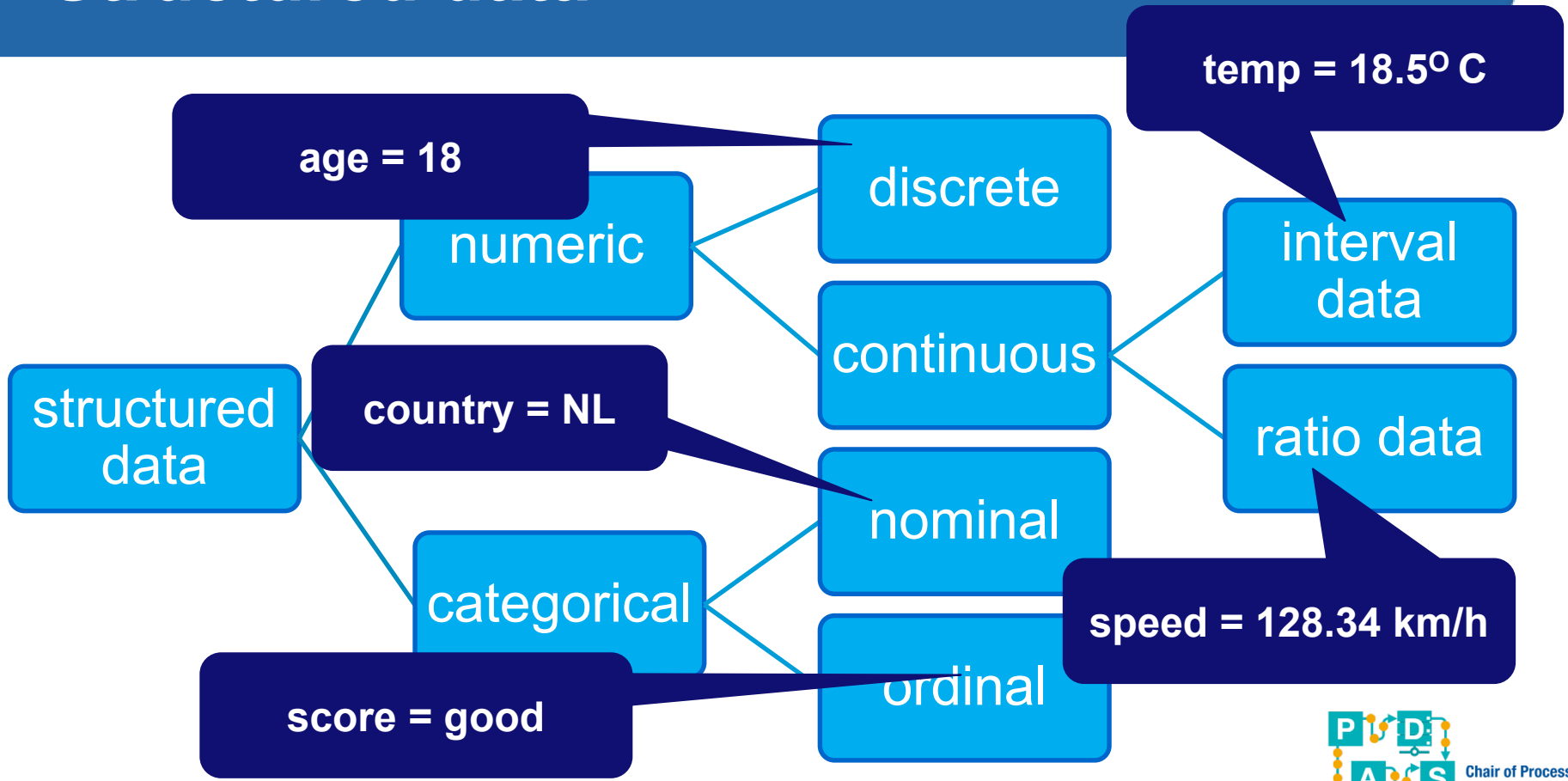
Chair of Process and Data Science

Structured data



Chair of Process and Data Science

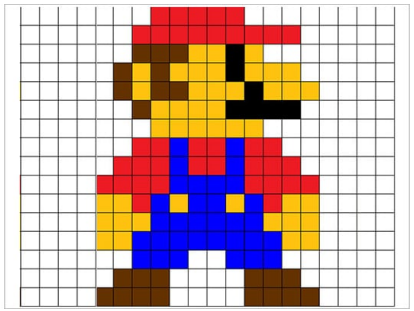
Structured data

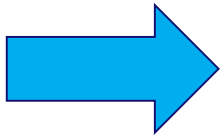


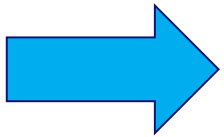
Chair of Process and Data Science

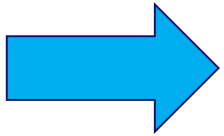
Unstructured data

- Text
- Audio
- Image
- Signal
- Video
- etc.









11111001000100111101

01001001000101110 1

11001001011101010 1

Chair of Process and Data Science

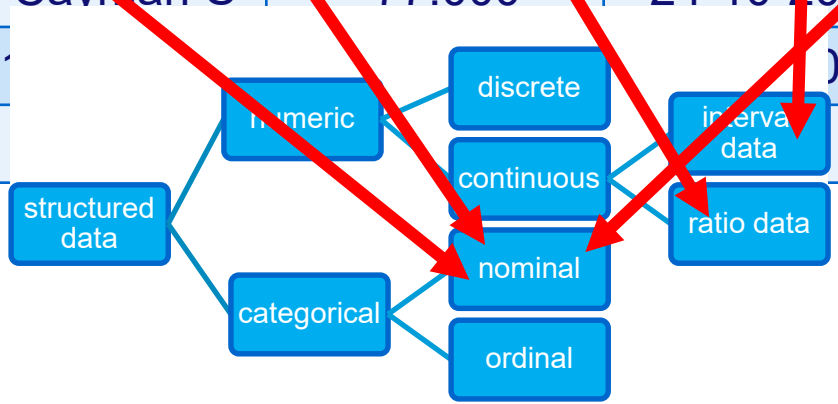
Tabular data

order id	product	price	date	complaint
32424	718 Cayman	66.000	21-10-2018	no
34535	911 Carrera	102.000	22-10-2018	yes
43555	911 Turbo	154.000	24-10-2018	yes
64564	718 Cayman S	77.000	24-10-2018	no
23424	911 Targa	143.000	26-10-2018	yes
…	…	…	…	

columns are features and rows are instances

Tabular data

order id	product	price	date	complaint
32424	718 Cayman	66.000	21-10-2018	no
34535	911 Carrera	102.000	22-10-2018	yes
43555	911 Turbo	154.000	24-10-2018	yes
64564	718 Cayman S	77.000	24-10-2018	no
23424	91		018	yes
…				



Chair of Process and Data Science

Tabular data

from	to	message	image	date
Sue	Pete	“How are you?”		21-10-2018
Pete	Sue	“I’m busy!”		22-10-2018
Pete	Mary	“Let’s go out.”		24-10-2018
Mary	Sue	“Pete joins us.”		24-10-2018
Mary	Kim	“We will go now.”		26-10-2018
…	…	…	…	…

columns are features and rows are instances

Features

- Features are raw or derived (max, min, average, rank, bin, etc.).
- Time plays a special role: time cannot decrease and often we want to predict the future based on the past.
- In case of labeled data there are descriptive features and a target feature.

Labeled tabular data

order id	product	price	date	complaint
32424	718 Cayman	66.000	21-10-2018	no
34535	911 Carrera	102.000	22-10-2018	yes
43555	911 Turbo	154.000	24-10-2018	yes
64564	718 Cayman S	77.000	24-10-2018	no
23424	911 Targa	143.000	26-10-2018	yes
…	…	…	…	

target feature

descriptive features

Chair of Process and Data Science

Labeled tabular data

- Alternative names for descriptive features
  - predictor variables
  - independent variables
- Alternative names for target feature
  - response variable
  - dependent variable
- Alternative names for instances: individuals, entities, cases, objects, or records.

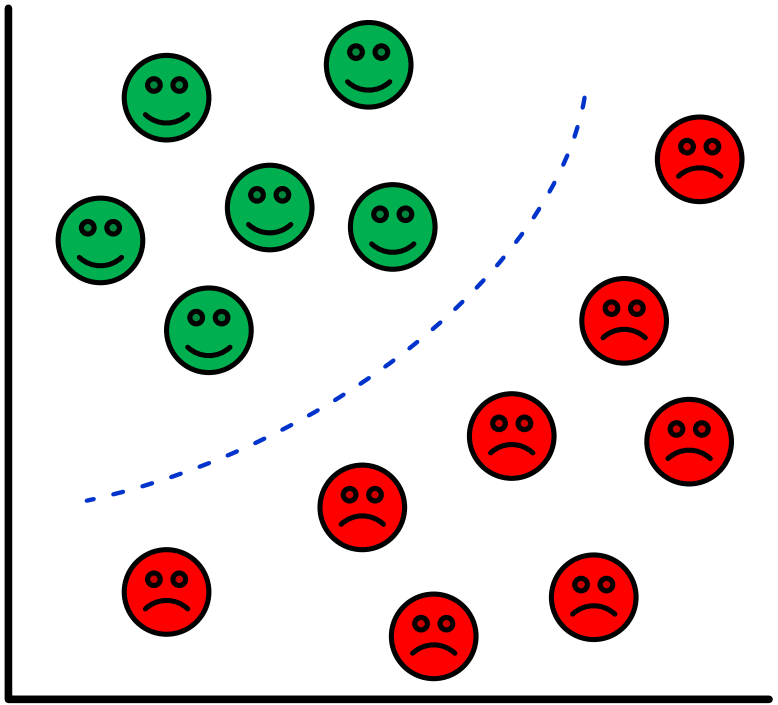
Unlabeled tabular data

from	to	message	image	date
Sue	Pete	“How are you?”		21-10-2018
Pete	Sue	“I’m busy!”		22-10-2018
Pete	Mary	“Let’s go out.”		24-10-2018
Mary	Sue	“Pete joins us.”		24-10-2018
Mary	Kim	“We will go now.”		26-10-2018
…	…	…	…	…

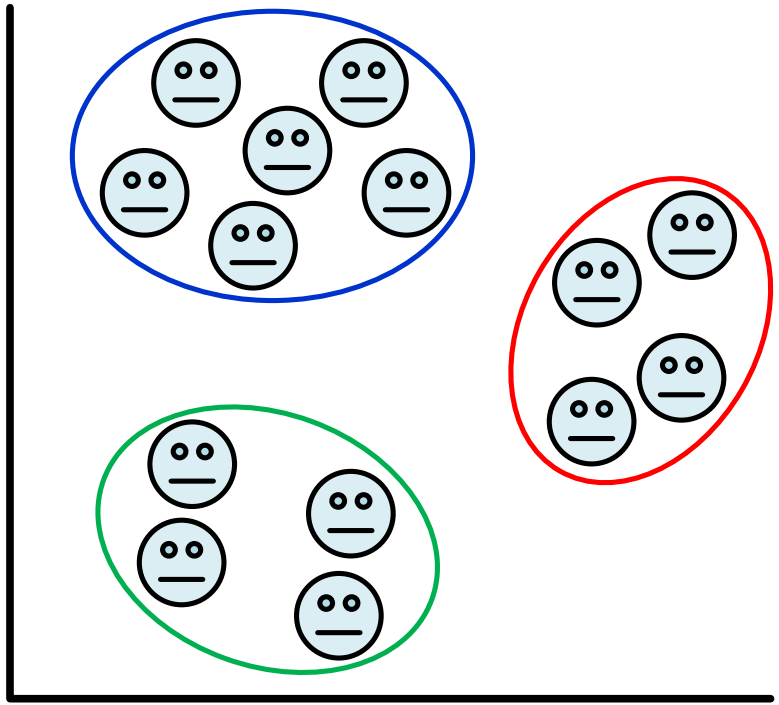
No target feature has been selected

Chair of Process and Data Science

Supervised versus unsupervised learning



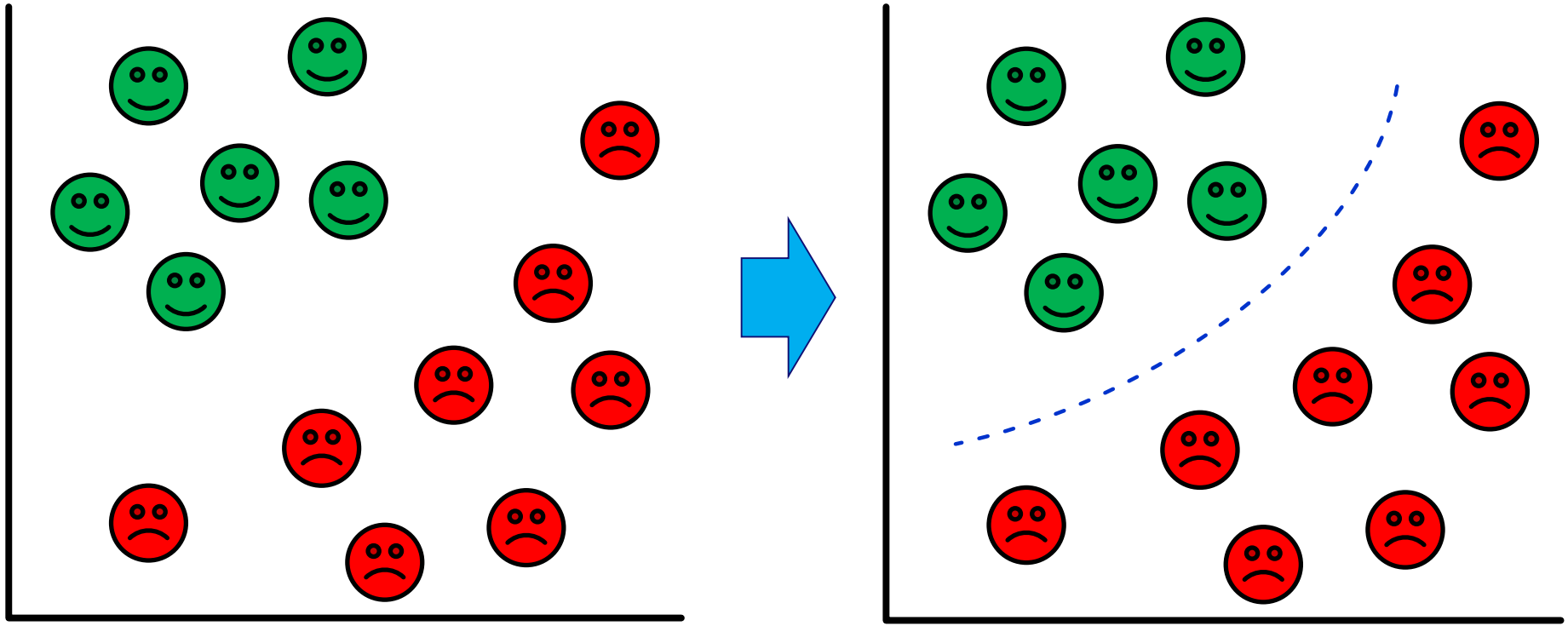
supervised



unsupervised

Chair of Process and Data Science

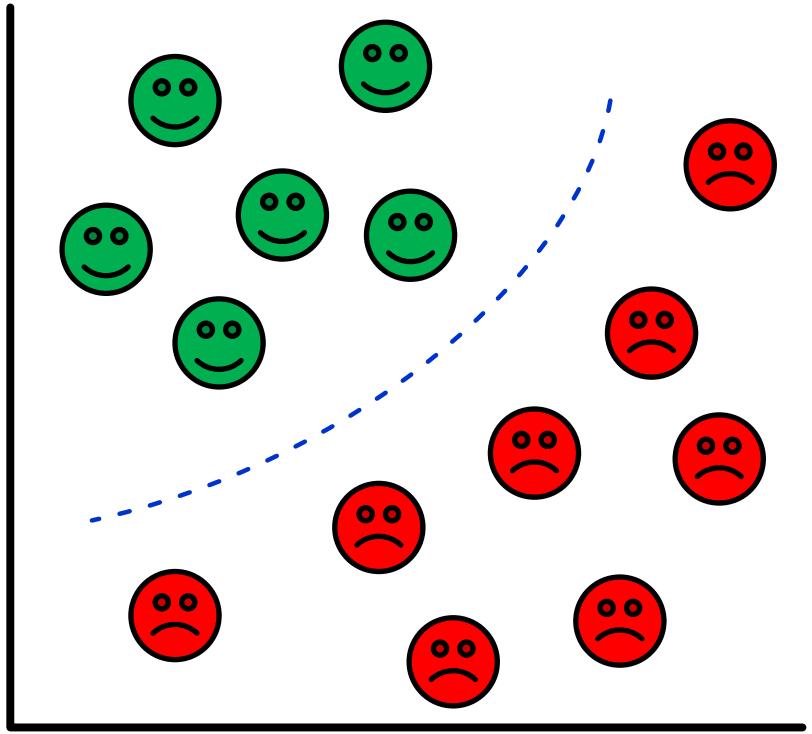
Supervised learning using labeled data



All instances have a target label (here color).

Chair of Process and Data Science

Supervised learning using labeled data

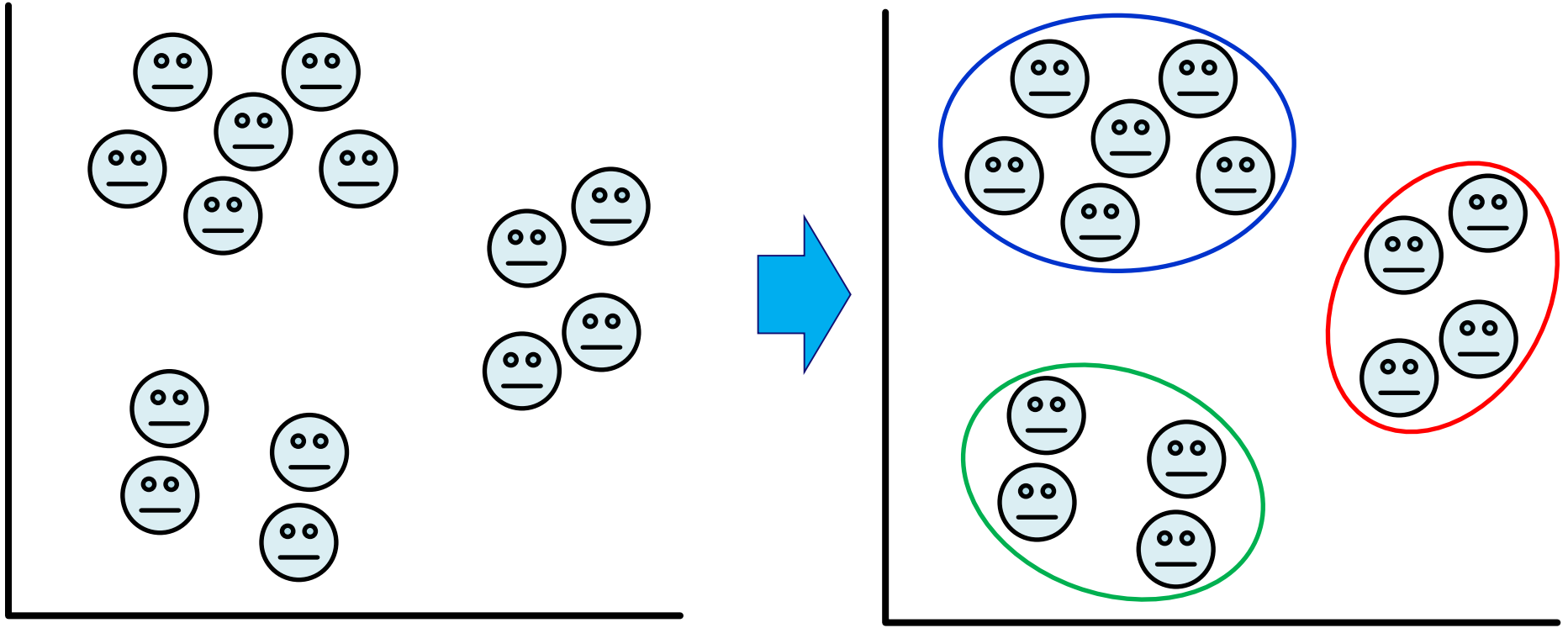


The goal is to find a “rule” in terms descriptive features that explains the target feature as good as possible.

Examples of supervised learning

- **Hospital:**
  - target variable = recover (Y/N)
  - descriptive variables = age, gender, smoking, …
- **University**
  - target variable = drops out (Y/N)
  - descriptive variables = mentor, prior education, …
- **Production**
  - target variable = order is delivered in time (Y/N)
  - descriptive variables = product, agent, …

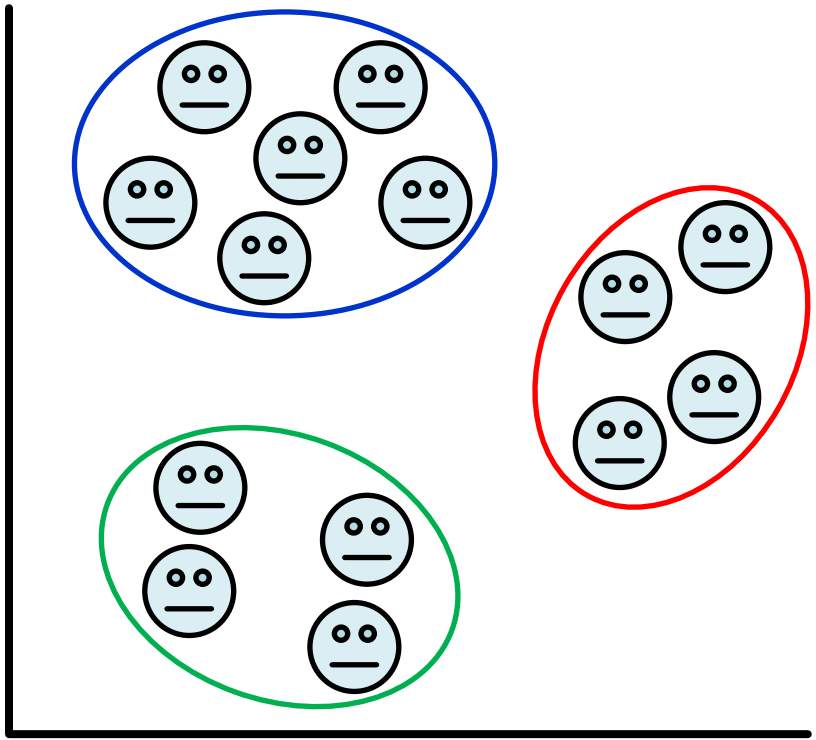
Unsupervised learning



Instances do not have a target label.

Chair of Process and Data Science

Unsupervised learning



The goal is to find clusters or patterns.

- **Clusters are homogeneous sets of instances.**
- **Patterns reveal hidden structures in the data (unknown unknowns).**

Examples unsupervised learning

- Find similar groups of patients, students, customers, orders, cars, companies, etc.
- Find rules of the form (unknown unknowns):
  - Customers that buy bread and butter typically pay by cash.
  - Patients that drink and smoke typically pay the hospital bill earlier than others.
  - Products produced by team A on Monday tend to be returned more frequently by the customer.

Terminology

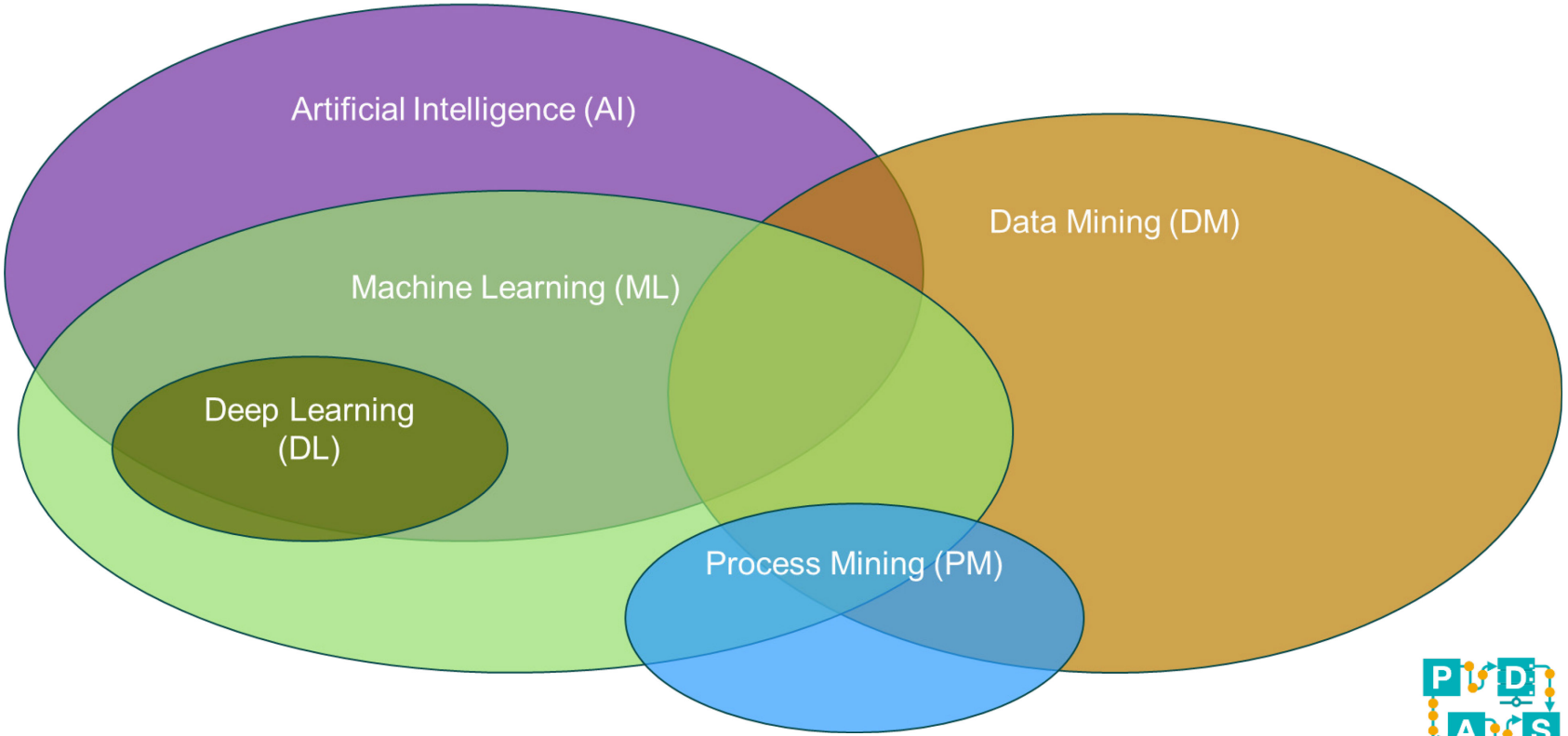
Terminology

- **Many different names** (statistics, data analytics, data mining, machine learning, artificial intelligence, predictive analytics, process mining, etc.) **are used to refer to the key disciplines that contribute to data science.**
- **Unfortunately, the areas these names describe are heavily overlapping and context dependent.**

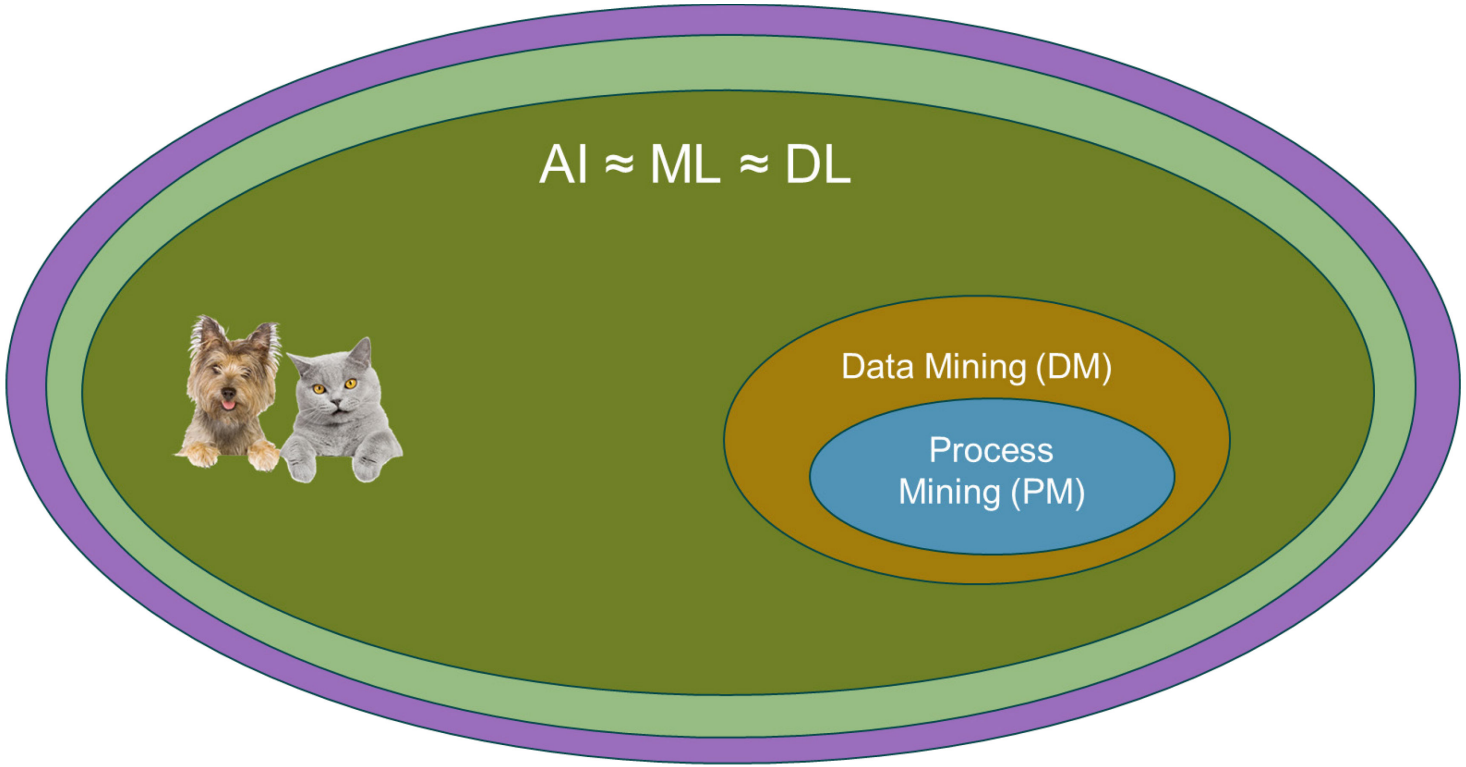
Example: Scoping machine learning

- Sometimes machine learning is used as a synonym for deep neural networks and sometimes to cover the entire spectrum of learning techniques (including mining, etc.).
- The fact that a neural network can be used as a classifier does not imply that the numerous classification techniques developed in data mining are part of machine learning (in the narrow sense).

A more balanced view



Perception by outsiders



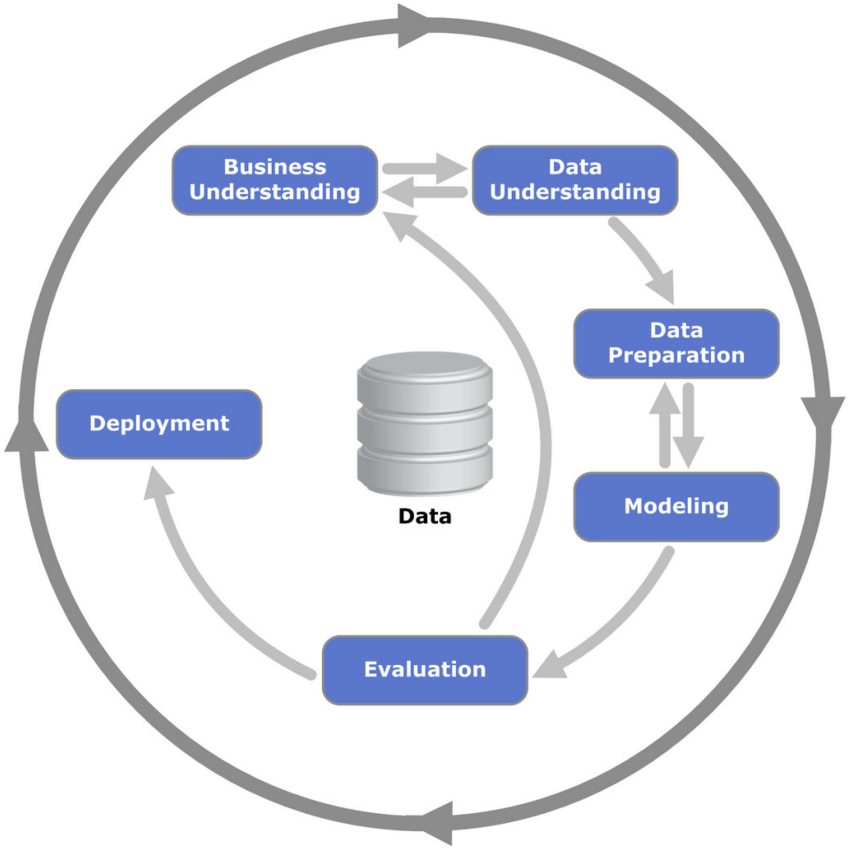
Chair of Process and Data Science

Data science process

CRISP-DM

Cross-industry standard process for data mining

- **CRISP-DM was developed in the late 1990-ties involving SPSS, Teradata, Daimler AG, NCR Corporation and Ohra.**
- **Quite obvious**



Chair of Process and Data Science

CRISP-DM

Cross-industry standard process for data mining

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

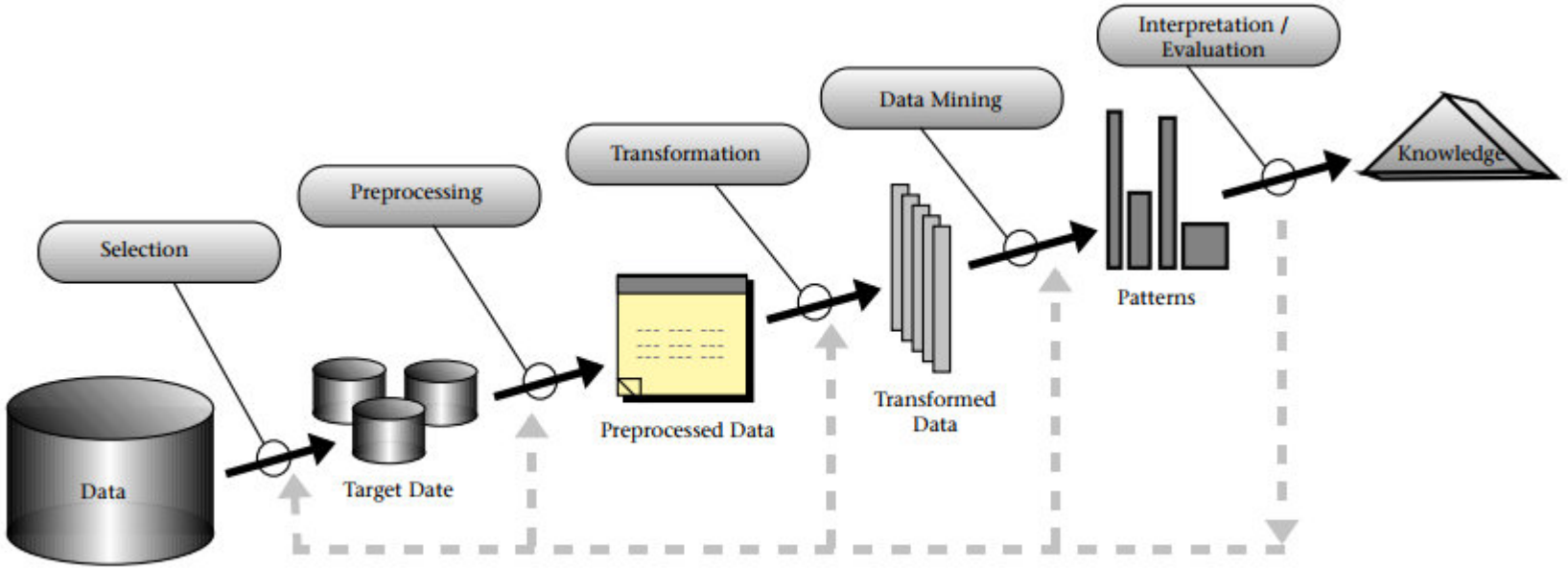
Review Project
Experience Documentation

The term “modeling” is sometimes a bit misleading: selection and assumptions (human) and automated learning by tool/algorithm are combined.

Taken from Pete Chapman (1999) The CRISP-DM User Guide.

Chair of Process and Data Science

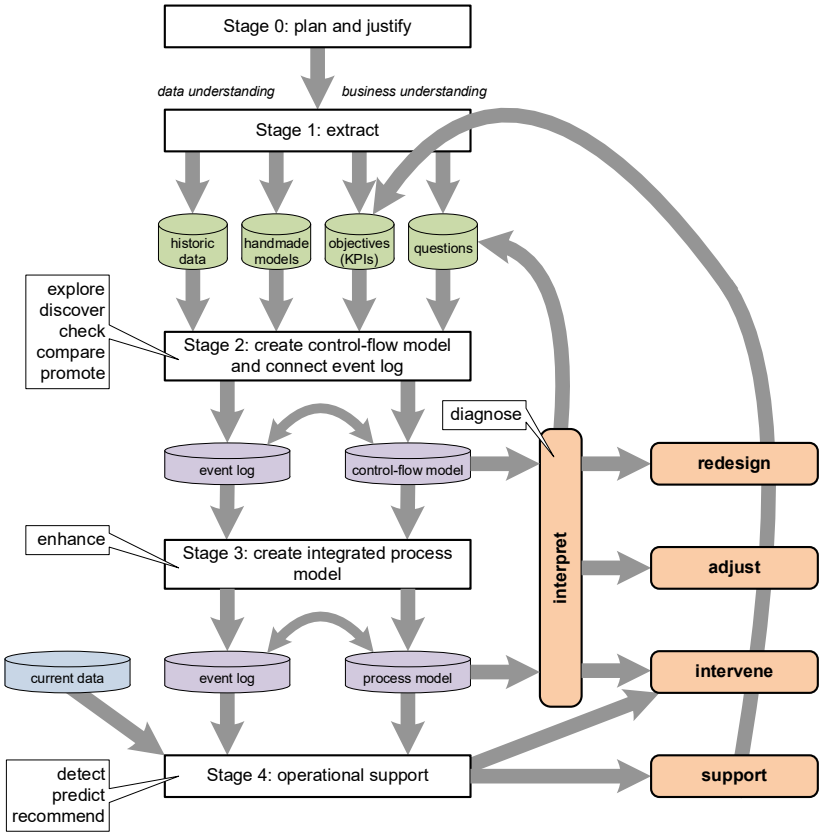
KDD (Knowledge Discovery in Databases) Process (by Fayyad, Piatetsky-Shapiro, and Smyth)



SEMMA (with the phases Sample, Explore, Modify, Model, and Assess) is another process model developed by SAS Institute.

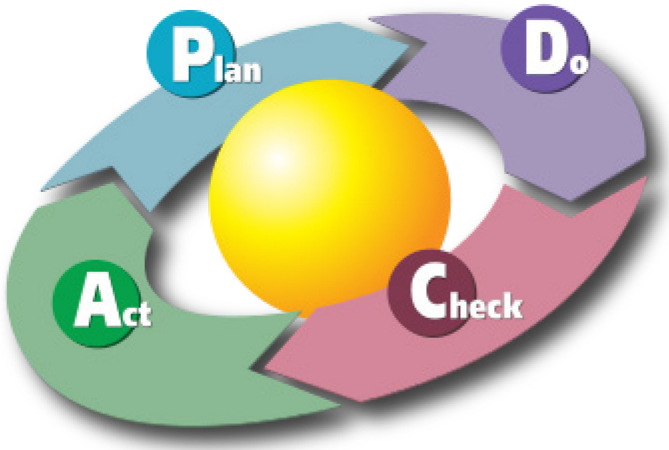
Chair of Process and Data Science

L* lifecycle model (specific for process mining)



Chair of Process and Data Science

Related to PDCA and DMAIC methodology

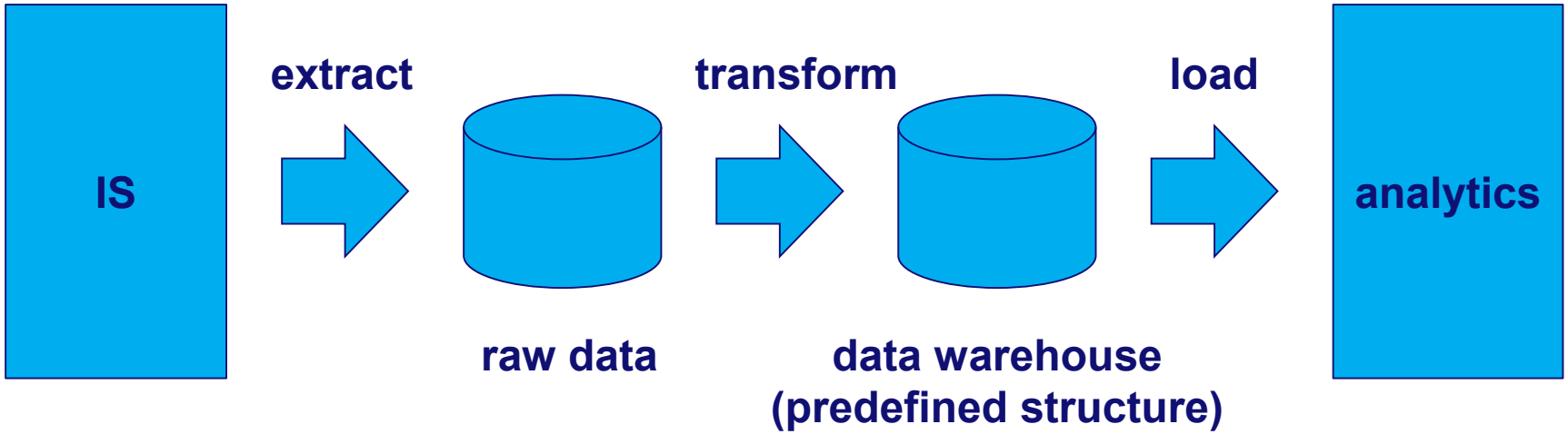


PDCA (Plan–Do–Check–Act) methodology by William Deming

DMAIC (Define, Measure, Analyze, Improve and Control) methodology used in Six Sigma projects

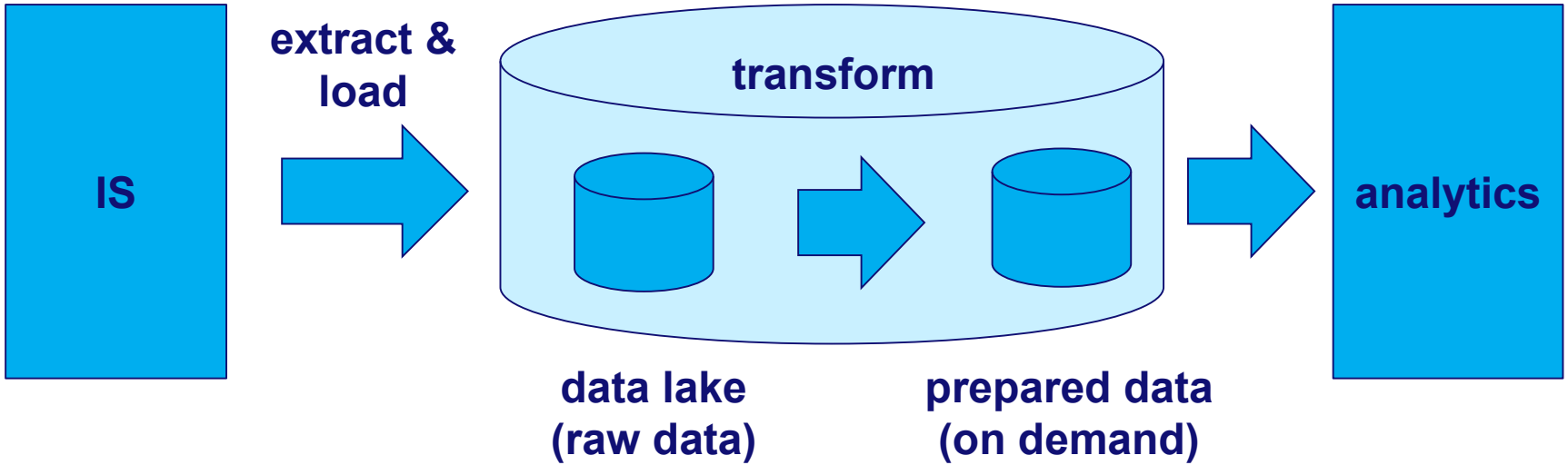
Chair of Process and Data Science

Extract, Transform, Load (ETL)



Chair of Process and Data Science

Extract, Load, Transform (ELT)



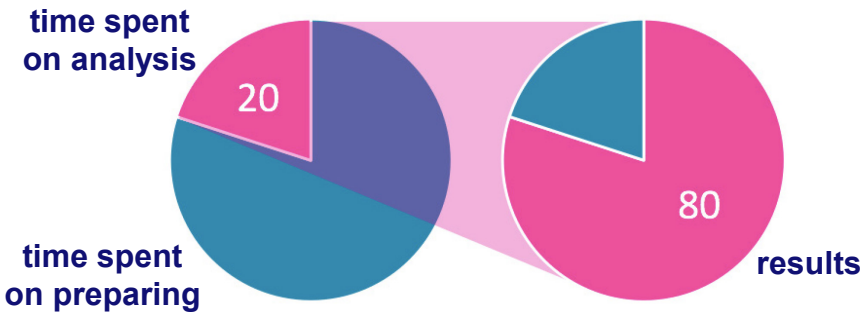
Chair of Process and Data Science


data lakes (ELT)


data warehouses
(ETL)

Another 80/20 rule

- 80 percent of a data scientist's time is spent on finding, cleaning, preprocessing, and organizing data, leaving only 20 percent to actually perform analysis.
- However, the 20 percent effort determines 80% of the results.



Chair of Process and Data Science

Challenges

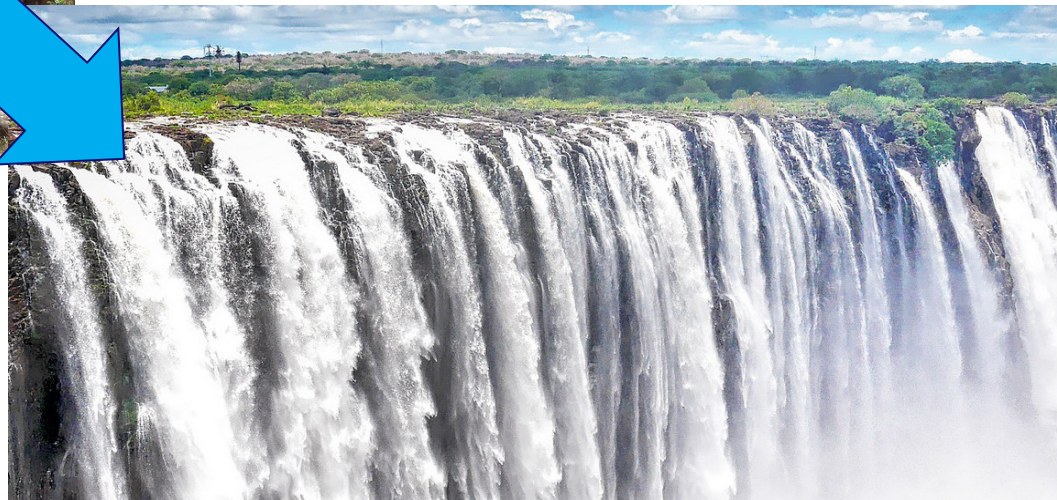
Finding data

- There may be hundreds or thousands of tables.
- There many be many different entities that are less relevant.

Transforming data

- **Reorganizing data.**
- **Extracting relevant features.**

Dealing with Big data



Dealing with streaming data

Data quality

- Data may be incomplete, invalid, inconsistent, imprecise, and/or outdated.
- Example timestamps:
  - Incomplete (missing event)
  - Invalid (14-14-2018)
  - Inconsistent (14-7-2018 => 7-14-2018)
  - Imprecise (2018-09-21'T'13.20.10)

Overfitting / Underfitting

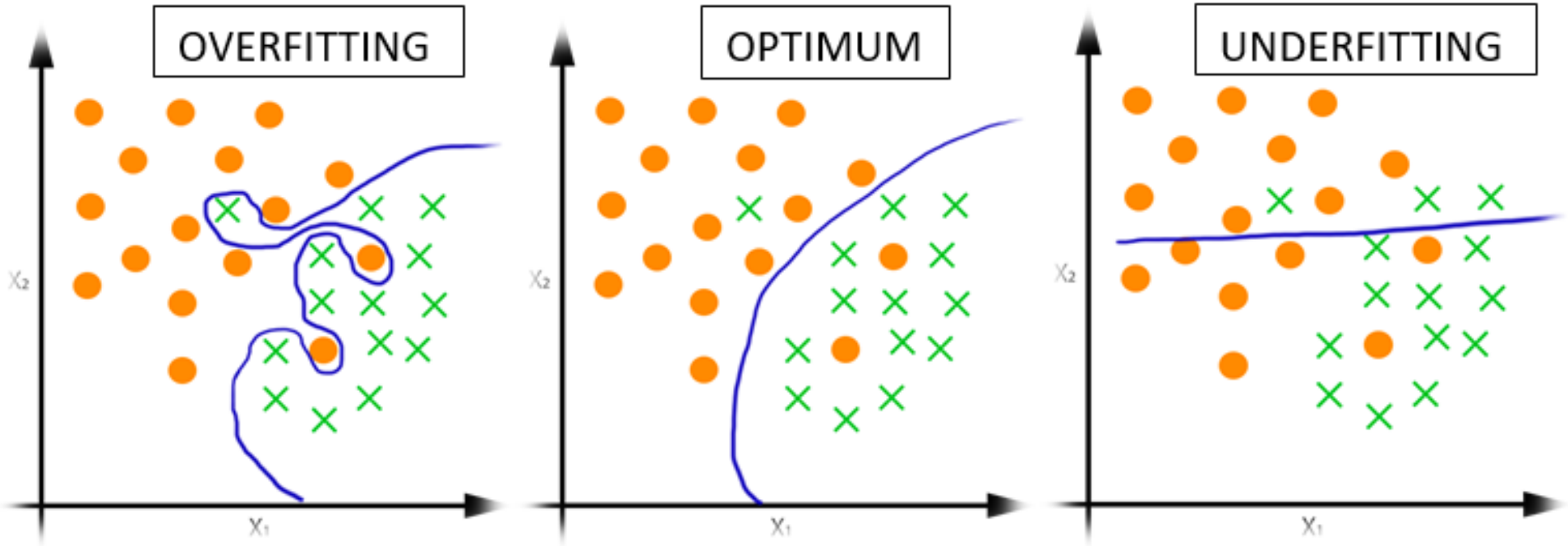
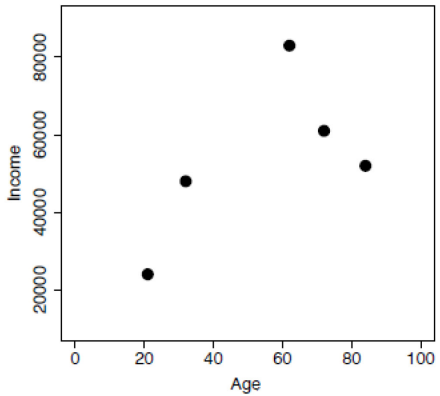
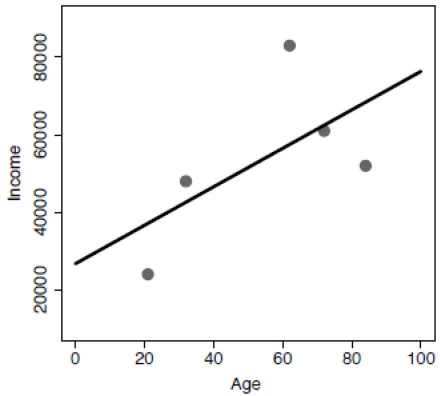


Diagram by Sachin Joglekar (Google).

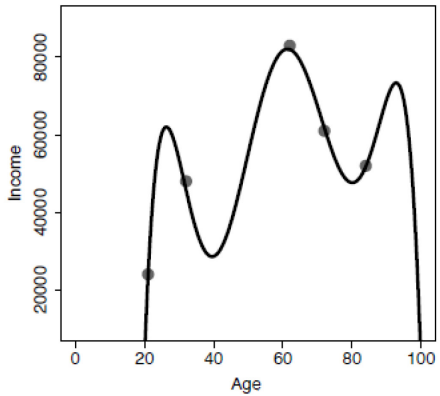
Overfitting / Underfitting



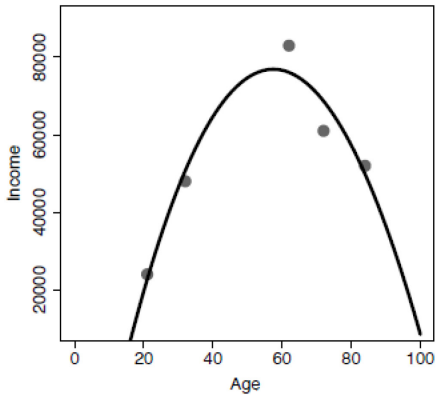
(a) Dataset



(b) Underfitting



(c) Overfitting



(d) Just right

Diagram taken from Fundamentals of Machine Learning for Predictive Data Analytics by J. Kelleher, B. Mac Namee and A. D'Arcy.

Dealing with concept drift

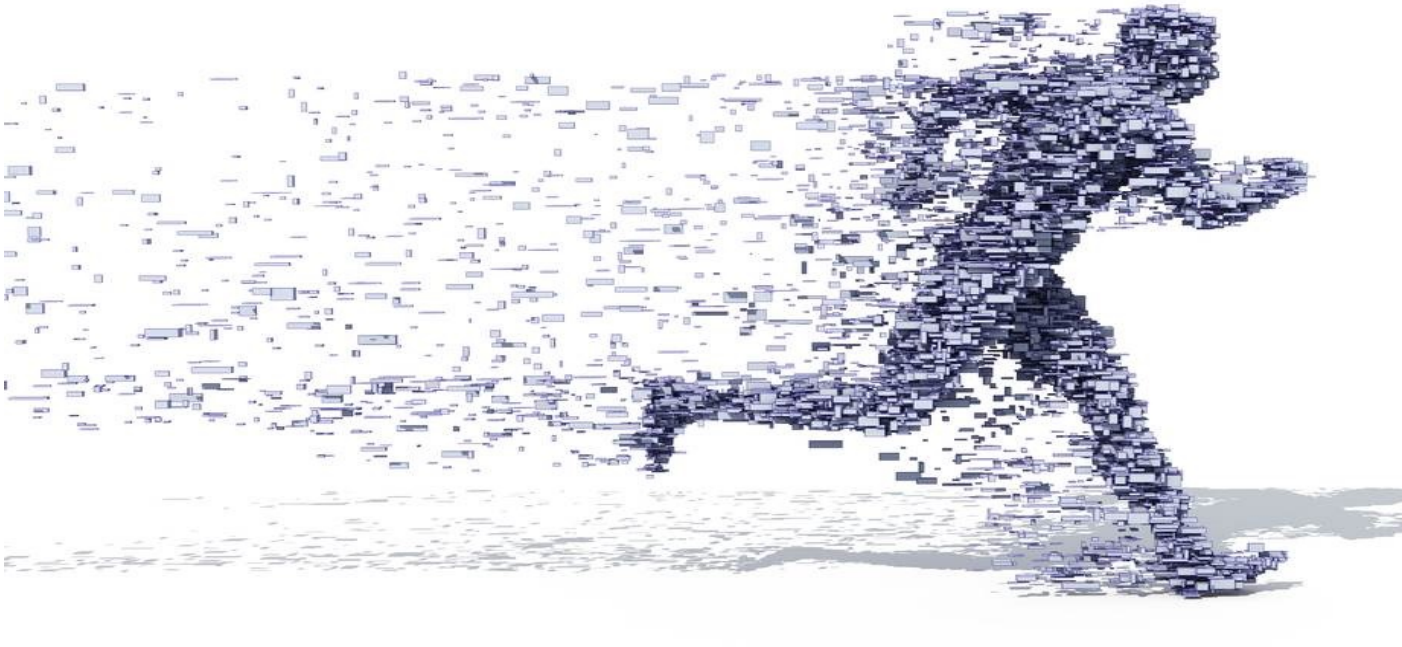
Chair of Process and Data Science

Making results actionable

- **Analysis results need to be relevant, specific, novel and clear.**

Ensuring fairness

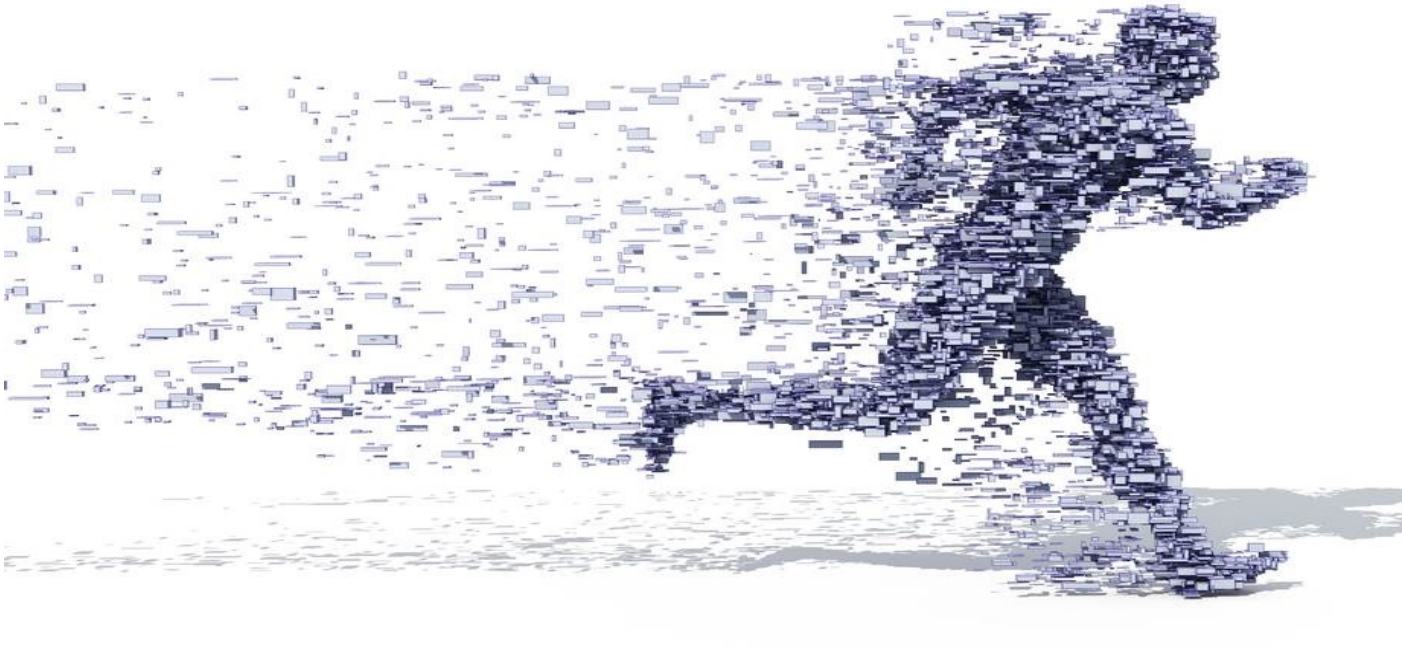
Fairness: Data Science without prejudice: How to avoid unfair conclusions even if they are true?



Chair of Process and Data Science

Ensuring accuracy

Accuracy: Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?



Chair of Process and Data Science

Ensuring confidentiality

Confidentiality:
Data Science that ensures confidentiality: How to answer questions without revealing secrets?

Chair of Process and Data Science

Ensuring transparency

Transparency: Data Science that provides transparency: How to clarify answers such that they become indisputable?

Chair of Process and Data Science

Ill-posed problems

- A problem is well-posed if
  - a solution exists and
  - the solution is unique.
- Problems in data science are often ill-posed:
  - there may be many possible models explaining observed phenomena,
  - the (training) data set is just a sample,
  - there may be noise (exceptional or incorrectly recorded instances) in the data set, and
  - the result needs to generalize to have predictive/ explanatory value.

Conclusion

Conclusion

- **Data science**
  - **has many faces,**
  - **is “super relevant”, and**
  - **provides many challenges.**
- **Next lecture:**
  - **Crash course in Python**

#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2	*Decision trees and data visualization/exploration*	*19/10/2018*	*Friday*
Lecture 5	Regression	24/10/2018	Wednesday
Lecture 6	Support vector machines	25/10/2018	Thursday
Instruction 3	*Regression and support vector machines*	*26/10/2018*	*Friday*
Lecture 7	Neural networks (1/2)	31/10/2018	Wednesday
Instruction 4	*Neural networks and supervised learning*	*02/11/2018*	*Friday*
Lecture 8	Neural networks (2/2)	07/11/2018	Wednesday
Lecture 9	Evaluation of supervised learning problems	08/11/2018	Thursday
Instruction 5	*Neural networks and supervised learning*	*09/11/2018*	*Friday*
Lecture 10	Clustering	14/11/2018	Wednesday
Lecture 11	Frequent items sets	15/11/2018	Thursday
Lecture 12	Association rules	21/11/2018	Wednesday
Lecture 13	Sequence mining	22/11/2018	Thursday
Instruction 6	*Clustering, frequent items sets, association rules*	*23/11/2018*	*Friday*
Lecture 14	Process mining (unsupervised)	28/11/2018	Wednesday
Lecture 23	Big data (2/2)	17/01/2019	Thursday
Instruction 11	*Big data*	*18/01/2019*	*Friday*
Lecture 24	Closing	23/01/2019	Wednesday
backup		24/01/2019	Thursday
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*
backup		30/01/2019	Wednesday
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*

Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2	*Decision trees and data visualization/exploration*	*19/10/2018*	*Friday*