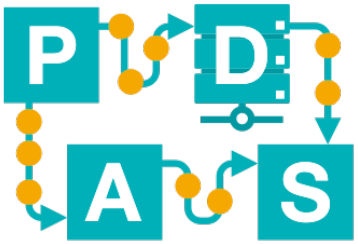
Introduction to Data Science

Lecture 1

IDS-L1



Chair of Process and Data Science

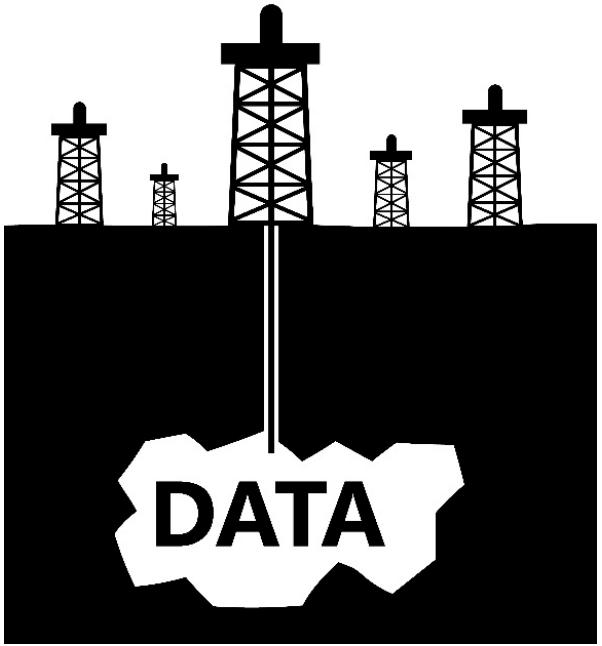
RWTH AACHEN UNIVERSITY

Outline of the lecture

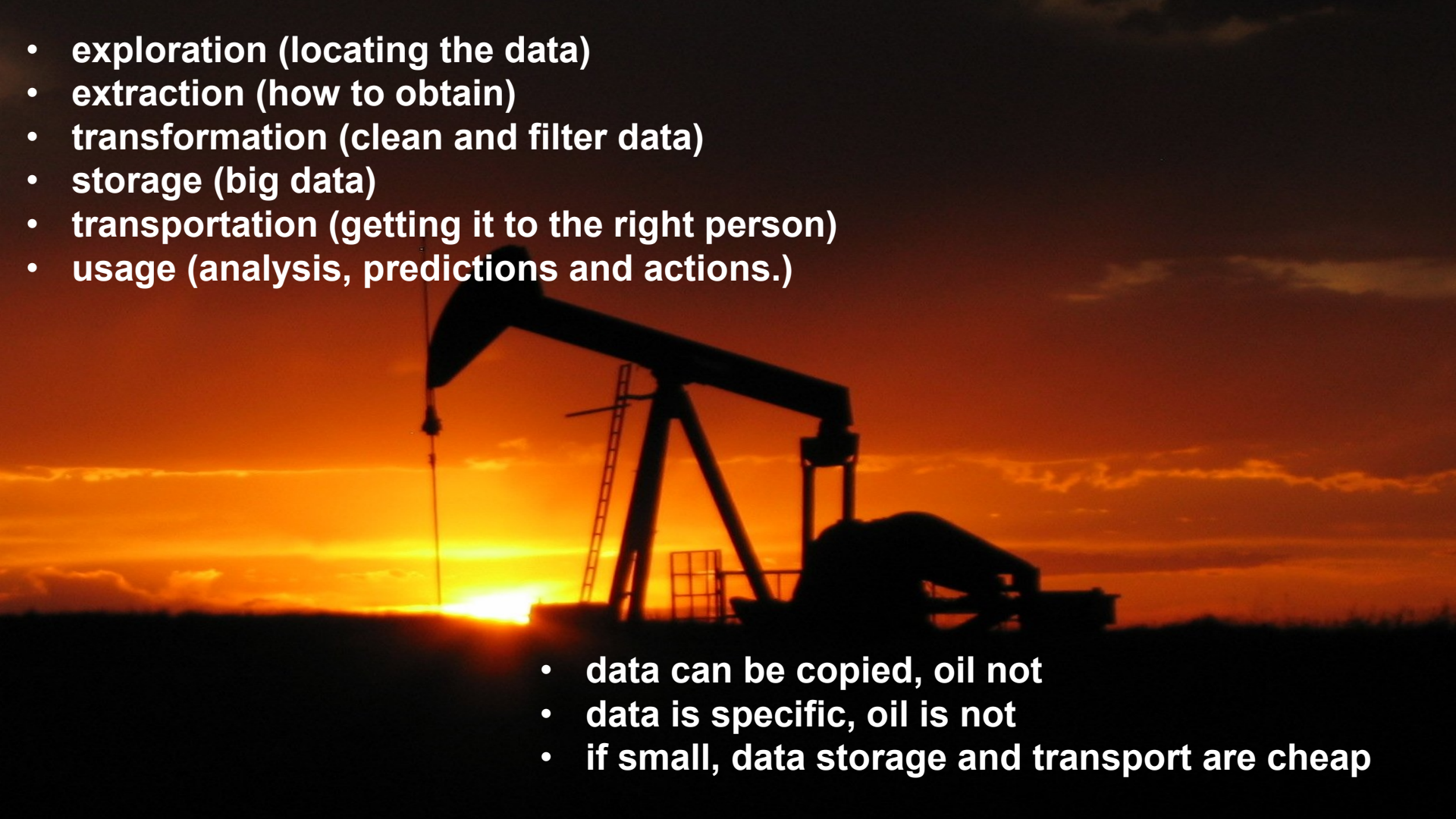
- Outline of the course
- Organization and exam regulations
- Motivation (Big data & skills)
- Data Science pipeline
- Data types
- Terminology
- Data science process
- Challenges

Outline of the 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.
- Allows to create new and develop existing products and services.



- exploration (locating the data)
- extraction (how to obtain)
- transformation (clean and filter data)
- storage (big data)
- transportation (getting it to the right person)
- usage (analysis, predictions and actions.)

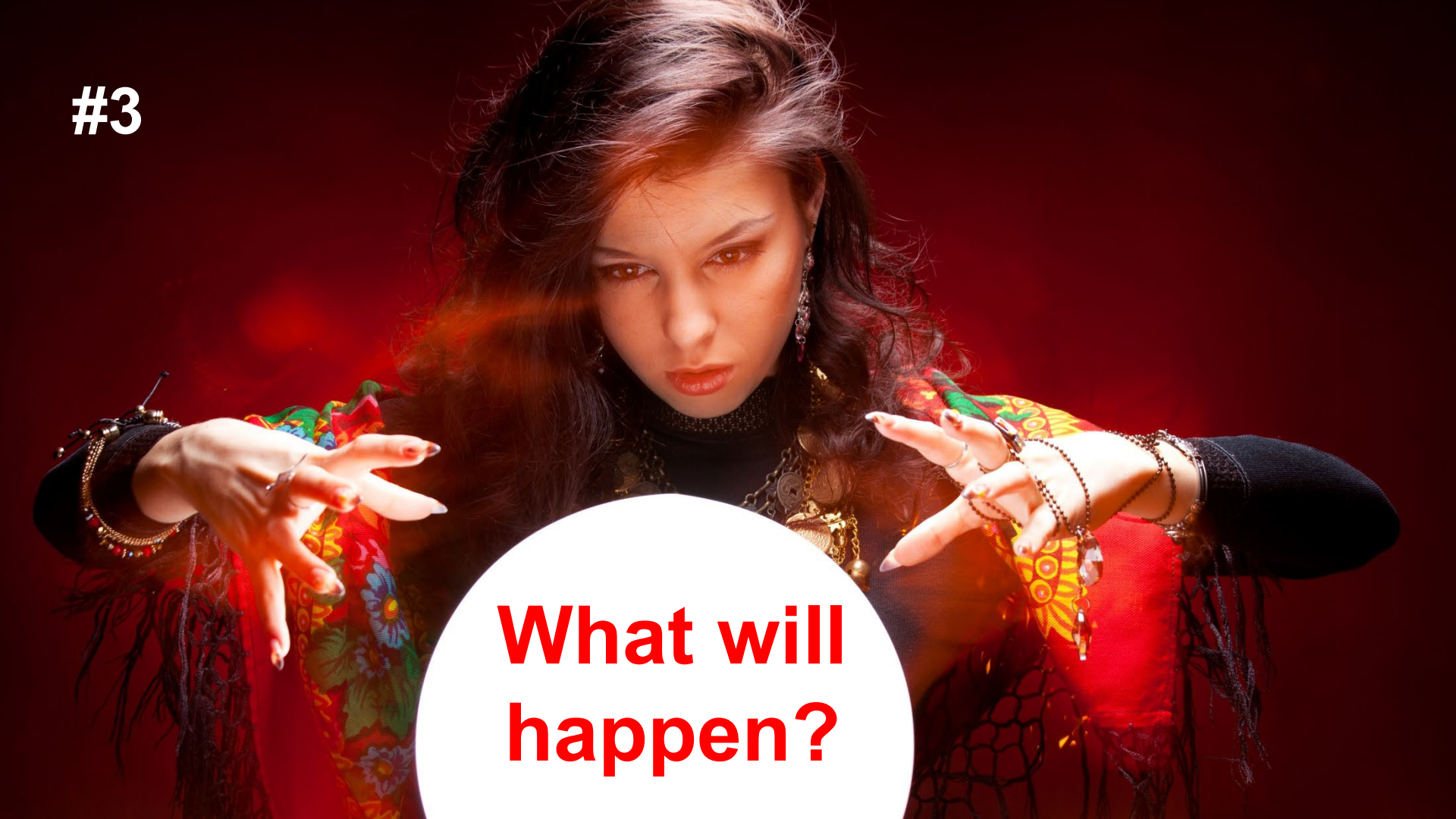
- data can be copied, oil not
- data is specific, oil is not
- if small, data storage and transport are cheap


Four generic
data science
questions

#1

#2

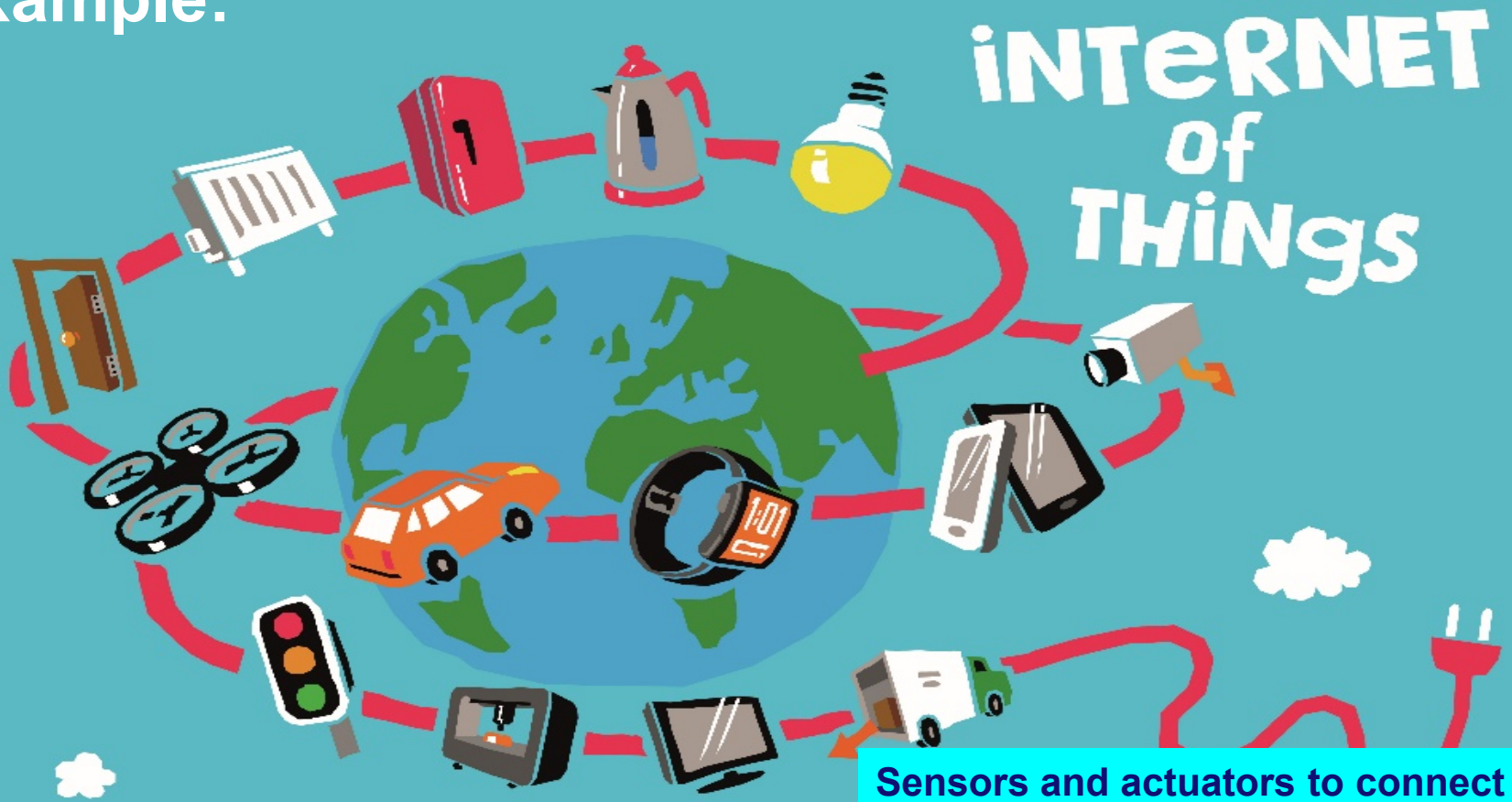
#3



#4

Why now? The Turing award winner Peter Naur used the term “data science” already in 1974

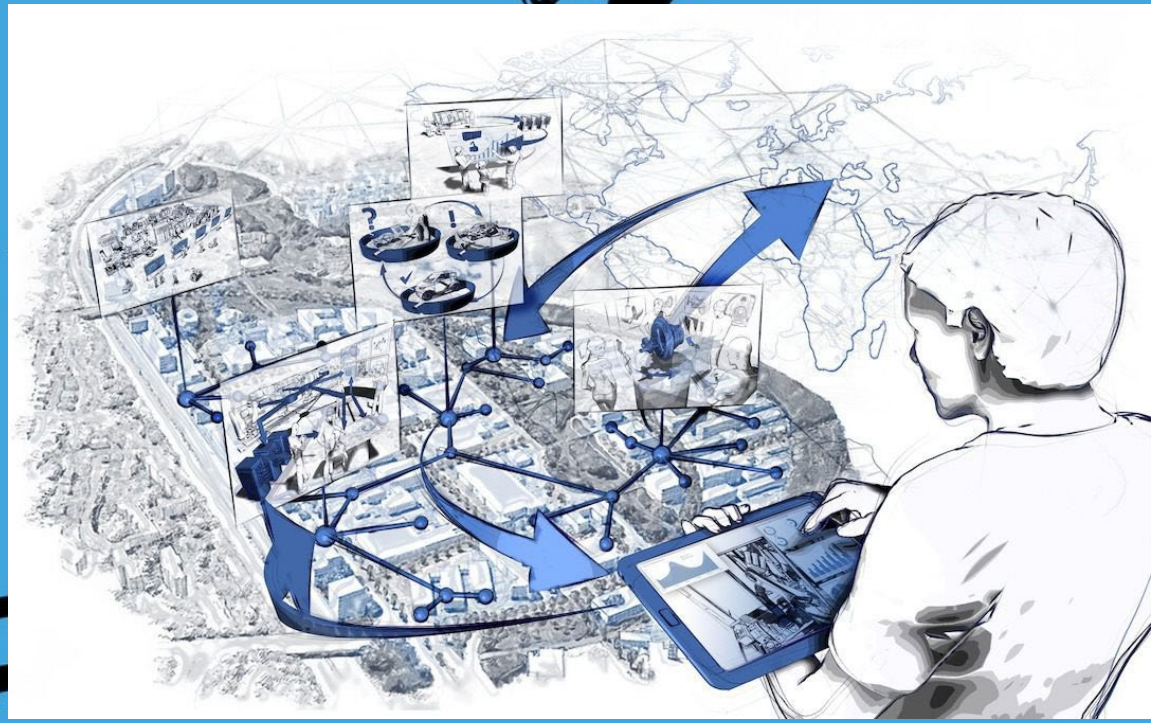
Example:



Sensors and actuators to connect the digital and physical world!

Our digital shadows are catching up!

LUCKY LUKE



PAULO FIGUEIREDO 2016

Dimensions

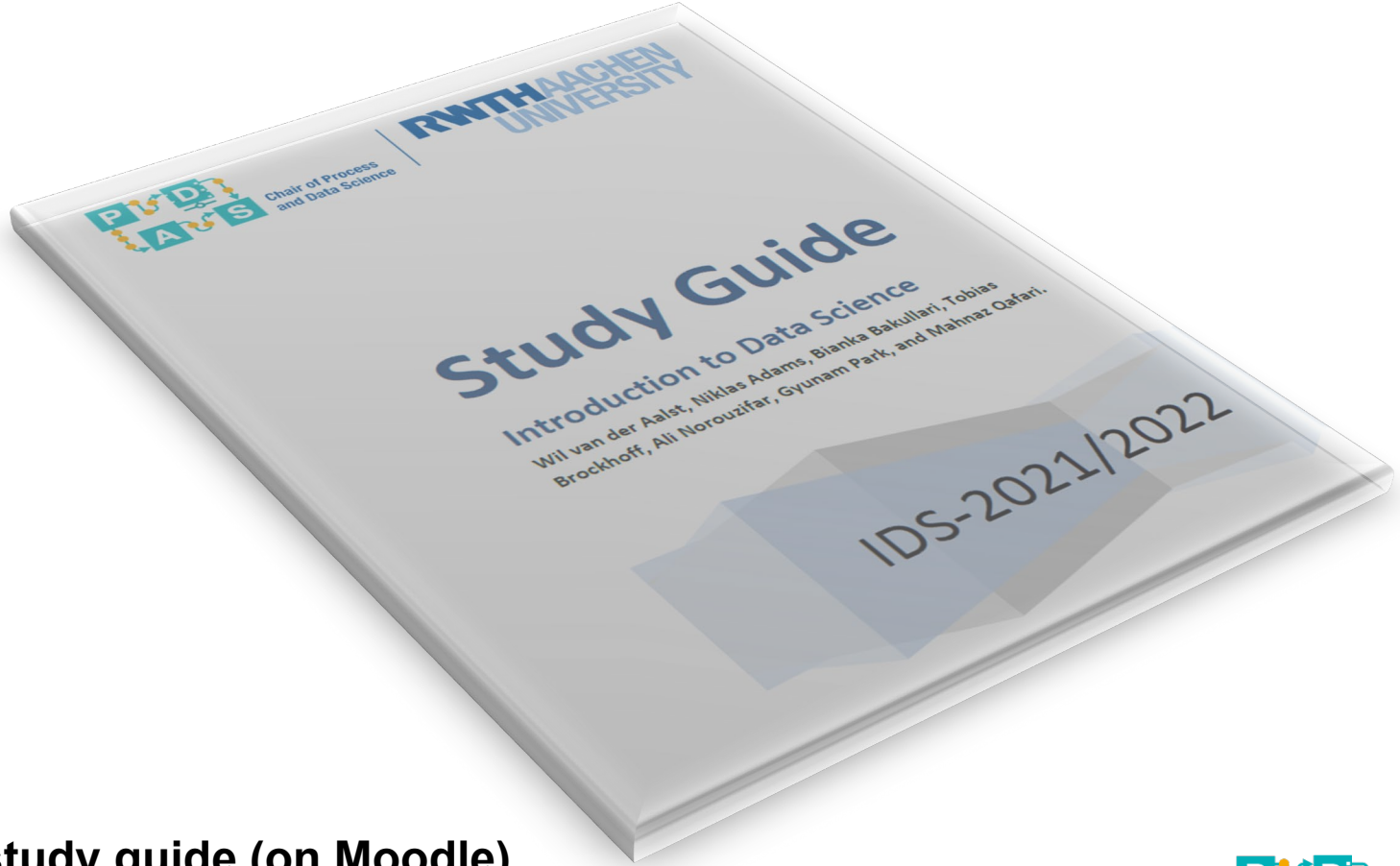
- Different types of data: structured, unstructured, text, images, events, etc.
- Different types of tasks: supervised, unsupervised.
- Human versus machine: who does what?
- Algorithm versus visualization.
- Flexibility versus usability: e.g., Python versus Celonis.
- Scalability versus quality.
- Responsibility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

Topics covered in lectures and instructions

	Lecture	date	day
Lecture 1	Introduction	13/10/2021	Wednesday
Instruction 1	Python	14/10/2021	Thursday
Instruction 2	Crash Course in Python	15/10/2021	Friday
Lecture 2	Basic data visualization/exploration	20/10/2021	*Wednesday*
Lecture 3	Decision trees	21/10/2021	Thursday
Instruction 3	Decision trees and data visualization/exploration	22/10/2021	Friday
Lecture 4	Regression	27/10/2021	Wednesday
Lecture 5	Support vector machines	28/10/2021	Thursday
Instruction 4	Regression and support vector machines	29/10/2021	Friday
Lecture 6	Neural networks (1/2)	03/11/2021	Wednesday
Lecture 7	Neural networks (2/2)	04/11/2021	Thursday
Lecture 8	Evaluation of supervised learning problems	10/11/2021	Wednesday
Instruction 5	Neural networks	11/11/2021	Thursday
Instruction 6	Neural networks and evaluation	12/11/2021	Friday
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets	18/11/2021	Thursday
Instruction 7	Clustering and frequent item sets	19/11/2021	Friday
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining	25/11/2021	Thursday
Instruction 8	Association rules and sequence mining	26/11/2021	Friday
Lecture 13	Process mining (unsupervised)	01/12/2021	Wednesday
Lecture 14	Process mining (supervised)	02/12/2021	Thursday
Instruction 9	Process Mining	03/12/2021	Friday
Lecture 15	Text Mining (1/2)	08/12/2021	Wednesday
Lecture 16	Text Mining (2/2)	09/12/2021	Thursday
Instruction 10	Q&A Assignment 1	10/12/2021	Friday
Lecture 17	Data preprocessing, data quality, binning, etc.	15/12/2021	Wednesday
Lecture 18	Visual analytics & information visualization	16/12/2021	Thursday
Instruction 11	Text Mining	17/12/2021	Friday
Lecture 19	Responsible data science (1/2)	22/12/2021	Wednesday
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data	12/01/2022	Wednesday
Instruction 12	Preprocessing and visualization	13/01/2022	Thursday
Instruction 13	Q&A Assignment 2	14/01/2022	Friday
Lecture 22	Closing	19/01/2022	Wednesday
Instruction 14	Big Data (1/2)	20/01/2022	Thursday
Instruction 15	Responsible data science	21/01/2022	Friday
Instruction 16	Big Data (2/2)	27/01/2022	Thursday
Instruction 17	Example Exam Questions	28/01/2022	Friday
Instruction 18	Questions	02/02/2022	Wednesday

- **Introduction**
- **Basic data exploration and visualization**
- **Python installation and demo**
- **Decision trees**
- **Regression**
- **Support vector machines**
- **Neural networks**
- **Evaluation of supervised learning problems**
- **Clustering**
- **Frequent items sets**
- **Association rules**
- **Sequence mining**
- **Process mining**
- **Text mining**
- **Data preprocessing, data quality and binning.**
- **Visual analytics and information visualization**
- **Responsible data science**
- **Big data technologies**

Organization and exam regulations



Read your study guide (on Moodle)

Chair of Process and Data Science

People involved ids@pads.rwth-aachen.de

1. Wil van der Aalst
2. Niklas Adams
3. Bianka Bakullari
4. Tobias Brockhoff
5. Ali Norouzifar
6. Gyunam Park
7. Mahnaz Qafari

Chair of Process and Data Science

A few words about Wil van der Aalst

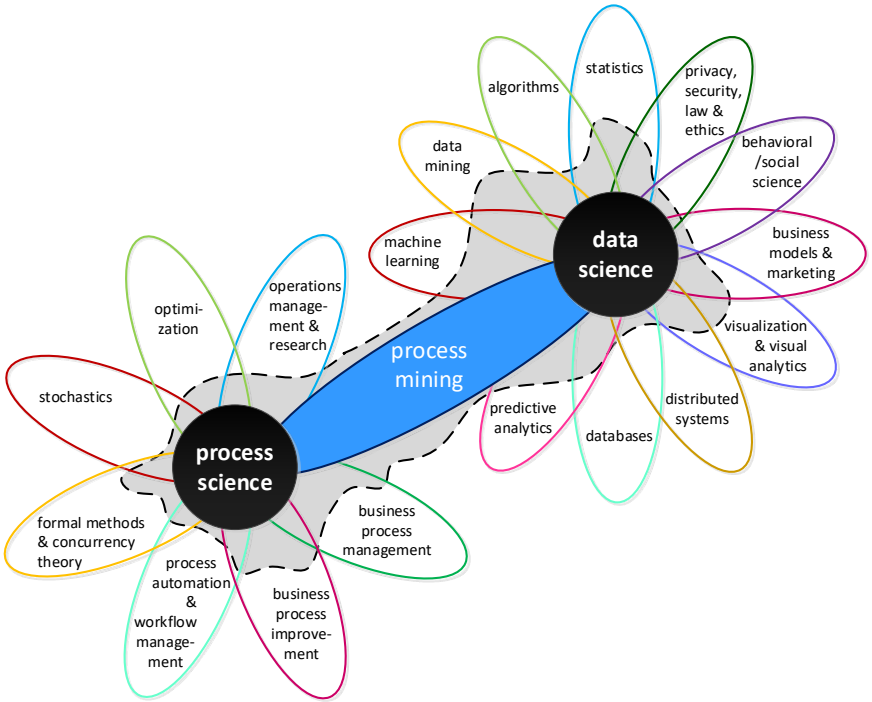
With Anja Karliczek, German Minister of Education and Research

Chair of Process and Data Science

A few words about PADS

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY



- **Focus of PADS is on the interplay between process science and data science.**
- **The leading process mining group in the world.**
- **Rapidly growing interest in industry (35+ software vendors based on process mining research done before).**
- **Variety of scientific challenges and a huge practical relevance.**
- **Where you can make a difference (HiWi, BSc, MSc and PhD positions).**
- **IDS has a broader focus, but also note the process mining lectures.**

Chair of Process and Data Science

Lectures & Instructions

- **Note that the course is held mostly online!**
- **Lectures are on Wednesdays and Thursdays from 8:30 to 10:00 and 08:30 to 10:00 (prerecorded).**
- **Instructions are on Fridays from 8.30 to 10.00 (live/online). Q&A parts of instructions are not recorded!**

	Lecture	date	day
Lecture 1	Introduction	13/10/2021	Wednesday
Instruction 1	Python	14/10/2021	Thursday
Instruction 2	Crash Course in Python	15/10/2021	Friday
Lecture 2	Basic data visualization/exploration	20/10/2021	*Wednesday*
Lecture 3	Decision trees	21/10/2021	Thursday
Instruction 3	Decision trees and data visualization/exploration	22/10/2021	Friday
Lecture 4	Regression	27/10/2021	Wednesday
Lecture 5	Support vector machines	28/10/2021	Thursday
Instruction 4	Regression and support vector machines	29/10/2021	Friday
Lecture 6	Neural networks (1/2)	03/11/2021	Wednesday
Lecture 7	Neural networks (2/2)	04/11/2021	Thursday
Lecture 8	Evaluation of supervised learning problems	10/11/2021	Wednesday
Instruction 5	Neural networks	11/11/2021	Thursday
Instruction 6	Neural networks and evaluation	12/11/2021	Friday
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets	18/11/2021	Thursday
Instruction 7	Clustering and frequent item sets	19/11/2021	Friday
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining	25/11/2021	Thursday
Instruction 8	Association rules and sequence mining	26/11/2021	Friday
Lecture 13	Process mining (unsupervised)	01/12/2021	Wednesday
Lecture 14	Process mining (supervised)	02/12/2021	Thursday
Instruction 9	Process Mining	03/12/2021	Friday
Lecture 15	Text Mining (1/2)	08/12/2021	Wednesday
Lecture 16	Text Mining (2/2)	09/12/2021	Thursday
Instruction 10	Q&A Assignment 1	10/12/2021	Friday
Lecture 17	Data preprocessing, data quality, binning, etc.	15/12/2021	Wednesday
Lecture 18	Visual analytics & information visualization	16/12/2021	Thursday
Instruction 11	Text Mining	17/12/2021	Friday
Lecture 19	Responsible data science (1/2)	22/12/2021	Wednesday
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data	12/01/2022	Wednesday
Instruction 12	Preprocessing and visualization	13/01/2022	Thursday
Instruction 13	Q&A Assignment 2	14/01/2022	Friday
Lecture 22	Closing	19/01/2022	Wednesday
Instruction 14	Big Data (1/2)	20/01/2022	Thursday
Instruction 15	Responsible data science	21/01/2022	Friday
Instruction 16	Big Data (2/2)	27/01/2022	Thursday
Instruction 17	Example Exam Questions	28/01/2022	Friday
Instruction 18	Questions	02/02/2022	Wednesday

Lecture Videos

The lectures are available via RWTH Moodle. If there are problems, you can alternatively watch the videos via:

- **Video AG**
  https://video.fsmpi.rwth-aachen.de/20ws-ids (URL new videos will be shared later)
- **YouTube**
  https://youtube.com/playlist?list=PLG_1ZxIPXO0vTTfheRNDhq4vNYAdJBUyC &
  https://youtube.com/playlist?list=PLG_1ZxIPXO0vReKHuzL-n--f4iO2JljcJ

Note that the lectures were mostly recorded last year. The content did not change, just ignore the slides with information about dates. There are face-to-face sessions to provide additional explanations and extra possibilities to interact.

Extra service: Face-to-Face Sessions

Additionally, throughout the semester around half of the lecture slots will be used to offer "face-to-face sessions" in presence. These are planned as face-to-face discussions with prof. van der Aalst where you can ask questions covering the topics handled up until that point. These lectures will start with a short summary. At the end of this study guide there is a provisional schedule. Please note that these face-to-face sessions will not be recorded and no material will be uploaded to Moodle. This is an extra service to ask questions and to get a summary of the material to keep a good overview. You should use these face-to-face sessions as a chance to ask questions and get to know us and each-other better. It is very important that you come prepared to the instructions and to the face-to-face discussions by having studied the lecture topics handled up until that point.

Please note that due to the uncertainty related to the pandemic, new regulations and rules appointed from the university and/or government may affect the lecture plan and schedule. Currently, the capacity of the lecture hall is limited to approx. 150 participants. Therefore, you need to register before via RWTH Moodle. We cannot guarantee that everybody will be able to attend the face-to-face sessions, but based on prior experiences with recorded lectures and such sessions there should not be a problem. Students that did not register will not be allowed in. We will also take note of students that register, but do not attend.

AH IV (2354|030) from 8.30-10.00 on Wednesday of Thursday

	Lecture	date	day
Lecture 1	Introduction		
Lecture 2	Basic data visualization/exploration	20/10/2021	*Wednesday*
Lecture 3	Decision trees		
Lecture 4	Regression	27/10/2021	Wednesday
Lecture 5	Support vector machines		
Lecture 6	Neural networks (1/2)		
Lecture 7	Neural networks (2/2)	10/11/2021	Wednesday
Lecture 8	Evaluation of supervised learning problems		
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets		
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining		
Lecture 13	Process mining (unsupervised)		
Lecture 14	Process mining (supervised)	02/12/2021	Thursday
Lecture 15	Text Mining (1/2)		
Lecture 16	Text Mining (2/2)	09/12/2021	Thursday
Lecture 17	Data preprocessing, data quality, binning, etc.		
Lecture 18	Visual analytics & information visualization	16/12/2021	Thursday
Lecture 19	Responsible data science (1/2)		
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data		
Lecture 22	Closing	19/01/2022	Wednesday

Extra service: Face-to-Face Sessions

Additionally, throughout the semester around half of the lecture slots will be used to offer "face-to-face sessions" in presence. These are planned as face-to-face discussions with prof. v… topics … with a … a provi… sessio… upload… and to … overview. You should use these face-to-face sessions as a chance to ask questions and get to know us and each-other better. It is very important that you come prepared to the instructions and to the face-to-face discussions by having studied the lecture topics handled up until that point.

You need to register before via RWTH Moodle!

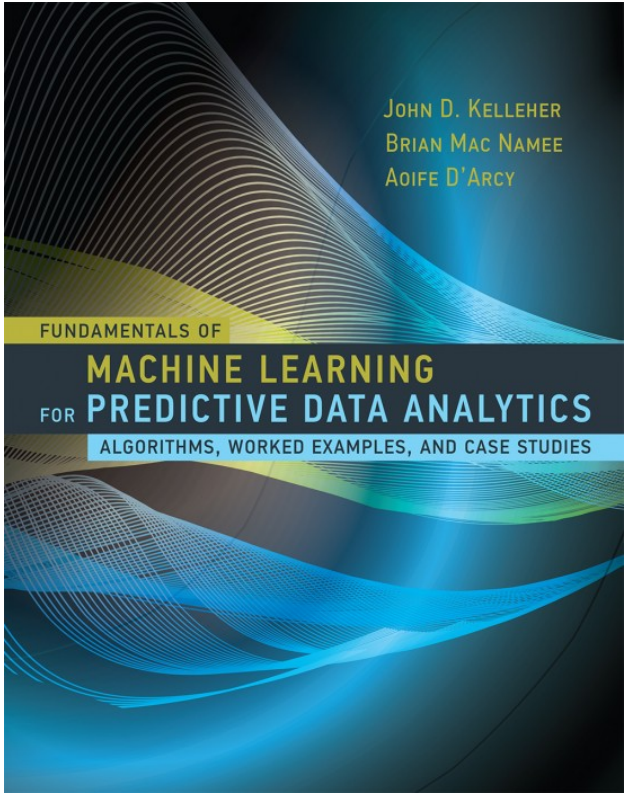
Please note that due to the uncertainty related to the pandemic, new regulations and rules appointed from the university and/or government may affect the lecture plan and schedule. Currently, the capacity of the lecture hall is limited to approx. 150 participants. Therefore, you need to register before via RWTH Moodle. We cannot guarantee that everybody will be able to attend the face-to-face sessions, but based on prior experiences with recorded lectures and such sessions there should not be a problem. Students that did not register will not be allowed in. We will also take note of students that register, but do not attend.

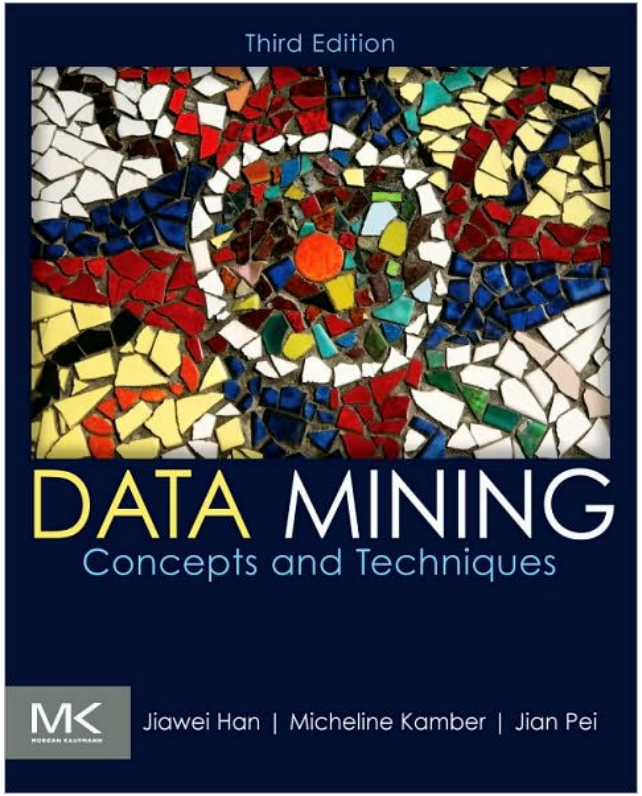
How well this works depends on you!

AH IV (2354|030) from 8.30-10.00 on Wednesday of Thursday

	Lecture	date	day
Lecture 1	Introduction		
	Basic data visualization/exploration	20/10/2021	*Wednesday*
	Decision trees		
	Regression	27/10/2021	Wednesday
	Support vector machines		
	Neural networks (1/2)		
	Neural networks (2/2)	10/11/2021	Wednesday
Lecture 8	Evaluation of supervised learning problems		
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets		
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining		
Lec…			
Lec…			Thursday
Lec…			
Lec…			Thursday
Lec…			
Lec…			Thursday
Lecture 19	Responsible data science (1/2)		
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data		
Lecture 22	Closing	19/01/2022	Wednesday

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 (You can also use the 2020 version!) (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 study material will be provided during the course.
- Given the nature of the course and the rapidly developing field of data science, there is no a single textbook.
- Distribution of slides and assignments via **RWTHmoodle**. (Register ASAP)

Software
(see the instructions)

- During the first instruction, we will demonstrate how to install software. You can also find the guidance on the RWTH Moodle.
- For the course, you will need a working installation of Python with some common data science software packages, such as
    - numpy
    - scipy
    - pandas
    - matplotlib
    - scikit-learn
    - nltk
- Detailed information is given in the first two instructions.
- In this course, we use a single, unified virtual environment for python programming. It is highly recommended for you to practice and work on your assignment with the virtual environment. Follow instructions to avoid installation issues.

Examinations and assignments

- The exam consists of two parts:
  - an assignment consisting of two parts each counting for 20% of the final result,
  - and the final written test that counts for remaining 60% of the final result.
- Participation in the both parts of the assignment is required for taking the final test.
- The assignment is done in groups of 2-3 students.
- To pass the course, it is required to pass both the assignment and the written test. This implies that you should obtain a minimum score of 50% in the assignment and a minimum score of 50% in the final written test.
- Only the final test can be retaken in this semester (there will be one re-exam). Assignments can be redone only in the next academic year.

Examinations and assignments

- **Important dates are:**
  - Final written test (60%) Questions to test the theoretical knowledge of the algorithms and techniques learned:
    - First option (PT1): 9/2/2022 16.00-18.00 (tentative, check later)
    - Second option (PT2): 21/3/2022 16.00-18.00 (tentative, check later)
  - Assignment Part 1 (20%): analysis of the real-life and synthetic datasets using techniques and tools discussed in the course. This part is used to test an understanding of the material given in the lectures 1-8. The deadline is on Wednesday 14/12/2021 23:59 (strict).
  - Assignment Part 2 (20%): analysis of complex datasets using various data science techniques and interpreting the results. The focus is on the lectures 9-21. The deadline is on Friday 19/01/2022 23:59 (strict).

Plagiarism (no excuses)

- We systematically check for plagiarism.
- <u>All group members</u> are responsible to submit an individual piece of work and avoid unfair academic practice.
- In case of a proven plagiarism, all members fail the assignment.
- The case will be reported, and this may lead to your expulsion from the university.

Who can take the course?

- **We want you to join this course! But …**
- **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, Software Engineering, Computational Social Systems, etc.**
- **Other students are also invited to participate, but it is up to the management and rules of the corresponding programs to decide whether the course "counts".**
- **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.**
- **Please formulate questions explicitly (what we should do).**
- **We can add “viewers” on Moodle (e-mail ids@pads.rwth-aachen.de), but this is only a temporary solution.**

Q&A

- Use the Question & Answers feature on the RWTHmoodle to address questions about the course: content related or about your administrative issues.
- Assuming high number of students, please use this Q&A function instead of sending e-mails.
- In case of urgent personal questions, please use ids@pads.rwth-aachen.de.
- When having “administrative problems”: make sure you are asking the right person and propose an action what we could do (not just “I cannot register”).

Course certificates and exchange students

- If you are attending this course, but you do not need the credits for your course of studies, you can request a certificate of attendance.
  - Typical case: you are a PhD student
- If you are an exchange student and your program requires you to check back at your home university before the date of the final exam, we will take the necessary arrangements for a remote exam.
- In any of these two cases, please contact ids@pads.rwth-aachen.de.

Make sure to…

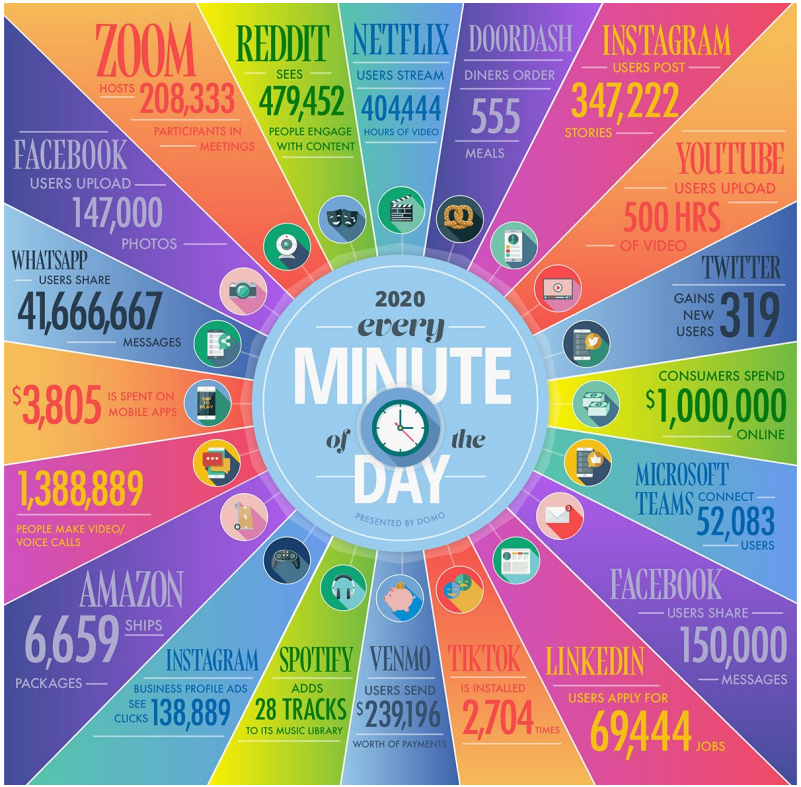
- **Register for the course and the final exam.**
- **Submit the assignment (its both parts) in time (German time ☺).**

Warning

- Slides are intended to be self-contained, but …
- May overlap with earlier or later courses because it is an introductory course and provides a broad overview.
- The diversity of the course will make it tricky, so stay synchronized.

Motivation
(Big data & skills)

Impact and size of data



Source: domo.com

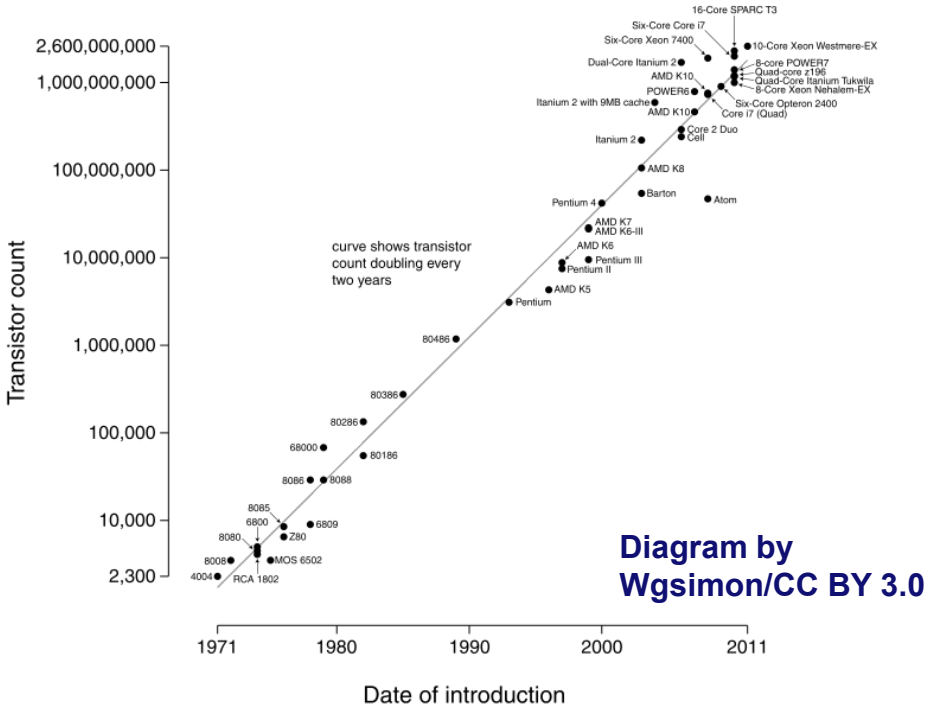
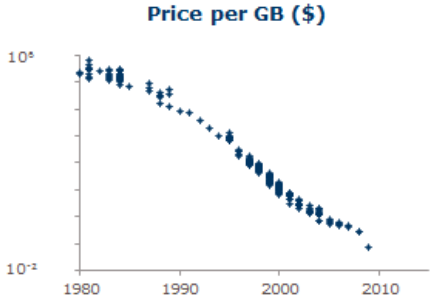
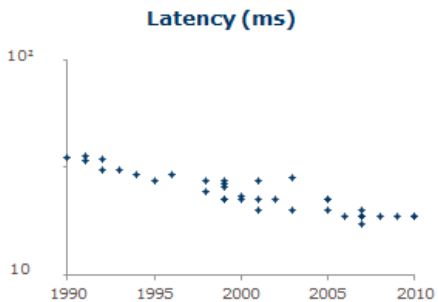
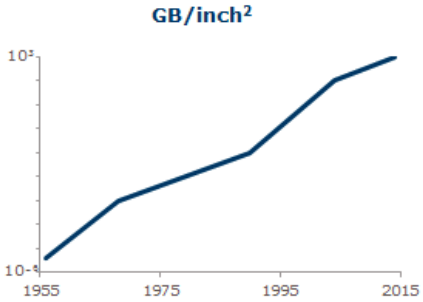
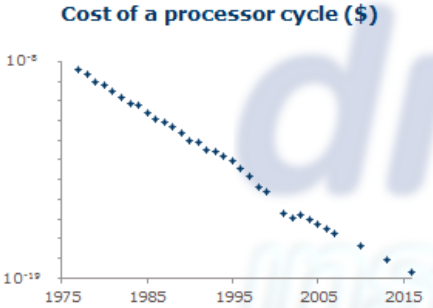
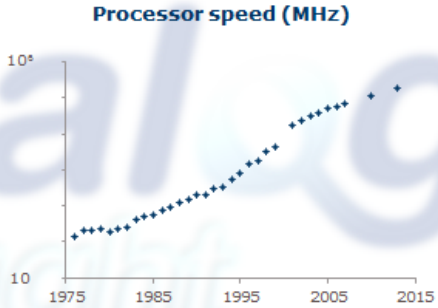
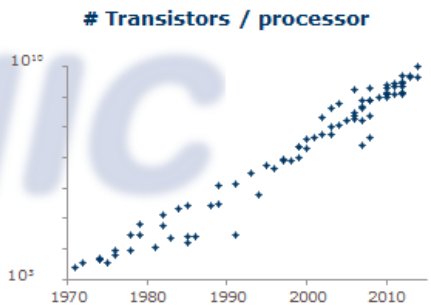
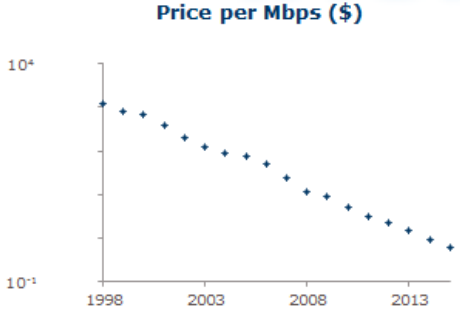
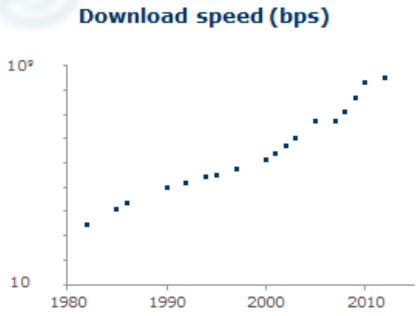
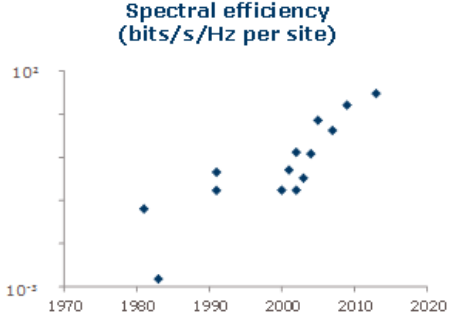


Diagram by Wgsimon/CC BY 3.0

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?

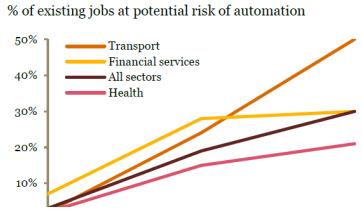
An international analysis of the potential long term impact of automation

pwc

Key findings: impact of automation

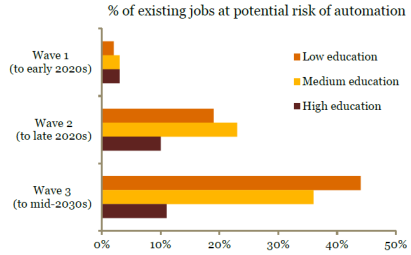
Financial services jobs could be relatively vulnerable to automation in the shorter term, while transport jobs are more vulnerable to automation in the longer term

Figure 1 – Potential job automation rates by industry across waves



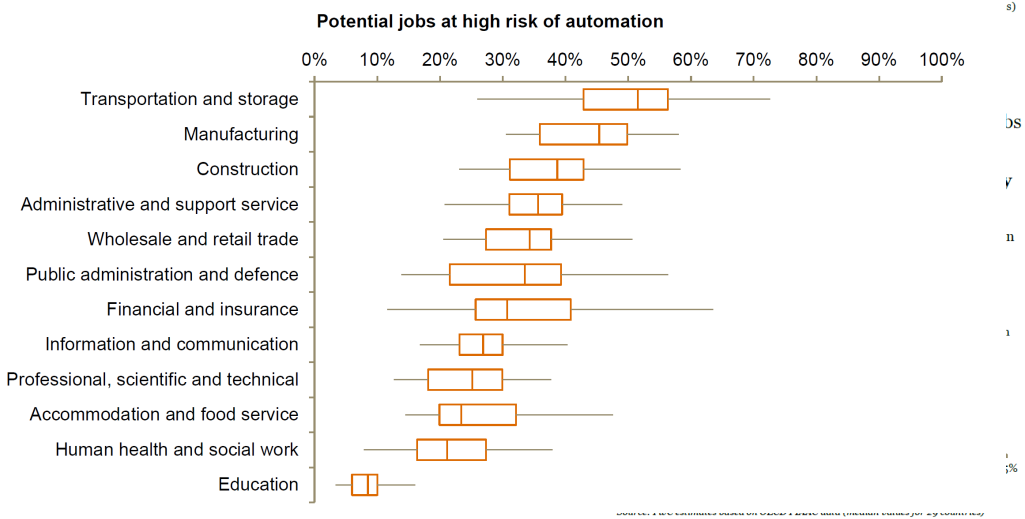
In the long run, less well educated workers could be particularly exposed to automation, emphasising the importance of increased investment in lifelong learning and retraining

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.

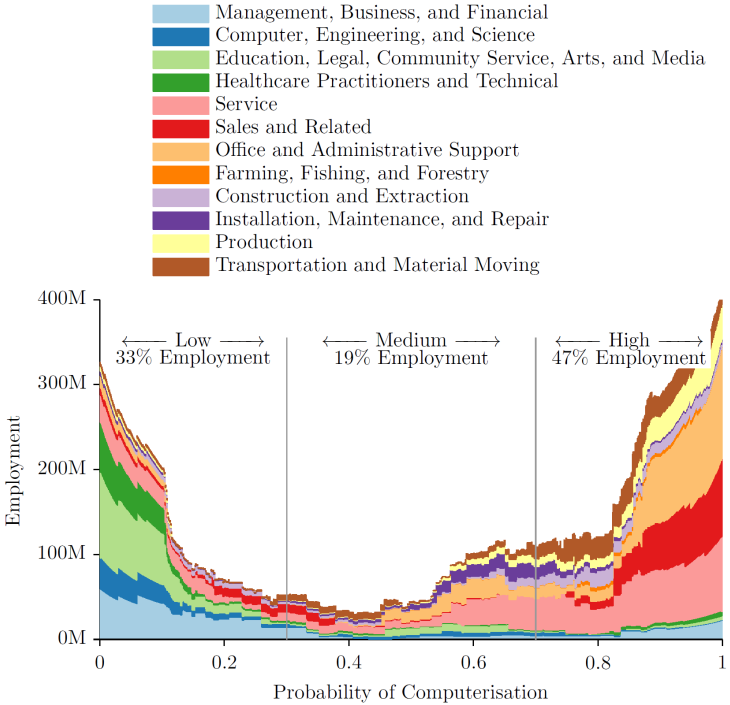
“According to our estimates around 47 percent of total us employment is in the high risk category. We refer to these as jobs at risk – i.e. jobs we expect could be automated relatively soon, perhaps over the next decade or two.”

WORKING PAPER

OXFORD MARTIN SCHOOL

UNIVERSITY OF OXFORD

Published by the Oxford Martin Programme on Technology and Employment

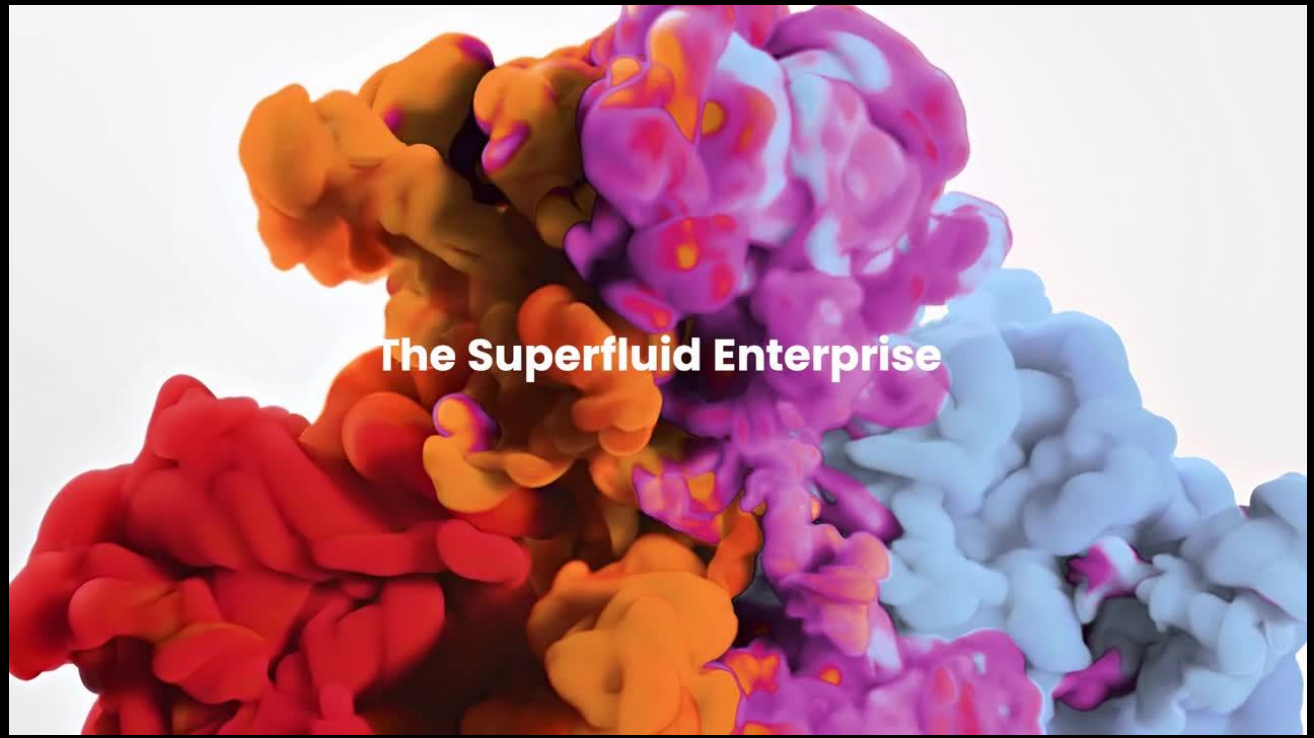


Occupations and probability of computerization (sample from 702 occupations):

- **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

Data used to remove all “friction” in processes



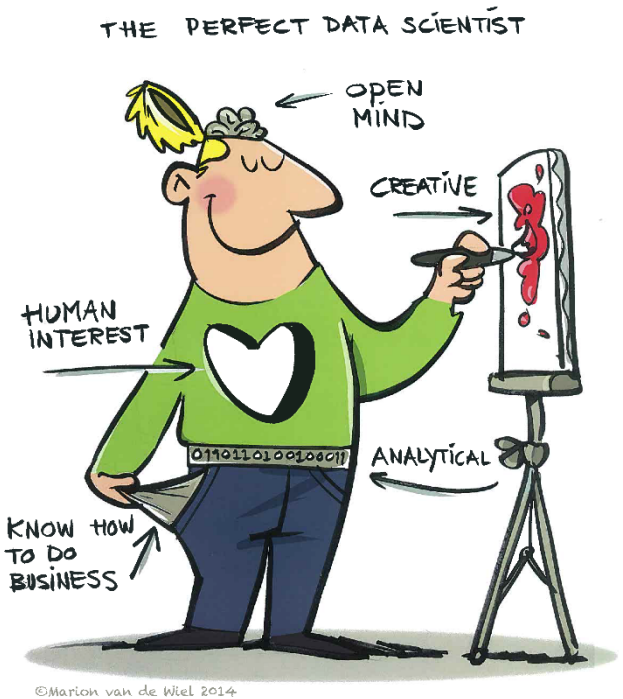
https://www.youtube.com/watch?v=_4OkKpAw1WM

Scope and impact of the growth of data are undeniable

data scientist

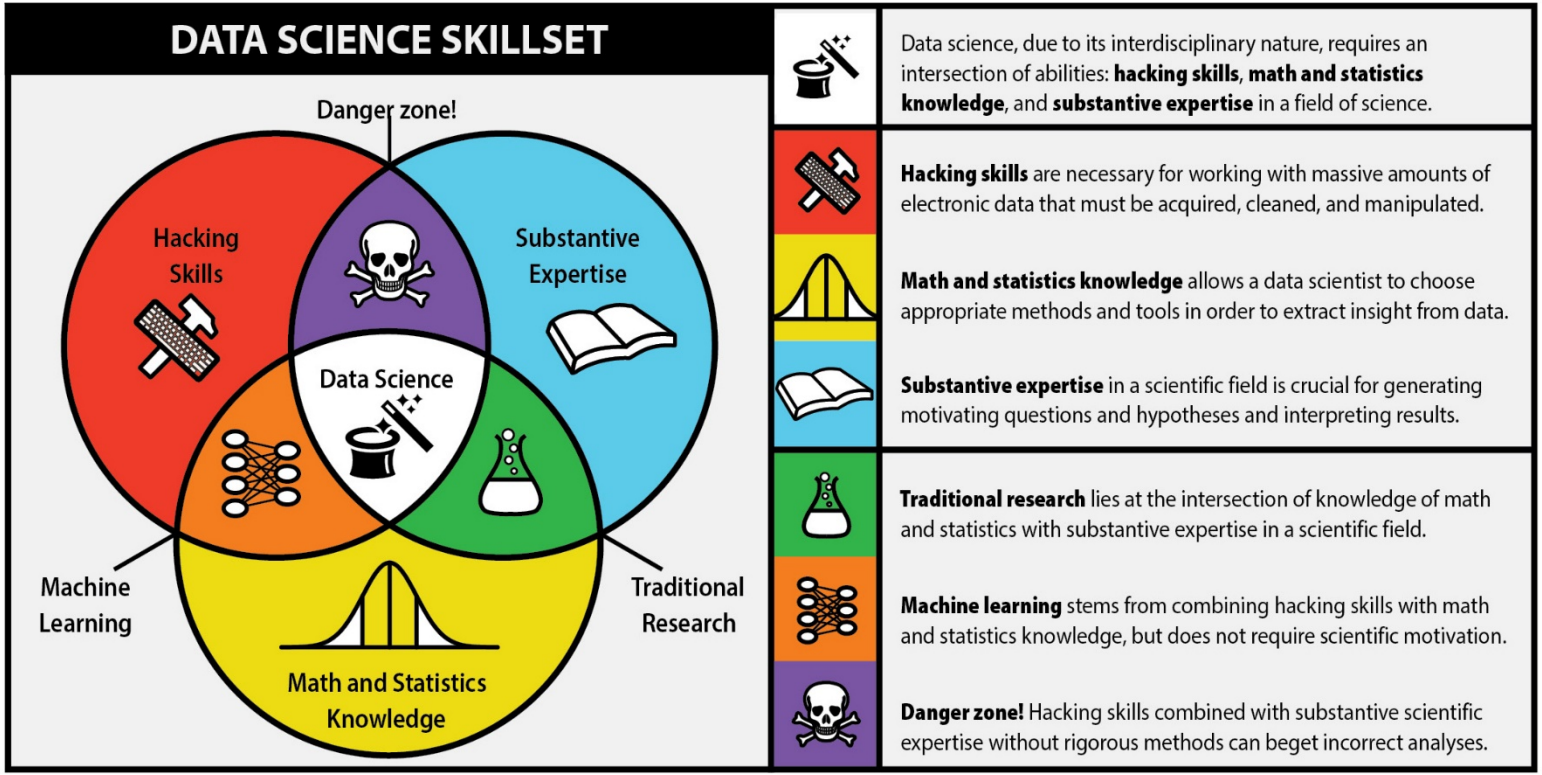
Chair of Process and Data Science

Data scientists need to combine different skills



Chair of Process and Data Science

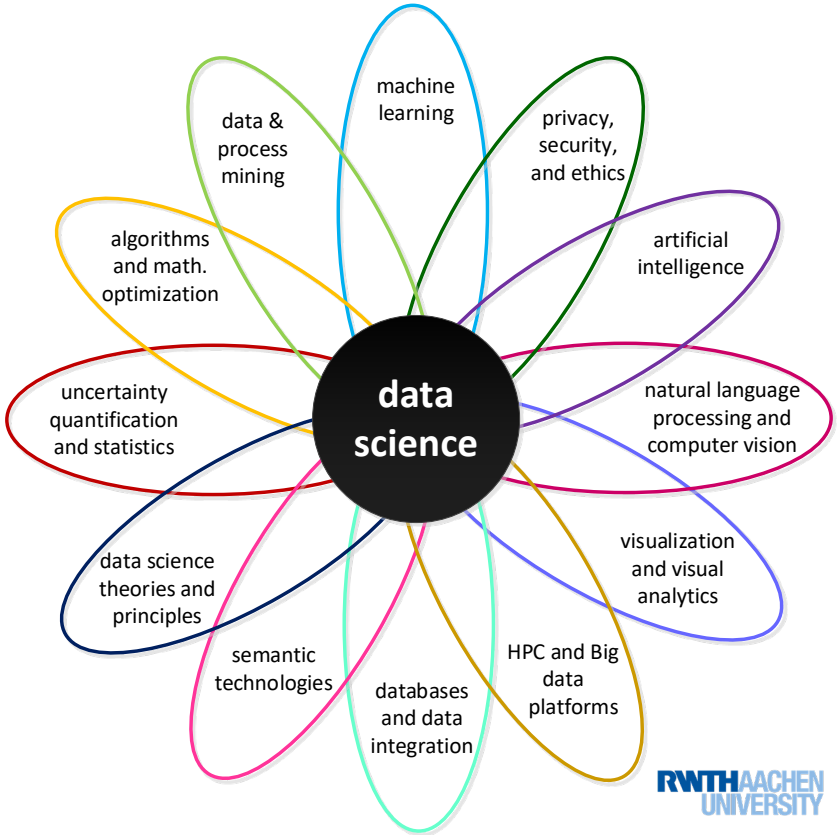
Another characterization



Picture by Natalia Bilenko, Drew Conway, et al.

Chair of Process and Data Science

The RWTH Data science flower



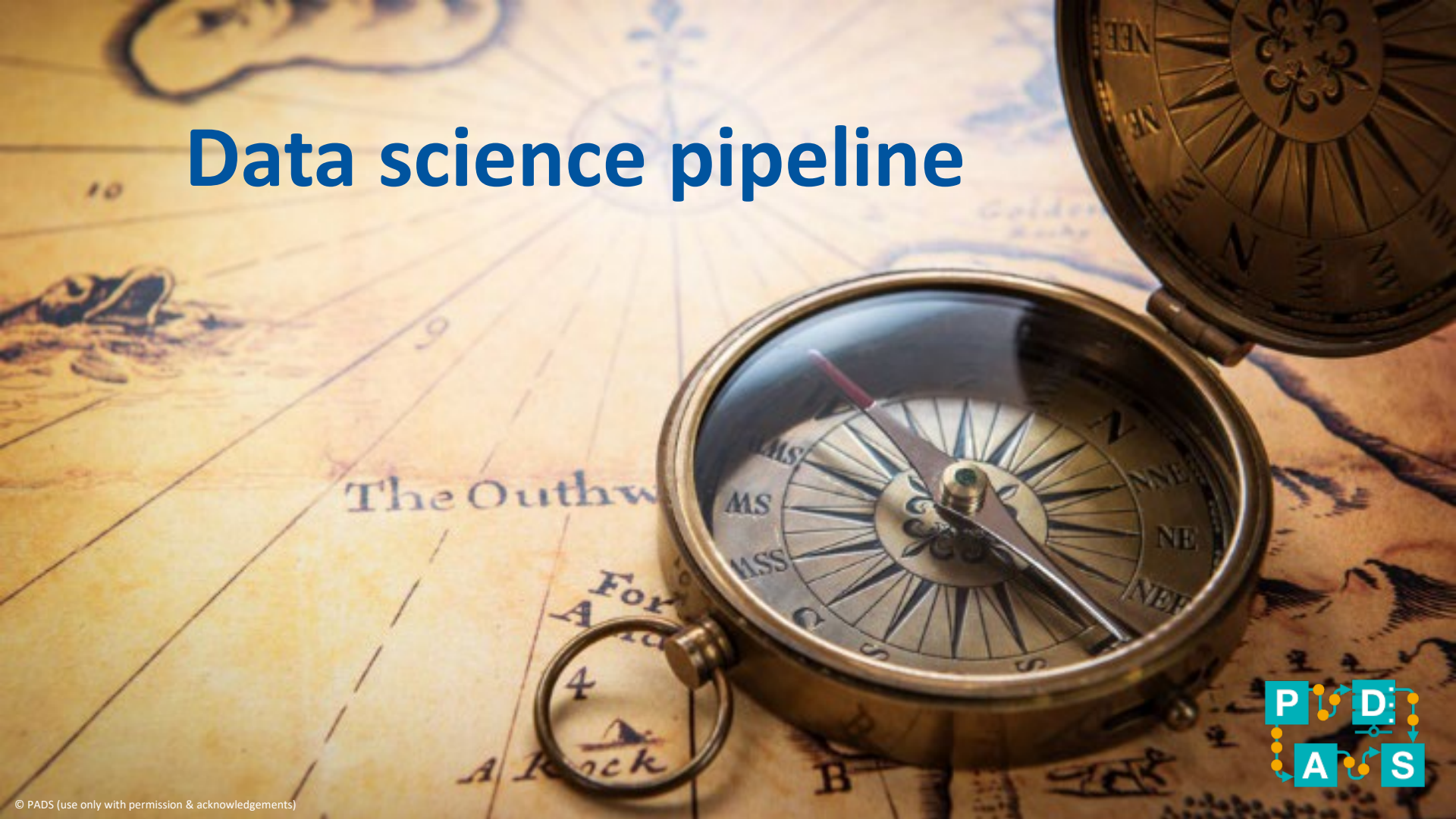
RWTH AACHEN UNIVERSITY

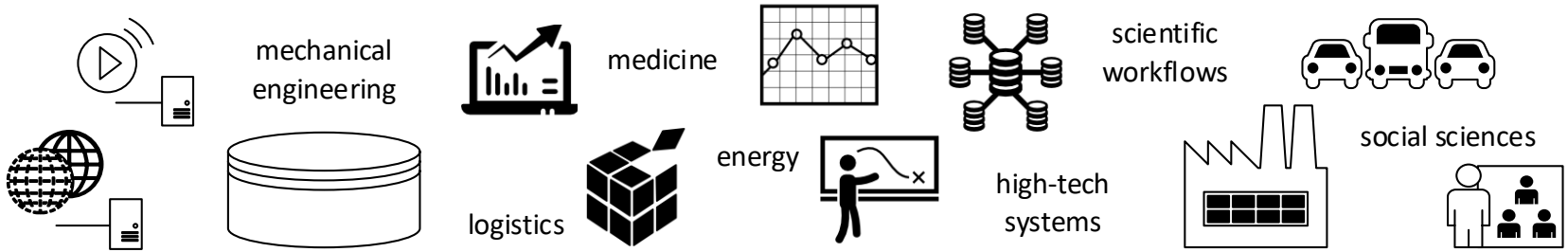
Chair of Process and Data Science

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"

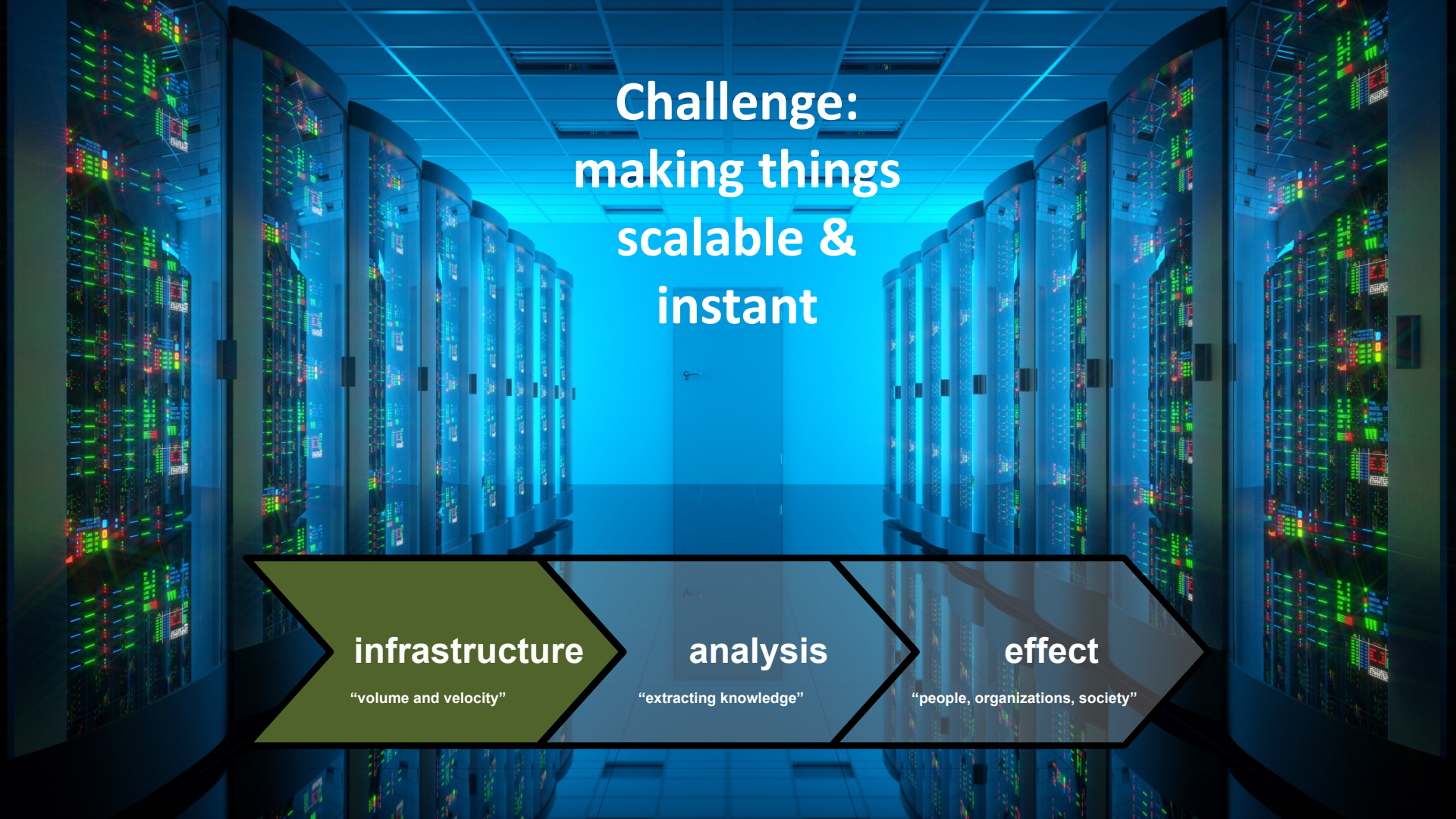
- statistics
- data/process mining
- machine learning
- artificial intelligence
- visualization
- ...

effect

"people, organizations, society"

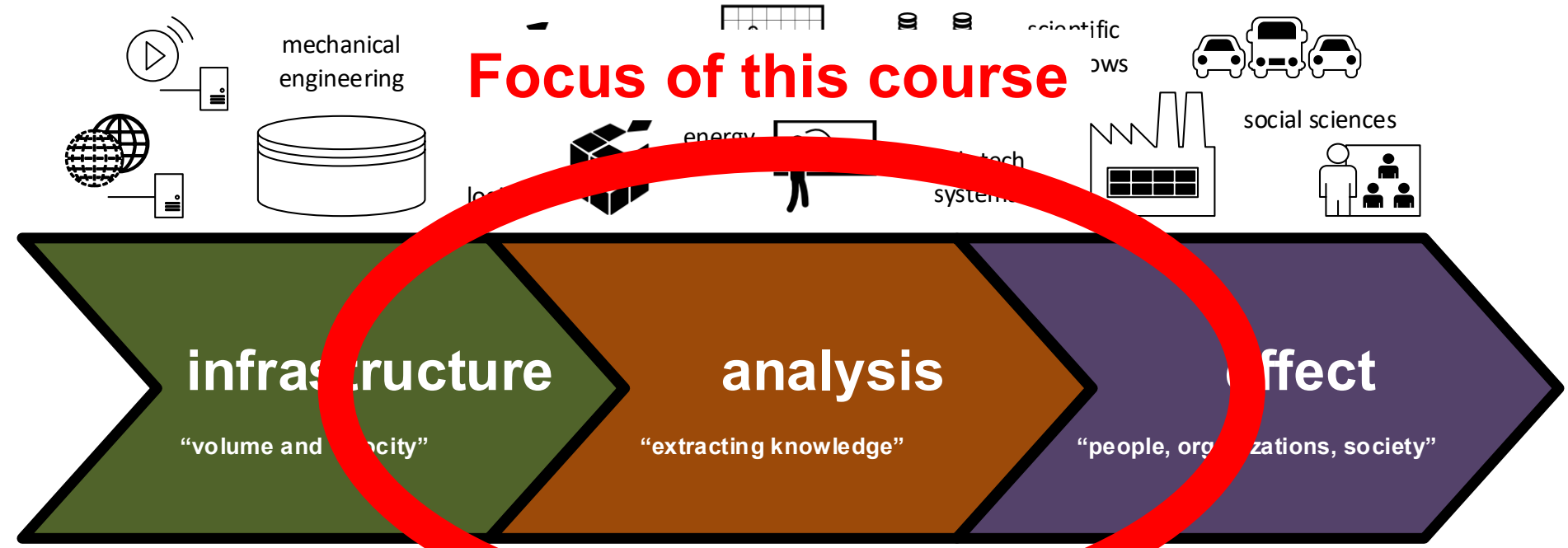
- ethics & privacy
- IT law
- operations management
- business models
- entrepreneurship
- ...

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
infrastructure
"volume and velocity"
analysis
"extracting knowledge"
effect
"people, organizations, society"
United States Secretary of Defense
Donald Rumsfeld (12-2-2002)

Challenge: doing all of this in a responsible manner!



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

- statistics
- data/pr
- machine learning
- artificial intelligence
- visualization
- ...

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

Chair of Process and Data Science

Types of data

Types of data

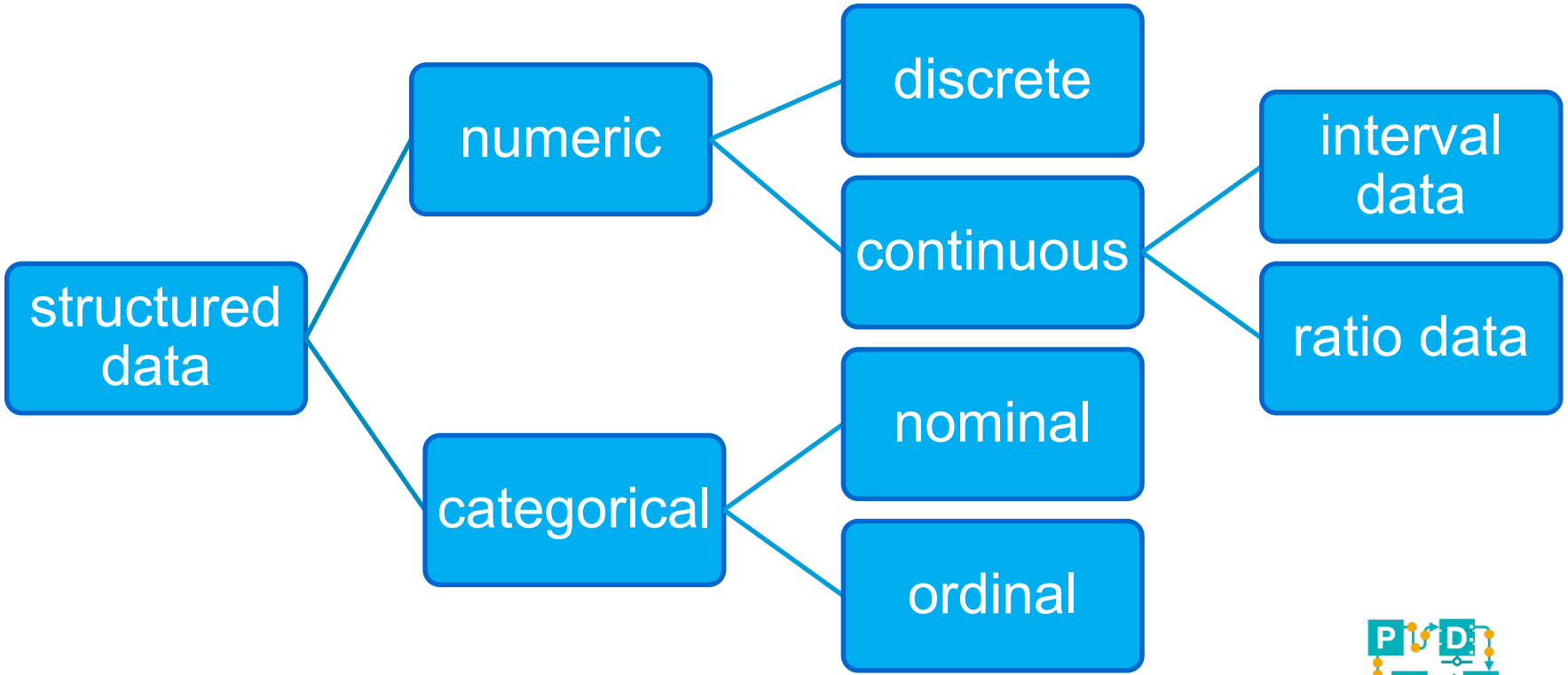
Structured data

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

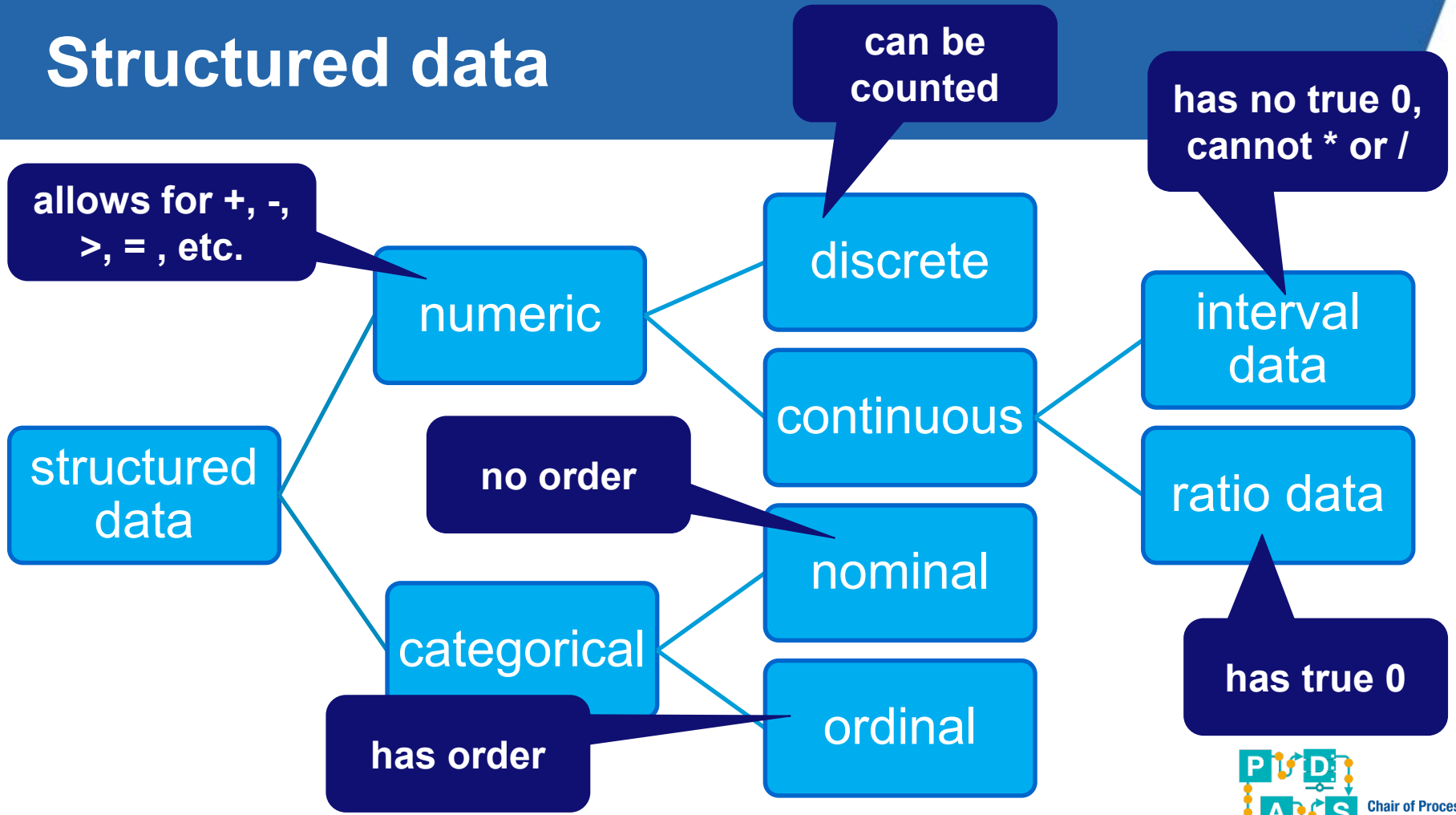
Unstructured data

- Text
- Audio
- Image
- Signal
- Video

Structured data

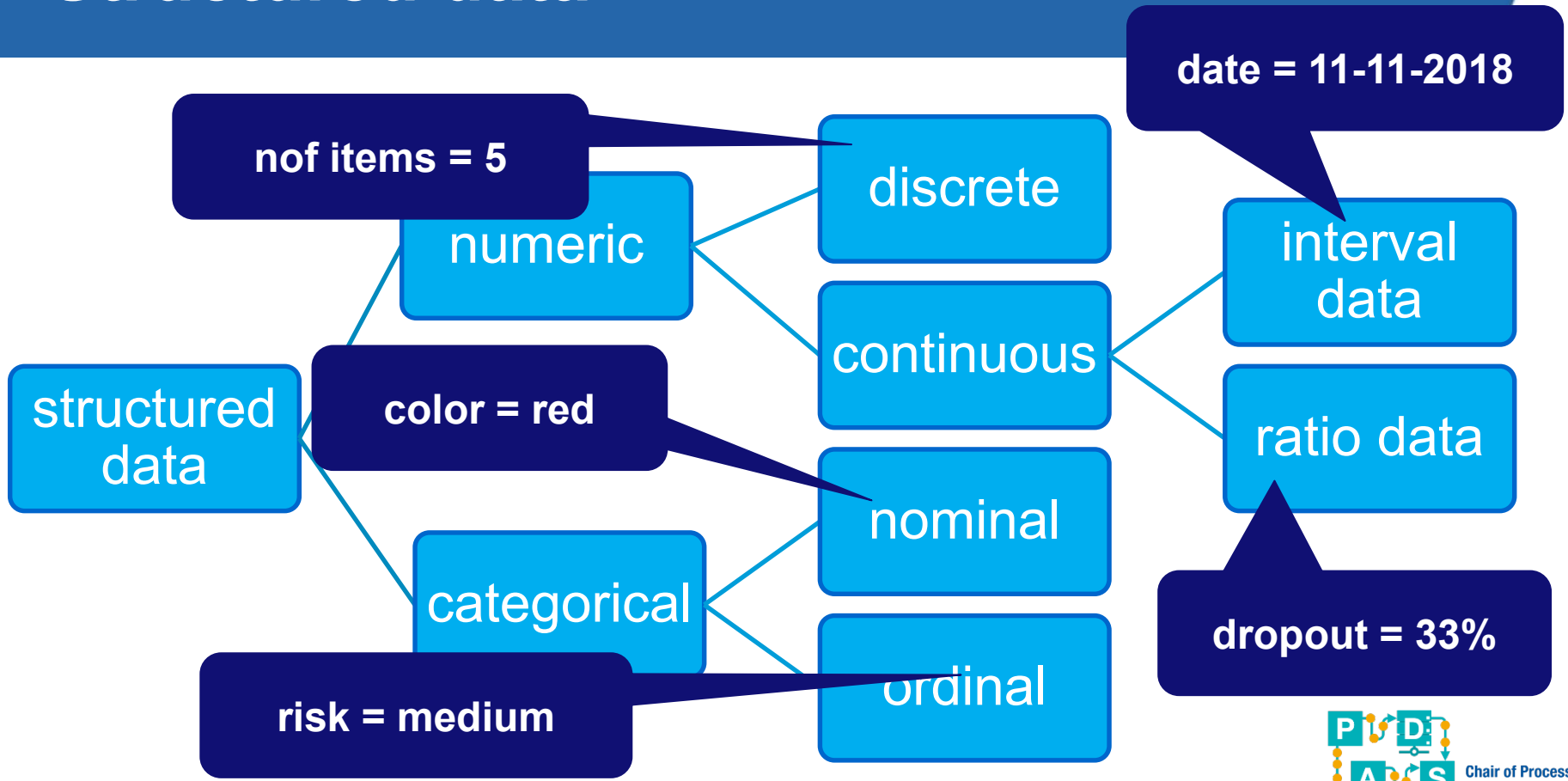


Structured data



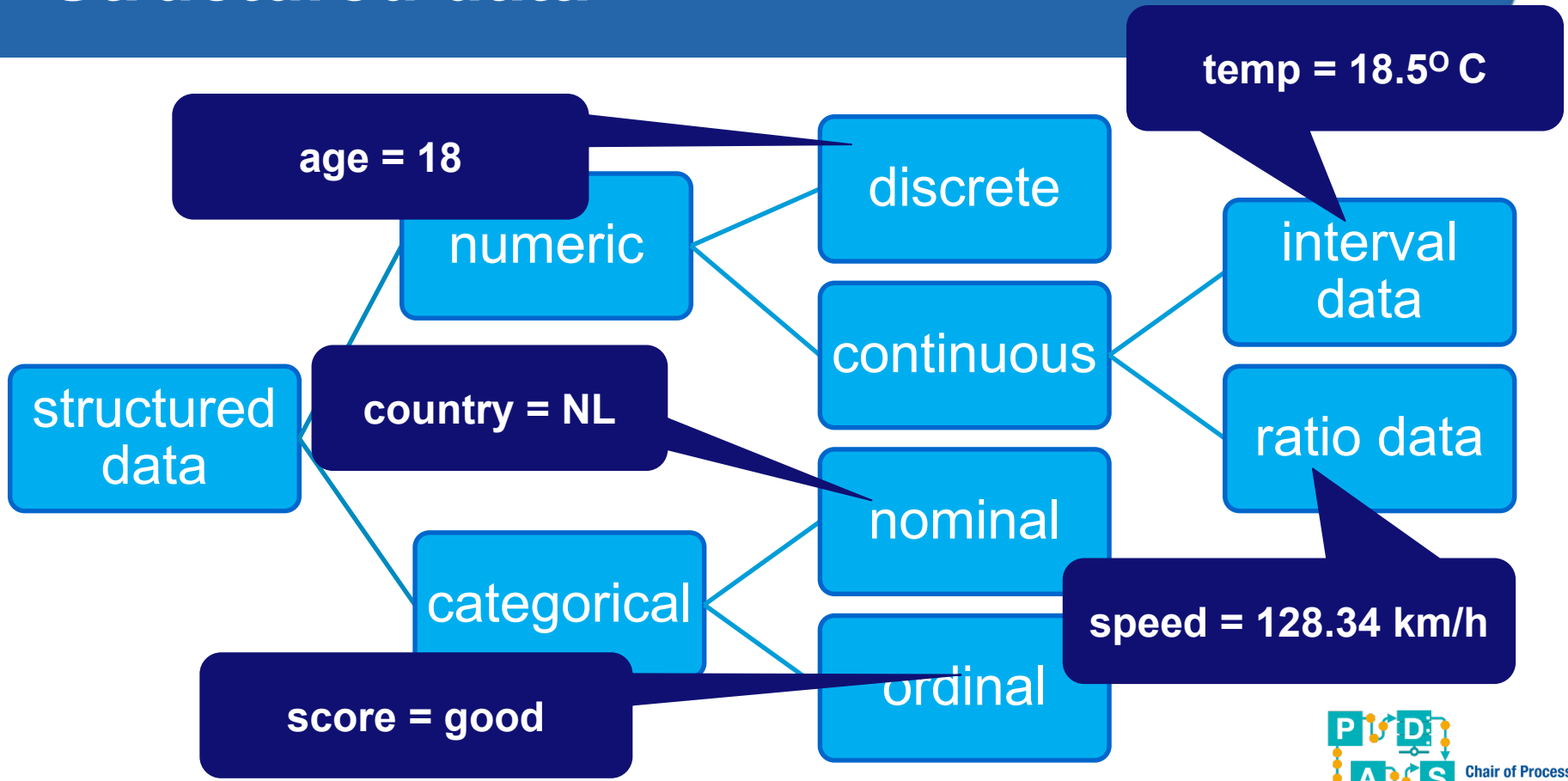
For interval data zero is arbitrary, for ratio data, zero is absolute.

Structured data



Chair of Process and Data Science

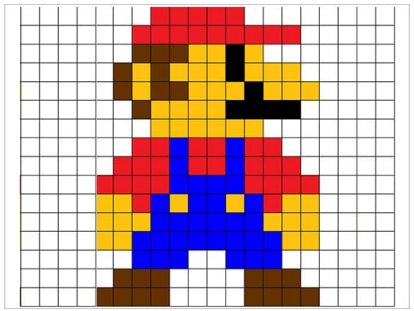
Structured data



Chair of Process and Data Science

Unstructured data

- Text
- Audio
- Image
- Signal
- Video



11111001000100111101

01001001000101110 1

11001001011101010 1

Chair of Process and Data Science

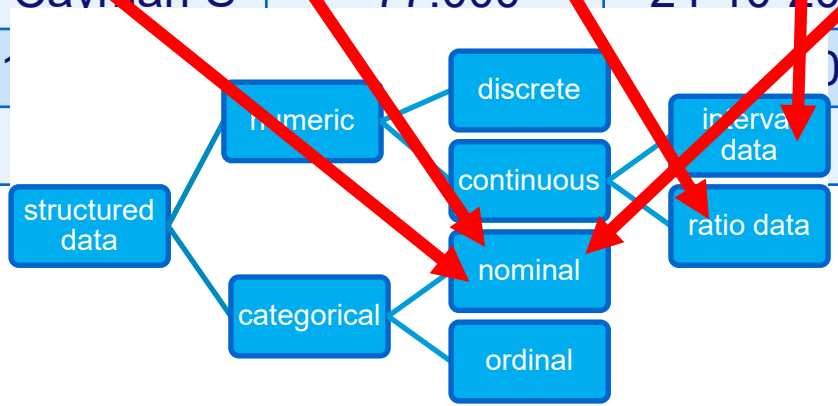
Tabular data

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.

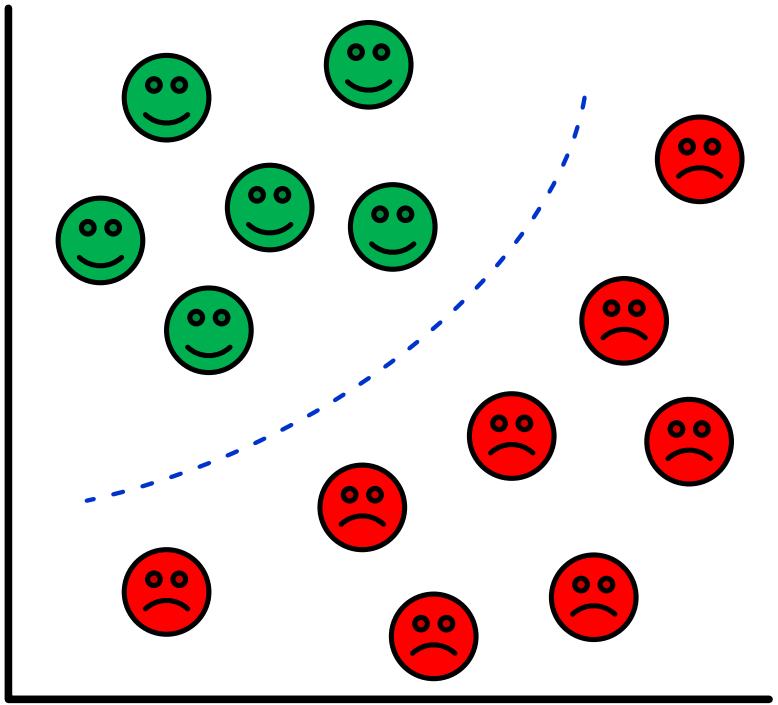
Unlabeled tabular data

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
…	…	…	…	…

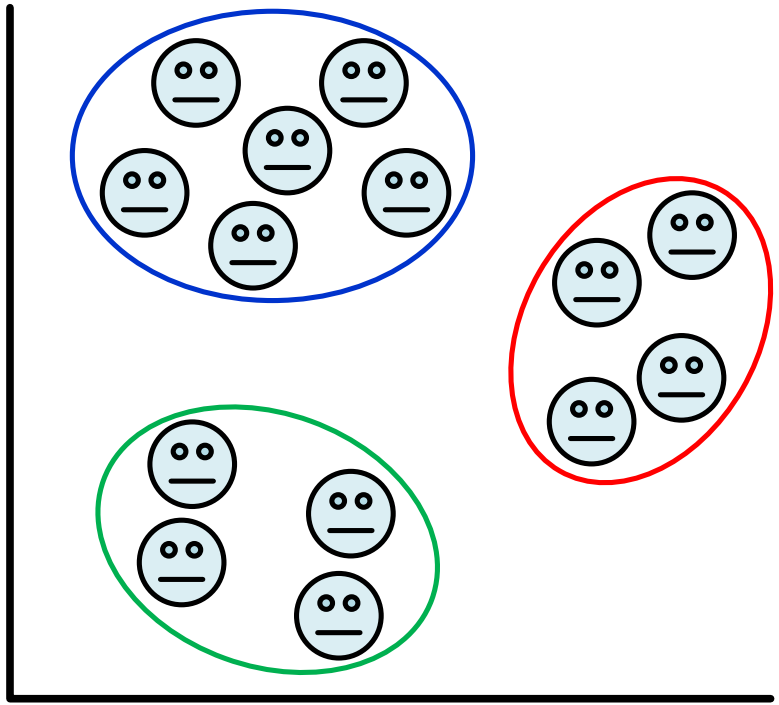
No target feature has been selected

Chair of Process and Data Science

Supervised versus unsupervised learning



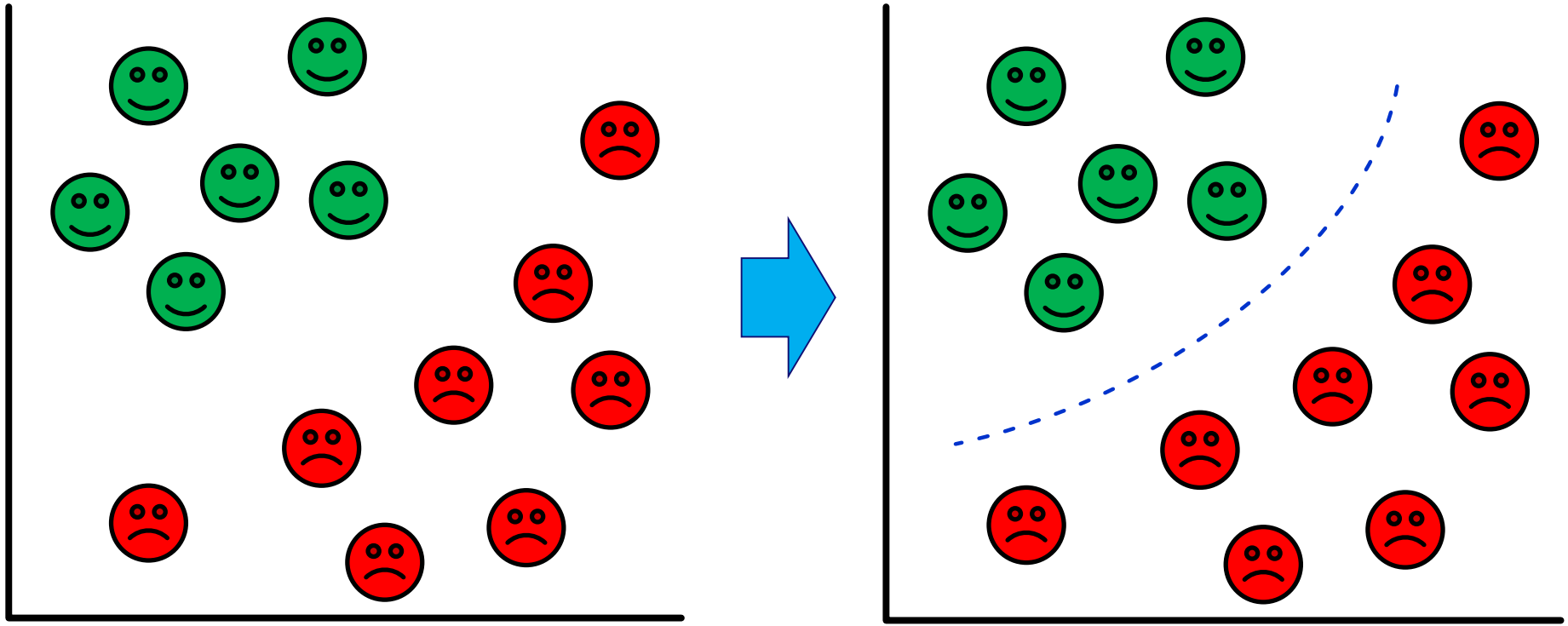
Supervised



Unsupervised

Chair of Process and Data Science

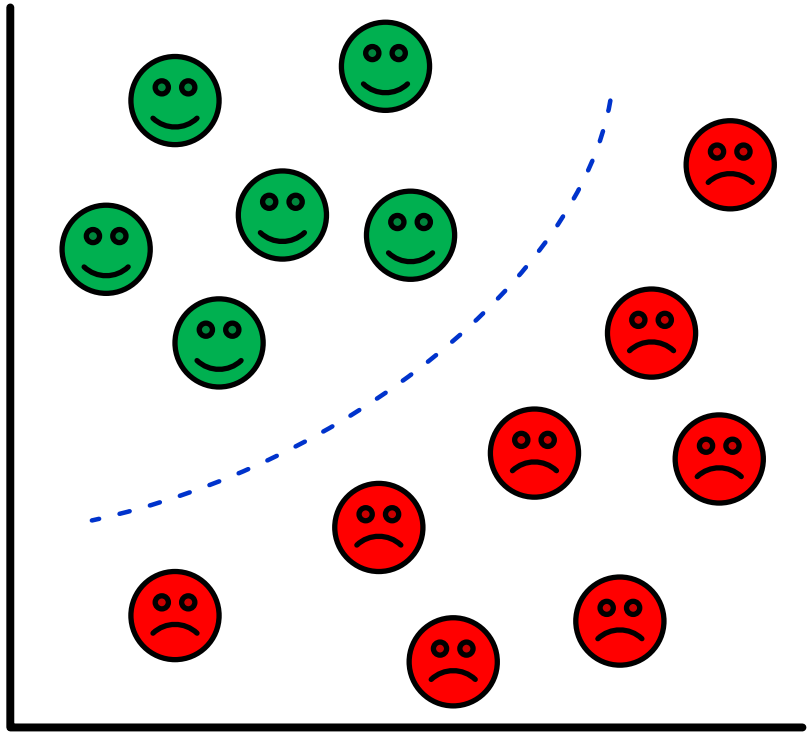
Supervised learning using labeled data



All instances have a target label (here color).

Chair of Process and Data Science

Supervised learning using labeled data

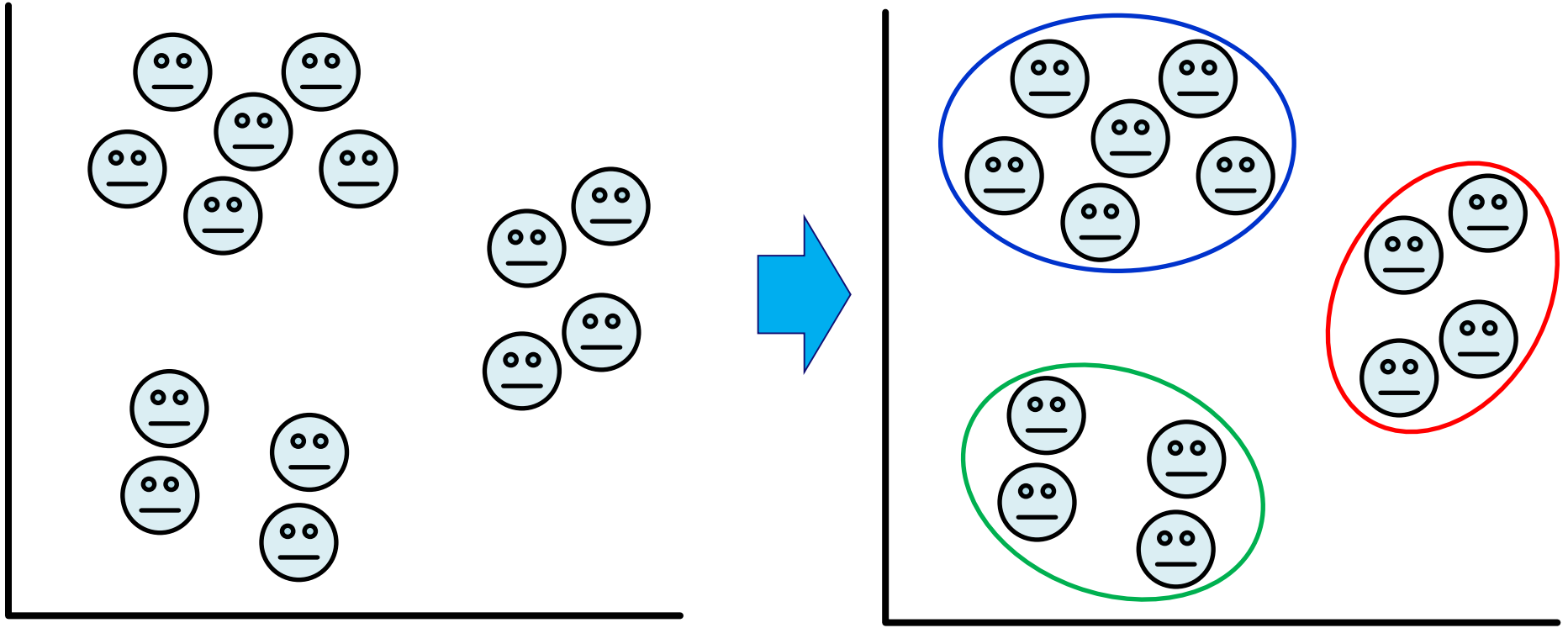


The goal is to find a “rule” in terms descriptive features that explains the target feature as good as possible.

Examples of supervised learning

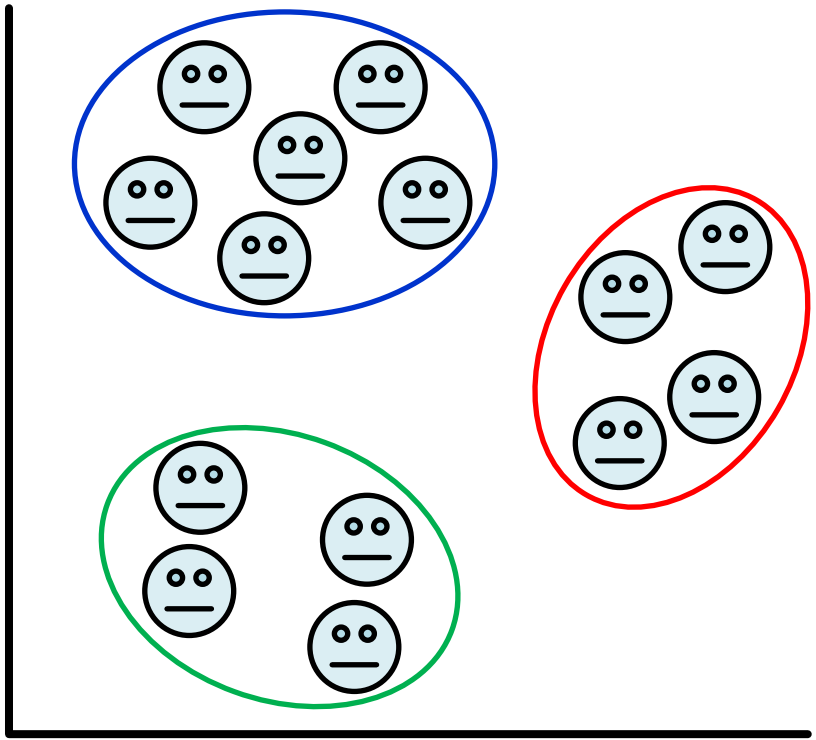
- **Hospital**:
  - target variable = recover (Y/N)
  - descriptive variables = age, gender, smoking, …
- **University**:
  - target variable = drops out (Y/N)
  - descriptive variables = mentor, prior education, …
- **Production**:
  - target variable = order is delivered in time (Y/N)
  - descriptive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

Unsupervised learning



Instances do not have a target label.

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 a form (unknown unknowns):
  - Customers who buy bread and butter typically pay by cash.
  - Patients who 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 customers.

Process discovery is also a form of unsupervised learning (process model = very sophisticated rule)

Case ID	Activity	Resource	Timestamp	product	prod-price	quantity	address
...	...	....	...	....	...	...	...
6350	place order	Aiden	2018/02/13 14:29:45.000	APPLE iPhone 6 16 GB	639,00 €	5	NL-7751DG-21
6283	pay	Lily	2018/02/13 14:39:25.000	SAMSUNG Galaxy S6 32 GB	543.99	3	NL-7828AM-11a
6253	prepare delivery	Sophia	2018/02/13 15:01:33.000	APPLE iPhone 6 16 GB	639,00 €	3	NL-7887AC-13
6257	prepare delivery	Aiden	2018/02/13 15:03:43.000	SAMSUNG Galaxy S6 32 GB	543.99	1	NL-9521KJ-34
6185	confirm payment	Emily	2018/02/13 15:05:36.000	SAMSUNG Galaxy S4	329,00 €	1	NL-9521GC-32
6218	confirm payment	Emily	2018/02/13 15:08:11.000	APPLE iPhone 6s Plus 64 GB	969,00 €	2	NL-7948BX-10
6245	make delivery	Michael	2018/02/13 15:14:04.000	APPLE iPhone 6 16 GB	639,00 €	3	NL-7905AX-38
6272	pay	Emily	2018/02/13 15:20:36.000	APPLE iPhone 6 16 GB	639,00 €	1	NL-7821AC-3
6269	pay	Charlotte	2018/02/13 15:25:21.000	SAMSUNG Galaxy S4	329,00 €	1	NL-7907EJ-42
6212	prepare delivery	Sophia	2018/02/13 15:43:39.000	HUAWEI P8 Lite	234,00 €	1	NL-7905AX-38
6323	send invoice	Alexander	2018/02/13 15:46:08.000	APPLE iPhone 6 16 GB	639,00 €	1	NL-7833HT-15
6246	confirm payment	Jack	2018/02/13 15:56:03.000	SAMSUNG Galaxy S4	329,00 €	3	NL-7833HT-15
6347	send invoice	Jack	2018/02/13 15:57:42.000	SAMSUNG Galaxy S4	329,00 €	3	NL-7905AX-38
6351	place order	Zoe	2018/02/13 16:17:37.000	APPLE iPhone 5s 16 GB	449,00 €	3	NL-9521GC-32
6204	prepare delivery	Sophia	2018/02/13 16:31:28.000	SAMSUNG Core Prime G361	135,00 €	1	NL-7828AM-11a
6204	make delivery	Kaylee	2018/02/13 16:51:54.000	SAMSUNG Core Prime G361	135,00 €	1	NL-7828AM-11a
6265	confirm payment	Lily	2018/02/13 16:55:55.000	SAMSUNG Galaxy S4	329,00 €	4	NL-9521GC-32
6250	confirm payment	Jack	2018/02/13 17:03:26.000	MOTOROLA Moto G	199,00 €	4	NL-7942GT-2
6328	send invoice	Lily	2018/02/13 17:30:16.000	APPLE iPhone 6s 64 GB	858,00 €	4	NL-9514BV-16
6352	place order	Aiden	2018/02/13 17:53:22.000	APPLE iPhone 6 16 GB	639,00 €	2	NL-9514BV-16
6317	send invoice	Jack	2018/02/13 18:45:30.000	APPLE iPhone 6s 64 GB	858,00 €	5	NL-7907EJ-42
6353	place order	Sophia	2018/02/13 20:16:20.000	APPLE iPhone 5s 16 GB	449,00 €	4	NL-7751AR-19
...	...	....	...	...	...	...	...

71,043 events
12,666 cases
7 activities

place order
send invoice
pay
prepare delivery
make delivery
confirm payment

8016 x

place order
send invoice
cancel order

1651 x

place order
send invoice
pay
prepare delivery
confirm payment
make delivery

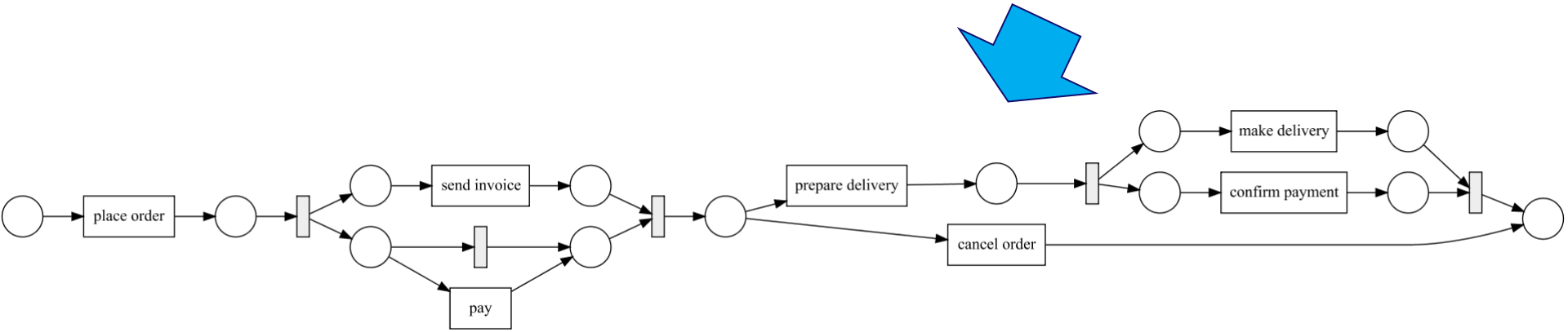
2962 x

place order
pay
send invoice
prepare delivery
make delivery
confirm payment

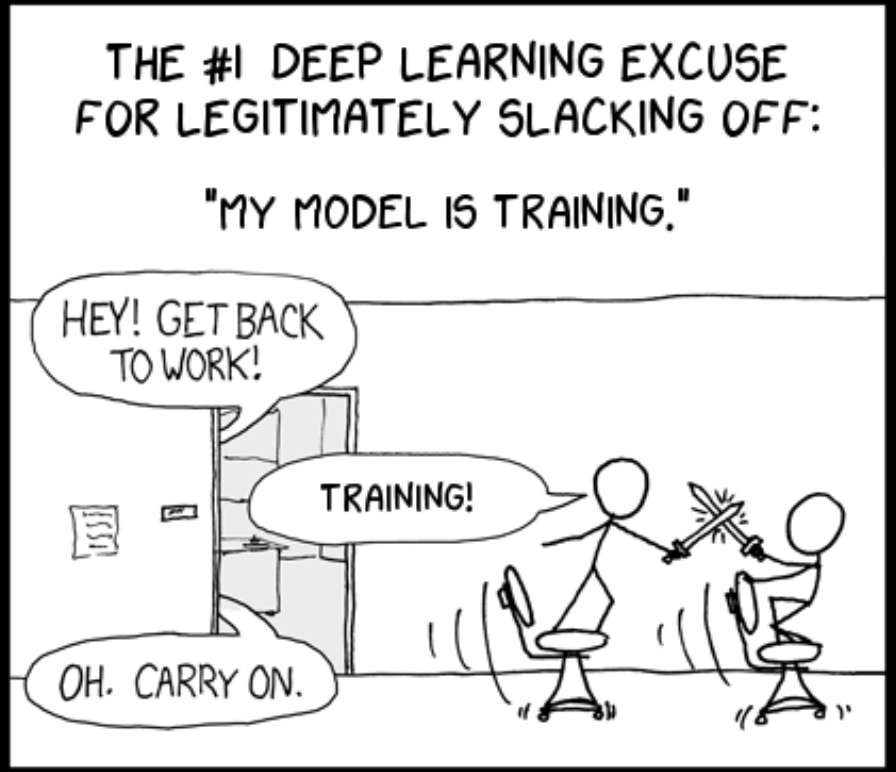
30 x

place order
pay
send invoice
prepare delivery
confirm payment
make delivery

7 x



Terminology



Cartoons by Randall Munroe xkcd.com

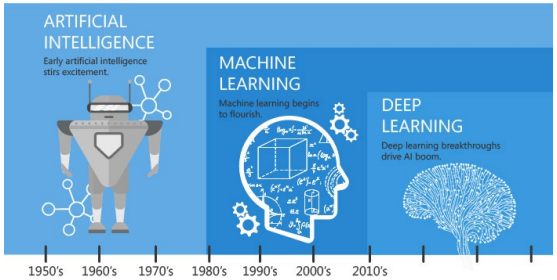
Chair of Process and Data Science

Terminology

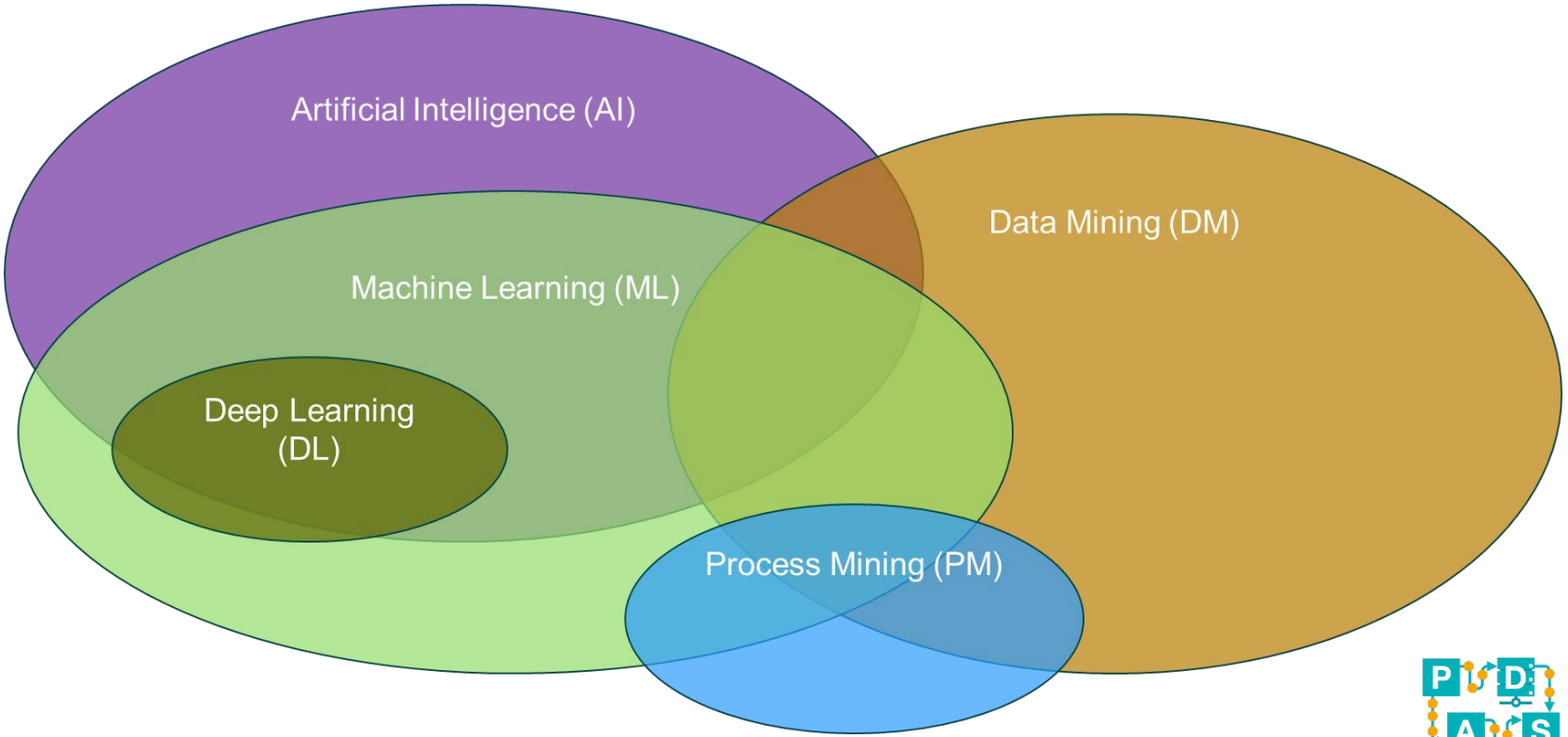
- **Many different names** (statistics, data analytics, data mining, machine learning, artificial intelligence, predictive analytics, process mining, etc.) **are used to refer to the key disciplines that contribute to data science.**
- **Unfortunately, the areas these names describe are heavily overlapping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

Example: scoping machine learning

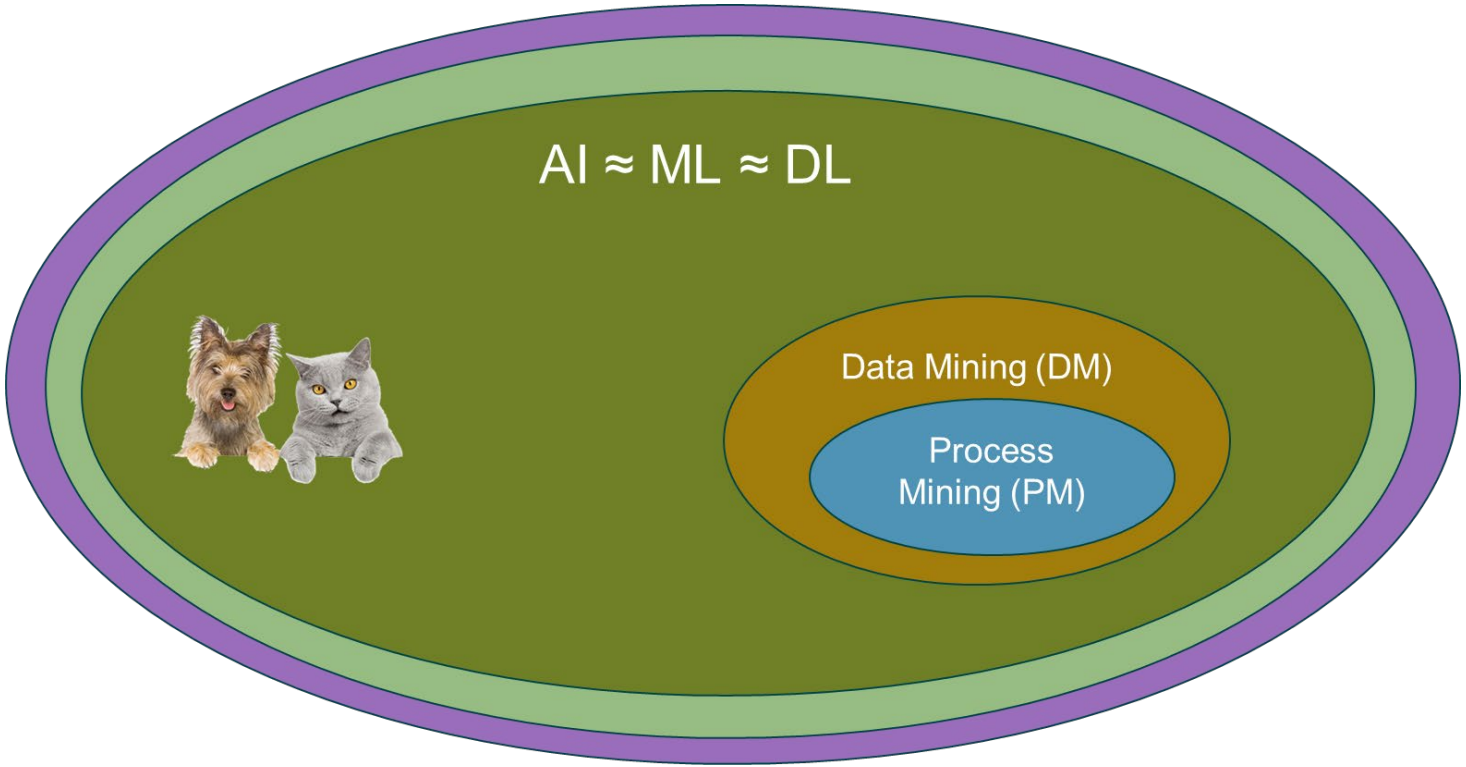
- Sometimes machine learning is used as a synonym for deep neural networks and sometimes to cover the entire spectrum of learning techniques.
- The fact that a neural network can be used as a classifier does not imply that the numerous classification techniques developed in data mining are part of machine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



A more balanced view



Perception by outsiders



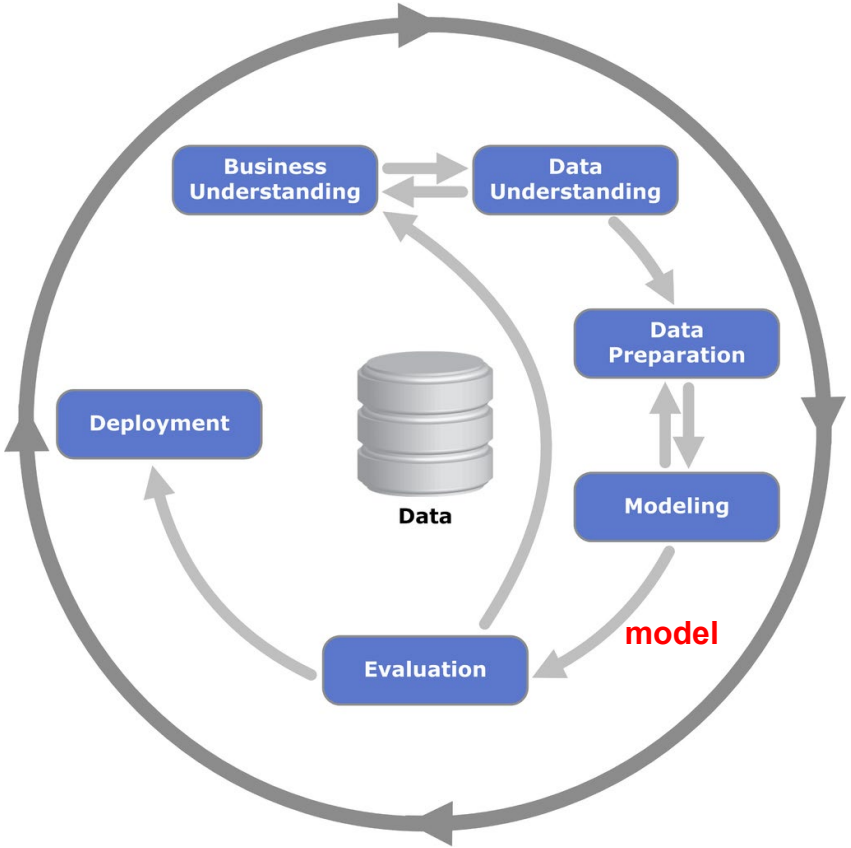
Chair of Process and Data Science

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

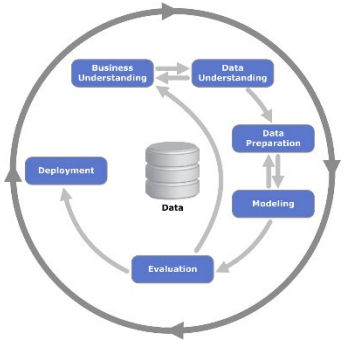
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

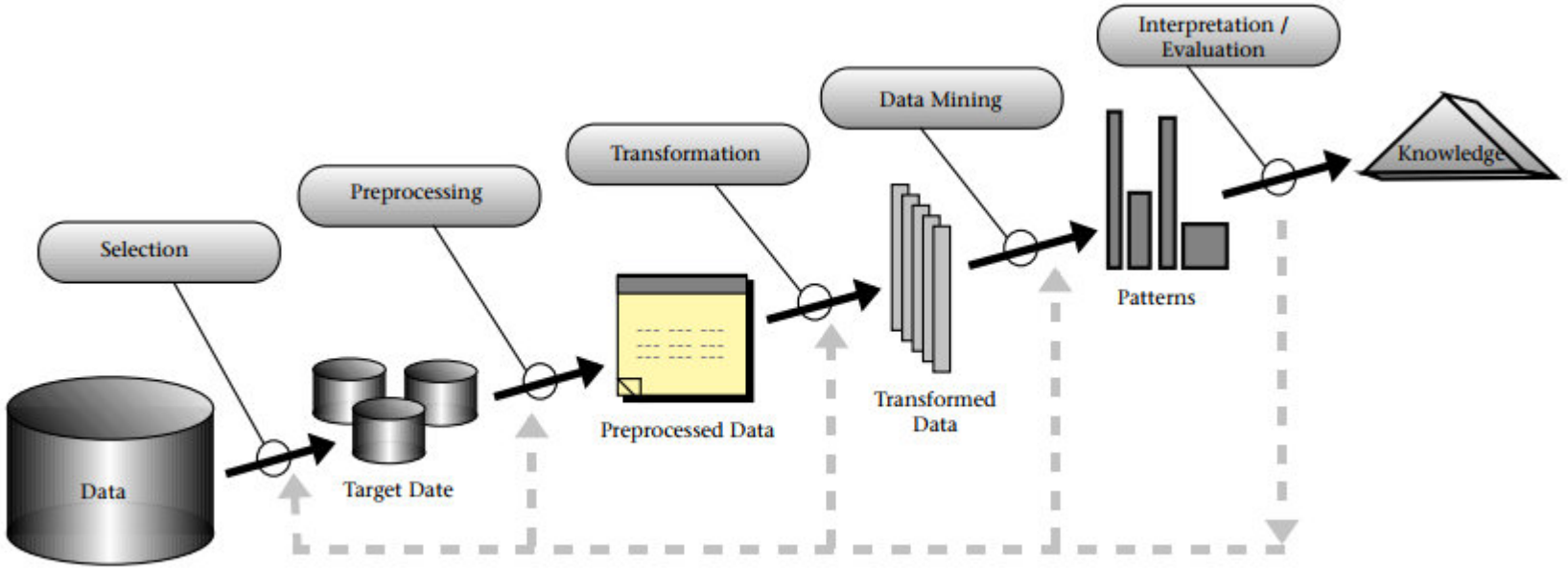
The term “modeling” can be misleading: selection and assumptions (human) or automated learning by a tool or a algorithm.

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



Chair of Process and Data Science

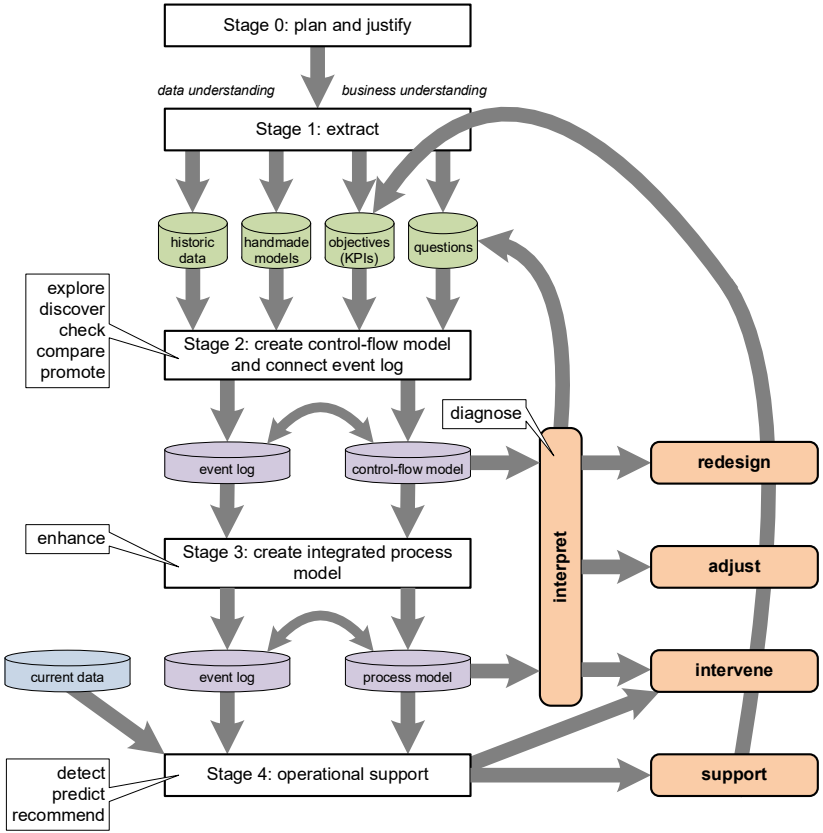
KDD (Knowledge Discovery in Databases) Process (by Fayyad, Piatetsky-Shapiro, and Smyth)



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

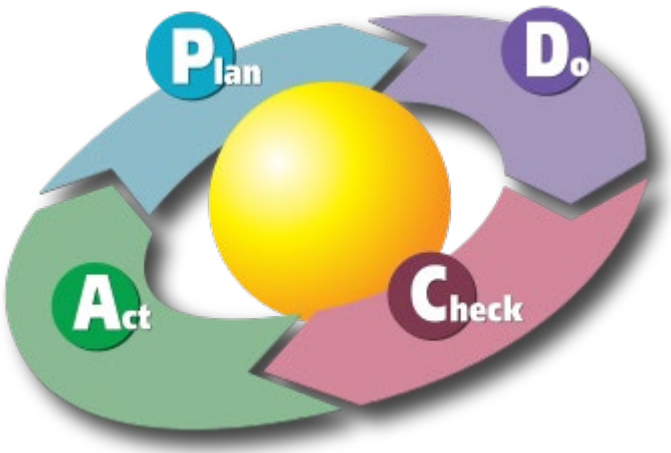
Chair of Process and Data Science

L* lifecycle model (specific for process mining)



Chair of Process and Data Science

Related to PDCA and DMAIC methodology



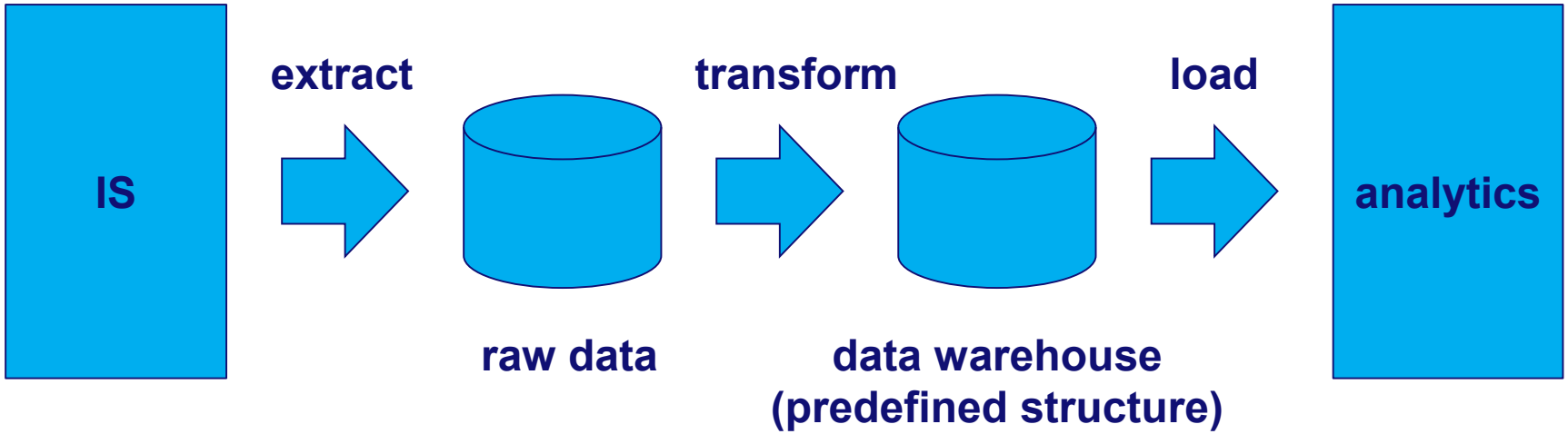
PDCA (Plan–Do–Check–Act) methodology by William Deming

Deming (1900-1993) is widely acknowledged as the leading management thinker in the field of quality. He was a statistician and business consultant whose methods helped Japan's recovery after the Second World War and beyond. Deming cycle dates from the 1950-ties and evolved into DMAIC.

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

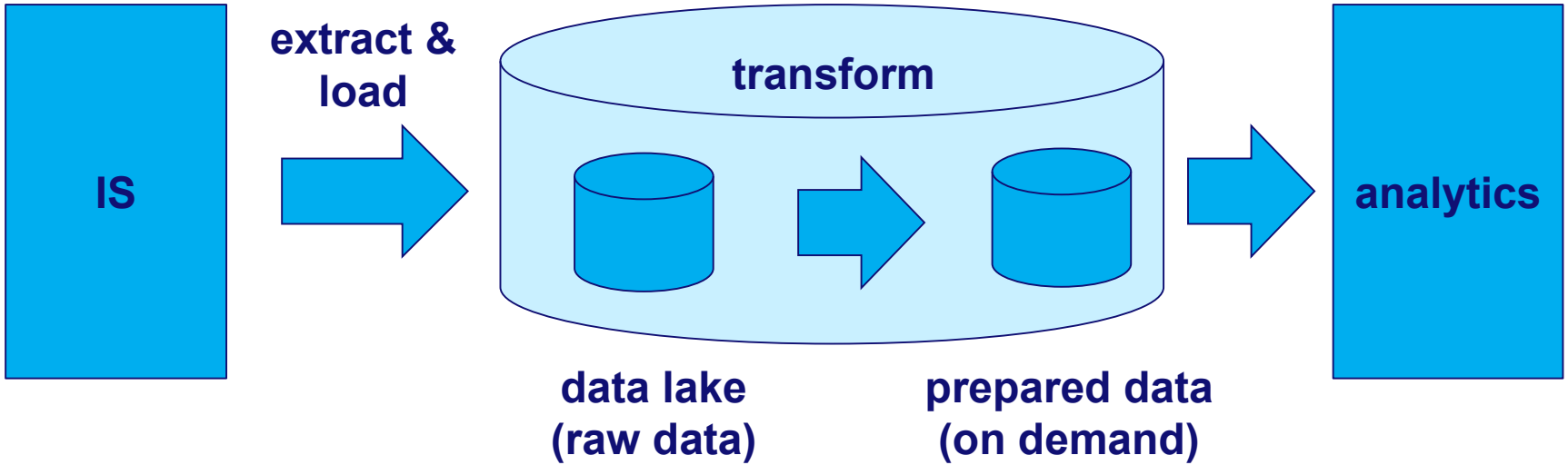
Chair of Process and Data Science

Extract, Transform, Load (ETL)



Chair of Process and Data Science

Extract, Load, Transform (ELT)



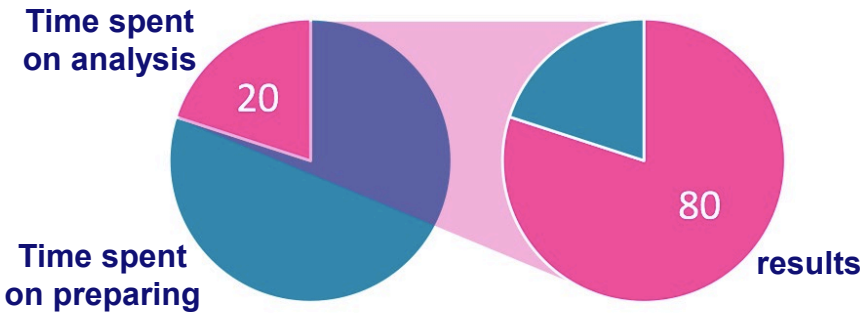
Chair of Process and Data Science


Data lakes (ELT)

Data warehouses (ETL)

Another 80/20 rule

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



Chair of Process and Data Science

Challenges

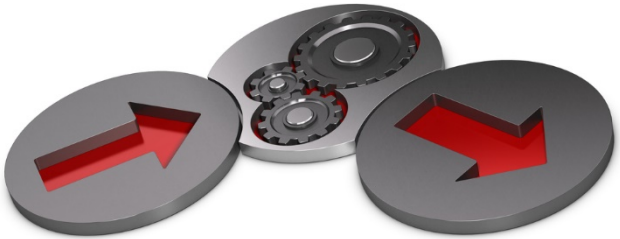
Finding data

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

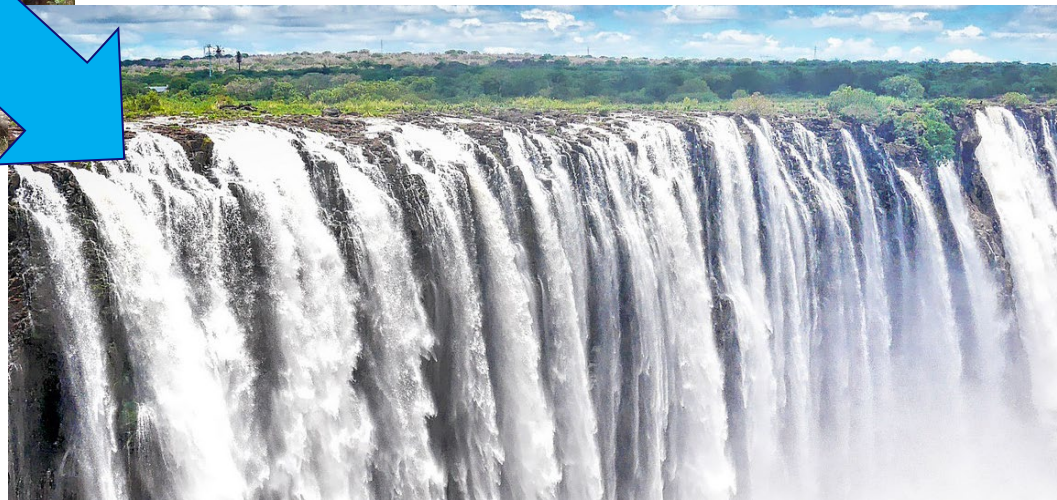
Chair of Process and Data Science

Transforming data

- Reorganizing data, filtering, etc.
- 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

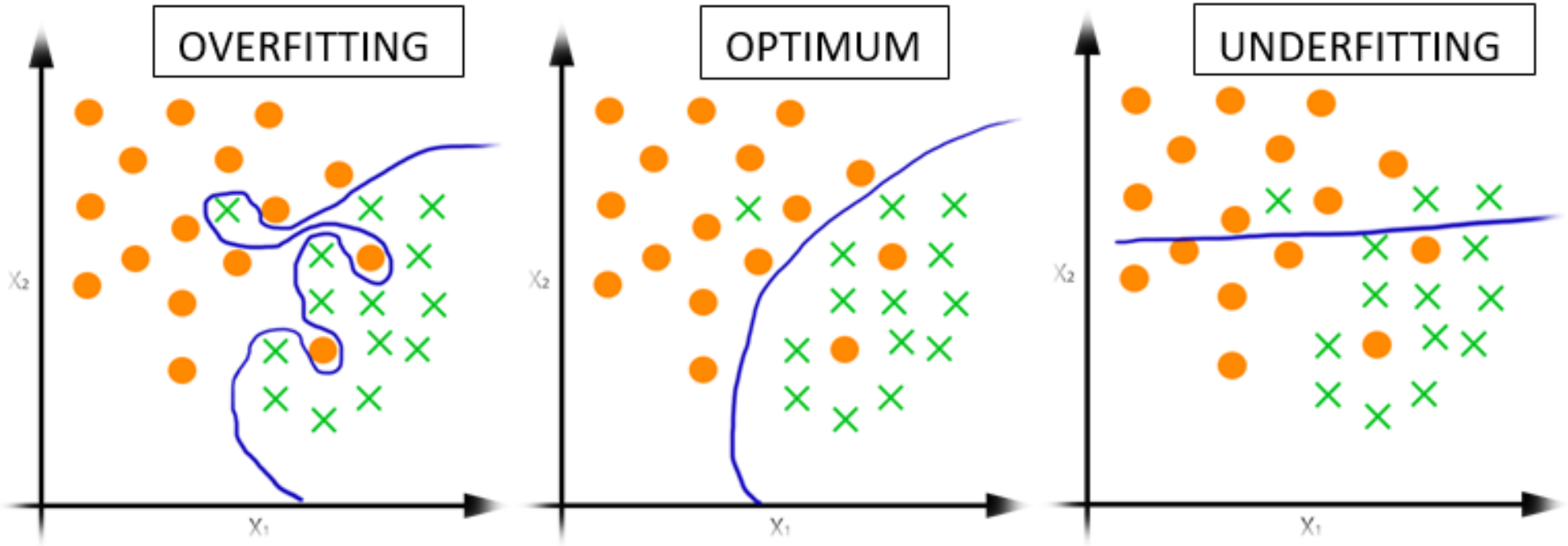
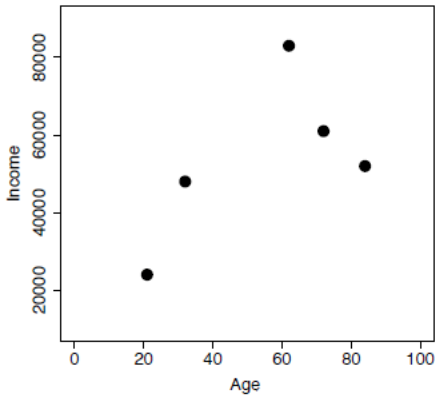
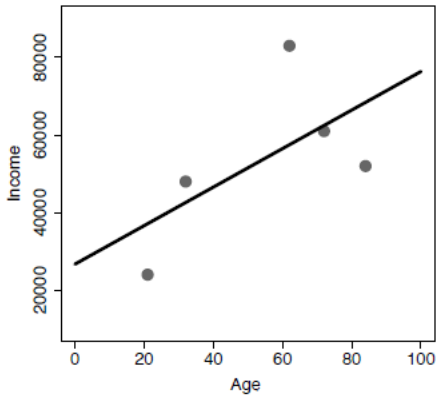


Diagram by Sachin Joglekar (Google).

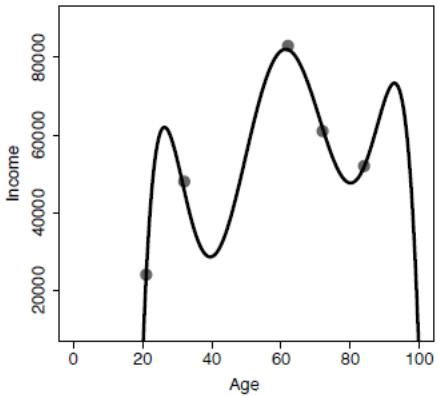
Overfitting / underfitting



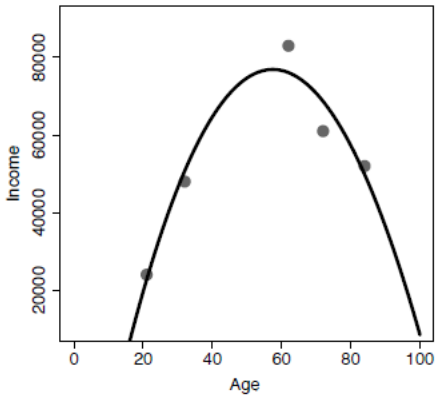
(a) Dataset



(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 a concept drift

Chair of Process and Data Science

Making results actionable

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

You are in a traffic jam …

It’s raining …

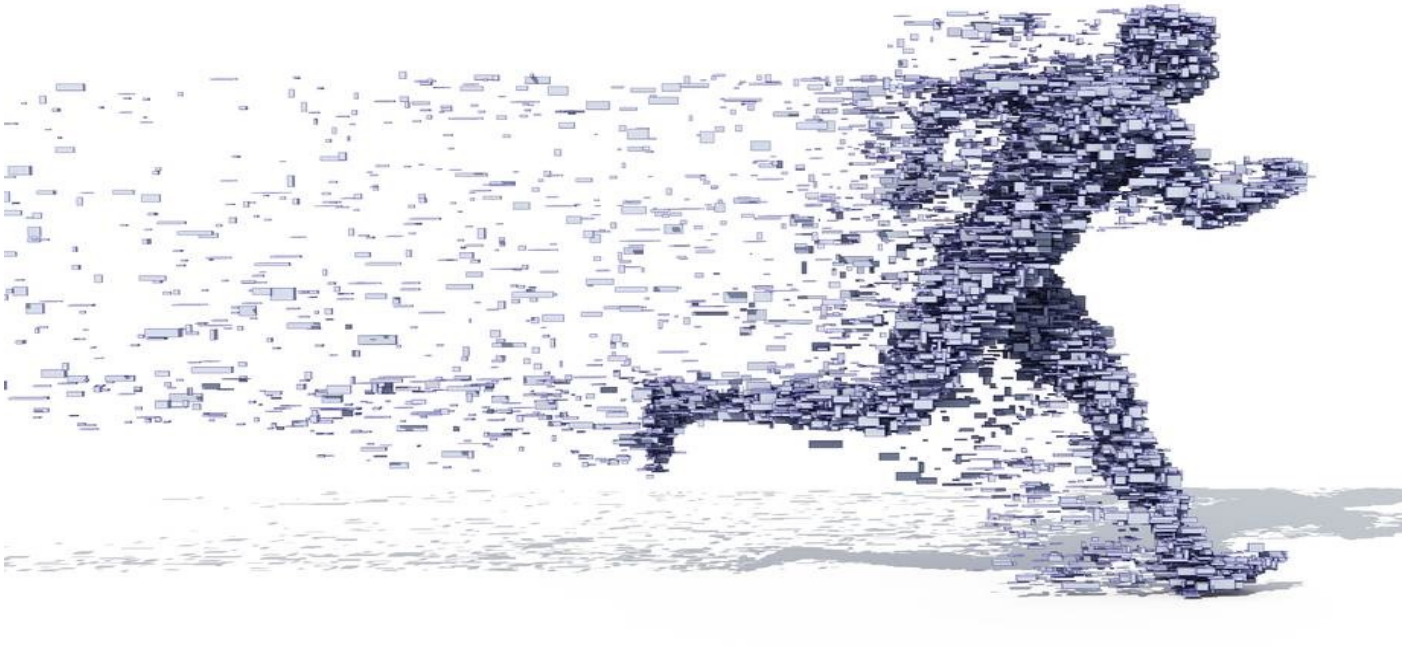
Concerns …

Responsible Data Science (RDS)

Ensuring fairness

Fairness

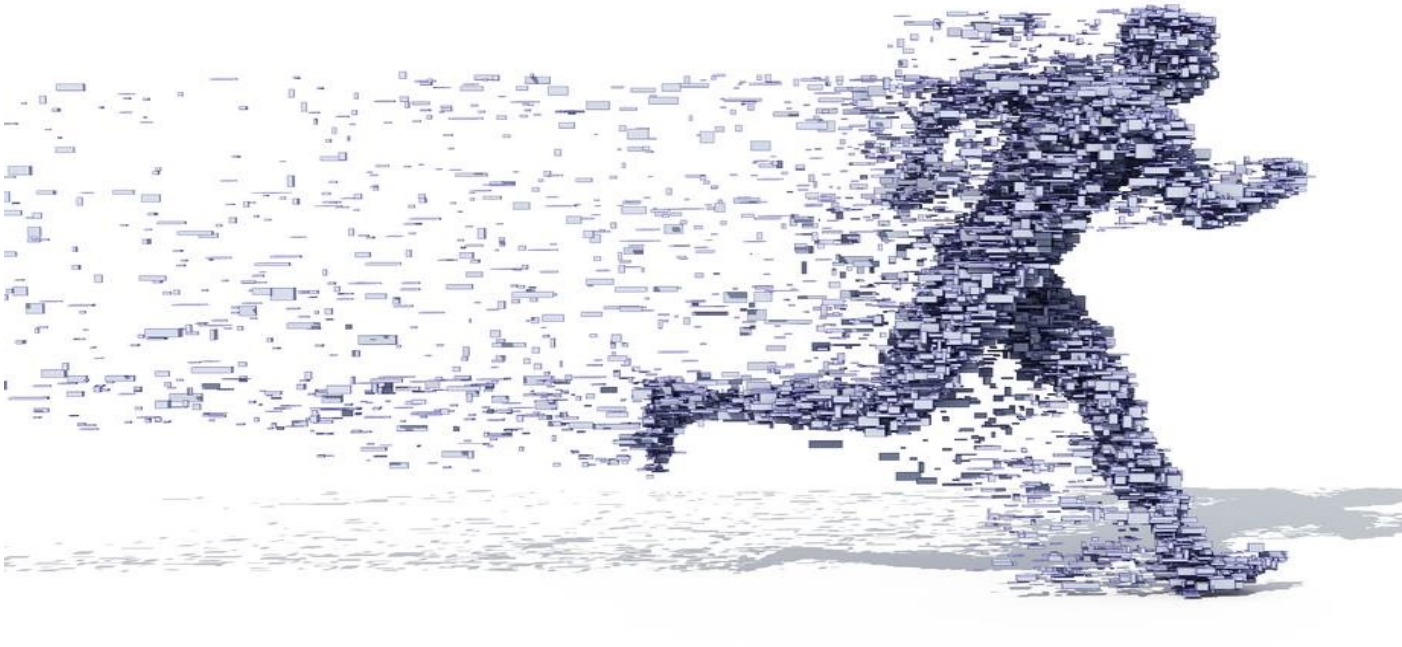
Data science without prejudice: how to avoid unfair conclusions even if they are true?



Ensuring accuracy

Accuracy

Data science without guesswork: how to answer questions with a guaranteed level of accuracy?



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

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 or explanatory value.

Conclusion

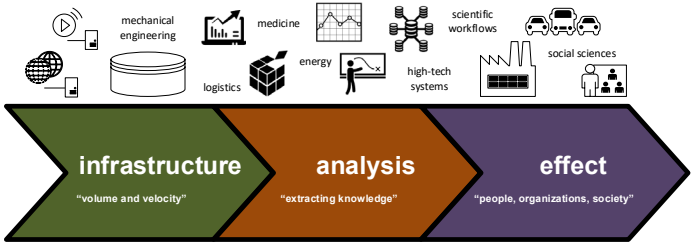
Conclusion

- **Data science**
  - **has many faces,**
  - **is “super relevant”, and**
  - **provides many challenges.**

- **Next lecture: basic data exploration and visualization**
- **Next instruction: Python installation and demo**

	Lecture	date	day
Lecture 1	Introduction	13/10/2021	Wednesday
Instruction 1	Python	14/10/2021	Thursday
Instruction 2	Crash Course in Python	15/10/2021	Friday
Lecture 2	Basic data visualization/exploration	20/10/2021	*Wednesday*
Lecture 3	Decision trees	21/10/2021	Thursday
Instruction 3	Decision trees and data visualization/exploration	22/10/2021	Friday
Lecture 4	Regression	27/10/2021	Wednesday
Lecture 5	Support vector machines	28/10/2021	Thursday
Instruction 4	Regression and support vector machines	29/10/2021	Friday
Lecture 6	Neural networks (1/2)	03/11/2021	Wednesday
Lecture 7	Neural networks (2/2)	04/11/2021	Thursday
Lecture 8	Evaluation of supervised learning problems	10/11/2021	Wednesday
Instruction 5	Neural networks	11/11/2021	Thursday
Instruction 6	Neural networks and evaluation	12/11/2021	Friday
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets	18/11/2021	Thursday
Instruction 7	Clustering and frequent item sets	19/11/2021	Friday
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining	25/11/2021	Thursday
Instruction 8	Association rules and sequence mining	26/11/2021	Friday
Lecture 13	Process mining (unsupervised)	01/12/2021	Wednesday
Lecture 14	Process mining (supervised)	02/12/2021	Thursday
Instruction 9	Process Mining	03/12/2021	Friday
Lecture 15	Text Mining (1/2)	08/12/2021	Wednesday
Lecture 16	Text Mining (2/2)	09/12/2021	Thursday
Instruction 10	Q&A Assignment 1	10/12/2021	Friday
Lecture 17	Data preprocessing, data quality, binning, etc.	15/12/2021	Wednesday
Lecture 18	Visual analytics & information visualization	16/12/2021	Thursday
Instruction 11	Text Mining	17/12/2021	Friday
Lecture 19	Responsible data science (1/2)	22/12/2021	Wednesday
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data	12/01/2022	Wednesday
Instruction 12	Preprocessing and visualization	13/01/2022	Thursday
Instruction 13	Q&A Assignment 2	14/01/2022	Friday
Lecture 22	Closing	19/01/2022	Wednesday
Instruction 14	Big Data (1/2)	20/01/2022	Thursday
Instruction 15	Responsible data science	21/01/2022	Friday
Instruction 16	Big Data (2/2)	27/01/2022	Thursday
Instruction 17	Example Exam Questions	28/01/2022	Friday
Instruction 18	Questions	02/02/2022	Wednesday

Disclaimer: always check RWTH Moodle for last minute changes



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

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

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

	Lecture	date	day
Lecture 1	Introduction		
Lecture 2	Basic data visualization/exploration	20/10/2021	*Wednesday*
Lecture 3	Decision trees		
Lecture 4	Regression	27/10/2021	Wednesday
Lecture 5	Support vector machines		
Lecture 6	Neural networks (1/2)		
Lecture 7	Neural networks (2/2)	10/11/2021	Wednesday
Lecture 8	Evaluation of supervised learning problems		
Lecture 9	Clustering	17/11/2021	Wednesday
Lecture 10	Frequent item sets		
Lecture 11	Association rules	24/11/2021	Wednesday
Lecture 12	Sequence mining		
Lecture 13	Process mining (unsupervised)		
Lecture 14	Process mining (supervised)	02/12/2021	Thursday
Lecture 15	Text Mining (1/2)		
Lecture 16	Text Mining (2/2)	09/12/2021	Thursday
Lecture 17	Data preprocessing, data quality, binning, etc.		
Lecture 18	Visual analytics & information visualization	16/12/2021	Thursday
Lecture 19	Responsible data science (1/2)		
Lecture 20	Responsible data science (2/2)	23/12/2021	Thursday
Lecture 21	Big data		
Lecture 22	Closing	19/01/2022	Wednesday