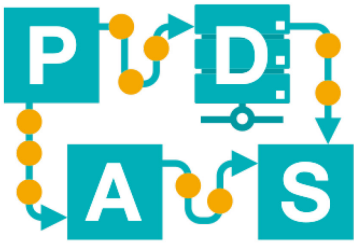
Closing Summary, Next Steps, Questions

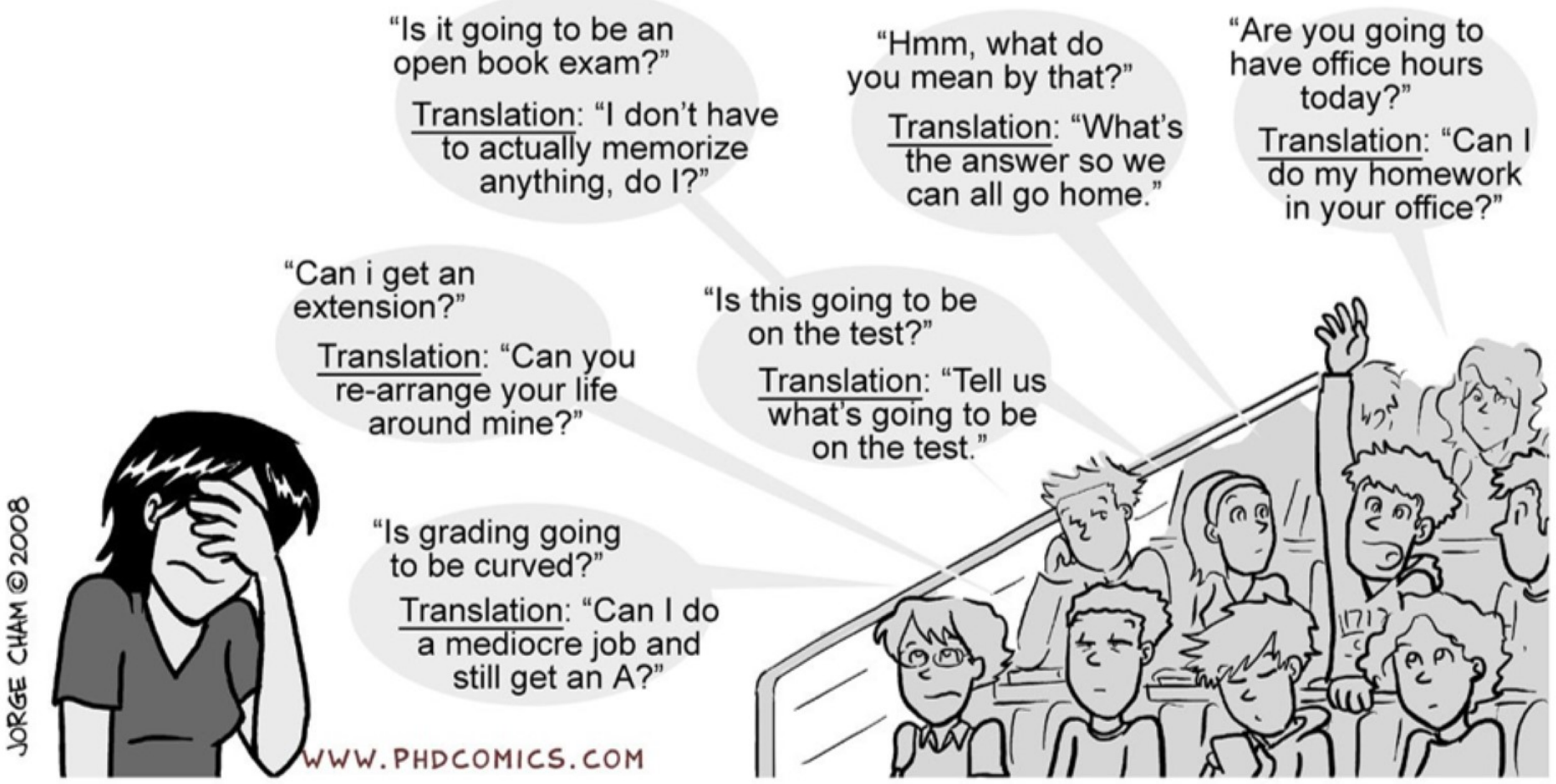
Lecture 24

IDS-L24



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY


"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM

Outline

- Rules of the Game
- The Bigger Picture
- Topics covered in the 24 lectures
- Possible questions
- Next steps
- Questions

Rules of the Game

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.

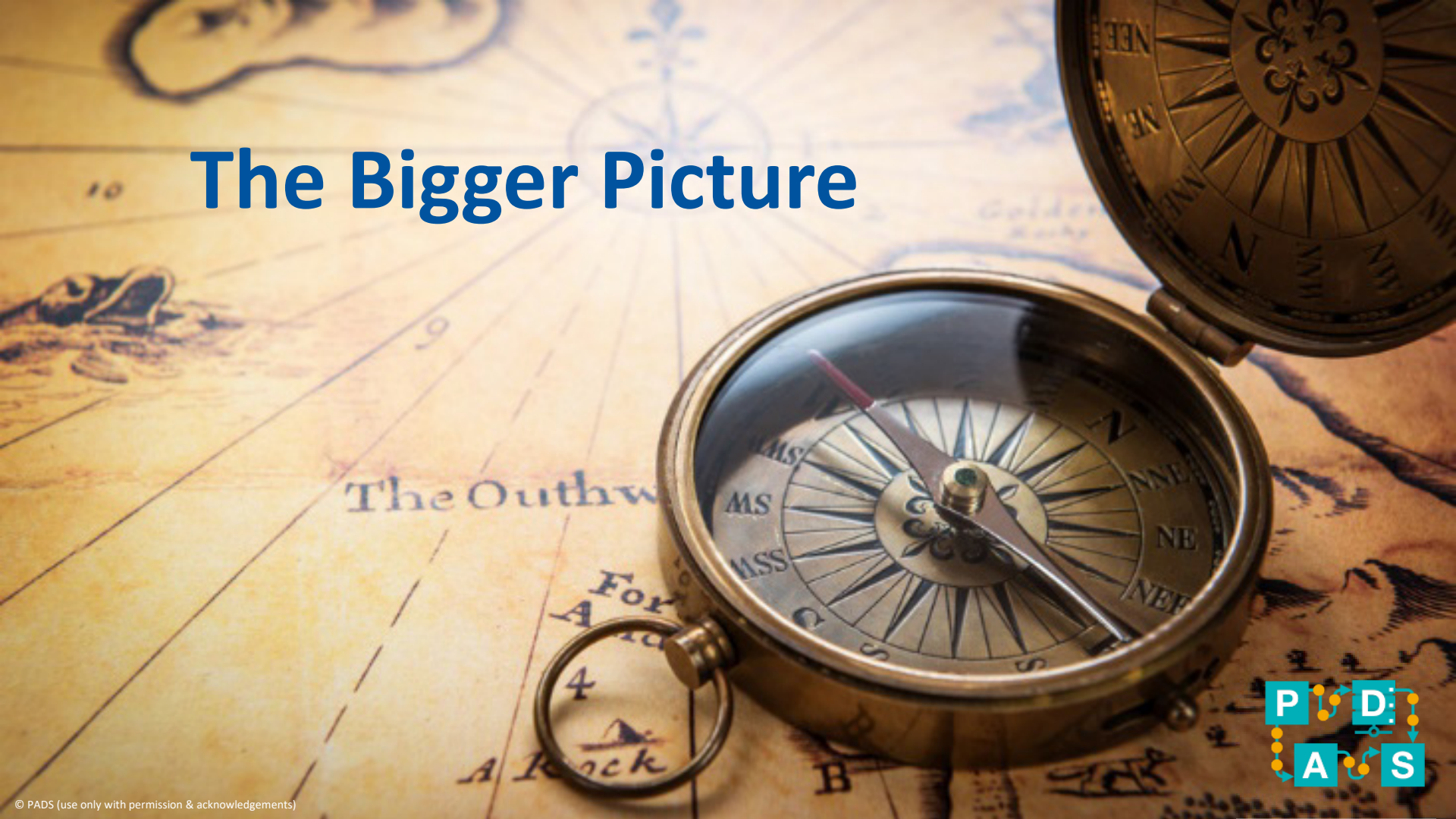
Final written test: 25-02-2019 09:00 – 11:00

- **You can use a basic calculator, but no books, notes, phones, laptops, etc.**

Two instructions remaining

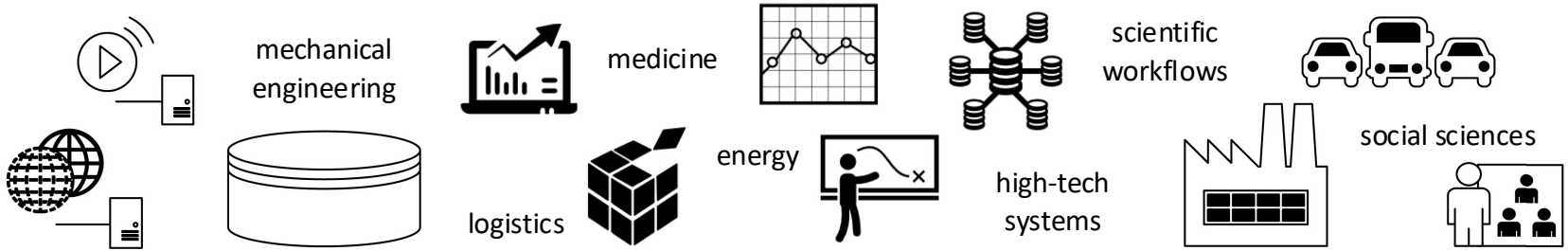
- After this closing lecture, there will be two more instructions (25-1-2019 and 1-2-2019).
- During the first instruction (25-1-2019), two sample exams will be handed out and it is still possible to ask questions.
- During the last instruction (1-2-2019), solutions will be given.
- It is vital to make the exams yourself beforehand (understanding solutions is not the same as creating them).

The Bigger Picture



The data scientist

- Everyone needs her/him.
- No so easy: Combining different skills sets.
- Involves a range of topics and techniques.
- You only saw the tip of the iceberg.



infrastructure

"volume and velocity"

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

analysis

"extracting knowledge"

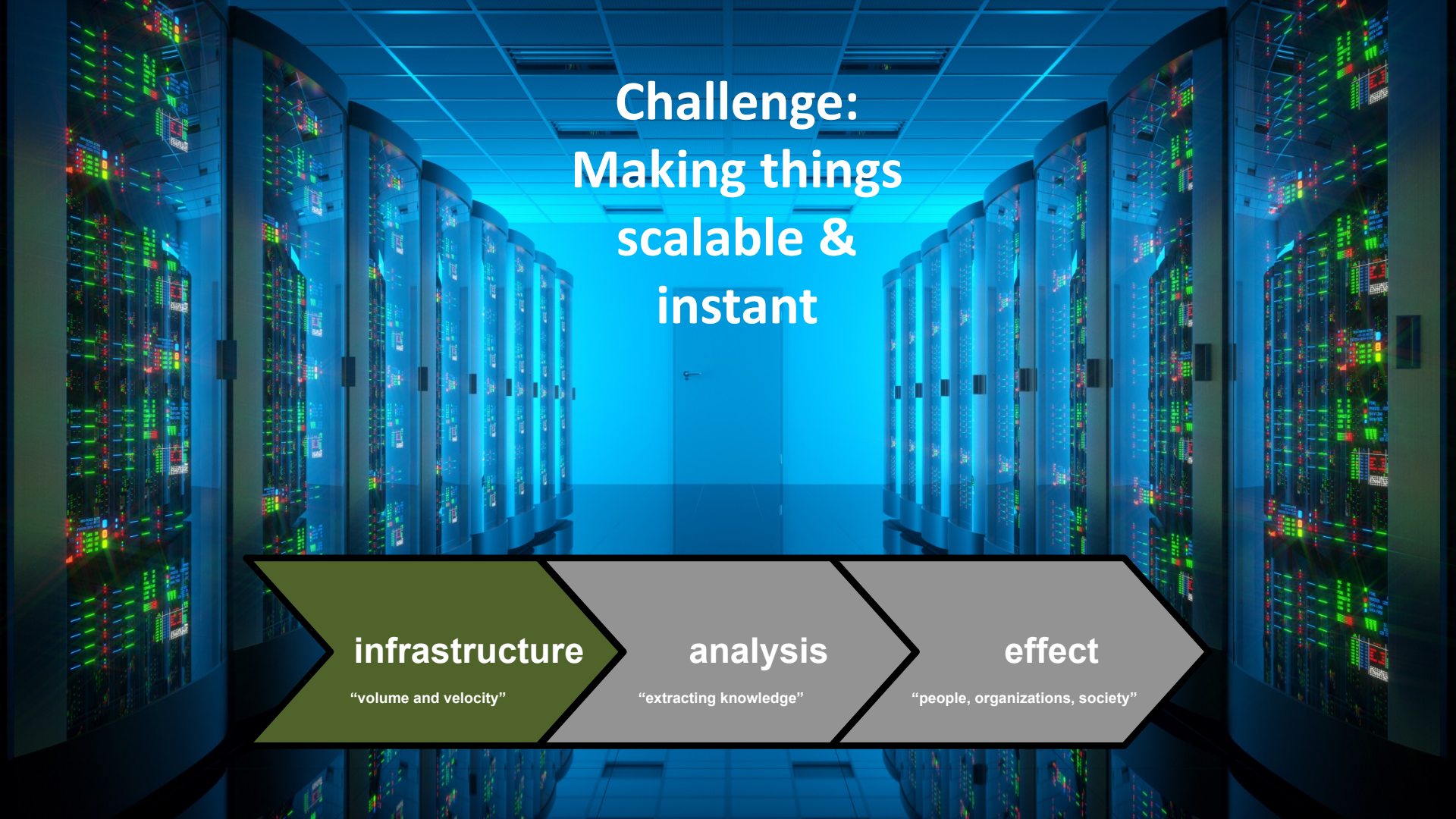
- statistics
- data/process mining
- machine learning
- artificial intelligence
- visualization
- ...

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!

Software for Sheep with Five Legs

- Infrastructure (distributed computing and storage, e.g., Hadoop).
- Programming / scripting languages tailored towards data analysis: Python, R, etc.
- Analysis tools using a visual workflow: Knime, RapidMiner, etc.
- BI tools: Tableau, Qlikview, PowerBI, etc.
- Statistical software: SAS, SPSS, etc.
- Traditional data mining: WEKA, etc.
- Specialized tools: ProM, Disco, Celonis, QPR, etc.

Reminder: Data science is like a making cocktail. Mix the proper ingredients in the right way!

Python
Visualization
Decision trees
Regression
Support vector machines
Neural networks
Evaluation
Clustering
Frequent items sets
Association rules
Sequence mining
Process mining
Process discovery
Conformance checking
Text mining
Preprocessing
Visual analytics
Encryption
Anonymization
Big data infra
Distribution

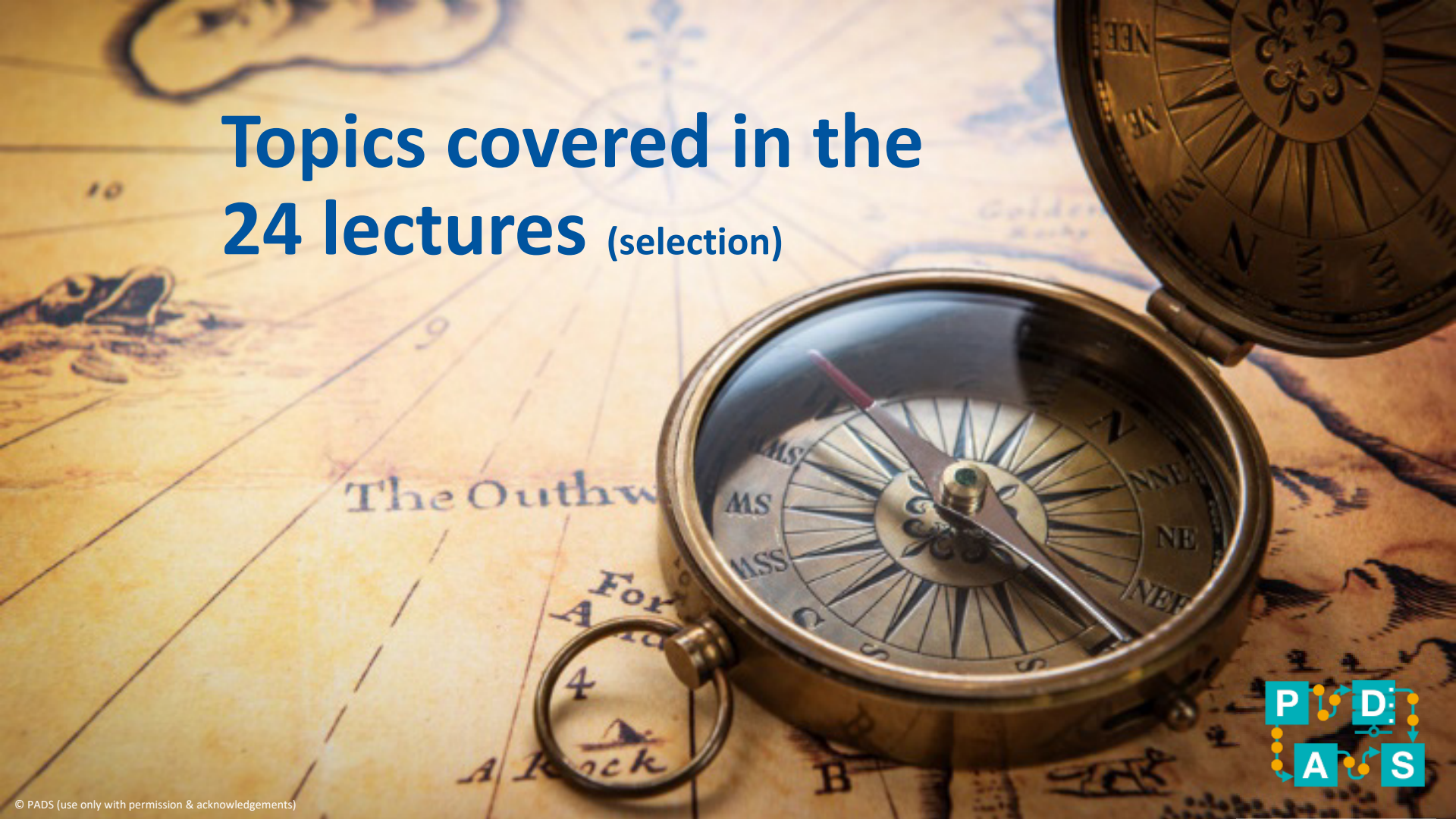
Process mining

Clustering

Text mining

Anonymization

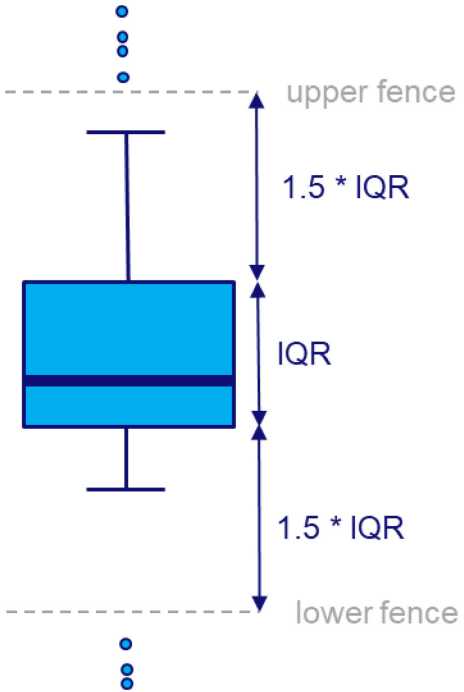
Topics covered in the 24 lectures (selection)

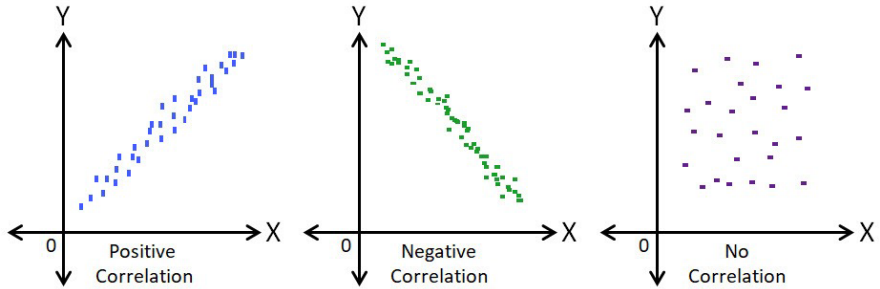


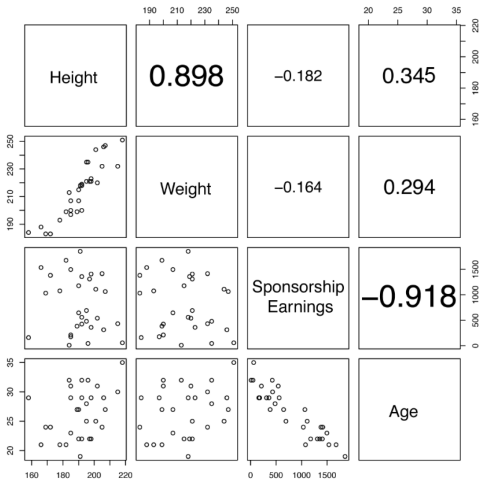
#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2	*Decision trees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

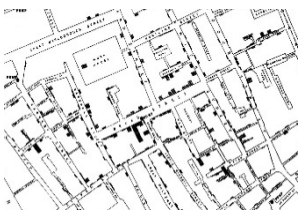
Basic data visualization/exploration

Basic visualizations (boxplots, scatterplots, descriptive statistics)



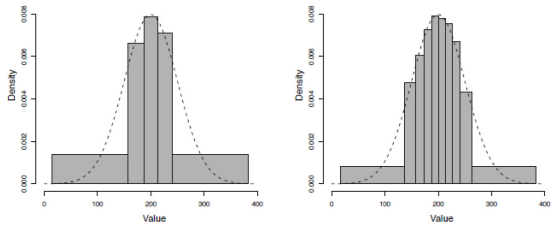




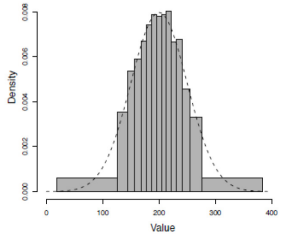


Chair of Process and Data Science

Data preprocessing (binning, sampling, etc.)

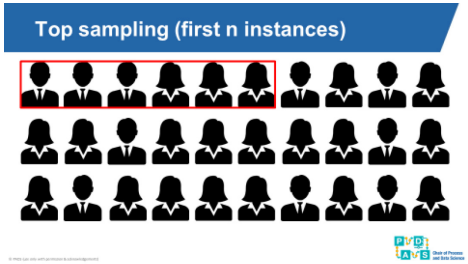


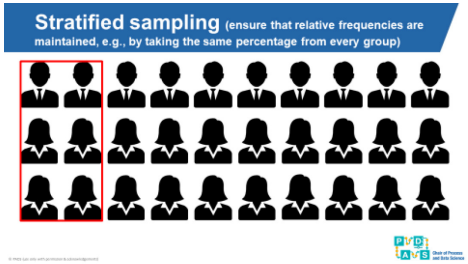
(k) 5 Equal-frequency bins (l) 10 Equal-frequency bins

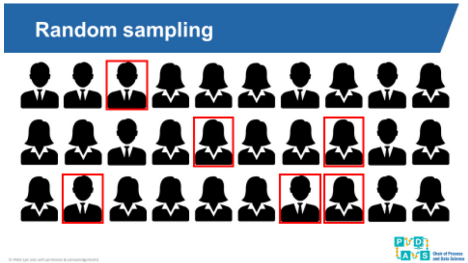


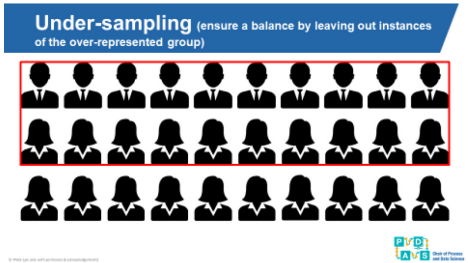
(m) 15 Equal-frequency bins

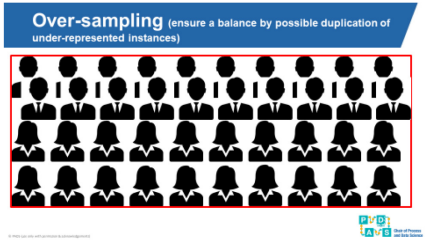
equal width versus equal frequency











Chair of Process and Data Science

Example questions

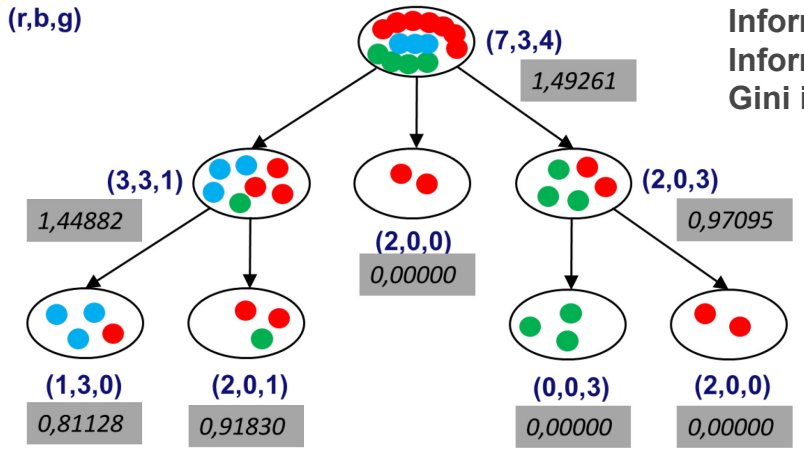
- Interpret boxplots, scatterplots, etc.
- Normalize data (e.g., scale to [0,1]).
- Perform binning on a given data set.

Decision trees

Decision trees (Supervised Learning)

Algorithm 4.1 Pseudocode description of the ID3 algorithm.

Require: set of descriptive features **d**
Require: set of training instances $\mathcal{D}$
1: **if** all the instances in $\mathcal{D}$ have the same target level C **then**
2: **return** a decision tree consisting of a leaf node with label C
3: **else if d** is empty **then**
4: **return** a decision tree consisting of a leaf node with the label of the majority target level in $\mathcal{D}$
5: **else if** $\mathcal{D}$ is empty **then**
6: **return** a decision tree consisting of a leaf node with the label of the majority target level of the dataset of the immediate parent node
7: **else**
8: $\mathbf{d}[best] \leftarrow \arg\max_{d \in \mathbf{d}} IG(d, \mathcal{D})$
9: make a new node, $Node_{\mathbf{d}[best]}$, and label it with $\mathbf{d}[best]$
10: partition $\mathcal{D}$ using $\mathbf{d}[best]$
11: remove $\mathbf{d}[best]$ from **d**
12: **for** each partition $\mathcal{D}_i$ of $\mathcal{D}$ **do**
13: grow a branch from $Node_{\mathbf{d}[best]}$ to the decision tree created by rerunning $ID3$ with $\mathcal{D} = \mathcal{D}_i$



Information gain (IG)
Information gain ratio (GR)
Gini index (Gini)

Pre-pruning (early stopping/forward)
Post-pruning (reduced error/backward)

Focus on categorical variables, but extensions to continuous variables possible.
Many variations possible based on the basic ID3 algorithm (pruning, ensembles, different purity measures, etc.).

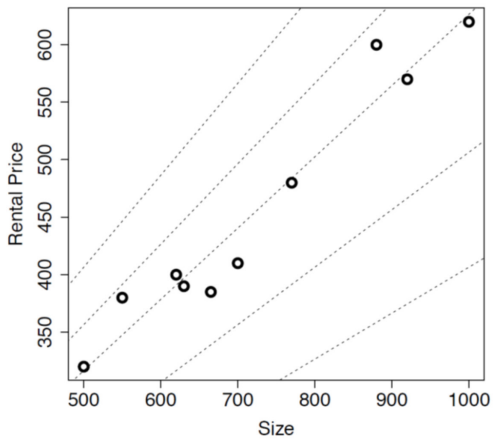
Chair of Process and Data Science

Example questions

- Apply the ID3 algorithm to a data set.
- Answer questions about entropy / information gain.
- Compare Information Gain (IG), Information gain ratio (GR), Gini index (Gini).
- Evaluate the quality of a decision tree, both for a categorical target feature and a numerical target feature.

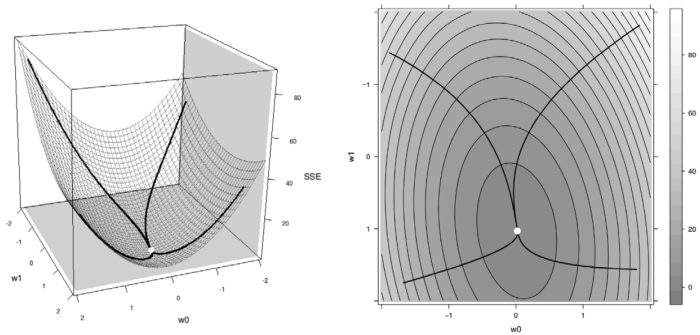
Regression

Linear regression (numerical target feature)



$$L_2(\mathbb{M}_{\mathbf{w}}, \mathcal{D}) = \frac{1}{2}\sum_{i=1}^{n}(t_i - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i[1]))^2$$

w[1]	error
w[1] = 0.80	90978
w[1] = 0.70	20092
w[1] = 0.62	**2837**
w[1] = 0.50	42712
w[1] = 0.40	136218



Minimizing the error while walking downhill

Categorical descriptive features: one-hot-encoding

Chair of Process and Data Science

Simple Linear Regression Example

Formula given at instruction to compute regression for one independent feature.

$$\sum_{i=1}^{M} \left(y_i - (b + mx_i)\right)^2$$

By minimizing this equation we get the least squares estimates for m and b.

$$b = \bar{y} - m\bar{x}$$

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

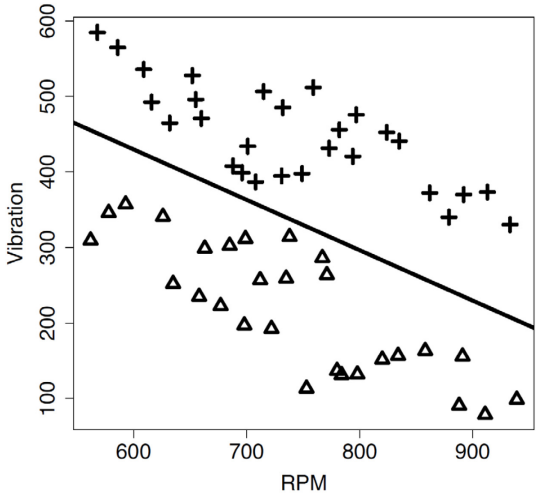
y_i : observed target value for data point i

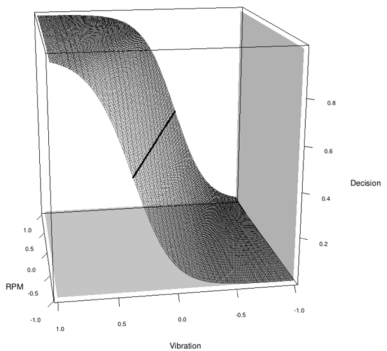
x_i : value of independent variable of data point i

$\bar{y}$: mean of all observed y-values

$\bar{x}$: mean of all observed x-values

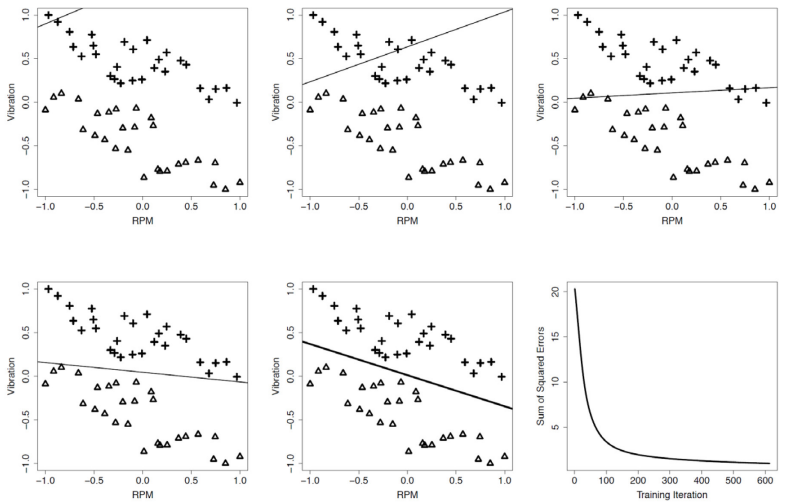
Logistic regression (categorical target feature)

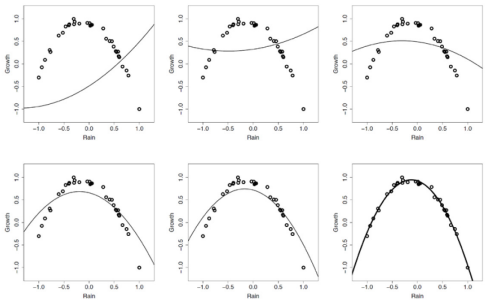




$$\mathbb{M}_{\mathbf{w}}(\mathbf{d}) = Logistic(\mathbf{w} \cdot \mathbf{d}) = \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{d}}}$$

sigmoid function is a squashing function





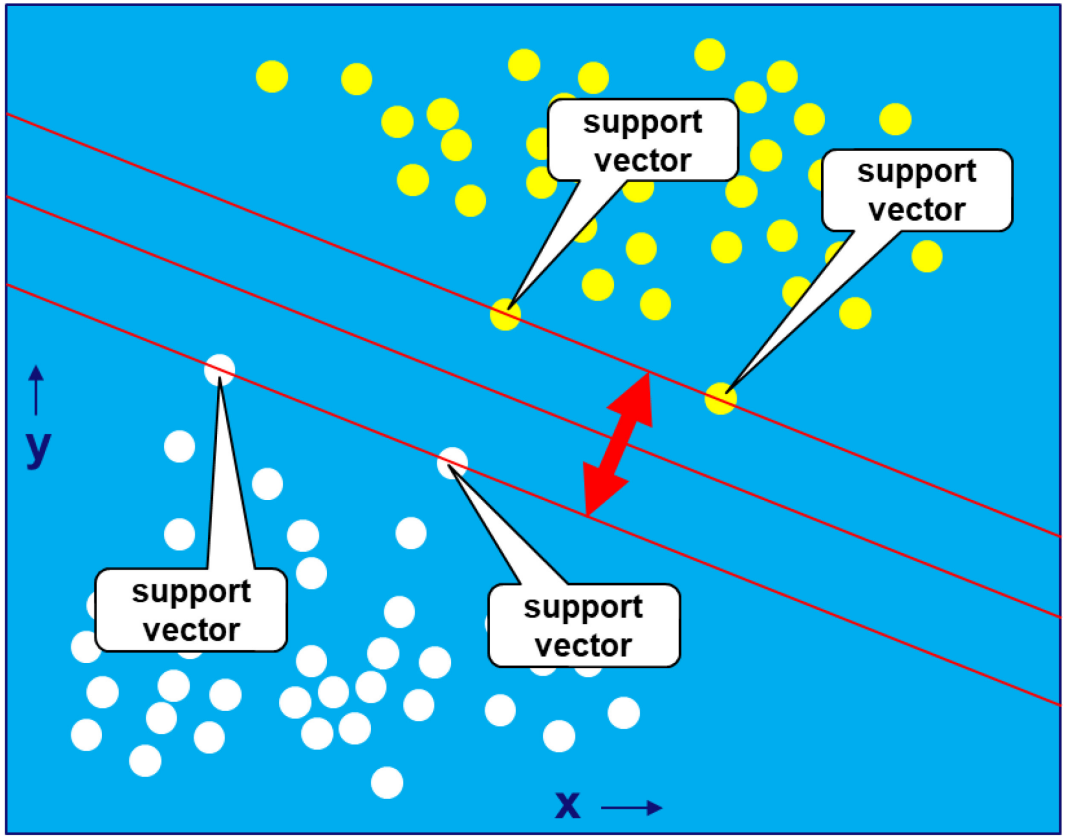
To handle non-linear relationships, the data is transformed before the "linear machinery" is used

Possible questions

- Reason about the error in regression.
- Perform linear regression on a small data set.
- Interpret the outcome of linear or logistic regression.

Support vector machines

Support vector machines



Given a set of m instances

$$\{(\vec{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} | 1 \leq i \leq m\}$$

$$\min_{\vec{w},b} \frac{1}{2} \|\vec{w}\|^2 \text{ such that}$$

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \text{ for any } i$$

With soft margin:

Given a set of m instances

$$\{(\vec{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} | 1 \leq i \leq m\}$$

$$\min_{\vec{w},b,\epsilon} \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{m} \epsilon_i$$

such that

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \epsilon_i \text{ for any } i$$

Chair of Process and Data Science

Possible questions

- Be able to identify the support vectors.
- Able to explain the role of the soft margin.
- Given a visualization of a data set with one target feature and two descriptive features, be able to explain the difference between SVMs and logistic regression.

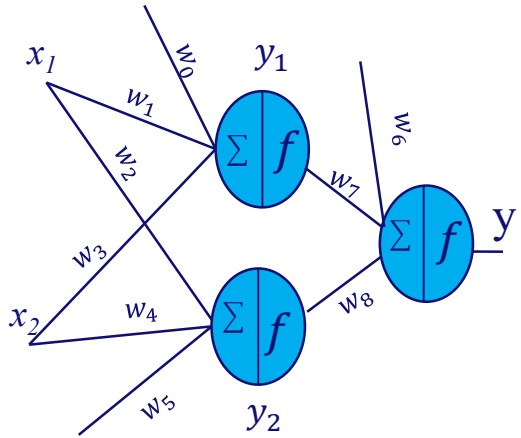
Neural networks

Example

Logical XOR function

x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0

Hidden layer of neurons



$$f(w_1 * x_1 + w_3 * x_2 + w_0) = y_1$$

$$f(w_2 * x_1 + w_4 * x_2 + w_5) = y_2$$

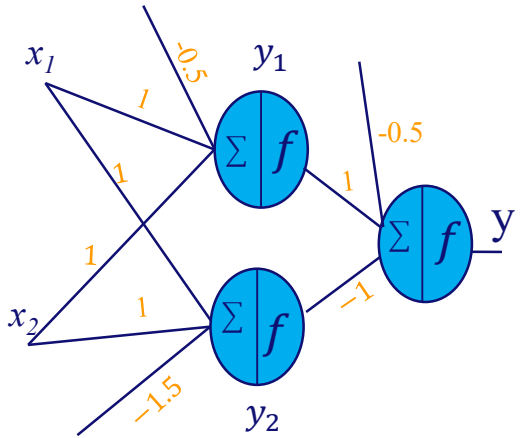
$$f(w_7 * y_1 + w_8 * y_2 + w_6) = \mathrm{y}$$

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

Multiple neurons are needed to encode XOR!

Chair of Process and Data Science

Example



$w_0 = -0.5$
$w_1 = 1$
$w_2 = 1$
$w_3 = 1$
$w_4 = 1$
$w_5 = -1.5$
$w_6 = -1$
$w_7 = 1$
$w_8 = -0.5$

$$f(1 * x_1 + 1 * x_2 - 0.5) = y_1 \quad \text{“or”}$$

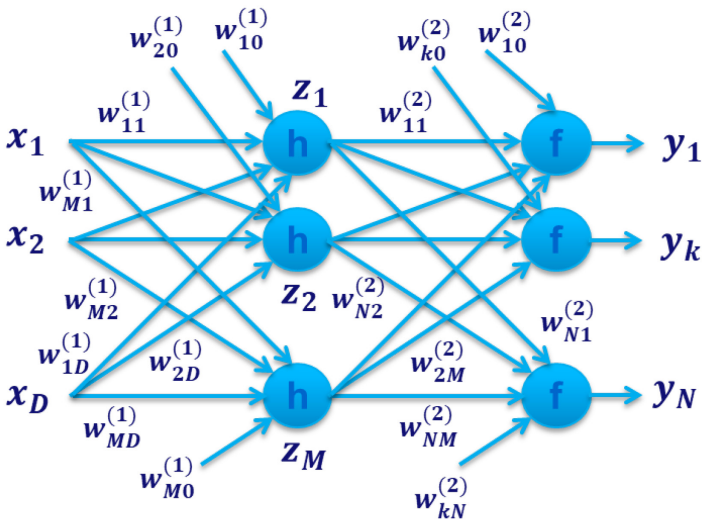
$$f(1 * x_1 + 1 * x_2 - 1.5) = y_2 \quad \text{“and”}$$

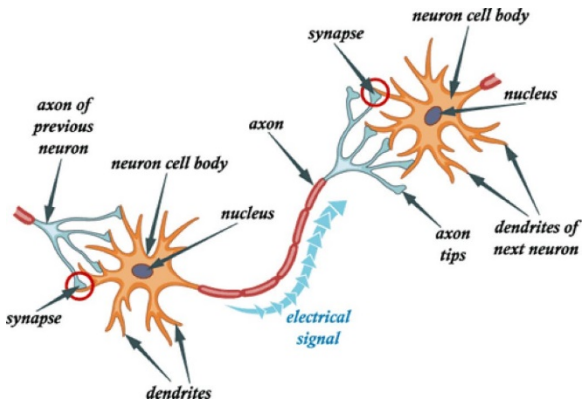
$$f(1 * y_1 - 1 * y_2 - 0.5) = y \quad \text{“and not”}$$

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

Neural networks

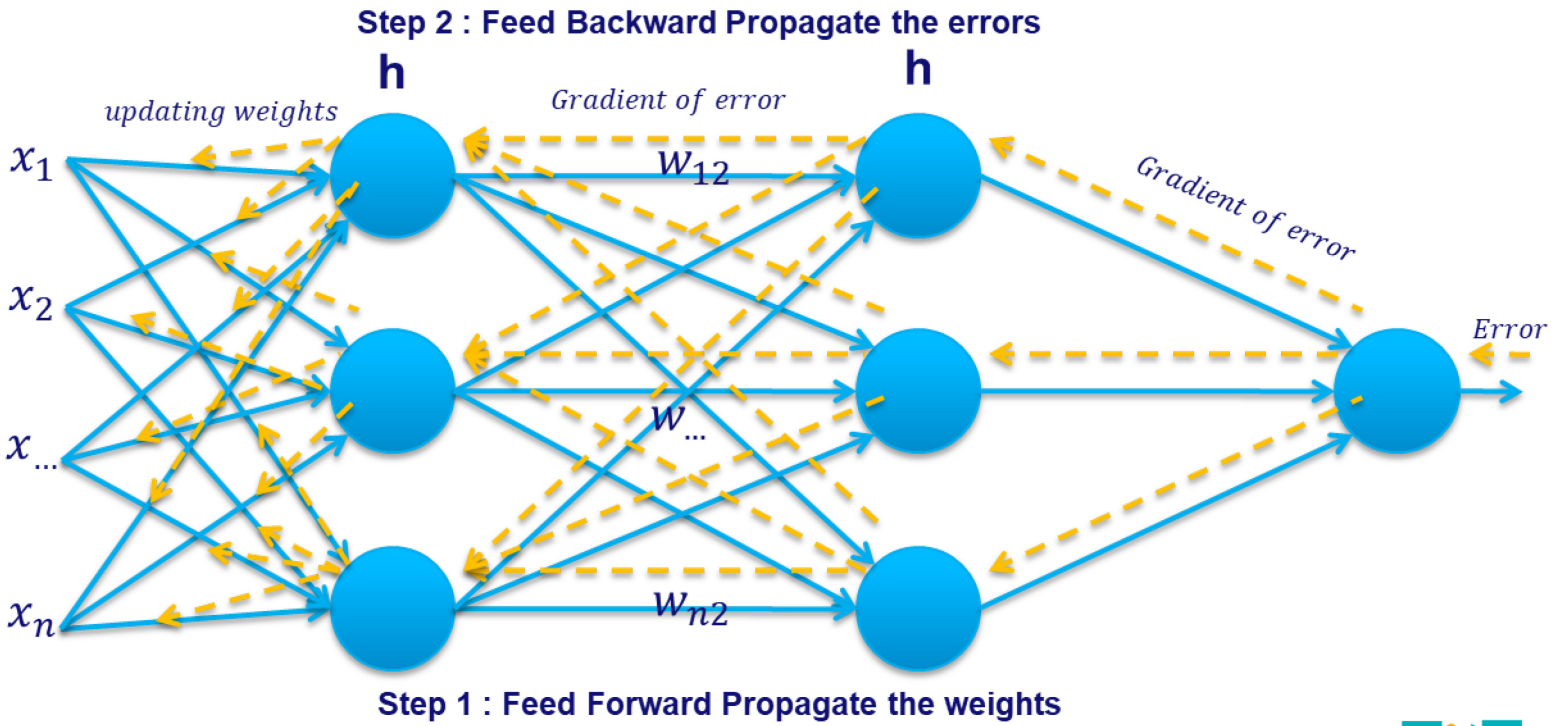
$$y_k(\boldsymbol{x},\boldsymbol{w}) = f\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$





Chair of Process and Data Science

Backpropagation



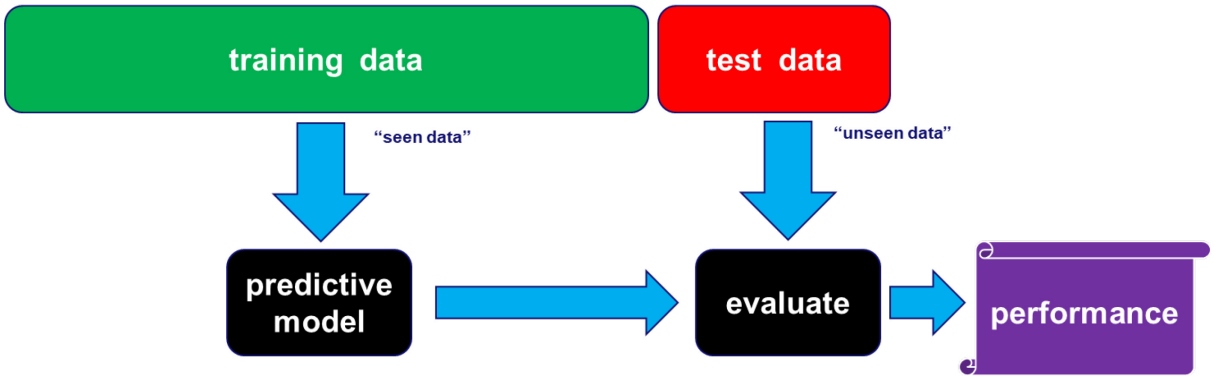
Chair of Process and Data Science

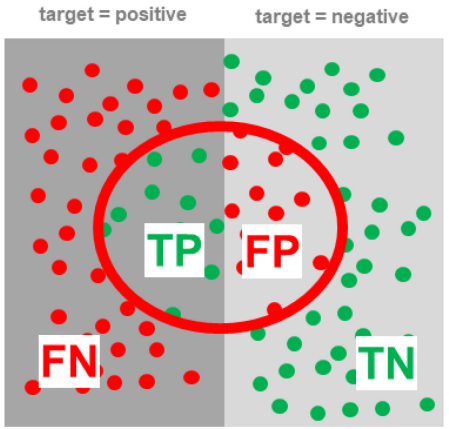
Example questions

- Provide a small neural network to perform a given Boolean function (how many neurons are needed?).
- Show that operators like the XOR cannot be encoded in a single neuron.
- Given a neural network, provide the predicted value and/or compute the error.

Evaluation of supervised learning problems

Evaluation of supervised learning problems





Always test on unseen data to avoid overfitting the training data

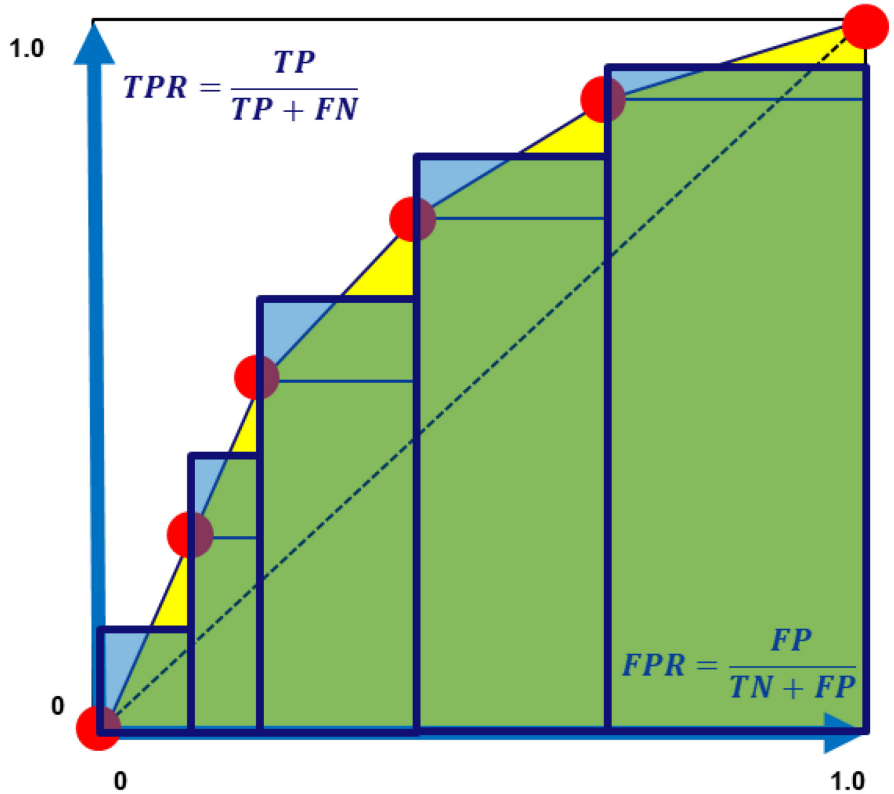
Precision: $precision = \frac{TP}{TP+FP}$

Recall: $recall = \frac{TP}{TP+FN}$

F1-measure, accuracy, etc.

Chair of Process and Data Science

Area Under the Curve (AUC)



(a) Threshold: 0.75

		Prediction 'spam'	Prediction 'ham'
Target	'spam'	4	4
	'ham'	2	10

(b) Threshold: 0.25

		Prediction 'spam'	Prediction 'ham'
Target	'spam'	7	2
	'ham'	4	7

confusion matrix using different parameters

Measuring error (numerical)

$$\text{root mean squared error} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - \mathbb{M}(\mathbf{d}_i))^2}{n}}$$

$$\text{mean absolute error} = \frac{\sum_{i=1}^{n} abs(t_i - \mathbb{M}(\mathbf{d}_i))}{n}$$

Torture the data, and it will confess to anything!

Ronald Harry Coase (1910-2013)
British economist.

$$R^2 = 1 - \frac{\text{sum of squared errors}}{\text{total sum of squares}}$$

$$\text{sum of squared errors} = \frac{1}{2}\sum_{i=1}^{n}(t_i - \mathbb{M}(\mathbf{d}_i))^2$$

$$\text{total sum of squares} = \frac{1}{2}\sum_{i=1}^{n}(t_i - \bar{t})^2$$

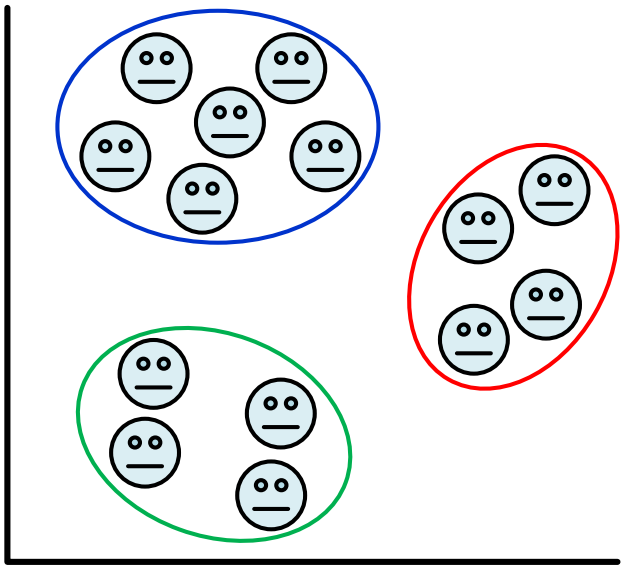
Example questions

- Explain concepts such as overfitting, etc.
- Give the confusion matrix.
- Compute precision, recall, F1-score, etc.
- Understand the problems when the classes are very unbalanced. ("majority vote" becomes a good predictor)
- Interpret the Area Under the Curve (AUC).
- Compute Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2.

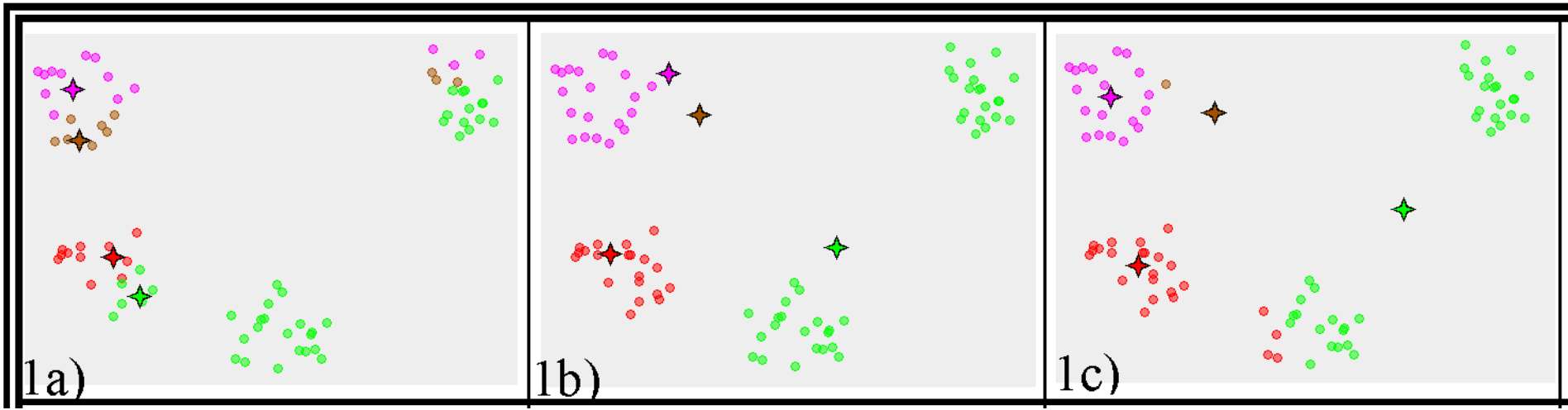
Clustering

Clustering

- k-means clustering
- k-medoids clustering
- Agglomerative Hierarchical Clustering
- Density-based clustering using DBSCAN
- Self-organizing maps

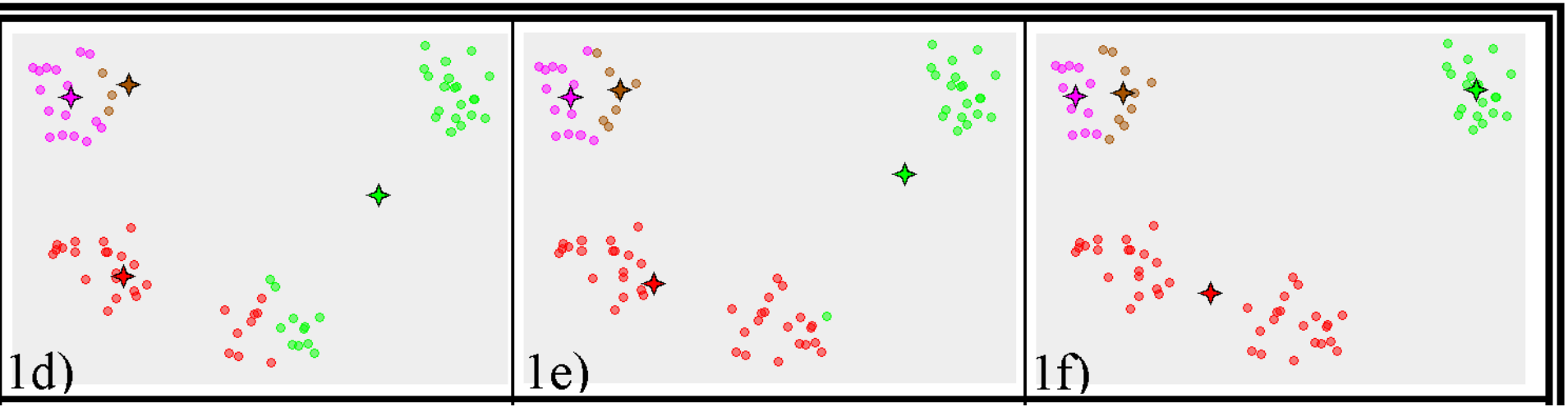


Example *k*-means



The illustration was prepared with the Java applet, E.M. Mirkes, K-means and K-medoids: applet. University of Leicester, 2011.

Example *k*-means



The illustration was prepared with the Java applet, E.M. Mirkes, K-means and K-medoids: applet. University of Leicester, 2011.

Example questions

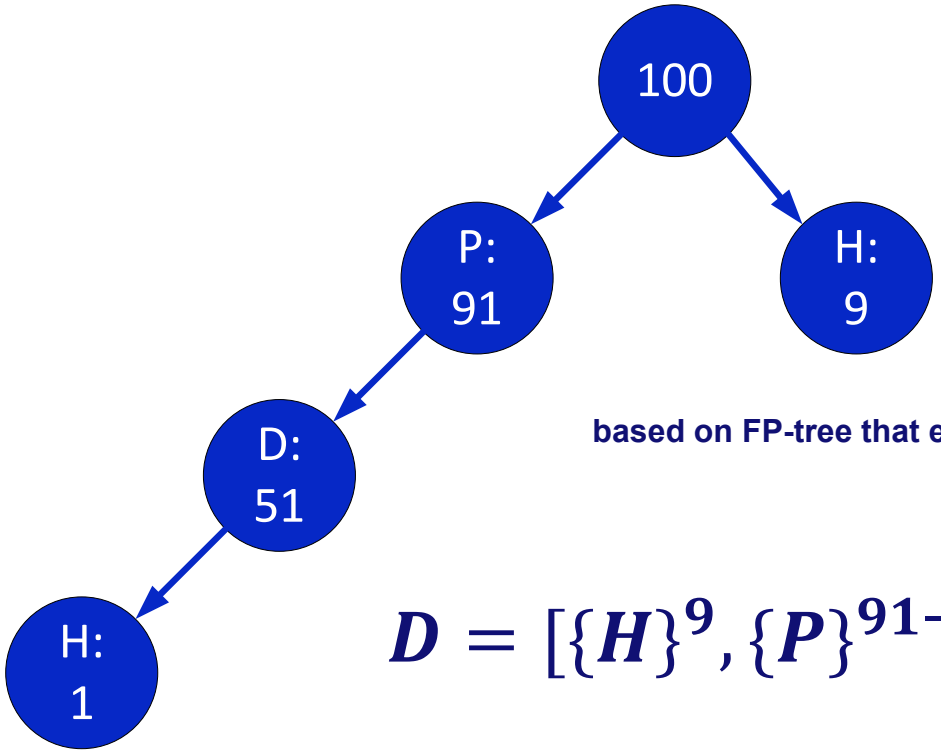
- Apply k-means to a given data set.
- Apply k-medoids to a given data set.
- Given the n-th iteration of the k-means/k-medoids, give the outcome of the next iteration (which elements belong to which cluster and what are the new centroids/medoids).
- Describe the differences between the different clustering techniques (e.g., when to use DBscan rather than k-means).

Frequent items sets

Apriori algorithm

- {a,b,c} can only be frequent if {a},{b},{c},{a,b},{a,c}, and {b,c} are all frequent.
- This can be exploited as follows: It is possible to compute the C_k candidates from the previous generation L_{k-1} and prune C_k based on L_{k-1}.

FP-growth algorithm



based on FP-tree that encodes the whole data set

$$D = [\{H\}^{9}, \{P\}^{91-51}, \{P, D\}^{51-1}, \{P, D, H\}^{1-0}]$$

Chair of Process and Data Science

Possible questions

- Apply the apriori algorithm to a given dataset.
- Compute the C_k candidates from the previous generation L_{k-1} and prune C_k based on L_{k-1}.
- Give the FP-tree

Association rules

Support, Confidence, and Lift

$$support(A \Rightarrow B) = support(A \cup B) = \frac{support_{count}(A \cup B)}{support_{count}(\emptyset)}$$

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{support_{count}(A \cup B)}{support_{count}(A)}$$

$$lift(A \Rightarrow B) = \frac{support(A \cup B)}{support(A) \cdot support(b)} = \frac{P(A \cup B)}{P(A) \cdot P(B)}$$

Compute using frequent item sets

{Bitburg, Pumpernickel} ⇒ {Merlot}
(people that buy Bitburg pilsner and Pumpernickel bread also tend to buy Merlot wine)

{Bitburg} ⇒ {Heineken, Palm}
(people that buy Bitburg pilsner tend to buy both Heineken and Palm pilsner)

{Carbonara, Margherita } ⇒ {Espresso, Tiramisu}
(people that buy Bitburg pilsner and Pumpernickel bread, also tend to buy Merlot wine)

{BPI, IDS} ⇒ {APM}
(students that take the BPI and IDS courses also tend to take the APM course)

{part-245, part-345, part-456} ⇒ {part-372}
(when parts 245, 345, and 456 are replaced, then often also part 372 is replaced)

Possible questions

- Use apriori to compute frequent itemsets (as before) to create association rules.
- Compute support, confidence, and lift for given rules or select a rule based on this (e.g., provide a rule with support, confidence, and lift above a given level).

Sequence mining

Sequence mining

Ideas from frequent item sets and association rules applied to temporal data

$$[\langle\{a\},\{a,b\},\{b,c\},\{c\}\rangle,\langle\{a\},\{a\},\{a,b\},\{b,c\},\{c\}\rangle,$$
$$\langle\{a\},\{a\},\{a\},\{a\}\rangle,\langle\{a,b\},\{a,b\},\{a,b\},\{a,b\}\rangle,$$
$$\langle\{a,b\},\{b,c\},\{c,d,e\}\rangle,\langle\{a\},\{a,b\},\{b,c\},\{c\}\rangle]$$

Can use a-priori style algorithm

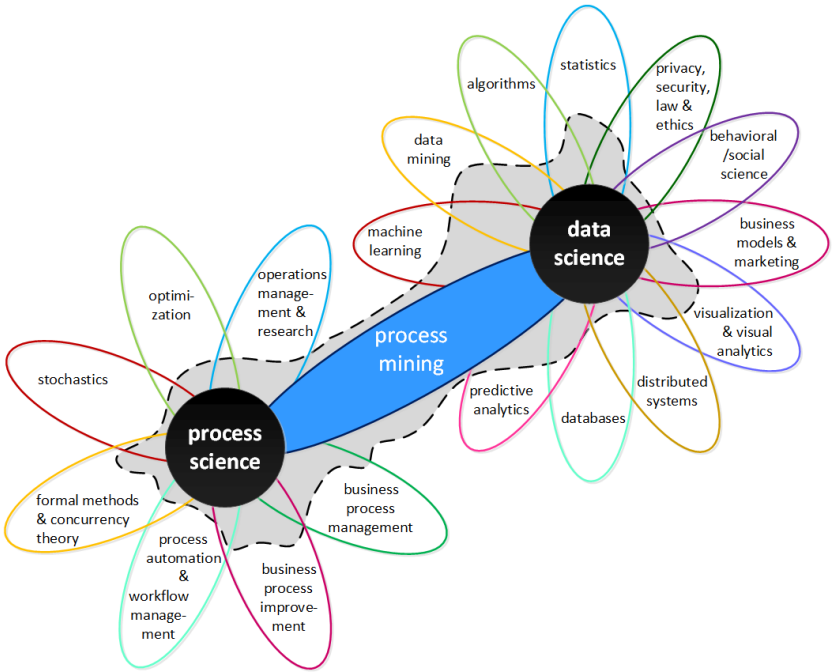
$$\langle\{beer\},\{red,white\}\rangle \Rightarrow \langle\{beer\},\{red,white\},\{wodka\}\rangle$$

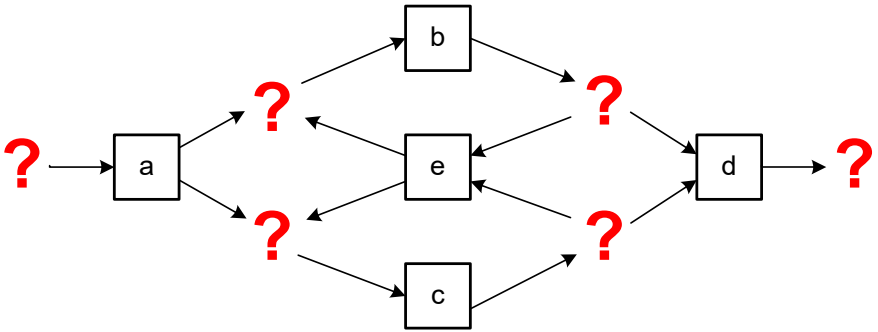
Possible questions

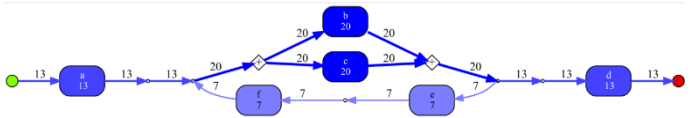
- Determine whether a subsequence is frequent or not.
- Compute support, confidence, and lift for a given rule.

Process mining

Process Mining







3x a b c d
4x a c b d
2x a b c e f b c d
2x a c b e f b c d
1x a b c e f c b d
1x a c b e f b c e f c b d

13x a

11x b c
9x c b

3x b c
4x c b
2x b c e f b c
2x c b e f b c
1x b c e f c b
1x c b e f b c e f c b

13x d

7x e f

20x b

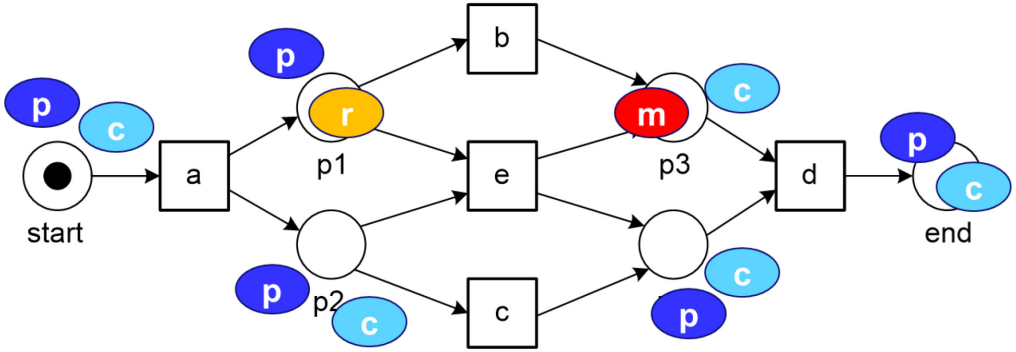
20x c

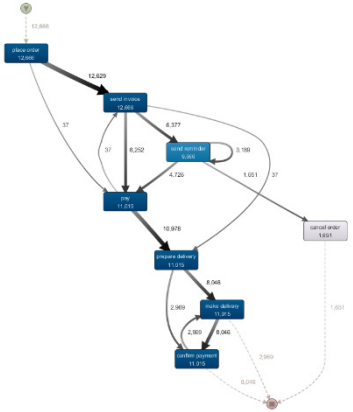
7x e

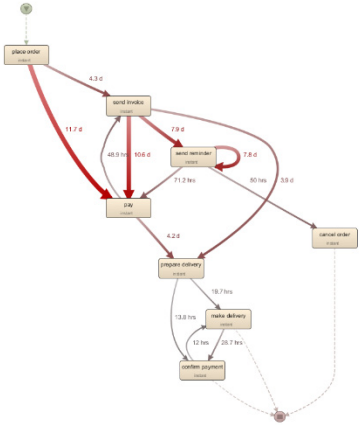
7x f

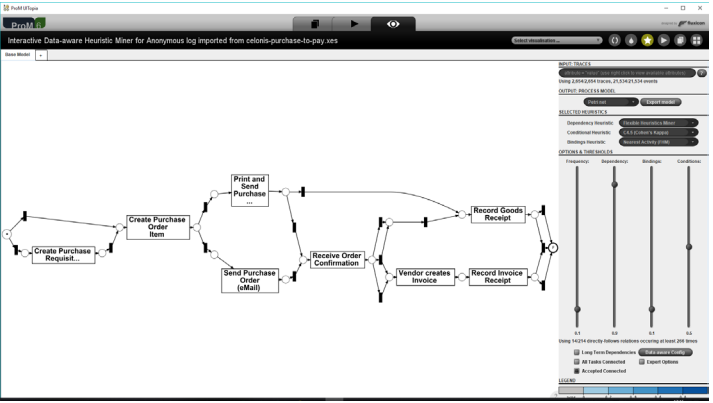
Chair of Process and Data Science

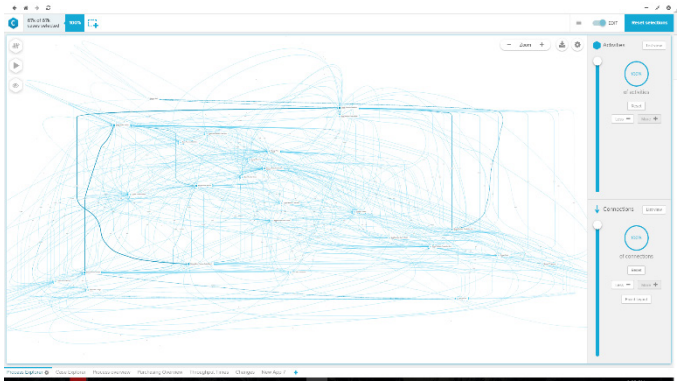
Process Mining











Chair of Process and Data Science

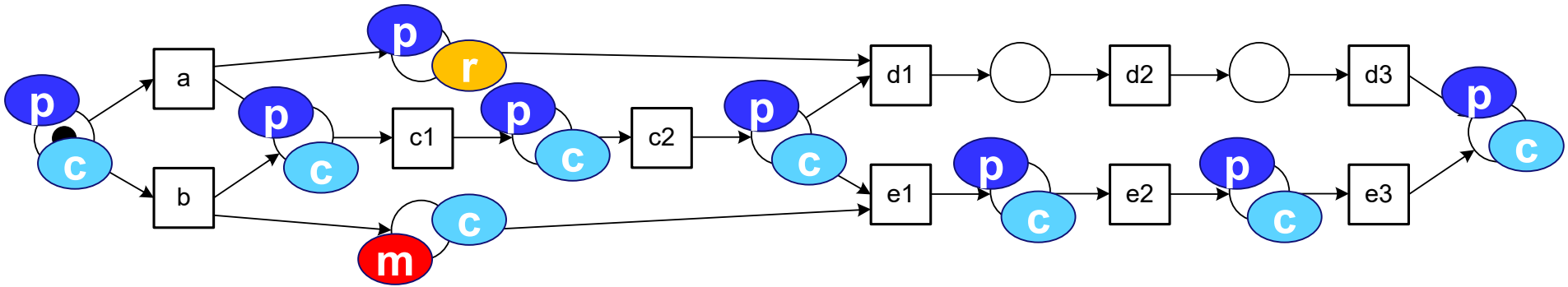
Possible questions

- Transform event data into multisets of traces.
- Apply the Inductive Mining algorithm to a small event log.
- Reason about the possibility to have a particular place in a Petri net given a log.
- Given a Petri net and an event log, compute fitness using produced, consumed, missing, and remaining tokens.

Example (just one trace, you should also be able to do this on a small event log)

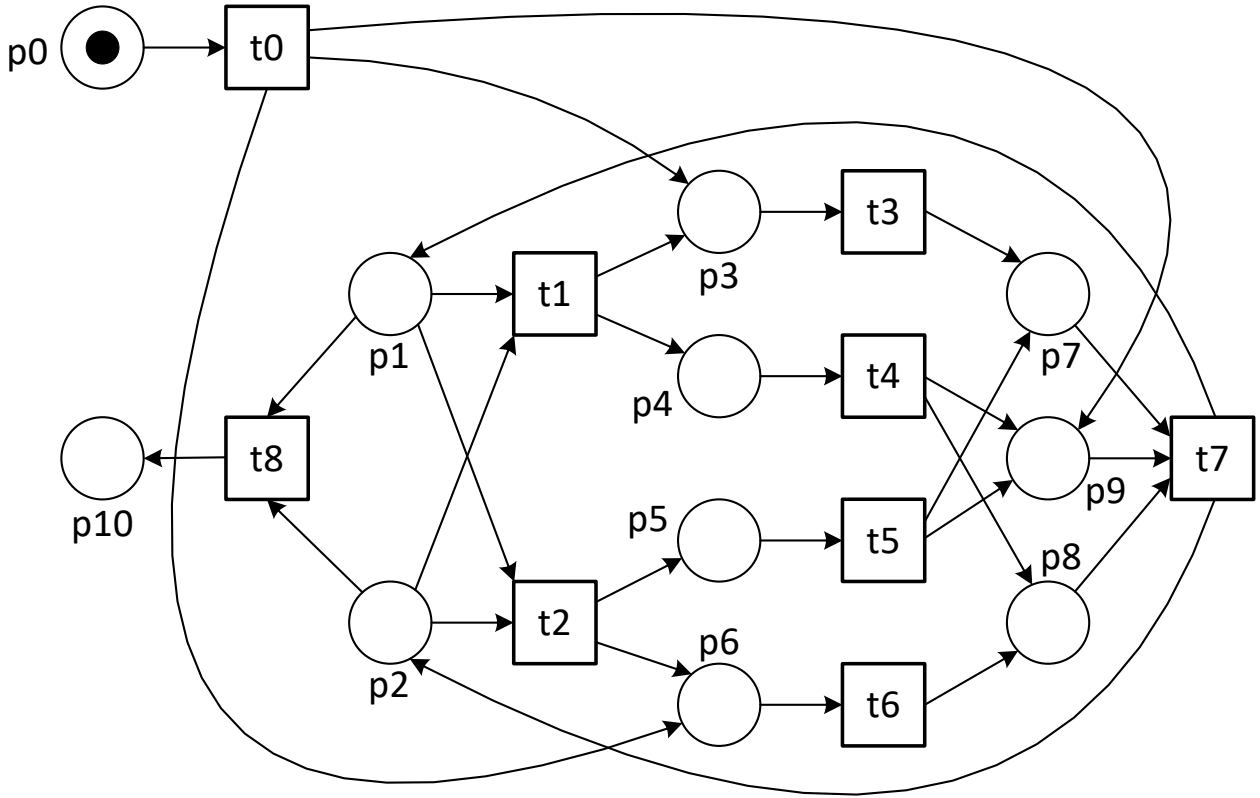
p = 8
c = 8
m = 1
r = 1
f = 0.875

⟨a,c1,c2,e1,e2,e3⟩



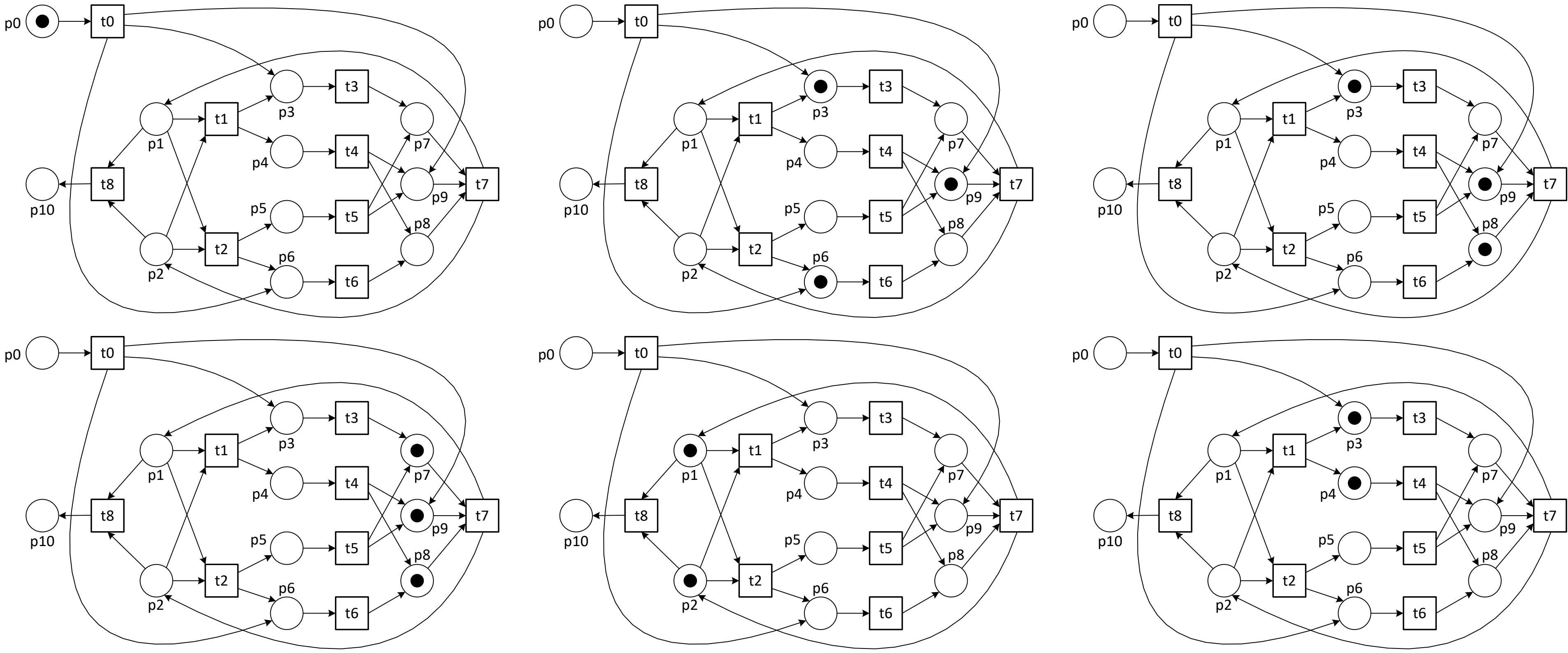
Chair of Process and Data Science

Understanding a complex Petri net



Chair of Process and Data Science

Understanding a complex Petri net



Text Mining

Text mining

Necessary to preprocess text to get structured data.

1. **Dividing the text in discrete units (tokenization)**
2. **Removing irrelevant tokens (stopwords)**
   (remove 'the', 'as', 'in', 'me', 'you', 'which', 'on', etc.)
3. **Normalize the tokens so that the same concept is always represented by the same token**
   1. **Represent concepts with stems**
   2. **Represent concepts with lemmas**

Word	Stem	Lemma
bakery	baker	bake
bakeries	baker	bake
police	polic	police
policy	polic	policy
numerical	numer	numerical

$$tf(w,d) = \#of\ occurrences\ of\ word\ w\ in\ document\ d$$

$$idf(w) = log_2(\frac{N}{\#of\ documents\ that\ contain\ w\ at\ least\ once})$$

$$tfidf(w,d) = tf(w,d) * idf(w)$$

bag of words

N-grams

2-gram
If you wait too long for the perfect moment, the perfect moment will pass you by.

3-gram
If you wait too long for the perfect moment, the perfect moment will pass you by.

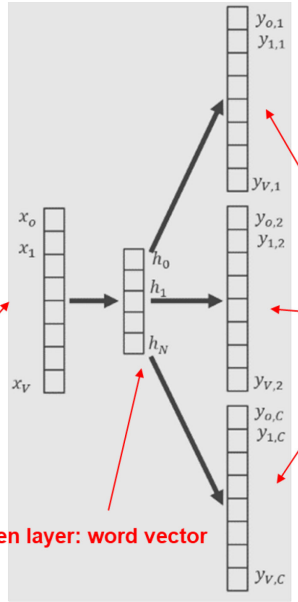
4-gram
If you wait too long for the perfect moment, the perfect moment will pass you by.

5-gram
If you wait too long for the perfect moment, the perfect moment will pass you by.

6-gram
If you wait too long for the perfect moment, the perfect moment will pass you by.

Chair of Process and Data Science

word2vec

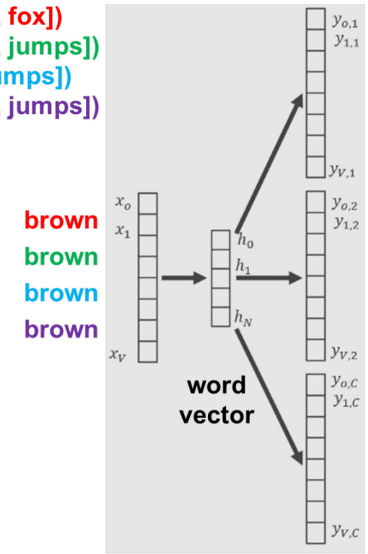


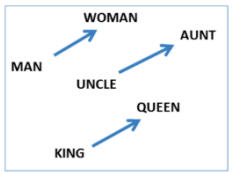
Input word w (one-hot encoded) as training data

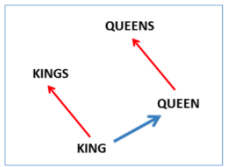
Hidden layer: word vector

Target: context of the word w (skip-gram)

(brown, [the, quick, fox])
(brown, [the, quick, jumps])
(brown, [the, fox, jumps])
(brown, [quick, fox, jumps])







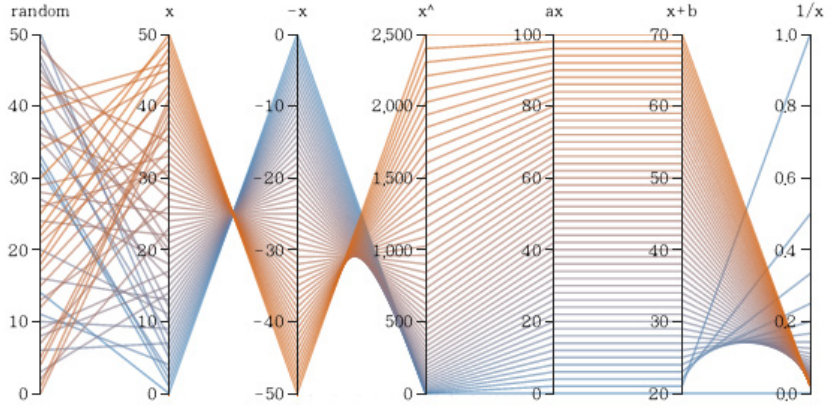
Chair of Process and Data Science

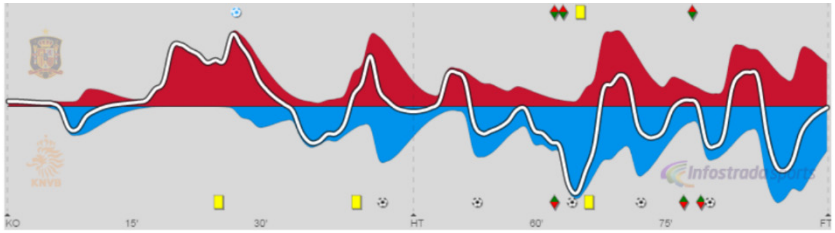
Possible questions

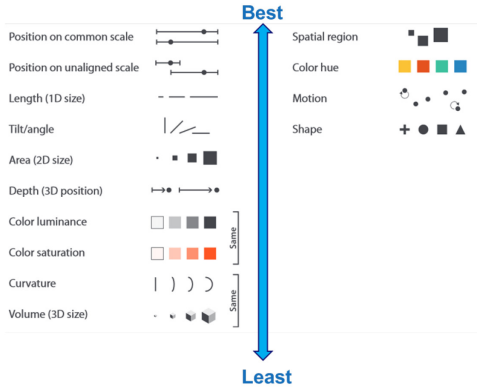
- Given a set of documents and a word (or a query), compute the tf-idf score.
- Compute the probability of an n-gram given a set of documents.

Data preprocessing and visualization

Data preprocessing and visualization





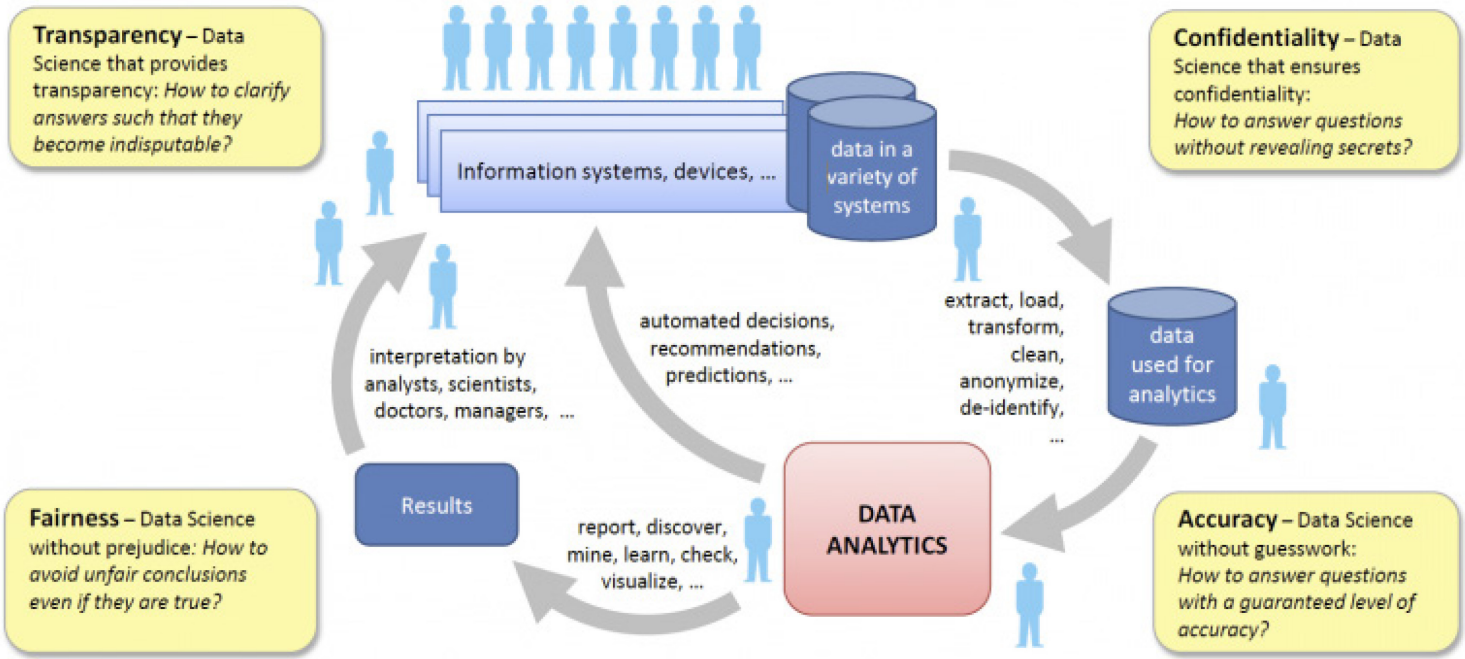


Possible questions

- What functions best as a magnitude channel?
- What functions best as an identity channel?
- Name graphical primitives (position, length, tilt, area, depth, color, curvature, volume, shape, etc.) that do not work well in certain situations.

Responsible data science

Responsible data science



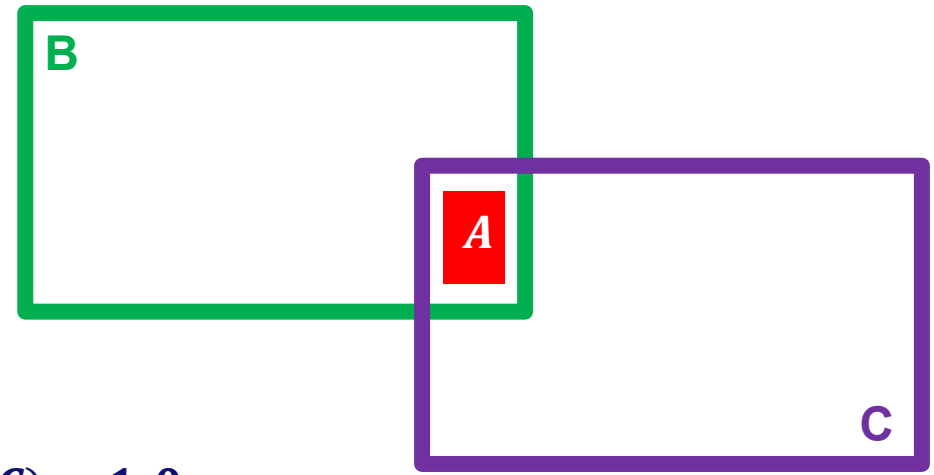
Chair of Process and Data Science

Fairness

How to measure? How to build fair trees? How to make trees more fair?

$$conf(A, B \rightarrow C) = 1.0$$

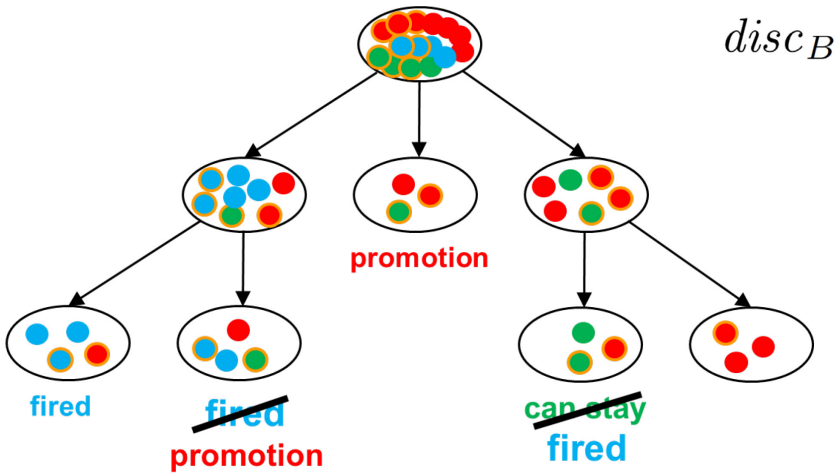
$$conf(B \rightarrow C) = 0.1$$



$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{1.0}{0.1} = 10$$

Fairness

How to measure? How to build fair trees? How to make trees more fair?



$$disc_B(C, D) := \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

Two approaches:

- **Balancing information gain and discrimination while building the tree.**
- **Selectively changing labels to minimize loss in classification error and discrimination.**

Confidentiality

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Quasi-identifiers

Sensitive attribute

Equivalence class
With **k=3** records

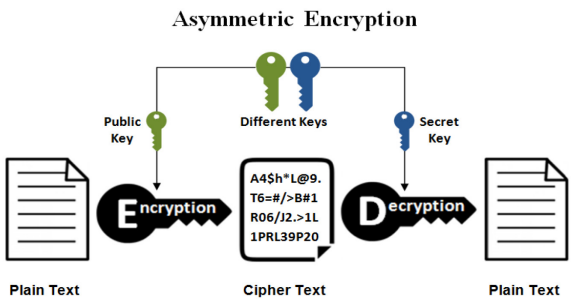
	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

(partial) suppression

generalization

(partial) suppression

k-anonymity
l-diversity
t-closeness



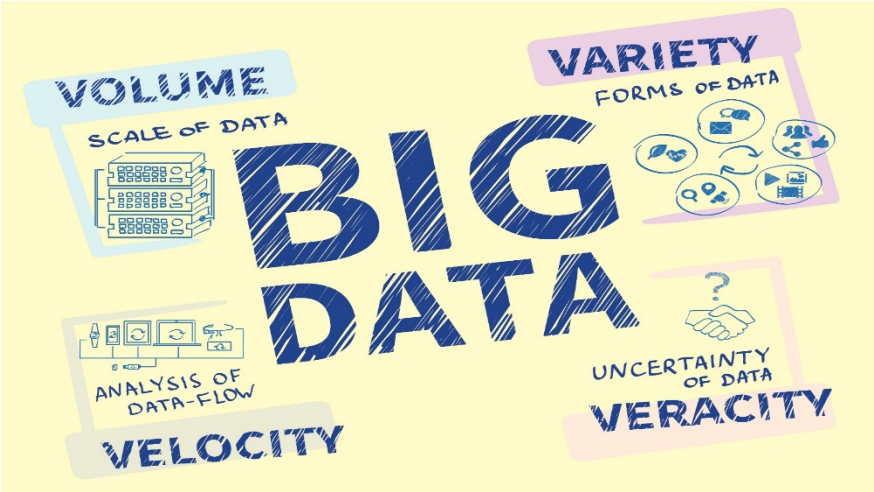
Chair of Process and Data Science

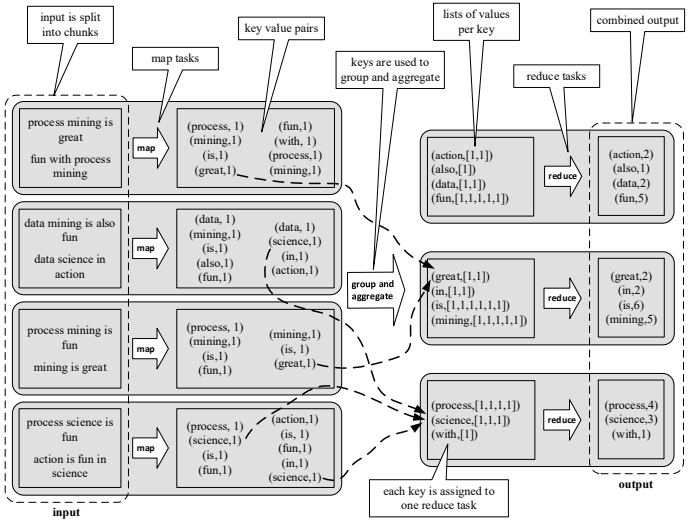
Possible questions

- Compute the level of discrimination of a base rule using elift().
- Compute the level of discrimination using the difference in fractions getting a positive outcome (disc(C,D)).
- Relabel a given decision tree to reduce discrimination.
- Given a data set determine k-anonymity, distinct l-diversity, entropy l-diversity, recursive (c,l)-diversity.

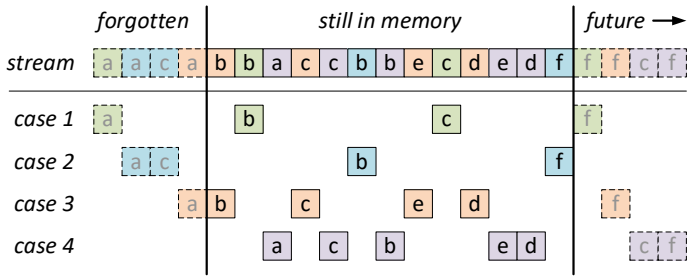
Big data

Big data





MapReduce



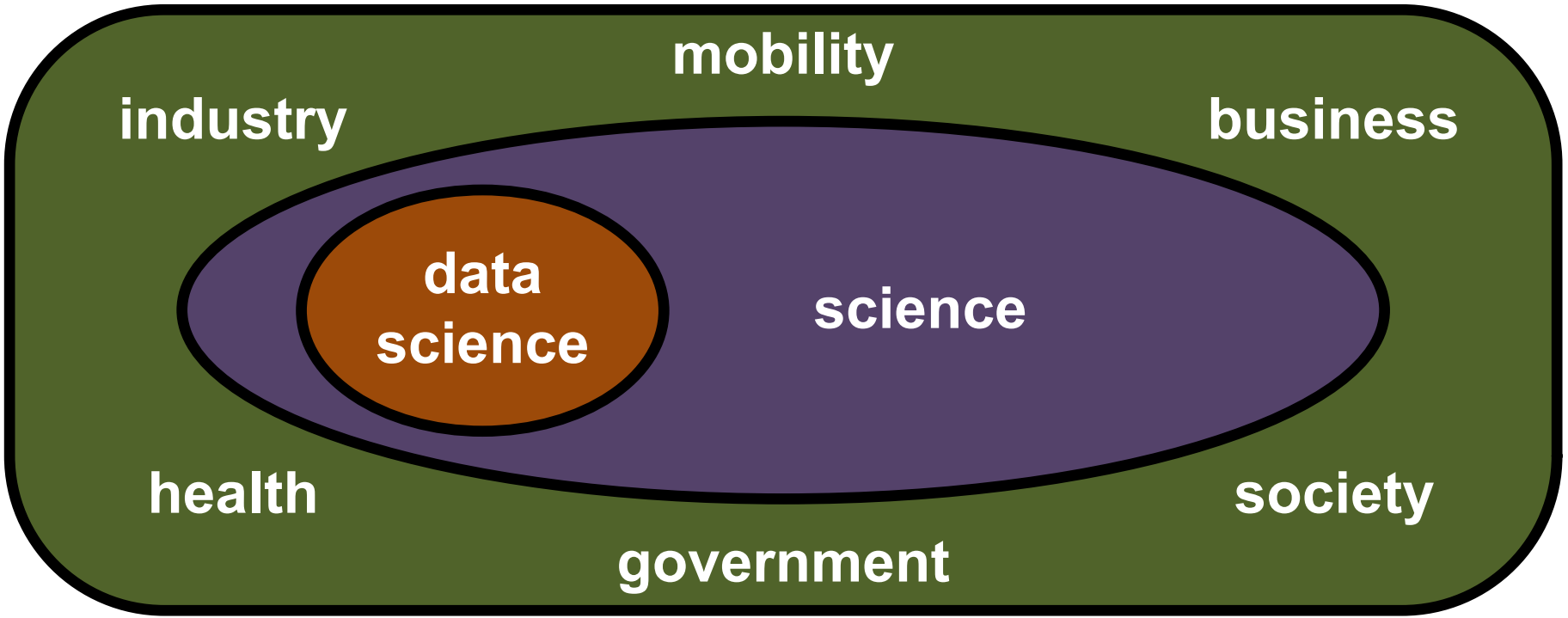
Streaming

Chair of Process and Data Science

Next Steps

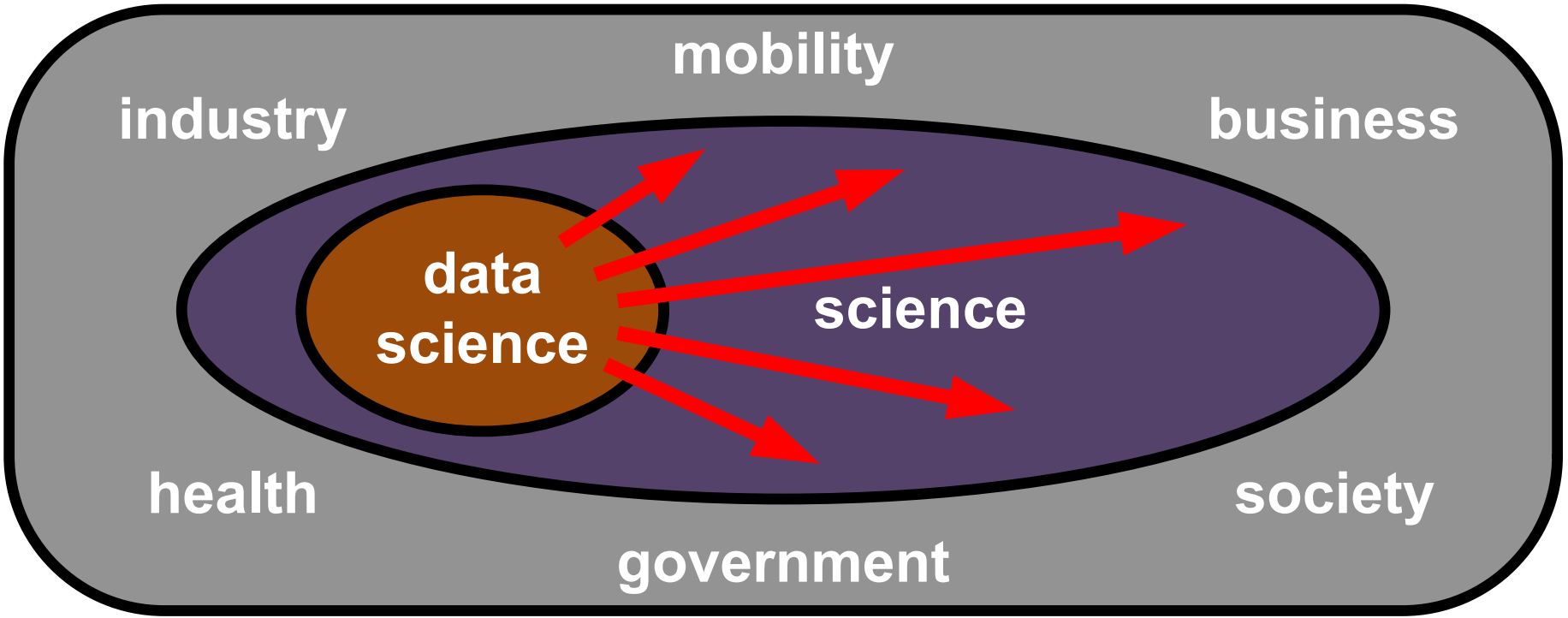

this course

Data Science



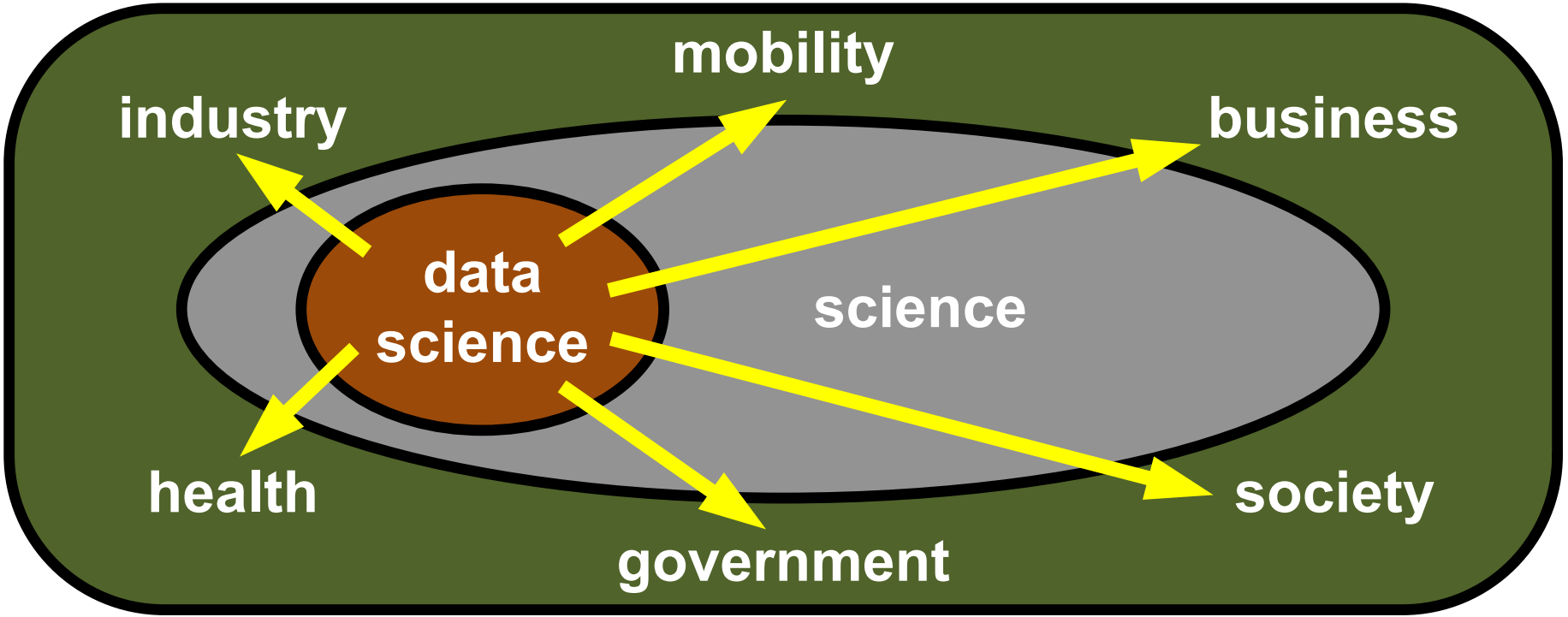
Chair of Process and Data Science

Data Science for Science



Chair of Process and Data Science

Data Science for Non-Science



Chair of Process and Data Science

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Data Science, Process Science, Process Mining, Business Process Management, Data Mining, Process Discovery, Conformance Checking, Simulation, and Responsible Data Science.

Alexander von Humboldt Stiftung/Foundation

Fraunhofer FIT

www.pads.rwth-aachen.de

Learning more about Process Mining

(All taking place the next semester!!)

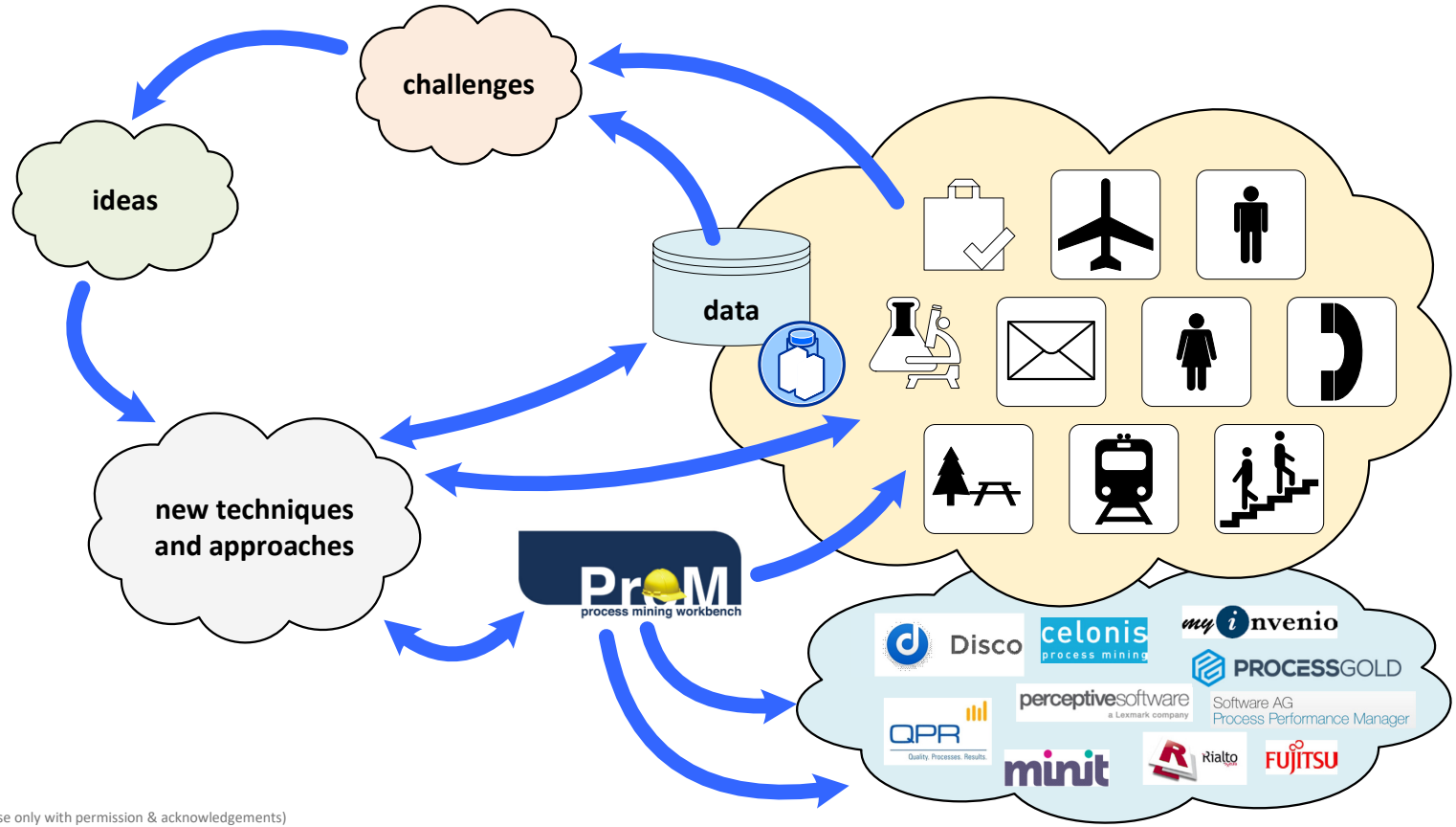
- **Business Process Intelligence (Bachelor/Master)**
- **Advanced Process Mining (Master)**
- **Selected Topics in Process Mining (Master Seminar)**
- **Introduction to Feature Prediction on Running Process Instances (Master Seminar)**
- **Process Discovery using Python (Praktikum)**
- **Conformance Checking Using Python (Praktikum)**

More about the “process mining iceberg”

- Many possibilities to go deeper.
- Many research challenges.
- Exciting projects.
- Industry is “screaming” for process mining experts.

Chair of Process and Data Science

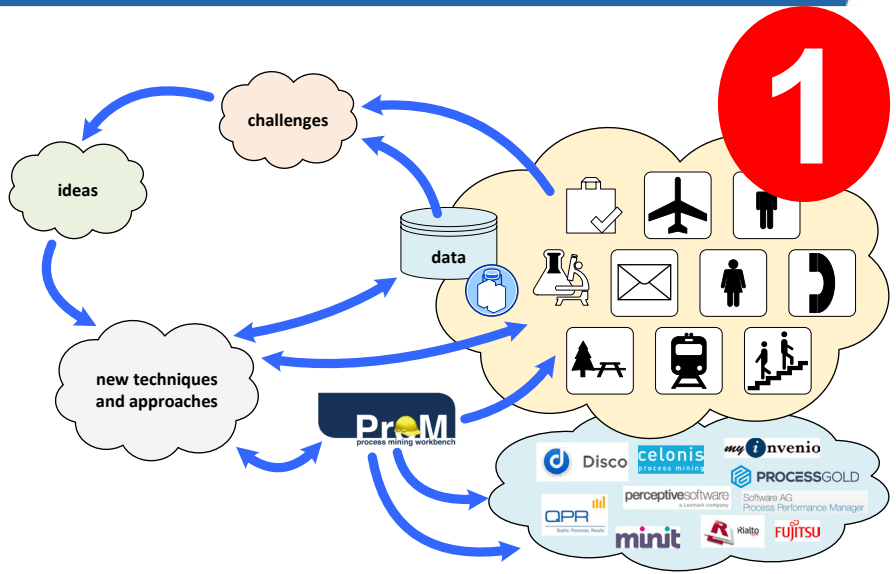
Interaction with industry



Chair of Process and Data Science

Organizations that apply process mining

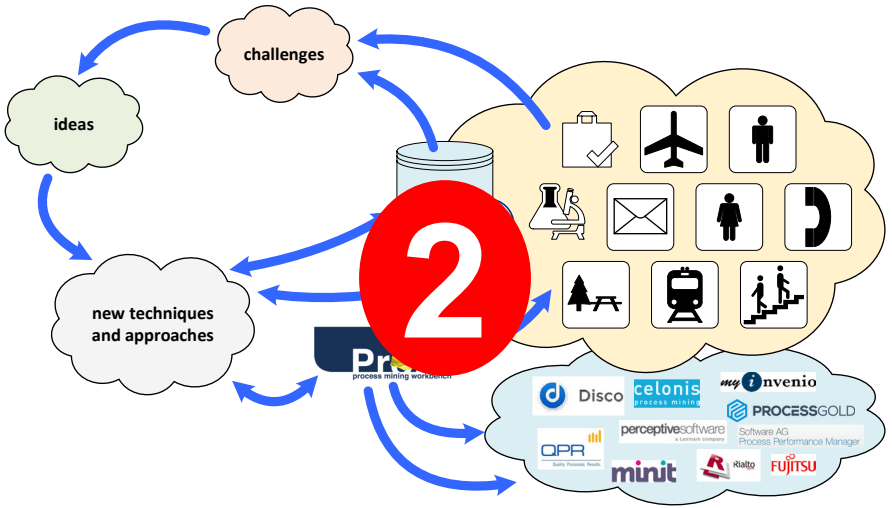
- Sales & procurement (SAP, etc.)
- Finance (banks, etc.)
- Insurance (Suncorp, etc.)
- Healthcare (hospitals, etc.)
- Government (municipalities, etc.)
- High-tech systems manufacturers (ASML, etc.)
- E-learning (Coursera, etc.)
- Etc.



Chair of Process and Data Science

Organizations that provide services based on process mining

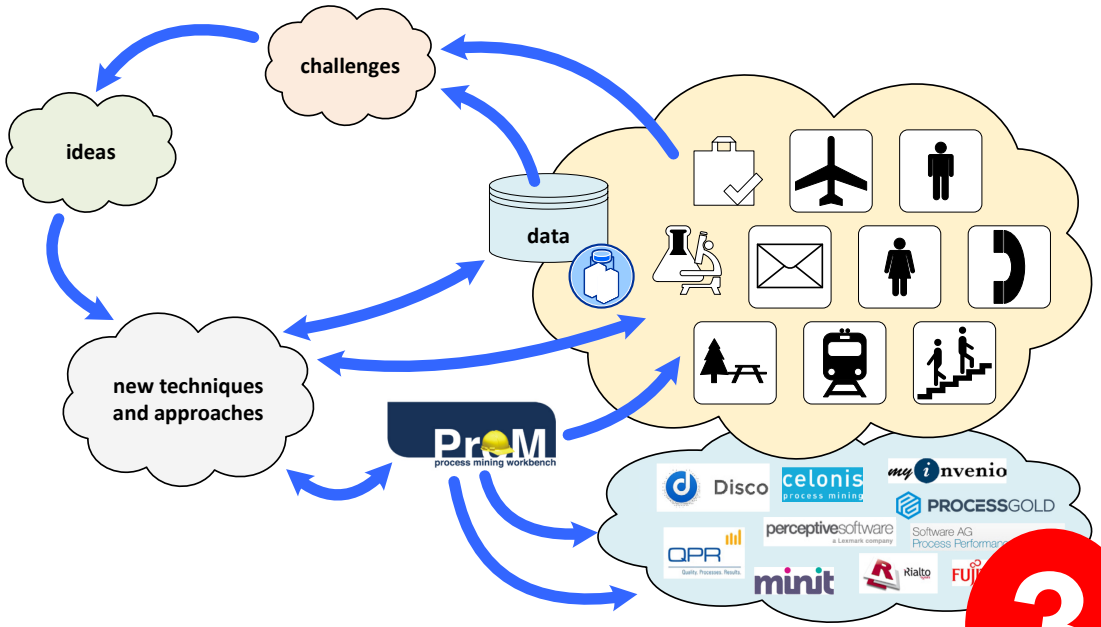
- Ernst & Young
- Deloitte & Touche
- KPMG
- PricewaterhouseCoopers
- Inform
- Zapliance
- Many smaller/specialized consulting firms
- Etc.



Chair of Process and Data Science

Organizations that develop process mining technology

- Celonis
- Fluxicon
- ProcessGold
- Minit
- QPR
- Lana
- Signavio
- Software AG
- + 20 more



3

Chair of Process and Data Science

Example: Celonis

3

Started in 2011 with three students using ProM, now over $1 billion company.

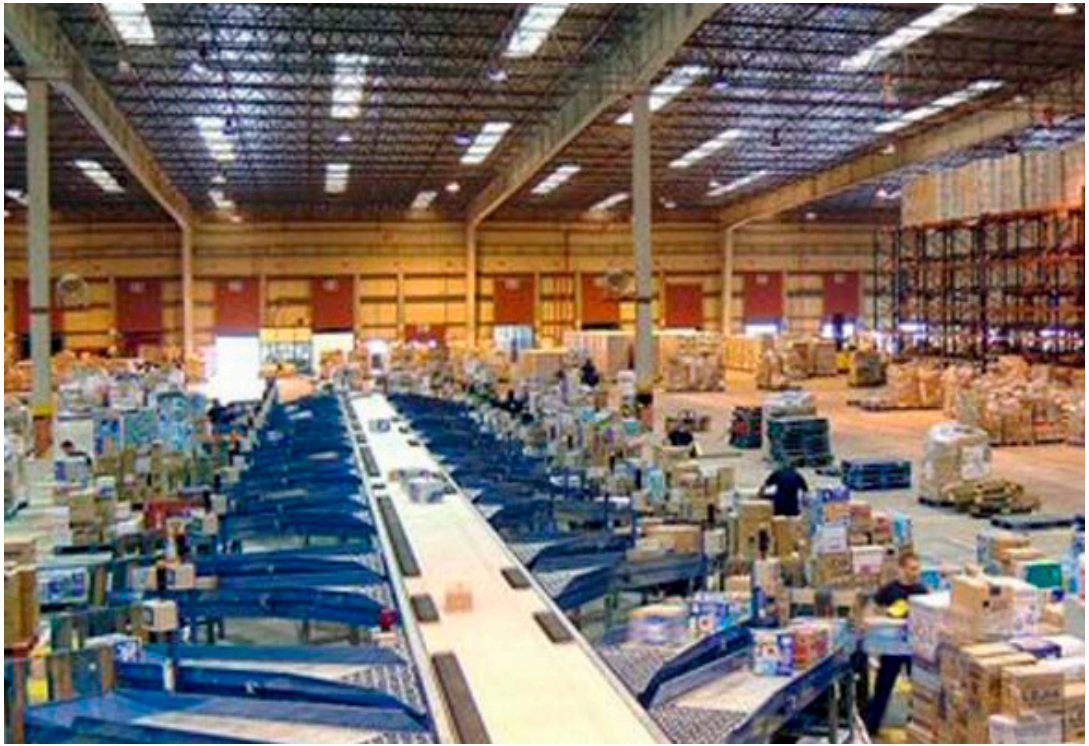
Chair of Process and Data Science

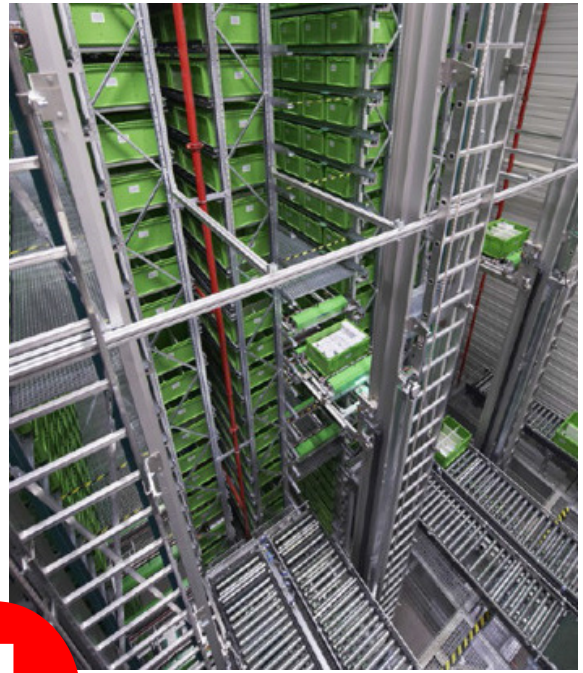
Example: Philips Healthcare

1

Chair of Process and Data Science

Example: Vanderlande Industries





1

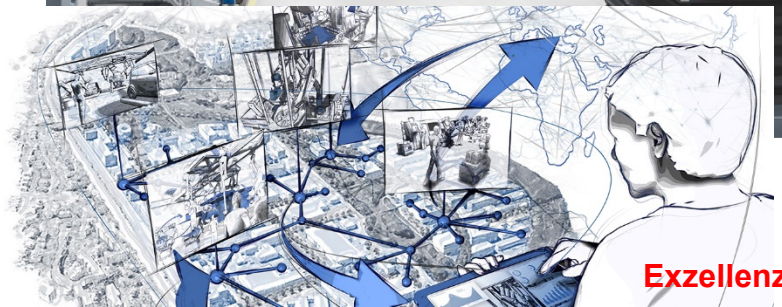
Chair of Process and Data Science

Example: Uniklinik RWTH Aachen

1

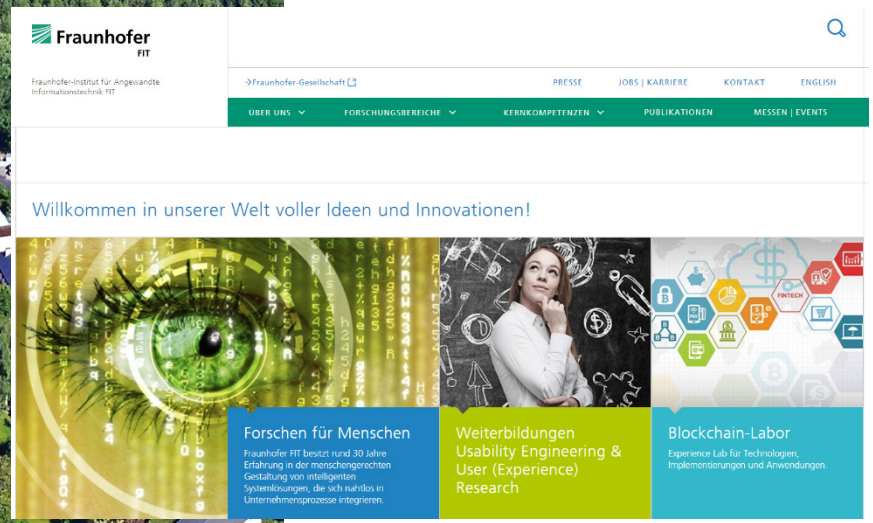
Chair of Process and Data Science

Example: e.GO Mobile



1

Example: Fraunhofer FIT



2

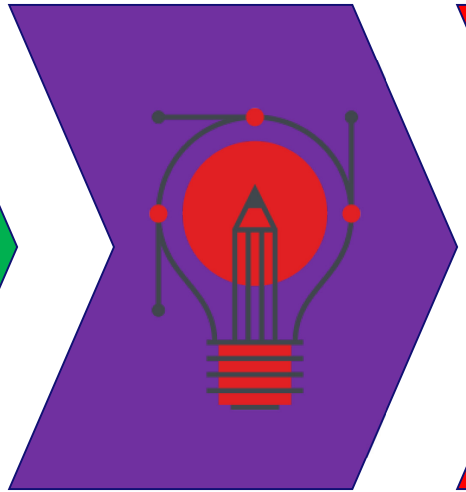
Chair of Process and Data Science

Example: Responsible Data Science

Chair of Process and Data Science

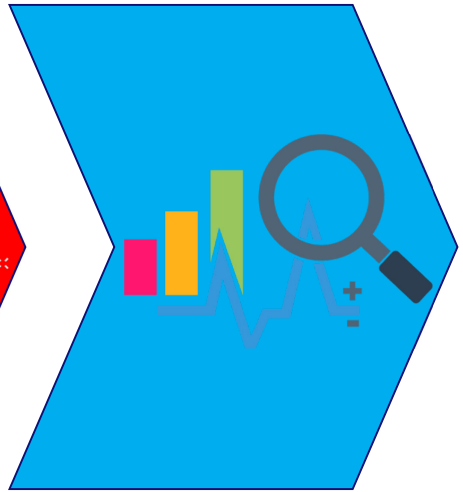
A typical process mining thesis project

novel idea



conceptualize/ formalize

design/ implement



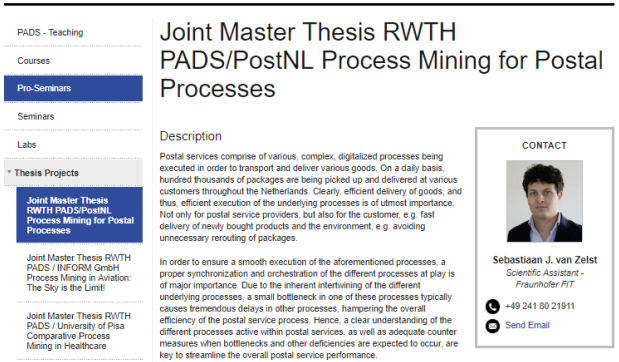
experiment/ evaluate

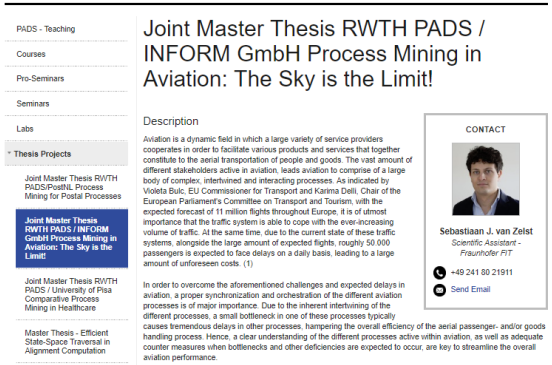
Chair of Process and Data Science

How to apply?

- Visit http://www.pads.rwth-aachen.de (link Thesis projects).
- Please fill out the thesis inquiry form and send it with a short motivation, your CV, and grades to Dr. Seran Uysal thesis@pads.rwth-aachen.de.
- Also PhD, SE, and HiWi positions (but make sure you have relevant experience first).

Example Projects





Many more.

Apply via Seran Uysal
thesis@pads.rwth-aachen.de.

Chair of Process and Data Science

Also check out internships (see posting before)

Data Scientist
(m/f/x) full-time, Munich

Are you ready for a new challenge?

Actively drive Celonis' expansion and work in project teams to kick-start our customers' process mining journey. You make digital business processes transparent and harmonize our customers' gigantic data flows by using our Celonis Process Mining Technology and by applying the most up-to-date extraction and transformation methods. You are involved in implementation projects with process data of varying degrees and complexity and for customers across industries.
After the successful evaluation and preparation of the process data, you connect the respective on-premise/ Cloud systems with our software. You extract and transform customers' data and design process- and customer-specific analyses.
Over the course of our projects, you expertly handle our customers' individual needs and actively participate in customer workshops.

YOU...

...have successfully completed your studies in Business Informatics, Computer Science / Mathematics / Physics or a comparable degree program

...have gained prior knowledge in Business Intelligence and data analysis

...are already experienced in programming, ETL and working with databases

... have excellent analytical skills and are always well-organized and known for being a quick learner

...enjoy evaluating complex data and working with complicated processes

...are excited by Big Data, Data Mining and Process Mining seek continuous improvement of your know-how

...are very dedicated, visionary and looking for a product that you can develop with passion and determination

...have very good English and German skills, other languages are an advantage

WE...

...are visionary and one of the fastest growing Tech-Unicorns in the world

...offer our customers the Intelligent Business Cloud which is the world's most powerful tool for analyzing, optimizing, and transforming all IT-supported business processes

...are pioneers and market leader in the area of Process Mining

...distinguish ourselves through a unique combination of innovative start-up atmosphere paired with great professionalism and self-responsible work

Interested? Apply now!

Laura Nagl | Talent Acquisition Manager
Theresienstraße 6 | 80333 Munich
l.nagl@celonis.de
+49 152 0911 4918

www.celonis.com/careers/

Data Science & Management Consultant
(m/f/x) full-time, Munich

Are you ready for a new challenge?

You analyze real data from our customers to increase transparency in business processes of any kind. Thereby, you use our Celonis Process Mining Technology in different proof-of-concept projects to demonstrate the value our software generates. Thanks to your conceptional working method, you generate conclusive examples and scenarios during this trial project and focus on the most relevant KPIs to convince our customers of its performance and benefits. In doing so, you react to individual needs of customers from various industries and in conclusion present your findings in front of managers and decision makers. Additionally - by outlining and mapping strategic concepts and use cases - you guide our international expansion, support the further development of our team and help prepare for our IPO.

YOU...

... have graduated with an exceptional degree in the area of mathematics / physics / industrial engineering / business informatics or of a comparable program

... already have gained experience in consulting and are highly interested in Data Science and Big Data

... think analytically, work in a structured manner and quickly comprehend new topics and interrelationships

... possess a thorough understanding of complex business processes of any kind

... are characterized by strong communicational skills and enjoy interacting with various customers and contact persons

... are very engaged and visionary and want to actively drive the Celonis Process Mining technology forward

... search for a job to think outside the box and continuously broaden your knowledge while having a steep learning curve

... are looking forward to discussing the latest technology with your colleagues over a match of table football

... have very good knowledge of spoken and written English, other languages are a bonus

WE...

...are visionary and one of the fastest growing technology-unicorn in the world

...offer the world's most powerful tool for analyzing and optimizing IT-supported business processes and data volumes

...are pioneers and market leader in the area of Process Mining

...are distinguished by an unique combination of innovative start up atmosphere combined with great professionalism and self-responsible work

Interested? Apply now!

Anouk Fechner | Talent Acquisition Manager
Theresienstraße 6 | 80333 Munich
a.fechner@celonis.de
+49 162 2894551

www.celonis.com/careers/

Business Process Analyst
(m/f/x) full-time, Munich

Are you ready for a new challenge?

As a member of our content store team you work on scalable business solutions for Celonis Process Mining, which we provide to our customers worldwide. As a business analyst, you develop new use cases and leverage existing solutions according to the demands of our customers or specific industries. In addition, you collect digital data from unknown IT Systems and bring transparency into underlying business processes. You implement smart algorithms and develop innovative applications that generate a value for our customers. For this, you combine your Data Science Skills with your business know-how in order to take our process mining technology to a new level.

YOU...

... possess an above-average university degree in Economic Computer Science / Information-oriented Business Administration / Mathematics or equivalent

... have a sound understanding of all kinds of business processes and have many ideas how to improve them

... have a good knowledge of SQL and are familiar with the relational databases

... are enthused by Big Data, Data Science or Business Intelligence and have ideally already gained experience in this field

... have an analytical mind, work in a structured manner and are a quick learner

... want to have an impact and are looking for a product that you can develop with passion and drive

... have very good English skills, German is a plus

WE...

... are visionary and one of the fastest growing technology-unicorn in the world

... offer the world's most powerful tool for analyzing and optimizing IT-supported business processes and data volumes

... are pioneers and market leader in the area of Process Mining

... are distinguished by an unique combination of innovative start-up atmosphere combined with great professionalism and self-responsible work

Interested? Apply now!

Alexandra Haberkern | Senior Talent Acquisition Manager
Theresienstraße 6 | 80333 Munich
a.haberkern@celonis.de
+49 89 4161596-742

www.celonis.com/careers/

Much more

Bundesministerium für Verkehr und digitale Infrastruktur

BMVI DATA-RUN

Einladung an Teilnehmerinnen und Teilnehmer zum BMVI Data-Run am 22. und 23. März 2019 in Berlin

Das **Bundesministerium für Verkehr und digitale Infrastruktur (BMVI)** lädt herzlich ein zum

BMVI Data-Run
am
22. und 23. März 2019
im
Bundesministerium für Verkehr und digitale Infrastruktur in Berlin!

Wir suchen Programmierer, Gründer, Entwickler und sonstige Interessierte! Egal ob Ihr aus der IT oder aus anderen Bereichen kommt: Lasst Eurer Kreativität freien Lauf und zeigt uns, welche Lösungen Ihr mit unseren Daten entwickeln könnt – und das alles innerhalb von 24 Stunden!

Einen ersten Überblick über die Daten des BMVI-Geschäftsbereichs sowie die Daten Dritter erhaltet Ihr über das offene Datenportal mCLOUD unter
www.mcloud.de

Nähere Informationen zu den im Rahmen der Veranstaltung bereitgestellten Daten, unseren Challenges, zur Anmeldung und zum Ablauf des Hackathons findet Ihr unter:
www.bmvi-data-run.de

Ob als komplettes Team oder als Einzelperson: Merkt Euch den Termin vor, und seid am 22./23. März 2019 mit dabei! Die Teilnahme am BMVI Data-Run ist kostenlos.

Wir freuen uns, Euch in Berlin begrüßen zu dürfen!

Das mFUND Team im
Bundesministerium für Verkehr und digitale Infrastruktur

Der BMVI Data-Run wird veranstaltet im Rahmen des Förderprogramms mFUND, mit dem das BMVI seit 2016 Forschungs- und Entwicklungsprojekte rund um datenbasierte digitale Anwendungen für die Mobilität 4.0 fördert. Neben der finanziellen Förderung unterstützt der mFUND mit verschiedenen Veranstaltungsformaten die Vernetzung zwischen Akteuren aus Politik, Wirtschaft und Forschung. Weitere Informationen zum mFUND findet Ihr unter www.mfund.de.

- **Data Science Hackathons.**
- **Various general contests (e.g., Kaggle competitions).**
- **Specific contests:**
  - **Process Discovery Contest**
  - **BPI Challenge**
  - **etc.**

icpmconference.org

https://www.bmvi.de/SharedDocs/DE/Termine-mFUND/bmvi-data-run.html

https://meet.celonis.com/engie-hackathon/

Chair of Process and Data Science


Data
Scientist

Questions?

Good Luck!