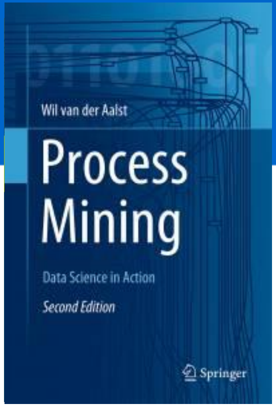
Process Mining: Data Science in Action

Evaluating Mining Results

prof.dr.ir. Wil van der Aalst

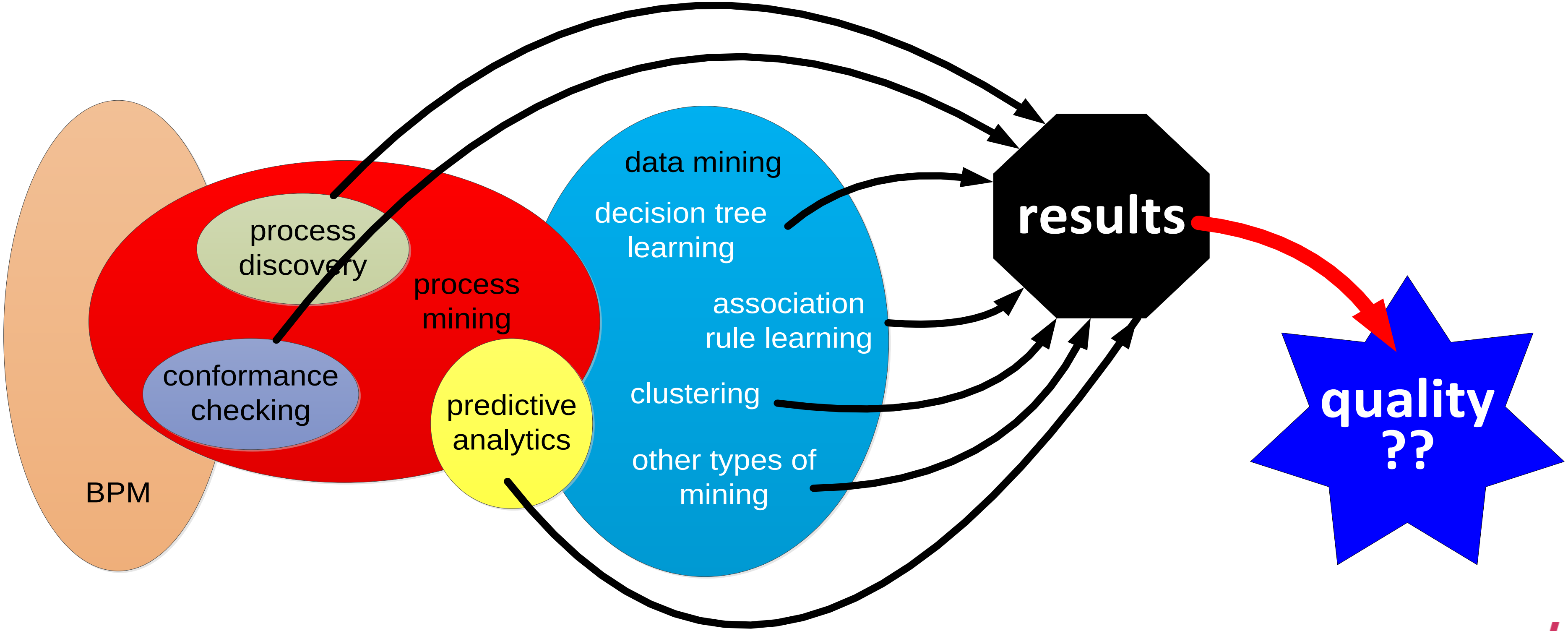
www.processmining.org



Technische Universiteit
Eindhoven
University of Technology

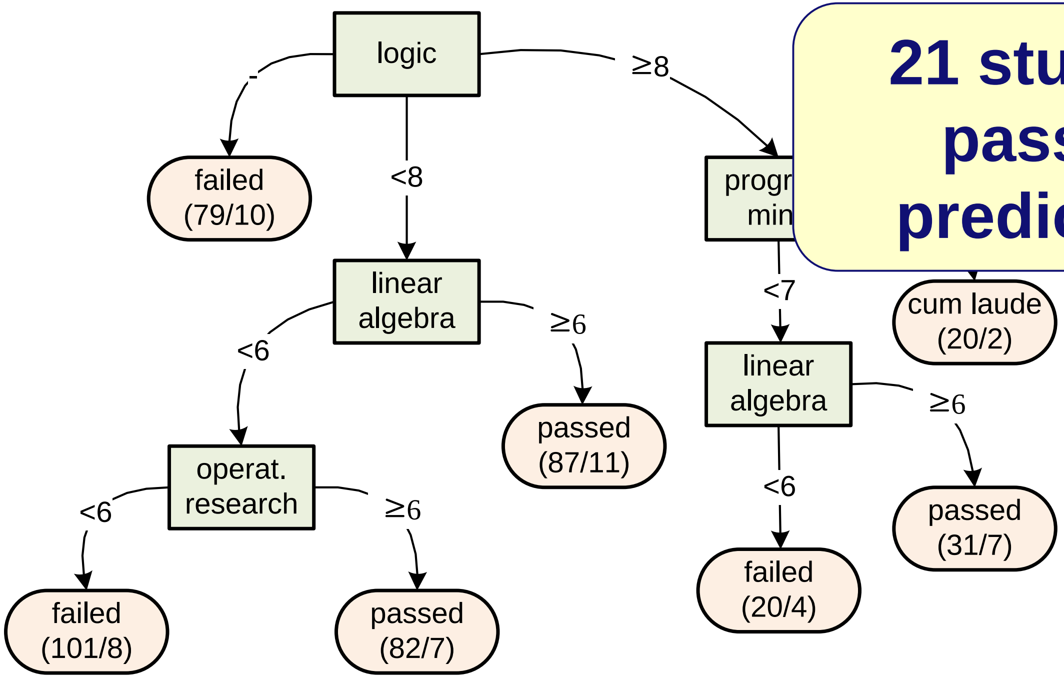
Where innovation starts

Evaluating (data/process) mining results



Confusion matrix and related measures

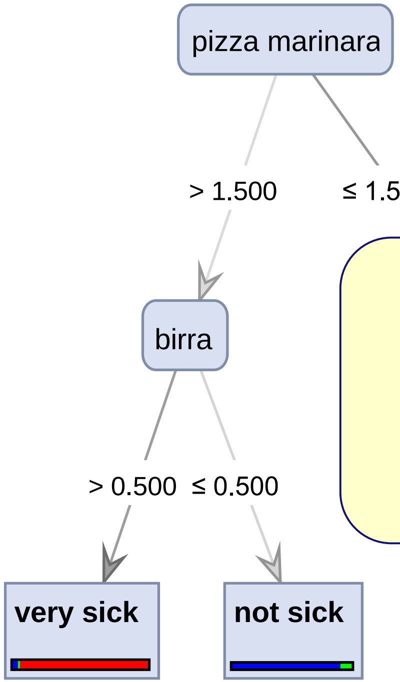
Confusion matrix for decision tree

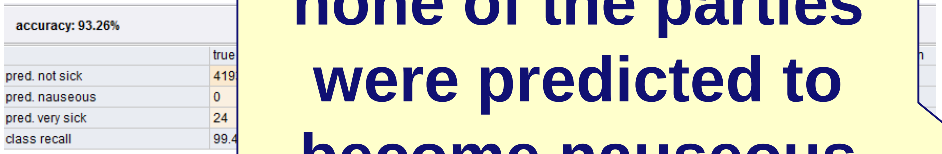


21 students that passed were predicted to fail

		predicted class		
		failed	passed	cum laude
actual class	failed	178	22	0
	passed	21	175	2
	cum laude	1	3	18

Another example





none of the parties were predicted to become nauseous

all 470 parties that became very sick were predicted to become sick

		predicted class		
		not sick	nauseous	very sick
actual class	not sick	4193	0	24
	nauseous	307	0	6
	very sick	0	0	470

TU/e

Confusion matrix for binary classification

- **True Positives (TP)**: positive instances predicted to be positive .
- **True Negatives (TN)**: negative instances predicted to be negative.
- **False Positives (FP)**: negative instances predicted to be positive.
- **False Negatives (FN)**: positive instances predicted to be negative.

		predicted class	
		+	-
actual class	+	TP	FN
	-	FP	TN

- **True Positives (TP): sick customers predicted to be sick.**
- **True Negatives (TN): non-sick customers predicted to be not sick.**
- **False Positives (FP): non-sick customers predicted to be sick.**
- **False Negatives (FN): sick customers predicted to be not sick.**

		predicted class	
		+	-
actual class	+	TP	FN
	-	FP	TN

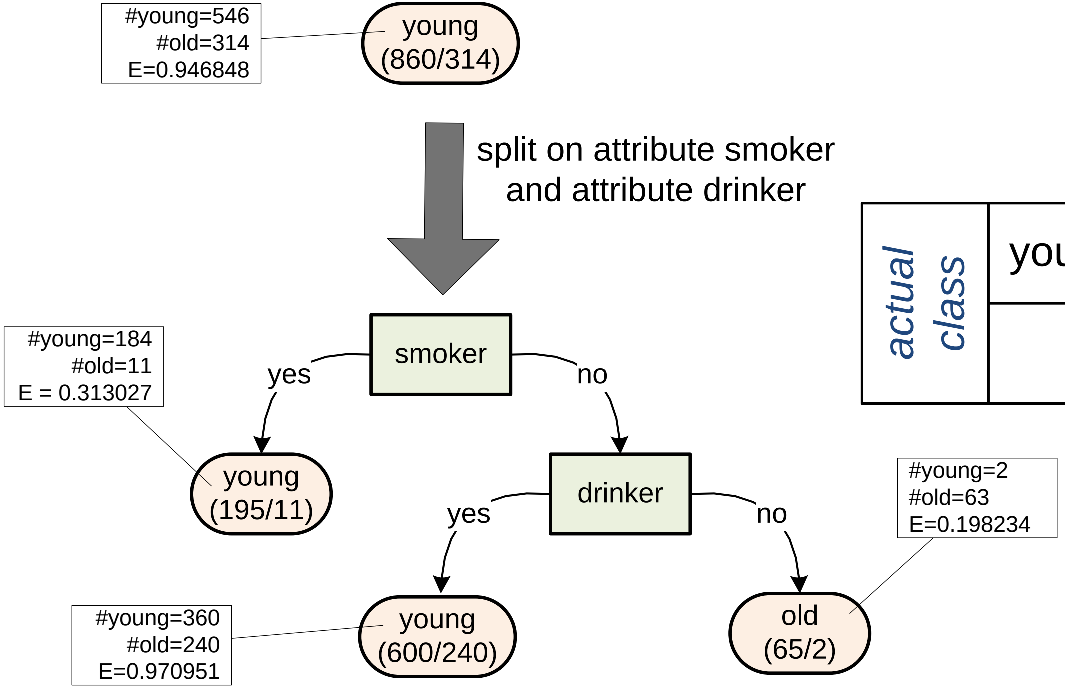
Quality measures
(based on confusion matrix)

- **error** = (FP+FN)/K
- **accuracy** = (TP+TN)/K
- **precision** = TP/P' = TP/(TP+FP)
- **recall** = TP/P = TP/(TP+FN)
- **F1-score** =
  (2 x precision x recall)/(precision + recall)
  (harmonic mean of precision and recall)

		predicted class		
		+	-	
actual class	+	TP	FN	P = TP+FN
	-	FP	TN	N = FP+TN
		P' = TP+FP	N' = FN+TN	K = TP+FN+FP+TN

Question: Compute precision, recall, and the F1-score before and after splitting

young = positive
old = negative



		predicted class: young	predicted class: old
actual class	young	546	0
	old	314	0

		predicted class: young	predicted class: old
actual class	young	544	2
	old	251	63

Answer

precision = 546/(546+314) = 0.635
recall = 546/(546+0) = 1.000
F1-score = 0.777

		predicted class: young	predicted class: old
actual class	young	546	0
	old	314	0

		predicted class: young	predicted class: old
actual class	young	544	2
	old	251	63

precision = 544/(544+251) = 0.684
recall = 544/(544+2) = 0.996
F1-score = 0.811

Cross-validation

Consider your 10 best friends

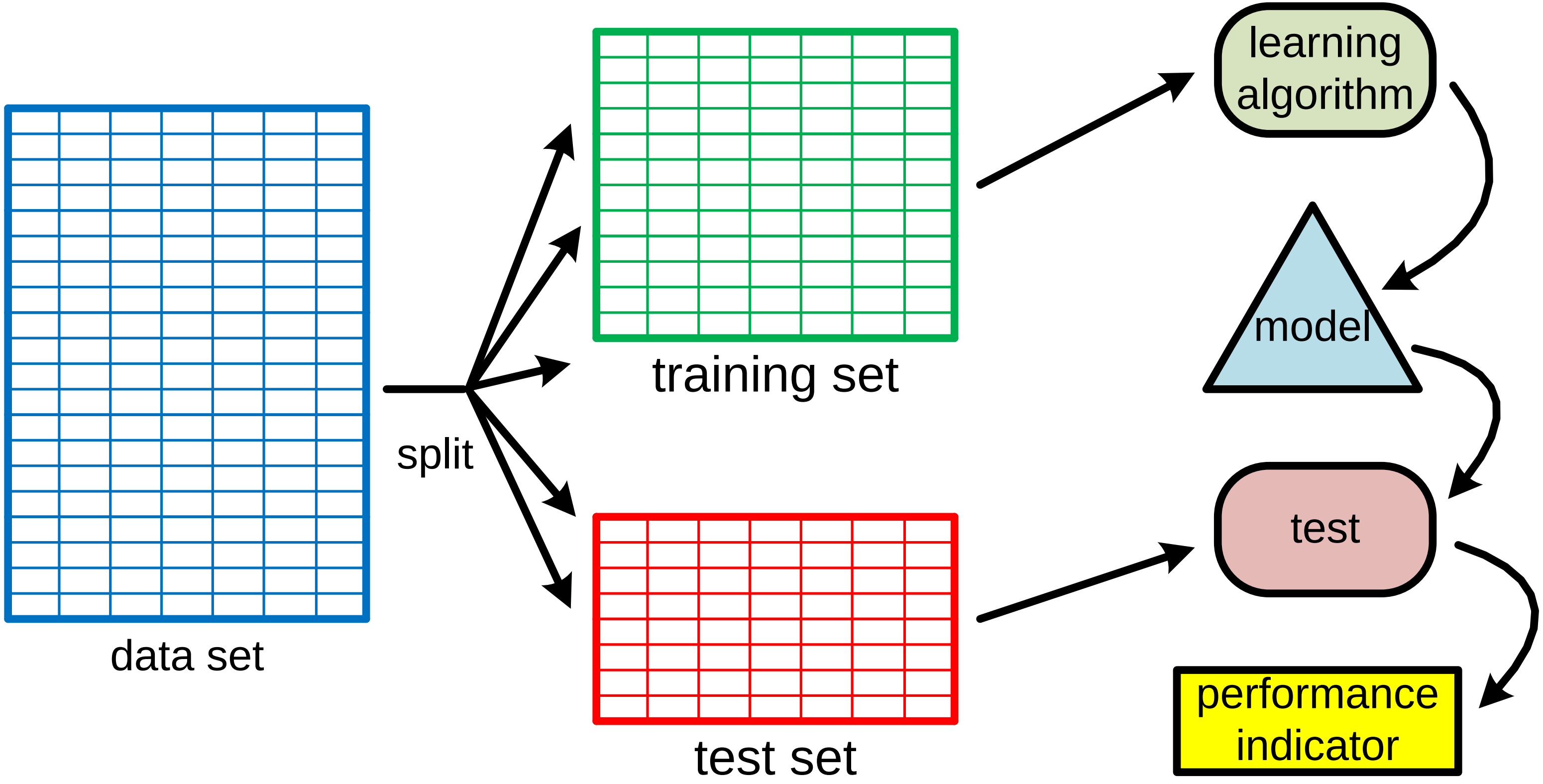
You can create a decision tree that accurately predicts the length of a friend based on his/her birth date and eye color.

… but the model will be overfitting the data set and will most likely not apply to any new friends.

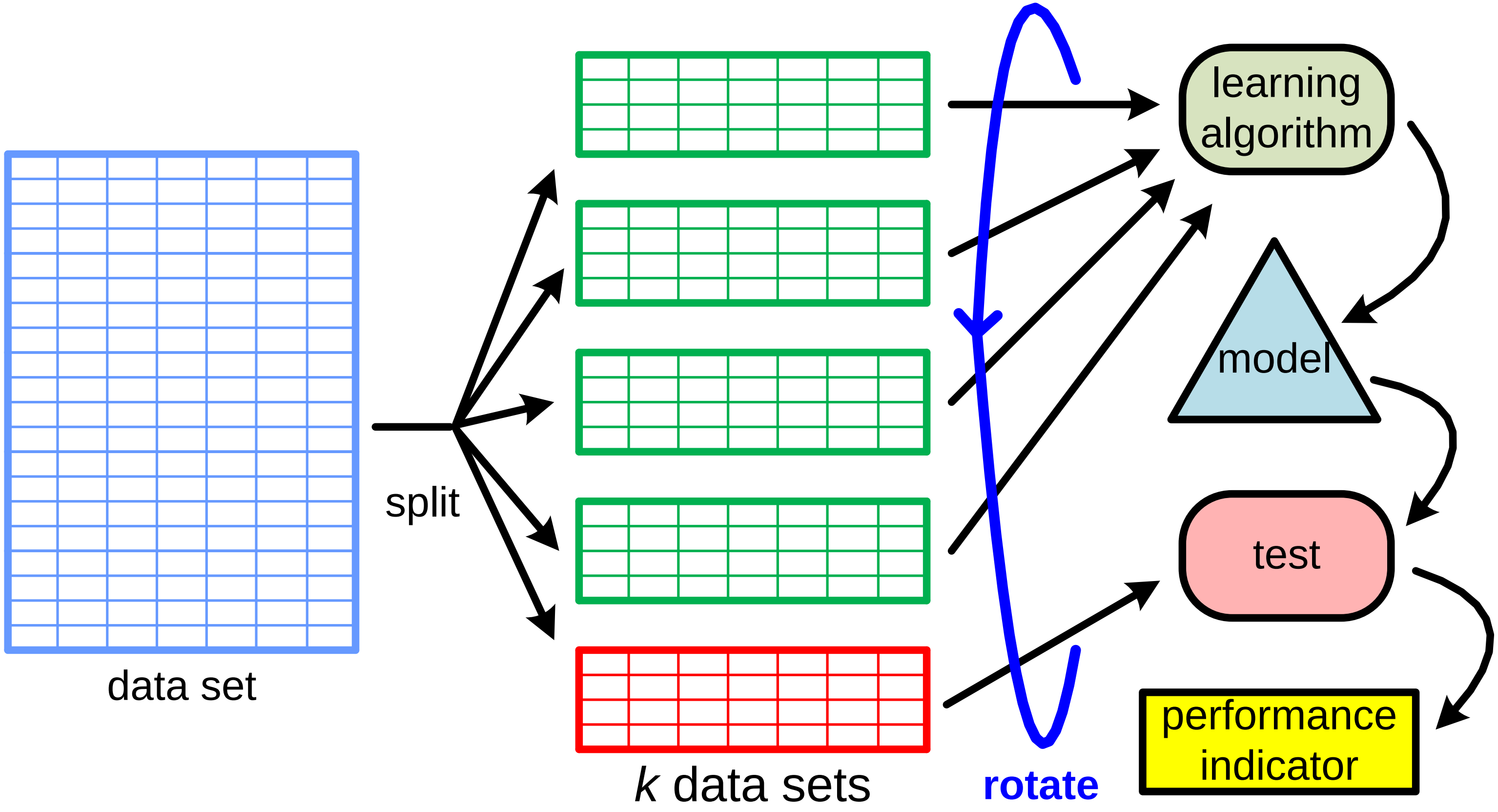
Overfitting and underfitting

- **Overfitting**: the model is too specific for the data set used to learn the model and performs poorly on new instances.
  - If birth date = 16-05-1998 and eye color = blue, then length = 172.8 cm.
- **Underfitting**: the model is too general and does not exploit the data.
  - If gender = male, then length > 1 meter.

Cross-validation



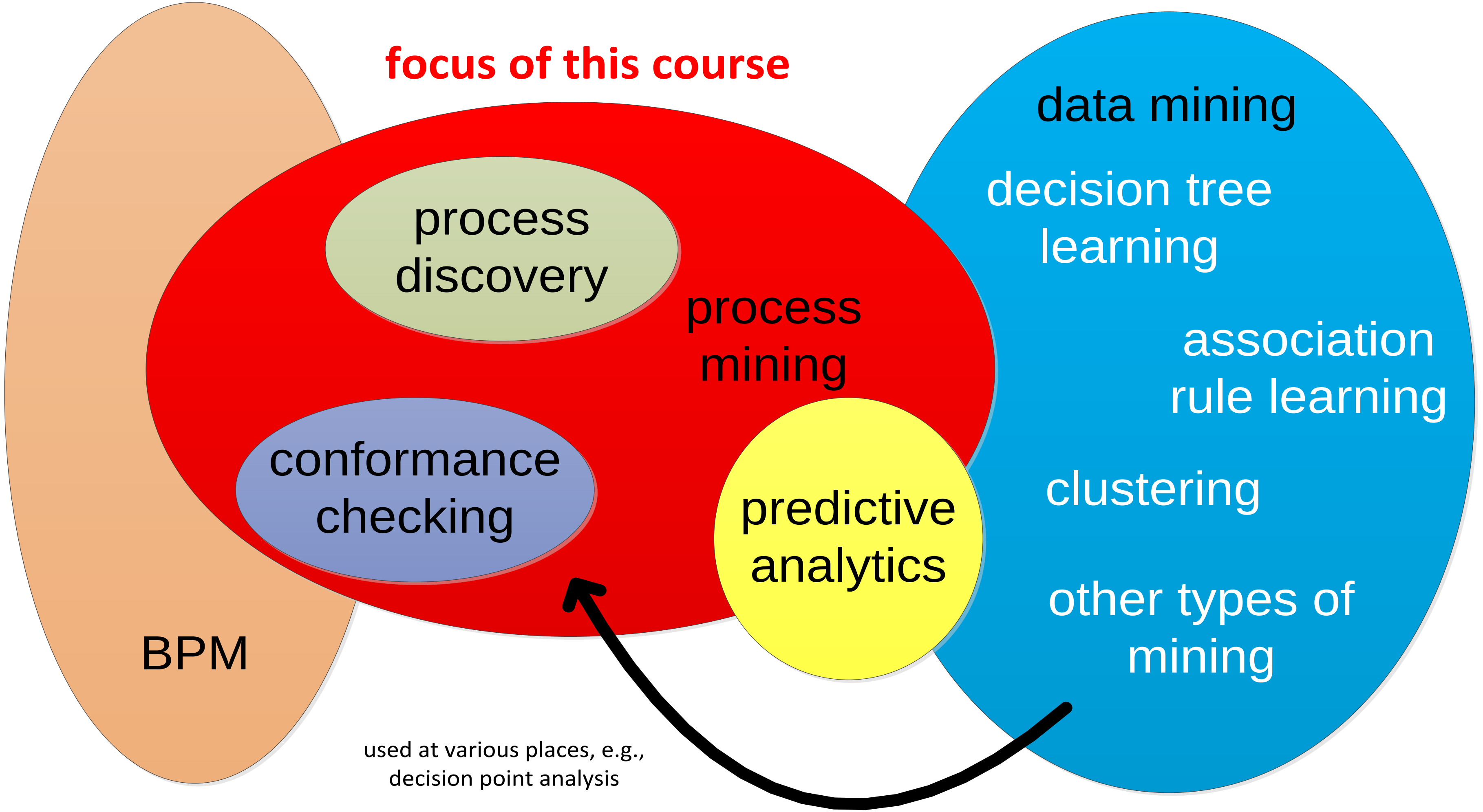
k-fold cross-validation

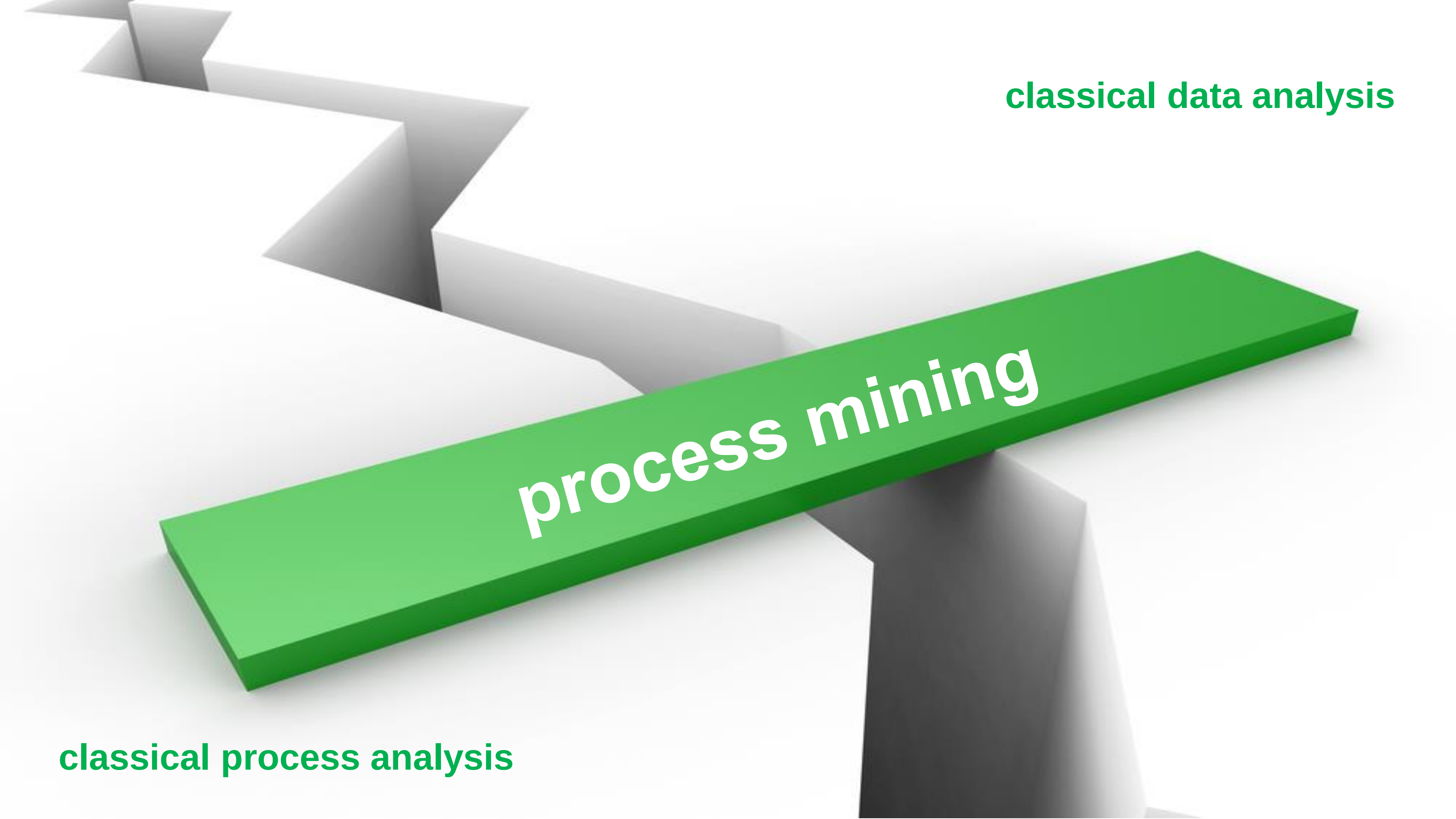


TU/e

Possible complications

- Concept drift (model should change over time).
- No negative examples (we only know about sick customers that complained afterwards).
- …


focus of this course
data mining
decision tree learning
association rule learning
clustering
other types of mining
process discovery
process mining
conformance checking
predictive analytics
BPM
used at various places, e.g., decision point analysis


classical data analysis
process mining
classical process analysis

Part I: Introduction

- **Chapter 1** Data Science in Action
- **Chapter 2** Process Mining: The Missing Link

Part II: Preliminaries

- **Chapter 3** Process Modeling and Analysis
- **Chapter 4** Data Mining

Part III: From Event Logs to Process Models

- **Chapter 5** Getting the Data
- **Chapter 6** Process Discovery: An Introduction
- **Chapter 7** Advanced Process Discovery Techniques

Part IV: Beyond Process Discovery

- **Chapter 8** Conformance Checking
- **Chapter 9** Mining Additional Perspectives
- **Chapter 10** Operational Support

Part V: Putting Process Mining to Work

- **Chapter 11** Process Mining Software
- **Chapter 12** Process Mining in the Large
- **Chapter 13** Analyzing “Lasagna Processes”
- **Chapter 14** Analyzing “Spaghetti Processes”

Part VI: Reflection

- **Chapter 15** Cartography and Navigation
- **Chapter 16** Epilogue