Combining Character Recognition, Text Analysis and Machine Learning for Automatic Document Analysis.

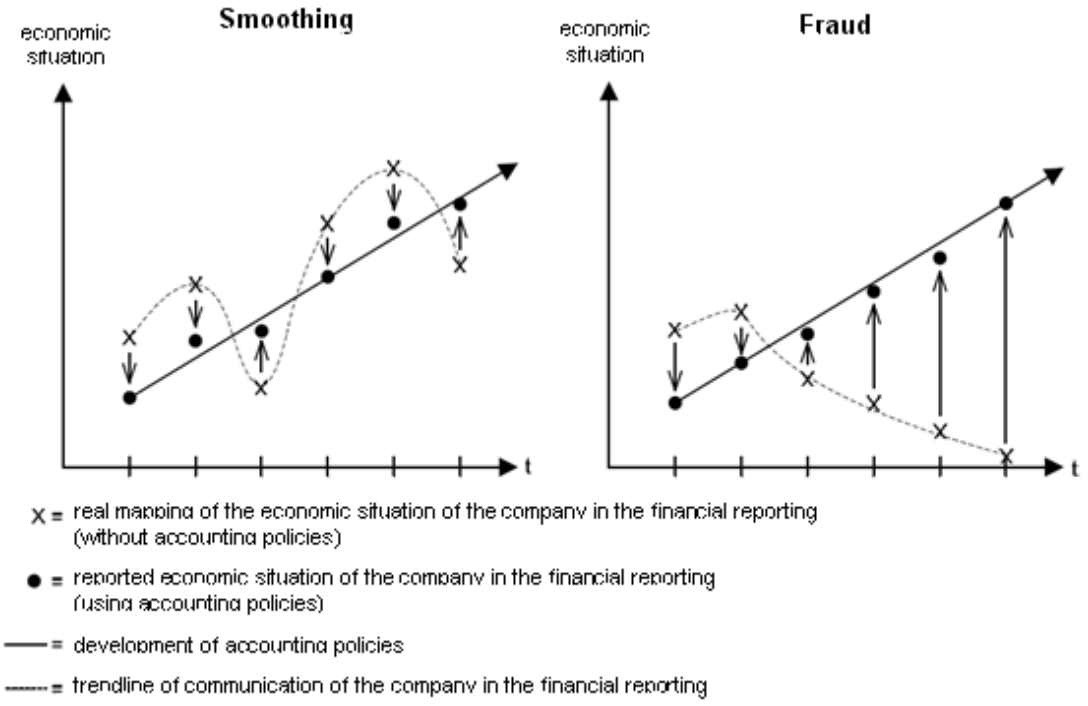
Dennis Weinbender

Disclaimer

This presentation should not be reported as representing
the views of the Commerzbank.
The views expressed are those of the author
and do not necessarily reflect those of the Commerzbank.

Fraud Definition: When an Incorrect Credit Decision is Based on Manipulated Information.

Accounting Policy vs. Accounting Manipulation



Progressive accounting policy may turn into accounting manipulation [1)]

- Legal accounting options are exercised in order to impact the perception of the company
- Most companies have no intention to commit fraud when becoming our customers. In case of a deteriorating economic situation the gap between the reported and real situation of a company increases and the companies eventually become fraudsters
- Progressive accounting policy can be a first signal for fraud

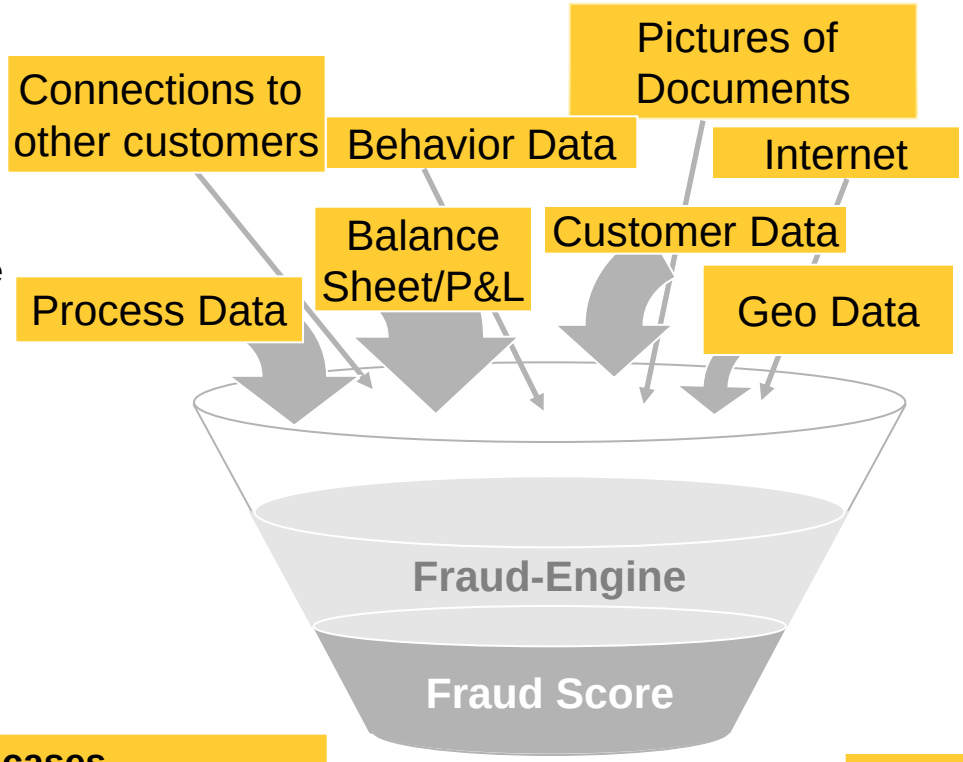
- **Fraud exists at the latest, when an incorrect credit decision is based on manipulated accounts, disinformation or withheld information**

[1)] For the main ideas and graphs we refer to Obermann, M., Bilanzpolitik und Kreditvergabeentscheidungen, Lüneburg 2010

Fraud-Engine, its Data Sources and the Creation of the Fraud-Score

Flexible Data Input

- Various formats: pictures, diagrams, text, numbers
- Different structured and unstructured features are analyzed by the fraud engine
- Fraud Engine extracts important features and determines whether they improve the overall forecast quality
- Summaries for suspicious cases are created, containing hints for investigation.



Statistical Model

- The basis are three groups: non fraud customers, fraud customers and defaulted customers without fraud.
- The goal is to bundle a large number of frauds in a small part of a portfolio
- Fraud-Score quantifies the similarity of a case to known fraud patterns.

Suspicious cases

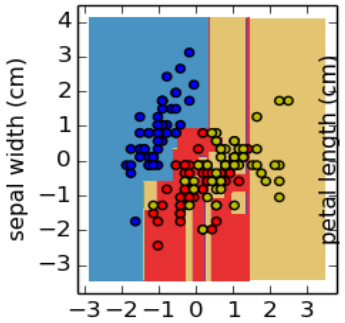
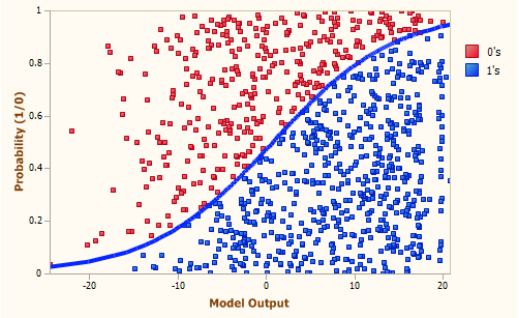
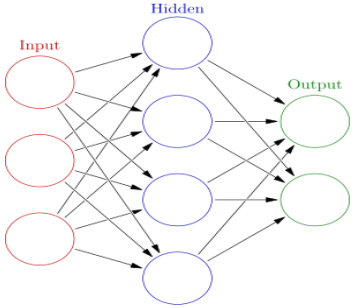
- Automatically generated first-check-questions help to adjust the focus to most important points

The final decision is taken by a human.

Unsuspicious cases

- no further actions are required for most cases

Fraud-Engine Combines the Advantages of Different Machine Learning Algorithms

Method	Rules	Regression based (Logit)	Black Box (Neuron Net)
Classical toolbox			
Example	▪ Companies with a moderate cash flow, high profit increase and no provisions have increased fraud probability.	▪ Fraud probability decreases with each additional percent of provision	▪ Several neurons are active, which correspond mostly to cash flow related variables
Machine Learning extensions	▪ Support Vector Machine (SVM) ▪ Random Forest ▪ Gradient Tree Boosting	▪ Multivariate Adaptive Splines ▪ Elastic net ▪ Sparse PLS	▪ Stacked Autoencoders ▪ Deeper Neuron Nets with dropout ▪ t-SNE

- Machine Learning extends the classical toolbox, but every estimator still has own strengths and weaknesses. **(No Free Lunch Theorem)**
- **Ensemble Blending** overcomes this limitations by combining diverse views into one and learning to correct different errors.

Our Approach

- Blackbox-Methods are used as benchmarks and help to detect hidden properties of the data.
- Whitebox-Methods are the core of the application. The signal is decomposed into simple triggers, which deliver first points for fraud investigation.

Documents Processing: Blending of Deep Neural Nets for OCR Recognition and Using Statstical Postprocessing to Recognize Geopattern

0 Document picture is provided

- Automatic engine start

1 Imagemagick picture transformations and noise reduction

- preprocess image from any format to png and apply filters (i.e. rotations, contrast, scaling) for quality improvement.

2 Apply Ensemble of OCR Tesseract Engines (Deep Neural Net)

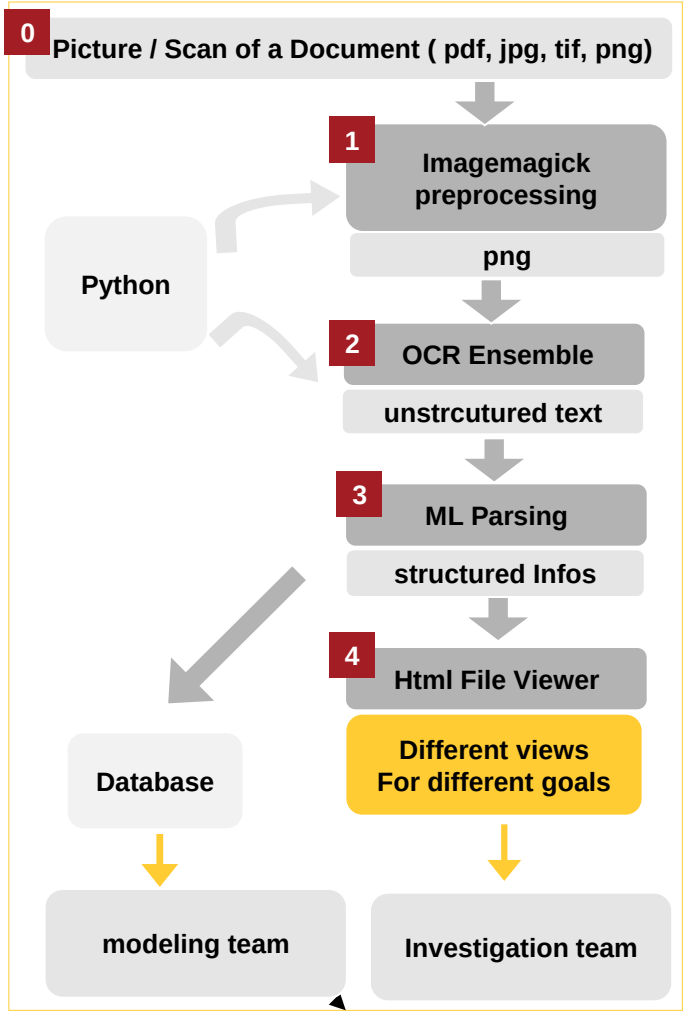
- An ensemble of different OCR machines extract diverse views of the data
- Determine the class of the document.
- Given the class, we use our own dictionaries of words, skip(n,k) grams and more generalized contexts to improve word recognition.

3 Machine Learning Based Text Parsing (ML Parsing)

- For every view and every set of the document a consistency analysis is run
- Known error types are corrected on the fly
- Country / City detection, Entity (known customer) detection, Good Detection.

4 Html File Viewer allows to visualize extracted information

- The tool provides COs with specialized views. The final decision is made by human.
- Data is saved and used steadily to improve the process



Application Example : Recognizing relevant Geopattern.

For every single document:

1 Doc Class Recognition

A document is cut in single pages. In order to exploit the different structure behind documents, we first estimate the document class of single pages and merge the pages to documents together. Results:

- ~100 document classes, each in several languages.
- 98% overall precision, most classes are at 100% precision
- 96% recall

2 Word Correction given Doc Class

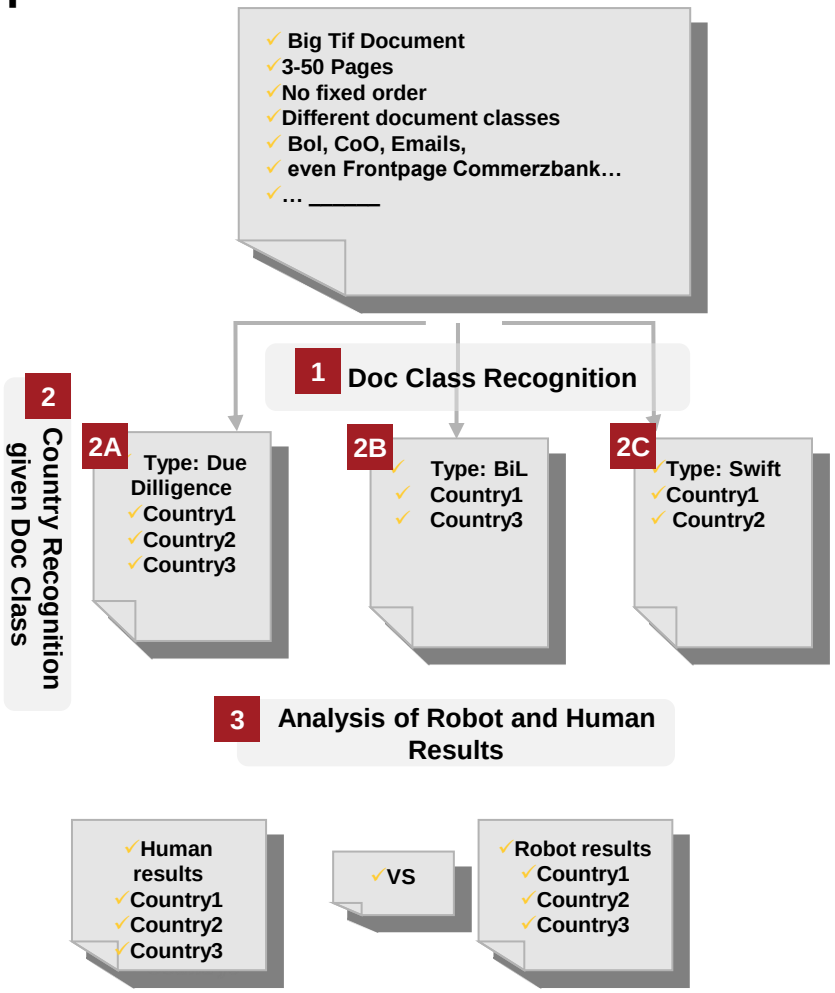
Given the document class, different logic engines are used. Some documents use only capital letter words, others have an extra vocabulary.

- Approx 93% of words are recognized before postprocessing
- N-grams, skip(n,k)-grams are extracted from the pseudo documents generated by a window of words is used
- Approx 97% of words are recognized after postprocessing

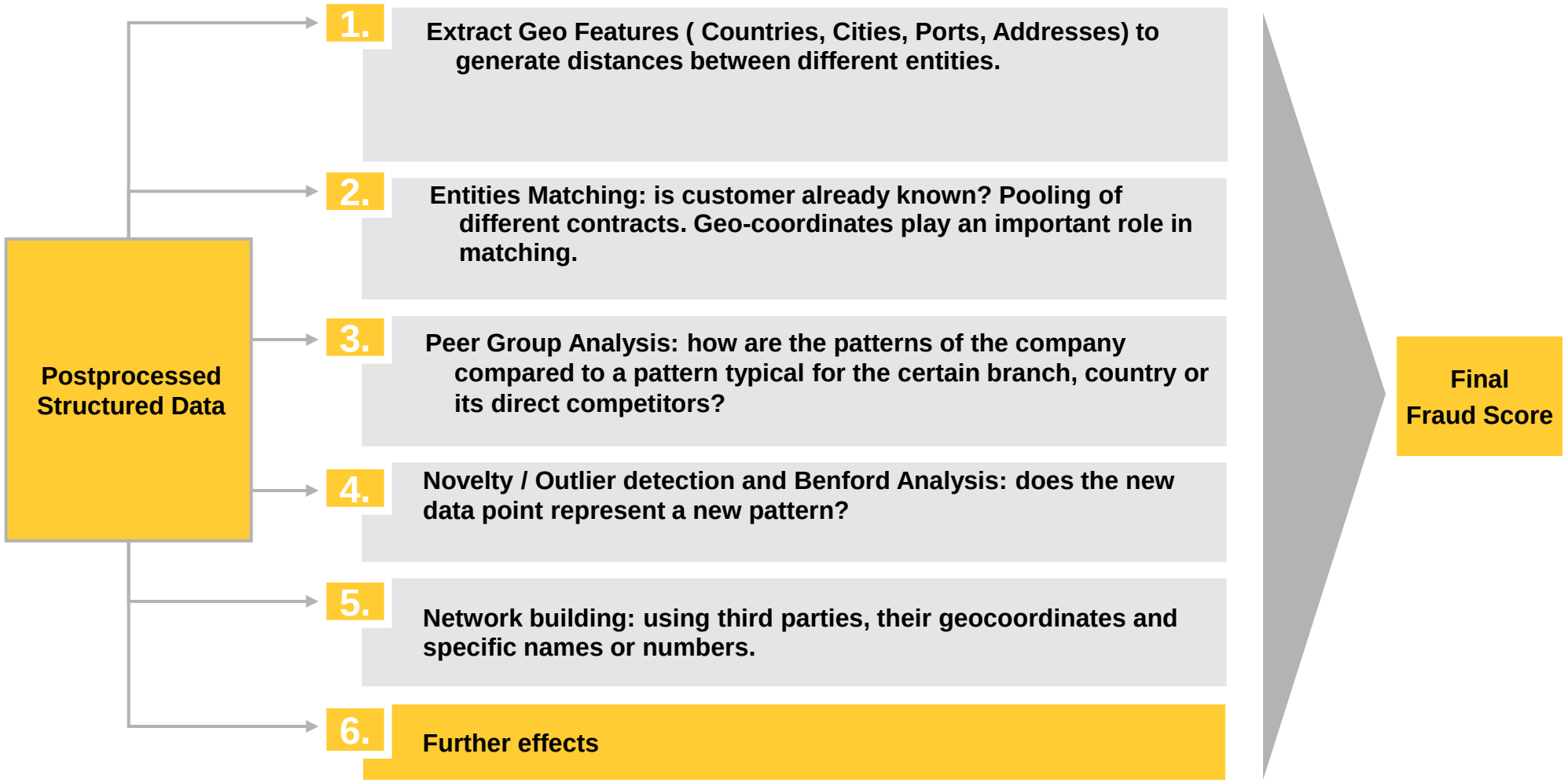
3 Country Recognition and Comparison Vs Human Results

The results of the machine are highlighted in the pictures and can be verified by a human observer.

- 190 Countries, 100% recall and 80%precision
- The tool already found human mistakes in 3-4% of documents



Using Geo Information to Extract Triggers



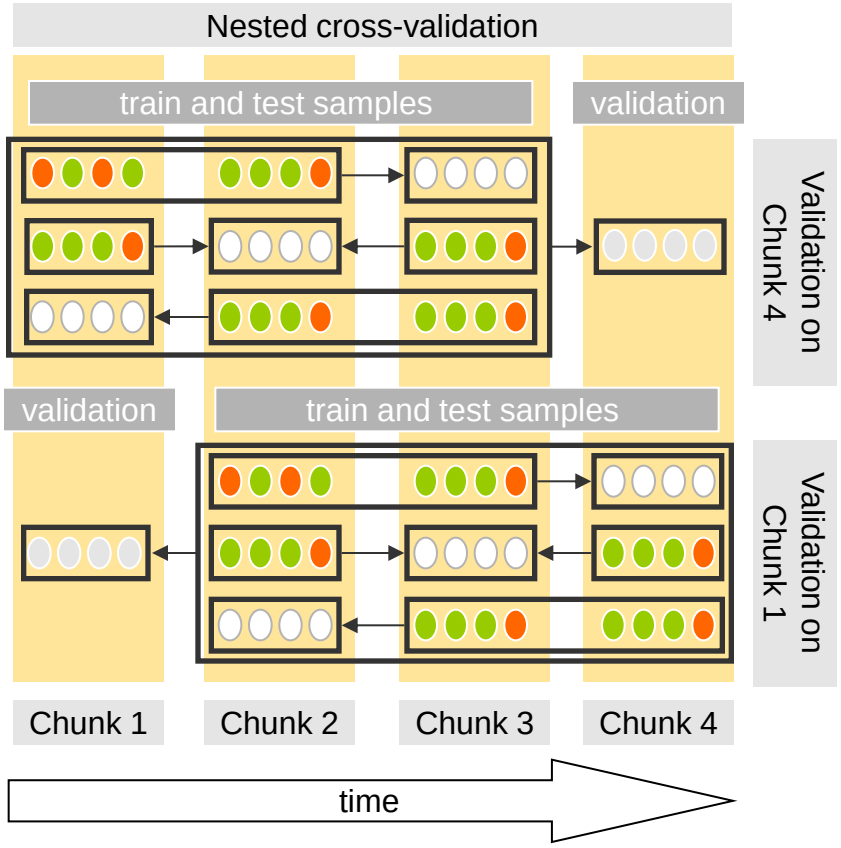
Thank you for listening

Dennis Weinbender

Chief Data Scientist
Fraud Management
Group Credit Risk Management

Tel.: +49 69 136 - 21 733
Mail: Dennis.Weinbender@commerzbank.com

Backup: Fraud-Engine uses State-of-the-Art Algorithms for Model Estimation and Validation



Repeated Nested cross-validation: prevents overfitting and hinders the forming of artifacts

- The inner layer is iterated several times and chooses a model which predicts new data and therefore generalizes fraud patterns well.
- Data cleaning, data subsampling and oversampling techniques are building blocks of the models, which need to be tested
- Different feature selection algorithms are also tested here
- After the final model is chosen, the outer layer ensures the robustness of the operating model
- Both layers are repeated several times to ensure stability with respect to data splits

Further computational advances ensure computability of the models, while maintaining high precision

- Downsampling of the majority class and structural upsampling of the minority class (SMOTE, ROSE)
- Adaptive hyperparameter search
- Subsample ensembling of the models (Subsembling)
- Using of the results of the models building to predict further approximation of the 0 / 1 signs.