Selection and optimization of historical data for the training of artificial intelligence in power plant engineering

Dipl.-Ing. Frank Gebhardt
Uniper Technologies GmbH

Facts about Uniper

Our operations

Power **Generation** | Commodity **Trading** | Energy **Storage**
Energy **Sales** | Energy **Services**

Main activities

Gas-fired plants
19.2 GW

Wind

Nuclear plants
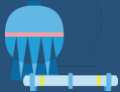1.9 GW

Energy sales

Coal-fired plants
10.5 GW

Gas pipelines & infrastructure

Hydroelectric plants
3.6 GW

Engineering services

Energy storage
Gas: 8.5 bn m^3

Liquefaction & regasification

Trading

Digital solutions

Content of today's presentation

- **Introduction** to the lecture / Focus

- **Challenges** of "Big Data"

- **Theory** of predictions of process variables with NN

- Integral **prediction**

- **Selection** and **evaluation** of training and test data

- **Case study**

Introduction to the lecture / Focus

- Application of artificial intelligence (AI) in power plant process engineering

  ►Encounter with "Big Data"!

- Procedures for the selection and optimization of historical data for the training of AI

- Applications for predicting process variables:

  - Steamgeneration

  - Temperature boiler ceiling

  - CO

  - etc…

- High expenditure of time for selection of learning patterns

  ► > 64,0000 learning patterns corresponds to > 1.5 Gbyte of data

Challenges of "Big Data"

- Data amount
  - cannot be processed because, for example, there are upper limits on the number of rows in Excel

- Performance Issues
  - is available with e.g. Excel, if you want to open a file > 1 GByte
  - may mean several minutes of waiting time

- Search for alternatives is required

- Algorithms for preprocessing the training data very helpful!
 ► Professional processing required for cost-effectiveness!

Theory of predictions of process variables with NN 1v2

Different approaches and possibilities for predicting process variables:

- Dependence on only **one** variable:
  - Only one signal is used to predict future behavior, e.g. only steam
  - This dependence is called "univariate" in mathematics
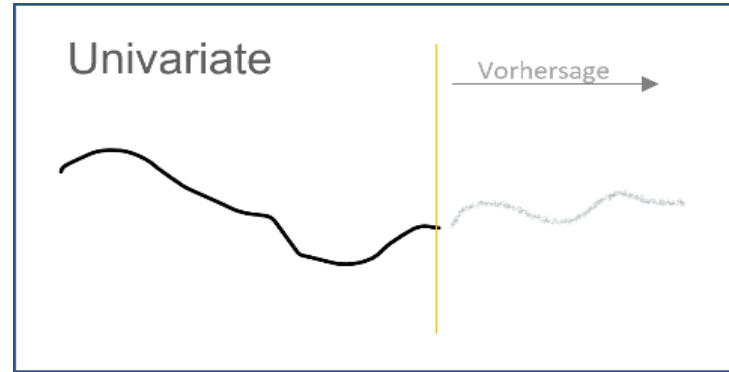
- Dependence on **several** variables:
  - Several signals are used to predict future behavior, e.g. steam, temperature, CO, air, ...
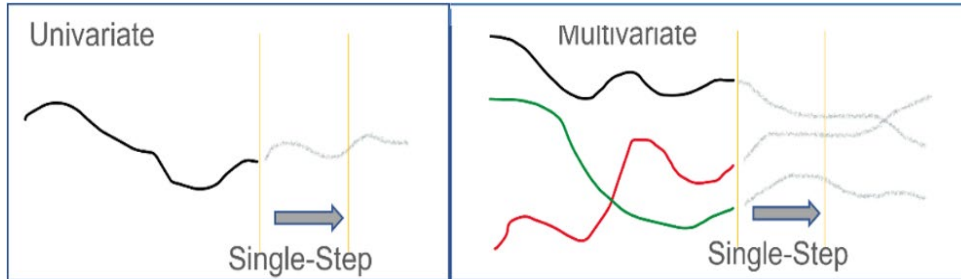 - This dependence is called "multivariate" in mathematics
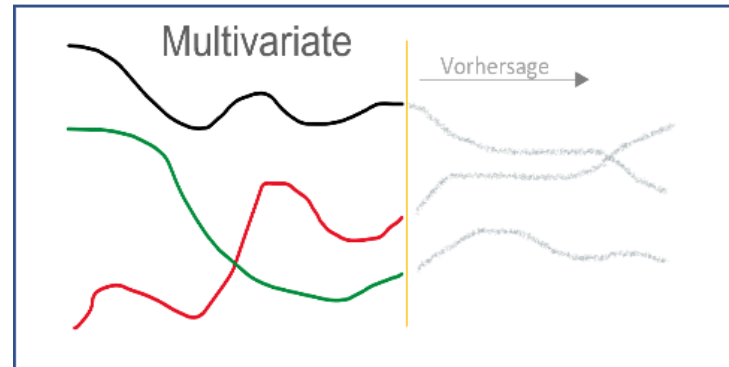
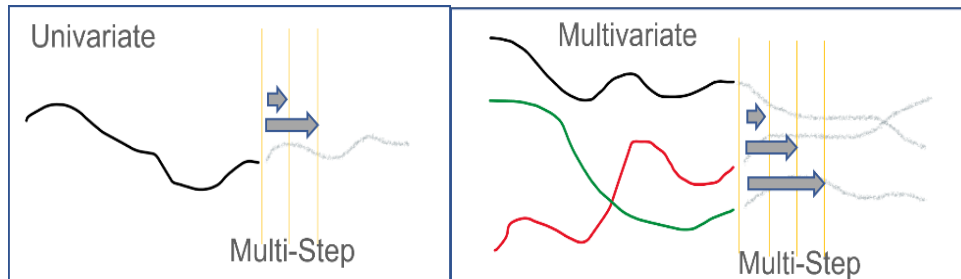# Theory of predictions of process variables with NN 2v2

Forecast horizons

There are different forecast horizons for predicting process variables:

- Single-Step: only one forecast horizon, e.g. only 5 minutes:

- Multi-Step: Predict multiple forecast horizons, e.g. 5 minutes, 10 minutes, ...

Integral prediction

- Multivariate and multi-level forecasting method

- Averaging over forecast horizon

- **Averages:**
  - easier to predict
  - contain the important process engineering information

- Confirmation by practice
  - ▶ Prediction of steam mass flows

Mittelwertbildungen
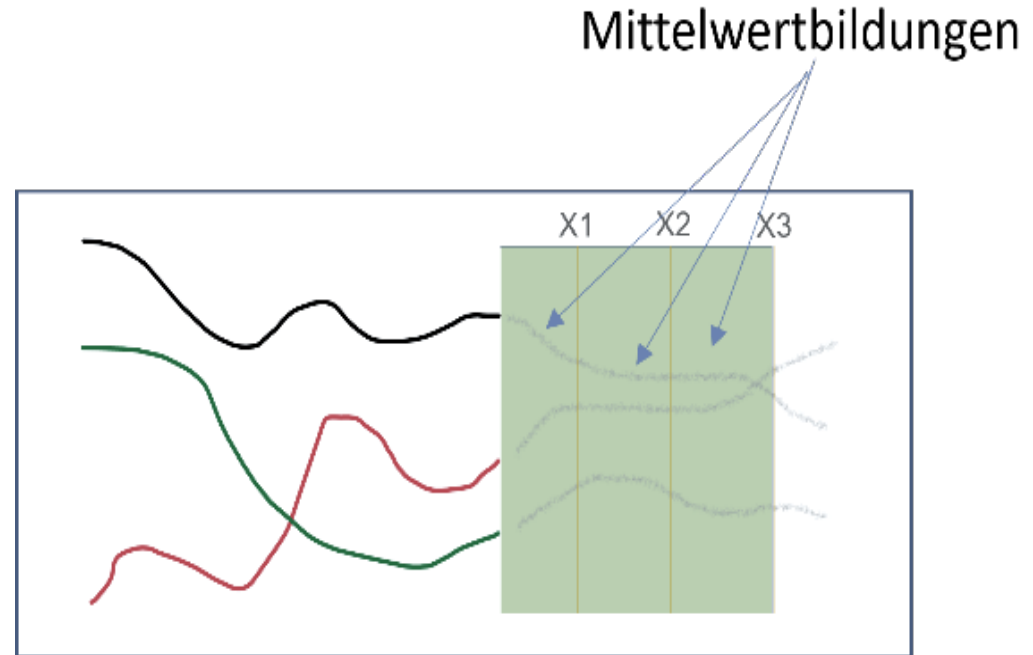
Selection and evaluation of training and test data 1v5

- Training dataset
  - 4 to 6 months for predictive model, e.g. steam prediction
  - Elimination of plant downtimes and disruptive process events
  - Storage of as much data as possible

- Test dataset
  - Selection of a small test set of a few weeks

    ► These are the data that the neural network has not learned, i.e. does not know

- Overfitting
  - Typical strategies to avoid overfitting don't seem to work here
  - Large datasets are the key!
  - Practical confirmation of steam prediction and temperature prediction boiler ceiling

    ► Use of very, very large data sets required!

uni
per

Selection and evaluation of training and test data 2v5

Complexity optimization (with the same NN topology)

- **Large** data set

    In case of underfitting (poor learning), the complexity of the NN is too low to model the complexity of the data set

    ► Solution: Reduce the number of trainable outputs to increase the available complexity per output or reduce variarity

- **Small** data set

    - With **Overfitting** the complexity of the NN is too great
    - The NN remembers each sample of the training data set, the generalization of test data will be poor in most of these cases
    - ► Solution: Increase the number of trainable outputs or increase training dataset or variarity

Selection and evaluation of training and test data 3v5

**Uneven** distribution of data in a data set

- "Unbalanced regression"
- Large 6-month data sets reflect typical distribution of actual plant operation

Number of input/output pairs	Spatial density	Probability of alternative similar input/output pairs
Accumulation	high	high
Shortage	low	low

Selection and evaluation of training and test data 4v5

Example of a data set distribution
►Negative correlation (-0.47) between the number of samples for a given operating behaviour and a test error

- **Frequent** data in a particular operating behavior results in a **small test error** for that behavior

- **Rare** data in a certain operating behavior leads to a **high test error** for this operating behavior
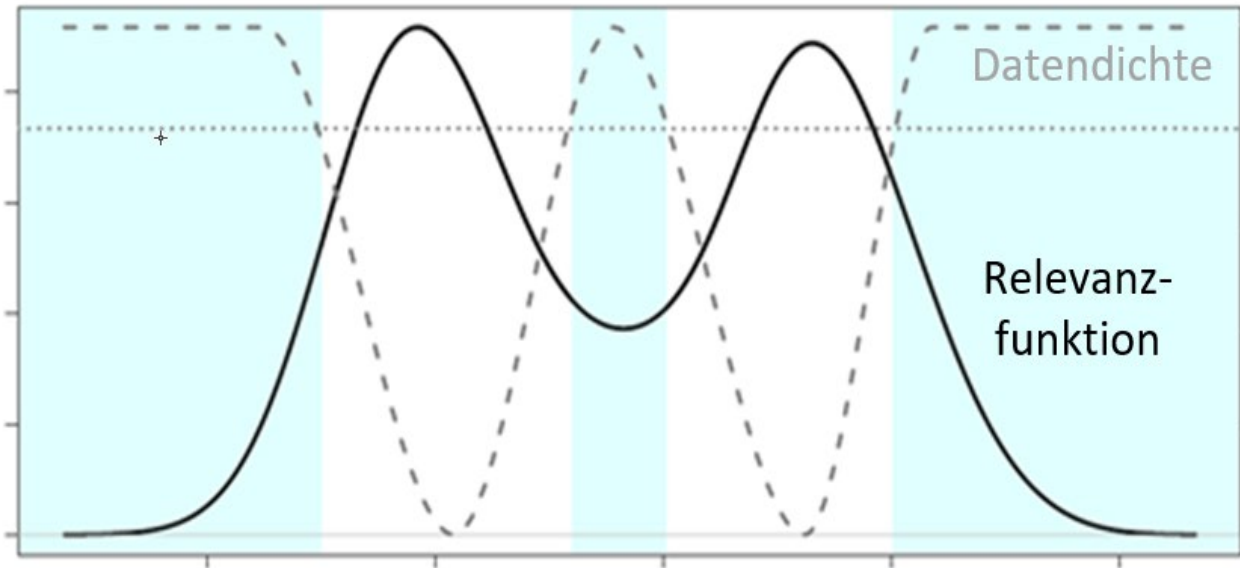

Pearson correlation: -0.47
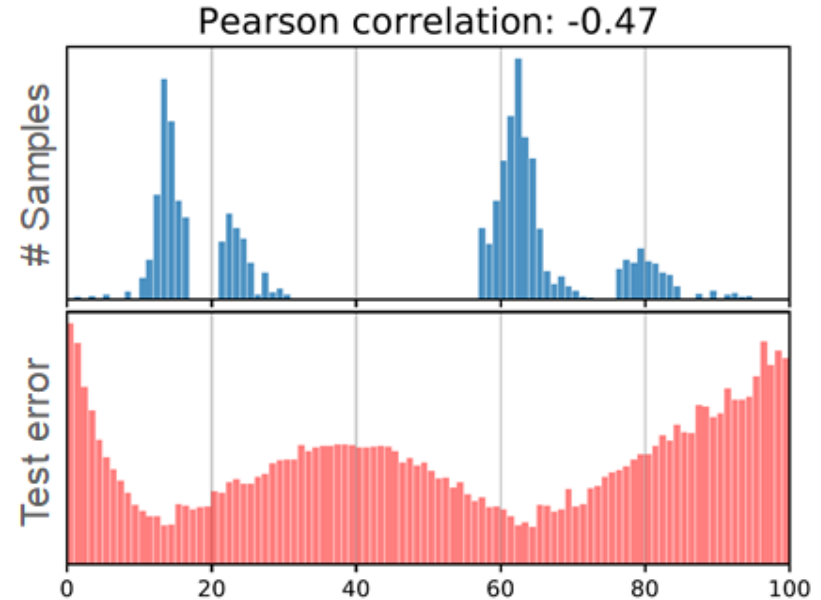
Selection and evaluation of training and test data 5v5

Relevance function

To improve the learning success of the NN:

▶ **Inverted** data density is used as a relevance function

There are many theoretical methods from the literature:

- SMOTER (2013)
- SMOGN (2017)
- **WERCS (2018)**
- Dense Loss (2021)

Steam prediction:
- Method WERCS
- Specific weighting of synapse weights
- Strong improvement
- Practical!

Practical example - Procedure

Standard procedure for balancing algorithms for balancing learning patterns according to the previous mentioned methods:
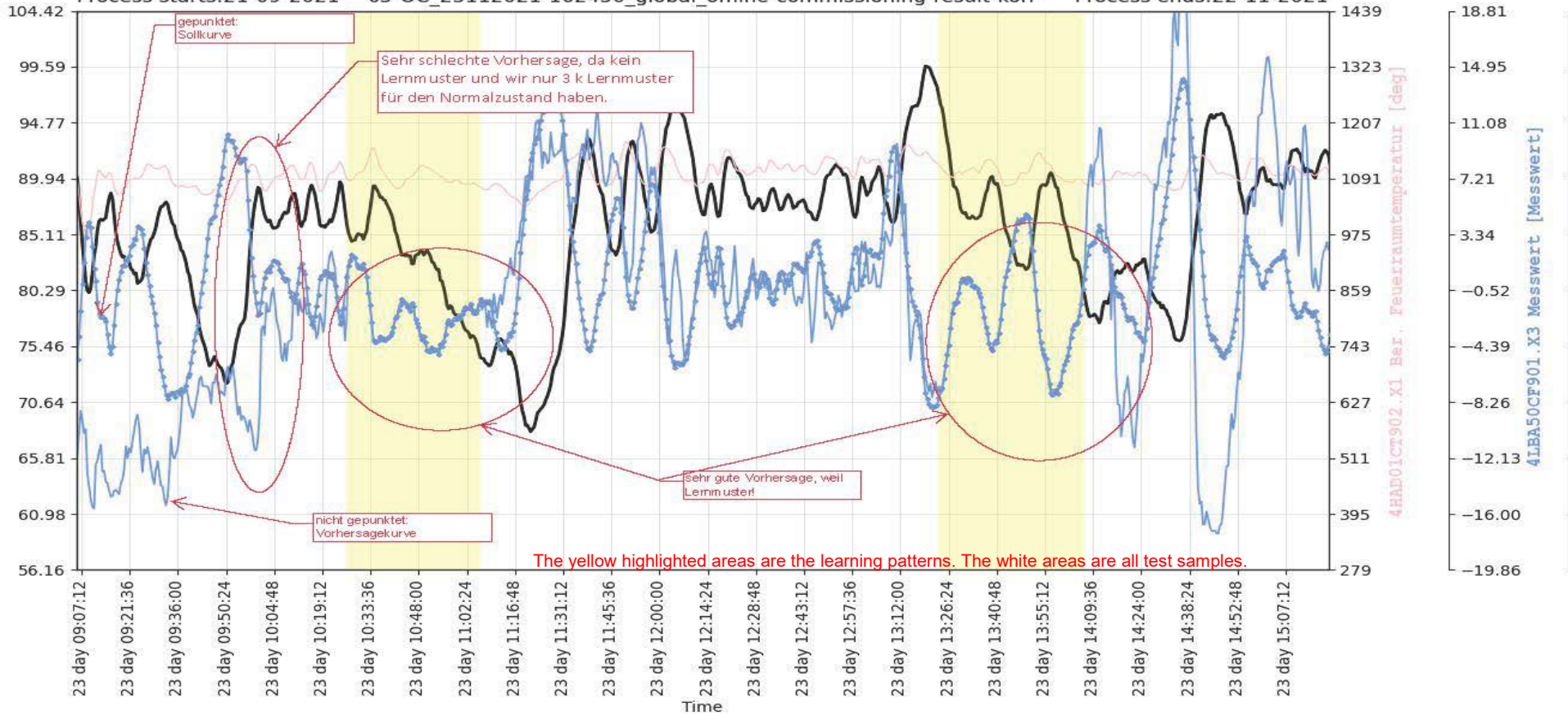
1. **Analysis** of the complete data set for common and rare data

2. **Balancing** the common and rare cases by using a combination of:

   - **Oversampling <u>rare</u>** data by creating synthetic data, such as adding random noise to existing data
   - **Undersampling for data clusters by finding data that:**
        - are close to each other and therefore approximately the same
        - are staggered in time and similar
        and can therefore be deleted.

Practical example - steam forecast WIP (waste incineration plant)   2v8

4k-Learning-Pattern-3k-Normal-1k-Steamdrop

The yellow highlighted areas are the learning patterns. The white areas are all test samples.

The following test pattern was found for the normal state (no steamdrop)

Process starts:21-09-2021   03-OC_23112021-162450_global_offline-commissioning-result-korr   Process ends:22-11-2021
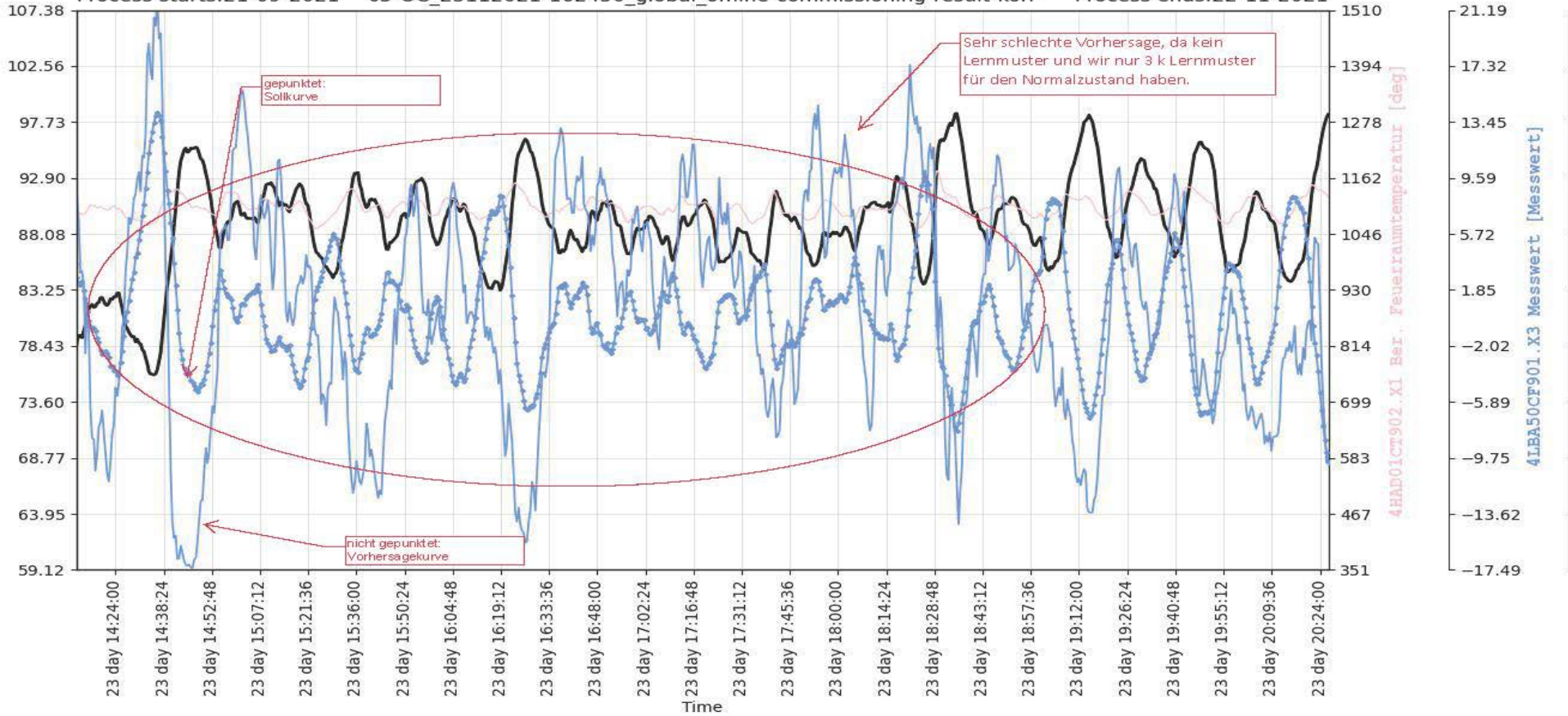
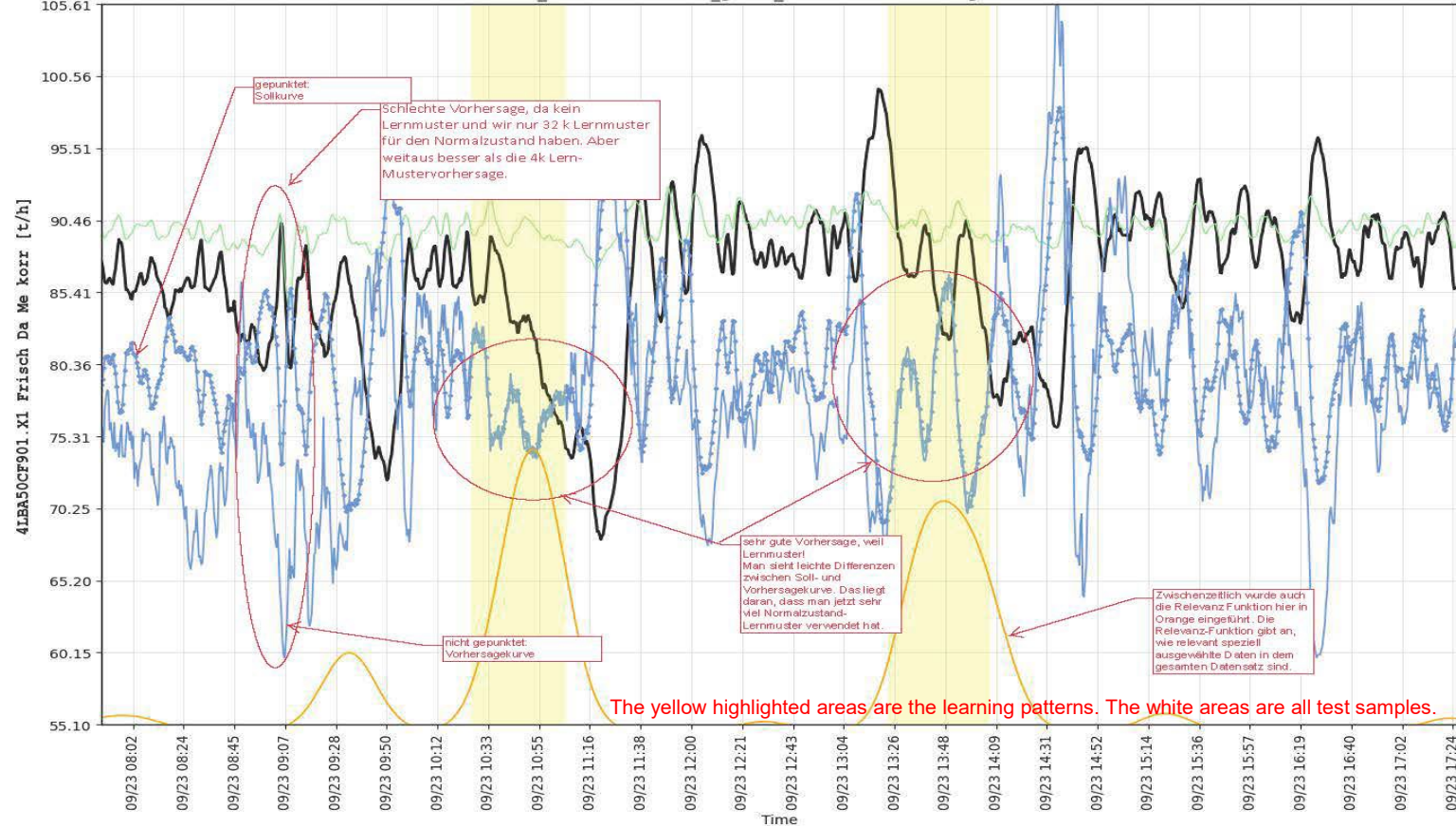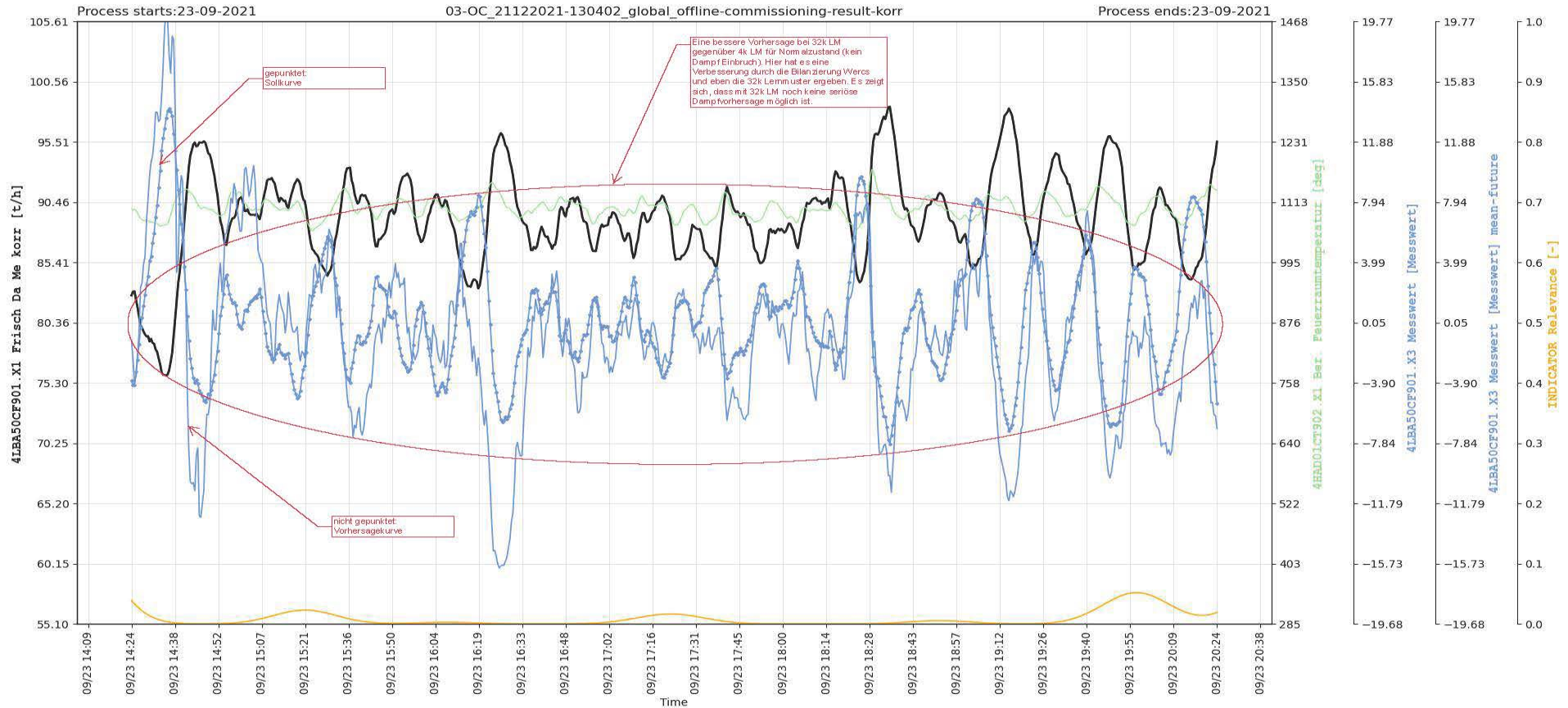Practical example - steam forecast WIP (waste incineration plant)   4v8

32k-Learning-Pattern-30k-Normal-2k-Steamdrop:

Relevance Function:
- Trend in Orange
- High relevance values lead to the reinforcement of learning patterns!

The yellow highlighted areas are the learning patterns. The white areas are all test samples.

The following test pattern was found for the normal state (no steamdrop)

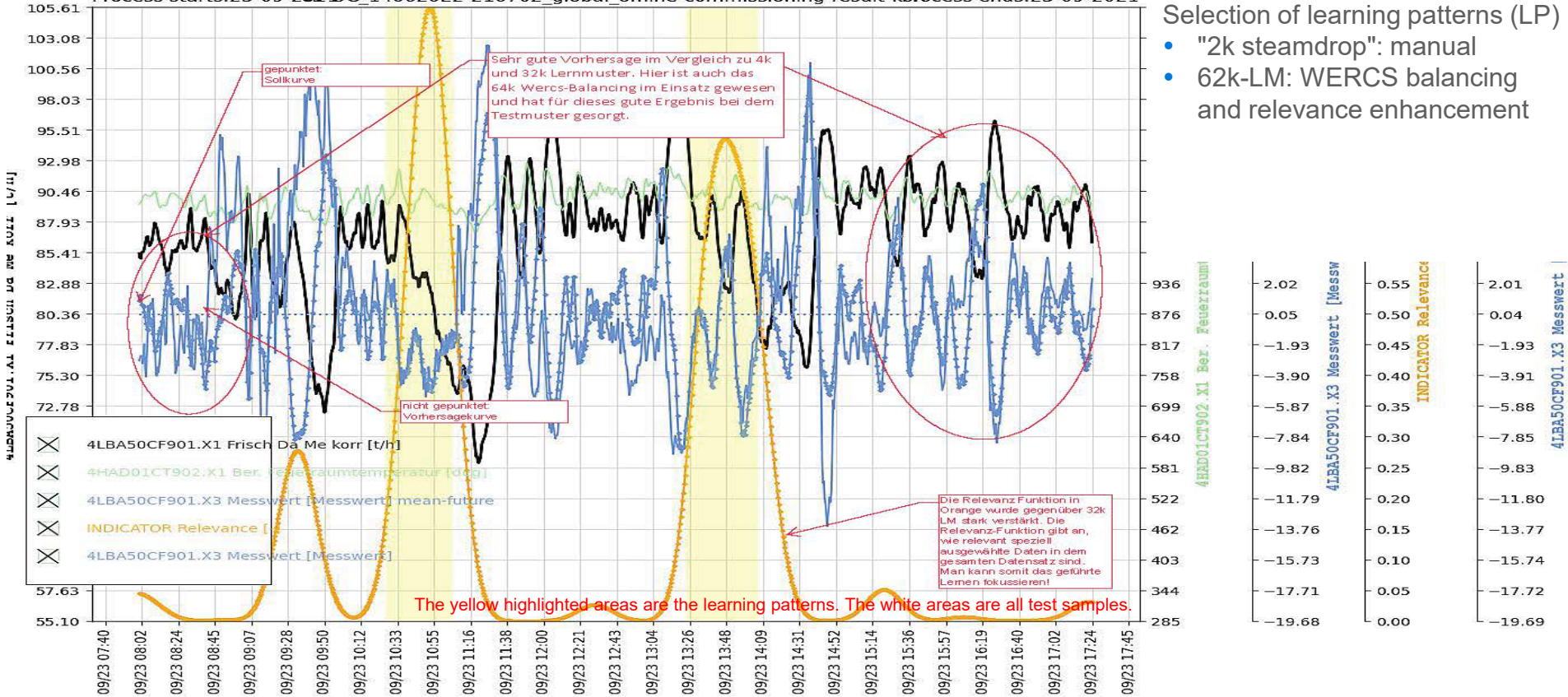# Practical example - steam forecast WIP (waste incineration plant)   6v8

64k-Learning-Pattern-62k-Normal-2k-Steamdrop:

Selection of learning patterns (LP)
- "2k steamdrop": manual
- 62k-LM: WERCS balancing and relevance enhancement

The following test pattern was found for the normal state (no steamdrop):

Practical example - steam forecast WIP _(waste incineration plant)   8v8

Relevance / Process Engineering Indicator – Definition

- The relevance was also used as a process engineering indicator, which was called here "Process category 1: Waste grate overfill" (double function!)

- Trends in measurements:

  - Primary air pressures: trend upwards
  - Flue gas O2 content: trend upwards
  - Fire rate control grate feed: trend upwards
  - Combustion chamber temperature: trend downwards
  - Moisture measurement: trend upwards

- Future values of steam production

  - 5 minutes: Trend downward
  - 15 minutes: Trend downward
  - 30 minutes: Trend downward

- Calculation of relevance / process engineering indicator

  - Mathematical linking of:
    - Trends in measurements
    - Future values of steam production

uni per

Frank Gebhardt

Project Leader Electrical & Instrumentation

Lead Engineer Artificial Intelligence

Frank.Gebhardt@uniper.energy

+49 177 611 5983