Karlsruhe Institute of Technology

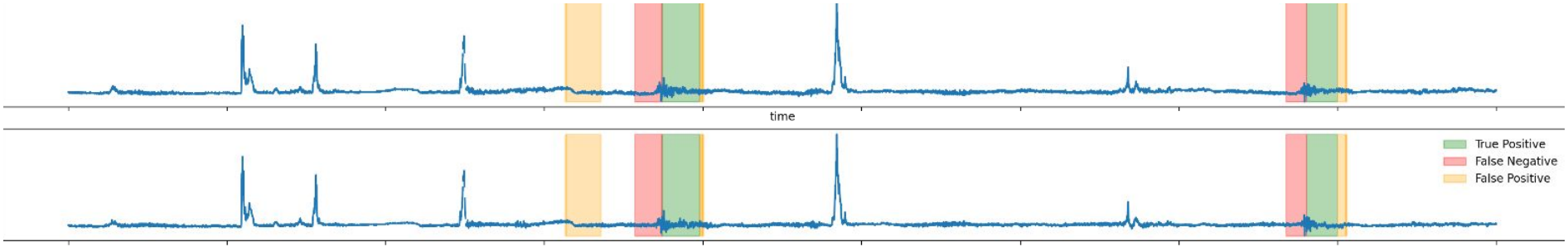
UFZ HELMHOLTZ Centre for Environmental Research

KIT-Campus Alpin
IMK-IFU: Atmospheric Environmental Research

Supervised and unsupervised machine-learning for automated quality control of environmental sensor data

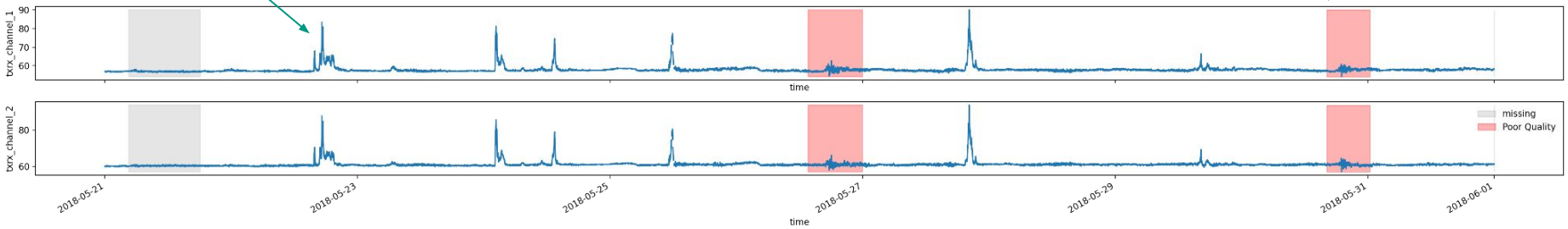
Julius Polz[1], Lennart Schmidt[3], Luca Glawion[1], Maximilian Graf[1], Christian Werner[1], Christian Chwala[1,2], Hannes Mollenhauer[3], Corinna Rebmann[4],Harald Kunstmann[1,2], and Jan Bumberger[3]

[1] *Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany*
[2] *Institute of Geography, University of Augsburg, Augsburg, Germany*
[3] *Department of Monitoring and Exploration Technologies, Helmholtz-Centre for Environmental Research (UFZ), Leipzig, Germany*
[4] *Department of Computational Hydrosystems, Helmholtz-Centre for Environmental Research (UFZ), Leipzig, Germany*



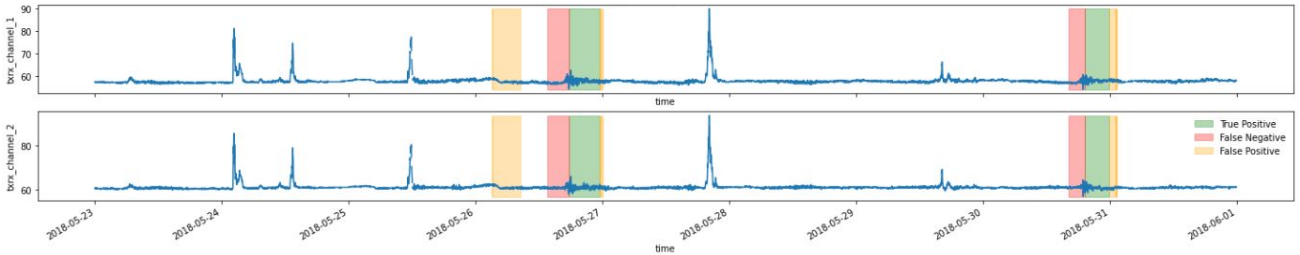
The general problem

Sensor networks for monitoring environmental variables like moisture, temperature or precipitation need quality control to provide usable data. The amount of data gathered grows beyond what manual quality control by trained staff can handle.



Thus, fully-automated solutions are required. Next to manually defined tests, which are tedious to parametrize and require expert knowledge, machine learning algorithms can be used to separate "good" from "bad", i.e. erroneous, data.

Example:
Automatic classification of erroneous data through a Convolutional neural network (CNN)



Example dataset 1: Commercial Microwave Links

Dataset
- 400 CMLs
- 1 month of flagged data
- Distributed all over Germany

Used for:
Path averaged **precipitation** intensity

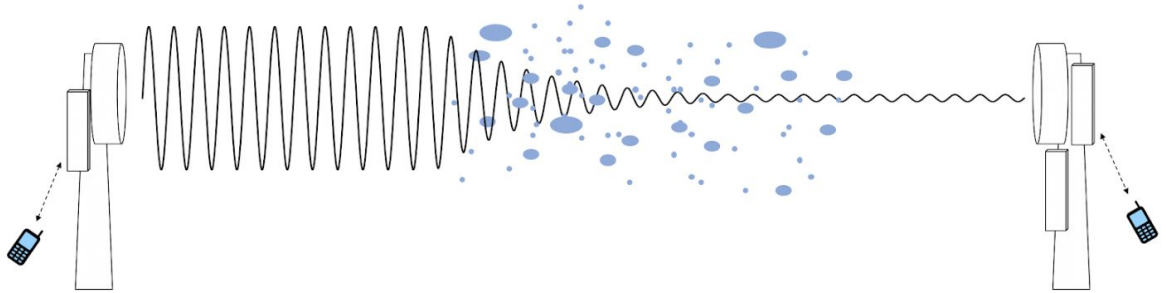


Fig.: Rainfall attenuates microwave signal of a commercial microwave link. Graf et. al (HESS 2020)

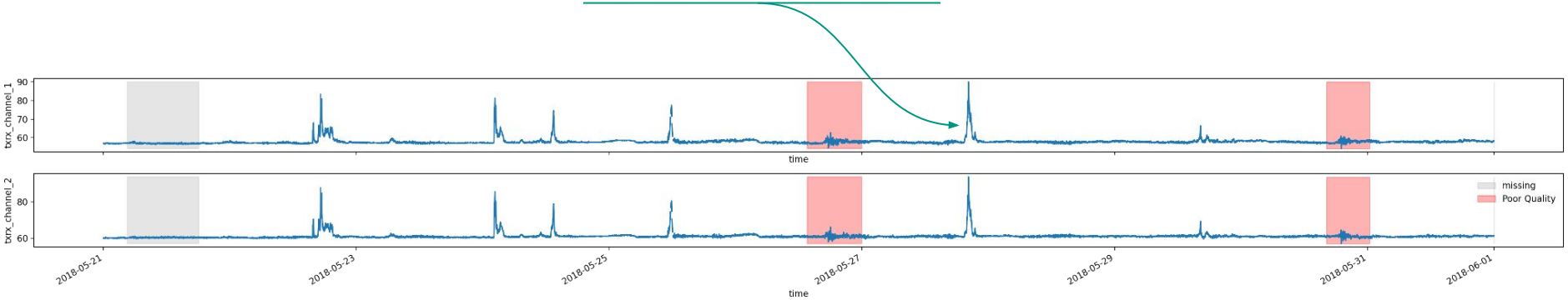


Fig.: 1 min resolution time series of: **Transmitted minus received signal level (TRSL);** 2 channels, one for each transmission direction between 2 antennas.

Example dataset 2: Distributed soil moisture/temperature network

Dataset
- 234 Sensors
- 2 years of flagged data
- TERENO-site “Hohes Holz”

Used for:
Soil **moisture**
Soil **temperature**

Fig.: 6 SoilNet sensors in one location sharing a battery. Later buried at various depths.

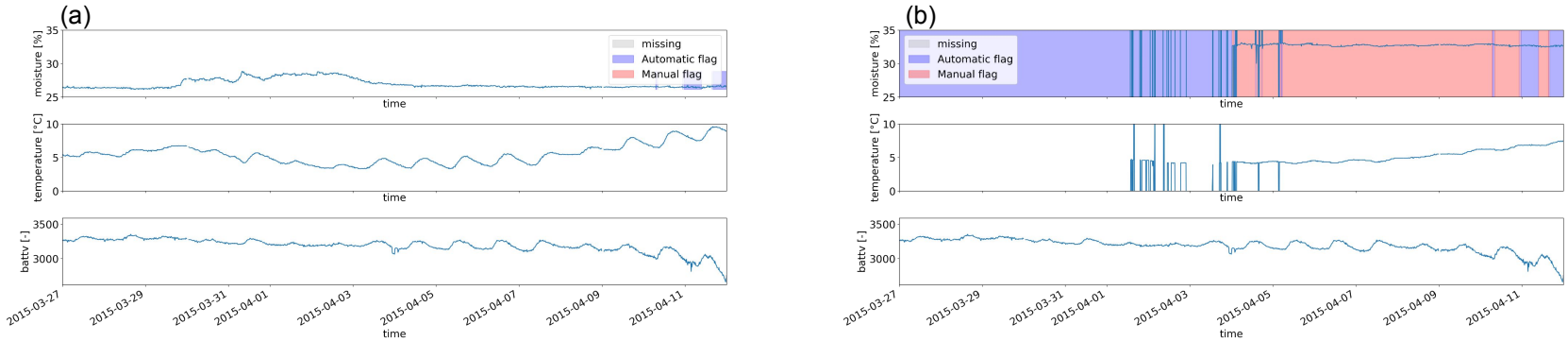
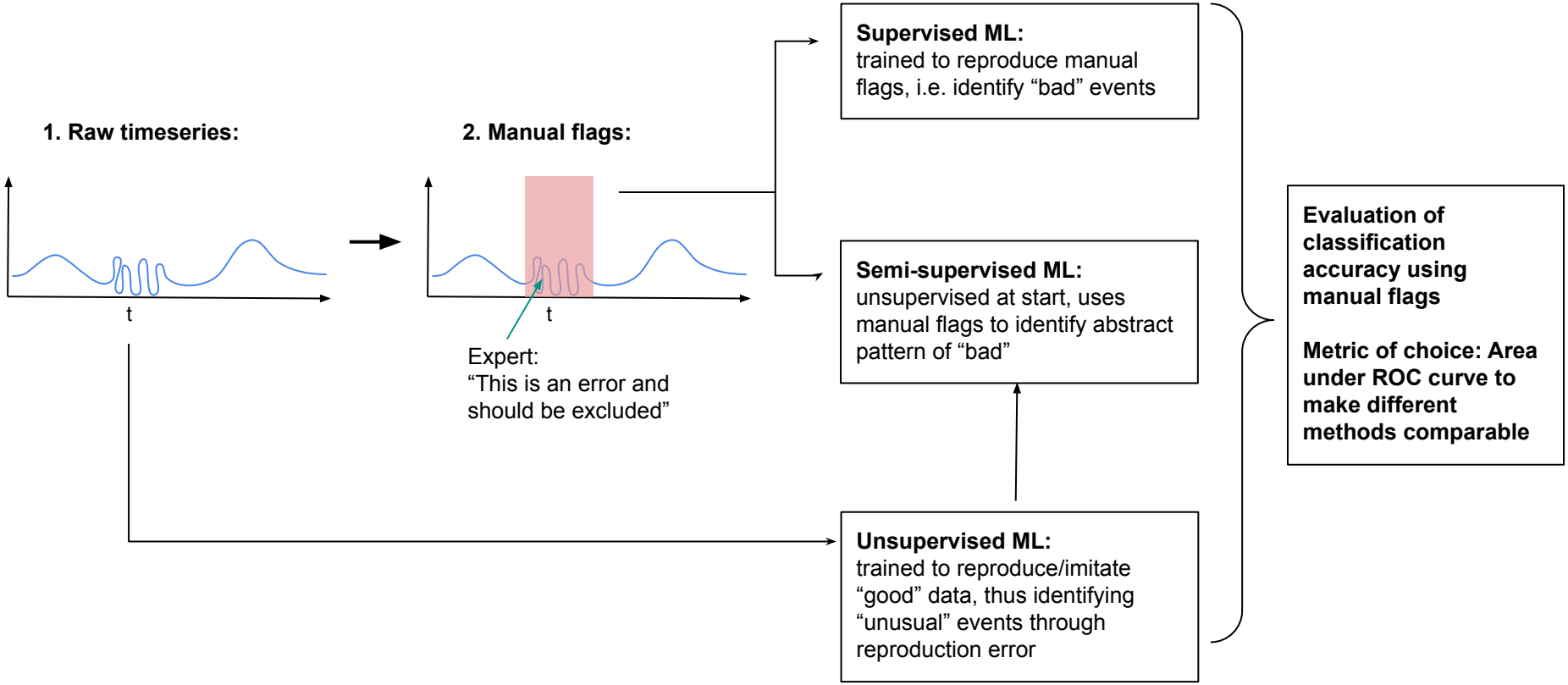
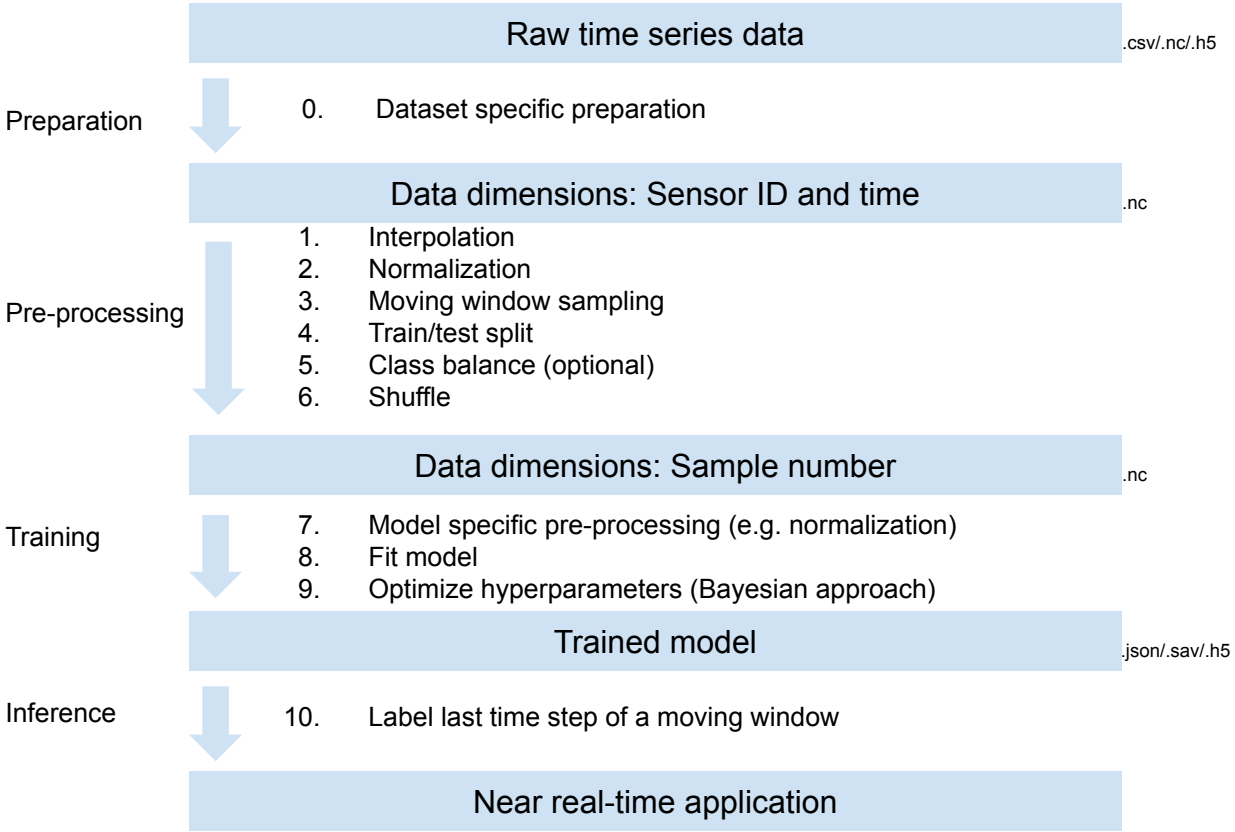


Fig.: 15 min resolution time series of: **soil moisture + soil temperature + battery voltage;** good data (a) and erroneous data (b)

Automated quality control: Workflow

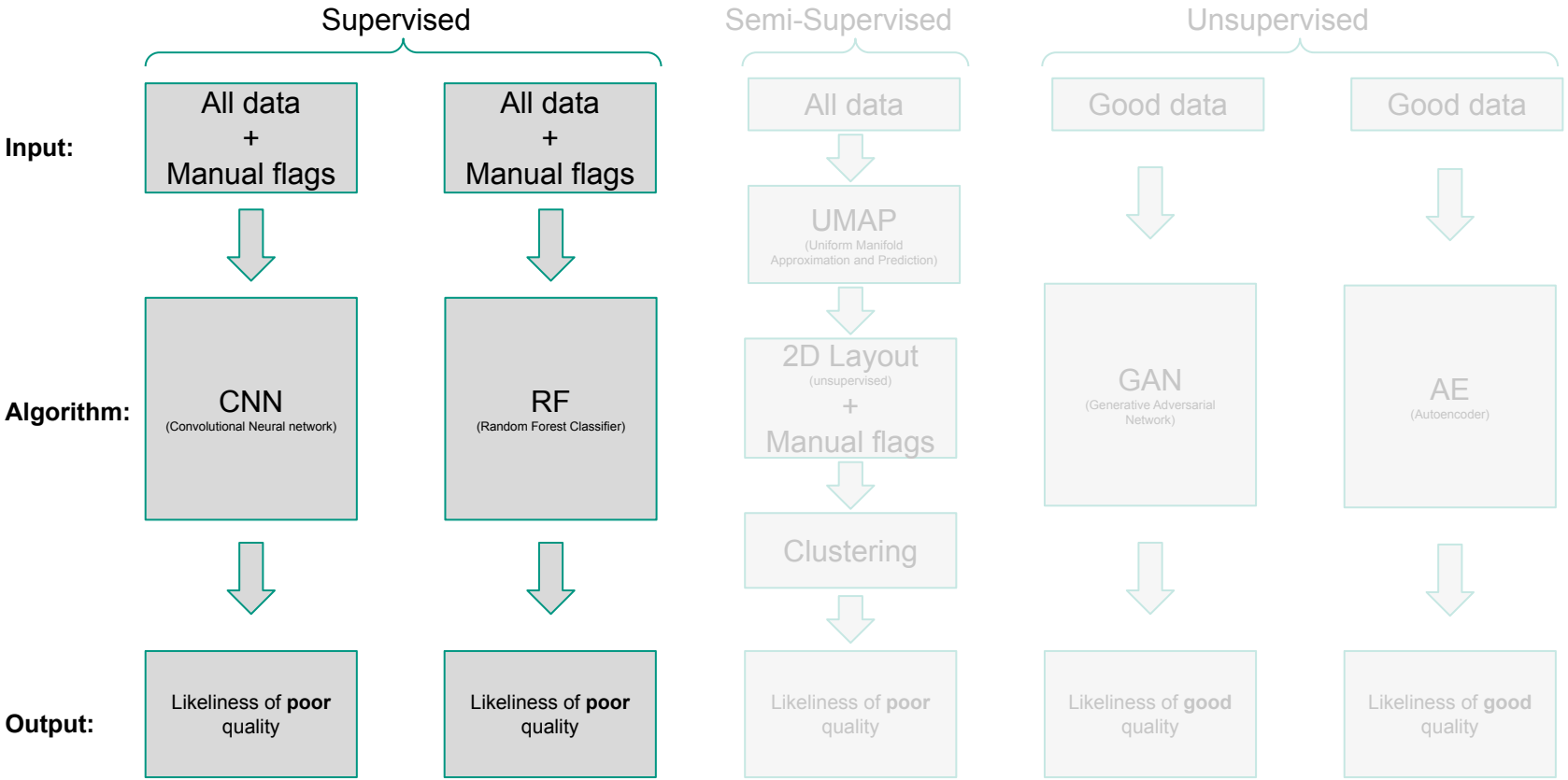


Automated quality control: Processing



- The reference of our results and the labels for supervised training are generated by manual quality control through expert knowledge.
- Although using this subjective reference is tricky, our research topic is the reproducibility of these quality flags.

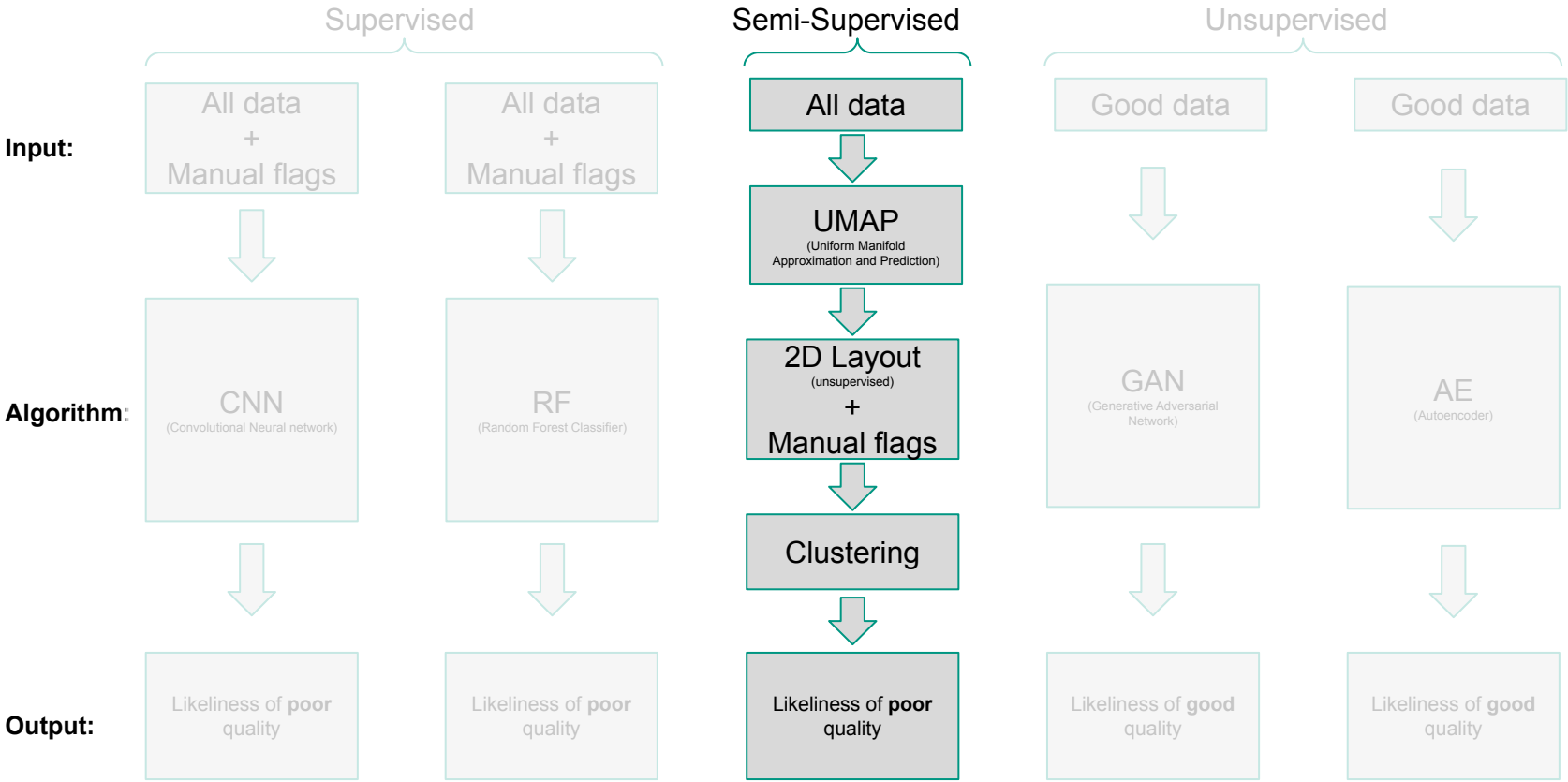
Automated quality control: Models



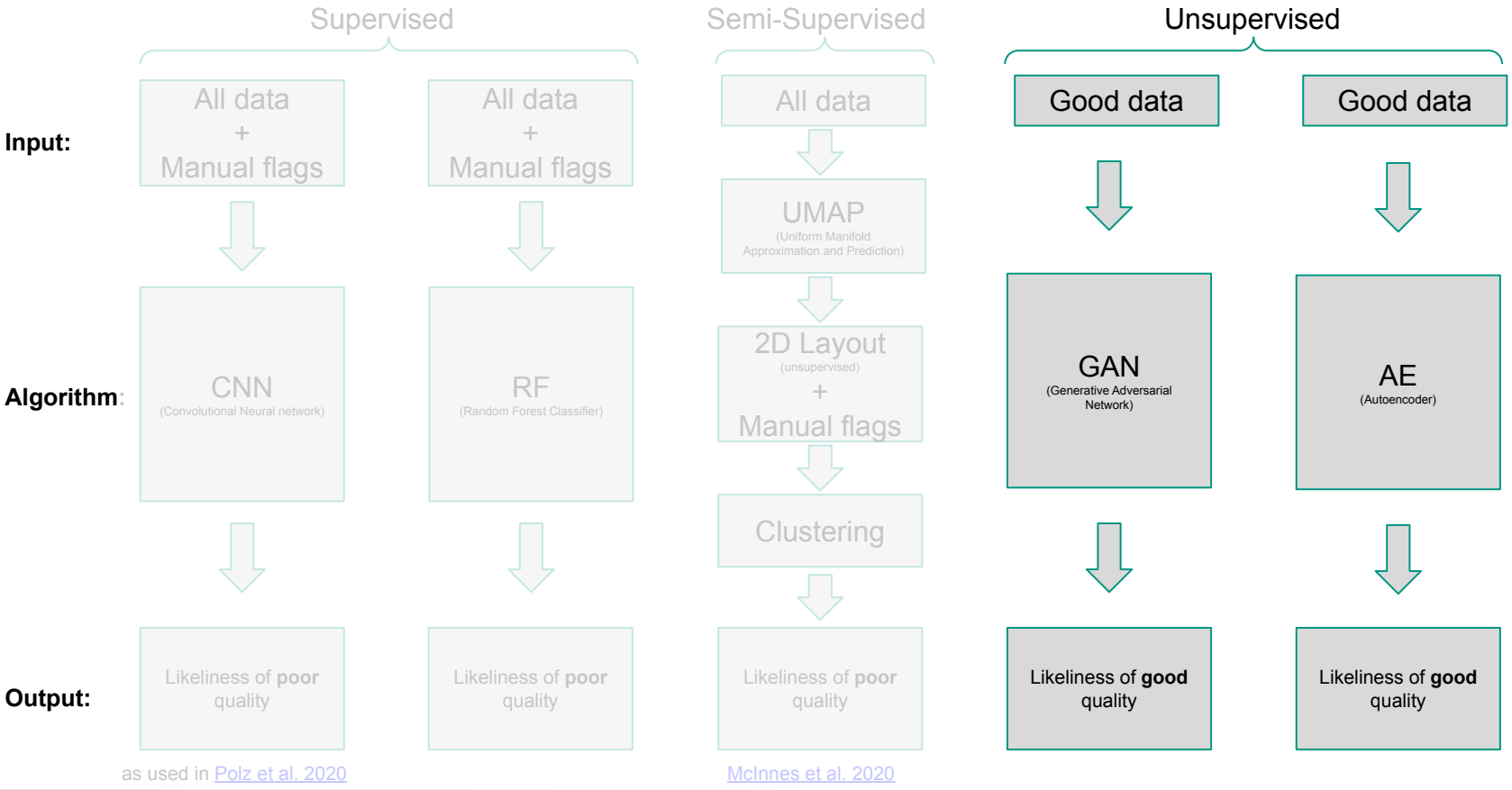
as used in Polz et al. 2020

McInnes et al. 2020

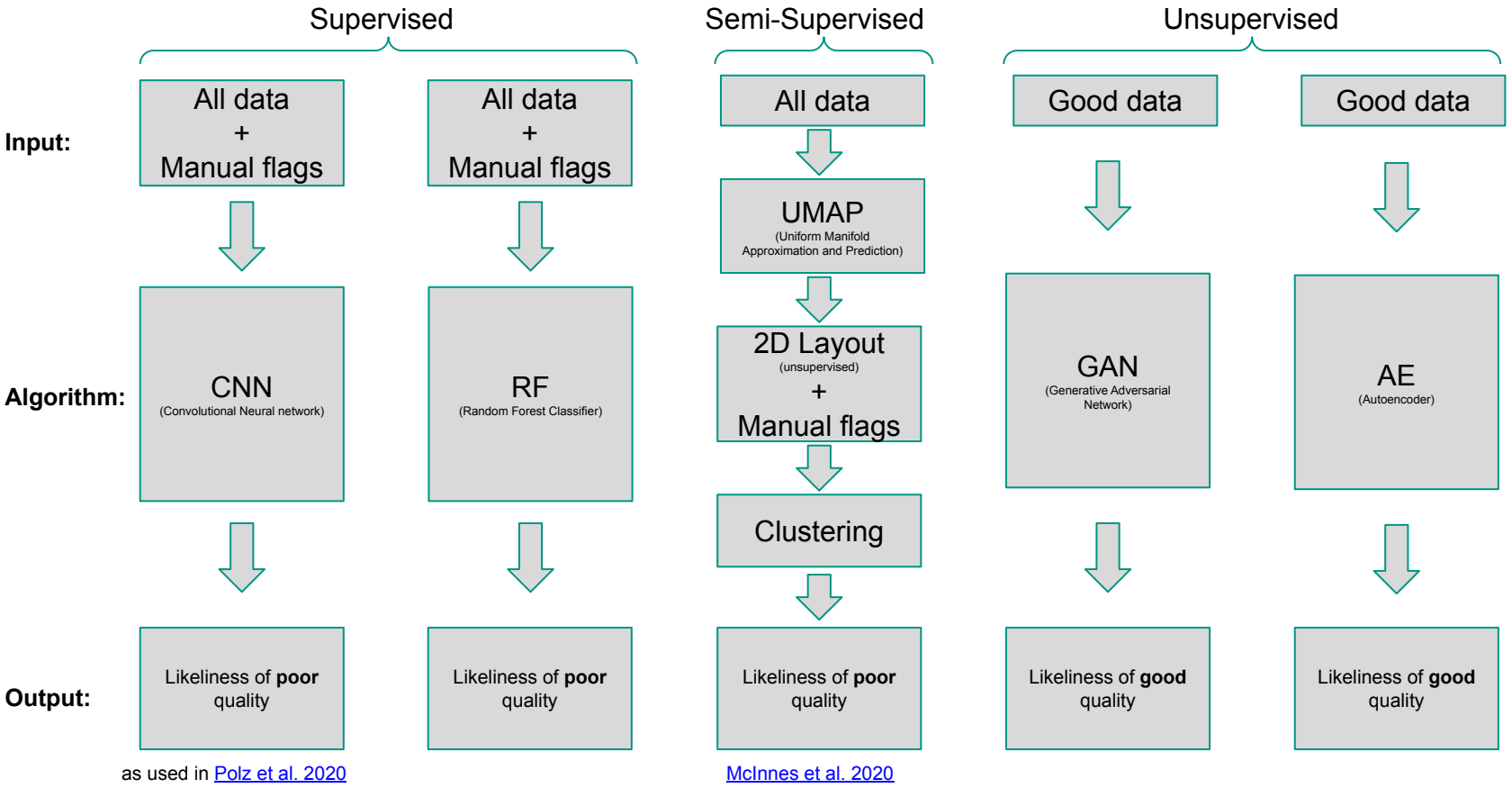
Automated quality control: Models



Automated quality control: Models



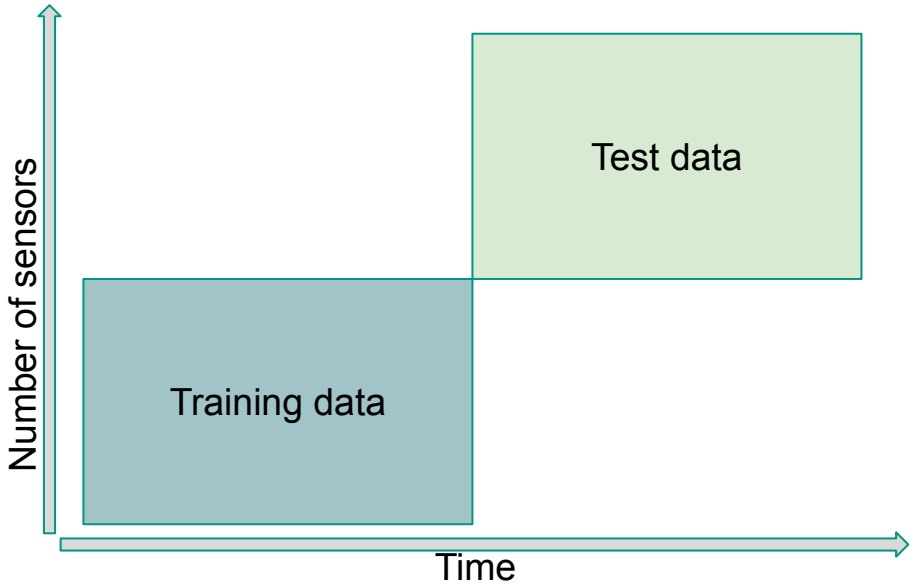
Automated quality control: Models



as used in Polz et al. 2020

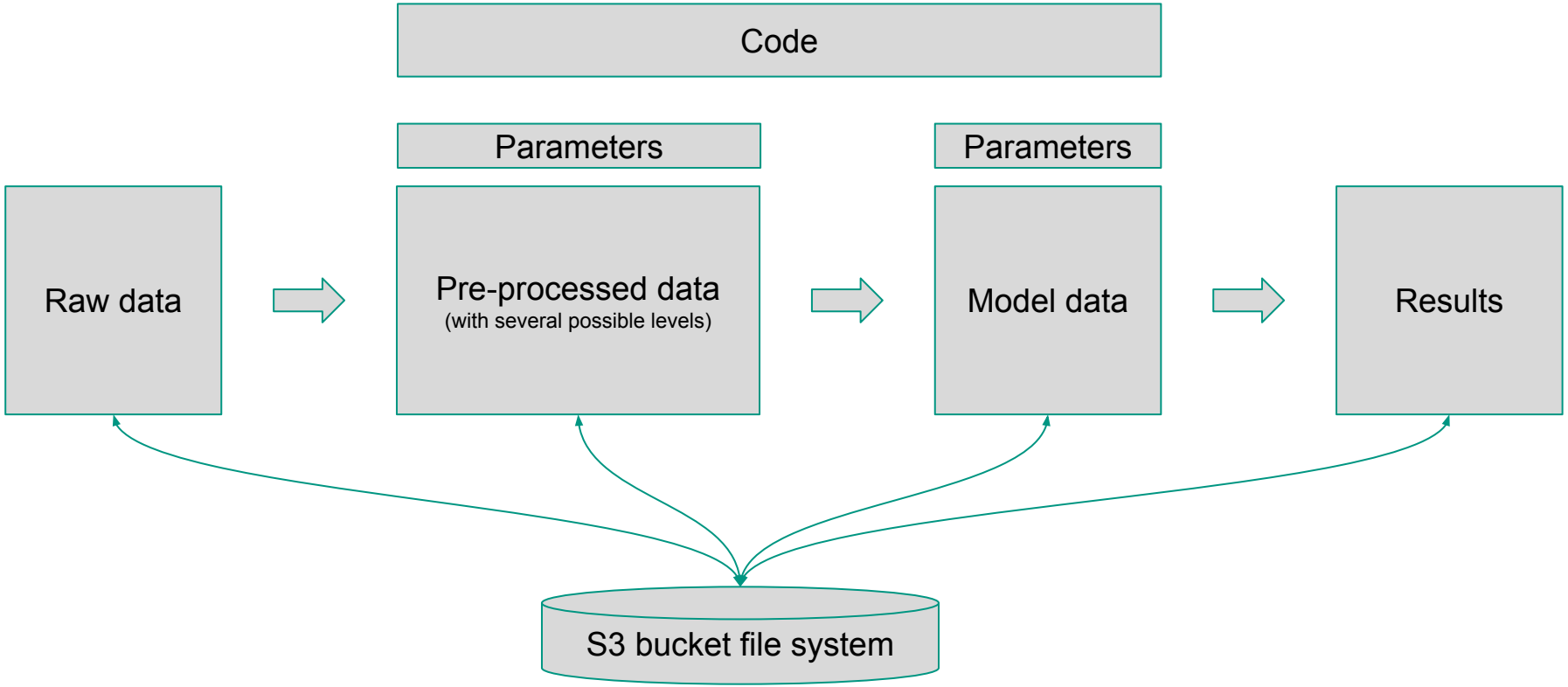
McInnes et al. 2020

Automated quality control: Train/Test split



- To avoid ‘information leakage’ from test to train and to test the model performance under changing conditions, we separate the data sets along both the time and the sensor number dimensions.
- Our split-set configuration: 50 - 50 in time and sensor space.
- This approach is also a realistic setup for near real-time applications.

Model and data versioning

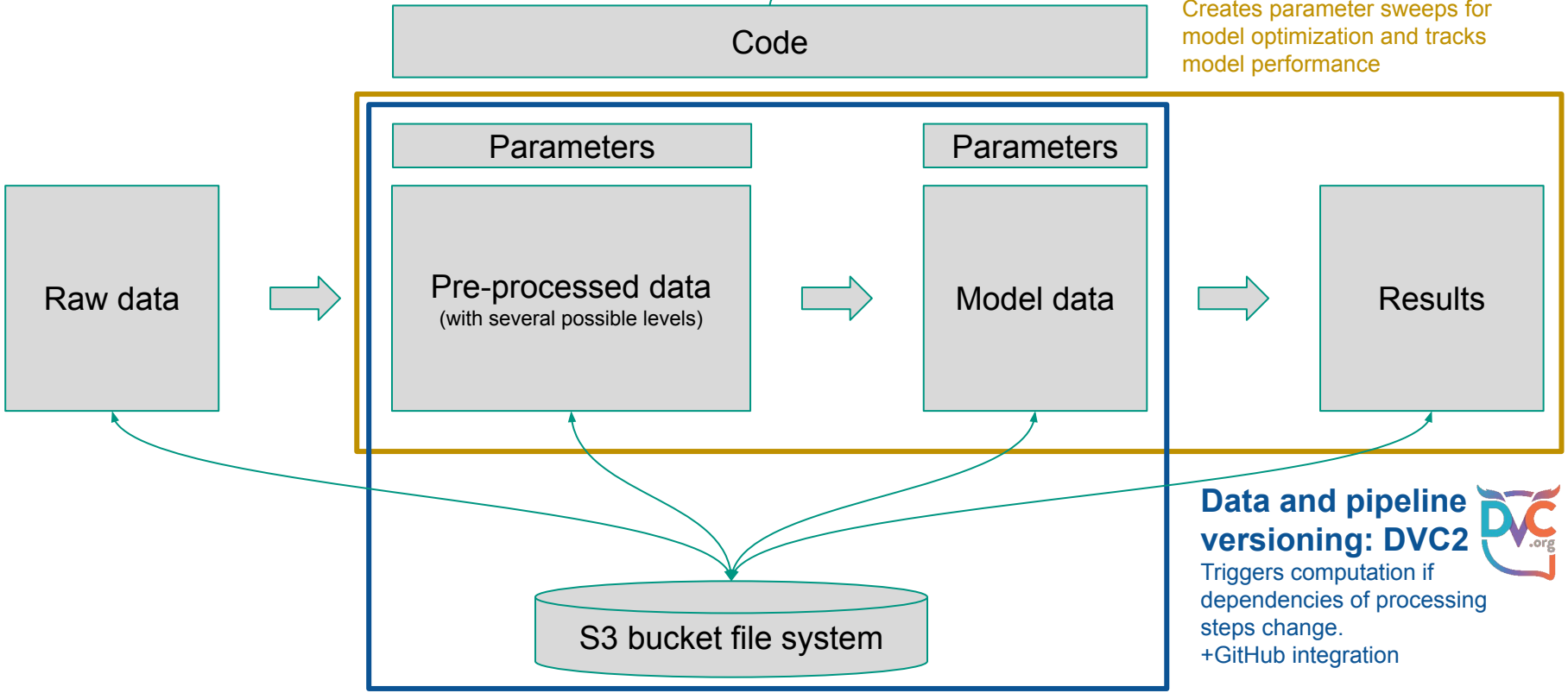


Model and data versioning

Source control: GitHub

Experiment tracker: Weights & Biases

Creates parameter sweeps for model optimization and tracks model performance



Data and pipeline versioning: DVC2

Triggers computation if dependencies of processing steps change.
+GitHub integration

Automated quality control: Results

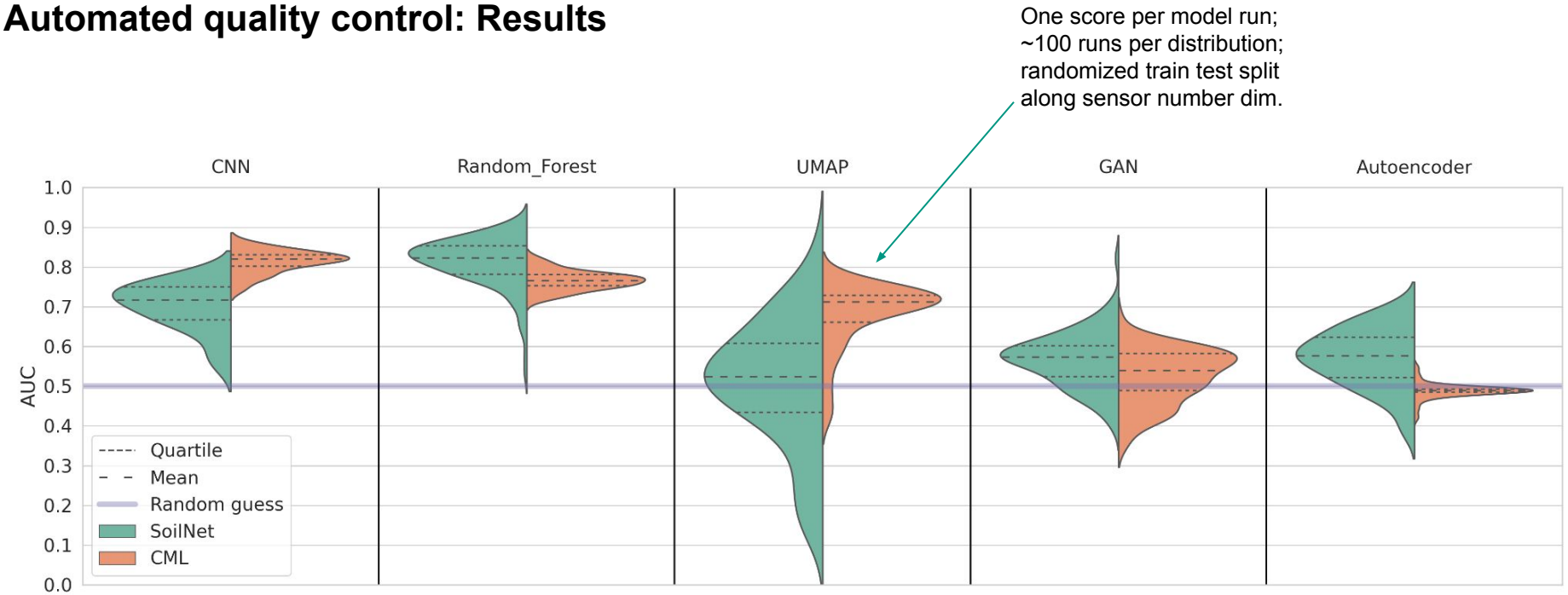


Fig.: Distribution of area under roc curve (AUC) scores for multiple hyper parameter optimization runs per data set and model type.

- As expected the supervised algorithms perform better than the unsupervised.
- Still, it is possible to achieve reasonable scores for all models and datasets.
- The CML-Autoencoder case is currently under investigation.

Automated quality control: Results

Data variables

TRSL 1

TRSL 2

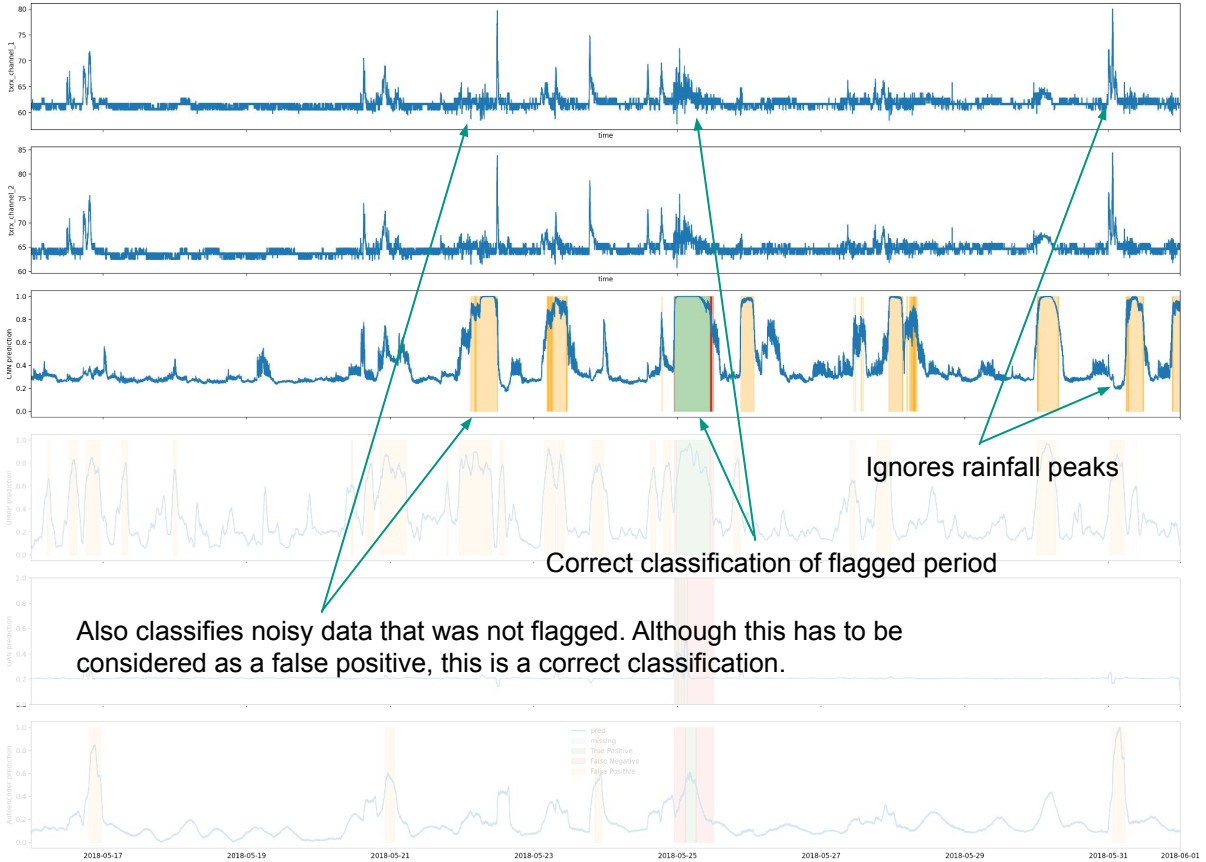
Probability of erroneous data

CNN

UMAP

GAN

Autoenc.



True positive

False negative

False positive

True negative

Fig.: Example CML time series illustrating the behavior of different methods

Automated quality control: Results

Data variables

TRSL 1

TRSL 2

Probability of erroneous data

CNN

UMAP

GAN

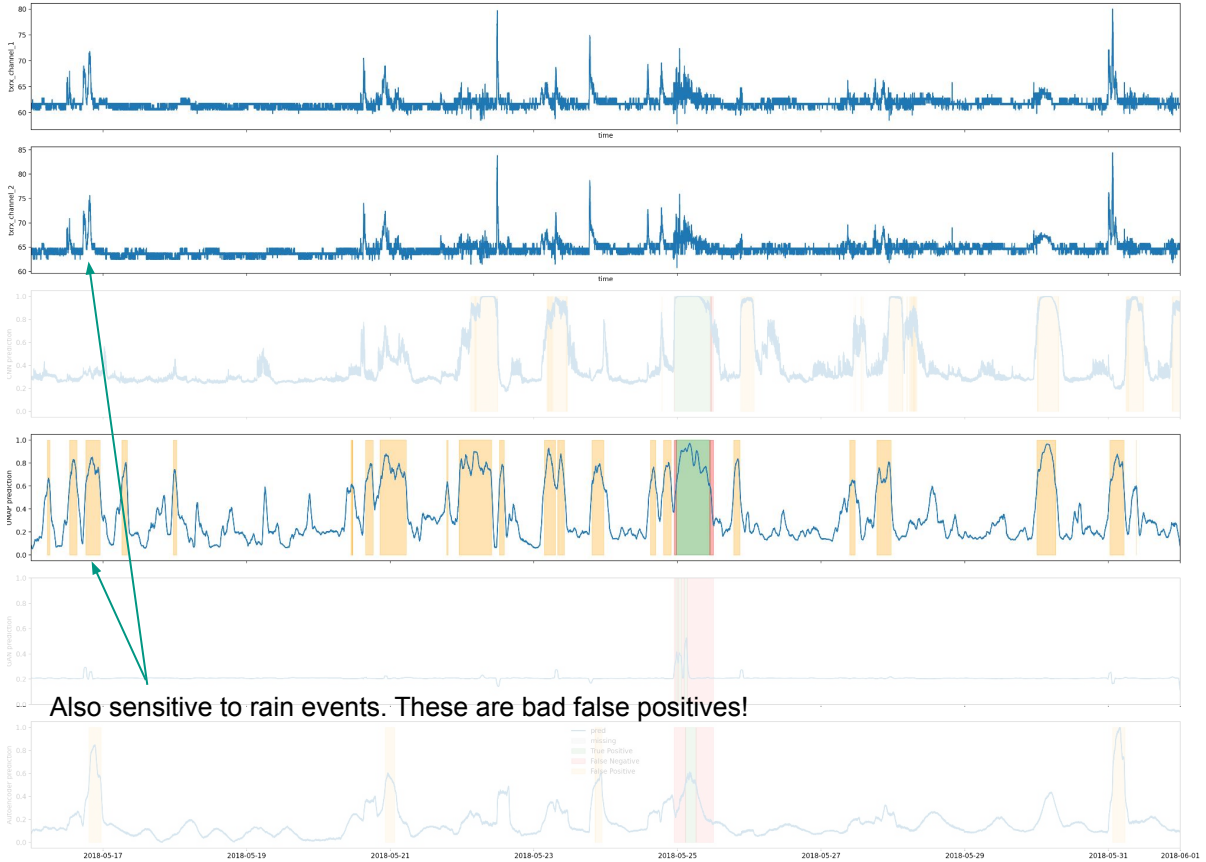
Autoenc.

True positive

False negative

False positive

True negative



Also sensitive to rain events. These are bad false positives!

Fig.: Example CML time series illustrating the behavior of different methods

Automated quality control: Results

Data variables

TRSL 1

TRSL 2

Probability of erroneous data

CNN

UMAP

GAN

Autoenc.

True positive

False negative

False positive

True negative

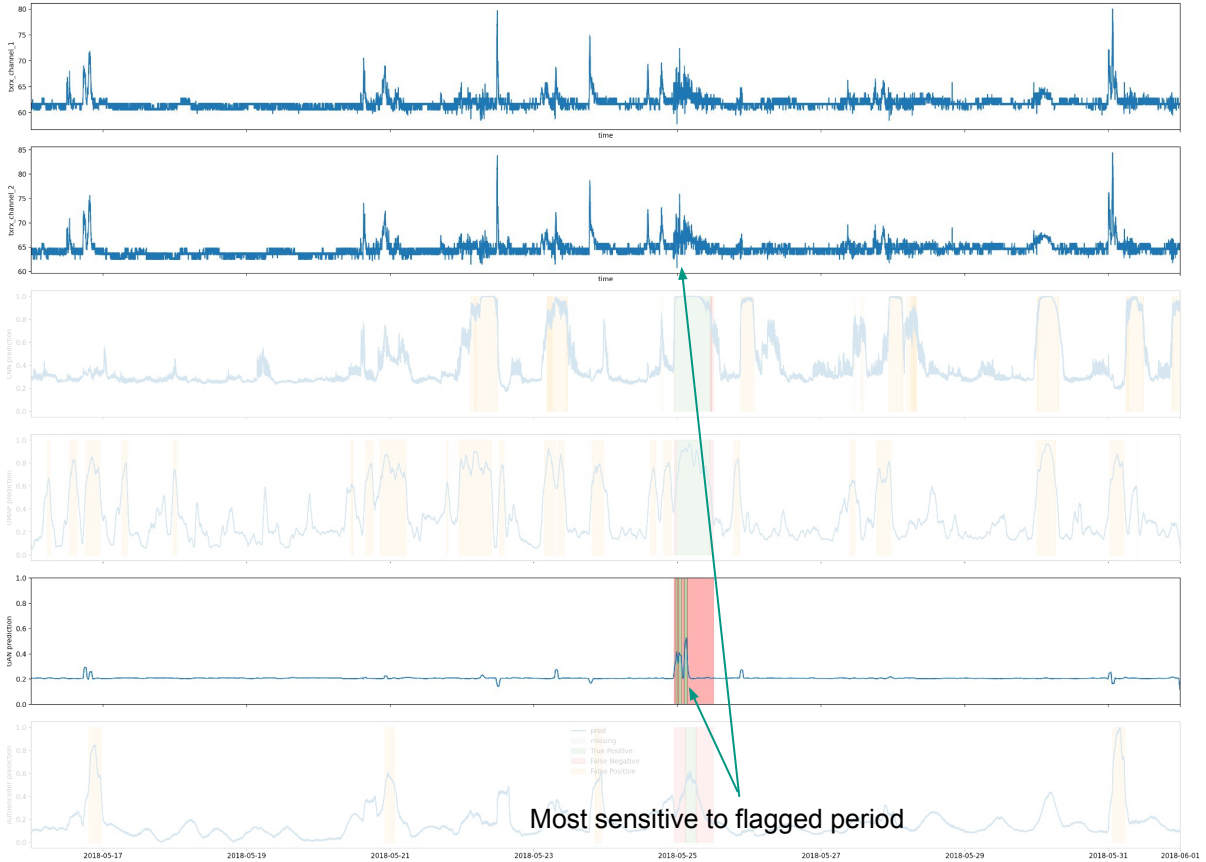


Fig.: Example CML time series illustrating the behavior of different methods

Automated quality control: Results

Data variables

TRSL 1

TRSL 2

Probability of erroneous data

CNN

UMAP

GAN

Autoenc.

Also sensitive to rain events

True positive

False negative

False positive

True negative

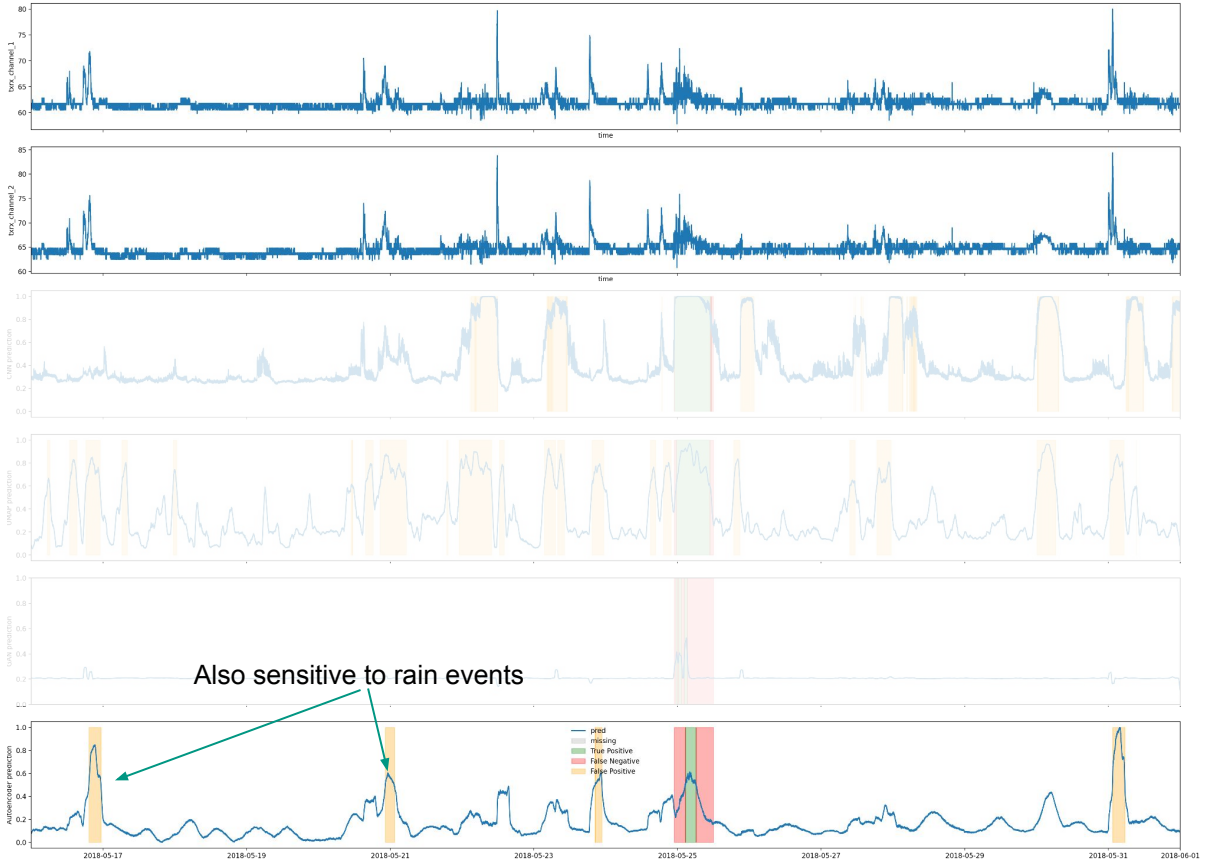


Fig.: Example CML time series illustrating the behavior of different methods

Automated quality control: Results

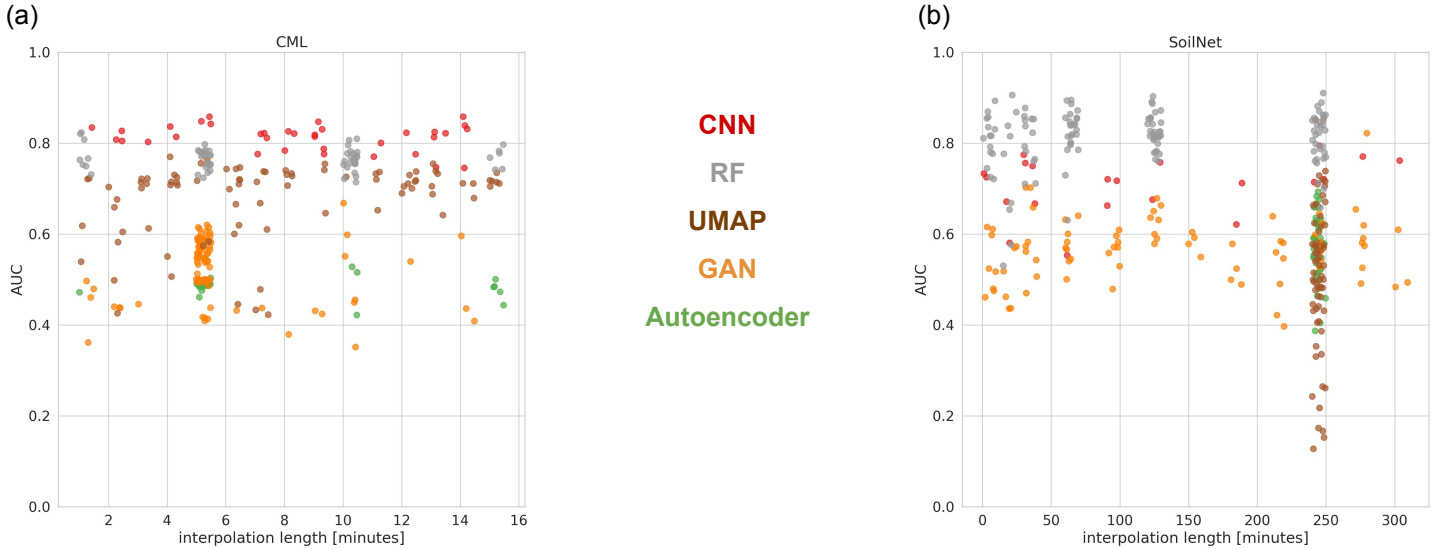


Fig.: Score vs. length of linear time series interpolation for the CML (a) and SoilNet (b) dataset. No obvious dependency could be observed for one of the models. An explanation could be that the missing periods themselves are not subject to quality control, but serve as an optional source of information.

Automated quality control: Discussion

- All algorithms are **insensitive to linear interpolation** of missing periods, increasing robustness to missing data.
- Classification accuracy of supervised algorithms is higher than that of semi-/unsupervised ones.
- Semi-supervised algorithm shows satisfactory results considering the potential practical benefits. For the selection of clusters, **training data can be reduced** drastically. (preliminary result not shown)
- Unsupervised algorithms' **sensitivity to good data**, such as rainfall spikes, **can be reduced** by increasing their frequency in the training data. (preliminary result not shown)
- Training with and comparison to the manual flags by experts introduces subjectivity into accuracy metrics. Revisiting the data shows that:
  - \+ Algorithms performance can surpass the experts flagging by higher temporal precision.
  - \- Performance of algorithms can be hard to interpret without a good reference.

Automated quality control: Outlook

- Methodology
  - Test pipeline for new data sets
  - Compare hyperparameters between many datasets to derive best practices
  - Include spatial variables in the unsupervised case (i.e. neighboring sensors)
  - Further improvements with GANs and Autoencoders
  - Compare to non machine learning approaches
  - Add domain adaptation techniques

- Practical Application
  - Test and improve robustness to missing data by refining interpolation
  - Identify minimum amount of manually flagged training data needed
  - Test transferability between measurement sites
  - Test generalization to changes in environment (experimental set-up, climate..)

- Technical Vision:
  - Entirely reproducible pipeline, i.e. dataset, model + code versioning
  - Python package with easy application to new datasets

Automated quality control: Key messages

+ Supervised algorithms provide good and stable performance, ...	− … but they rely on a lot of training data.
+ Unsupervised algorithms can lead to good classification, …	− … but the performance is less stable and needs parameter optimization. At least a sanity check is always needed.
+ Semi- or unsupervised algorithms need substantially less training data, ...	− … but if the training data is available, supervised learning is a better choice.

Feel free to contact us





	@jpolz3	@lschm1dt	@glawion_l	@Max_Grave_	@cwerner76
GitHub	jpolz	schmidtlennart	LGlawion	maxmargraf	cwerner

julius.polz@kit.edu