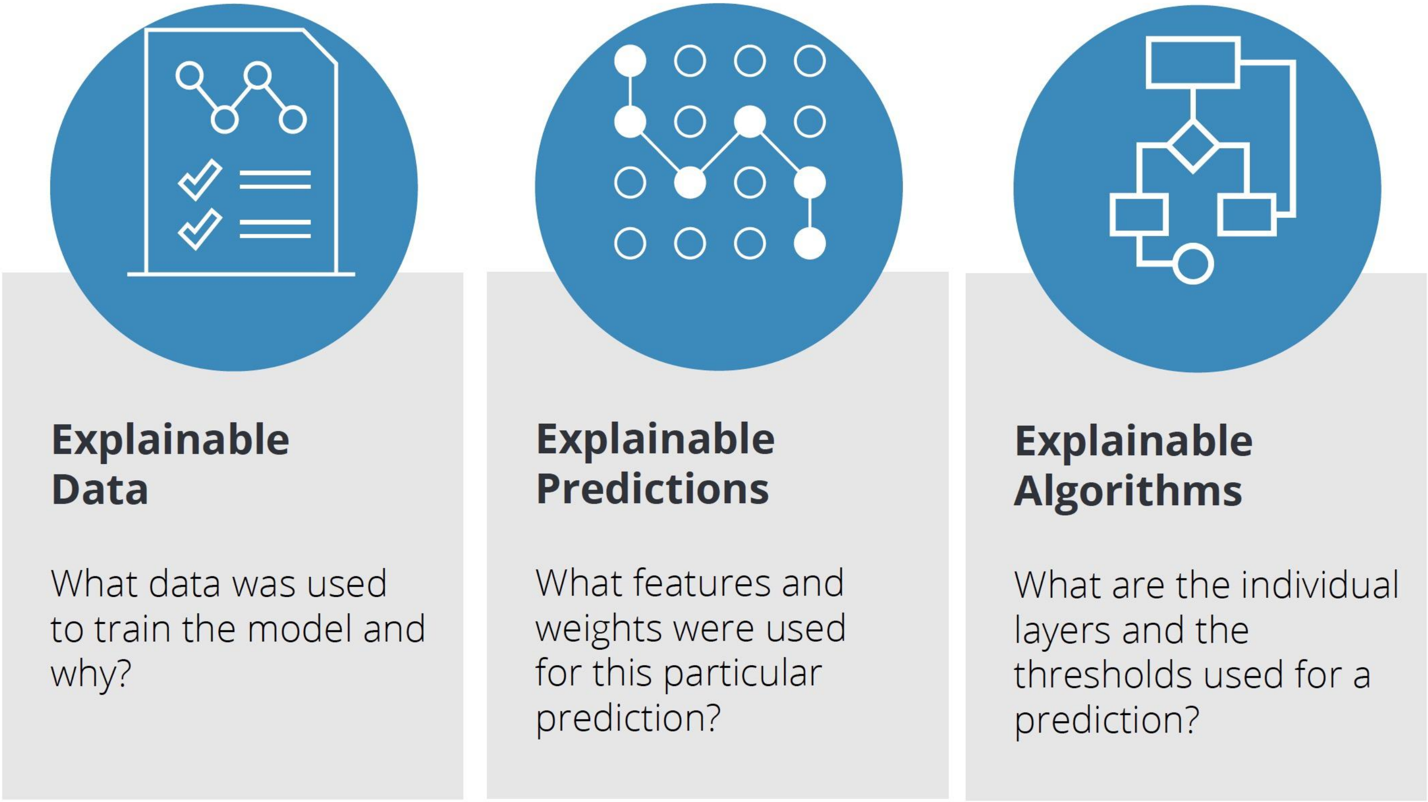
… STEFAN BOSSE[1,2*]

[1] Institute of Computer Science
Researchgroup Practical Computer Science
University of Koblenz

[2] Department of Mechanical Engineering
Lehrstuhl für Materialkunde und Werkstoffprüfung
University of Siegen



WHY TRACEABILITY AND EXPLAINABILITY OF AI/ML MODELS IN MEASUREMENT AND TESTING TECHNOLOGY ARE MORE IMPORTANT THAN ACCURACY AND PRECISION

Stefan Bosse

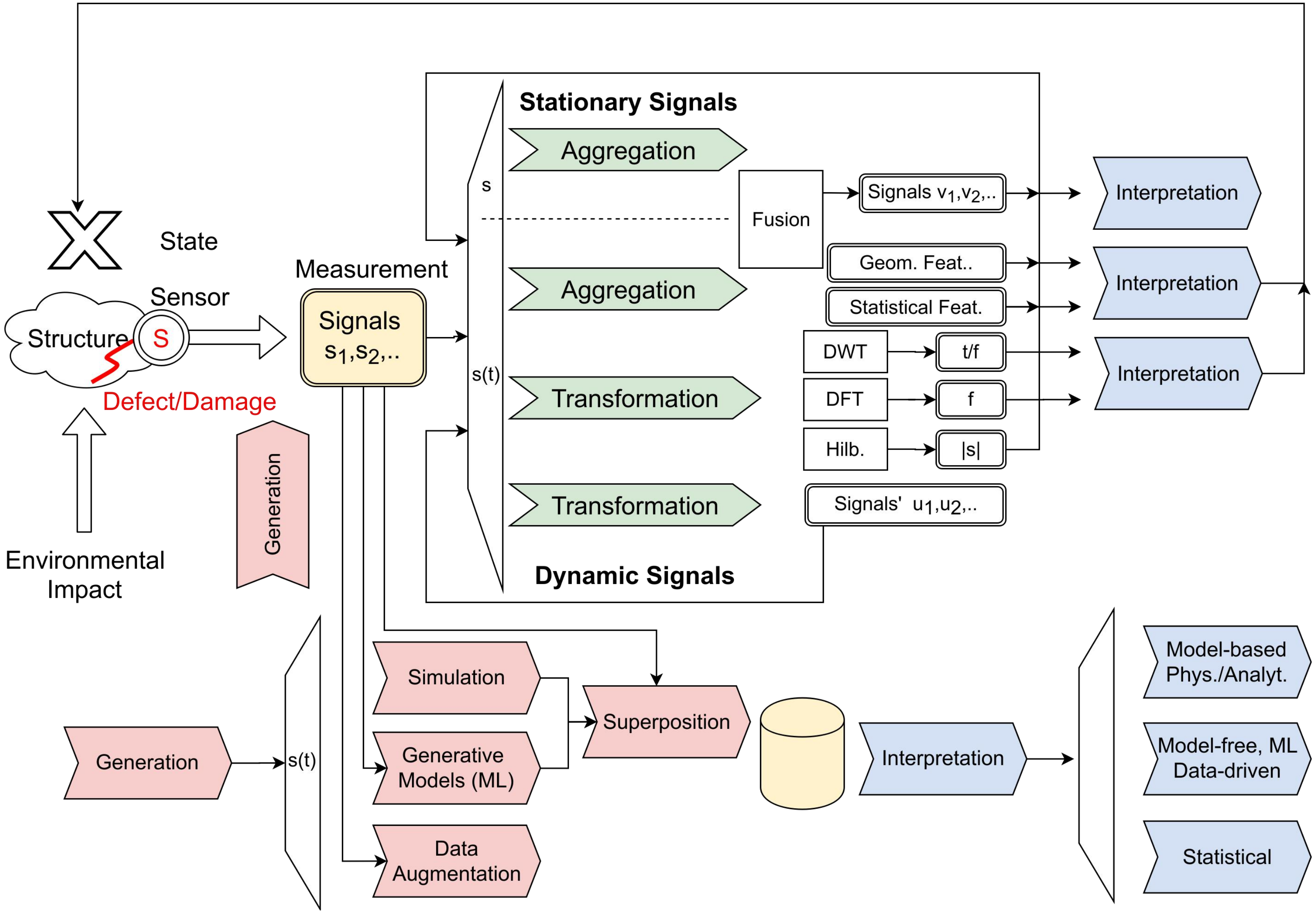
LMW Lehrstuhl für Materialkunde und Werkstoffprüfung

university of koblenz Computer Science

CONTENT

BASICS

- **Data-driven Methods for**
  - Structural Health Monitoring
  - Damage Diagnostics
  - Material Testing
- **Data-driven Methods in**
  - Measurement - Sensors
  - Signal Processing
  - Aggregation
  - Transformation
  - Fusion
  - **Interpretation (Prediction)**
  - **Generation**



university of koblenz
Computer Science

BASICS

Accuracy versa Precision

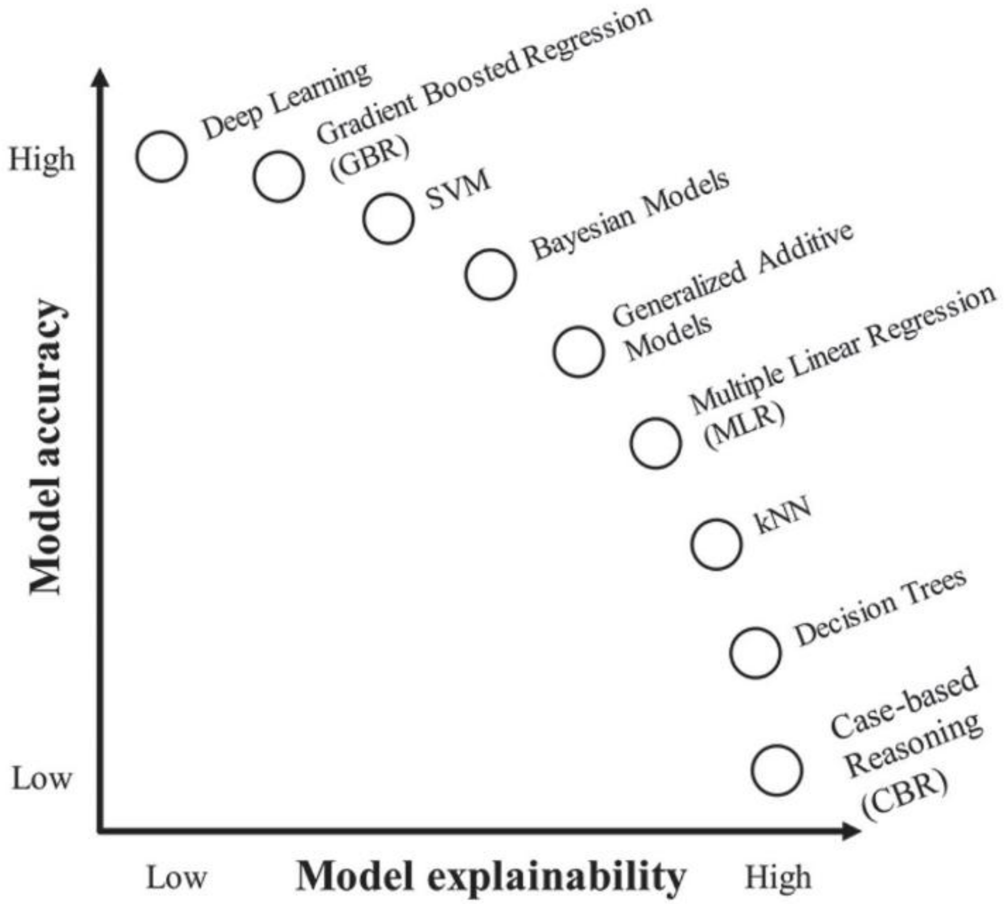
- Strong. Accuracy: Close to expectation value (ground truth)
- Weak. Precision: Low variance

Explainability versa Traceability or Tracktability

- Strong. Explainability: An inductive model relationship between x and y based on knowledge
- Weak. Traceability: Which input contrbutes to output?

Generalization: Specific or more general?

- Interpolation versa Extrapolation



We have an opposite relationship between model accuracy and model explainability! And what‘s about model generalization?

university of koblenz
Computer Science

Hammann, D., & Wouters, M. (2025). Explainability Versus Accuracy of Machine Learning Models: The Role of Task Uncertainty and Need for Interaction with the Machine Learning Model. European Accounting Review, 1–34. https://doi.org/10.1080/09638180.2025.2463961

BASICS

Interpolation versa Extrapolation: Prediction Errors?

Models

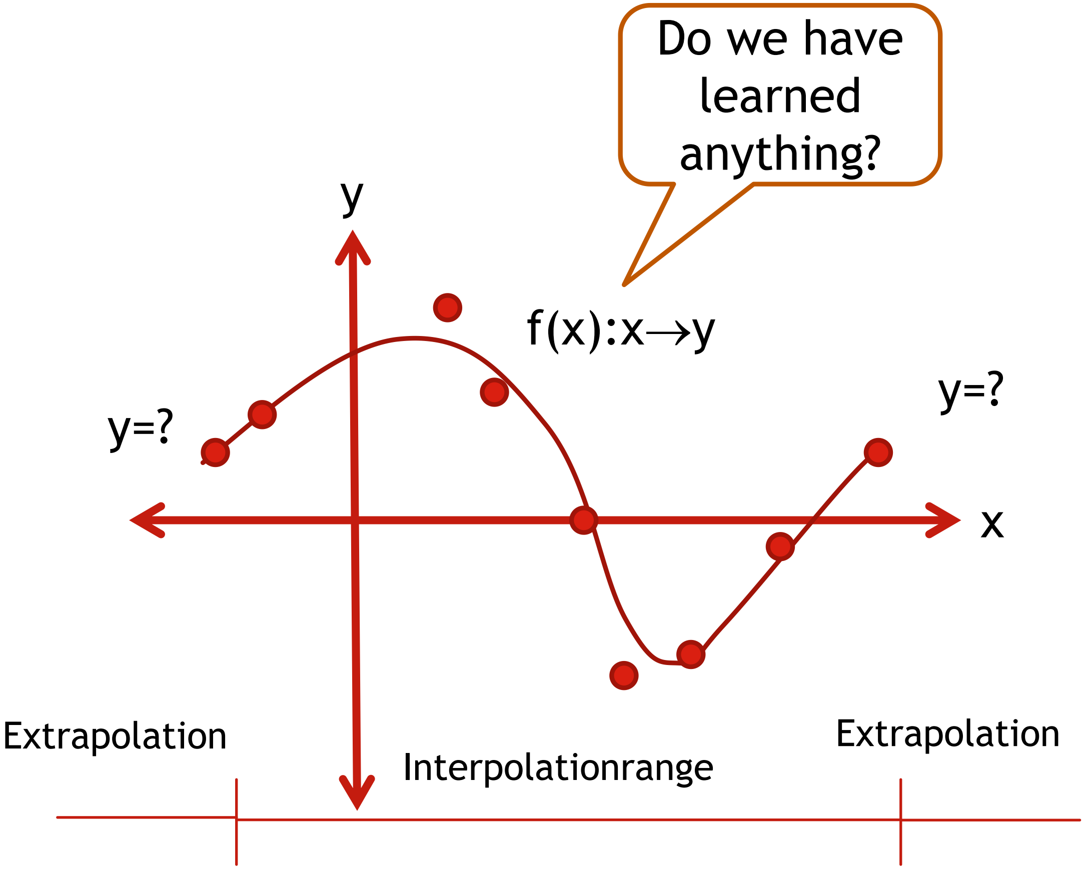
- Trees: I/E NA (partly by regression trees)
- Functions and Functional Graphs (SVM, ANN, CNN, ..)
- Training = Definition from Data ↔ Def. Parameter Space

Interpolation

- Input values between training points but inside parameter space

Extrapolation

- Input (and output) values outside the parameter space



Interpolation should be possible with all functional models, extrapolation fails commonly! Generalization provides interpolation (degree 1) and extrapolation (degree 2). A specialized model will provide neither nor.

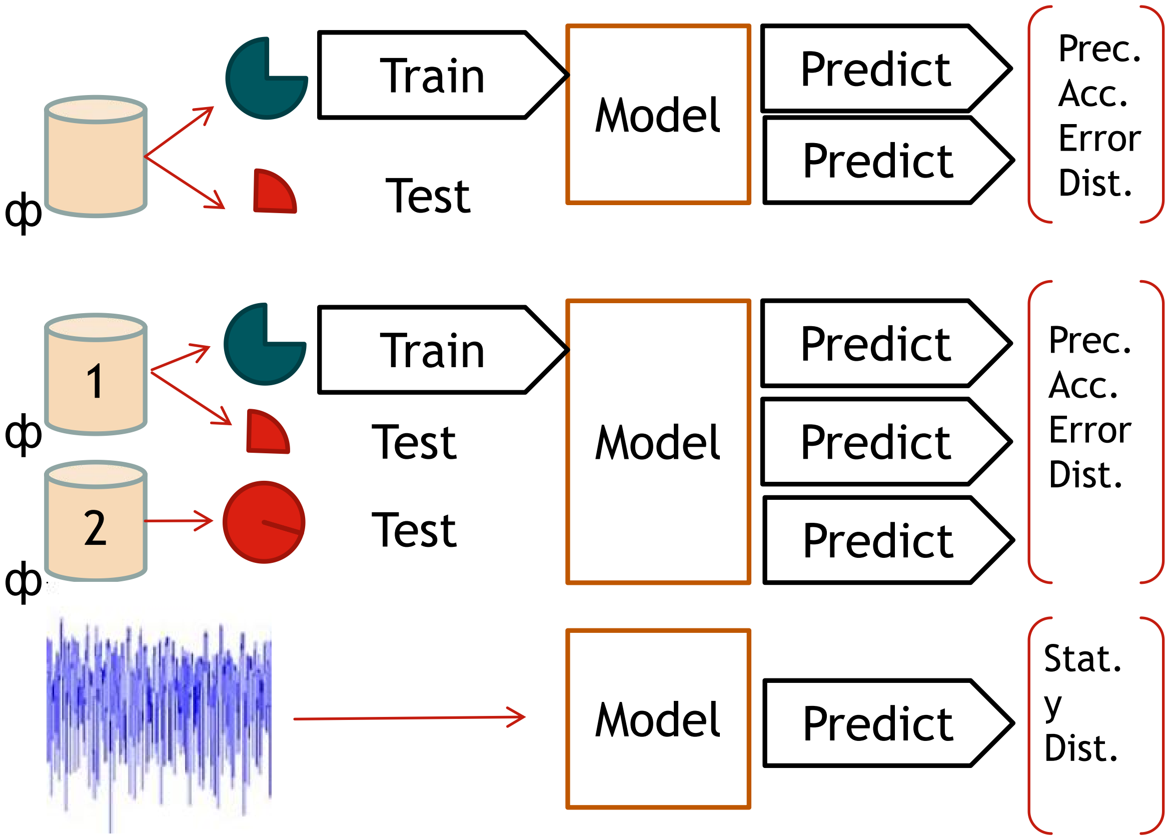
BASICS

Monitoring / Observation

- Interpolation: Test with sample set not included in training (classical validation) but same parameters ф
- Extrapolation: Different sample distributions ф for test and training data
- Training Data versa Test Data: Accuracy, Precision, ...

Robustness

- What is the output if the input is ouside of the definition parameter space or non-sense data?
- Noise sensitivity?

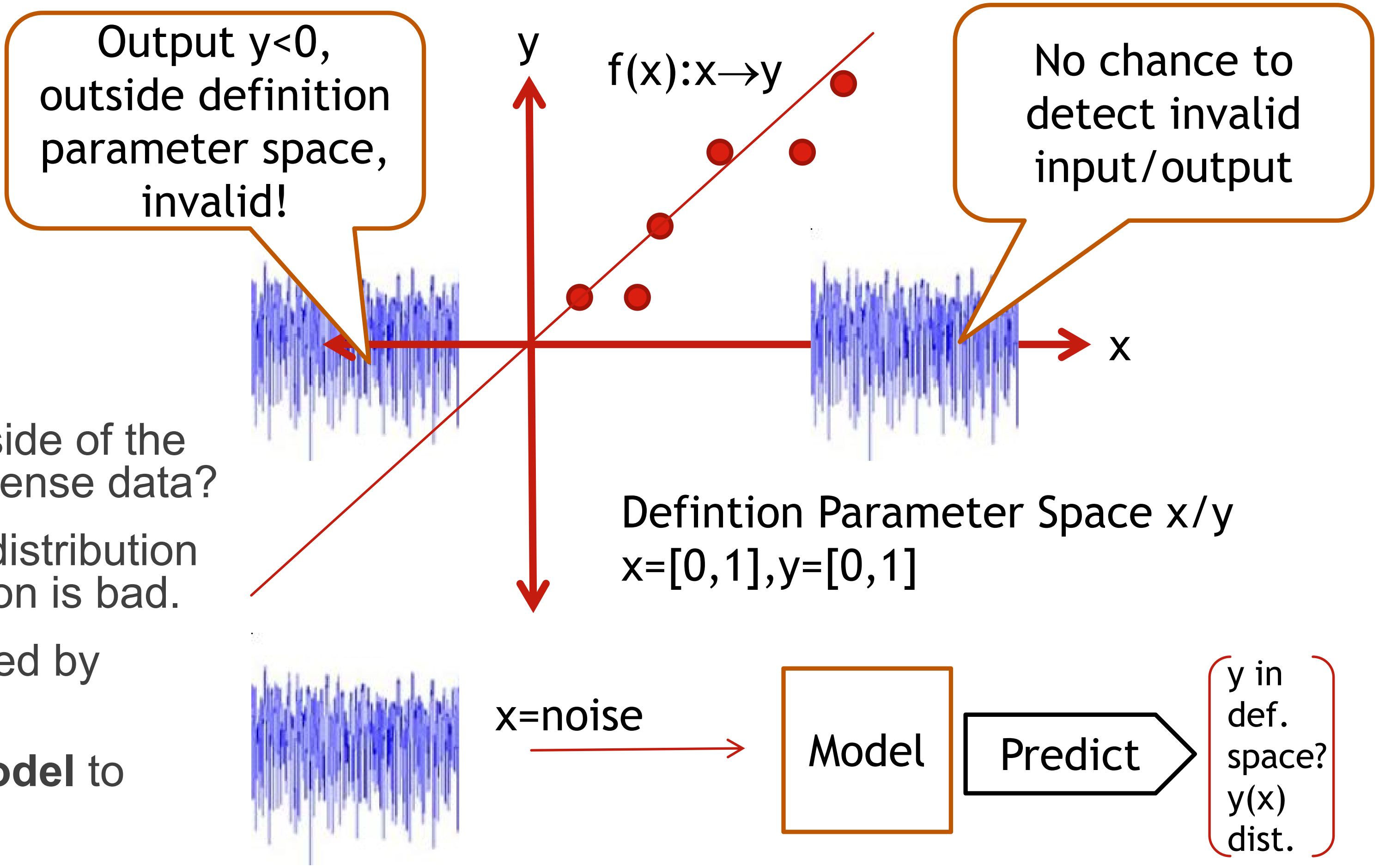


Training error low, test error high: Specialized Model ▶ Nice to have, but useless!
Training and test error low: Interpolating Model ▶ Degree 1 of Generalization
Different sample distributions covering different parameter spaces:
Extrapolation Test, low error outside training distribution ▶ Degree 2 of Generalization

BASICS

Robustness

- What is the output if the input is ouside of the definition parameter space or non-sense data?
- Noise sensitivity? Make a y(noise) distribution test - high variant y(noise) distribution is bad.
- Can we detect **invalid output** caused by **invalid input**?
- Do we need a separate **scoring model** to classify input data validity? Or the „noise“ class as additional output?



With simple models we can maybe recognize and detect invalid input and invalid output. With any complex, highly non-linear, multi-variate and deep (nested) models this is mostly impossible!

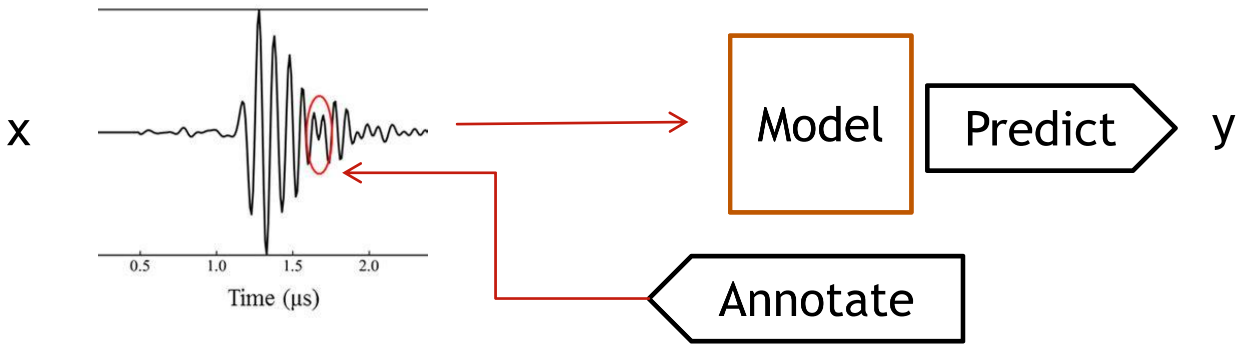
BASICS

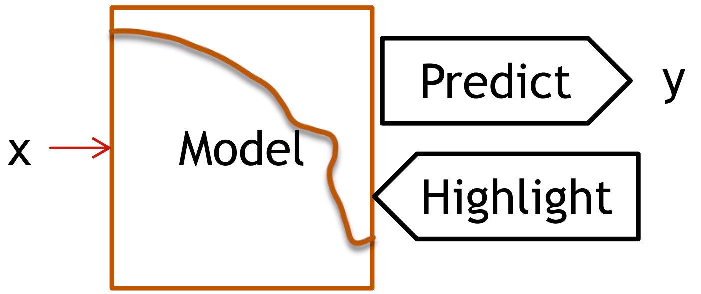
Traceability

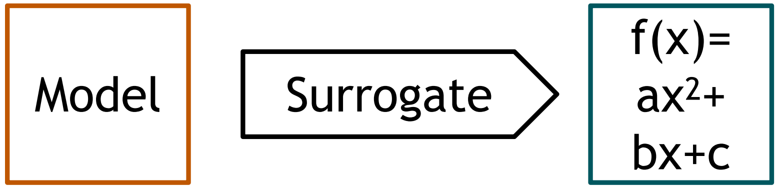
- Which part of input is relevant?
- Which functional terms or paths in graphs are contributing to output (or decision making)?

Explainibility

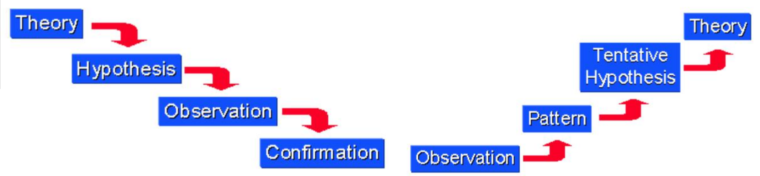
- Anything learned? Generalization towards a physical law?
- Analytical explanations $x \rightarrow y$?
- Surrogate Modeling
- Explainable Models (e.g. XANN)







The most important methodology: Explain the model behavior or extraxt knowledge from a data-driven model. Induction is better than Deduction.

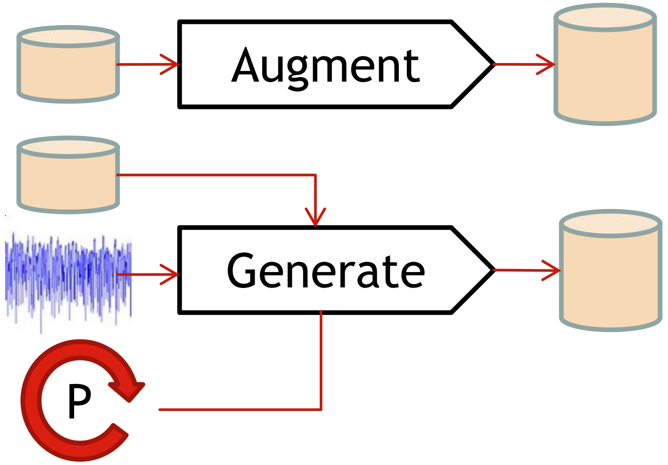


PREDICTION VERSA GENERATION

Signal Data Generation: Output is Signal

- Model-driven or random Augmentation from real measured data (linear independent data?), additive and multiplicative noise, super-position (real-synthetic)
- Model-based or Model-driven Simulation Methods (Physics-drivem but Reality gap!)
- Model-free Random-process Generative Models (data-driven), e.g. Generative Adversarial Networks or Variational Autoencode Models
- Model-free (?) or Model-driven Parameterizable Generative Models (data-driven)

Motivation: Lack of parameter variance and sparse paramater space coverage of experimental data.



The most important quesion of signal data generation: Produces the generator what we want - is it physically correct, and is the parameter space broadly covered?

university of koblenz
Computer Science

EXAMPLE 1: USELESS DATA-DRIVEN MODEL

Ultrasonic High-frequncy Pulse-Echo Measurements for Porosity Detection in Die-casted Aluminum Plates

Input

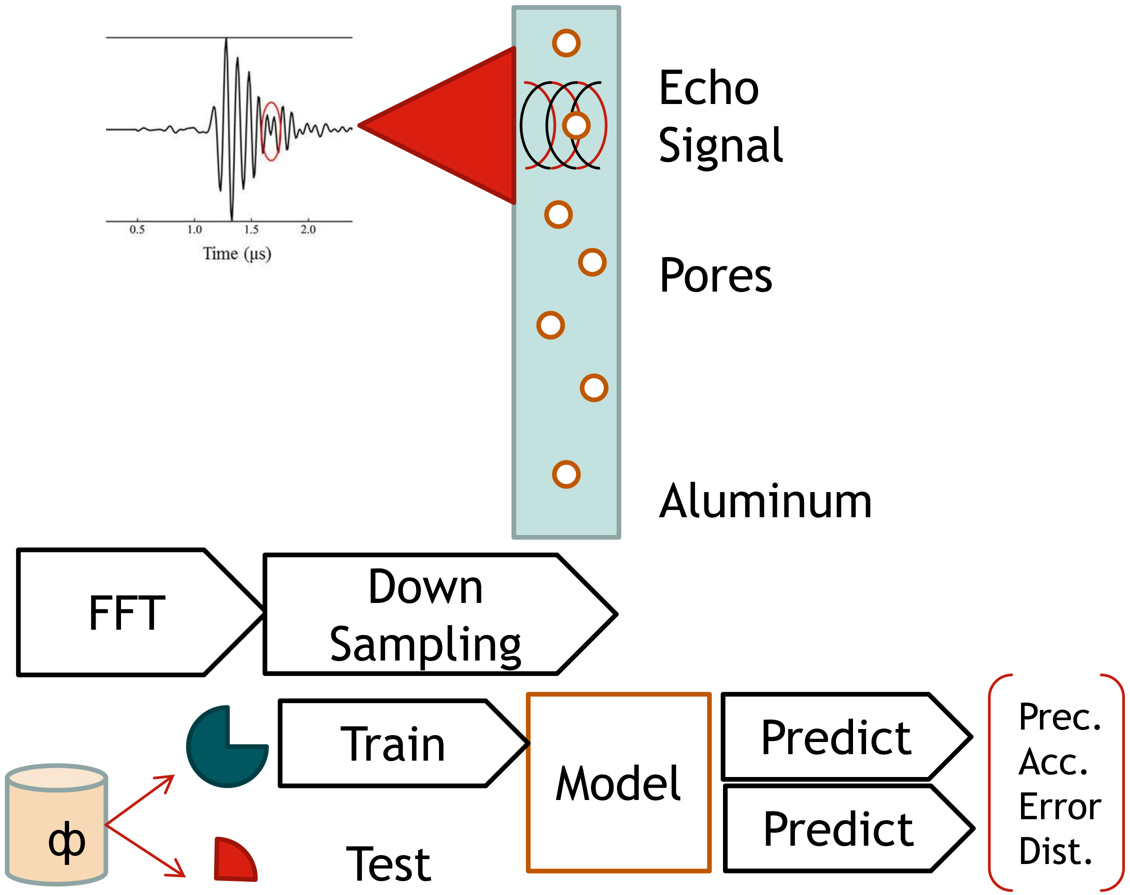
- Time-resolved US signal (stimuls: bipolar pulse, broad frequency spectrum) , 50 specimens, 3 measuring locations
- Transducer: Dual-piezo-crystal, 5 MHz, 10 mm Dia.
- Material: Aluminum alloys (primary, secodary 58/89%)

Features

- Frequency spectrum of response signal (FFT)
- Assumption: Attenuation is frequency dependent and frequency spectrum is depending on pores (size, density)

Output

- Pore grade classification (A,B,C,D)
- Simple ANN (two layers [10,5], sigmoid activation function softmax output layer)

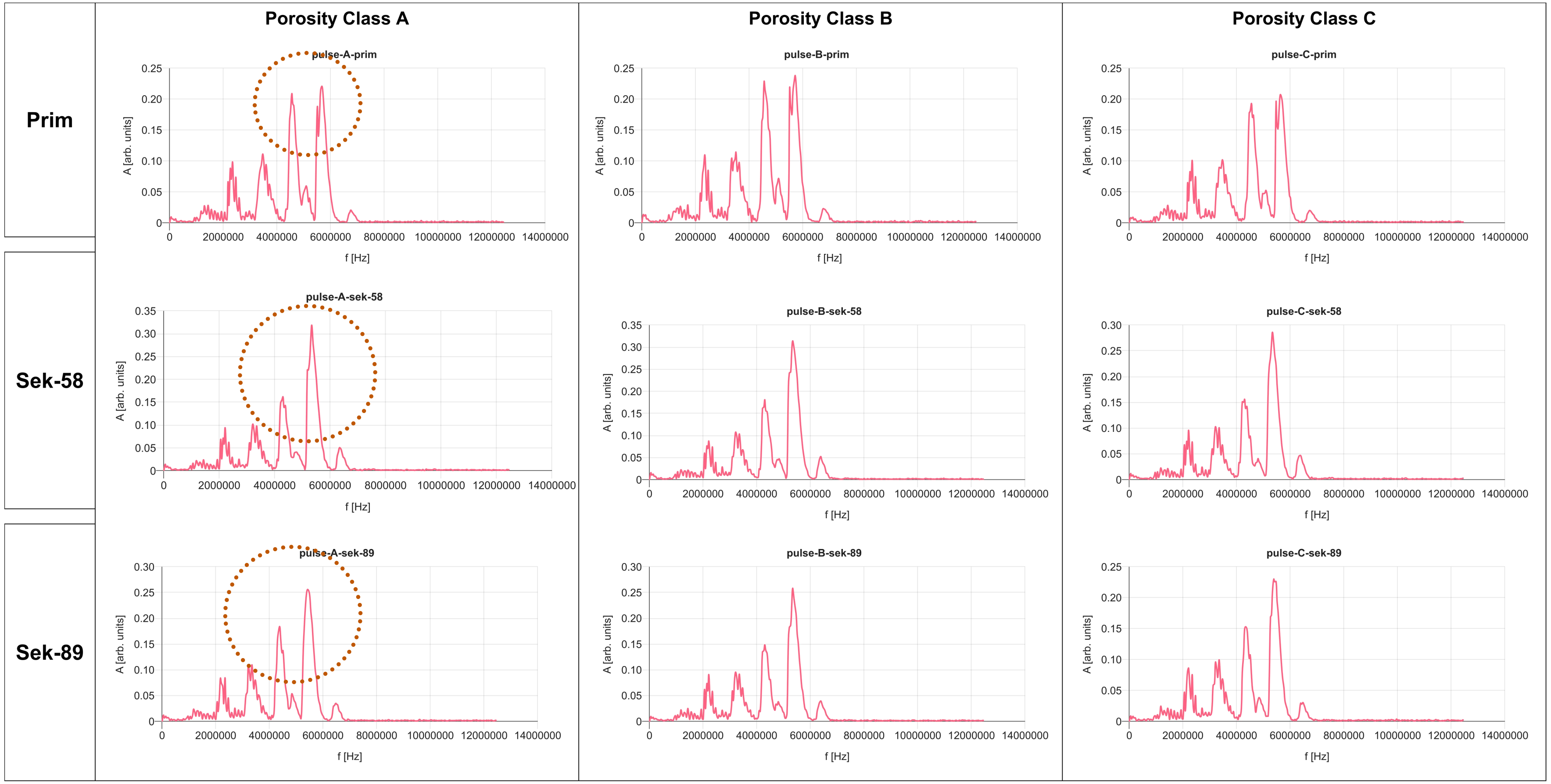


EXAMPLE 1: USELESS DATA-DRIVEN MODEL

Ultrasonic High-frequncy Pulse-Echo Measurements for Porosity Detection in Die-casted Aluminum Plates

No porosity class correlation visible. Hidden?

Classification of alloy class is directly visible! Strong feature correlation



EXAMPLE 1: USELESS DATA-DRIVEN MODEL

Ultrasonic High-frequncy Pulse-Echo Measurements for Porosity Detection in Die-casted Aluminum Plates

Data

- φ: 50 specimen x 3 positions x 5 augmentation (multiplciative normal distrubuted noise 10%), mixed 3 alloys
- Training/Test split: 80/20% (random)

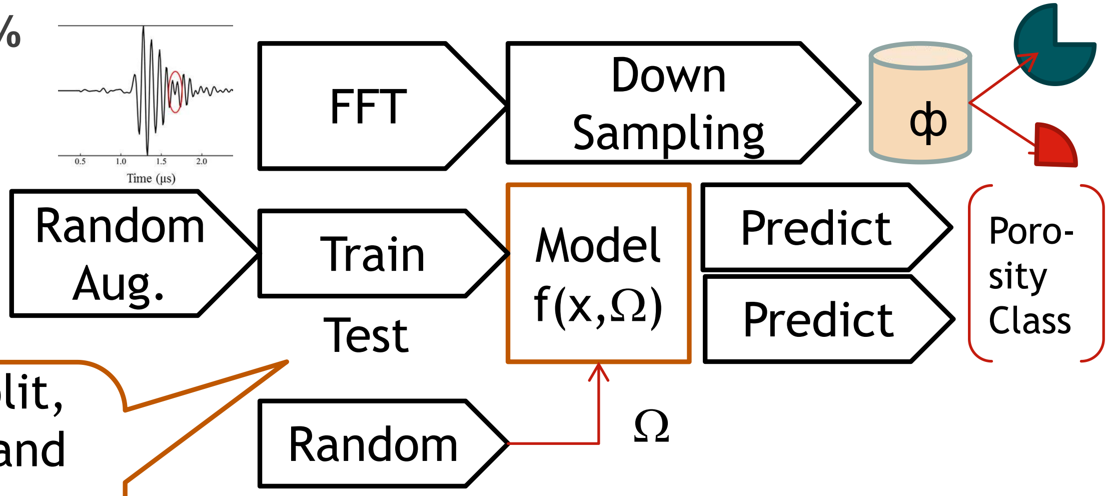
Results

- Training of ANN (adam optimizer) with augmented data: Smooth and convergent!
- Classification error: **Training Data=0%..3% (!), Test=40%±20%**

Explainibility

- Large Training/Test error ratio: **Specialized Model**! No I/E
- Very small/weak input features were amplified resulting in a practical unusable, unpredictable and instable model
- **Unknown functional x-y relation**

Interpolation (I)/ Extrapolation (E) Tests not possible



Random data split, augmentation, and model training was repeared N times (Statistiscs)

Computer Science

EXAMPLE 2: GENERATIVE MODEL

Random-process and Data-driven Generative Model for Guided Ultrasonic Wave Data

Signals

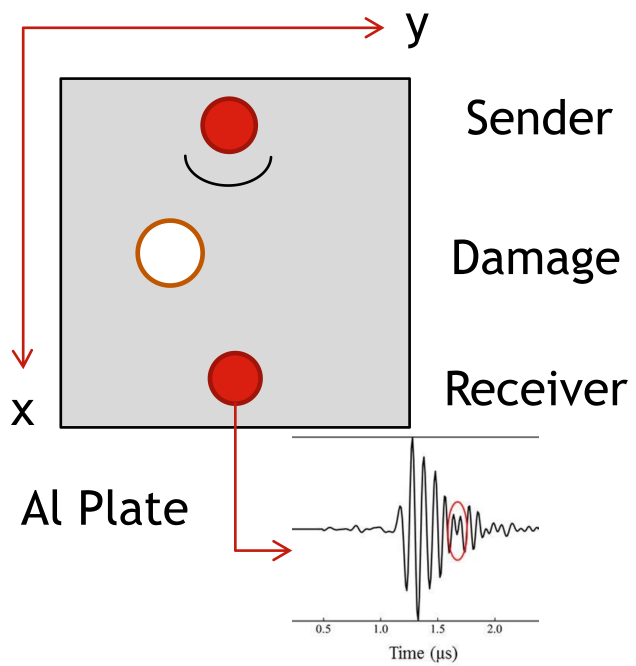
- Time-resolved US signal (stimuls: gaussian-masked sind pulse, narrow frequency spectrum)
- Assumed measuring set-up: On sender transduce, one receiver transducer, single and straight path
- Material: Solid (e.g., Aluminum), Damage: Air, e.g. a circular hole

Simulation and Ground-truth (GT) Data

- Parameter space is limited - simulation using visco-elastic wave equiation is used
- A large set of data can be generated with exact labelling (GT)

Generative Model

- Generative Adversarial Network (GAN)
- Input: Random vector, Output: GUW signal, Training: GUW signals from simulation



$$\phi=\{f_{wave},dim.,pos_{dam},size_{dam},T,..\}$$

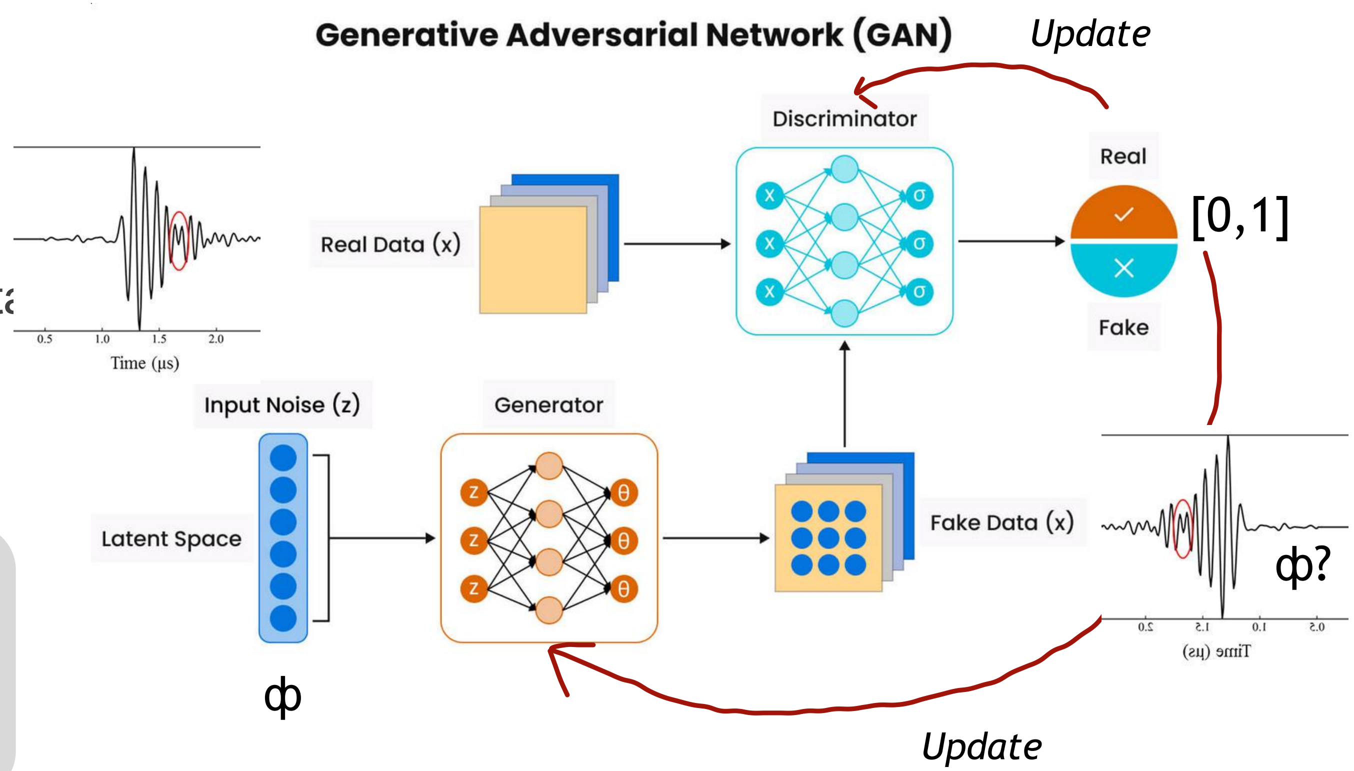
EXAMPLE 2: GENERATIVE MODEL

Random-process and Data-driven Generative Model for Guided Ultrasonic Wave Data

Generative Adversarial Network Model

- Generator
- Discriminator (only Training)
- The generator nevers sees the orginal data
- Feedback only from discriminator which predicts a fake score [0,1]

How can we evaluate the signal generator quality?
Physical correctness? Covered Generator parameter space?



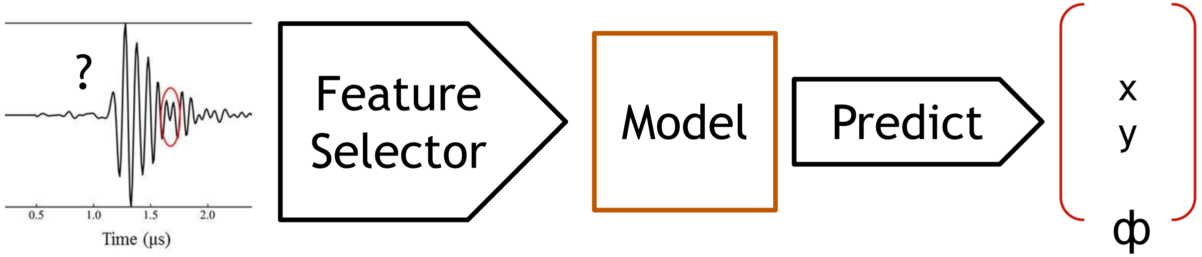
Computer Science

EXAMPLE 2: PARAMETER PREDICTOR MODEL

Random-process and Data-driven Generative Model for Guided Ultrasonic Wave Data

Parameter Predictor Model

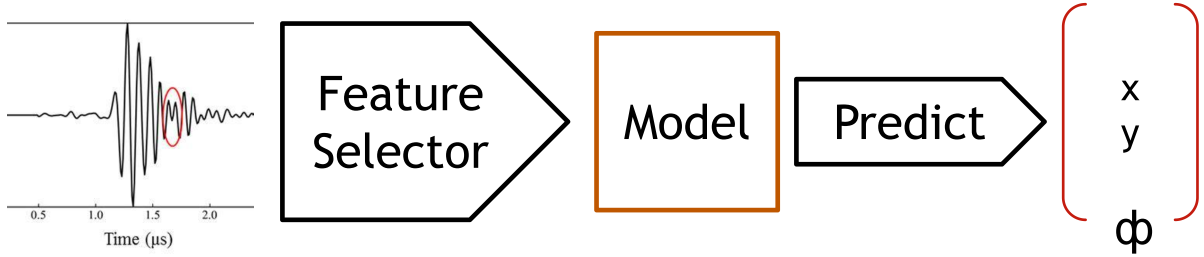
- Input: GUW signal (real/simu./synth.)
- Output: Parameter vector
- Here: Damage position $\mathbf{p}_{dam}$=(x,y), trained with simulation data (GT)
- Interpolation and Extrapolation required

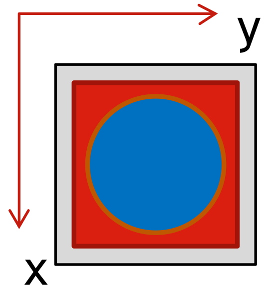


Use another data-driven predictor model for the data parameter space! Good idea? How good is the predictor? Can we trust the model?

EXAMPLE 2: PARAMETER PREDICTOR MODEL

Random-process and Data-driven Generative Model for Guided Ultrasonic Wave Data

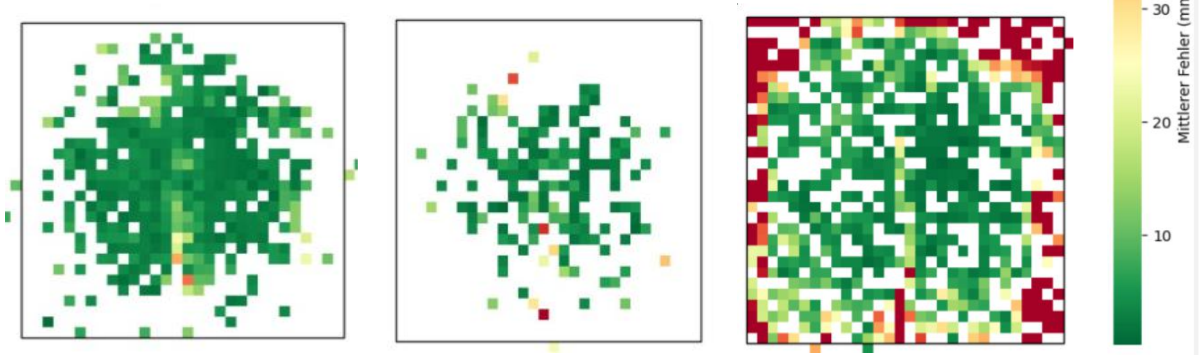




Parameter Predictor Model: IE Test

- Two sample sets (ϕ is damage location): Normal (N) and uniform (U) random distribution of damage location (x,y)
- Training with normal distributed data (80%), Test with uniform distributed data (100%)

It seems the model is I and partly E capable, at least we can hope!



[Sidar Kilinc, 2025]

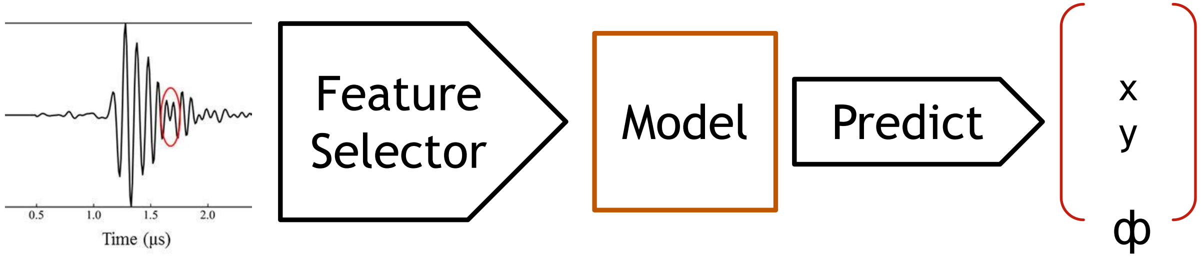
EXAMPLE 2: PARAMETER PREDICTOR MODEL

Random-process and Data-driven Generative Model for Guided Ultrasonic Wave Data

Parameter Predictor Model: IO Analysis

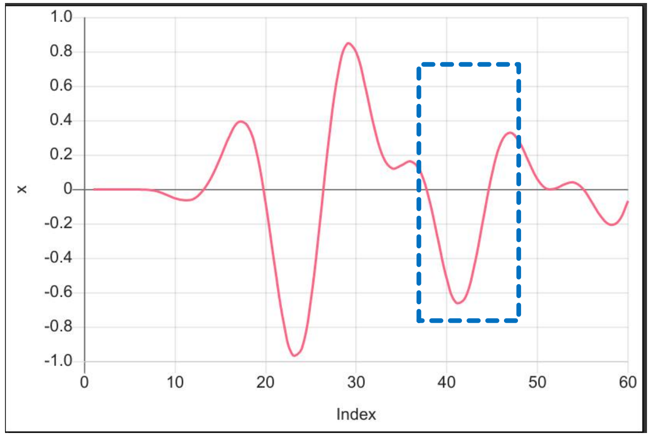
- Identify major input elements / regions contributing to the output (strong feature analysis), Input: signal *s*, Output: location coordinates (*x*,*y*)
- Gradient is a measure of activation:
  $\frac{\partial x}{\partial s_i} \approx \frac{\Delta x}{\Delta s_i}, \frac{\partial y}{\partial s_i} \approx \frac{\Delta y}{\Delta s_i}$

If the prediction / regression error is low then stimulus signal regions can be identified. If the error is high, no clear correlation is visible!

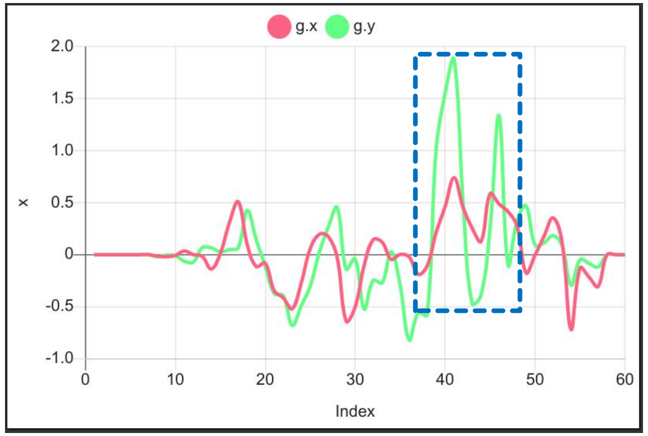


Ground Truth	x=352	y=307
Predicted	x=226 (-25%)	y=308 (0.3%)

(CNN Model)



Signal

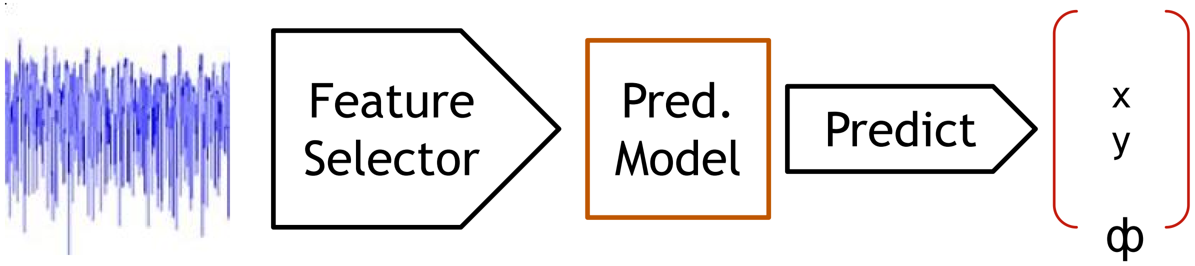


Gradients

university of koblenz
Computer Science

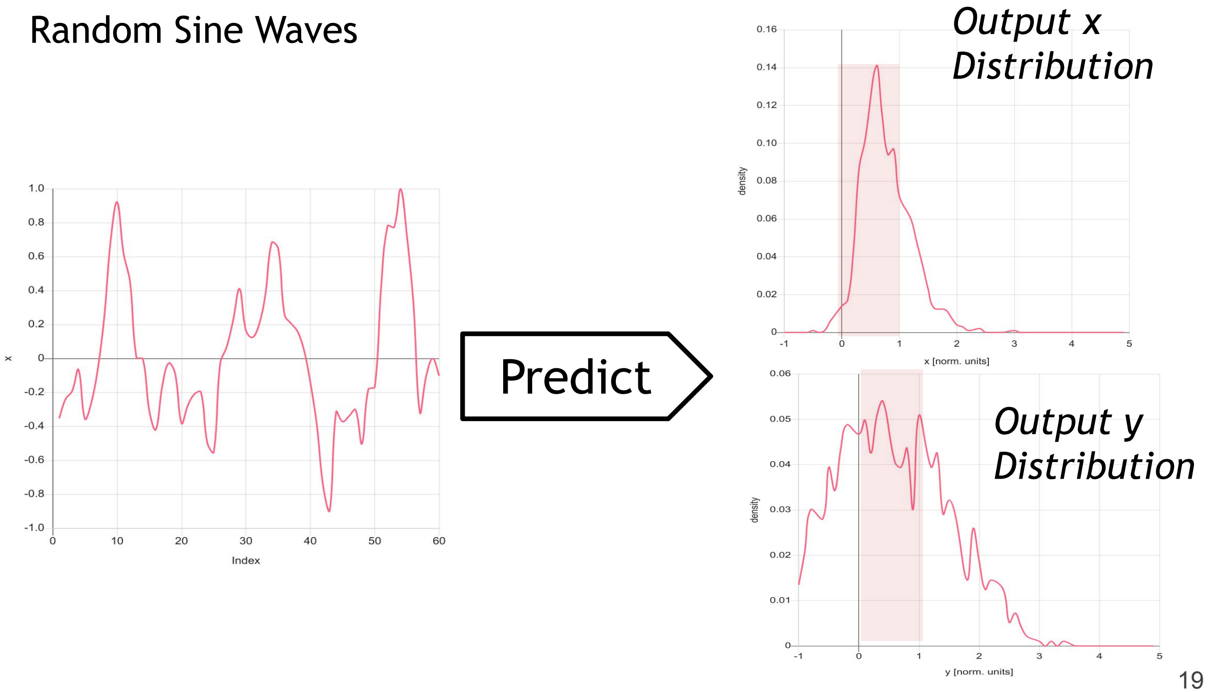
EXAMPLE 2: SCORER MODEL

Random-process and Data-driven Generative Model for Guided Ultrasonic Wave Data



Parameter Predictor Model: Noise Test

- Invalid and noise data test: Feed model with pure random noise and random sine waves.
- Train an additional scorer model that tests the input signal data for validity.



> **What happens if the predictor gets noise or random data? Make a test... The predictor model outputs a broad parameter range..**

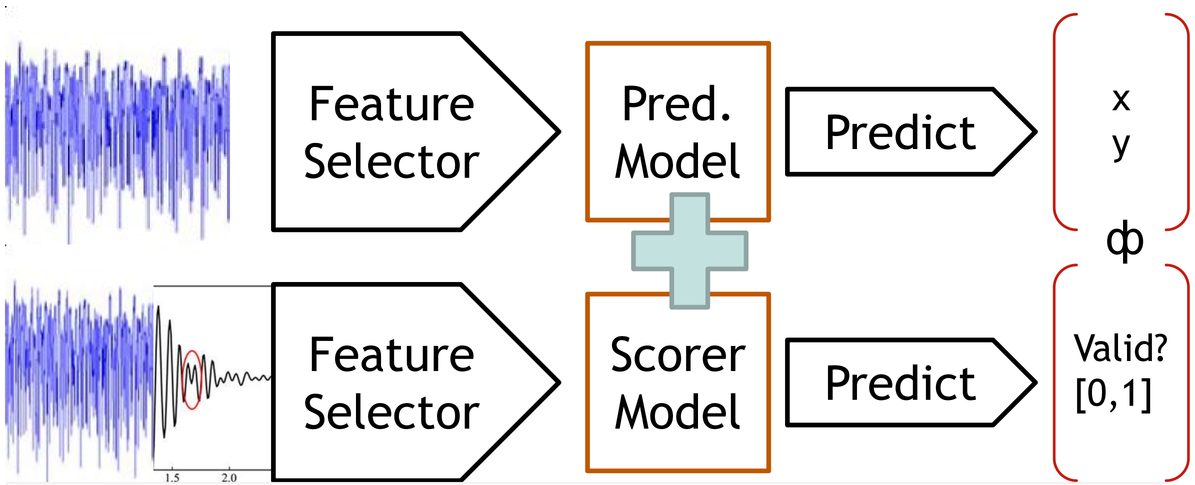
EXAMPLE 2: SCORER MODEL

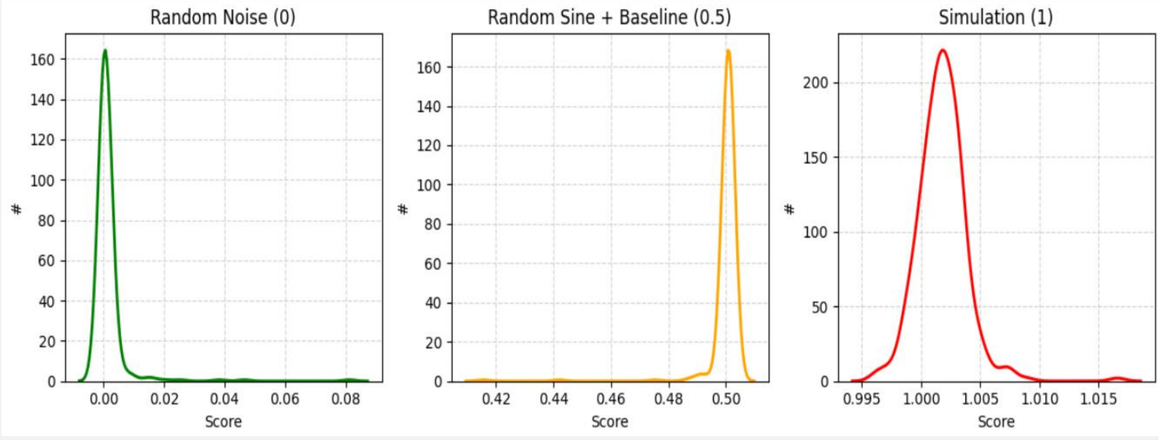
Random-process and Data-driven Generative Model for Guided Ultrasonic Wave Data

Parameter Predictor Model: Noise Test

- Invalid and noise data test: Feed model with pure random noise and random sine waves.
- Train an additional scorer model that tests the input signal data for validity.
- Hierarchical model: 1. Scorer 2. Predictor

What happens if the predictor gets noise or random data? Make a test... The predictor model outputs a broad parameter range. A scorer is required.





Noise	Random Sine Waves	GUW
0	0.3 /0.5 (+ GUW)	1.0

CONCLUSIONS

Generalization

- Degree 1: Model can be used for Interpolation w/o large deviations within defined parameter Space
- Degree 2: Model can be used outside (trained) parameter space w/o large deviation
- Degree 3: The model is stable against noise and invalid data
- If we want to test and evaluate generative models (random process) we need degree 3!

Traceability

- Effect of input on model activation (paths inside a model) / Activation paths
- Which part or region of model input is relevant for a specific output?
- Model behavior with invalid or pure noise data, model selection

Explainability

- Do we have learned something from the data-driven model? Induction versa deduction?
- Why is the model giving a specific output for a specific input?
- Surrogate models can help to reduce a complex model to simplified and explainable well known functional laws
- Can we solve the inverse problem with a specific model?
- Can we explain generative models?

THANK YOU

Stefan Bosse

sbosse@uni-koblenz.de

www.edu-9.de

university of koblenz
Computer Science

Praktische Informatik