Time Series based Electricity Price Forecasting with Artificial Neural Networks: Methodology, Benchmarking and Coding

How important decisions about the approach, IT architecture, tools and methods can be made in order to successfully carry out a data science project

Illustrate the interaction of data, methods and tools for analysis and visualization

Dr. Michael Thrun

What the talk is going to be about…

- Only a „bird‘s-eye view“ is possible in the given time
  - Requires several simplifications
  - All required code will be shown but remains largely unexplained

  -> Please look into the documented source code as „homework“

Learning targets

1. Success factors for data science are understood
2. Challenges in forecasting are known based on ML theory
3. Benchmarking in Forecasting can be performed (on the example of electricity price forecasting)

$\Rightarrow$You will be able to spot the pitfalls of my live coding in forecasting
(Using the concepts presented)
(Without the necessity to understand the details of the code)

What is the most important Skills for a Data Scientist?

- Programming (e.g. Python/R)?
- Mathematics or Statistics or Machine Learning?
- Technology Stack?
  - Acquaintance with Data Science Tools (e.g. Tableau, Pandas/Scikit-learn, Databricks)?
  - Big Data Technologies (e.g Spark)?
- Data Wrangling?
- Soft-Skills? (e.g. Communication skills, social intelligence, people skills)?
- Software-Engineering

Relevance of Skills in Data Science

- **Communication, especially with "domain expert"**
- Intuition about Data (Experience)
- Data Wrangling
- Statistics
- Machine Learning
- Programming
- Visualization
- Software Engineering
- Big Data

- 80 % of work is preprocessing:
  - Data wrangling (i.e. generating structured data set
  - Redefining the objective until it fits the request
  - Then, Restarting preprocessing after talking to the domain expert
  - More preprocessing ☺
- 15% is applying statistics and machine learning
- 5% finding understandable visualization/Presenting the results

Perspectives of a Data Science Project

How you'll feel as Data Scientist

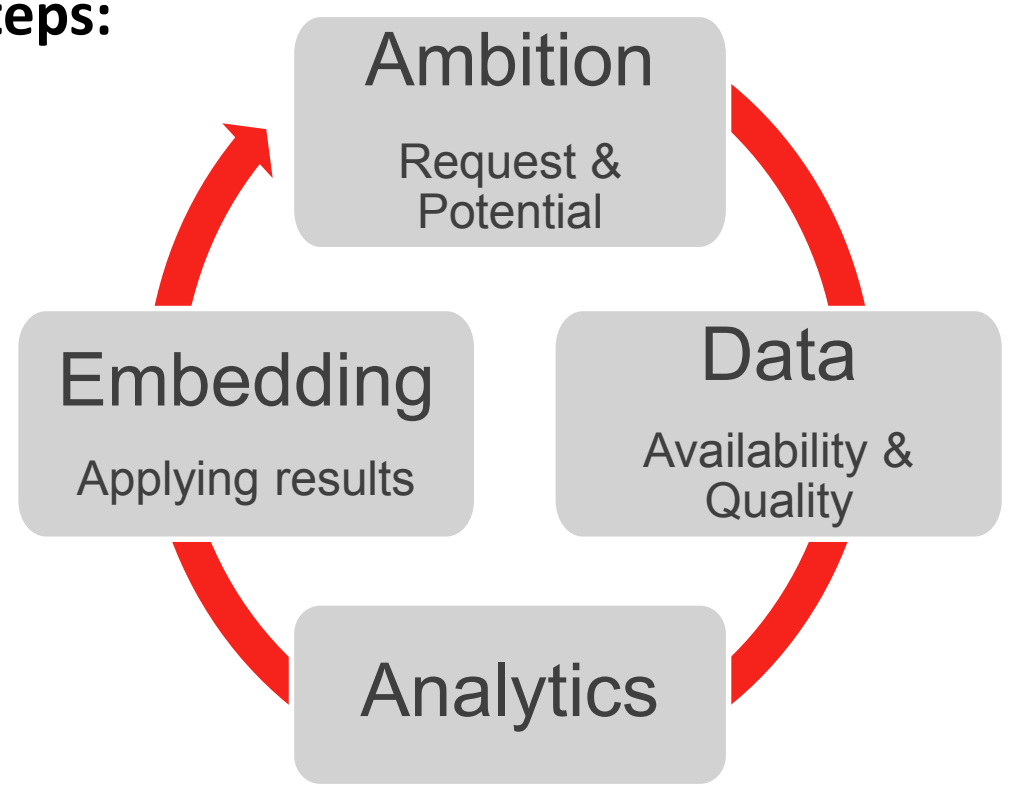
Source: systemkameraforum.de

⇒ 300+ PS on a dirt road

Overall objective:

- Generate new turnover with smart services
- Lower costs with automation and insights
- Support Decisions through advanced analytics/ML/AI

Process Steps:



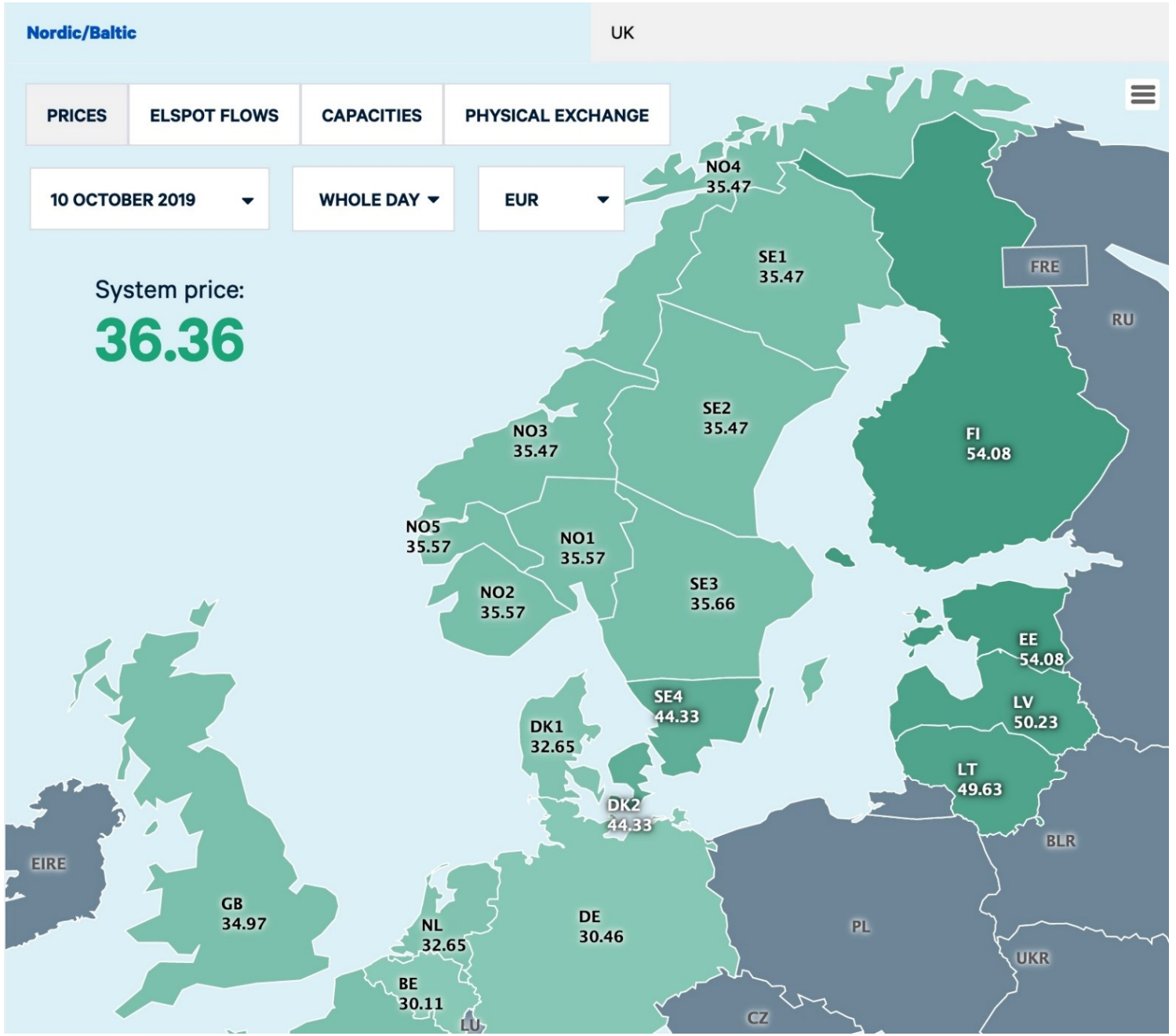
Roles in Data Science

- Solution Architect and Domain Expert
  - Deep knowledge in the data sources of the company/research field
  - Contact points for analytic requests
  - Translator of domain specific issues to data science objectives **(interdisciplinary)**
- Data Scientist
  - Translator of to data science to domain specific objectives **(interdisciplinary)**
  - Deep knowledge in Machine Learning and Statistics
  - Performs proof-of-concepts and builds prototypes
    - Develops models for insights and automated decisions
- Data Engineer
  - Deep knowledge in IT architecture
  - Extracts Data from Source (e.g. IoT, database,…)
  - Embedding, i.e. Builds the minimal viable product from the prototype

Dataset

- https://www.nordpoolgroup.com/historical-market-data/
- 7 years of data in hourly resolution
- Regions SE1;SE2;SE3;SE4;FI;DK1;DK2;Oslo;Kr.sand;Bergen;Molde;Tr.heim;Tromse;EE;LV;LT
- Sys: system price for each hour is based on the intersection of the aggregated supply and demand curves representing all bids and offers in the entire Nordic market

  -> Target feature



Typical Starting Point

- Communication with Domain expert is the key for success!

Typical Starting Point

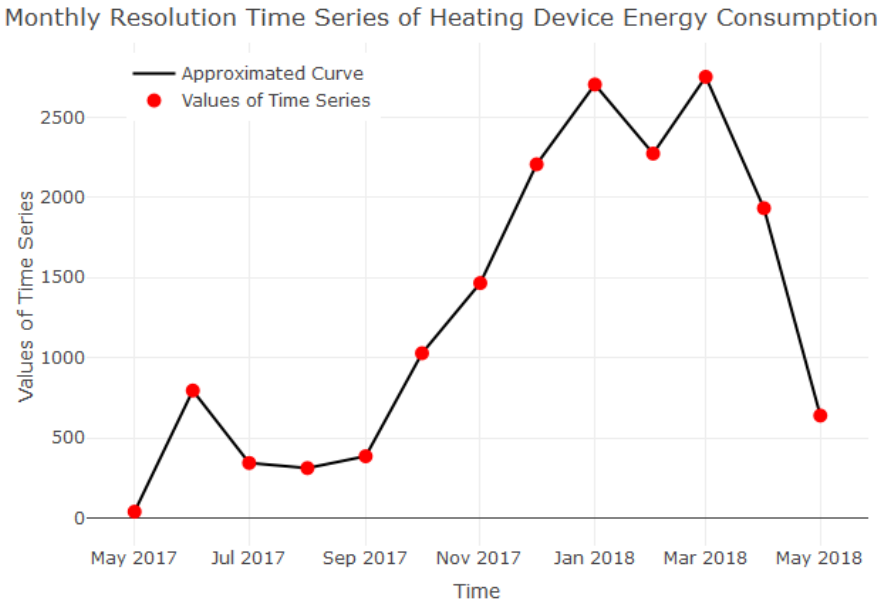
- Communication with Domain expert is the key for success!

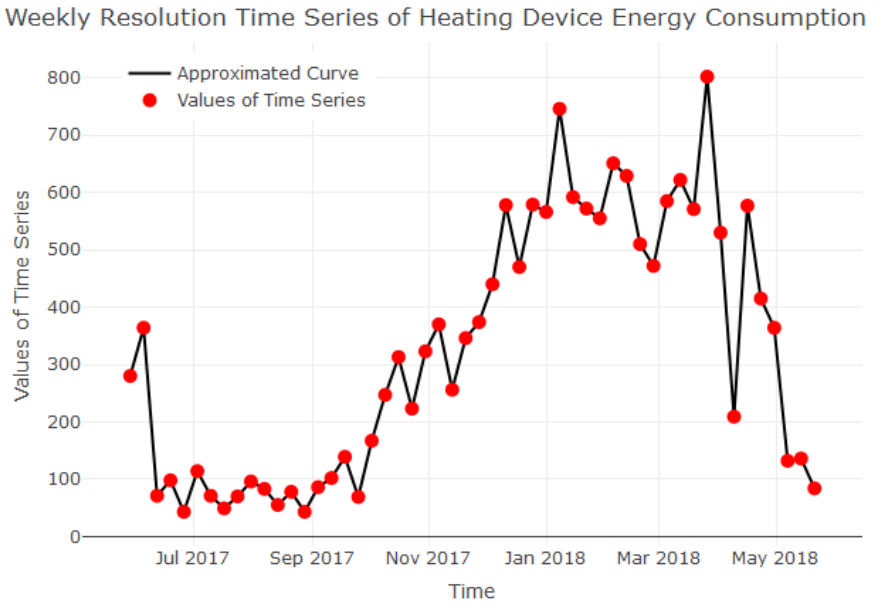
Always assume:

-> On the example of Electricity Price Forecasting I will outline why

Recapitulation: Basic Definitions

- A structured data set is generated/available
  - Columns =Variables/characteristics/properties/features
  - Lines: Cases, e.g. measured points
  - Key: Unique assignment, e.g. time with a resolution (e.g. Days, Weeks, Months...)
- One feature Y_t shall be forecasted
  - Other features can be used as predictors
    - If and only if Y_t depends on them!
- Y_t is equidistant and a series of measurements visualized as curve with a specific resolution





Recapitulation: Forecasting

Given Y_f, an h step forecasts is a prediction of the values of $Y_{f+1}, \ldots, Y_{f+h}$ if it is based only on the information available at time t=f

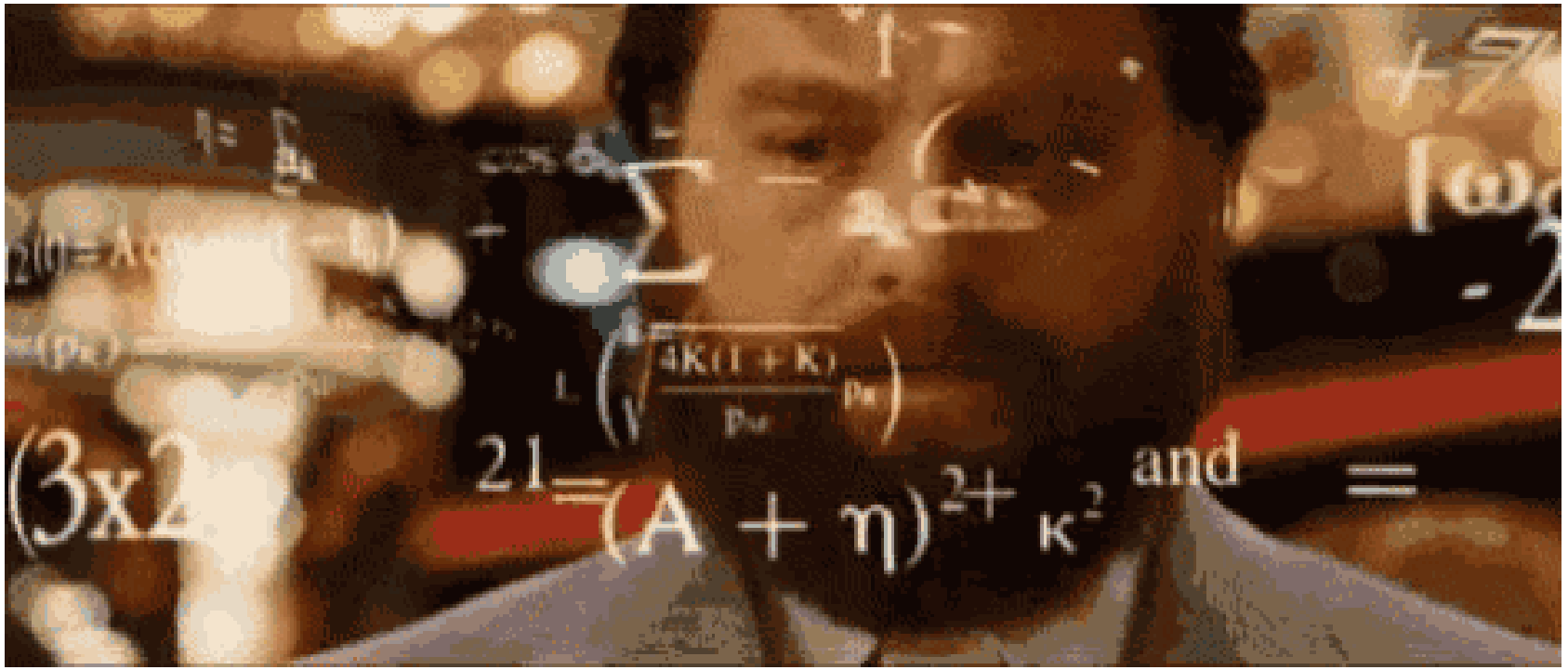
- f is the forecast origin
- h is the forecast horizon
- $\widehat{Y}_i$ is the forecasted time series in f+1,…,f+h

Then the quality of forecast w.r.t. given time series can be defined as the mean absolute error (MAE) with

$$\mathbf{D}(\widehat{\boldsymbol{Y}},\mathbf{Y}) = \frac{1}{h}\sum_{i=1}^{h} |\widehat{Y}_i - Y_i|$$

Optimizing Mean Absolute Error (MAE)?

$$\mathbf{D}(\widehat{\boldsymbol{Y}},\mathbf{Y}) = \frac{1}{h}\sum_{i=1}^{h} |\widehat{Y}_i - Y_i|$$



=> Minimizing the error D maximizes the forecast quality?

Can we perform forecasts for every problem using standard optimization techniques?

Outlining Pitfall and Challenges using ML Theory

- Ugly Duckling Theorem (UDT) [Watanabe, 1969]
- Uncertainity (e.g. [Silver, 2012])
- Pitfall of Learning Behavior of ML Models [Ultsch Lectures; Geman et al., 1992; Gigerenzer/Brighton, 2009; Ben-David et al., 2018]
- No Free Lunch Theorem (NFL) [Wolpert, 1996]

What do the theorems mean?

No, but a lot of people thought so in various examples illustrated by [Silver, 2012] ☺

What does D($\hat{Y}$,Y) mean?

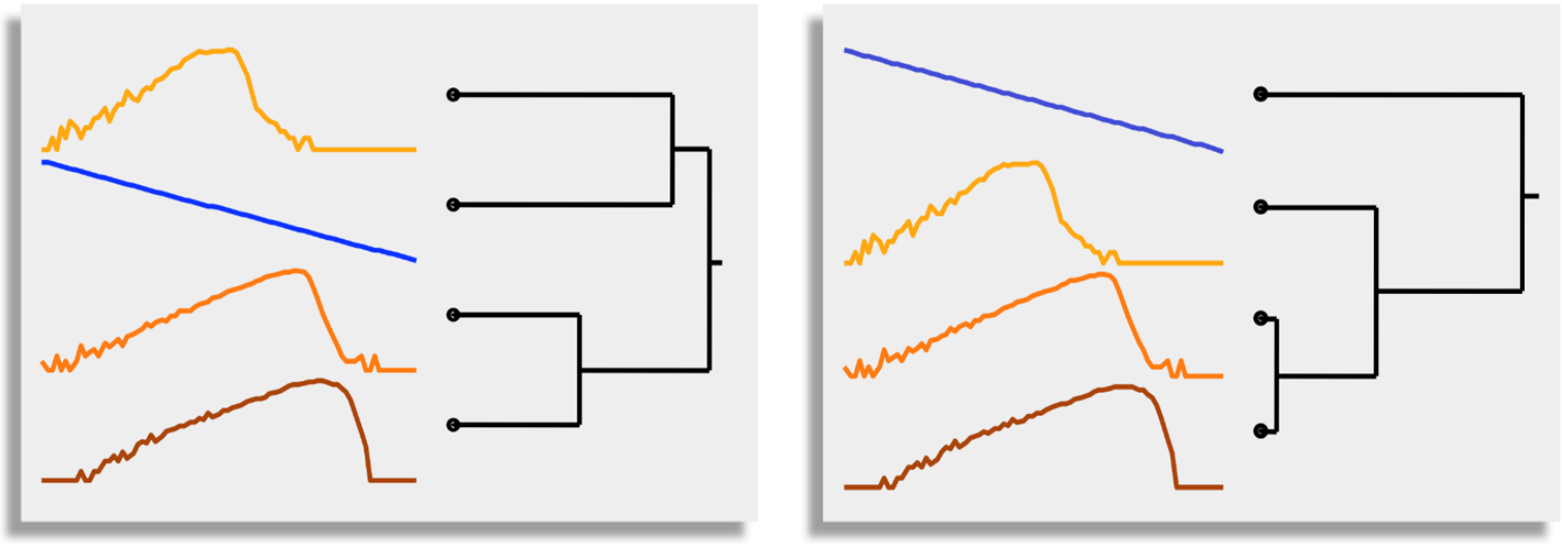
- Measuring Forecasting Quality is depicted with the error measure

  „ why?

=> Quality Measurement is a type of similarity/distance measurement:

- Comparison of curve of forecasted values with curve of historical data

- What is the consequence?



The definition of similarity depends on the user, the domain and the task at hand. We need to be able to handle this subjectivity.

The Ugly Duckling Theorem (UDT) [Watanabe, 1969]

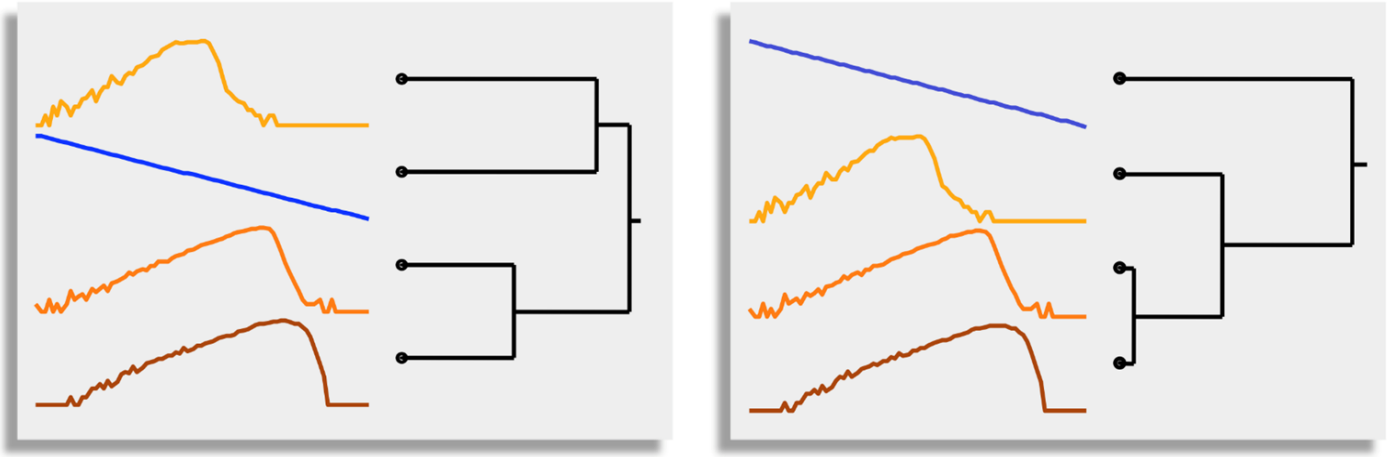
- One of the key question in Data Science is similarity
  - Relevant for every pattern recognition algorithm
  - (Dis-)similarity is most often Euclidean distance, but thats most often incorrect
- UDT states: **classification is impossible without some sort of bias**
  - Depends on the features chosen
    - Example on the right

⇒Similarity depends on the representation of data

Are this two pictures is similar to each other?

Consequence of UDT

- ⇒ UDT: Evaluation of forecasting results is always biased if a quality measure (QM) is seen as a similarity measure between the forecast curve and the test set curve of data [Thrun et al., 2019]
- ⇒ UDT: any two different curves share the same number of properties (c.f. [Watanabe, 1969])
- ⇒ QM should be chosen accordingly to the goal
- ⇒ Select the bias relevant to your problem (e.g. [Thrun et al., 2019])



The definition of similarity depends on the user, the domain and the task at hand. We need to be able to handle this subjectivity.

The signal and the noise [Silver, 2012]

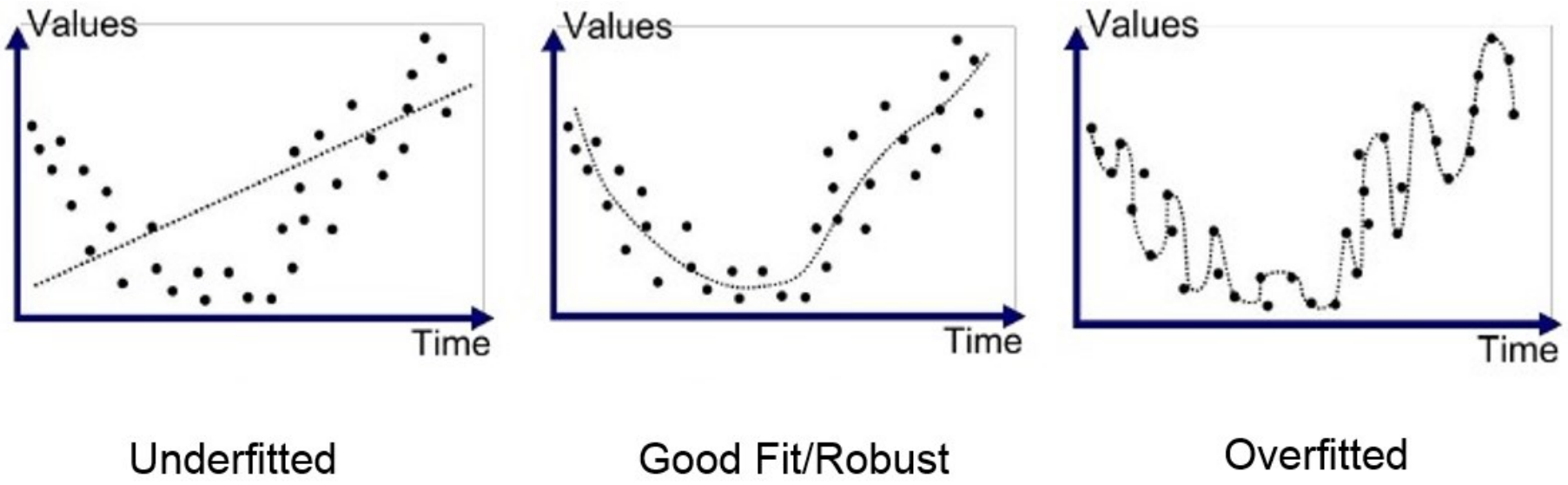
- In time series analysis, the data-generating process is unknown so that we must use the information in Y_f to build a model

=>Model itself is uncertain

$\Rightarrow$BUT its hard to handle model uncertainty, e.g. [Silver, 2012]

What unfavorable effects can happen?

Effects of Uncertainty



- Underfitting refers to a model that can neither model the training data nor generalize to new data

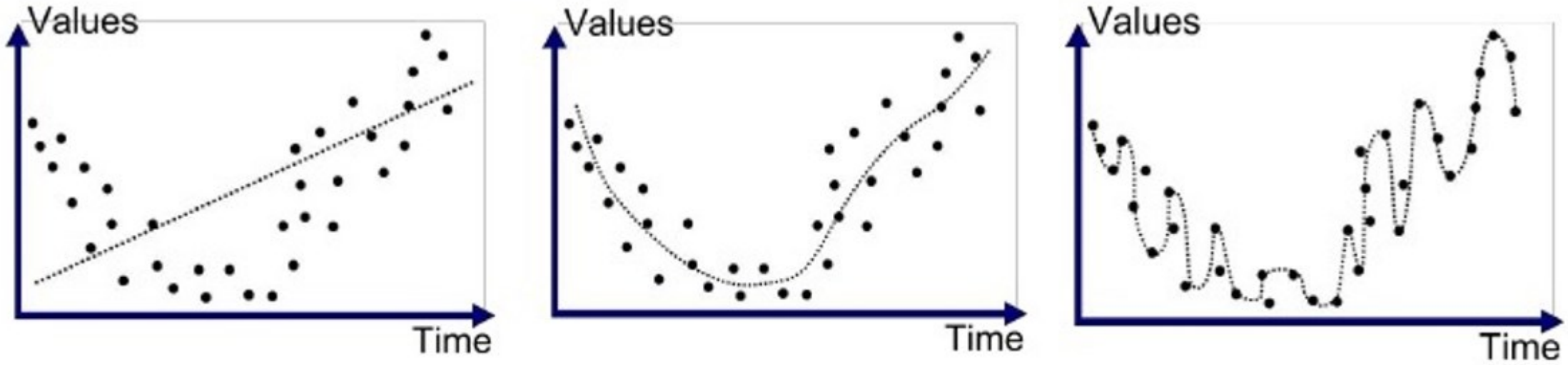
  -> Knowledge Discovery approaches on residuals and temporal structures

- Overfitting refers to a model that learns also noise instead of only learning the signal

  -> Should be investigated with statistical approaches using specific cross-validation

  -> More likely with nonlinear models like neural networks

Interpretation of Underfitting versus Overfitting



Underfitted | Good Fit/Robust | Overfitted

- Underfitting ⇔ Model has high bias
- Overfitting ⇔ model has a high variance (especially on the test set)

Limiting Uncertainty by “More Learning”?

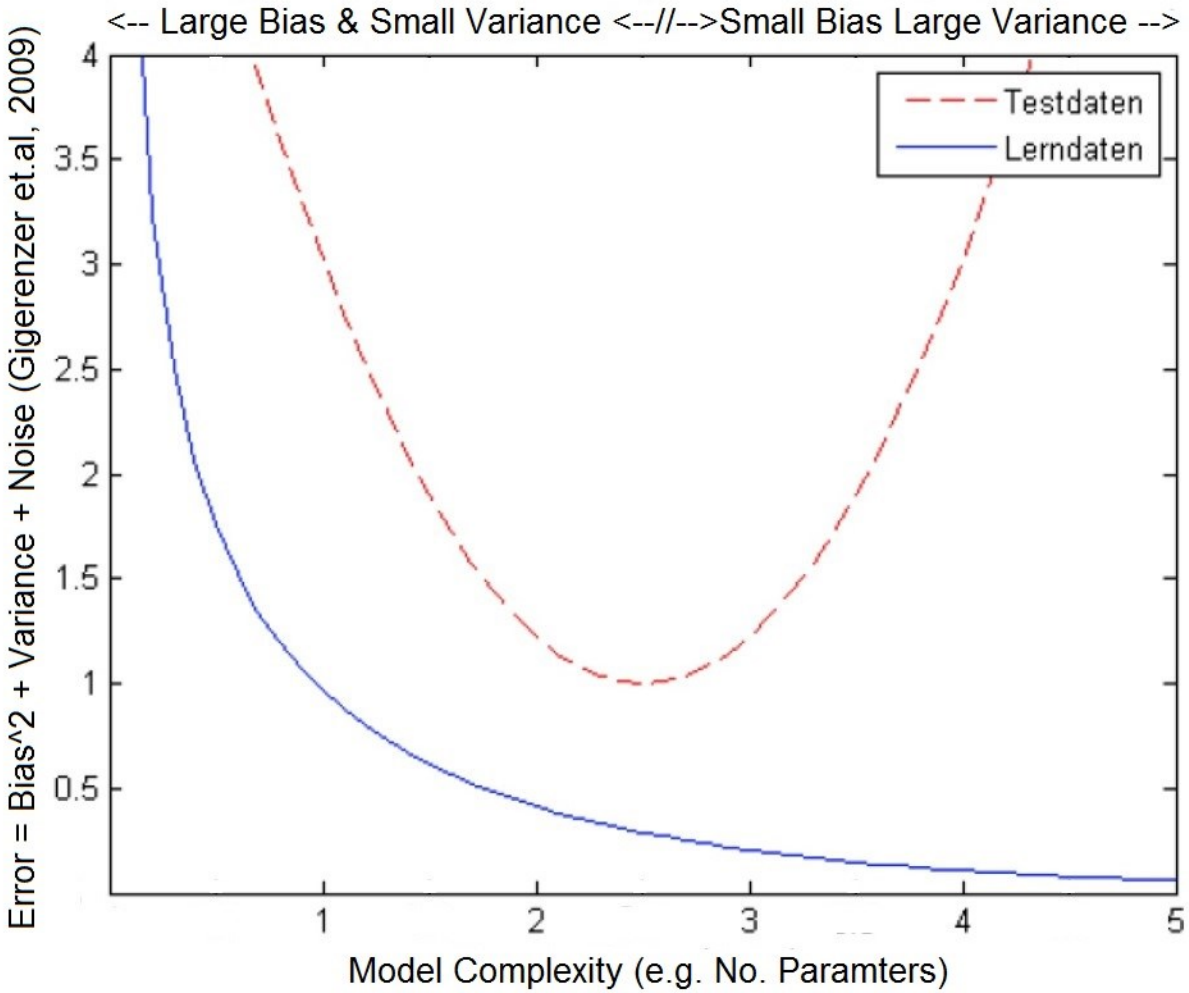
- Uncertainty can lead to either overfitting or underfitting

- Can we assume that larger training datasets or more complex models resolve the problem?
  - E.g. deep learning on large data sets

Pitfall of the Learning Behavior of ML Models

- The more data the method “learns”, the more complex the model and more precisely the adaptation
- However, above a certain limit, the underlying principles are not well mapped, i.e. the test data set is forecasted with large errors

- Sources: [Ultsch Lectures; Geman et al., 1992; Gigerenzer/Brighton, 2009; Ben-David et al., 2018]



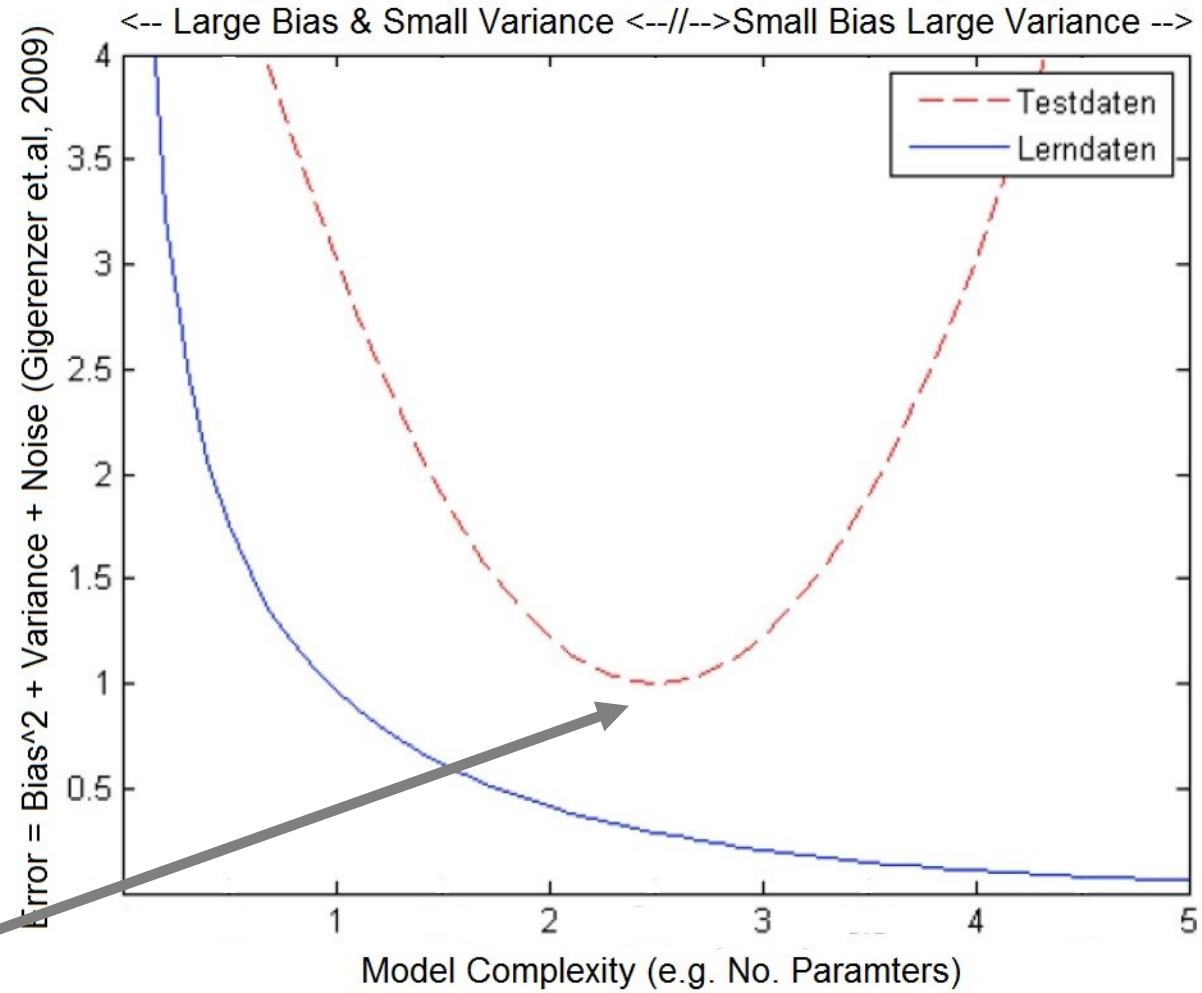
Learning behavior of forecasting models when comparing test data and learning data

Pitfall of the Learning Behavior of ML Models

- The more data the method “learns”, the more complex the model and more precisely the adaptation
- However, above a certain limit, the underlying principles are not well mapped, i.e. the test data set is forecasted with large errors

-> Optimal „Approximation of a target concept given a bounded amount of data” (“Learning”) cannot be proven [Ben-David et al., 2018]

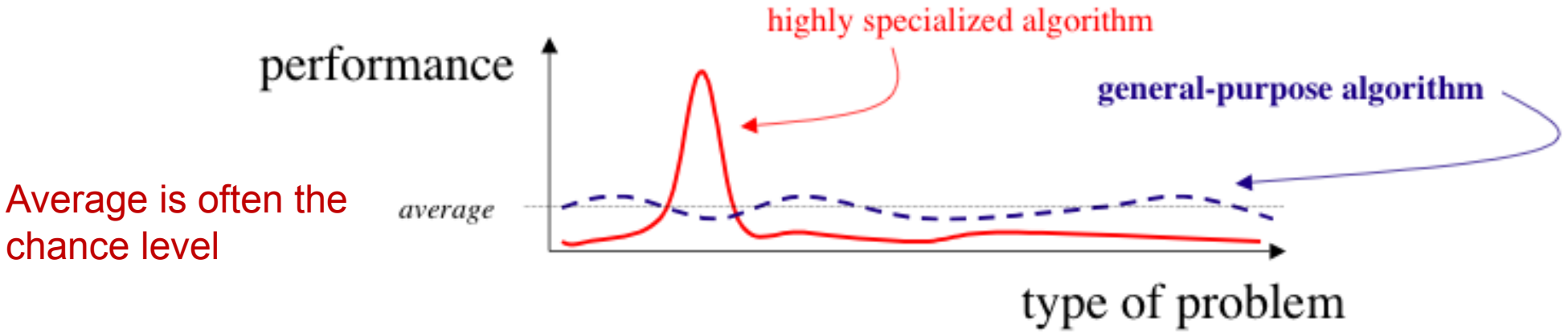
=> „Optimum unknown“



Learning behavior of forecasting models when comparing test data and learning data

No Free Lunch Theorem (NFL) - [Wolpert, 1996]

- Let X Time series of electricity prices should be forecasted, one per region, then
- **No solution given by an algorithm can be better than any other if the number of problems is high enough**



⇒Choose the "right" representation of data meaning that you answer the question:
„What are my objects/smallest units in which I operate"?

⇒Select algorithm(s) problem-specific and adapted to your data representation(e.g. [Thrun et al., 2019])

⇒Requires discussion with domain expert

Summary of Challenges

- Underfitting can be restricted with typical knowledge discovery approaches
  - investigate residuals and their temporal structures
- Overfitting can be handled with Crossvalidation
    - Test-sample bigger than one season (repeating cycle)
    - At least 100 times for statistical relevance in order to handle uncertainty
- The Ugly Duckling Theorem (UDT)
  - Quality measure (similarity) has to be defined problem-specific

    -> communicate with domain expert
- No Free Lunch Theorem (NFL)
  - No algorithm can outperform all others in every application

=> Data preprocessing using knowledge discovery and communication with the domain expert is critical

Benchmarking Various Methods

- Premise is that data source (Nord Pool) is the appropriate market providing useful data for forecasting
  - c.f. Prediction of Quarterly Financial Data of Stocks [Thrun, 2019a])

- Lets try to spot the implicit assumption I made
- Maybe even spot mistakes I made ☺

-> Benchmarking in Rmarkdown….

Explicit Assumptons are the Key Factors for Success

- Well-selected forecasting horizon/interval, data & target feature(s)
- Is the appropriate unit: Euro/MWh of hourly time series?
  - Even if the region lie in scandinavian countries having other currencies
  - No systematic change between weekdays (e.g. [Taylor/Letham, 2018]) or day and night (e.g. [Locarek-Junge, 2019]) exist
- Does the regional location of time series not result in structural differences?
  - Contrary example in country-based GDP [Thrun, 2019b]
- Overfitting is accounted for well because time interval of test set is large enough (at least one season)
- Quality measure with an appropriate bias is chosen

1. Unambiguous Objective

2. Proper Data Reprensentation

3. Limited Uncertainty

4. Choice of Similarity

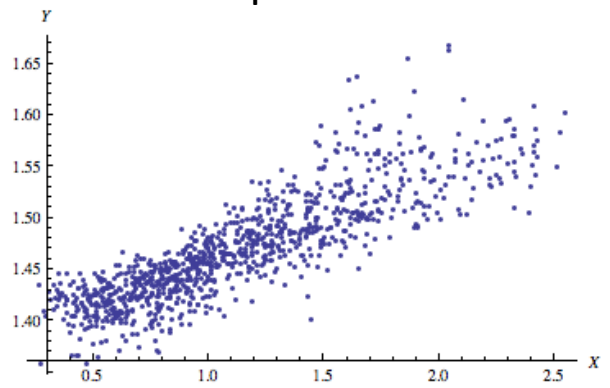
Implicit Assumptions for Working with ANNs

I. Less complex but „understandable“ ML approaches approaches (like Facebook‘s prophet, [Taylor/Letham, 2018]) failed

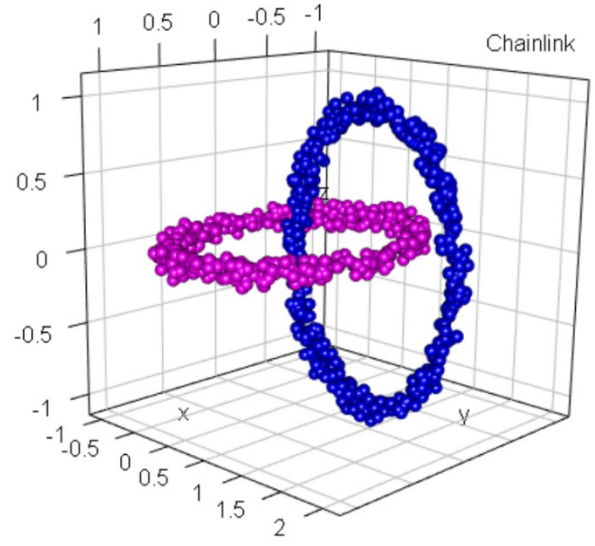
II. Knowledge discovery performed extensively
   1. Structure and quality of the data must be statistically clarified PRIOR to coding
   2. Problem is multivariate
      - Relevant information is contained in more than one variable
      - Information is nonlinear and complex interrelated
   3. Causality (cause -> effect)
      - No other predictors majorly influence the price market significantly (e.g. energy production, hydro reservoirs in Scandinavia, …)
      - Used predictors are available BEFORE period of forecasting

=> Then, and only then we should use artificial neural networks

Simple Interrelation



Complex Interrelation



Pitfalls for Artificial Neural Networks (ANN)

- Too much data can lead to memorization of each data point
- Too complex models will model everything leading to complete deterministic worldview
- “Which ANN with which architecture is suitable for a given problem remains largely unsolved” [Ultsch Lectures on ANN]
- Artificial neural networks usually learn based on errors and are not human-understandable (interpretable)

$\Rightarrow$Learning is self-restricted (no more errors no more optimization)

$\Rightarrow$But one can learn more/better if one uses knowledge and understands the problem (interpretable ML)

$\Rightarrow$Use Ockham's razor to decided between models

Summary

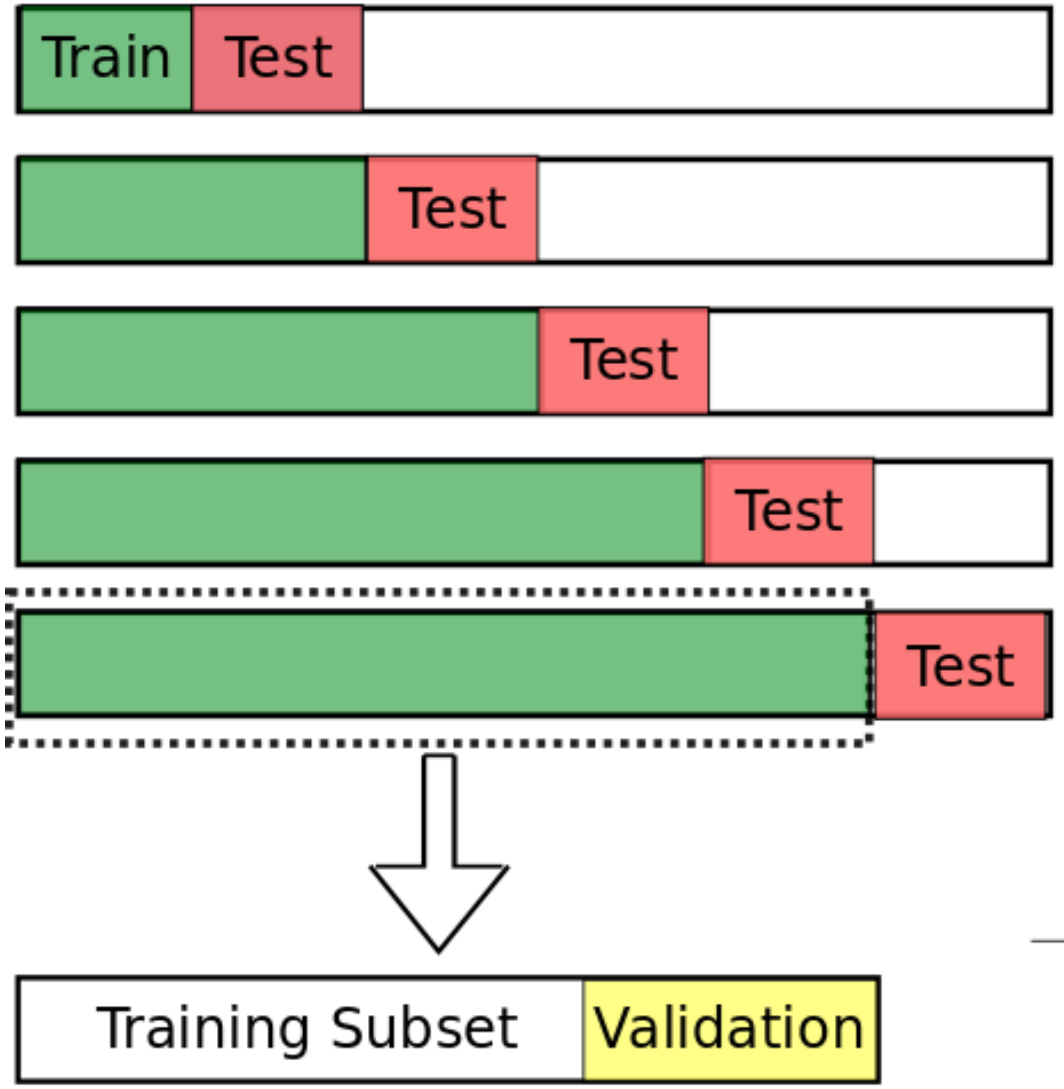
- Data Science Projects
  - Are interdisziplinary –> require team work and communication
  - Are successfully performed if an domain expert is available
  - Should only start if they have a clear objective
- Forecasting has the following principal challenges
  - Problem-specific preprocessing is the key to success and requires knowledge discovery
  - Uncertainty can lead to either overfitting or underfitting
    - Underfitting has to be investigated with methods of knowledge discovery
    - Overfitting should be investigated statistical approaches
  - Biases of forecasting methods and quality measure have to be exploited depending on the structures/patterns of the time series relevant to the problem

Sources

- **Watanabe, S.**: *Knowing and Guessing: A Quantitative Study of Inference and Information*, New York, USA, John Wiley & Sons Inc., ISBN: 9780471921301, **1969**.
- **Wolpert, D. H.**: The lack of a priori distinctions between learning algorithms, *Neural Computation, Vol. 8*(7), pp. 1341-1390. **1996.**
- **Geman, S., Bienenstock, E., & Doursat, R.**: Neural networks and the bias/variance dilemma, *Neural Computation, Vol. 4*(1), pp. 1-58. **1992.**
- **Gigerenzer, G., & Brighton, H.**: Homo heuristicus: Why biased minds make better inferences, *Topics in cognitive science, Vol. 1*(1), pp. 107-143. **2009.**
- **Silver, Nate**: *The signal and the noise: why so many predictions fail--but some don't*. Penguin, **2012.**
- **Locarek-Junge, H.** : A Night in the Stock Exchange. What Happens Between Dusk and Dawn to the WIG20 Index?, Proc. 5th German-Polish Symposium on Data Analysis and Applications (GPSDAA), p. 3, Germany, **2019.**
- **Taylor, S. J., & Letham, B.**: Forecasting at scale, *The American Statistician, Vol. 72*(1), pp. 37-45. **2018.**
- **Thrun, M. C.**: *Knowledge Discovery in Quarterly Financial Data of Stocks Based on the Prime Standard using a Hybrid of a Swarm with SOM*, in Verleysen, M. (Ed.), European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Vol. 27, pp. 397-402, Ciaco, 978-287-587-065-0, Bruges, Belgium, **2019a.**
- **Thrun, M. C.**: Cluster Analysis of Per Capita Gross Domestic Products, *Entrepreneurial Business and Economics Review (EBER), Vol. 7*(1), pp. 217-231. doi 10.15678/EBER.2019.070113, **2019b.**
- **Thrun, M. C., Märte, J., Böhme, P., & Gehlert, T.**: Applying Two Theorems of Machine Learning to the Forecasting of Biweekly Arrivals at a Call Center, Proc. European Conference on Data Analysis (ECDA), pp. 36, Bayreuth, Germany, **2019**.
- **Behnisch, M., Ultsch, A.:** Knowledge Discovery in Spatial Planning Data - A Concept for Cluster Understanding, *in: Helbich, M., Arsanjani, J. J., Leitner, M. (eds.): Computational Approaches for Urban Environments, in: Gatrell, J.D., Jensen, R.R.: Geotechnologies and the Environment Series, Vol, 13, Springer, Berlin, pp. 49-75,* ***2015.***
- ***Ultsch Lectures****: „Neuronale Netze“, Lehrstuhl Neuroinformatik und Künstliche Intelligenz, Universität Marburg.*
- **Anastasiadis, A. D., Magoulas, G. D., & Vrahatis, M. N.**: New globally convergent training scheme based on the resilient propagation algorithm, *Neurocomputing, Vol. 64*, pp. 253-270. **2005.**
- **Huang, G.-B., Wang, D. H., & Lan, Y.**: Extreme learning machines: a survey, *International journal of machine learning and cybernetics, Vol. 2*(2), pp. 107-122. **2011.**
- **Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K.**: Extreme learning machine: theory and applications, *Neurocomputing, Vol. 70*(1-3), pp. 489-501. **2006**
- **Ben-David, S., Hrubeš, P., Moran, S., Shpilka, A., & Yehudayoff, A.**: Learnability can be undecidable, *Nature Machine Intelligence, Vol. 1*(1), pp. 44. **2019.**

Excurse: Nested Cross-validation

- Nested Cross-validation can account for overfitting in complex models



Outer Loop

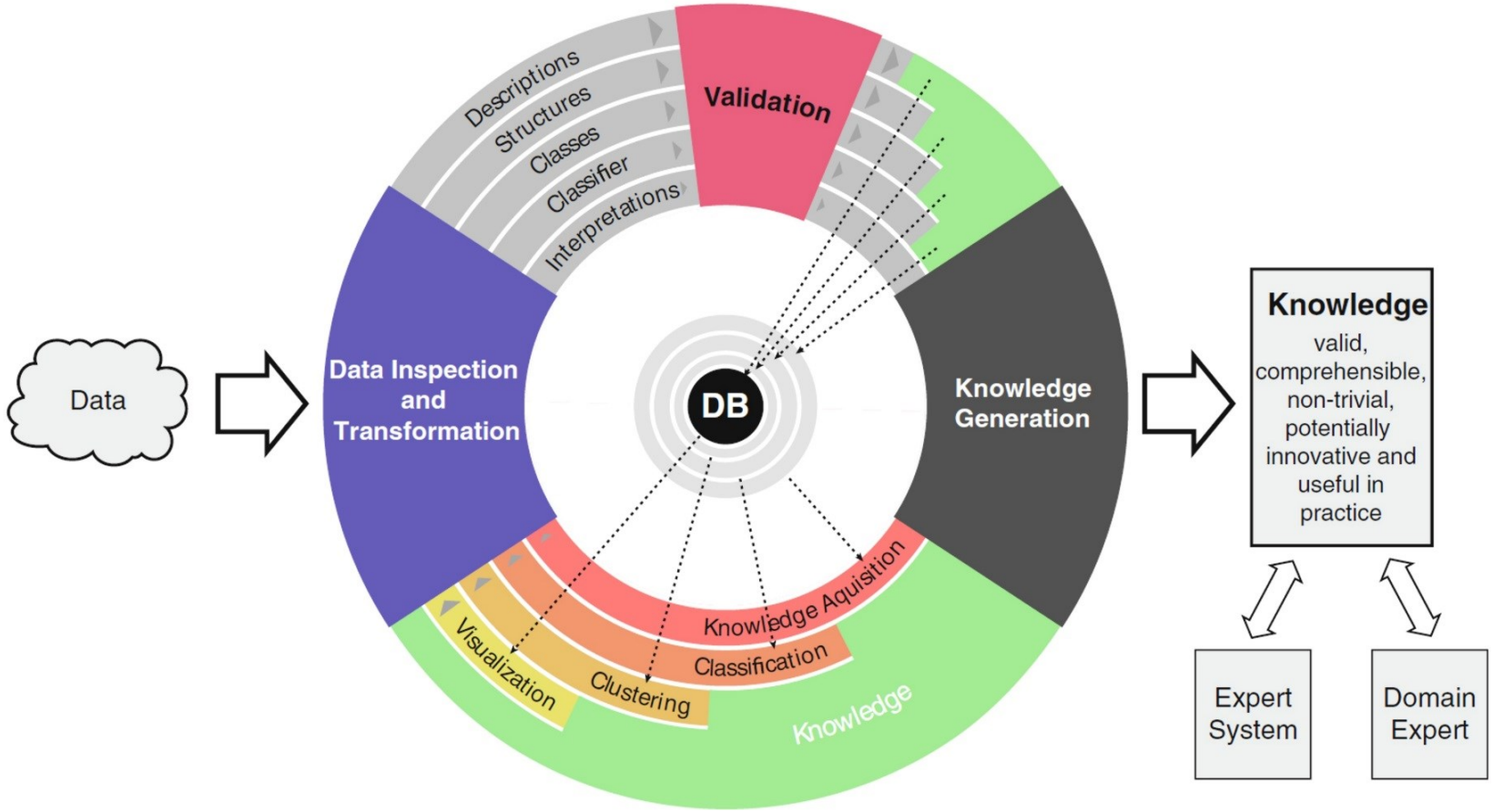
1. Train each split with optimal parameters
2. Average each split's test error

Inner Loop

Tune hyperparameters

Excurs: Knowledge Discovery

1. Descriptions: Modeling the distribution of each variable separately
2. Structures: Finding structures in high-dimensional space
3. Classes: Finding intrinsic groups, called clusters, in a data set
4. Classifier: Symbolic classifiers to assist human skills of comprehension
5. Interpretations: Human understanding of clusters



Source: [Ultsch/Behnisch, 2015]

Key Factors for Success in Data Science

1. Unambiguous Objective

- In Forecasting: Well-selected forecasting horizon/interval, data & target feature(s) and algorithms (interpretable ML vs ANN)
- Choice of appropriate resolution and units/objects
  - For example is there a systematic change between weekdays (e.g. [Taylor/Letham, 2018]) or day and night (e.g. [Locarek-Junge, 2019])

  => Disaggregate one time series into components

2. Proper Data Reprensentation

- Does the regional location of time series has structural differences?
  - Perform cluster analysis (e.g. [Thrun, 2019b])
- Can problems be summarized? ->Aggregate multivariate problem to univariate time series (e.g. [Thrun et al., 2019]))
- Preprocessing (e.g. detrending, standardization,…)

3. Limited Uncertainty

- Overfitting is accounted for well if time interval of test set is at least one season with a statistically large enough sample (>100 cases)

4. Choice of Similarity

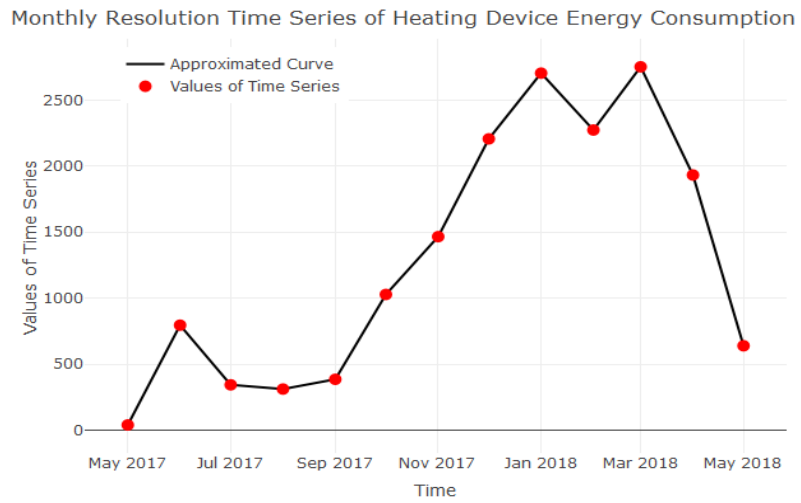
- Quality measure with an appropriate bias is chosen
- Algorithm with appropriate bias is selected

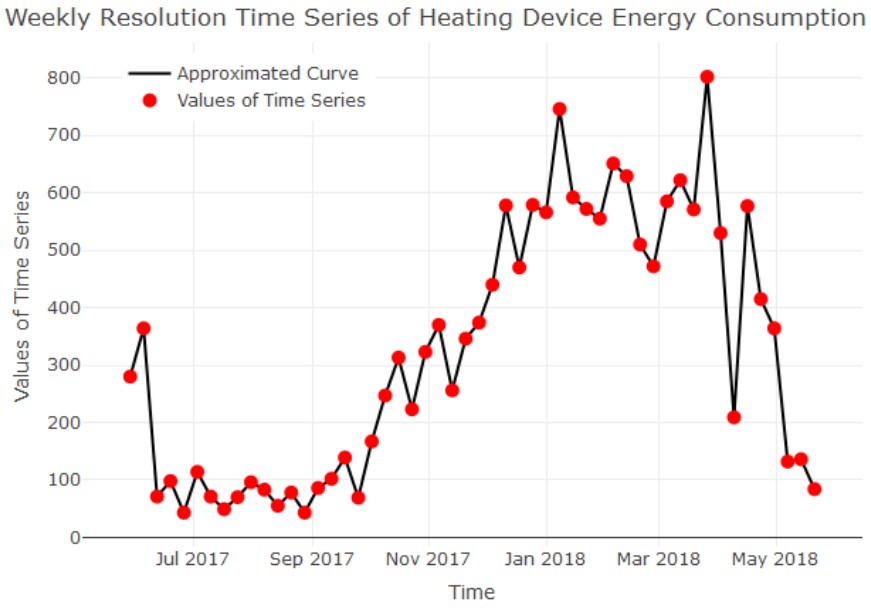
Recapitulation: Forecasting

Y_t is equidistant and a series of measurements visualized as curve with a specific resolution

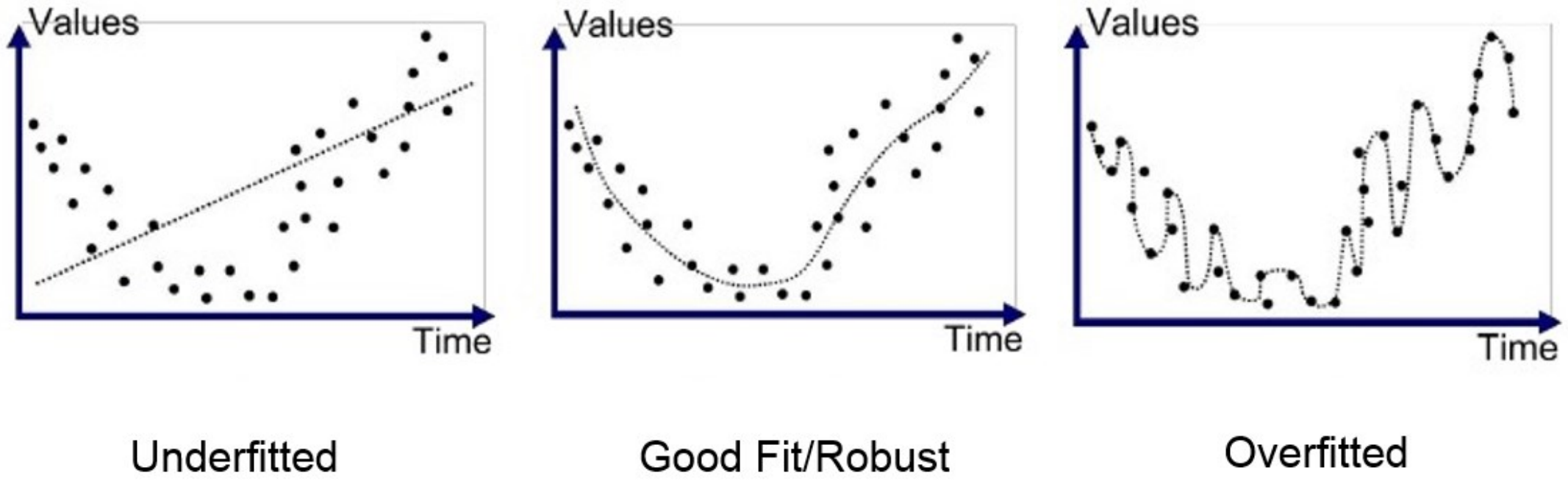
Given Y_f, an h step forecasts is a prediction of the values of $Y_{f+1}, \ldots, Y_{f+h}$ if it is based only on the information available at time t=f

- f is the forecast origin
- h is the forecast horizon
- $\widehat{Y}_i$ is the forecasted time series in f+1,…,f+h





Effects of Uncertainty



- Underfitting refers to a model that can neither model the training data nor generalize to new data

  -> Knowledge Discovery approaches on residuals and temporal structures

- Overfitting refers to a model that learns also noise instead of only learning the signal

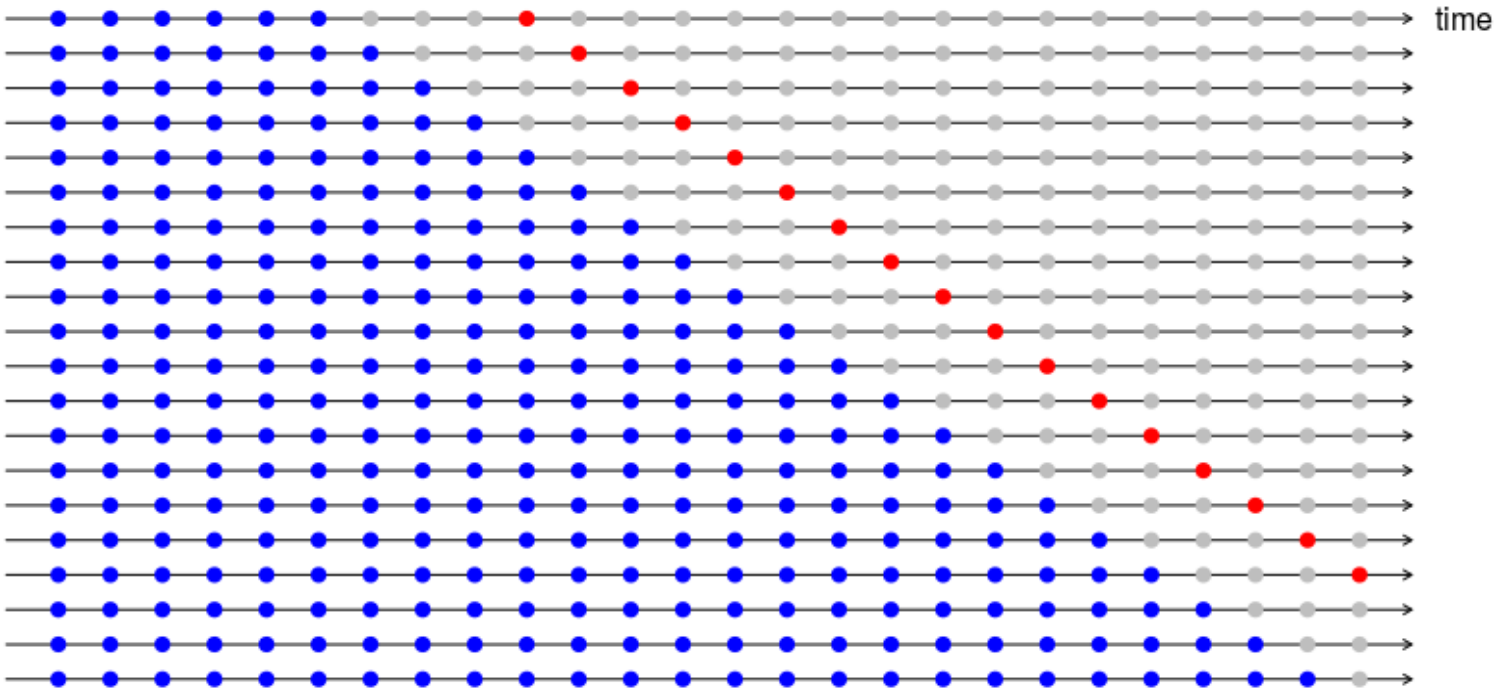
  -> Should be investigated with statistical approaches using specific cross-validation prozedure

  -> More likely with nonlinear models like neural networks

Best Practice: Cross-Validation

- Cross-validation
  - Use more than one round of out-of-sample forecasting
  - Account for temporal structures by rolling forecast
  - Multi-step forecasting horizon with unit point h in red
  - See coding in Rmarkdown for details (presented later)

- Blue: training set
- grey: test set (sample not used in model building)
- red: forecast horizon



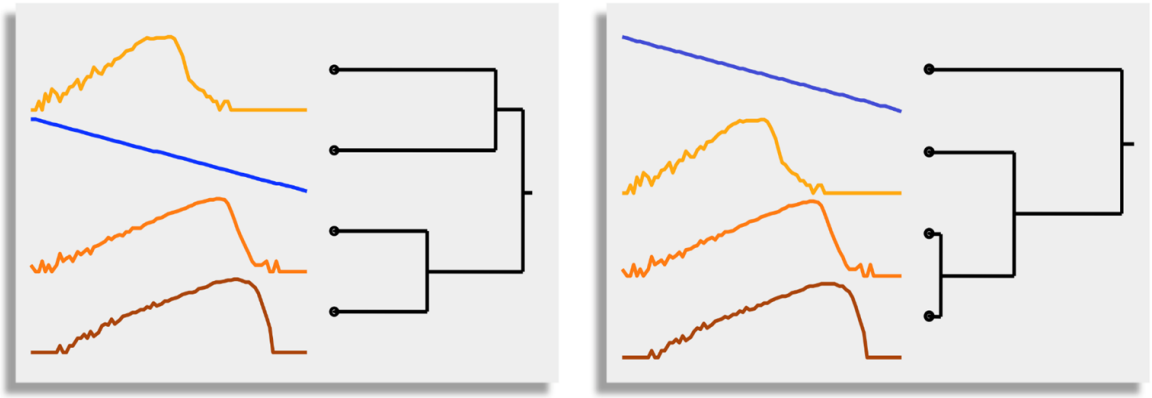
The Ugly Duckling Theorem (UDT) [Watanabe, 1969]

- One of the key question in Data Science is similarity
  - Relevant for every pattern recognition algorithm
  - (Dis-)similarity is most often Euclidean distance, but thats most often incorrect
- UDT states: **classification is impossible without some sort of bias**
  - Depends on the features chosen

⇒Similarity depends on the representation of data

=> Quality Measurement can be regarded as a similarity measure [Thrun et al., 2019]

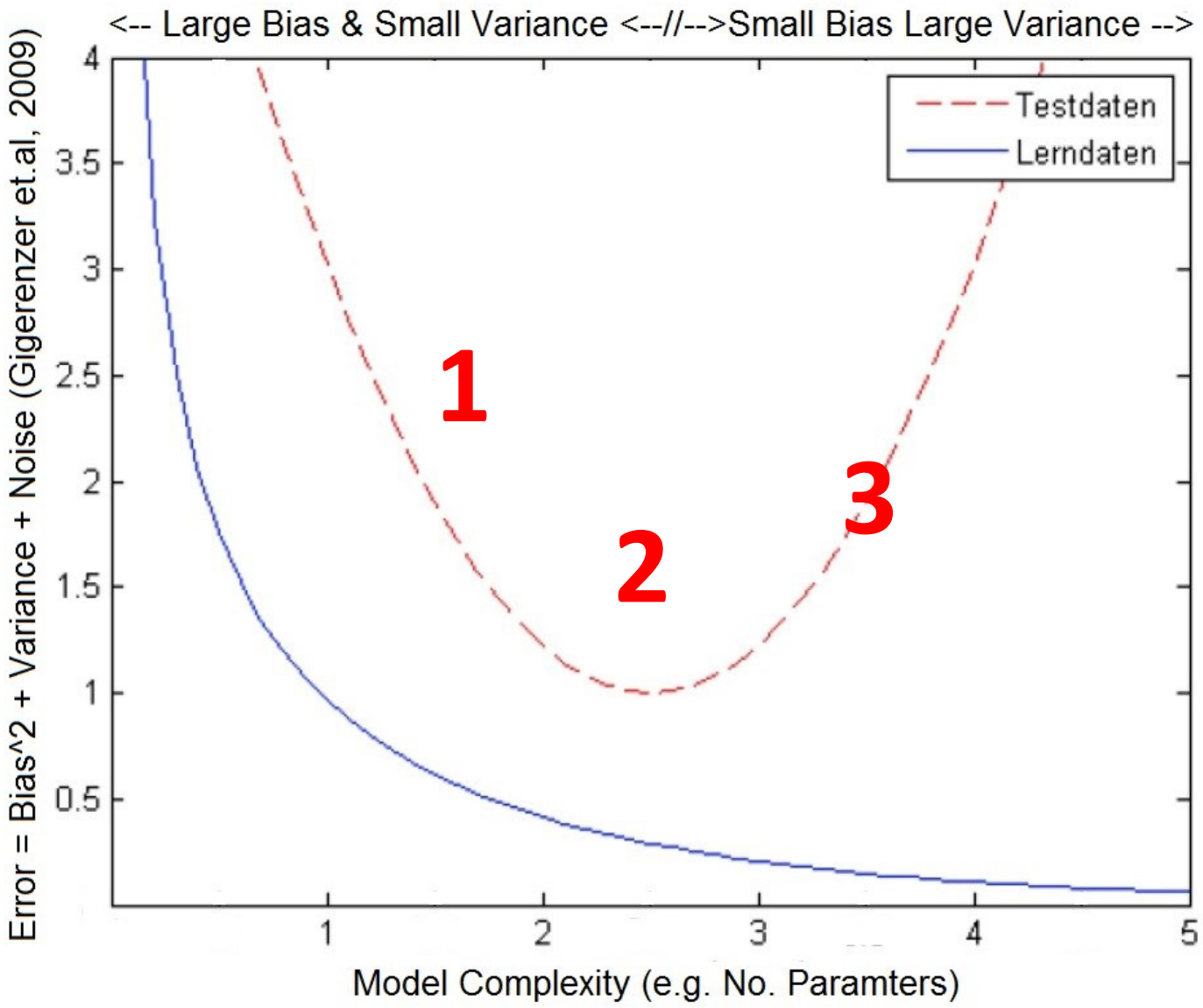
Are this two pictures is similar to each other?



The definition of similarity depends on the user, the domain and the task at hand. We need to be able to handle this subjectivity.

Pitfall of the Learning Behavior of ML Models

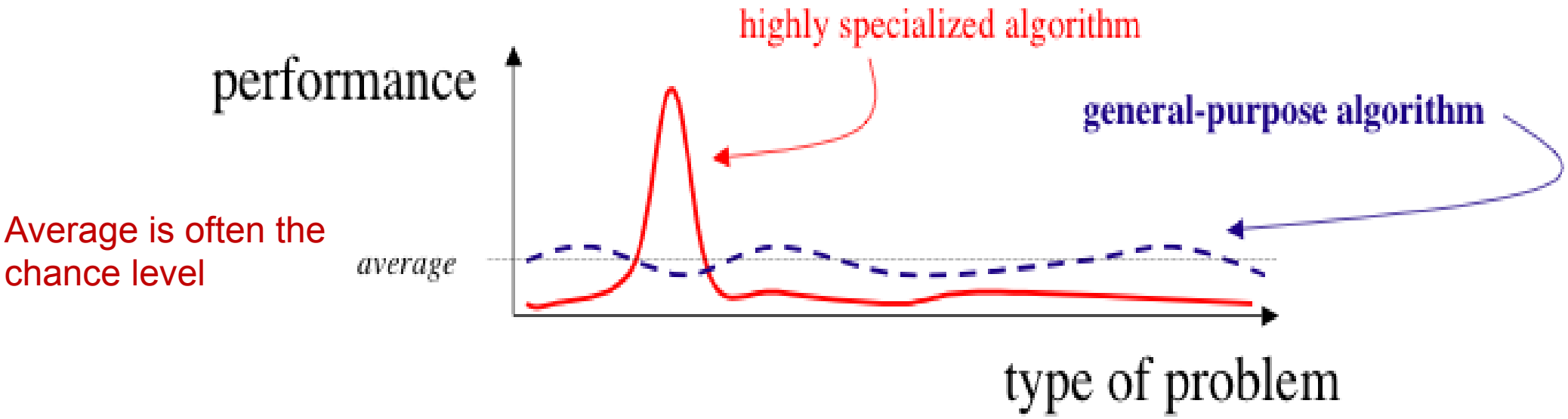
1. Bias of algorithm is reduced and variance is increased in relation to model complexity
2. As more and more parameters are added to a model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls
3. However, above a certain generally unknown limit the underlying principles are not well mapped, i.e. the test data set is forecasted with large errors



Sources: [Ultsch Lectures; Geman et al., 1992; Gigerenzer/Brighton, 2009; Ben-David et al., 2018]

No Free Lunch Theorem (NFL) - [Wolpert, 1996]

- Let X Time series of electricity prices should be forecasted, one per region, then
- **No solution given by an algorithm can be better than any other if the number of problems is high enough**



⇒Choose the "right" representation of data meaning that you answer the question:
„What are my objects/smallest units in which I operate"?

⇒Select algorithm(s) problem-specific and adapted to your data representation(e.g. [Thrun et al., 2019])

⇒Requires discussion with domain expert

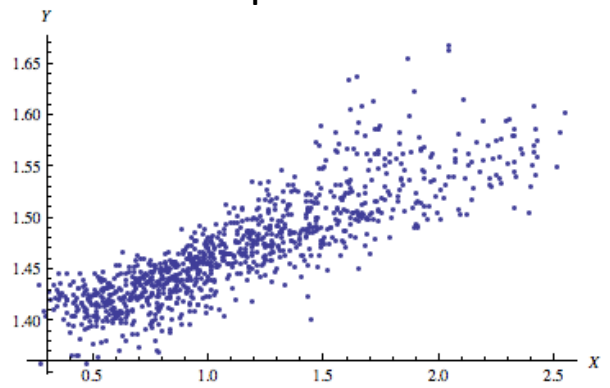
Implicit Assumptions for Working with ANNs

I. Less complex but „understandable“ ML approaches approaches (like Facebook‘s prophet, [Taylor/Letham, 2018]) failed

II. Knowledge discovery was performed extensively
   1. Structure and quality of the data were statistically clarified PRIOR to coding
   2. Problem is multivariate
      - Relevant information is contained in more than one time series
      - Information is nonlinear and complex interrelated
   3. Causality (cause -> effect)
      - No other predictors majorly influences the price market significantly (e.g. energy production, hydro reservoirs in Scandinavia, …)
      - Used predictors are available BEFORE period of forecasting

=> Then, and only then we should use artificial neural networks

Simple Interrelation



Complex Interrelation