Data-driven methods for chemists & chemical engineers

Validation problems of data-based models

Sebastian Werner, Mar/30th 2021

Code & slides are at https://github.com/blackw1ng/data-validation-lecture

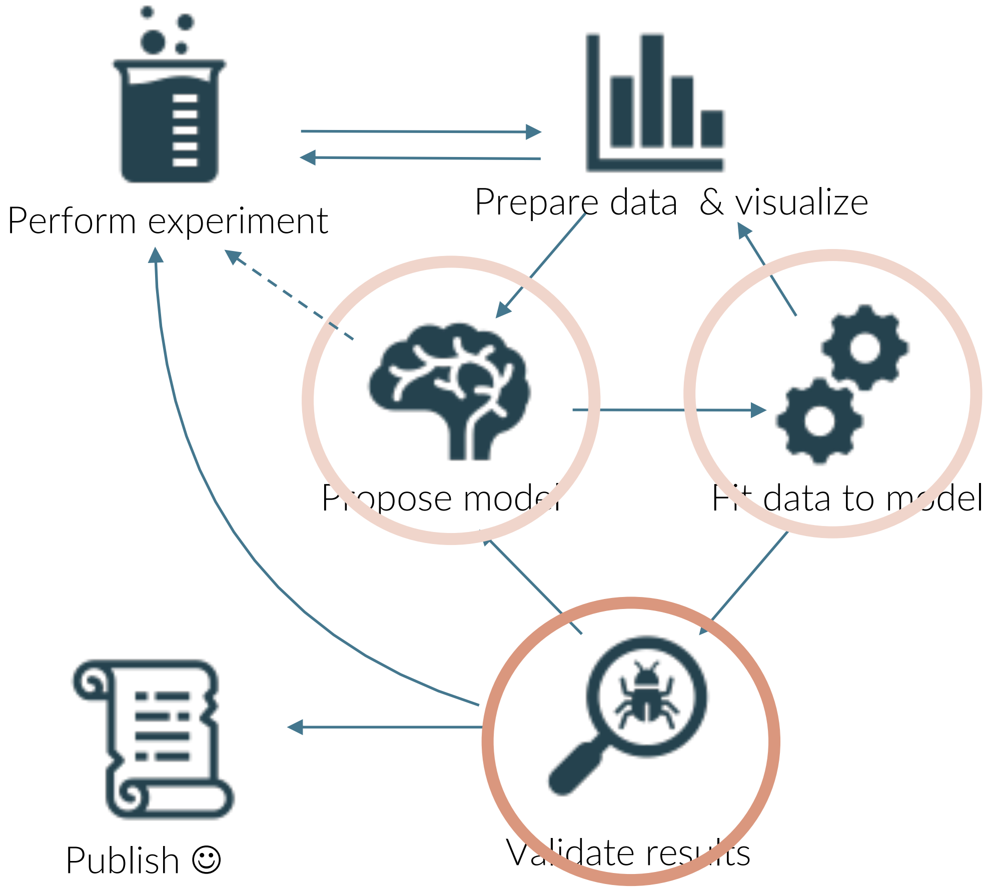
Underlying principles of data-based models

- A model has a **structure**, as well as **parameters**
- Models link **inputs** to a system to **outputs**

$$y = m \cdot x + b$$

- Common challenge in chemistry & chemical engineering:
    - Fitting **measurement** data to a **known model**
    - Even then, validation is important

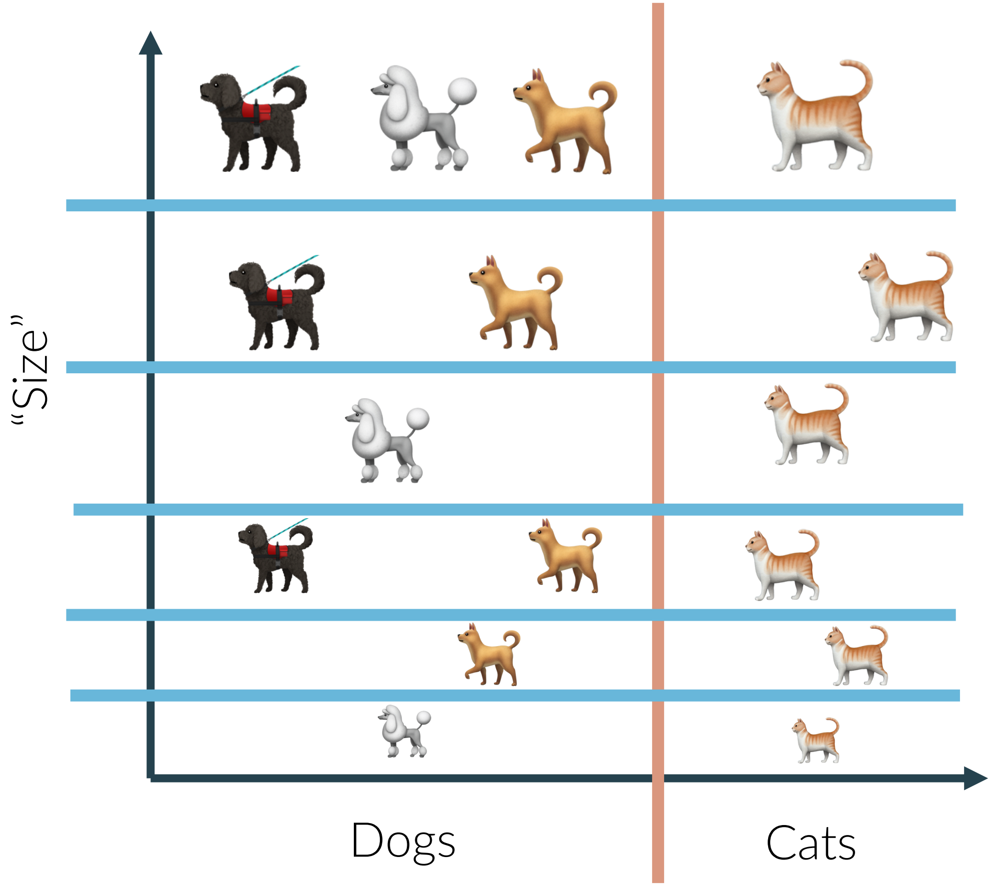
Today: Fit data where we have neither know the **structure** nor the **parameters**.
We just have **inputs** and **outputs**!



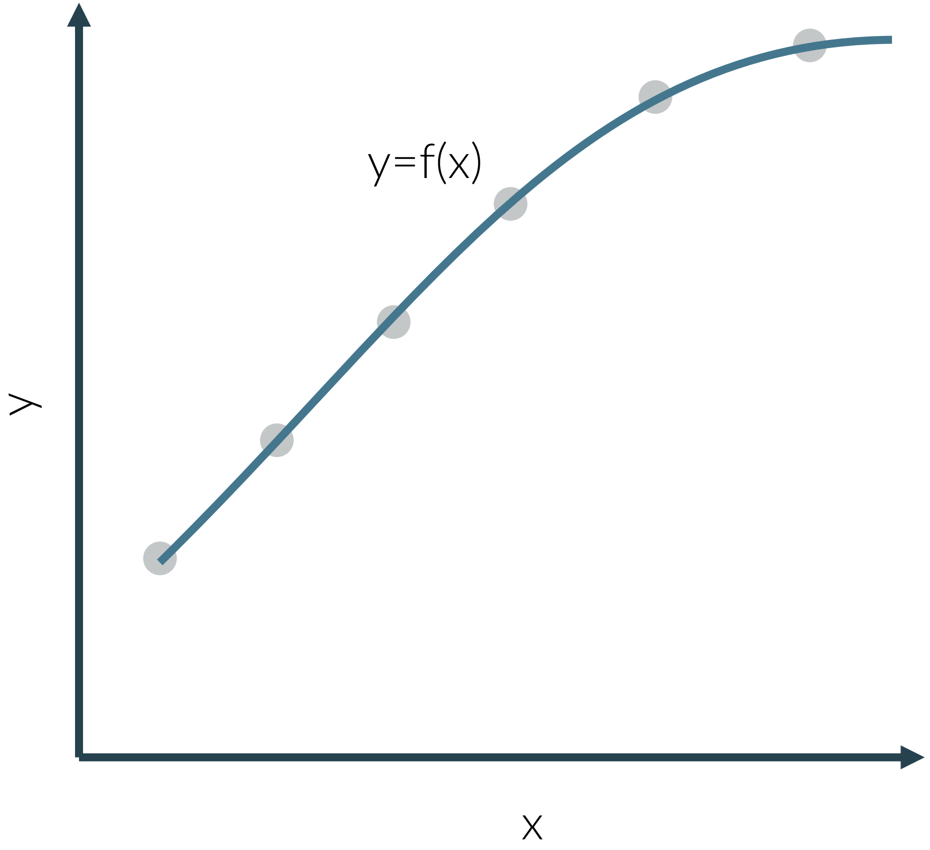
Modified CRISP-DM process for experimental data analysis

Know your challenge: Classification vs. Regression problems

Assignment of a **label** based on input



Assignment of a **quantity** based on input

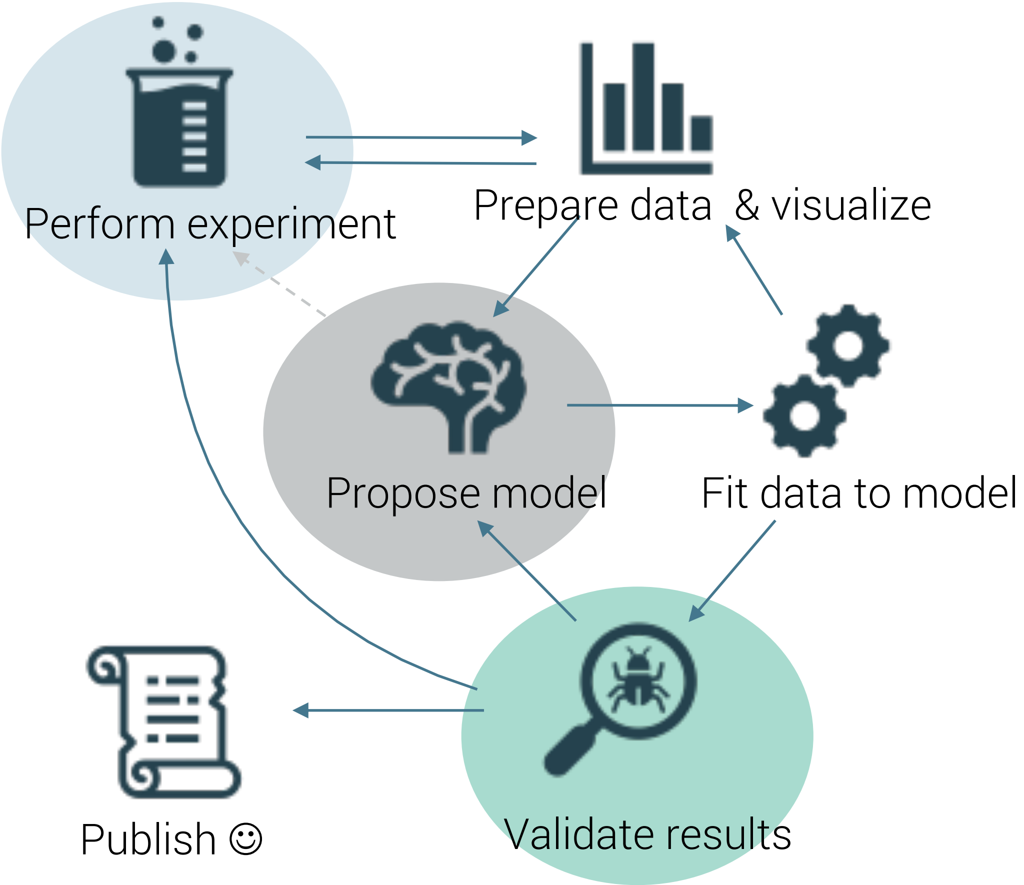


Focus of today a data-based regression models

Models & validation: A chicken and egg problem

Observations can be used to inspire a model structure… and you validate them with more experiments

- Newton’s first law
- Stefan-Boltzmann law
- Transport-resistance laws:
  Fourier’s, Ohm’s, Fick’s & Darcy’s law



Models postulated based on theory and then subsequently proof / validate with experiments

- Einstein theory on relativity
- Higg’s boson

Validation of models versus model verification

Verification

Making sure a model structurally fits the training data

Validation

Assessment of predictive quality outside the testing regime

Verification answers the questions, whether the model was **built right**.
Validation answers the question, that the **right model** was built.

D Cook, J.Skinner CrossTalk 2005 18(5), 20-24.

➔ It makes little sense to validate a model without prior verification!

Validation challenges of data based-models

Data-based “soft” models exhibit **several pitfalls**:

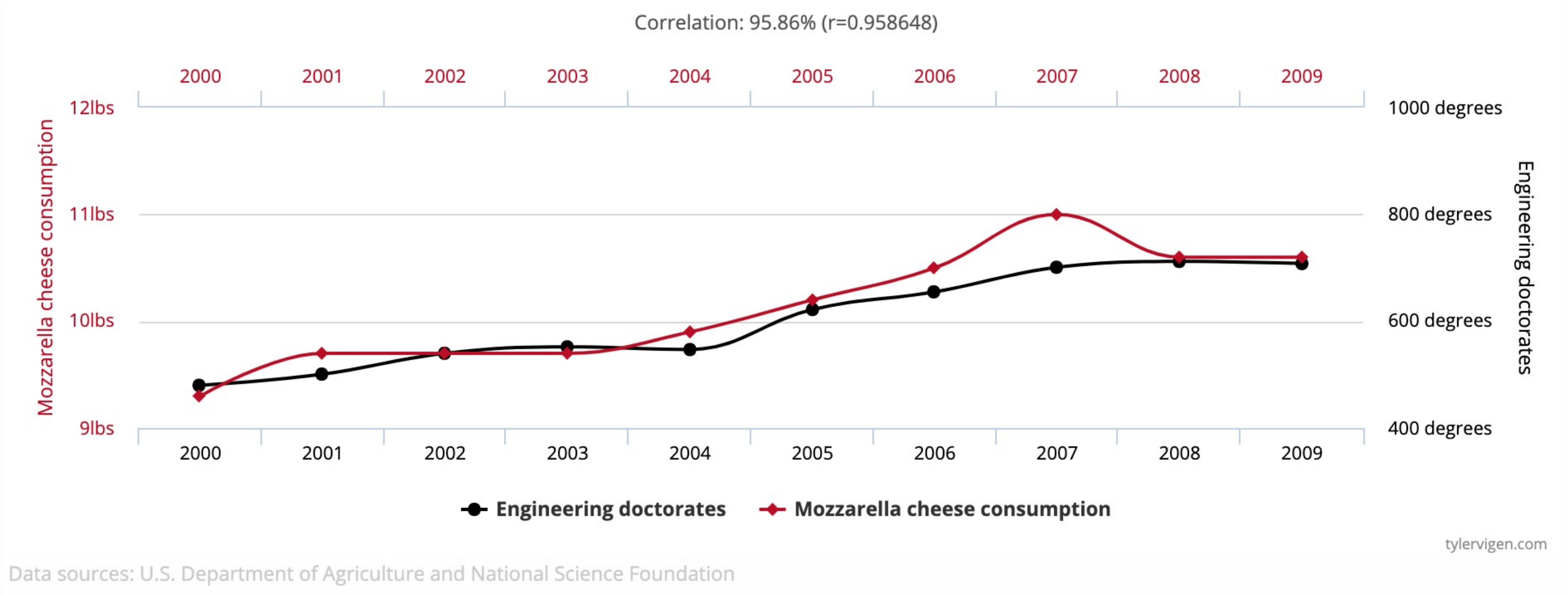
- **<u>Overfitting</u>**: That may describe available data well, but fail to extrapolate
- **<u>Underfitting</u>:** Inputs that may influence the modelled output are **not considered**
- **<u>Not robust</u>**: Model changes significantly depending on training
- **<u>Cause-effect</u>** relationships are not necessarily correctly described

Validation of a **model** implies a previous **validation** of **experimental setup** that generated this **data**!
(Basically: Calibrate / Validate input and output)

Can the model distinguish between cause and effect?

Causal: Clear relationship between input and output.

Descriptive: cannot give conclusive evidence about cause and effect.



Quality measures for regressions

- **R-squared** (R^2): *For linear models only!*
  Prediction error divided by deviation from mean.

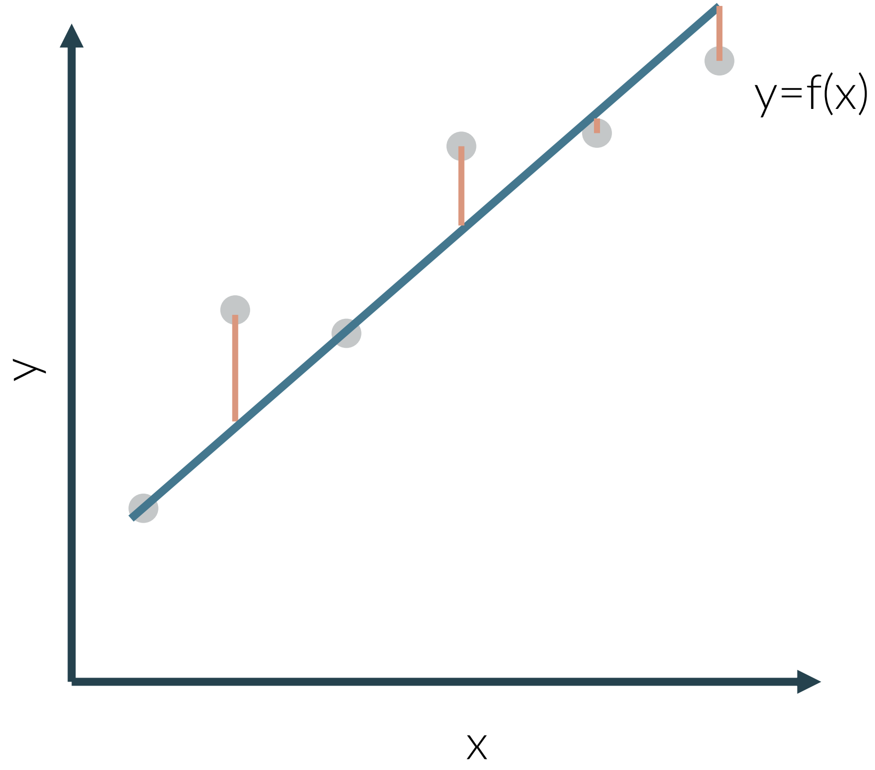
$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- **Mean Absolute Error** (MAE)
  Normalized sum of absolute prediction errors

$$MAE = \frac{1}{N} \sum_i^N |y_i - f_i|$$

- **(Root) Mean Squared Error** (RMSE):
  (Square root of) normalized sum of squared distance of real value and prediction

$$RMSE = \sqrt{\frac{1}{N} \sum_i^N (y_i - f_i)^2}$$



RMSE gives large penalty to big prediction error (e.g. outliers) by square it while MAE treats all errors the same.

Split available datasets: Train & test splits

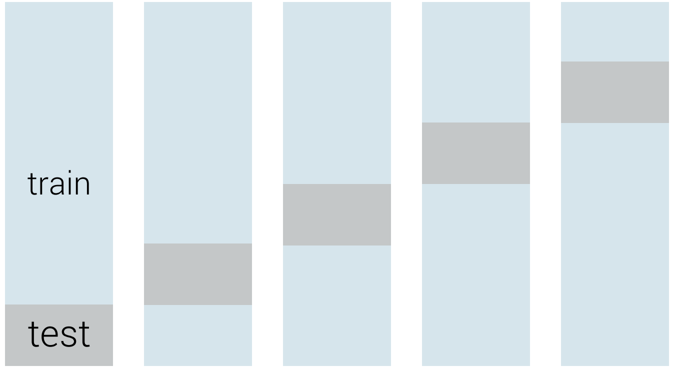
- Take most of the data for “**training**” the model – and **verification**
- Spare some data to **test** the model & **validate** it
- This is specifically important for data-based models

- For larger datasets, a common-technique is k-fold **cross-validation**
  - Compare results from each “fold”
  - You do it k times
- In case results vastly differ, it indicates a ill-defined model

train/test splitting

t	x	y	
13:01	23	2	train
13:02	23.5	5	train
13:03	42	6	train
13:04	21.2	2	train
13:05	42	3	test
…			test

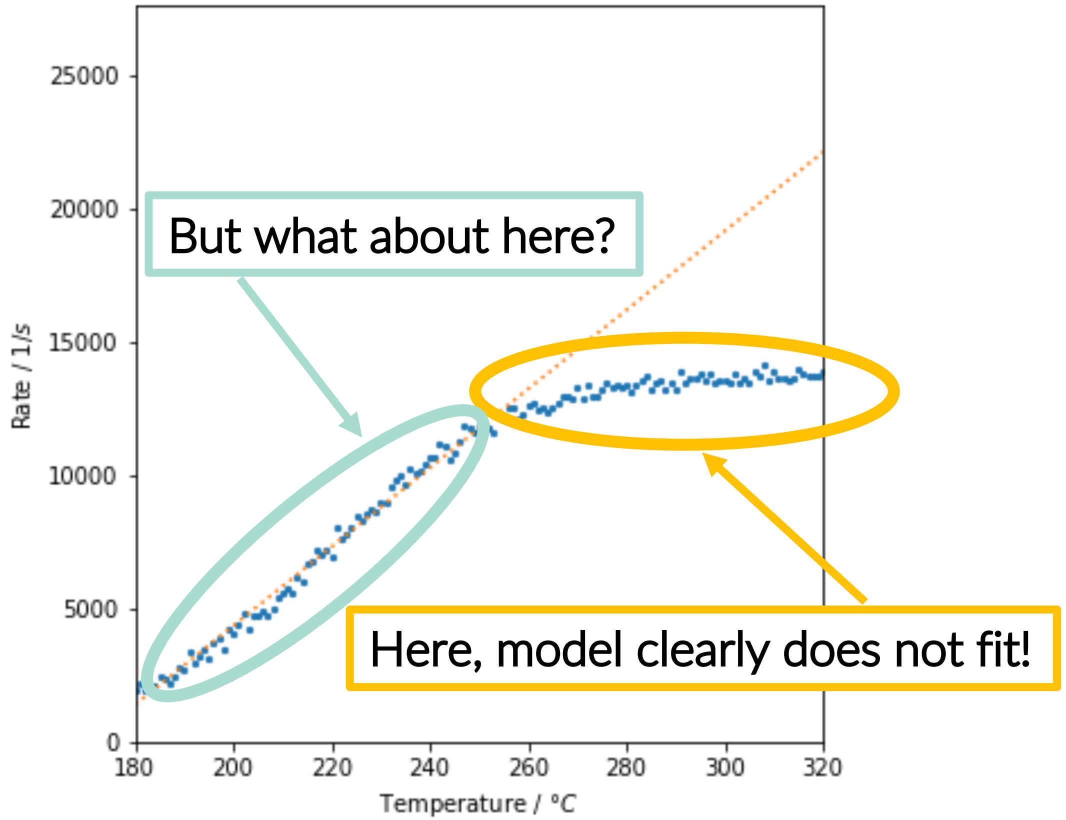
k-fold cross-validation



Rate measurements over temperature

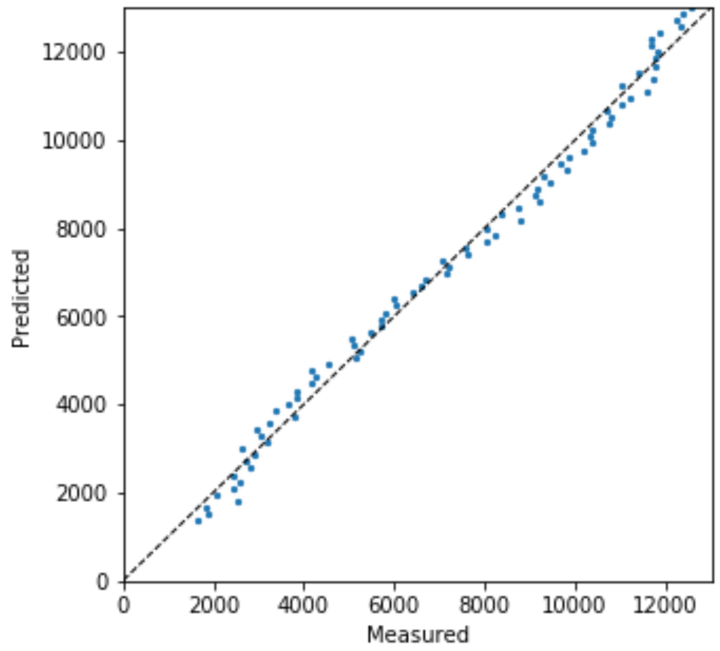
- Observations of rate over a range of temperatures

- "It almost looks linear!"
- R^2 is in the range of 0.99!



Techniques to visually inspect quality: parity plots.

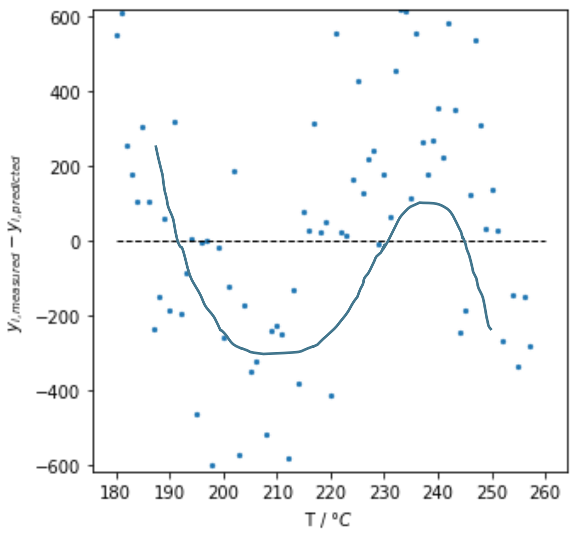
- Visualizes model quality transparently
  - Works for linear & non-linear models
- Builds on reangular plot
  - **Measured** value on **x**-axis
  - **Modeled** value on **y**-axis
- Visually inspect model quality
  - Ideal model should follow diagonal over whole validity range
- Mathematically: Residual analysis



Common tools: Variance / residual analysis

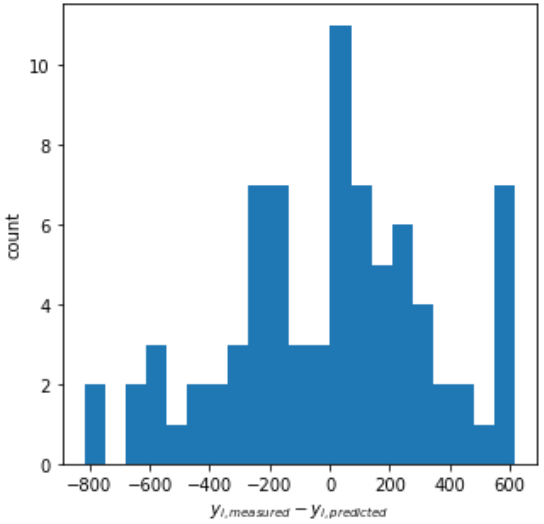
- Formalizes the parity plot approach
- Allows you to assess heteroscedacticity
  - Watch out for a slope in variance
  - Patterns like “waves”
- Histogram shape should match expected error model (e.g. Gaussian, Poisson)
- Additional methods that work best with train/test sets
  - Student’s t-test: compare variance of subsets
  - F-test: compare means / std of subsets
  - χ^2 test: statistical significance tests

All of those are methods to support verification and subsequent validation



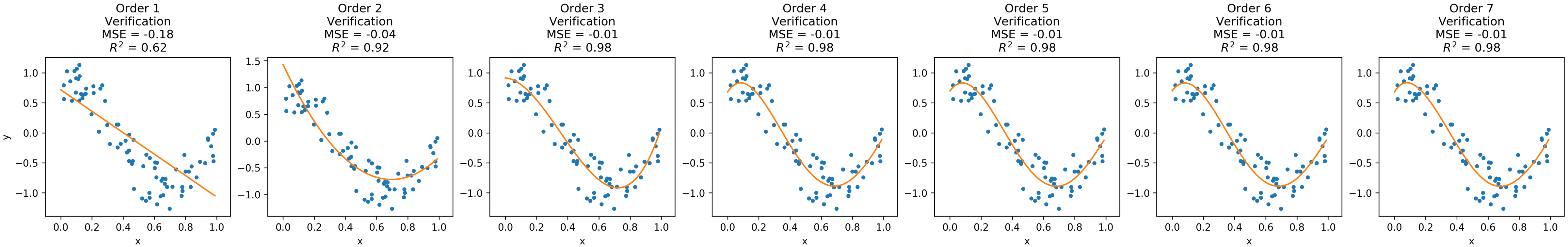
?

$$k = k_0 \exp\left(\frac{-E_A}{RT}\right)$$



Overfitting and underfitting: Finding the right number of paramters

- Based on just “data”, the functional relationship can only be inferred
- Let us try to fit this set of data with a n-th order polynomial

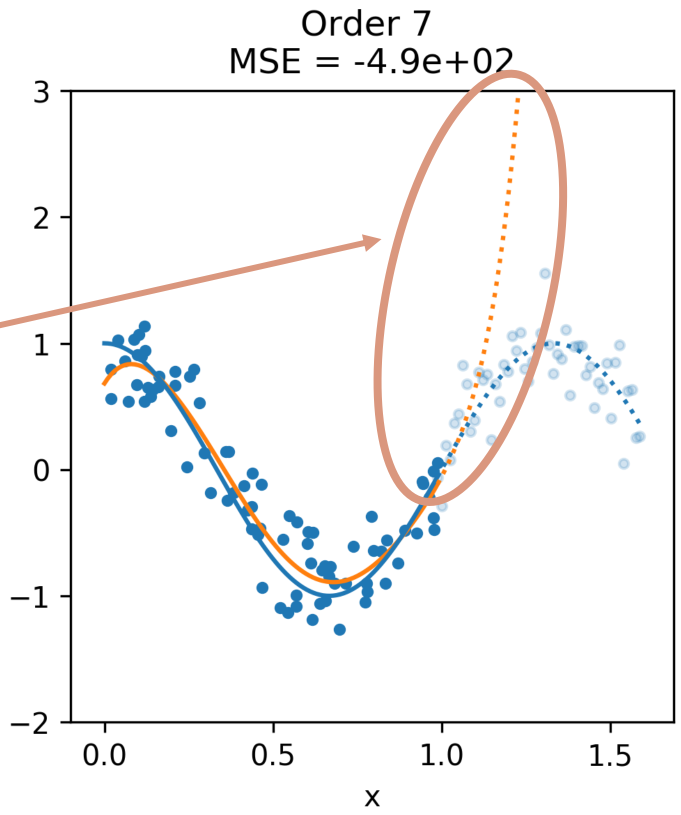


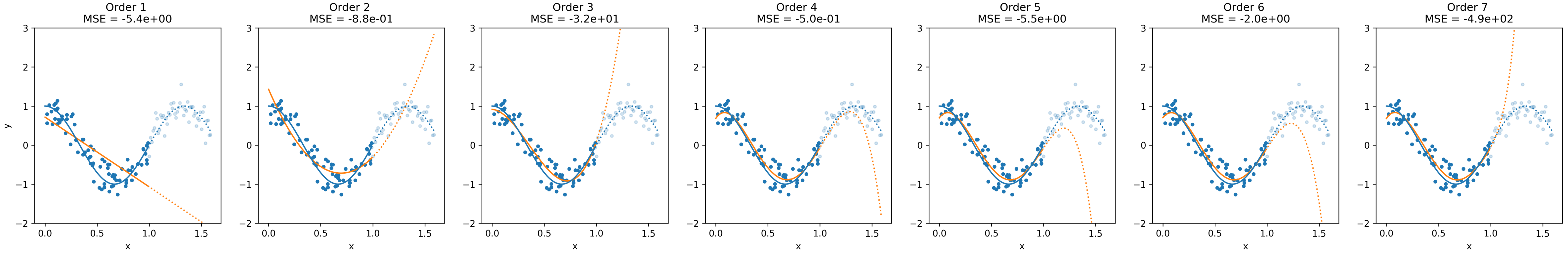
- MSE is drecreasing
- R2 is increasing, indicating a “better fit”

- Low order polynomial is not capturing the all the effects: Underfitting

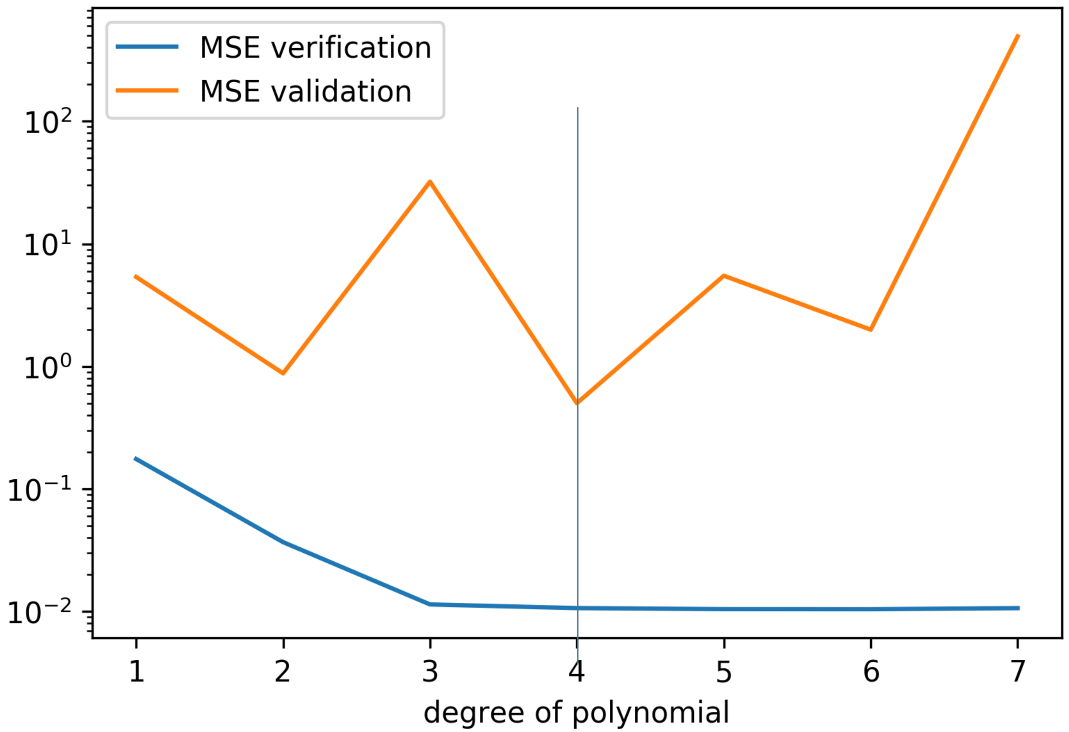
Checking outside of the verification range

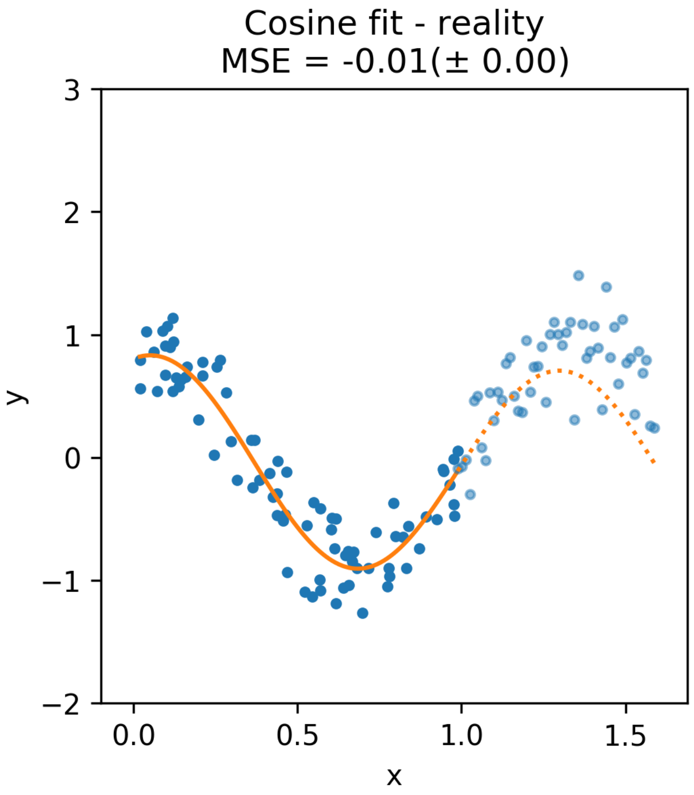
- Outside of the training range you can really see
- Starting from a certain point, the MSE really "takes off"
- This typically indicates overfitting





Analyze MSE for verification and validation to determine the "most suitable" – but this still may not be the "right one"



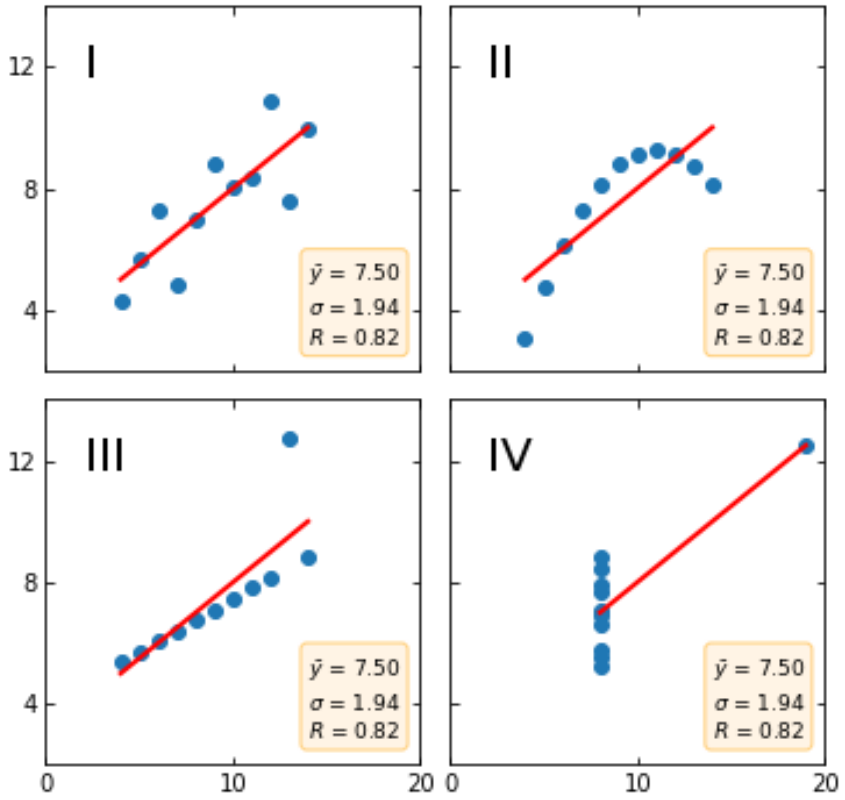


In this case, a 4th order seems to be the best compromise

Finding the ground truth in data-based models is often an iterative trial & error process

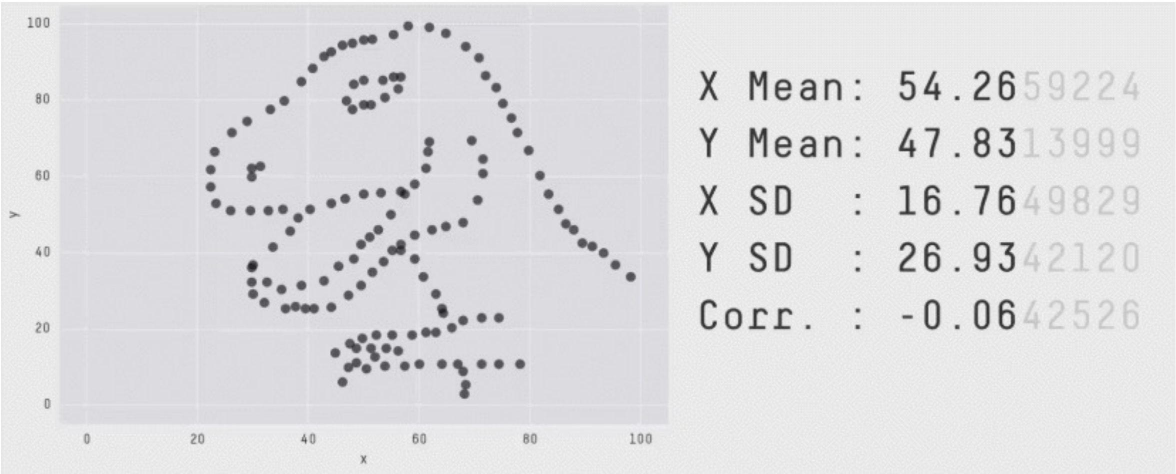
Do not trust a single statistical numbers alone!

- Anscombe's quartet



https://matplotlib.org/3.2.1/gallery/specialty_plots/anscombe.html

- Datasaurus dozen



https://www.autodeskresearch.com/publications/samestats

Summary on validation of data-based models

Common things we can do to assist validation:

- Use **data-sets** that are **artifact free** and from validated experiments ☺
- Carefully **verify** all model candidates using statistics
- Check for **under-/overfitting**
- **Cross-validate** models on available data
- Take care of **validity ranges** based on trained data
- Clearly **document assumptions** and boundary conditions

Fully validating a data-based model may result in commonly-accepted relationship!
(cf. first principle model / law)

Next steps in our journey

- Validation of classification models
- Fitting beyond least squares: Likelihood based fitting of noisy data
- Advanced goodness-of-fit tests
  - AIC: Akaike information criterion
  - Chi-squared test
  - Bayes information criterion
- Preparation of datasets for parameter estimations
- Model construction, selection & generation criteria

Validation & verification are almost “never-ending” tasks, unless you deal with a hard, first-principle model… and even then, you have to verify your measurement data!

Literature for further study

- Ross, S: Introduction to probability and statistics for engineers and scientists, 5th ed, Elsevier, 2014
- Raasch, J: Statistik für Verfahrenstechniker und Chemie-Ingenieure, 2010
- Bruce, A. & Bruce, P: Practical Statistics for Data Scientists, O'Reilly, 2017
- Strutz, T.: Data fitting & uncertainty, 2nd ed, Springer, 2016

Code & slides are at https://github.com/blackw1ng/data-validation-lecture

Any questions?