Sample lecture

Validation problems of data-based models

Sebastian Werner, Mar/30th 2021

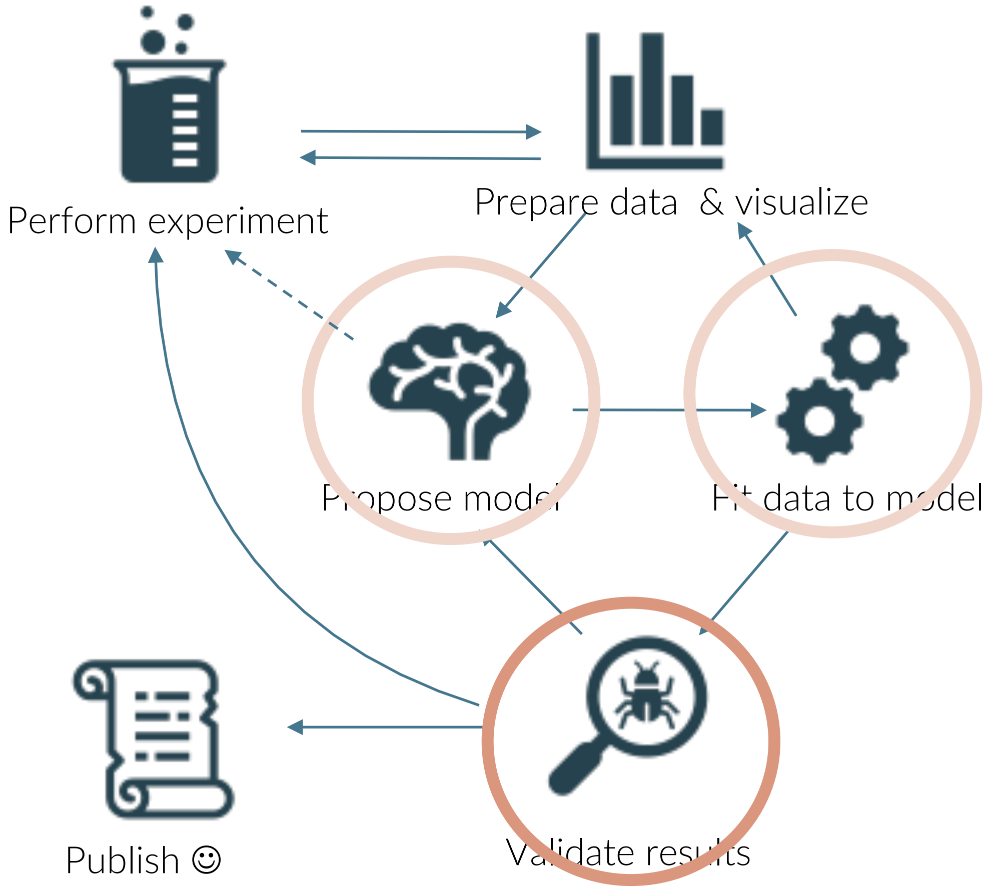
Code & slides are at https://github.com/blackw1ng/data-validation-lecture

Underlying principles of data-based models

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

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

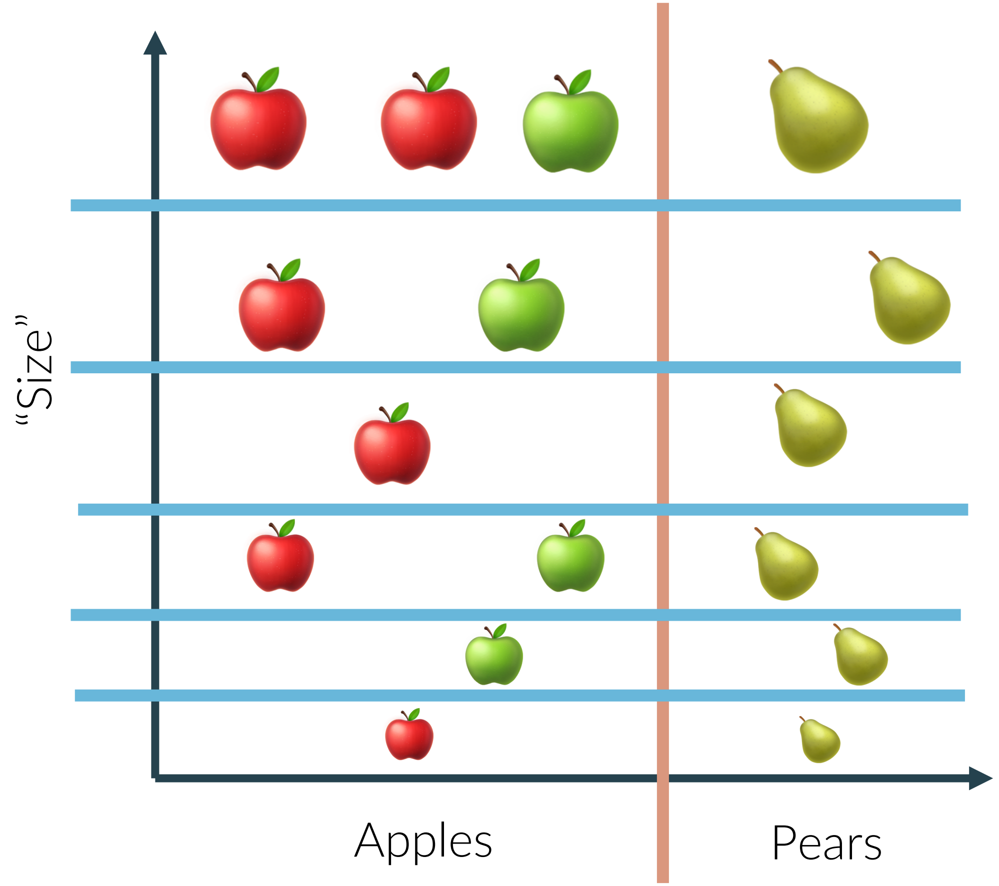
- Common challenge in chemistry & chemical engineering:
    - Fitting measurement data to a known model
    - Even then, validation is important
- Validation of a model implies a previous validation of experimental setup that generated this data!



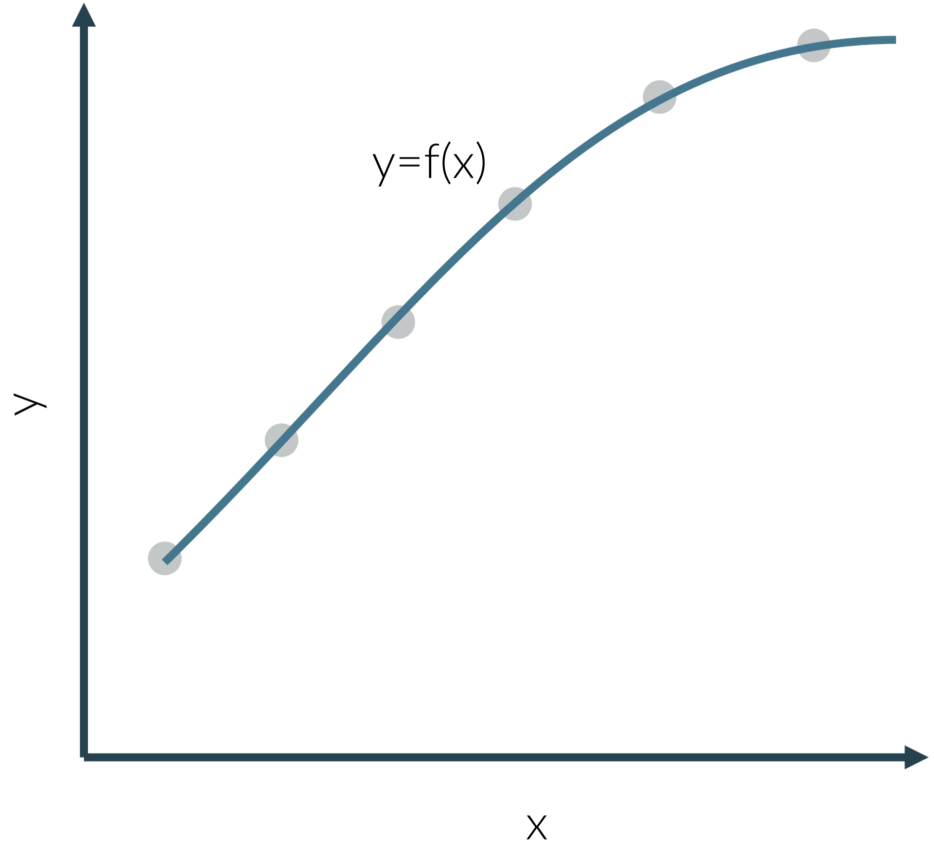
Modified CRISP-DM process for experimental data analysis

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

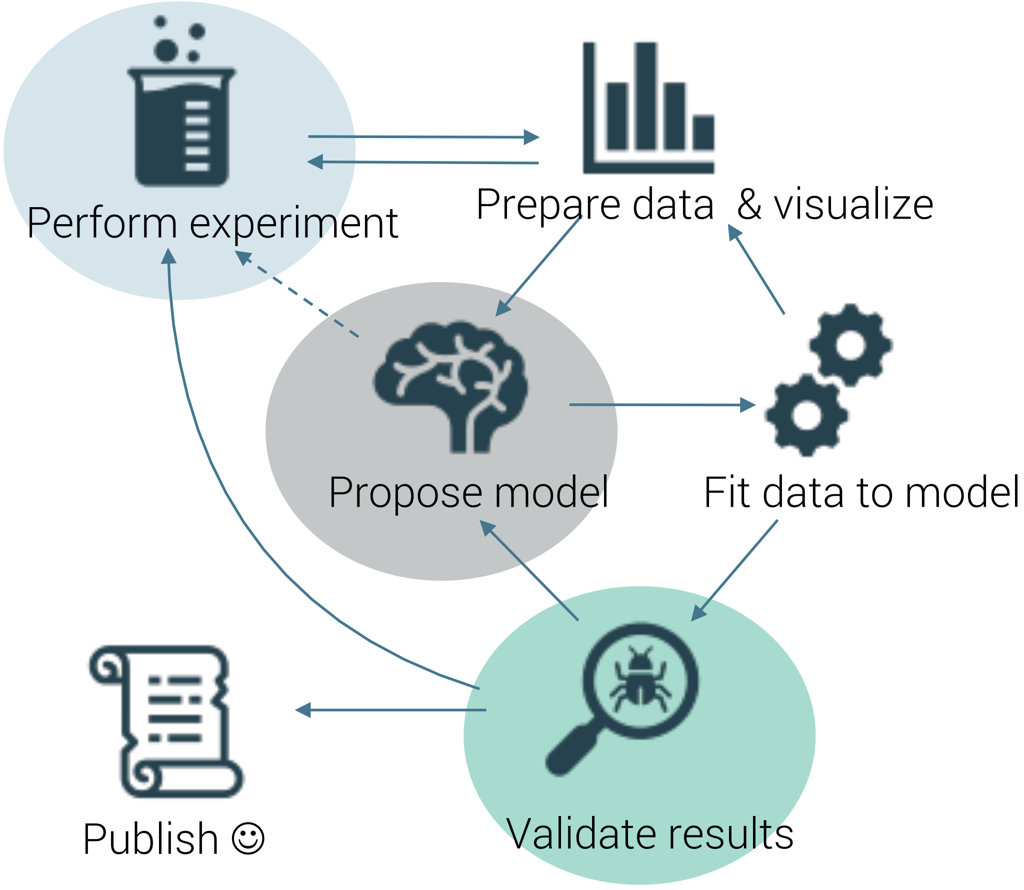
Models & validation: A chicken and egg problem ☺

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!

How can you do validation on a finite data set?

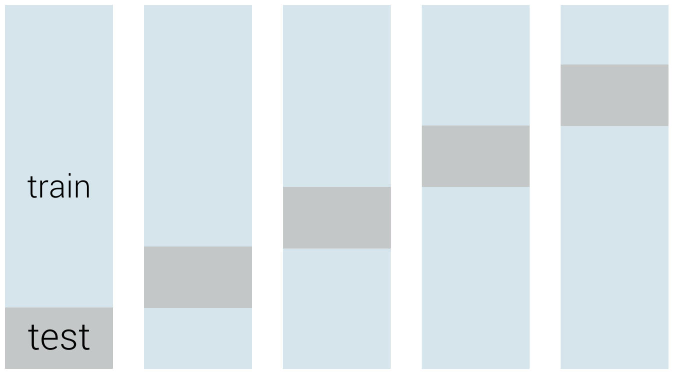
- Split available datasets!
- Take most of the data for "**training**" the model – and **verification**
- Spare some data to **test** the model & **validate** it
- This is specifically important for data-based models

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

train/test splitting

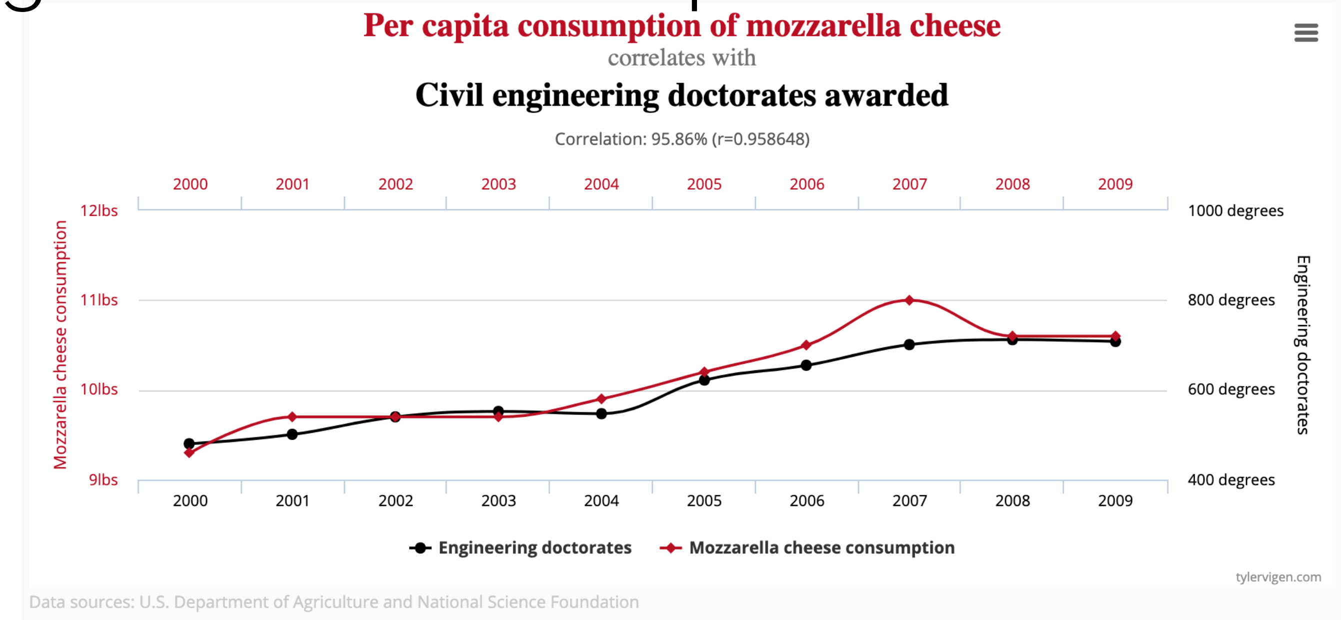
t	x	y	
13:01	23	2	train
13:02	23.5	5	
13:03	42	6	
13:04	21.2	2	
13:05	42	3	test
…			

k-fold cross-validation



Type & background of model impact validation strategy

Chemie beispiel



Logical type

"Can they distinguish between cause and effect?" (correlation vs. causation)

- **Causal**: Clear relationship between input and output. Can be inverted.
- **Descriptive**: cannot give conclusive evidence about cause and effect.

Model background

- **Hard models**: Based on well-established first principle knowledge.
- **Soft models**: Constructed without a-priori knowledge. **Purely data-based**! Often many factors, and complex systems.

How can we validate models?

Quality measures for regressions

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

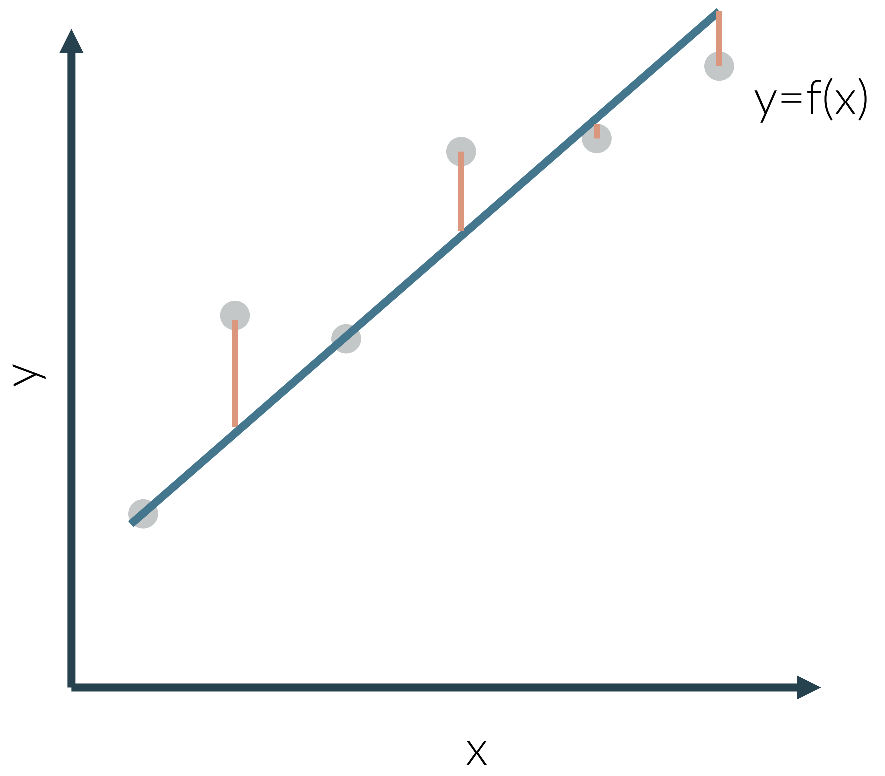
$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

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

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

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

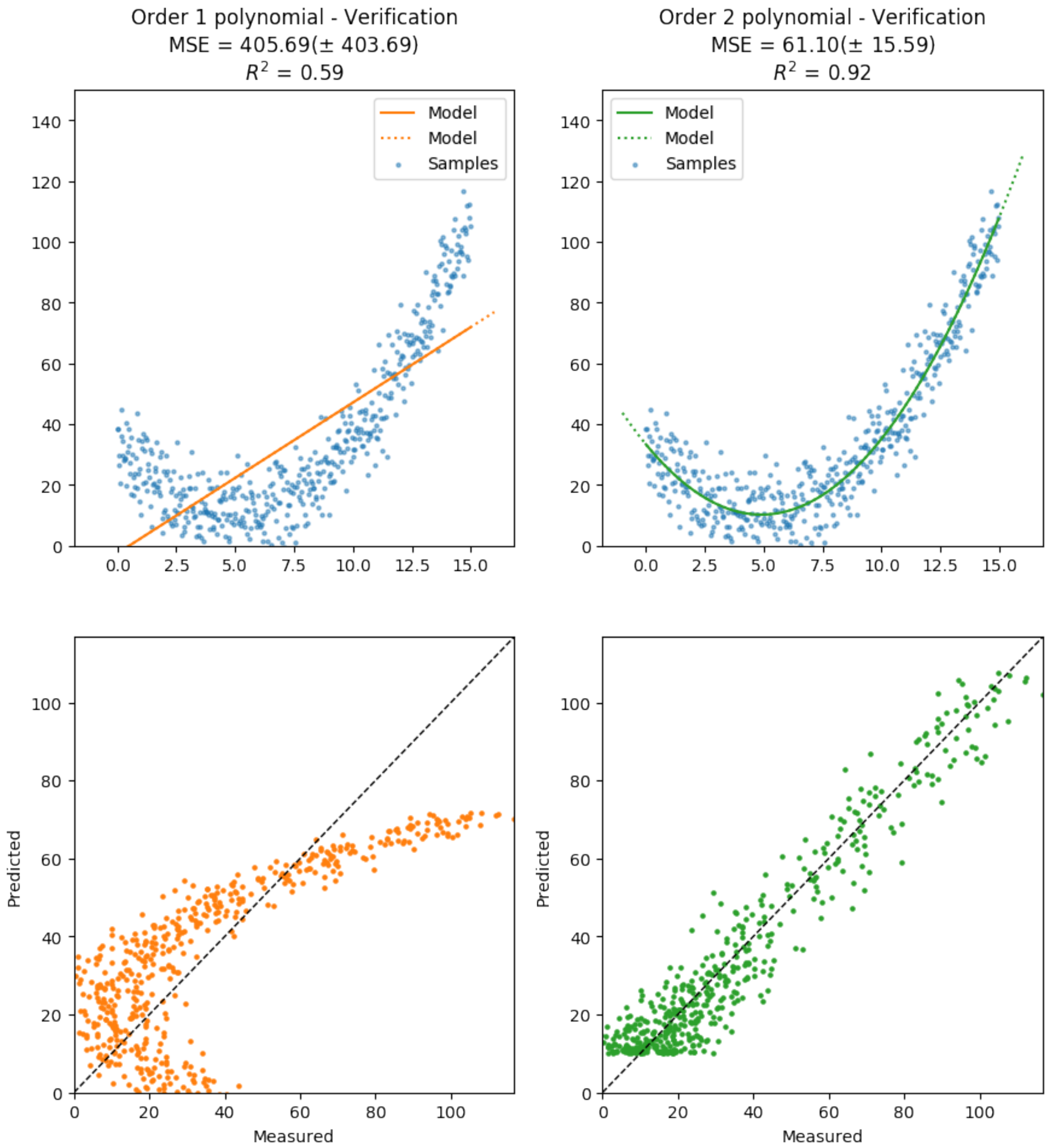
$$RMSE = \sqrt{\frac{1}{N}\sum_i^N (y_i - f_i)^2}$$



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

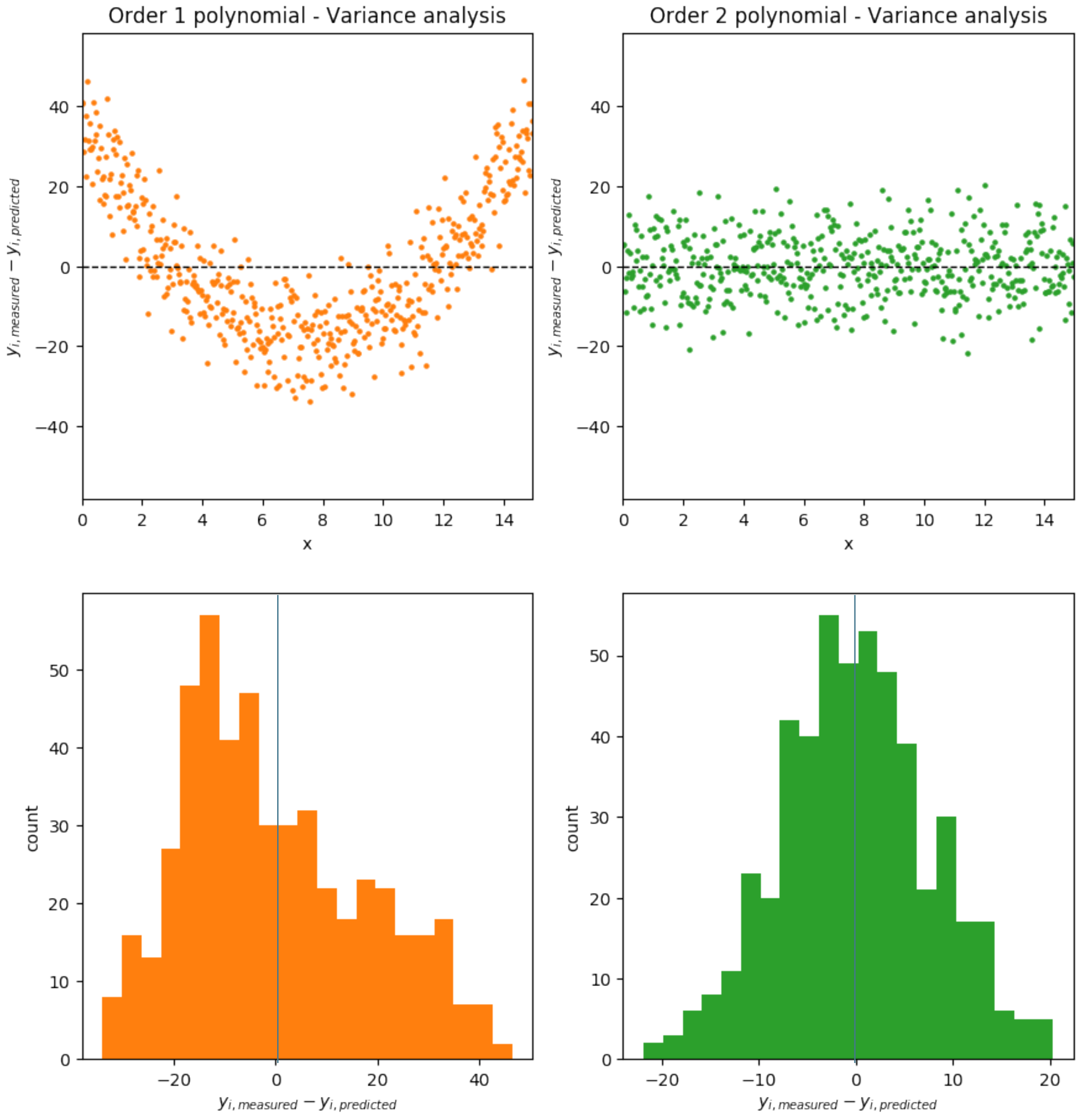
Techniques to visually inspect quality: parity plots.

- 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



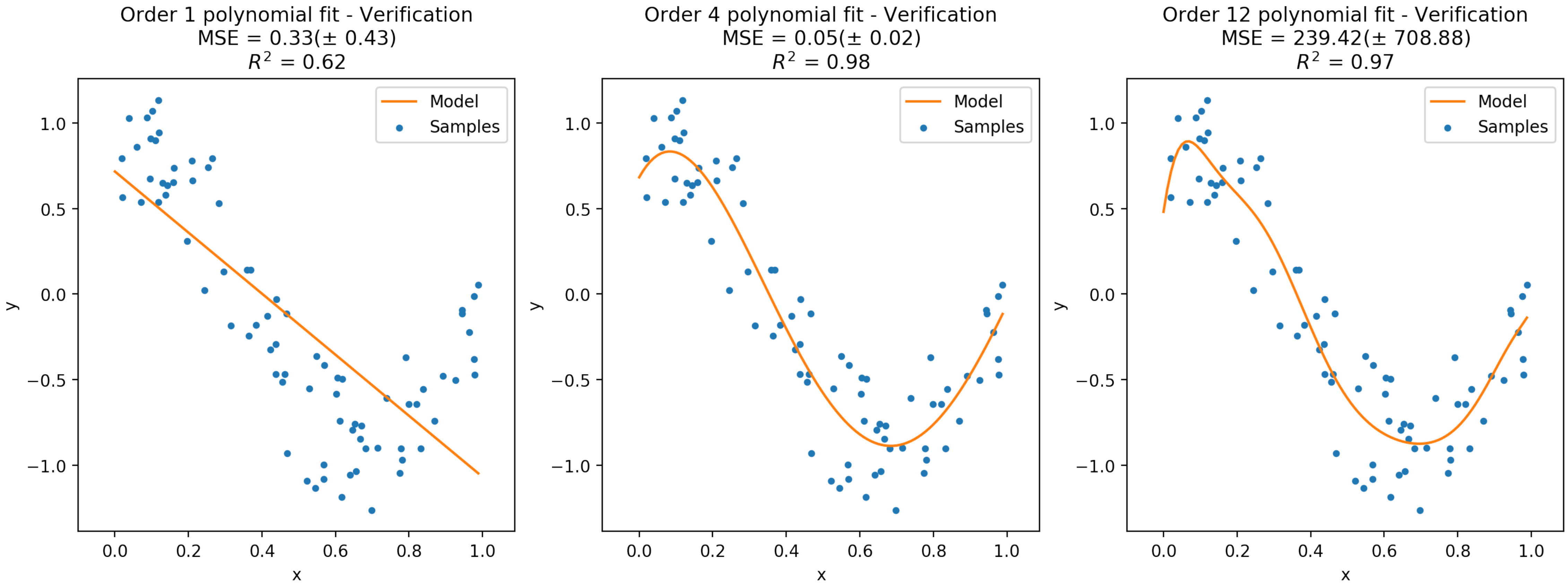
Common tools: Variance analysis

- Formalizes the parity plot approach
- Allows you to assess heteroscedacticity
  - Watch out for a slope in variance
- 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



Pitfalls: Verify models for overfitting and underfitting

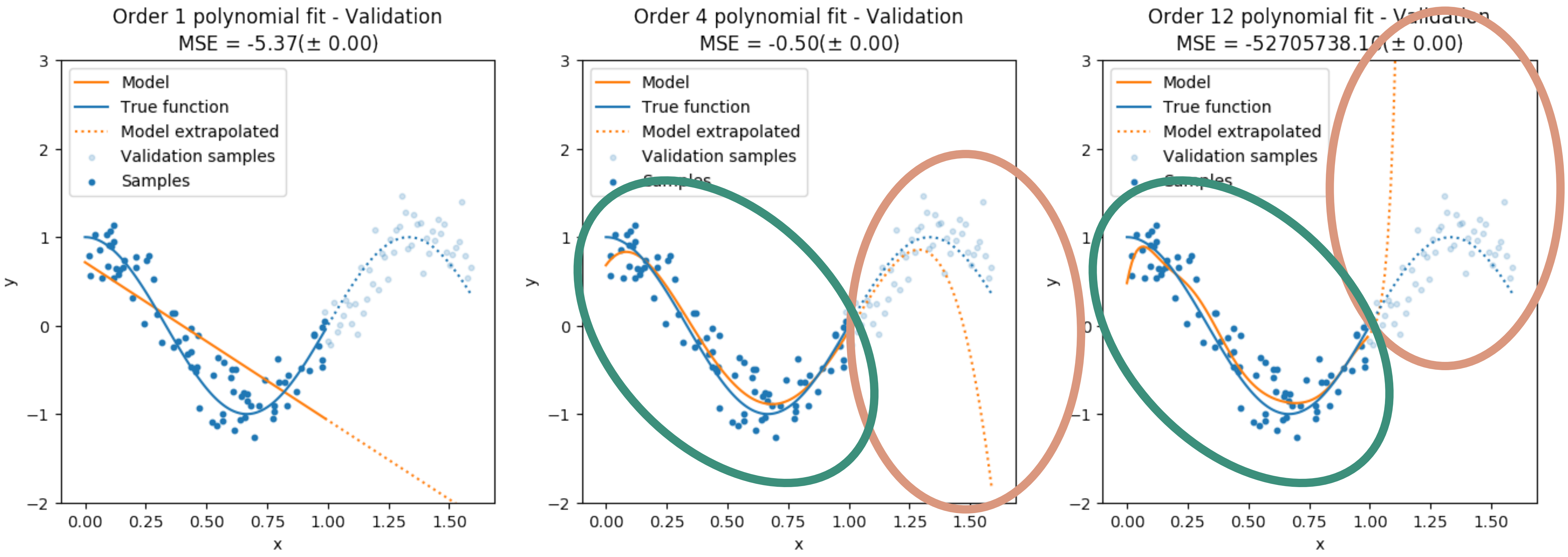
- Based on just “data”, the functional relationship can only be inferred



- Model with 4th order polynomial seem to check out well based on MSE & R^2!

Overfitting / Underfitting become more obvious in validation

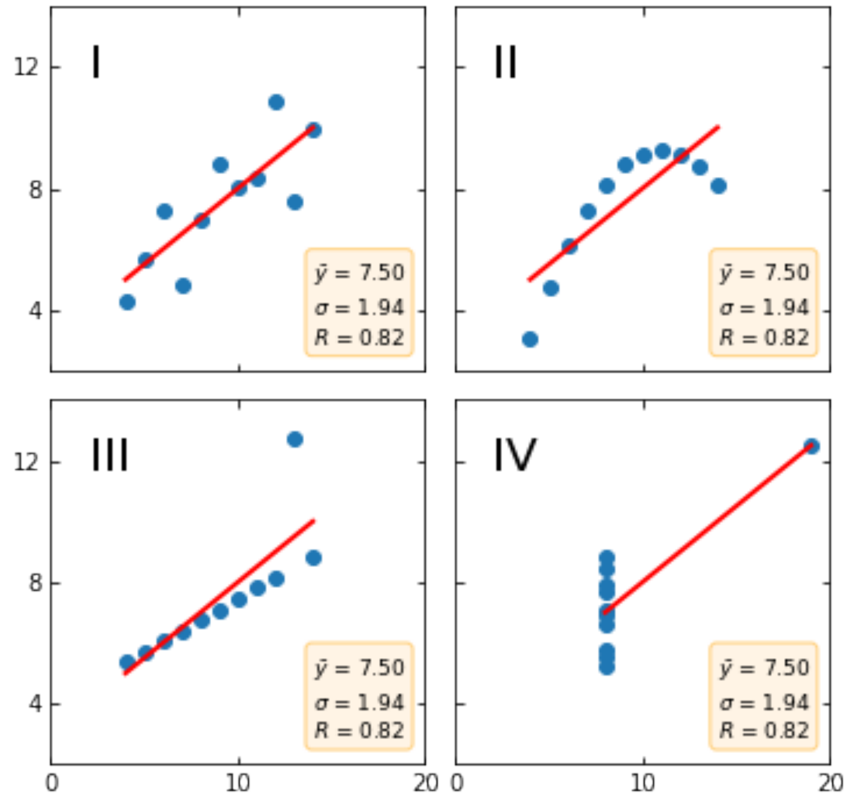
- Left graph is “underfitting”- and the other two also do not fit well.



Ideally this is combined with k-fold cross-validation!

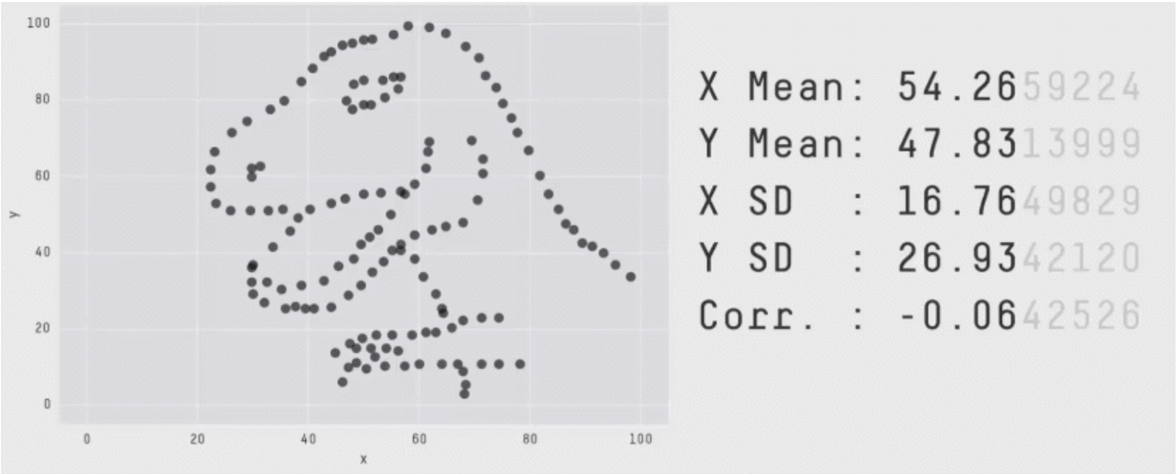
Do not trust a single statistical numbers alone!

- 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

Data-based models, constructed without prior system knowledge have **several pitfalls**:

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

Common things we can do to assist validation

- Use **data-sets** that are **artifact free** and from validated experiments ☺
- **Cross-validate** models
- Take care of **validity ranges** based on trained data
- Check for **under-/overfitting**
- Clearly **document assumptions** are 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

- 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

Thank you very much for your attention.

Any questions?