Data science for business

Sebastian Sauer

2019-05-14

Five Questions

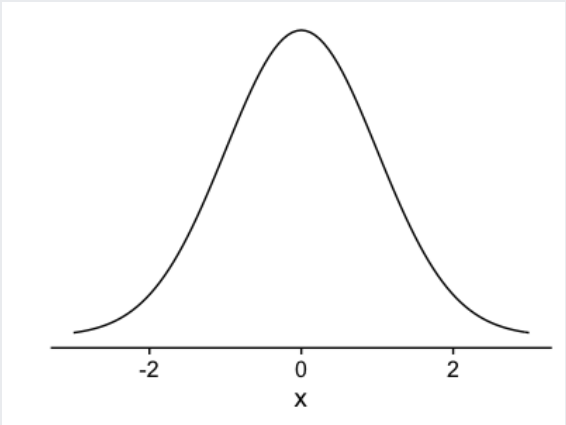
on the use of data science for business

1. What's the meaning of *data science*, *machine learning*, and all these fancy terms?
2. What's the best model out there?
3. How do I know my model is doing good or bad?
4. Can you give me a cook book for data science?
5. ~~What are all the core concepts of the field~~?

1. What's the meaning of *data science*, *machine learning*, and all these fancy terms?

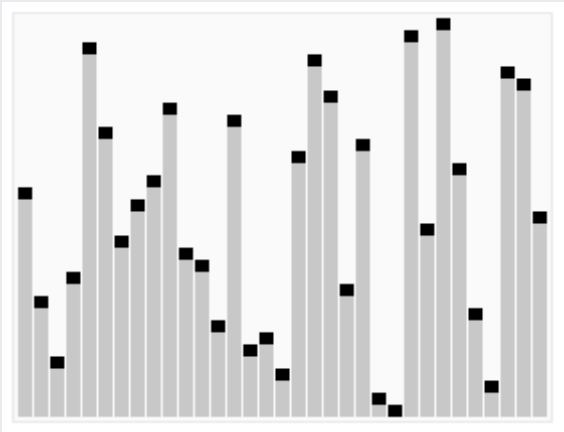
statistical models:

probability theory



machine learning:

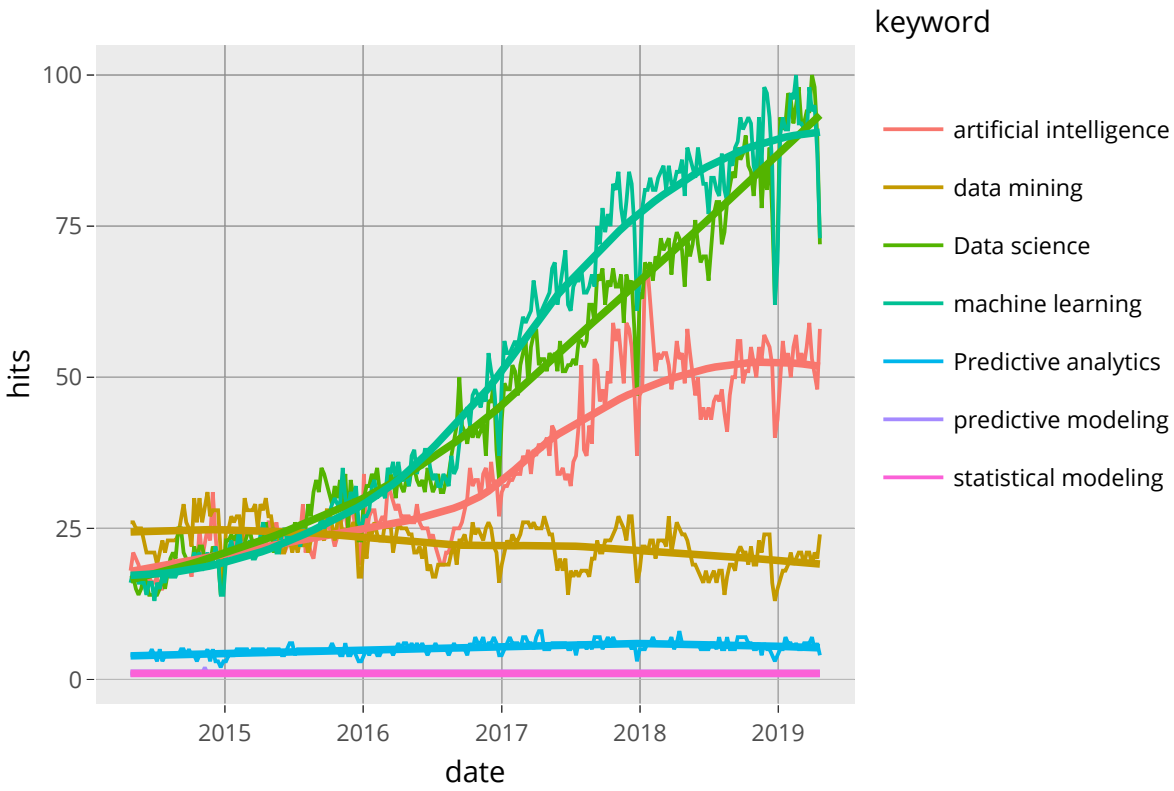
algorithmic models



Source: Wikipedia by en:User:RolandH

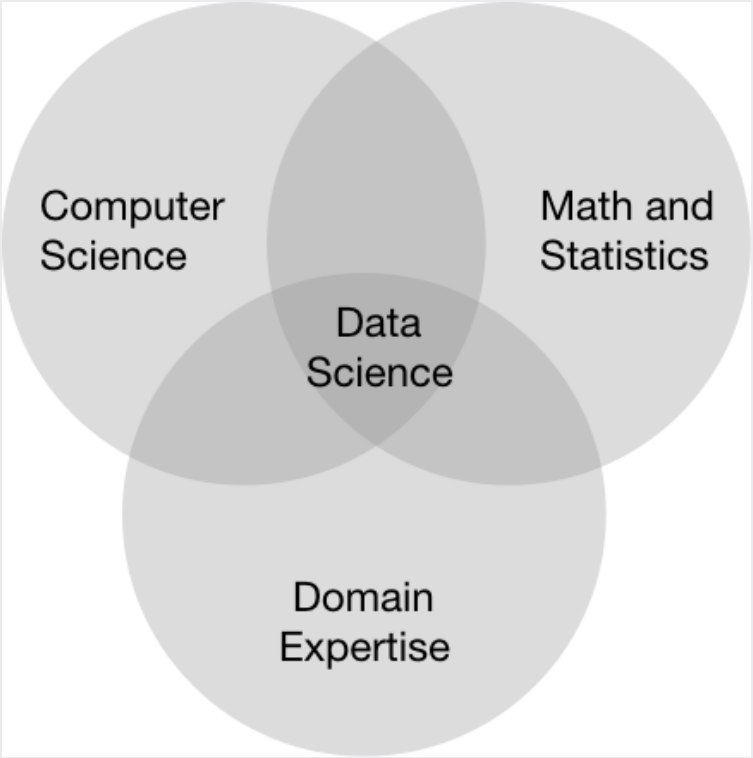
'data science' is a popular term

Google Trends (2019-04-32) of data analysis jargon



What's data science?

Depends on whom you ask.

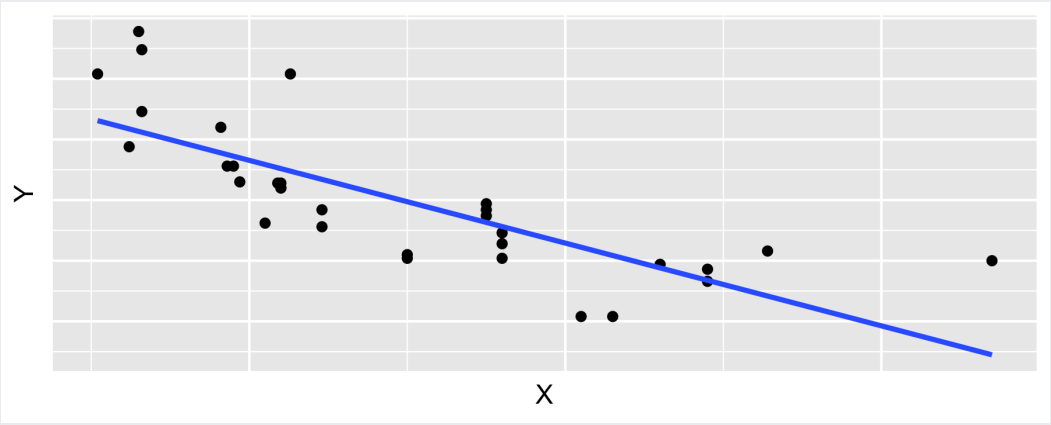


Common theme

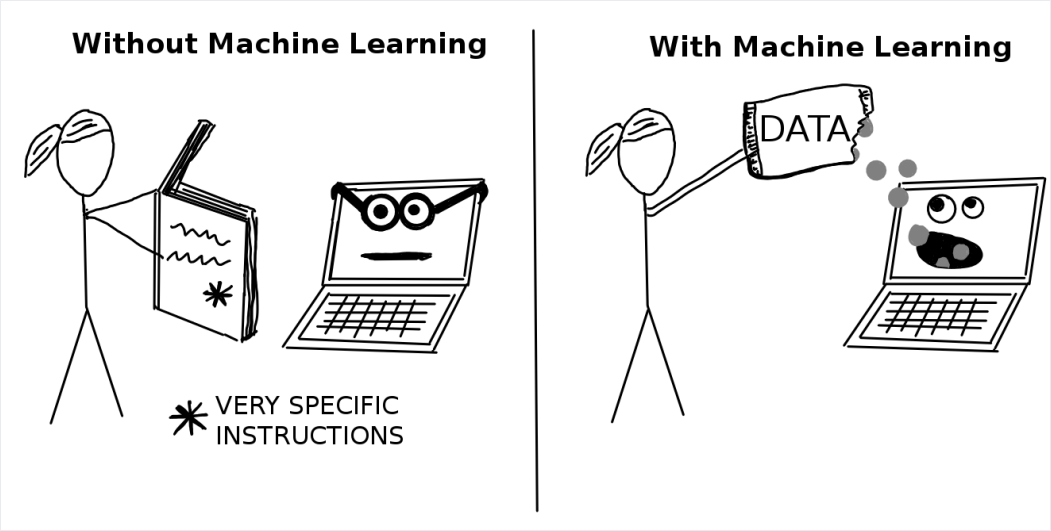
Art and science of learning from data

$$Y = f(X) + \epsilon$$

$$\hat{Y} = \hat{f}(X)$$



Machine learning: Feed the computer data, not rules



Source: Molnar, C. (2019). Interpretable Machine Learning [ePub Book]. Morrisville, NC: Christoph Molnar.

2. What's the best model out there?

A lot of models out there

package caret

```
getModelInfo() %>%
  names() %>%
  length()
## [1] 238
```

Show 5 entries

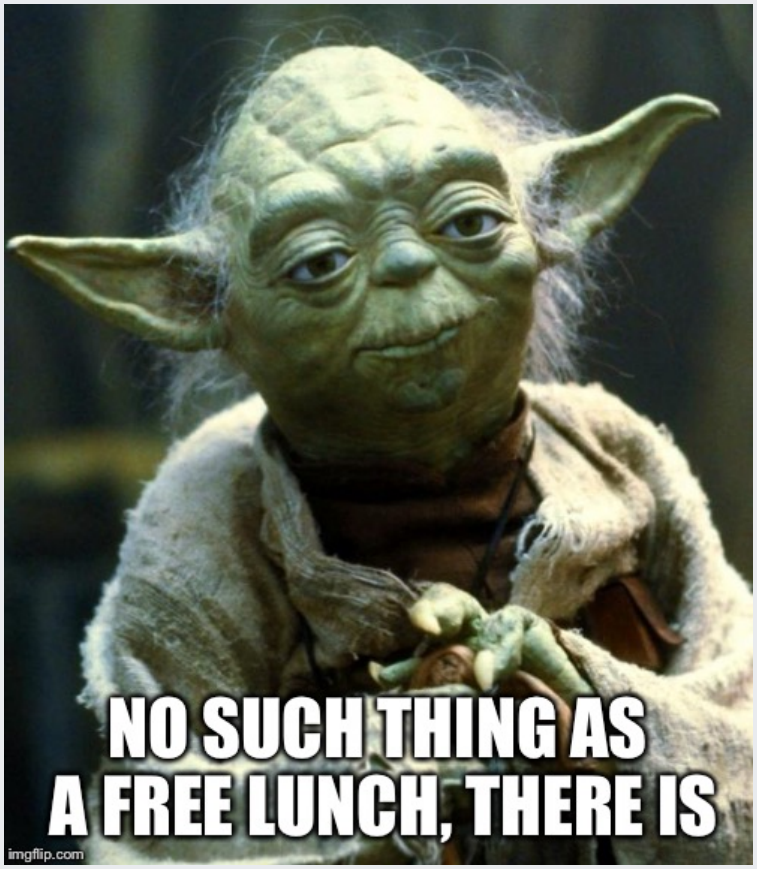
Search:

	name	value
1	1	ada
2	2	AdaBag
3	3	AdaBoost.M1
4	4	adaboost
5	5	amdai

Showing 1 to 5 of 238 entries

Previous 1 2 3 4 5 ... 48 Next

Wait, tell me which model is *best*



There is no single best model

Black box models

- Random forests
- Support vector machines
- Neural networks
- ...

less interpretable

more accurate (at times)

less robust

"White box" models

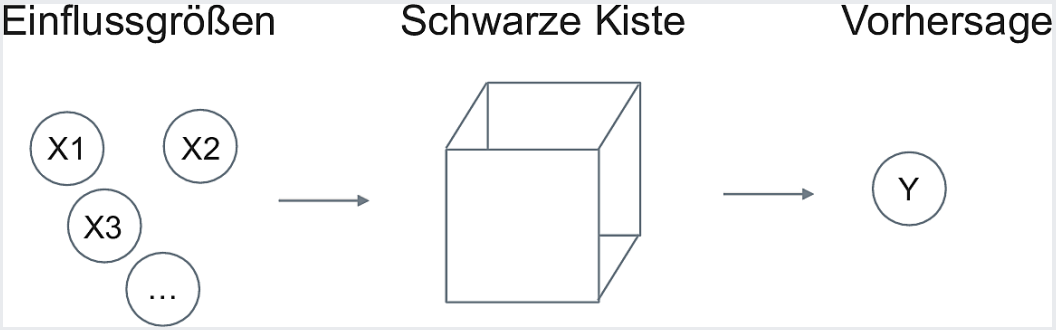
- Linear regression
- k-nearest neighbours
- Decision trees
- ...

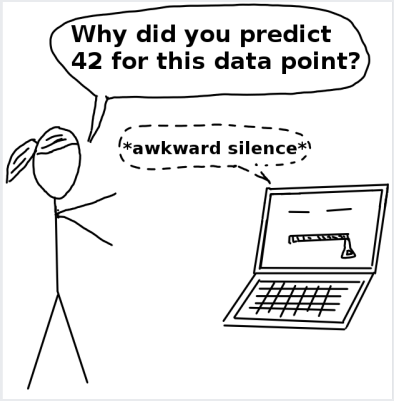
more interpretable

less accurate (at times)

more robust

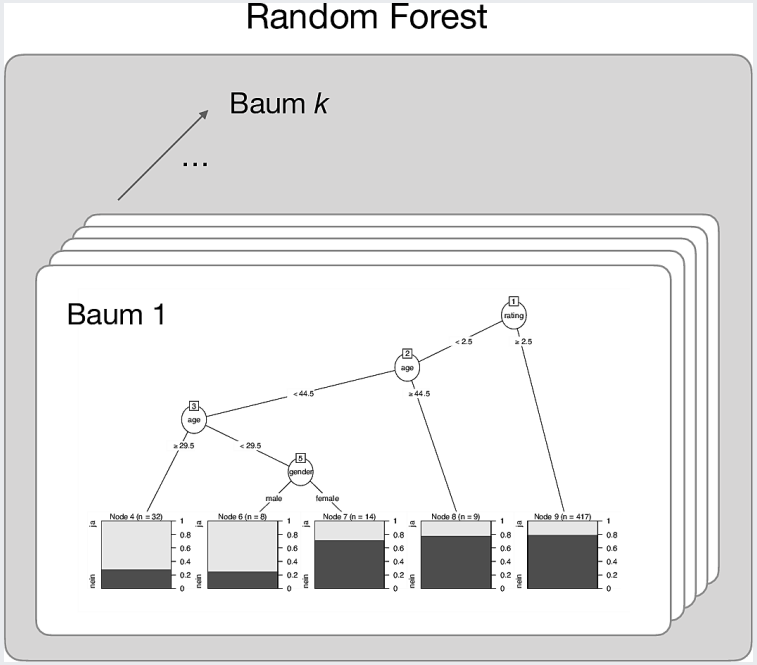
Blackbox models do not explain





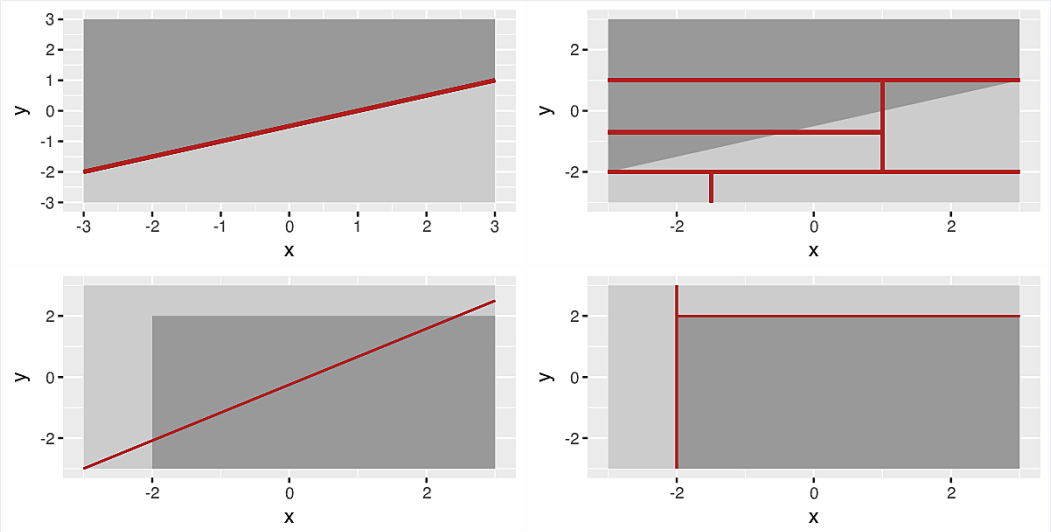
Source: Molnar, C. (2019). Interpretable Machine Learning [ePub Book]. Morrisville, NC: Christoph Molnar.

Ensemble learners show a good track record



Source: Sauer, S. (2018). Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren. Wiesbaden: Springer.

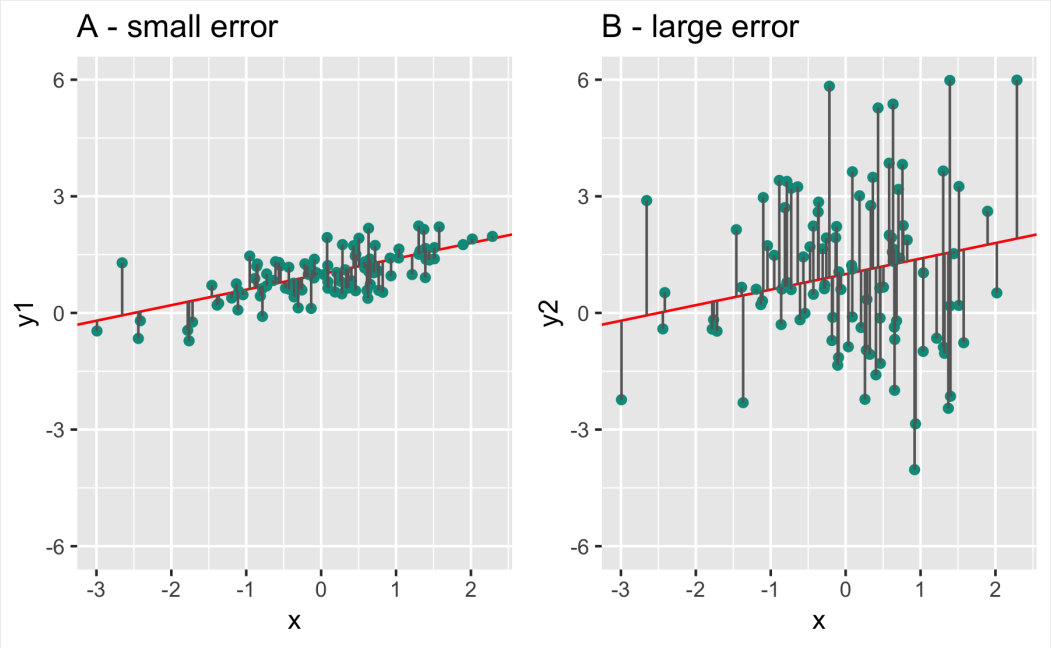
The fit of a model depends on eg the linearity of associations



Oben: Eine zweidimensionale Klassifizierung mit linearen Klassengrenzen;
eine klassische Regression ist besser darauf ausgelegt als ein Entscheidungsbaum.
Unten: In diesem Fall sind die Klassengrenzen nicht linear;
ein Entscheidungsbaum kann damit besser umgehen als eine Regression

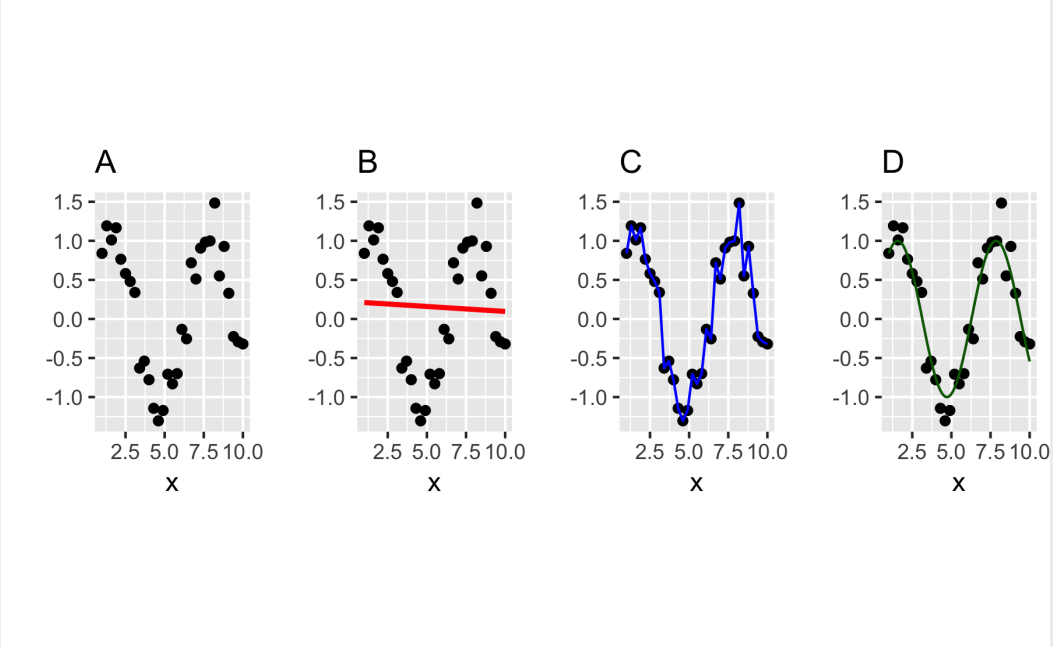
3. How do I know my model is doing good or bad?

Short answer: The less error, the better the model



Wait ...

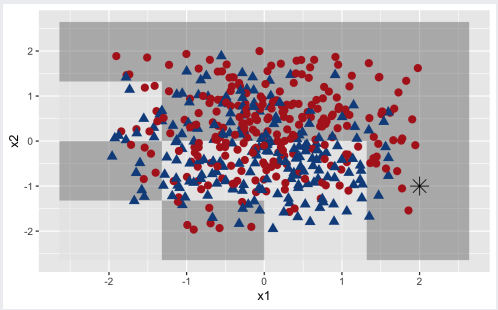
Which model do you prefer?



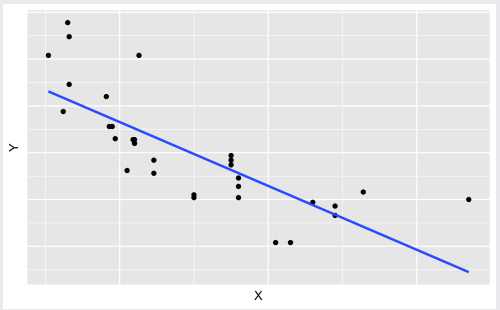
4. Can you give me a cook book for data science?

Step 1: Choose your model(s)

Classify stuff

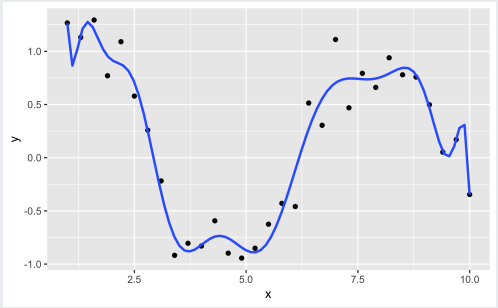


Estimate stuff

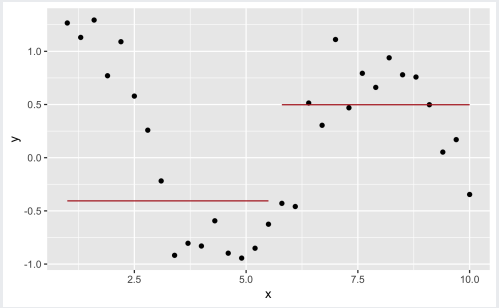


Step 2: Build model fed on historical data

Overfitting

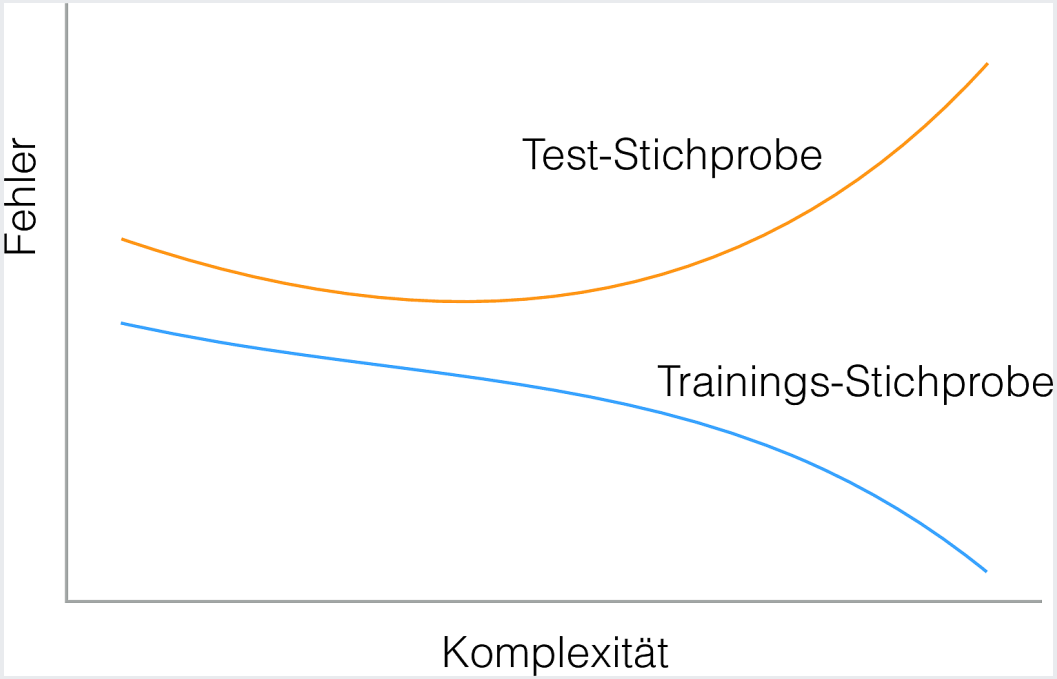


Underfitting

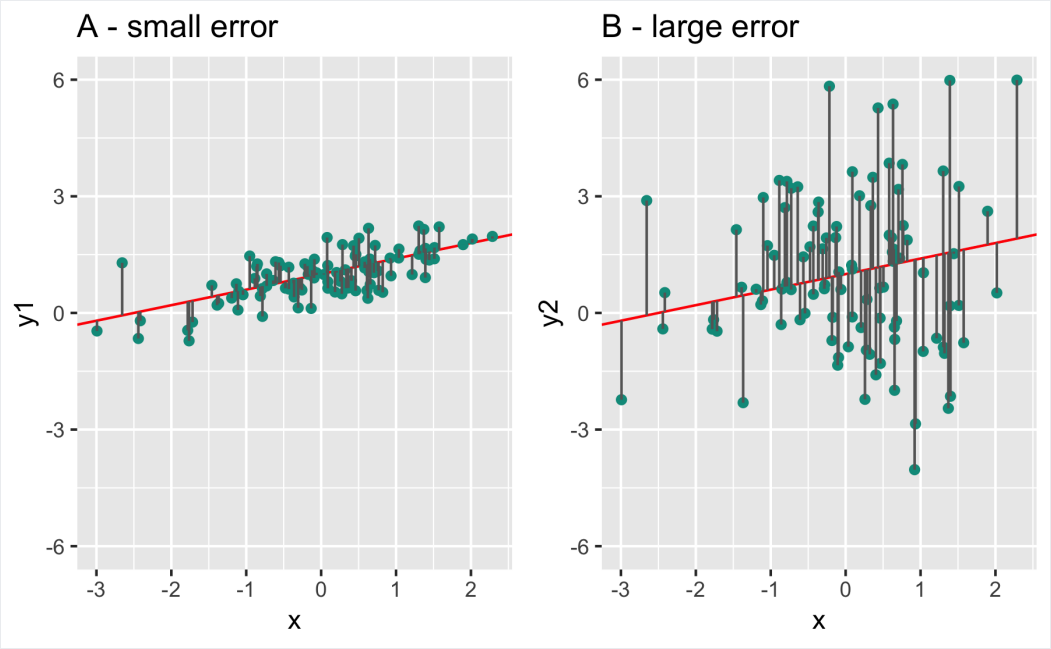


Step 3: Predict the future

Run the model on *new* data



Step 4: Evaluate the model



Here's one way how to get going

Source: https://www.williamrchase.com/slides/slide_img/throw_into_pool.gif

Some literature explaining core concepts of data science

Grolemund, G., & Wickham, H. (2016). R for Data Science. Retrieved from https://books.google.de/books?id=aZRYrgEACAAJ

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York City, NY: Springer.

Sauer, S. (2019). Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren (1. Auflage 2019). Wiesbaden: Springer.

Sebastian Sauer

sebastiansauer

https://data-se.netlify.com/

sebastian.sauer@data-divers.com

sauer_sebastian

Get slides here: https://data-se.netlify.com/slides/afd_ecda2019/afd-modeling-ECDA-2019.pdf

Reproducibility

- Versions of employed software as of 2019-05-02, running this OS: macOS Mojave 10.14.4.
- Built with R, R version 3.5.3 (2019-03-11), RStudio 1.2.1335, xaringan, on the shoulders of giants
- Source Code: XXX
- Icons are from FontAwesome, licenced under CC-BY-4 (details)
- R-Packages used: assertthat_0.2.1, backports_1.1.4, broom_0.5.2, Cairo_1.5-10, caret_6.0-82, cellranger_1.1.0, class_7.3-15, cli_1.1.0, codetools_0.2-16, colorspace_1.4-1, crayon_1.3.4, crosstalk_1.0.0, data.table_1.12.2, digest_0.6.18, dplyr_0.8.0.1, DT_0.5, evaluate_0.13, forcats_0.4.0, foreach_1.4.4, generics_0.0.2, ggplot2_3.1.1, ggrepel_0.8.0, glue_1.3.1.9000, gower_0.2.0, gridExtra_2.3, gtable_0.3.0, gtrendsR_1.4.2, haven_2.1.0, hms_0.4.2, htmltools_0.3.6, htmlwidgets_1.3, httpuv_1.5.1, httr_1.4.0, icon_0.1.0, ipred_0.9-8, iterators_1.0.10, jsonlite_1.6, knitr_1.22, labeling_0.3, later_0.8.0, lattice_0.20-38, lava_1.6.5, lazyeval_0.2.2, lubridate_1.7.4, magrittr_1.5, MASS_7.3-51.1, Matrix_1.2-15, mime_0.6, ModelMetrics_1.2.2, modelr_0.1.4, munsell_0.5.0, nlme_3.1-137, nnet_7.3-12, pillar_1.3.1, pkgconfig_2.0.2, plotly_4.9.0, plyr_1.8.4, prodlim_2018.04.18, promises_1.0.1, purrr_0.3.2, R6_2.4.0, Rcpp_1.0.1, readr_1.3.1, readxl_1.3.1, recipes_0.1.5, reshape2_1.4.3, rlang_0.3.4, rmarkdown_1.12.6, rpart_4.1-13, rprojroot_1.3-2, rstudioapi_0.10, rvest_0.3.3, scales_1.0.0, sessioninfo_1.1.1.9000, shiny_1.3.1, stringi_1.4.3, stringr_1.4.0, survival_2.43-3, tibble_2.1.1, tidyr_0.8.3, tidyselect_0.2.5, tidyverse_1.2.1, timeDate_3043.102, viridisLite_0.3.0, withr_2.1.2, xaringan_0.9, xaringanthemer_0.2.0, xfun_0.6, xml2_1.2.0, xtable_1.8-3, yaml_2.2.0