ABSCHNITT 5

DIE DATA SCIENCE PIPELINE

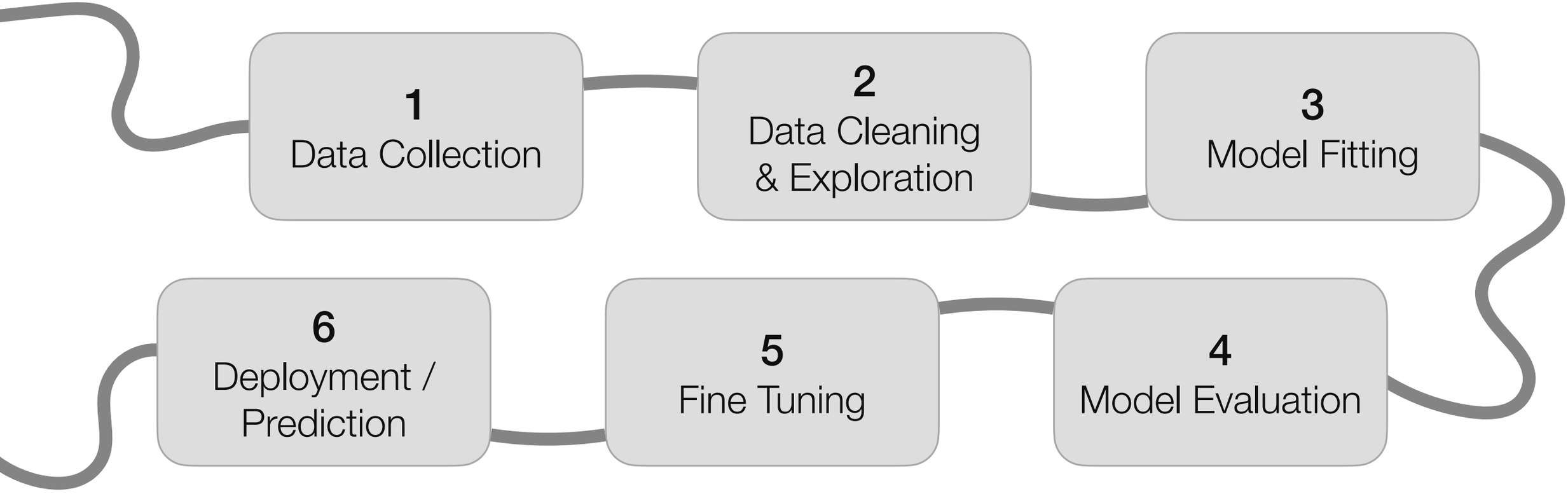
EIN REZEPT FÜR MACHINE LEARNING PROJEKTE

Alina Herderich

WAS IST EINE DATA SCIENCE PIPELINE?

Eine Data Science Pipeline umfasst die prototypischen **Schritte**, die es braucht, um ein **Machine Learning** Modell erfolgreich zu **implementieren**.

EINE DATA SCIENCE PIPELINE UMFASST (UNGEFÄHR) SECHS SCHRITTE:



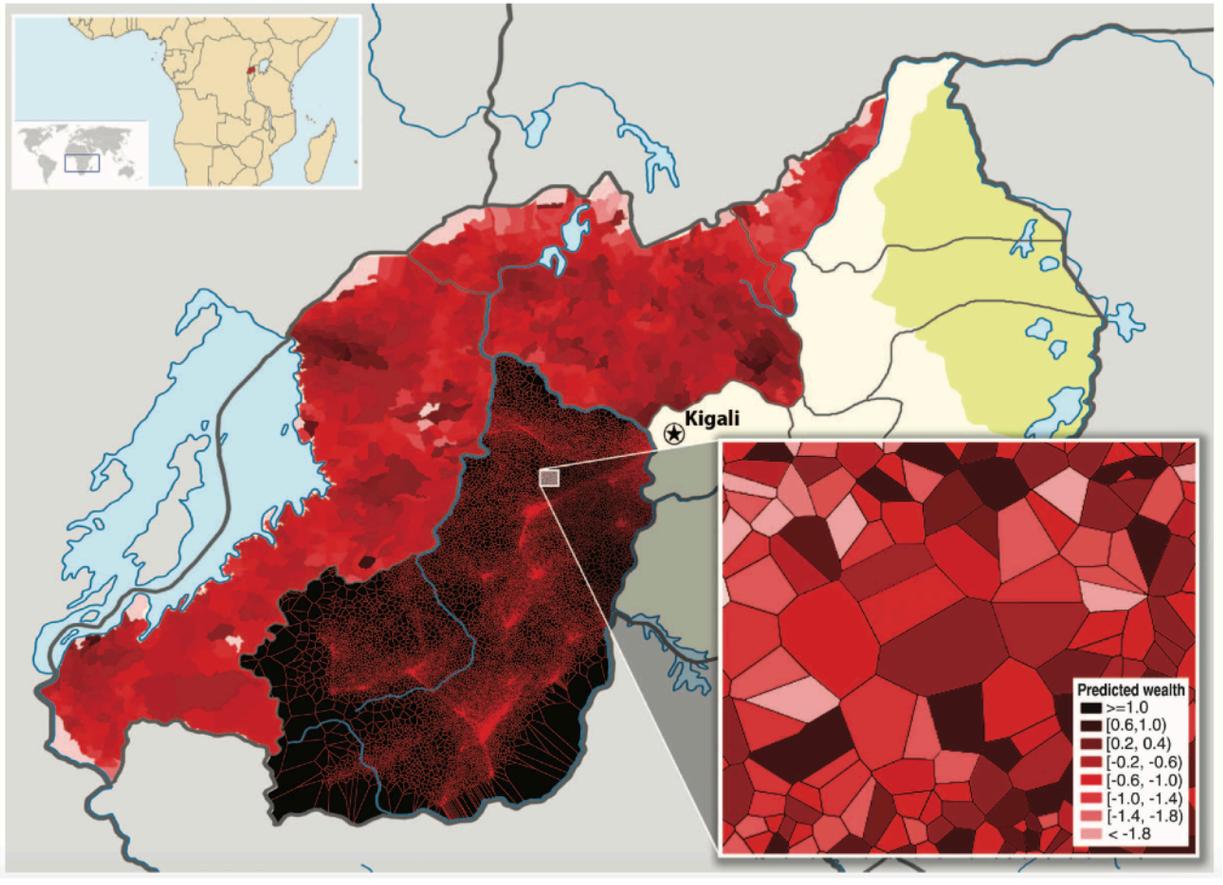
1 // DATA COLLECTION

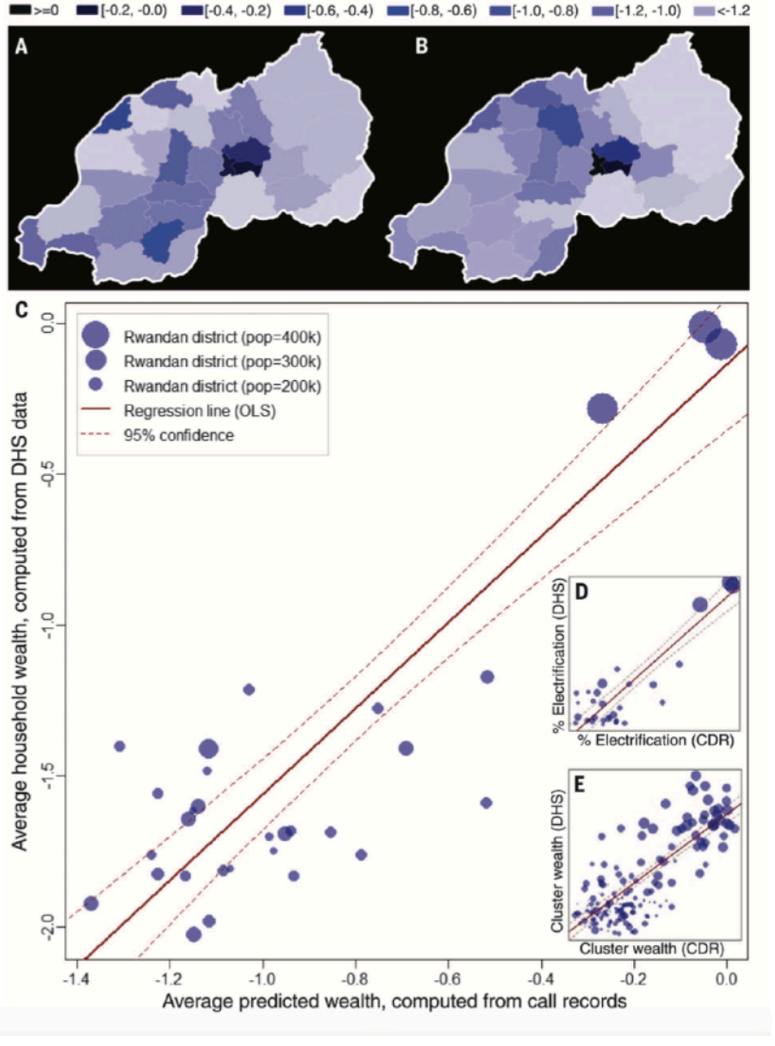
custom-made

ready-made

(Salganik, 2018)

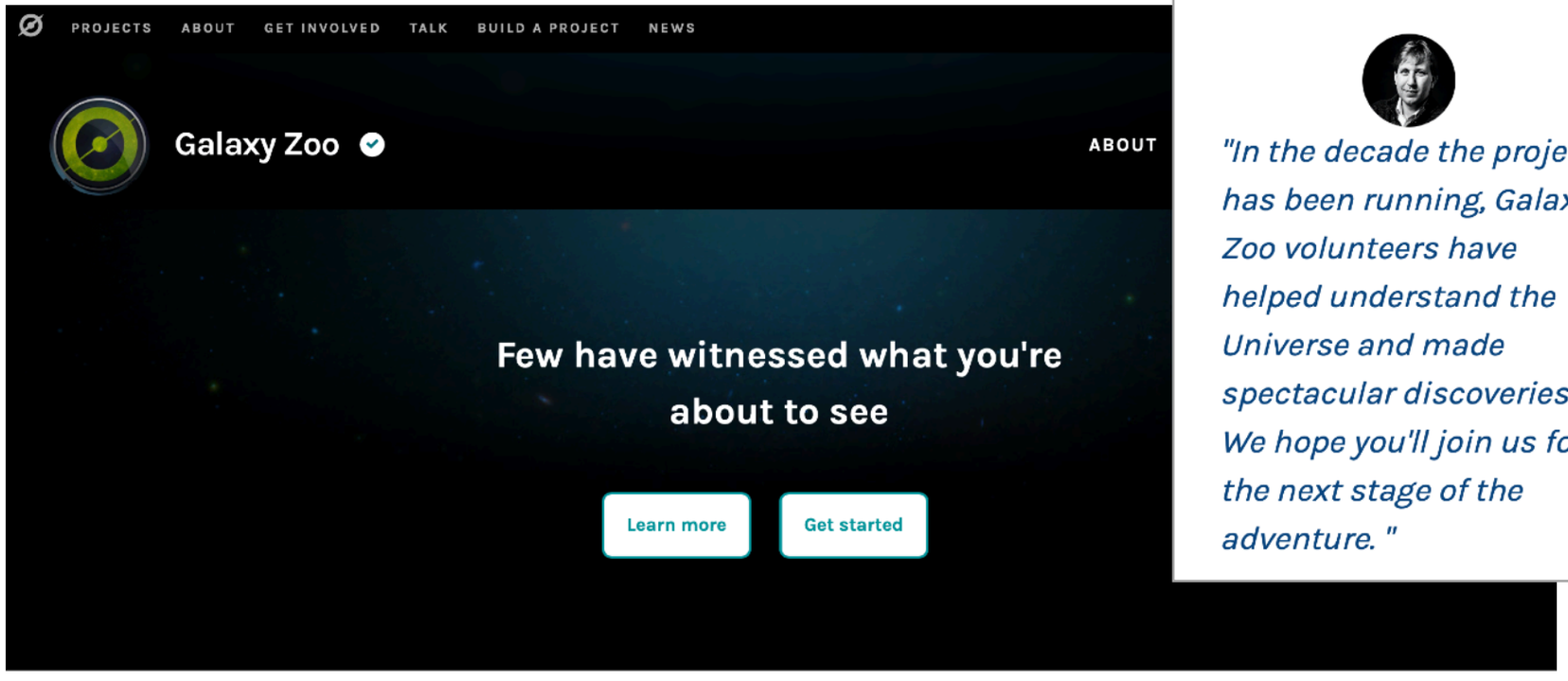
1 // EXKURS: READY UND CUSTOM MADE DATEN KOMBINIEREN.





(Blumenstock, Cadamuro, & On, 2015)

1 // EXKURS: CUSTOM MADE DATEN IM GROSSEN STIL.



WORDS FROM THE RESEARCHER

"In the decade the project has been running, Galaxy Zoo volunteers have helped understand the Universe and made spectacular discoveries. We hope you'll join us for the next stage of the adventure. "

ABOUT GALAXY ZOO

To understand how galaxies formed we need your help to classify them according to their shapes. If you're quick, you may even be the first person to see the galaxies you're asked to classify.

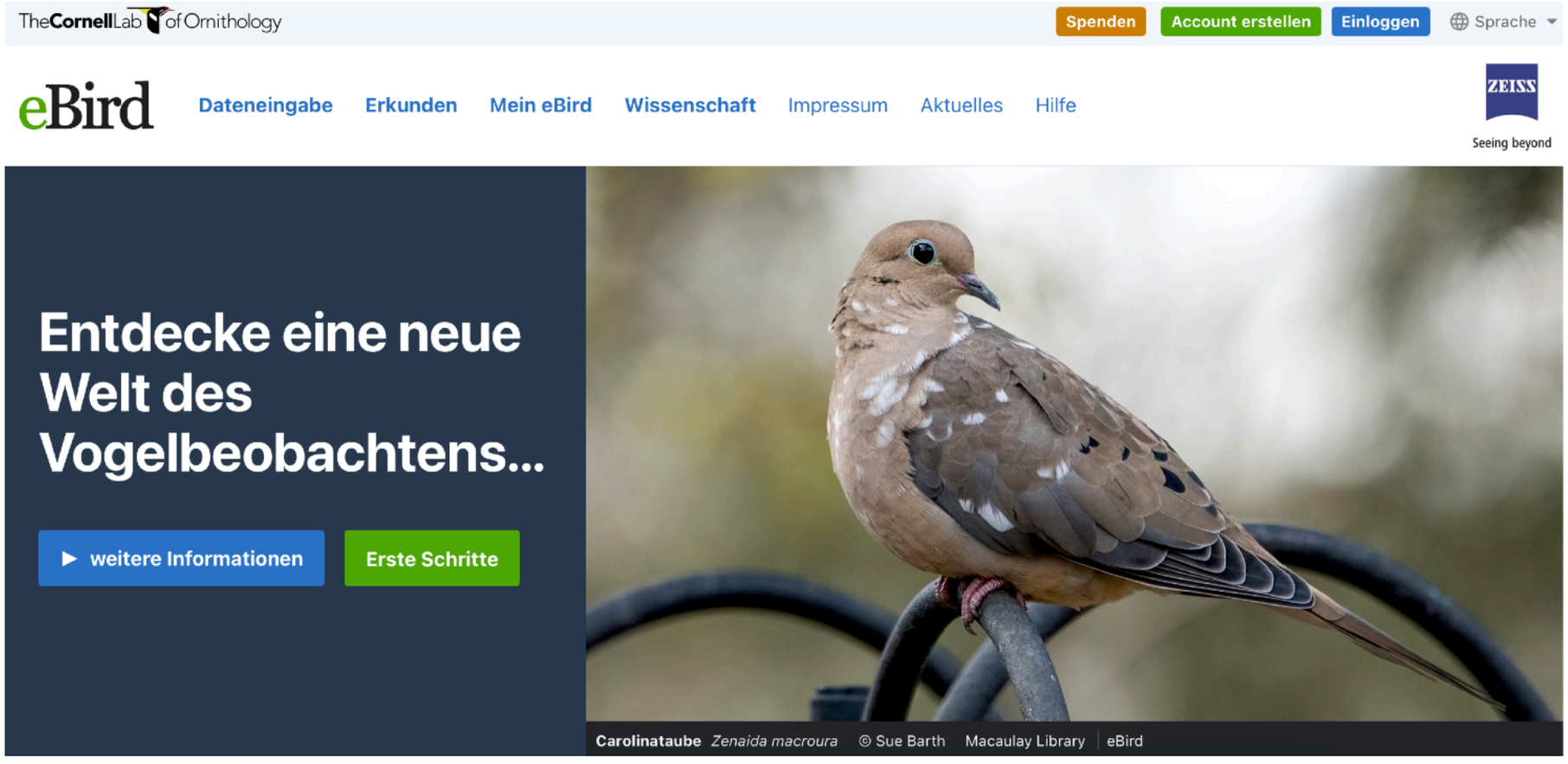
Look at telescope images of distant galaxies.

Explore the sky. What will you find?

76.647	3.654.869	287.169	273.174
Volunteers	Classifications	Subjects	Completed Subjects

https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/

1 // EXKURS: READY MADE DATA NUTZBAR MACHEN.



https://ebird.org/home

1 // EXKURS: WIE VIELE DATEN BRAUCHE ICH EIGENTLICH?

Es kommt drauf an.

Auf das Modell. Auf das zu lösende Problem. Auf die Qualität der Daten.

Mehr Parameter, mehr Daten.

Mehr Klassen vorherzusagen, mehr Daten.

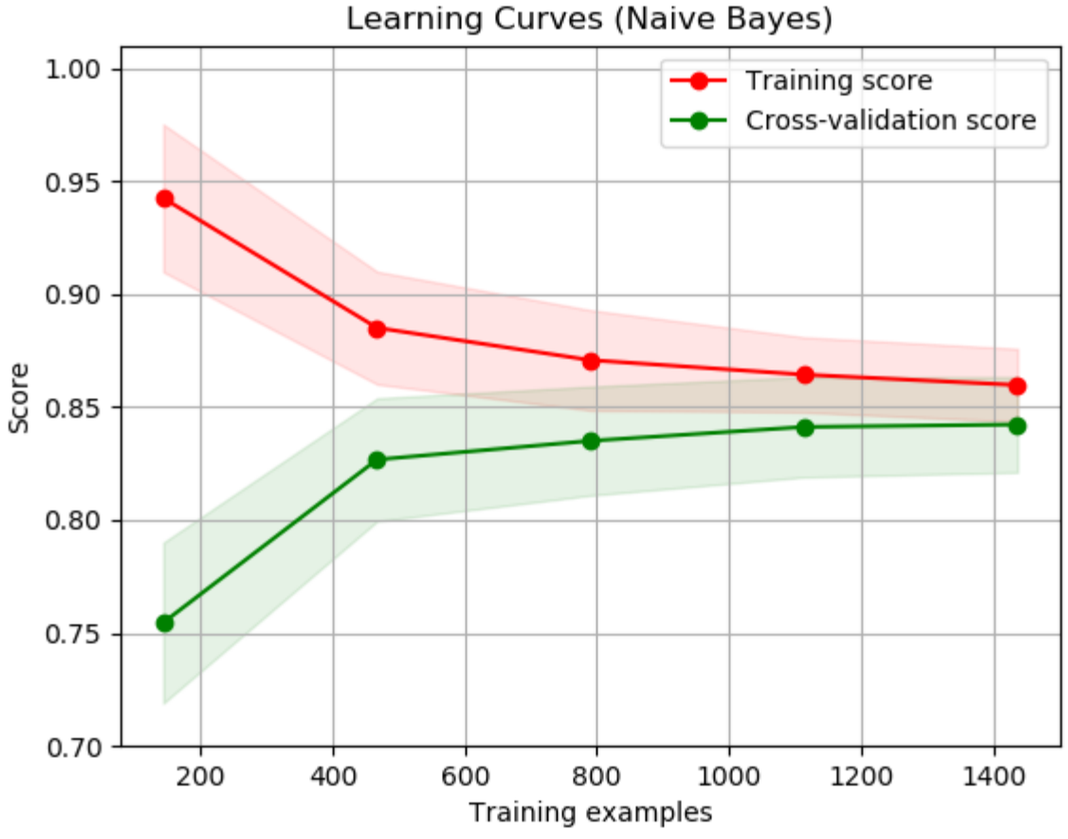
Saubere Daten, weniger Daten. Ungenaue Daten, mehr Daten.

signal-to-noise-ratio

Die richtigen Daten, weniger Daten. Die falschen Daten, mehr Daten.

https://de.coursera.org/lecture/data-machine-learning/how-much-data-do-i-need-VolCP

1 // EXKURS: WIE VIELE DATEN BRAUCHE ICH EIGENTLICH?



Sich an ähnlichen Problemen in der Literatur orientieren und die Lernkurve (**Learning Curve**) analysieren.

https://en.wikipedia.org/wiki/Learning_curve_(machine_learning)

2.A // DATA CLEANING

Randomisieren.

Deduplizieren.

Normalisieren.

Imputieren.

Data Augmentation.

Dimensionsreduktion.

Feature Engineering.

2.B // EXPLORATORY DATA ANALYSIS

Datenvisualisierung.

Deskriptive Statistiken.

Plausibilitätsprüfung.

Muster erkunden.

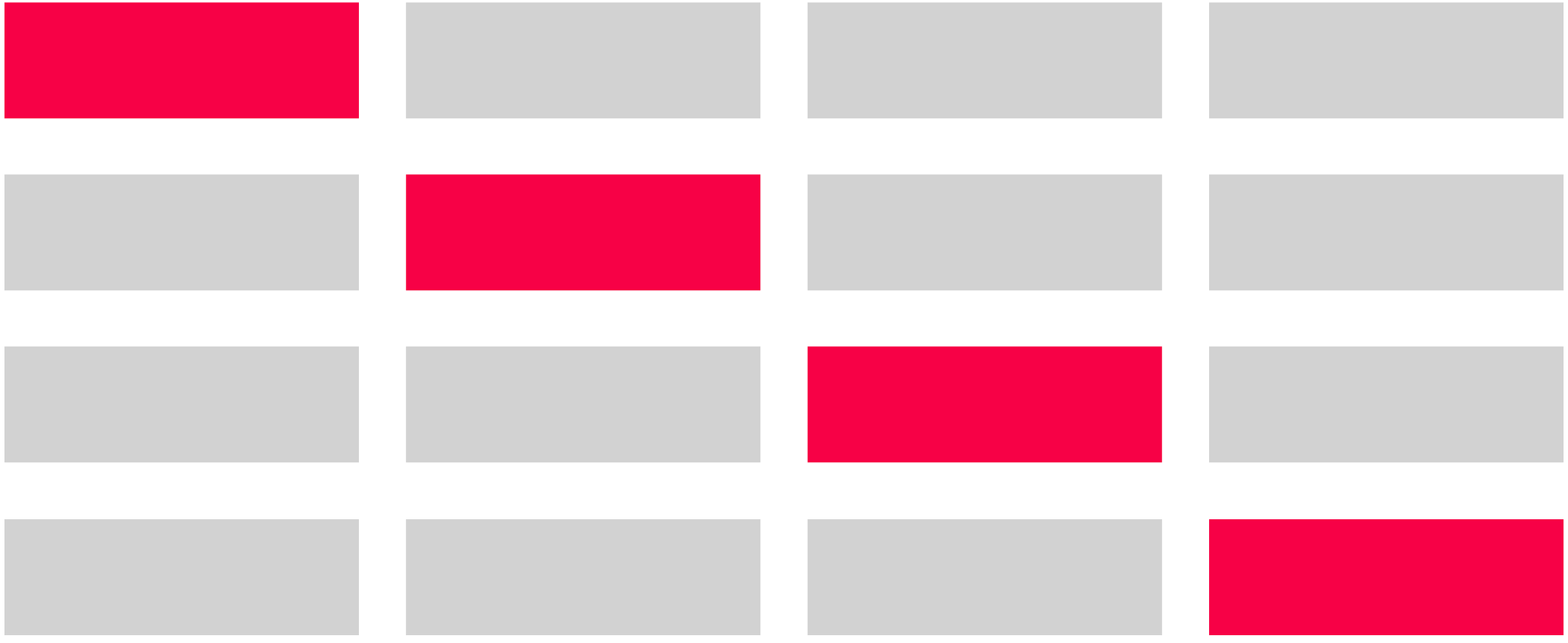
VOR DER MODELLANPASSUNG: DIE CROSS-VALIDIERUNG VORBEREITEN

Training Set: wird verwendet, um die Modellparameter zu schätzen

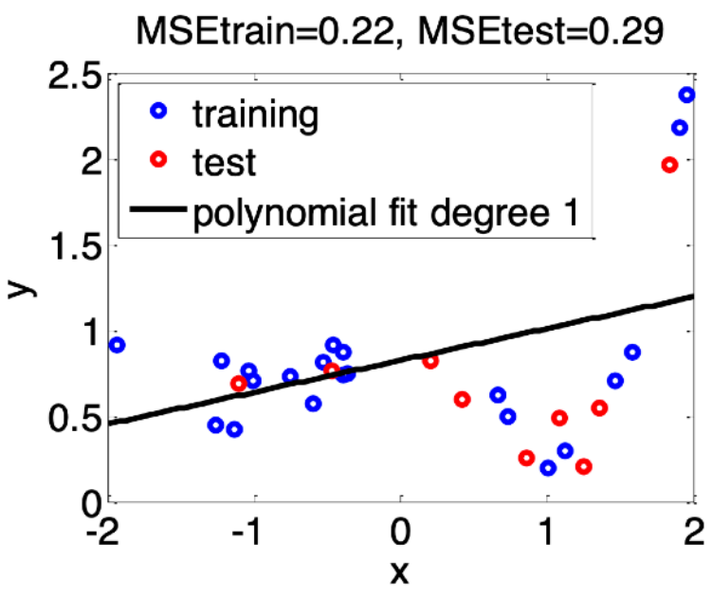
Validation Set: wird verwendet, um das Modell bereits in der Trainingsphase auf dessen Güte zu testen; das Modell mit den besten Evaluierungsergebnissen im Validation Set wird ausgewählt.

Test Set: wird verwendet, um die tatsächliche Güte des Modells auf ungesehenen Daten zu testen

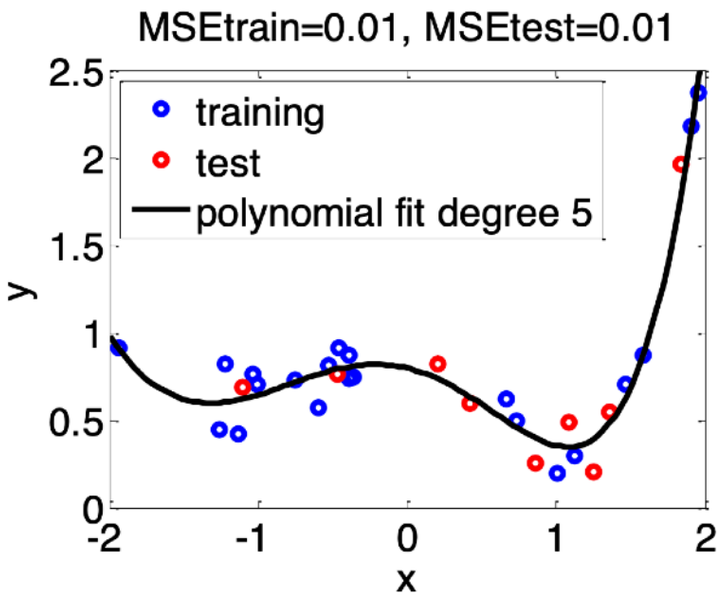
VOR DER MODELLANPASSUNG: DIE CROSS-VALIDIERUNG VORBEREITEN



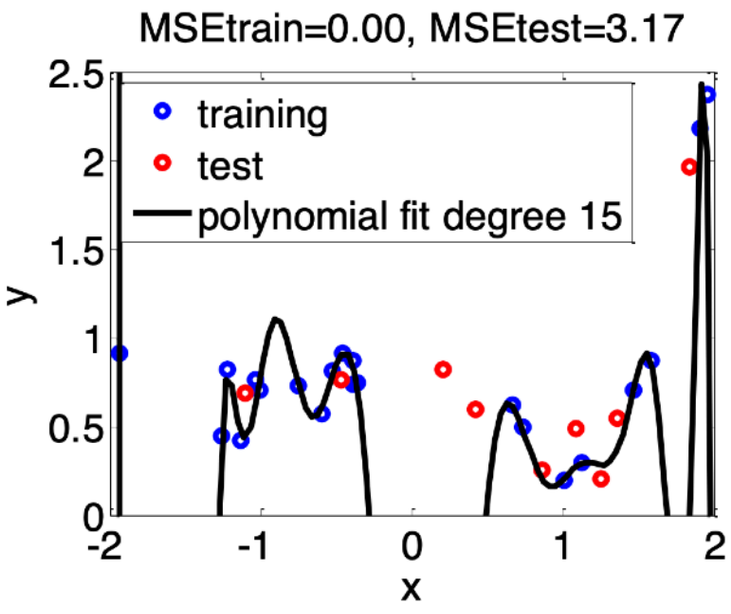
DIE CROSS-VALIDIERUNG VERHINDERT OVER- UND UNDERFITTING.



underfitting



„just right“



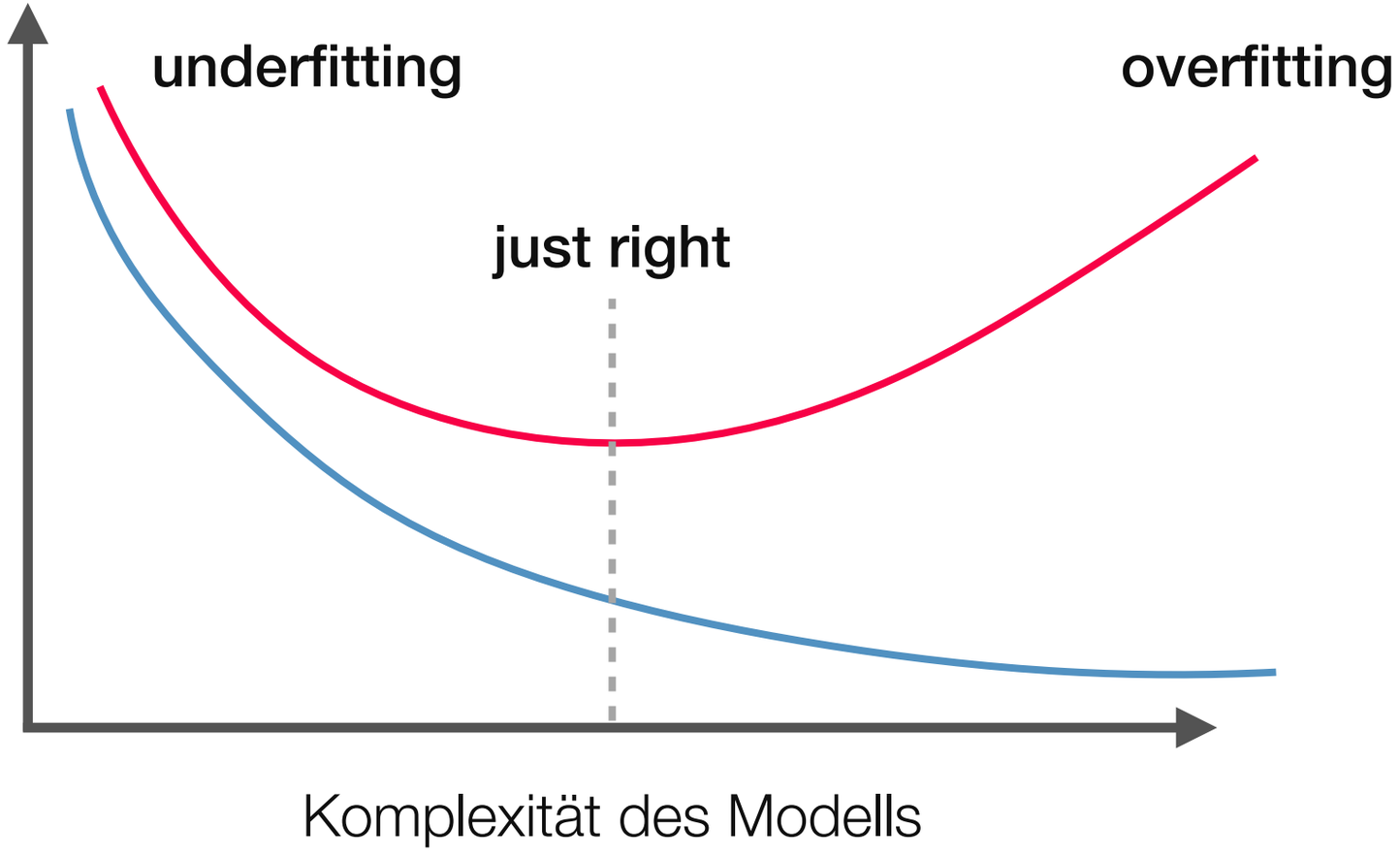
overfitting

(Subramoney & Bellec, TU Graz, 2019)

DIE CROSS-VALIDIERUNG VERHINDERT OVER- UND UNDERFITTING.

Fehler im Validation Set

Fehler im Training Set



(Subramoney & Bellec, TU Graz, 2019)

FAUSTREGEL:
80% VORBEREITUNG
20% MODELLANPASSUNG

3 // MODEL FITTING

lineare & logistische Regression

Decision Trees

PCA

SVM

k-means

…

4 // MODEL EVALUATION: FÜR REGRESSION

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

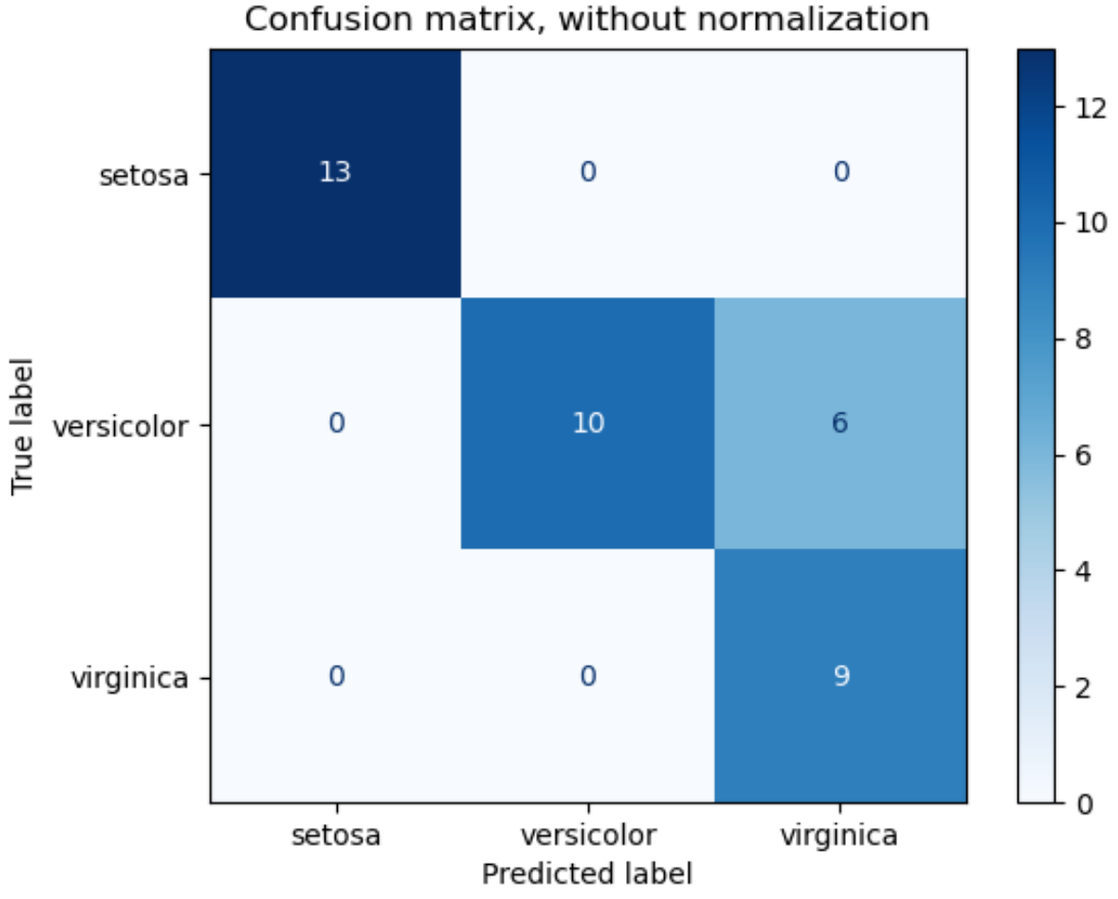
Mean Squared Error

$$R^2 = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}}$$

Coefficient of Determination

4 // MODEL EVALUATION: FÜR KLASSIFIZIERUNG

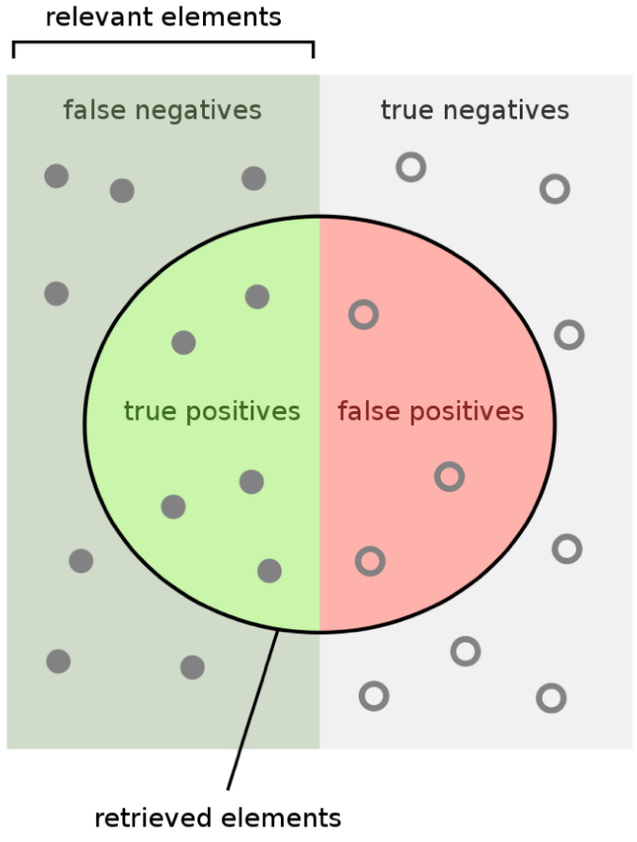
Die Confusion Matrix



https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

4 // MODEL EVALUATION: FÜR KLASSIFIZIERUNG

Der F1-Score



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

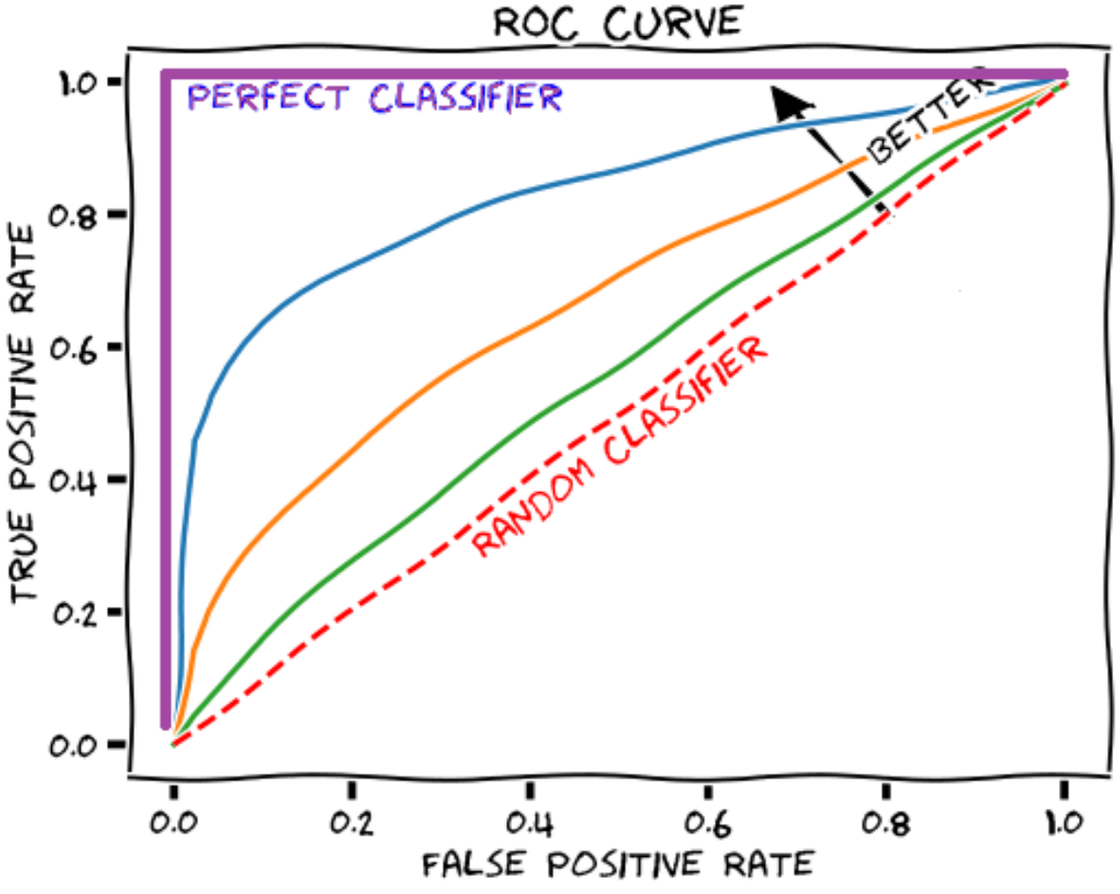
https://en.wikipedia.org/wiki/F-score

4 // MODEL EVALUATION: FÜR KLASSIFIZIERUNG

ROC-AUC Kurve

Receiver Operating Characteristics

Area Under the Curve



https://medium.datadriveninvestor.com/understanding-auc-roc-clearly-explained-74c53d292a02

5 // FINE TUNING

Die Hyperparameter des Modells anpassen.

(z.B. die Anzahl der Trainingsschritte oder die Learning Rate)

TRIAL AND ERROR!

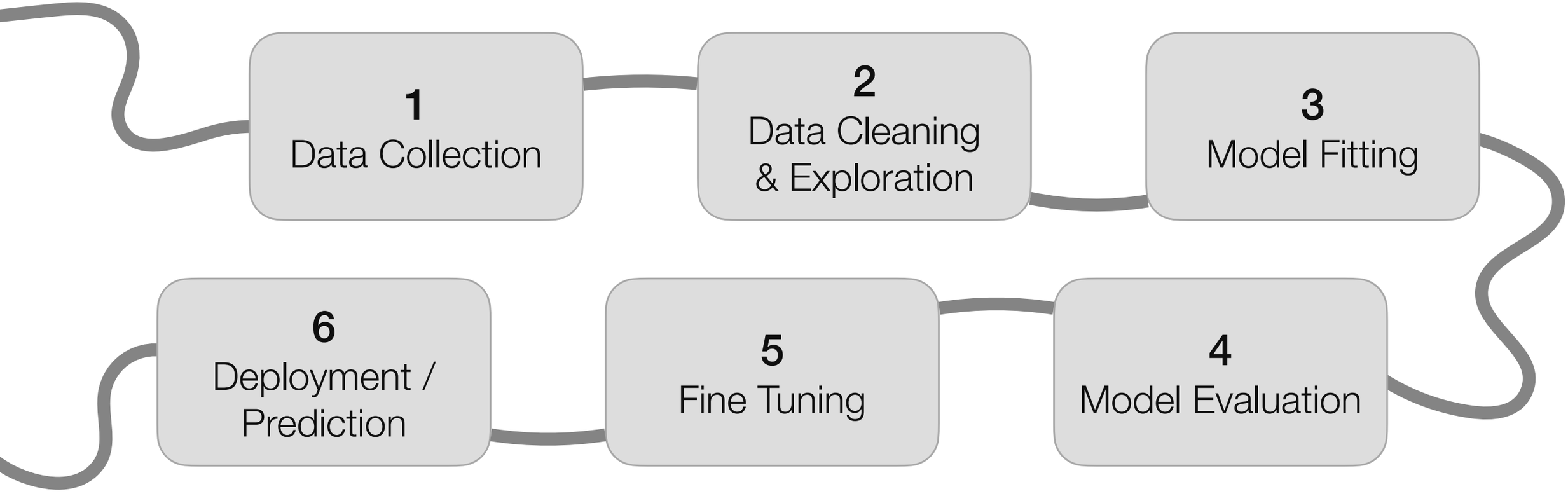
X. // BENCHMARKING

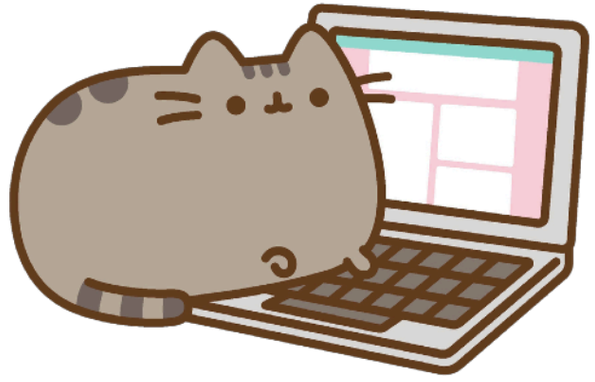
SVMs

eine tolle neue Neural Network Architektur

6 // PREDICTION

DIE DATA SCIENCE PIPELINE UMFASST (UNGEFÄHR) SECHS SCHRITTE:





THANK YOU!