

# TECHNICAL APPLICATIONS AND DATA MANAGEMENT. WS 2020/2021.

## VORLESUNG 4

05.10.2021

MÜNCHEN

STUDIENGANG  
DIGITAL  
MANAGEMENT.




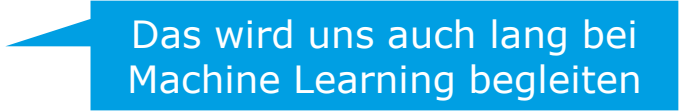
## AGENDA

1. Case Study Descriptive and Explorative Statistik
2. Inferenzstatistik
  1. Bayes'sche Inferenz
  2. Regression
  3. Modellevaluation und -Güte
3. Ausblick auf Maschinelles Lernen

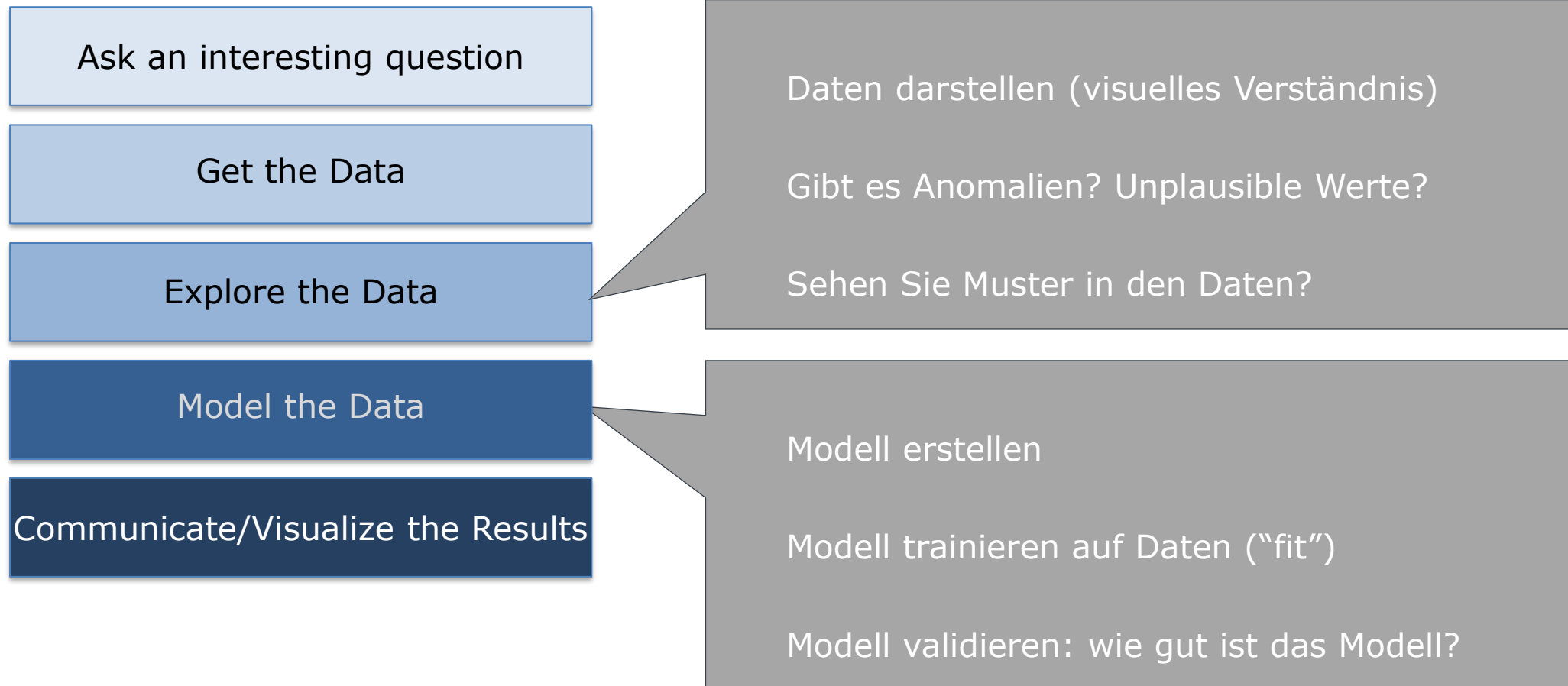
## WAS HABEN WIR BIS JETZT GEMACHT?

ROADMAP	WAS HABEN WIR GEMACHT?
Vorlesung 1	Workflow Data Management, Datentypen und Datenqualität
Vorlesung 2	Einführung Data Science und Data Science Workflow, Grundlagen Data Management
Vorlesung 3	Grundlagen Stochastik: Wahrscheinlichkeitsrechnung, deskriptive und explorative Statistik

# WAS MACHEN WIR HEUTE?

- Case Study deskriptive und explorative Statistik
- Statistische Inferenz mit
  - Bayes Theorem/ Bayes Inferenz
  - ~~Hypothesentest~~  Aus zeitlichen Gründen
  - Regression
  - Bewertung Modellgüte/ Modellvalidierung  Das wird uns auch lang bei Machine Learning begleiten
- Detaillierung Modellbildung und Prüfen Modellgüte anhand Titanic-Notebook
- Ausblick maschinelles Lernen
- Verteilen Projektarbeit

## EINORDNUNG DER HEUTIGEN INHALTE IM DATA SCIENCE WORKFLOW.



# WIEDERHOLUNG: WAS IST STOCHASTIK?

Stochastik<sup>1</sup> besteht aus folgenden Teilgebieten:

- Wahrscheinlichkeitstheorie: mathematisches Erfassen und Analyse zufälliger (nicht-deterministischer) Ereignisse
- Mathematische Statistik<sup>2</sup>:
  - Deskriptive Statistik: Daten durch Graphiken oder Tabellen visuell beschreiben.
  - Explorative Statistik: Zusammenhänge/ Muster zwischen Daten finden und bewerten, Entdecken von Hypothesen
  - Inferenzstatistik: aus einzelnen Eigenschaften einer Menge Eigenschaften über Gesamtmenge ableiten, Hypothesen testen



# 1. CASE STUDY

## FALLBEISPIEL IN GRUPPENARBEIT: DESKRIPTIVE UND EXPLORATIVE STATISTIK.

**Amazon:** 50 bestselling novels on Amazon each year from 2009 to 2020.

Datensatz verfügbar unter: [Link](#)

**Loan\_Data:** Daten von Privatkrediten.

Datensatz verfügbar unter: [Link](#).

Erklärung Datensatz: [Link](#)

Bisschen schwieriger, aber  
probieren Sie es aus!

**IMDB:** Top 1000 Filme auf IMBDB

Datensatz verfügbar unter: [Link](#)

Empfehlung zum Einstieg



## CASE STUDY IN GRUPPENARBEIT

1. Erstellen Sie ein Notebook in Google Colab mit Namen ihres Fallbeispiels (analog Template).
2. Laden Sie die Standard-Bibliotheken (analog Titanic).
3. Laden Sie den Datensatz per Pandas-Funktion `read_csv()`.
4. Wenden Sie die gelernten deskriptiven Statistik-Methoden auf den Datensatz an (Tip: Pandas-Describe Funktion). Welche Ergebnisse und Hypothesen können Sie daraus generieren?
5. Plotten Sie für jede der vorgestellten Plot-Kategorien der explorativen Statistik je ein Beispiel. Denken Sie dabei an die jeweiligen Datentypen (Ganz-/ reelle Zahlen, Kategorien).
6. Leiten Sie aus den Plots Hypothesen oder Ergebnisse ab (wie letzte Vorlesung).
7. Können Sie die Hypothesen durch weitere Analysen bestätigen oder widerlegen? Sind die Ergebnisse statistisch belastbar??
8. Stellen Sie Ihre Ergebnisse, Hypothesen und Plots vor.


~60 Minuten

## BEISPIELHAFTE PLOTS FÜR DAS IMDB-DATASET.

### Allgemein:

- Was ist die häufigste Länge eines Filmes? (Histogram/ KDEPlot: duration)
- Was ist das häufigste Rating eines Filmes? (Histogram/ KDEPlot: star\_rating)
- Gibt es einen Zusammenhang zwischen Filmlänge und Rating? (Violinplot/ Scatterplot: star\_rating vs. duration)
- Gibt es einen Zusammenhang zwischen Filmgenre und Rating? (Stripplot/ Boxplot: genre vs. star\_rating)
- Welches Genre hat die meisten Filme unter den Top 1000?
- Gibt es Zusammenhänge zwischen Länge, Rating und Genre? (Violinplot/Scatterplot: star\_rating vs. Duration mit hue=content\_rating)
- Plotten Sie einen Pairplot. Was kann man für Auffälligkeiten sehen?

### Detailanalysen:

- Genre: Schauen Sie sich ein beliebiges Genre an, z.B. Crime. Machen Sie die gleichen Auswertungen: Gibt es Unterschiede?  
ACHTUNG: hierfür müssen Sie das Dataset filtern. Dafür müssen Sie Sie statt data=IMDB\_df folgendes einsetzen:  
data=IMDB\_df[IMDB\_df['genre']=='Crime'].  Das ist ein sogenannter Filter in Pandas.
- Rating: Schauen Sie sich ein beliebiges Rating an. Wie heißt der Filter? Was sehen Sie für Erkenntnisse?  
ACHTUNG: hierfür müssen Sie wie bei der obigen Frage die Menge nach dem gewählten Rating filtern.....

## 2. STATISTISCHE INFERENZ

# STATISTISCHE INFERENZ.

## Explorative Statistik:

- Auf Basis von Visualisierungen von Daten werden Hypothesen oder Annahmen definiert.
- Die Daten basieren dabei häufig auf mehrfachen Experimenten, stellen also keine Gesamtmenge dar.



## Statistische Inferenz

- Ermöglicht das Validieren der Übertragbarkeit von Erkenntnissen von Experimenten auf Gesamtmenge mit Wahrscheinlichkeitsrechnung
- Ermöglicht Berechnung, wie gut die Übertragbarkeit ist (statistisch gesehen).



## 2.1 BAYES THEOREM UND BAYES'SCHE INFERENZ

# BAYES THEOREM UND BAYES INFERENZ.

- Eine der wichtigsten und ältesten Methoden (1763!) für probabilistische Inferenz.
- Kann kontinuierliche und dynamische Aktualisierungen von Wahrscheinlichkeiten berechnen (d.h. über Verlauf der Zeit).
- Einsatzgebiete:
  - Autopilot Flugzeug oder autonomes Fahren: Fusion verschiedenster Sensoren für Lokalisieren inkl. Unschärfe.
  - Spam-Filter für Mails.
  - Expertensysteme inkl. Schlussfolgerung(en), bspw. für Medizin.

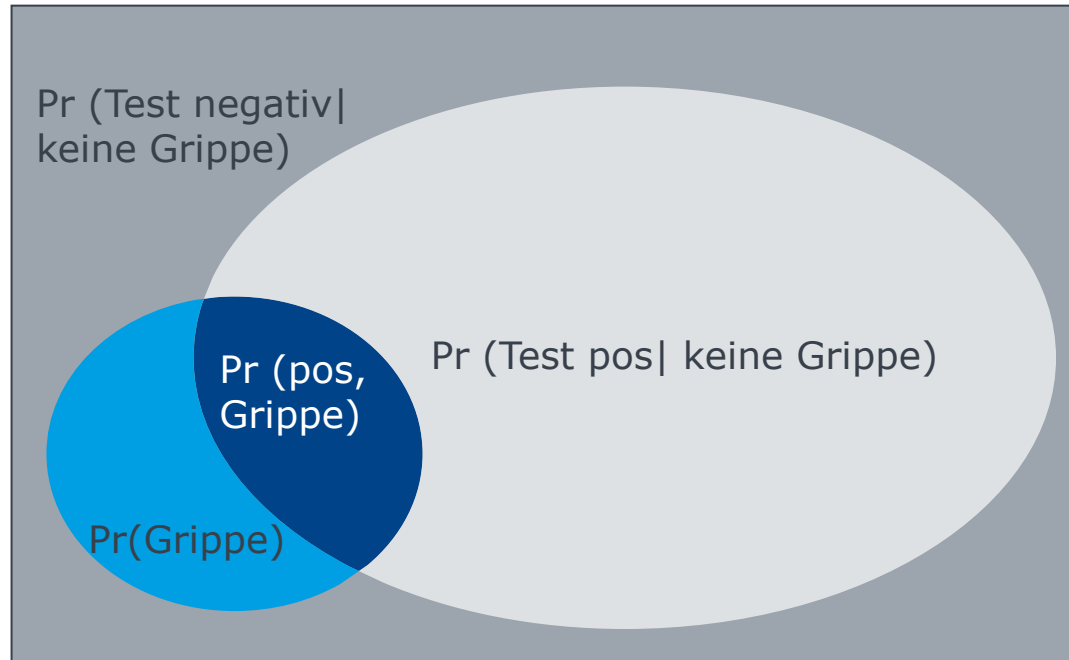
Bayes-Formel entsteht aus Umformungen der bedingten Wahrscheinlichkeit:

- $\Pr(A|B)$  := bedingte Wahrscheinlichkeit von A gegeben Evidenz B
- $\Pr(B|A)$  := bedingte Wahrscheinlichkeit von B gegeben Evidenz A
- $\Pr(A)$  := a priori (vorherige) Wahrscheinlichkeit von A
- $\Pr(B)$  := a priori Wahrscheinlichkeit von B

$$\Pr(A|B) = \frac{\Pr(B|A) * \Pr(A)}{\Pr(B)}$$

# WIESO IST BAYES/ BAYES-INFERENZ SO WICHTIG? ODER: GRAPHISCHE VERANSCHAULICHUNG ANHAND DIAGNOSE.

## Menge aller Menschen



## Prior/ Vorher bekannte Wahrscheinlichkeiten:

- Prävalenz<sup>1</sup>: Häufigkeit Grippe  $\Pr(\text{Grippe})$
- Sensitivität<sup>2</sup> Test:  $\Pr(\text{Test pos} \mid \text{Grippe})$
- Spezifität<sup>3</sup> Test:  $\Pr(\text{Test neg} \mid \text{keine Grippe})$

## Uns interessieren die unbekannten Wahrscheinlichkeiten:

- Grippe bei positivem Test:  $\Pr(\text{Grippe} \mid \text{Test positiv})$
- Fehllarm Test:  $\Pr(\text{keine Grippe} \mid \text{Test positiv})$

### Beispiel Corona-Antigen-Tests<sup>4</sup>:

- Sensitivität  $\geq 80\%$
- Spezifität  $\geq 97\%$

1 Prävalenz: Wie viele Menschen haben die Krankheit (oft nur geschätzt!!!)

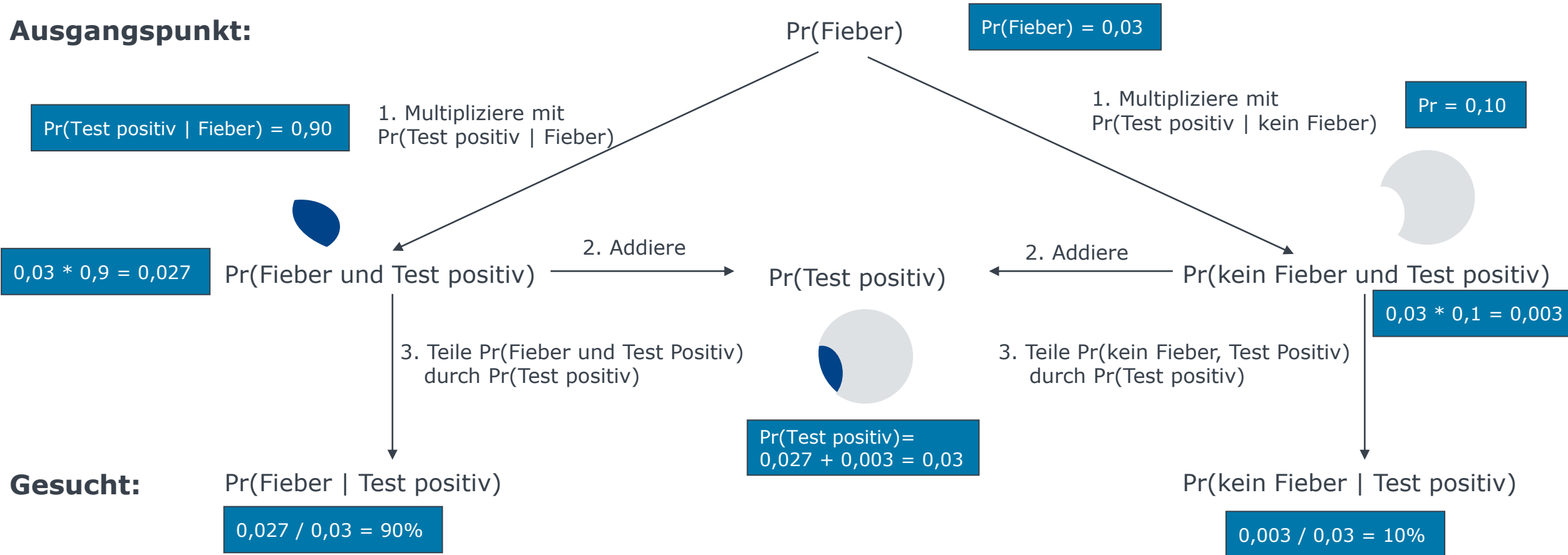
2 Sensitivität: Anteil an tatsächlich Kranken, bei denen auch eine Krankheit diagnostiziert wird

3 Spezifität: Anteil der Gesunden, bei denen tatsächlich keine Krankheit diagnostiziert wird

4 Quelle: [Link](#)

# WIESO IST BAYES/ BAYES-INFERENZ SO WICHTIG? ALGORITHMUS BAYES INFERENZ.

## Ausgangspunkt:



## Gesucht:

**Nicht jede positiv getestete Person ist auch wirklich krank bei geringer Prävalenz!  
Das ist wichtig und wird oft falsch verstanden: [Link](#), [Link](#) (und viele weitere mehr ☹)**



## 2.2 REGRESSION

# MOTIVATION REGRESSIONSANALYSE.

- Bisher: Untersuchen Zusammenhänge zwischen Variablen von Hand (bspw. Passagierklasse und Überlebensrate).
- Regression basiert auf Annahme Beziehung zwischen abhängiger<sup>1</sup> **Variable** und weiteren **Größen**. Sie ermöglicht automatisiert:
  - Schätzen und Vorhersagen von abhängiger<sup>1</sup> Variablen → Überschneidung mit Machine Learning
  - Untersuchen kausaler Zusammenhänge zwischen Variablen (Achtung: Abhänge müssen begründbar sein!)
- Wir befassen uns in dieser Vorlesung (aus Zeitgründen) nur mit der linearen und logistischen Regression.
  - Lineare Regression: Zielvariable wird per Addition und Multiplikation einer/ mehreren Variablen per Gerade<sup>2</sup> angenähert
  - Logistische Regression: Spezialfall der lineare Regression, Zielvariable hat entweder Wert 0 oder 1

ACHTUNG: IM MACHINE  
LEARNING HAT REGRESSION  
ABER GANZ BESTIMMTE  
BEDEUTUNG. Später mehr...

Regressionsanalyse ist eines der bekanntesten und einfachsten statistischen Verfahren

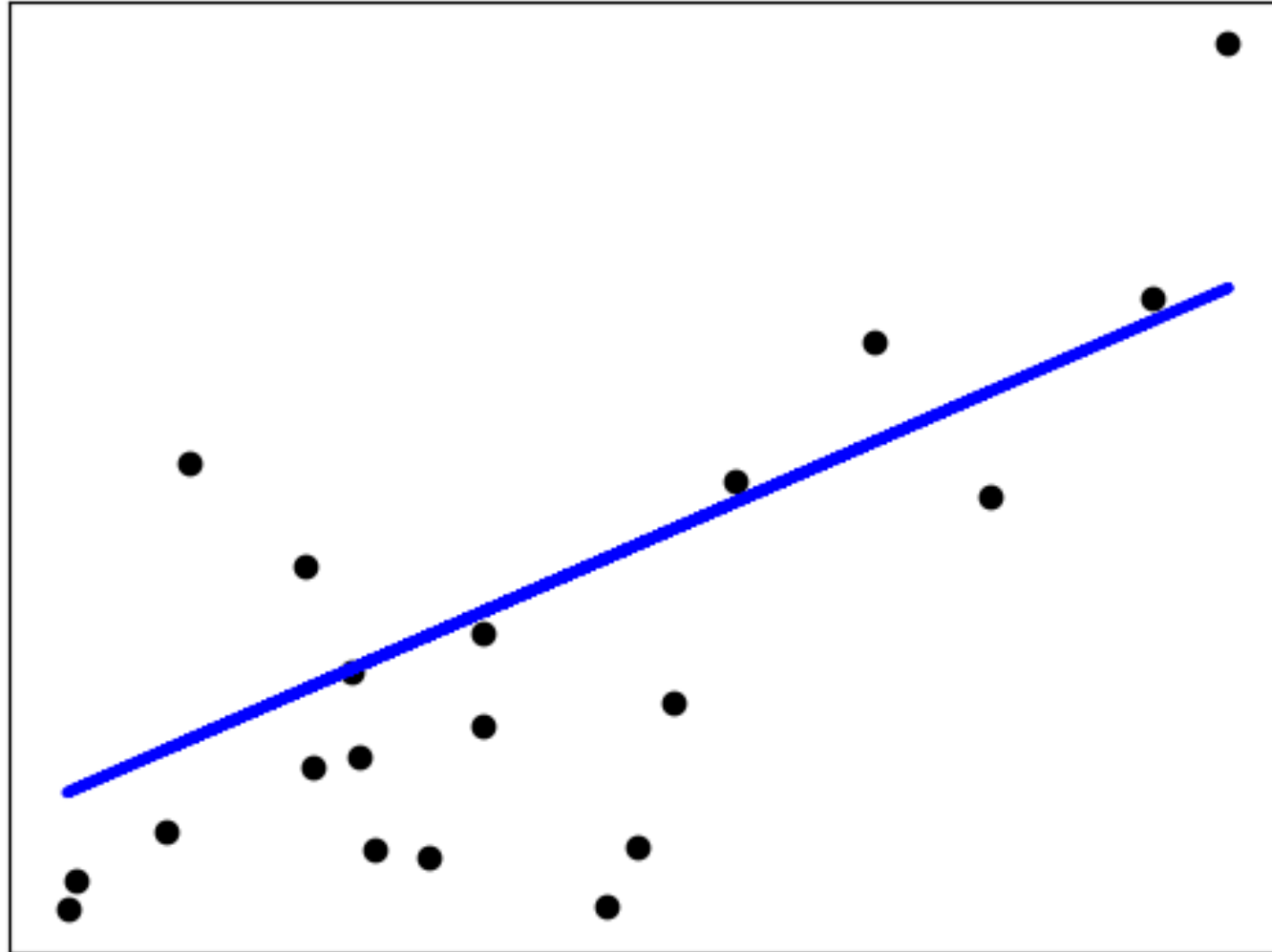
## REGRESSION: WIE FUNKTIONIERT DAS?

1. Datenaufbereitung/ -bereinigung: Plausibilisieren Werte, Beseitigen fehlerhafte/ fehlende Werte, Typumwandlung, ...  
→ haben wir uns schon angesehen
2. Modellwahl und -anpassung: Wahl eines Algorithmus sowie Verfahren für Anpassen Modell.  
→ wir schauen uns Lineare Regression sowie Least Squares an.
3. Modellvalidierung: Prüfen, wie gut Modell nach Schritt 2 ist.  
→ wir schauen uns Standardmetriken an
4. Vorhersage von Werten

Vgl. Jupyter Notebook Titanic

Ausblick: Diese Schritte sind bei allen Machine Learning- Verfahren gleich

## LINEARE REGRESSION: BEISPIELHAFTE DARSTELLUNG.



## LINEARE REGRESSION: FORMALE DEFINITION.

- Grundlage Regression: Annahme Beziehung zwischen abhängiger **Zielvariable Y** und **Größen/ Features X<sub>i</sub>**
- Ziel: Vorhersage einer Zielvariablen  $Y_i$  durch Kombination einzelner Features in Form von

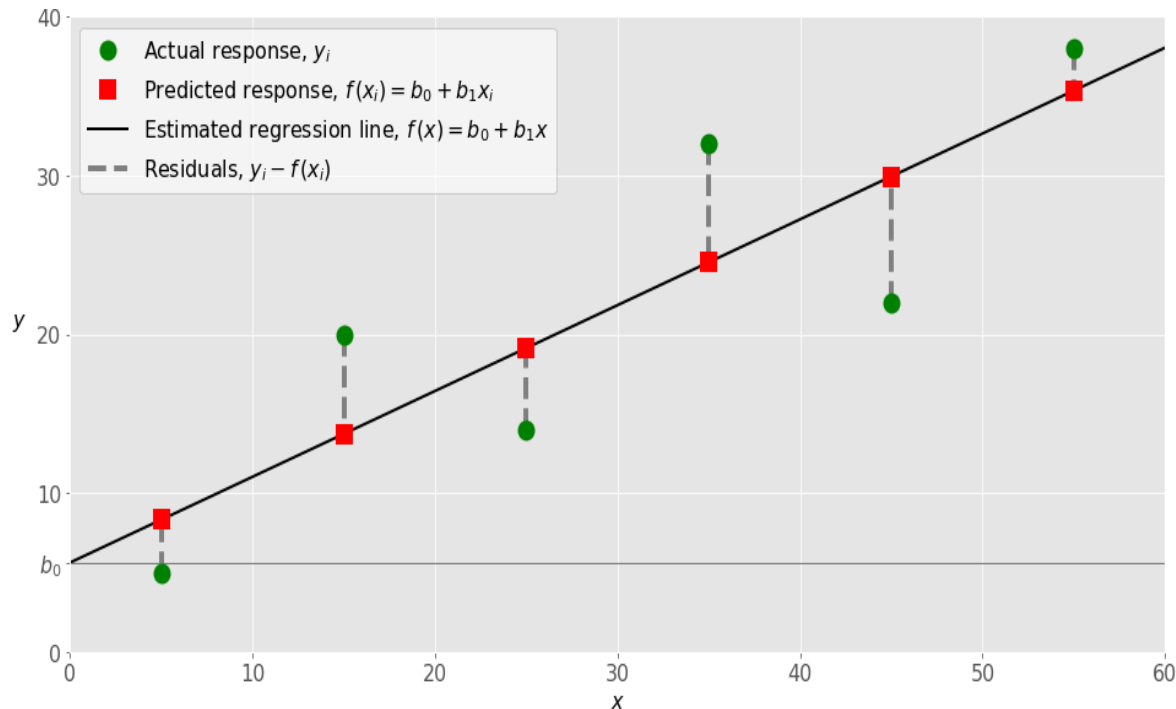
$$\widehat{Y}_i = w_i X_i + b_i + \varepsilon_i$$

Dabei ist:

- $\widehat{Y}_i$  die vorherzusagende Größe/ Label
- $X_i$  die einzelnen Einflußgrößen/ Features
- $w_i$  das Gewichte für Feature  $X_i$ . Bei linearer Regression auch oft Slope (Steigung) genannt
- $b_i$  die Abweichung/ Bias. Bei linearer Regression auch Intercept genannt.
- $\varepsilon_i$  eine stochastische Komponente bzw. Messfehler. Im Machine Learning wird dieser meist nicht explizit aufgeführt.

Wir vereinfachen die Regression aus der Statistik, auch um den Übergang zu Machine Learning zu vereinfachen

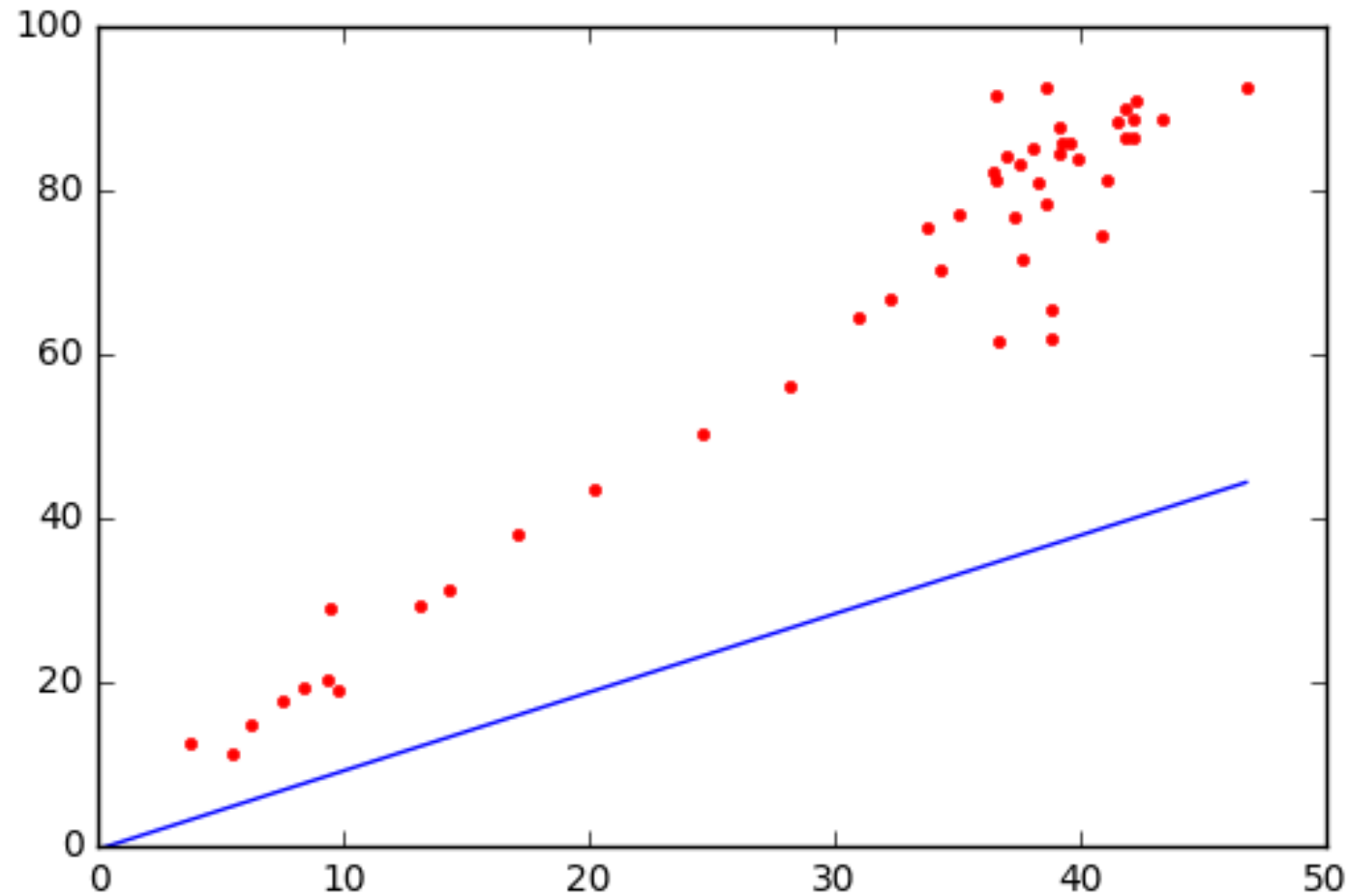
## WIR SETZEN LEAST SQUARES EIN, UM DIE REGRESSION MÖGLICHST EXAKT WERDEN ZU LASSEN.



- Differenz zwischen einem durch das Modell prognostizierten und einem realen y-Wert wird **Residual** genannt.
- Da die Punkte sowohl über als auch unter der Gerade liegen können, quadrieren wir das Residual. Damit „bestrafen“ wir alle die prädizierten Punkte, die weit entfernt vom realen Punkt liegen.
- Ziel ist es, durch Wahl der Parameter des Regressionsmodells die **Summe aller Residuale zu minimieren (Sum of Least Squares)**.
- Dies geschieht über Berechnen der Ableitung sowie Setzen der Ableitung auf 0. Dies erledigt aber der Computer für uns.

Verfahren ist ca. 200 Jahre alt und wurde zuerst von Gauss für die Bestimmung der Bahn des Planeten Ceres eingesetzt

## LINEARE REGRESSION: BEISPIELHAFTER ABLAUF ALGORITHMUS.



# LINEARE REGRESSION: BEWERTUNG

Lineare Regression ist:

- Einfach zu verstehen.
- Schnell zu trainieren.
- Guter Einstieg in Machine Learning, da Vorgehensweise ähnlich vieler anderen Algorithmen aus ML ist.

Aber der Einsatz erfordert höheren Aufwand für Datenbereinigung und Datenaufbereitung, denn Lineare Regression:

- Kann nicht mit unabhängigen Variablen umgehen (versucht diese zu integrieren).
- Erkennt irrelevante Features nicht und paßt Modell an diese fälschlicherweise an (Overfitting).
- Nicht robust gegenüber Ausreißer/ Outlier, Genauigkeit Vorhersage wird enorm verschlechtert  
→ Normieren Werte meist erforderlich.
- Kann kategorische Variablen nicht direkt verwenden.





## **2.3 MODELLEVALUATION: WIE GUT IST DAS MODELL?**

# MODELLGÜTE: ÜBERSICHT AUSGESUCHTER, WICHTIGER METRIKEN.

## Klassifizierungsmetriken

- **Confusion Matrix**
- **F1-Score**
- **Accuracy**
- **Precision**
- **Recall**
- **ROC/ AUC**

## Regressionsmetriken

- **MAE**
- **MSE**
- **RSME**
- **R-squared**
- Adjusted R squared
- Explained Variance

## Statistische Metriken

- Korrelation
- Abhängigkeit
- Chi-Quadra
- Statistischer Test

Aus Zeitgründen nicht  
im Fokus

Es gibt zudem weitere spezifische Metriken für bestimmte Machine Learning Gebiete. Diese sind aber nicht Fokus Vorlesung

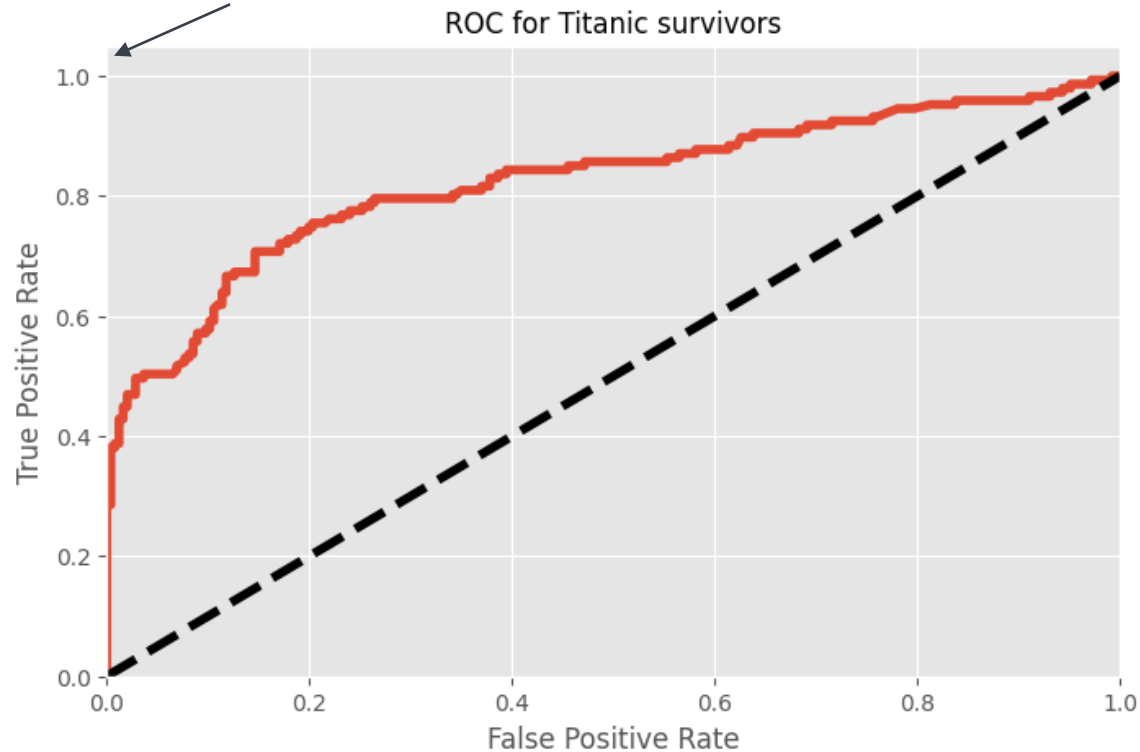
## DETAILLIERUNG KLASSIFIZIERUNGSMETRIKEN. CONFUSION MATRIX.

		Predicted	
		Ja	Nein
Tatsächlich	Ja	True Positive	False Negative
	Nein	False Positiv	True Negative

- **Recall/ Sensitivity** = wie viele Elemente korrekt vorhergesagt wurden ( $\frac{\text{True positive}}{\text{True Positive} + \text{False negative}}$ )
- **Precision** = wie viele von den als "wahr" vorhergesagten wirklich wahr waren ( $\frac{\text{True positive}}{\text{True Positive} + \text{False positive}}$ )
- **Accuracy** = korrekt vorhergesagte Elemente geteilt durch die Gesamtzahl ( $\frac{\text{True positive} + \text{false negative}}{\text{Gesamtzahl}}$ )
- **F1-Score** = je höher, desto besser kann das Modell vorhersagen ( $\frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$ )

## DETAILLIERUNG KLASSIFIZIERUNGSMETRIKEN. ROC/ AUC.

**Perfekter Klassifikator** (100% true pos., 0% false pos.)



- **ROC**(Receiver-Operating-Characteristics): Darstellung des prozentualen Verhältnisses zwischen den richtig sowie den falsch als wahr prädizierten Werten.
- **AUC**(Area under ROC Curve): Integral der ROC-Kurve. Je höher die AUC-Fläche, desto besser ist der Klassifikator.

## DETAILLIERUNG REGRESSIONSMETRIKEN

- **MAE (Mean absolute error)**: gemittelte Abweichung des prädizierten vom realen Wert. Gut für Feintuning.  
→ je kleiner, desto besser das Modell.

$$\frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

- **MSE (Mean Squared Error)**: gibt an, wie weit die Prognosewerte um den erwarteten/ realen Wert streut. Durch Quadrierung werden starke Abweichungen besonders "bestraft".  
→ Je größer, desto schlechter das Modell.

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

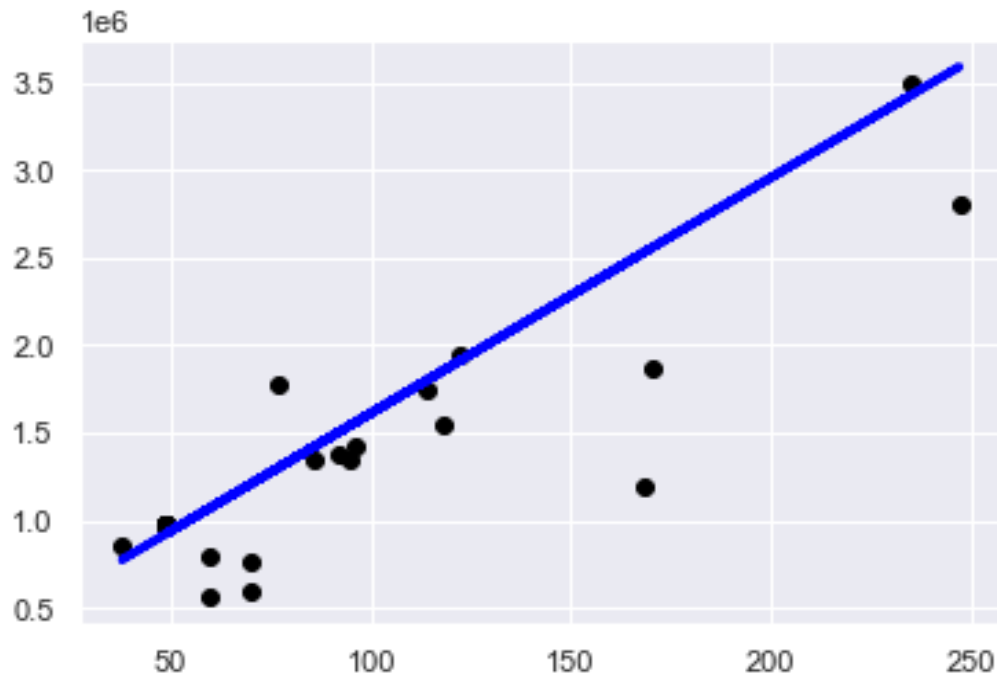
- **RSME (Rooted Mean Squared Error)**: Wurzel aus Mean Squared Error.  
→ Je größer, desto schlechter Modell.

$$\sqrt{MSE}$$

- **R-squared** = gibt an, wie gut das Modell die Daten erklären kann; d.h. wie gut das Modell die Y-Werte abdeckt. Wert ist zwischen 0 und 1  
→ Wert von 1 bedeutet perfekte Abdeckung, jeder reale Wert wird durch prädizierten Wert abgedeckt.

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

## DETAILLIERUNG METRIKEN: BEISPIEL REGRESSIONSMETRIKEN.



- MSE: hoch oder niedrig?
- RMSE: hoch oder niedrig?
- $R^2$ : hoch oder niedrig?
- Wenig oder viele Ausreißer?

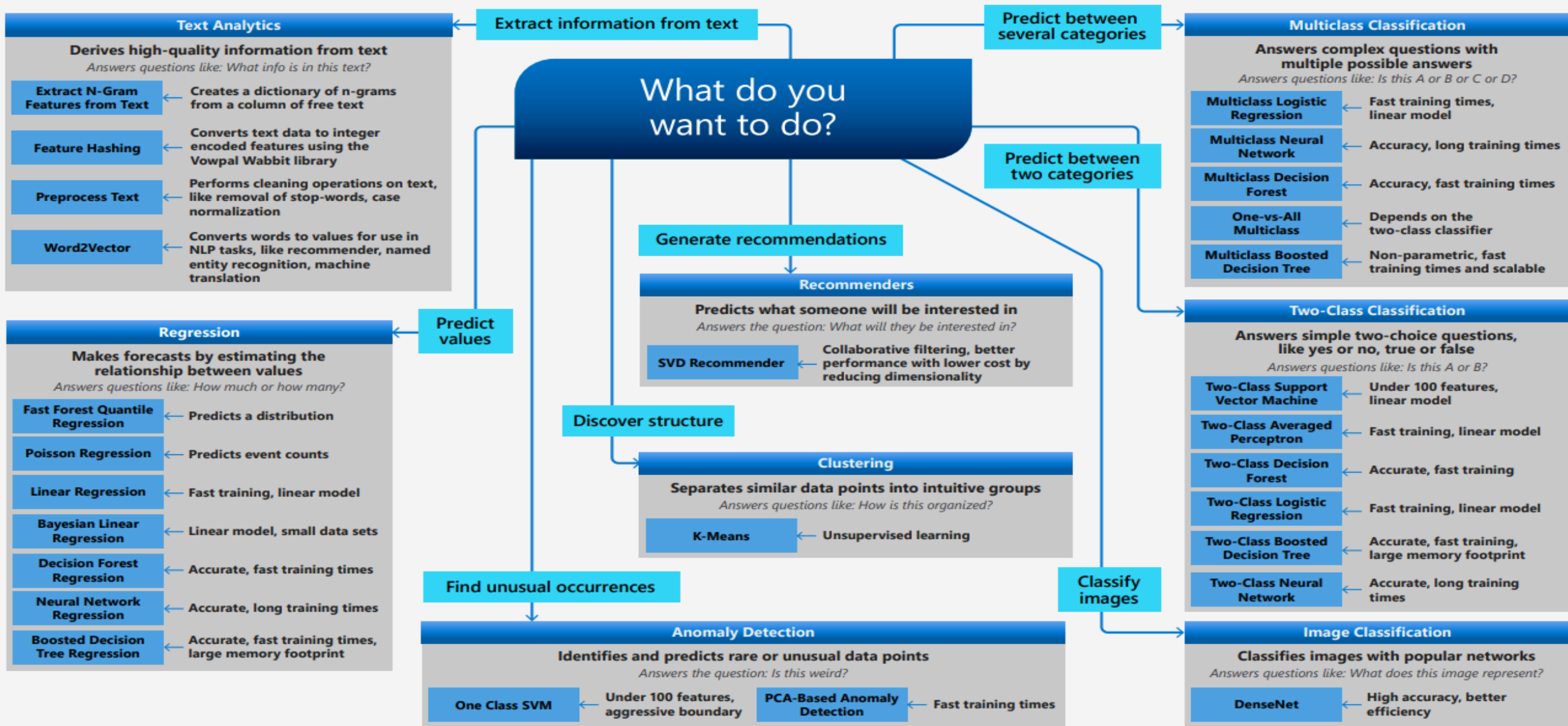


## 3. AUSBLICK MASCHINELLES LERNEN



# Microsoft Azure Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.





## LITERATUR UND WEITERE QUELLEN (AUSZUG).

### Statistik:

- Schickinger, Steger: Diskrete Strukturen 2 – Wahrscheinlichkeitstheorie und Statistik.
- James, Witten, Hastie and Tibshirani: An Introduction to Statistical Learning (freies Ebuch unter: [Link](#))
- Spiegelhalter: The Art of Statistics: Learning from Data
- Witte: Statistics (10<sup>th</sup> Edition)
- Silver: The Signal and the noise
- Taleb: Black Swan
- Huff: How to Lie with Statistics
- Wheelan: Naked statistics

### Kostenfreie Online-Kurse (bei Interesse):

- Python-Kurse
  - Python for Everybody ([Link](#))
  - Udacity Python Course ([Link](#))
  - Kaggle Courses:
    - Python ([Link](#))
    - Python Library Pandas ([Link](#))
    - PythonData Visualization ([Link](#))
    - Intro to Machine Learning ([Link](#))

### Interessante Titanic-Notebooks:

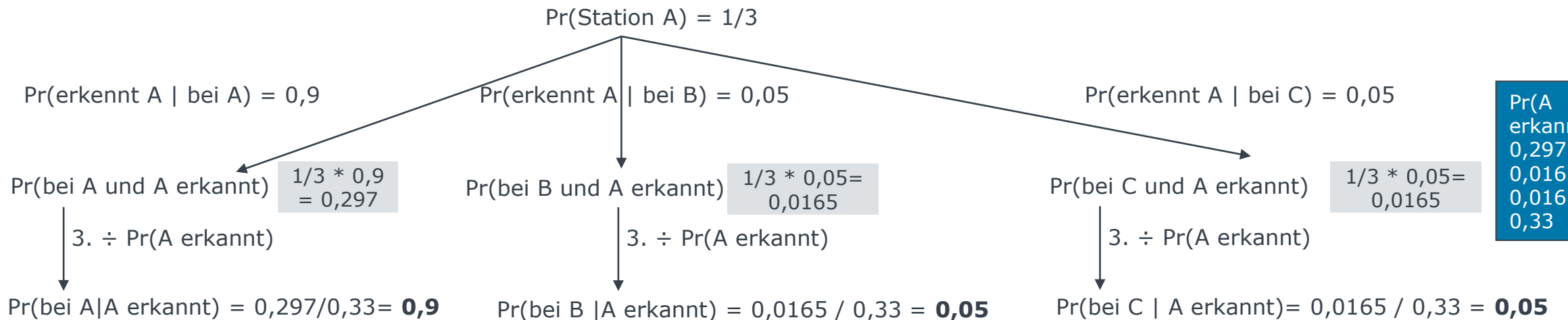
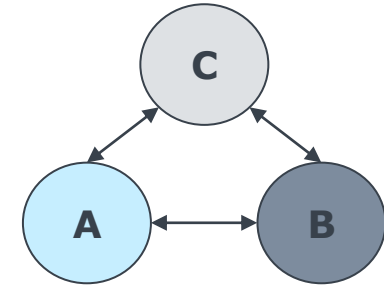
- Sehr umfangreiches Notebook inkl. AI: [Link](#)
- Fokus "schwierigere" Data Science Fragen: [Link](#)
- Data Science Workflow und AI: [Link](#)



**BACKUP**

## DETAILLIERUNG BAYES ALGORITHMUS ANHAND FALLBEISPIEL LOKALISIERUNG AUTONOMER ROBOTER.

- Ein autonomer Roboter fährt kontinuierlich 3 verschiedene Stationen A, B, C an.
- Der Roboter hat (fehlerhafte) Sensoren zur Erkennung an welcher Station er ist.
- Folgende Werte sind initial bekannt:
  - Da Ausgangslage unbekannt, nehmen wir  $\Pr(A) = \Pr(B) = \Pr(C) = 1/3$  an.
  - Erkennungsgüte Sensoren einer Station liegt bei 90%, also bspw.  $\Pr(\text{erkennt A} \mid \text{ist bei A})$ .
- **Gesucht wird die Wahrscheinlichkeit bei welcher Station der Roboter ist, wenn er A erkannt hat.**



Autonome Roboter nutzen Algorithmus kontinuierlich für Integration aller Sensorenwerte und leiten daraus Aktionen ab