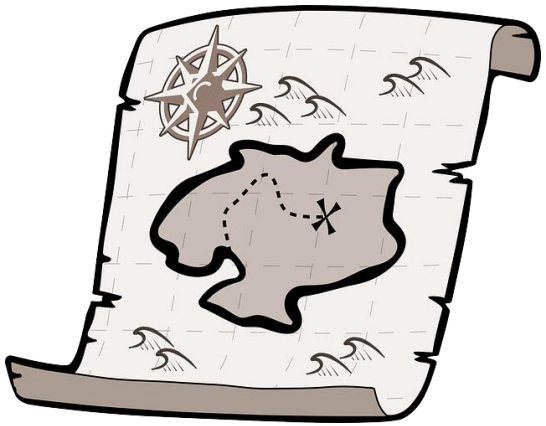
Universität Regensburg

Big Data und Machine Learning in der psychologischen Forschung

Im Spannungsfeld zwischen Erklärung und Vorhersage

Themenfelder

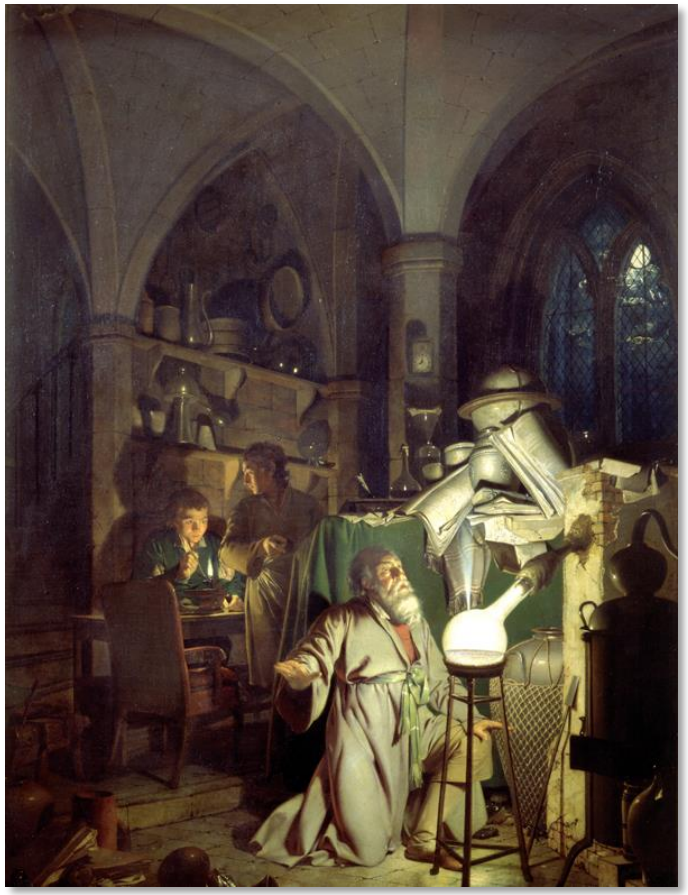
- Ziele einer empirischen Wissenschaft
- Kleiner Überblick Machine Learning
- Kleiner Überblick Big Data
- Vergleich moderner und klassischer statistischer Methoden in der Forschung
  - Die Rolle der Psychologischen Forschung in diesem Feld
- Beispiele psychologischer Studien
  1. Big Data: Persönlichkeit und Smartphonenutzung
  2. Machine Learning: Klassifizierung von Persönlichkeitstypen
  3. Big Data und Machine Learning: Schreiben lernen per App
- Zusammenfassung: Verbindung der Felder



Ziele einer empirischen Wissenschaft

1. **Beschreibung**
   - **Deskriptive Statistik:** Zusammenfassende Maße und Grafiken, um die Daten verständlich zu machen
2. **Erklärung**
   - **Inferenzstatistik:** Schätzen von Modellparametern, welche Zusammensetzung und –hänge der Daten modellieren
3. **Vorhersage**
   - **Machine Learning:** Vorhersage neuer Daten, nachdem ein Modell durch Resampling und einen Lernalgorithmus trainiert wurde

➢Normalerweise in genau dieser Reihenfolge



https://eclecticlight.co/2016/02/17/the-story-in-paintings-enlightened-by-science/

Drei Phasen nach Efron & Hastie

Die Phasen der statistischen Modelle im 20. und 21. Jahrhundert (grob eingeteilt):

1. Klassische Inferenzstatistik
   - ALM, GLM etc.

2. Computationale Methoden
   - Bayesianische Statistik, Bootstrap

3. Computerintensive Methoden
   - Machine Learning

Kleiner Überblick Machine Learning

Modellarten

- Baumbasierte Modelle
  - Random Forest, Boosting
- Kernelbasierte Modelle
  - Support Vector Machines
- Deep Learning
  - Neuronale Netzwerkmodelle

Eigenschaften

- Nutzen von **Resampling**
- **Optimiert auf Vorhersage** neuer Daten
- Weitgehend ohne direkt interpretierbare Parameter
- Hohe Funktionalität mit vielen Features (Variablen) und Daten –> **Big Data**

http://thecontextofthings.com/2016/04/18/transparency-in-business/

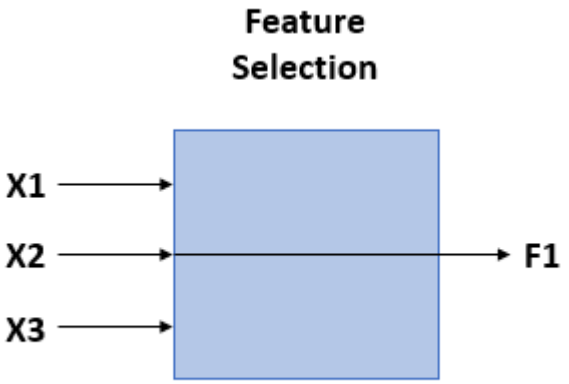
Kleiner Überblick Big Data

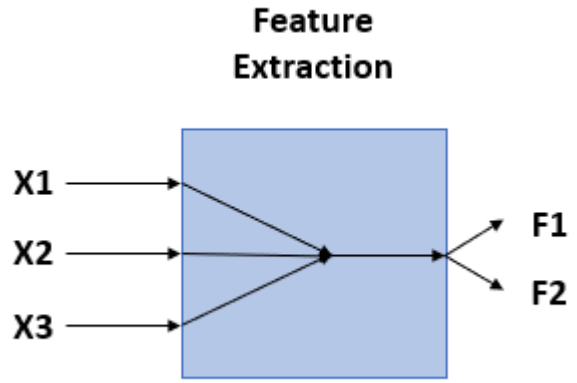
- Geprägt einem Vortrag von Roger Magoulas (2005)
  - Beschreibt einen Datensatz, der so groß ist, dass er praktisch nicht mit klassischen Methoden überblickt und bearbeitet werden kann
  - Zu divers, schnell wachsend und gewaltig, um mit bloßem Auge überblickt zu werden
- Beispiele:
  - Typischerweise **Browserdaten**, **Sensordaten** von Smartphones und –watches, **Kundendaten** von Banken, Versicherungen, Supermarktketten etc.
  - Auch in der **Medizin** (vor allem der **Bilderkennung** und –verarbeitung) und Schrift- und **Spracherkennung** im Einsatz
- Die immense Menge an Daten (vor allem Variablen) überfordert klassische statistische Modelle und prädestiniert für Machine Learning Verfahren
  - ➢ Allerdings müssen weiterhin essenzielle Entscheidungen von Menschen getroffen werden

Wichtige Entscheidungen bei Machine Learning

- Resampling
  - Muss eine Schichtung der Stichprobe berücksichtigt werden?
  - Wie gut soll sich der Algorithmus auf den Trainingsdaten anpassen?
- Feature Selection
- Feature Extraction
  - In vielen Fällen ergeben Features erst in aufbereiteter Form Sinn

➢Theoretische Basis ist essenziell





Phone Study: Ausgangslage

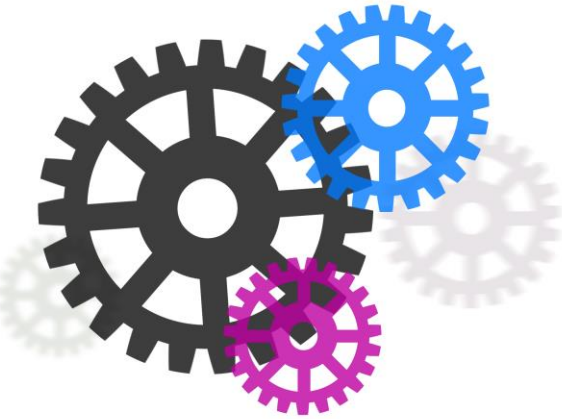
- Vorhersage der Nutzung von Smartphone Apps mit Persönlichkeitsfaktoren und -facetten
- Extrem großer Datensatz mit einer großen Anzahl an Variablen und Datenpunkten
  - ➢ *N* = 137 Personen
  - ➢ *p* = 2835 verschiedene Apps
  - ➢ Über eine Million Events

➢ Hohe Kollinearität durch Vorhersage mit allen 30 Persönlichkeitsfacetten

Phone Study: Preprocessing

Universität Regensburg

- Feature Engineering
  - Labeling der Kategorien
  - Kategorisierung der Einzelnutzungsdauer
    - Bis zum Start des nächsten Nutzungsevent, das keine Bloatware ist



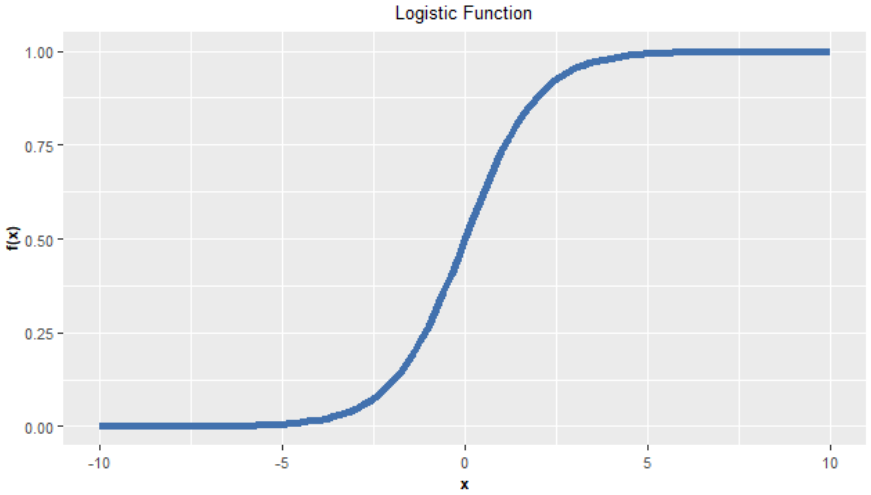
- Winsorizing von Ausreißern
- Variablenselektion bei hoher Kollinearität
  - Stability Selection (Hofner, Boccuto & Göker, 2015)
  - Least Angular Shrinkage Selection Operator (LASSO) Regression (Friedman, Hastie & Tibshirani, 2010)
- Auswertung durch Quasipoisson Regression mit Resampling

Phone Study: Zusammenfassung Ergebnisse

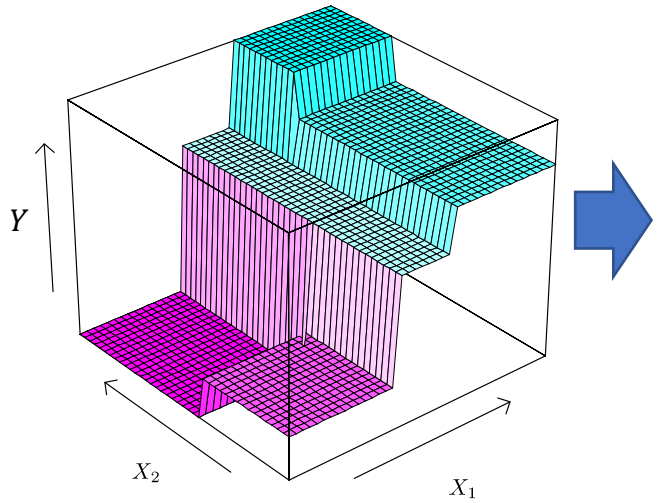
- **Extraversion** (besonders Geselligkeit) sagt die Häufigkeit von Anrufen und Kamera positiv vorher
  - ➢Höhere Extraversion, häufigere Nutzung
- **Gewissenhaftigkeit** sagt die Häufigkeit der Nutzung von Spielen vorher
  - ➢Höhere Gewissenhaftigkeit, seltenere Nutzung
- **Verträglichkeit** sagt die Häufigkeit der Nutzung von Transport Apps vorher
  - ➢Höhere Verträglichkeit, häufigere Nutzung

Vergleich Klassifizierungsmodelle

- Klassischerweise wird die logistische Regression zur (dichotomen) Kategorisierung genutzt
  - Interpretierbare Parameter, allerdings wenig flexibel beim Datenfit
- Baumbasierte Machine Learning Modelle sind flexibler
  - Allerdings Interpretierbarkeit oft schwierig und Extrapolation begrenzt



$$\frac{e^x}{1+e^x}$$



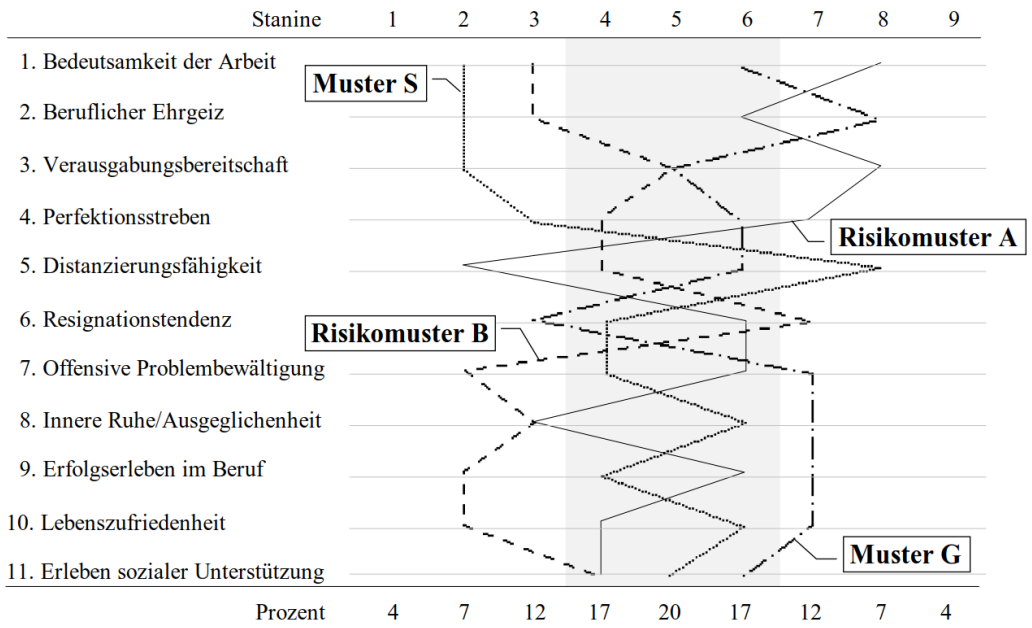
?

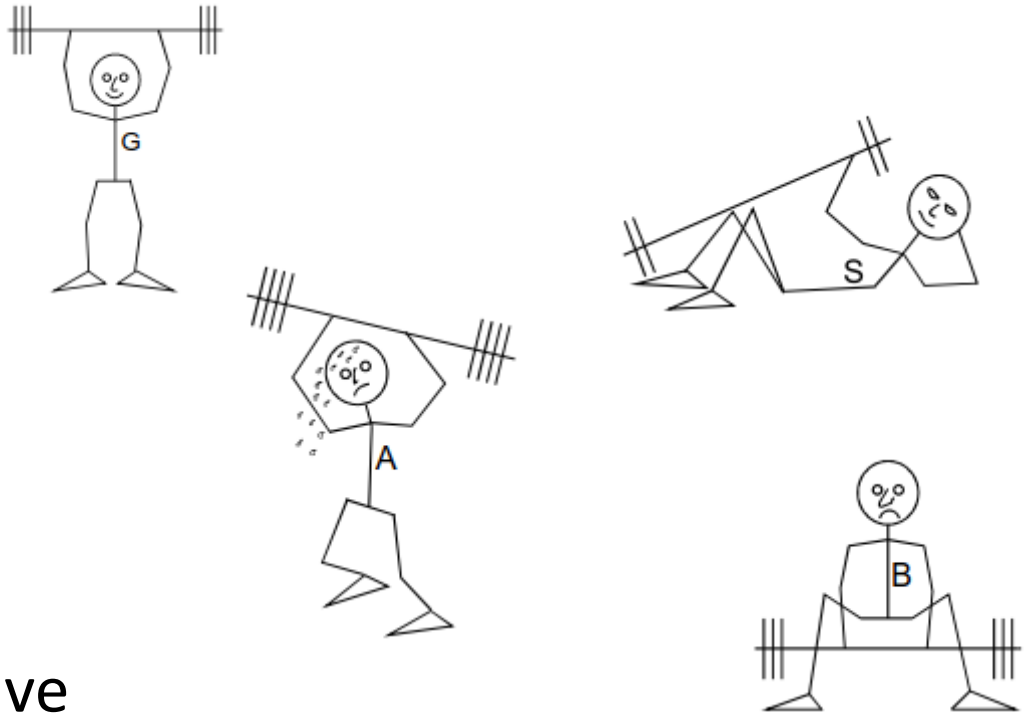
Grafik: Efron, B., & Hastie, T. (2016). *Computer age statistical inference*(Vol. 5). Cambridge University Press., S. 125

Exemplarische Studie Persönlichkeitstypen

Universität Regensburg

- AVEM: arbeitsbezogene Verhaltens- und Erlebnismuster (Schaarschmidt & Fischer, 1996) Modell an *N* = 478 Lehrern erhoben





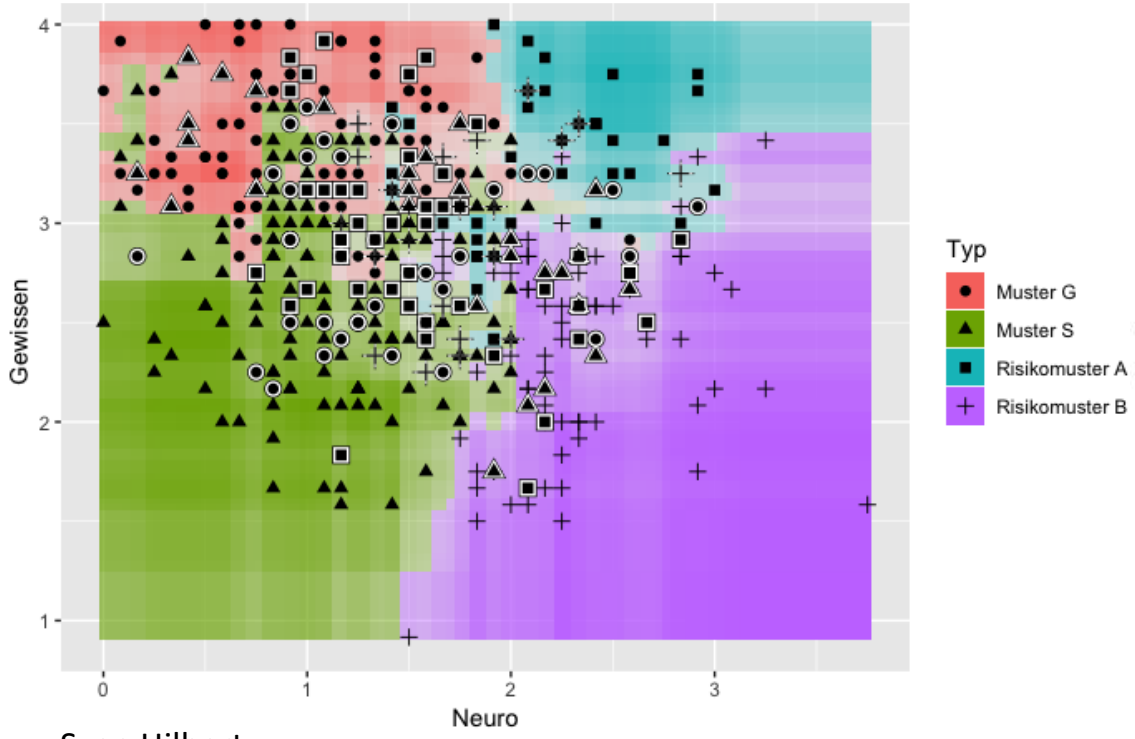
➢Versuch der Vorhersage durch die Big Five

Vorhersage und Erklärung bei Persönlichkeitstypen

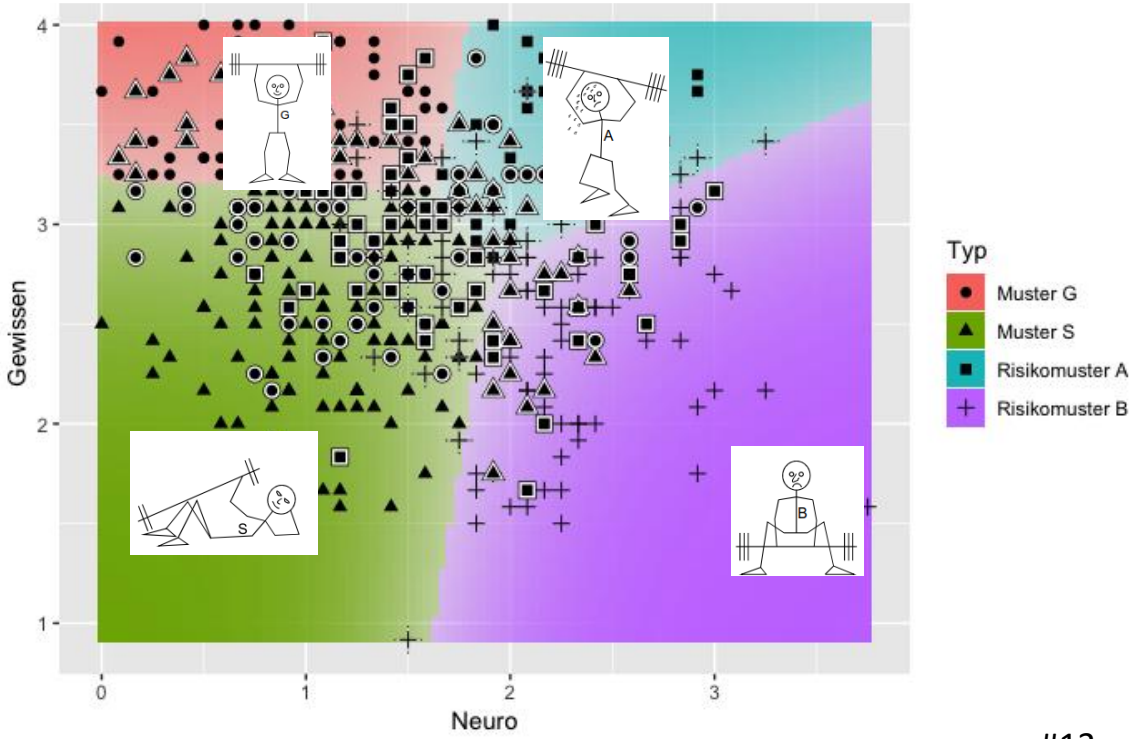
Universität Regensburg

Zwei Modellierungen: **Random Forest** und **Regressionsmodell**

Modell mit höchster Vorhersagekraft



Modell mit bester Interpretierbarkeit



Verbindung psychologische Forschung und Machine Learning

- Verständnis beider Felder
- Theoriestärke der Psychologie nutzen
  - Preprocessing, besonders Feature Extraction
  - ➢ **Interpretable Machine Learning** betreiben
- Reines Fokussieren auf die Vorhersage bringt Probleme mit sich
  - Das Phänomen wird nicht erklärt
  - Extrapolieren ist bei reiner Prädiktion nicht möglich
- Eine solide Theorie wird als Grundlage für interpretierbare (oft auch für funktionierende) Algorithmen benötigt

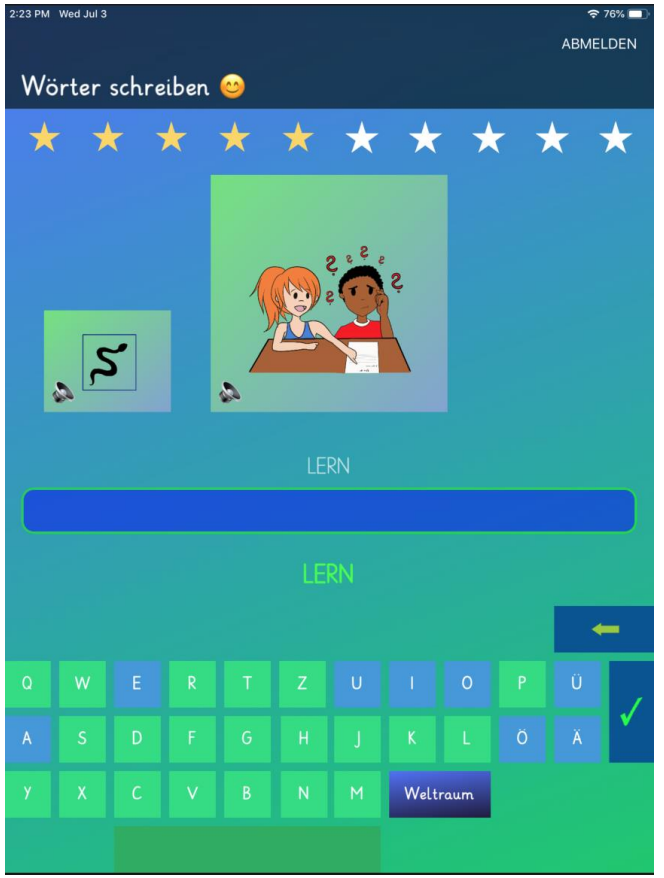
https://www.britannica.com/biography/Wilhelm-Wundt

Beispiel 3: Schreiben Lernen

- Datenerhebung mit iPads
  - Big Data
    - Jeder einzelne Buchstabe wird geloggt (586 Wörter, 2828 Buchstaben, ca. 900 Kinder)
    - Jede Reaktionszeit wird geloggt
  - Demografische Variablen
    - Geschlecht, Alter, Muttersprache etc.
  - Motivationsitems
- Wörter werden automatisch vorgegeben
- Leistung der Kinder wird online geloggt

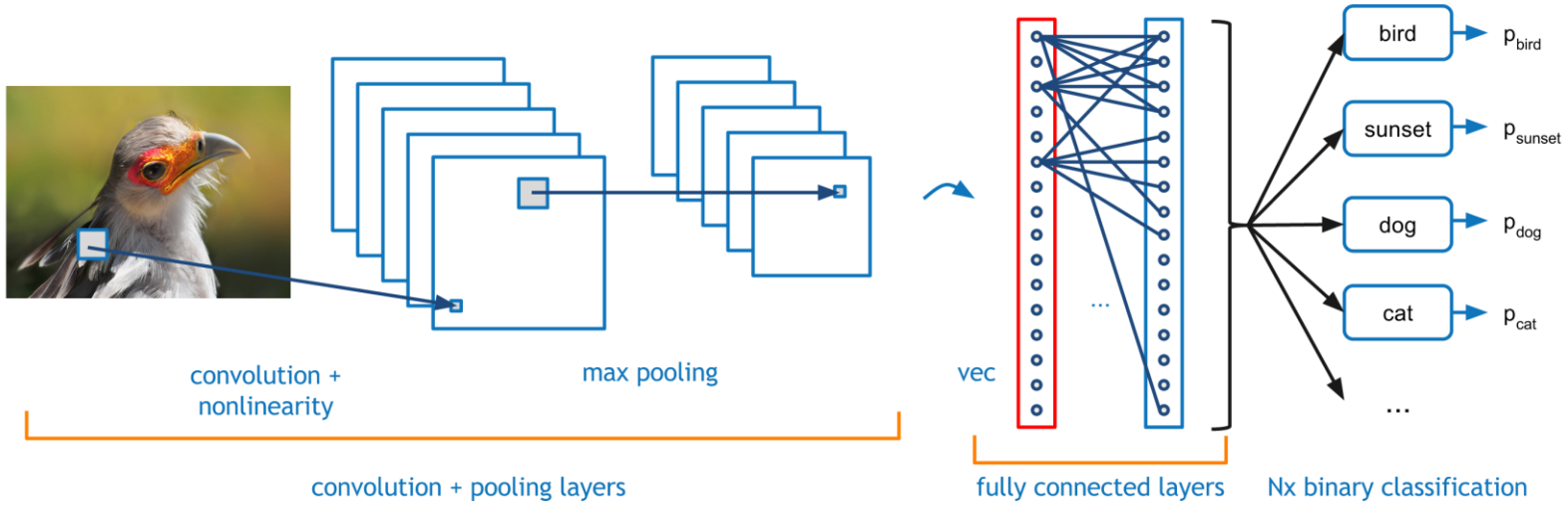
App: Schreiben lernen

- **Ziel:** Erstellung eines Algorithmus zur Vorhersage von Schwächen bei spezifischen Rechtschreibregeln
  - Algorithmus kann genutzt werden, um Kindern die korrekten Items vorzulegen, um an ihren spezifischen Schwächen zu arbeiten
  - Hierbei können Motivation und demografische Variablen einbezogen werden
    - Items sollten nicht zu leicht und nicht zu schwer sein, was allerdings auch von der Motivation beeinflusst werden kann



Algorithmus: Schreiben lernen

- Als Algorithmus wird ein konvolutionales Neuronales Netzwerk (cNN) genutzt
  - Konvolutional wird das Netzwerk aufgrund der Zusammenfassung von Einheiten zu weniger differenzierten Ebenen
    - ➢ Einbezug von verschiedenen Features (Buchstaben, Graphemkategorien, motivationale Variablen etc.)



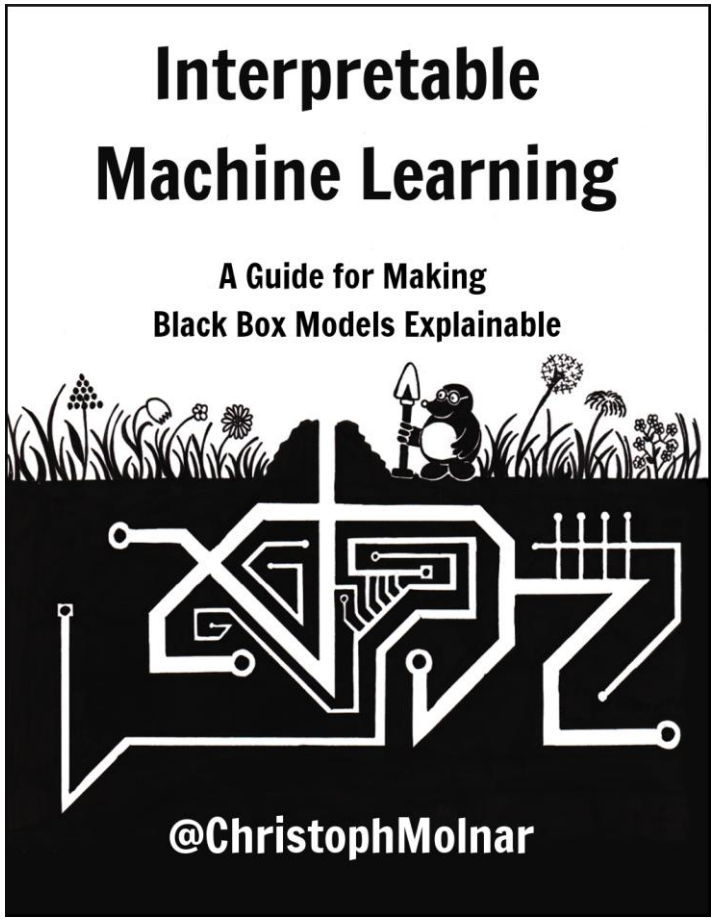
Feature Engineering: Schreiben Lernen

- **Features** müssen vorher gewählt werden
  - Demografische Variablen
  - Motivationsfragen
- **Theoretisches Wissen** zu Rechtschreibschwächen muss implementiert werden
  - Einteilung der Wörter in Grapheme
  - Kategorisierung und Zuordnung der Grapheme zu Rechtschreibregeln
- Algorithmus: Vermeidung negativer **Feedback Loops**
  - Schwache Kinder nicht noch schwächer machen

Verletzung		Wort 1	Wort 2	Wort 3	Wort 4
Regel 1	Graphem 1	X			X
Regel 2	Graphem 2		X		
Regel 1	Graphem 3			X	
Regel 2	Graphem 4			X	X

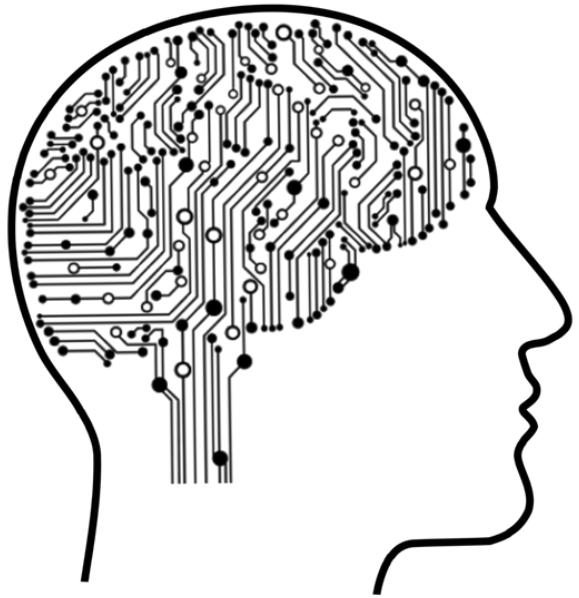
Zusammenfassung: Verbindung der Felder

- Die Vorteile von Big Data nutzen und die Probleme kennen
- **Theoriestärke** der Psychologie nutzen
  - Feature Engineering
  - Interpretation der Modelle
- **Vorhersage** mit **Beschreibung** und **Erklärung** verbinden
- **Integration** von Machine Learning in Lehre und Forschung
  - Zumindest rudimentäres Verständnis
- Weiterer wichtiger Punkt: **Open Science**



Universität Regensburg

Vielen Dank



Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Universität Regensburg

Literatur

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*(Vol. 5). Cambridge University Press.

Schaarschmidt, U., & Fischer, A. (1996). *AVEM: arbeitsbezogene Verhaltens- und Erlebnismuster*. Swets Test Services.

Stachl, C., Hilbert, S., Au, J. Q., Buschek, D., De Luca, A., Bischl, B., ... & Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701-722.

Universität Regensburg

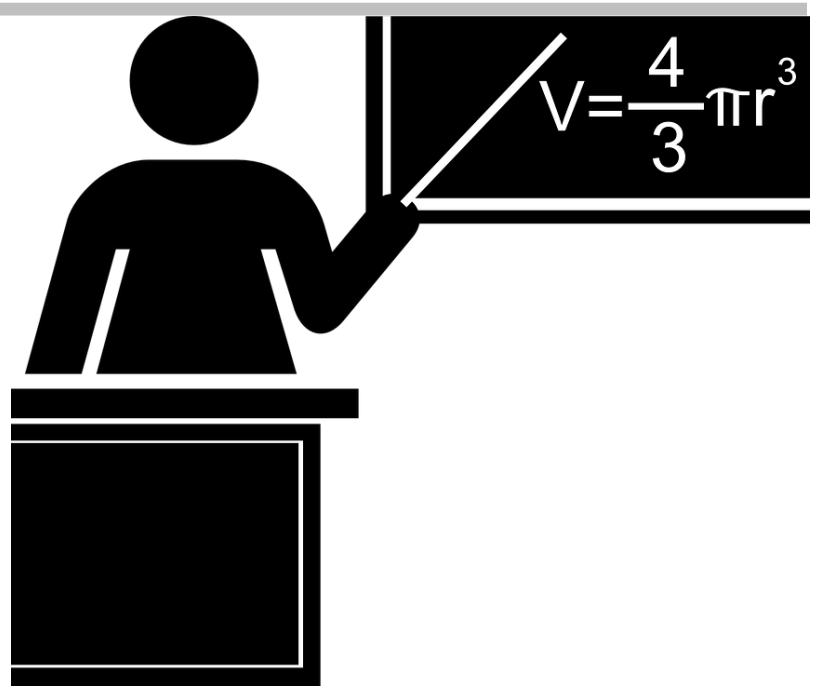
Appendix

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

COACTIV Sekundäranalyse

Universität Regensburg

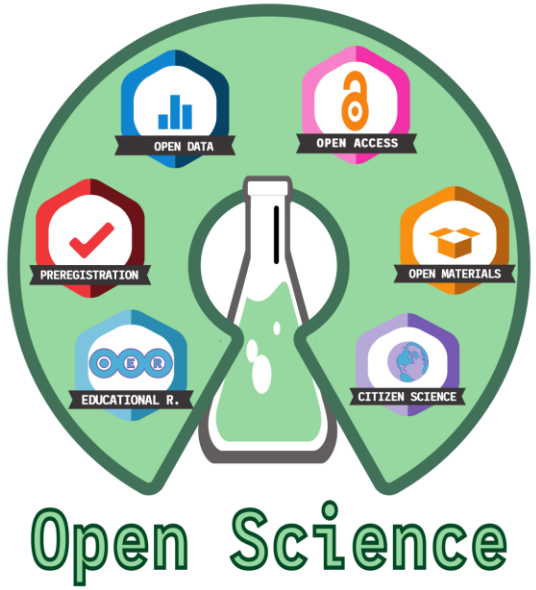
- Über 100 Artikel
- Mehr als zehn verschiedene Modelle
- Über 20 verschiedene Kontrollvariablen
- Stabilität der Modelle steht zur Disposition
  - Feature Engineering
  - Resampling



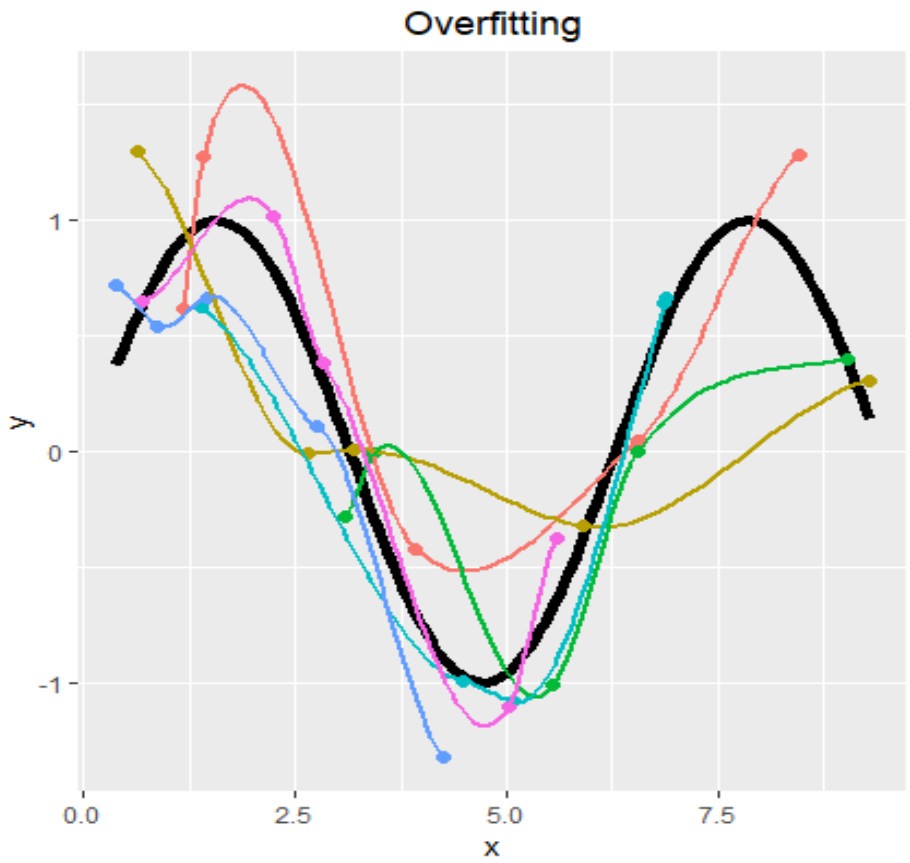
Open Science und Big Data

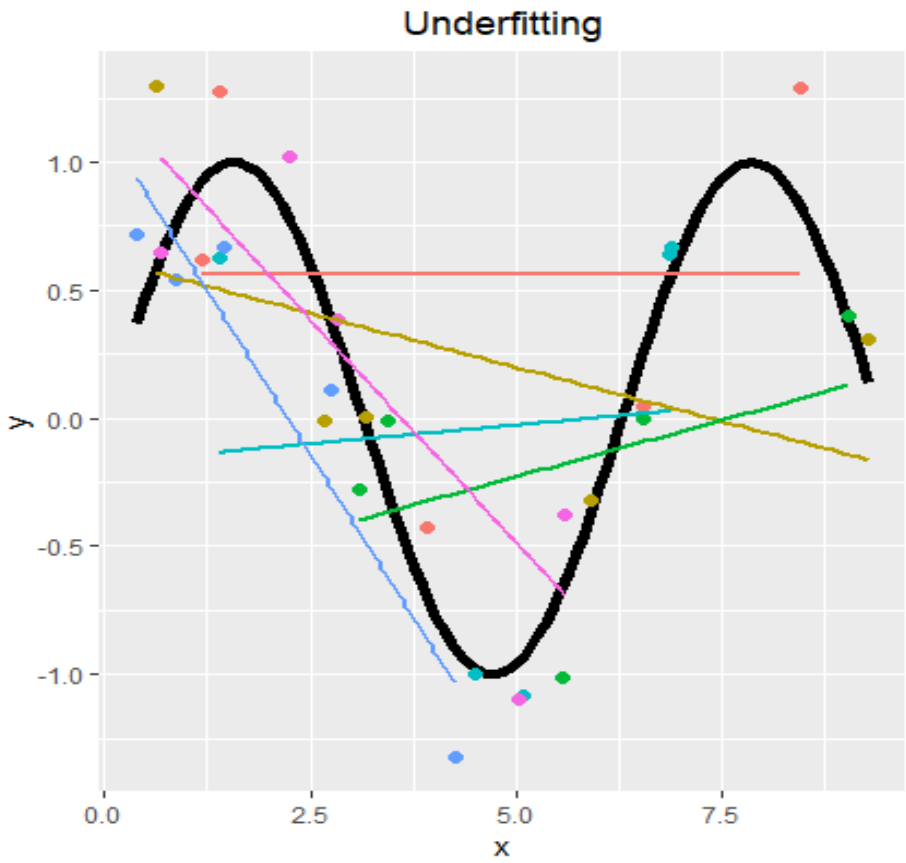
Universität Regensburg

- Open Science ist ein wichtiger Part einer glaubwürdigen psychologischen Forschung
- Bereitstellung der Daten in vielen Bereichen von Big Data allerdings extrem sensibel
  - Browserdaten, Smartphones, Smartwatches, Haushaltsgeräte
- Möglichkeit, der Aggregation von Daten vor der Versendung an den Auswertungsserver
  - Aggregation von Daten auf dem Endgerät
    - Kann suffiziente Statistiken liefern
  - Pseudonymisierung von Inhalten
    - Muss nur Isomorphie beachten

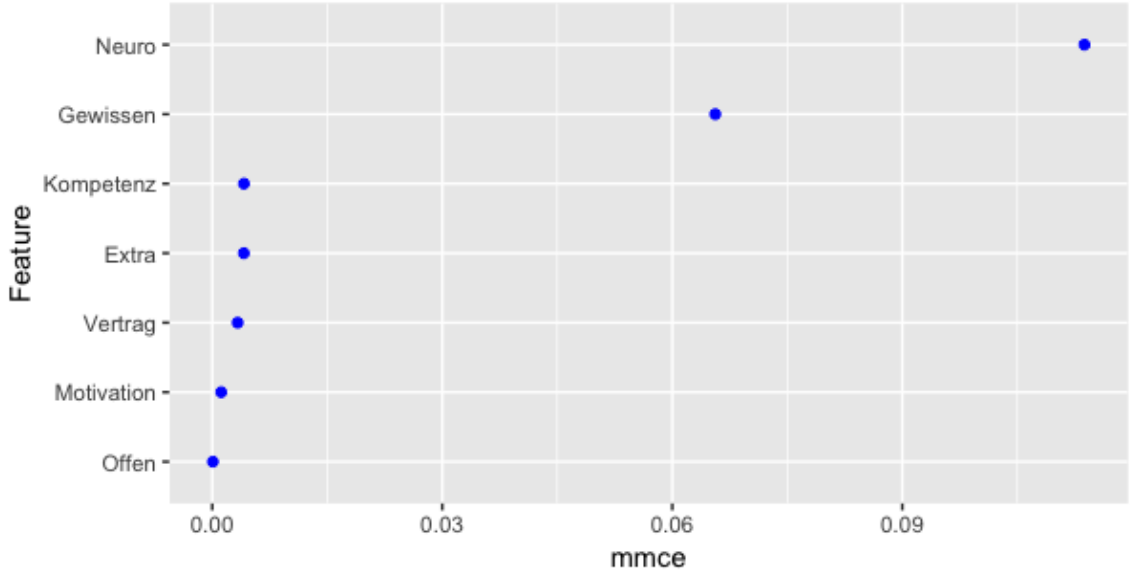


Over- und Underfit

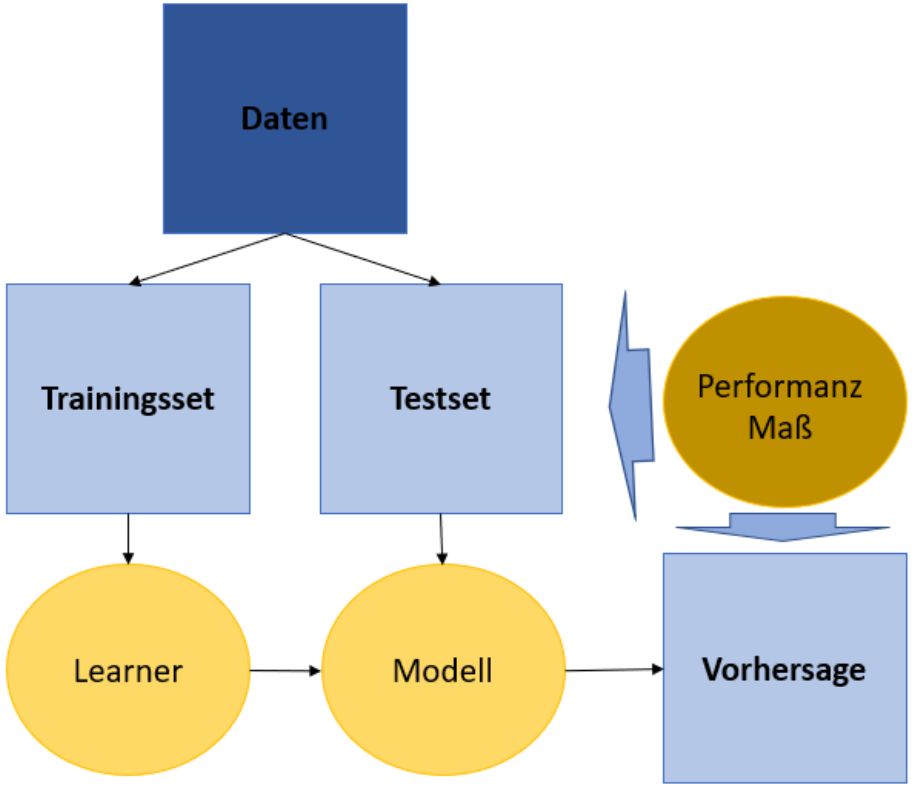




Variable Importance Plots



Resampling


Daten
Trainingsset
Testset
Performanz Maß
Learner
Modell
Vorhersage