Universität Regensburg

Educational Data Science

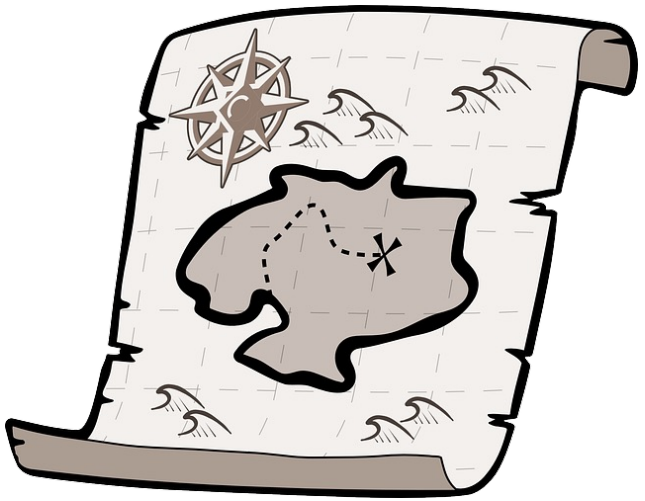
Nachwuchs- und Vernetzungstag ‚Lehrerinnen- und Lehrerbildungsforschung‘

Sven Hilbert

Themenfelder

Inhalte des Vortrags

1. Einführung Themenfelder
2. Kleiner Überblick Machine Learning & Big Data
3. Definitionen von Lernen
4. Als Disziplinen voneinander lernen
5. Zusammenfassung und Ausblick



Kleiner Überblick Machine Learning & Big Data

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

Drei Phasen nach Efron & Hastie (2016)

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**

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

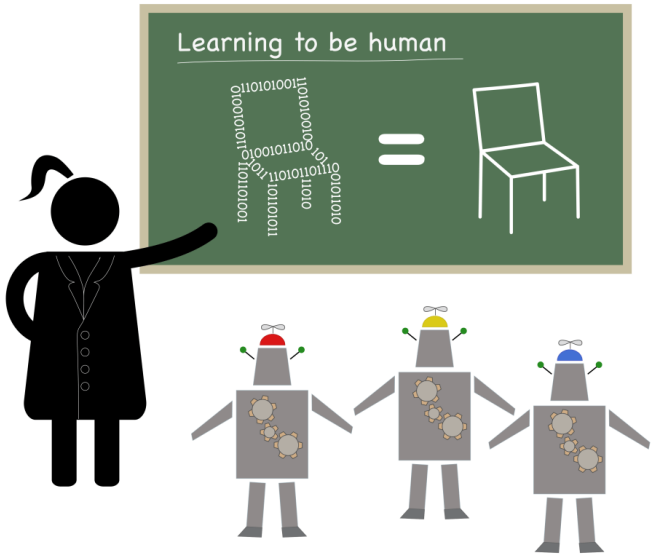
Definitionen von Lernen

Lernen in der Bildungsforschung

Universität Regensburg

Einführung in Grundbegriffe und Grundfragen der Erziehungswissenschaft (Treml, 2002):

- Grundbegriff der Pädagogik
- Wie alle Grundbegriffe, die dazu auch noch in der Alltagssprache vorkommen, vieldeutig
- Von einer einheitlichen, allgemein anerkannten und präzisen Definition kann man auch heute noch nicht sprechen



Hasselhorn & Gold (2013):

- Prozess, bei dem es zu überdauernden Änderungen im Verhaltenspotenzial als Folge von Erfahrungen kommt

Lernen in Data Science

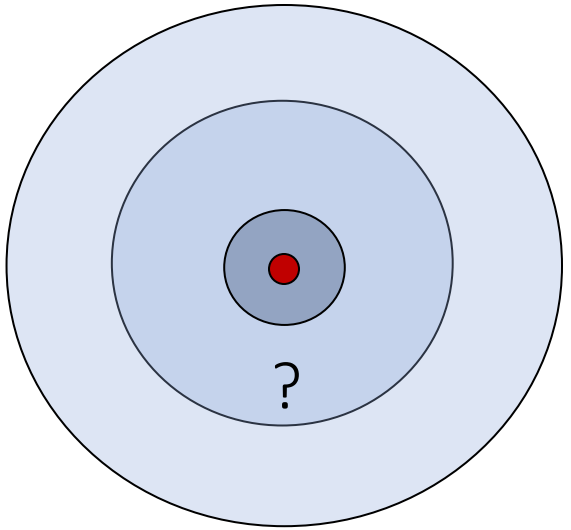
Definition aufbauend auf Domingos (2012):

1. **Repräsentation**
   - Ein Modell der Datenstruktur (z.B. LM)
2. **Evaluation**
   - Eine Funktion, um die Güte des Modells zu bestimmen (z.B. Erklärte Varianz)
3. **Optimierung**
   - Methode, um die Anpassung der Modellparameter zu determinieren, welche zu einer stärkeren Datenanpassung führen (z.B. Gradient Descent)

Lernen in Data Science

1. Repräsentation

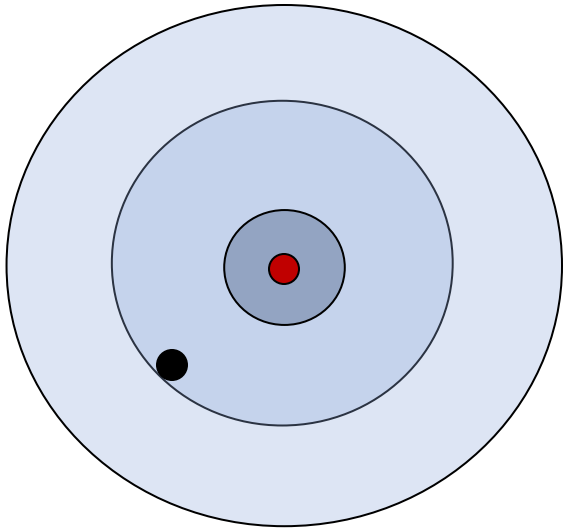
- Ein Modell wird aufgestellt um die Frage zu beantworten, wie die Daten zustande kommen
- Dabei wird definiert, welche Werte vorhergesagt werden
- Ein Bereich möglicher Werte ist hierbei Teil des Modells



Lernen in Data Science

1. Repräsentation

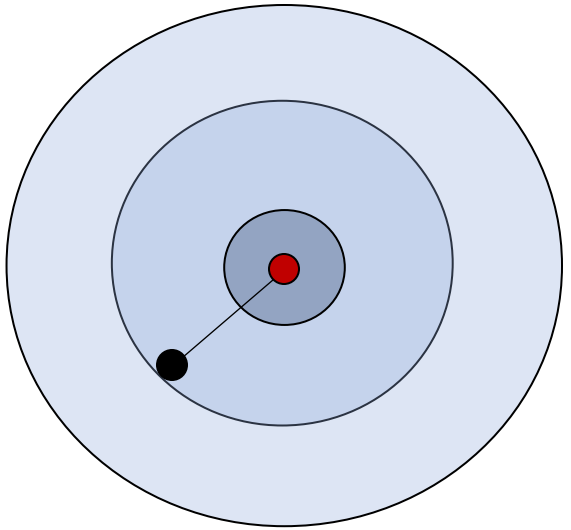
- Ein Modell wird aufgestellt um die Frage zu beantworten, wie die Daten zustande kommen
- Dabei wird definiert, welche Werte vorhergesagt werden.
- Ein Bereich möglicher Werte ist hierbei Teil des Modells
- Beispiel: $\hat{Y} = \alpha + \beta X$



Lernen in Data Science

2. Evaluation

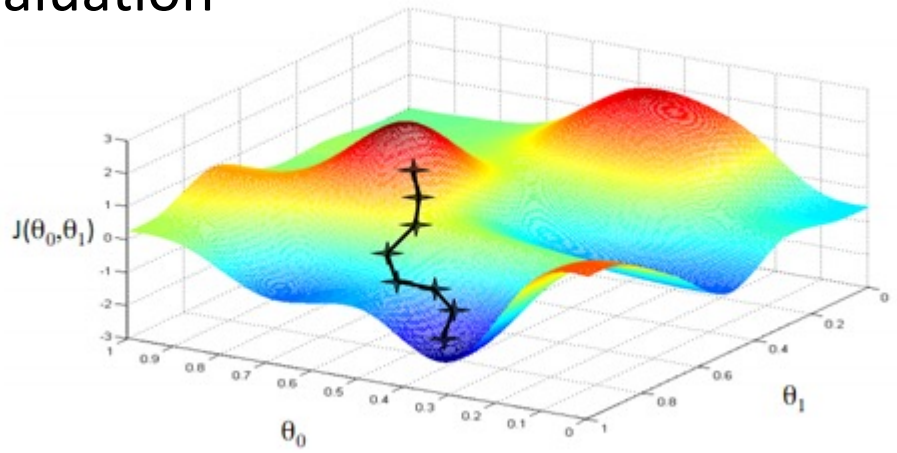
- Eine Vorschrift wird definiert, nach der die Distanz der Vorhersage zu den empirischen Werten berechnet wird
- Hierbei wird die Distanz der einzelnen vorhergesagten Werte zu ihren jeweiligen empirischen Pendants definiert und daraus ein einzelner ‚Abstandswert' gebildet
- Beispiel: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

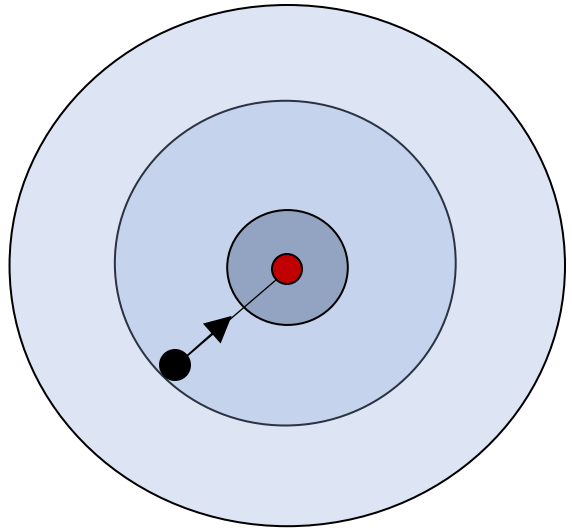


Lernen in Data Science

3. Optimierung

- Eine Vorschrift wird definiert, nach der die ‚Richtung' zum Vorhersageziel gewählt wird
- Ein Beispiel ist die Nutzung von Steigungen in Funktionen zur Modellevaluation
- Beispiel:

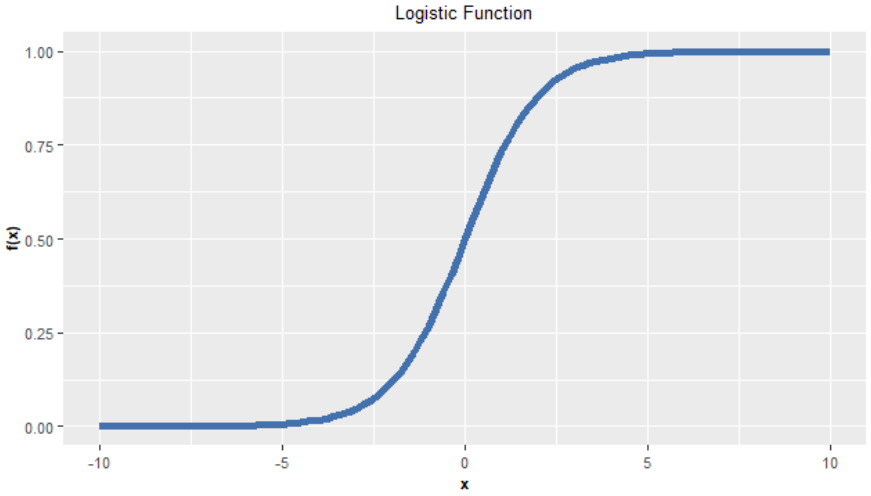




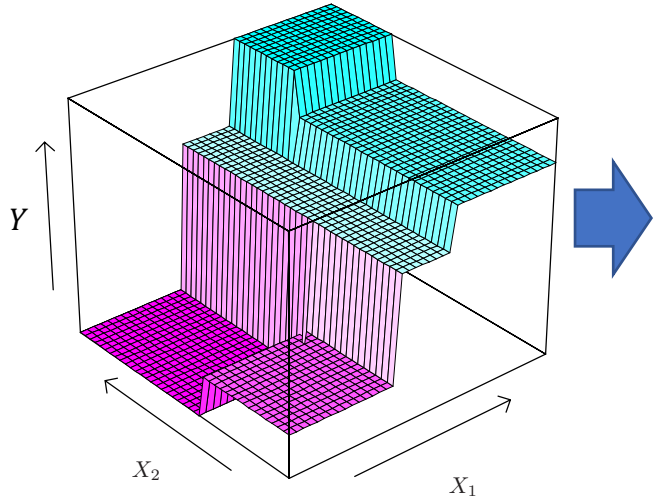
Als Disziplinen voneinander Lernen

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



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



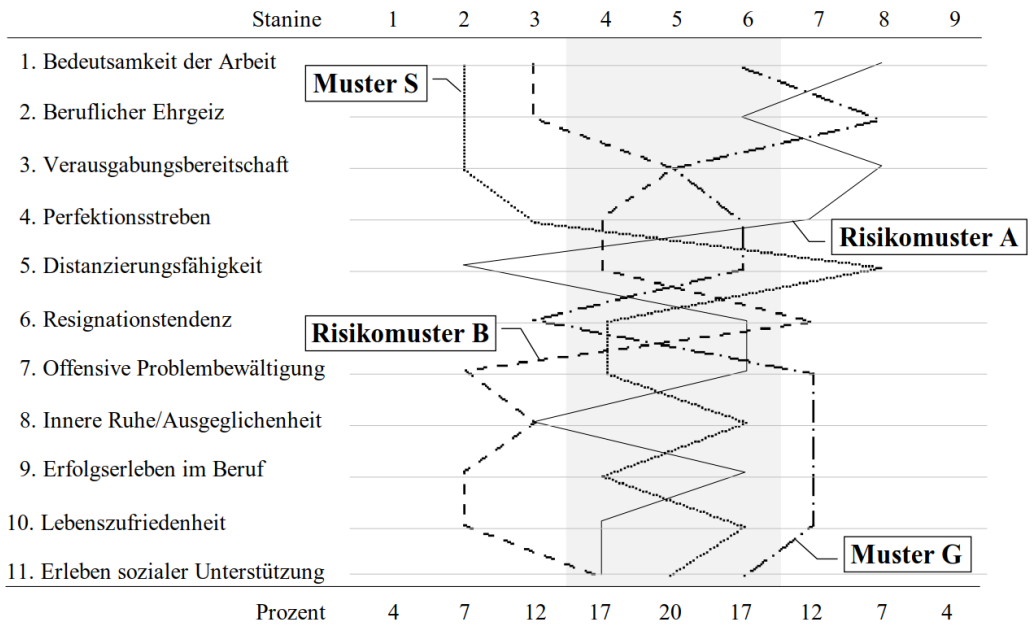
?

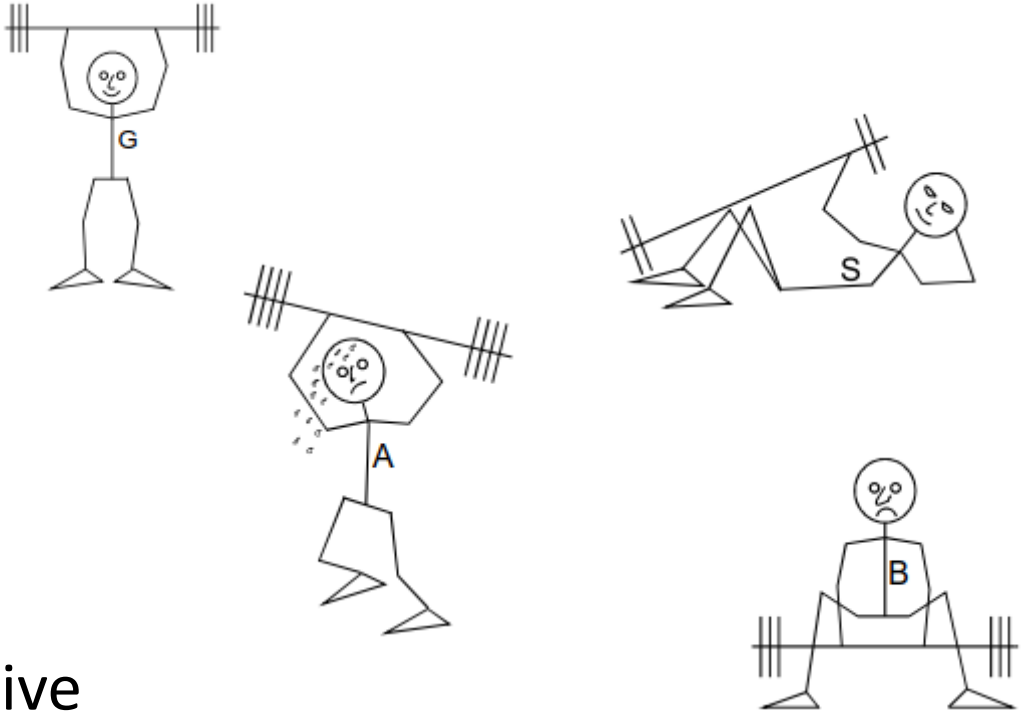
Grafik: Efron, B., & Hastie, T. (2016). *Computer age statistical inference*(Vol. 5). Cambridge University Press., S. 125

Beispiel: Studie Persönlichkeitstypen

Universität Regensburg

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





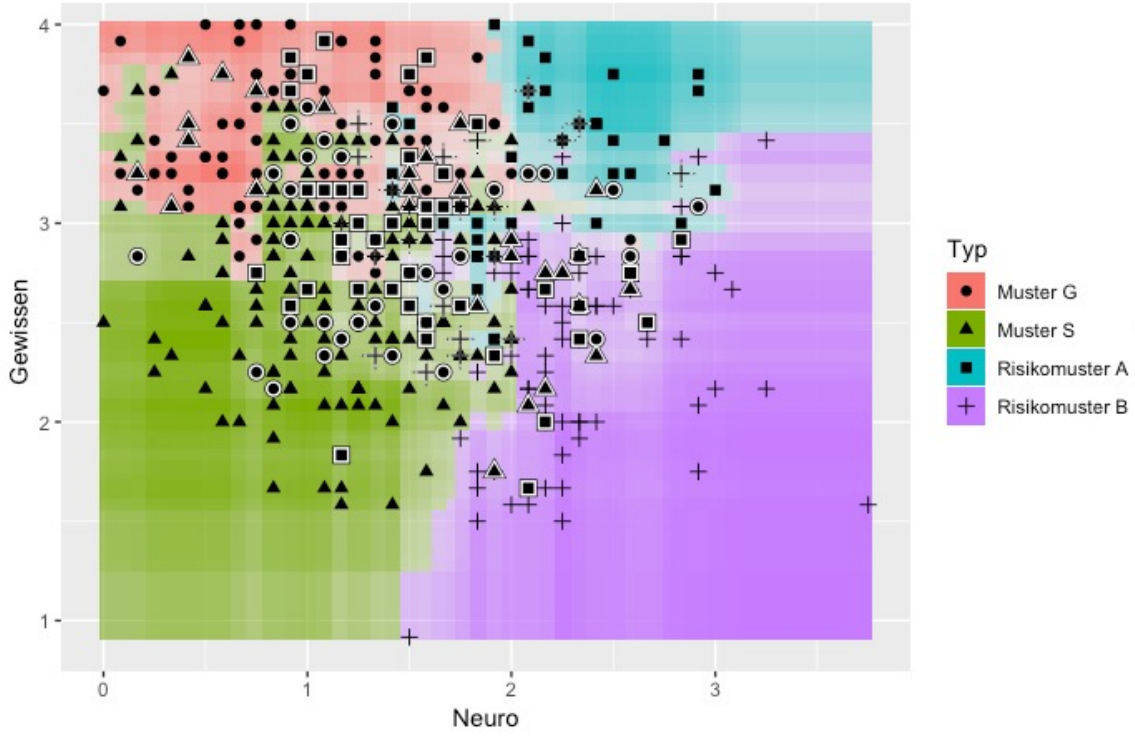
➢ Versuch der Vorhersage durch die Big Five

Vorhersage und Erklärung bei Persönlichkeitstypen

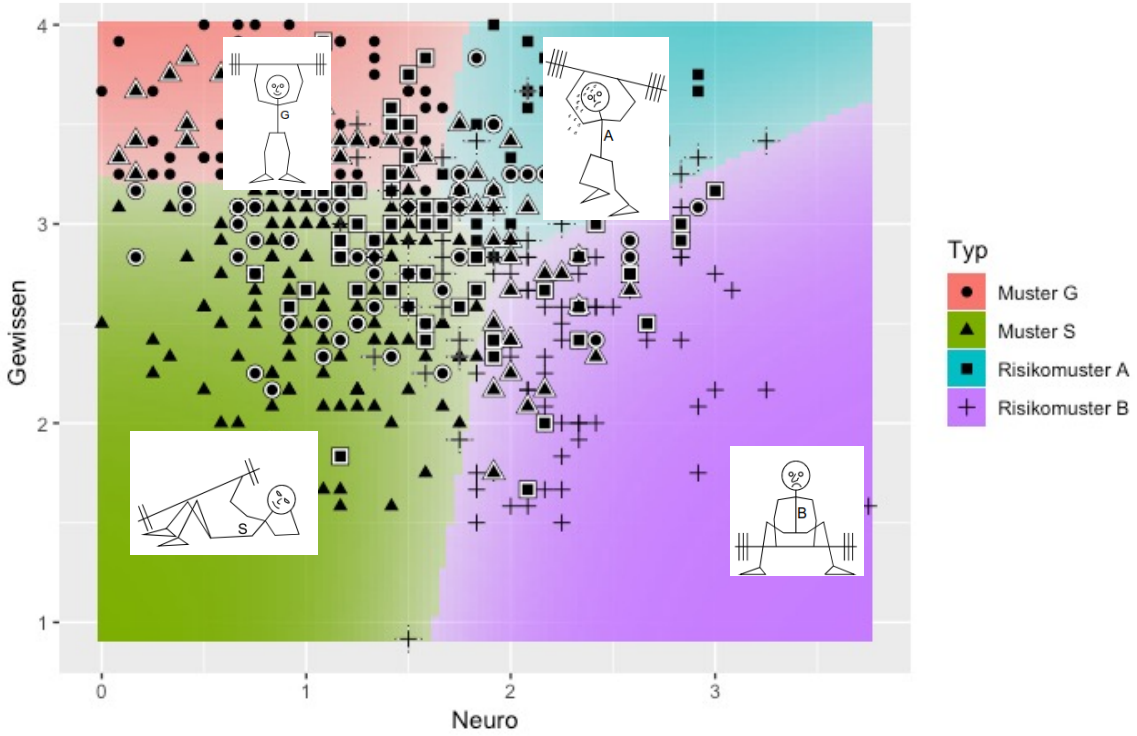
Universität Regensburg

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

Modell mit höchster Vorhersagekraft

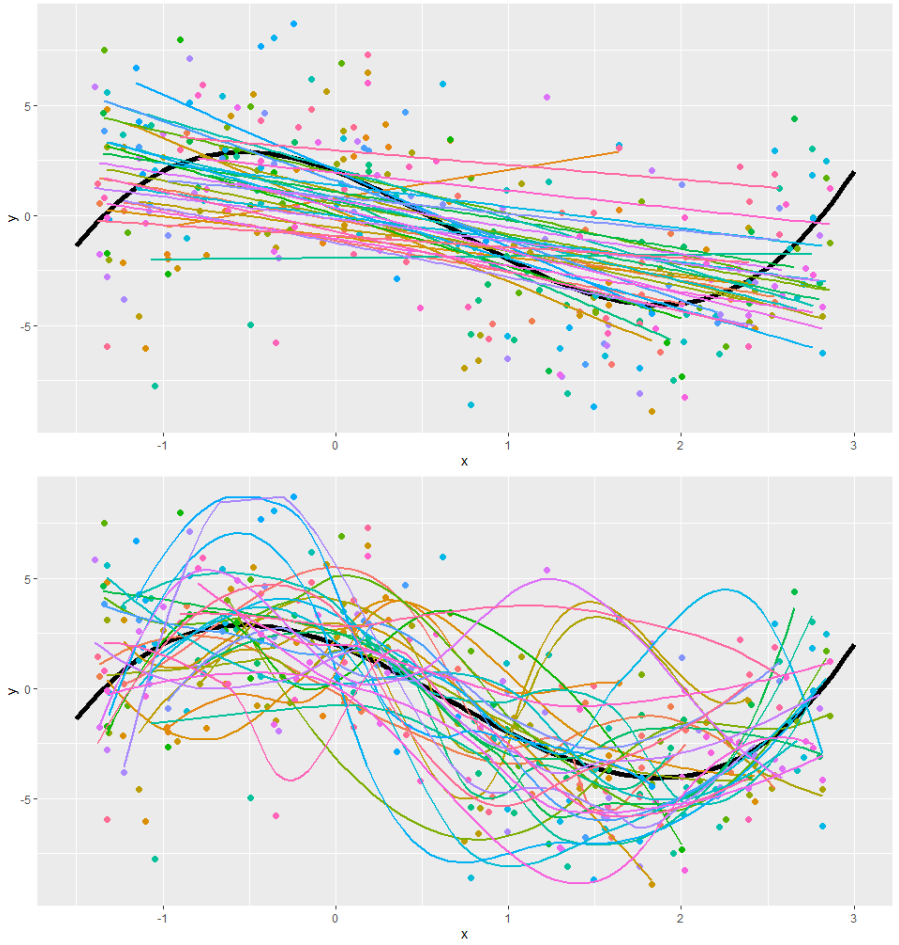


Modell mit bester Interpretierbarkeit



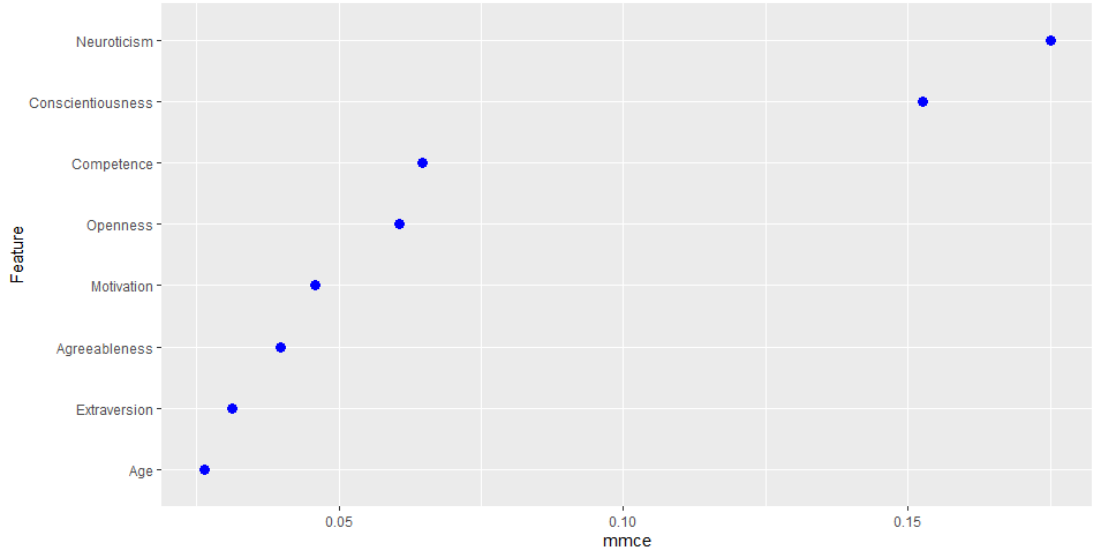
Bias-Variance Decomposition

- Das Signal erkennen
- Mehr als nur eine Methode für die Bildungswissenschaften
  - Eine Philosophie
- Aber oft geht es nicht vorrangig um Lernen, sondern um Lehren



Bias-Variance Decomposition

- Das Signal erkennen
- Mehr als nur eine Methode für die Bildungswissenschaften
  - Eine Philosophie
- Aber oft geht es nicht vorrangig um Lernen, sondern um Lehren
  - ➢Interpretable Machine Learning
  - ➢Explainable AI



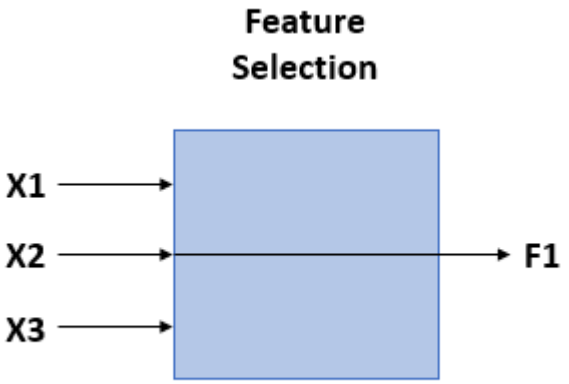
Beispiel: Phone Study

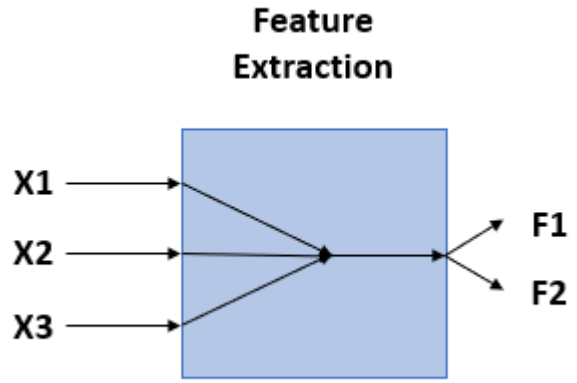
- Vorhersage der Nutzung von Smartphone Apps mit Persönlichkeitsfaktoren und –facetten
- **Extraversion** (besonders Geselligkeit) sagt die Häufigkeit von Anrufen und Kamera positiv vorher
- **Gewissenhaftigkeit** sagt die Häufigkeit der Nutzung von Spielen vorher
- **Verträglichkeit** sagt die Häufigkeit der Nutzung von Transport Apps vorher
- *Feature Engineering*
  - Labeling der Kategorien
  - Kategorisierung der Einzelnutzungsdauer
    - Bis zum Start des nächsten Nutzungsevent, das keine Bloatware ist

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





Zusammenfassung und Ausblick

Verbindung Bildungsforschung und Data Science

- Verständnis beider Felder
- Theoriestärke der Bildungsforschung nutzen
  - Preprocessing, besonders Feature Extraction
  - ➢ **Interpretable Machine Learning**
- 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
- Bildungsforschung und Data Science müssen **voneinander lernen**

Etablierung notwendiger Infrastruktur

Universität Regensburg

Entwicklung und und Bereitstellung von Kompetenzen

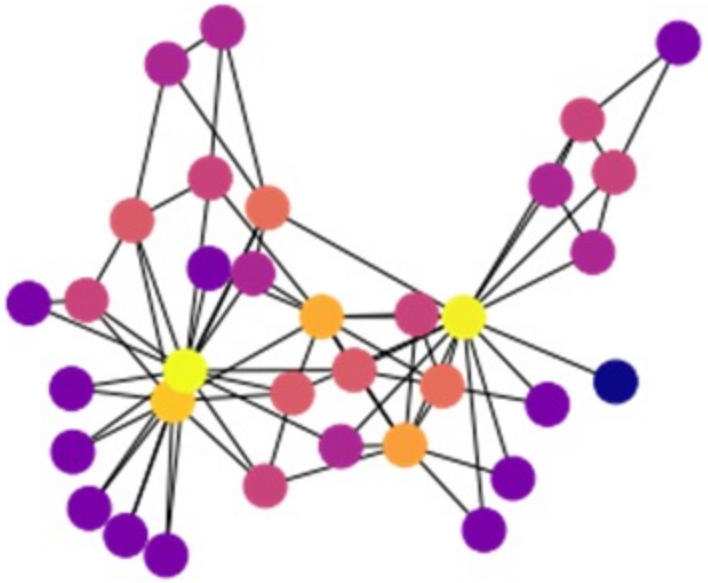
- Softwareentwicklung und Wartung von Programmen
- Online Erhebungstools
  - Fragebögen
- Lehr- und Lernplattformen
  - APIs, Container
- Medienschulung

Anschaffung von Technik

- Leistungsfähige Prozessoren
- Server zur Datenspeicherung und Analyse

Achtung des Datenschutz

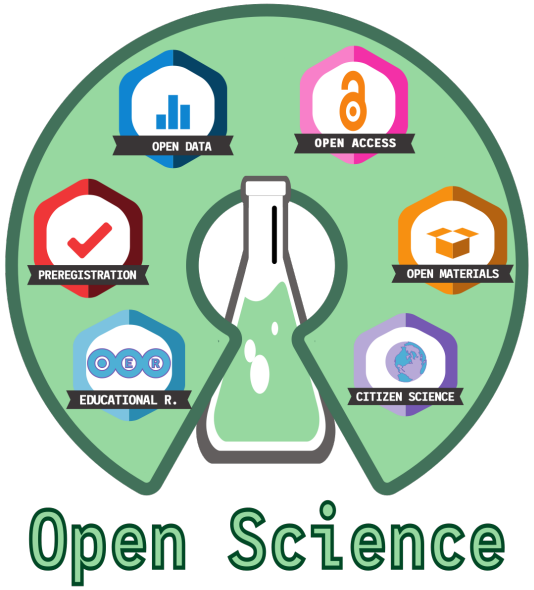
- Datenschutzbeauftragte
- Tragfähige Datenschutzkonzepte
- Aufklärung und Ausbildung in Data Literacy



Open Science und Big Data

Universität Regensburg

- Open Science ist ein wichtiger Part einer glaubwürdigen empirischen 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



Some final thoughts

McKinsey Global Institute (2011)

- Die Datenmenge in unserer Welt explodiert. Unternehmen erfassen **Billionen von Bytes an Informationen** über ihre Kunden, Lieferanten und Betriebe. Big Data – große Datenpools, die erfasst, kommuniziert, gespeichert und analysiert werden können – sind heute **Teil aller Sektoren und Funktionen der globalen Wirtschaft**.

➢ Vor diesem Hintergrund wird Machine Learning ein **Motor der nächsten Innovationswelle** sein.

Pedro Domingos (2012)

- In der letzten Dekade hat der Nutzen von Machine Learning sich **weit über die Computerwissenschaften hinaus** verbreitet
- Machine Learning wird genutzt in der Websuche, Spam Filtern, Werbung, Kreditwürdigkeitsschätzung, Betrugsdetektion, Recommender Systemen, Aktienhandel, Medizin und unzähligen weiteren Bereichen

Literatur

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87.

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

Hasselhorn, M., & Gold, A. (2022). *Pädagogische Psychologie: Erfolgreiches Lernen und Lehren*. Kohlhammer Verlag, Stuttgart.

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., ... & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, *9*(3), e3310.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

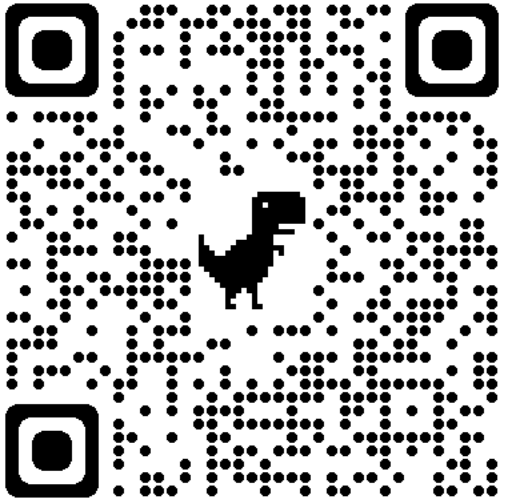
Schaarschmidt, U., & Fischer, A. (1996). *AVEM: arbeitsbezogene Verhaltens- und Erlebnismuster*. Swets Test Services.

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.

Treml, A. K. (2002). Lernen. In *Einführung in Grundbegriffe und Grundfragen der Erziehungswissenschaft* (pp. 93-102). VS Verlag für Sozialwissenschaften, Wiesbaden.

Universität Regensburg

Vielen Dank



Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., ... & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, *9*(3), e3310.

.