

# Einführung in den Themenbereich Datenbeschaffung und -aufbereitung

Erarbeitet von  
Dr. Ann-Kathrin Selker

|                         |   |
|-------------------------|---|
| <b>Lernziele</b> .....  | 1 |
| <b>Inhalt</b> .....     | 2 |
| Einstieg .....          | 2 |
| Was dich erwartet ..... | 2 |
| Abschluss .....         | 3 |
| <b>Disclaimer</b> ..... | 3 |

## Lernziele

- Du kannst Schritte nennen, die beim Sammeln, Bereinigen und Aufbereiten von Daten wichtig sind

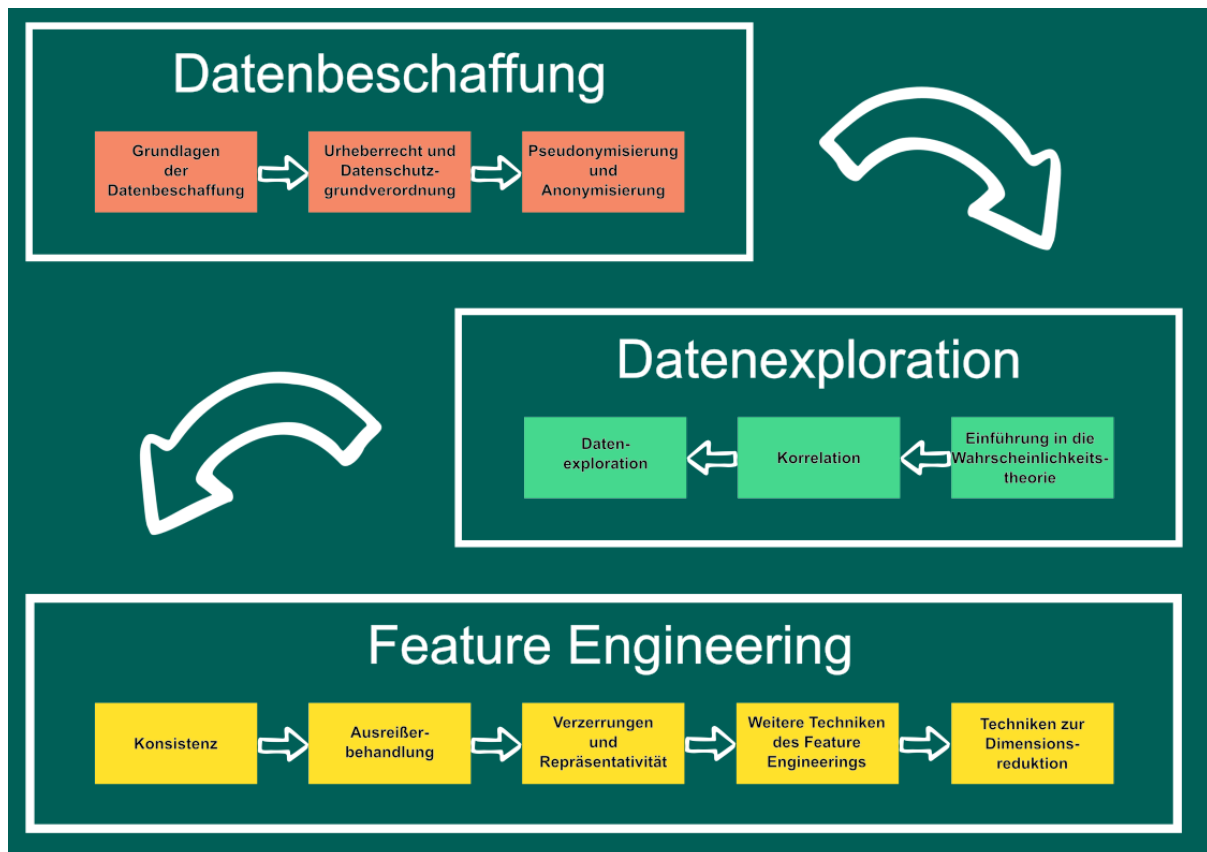
## Inhalt

### Einstieg

Der Erfolg eines Machine Learning Prozesses steht und fällt mit den Daten. Schließlich bilden die Daten die Grundlage, aus denen das Machine-Learning-Modell lernt und dann später seine Aufgabe erledigen kann. Damit ein Modell richtig aus Daten lernen kann, müssen sie gesammelt, bereinigt und auf die gewünschte Aufgabe zugeschnitten werden. Das ist mit Abstand die zeitaufwändigste Arbeit, die anfällt! Viele (wenn nicht sogar die meisten) Fehler passieren allerdings schon hier: Wenn Daten fehlen, zu einseitig sind oder gewisse Voreingenommenheiten beinhalten, wirkt sich das auf den gesamten Lernprozess und vor allem auf den Lernerfolg aus. In diesem Themenabschnitt beschäftigen wir uns daher ausschließlich mit Daten – von der Beschaffung über die Exploration bis hin zur Aufbereitung. Dabei zeigt sich eine Sache ganz besonders: Fehler bei der Datensammlung, -bereinigung und -aufbereitung kommen ständig vor, ob durch unsauberes Arbeiten, Probleme bei der Sammlung oder nicht beachteten Stolperfallen.

### Was dich erwartet

Damit dir das nicht auch passiert, geben wir dir einen Überblick über die relevanten Themen rund um das Thema Daten. Zuerst beschäftigen wir uns mit dem Beschaffen von Daten. Du erfährst, welche Datenquellen es gibt und worauf du beim Sammeln achten musst. Dabei spielen natürlich auch rechtliche Voraussetzungen eine Rolle. Dafür tauchen wir in die Welt des Urheberrechts und der Datenschutzgrundverordnung ein. Weiter geht es dann mit der explorativen Datenanalyse. Du siehst, mit welchen Techniken du den Überblick über deine Daten gewinnst, um das meiste aus ihnen herauszuholen. Sobald du dich in deinen Daten so gut auskennst wie in deiner Westentasche, gehen wir über zum letzten Schritt: Dem Feature Engineering. Feature Engineering ist ein mathematischer Prozess, der auf Konzepten aus der Statistik und Wahrscheinlichkeitstheorie aufbaut. Darunter fallen alle Aufbereitungsschritte, bei denen Daten verändert werden. Dies beinhaltet neben Fehlerbereinigungen und Ausreißerbehandlung auch zum Beispiel die Dimensionsreduktion, bei der die Anzahl der Features durch Entfernen oder Kombinieren reduziert wird.



## Abschluss

Im Folgenden zeigen wir dir nicht nur, was alles schiefgehen kann, sondern geben dir auch Hinweise, wie du richtig mit deinen Daten umgehst. Los geht's!

## Disclaimer

Transkript zu dem Video „04 Datenbeschaffung und -aufbereitung: Einführung in Datenbeschaffung und -aufbereitung“, Ann-Kathrin Selker.

Dieses Transkript wurde im Rahmen des Projekts ai4all des Heine Center for Artificial Intelligence and Data Science (HeiCAD) an der Heinrich-Heine-Universität Düsseldorf unter der Creative Commons Lizenz [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/) veröffentlicht. Ausgenommen von der Lizenz sind die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.