**Einheit 1 – Prozessmodell: Data Wrangling**

* Hier: Abb\_1\_1\_Titel\_Prozessmodell.jpg –

**Was ist *Data Wrangling?***

Im Rahmen von Studien, Umfragen, Experimenten etc., erheben wir Daten, um sie anschließend bezogen auf eine Fragestellung zu analysieren.

In Statistikkursen lernen Sie verschiedene statistische Analysemethoden kennen und anwenden. Hier werden meist Daten verwendet, die bereits vorbereitet sind. So könnte der Eindruck entstehen, dass Sie nach einer Erhebung gleich mit der Analyse loslegen können.

In der "echten" Welt sind Daten allerdings chaotisch. Sie sind oft unvollständig (Angaben fehlen, in Datensätzen von Dritten liegen sie eventuell nur in aggregierter Form vor), noch nicht bereinigt (der Datensatz enthält Fehler oder Ausreißer) oder inkonsistent (wie Diskrepanzen bei Codes oder Bezeichnungen, z.B.: wenn Alter und Geburtsjahr nicht zusammenpassen).

Je nachdem, was wir mit den Daten vorhaben, kann es sein, dass unsere Daten aus unterschiedlichen Quellen kommen. Hier haben die Forschenden vielleicht unterschiedliche Dateiformate, Bezeichnungen oder Ratingskalen verwendet.

Bevor wir Rechnungen vornehmen können, um unsere Hypothesen zu überprüfen, müssen wir unsere Daten aufräumen. Wir müssen sie sortieren, bereinigen, organisieren - Doppelungen entfernen, fehlende Werte ersetzen oder ausschließen und/oder vorab Berechnungen vornehmen, mit denen wir später arbeiten können. Anschließend müssen wir gewissenhaft überprüfen, ob in diesem Prozess etwas schiefgegangen ist, unser Ordnungssystem plausibel und verständlich ist und ob die Daten noch immer die Realität widerspiegeln oder eben verfälscht wurden. Dafür sind auch inhaltliche Überlegungen nötig. Bei der Aufbereitung Ihrer Daten müssen Sie darüber nachdenken, was hinter diesen Daten steckt, was Sie mit der Erhebung beabsichtigt haben und welche Qualität die Daten haben, mit denen Sie arbeiten wollen.

Das Sprichwort "Garbage in, garbage out", welches ursprünglich verwendet wurde um zu beschreiben, dass ein Computer nicht von sich aus gute Ergebnisse produziert, sondern gute Befehle und korrekte Eingaben benötigt, kann hier ebenfalls angewendet werden - auf der Grundlage einer mangelhaften Datenbasis können Sie keine hochwertigen Ergebnisse erzielen.

* Hier Abb\_1\_2\_Garbage\_in\_Garbage\_out.jpg –

Data Wrangling bezeichnet den Prozess der Aufbereitung von Rohdaten zur anschließenden Analyse. Je nachdem, in welchem Fachbereich Sie tätig sind, wird Ihnen dieser Prozess wahrscheinlich unter etwas anderen Namen begegnen. So werden Data Wrangling, Pre-Processing, Aufbereitung von Rohdaten, Datenaufbereitung etc., oft synonym verwendet. Im Allgemeinen ist mit Data Wrangling der Umgang mit einer größeren Menge von Daten gemeint. In diesem Kurs beschäftigen wir uns allerdings nicht gesondert mit großen Datensätzen.

**Unten sehen Sie das schematische Prozessmodell des Data Wranglings, welches im nächsten Abschnitt näher erklärt wird.**

**An diesem Modell orientieren wir uns über den gesamten Kurs hinweg.**

* Hier Abb\_1\_3\_Prozessmodell.jpg –

**Merke:**Die Aufbereitung von Daten kann einige Zeit in Anspruch nehmen, spart einem aber eine Menge Ärger. Wenn Sie nicht darauf achten, bei der Aufbereitung Ihrer Daten gewissenhaft vorzugehen, dann kann es sein, dass Sie bei der Analyse der Daten feststellen, dass Ihre Ergebnisse vielleicht keinen Sinn machen. Es kann natürlich sein, dass die Ergebnisse tatsächlich nicht interpretierbar sind, ein weiterer möglicher Grund ist aber auch eine falsche Aufbereitung der Rohdaten.

In manchen Bereichen, wie z.B. beim Data Mining (hier werden große Datenbestände auf Muster, Querverbindungen und Trends untersucht) nimmt die Aufbereitung von Daten einen Großteil (bis zu 90%) des gesamten Arbeitsprozesses ein.

* Hier neuer Abschnitt –

**Prozessmodell: Data Wrangling**

In diesem Abschnitt erhalten Sie einen kurzen Überblick über die einzelnen Schritte im Prozess der Datenaufbereitung.

1. **Überblick verschaffen**

* Hier Abb\_1\_4\_Überblick.jpg –
* Im ersten Schritt, bevor es an den tatsächlichen Umgang mit den (Roh-)Daten geht, müssen Sie sich einen Überblick verschaffen. Das gilt sowohl für den Umgang mit selbst erhobenen Daten als auch für den Umgang mit sogenannten Fremddaten.
* Es kann zum Beispiel sein, dass Sie für Ihr Forschungsprojekt, Ihre Bachelor- oder Masterarbeit oder auch einfach aus Interesse frei zugängliche Daten analysieren wollen. Gerade weil diese Daten von Dritten vorbereitet, transformiert und benannt wurden, müssen Sie sich die notwendigen Informationen beschaffen, die Ihnen dabei helfen, den Datensatz zu verstehen.
* Der Schritt "Überblick verschaffen" kann viele verschiedene Unterschritte beinhalten. Wie viele, hängt schließlich davon ab, welche Informationen verfügbar sind und wie gewissenhaft die Person(en), die den Datensatz vorbereitet haben (Data Wrangling) vorgegangen sind.
* In diesem Kurs werden Sie lernen, wie Sie einen eigenen Datensatz auf die anschließende Analyse vorbereiten - so lernen Sie außerdem, was die Ersteller\*innen eines frei verfügbaren Datensatzes wahrscheinlich vor der Veröffentlichung getan haben.

### **2. Strukturieren**

* Hier Abb\_1\_5\_Strukturieren.jpg –
* Im zweiten Schritt ordnen Sie Ihre Daten so an, dass Sie für Sie in idealer "Ordnung" vorliegen. So kann es zum Beispiel sein, dass Sie einen Fremddatensatz nutzen und Sie möchten die Bezeichnungen für die einzelnen Variablen verändern, damit Sie sie besser auf einen Blick erkennen können. Vielleicht möchten Sie auch, dass die Reihenfolge der Variablen in Ihrer Datei anders ist als im Originaldatensatz, weil Sie so einen besseren Überblick haben.
* Wie bereits erwähnt, werden Sie in diesem Kurs mit mehr oder weniger selbst erhobenen Daten umgehen. Das bedeutet, dass Sie sich z.B. sinnvolle Benennungen für die einzelnen Datenpunkte überlegen müssen, dass Sie die Anordnung der Variablen von Beginn an selbst festlegen und dass Sie entscheiden, welche Informationen in den Datensatz, der später für die Analyse genutzt werden soll, einfließen.

### **3. Bereinigen**

* Hier Abb\_1\_6\_Bereinigen.jpg –
* In einem Datensatz, ob selbst oder durch Dritte erhoben, befinden sich oft auch Variablen, die Sie z.B. nicht für Ihre Analyse benötigen (z.B. wird bei Online-Umfragen durch das Online-Tool häufig der verwendete Browser mit angegeben). Außerdem müssen Sie eventuell auch einzelne Teilnehmende aus der Analyse ausschließen, z.B. weil diese nur sehr lückenhaft geantwortet haben oder angeben, dass sie nicht wollen, dass ihre Daten verarbeitet werden.
* In diesem dritten Schritt bereinigen Sie Ihren Datensatz, sodass am Ende nur noch diejenigen Variablen enthalten sind, die Sie brauchen und mit denen Sie etwas anfangen können.

### **4. Anreichern**

* Hier Abb\_1\_7\_Anreichern.jpg –
* Oft müssen Sie Ihren Datensatz auch noch weiter anreichern, damit Sie die Analysen durchführen können, die Sie durchführen wollen.
* Sie müssen z.B. Punktzahlen in einem Test für die verschiedenen Teilnehmenden berechnen, fehlende Werte ersetzen oder eine Zusatzinformation darüber eintragen, in welcher Experimentalgruppe sich eine Person befunden hat.

### **5. Validieren**

* Hier Abb\_1\_8\_Validieren.jpg –
* Im letzten Schritt, sollten Sie überprüfen, ob die Veränderungen, die Sie an den Daten vorgenommen haben, auch plausibel sind. Können die Daten so noch genutzt werden? Ist Ihnen vielleicht ein Fehler unterlaufen? Haben Sie etwas übersehen?
* In diesem Kurs lernen Sie verschiedene Möglichkeiten kennen, mit denen Sie die Plausibilität Ihrer vorangegangenen Schritte überprüfen können.

### **Dokumentation**

* Hier Abb\_1\_2\_Garbage\_in\_Garbage\_out.jpg -
* Alle Schritte, die Sie bei der Datenaufbereitungen durchlaufen, sollten dokumentiert werden. Sowohl Sie als auch Andere sollten später genau nachvollziehen können, was mit den Rohdaten gemacht wurde.
* Je nachdem, welche(s) Programm(e) Sie zur Verarbeitung der Daten verwenden, bieten sich etwas unterschiedliche Tools zur Dokumentation an.
* **Merke:** Natürlich kann es immer sein, dass Sie an einem Punkt im Prozess auf eine vorherige Stufe zurückfallen.