Karlsruher Institut für Technologie

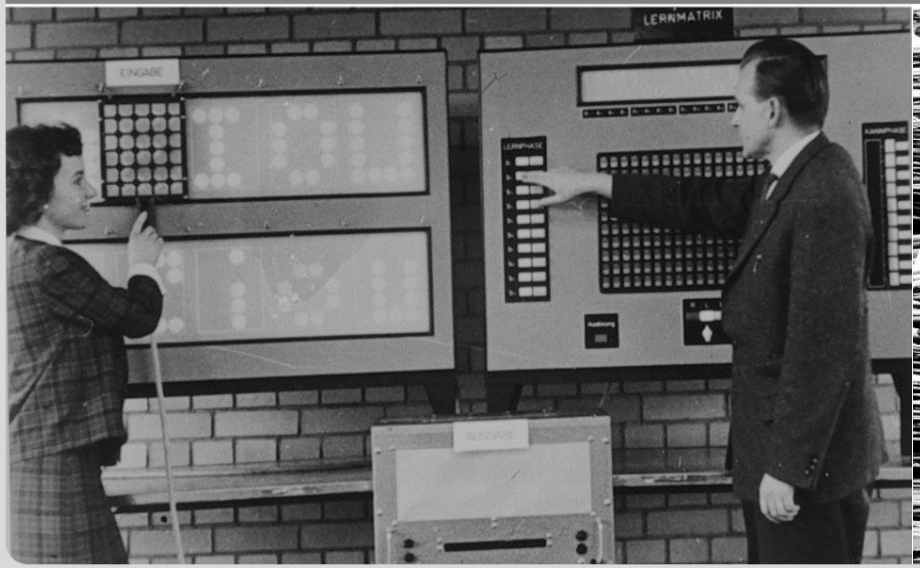
Institut für Technik der Informationsverarbeitung

Institutsleitung
Prof. Dr.-Ing. J. Becker
Prof. Dr.-Ing. E. Sax
Prof. Dr. rer. nat. W. Stork

Übung 5

Übung zu Informationstechnik II und Automatisierungstechnik – Nathalie Brenner

Prof. Dr.-Ing. Eric Sax



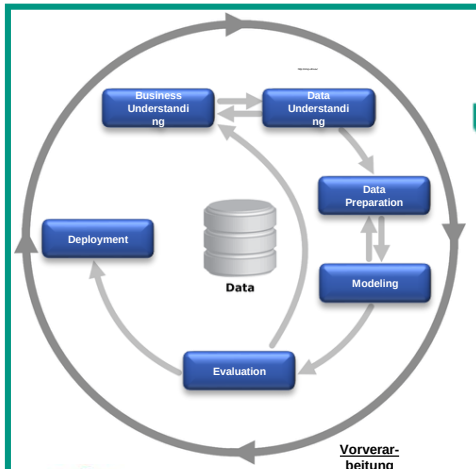
WIEDERHOLUNG ÜBUNG 4

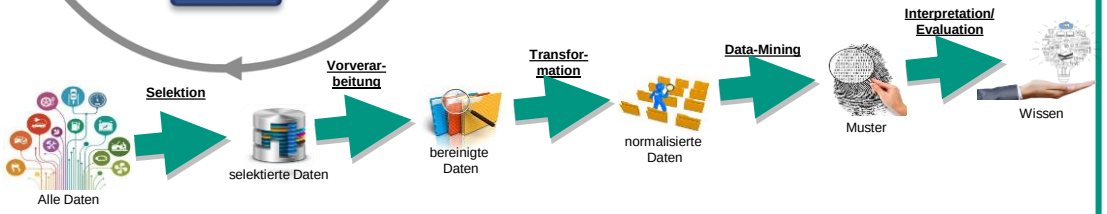
Wiederholung Übung 4

Charakteristika zur Analyse großer Datenbestände → Big Data

Big Data als Prozess

- Zyklisches/iteratives Vorgehen
  - KDD
  - CRISP-DM





Business Understanding

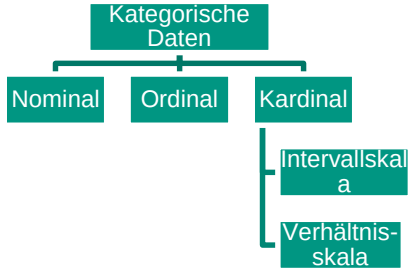
- Verständnis des Projekts
- Ziele und Anforderungen
- Projekplan

Data Understanding

- Datentypen

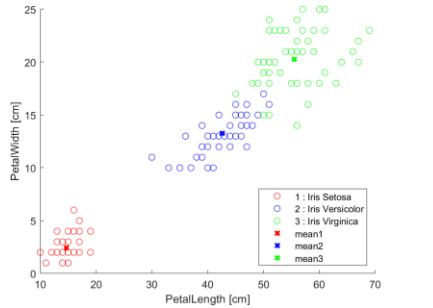
Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	51	35	14	2	Iris-setosa
2	49	30	14	2	Iris-setosa
3	47	32	13	2	Iris-setosa
4	46	31	15	2	Iris-setosa
5	50	36	14	2	Iris-setosa
6	54	39	17	4	Iris-setosa
7	46	34	14	3	Iris-setosa
8	50	34	15	2	Iris-setosa
9	44	29	14	2	Iris-setosa
10	49	31	15	1	Iris-setosa
11	54	37	15	2	Iris-setosa
12	48	34	16	2	Iris-setosa
13	48	30	14	1	Iris-setosa
14	43	30	11	1	Iris-setosa
15	58	40	12	2	Iris-setosa
51	70	32	47	14	Iris-versicolor
52	64	32	45	15	Iris-versicolor
53	69	31	49	15	Iris-versicolor
54	55	23	40	13	Iris-versicolor
55	65	28	46	15	Iris-versicolor
56	57	28	45	13	Iris-versicolor
57	63	33	47	16	Iris-versicolor
58	49	24	33	10	Iris-versicolor
59	66	29	46	13	Iris-versicolor
60	52	27	39	14	Iris-versicolor
61	50	20	35	10	Iris-versicolor
62	59	30	42	15	Iris-versicolor
63	60	22	40	10	Iris-versicolor
64	61	29	47	14	Iris-versicolor
65	56	29	36	13	Iris-versicolor

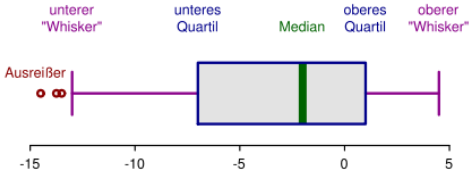
Kardinal | Kardinal | Kardinal | Kardinal | Label



- Statistik und Visualisierung

- Statistik:
  - Mittelwert $\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$
  - Varianz $\sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2$
  - Standardabweichung $\sigma = \sqrt{\sigma^2}$
  - Median $x = \begin{cases} x_{n+1/2} \; n \; ungerade \\ 1/2 \left(x_{n/2} + x_{(n/2)+1} \right) \; n \; gerade \end{cases}$
  - Minimum
  - Maximum





INHALT ÜBUNG 5

Ziele der heutigen Übung

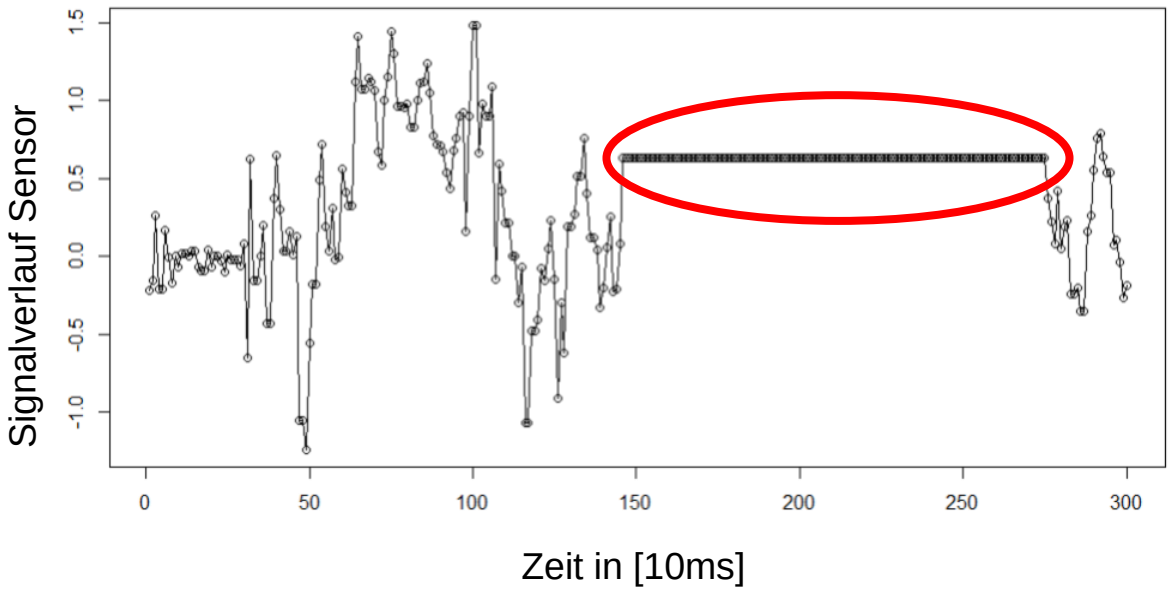
- Nach der heutigen Übung können Sie….

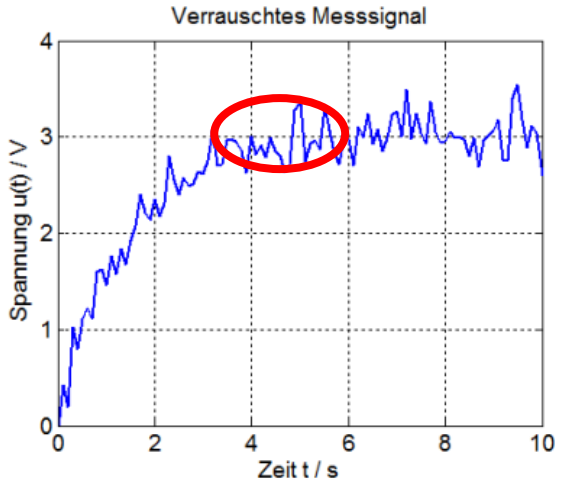
- …Ansätze zur Verwaltung und Analyse großer Datenbestände hinsichtlich ihrer Anwendbarkeit und Wirksamkeit einschätzen

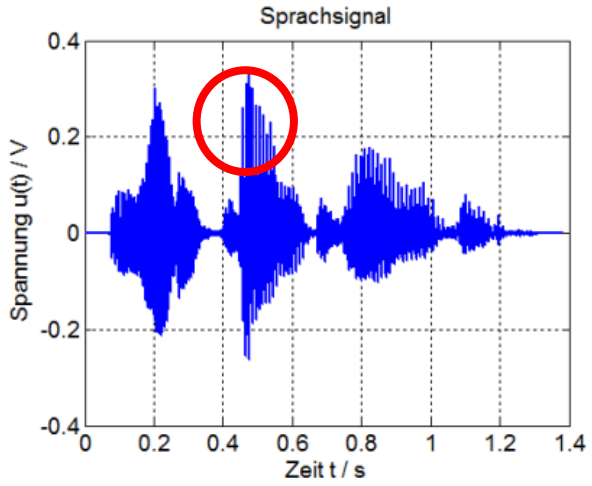
1. … die Bedeutung und den Nutzen von Datenvorverarbeitung erläutern
2. … das Vorgehen zur Datenvorverarbeitung aufzählen
3. … Verfahren zur Datenbereinigung zum Zweck der Vorverarbeitung nennen und anwenden
4. … Verfahren zur Datenmanipulation zum Zweck der Vorverarbeitung nennen und anwenden

Big Data als Prozess

Datenaufbereitung







- Fehlerhafte Sensordaten führen zu invaliden oder irreführenden Ergebnissen, daher müssen diese Daten zunächst aufbereitet werden
  - Datenbereinigung
- Für einen fehlerfreien Durchlauf des Algorithmus müssen die Daten für diesen angepasst werden
  - Datenmanipulation

Ist ein bestimmtes Attribut wichtig für die Data Mining Ziele?

Schließt die Qualität bestimmter Daten die Ergebnisgültigkeit aus?

Gibt es Beschränkungen bezüglich der Daten? (Datenschutz oder Ähnliches)

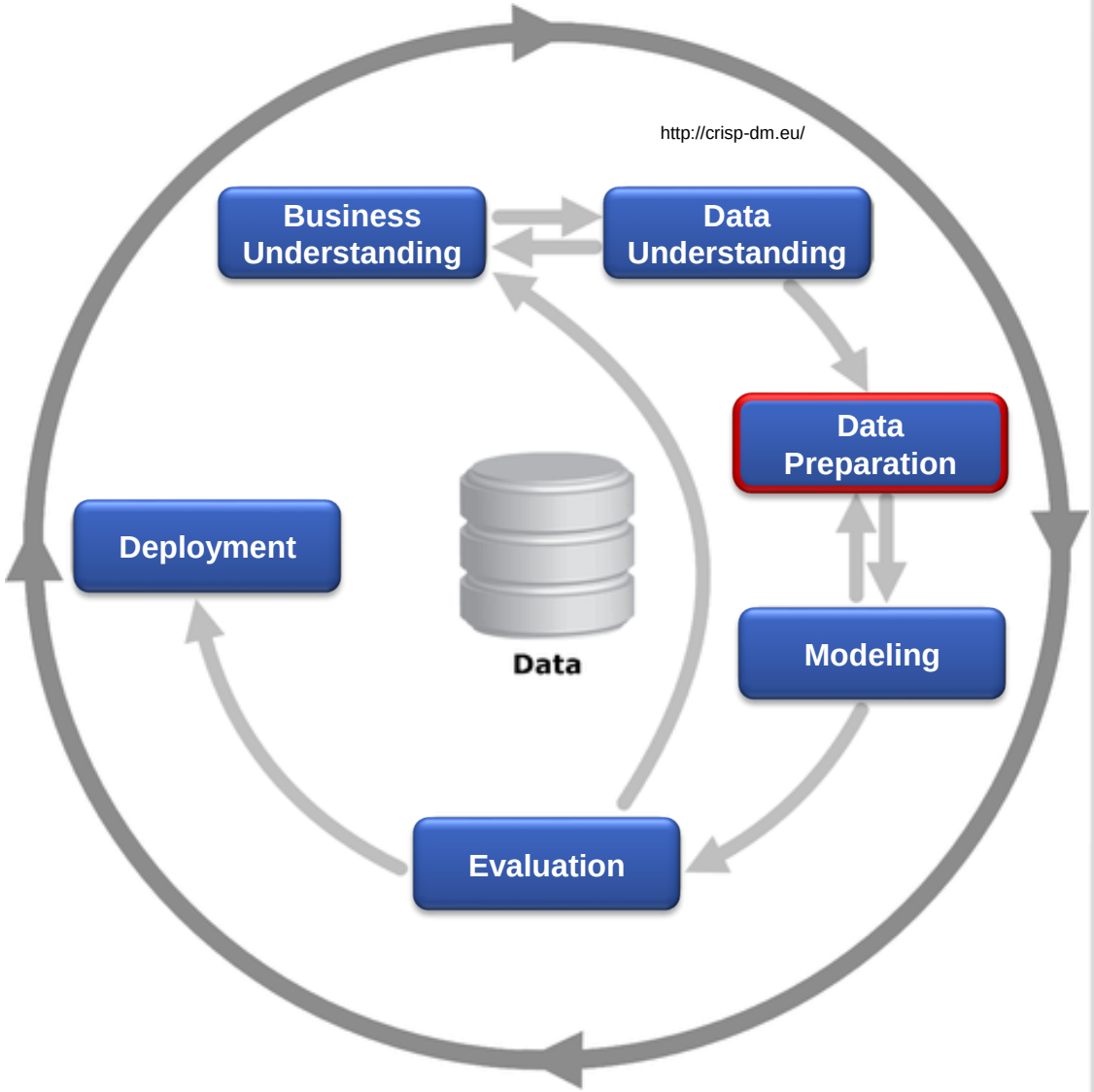
Big Data als Prozess – CRISP-DM

Datenaufbereitung

- Datenaufbereitung
  - Daten müssen bestimmte Voraussetzungen erfüllen
    - Konvertierung der Daten
    - „Glätten“ der Daten
    - Ausreißerdetektion
    - Normalisierung, Skalierung, etc.

> *„Im Allgemeinen müssen Data Scientists anfangs beträchtliche Zeit dafür aufwenden, die Variablen zu definieren, die später verwendet werden. Gerade hier kommen die menschliche Kreativität, der gesunde Menschenverstand und das Fachwissen ins Spiel. Die Qualität einer Data-Mining-Lösung beruht oft darauf, wie gut die Analysten die Aufgabenstellung strukturieren und die Variablen gestalten […]“* ~ *Tom Fawcett*

- Datenbereinigung
- Datenmanipulation



DATENBEREINIGUNG

Datenaufbereitung

Datenbeschaffung und -bereinigung

Acc_x	Acc_y	Acc_z
0.33150518	-0.036584496	-0.10886677
0.26663115	-0.043309471	-0.14096216
0.24042087	0.0010854188	-0.11888144
0.29112809	0.015567293	-0.10429218
0.35552927	-0.023789742	-0.12723972
0.28744253	-0.050273681	-0.13274829
0.1844141	-0.014817877	-0.089991964
0.25204831	-0.00060425372	-0.070452518
0.33806106	-0.0413713	-0.097384161
0.29030031	-0.044250443	-0.12549008
n/a	-0.0415713	-0.0962841
0.26086953	-0.015550952	-0.12094838
0.39276304	-0.059173884	-0.11092832
0.33322745	-0.024943465	-0.15891904
0.31615359	0.0012773598	-0.06545266
0.15366105	-0.010077719	-0.043894838
0.071035289	-0.015656183	-0.094263452
0.33304231	-0.01019258	-0.1220055
0.27579996	0.0018368697	-0.13628781
0.27138623	-0.042483426	-0.12821114
0.42604818	-0.058615109	-0.1170689
0.26938623	-0.039483426	n/a
0.2856233	-0.0060363793	-0.19809731
0.35480453	-0.018153403	-0.14403353
0.39558207	-0.012193647	-0.14801867
0.22715659	-0.022146672	-0.14521449
0.23483919	0.0081011057	-0.14108314
0.23820197	-0.0026928807	-0.12149269
0.27873663	-0.048279124	-0.12092488
0.26374279	-0.02958616	-0.050708835

- Daten „beschaffen“
  - Manuelle Eingabe
  - Textdateien einlesen
  - HTML (from web)
  - APIs
- Daten bereinigen
  - „defekte“ Daten erkennen und entfernen:
  1. Fehlende Werte löschen/ersetzen
  2. Fensterauswahl
  3. Anomaliedetektion

> *„In order to be a data scientist you need data. In fact, as a data scientist you will spend an embarrassingly large fraction of your time acquiring, cleaning and tranforming data. […]“*
> *~ Joel Grus*

Datenaufbereitung

Datenbereinigung - Fehlende Werte löschen/ersetzen

Acc_x	Acc_y	Acc_z
0.33150518	-0.036584496	-0.10886677
0.26663115	-0.043309471	-0.14096216
0.24042087	0.0010854188	-0.11888144
0.29112809	0.015567293	-0.10429218
0.35552927	-0.023789742	-0.12723972
0.28744253	-0.050273681	-0.13274829
0.1844141	-0.014817877	-0.089991964
0.25204831	-0.00060425372	-0.070452518
0.33806106	-0.0413713	-0.097384161
0.29030031	-0.044250443	-0.12549008
n/a	-0.0415713	-0.0962841
0.26086953	-0.015550952	-0.12094838
0.39276304	-0.059173884	-0.11092832
0.33322745	-0.024943465	-0.15891904
0.31615359	0.0012773598	-0.06545266
0.15366105	-0.010077719	-0.043894838
0.071035289	-0.015656183	-0.094263452
0.33304231	-0.01019258	-0.1220055
0.27579996	0.0018368697	-0.13628781
0.27138623	-0.042483426	-0.12821114
0.42604818	-0.058615109	-0.1170689
0.26938623	-0.039483426	n/a
0.2856233	-0.0060363793	-0.19809731
0.35480453	-0.018153403	-0.14403353
0.39558207	-0.012193647	-0.14801867
0.22715659	-0.022146672	-0.14521449
0.23483919	0.0081011057	-0.14108314
0.23820197	-0.0026928807	-0.12149269
0.27873663	-0.048279124	-0.12092488
0.26374279	-0.02958616	-0.050708835

- Fehlerhafte oder Fehlende Werte können das Programm bzw. den Algorithmus zum Absturz bringen und müssen daher identifiziert werden. Der Umgang mit solchen Werten ist zu entscheiden
  - Löschen
  - Interpolieren
  - None-Wert
- Identifikation beispielsweise über „try-funktion“

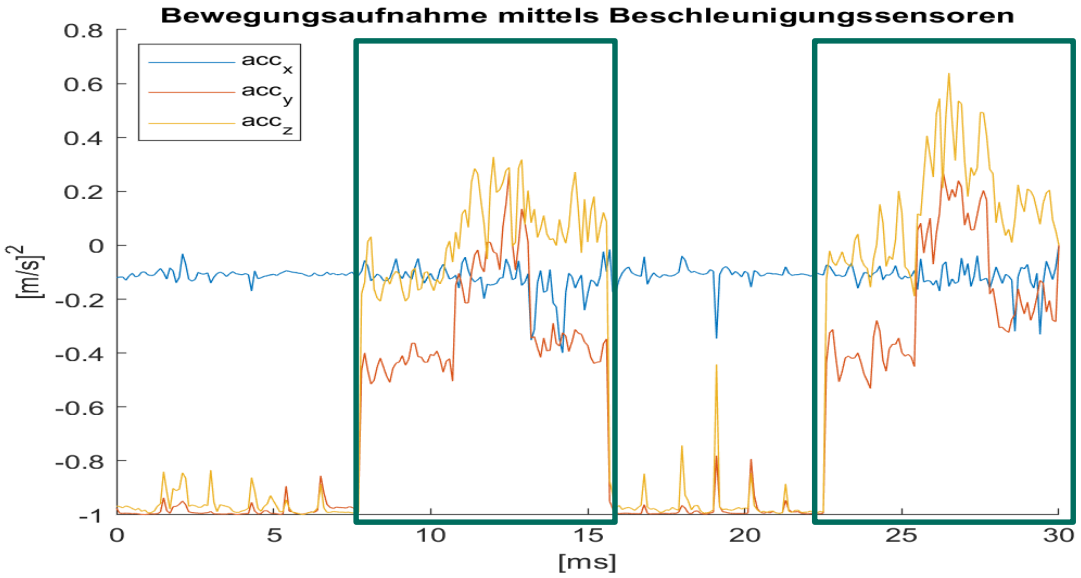
```
def try_or_none(f):
    def f_or_none(x):
        try: return f(x)
        except: return None
    return f_or_none
```

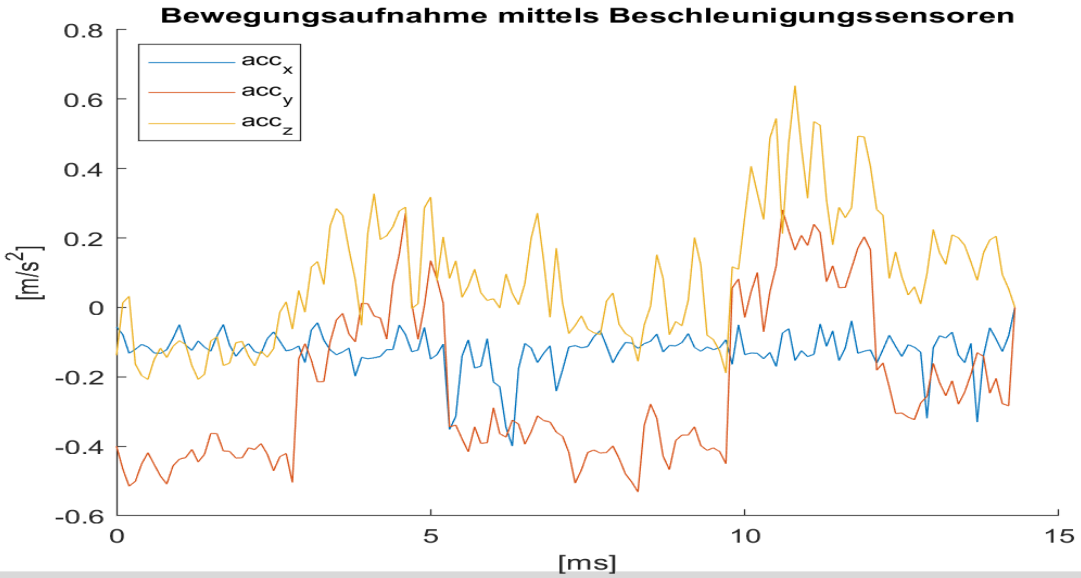
Datenaufbereitung

Datenbereinigung - Fensterauswahl

Acc_x	Acc_y	Acc_z
0.33150518	-0.036584496	-0.10886677
0.26663115	-0.043309471	-0.14096216
0.24042087	0.0010854188	-0.11888144
0.29112809	0.015567293	-0.10429218
0.35552927	-0.023789742	-0.12723972
0.28744253	-0.050273681	-0.13274829
0.1844141	-0.014817877	-0.089991964
0.25204831	-0.00060425372	-0.070452518
0.33806106	-0.0413713	-0.097384161
0.29030031	-0.044250443	-0.12549008
0.26086953	-0.015550952	-0.12094838
0.39276304	-0.059173884	-0.11092832
0.33322745	-0.024943465	-0.15891904
0.31615359	0.0012773598	-0.06545266
0.15366105	-0.010077719	-0.043894838
0.071035289	-0.015656183	-0.094263452
0.33304231	-0.01019258	-0.1220055
0.27579996	0.0018368697	-0.13628781
0.27138623	-0.042483426	-0.12821114
0.42604818	-0.058615109	-0.1170689
0.2856233	-0.0060363793	-0.19809731
0.35480453	-0.018153403	-0.14403353
0.39558207	-0.012193647	-0.14801867
0.22715659	-0.022146672	-0.14521449
0.23483919	0.0081011057	-0.14108314
0.23820197	-0.0026928807	-0.12149269
0.27873663	-0.048279124	-0.12092488
0.26374279	-0.02958616	-0.050708835

- Zu betrachtendes Fenster auswählen
  - Tag/Nacht?
  - Sommer/Winter?
  - Event getriggert?





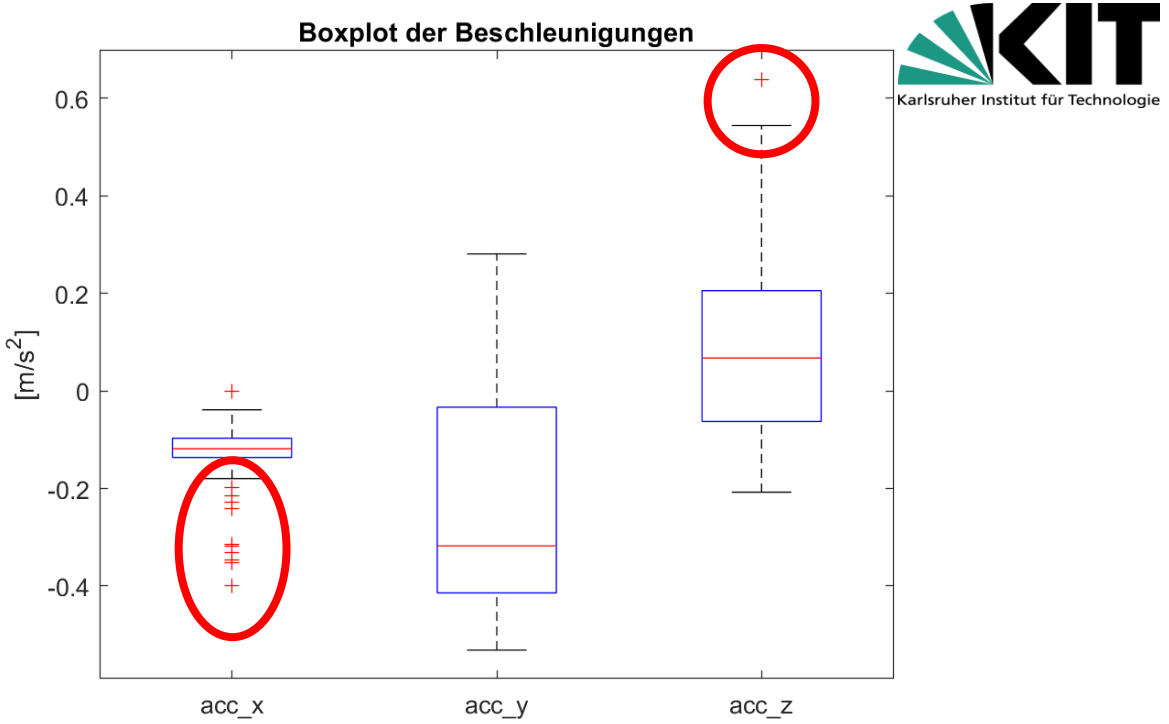
Datenaufbereitung

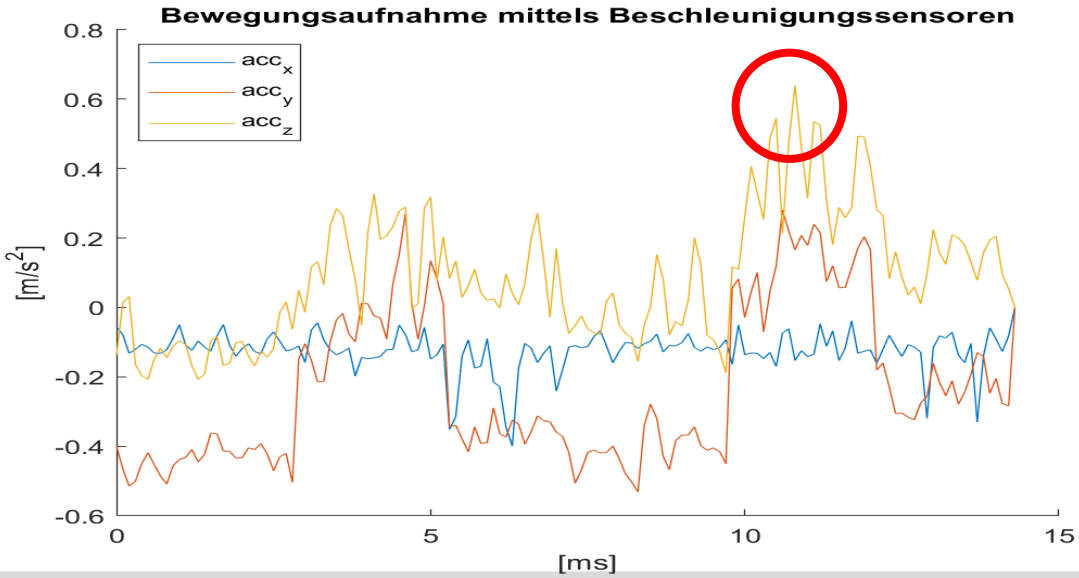
Datenbereinigung - Anomaliedetektion

- Anomalien führen zwar nicht zu einem Absturz des Programms, aber zu invaliden/verfälschten Ergebnissen → Anomaliedetektion
- Entscheidung über Umgang mit Anomalien
  - Löschen
  - Interpolieren
  - None-Wert

Inhalt in „Anwendungen aus dem Institut“

- Detektionsmöglichkeiten über Erkundung
  - Visualisierung
  - Anomaliemaß
  - Ad hoc





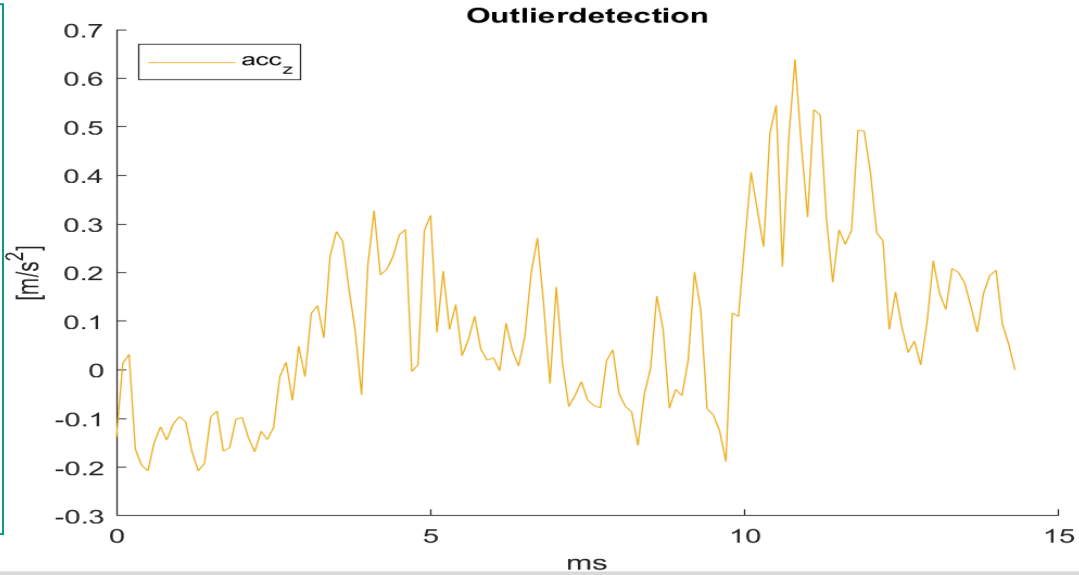
Datenaufbereitung – Anomaliedetektion über Anomaliemaß

Zwischenübung

- Führen Sie eine Anomaliedetektion über die **Standardabweichung** durch:
  - Schreiben Sie hierfür zunächst den Pseudocode
  - Ermitteln Sie die Anomalien im Anschluss grafisch

Statistik:

- Mittelwert $\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$
- Varianz $\sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2$
- Standardabweichung $\sigma = \sqrt{\sigma^2}$
- Median $x = \begin{cases} x_{n+1/2} & n\ ungerade \\ 1/2\left(x_{n/2} + x_{(n/2)+1}\right) & n\ gerade \end{cases}$
- Minimum
- Maximum



Datenaufbereitung - Anomaliedetektion

Zwischenübung - Lsg

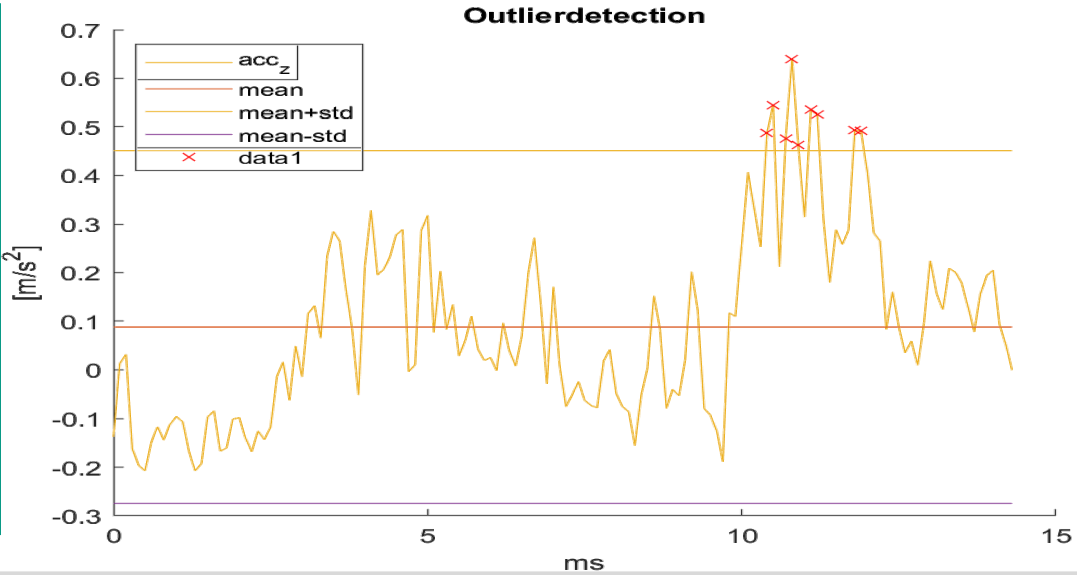
- Führen Sie eine Anomaliedetektion über die **Standardabweichung** durch:

```
outlier detect_outliers(data)
    mean = mean(data)
    std = std(data)
    top = mean+std*1.96
    bot = mean-std*1.96
    outlier_idx = find( (data>top) | (data<bot))
    outlier = data(outlier)
    return outlier
```

95% der Daten als „true“ anerkennen

Statistik:

- Mittelwert $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ = 0,0882
- Varianz $\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$ = 0,04694
- Standardabweichung $\sigma = \sqrt{\sigma^2}$ = 0,1850
- Median $x = \begin{cases} x_{n+1/2} & n\ ungerade \\ 1/2\left(x_{n/2} + x_{(n/2)+1}\right) & n\ gerade \end{cases}$ = 0,0676
- Minimum = -0,2076
- Maximum = 0,6390



- Datenbereinigung
  - Fehlende/Fehlerhafte Werte
  - Fensterung
  - Anomaliedetektion

DATENMANIPULATION

Datenaufbereitung
Datenmanipulation

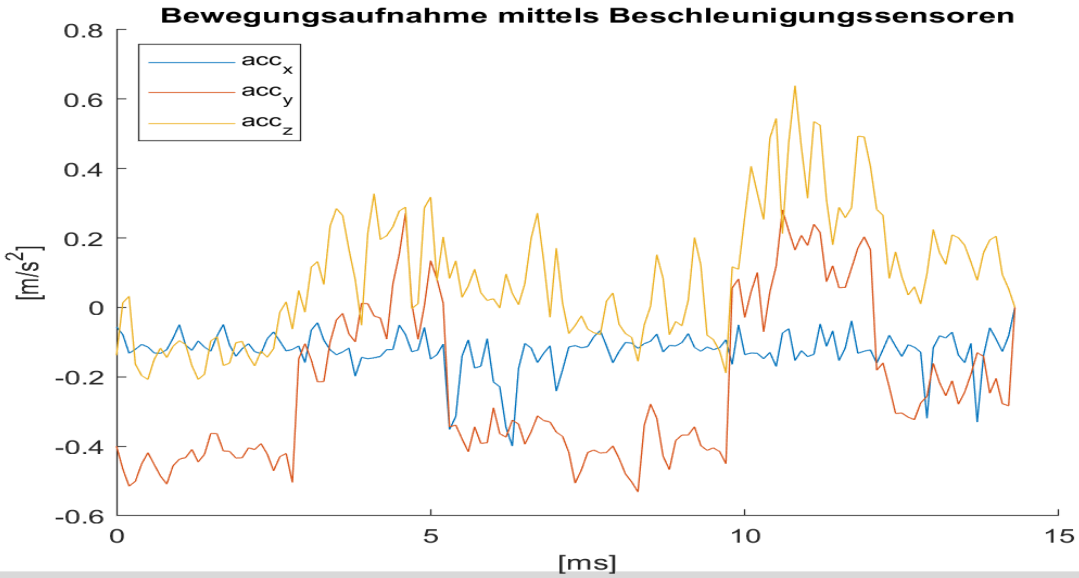
- Datenmanipulation ist ein allgemeiner Ansatz und keine bestimmte Technik
  Nach der Bereinigung der Daten werden diese anschließend weiterverarbeitet
- Datenmanipulation
  1. Umgang mit Ausreißern
  2. Konvertierung der Daten nach Use Case
     - Normierung, Standardisierung
     - Zeitsynchronisation
     - Anpassung von Einheiten
  3. Umgang mit fehlerhaften Werten
  4. Qualitätsverbesserung
  5. Merkmalsreduktion

Datenaufbereitung

Datenmanipulation – Umgang mit Ausreißern

- Anomalien führen zwar nicht zu einem Absturz des Programms, aber zu invaliden/verfälschten Ergebnissen → Anomaliedetektion
- Entscheidung über Umgang mit Anomalien
  - Löschen
  - Interpolieren
  - None-Wert
- Detektionsmöglichkeiten über Erkundung
  - Anomaliemaß
  - Ad hoc
  - Visualisierung

Inhalt in „Anwendungen aus dem Institut“



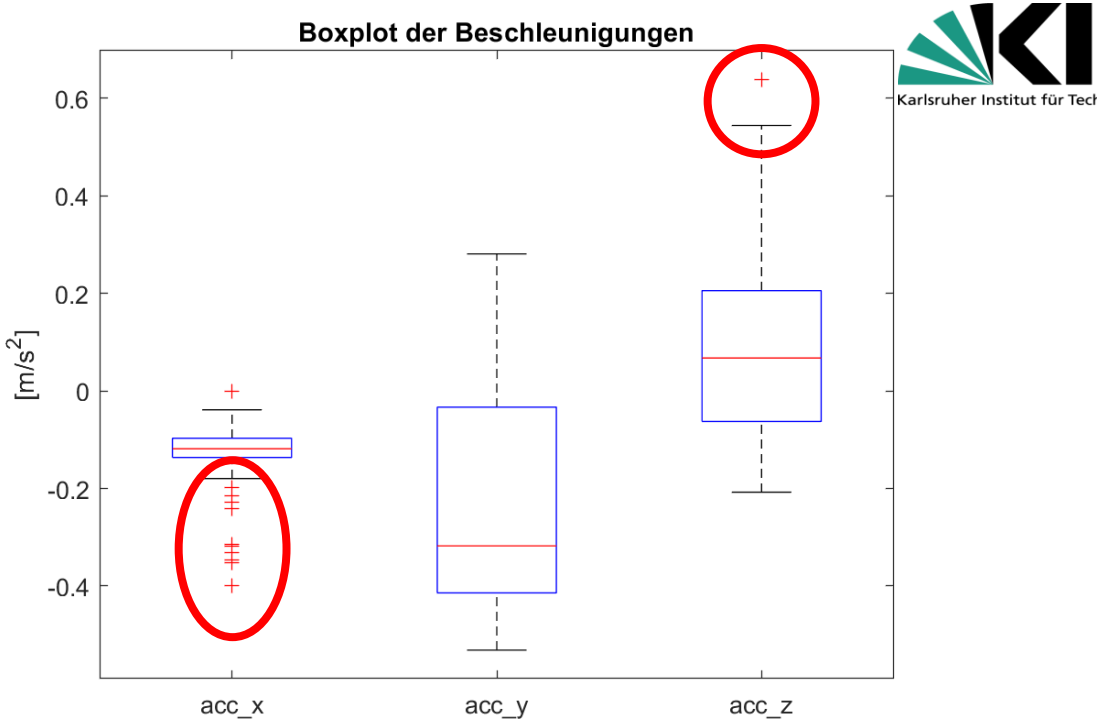
Datenaufbereitung

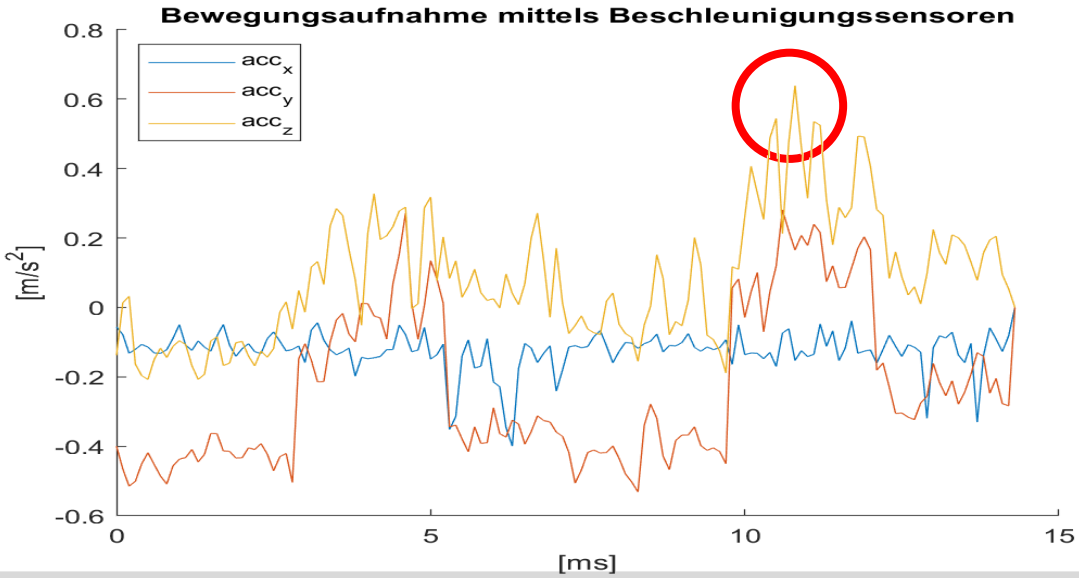
Datenmanipulation – Umgang mit Ausreißern

- Anomaliedetektion durch Boxplotdarstellung
- Umgang mit Anomalie *(hier)*
  - Interpolation
    - Aller Signale
    - Nur des Signals mit enthaltenem Fehler

Lineare Interpolation:

$$y_i = y_1 + (y_2 - y_1) * \frac{(x_i - x_1)}{(x_2 - x_1)}$$





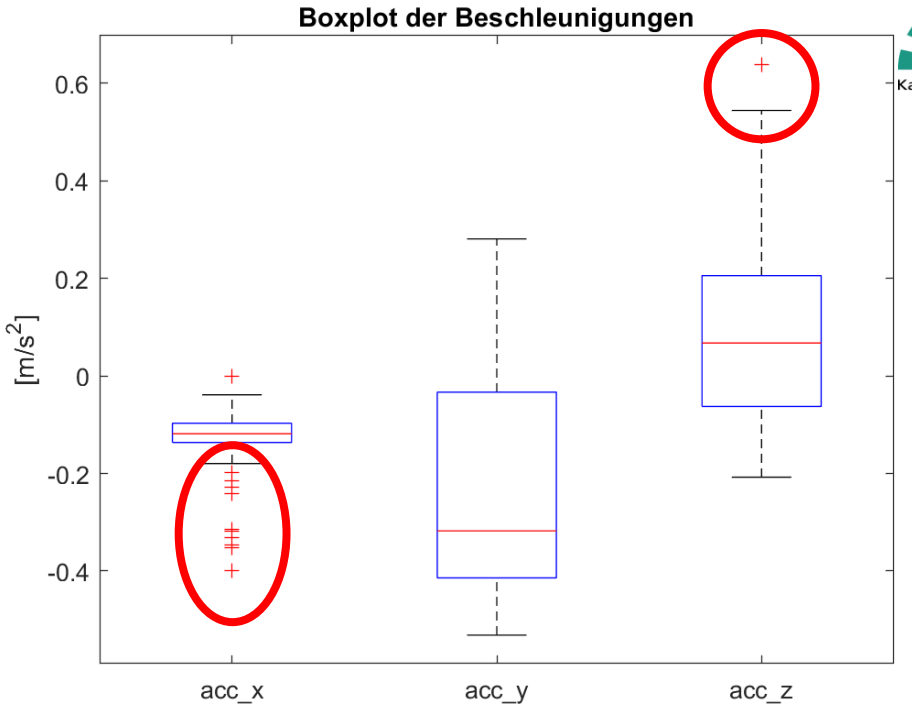
Datenaufbereitung

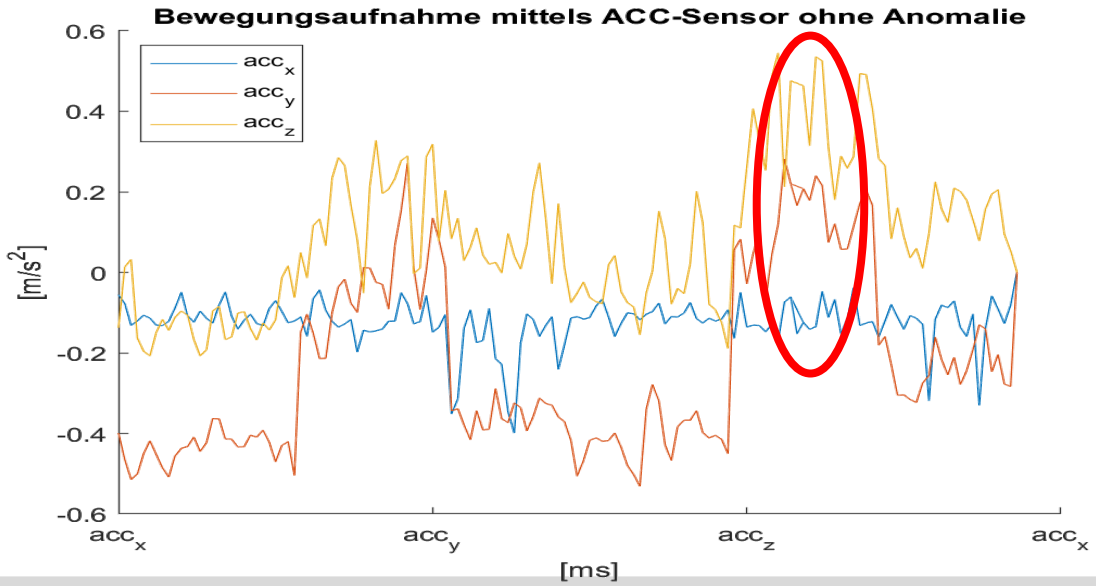
Datenmanipulation – Umgang mit Ausreißern

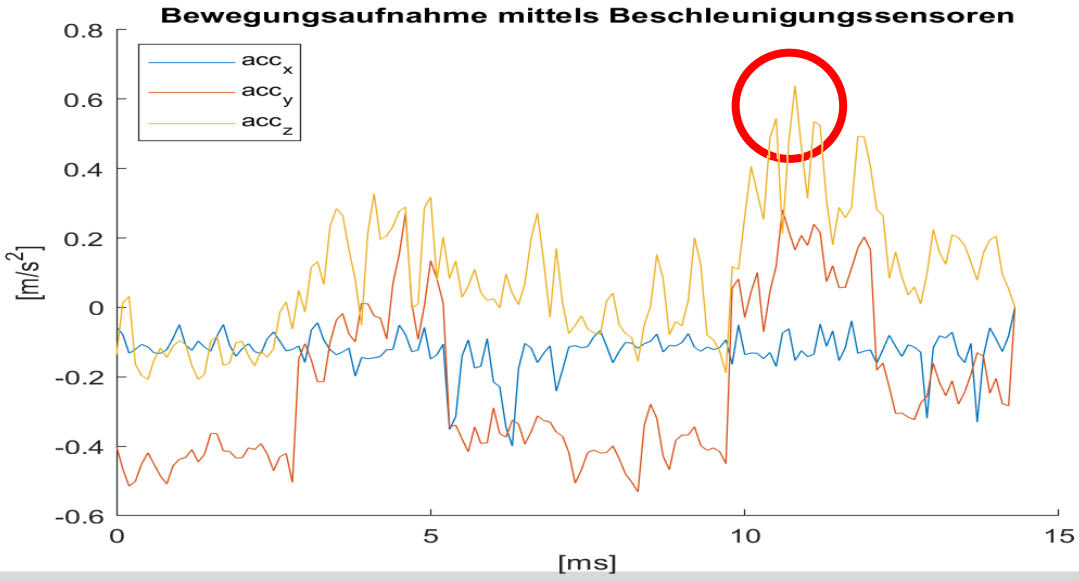
- Anomaliedetektion durch Boxplotdarstellung
- Umgang mit Anomalie *(hier)*
  - Interpolation
    - Aller Signale
    - Nur des Signals mit enthaltenem Fehler

Lineare Interpolation:

$$y_i = y_1 + (y_2 - y_1) * \frac{(x_i - x_1)}{(x_2 - x_1)}$$







Datenaufbereitung

Datenmanipulation - Konvertierung

- Standardisierung
  Werte an den benötigten Standard des Data Mining Algorithmus anpassen
  Beispiel: „verständliche“ Labels

Acc_x	Acc_y	Acc_z	Activity
0.33150518	-0.036584496	-0.10886677	SITTING
0.26663115	-0.043309471	-0.14096216	SITTING
0.28744253	-0.050273681	-0.13274829	SITTING
…	…	…	…
0.22715659	-0.022146672	-0.14521449	LAYING
0.23483919	0.0081011057	-0.14108314	LAYING
0.23820197	-0.0026928807	-0.12149269	LAYING
0.27873663	-0.048279124	-0.12092488	LAYING
0.26374279	-0.02958616	-0.050708835	LAYING
…	…	…	…
0.31615359	0.0012773598	-0.06545266	WALKING
0.15366105	-0.010077719	-0.043894838	WALKING
0.071035289	-0.015656183	-0.094263452	WALKING
0.33304231	-0.01019258	-0.1220055	WALKING

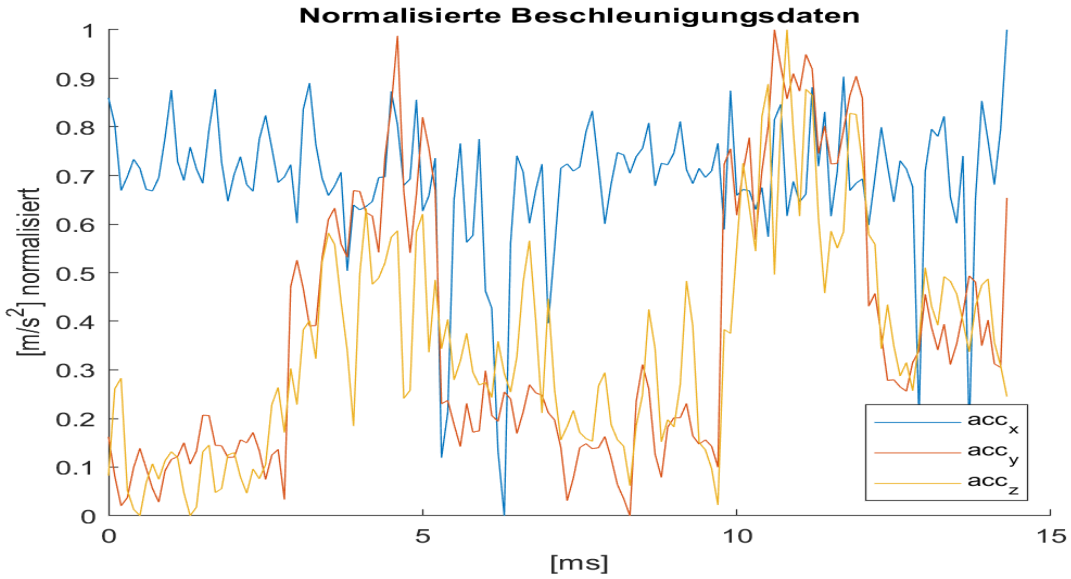
Activity
1
1
1
…
2
2
2
2
2
…
3
3
3
3

- Anpassung von Einheiten (meist in SI-Einheiten)
  Beispiel: hier schon gegeben mit m/s^2 und ms

Birnen mit Äpfeln vergleichen?

- Normierung
  - Min/Max-Normierung $x^{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$
  - Z-Score-Normierung $x^{new} = \frac{x - \mu}{\sigma_x}$
  - Dezimal-Skalierung $x^{new} = |x| * 10^a$,
    $a = max = i \in \mathbb{Z}, |x| * 10^i < 1$
  - Logarithmische Skalierung $x^{new} = \log_a x$

Beispiel: Beschleunigungen auf [0,1] normieren um Ausreißer durch Visualisierung zu erkennen



Datenaufbereitung

Datenmanipulation - Konvertierung

- Zeitsynchronisation durch Interpolation
  - Eine zeitsynchrone Gesamtdatei aus verschiedenen Datenquellen erstellen
  - Messzeitpunkte nicht exakt identisch

Sensor1

Time	Signal 1	…	Signal 25
$t_{1.1}$=0,01	Signal1 ($t_{1.1}$)	..	Signal25 (t_1)
$t_{1.2}$=0,02	Signal1 ($t_{1.2}$)	..	Signal25 (t_2)
$t_{1.3}$=0,03	Signal1 ($t_{1.3}$)	…	Signal25 (t_3)
…	…	…	…

Sensor2

Time	Signal 1	…	Signal 25
$t_{2.1}$=0,015	Signal1 (t_1)	..	Signal25 (t_1)
$t_{2.2}$=0,025	Signal1 (t_2)	..	Signal25 (t_2)
$t_{2.3}$=0,035	Signal1 (t_3)	…	Signal25 (t_3)
…	…	…	…

Sensor3

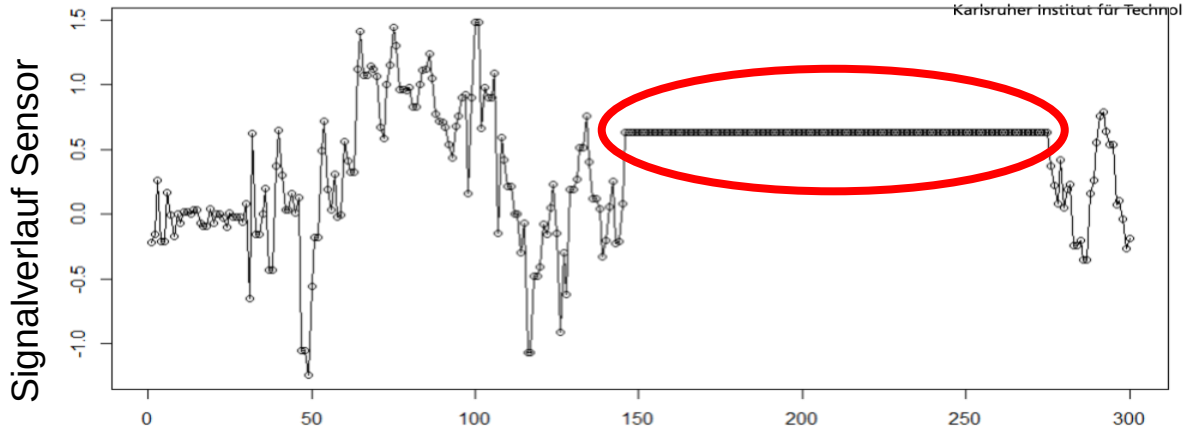
Time	Signal 1	…	Signal 3
$t_{3.1}$=0,005	Signal1 (t_1)	..	Signal3 (t_1)
$t_{3.2}$=0,010	Signal1 (t_2)	..	Signal3 (t_2)
$t_{3.3}$=0,015	Signal1 (t_3)	…	Signal3 (t_3)
…	…	…	…

Time	Signal 1.1	…	Signal 1.25	Signal 2.1	…	Signal 2.25	Signal 3.1	…	Signal 3.3
t_1	Signal1.1 (t_1)	..	Signal25 (t_1)	Signal2.1 (t_1)	..	Signal2.25 (t_1)	Signal3.1 (t_1)	..	Signal3.3 (t_1)
t_2	Signal1.1 (t_2)	..	Signal25 (t_2)	Signal2.1 (t_2)	..	Signal2.25 (t_2)	Signal3.1 (t_2)	..	Signal3.3 (t_2)
t_3	Signal1.1 (t_3)	…	Signal25 (t_3)	Signal2.1 (t_3)	…	Signal2.25 (t_3)	Signal3.1 (t_3)	…	Signal3.3 (t_3)
…	…	…	…	…	…	…	…	…	…

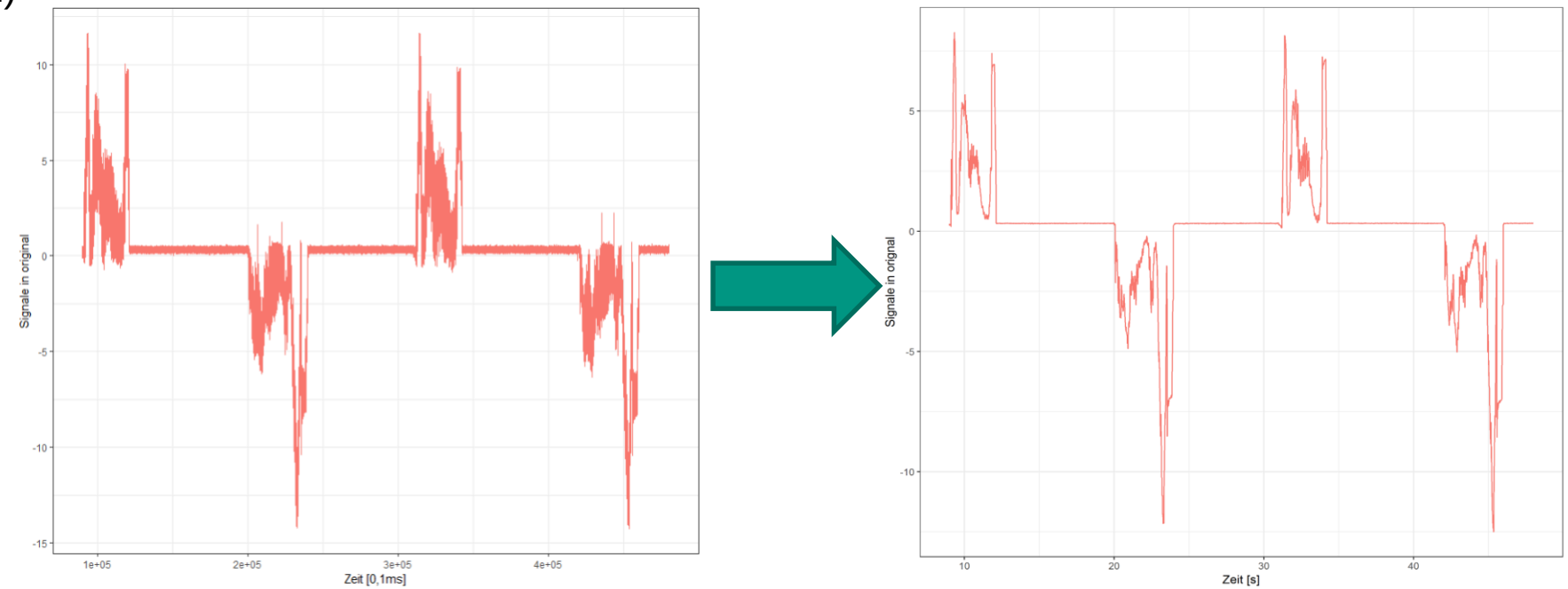
Datenaufbereitung

Datenmanipulation - Qualitätsverbesserung

- Entfernung von fehlerhaften Messungen



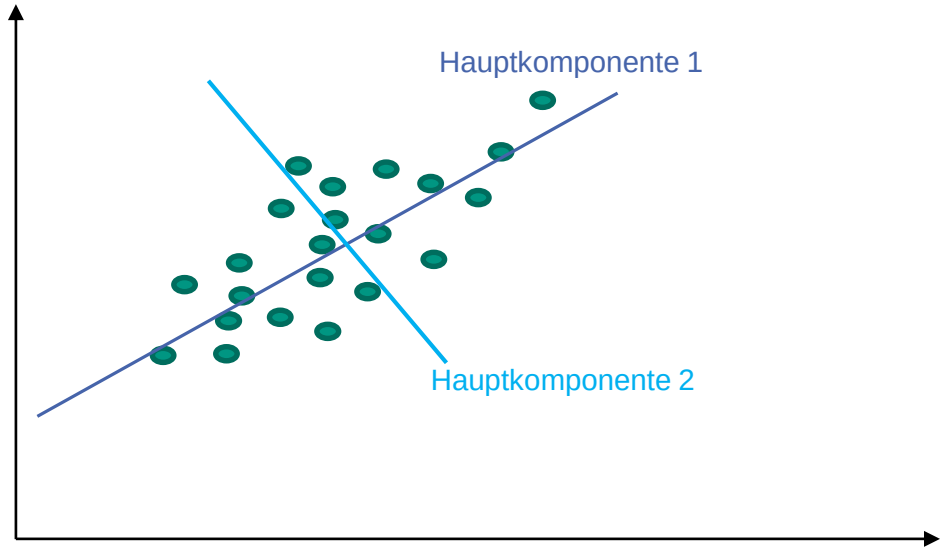
- Glättung von Messwerten
  (z.B. durch Interpolation)



Datenaufbereitung

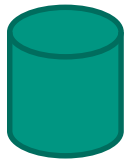
Merkmalsreduktion – Transformation durch PCA

- Datensatz liegt in n x m Matrix vor: als Punktwolke in einem Diagramm darstellbar
- Datenpunkte in einen q-dimensionalen Unterraum projizieren, so dass möglichst wenig Informationen verloren gehen
- Redundanz in Form von Korrelation in den Datenpunkten
- Mathematisch: Hauptachsentransformation
  - Minimierung der Korrelation mehrdimensionaler Merkmale durch Überführung in Vektorraum neuer Basis
  - Orthogonale Matrix, bestehend aus Eigenvektoren der Kovarianzmatrix
- Problemabhängig
- Gesucht:
  beste lineare Approximation
  1. diejenige Gerade, welche die Daten am besten approximiert. Fehler ist hierbei der Abstand zur Geraden.
  → Hauptkomponente 1 ist die Gerade, bei der die Summe der Quadrate aller Fehler minimal
  2. Hauptkomponente 2 ist die Gerade die den Mittelwert der Daten wiedergibt und orthogonal zur Hauptkomponente 1 liegt



Datenaufbereitung

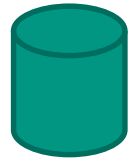
Merkmalsreduktion – Transformation durch PCA



Datenset als n x m Matrix Punktwolke

Mittelwert berechnen $\mu = \Sigma x_n / n$
Daten um Mittelwert verschieben

Normiertes Datenset

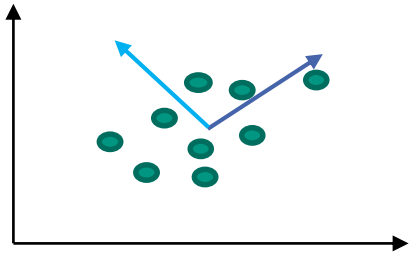


Kovarianzmatrix

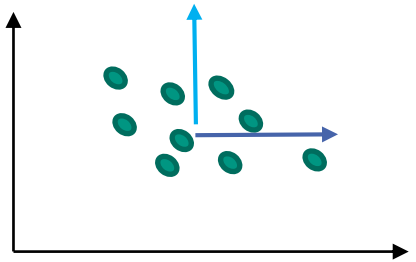
$$Cov(X,Y) := E\left[\left(X - E(X)\right) * \left(Y - E(Y)\right)\right]$$

Eigenwerte $(X - \lambda E) = 0$
Eigenvektoren $(X - \lambda_i E) * x_i = 0$
berechnen

Streuung in Richtung der Eigenvektoren, Gewichtet über die Eigenwerte



Transformationsmatrix aus Eigenvektoren auf Daten anwenden → Rotation der Daten



Daten in Abhängigkeit der Komponente mit der höchsten Abhängigkeit/ Information

Ziel: Reduzierung der Dimensionalität, aber Erhalt der Repräsentationsfähigkeit der Daten

Datenmanipulation – Lineare Interpolation

Zwischenübung

- Gegeben seien die folgenden, mit Fehler behafteten, Sensordaten. Die Daten enthalten fehlerhafte Daten und Anomalien. Bereinigen Sie die Daten, in dem Sie den angebrachten Umgang entscheiden und durchführen

Index	data1	data2	data3	data4
1	0,27	n/a	0,84	0,01
2	0,30	n/a	n/a	0,02
3	0,31	n/a	0,80	0,05
4	n/a	n/a	n/a	0,03
5	0,36	0	0,79	0,02

Lineare Interpolation:

$$y_i = y_1 + (y_2 - y_1) * \frac{(x_i - x_1)}{(x_2 - x_1)}$$

Datenmanipulation – Lineare Interpolation

Zwischenübung - Lsg

- Gegeben seien die folgenden, mit Fehler behafteten, Sensordaten. Die Daten enthalten fehlerhafte Daten und Anomalien. Bereinigen Sie die Daten, in dem Sie den angebrachten Umgang entscheiden und durchführen

Index	data1	data2	data3	data4
1	0,27	n/a	0,84	0,01
2	0,30	n/a	n/a	0,02
3	0,31	n/a	0,80	0,05
4	n/a	n/a	n/a	0,03
5	0,36	0	0,79	0,02

Index_n	data1	data3	data4
1	0,27	0,84	0,01
2	0,30	0,82	0,02
3	0,31	0,80	0,05
4	0,36	0,79	0,02

Lineare Interpolation:

$$y_i = y_1 + (y_2 - y_1) * \frac{(x_i - x_1)}{(x_2 - x_1)}$$

Lineare Interpolation:

$$y_i = 0{,}84 + (0{,}80 - 0{,}84) * (2-1)/(3-1)$$

- Datenmanipulation
  - Umgang mit Ausreißern
  - Konvertierung
  - Fehlerhafte Werte
  - Qualitätsverbesserung
  - Merkmalsredukation

Ziele der heutigen Übung

- Nach der heutigen Übung können Sie….

- …Ansätze zur Verwaltung und Analyse großer Datenbestände hinsichtlich ihrer Anwendbarkeit und Wirksamkeit einschätzen

1. … die Bedeutung und den Nutzen von Datenvorverarbeitung erläutern
2. … das Vorgehen zur Datenvorverarbeitung aufzählen
3. … Verfahren zur Datenbereinigung zum Zweck der Vorverarbeitung nennen und anwenden
4. … Verfahren zur Datenmanipulation zum Zweck der Vorverarbeitung nennen und anwenden