Karlsruher Institut für Technologie

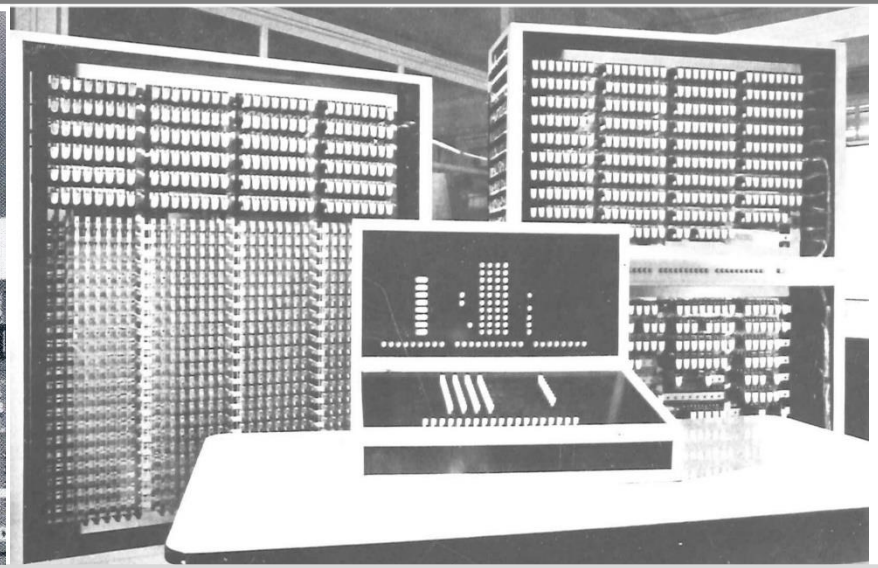
Institut für Technik der Informationsverarbeitung

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 – Felix Pistorius

Prof. Dr.-Ing. Eric Sax



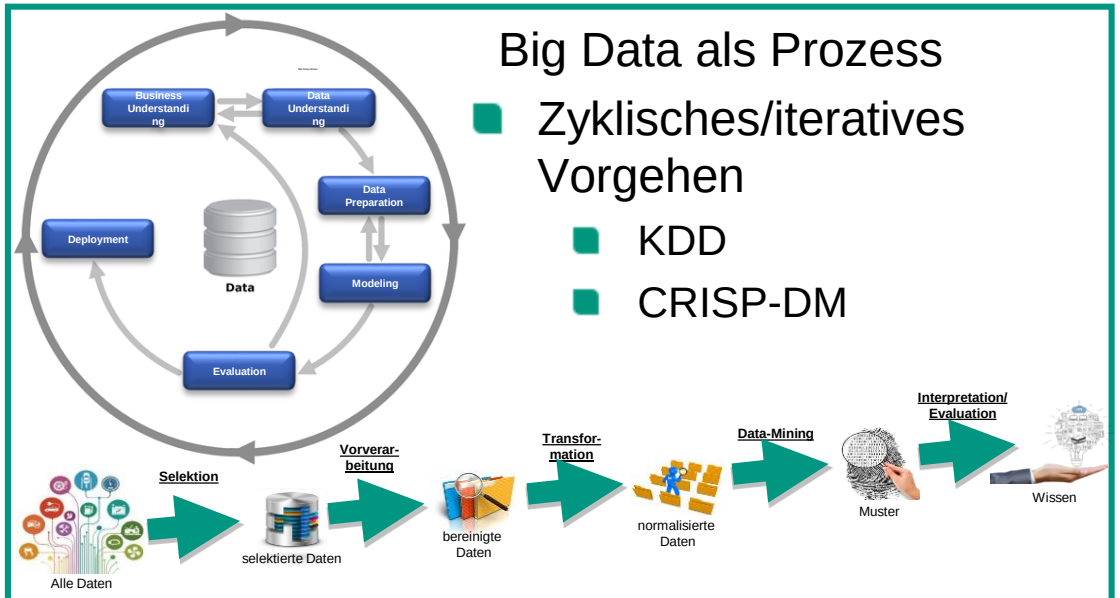
WIEDERHOLUNG ÜBUNG 4

Wiederholung Übung 4

Ansätze zur Verwaltung und Analyse großer Datenbestände

Big Data als Prozess

- Zyklisches/iteratives Vorgehen
  - KDD
  - CRISP-DM



Business Understanding

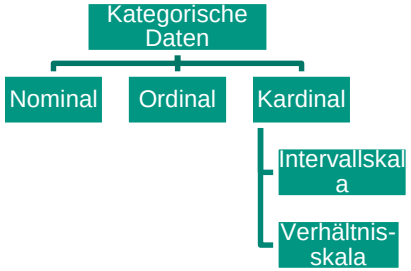
- Verständnis des Projekts
- Ziele und Anforderungen
- Projekplan

Data Understanding

- Datentypen

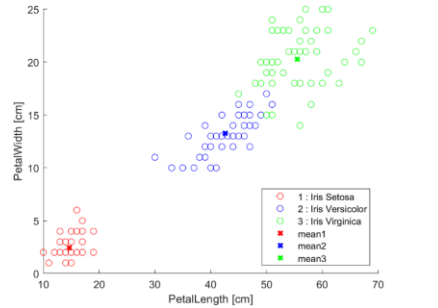
Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	51	35	14	2	Iris-setosa
2	49	30	14	2	Iris-setosa
3	47	32	13	2	Iris-setosa
4	46	31	15	2	Iris-setosa
5	50	36	14	2	Iris-setosa
6	54	39	17	4	Iris-setosa
7	46	34	14	3	Iris-setosa
8	50	34	15	2	Iris-setosa
9	44	29	14	2	Iris-setosa
10	49	31	15	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

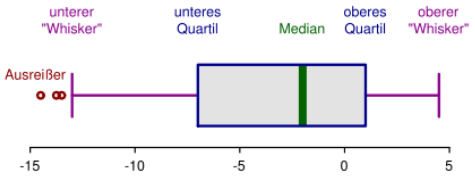
Kardinal | Kardinal | Kardinal | Kardinal | Label



- 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

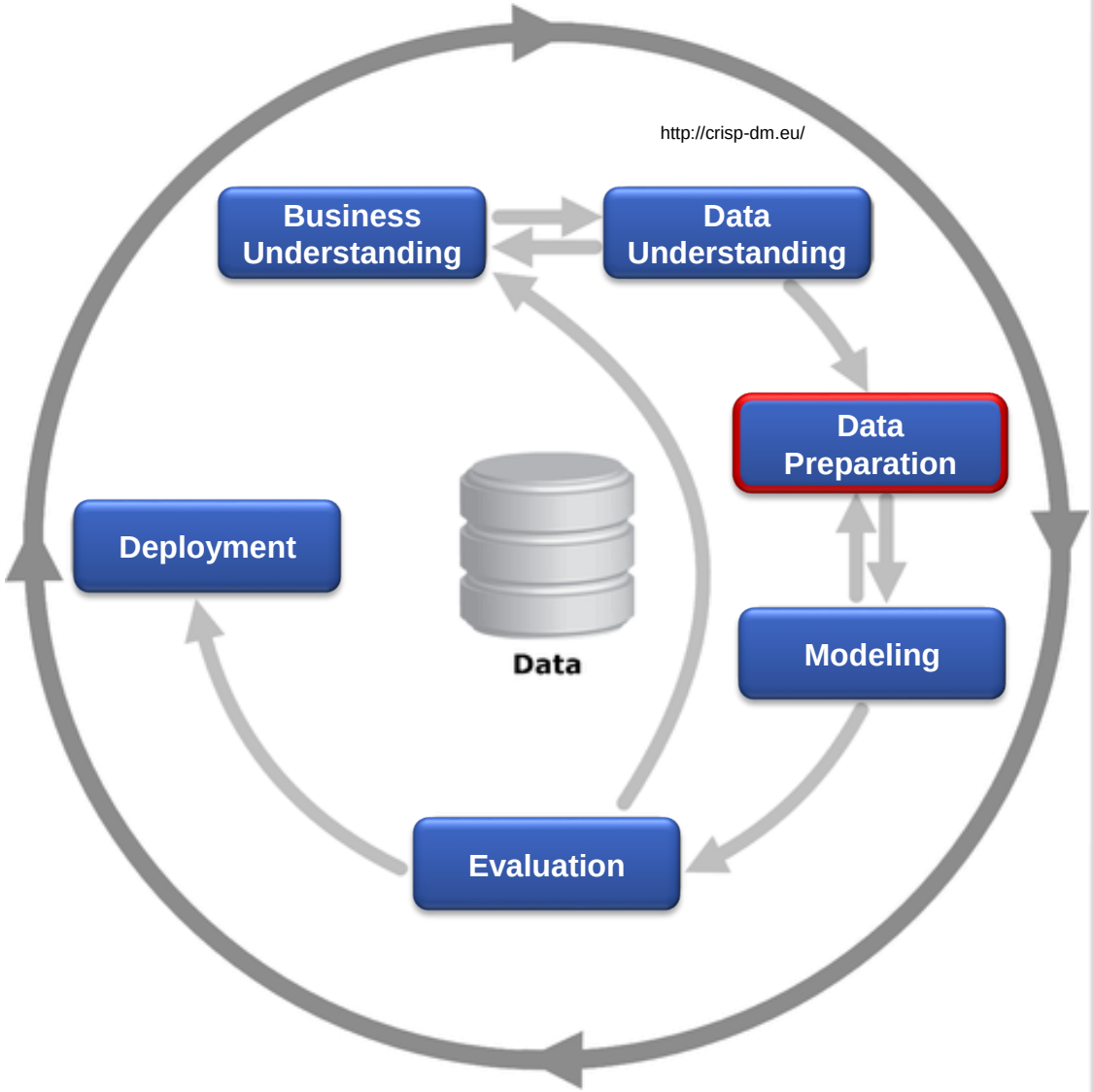
Big Data als Prozess – CRISP-DM

Datenaufbereitung

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

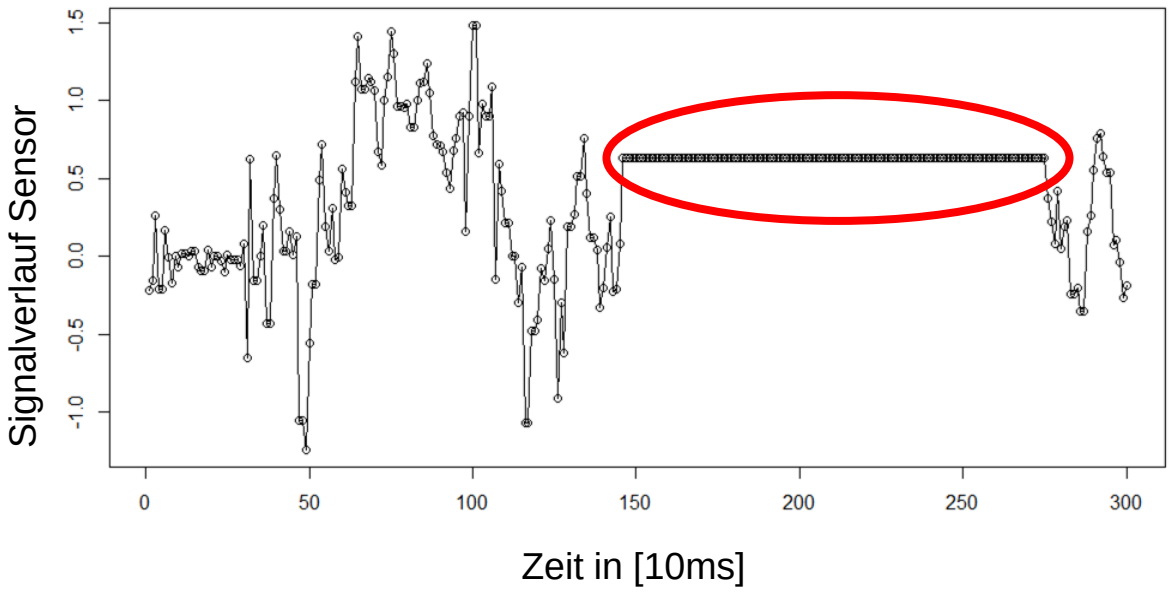
> *„Im Allgemeinen müssen Data Scientists anfangs beträchtliche Zeit dafür aufwenden, die Variablen zu definieren, die später verwendet werden. Gerade hier kommen die menschliche Kreativität, der gesunde Menschenverstand und das Fachwissen ins Spiel. Die Qualität einer Data-Mining-Lösung beruht oft darauf,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

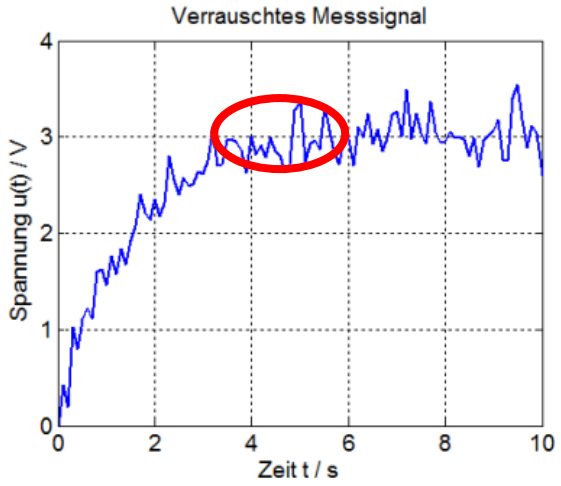
- Datenbereinigung
- Datenmanipulation

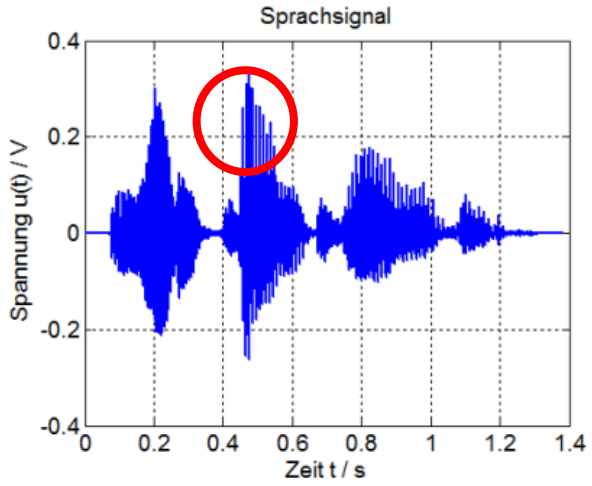


Big Data als Prozess

Datenaufbereitung







- Fehlerhafte („verschmutze“) Daten 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)

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

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

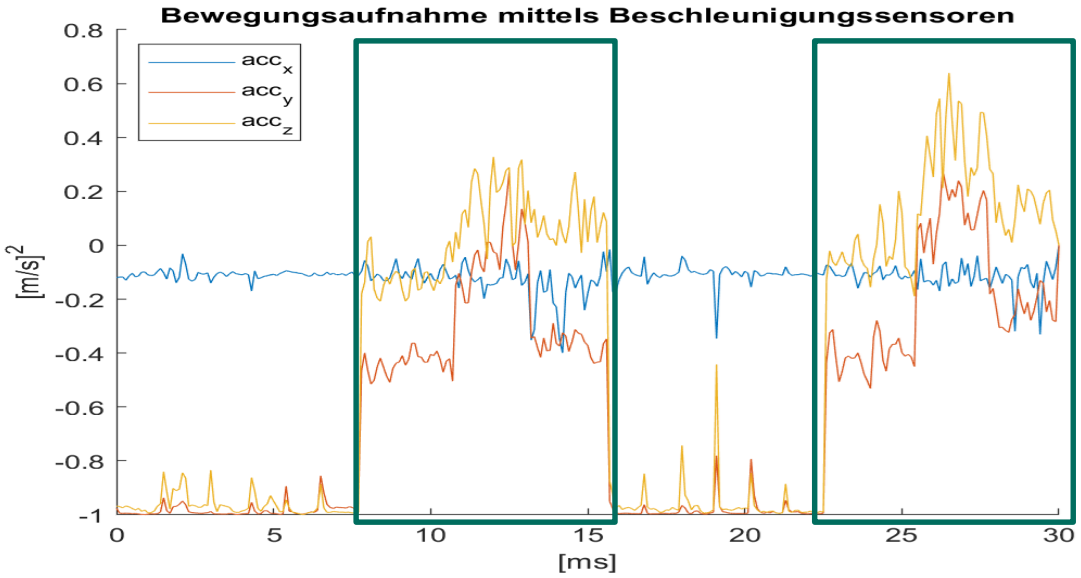
> *„In order to be a data scientist you need data. In fact, as a data scientist you will spend an embarrassingly large fraction of your time acquiring, cleaning and tranforming data. […]“*
> *~ Joel Grus*

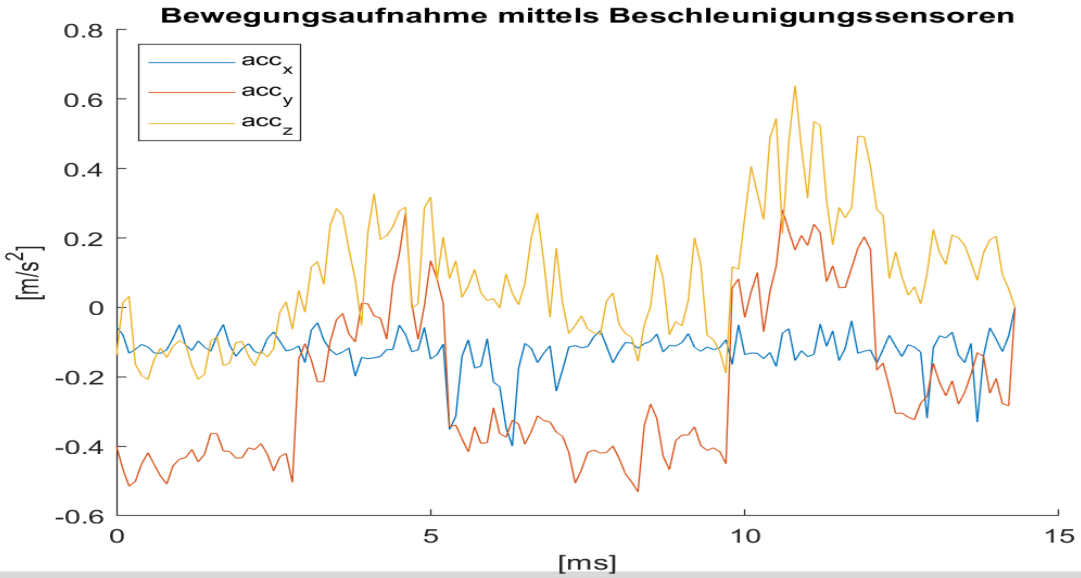
Datenaufbereitung

Datenbereinigung - 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?



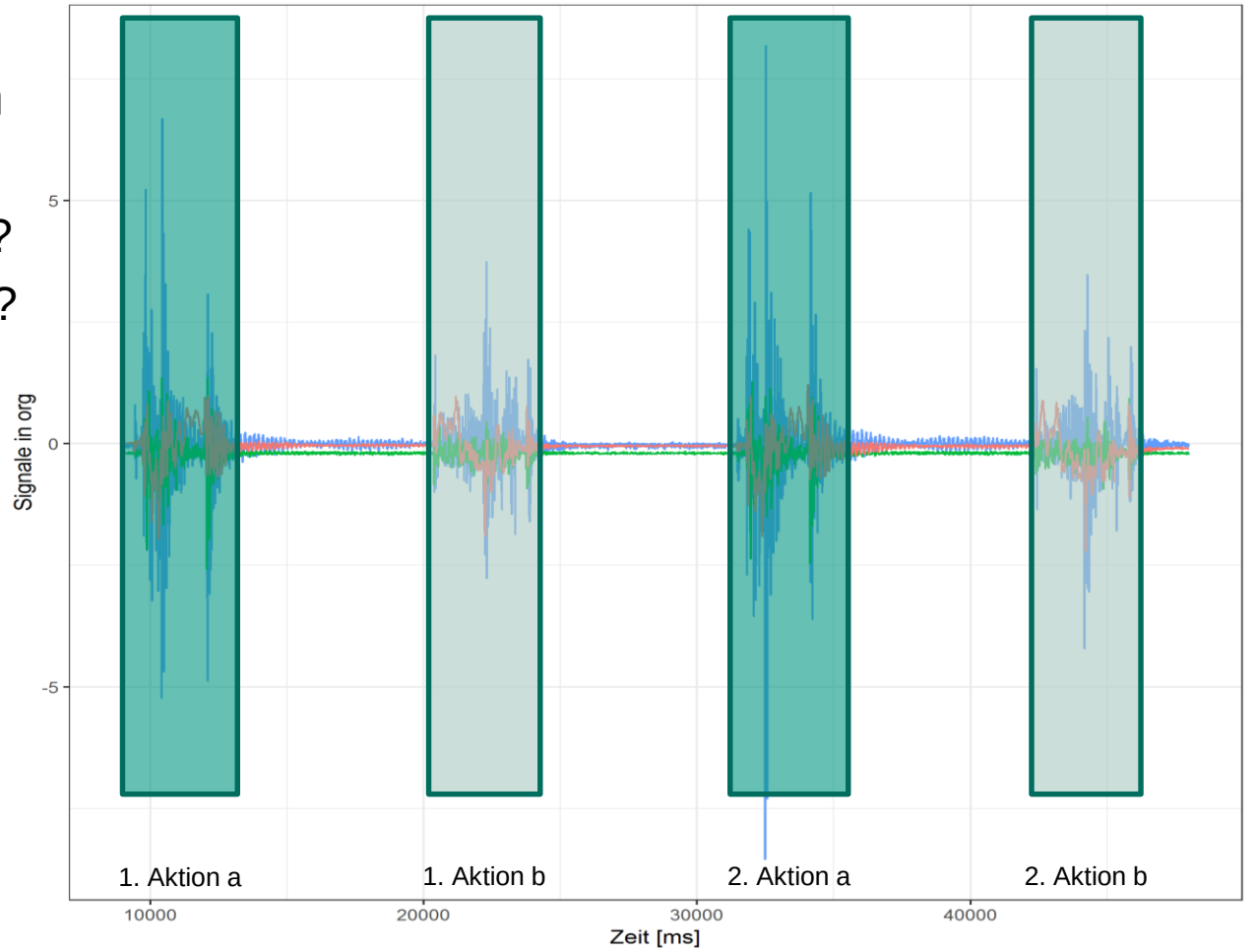


Datenaufbereitung

Datenbereinigung - Fensterauswahl

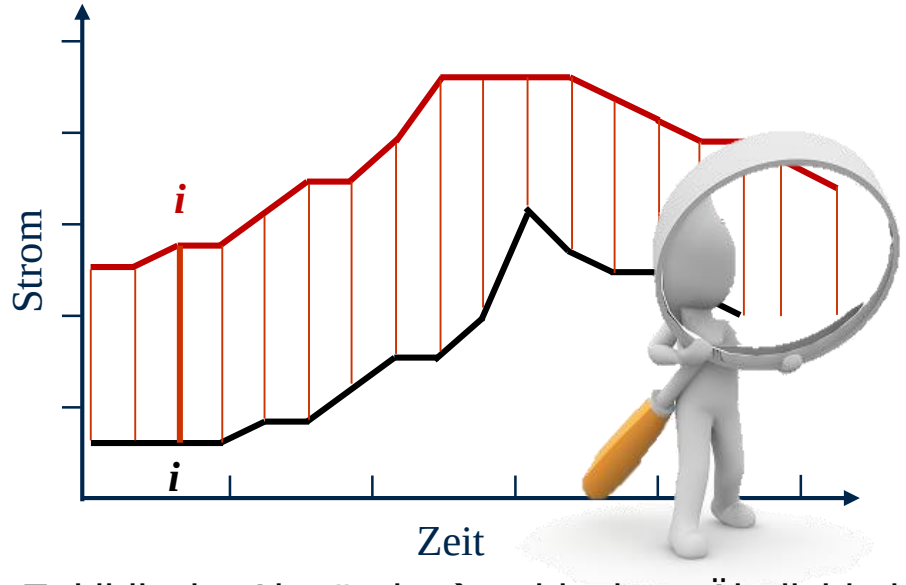
Acc_x	Acc_y	Acc_z
0.33150518	-0.036584496	-0.10886677
0.26663115	-0.043309471	-0.14096216
0.24042087	0.0010854188	-0.11888144
0.29112809	0.015567293	-0.10429218
0.35552927	-0.023789742	-0.12723972
0.28744253	-0.050273681	-0.13274829
0.1844141	-0.014817877	-0.089991964
0.25204831	-0.00060425372	-0.070452518
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?



Fensterauswahl
Methodische Unterstützung

- Reduktion der Daten auf relevante Situationen, um
  - invalide Ergebnisse zu vermeiden (Zusammenhang anderer Situationen)
  - Datenmenge für weitere Bearbeitung zu reduzieren
- Wiederfindung ähnlicher Situationen / Signalverläufe in Daten
  - Herausforderung: Streckung oder Stauchung gleicher Situationen

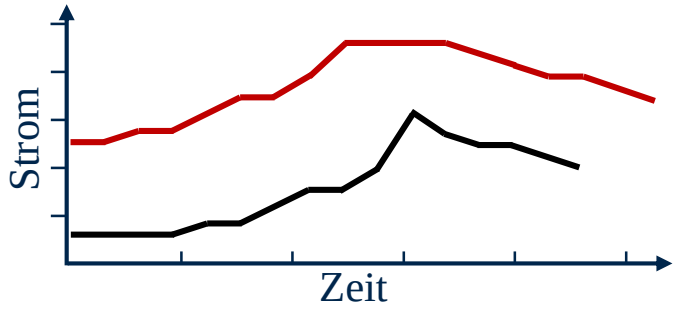


Euklidische Abstände → schlechtes Ähnlichkeitsmaß

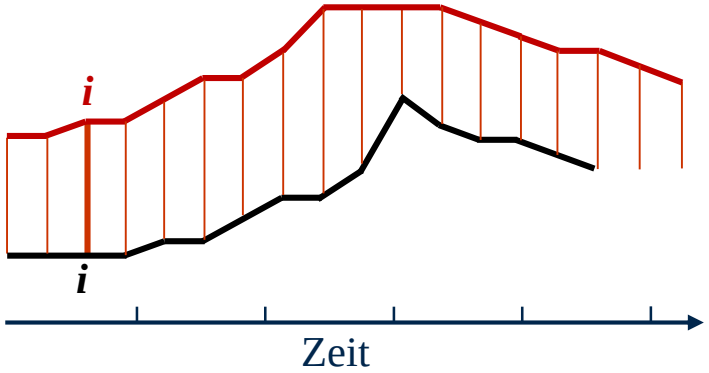
Fensterauswahl

Dynamic Time Warping: Idee

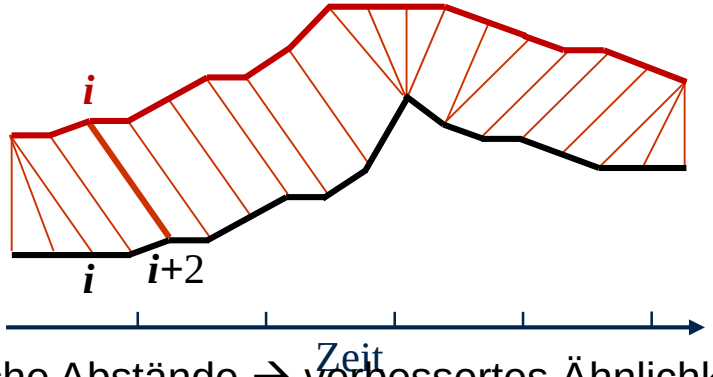
- Erkennung von Fehlerfällen anhand des Signalverlaufs
  - Herausforderung: ähnliche, aber unterschiedliche Verläufe
    - Z.B. zwei verschiedene Stromverläufe der selben Situation



- Methode: Dynamic Time Warping (DTW)



Euklidische Abstände → schlechtes Ähnlichkeitsmaß



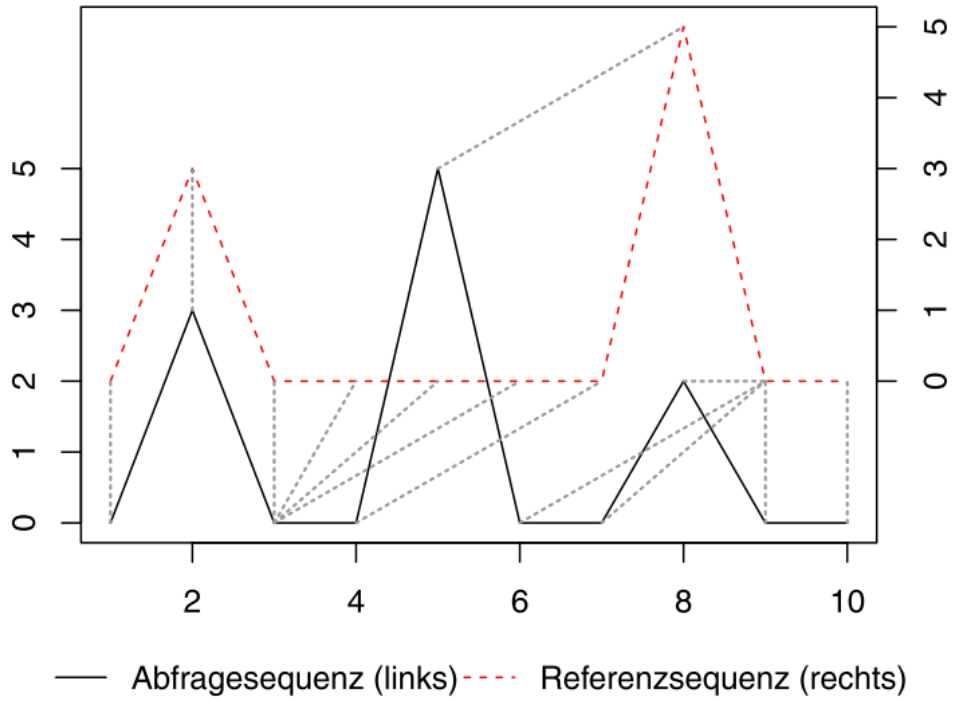
Elastische Abstände → verbessertes Ähnlichkeitsmaß

Fensterauswahl

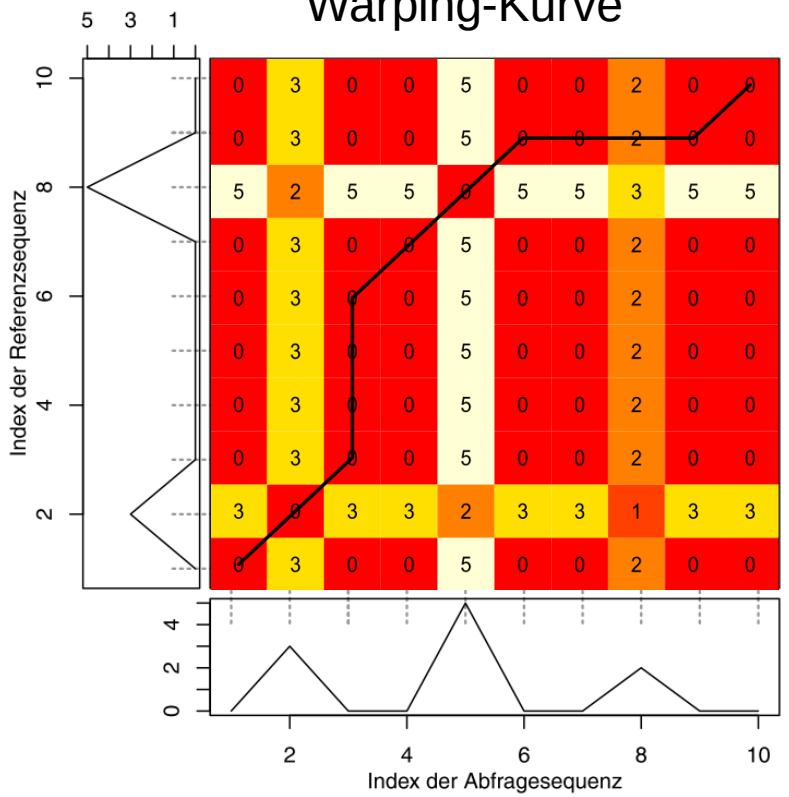
Dynamic Time Warping: Funktionsweise

- Vergleich zweier Signals
  - Berechnung des Abstandes der Signale anhand einer Warping-Kurve

Zuordnung anhand der Warping-Kurve



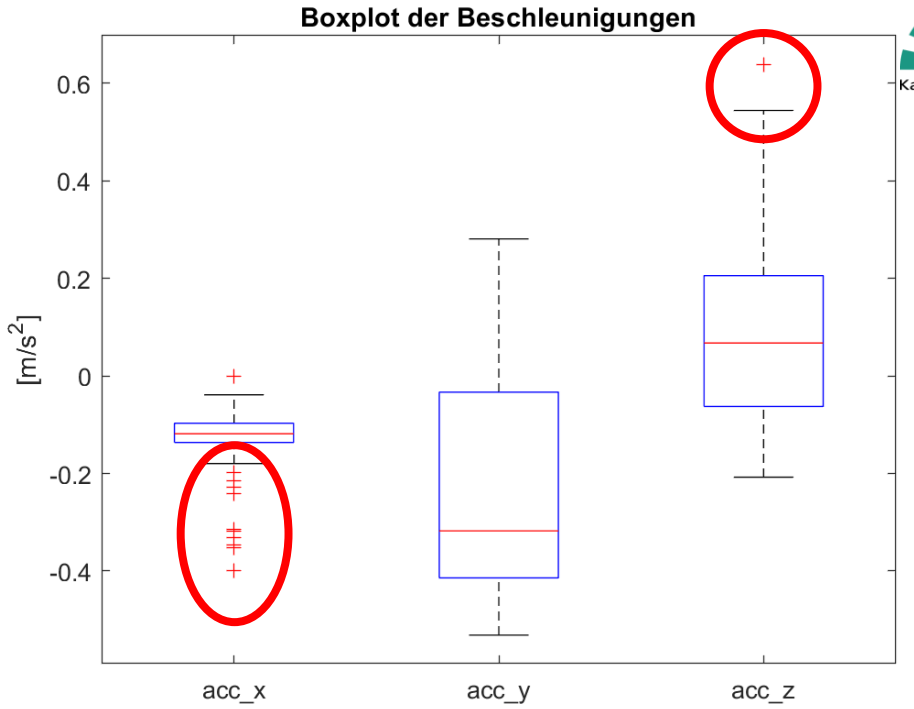
Warping-Kurve

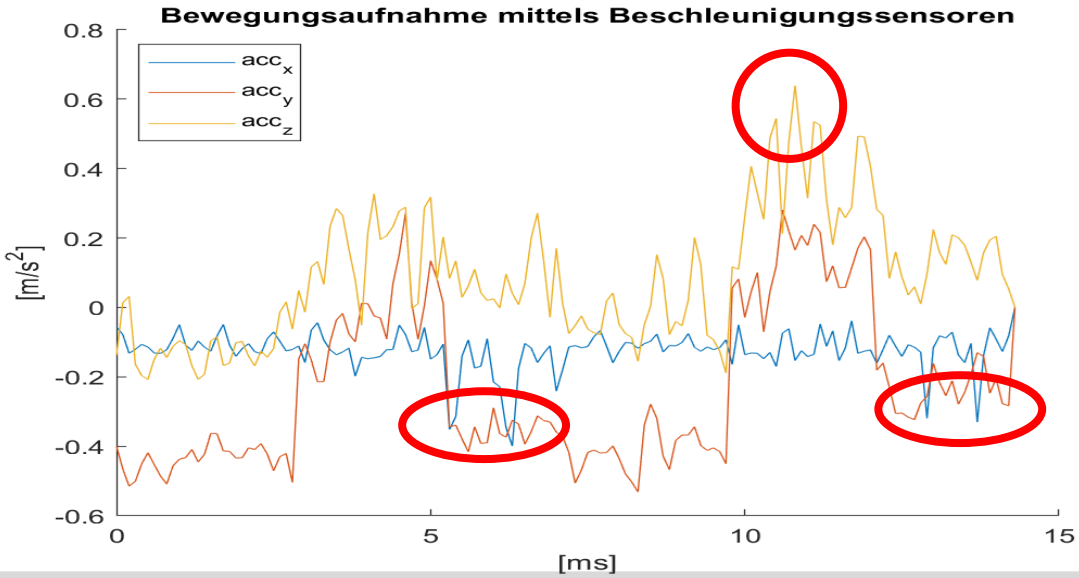


Datenaufbereitung

Datenbereinigung - Anomaliedetektion

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





Datenaufbereitung – Anomaliedetektion über Anomaliemaß

Zwischenübung

- Führen Sie eine Anomaliedetektion von Sensor 1 und 2 über die **Standardabweichung** und über die **Whisker eines Boxplots** durch:
  - Als Akzeptanzbereich für die **Standardabweichung** wurde 1,9σ festgelegt
  - Als **Whisker Grenze** wurde das 1,5 Fache des Interquartilsabstands (IQR) festgelegt

Zeit (h).	Sensor 1	Sensor 2
3	13	125
6	8	145
9	255	99
11	n/a	278
15	15	133
19	54	155
21	9	n/a
24	5	136
1534	-13	-12
1537	-6	58

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 $= \begin{cases} x_{n+1/2} & n\ ungerade \\ 1/2\left(x_{n/2} + x_{(n/2)+1}\right) & n\ gerade \end{cases}$

1. bzw. 3 Quartil ist der Median aus den Zahlen unterhalb bzw. oberhalb des 50%-Medians
(bei gerade Anzahl wird der 50%-Median hinzugezogen)

- Sensor 1:
  - Mittelwert = 22
  - Standardabweichung = 79,79
  - 1. Quartil = 5
  - 3. Quartil = 15
- Sensor 2:
  - Mittelwert = 124,11
  - Standardabweichung = 73,65
  - 1. Quartil = 99
  - 3. Quartil = 145

Datenaufbereitung – Anomaliedetektion über Anomaliemaß

Zwischenübung - Lsg

- Führen Sie eine Anomaliedetektion von Sensor 1 und 2 über die **Standardabweichung** und über die **Whisker eines Boxplots** durch:
  - Als Akzeptanzbereich für die **Standardabweichung** wurde 1,9σ festgelegt
  - Als **Whisker Grenze** wurde das 1,5 Fache des Interquartilsabstands (IQR) festgelegt

Zeit (h).	Sensor 1	Sensor 2
3	13	125
6	8	145
9	213	99
11	15	278
15	n/a	133
19	54	155
21	9	n/a
24	5	136
1534	-113	-12
1537	-6	58
$\bar{m}$	22	124,11
σ	79,79	73,65

- Standardabweichung:
  - $obere\ Grenze\ = \bar{m} + \sigma\ * 1{,}9$
  - $untere\ Grenze = \bar{m} - \sigma\ * 1{,}9$

→ Sensor 1 $obere\ Grenze\ = 22 + 79{,}79\ * 1{,}9 = 173{,}6$
untere $Grenze\ = 22 - 79{,}79\ * 1{,}9 = -129{,}6$

→ Sensor 2 $obere\ Grenze\ = 264{,}05$
untere $Grenze\ = -15{,}83$

Datenaufbereitung – Anomaliedetektion über Anomaliemaß

Zwischenübung - Lsg

- Führen Sie eine Anomaliedetektion von Sensor 1 und 2 über die **Standardabweichung** und über die **Whisker eines Boxplots** durch:
  - Als Akzeptanzbereich für die **Standardabweichung** wurde 1,9σ festgelegt
  - Als **Whisker Grenze** wurde das 1,5 Fache des Interquartilsabstands (IQR) festgelegt

Zeit (h).	Sensor 1	Sensor 2
3	13	125
6	8	145
9	213	99
11	15	278
15	n/a	133
19	54	155
21	9	n/a
24	5	136
1534	-113	-12
1537	-6	58
1.Q	5	99
3.Q	15	145

- Whisker eines Boxplots:
  - $obere\ Grenze\ = 3.Q + 1{,}5\ * IRQ$
  - $untere\ Grenze = 1.Q - 1{,}5\ * IRQ$

→ Sensor 1 $obere\ Grenze\ = 15 + 1{,}5\ * 10 = 30$

$\text{untere}\ Grenze\ = 5 - 1{,}5\ * 10\ = -10$

→ Sensor 2 $obere\ Grenze\ = 145 + 1{,}5\ * 46 = 214$

$\text{untere}\ Grenze\ = 99\ - 1{,}5\ * 46 = 30$

- Datenbereinigung
  - Fehlende/Fehlerhafte Werte
  - Fensterung
  - Anomaliedetektion

DATENMANIPULATION

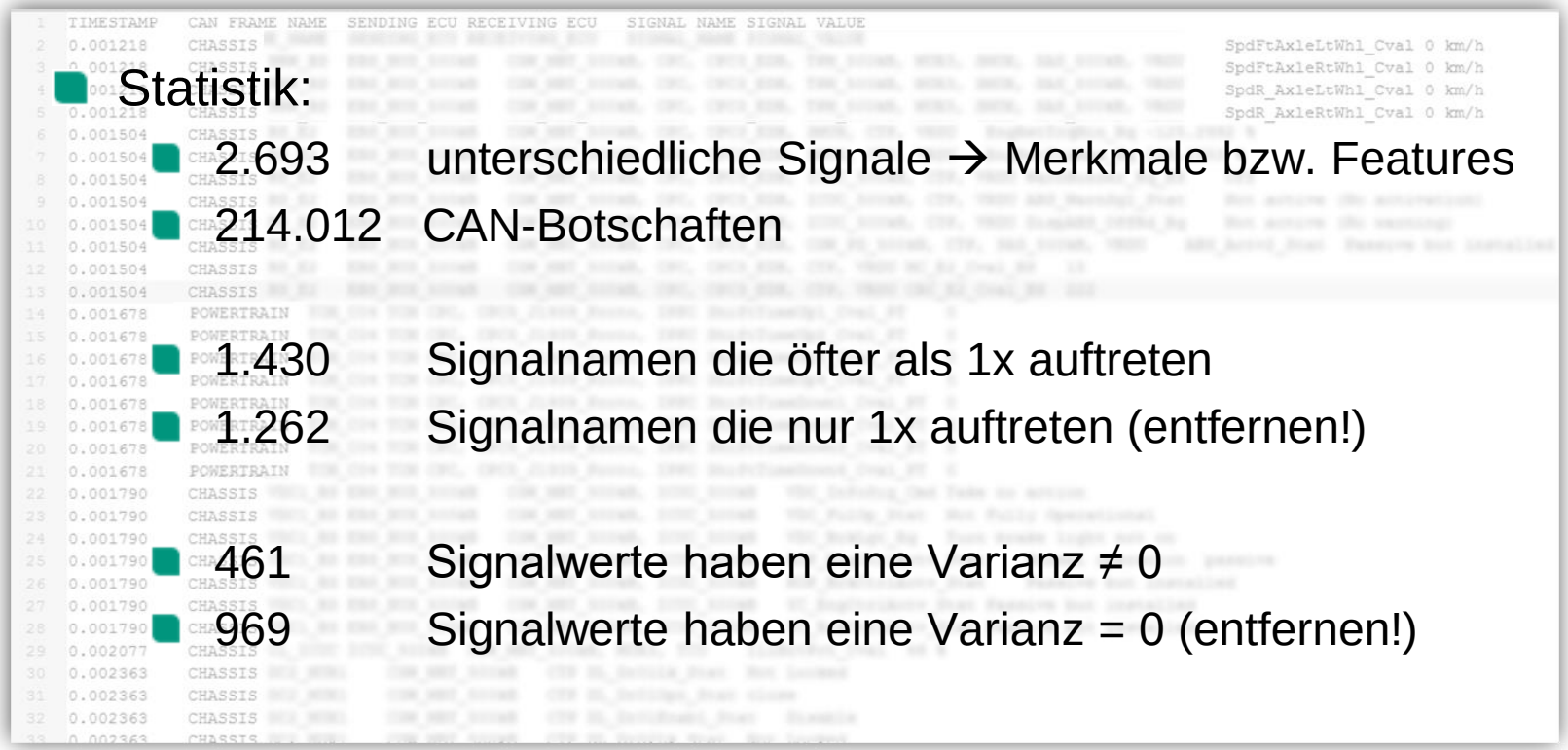
Datenaufbereitung
Datenmanipulation

- Datenmanipulation ist ein allgemeiner Ansatz und keine bestimmte Technik
  - Es ist ein kreativer Prozess, der je nach Daten und Anwendung unterschiedlich umgesetzt werden muss
  - Datenmanipulation
    1. Aufbereitung der Daten zur weiteren Verarbeitung
    2. Konvertierung der Daten nach Use Case
       - Normierung, Standardisierung
       - Zeitsynchronisation
       - Anpassung von Einheiten
    3. Umgang mit fehlerhaften Werten
    4. Qualitätsverbesserung
    5. Merkmalsreduktion

Datenaufbereitung

Bsp.: CAN Log aus einem Fahrzeug

- Steuergeräte im Fahrzeug, senden Signale nicht zeitsynchron, sondern nacheinander über einen Bus (z.b. CAN-Bus) → keine gemeinsame Zeitbasis zwischen den Signalen
  - Aufzeichnung von Chassis und Powertrain Bus
    - Zeitpunkt, Bus, Frame-Name, Sending, Receiving, Signal-Name, Signal-Wert
    - ca. 90,2 Sekunden → 1.706.860 Signal-Werte



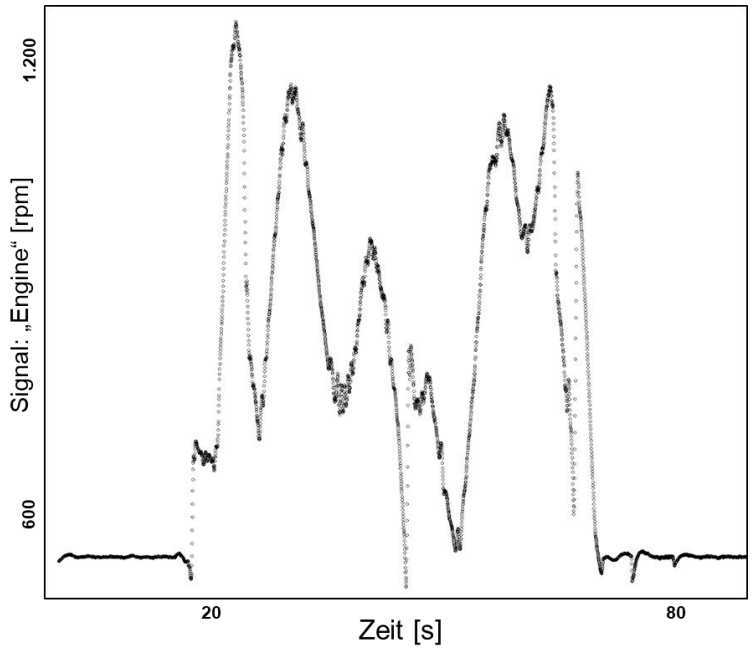
- Statistik:
  - 2.693 unterschiedliche Signale → Merkmale bzw. Features
  - 214.012 CAN-Botschaften
  - 1.430 Signalnamen die öfter als 1x auftreten
  - 1.262 Signalnamen die nur 1x auftreten (entfernen!)
  - 461 Signalwerte haben eine Varianz ≠ 0
  - 969 Signalwerte haben eine Varianz = 0 (entfernen!)

Datenaufbereitung

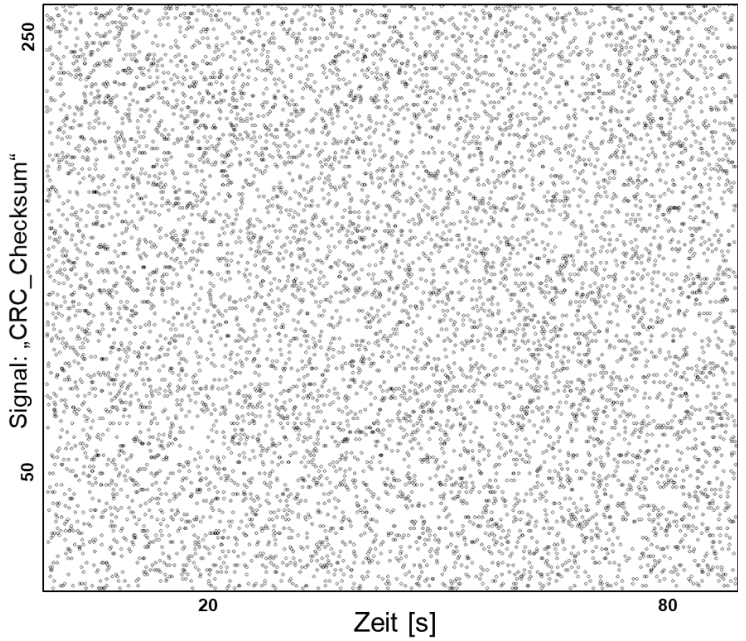
Beispiel einer Merkmalsselektion für CAN-Signale

- Ziel der Merkmalsselektion:
  - Reduzierung der Merkmale auf die, die den größten Informationsgewinn für die gewünschte Anwendung haben
    - ➢ Reduzierung des Rechenaufwands und somit des Trainings im „Modeling“
    - ➢ Reduzierung der Gefahr von „Scheinkorrelationen“ (z.B. Schnee bei der Klassifikation von Husky und Wolf)
- Frage: wie können die 461 Signale / Merkmale / Features weiter eingegrenzt werden?

Sensor-Signal



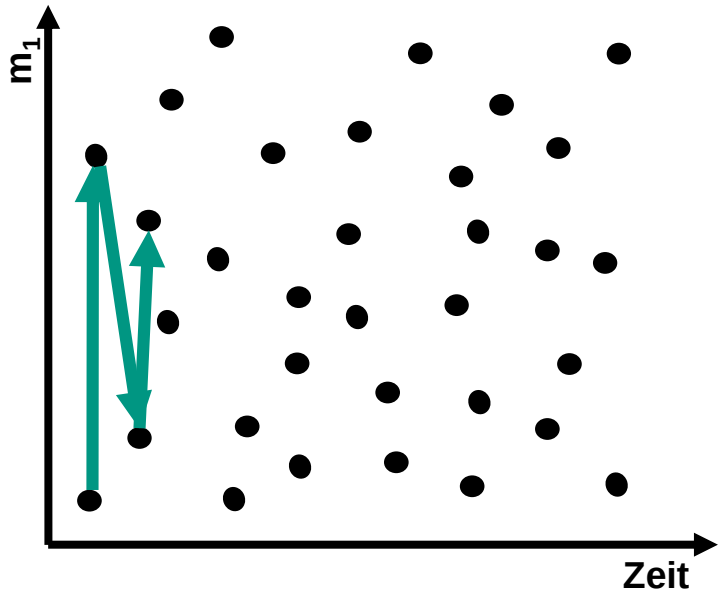
Kein Sensor-Signal



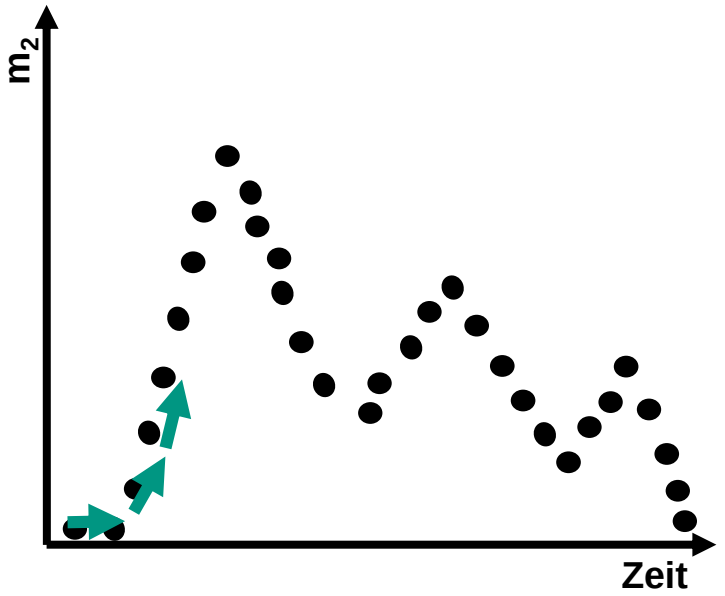
Datenaufbereitung

Beispiel einer Merkmalsselektion für CAN-Signale

- Filterung aller Signale mit physikalischem Bezug
  - Idee:



Signal ohne phy. Bezug



Signal mit phy. Bezug

  - Normierung der Signal-Werte
  - Berechnung des Ø der euklidischen Abstände über die Zeit
  - Reduzierung von 461 auf 119 Signale bei Schwellwert von 0,3 (empirisch)

Datenaufbereitung

Datenmanipulation - Konvertierung

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

CAN-Frame 1

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)
…	…	…	…

CAN-Frame 2

Time	Signal 1	…	Signal 13
$t_{2.1}$=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)
…	…	…	…

CAN-Frame 3

Time	Signal 1	…	Signal 4
$t_{3.1}$=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 2.13	Signal 3.1	…	Signal 3.4
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)
…	…	…	…	…	…	…	…	…	…

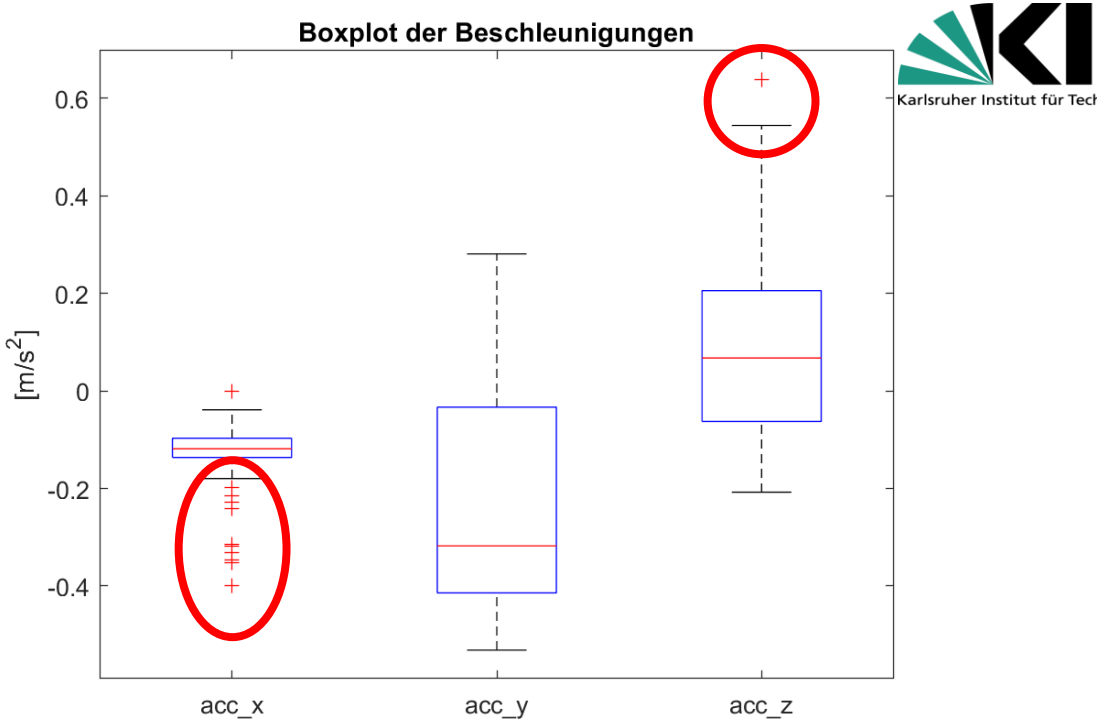
Datenaufbereitung

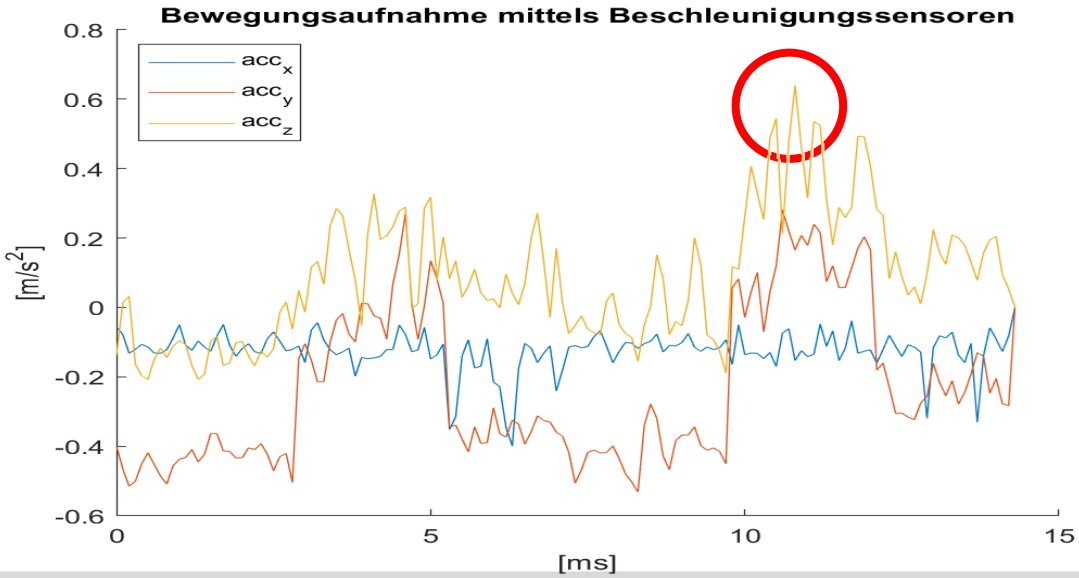
Datenmanipulation – Umgang mit Ausreißern

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

Lineare Interpolation:

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





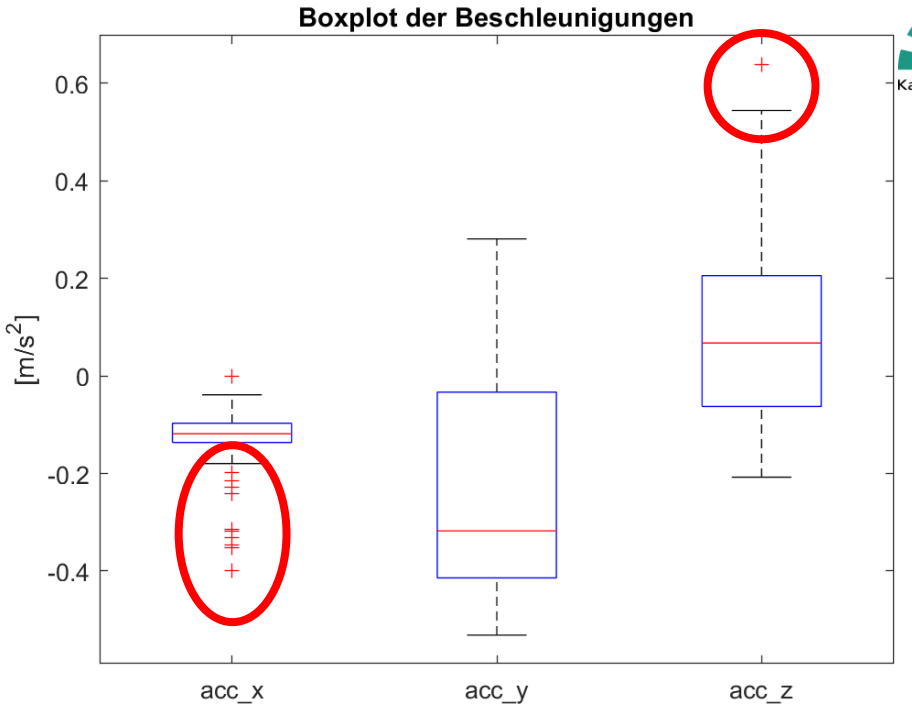
Datenaufbereitung

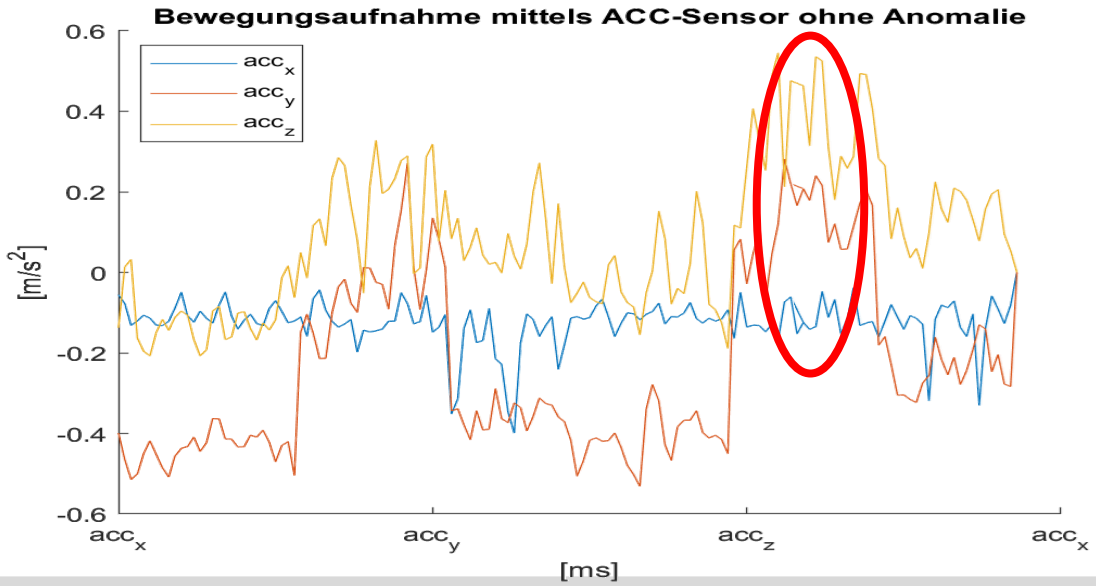
Datenmanipulation – Umgang mit Ausreißern

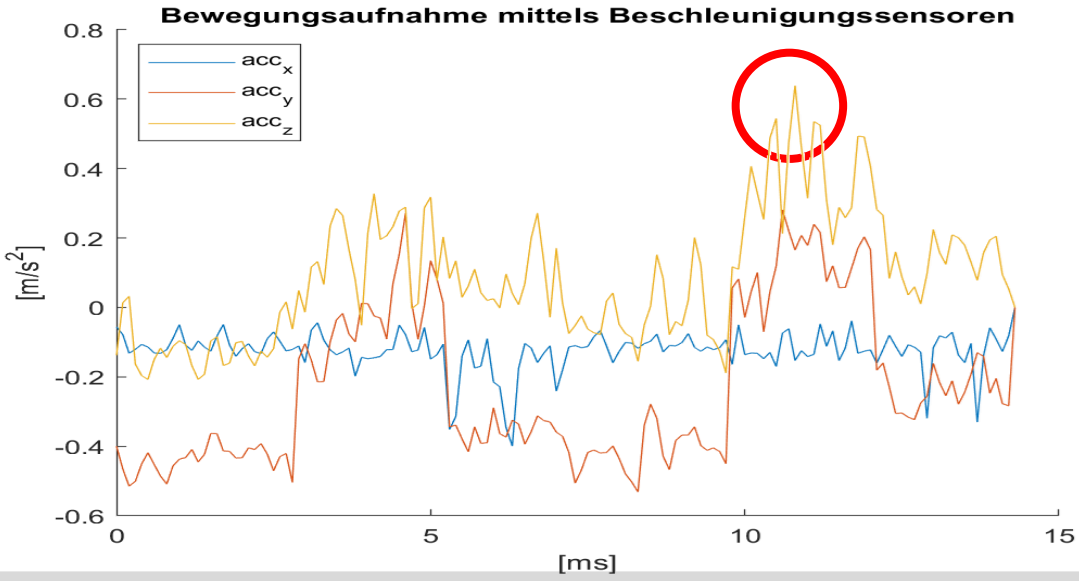
- Anomaliedetektion durch Boxplotdarstellung
- Umgang mit Anomalie *(hier)*
  - Interpolation
    - Aller Signale
    - Nur des Signals mit enthaltenem Fehler

Lineare Interpolation:

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







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:

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

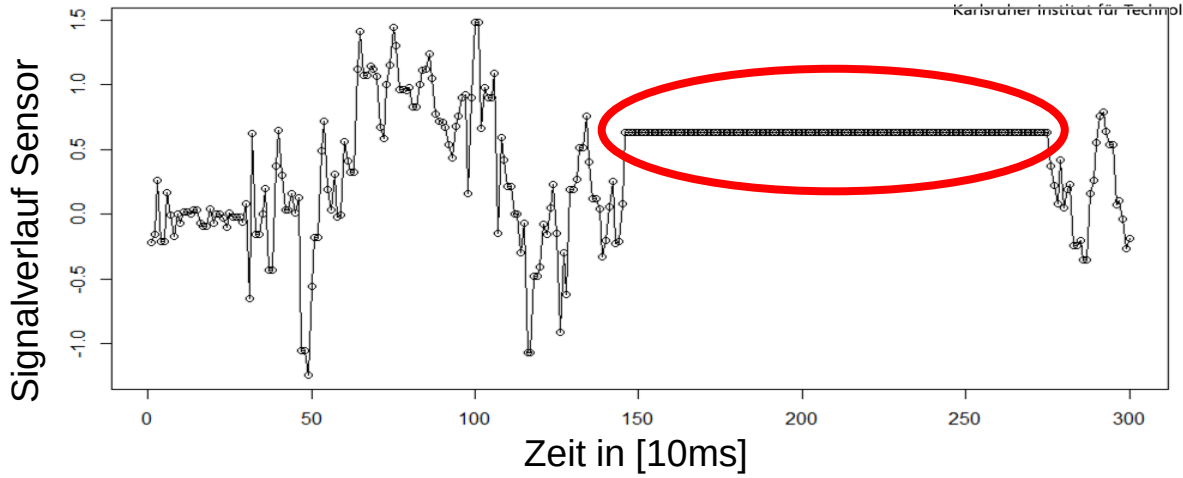
Lineare Interpolation:

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

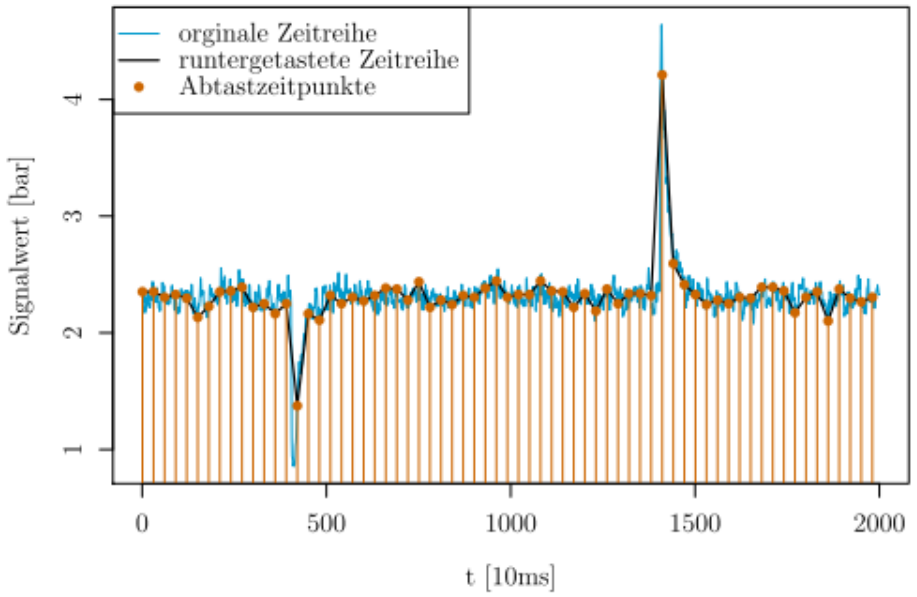
Datenaufbereitung

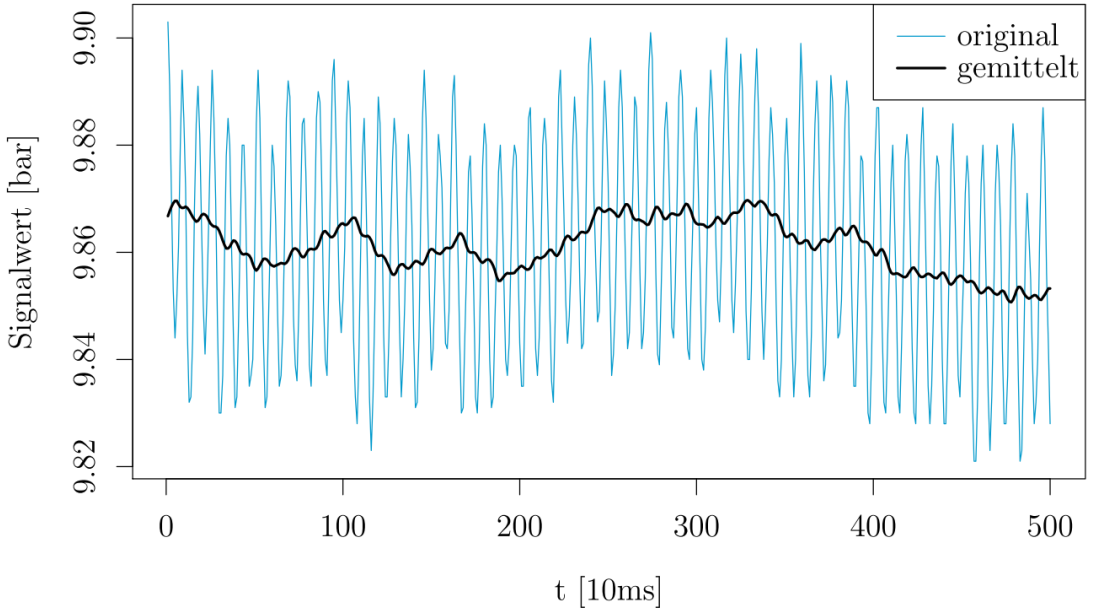
Datenmanipulation - Qualitätsverbesserung

- Entfernung von fehlerhaften Messungen



- Glättung von Messwerten
  (z.B. durch „Downsampling“ oder „Moving Average“)





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

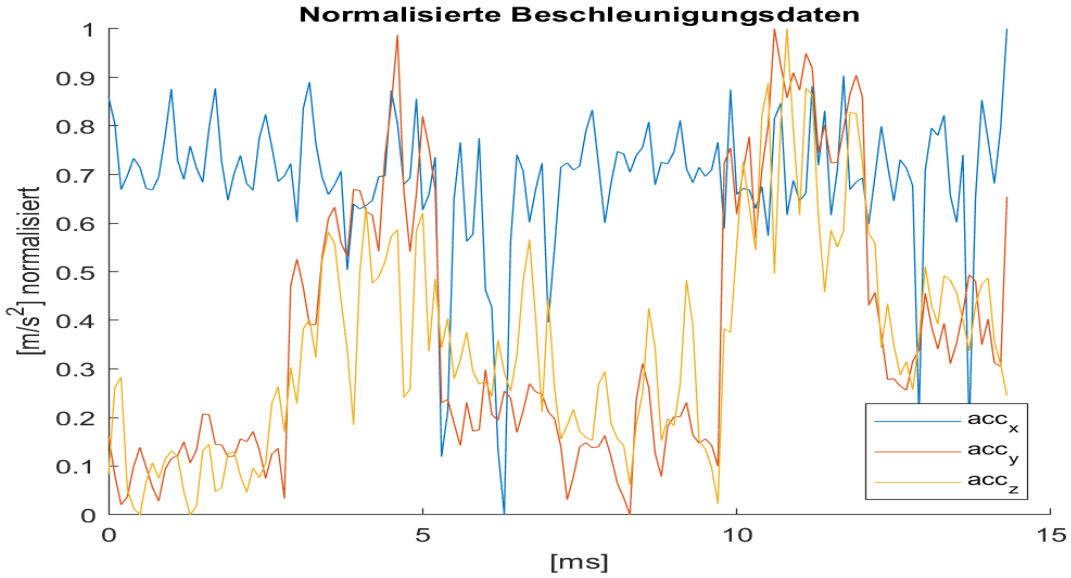
Activity
1
1
1
…
2
2
2
2
2
…
3
3
3
3

- 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 $x^{new} = \log_a x$

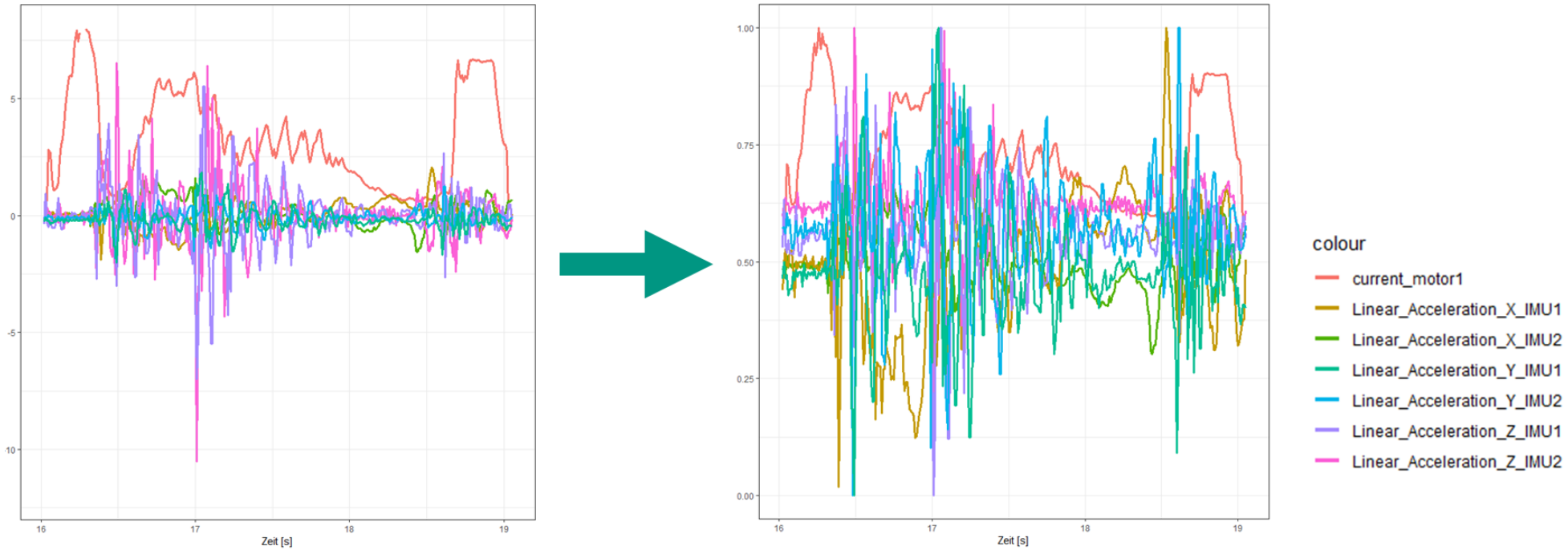
Beispiel: Beschleunigungen auf [0,1] normieren



Datenaufbereitung

Wichtigkeit der Normierung

- Beim maschinellen Lernen werden die Daten / Instanzen mittels mathematischen Funktionen miteinander verglichen
  - Um nicht Äpfel mit Birnen zu vergleichen, müssen die Merkmale untereinander vergleichbar sein! (Bsp. Durchschnittliche Drehzahl bei Diesel Motoren ist niedriger als die von Otto-Motoren)
  - Die Normierung ist somit ein wichtiger Schritt vor dem Modeling



Original Daten

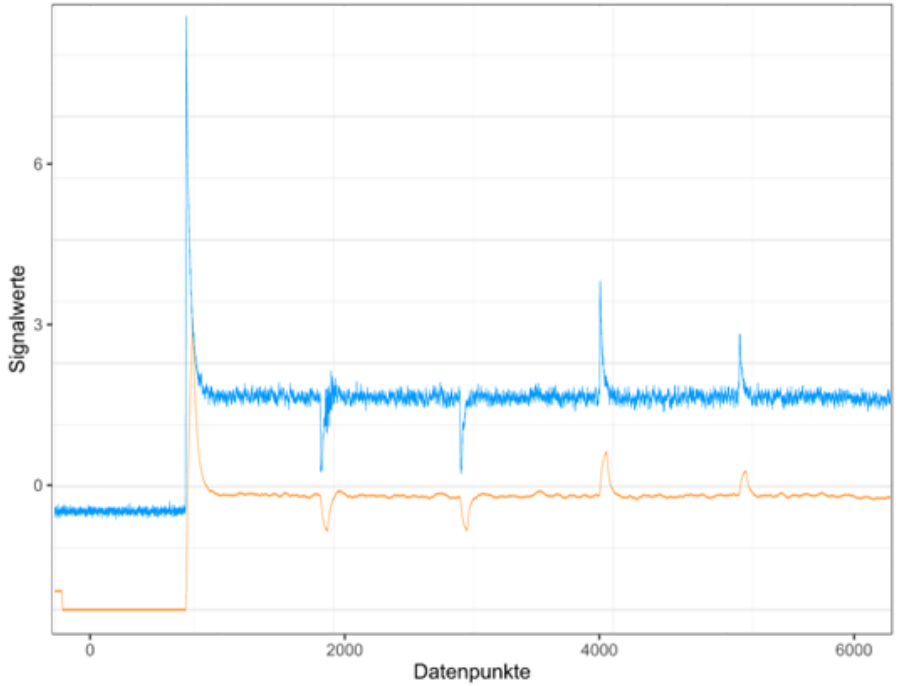
Daten nach Min/Max-Normierung

Datenaufbereitung

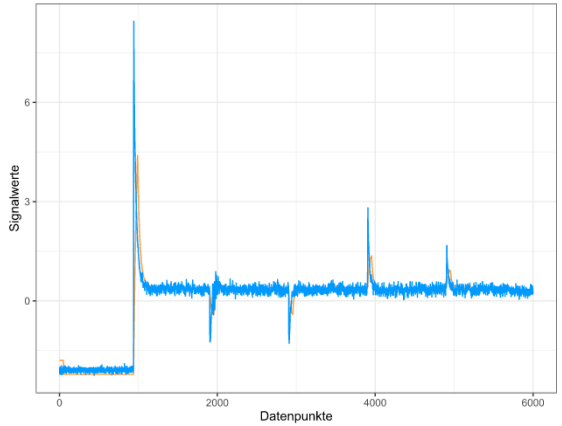
Wichtigkeit der Normierung

- Beim maschinellen Lernen werden die Daten / Instanzen mittels mathematischen Funktionen miteinander verglichen
  - Um nicht Äpfel mit Birnen zu vergleichen, müssen die Merkmale untereinander vergleichbar sein! (Bsp. Durchschnittliche Drehzahl bei Diesel Motoren ist niedriger als die eine Otto-Motors)
  - Die Normierung ist somit ein wichtiger Schritt vor dem Modeling

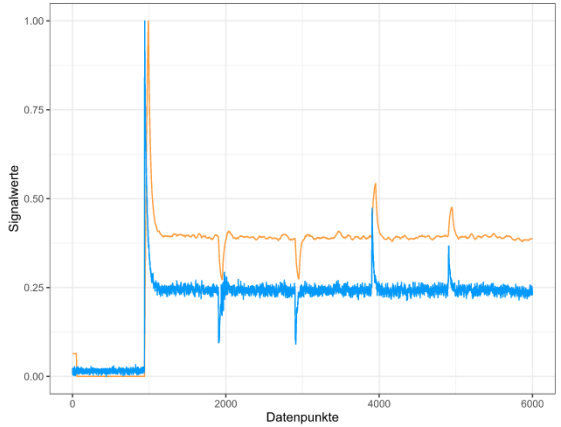
Ohne Normierung



Z-normalisiert



Min/Max-normalisiert

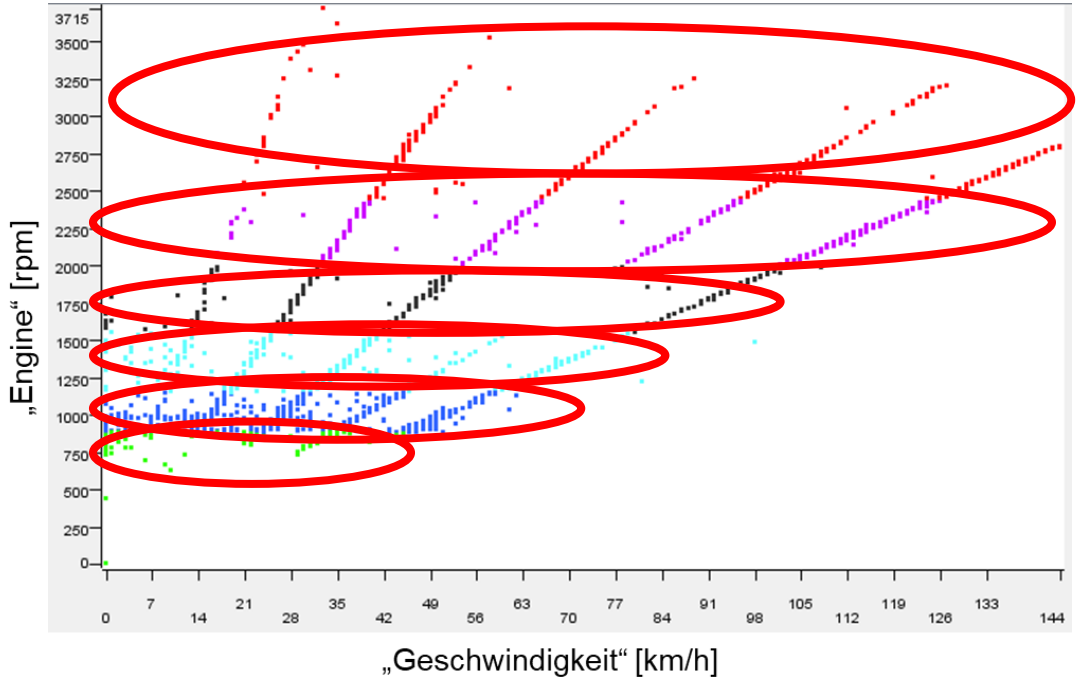


Datenaufbereitung

Wichtigkeit der Normierung anhand eines Bsp.

- Beim maschinellen Lernen werden die Daten / Instanzen mittels mathematischen Funktionen miteinander verglichen
  - Um nicht Äpfel mit Birnen zu vergleichen, müssen die Merkmale untereinander vergleichbar sein! (Bsp. Durchschnittliche Drehzahl bei Diesel Motoren ist niedriger als die eine Otto-Motors)
  - Die Normierung ist somit ein wichtiger Schritt vor dem Modeling

- Bespiel „Gangerkennung eines Fhzg.“
  - Verwendet wurde ein k-means Clustering (genaueres in Übung 7)
  - Die Cluster werden über eine Abstandsfunktion gebildet (z.B. euklidische Distanz)
  - Ergebnis ohne Normierung (Drehzahl zwischen 0 und 3.715; Geschwindigkeit zwischen 0 und 144

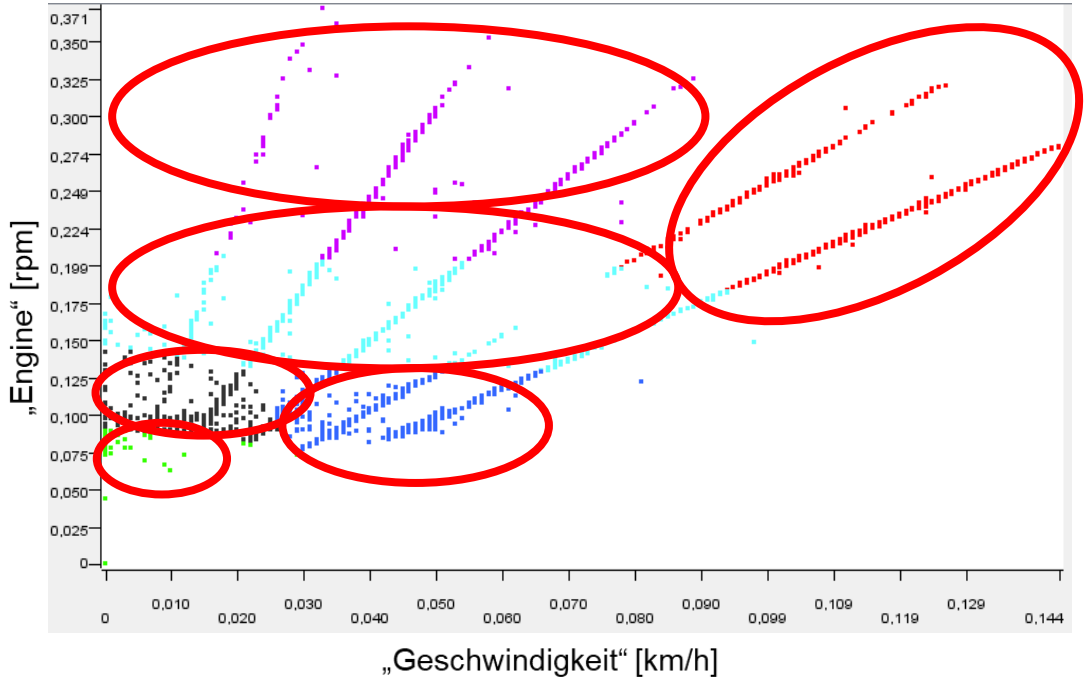


Datenaufbereitung

Wichtigkeit der Normierung anhand eines Bsp.

- Beim maschinellen Lernen werden die Daten / Instanzen mittels mathematischen Funktionen miteinander verglichen
  - Um nicht Äpfel mit Birnen zu vergleichen, müssen die Merkmale untereinander vergleichbar sein! (Bsp. Durchschnittliche Drehzahl bei Diesel Motoren ist niedriger als die eine Otto-Motors)
  - Die Normierung ist somit ein wichtiger Schritt vor dem Modeling

- Bespiel „Gangerkennung eines Fhzg.“
  - Verwendet wurde ein k-means Clustering (genaueres in Übung 7)
  - Die Cluster werden über eine Abstandsfunktion gebildet (z.B. euklidische Distanz)
  - Ergebnis mit Dezimal-Skalierung (Drehzahl zwischen 0 und 0,371; Geschwindigkeit zwischen 0 und 0,144

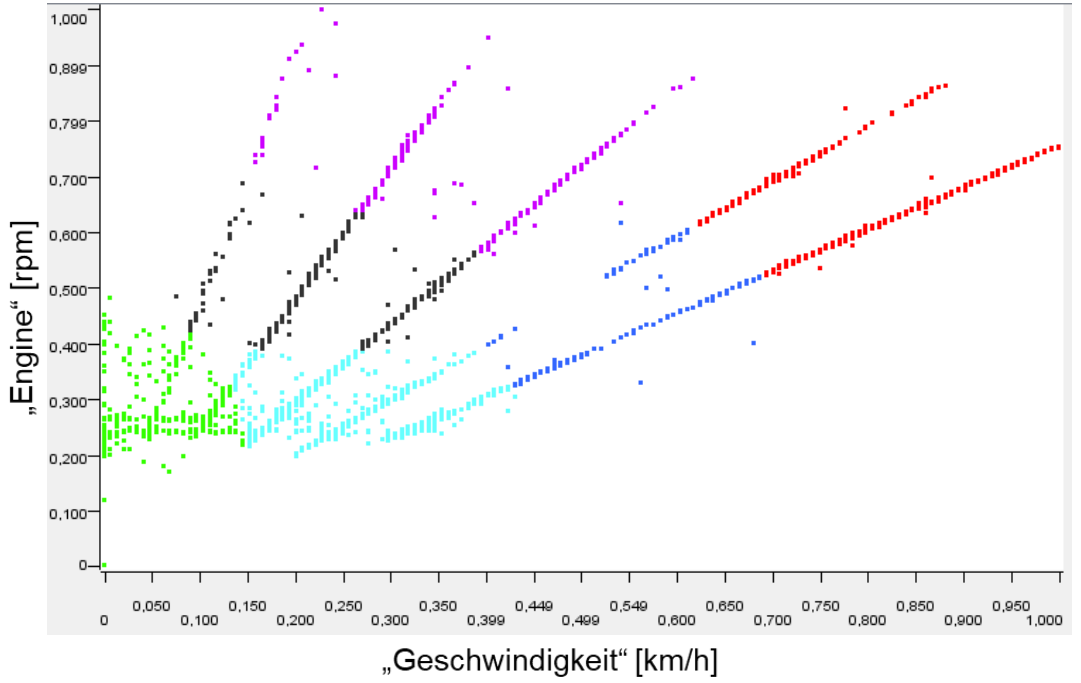


Datenaufbereitung

Wichtigkeit der Normierung anhand eines Bsp.

- Beim maschinellen Lernen werden die Daten / Instanzen mittels mathematischen Funktionen miteinander verglichen
  - Um nicht Äpfel mit Birnen zu vergleichen, müssen die Merkmale untereinander vergleichbar sein! (Bsp. Durchschnittliche Drehzahl bei Diesel Motoren ist niedriger als die eine Otto-Motors)
  - Die Normierung ist somit ein wichtiger Schritt vor dem Modeling

- Bespiel „Gangerkennung eines Fhzg.“
  - Verwendet wurde ein k-means Clustering (genaueres in Übung 7)
  - Die Cluster werden über eine Abstandsfunktion gebildet (z.B. euklidische Distanz)
  - Ergebnis mit Min/Max-Normierung (Drehzahl zwischen 0 und 1; Geschwindigkeit zwischen 0 und 1

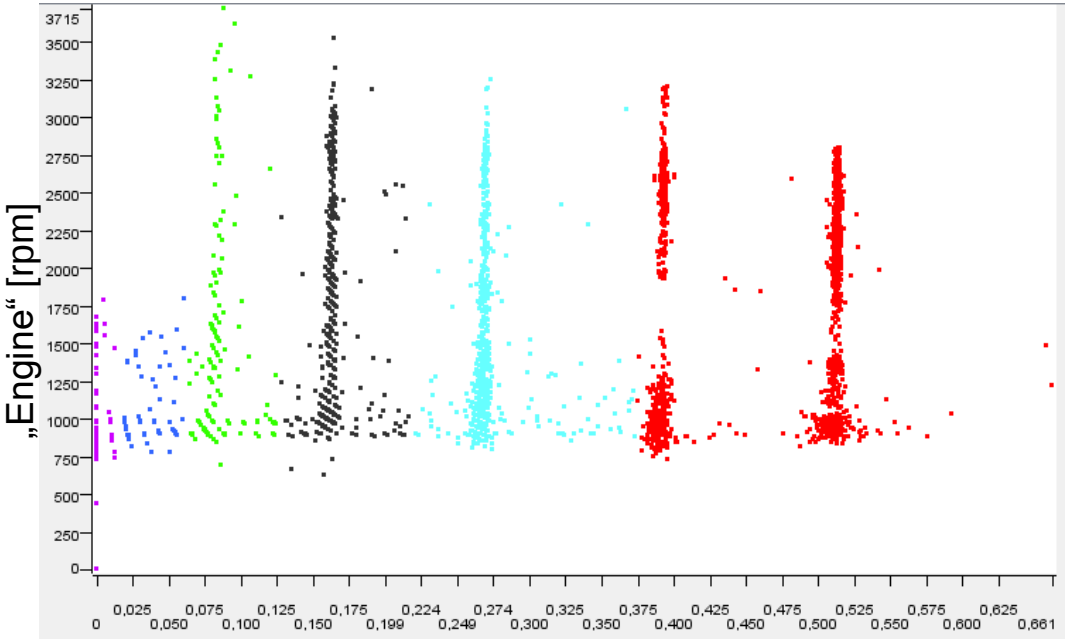


Datenaufbereitung

Wichtigkeit der Normierung anhand eines Bsp.

- Beim maschinellen Lernen werden die Daten / Instanzen mittels mathematischen Funktionen miteinander verglichen
  - Um nicht Äpfel mit Birnen zu vergleichen, müssen die Merkmale untereinander vergleichbar sein! (Bsp. Durchschnittliche Drehzahl bei Diesel Motoren ist niedriger als die eine Otto-Motors)
  - Die Normierung ist somit ein wichtiger Schritt vor dem Modeling

- Bespiel „Gangerkennung eines Fhzg.“
  - Verwendet wurde ein k-means Clustering (genaueres in Übung 7)
  - Die Cluster werden über eine Abstandsfunktion gebildet (z.B. euklidische Distanz)
  - Transformation der zwei Merkmale auf einen
    - Gangwinke = Arkustangens (Speed / RPM)

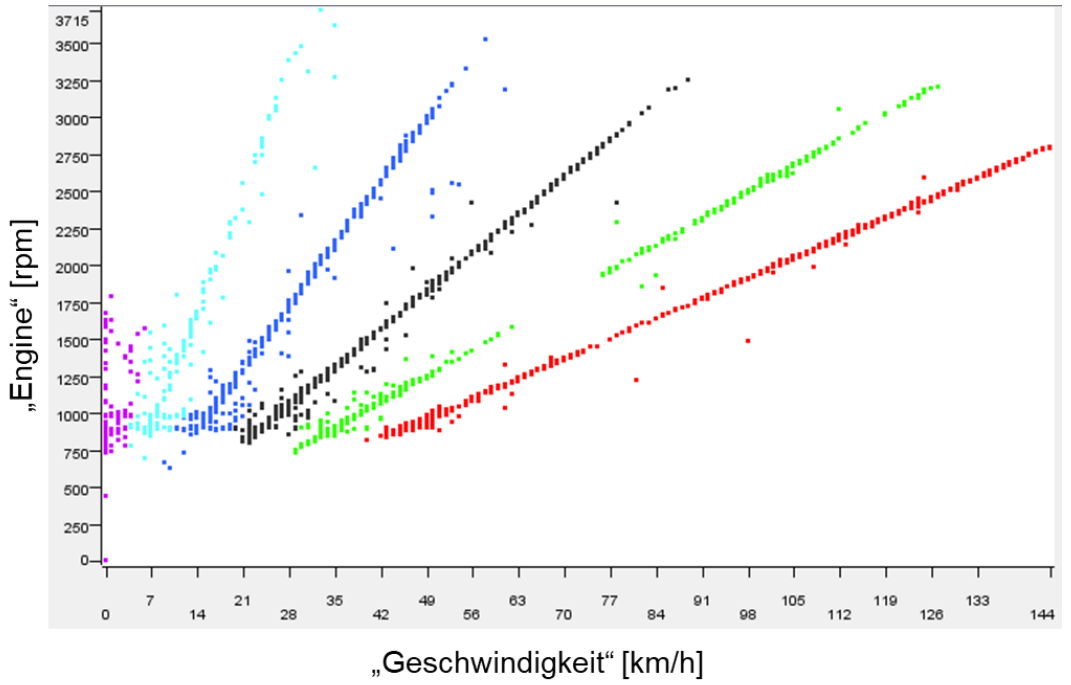


Datenaufbereitung

Wichtigkeit der Normierung anhand eines Bsp.

- Beim maschinellen Lernen werden die Daten / Instanzen mittels mathematischen Funktionen miteinander verglichen
  - Um nicht Äpfel mit Birnen zu vergleichen, müssen die Merkmale untereinander vergleichbar sein! (Bsp. Durchschnittliche Drehzahl bei Diesel Motoren ist niedriger als die eine Otto-Motors)
  - Die Normierung ist somit ein wichtiger Schritt vor dem Modeling

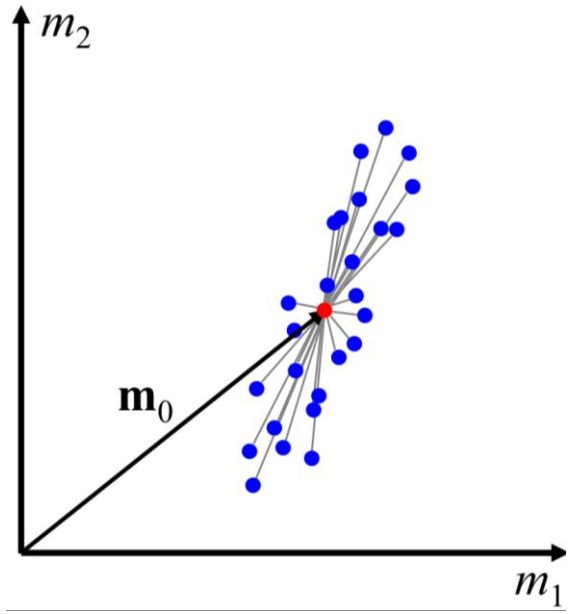
- Bespiel „Gangerkennung eines Fhzg.“
  - Verwendet wurde ein k-means Clustering (genaueres in Übung 7)
  - Die Cluster werden über eine Abstandsfunktion gebildet (z.B. euklidische Distanz)
  - Transformation der zwei Merkmale auf einen
    - Gangwinke = Arkustangens (Speed / RPM)
  - Weiterverbesserung über Optimierung der Anfangs-Seed im Clustering



Datenaufbereitung

Merkmalsreduktion – Transformation durch PCA

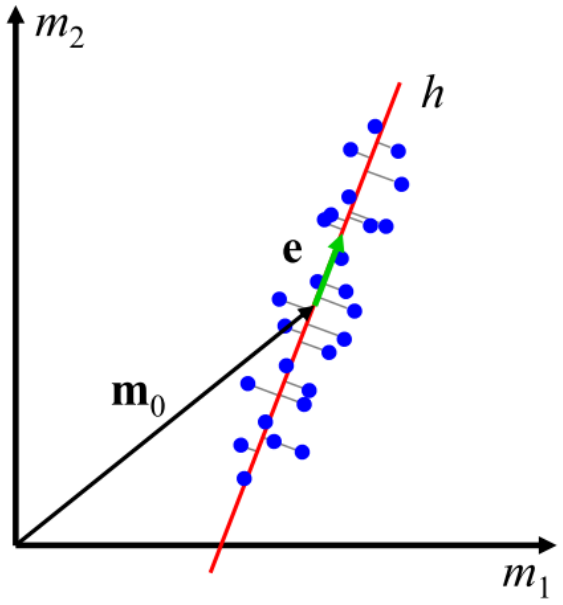
- Durch die Zeitsynchronisation kann der Datensatz inzwischen als n x m Matrix dargestellt werden, wobei n die jeweiligen Instanzen sind und m die Merkmale
- Der bereinigte und aufbereitete Datensatz kann so als Punktwolke in einem m-dimensionalen Raum dargestellt werden (jeder Punkt ist eine Instanz)
- Ziel der PCA ist es, die Datenpunkte so in einen q-dimensionalen Unterraum zu projizieren, dass dabei möglichst wenig Informationen verloren gehen (→ Reduktion der Merkmale ohne Informationsverlust)
- Mathematisch: Hauptachsentransformation
  - Minimierung der Korrelation mehrdimensionaler Merkmale durch Überführung in Vektorraum neuer Basis
  - Orthogonale Matrix, bestehend aus Eigenvektoren der Kovarianzmatrix
- Erster Schritt:
  Welcher Punkt m_0 im Merkmalsraum repräsentiert die Daten $D = \{p_1, ..., p_N\}$ mit minimalem quadratischen Fehler?



Datenaufbereitung

Merkmalsreduktion – Transformation durch PCA

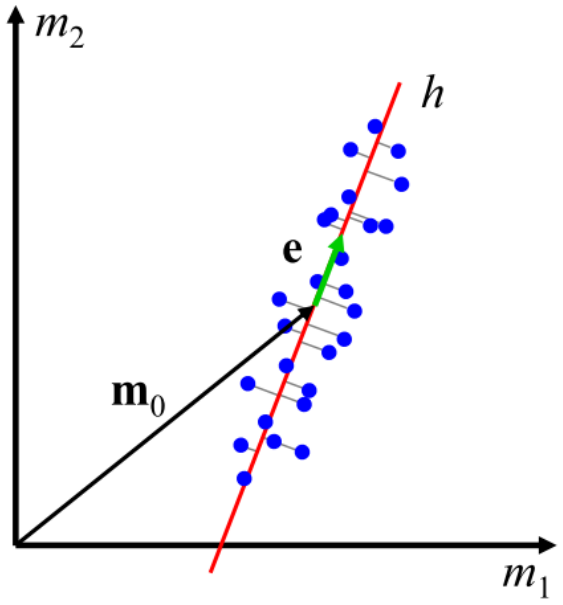
- Durch die Zeitsynchronisation kann der Datensatz inzwischen als n x m Matrix dargestellt werden, wobei n die jeweiligen Instanzen sind und m die Merkmale
- Der bereinigte und aufbereitete Datensatz kann so als Punktwolke in einem m-dimensionalen Raum dargestellt werden (jeder Punkt ist eine Instanz)
- Ziel der PCA ist es, die Datenpunkte so in einen q-dimensionalen Unterraum zu projizieren, dass dabei möglichst wenig Informationen verloren gehen (→ Reduktion der Merkmale ohne Informationsverlust)
- Mathematisch: Hauptachsentransformation
  - Minimierung der Korrelation mehrdimensionaler Merkmale durch Überführung in Vektorraum neuer Basis
  - Orthogonale Matrix, bestehend aus Eigenvektoren der Kovarianzmatrix
- Zweiter Schritt:
  Welche Gerade im Merkmalsraum repräsentiert die Daten D mit minimalem quadratischen Fehler?
  - Resultat ist die erste Hauptkomponente (PC 1)



Datenaufbereitung

Merkmalsreduktion – Transformation durch PCA

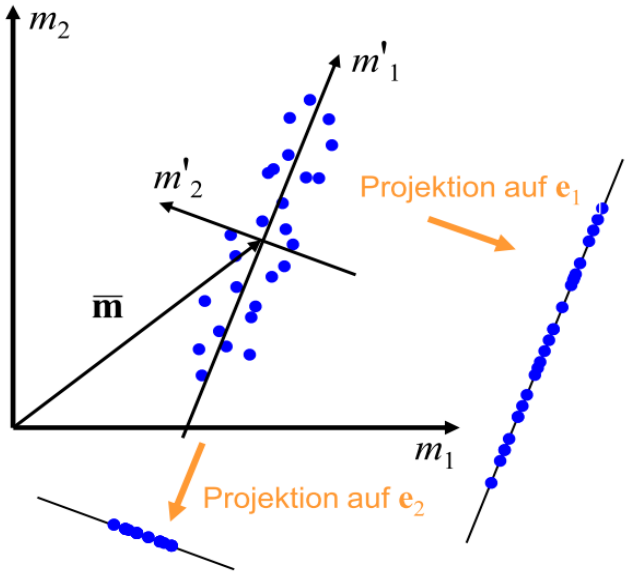
- Durch die Zeitsynchronisation kann der Datensatz inzwischen als n x m Matrix dargestellt werden, wobei n die jeweiligen Instanzen sind und m die Merkmale
- Der bereinigte und aufbereitete Datensatz kann so als Punktwolke in einem m-dimensionalen Raum dargestellt werden (jeder Punkt ist eine Instanz)
- Ziel der PCA ist es, die Datenpunkte so in einen q-dimensionalen Unterraum zu projizieren, dass dabei möglichst wenig Informationen verloren gehen (→ Reduktion der Merkmale ohne Informationsverlust)
- Mathematisch: Hauptachsentransformation
  - Minimierung der Korrelation mehrdimensionaler Merkmale durch Überführung in Vektorraum neuer Basis
  - Orthogonale Matrix, bestehend aus Eigenvektoren der Kovarianzmatrix
- Dritter Schritt:
  Welche Gerade, welche senkrecht zu den ermittelten Hauptkomponenten ist, repräsentiert die Daten weiter mit minimalem quadratischen Fehler?
- Vierter Schritt:
  Schritt drei so oft wiederholen, bis alle bzw. q Hauptkomponenten bestimmt sind



Datenaufbereitung

Merkmalsreduktion – Transformation durch Principal Component Analysis (PCA)

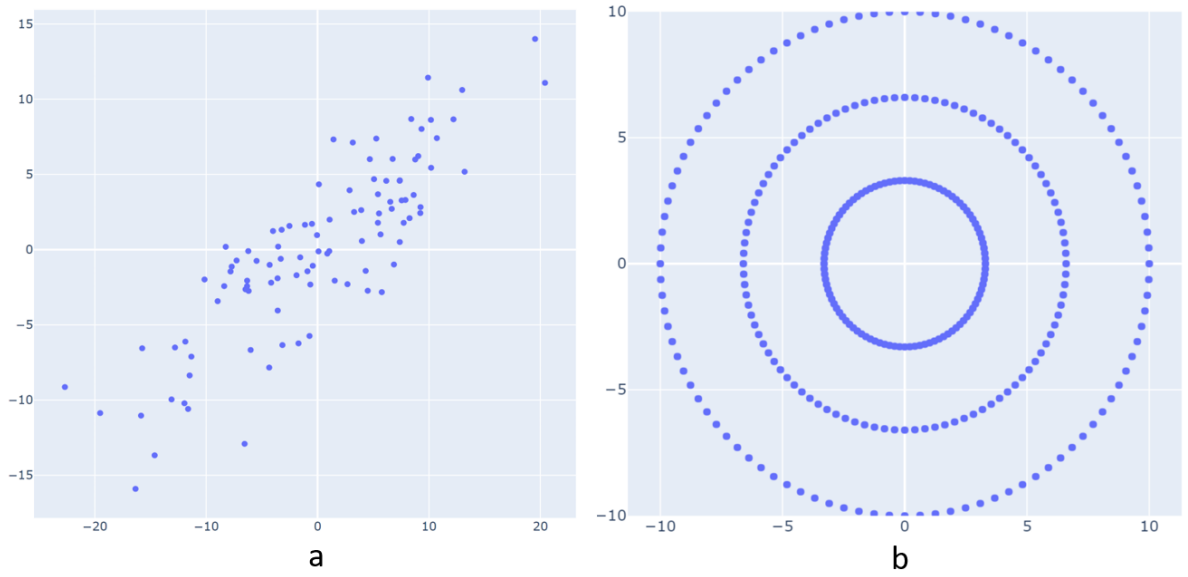
- Durch die Zeitsynchronisation kann der Datensatz inzwischen als n x m Matrix dargestellt werden, wobei n die jeweiligen Instanzen sind und m die Merkmale
- Der bereinigte und aufbereitete Datensatz kann so als Punktwolke in einem m-dimensionalen Raum dargestellt werden (jeder Punkt ist eine Instanz)
- Ziel der PCA ist es, die Datenpunkte so in einen q-dimensionalen Unterraum zu projizieren, dass dabei möglichst wenig Informationen verloren gehen (→ Reduktion der Merkmale ohne Informationsverlust)
- Mathematisch: Hauptachsentransformation
  - Minimierung der Korrelation mehrdimensionaler Merkmale durch Überführung in Vektorraum neuer Basis
  - Orthogonale Matrix, bestehend aus Eigenvektoren der Kovarianzmatrix
- **Resultat:**
  - Größter Anteil der Gesamtstreuung ist in der 1. PC, zweitgrößter Anteil in der 2. PC, usw.
  - In der Praxis werden dann nur die q PCs verwendet, die z.B. 90% der Gesamtvarianz ausmachen.
  - Nachteil: original Merkmale sind nur noch kombiniert repräsentiert → Darstellung nur schwer zu interpretieren



Datenaufbereitung – PCA

Zwischenübung

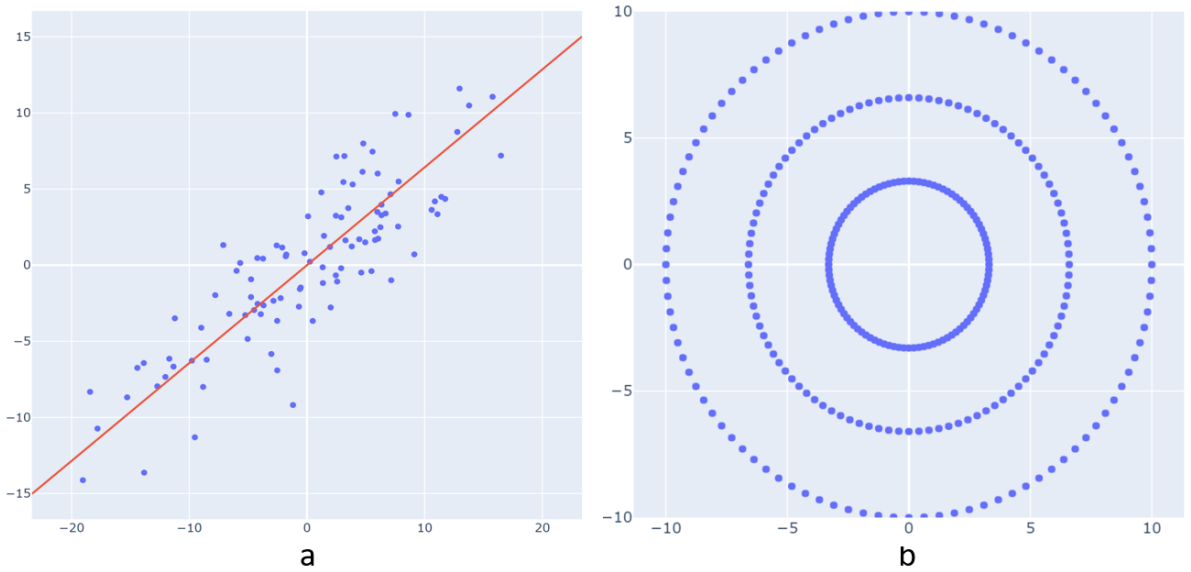
- Gegeben seien die Datensätze in der folgenden Abbildung, zeichnen Sie falls möglich, ohne exakte Berechnungen die erste Hauptkomponente nach der PCA-Methode. Begründen Sie jede Wahl!



Datenaufbereitung – PCA

Zwischenübung - Lsg

- Gegeben seien die Datensätze in der folgenden Abbildung, zeichnen Sie falls möglich, ohne exakte Berechnungen die erste Hauptkomponente nach der PCA-Methode. Begründen Sie jede Wahl!

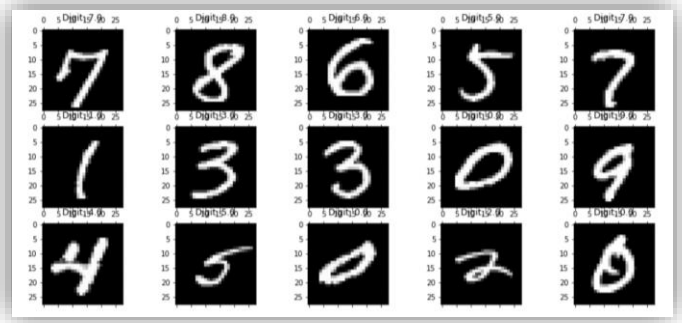


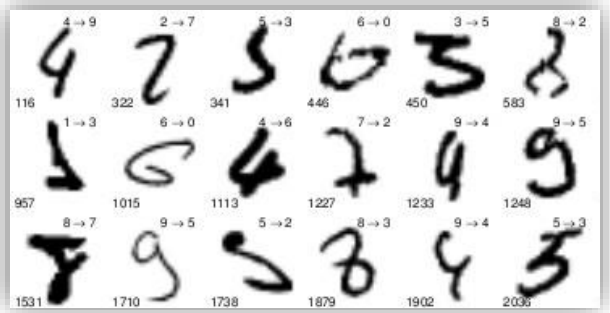
- Datensatz (a): Hauptkomponente bei der größten treuung/Standardabweichung.
- Datensatz (b): Da es kreisförmig ist, kann die Hauptkomponente in jede Richtung sein.

Datenaufbereitung

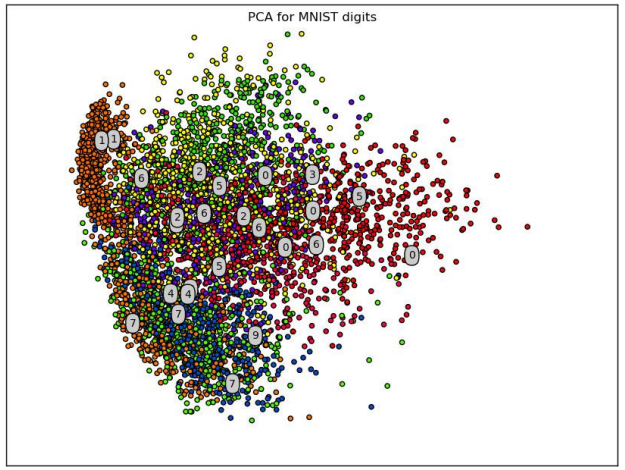
Merkmalsreduktion bzw. Visualisierung – weitere Methoden am Bsp. MNIST

- MNIST ist ein bekannte öffentlicher Datensatz
  - Er beinhaltet 70.000 handgeschriebene Ziffern, die jeweils in 28 x 28 Pixeln dargestellt sind.
    → 70.000 x 784 Matrix (= 54.880.000 Werten, 70k Instanzen und 784 Merkmalen)



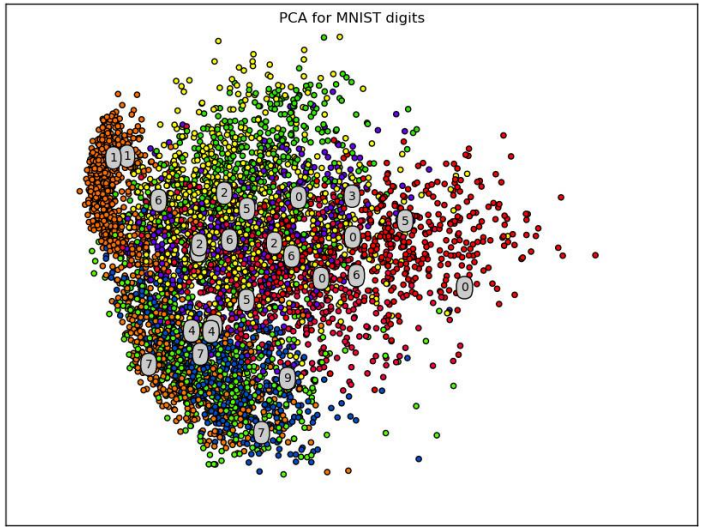


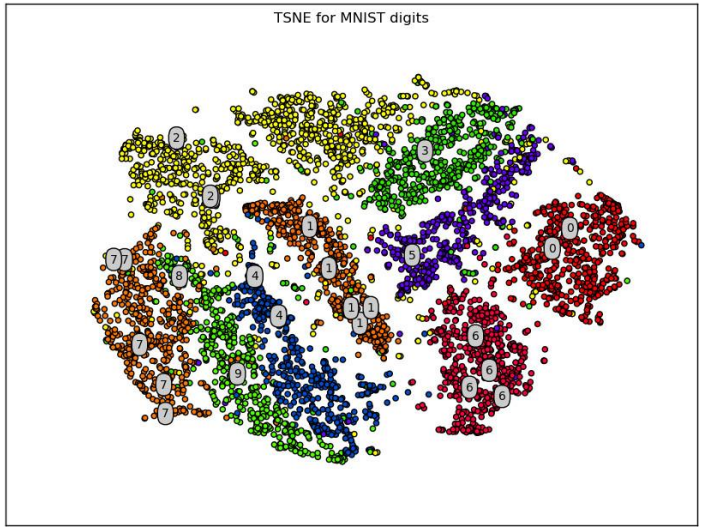
  - PCA-Darstellung in 2D (ersten beiden Hauptkomponenten):

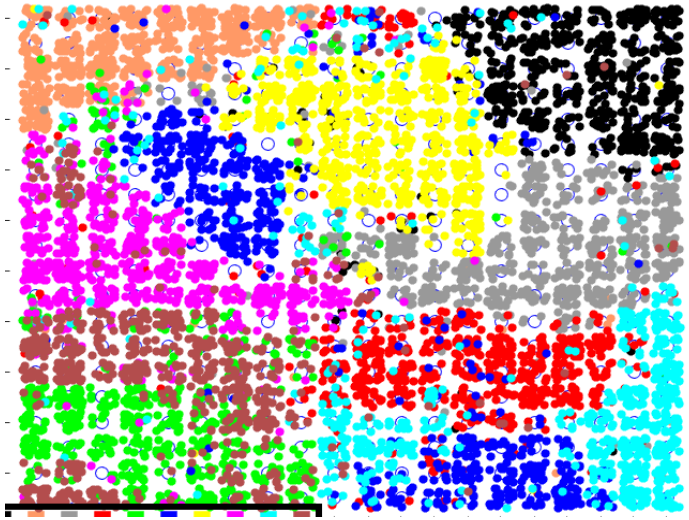


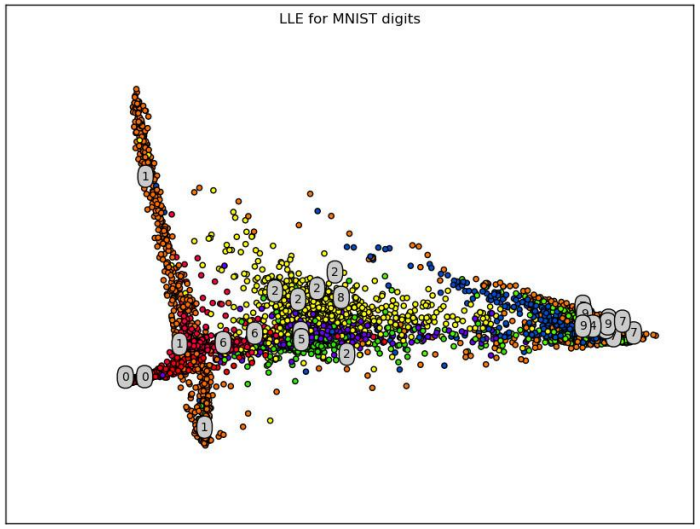
Datenaufbereitung

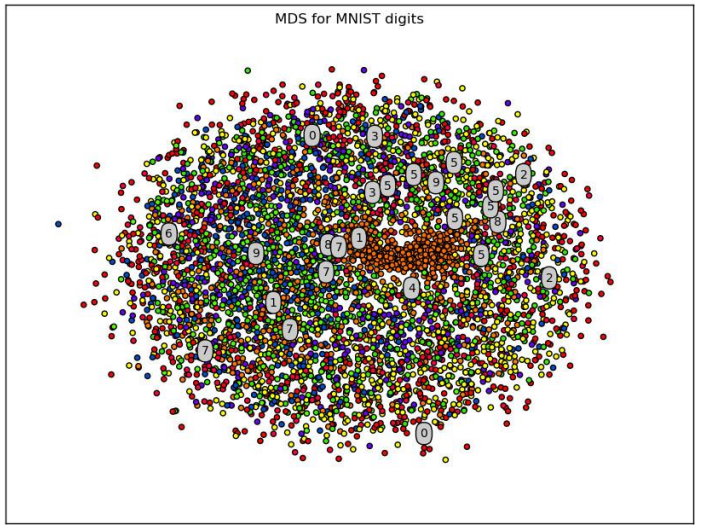
Merkmalsreduktion – weitere Methoden am Bsp. MNIST

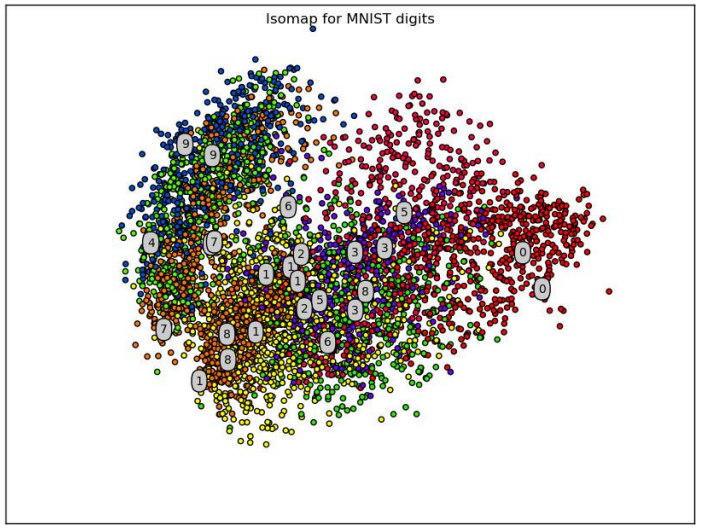












- Datenmanipulation
  - Umgang mit Ausreißern
  - Konvertierung
  - Fehlerhafte Werte
  - Qualitätsverbesserung
  - Merkmalsredukation

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