

Übung 04

Übung zu Informationstechnik II und Automatisierungstechnik – Nathalie Brenner

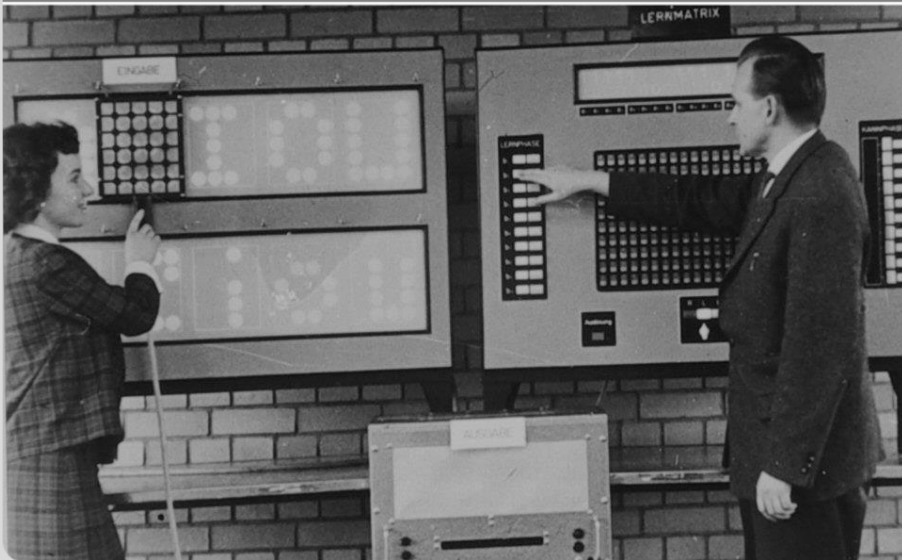
Prof. Dr.-Ing. Eric Sax

Institutsleitung

Prof. Dr.-Ing. J. Becker

Prof. Dr.-Ing. E. Sax

Prof. Dr. rer. nat. W. Stork



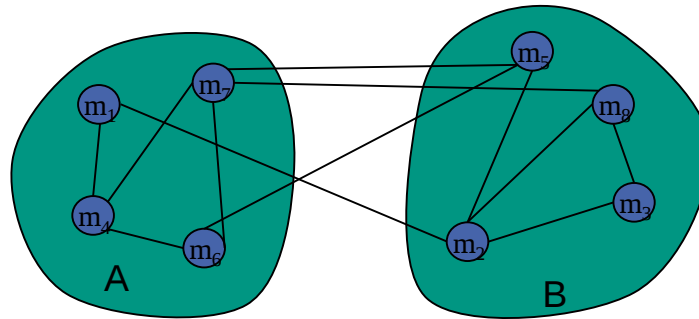
WIEDERHOLUNG ÜBUNG 3



Wiederholung Übung 3

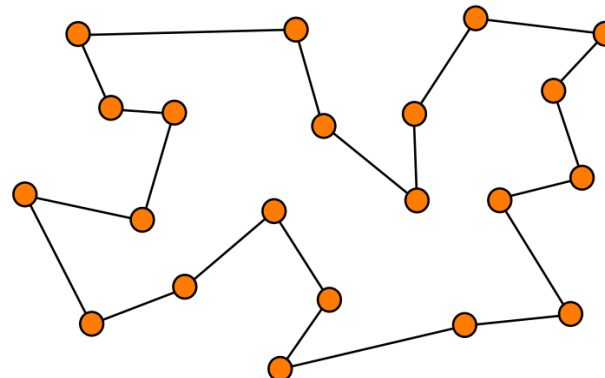
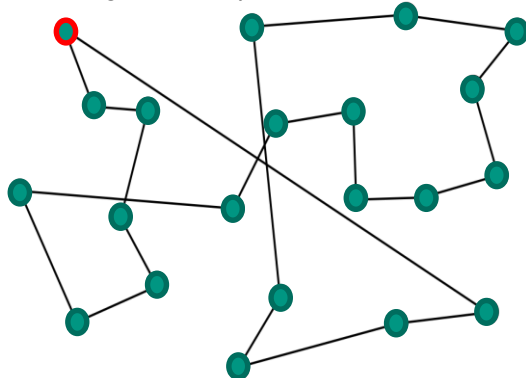
Optimierungsalgorithmen

Partitionierung und Kostenreduktion

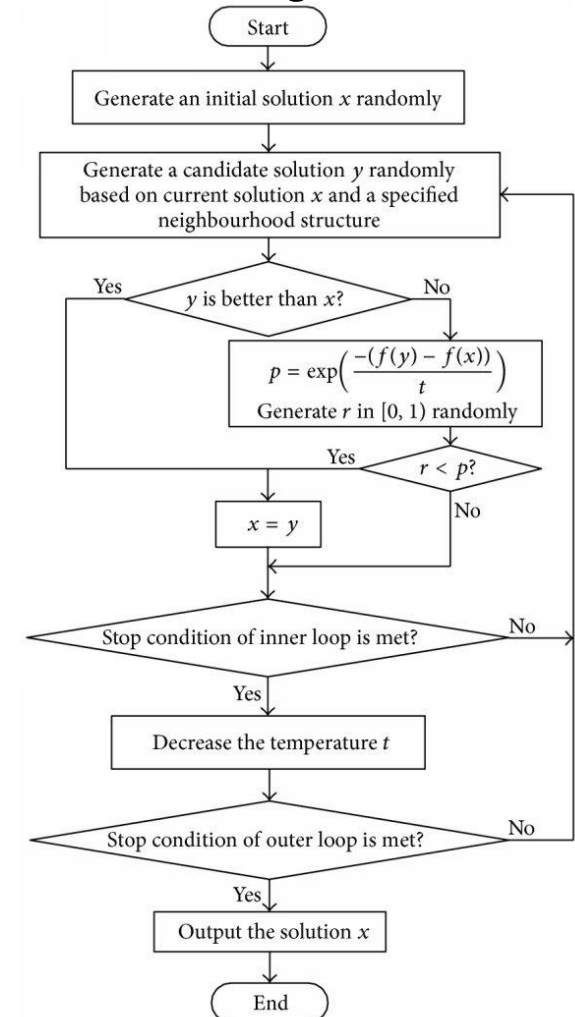


Greedy:

- Mit Greedy-Algorithmen ist eine Familie bzw. Klasse an Algorithmen gemeint, welche dem Grundprinzip folgen in jedem Verfahrensschritt diejenige Entscheidung zu treffen, die in diesem Moment am besten ist (ohne Berücksichtigung zukünftiger Schritte)



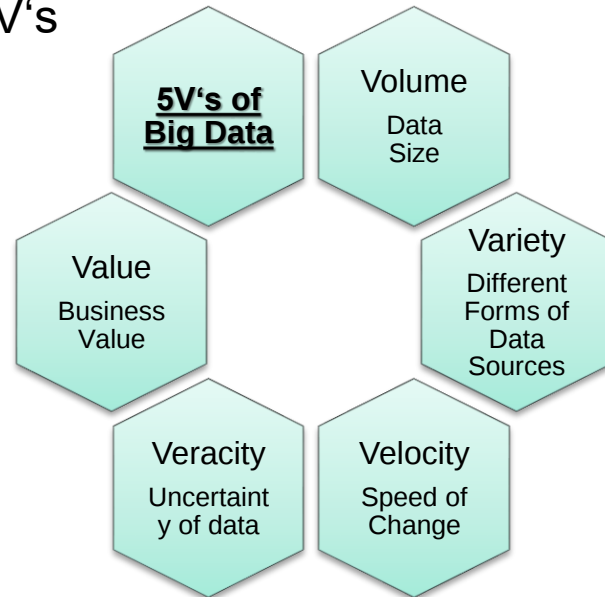
Simulated Annealing



Wiederholung Übung 3

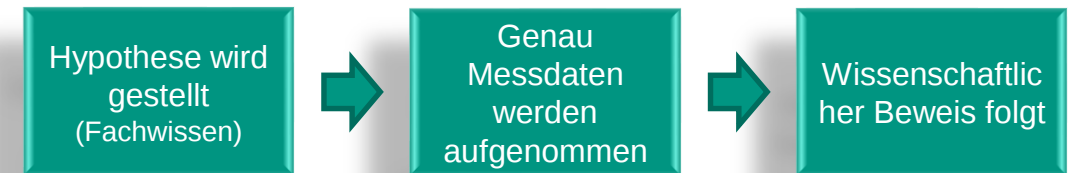
Charakteristika zur Analyse großer Datenbestände → Big Data

- Die drei grundlegenden V's
Die 5V's

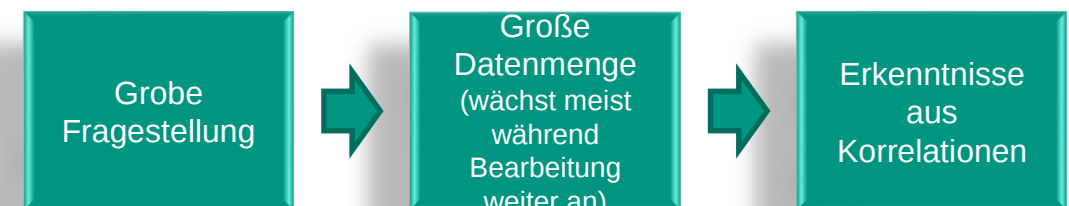


- Volume:**
Daten die aufgrund ihrer Menge bisher als kaum speicherbar, geschweige denn als auswertbar galten
- Variety:**
aus verschiedene Datenquellen strömen sortierte und unsortierte Datenmengen durch die Netze
- Velocity:**
Ergebnisse sollen möglichst schnell zur Verfügung stehen
- Veracity:**
Anspruch an hohe Datenqualität und Verlässlichkeit der Daten
- Value:**
Verwertbarkeit der gewonnen Erkenntnisse

- Hypothesengestützte Erkenntnisgewinnung
Beruht auf kausalem Ansatz



- Schlussfolgerung durch grobe Fragestellung
Beruht auf Auffinden von Korrelationen

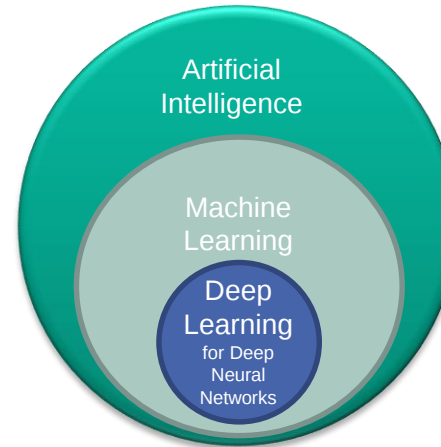


Wiederholung Übung 3

Data Science, Data Mining, Maschinelles Lernen und Trainieren

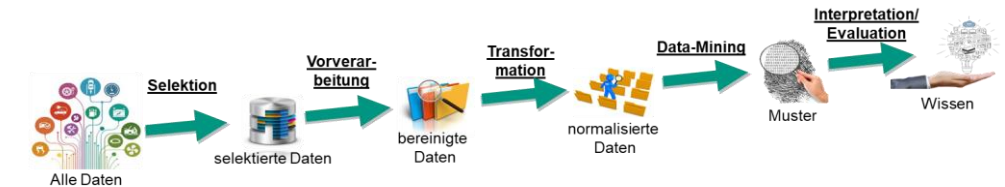
Grundbegriffe

- AI, ML, NNs:
Untergruppe von Data Mining
- Data Mining:
Untergruppe von Data Science
- Data Science:
Untergruppe von Big Data

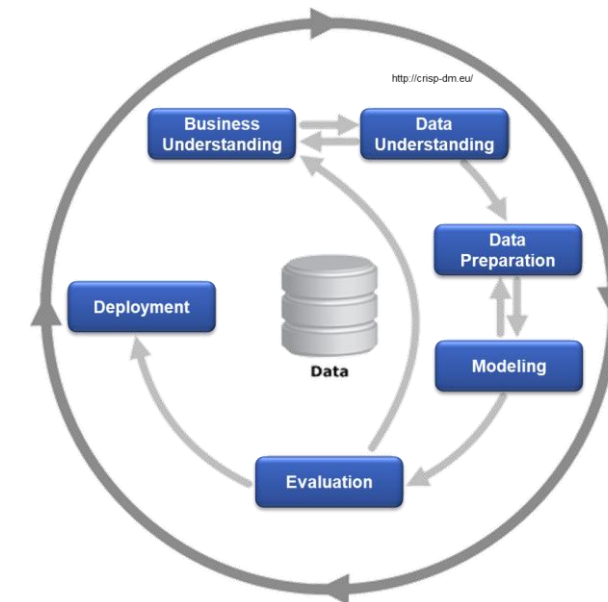


Big Data als Prozess

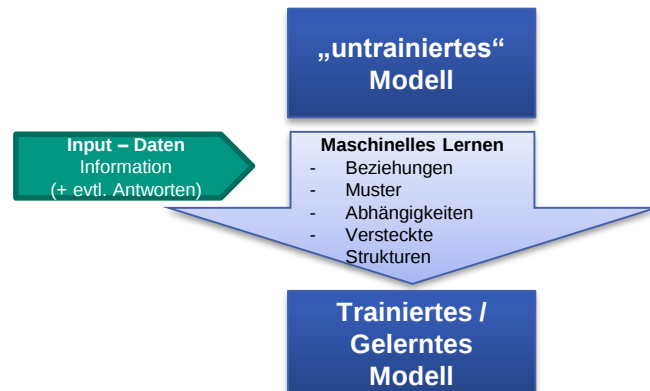
KDD (Knowledge Discovery in Databases)



CRISP-DM (Cross Industry Standard Process for Data Mining)



Modelle im Bezug auf Maschinelles Lernen



- Erstellung und Verwendung von Modellen, welche aus Daten gelernt haben
- Modelle nutzen um Ergebnisse für neue Daten voraussagen zu können

INHALT ÜBUNG 4



Ziele der heutigen Übung



■ Nach der heutigen Übung können Sie....

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

1

- ... Methoden zur Geschäftszielfindung beschreiben.

2

- ... Datentypen aufzählen und voneinander abgrenzen.

3

- ... „einfache“ Statistikwerte nennen und berechnen.

4

- ... Daten visualisieren.

CRISP-DM

BUSINESS UNDERSTANDING

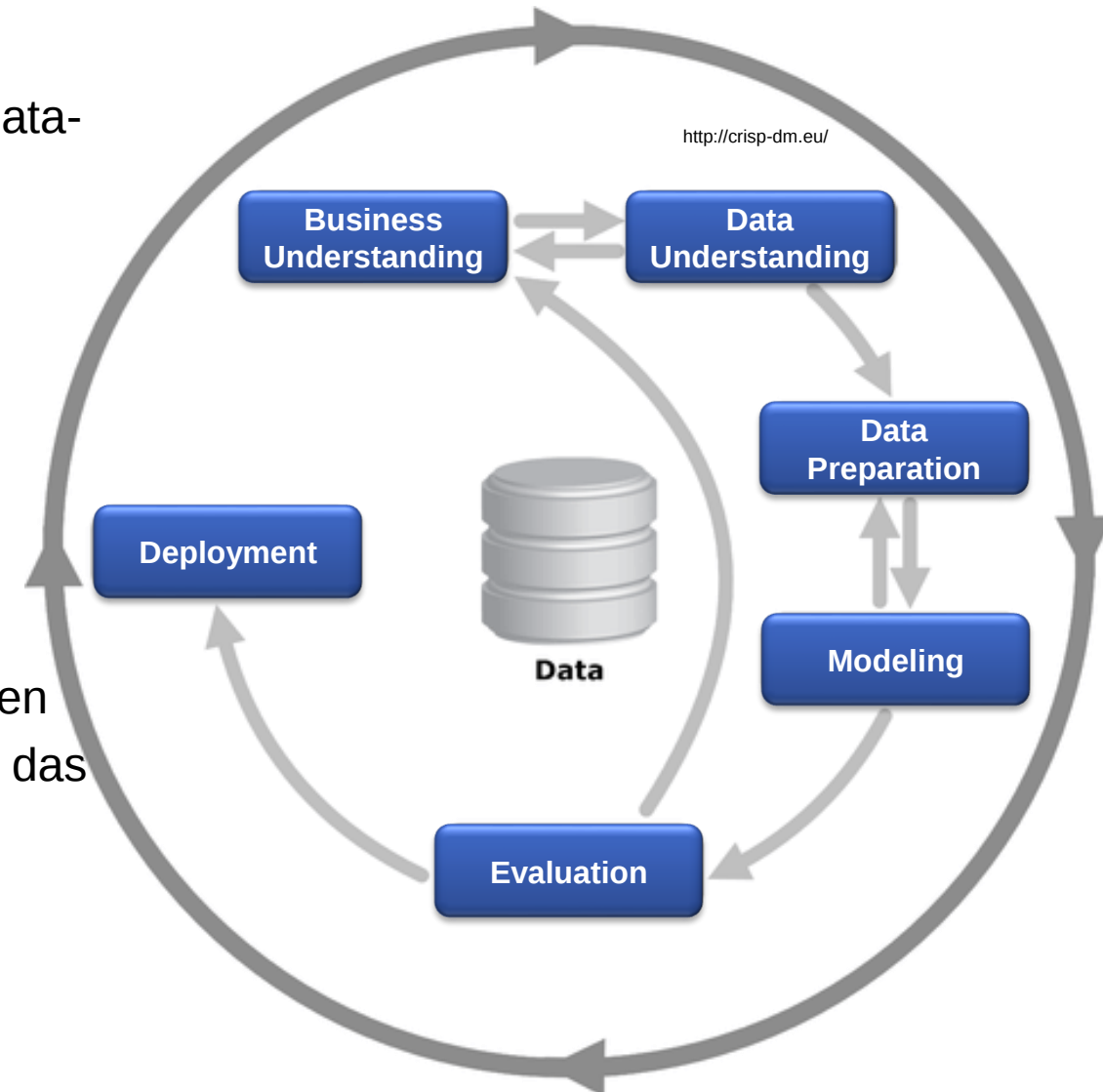
DATA UNDERSTANDING



Cross Industry Standard Process for Data Mining (CRISP-DM)

Business- and Data Understanding

- Aufgabenverständnis: „Formulierung der Aufgabe“
 - Aufteilung der Aufgabenstellung auf verschiedene Data-Science-Aufgaben
 - Kenntnisse von Data Mining hilfreich
 - Erkenntnisse von denkbaren Anwendungsfeldern
- Datenverständnis
 - Daten sind selten genau auf die Aufgabenstellung zugeschnitten (Datensammlung meist generell oder ursprünglich zu anderem Zweck)
 - Kosten-Nutzen bei Beschaffung neuer Daten beachten
 - Durch Zunahme des Datenverständnisses kann sich das Aufgabenverständnis wiederum ändern



Business Understanding

Von der geschäftlichen Aufgabe zum Data-Mining-Verfahren

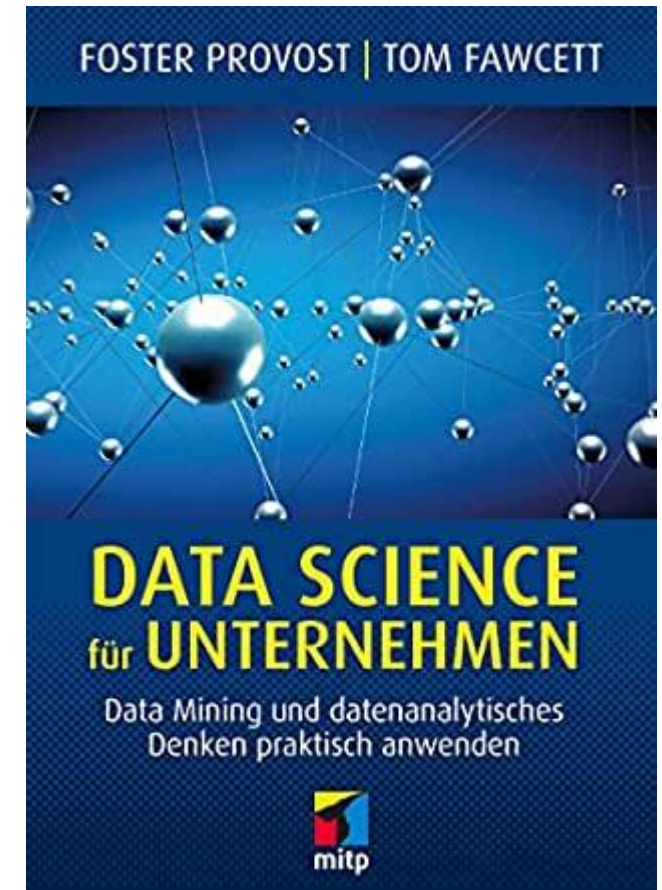
- Konzentration auf das Verständnis des Projekts
- Ziele und Anforderungen aus betriebswirtschaftlicher Sicht und die anschließende Umsetzung dieses Wissens in eine Data-Mining-Problemdefinition
- Aufstellung eines vorläufigen Projektplans

- Unterstützende Fragestellungen:

- Was genau ist das Ziel?
- Wie wollen wir es erreichen?
- Lohnt es sich zu investieren?
(Kosten-Nutzen Faktor des Projekts)
- Welche Teile des Anwendungsszenarios sind für Data-Mining Modelle geeignet?

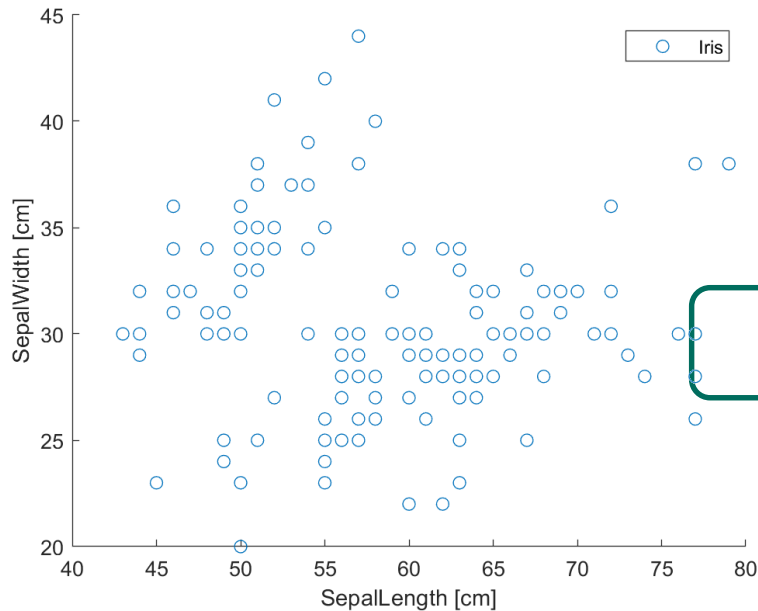


- Data Science für Unternehmen
Data Mining und datenanalytisches Denken praktisch anwenden
 - Die grundlegenden Konzepte der Data Science verstehen, Wissen aus Daten ziehen und für Vorhersagen und Entscheidungen nutzen
 - Die wichtigsten Data-Mining-Verfahren gezielt und gewinnbringend einsetzen
 - Zahlreiche Praxisbeispiele zur Veranschaulichung



Business Understanding

Anwendung auf den „Iris Datensatz“



Datenverständnis!

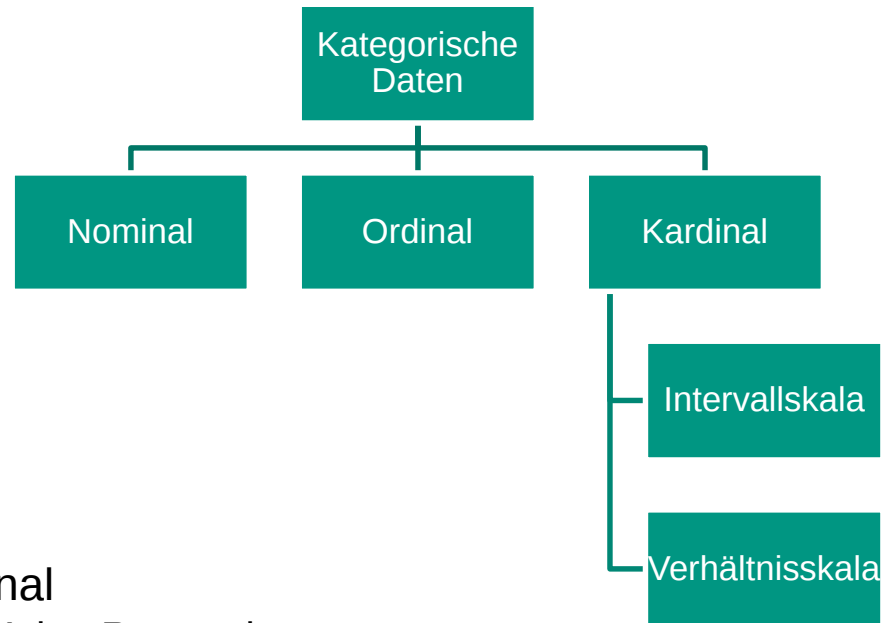
■ Unterstützende Fragestellungen:

- Was genau ist das Ziel?
- Wie wollen wir es erreichen?
- Lohnt es sich zu investieren?
(Kosten-Nutzen Faktor des Projekts)
- Welche Teile des Anwendungsszenarios sind für Data-Mining Modelle geeignet?

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
16	70	32	47	14	Iris-versicolor
17	64	32	45	15	Iris-versicolor
18	69	31	49	15	Iris-versicolor
19	55	23	40	13	Iris-versicolor
20	65	28	46	15	Iris-versicolor
21	57	28	45	13	Iris-versicolor
22	63	33	47	16	Iris-versicolor
23	49	24	33	10	Iris-versicolor
24	66	29	46	13	Iris-versicolor
25	52	27	39	14	Iris-versicolor
26	50	20	35	10	Iris-versicolor
27	59	30	42	15	Iris-versicolor
28	60	22	40	10	Iris-versicolor
29	61	29	47	14	Iris-versicolor
30	56	29	36	13	Iris-versicolor

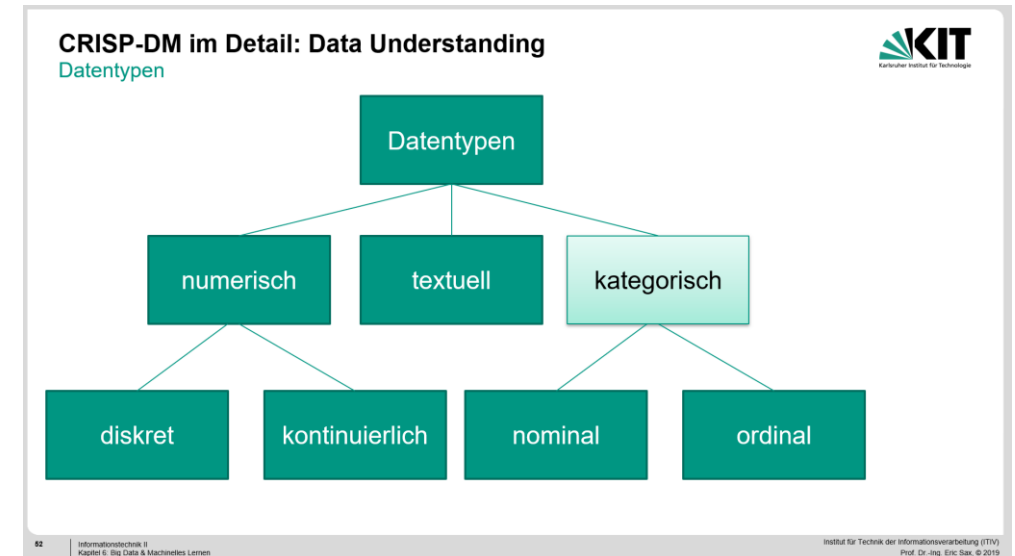
Data Understanding

Datentypen – Kategorische Daten



- **Nominal**
 - Keine Rangordnung
 - *Geschlecht, Studiengang*
- **Ordinal**
 - Rangordnung
 - Keine interpretierbaren Abstände
 - *Schulnoten, Steuerklassen*
- **Kardinal/metrische Skala**
 - Rangordnung
 - Interpretierbare Entfernungen
 - *Preise, Abstände*

Wiederholung Vorlesung:



Data Understanding

Merkmale/Features – Titanic Datensatz

■ 11 Features (Name, Alter, Geschlecht, Ticketklasse, Überlebt,...)

Feature, Merkmale, Attribute

1. Survival
2. Pclass
3. Name
4. Sex
5. Age
6. SibSp
7. Parch
8. Ticket
9. Fare
10. Cabin
11. Embarked

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2.	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male				3734		NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1		

Ordinal

Ordinal

Label

Nominal

Nominal

Kardinal

Kardinal

Fehlende Werte

Data Understanding

Merkmale/Features – Iris Datensatz

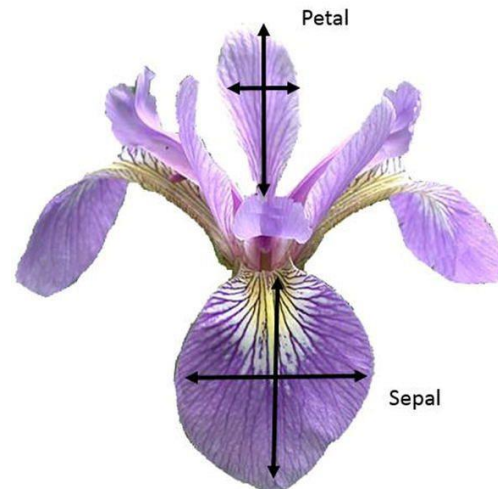
■ Gelabelte Daten:

- Iris Setosa
- Iris Versicolor
- Iris Virginica



■ Jeweils 4 Merkmale in [cm]

- SepalLength
- SepalWidth
- PetalLength
- PetalWidth

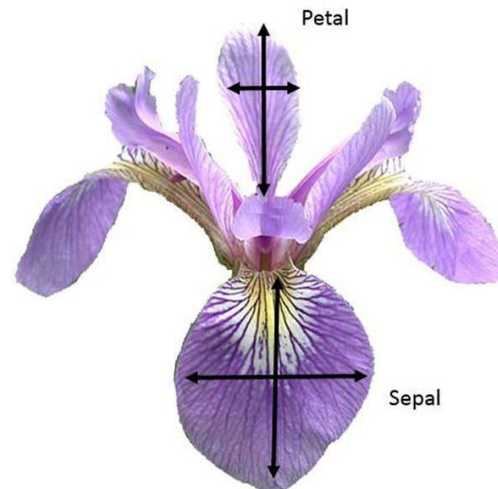


Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	Species
1	51	35	14		2 Iris-setosa
2	49	30	14		2 Iris-setosa
3	47	32	13		2 Iris-setosa
4	46	31	15		2 Iris-setosa
5	50	36	14		2 Iris-setosa
6	54	39	17		4 Iris-setosa
7	46	34	14		3 Iris-setosa
8	50	34	15		2 Iris-setosa
9	44	29	14		2 Iris-setosa
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

Data Understanding

Merkmale/Features – Iris Datensatz

- Nominal
 - Keine Rangordnung
 - *Geschlecht, Studiengang*
- Ordinal
 - Randordnung
 - Keine interpretierbaren Abstände
 - *Schulnoten, Steuerklassen*
- Kardinal/metrische Skala
 - Randordnung
 - Interpretierbare Abstände
 - *Preise, Abstände*



Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	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

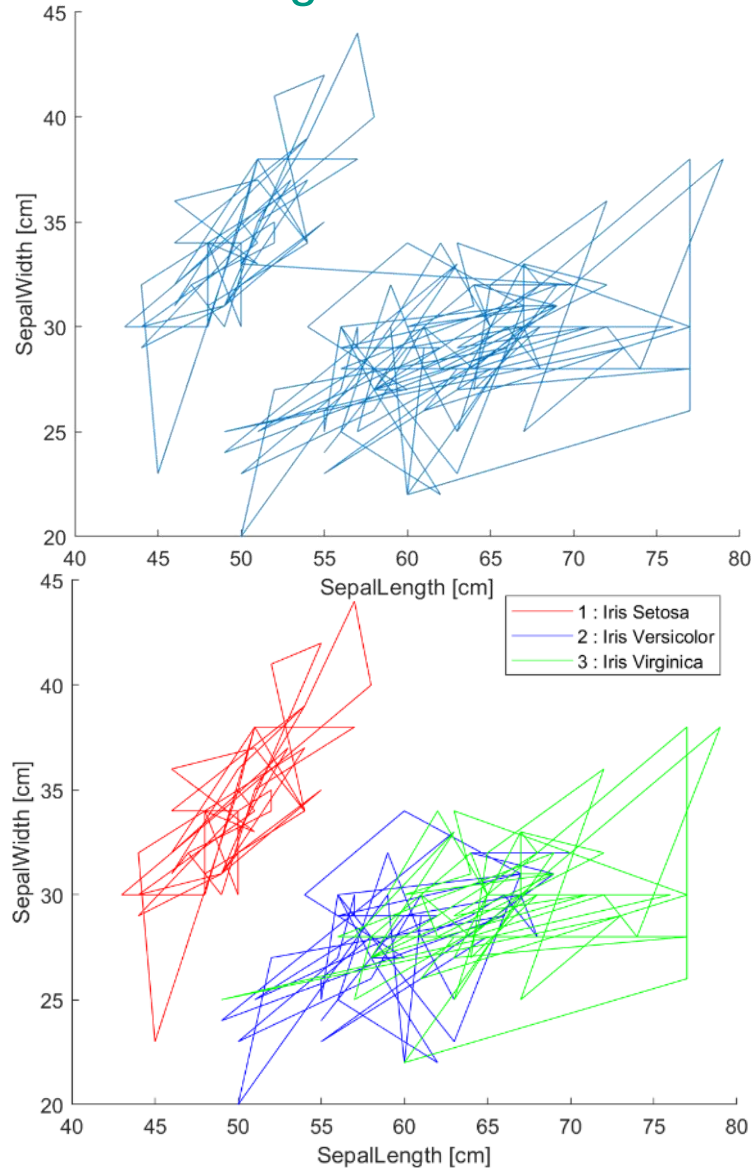
- Datenvisualisierung
 - Daten erkunden
 - Daten/Ergebnisse übermitteln
- Line Chart: gut geeignet um „Trends“ zu zeigen
- Scatter Charts: Beziehungen zwischen Datenpunkten gut erkennbar
- Bar Charts/Histogramme: Beziehung/Verhalten von Datenmengen zu einander
- Boxplots: Minimum, Maximum, Quantile, Median
Häufigkeitsverteilung der Daten innerhalb der Mengen

ABER: Nicht jede Möglichkeit zur grafischen Visualisierung ist für jede Problemstellung geeignet

	Line Chart	Scatter Chart	Histogramm	Boxplot
Iris Datensatz	?	?	?	?

Data Understanding

Visualisierung - Linechart



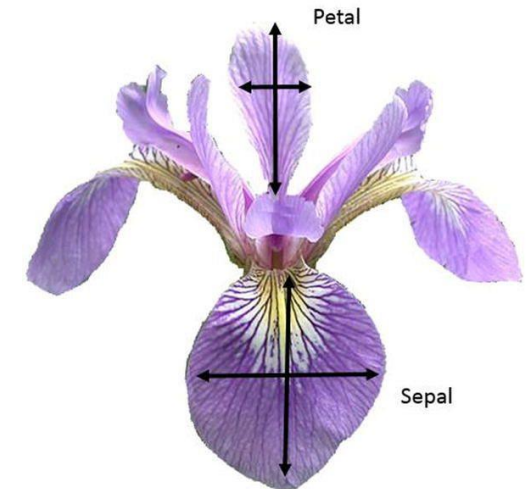
■ Gelabelte Daten:

- Iris Setosa
- Iris Versicolor
- Iris Virginica



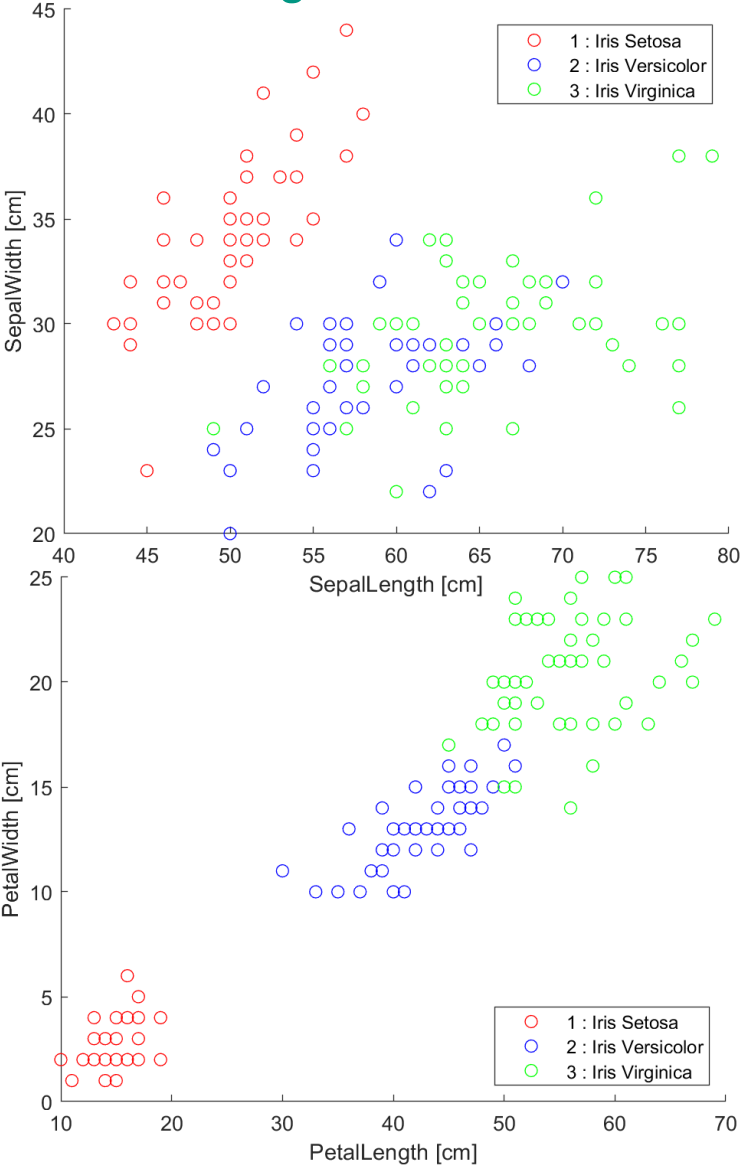
■ Jeweils 4 Merkmale in [cm]

- SepalLength
- SepalWidth
- PetalLength
- PetalWidth

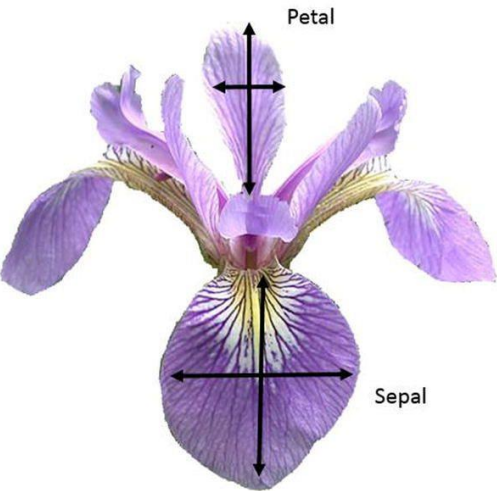


Data Understanding

Visualisierung - Scatterchart

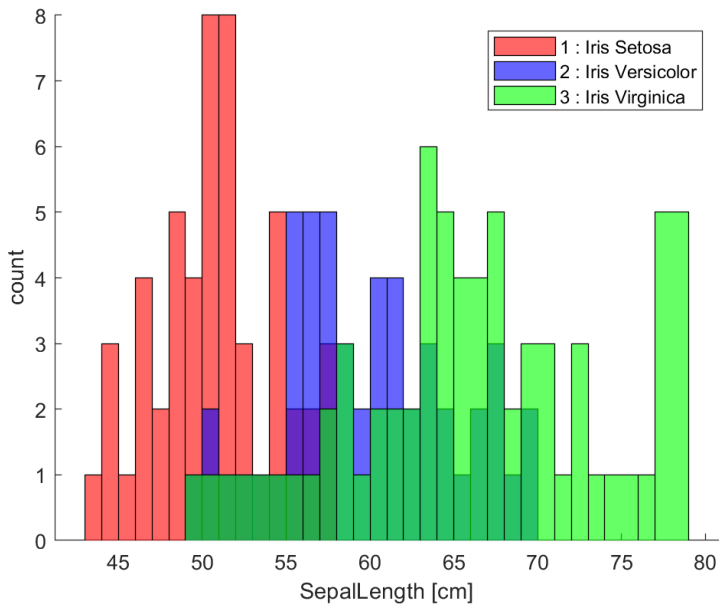


Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	Species
1	51	35	14	2	Iris-setosa
2	49	30	14	2	Iris-setosa
3	47	32	13	2	Iris-setosa
4	46	31	15	2	Iris-setosa
5	50	36	14	2	Iris-setosa
6	54	39	17	4	Iris-setosa
7	46	34	14	3	Iris-setosa
8	50	34	15	2	Iris-setosa
9	44	29	14	2	Iris-setosa
10	49	31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

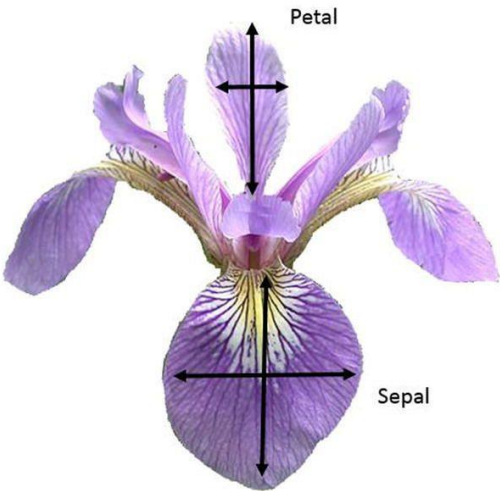


Data Understanding

Visualisierung - Histogramm



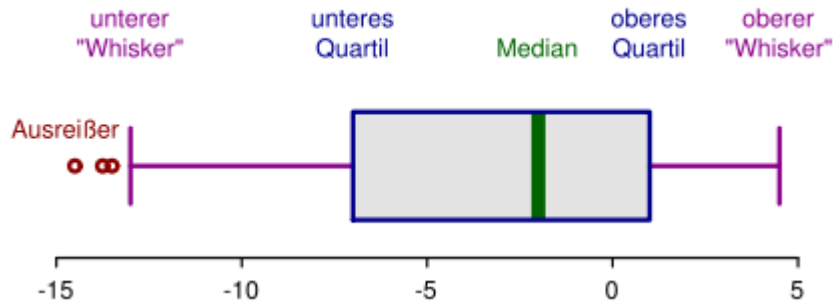
Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	Species
1	51	35	14	2	Iris-setosa
2	49	30	14	2	Iris-setosa
3	47	32	13	2	Iris-setosa
4	46	31	15	2	Iris-setosa
5	50	36	14	2	Iris-setosa
6	54	39	17	4	Iris-setosa
7	46	34	14	3	Iris-setosa
8	50	34	15	2	Iris-setosa
9	44	29	14	2	Iris-setosa
10	49	31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



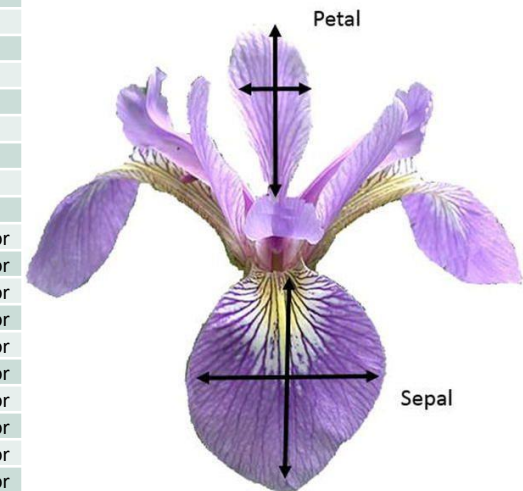
Data Understanding

Visualisierung - Boxplot

- Grafische Darstellung der Verteilung eines Merkmals
 - Wo liegt das Minimum
 - Wo liegt das Maximum
 - In welchem Bereich befinden sich die Daten
 - Wie verhält sich die Häufigkeitsverteilung der Datenpunkte?
 - „Box“ gibt 50% aller Daten an



Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	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



Data Understanding

Visualisierung

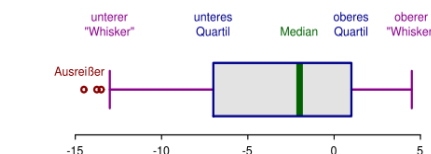
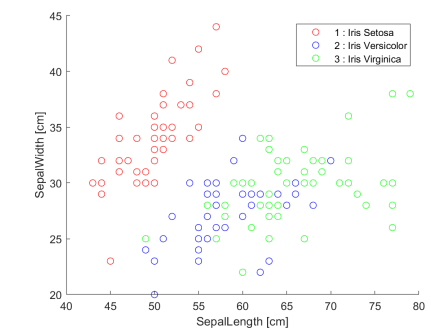
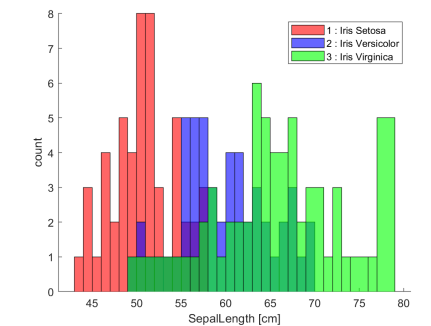
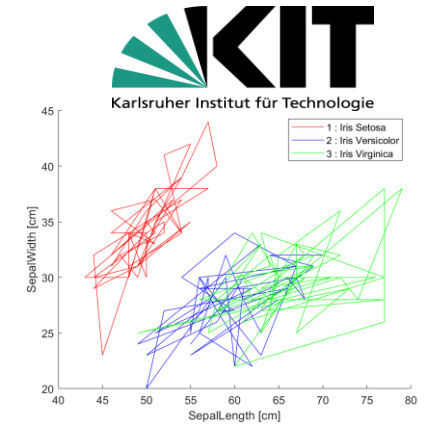
■ Datenvisualisierung

- Daten erkunden
- Daten/Ergebnisse übermitteln

- Line Chart: gut geeignet um „Trends“ zu zeigen
- Scatter Charts: Beziehungen zwischen Datenpunkten gut erkennbar
- Bar Charts/Histogramme: Beziehung/Verhalten von Datenmengen zu einander
- Boxplots: Minimum, Maximum, Quantile, Median
Häufigkeitsverteilung der Daten innerhalb der Mengen

ABER: Nicht jede Möglichkeit zur grafischen Visualisierung ist für jede Problemstellung geeignet

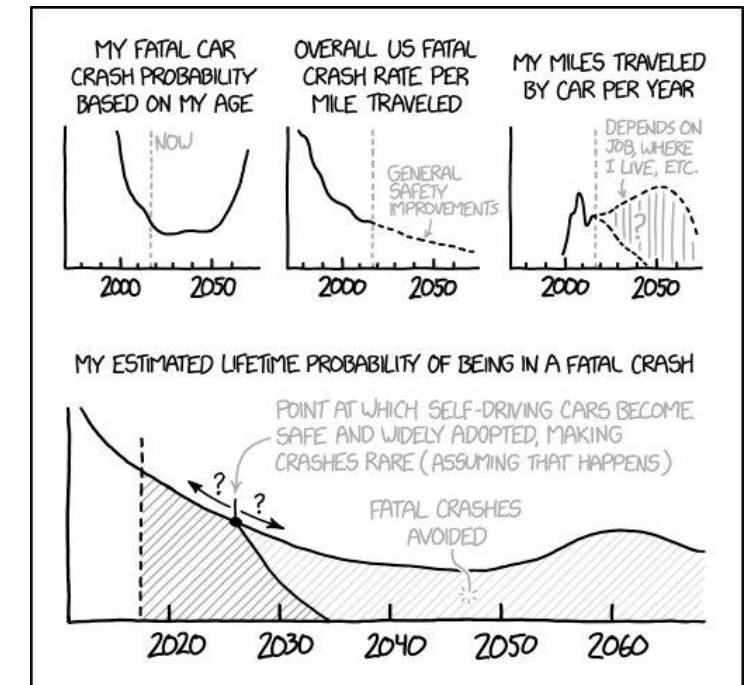
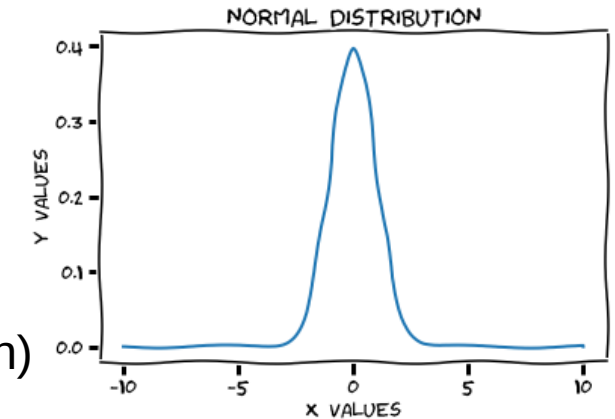
	Line Chart	Scatter Chart	Histogramm	Boxplot
Iris Datensatz	✗	✓	✓	✓



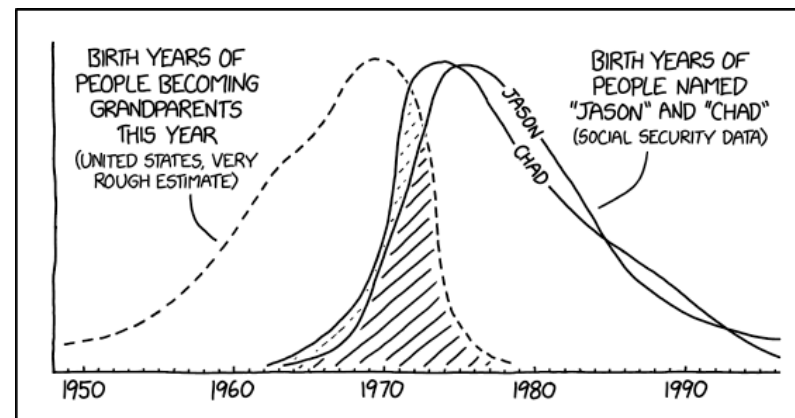
- Business Understanding
- Data Understanding - Visualisierung



- „Statistik“ hat in Datenanalyse zwei Bedeutungen
 - Zusammenfassende Statistik
Berechnung bestimmter numerischer Werte
 - Statistik (nach gleichnamigen Fachgebiet)
analytische Verfahren (statistische Verfahren um Hypothesen zu testen)
- IT2: Möglichkeit Daten zu beschreiben und zu visualisieren



„Facts are stubborn, but statistics are more pliable“
~ Marc Twain



FUN FACT: WE HAVE NOW ENTERED THE ERA OF "GRANDPA JASON" AND "GRANDPA CHAD."

IT FEELS WEIRD TO LOOK AT CAR CRASH STATISTICS AND WONDER WHETHER WE'LL ALL BE ABLE TO STOP DRIVING BEFORE I'M INVOLVED IN A FATAL CRASH.

Data Understanding

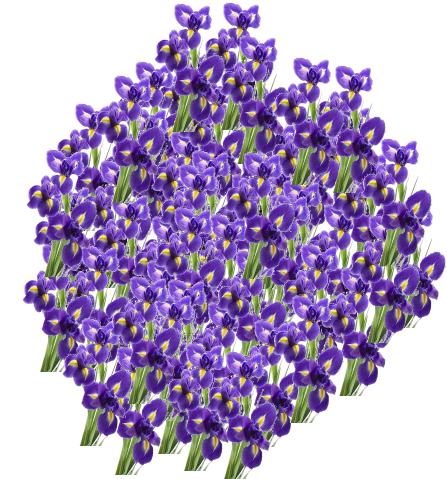
Statistik – Daten beschreiben I

■ Daten selbst

Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	Species
4	46	31	15		2 Iris-setosa
6	54	39	17		4 Iris-setosa
13	48	30	14		1 Iris-setosa



■ Was bei größeren Datensätzen? Hier „beginnt die Statistik“



- Anzahl
- Größter/ Kleinster Wert
- Sortieren nach?
Bubble~/ Insertion~/ Merge~/ Quick Sort
- Häufigster Wert

```
Anzahl = len( data[2])  
Min_value = min( data[2])  
Max_value = max( data[2])
```

```
count = Counter(data[2])  
max_count = max(count.values())
```

	Iris Setosa	
	SepalLength	SepalWidth
Anzahl	50	50
Minimum	43	23
Maximum	58	44
Sortieren	---	---
Häufigster Wert	50	34

■ Statistik:

■ Mittelwert

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

■ Median

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

■ Quantile

$$x_p = \begin{cases} \frac{1}{2} (x_{(np)} + x_{(np+1)}) & np \text{ ganzzahlig} \\ x_{(np+1)} & np \text{ nicht ganzzahlig} \end{cases}$$

■ Varianz

$$\sigma^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$$

■ Standard- abweichung

$$\sigma = \sqrt{\sigma^2}$$

Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	Species
1	51	35	14		2 Iris-setosa
2	49	30	14		2 Iris-setosa
3	47	32	13		2 Iris-setosa
7	46	34	14		3 Iris-setosa
8	50	34	15		2 Iris-setosa
9	44	29	14		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
59	66	29	46		13 Iris-versicolor
...	...	...	...	...	...
110	72	36	61		25 Iris-Virginica
111	65	32	51		20 Iris-Virginica
...	...	...	...	...	...

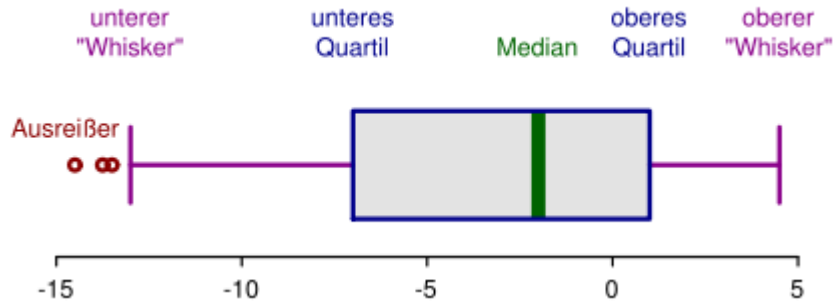
	Iris Setosa		Iris Versicolor		Iris Virginica	
	SepalLength	SepalWidth	SepalLength	SepalWidth	SepalLength	SepalWidth
Mittelwert	50,06	34,18	59,36	27,70	65,88	29,74
Median	50	34	59	28	65	30
25% Quantil	48	31	56	25	62	28
75% Quantil	52	37	63	30	69	31,75
Varianz	12,42	14,52	26,64	9,85	40,43	10,40
Standard- abweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

Data Understanding - Boxplot

Zwischenübung



- Zeichnen Sie ein detaillierter Box-Whisker-Plot, anhand dessen Ausreißer detektiert werden können. Berechnen sie gegeben falls fehlende, aber benötigte Werte.
 - Iris Setosa - SepalLength



	Iris Setosa	
	SepalLength	SepalWidth
Anzahl	50	50
Minimum	43	23
Maximum	58	44
Sortieren	---	---
Häufigster Wert	50	34

	Iris Setosa		Iris Versicolor		Iris Virginica	
	SepalLength	SepalWidth	SepalLength	SepalWidth	SepalLength	SepalWidth
Mittelwert	50,06	34,18	59,36	27,70	65,88	29,74
Median	50	34	59	28	65	30
25% Quantil	48	31	56	25	62	28
75% Quantil	52	37	63	30	69	31,75
Varianz	12,42	14,52	26,64	9,85	40,43	10,40
Standard-abweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

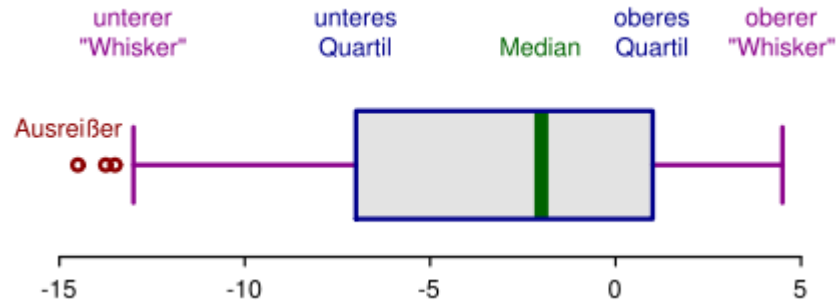
Data Understanding - Boxplot

Zwischenübung - Lsg

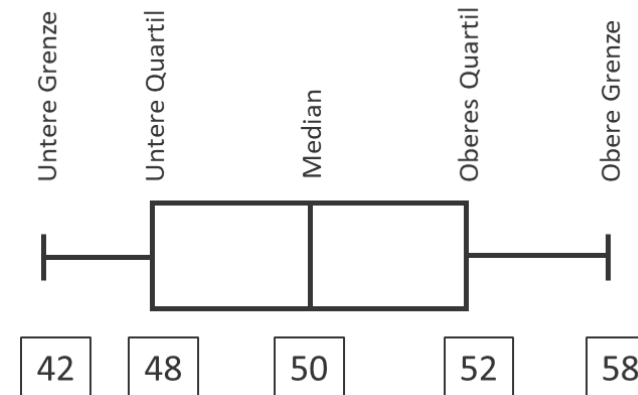
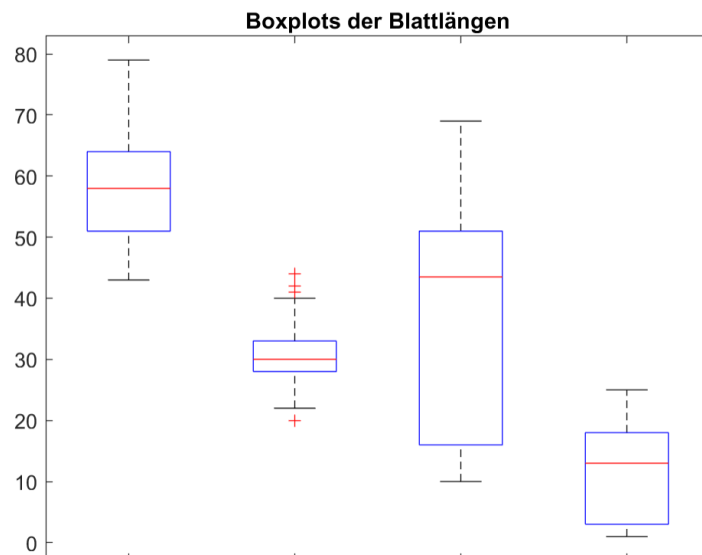


■ Zeichnen Sie (grob) ein Boxplot Diagramm des Iris Datensatzes zu:

- SepalWidth
- SepalLength
- PetalWidth
- PetalLength



Weitere Plots (mit und ohne Ausreißerererkennung):

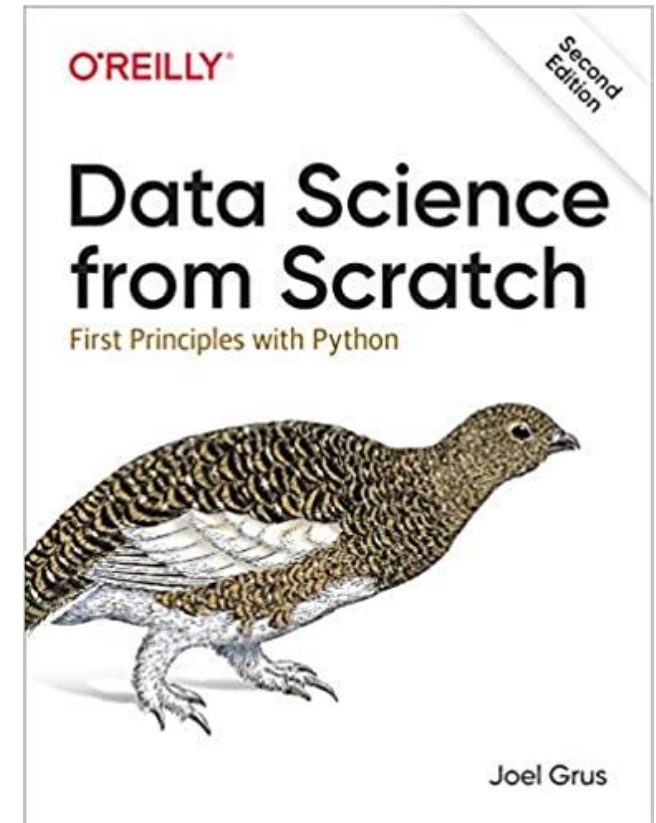


	Iris Setosa	
	SepalLength	SepalWidth
Mittelwert	50,06	34,18
Median	50	34
25% Quantil	48	31
75% Quantil	52	37
Minimum	43	23
Maximum	58	44
IQA	4	6
Obere Grenze	58	52
Untere Grenze	42	22

Data Understanding

Weiterführende Literatur

- Data Science from Scratch
 - First Principles with Python
- Grundlagen der linearen Algebra, der Statistik und der Wahrscheinlichkeitsrechnung - und wie und wann sie in der Datenwissenschaft verwendet werden (*Übung 4*)
- Daten sammeln, erforschen, bereinigen, vergraben und manipulieren (*Übung 5*)
- Implementierung von Modellen (*Übung 6 und 7*)



- Data Understanding
 - Visualisierung
 - Statistik



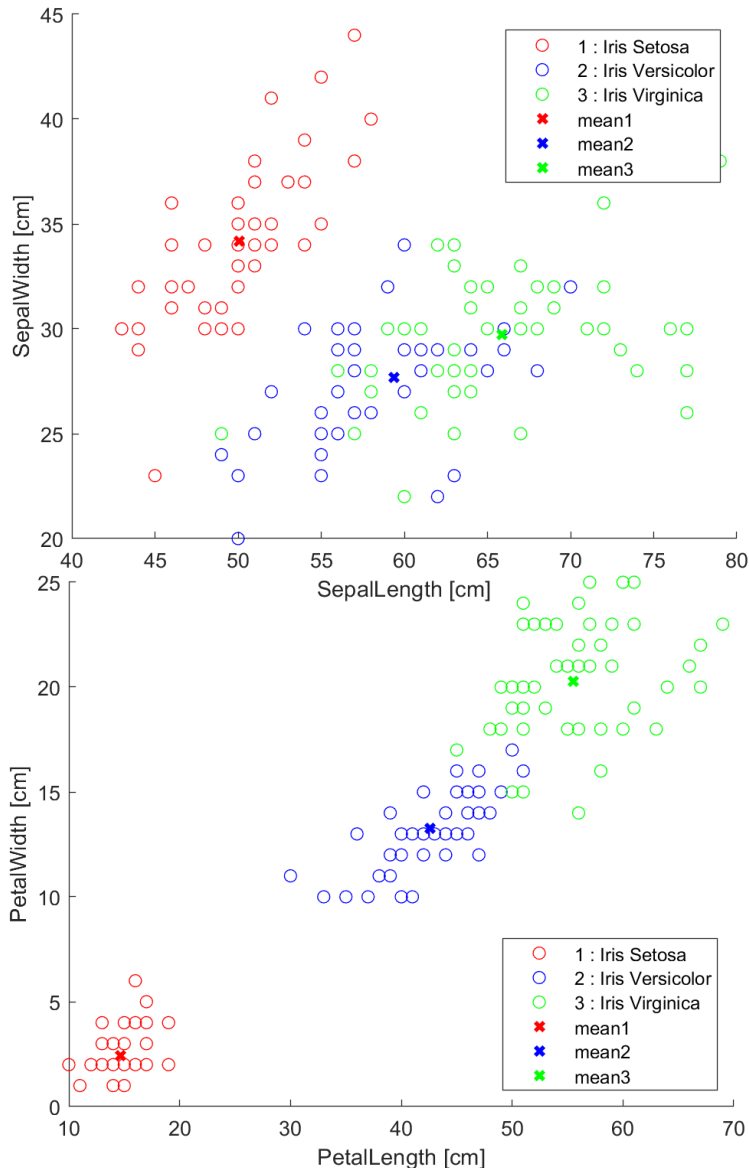
Fragen?

Business Understanding und Data Understanding

Ziele für den Iris Datensatz?



Jetzt kennen wir die Daten (grob)
Was könnte ein Business Ziel sein?



	Iris Setosa		Iris Versicolor		Iris Virginica	
	SepalLength	SepalWidth	SepalLength	SepalWidth	SepalLength	SepalWidth
Mittelwert	50,06	34,18	59,36	27,70	65,88	29,74
Median	50	34	59	28	65	30
25% Quantil	48	31	56	25	62	28
75% Quantil	52	37	63	30	69	31,75
Varianz	12,42	14,52	26,64	9,85	40,43	10,40
Standard-abweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

Business Understanding und Data Understanding

Ziele für den Iris Datensatz? -Lsg



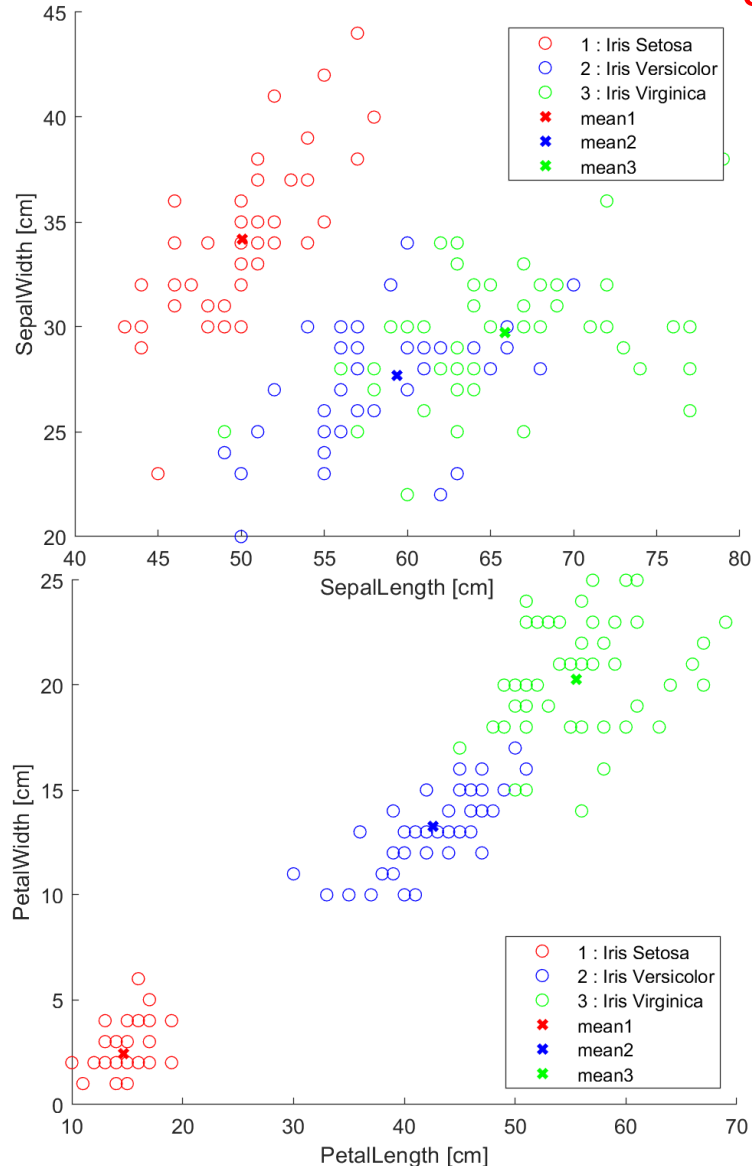
Jetzt kennen wir die Daten (grob)
Was könnte ein Business Ziel sein?

Es gibt nicht DIE EINE richtige Lösung,
aber zumindest wissen wir eine Richtung!

Regression?

Clustering?

Klassifikation?



	Iris Setosa		Iris Versicolor		Iris Virginica	
	SepalLength	SepalWidth	SepalLength	SepalWidth	SepalLength	SepalWidth
Mittelwert	50,06	34,18	59,36	27,70	65,88	29,74
Median	50	34	59	28	65	30
25% Quantil	48	31	56	25	62	28
75% Quantil	52	37	63	30	69	31,75
Varianz	12,42	14,52	26,64	9,85	40,43	10,40
Standard-abweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

Warum ist Data Understanding so wichtig?

Als Beispiel: Das Simpson Paradoxon

- 1. Beispiel: Bewertung verschiedener Gruppen fällt unterschiedlich aus, wenn man sie einzeln betrachtet oder gemeinsam

- Gerichtsstreit in Berkeley (1973)

	Bewerber	Davon zugelassen
Männlich	8442	44%
weiblich	4321	35%

- χ^2 – Verteilung: Bewerbungsanzahl war nicht gleich
- Es lag keine Diskriminierung vor (Frauen hatten generell eine niedrigere Zulassungsrate galt)

- 2. Beispiel: „More money makes you more li

Gesamtes Video unter:
<https://www.youtube.com/watch?v=ebEkn-BiW5k>



Ziele der heutigen Übung



■ Nach der heutigen Übung können Sie....

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

1

- ... Methoden zur Geschäftszielfindung beschreiben.

2

- ... Datentypen aufzählen und voneinander abgrenzen.

3

- ... „einfache“ Statistikwerte nennen und berechnen.

4

- ... Daten visualisieren.