Karlsruher Institut für Technologie

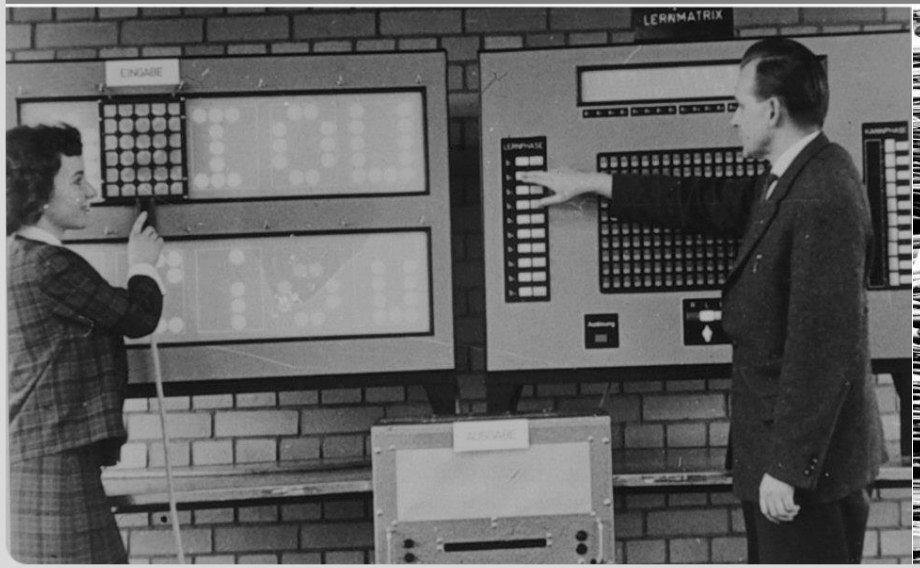
Institut für Technik der Informationsverarbeitung

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

Übung 04

Übung zu Informationstechnik II und Automatisierungstechnik – Nathalie Brenner

Prof. Dr.-Ing. Eric Sax

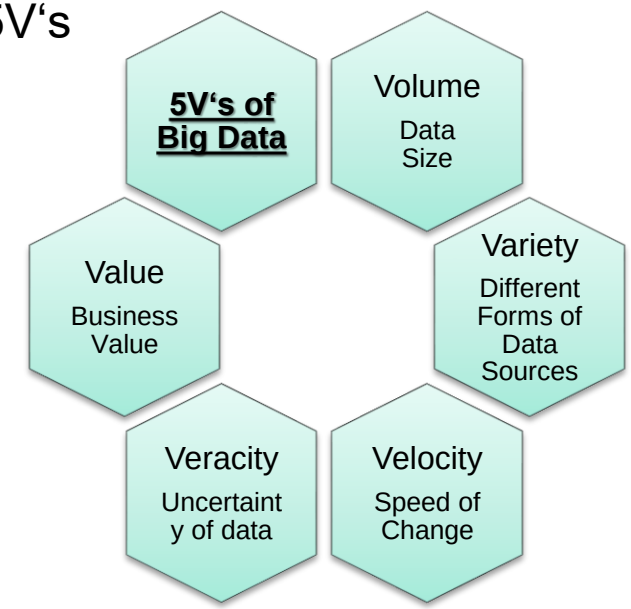


WIEDERHOLUNG ÜBUNG 3

Wiederholung Übung 2

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

Hypothese wird gestellt (Fachwissen) → Genau Messdaten werden aufgenommen → Wissenschaftlicher Beweis folgt

- Schlussfolgerung durch grobe Fragestellung
  Beruht auf Auffinden von Korrelationen

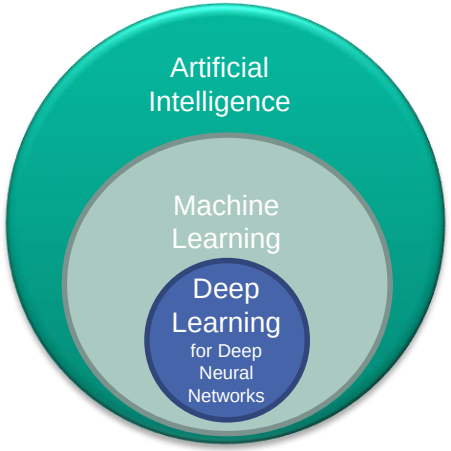
Grobe Fragestellung → Große Datenmenge (wächst meist während Bearbeitung weiter an) → Erkenntnisse aus Korrelationen

Wiederholung Übung 3

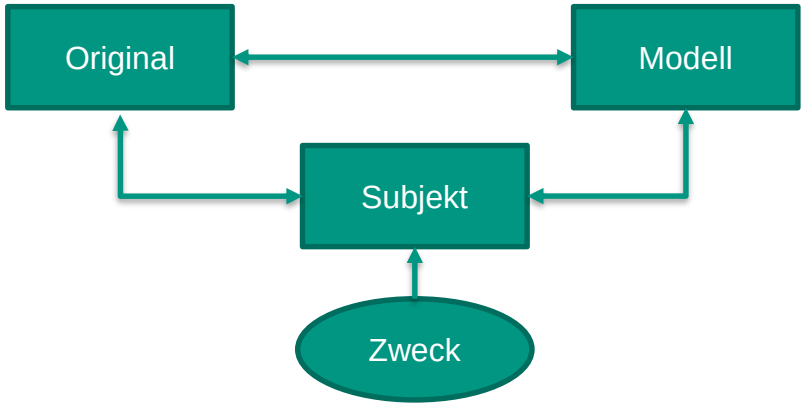
Data Science, Data Mining, Maschinelles Lernen und Trainieren

Grundbegriffe

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

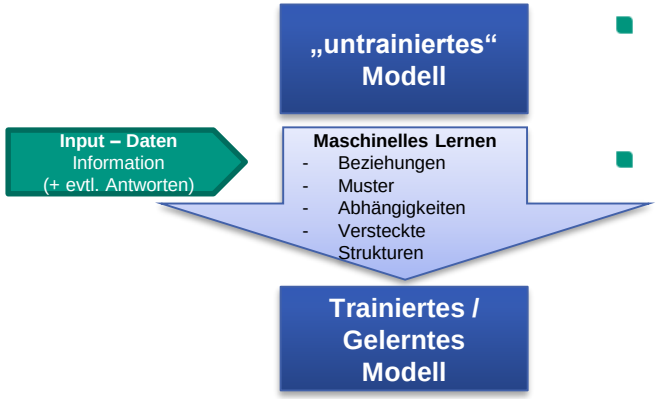


Definition eines Modells (nach Stachowiak)



Abbildung, Verkürzung, Pragmatismus

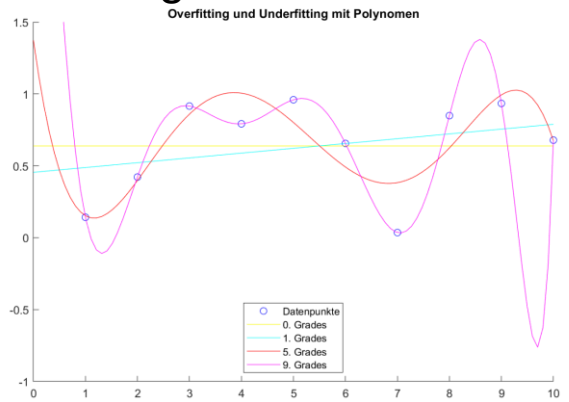
Modelle im Bezug auf Maschinelles Lernen



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

Gefahr des Overfitting und Underfitting

- **Overfitting**: Modell performiert gut auf Trainingsdaten, aber schlecht auf neuen Daten
  Rauschen kann ebenfalls mitgelernt werden
- **Underfitting**: Modell performiert sogar mit den Trainingsdaten schlecht
- *Schlussfolgerung: Modell ist nicht ausgereift/schlecht geeignet, neues Modell erstellen*



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. … gängige Prozessabläufe zur Analyse von Big Data Problemstellungen beschreiben
2. … Methoden zur Geschäftszielfindung beschreiben
3. … Datentypen aufzählen und voneinander abgrenzen
4. … „Methoden“ zur Datenverständnis nennen und anwenden

BIG DATA ALS PROZESS

Wir wollen eine Anomalie erkennen

Zwischenübung

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	57	44	15	4	Iris-setosa
17	54	39	13	4	Iris-setosa
18	51	35	14	3	Iris-setosa
19	57	38	17	3	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
66	67	31	44	14	Iris-versicolor
67	56	30	45	15	Iris-versicolor
68	58	27	41	10	Iris-versicolor
69	62	22	45	15	Iris-versicolor
70	56	25	39	11	Iris-versicolor
71	70	32	47	14	Iris-versicolor

- hierarchical cluster analysis
- DBSCAN
- One Class Support Vector Machine
- Isolation Forest

- (künstliche) Neuronale Netze
- Support Vector Machine
- Decision Tree
- Bayes-Klassifikation
- Random Forest
- Multivariate Adaptive Regressions-Splines
- Logistic Regression
- Linear Regression
- Harmonic Regression
- Diskriminanzanalyse
- Nächste-Nachbar-Klassifikation

- Regression
- Clustering
- Klassifikation

Lasst uns eine Anomalie erkennen!
Welchen Algorithmus würdet ihr verwenden?

Wir wollen eine Anomalie erkennen

Zwischenübung – „Lsg“ oder eben nicht…

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	57	44	15	4	Iris-setosa
17	54	39	13	4	Iris-setosa
18	51	35	14	3	Iris-setosa
19	57	38	17	3	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
66	67	31	44	14	Iris-versicolor
67	56	30	45	15	Iris-versicolor
68	58	27	41	10	Iris-versicolor
69	62	22	45	15	Iris-versicolor
70	56	25	39	11	Iris-versicolor
71	70	32	47	14	Iris-versicolor

hierarchical cluster analysis
DBSCAN
One Class Support Vector Machine
Isolation Forest
(künstliche) Neuronale Netze
Support Vector Machine
Decision Tree
Bayes-Klassifikation
Random Forest
Multivariate Adaptive Regressions-Splines
Logistic Regression
Linear Regression
Harmonic Regression
Diskriminanzanalyse
Nächste-Nachbar-Klassifikation

Regression

Clustering

Klassifikation

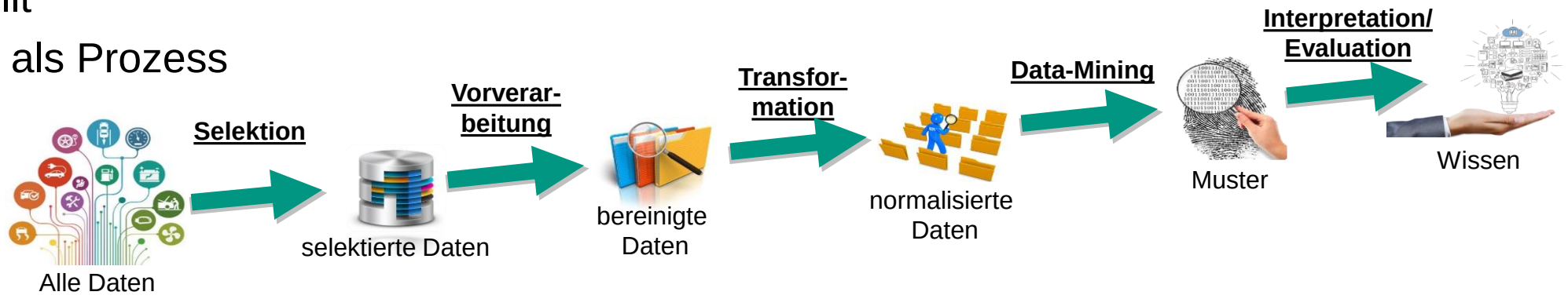
Lasst uns eine Anomalie erkennen!
Welchen Algorithmus würdet ihr verwenden?

Big Data als Prozess

Strukturierte Vorgehensweise

- Simpel einen beliebigen Maschinellen Lernalgorithmus auf vorliegende Daten anzuwenden ist nicht zielführend!
- ➢ Strukturelle Vorgehensweise wird benötigt!
- Data Mining erfordert die Anwendung von beträchtlichen wissenschaftlichen und technischen Kenntnissen
- Es wird ein wohlverstandenes Verfahren für die Strukturierung benötigt, welches
  - Einheitlichkeit
  - Reproduzierbarkeit
  - Objektivität

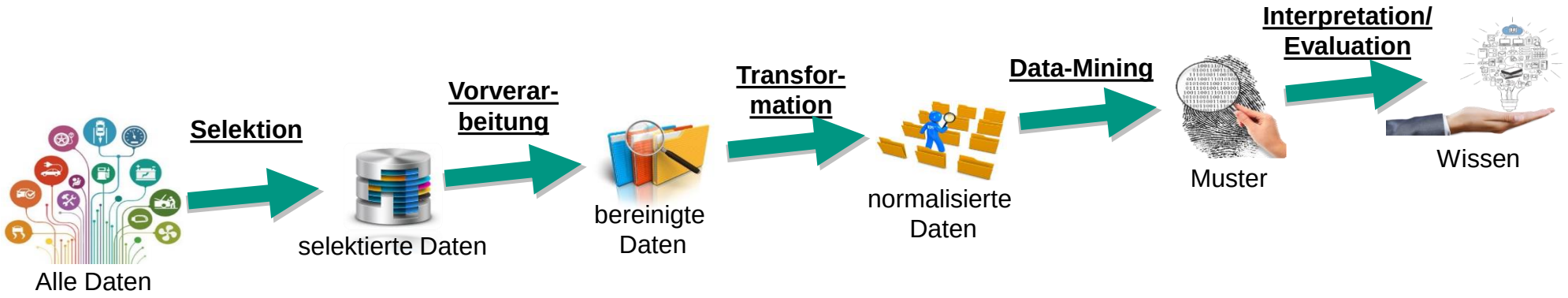
  sicherstellt
- ➢ Big Data als Prozess



Big Data als Prozess

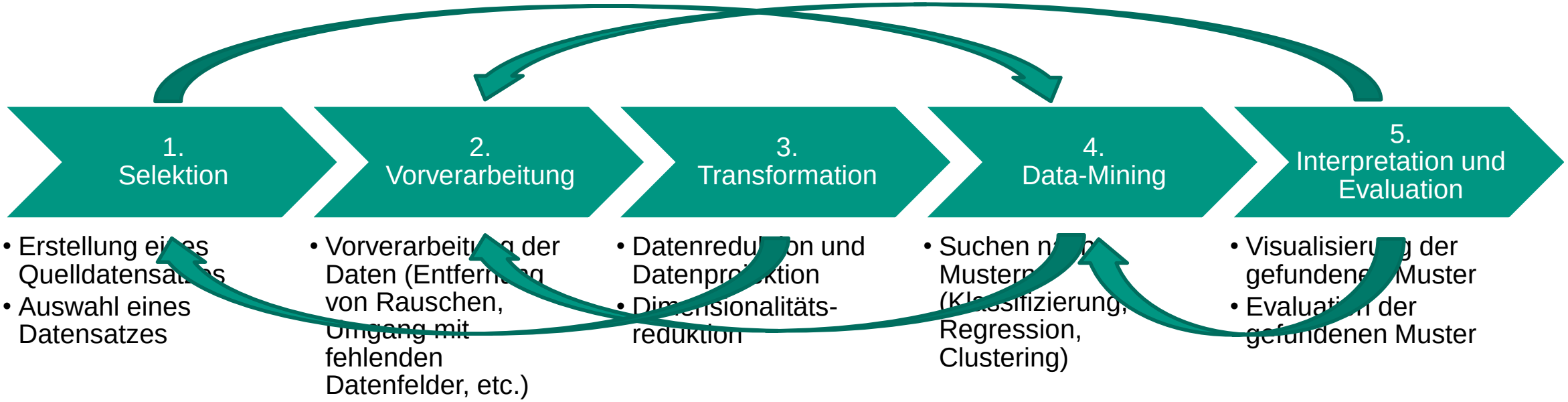
Knowledge Discovery in Databases (KDD)

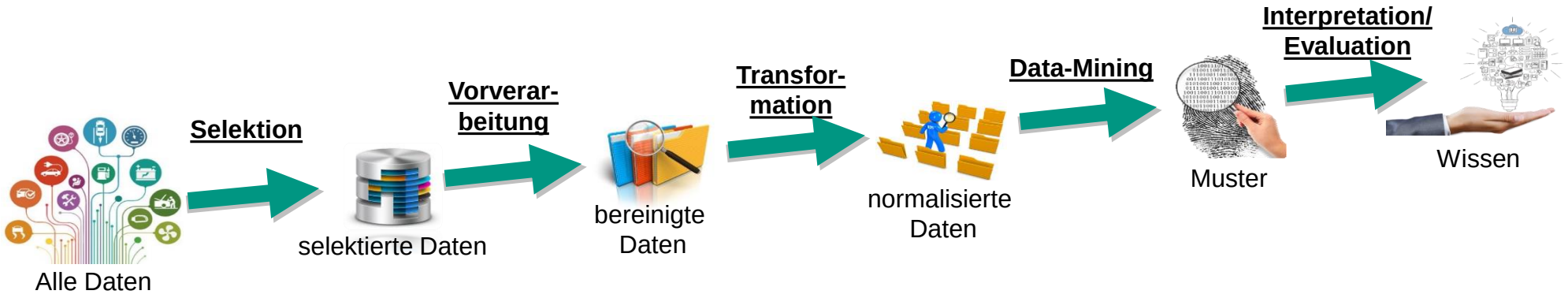
1. Selektion	2. Vorverarbeitung	3. Transformation	4. Data-Mining	5. Interpretation und Evaluation
• Erstellung eines Quelldatensatzes • Auswahl eines Datensatzes	• Vorverarbeitung der Daten (Entfernung von Rauschen, Umgang mit fehlenden Datenfelder, etc.)	• Datenreduktion und Datenprojektion • Dimensionalitäts-reduktion	• Suchen nach Mustern (Klassifizierung, Regression, Clustering)	• Visualisierung der gefundenen Muster • Evaluation der gefundenen Muster



Big Data als Prozess

Knowledge Discovery in Databases (KDD) als iterativer und zyklischer Prozess





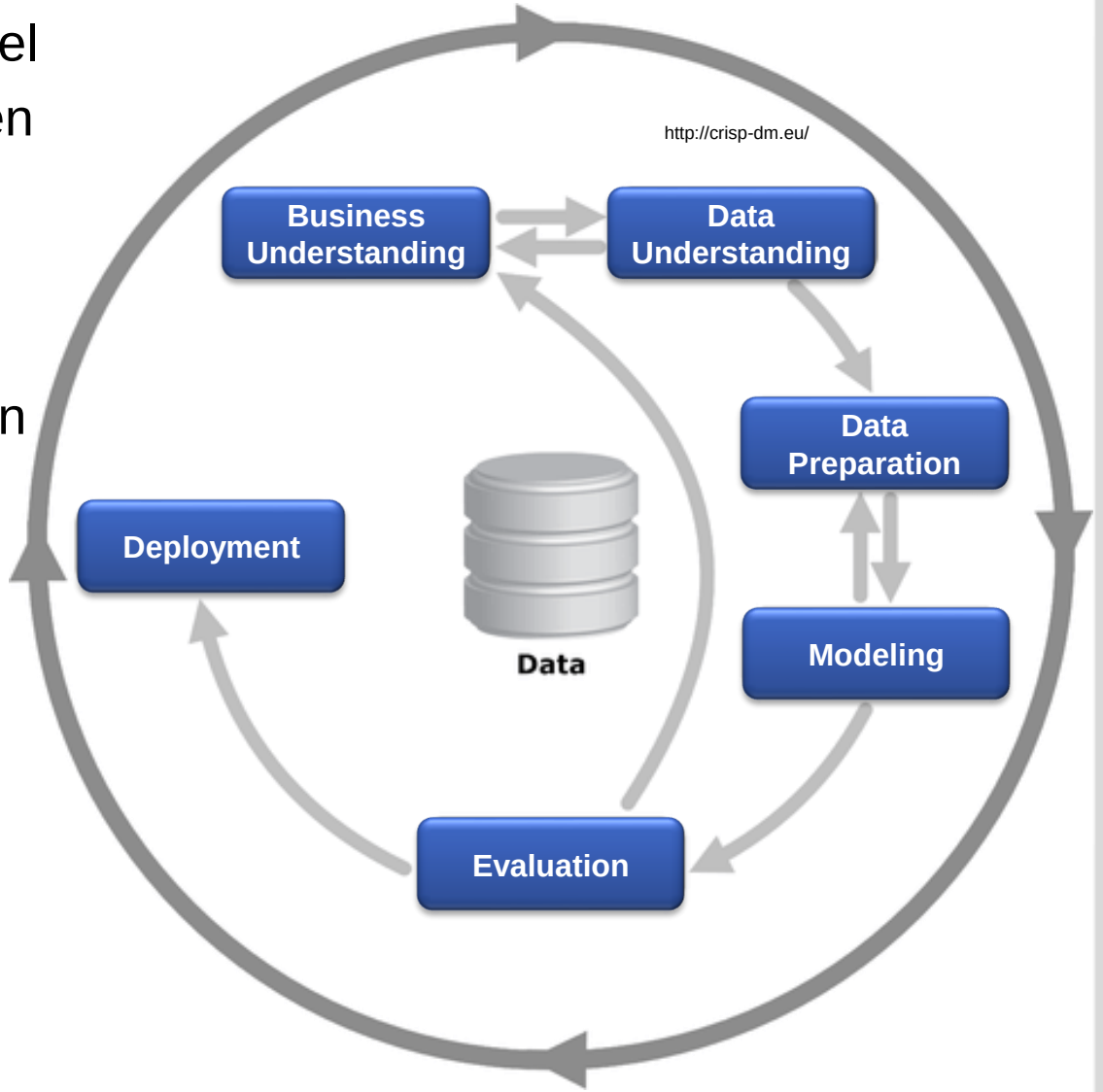
Big Data als Prozess

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

- Iterationen sind keine Ausnahmen, sondern die Regel
- erster Durchlauf dient meist der Erkundung der Daten

- Zyklischer Charakter ist im CRISP-DM enthalten

- Phasen nehmen unterschiedlichen Arbeitsaufwand in Anspruch
  - 20-30% Data Understanding
  - 50-70% Data Preparation
  - 10-20% Modeling and Evaluation
  - 5-10% Deployment

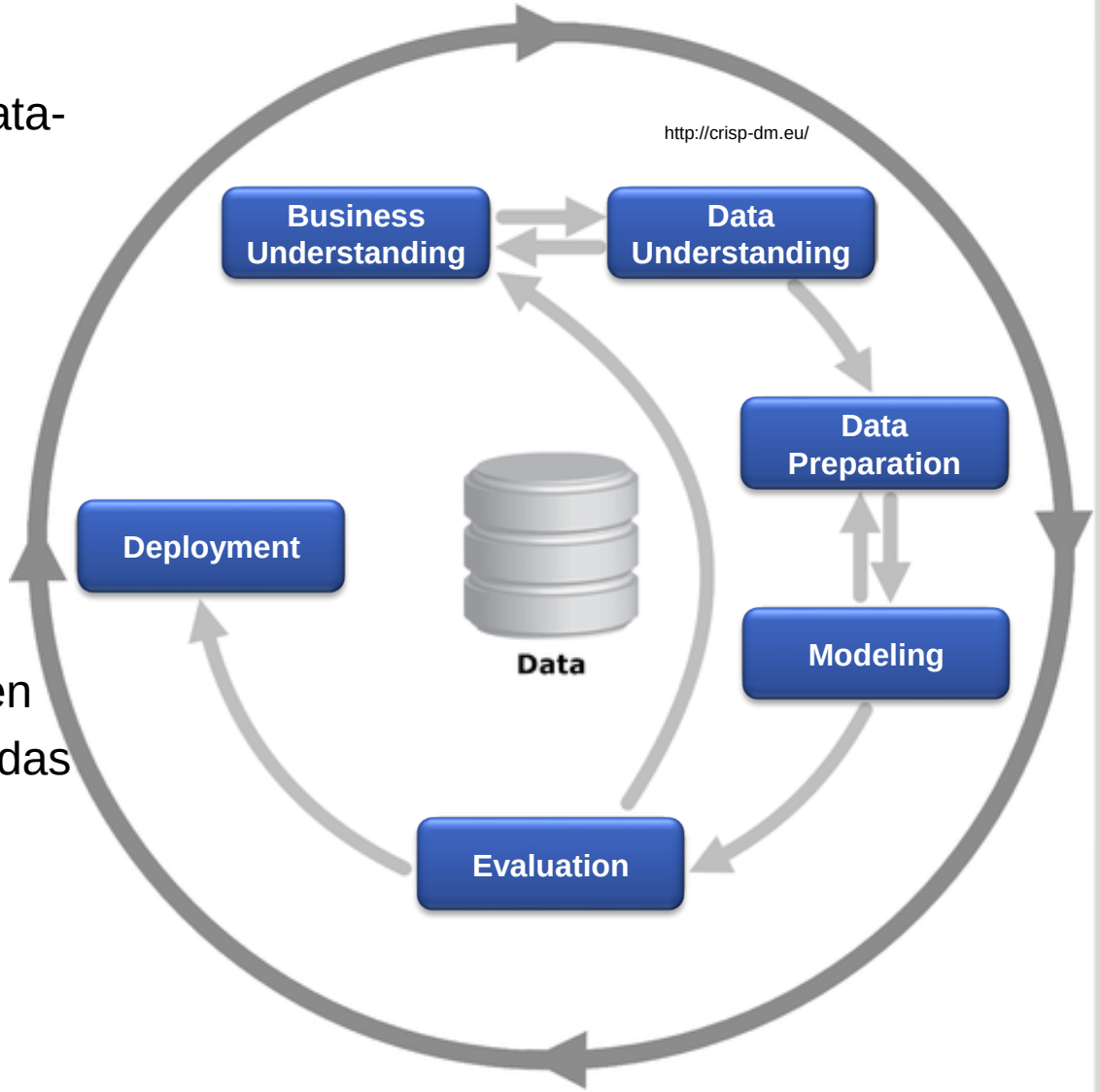


Cross Industry Standard Process for Data Mining (CRISP-DM)

Business- and Data Understanding

- Aufgabenverständnis: „Formulierung der Aufgabe“
  - Aufteilung des 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

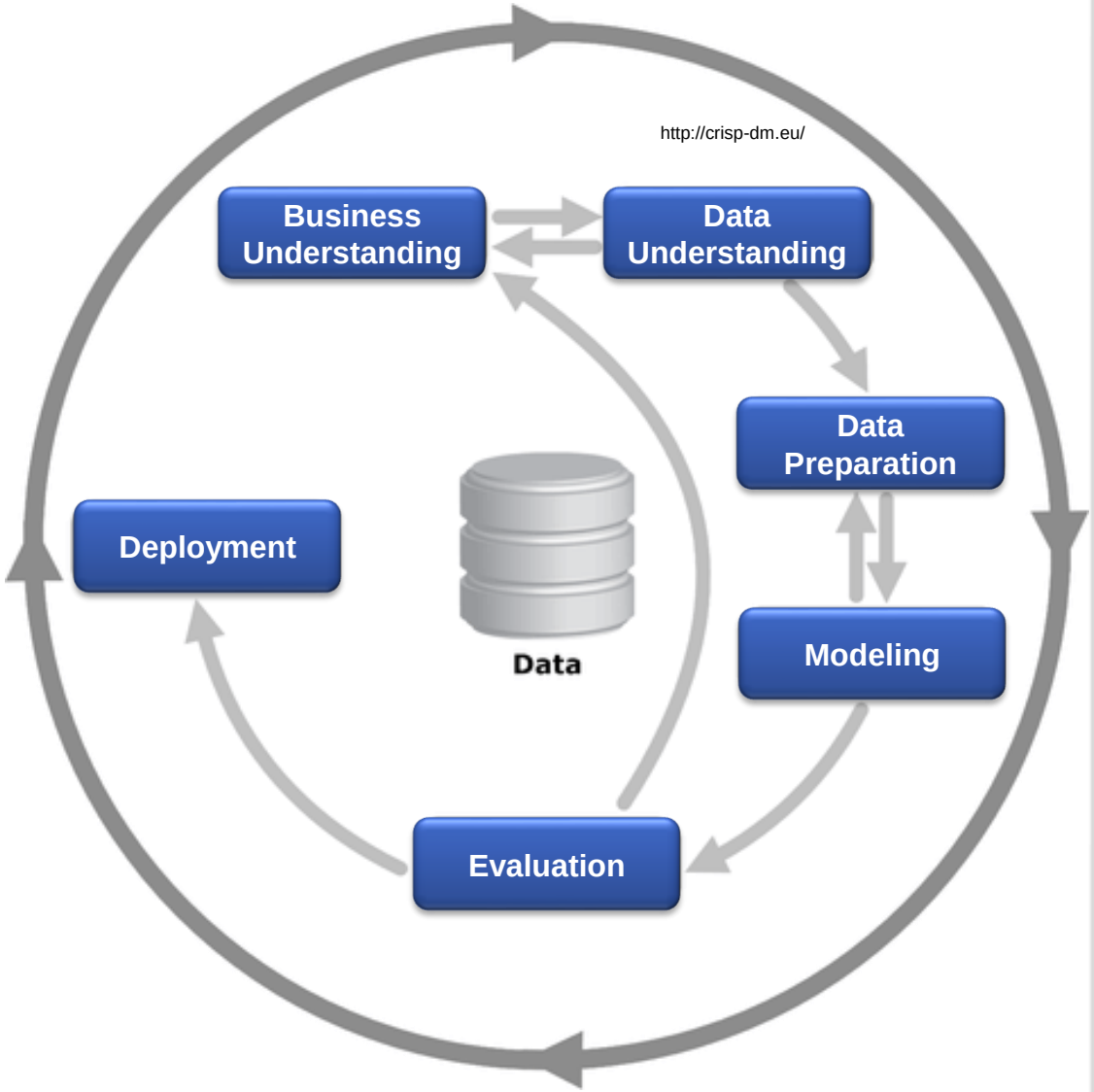


Cross Industry Standard Process for Data Mining (CRISP-DM)

Data Preparation

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

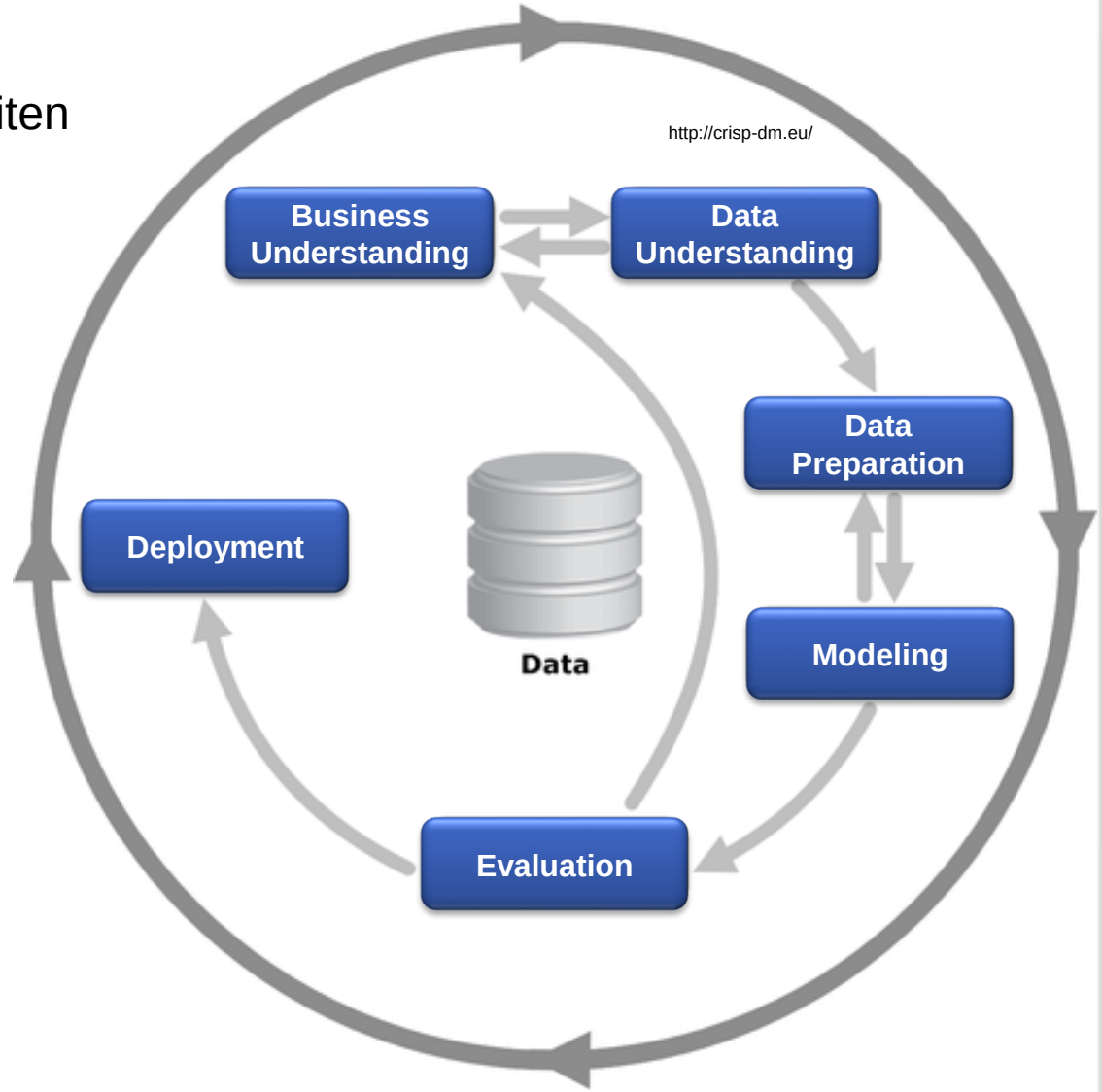
„Im Allgemeinen müssen Data Scientists anfangs beträchtliche Zeit dafür aufwenden, die Variablen zu definieren, die später verwendet werden. Gerade hier kommen die menschliche Kreativität, der gesunde Menschenverstand und das Fachwissen ins Spiel. Die Qualität einer Data-Mining-Lösung beruht oft darauf,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



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

Modeling

- Modellbildung
  - Suche nach Modellen, Mustern oder Gesetzmäßigkeiten in den vorliegenden Daten

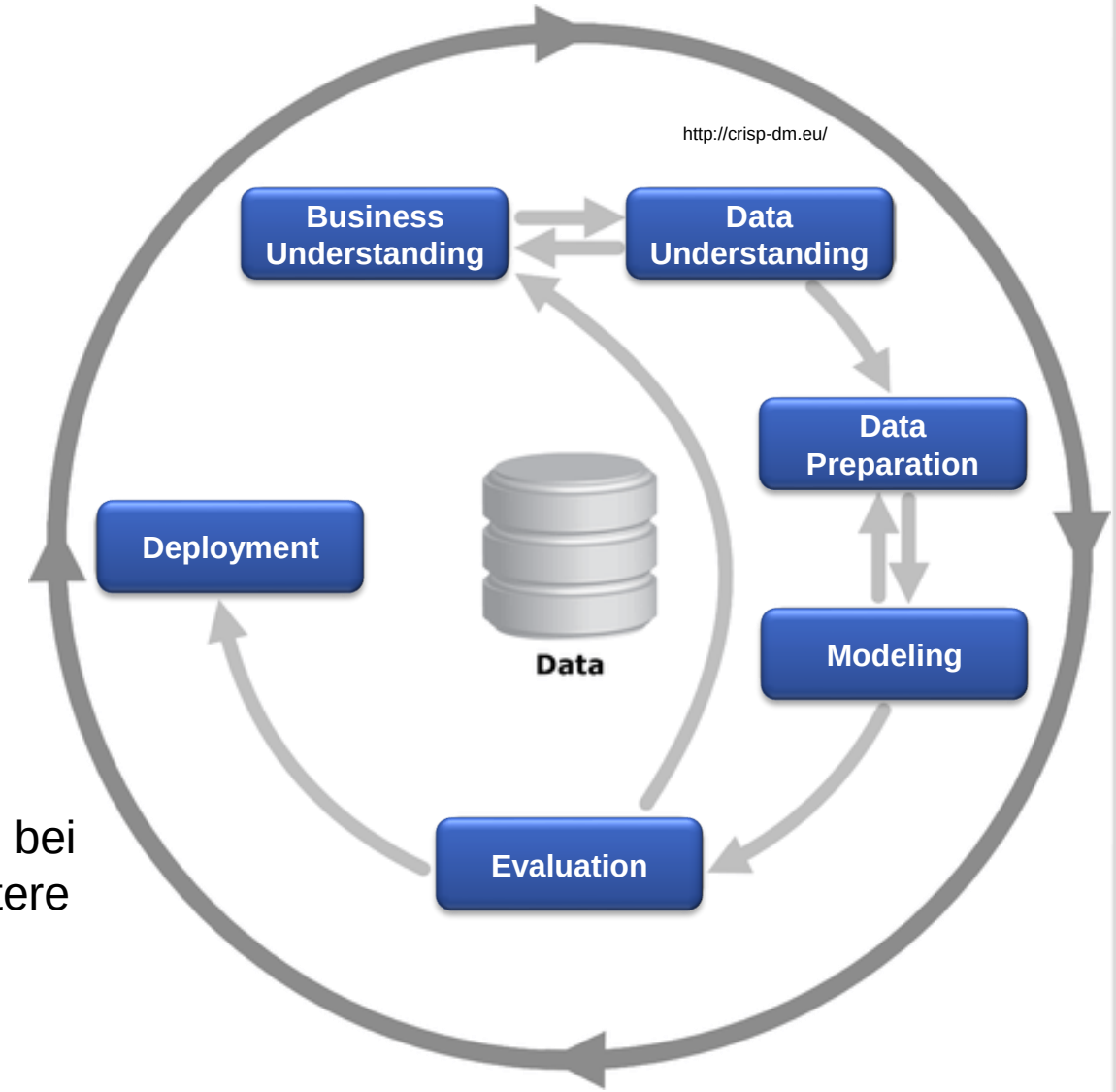


Cross Industry Standard Process for Data Mining (CRISP-DM)

Evaluation und Deployment

- Beurteilung
  - Bewertung der Ergebnisse der Modellbildung (Zuverlässigkeit, Gültigkeit, etc.)
  - Erfüllung der Aufgabenstellung prüfen
  - Testen der Praxistauglichkeit
- Einsatz
  - Anwenden der Ergebnisse der Modellbildung
  - Implementierung eines Vorhersagemodells für Informationssysteme/Geschäftsvorgänge

Risiko: *„Das Modell ist nicht das, was die Data Scientists entwerfen, sondern das was die Ingenieure herstellen.“*



- Häufig wird nach Durlaufens des Deployment nun wieder bei Phase des Aufgabenverständnisses begonnen. Eine weitere Iteration kann eine verbesserte Lösung hervorbringen (Zykluseigenschaft des CRISP-DM)

CRISP-DM und kdd

Was sagt „die Industrie“ dazu?

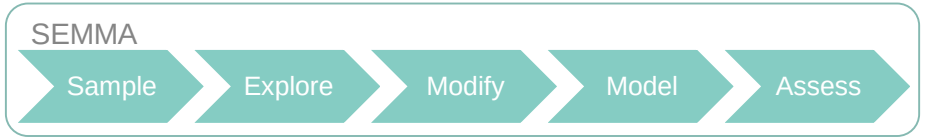
- **CRISP-DM** ist eine Beschreibung des „Workflows“ in Data-Mining-Projekten
  „the first step towards defining a data science methodology“
  ~ Saltz J.
- **Kdd**-Ansätze konzentrieren sich auf die Schritte der Durchführung von Data Mining als auf die Beschreibung eines umfassenden Projektmanagement-Konzepts

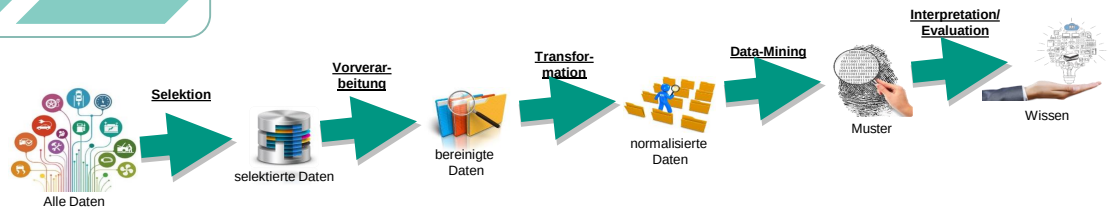
http://www.datascience-pm.com

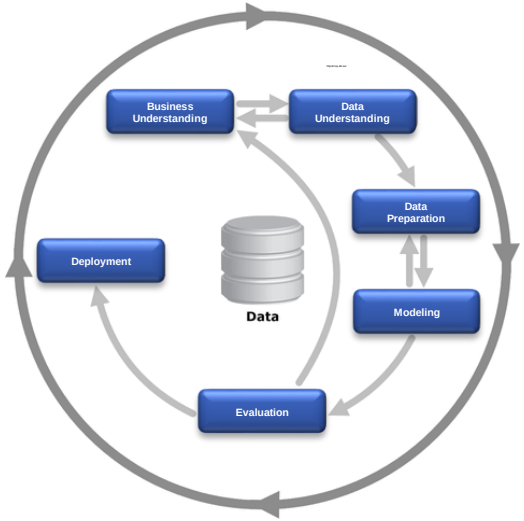
What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

	2014 poll	2007 poll
CRISP-DM (86)	43%	42%
My own (55)	27.5%	19%
SEMMA (17)	8.5%	13%
Other, not domain-specific (16)	8%	4%
KDD Process (15)	7.5%	7.3%
My organizations' (7)	3.5%	5.3%
A domain-specific methodology (4)	2%	4.7%
None (0)	0%	4.7%

- CRISP-DM ist der am häufigsten in der Industrie genutzte Standard



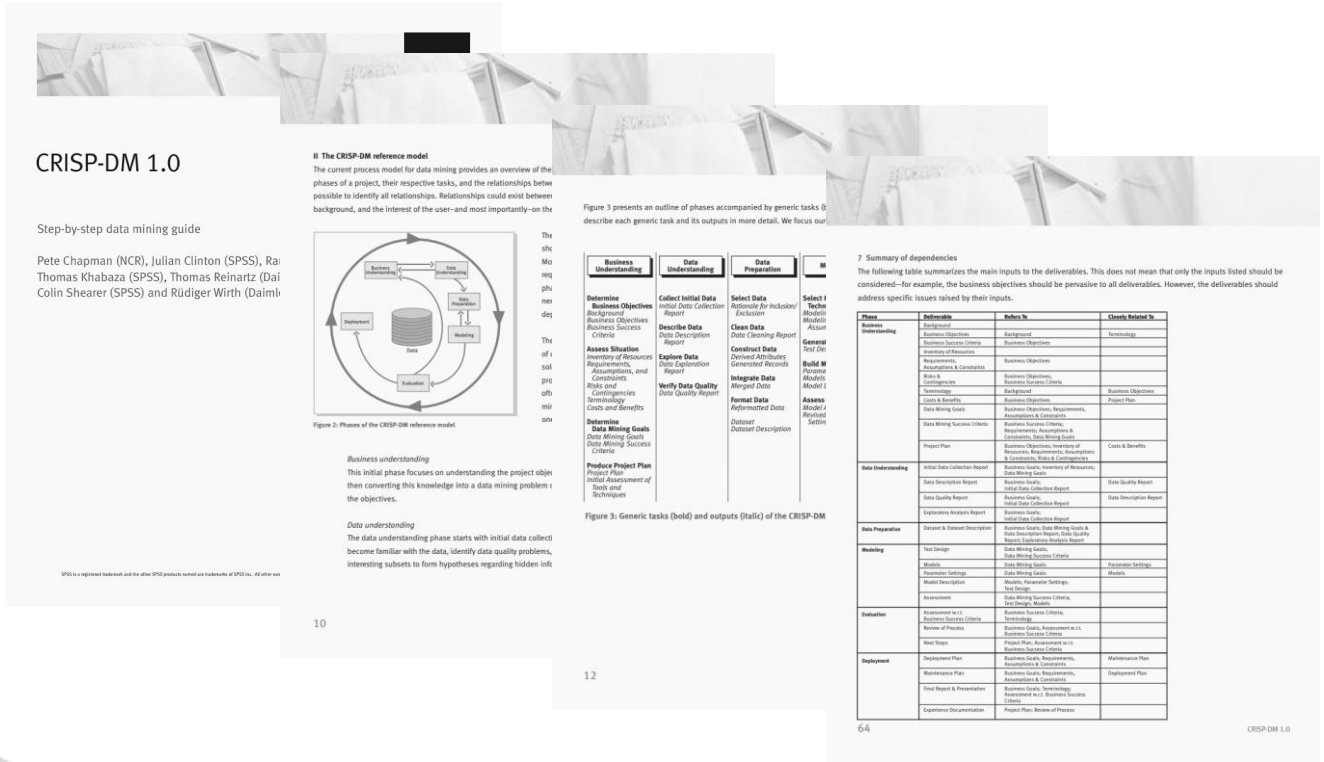




CRISP-DM und kdd

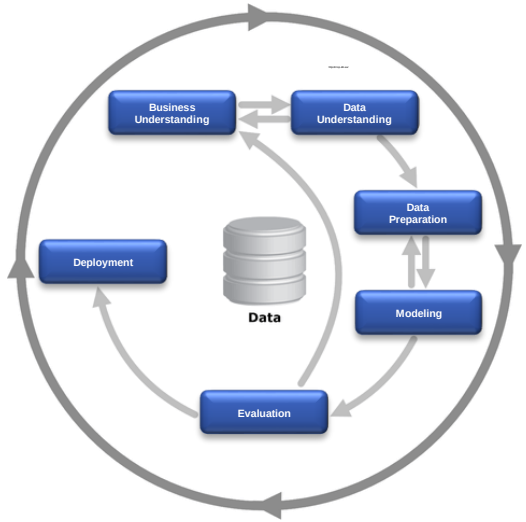
Was sagt „die Industrie“ dazu?

- **CRISP-DM** ist eine Beschreibung des „Workflows“ in Data-Mining-Projekten
  „the first step towards defining a data science methodology“
  ~ Saltz J.
- CRISP-DM bietet gute Dokumentation durch das bereit-gestellte „Handbook“



http://www.datascience-pm.com

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]	2014 poll	2007 poll
CRISP-DM (86)	43%	42%
My own (55)	27.5%	19%
SEMMA (17)	8.5%	13%
Other, not domain-specific (16)	8%	4%
KDD Process (15)	7.5%	7.3%
My organizations' (7)	3.5%	5.3%
A domain-specific methodology (4)	2%	4.7%
None (0)	0%	4.7%



- Big Data als Prozess
  - KDD und CRISP-DM
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modeling
    - Evaluation
    - Deployment

CRISP-DM
BUSINESS UNDERSTANDING
DATA UNDERSTANDING

Business Understanding

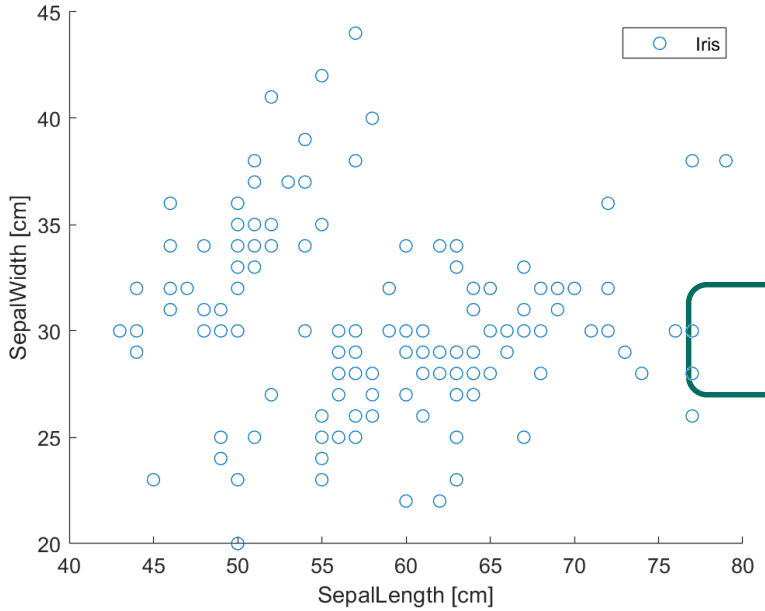
Von der geschäftlichen Aufgabe zum Data-Mining-Verfahren

- 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?

Business Understanding

Anwendung auf den „Iris Datensatz“



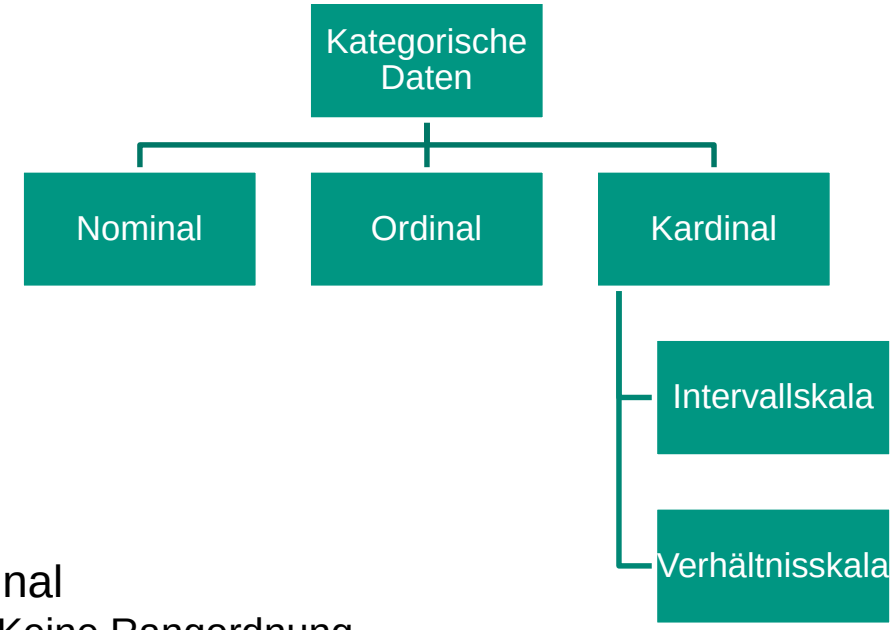
Datenverständnis!

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
	46	31	15	2	Iris-setosa
	50	36	14	2	Iris-setosa
	54	39	17	4	Iris-setosa
	46	34	14	3	Iris-setosa
	50	34	15	2	Iris-setosa
	44	29	14	2	Iris-setosa
	49	31	15	1	Iris-setosa
11	54	37	15	2	Iris-setosa
12	48	34	16	2	Iris-setosa
	48	30	14	1	Iris-setosa
	43	30	11	1	Iris-setosa
	58	40	12	2	Iris-setosa
	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

- 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?

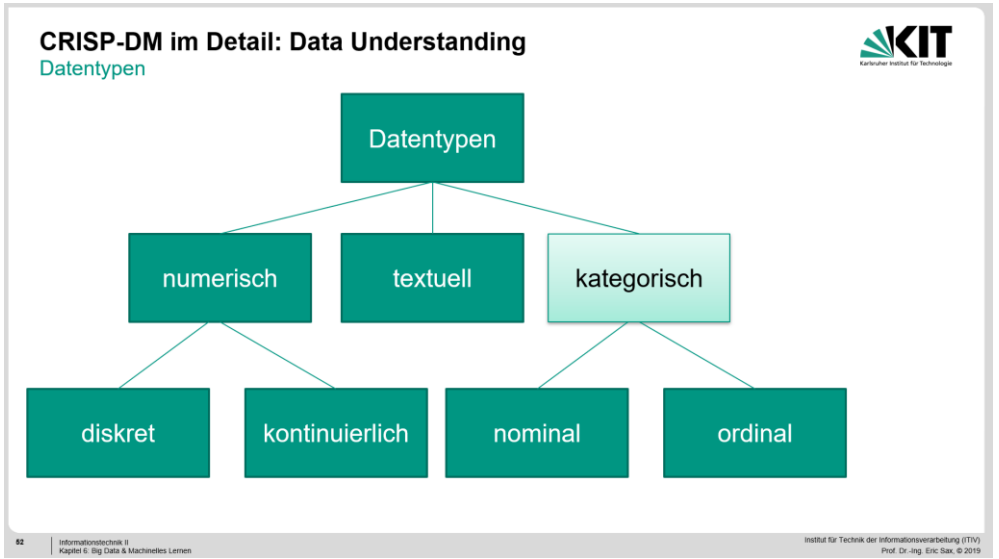
Data Understanding

Datentypen – Kategorische Daten



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

Wiederholung Vorlesung:



Data Understanding

Merkmale/Features – Titanic Datensatz

- 11 Features (Name, Alter, Geschlecht, Ticketklasse, Überlebt,…)
  1. Survival
  2. Pclass
  3. Name
  4. Sex
  5. Age
  6. SibSp
  7. Parch
  8. Ticket
  9. Fare
  10. Cabin
  11. Embarked

Feature, Merkmale, Attribute

	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	umings, 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		aN	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

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

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-s
13	48	30	14	1	Iris-s
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

Data Undestanding
Statistik und Visualisierung

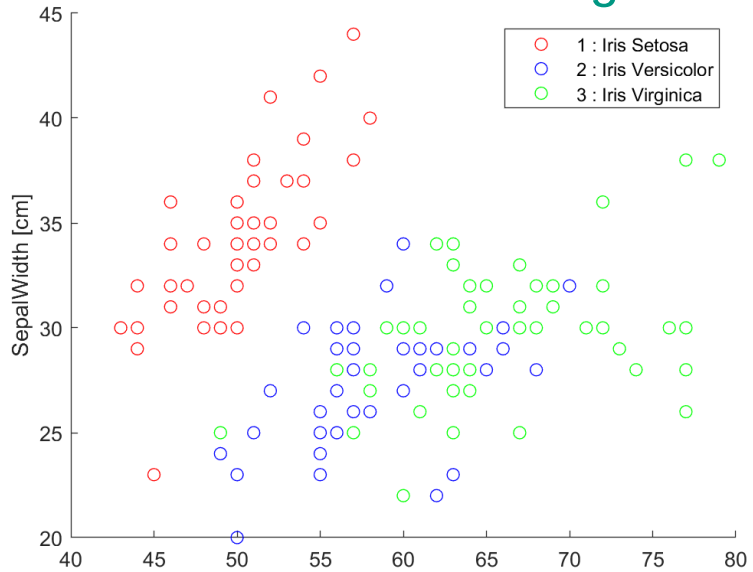
- Datenvisualisierung
  - Daten erkunden
  - Daten/Ergebnisse übermitteln

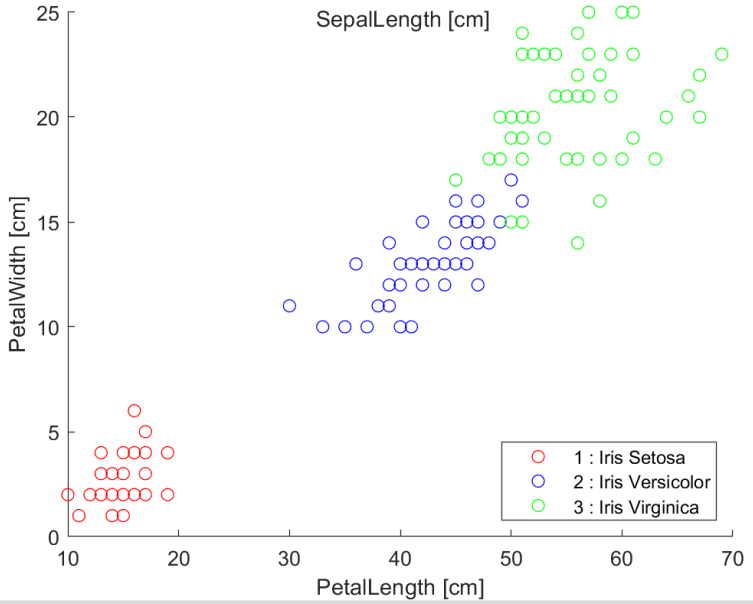
- Line Chart: gut geeignet um „Trends“ zu zeigen
- Scatter Charts: Beziehungen zwischen Datenpunkten gut erkennbar
- Bar Charts/Histogramme: Beziehung/Verhalten von Datenmengen zu einander
- Boxplots: Minimum, Maximum, Quantile, Median
  Häufigkeitsverteilung der Daten innerhalb der Mengen

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

Data Understanding

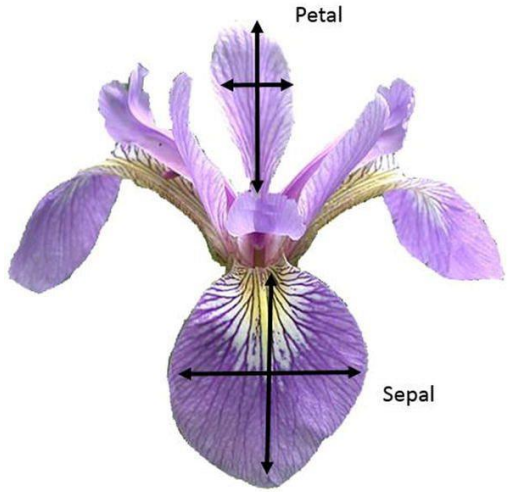
Statistik und Visualisierung - Scatterplot





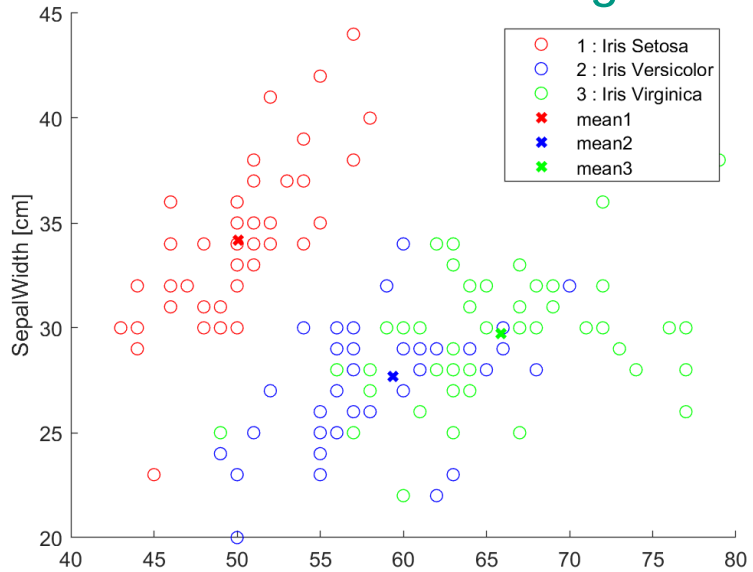
- Gelabelte Daten:
  - Iris Setosa
  - Iris Versicolor
  - Iris Virginica

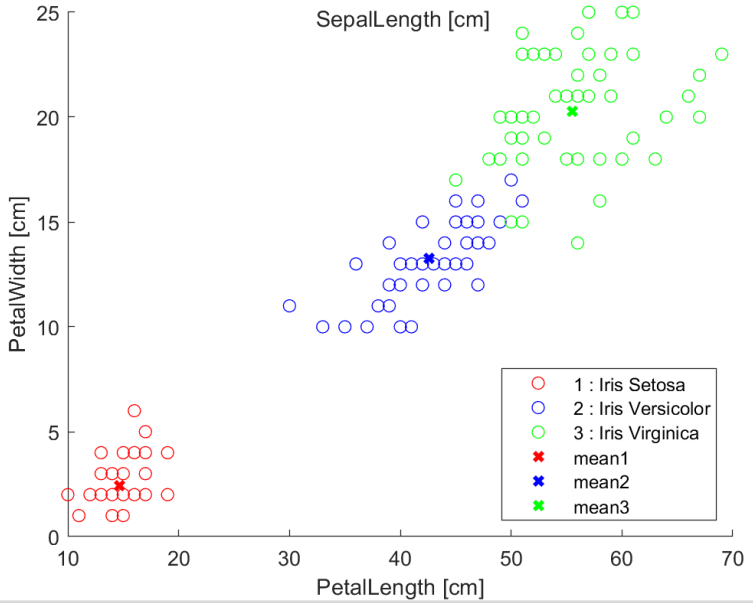
- Jeweils 4 Merkmale in [cm]
  - SepalLength
  - SepalWidth
  - PetalLength
  - PetalWidth



Data Understanding

Statistik und Visualisierung - Scatterplot





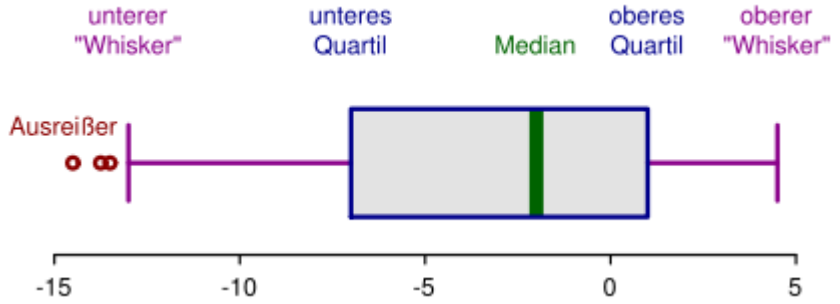
- Statistik:
  - Mittelwert $\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$
  - Varianz $\sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2$
  - Standardabweichung $\sigma = \sqrt{\sigma^2}$
  - Median $x = \begin{cases} x_{n+1/2} & n\ ungerade \\ 1/2\ (x_{n/2} + x_{(n/2)+1}) & n\ gerade \end{cases}$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

	Iris Setosa		Iris Versicolor		Iris Virginica	
	SepalWidth	SepalLength	SepalWidth	SepalLength	SepalWidth	SepalLength
Mittelwert	34,18	50,06	27,70	59,36	29,74	65,88
Varianz	14,52	12,42	9,85	26,64	10,40	40,43
Standard-abweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$
Minimum	23	43	20	49	22	49
Maximum	44	58	34	70	38	79

Data Understanding

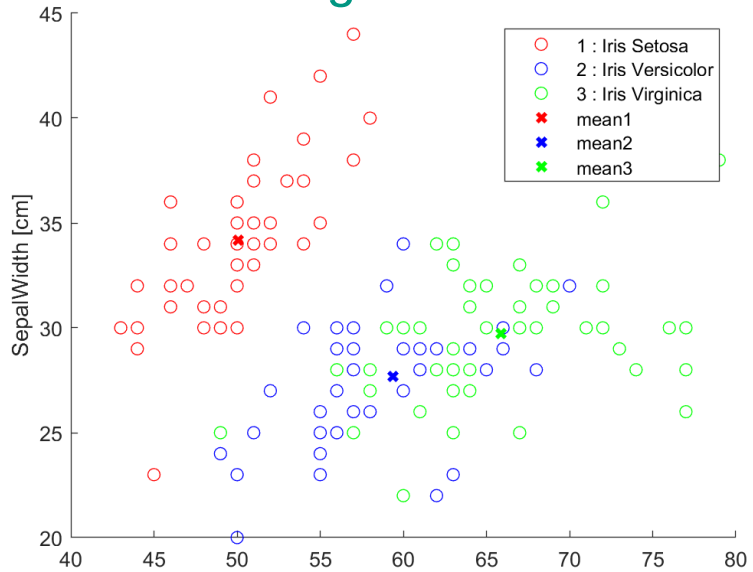
Statistik und Visualisierung - Boxplot

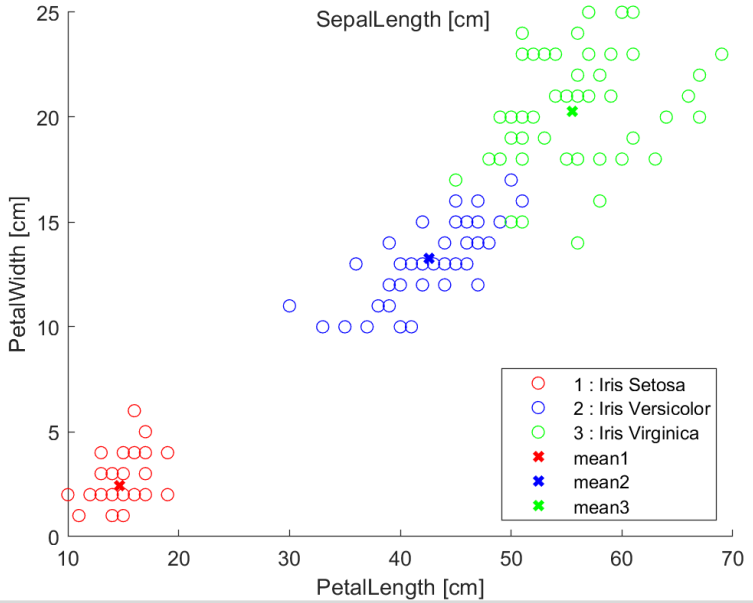
- Grafische Darstellung der Verteilung eines Merkmals
  - Wo liegt das Minimum
  - Wo liegt das Maximum
  - In welchem Bereich befinden sich die Daten
  - Wie verhält sich die Häufigkeitsverteilung der Datenpunkte?
  - „Box“ gibt 50% aller Daten an



Data Understanding - Boxplot

Zwischenübung

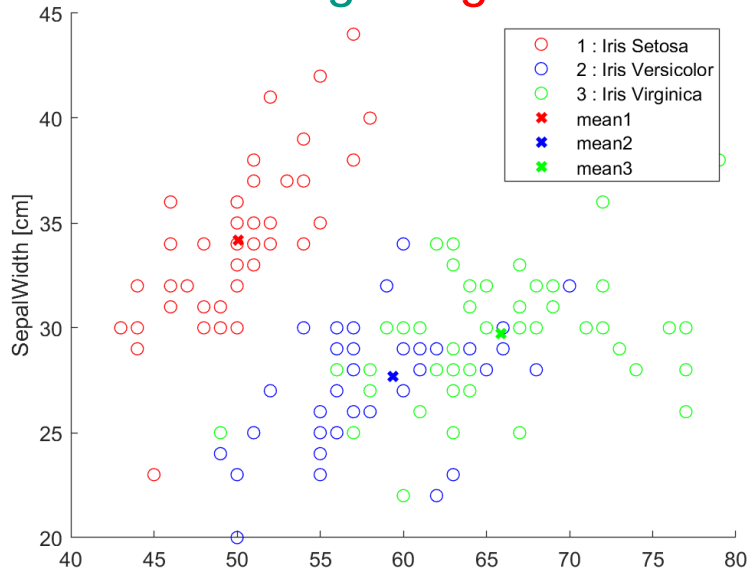


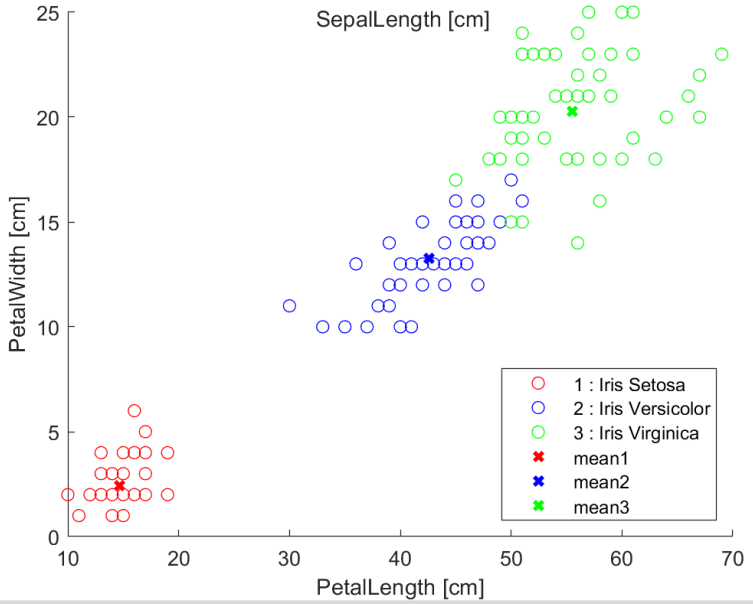


- Zeichen Sie (grob) ein Boxplot Diagramm des Iris Datensatzes zu:
  - SepalWidth
  - SepalLength
  - PetalWidth
  - PetalLength

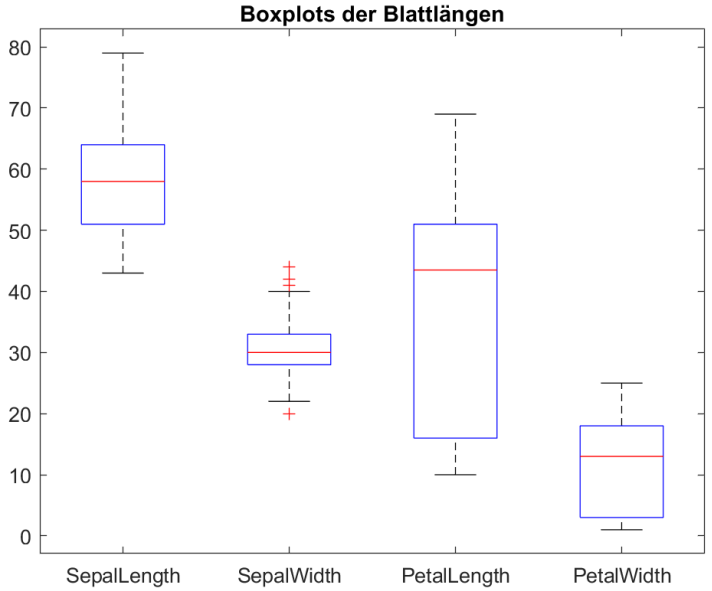
Data Understanding - Boxplot

Zwischenübung - Lsg





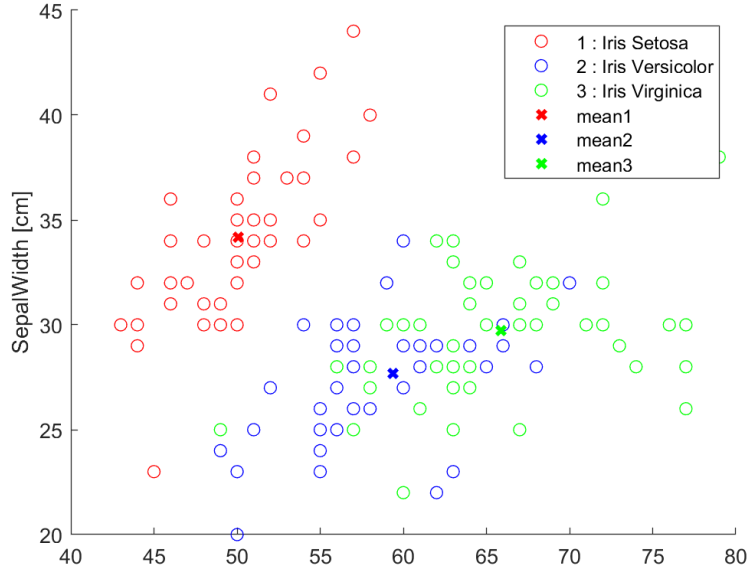
- Zeichen Sie (grob) ein Boxplot Diagramm des Iris Datensatzes zu:
  - SepalWidth
  - SepalLength
  - PetalWidth
  - PetalLength

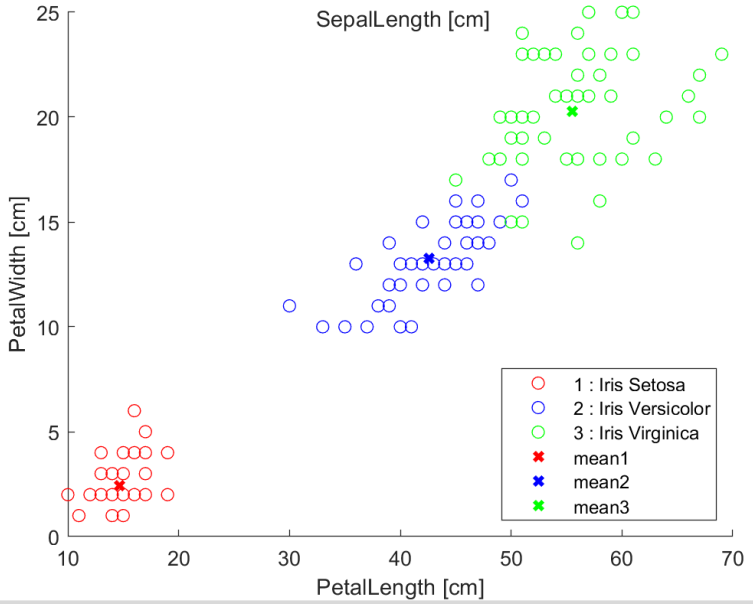


- Business Understanding
- Data Understanding

Business Understanding und Data Understanding

Ziele für den Iris Datensatz?





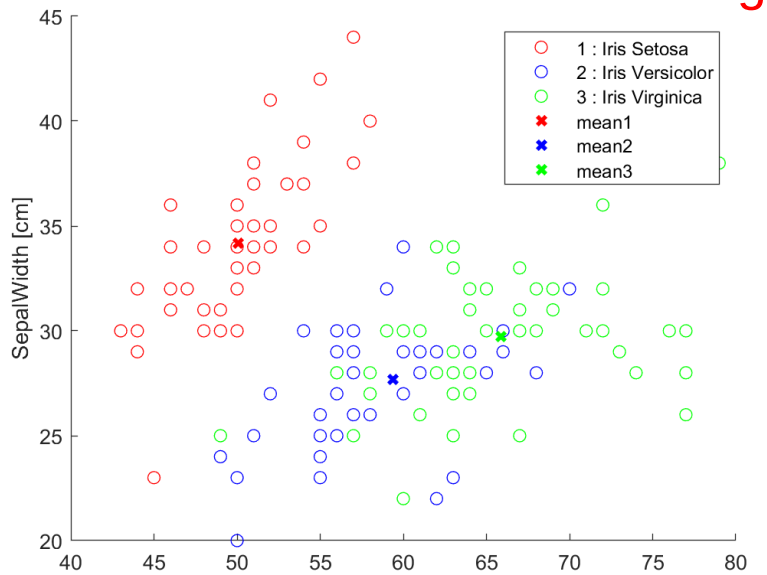
Jetzt kennen wir die Daten (grob)

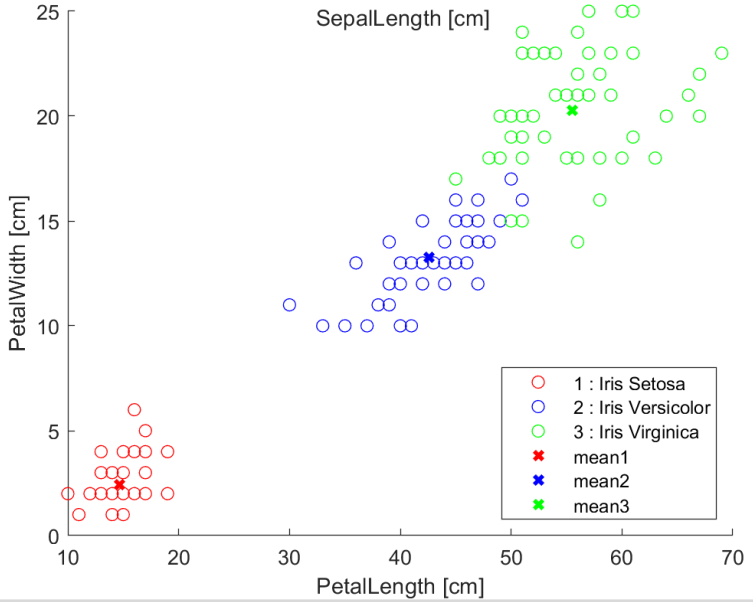
Was könnte ein Business Ziel sein?

	Iris Setosa		Iris Versicolor		Iris Virginica	
	SepalWidth	SepalLength	SepalWidth	SepalLength	SepalWidth	SepalLength
Mittelwert	34,18	50,06	27,70	59,36	29,74	65,88
Varianz	14,52	12,42	9,85	26,64	10,40	40,43
Standard-abweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$
Minimum	23	43	20	49	22	49
Maximum	44	58	34	70	38	79

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	
	SepalWidth	SepalLength	SepalWidth	SepalLength	SepalWidth	SepalLength
Mittelwert	34,18	50,06	27,70	59,36	29,74	65,88
Varianz	14,52	12,42	9,85	26,64	10,40	40,43
Standard-abweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$
Minimum	23	43	20	49	22	49
Maximum	44	58	34	70	38	79

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%

  - $X^2 - Verteilung$: Bewerbungsanzahl war nic
  - Es lag keine Diskriminierung vor (Frauen ha
    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. … gängige Prozessabläufe zur Analyse von Big Data Problemstellungen beschreiben
2. … Methoden zur Geschäftszielfindung beschreiben
3. … Datentypen aufzählen und voneinander abgrenzen
4. … „Methoden“ zum Datenverständnis nennen und anwenden