Einführung in Data Science

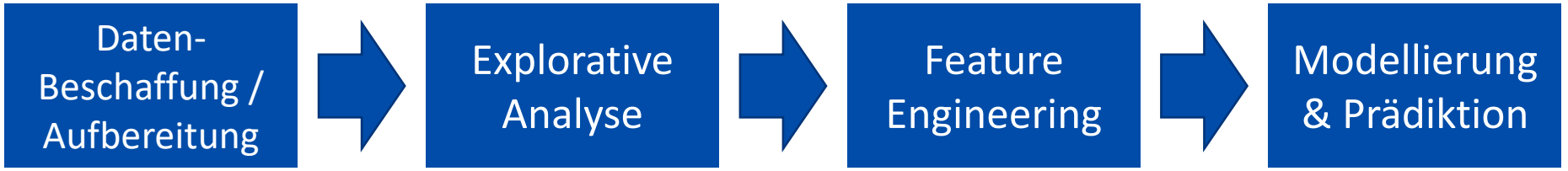
Unser Plan für heute:

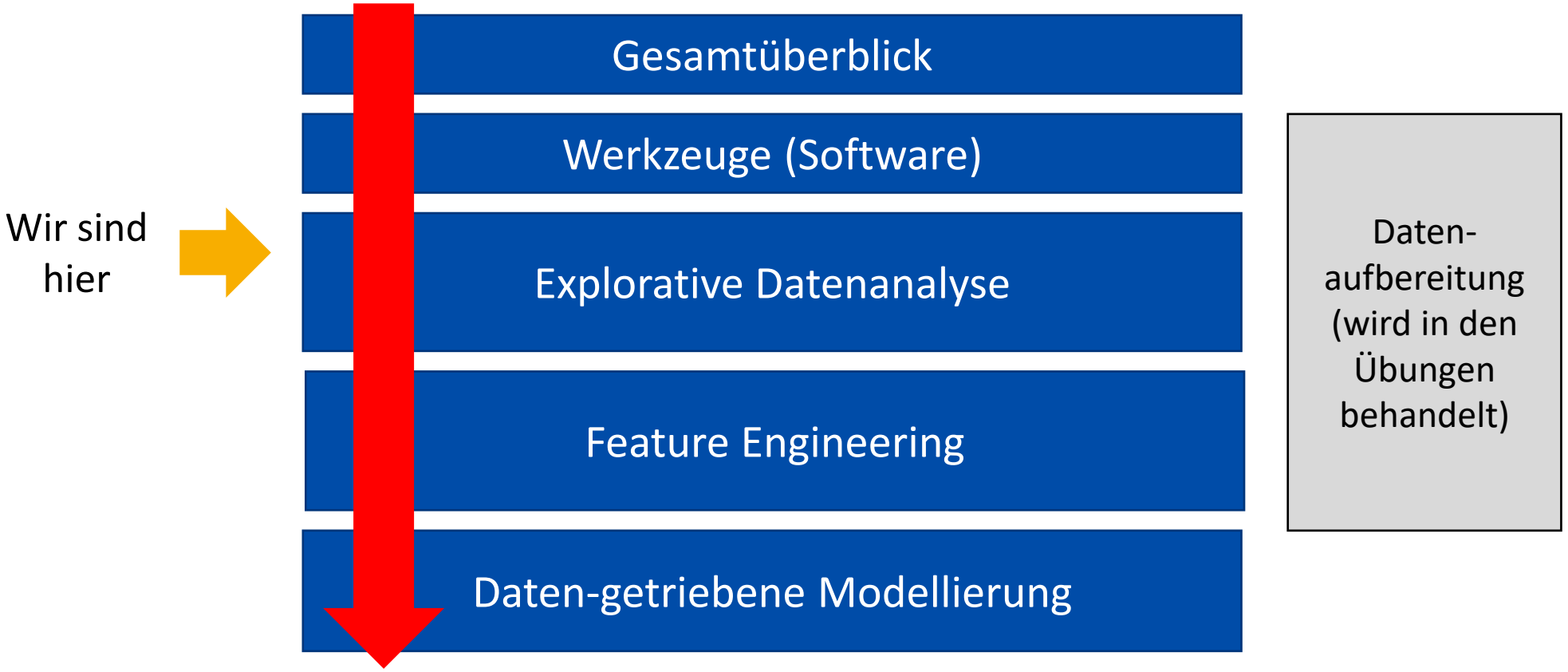
1. Wiederholung

2. Visualisierung – Teil 2

3. Multivariate explorative Analyse
   - Zusammenhangsmaße (Interdependenzmaße)

Data Science





Curriculum

1. Grundbegriffe / Überblick
2. Zentrale Softwarebibliotheken

3. Univariate explorative Analyse (EDA) Visualisierung (Teil 1)
4. Visualisierung (Teil 2), Multivariate explorative Analyse (Teil 1)
5. Multivariate explorative Analyse (Teil 2)
6. Dimensionsreduktion (Teil 1): PCA
7. Dimensionsreduktion (Teil 2): MDS, Isomap
8. Clustering: K-Means, HCA
9. Clustervalidierung

10. Probeklausur

11. Feature Engineering, Datengetriebene Modellierung (Teil 1)
12. Datengetriebene Modellierung (Teil 2)

Überblick / Begriffe

Explorative Analyse (EDA)

Feature Engineering & Modellierung

Wdh | Explorative Datenanalyse

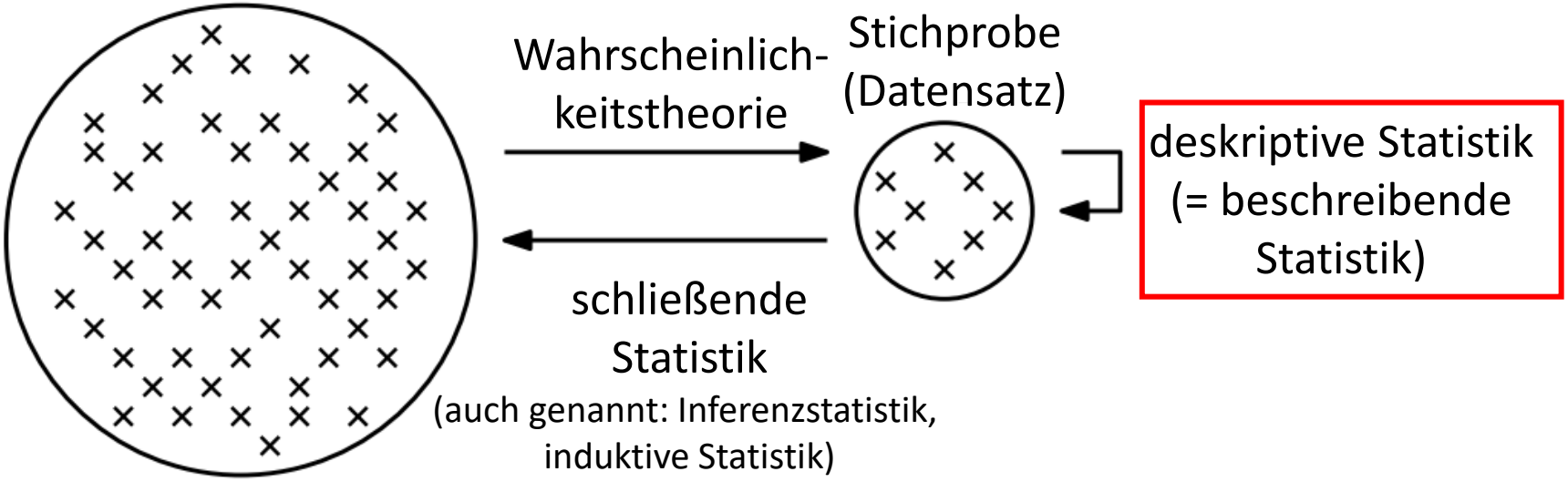
… ist die Erkundung von Daten mit folgenden Zielen:

1. Identifikation von Problemen im Datensatz
2. Prüfung, ob initiale Fragen beantwortbar sind
3. Erzeugung erster Antwortskizzen
4. **Erzeugung neuer Fragen / Hypothesen**

Typische Werkzeuge:

- Erwartungshaltung (Fragen formulieren)
- Deskriptive Statistik
- Visualisierung
- Dokumentation des Erkenntnisweges

Wdh | Explorative Datenanalyse



Wdh | Explorative Datenanalyse | Deskriptive Statistik

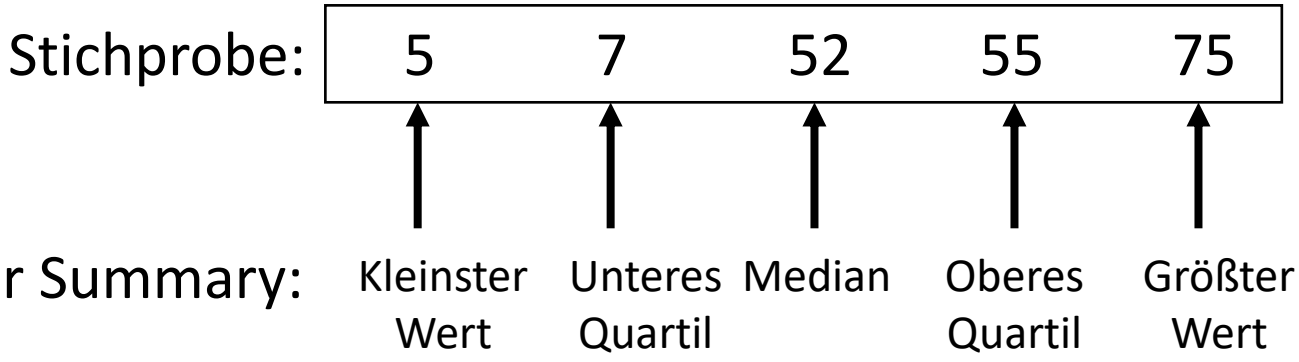
Kennzahlen (englisch: Summary Statistics)
(auf Deutsch auch genannt: aggregierende Parameter, Maßzahlen)

- beschreiben eine Stichprobe (Häufigkeitsverteilung) in wenigen Zahlen

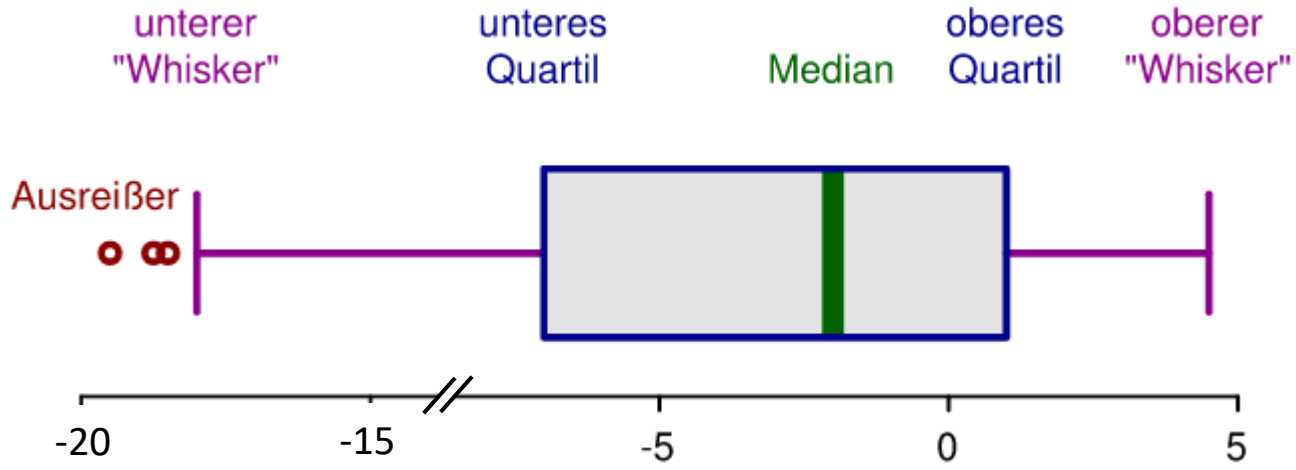
Typ	**Verwendung**	**Beispiele**
Lageparameter	beschreiben zentrale Tendenz der Stichprobe (z.B. wo die meisten Werte der Stichprobe sind)	▪ Modus (= häufigster Wert) ▪ Median ▪ Quartile ▪ p-Quantile ▪ arithmetisches Mittel
Streuungsparameter	beschreiben Streubreiten von Stichprobe	▪ Varianz, ▪ Standardabweichung ▪ Interquartilsabstand

Explorative Datenanalyse | Deskriptive Statistik

„5 Number Summary“ (*5-Punkte-Zusammenfassung)*
nach John W Tukey



Tukey Boxplot



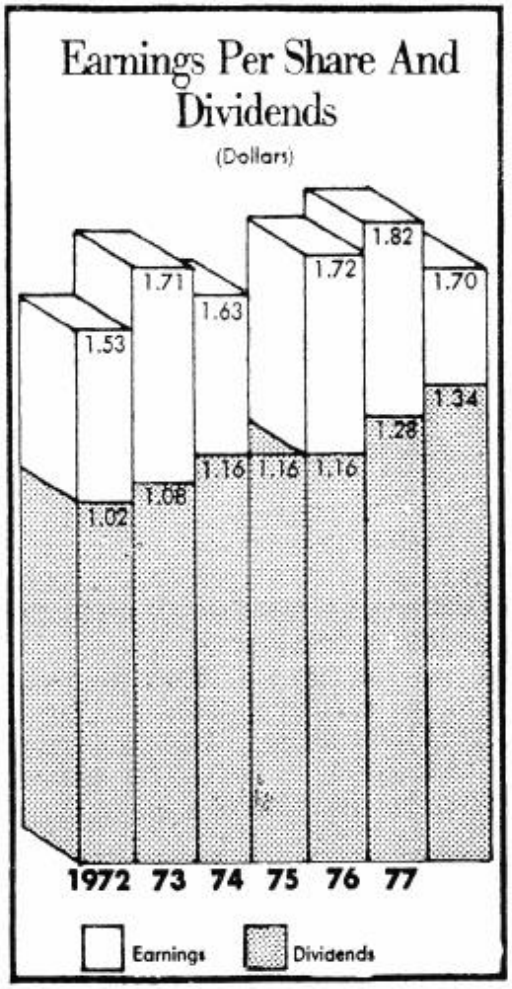
Wdh | Deskriptive Statistik

- Summary Statistics (Kennzahlen) hilfreich zur ersten Charakterisierung von Daten
- Aber: *Ähnliche Kennzahlen für unterschiedliche Datensätze*
  → Wir benötigen Visualisierung für die Erkundung von Daten!

1) Matejka J, Fitzmaurice GW. PCHFCS, 1290-1294, 2017, DOI: 10.1145/3025453.3025912

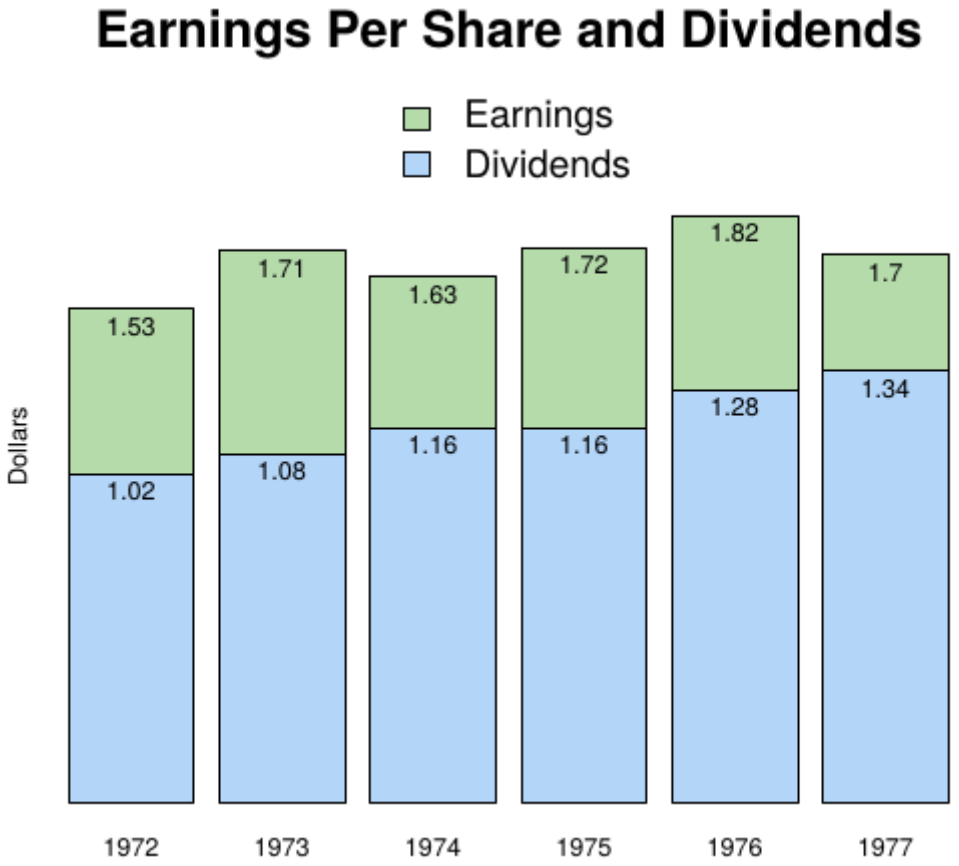
Wdh | Visualisierung

Beispiele für problematische Abbildungen



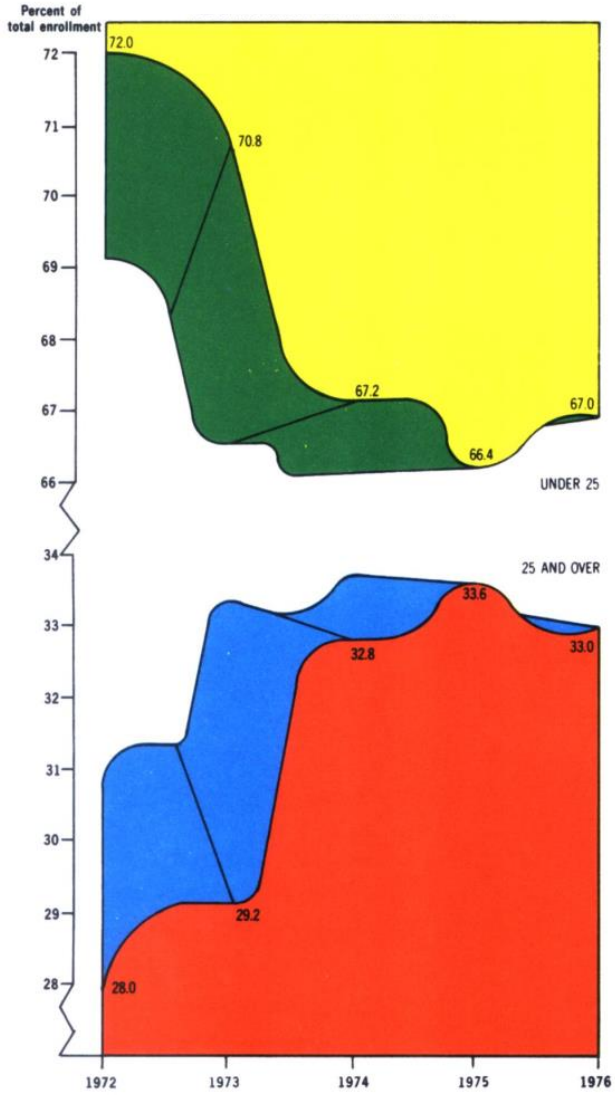
The Washington Post, 1979

besser



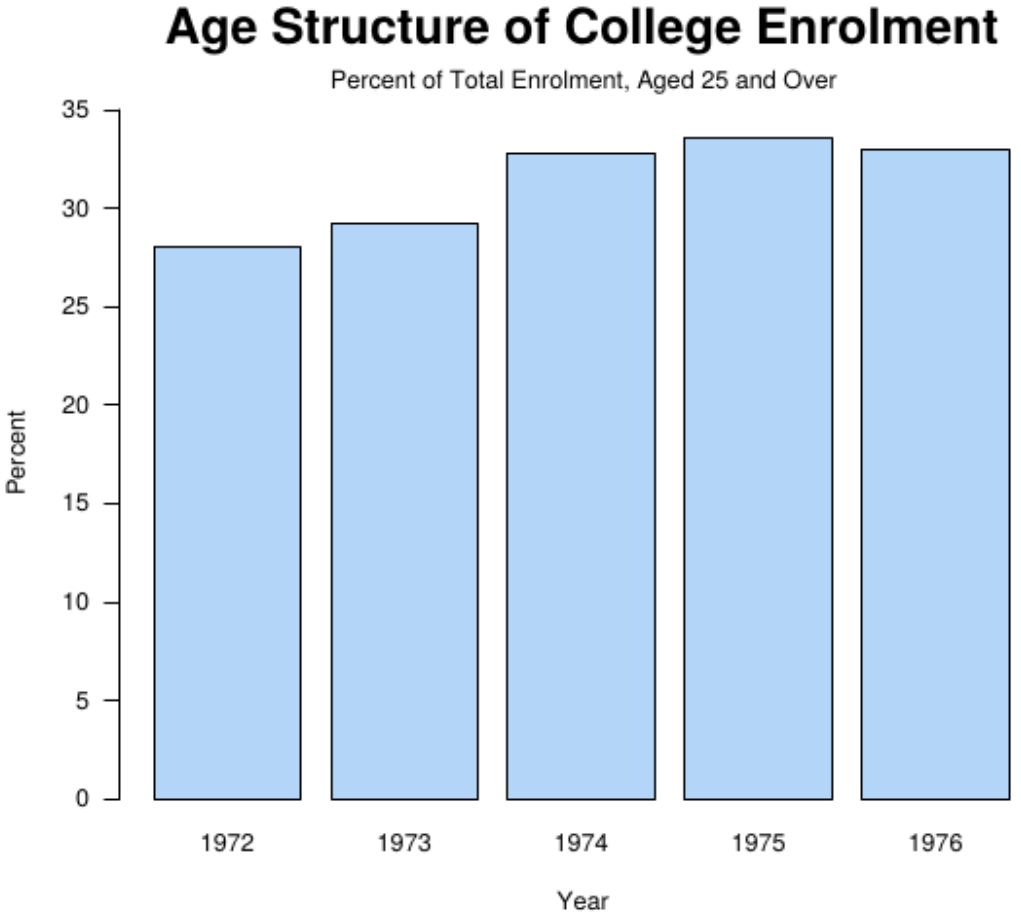
Wdh | Visualisierung

Age Structure of College Enrolment (1972-76)



American Education Magazine

besser

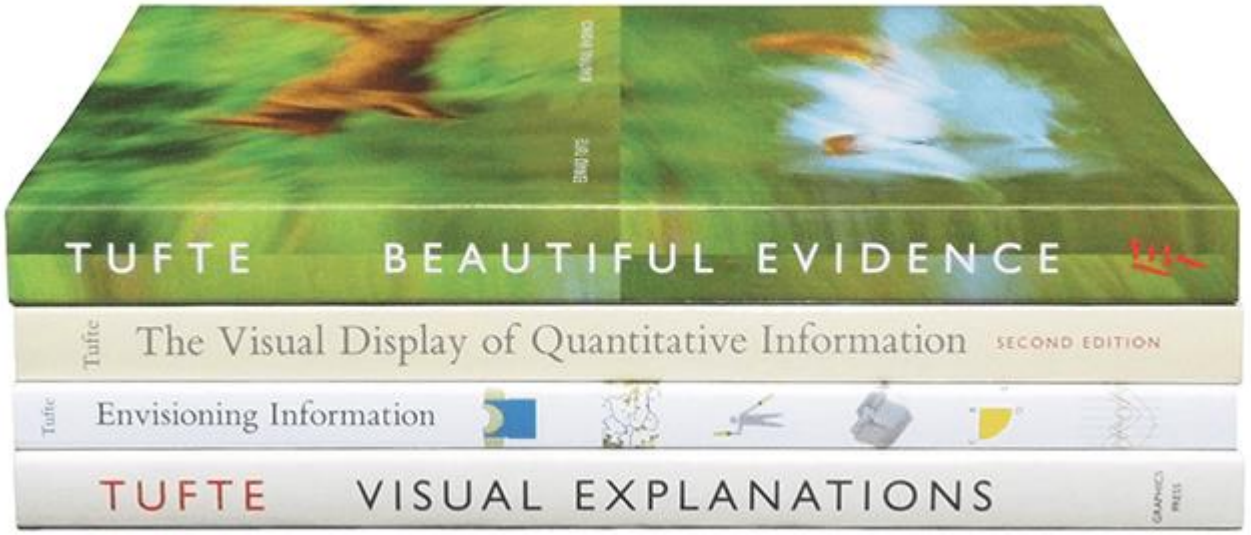


Explorative Analyse | Visualisierung

Edward Tufte

- US amerikanischer Statistiker
  (seit 1999 Professur an der Yale University,
  vorher Princeton)
- Pionier im Bereich Datenvisualisierung
- *Chartjunk*: visuelle Elemente in Abbildungen, die unnötig sind oder das Verständnis erschweren

Klassiker



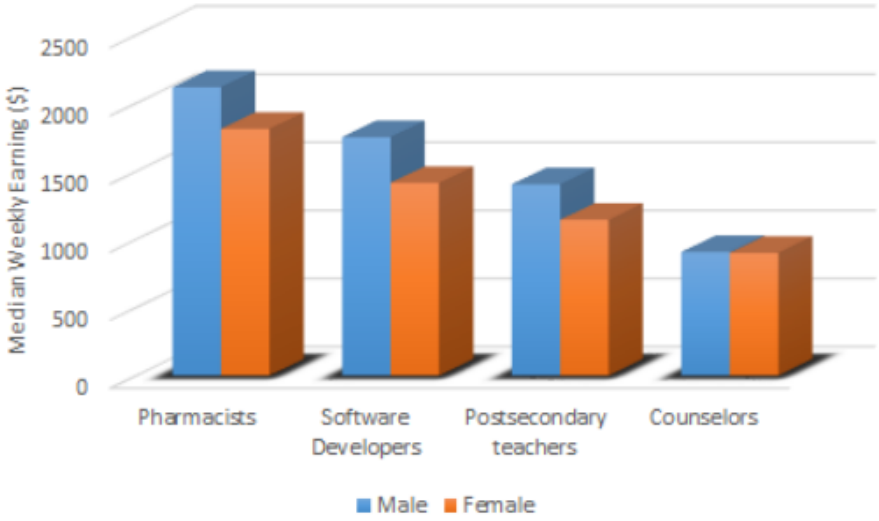
Explorative Analyse | Visualisierung

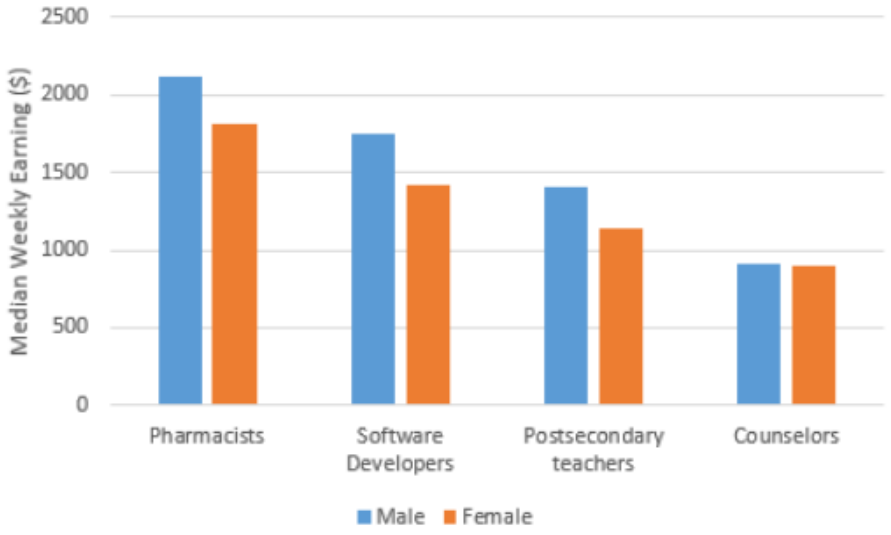
Prinzipien guter Visualisierung (nach Edward Tufte)

1. Maximieren Sie das Daten-Tinte Verhältnis (*data-ink ratio*)

$$q = \frac{\text{Tintenmenge für die Darstellung von Daten}}{\text{Gesamttintenmenge für die Darstellung der Abbildung}}$$

auf Deutsch: sowenig Tinte für so viele Daten wie möglich





Explorative Analyse | Visualisierung

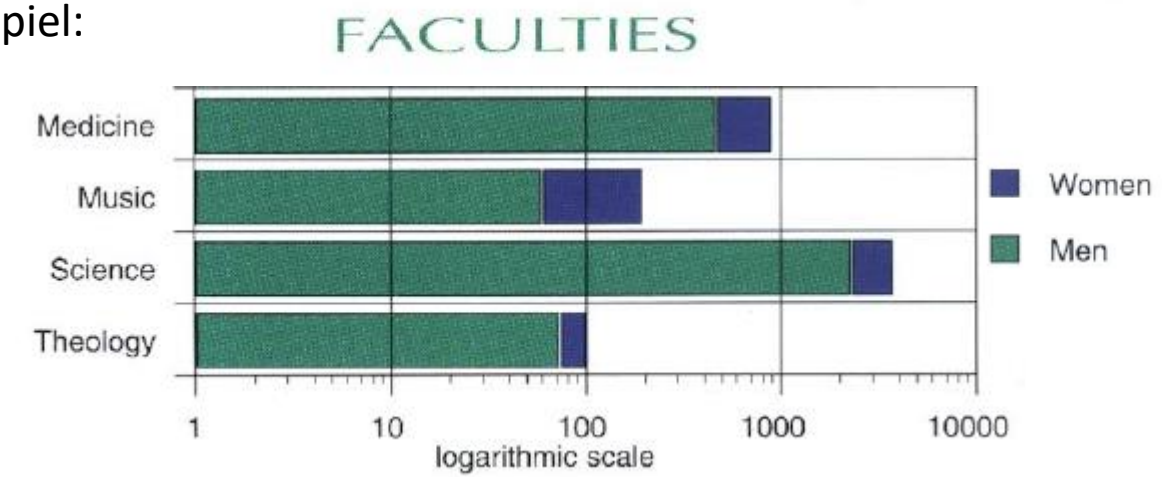
Prinzipien guter Visualisierung (nach Edward Tufte)

2. Minimieren Sie den Lügenfaktor (*lie factor*)

$$l = \frac{\text{Größe eines Effekts in der Abbildung}}{\text{Größe des Effekts in den Daten}} \qquad (*)$$

auf Deutsch: Leiten Sie Ihre Publikum nicht fehl.

Negativ-Beispiel:



(logarithmische Darstellung suggeriert kleineren Frauenanteil)

(*) Die „Gleichung“ ist nicht berechnbar. Sie dient Ihnen nur als Gedächtnisstütze.

Explorative Analyse | Visualisierung

Prinzipien guter Visualisierung (nach Edward Tufte)

2. Minimieren Sie den Lügenfaktor (*lie factor*)

$$l = \frac{\text{Größe eines Effekts in der Abbildung}}{\text{Größe des Effekts in den Daten}} \qquad (*)$$

auf Deutsch: Leiten Sie Ihre Publikum nicht fehl.

Fragwürdige Praktiken:

- Mittelwerte ohne Standardabweichung zeigen
- Interpolierte Werte ohne tatsächliche Daten zeigen
- Seitenverhältnis manipulieren, um Daten zu dramatisieren (siehe Vorlesung 3; besser: Goldener Schnitt als Seitenverhältnis)
- Koordinatenursprung nicht zeigen
- *Tick Labels* bei numerischen Achsen nicht darstellen
- Achsen nicht oder irreführend beschriften

(*) Die „Gleichung“ ist nicht berechenbar. Sie dient Ihnen nur als Gedächtnisstütze.

Explorative Analyse | Visualisierung

Prinzipien guter Visualisierung (nach Edward Tufte)

3. Minimieren Sie *Chartjunk*

   Chartjunk: visuelle Elemente in Abbildungen, die unnötig sind oder das Verständnis erschweren

   Typisch: unnötige 3D Effekte

4. Nutzen Sie angemessene Skalen und Beschriftungen.
   z.B. Beschriftung der Achsen

5. Farbe zum Darstellen von Eigenschaften Ihrer Daten – und nicht, um eine künstlerische (ästhetische) Aussage zu machen.

Weitere Hinweise zu Visualisierung finden Sie unter anderem im Paper „Ten Simple Rules for Better Figures“ von N.P. Rougier et al, PLOS Comput. Biol. 10, e1003833, 2014

Explorative Analyse | Visualisierung

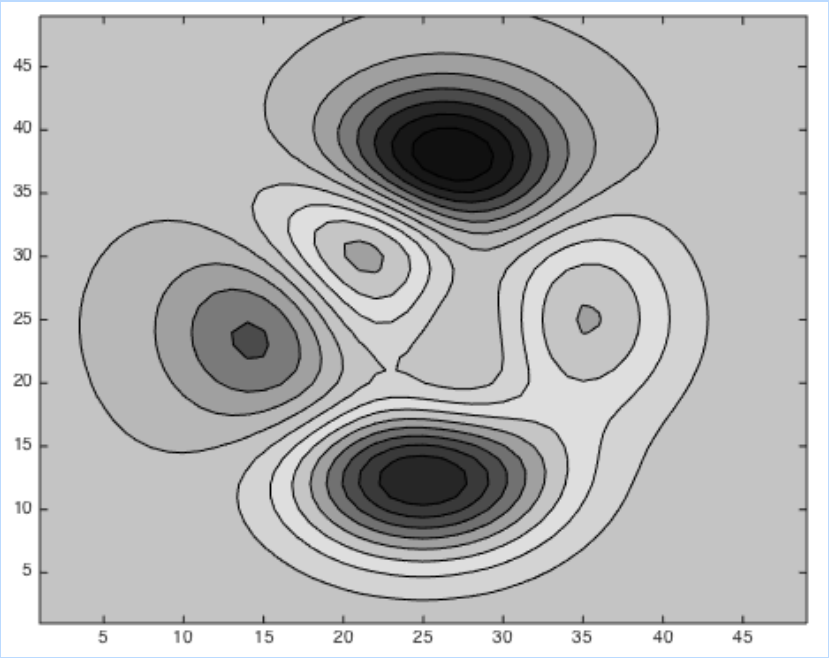
Aktivität

Arbeiten Sie mit Ihrem Banknachbarn zusammen.

1. Betrachten Sie den folgenden Farbverlauf namens *„Jet“*

niedrig  hoch

Die rechts stehende Abbildung wurde mit dem Farbverlauf erzeugt und dann in Graustufen ausgedruckt.



2. Identifizieren Sie: Wo sind die Minima, wo die Maxima in der rechten Abbildung?

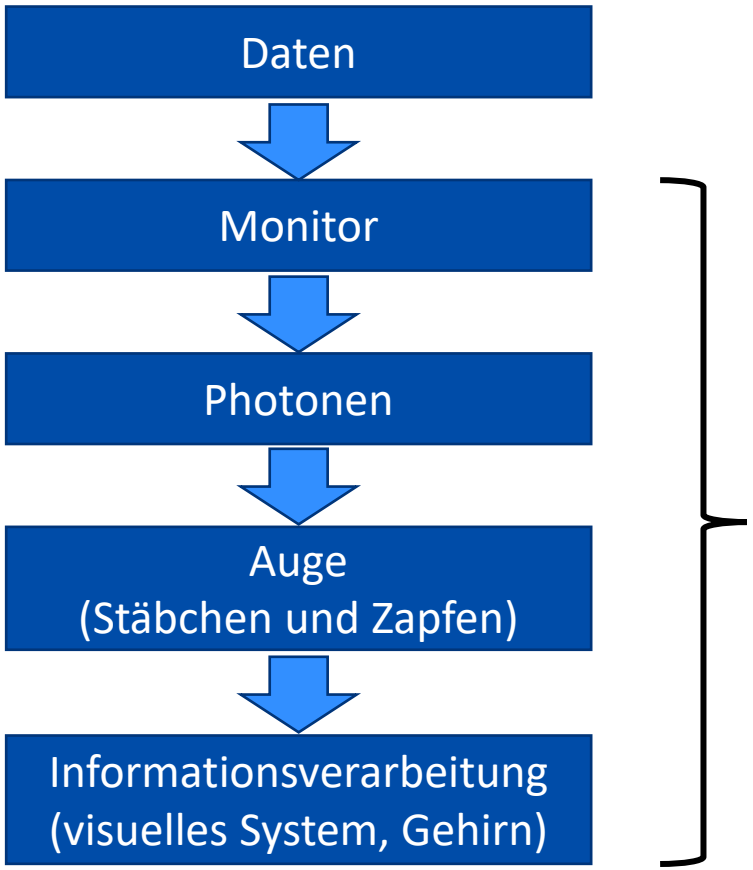
3. Sammeln Sie: Was finden Sie problematisch?

Explorative Analyse | Visualisierung

Explorative Analyse | Visualisierung

Explorative Analyse | Visualisierung

Warum sehen Sie mit Farbverlauf *Jet* (und einigen anderen) vermeintliche Eigenschaften in den Daten? → menschliche **Farbwahrnehmung**



Was wir eigentlich wollen:
Gleiche Unterschiede zwischen Datenpunkte sollen sich in gleichen *perzeptuellen (wahrnehmbaren)* Unterschieden in den Farben widerspiegeln.

Modell f der menschlichen Farbwahrnehmung.

Beispiel:
Farberscheinungsmodell „CAM 02 UCS“[1] der internationalen Beleuchtungskommission *CIE* in Wien

1) Moroney et al, “The CIECAM02 color appearance model“, 2002

Explorative Analyse | Visualisierung

Kernidee:

- Gleiche *wahrnehmbare* Farbunterschiede für gleiche Unterschiede in den Daten.

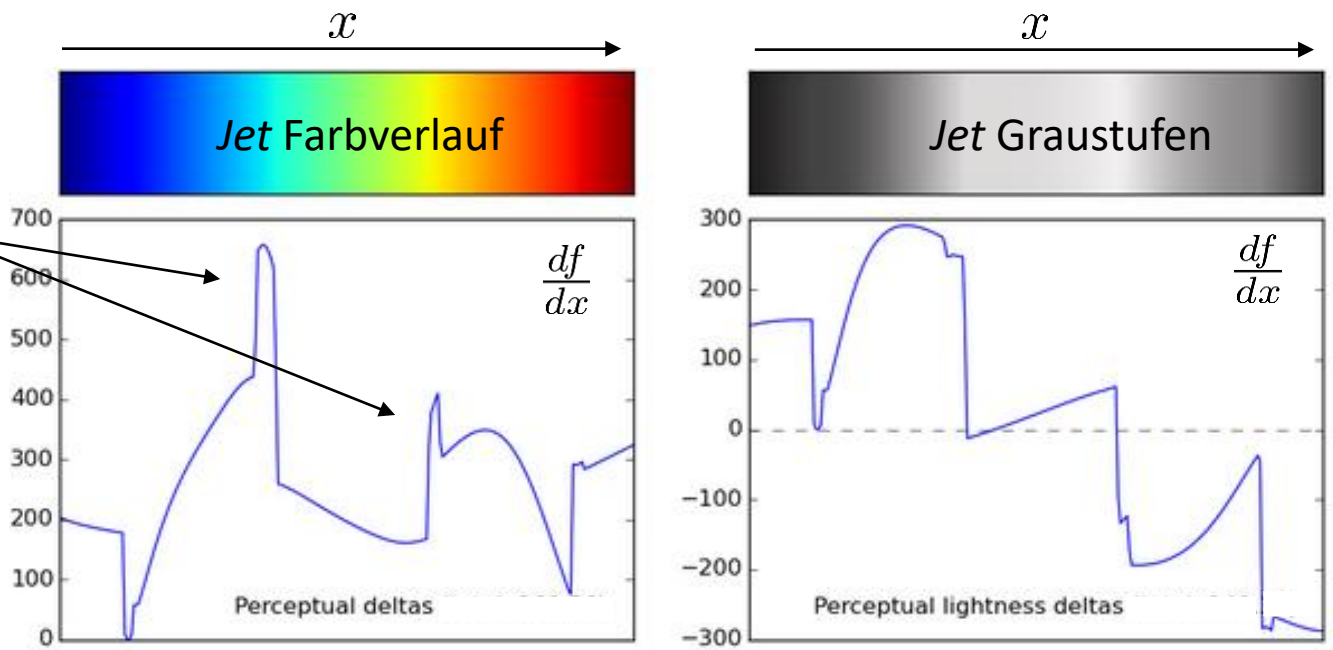
$$\frac{\Delta f}{\Delta x} = \frac{df}{dx} \overset{!}{=} \text{const}$$

Beispiel:

- Daten $\mathbf{x} = [0.1, 0.2, 0.8, 0.9]$
- Unterschiede $\Delta x^{(1)} = x_2 - x_1 = 0.1$ und $\Delta x^{(2)} = x_4 - x_3 = 0.1$ sind identisch und sollen zum identisch großen Farbunterschied Δf führen.

Peaks bei cyan und gelben Bereichen

Beobachtung:
$\frac{df}{dx} \neq \text{const}$



Explorative Analyse | Visualisierung

Kernidee:

- Gleiche *wahrnehmbare* Farbunterschiede für gleiche Unterschiede in den Daten.

$$\frac{\Delta f}{\Delta x} = \frac{df}{dx} \overset{!}{=} \text{const}$$

Beispiel:

- Daten $\mathbf{x} = [0.1, 0.2, 0.8, 0.9]$
- Unterschiede $\Delta x^{(1)} = x_2 - x_1 = 0.1$ und $\Delta x^{(2)} = x_4 - x_3 = 0.1$ sind identisch und sollen zum identisch großen Farbunterschied Δf führen.

Beobachtung:

$$\frac{df}{dx} = \text{const}$$

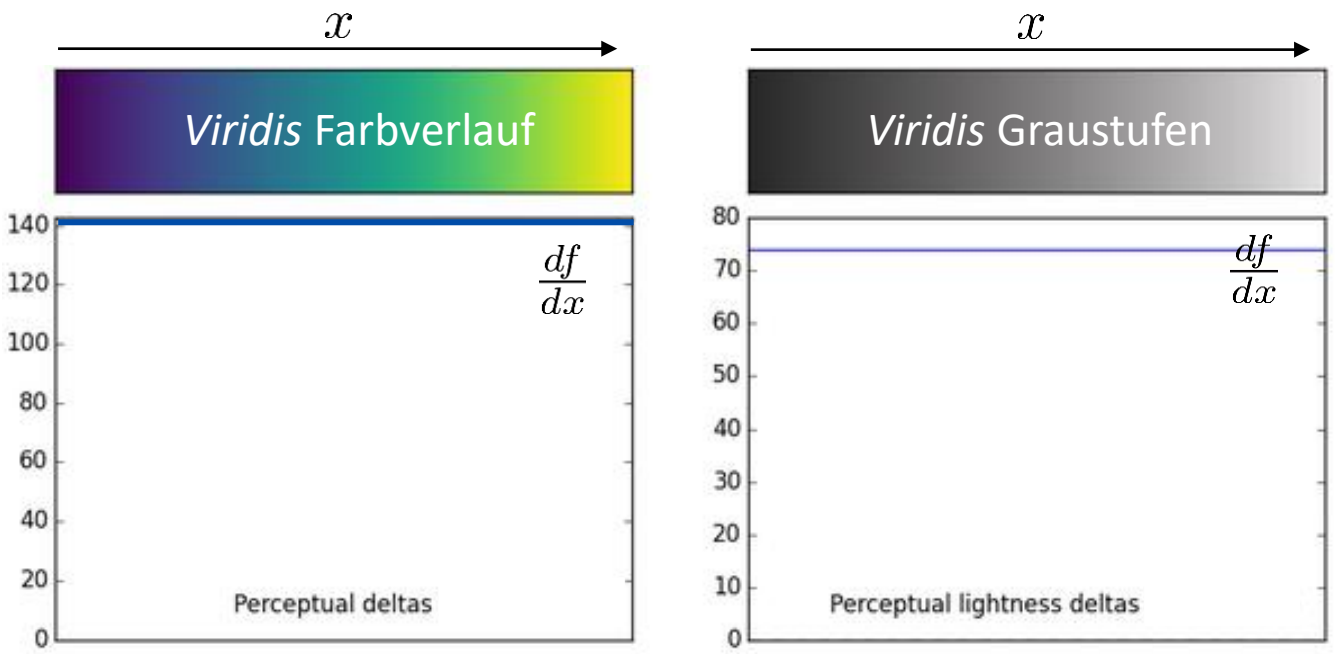


 Abbildung: Van der Walt S, Smith N. Berkely University of California, Data Science Institute, bids.github.io/colormap

Explorative Analyse | Visualisierung

F

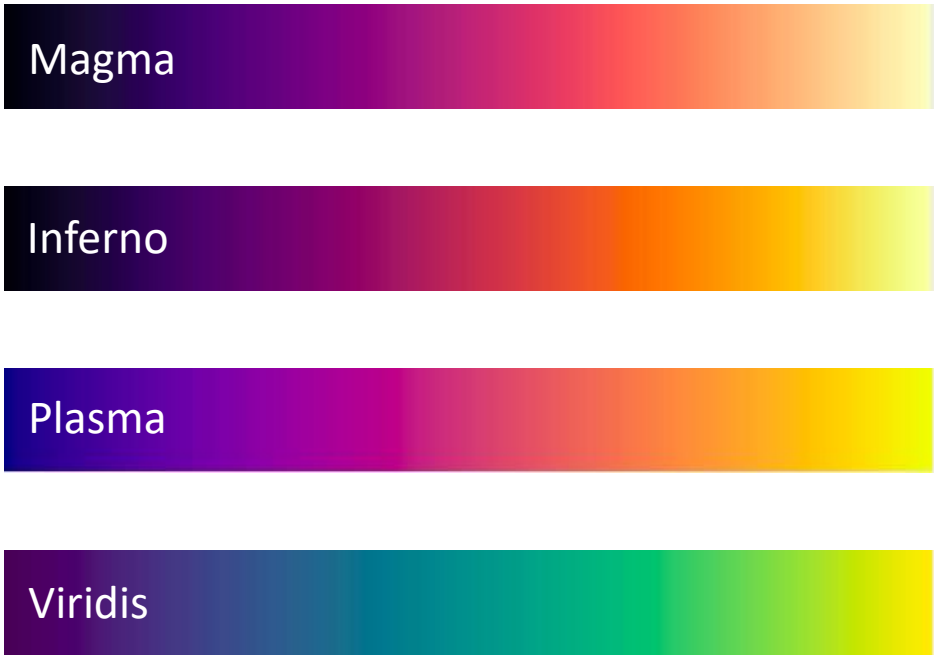
Viridis und andere Farbverläufe:

- optimiert hinsichtlich Farbwahrnehmung
- nutzbar auch bei Graustufendruck
- optimiert hinsichtlich verschiedener Farbfehlsichtigkeiten (beispielsweise Rot/Grün Sehschwäche)

Frage

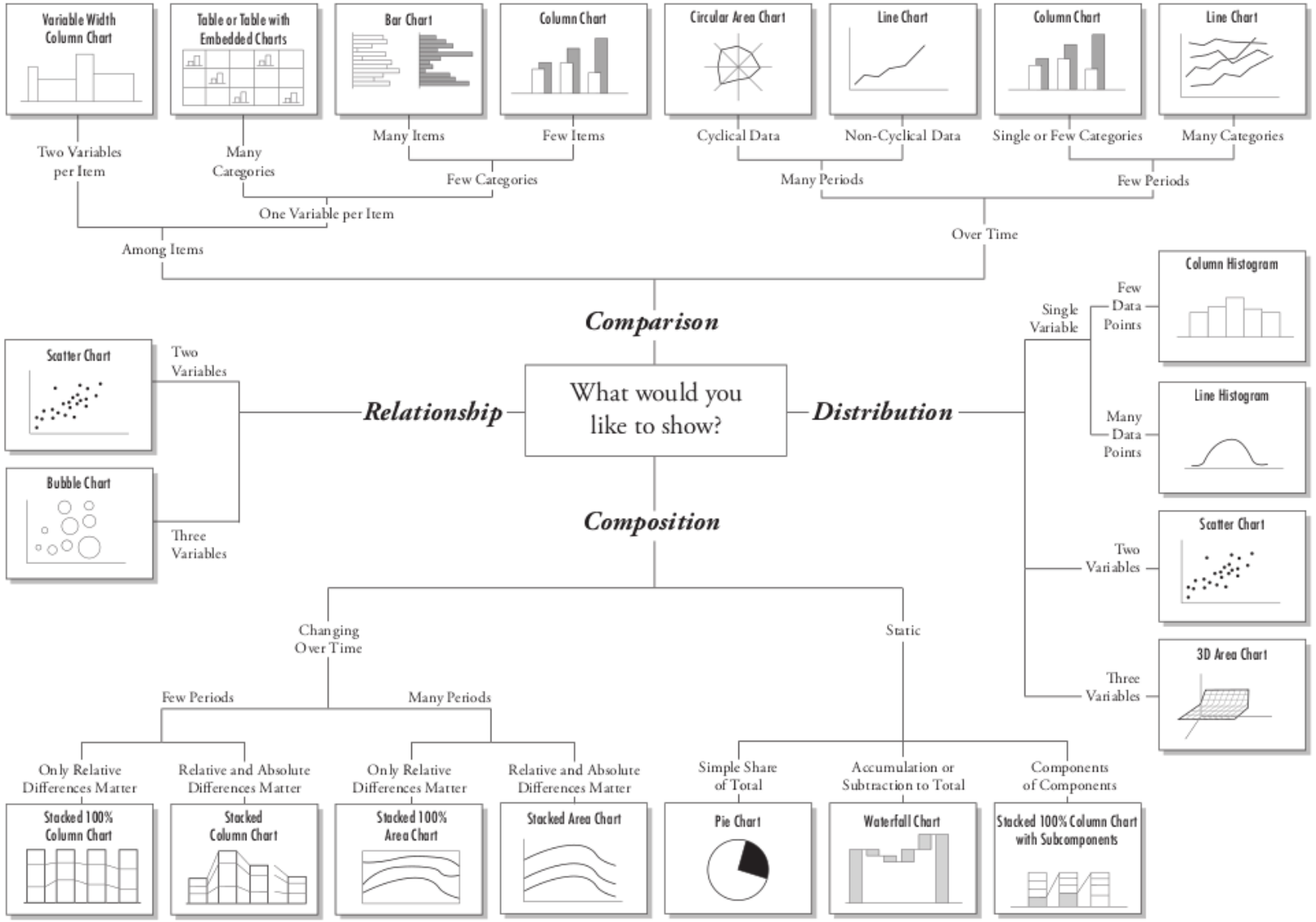
- Wieviel Prozent aller Menschen haben eine Form der Farbfehlsichtigkeit?

- Wieviel Prozent aller Menschen haben eine Rot/Grün Farbfehlsichtigkeit?



1) Kalloniatis M, Luu C (2007). "The Perception of Color". In Kolb H et al (eds.). The Organization of the Retina and Visual System

Explorative Analyse | Visualisierung



Multivariate explorative Analyse

Multivariate Explorative Analyse

Bisher:

- Methoden der *univariaten* deskriptiven Statistik und explorativen Analyse (Kennzahlen für Lage und Streuung, Boxplots, …)
- *univariate* Daten: eindimensionale Daten (=ein Merkmal bzw. Feature)

Jetzt:

- Methoden der *multivariaten* deskriptiven Statistik und explorativen Analyse
- *multivariate* (= mehrdimensionale) Daten – also Daten, die mehr als ein Feature (Merkmal) enthalten.

Beispiel:

- Charakterisierung von Mietpreisen für Wohnungen (univariat)
- Charakterisierung von Abhängigkeiten zwischen Mietpreisen und Zimmeranzahl (multivariate Beschreibung)

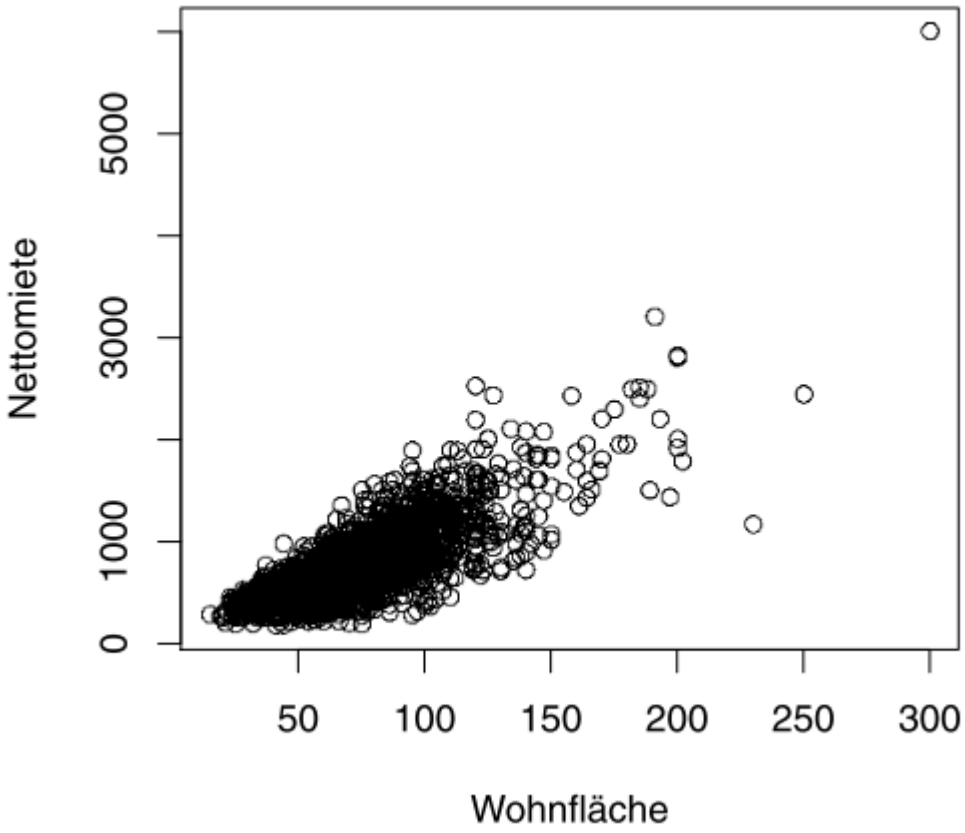
Multivariate Explorative Analyse

Scatterplot (Streudiagramm)

- Darstellung der Wertepaare (x_i, y_i) zweier Features (Merkmale) X, Y
  → bivariate Darstellung

- Erleichtert Identifikation von Abhängigkeitsstrukturen zwischen Merkmalen

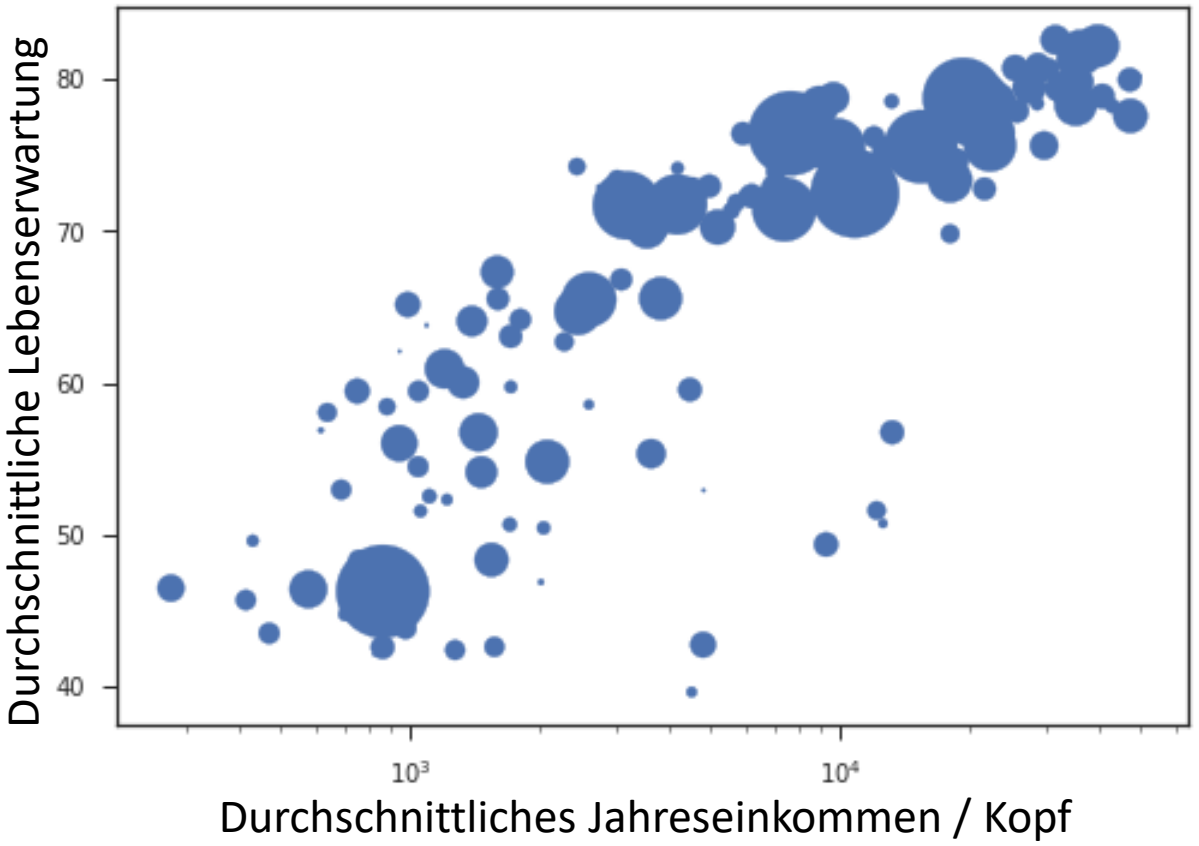
Daten: Münchener Mietspiegel (2015)



Multivariate Explorative Analyse

Bubble Chart (Blasendiagramm)

- Darstellung der Wertetupel (x_i, y_i, z_i) dreier Features (X, Y, Z)
  → trivariate Darstellung



Kreisflächen: proportional zur Bevölkerungsanzahl im jeweiligen Land

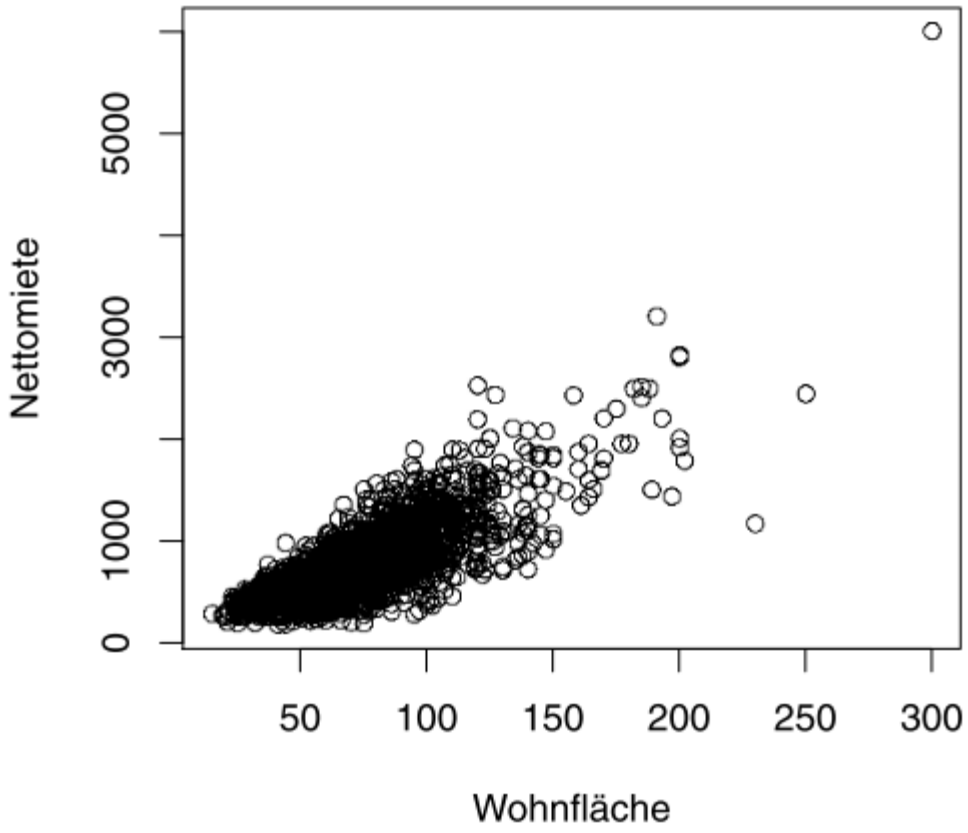
M‘variate Explorative Analyse | Zusammenhangsmaße

Charakterisierung von Zusammenhängen durch Zusammenhangsmaße (Interdependenzmaße)

Klassen von bivariaten Maßen:

1. Maße zur Quantifikation der **Stärke** eines Zusammenhangs
2. Maße zur Charakterisierung der **Richtung** eines Zusammenhangs
3. Maße zur Charakterisierung der **Direkt-/Indirektheit** eines Zusammenhangs [meist tri- oder n-variate (n≥3) Maße]

Daten: Münchener Mietspiegel (2015)



M‘variate Explorative Analyse | Zusammenhangsmaße

Bivariate Maße zur Quantifikation der **Stärke** eines Zusammenhangs

1. Pearsons Korrelationskoeffizient
   Charakterisierung der Stärke linearer Zusammenhänge

2. Spearmans Korrelationskoeffizient
   Charakterisierung der Stärke monotoner Zusammenhänge

... und hunderte weitere Maße aus verschiedenen Domänen
(von der Chaosforschung bis hin zur Charakterisierung
feuernder Neuronen)

M‘variate Explorative Analyse | Korrelationskoeffizient

Pearsons Korrelationskoeffizient

Seien *X* und *Y* Merkmale (Features). Seien (x_i, y_i) die Werte der Merkmale für eine Stichprobe der Größe *n,* und bezeichnen $\bar{x}, \bar{y}$ die jeweiligen empirischen Mittelwerte.

Dann ist der Pearson Korrelationskoeffizient *r* definiert als

$$r = r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\tilde{s}_{XY}}{\tilde{s}_X \tilde{s}_Y}$$

← Kovarianz

← Standardabweichungen der Stichproben

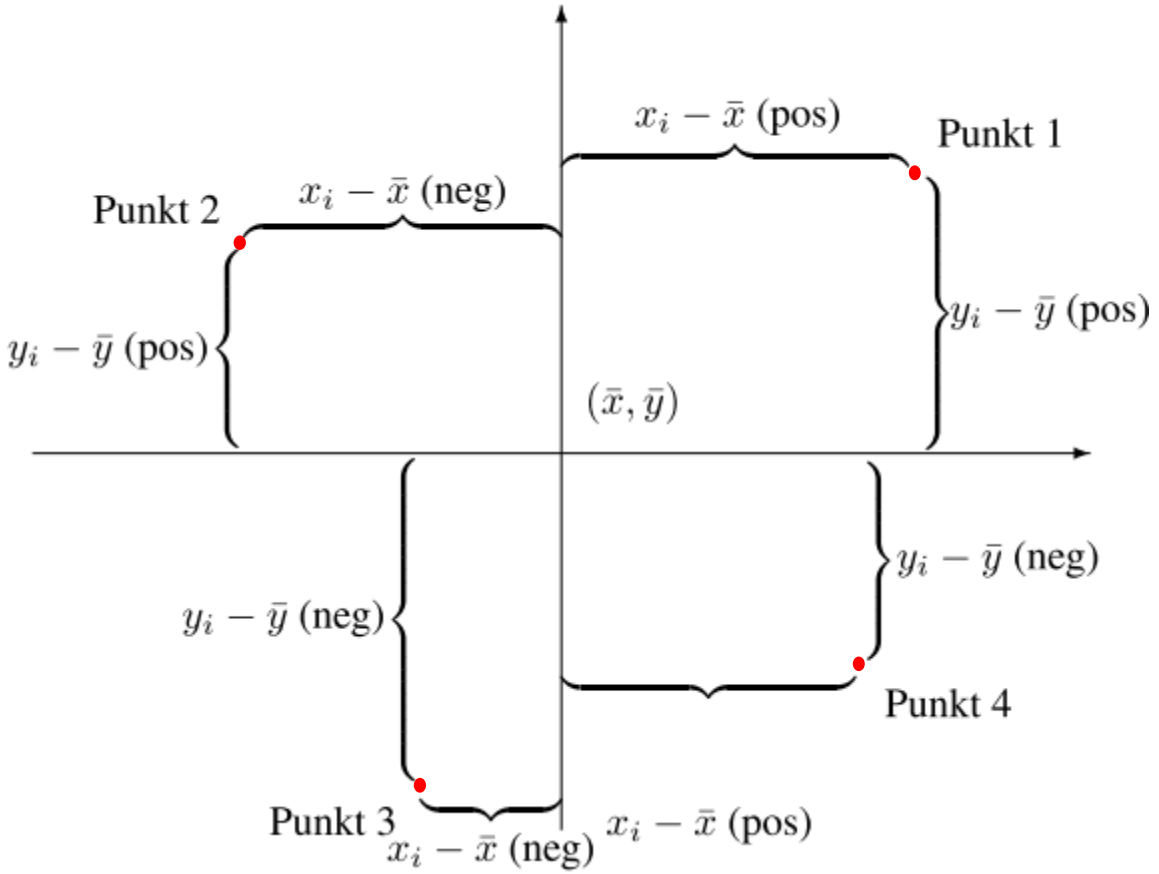
mit $\tilde{s}_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad \tilde{s}_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$

$$\tilde{s}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

M‘variate Explorative Analyse | Korrelationskoeffizient

Kovarianz $\tilde{s}_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{(x_i - \overline{x})(y_i - \overline{y})}$

Vorzeichen jedes Summanden ergibt sich aus Lage von (x_i, y_i) gegenüber $(\bar{x}, \bar{y})$



M‘variate Explorative Analyse | Korrelationskoeffizient

Bemerkungen:

1. *r* ist normiert: $-1 \leq r \leq 1$

2. *r* misst Stärke des linearen Zusammenhangs:

   Je näher die Messwerte an einer Geraden liegen,
   desto näher liegt *r* bei 1 (Gerade hat positive Steigung)
   bzw. bei -1 (Gerade hat negative Steigung)

3. schwache Korrelation: $|r| < 0.5$
   mittlere Korrelation: $0.5 \leq |r| < 0.8$
   starke Korrelation: $|r| \geq 0.8$

M‘variate Explorative Analyse | Korrelationskoeffizient

Frage
Welche ungefähren Werte für den Korrelations-koeffizienten erhalten Sie in nebenstehenden Beispielen?

M‘variate Explorative Analyse | Zusammenhangsmaße

Bivariate Maße zur Quantifikation der **Stärke** eines Zusammenhangs

1. Pearsons Korrelationskoeffizient
   Charakterisierung der Stärke linearer Zusammenhänge

2. Spearmans Korrelationskoeffizient
   Charakterisierung der Stärke monotoner Zusammenhänge

... und hunderte weitere Maße aus verschiedenen Domänen
(von der Chaosforschung bis hin zur Charakterisierung
feuernder Neuronen)

M‘variate Explorative Analyse | Korrelationskoeffizient

Spearmans Korrelationskoeffizient

… entspricht dem Pearson Korrelationskoeffizienten, allerdings für die Ränge der Werte.

Seien die Werte x_i einer Stichprobe der Größe *n* des Merkmals *X* geordnet:

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

Dann ist der Rang eines Wertes definiert als:

$$rg(x_i) = i$$

Tritt ein und derselbe Wert mehrfach auf, dann spricht man von *Tie* bzw *Bindung*. Der Rang aller identischen Werte wird mit dem Mittelwert der Ränge dieser Werte ersetzt.