CONSULTANT DATA SCIENCE

DATEN, HYPOTHESEN, STATISTIK…

Eva Schabedoth 12.8.-22.8.

Inhalt

10.01.20	Vorstellung; Organisatorisches Einführung / Überblick Berufsfeld Data Science Der idealtypische Researchprozeß Vorstellung des Übungsdatensatzes; ggf. STATISTICA
11.02.20	Recall: Mathematische Basics Univariate / bivariate Statistik Inferenzstatistik: Stichprobe, Stichprobenschätzer, Verteilungen
12.02.20	ggf. Fortsetzung/Wiederholung Inferenzstatistik: Signifikanz /Konfidenzintervalle/Hypothesentests
13.02.20	ggf. Fortsetzung/Wiederholung Data Cleansing, Fehlende Werte, Datenaufbereitung Vorbereitung/Aufgabe Projekttag
14.02.20	*Projekttag*
17.02.20	Wiederholungen; komplexere Datensätze, gemeinsame praktische Übungen
18.02.20 19.02.20	Multivariate Analyseverfahren Regression; Faktorenanalyse; Clusteranalyse, Diskriminanzanalyse
20.02.20	Offene Fragen; Vorbereitung/Aufgabe Projekttag Datenschutz
21.02.20	*Projekttag*

vermittelt einen Eindruck vom Berufsfeld des Data Scientist und den grundsätzlichen Anforderungen

bildet Sie NICHT umfassend zum Data Scientist aus

Setzt Grundkenntnisse der Statistik sowie Erfahrungen mit klassischer statistischer Datenanalyse voraus

ersetzt keine Ausbildung im Bereich Statistik und Datenanalyse

hilft Ihnen, entsprechende Wissenslücken zu erkennen und ggf. selbst zu schließen

bringt Ihnen hoffentlich nahe, dass Datenanalyse eine spannende Sache ist

Just do it.

Fragen? Fragen!

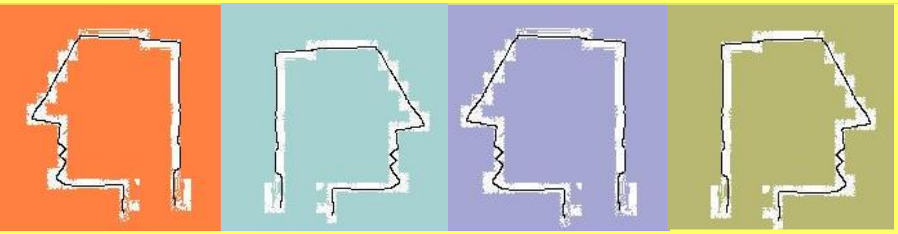
die selbstständigen Projekttage gut nutzen

sich mit den zur Verfügung gestellten Daten aktiv vertraut machen, mit ihnen „spielen“

Probleme aller Art sowie Verständnisschwierigkeiten möglichst *sofort* thematisieren, gerne auch privat:

eva.schabedoth@t-online.de

Spaß haben.



BIG DATA !

Oder:

Alles wird gut (eindeutig zu erklären, perfekt zu prognostizieren), wenn wir nur genug (alle!) Daten hätten...

Was sind denn „Daten“?

Über was oder wen machen sie welche Aussagen?

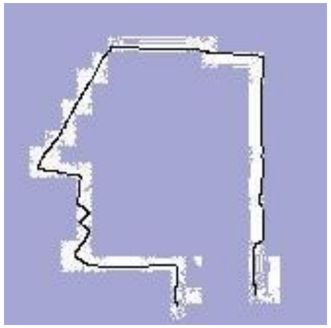
BIG DATA

07356 25 000000000000000111111110001111111000000000000000000001110001111111100001111111111110000
07356 03 1111111000000000000000000000000000000011100000000000000000000000000000000000000001000
07356 12 00000001111100000000000000000000000111000000000000000000011100000000011100000000000000
07356 07 00000000000011100000000000000000000000000000000011000000000000000000000000000000010000
07356 18 000000000000000000000010000000000000010000000000000000000100000000000000000000000001011

Teilnehmer-ID Sender-ID Zeit: Messung pro Sekunde

TV-Reichweitenmessung

ca. 5.100 Haushalte = ca. 11.500 Personen
pro Sekunde: Erfassung des eingeschalteten Programms
Rating (Quote) = 60 sec. konsekutiv



Andere klassische *Big Data*-Bereiche:

z.B.
Meteorologie / Klimaforschung
DNA-Sequenzierung
Börsendaten
Mikrozensus / Sozioökonomisches Panel

BIG DATA

```
07356 25 0000000000000001111111100011111110000000000000000000111000111111110000111111111110000
07356 03 1111110000000000000000000000000000011100000000000000000000000000000000000000001000
07356 12 0000000111110000000000000000000000111000000000000000000111000000000111000000000000000
07356 07 000000000000111000000000000000000000000000000001100000000000000000000000000000000010000
07356 18 0000000000000000000000100000000000000100000000000000000001000000000000000000000000001011
```

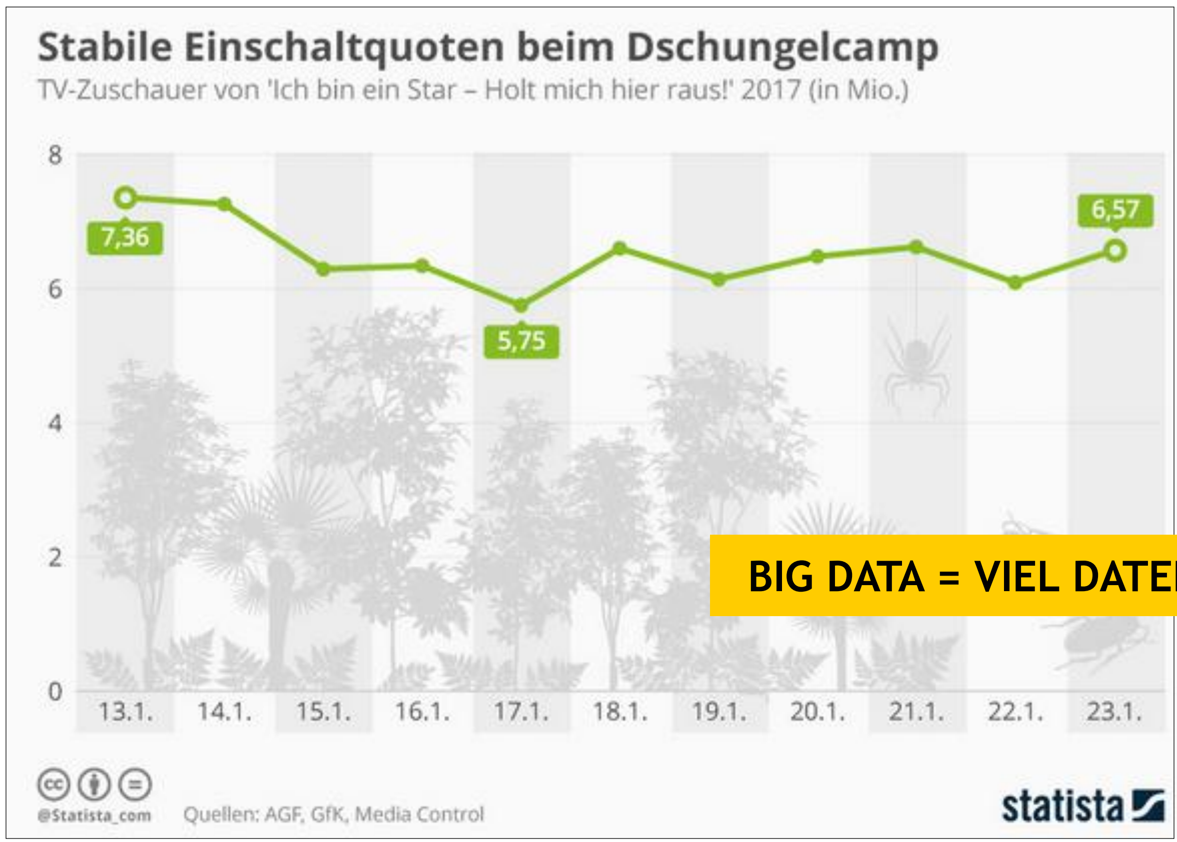
Teilnehmer-ID Sender-ID

Auswertung:

- **kumulierte Dauer**
- **Prozentuierung**
- **Zeitverlauf**

Reporting:

- **Tabelle**
- **Liniendiagramm**



BIG DATA = VIEL DATEN...

WHAT'S NEW..?

VOLUME

Laut IBM werden weltweit täglich 2,5 Trillionen Byte an Daten produziert.

Andere Quellen: ca. 9 Billionen Gigabyte in 2016

große Mengen an Daten mit geringer Dichte

VARIETY

z.B.

- Technische Messwerte
- Social Media-Inhalte
- Machine-to-Machine-Kommunikation
- Umfragedaten

VELOCITY

Permanenter Echtzeit-Datenstream

zunehmend Echtzeit-Verarbeitung und -analyse

Business Intelligence-Datenbestände
mobile Apps
Location-based Services
Cloud Computing
Web/Social Media
Sensoren
M-2-M-Kommunikation

z.B.

Zusammenführen
Selektieren
Bereinigen

QUELLEN → **BIG DATA** ← ANALYSE

verschiedene Arten strukturierter Daten (z.B. aus Datenbanken)

verschiedene Arten unstrukturierter Daten (z.B. Social Media-Inhalte)

Deskription
Mustererkennung
Inferenzanalyse
Prognose

Ein ganz besonderer Fall

„Unsere Risiko- und Investmentplattform **Aladdin®** verbindet skalierbar Portfolioanalyse und Risikomanagement mit einer vollständigen Handelsplattform. Auf **Aladdin®** werden derzeit Anlagen von über 14 Billionen USD für mehr als 160 Kunden und insgesamt 30.000 Investmentportfolios verwaltet – von **BlackRock**, aber auch von Wettbewerbern, Banken, Vorsorgeeinrichtungen und Versicherern.“

KI zur Abschätzung von Investmentrisiken und Entwicklung von Szenarien

vier Rechenzentren mit jeweils bis zu 6.000 Computern

Analyse globaler Wirtschaftsdaten, Börsenkursen und anderer potentieller Einflussfaktoren (z.B. politische Lagen, Wetterereignisse)

Außerdem: Daten von Firmen und Privatpersonen (z.B. Social Media-Aktivitäten, Parkplatzkameras von Firmen/Geschäften...)

Produkt, das ggf. auch für Fragestellungen außerhalb Investmentthemen einsetzbar ist /bereits eingesetzt wird

Neue industrielle Revolution

Verwandlung von Daten in ein Produkt

aktuell rund 1 Milliarde Umsatz in Deutschland

Wachstumschancen für Start Ups

Datenauswertung und -veredlung

Content Mining

Social Media Monitoring

Neue Art von Spezialisten

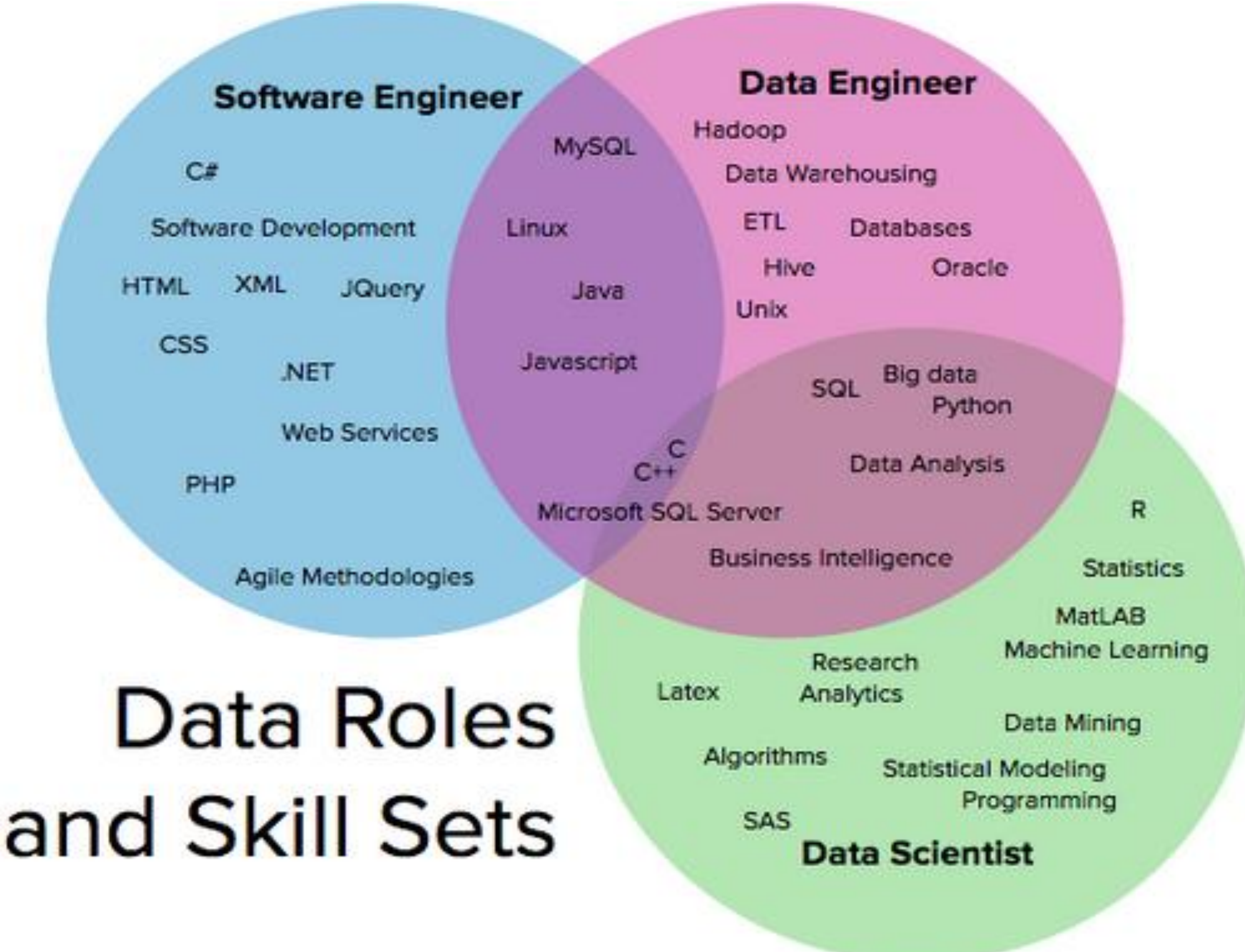
aktuell ca. 8.000 offene Stellen in diesem Bereich

neue Berufsbilder: Data Engineer, Data Analyst, *Data Scientist*

Durchschnittseinkommen €47.000-52.000,

je nach Unternehmen, Einstiegsalter, Berufserfahrung etc.

ROLLEN



Graphik nach Benjamin Aunkofer, CIO-Expertennetzwerk

Typisches Stellenangebot

Data Scientist (m/w/d)

Aufgaben:

Durchführung von qualitativen und quantitativen Analysen mit dem Ziel einer Optimierung der Geschäftsplanung und Prognose, z.B. durch die Entwicklung innovativer Modelle

Identifizieren und analysieren von relevanten internen und externen Datenquellen

Interpretation, Visualisierung und Kommunikation von Datenanalyseergebnissen sowie Ableitung von Handlungsempfehlungen

Dokumentation und Qualitätssicherung der implementierten Softwaresysteme und Prozesse

Typisches Stellenangebot

Anforderungen:

Abgeschlossenes Studium der Mathematik, Informatik oder eine vergleichbare Qualifikation mit statistischem Schwerpunkt

2-3 Jahre Berufserfahrung im Bereich Data Science / Business Analytics

Hohe Affinität zum Thema Data Science und Statistik sowie ausgeprägte strategische, analytische und konzeptionelle Fähigkeiten

Fundierte Software-/ Programmierkenntnisse im Bereich Statistik, Datenanalyse und -modellierung großer Datenmengen (z.B. SQL, Python, R oder SAS)

Erfahrungen mit stochastischen Methoden und statistischen Modellen

Idealerweise Kenntnisse in Big-Data-Infrastruktur und Cloud-basierten Analyseinstrumenten wie AWS, Azure, Google Cloud, Spark, Hadoop, AMLS oder BigQuery

Eigeninitiative sowie Kommunikationsstärke und Teamfähigkeit

Sehr gute Deutsch- sowie Englischkenntnisse

Typisches Stellenangebot

Benefits:

Attraktive Vergütung und umfangreiche Sozialleistungen nach Chemietarif

30 Tage Urlaub pro Jahr, Urlaubs- sowie Weihnachtsgeld

Betriebliche Altersvorsorge und weitere Mitarbeitervorteile (z.B. Eigenprodukte, Mitarbeiterrabatte, Auto-Leasing etc.)

Kantine mit Salatbar sowie kostenlose Bereitstellung von Kaffee, Wasser und frischem Obst

Kita

Große Auswahl an Betriebssportangeboten, gemeinsame Teilnahme an Sportevents, professionelles Rückentraining, Betriebsarzt und Gesundheitstage

Top-moderne Arbeitsbedingungen

Familiäre Arbeitsatmosphäre, flache Hierarchien, eigenverantwortliches Arbeiten und Freiraum für eigene Ideen

Elektrotankstelle

Gemeinsame Sommer- und Winterevents

DATA SCIENTIST

Bild: Randy Lao, Towards Data Science

The sexiest job of the 21st century!

Harvard Business Review

SKILLS

- Kommunikationskompetenz
- interdisziplinäres Denken
- detektivisches Denken
- Expertenwissen aus unterschiedlichen Bereichen
- Identifikation mit dem Unternehmen
- Statistik
- Mathematik
- Wirtschaftsinformatik
- Informatik

DATA SCIENTIST

Ein DATA SCIENTIST

ist eine Kombination aus Statistiker, Software-Entwickler und Datenanalyst

selektiert, generiert und analysiert Daten aus unterschiedlichsten Quellen

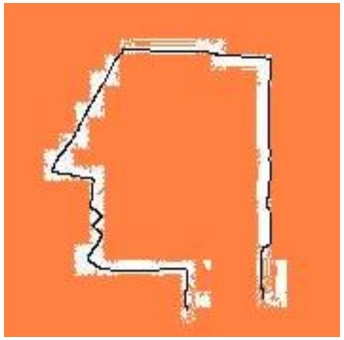
stellt Informationen für Unternehmen bereit, die es ihnen ermöglichen, effizienter zu arbeiten

ermittelt Verbesserungspotentiale und stellt Prognosen

entwickelt ggf. neue innovative Modelle und Analysetools

versteht die Erfordernisse unterschiedlichster Branchen

kann seine Ergebnisse klar und verständlich präsentieren



Ein DATA SCIENTIST

„versteht die Gegenwart und steuert die Zukunft."

KNOWLEDGE STACK

Business: Finance, HR, Customer
Engineering
Naturwissenschaften
Medizinwissenschaften

Statistik
Visualisierung
Deep/Maschine Learning

z.B. Apache Spark, Apache
Flink, Analyse-Software-Pakete

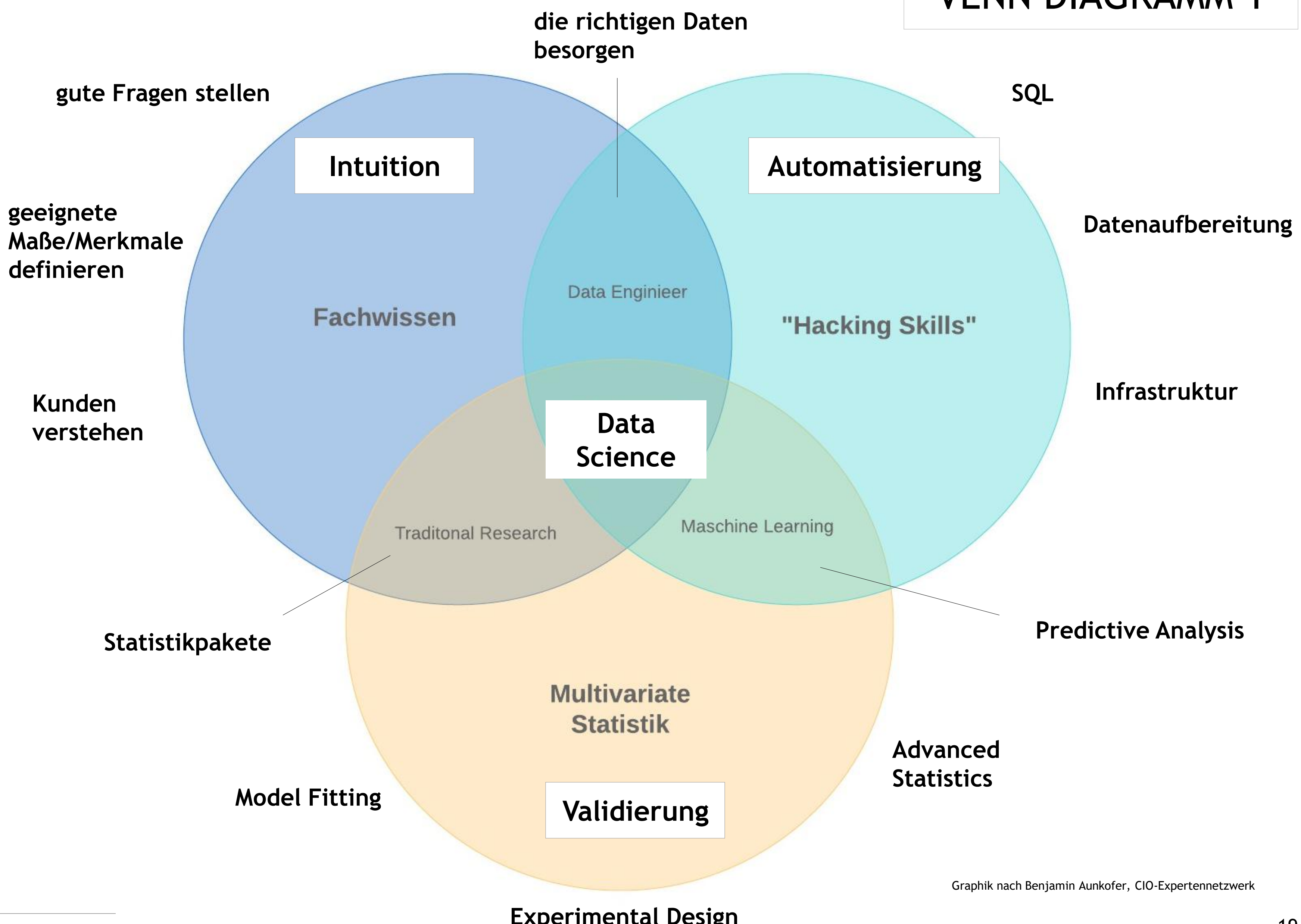
z.B. Python, R, Julia, C++, Java

Extrahieren, transformieren, laden
Netzwerk-Architektur
Datensicherheit

SQL / ERM / Normalisation
NoSQL
File Formate

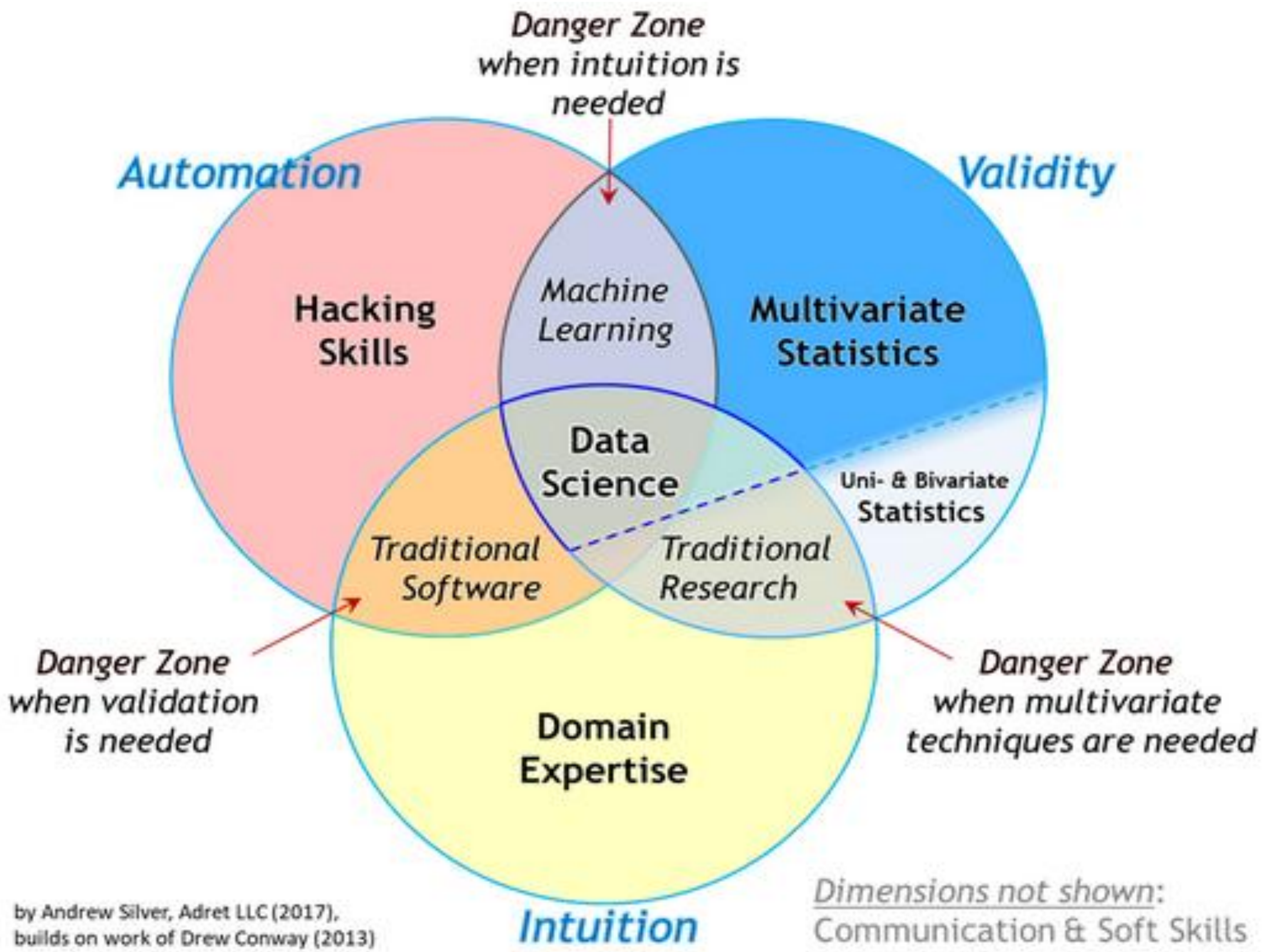
Graphik nach Benjamin Aunkofer, CIO-Expertennetzwerk

VENN DIAGRAMM 1

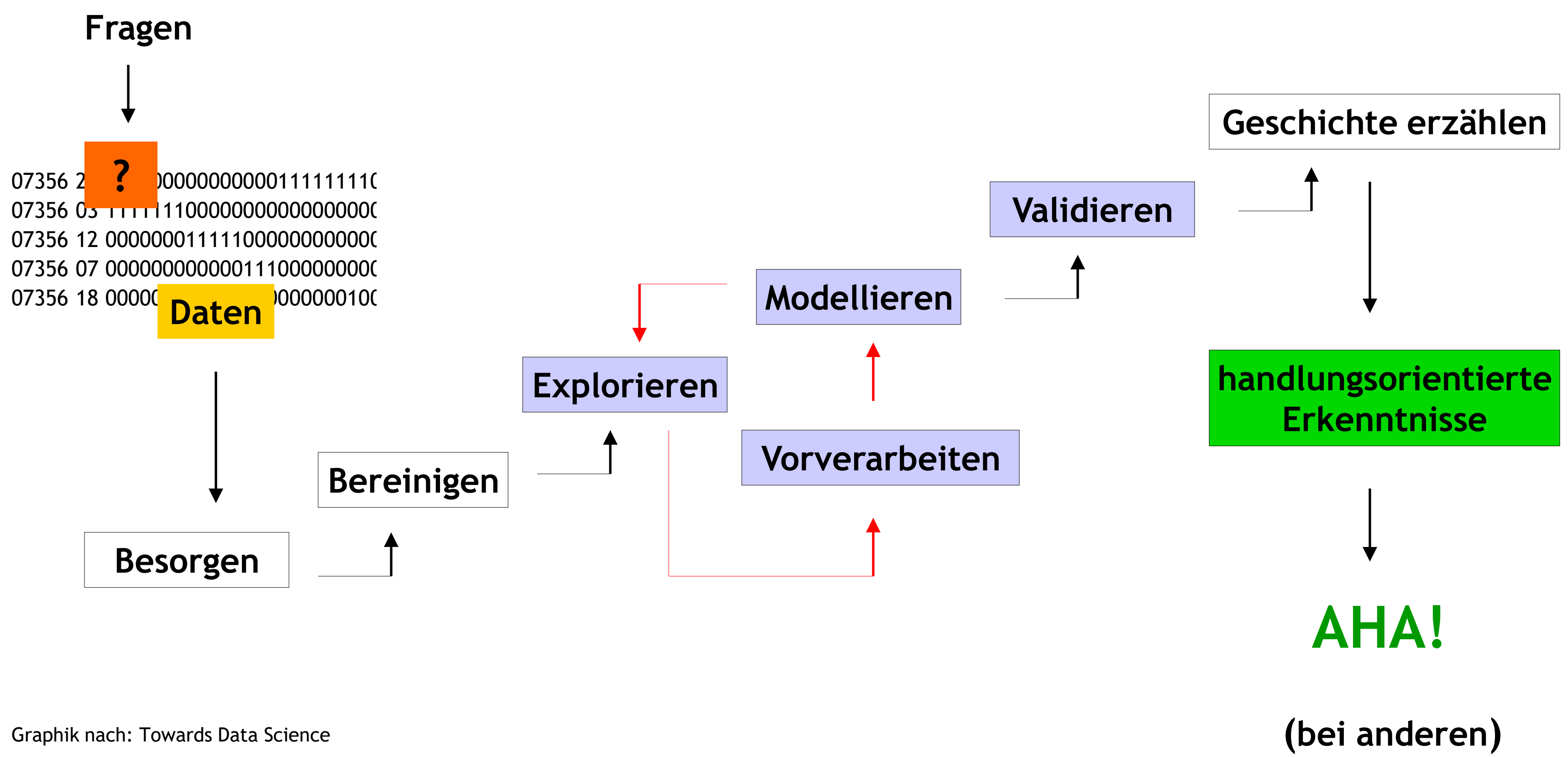


Graphik nach Benjamin Aunkofer, CIO-Expertennetzwerk

VENN DIAGRAMM 2

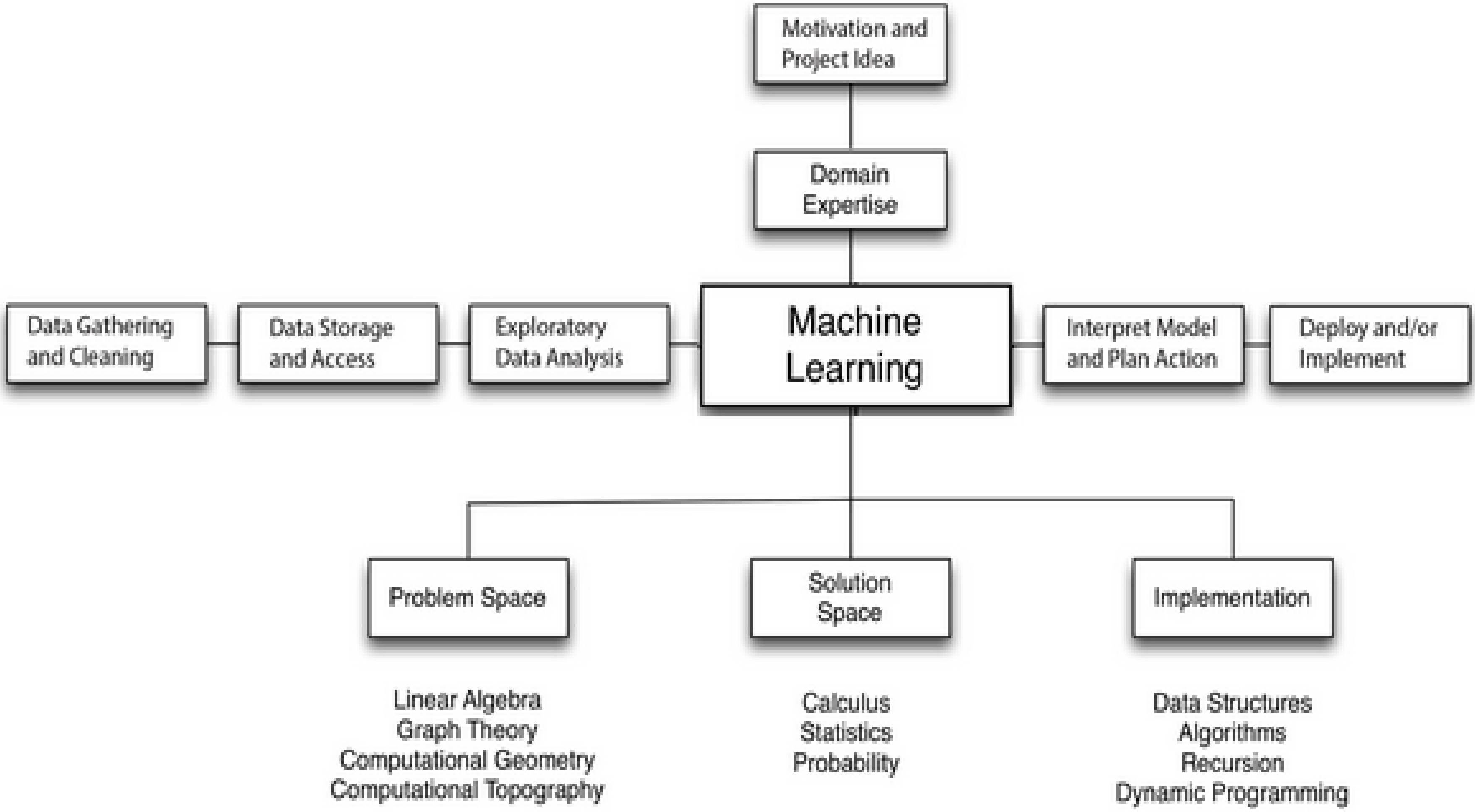


PIPELINE 1



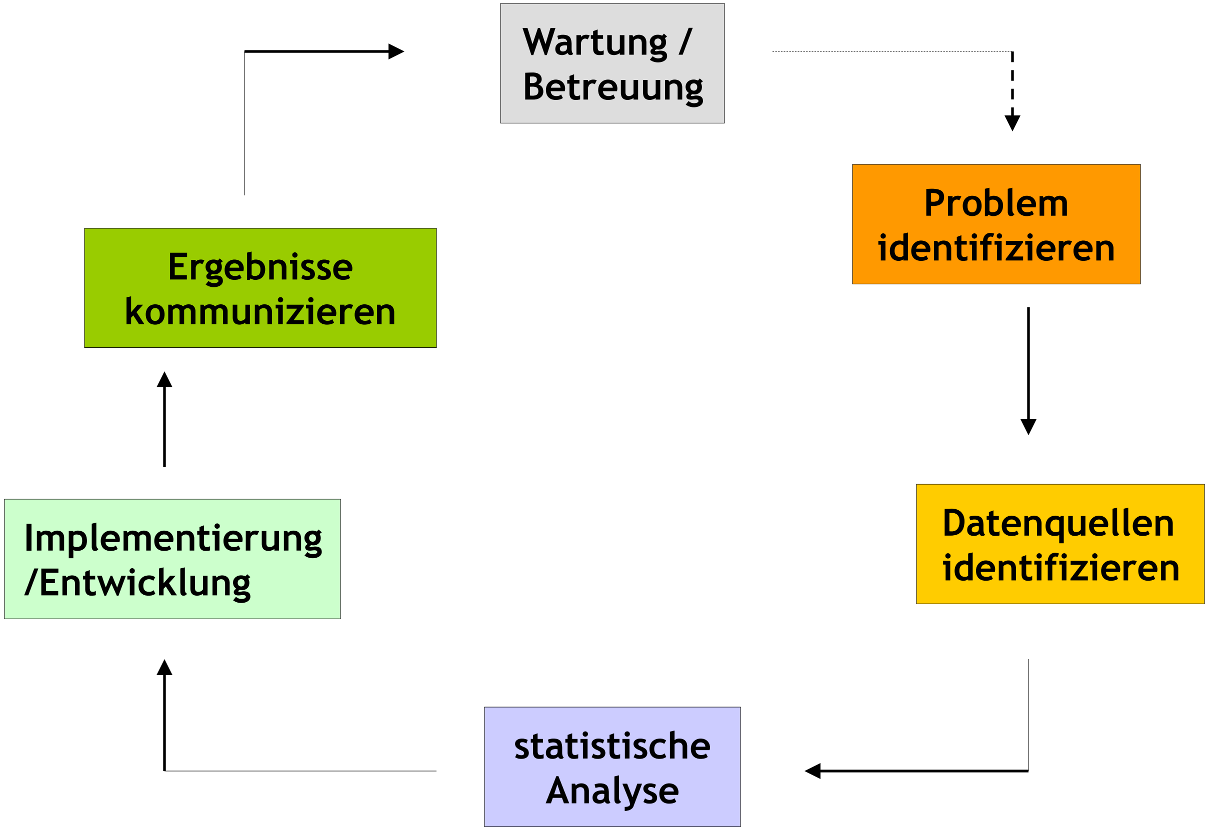
Graphik nach: Towards Data Science

PIPELINE 2



Graphik nach Benjamin Aunkofer, CIO-Expertennetzwerk

LIFE CIRCLE


Wartung / Betreuung
Problem identifizieren
Datenquellen identifizieren
statistische Analyse
Implementierung /Entwicklung
Ergebnisse kommunizieren

DOCH LIEBER „SMART DATA“?

Neuer Trend ?

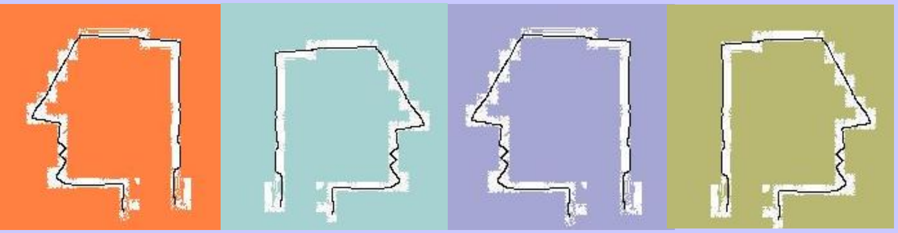
Welche Daten sind für welche Analysen wirklich von Nutzen?

Wäre es nicht sinnvoller / effektiver, anstatt möglichst große Datenmengen zu verarbeiten, die relevanten Daten bereits am Ort der Entstehung erkennen?

Muss man nicht überhaupt zunächst viel eher lernen,
gute Fragen an die Daten zu stellen?

„That‘s like grabbing data by the throat shouting ‘speak to me‘!“

(mein Statistikprofessor zu Analysen ohne konkrete Frageformulierung/ Hypothesen)



Das bringt uns zum idealtypischen Forschungsprozess...

ERKENNTNIS

Empirie

Fragen

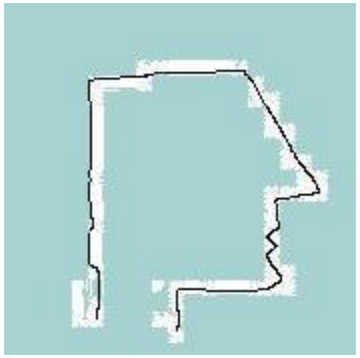
Beobachten

Messen

Testen

systematisch
nachvollziehbar
wiederholbar

theoriebasiert
hypothesengeleitet



Daten

Analysieren

Verstehen

[Anwenden]

Idee / Interesse / Problem

Theorie / Forschungsstand

↓

Konkretisierung der Untersuchungsfragen

Formulierung von Hypothesen

↓

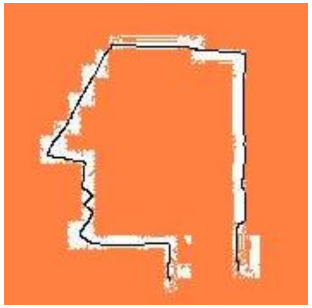
Bestimmung der Vorgehensweise

Dimensionen / Operationalisierung

Instrument / Tools

↓

Durchführung, Auswertung, Bericht

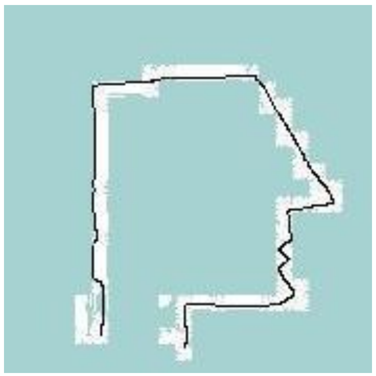


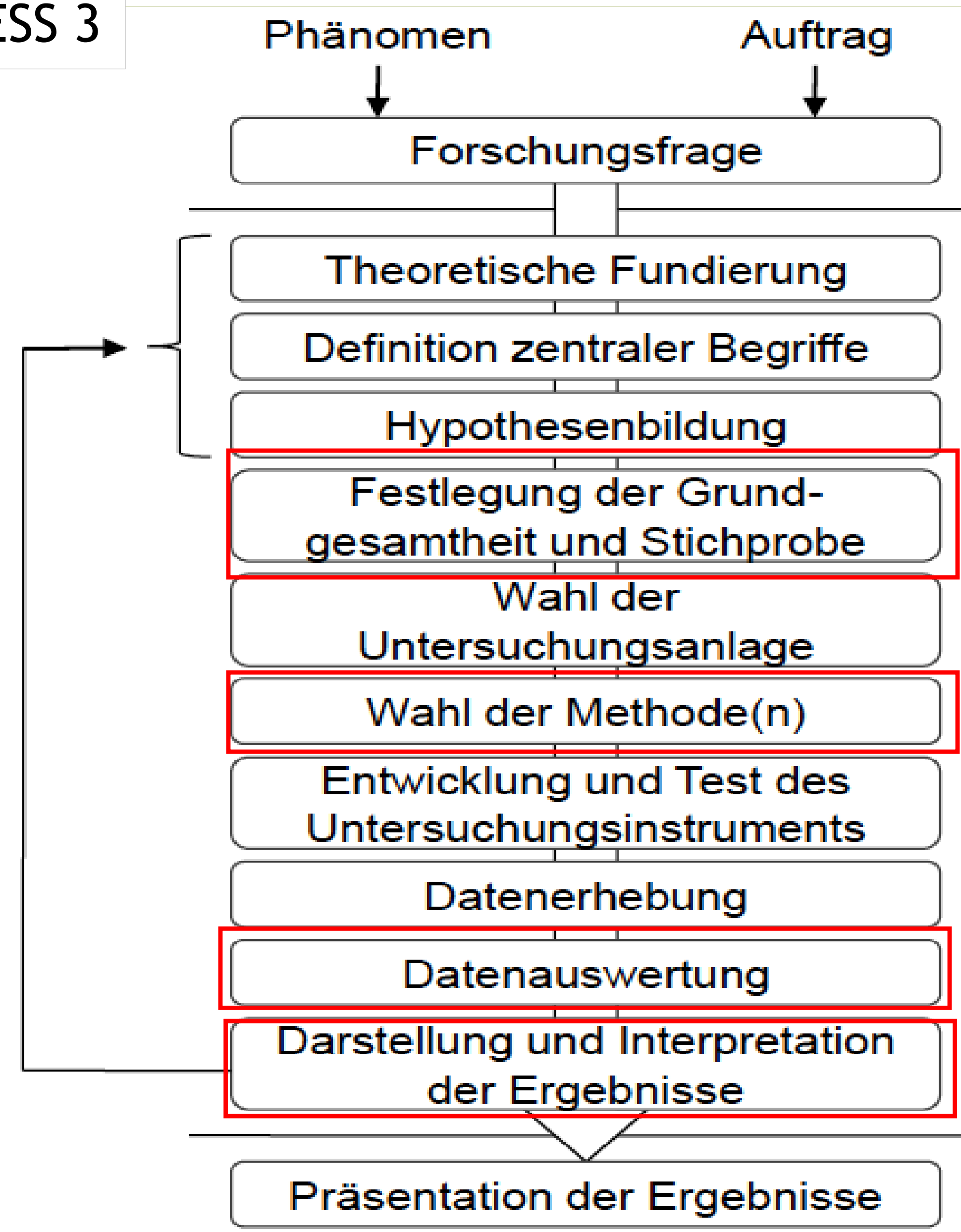
FORSCHUNGSPROZESS 2

Bestimmung der Vorgehensweise

Wer oder was?	Wie?	Reichweite?
Untersuchungsobjekt	Methode	exemplarisch
Untersuchungseinheit		Case Study
		explorativ
	Befragung Beobachtung Test Messen Text- / Bildanalyse	repräsentativ
Sample		
Größe Verfahren		

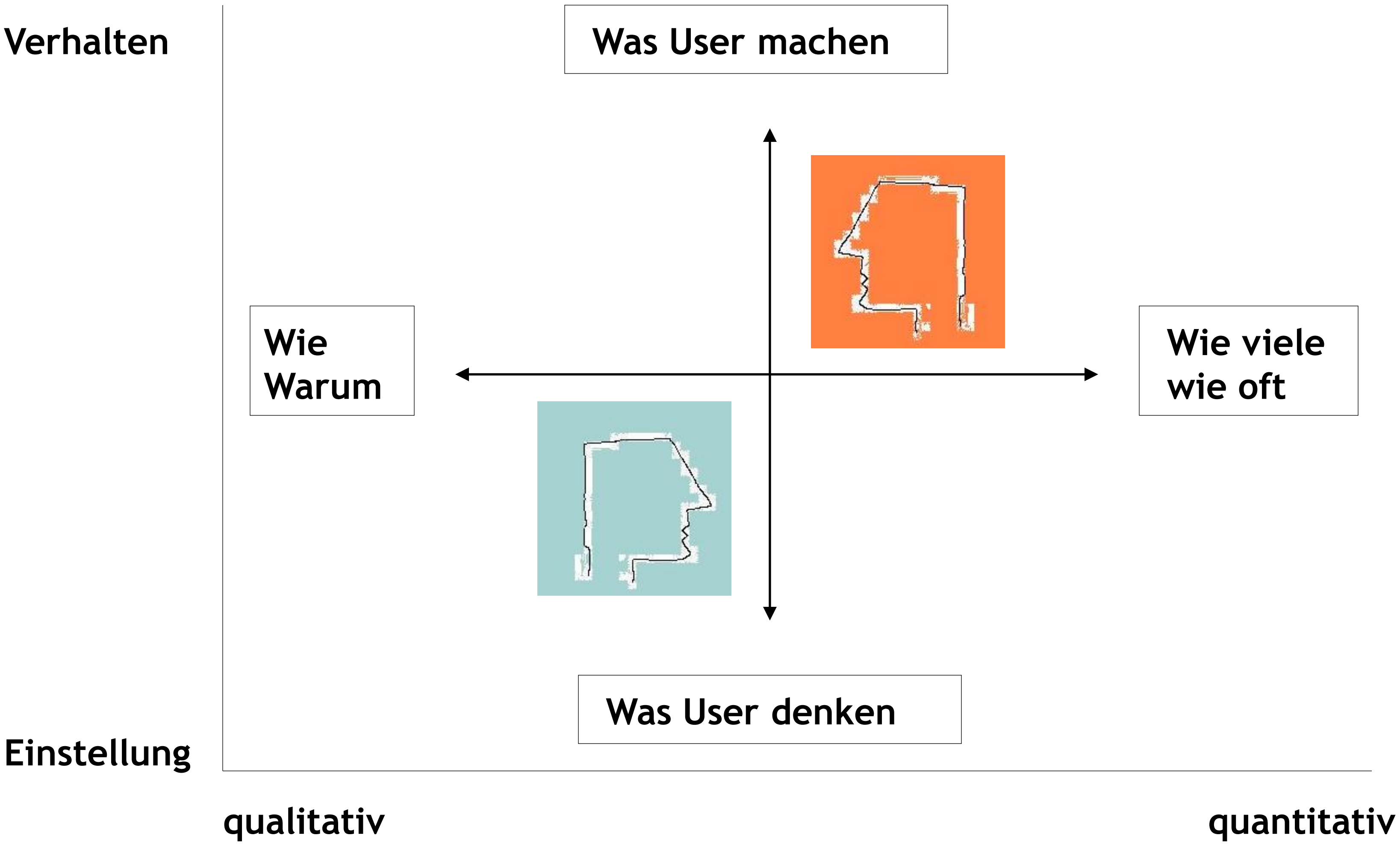
FORSCHUNGSPROZESS 3





Scheufele / Engelmann (2009)

QUANTITATIV vs. QUALITATIV



Quelle: Christian Rohrer

SONDERFALL TEXT

Quellen z.B. Interviews
Social Media
Tagebücher

Ziel Feststellung der Häufigkeit bestimmter Aussagen / Wörter
Clusteranalyse

Konventionelle Inhaltsanalyse	Konventionelle Softwaresysteme	KI - maschinelles Lernen-Systeme
Transkripte Merkmale festlegen ↓ Zahlenwerte zuweisen: codieren	Digitaltextdatei Tagging der Suchwörter ↓ Automatische Codierung	Digitaltextdatei Systemtrainingphase: Textsample Systemverarbeitungsphase ↓ Automatische Codierung

ANALYSE-TOOL

Häufigkeitsauszählung
Korrelationen
Cluster

Subjektivität / Zuverlässigkeit?

Zeitersparnis

bisher nur sehr einfache Texte

Unser Thema:
klassische quantitative statistische Analyse

Konfidenzbereich

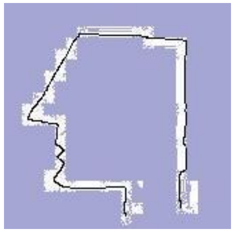
Wahrscheinlichkeit

Randomisierung

0-Hypothese

Häufigkeiten

Quantifizieren



Stichprobe

Prozentanteile

Varianz

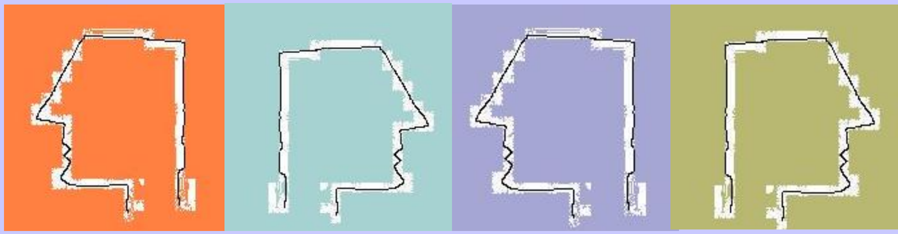
Korrelation

Signifikanz

Mittelwerte

Effektstärke

Dafür ein bisschen mathematische Basic...



Gleichungen

1. Binomische Formel $(a + b)^2 = a^2 + 2ab + b^2$

2. Binomische Formel $(a - b)^2 = a^2 - 2ab + b^2$

3. Binomische Formel $(a - b) * (a + b) = a^2 - b^2$

Rechenregeln

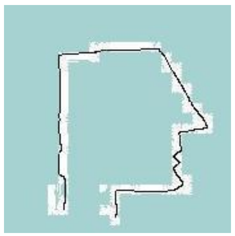
$a + b + c = (a + b) + c$ $a * b * c = (a * b) * c$	**Assoziativgesetz**
$a + b = b + a$ $a * b = b * a$	**Kommutativgesetz**
$(a + b) * c = a * c + b * c$	**Distributivgesetz**

Beispiele: Lineare Gleichungen

Ein bisschen auflösen…

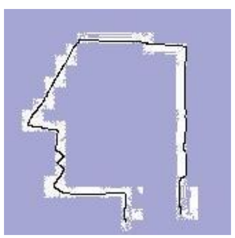
(1) $5x = 15$, $x = ?$

$$5x = 15 \quad | :5$$
$$x = 3$$



(2) $40 + 20x = 20$, $x = ?$

$$40 + 20x = 20 \quad | :20$$
$$2 + x = 1 \quad | -2$$
$$x = -1$$



(3) $4 - 3 + x = 5 - 2$, $x = ?$

$$4 - 3 + x = 5 - 2$$
$$1 + x = 3 \quad | -1$$
$$x = 2$$

(4) $3 + 5 * 2 + 5x = 10$, $x = ?$

$$3 + 5 * 2 + 5x = 10$$
$$3 + 10 + 5x = 10$$
$$13 + 5x = 10 \quad | -13$$
$$5x = -3 \quad | :5$$
$$x = -0{,}6$$

Lineare Gleichungssysteme 1

Definition

Ein lineares Gleichungssystem ist eine Menge von linearen Gleichungen mit einer oder mehrerer Unbekannten.

Es hat entweder genau eine, keine oder unendlich viele Lösungen.

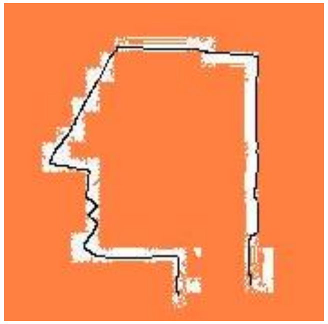
Lösungsverfahren

Gleichsetzung

Addition

Einsetzung

Gauss-Verfahren



Beispiele

(1)
(1) $y = x - 5$
(2) $y = 2x + 3$

(2)
(1) $5x + 3y = 14$
(2) $2x - 2y = -4$

(3)
(1) $y = 6x$
(2) $y/3 + x = 33$

(4)
(1) $5x - 4y + z = -3$
(2) $2x - y - 3z = 10$
(3) $3x - y - z = 4$

Lösungen

(1) Gleichsetzung

(1) y = x - 5

(2) y = 2x + 3

x - 5 = 2x + 3 (- x / + 5)

x = - 8

y = - 8 - 5 = - 13

L [-8 -13]

(2) Addition

(1) 5x + 3y = 14 | * 2

(2) 2x - 2y = - 4 | * 3

kleinstes gemeinsames Vielfaches suchen:
x (10); y (6)

(1) 10x + 6y = 28

(2) 6x - 6y = -12

→ **Addition**

(1+2) 16x = 16 | : 16

x = 1 → Einsetzen (in 1 oder 2)

5 + 3y = 14 | - 5

3y = 9 | : 3

y = 3

L [1 3)]

(3) Einsetzung

(1) y = 6x

(2) y/3 + x = 33

6x/3 + x = 33

2x + x = 33

3x = 33 | : 3

x = 11

y = 66

L [11 66]

Lösungen

(4) Gauss-Verfahren

(1) 5x - 4y + z = -3

(2) 2x - y -3z = 10

(3) 3x - y - z = 4

Gesucht: Lösungstripel (x y z), das alle 3 Gleichungen erfüllt

+

x	y	z	rechte Seite	
5	-4	1	-3	$\vert \cdot (-2)$
2	1	-3	10	$\vert \cdot 5$
3	-1	-1	4	

x	y	z	rechte Seite	
5	-4	1	-3	$\vert \cdot (-3)$
0	13	-17	56	
3	-1	-1	4	$\vert \cdot 5$

+

x	y	z	rechte Seite	
5	-4	1	-3	$\vert \cdot (-3)$
0	13	-17	56	
3	-1	-1	4	$\vert \cdot 5$

x	y	z	rechte Seite	
5	-4	1	-3	
0	13	-17	56	$\vert \cdot (-7)$
0	7	-8	29	$\vert \cdot 13$

+

Lösungen *(Fortsetzung)*

x	y	z	rechte Seite	
5	-4	1	-3	
0	13	-17	56	$\vert\cdot(-7)$
0	7	-8	29	$\vert\cdot 13$

\+

„Dreiecksgestalt“

x	y	z	rechte Seite
5	-4	1	-3
0	13	-17	56
0	0	15	-15

Rückwärts einsetzen, d.h. ab 3. Zeile aufwärts

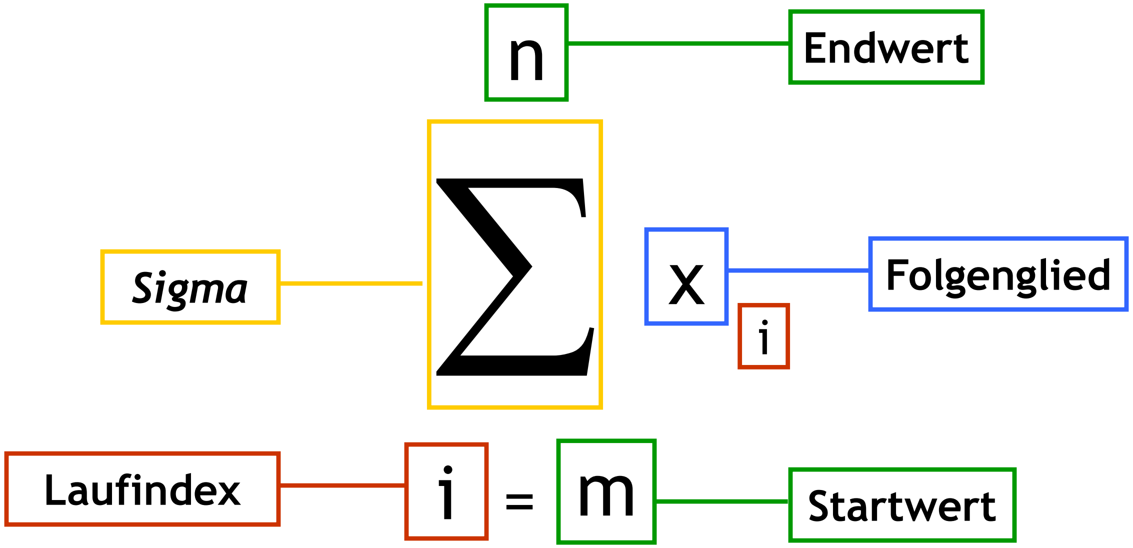
15z = -15 **z = -1**

13y - 17z = 56 **13y + 17 = 56** **13y = 39** **y = 3**

5x - 4y + z = -3 **5x -12 -1 = -3** **5x - 13 = -3** **5x = 10** **x = 2**

L [2 3 -1]

Summenzeichen



Beispiele:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

$$\sum_{i=3}^{8} x_i = x_3 + x_4 + x_5 + x_6 + x_7 + x_8$$

Allgemein:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + (\ldots) + x_n$$

Konstanter Faktor (Distributivgesetz)

konstanter Faktor c

$$\sum_{i=1}^{n} \boxed{c} * x_i = \boxed{c} * \sum_{i=1}^{n} x_i$$

$$= (\boxed{c}x_1 + \boxed{c}x_2 + \boxed{c}x_3 + (\ldots) + \boxed{c}x_n)$$

$$= \boxed{c} * (x_1 + x_2 + x_3 + (\ldots) + x_n)$$

Punktrechung vor Strichrechnung!

Summe (Assoziatives/Kommutatives Gesetz)

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$= ((x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) + (\ldots) + (x_n + y_n))$$

$$= ((x_1 + x_2 + x_3 + (\ldots) + x_n) + (y_1 + y_2 + y_3 + (\ldots) + y_n))$$

Aber:

$$\sum_{i=1}^{n} (x_i * y_i) \neq \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i$$

Lineare Funktionen

Definition

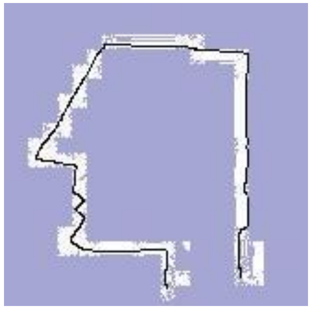
Eine lineare Funktion ist eine Funktion, deren Funktionsgraph eine Gerade ist.

Allgemeine Form

$y = m * x + b$

oder:

$f(x) = m * x + b$



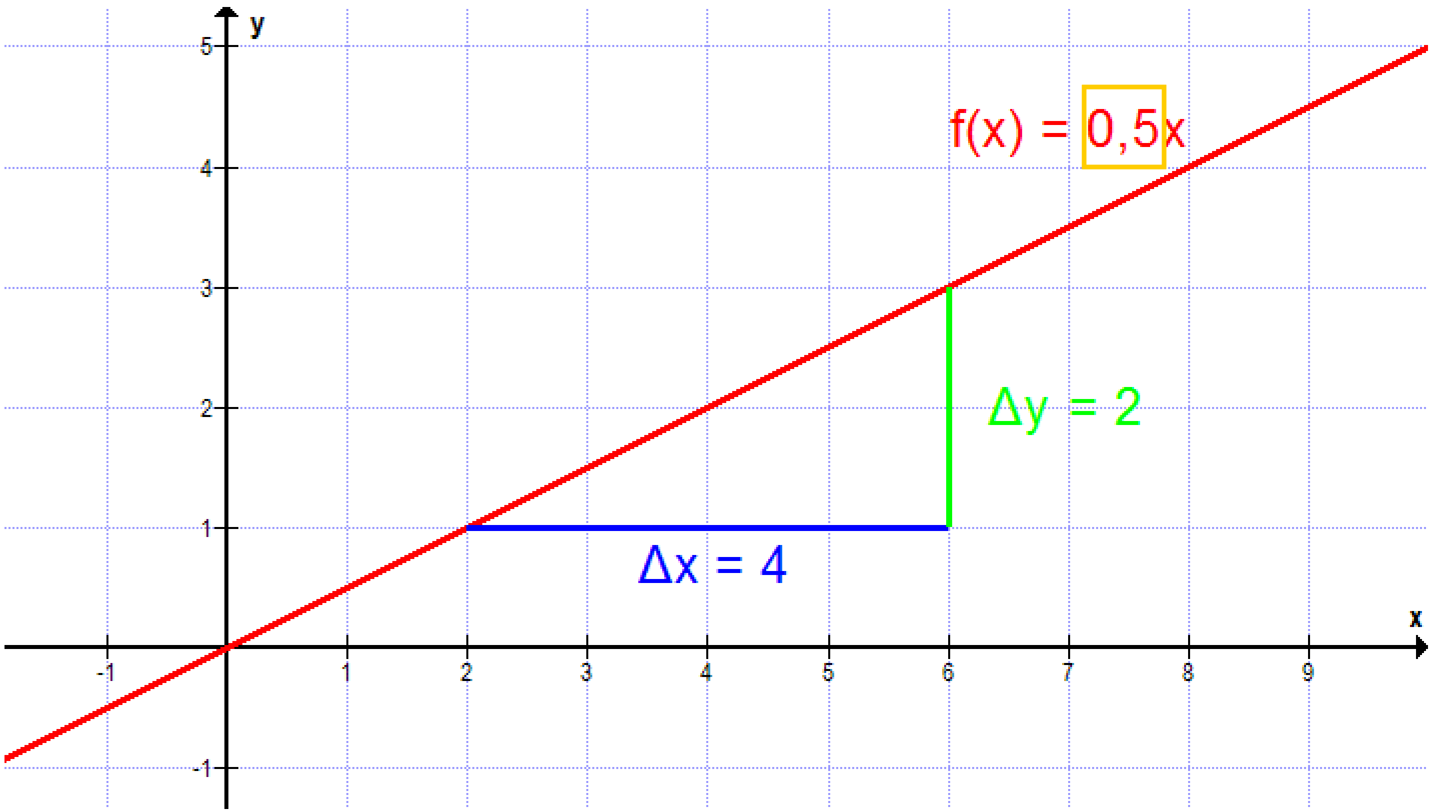
Die Variable Y ist eine Funktion der Variable X; f(x)=y.

m ist der Faktor (Koeffizient) von x und bezeichnet die **Steigung**.

b ist eine **Konstante** und bezeichnet die **Schnittstelle** mit der y-Achse.

Regressionsanalyse

Beispiel Gerade → **Lineare Funktion**



P1(x1|y1)
P2(x2|y2)

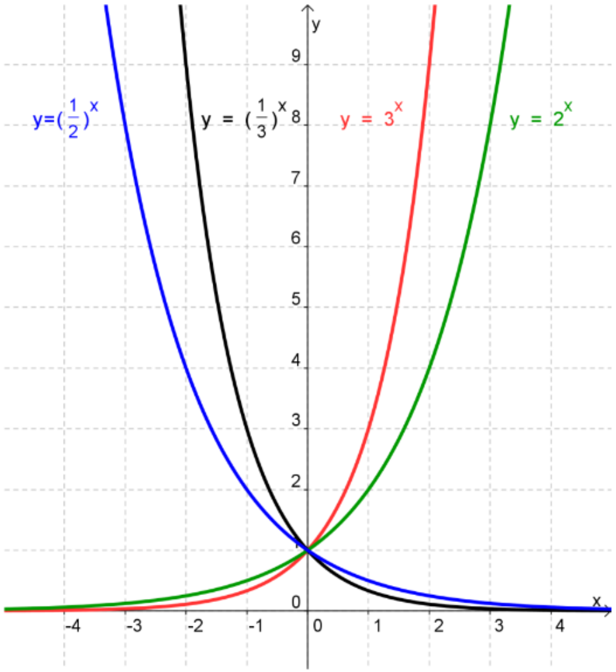
Steigung m $= \dfrac{y_2 - y_1}{x_2 - x_1} = \dfrac{\Delta y}{\Delta x}$ → **an jedem Punkt P der Geraden gleich**

Definition

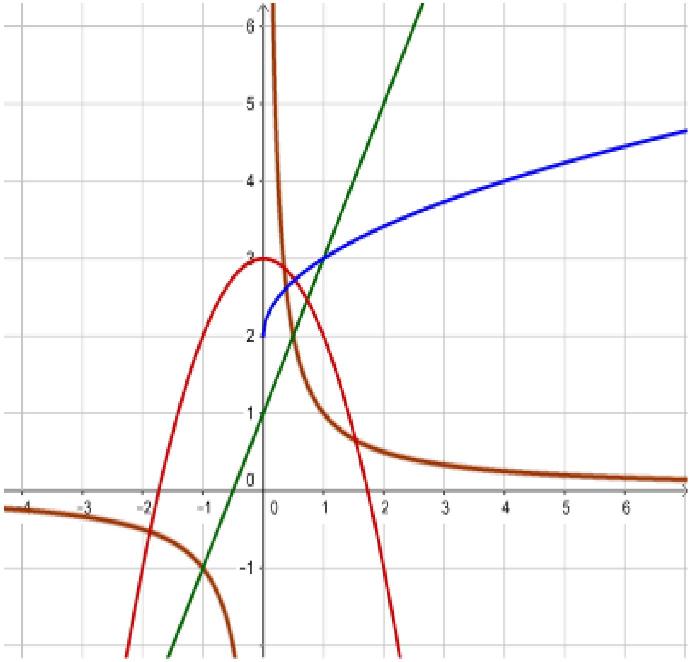
Eine nichtlineare Funktion ist eine Funktion, deren Funktionsgraph eine Kurve ist.

Formen

Exponentialfunktion: $y = a^x$

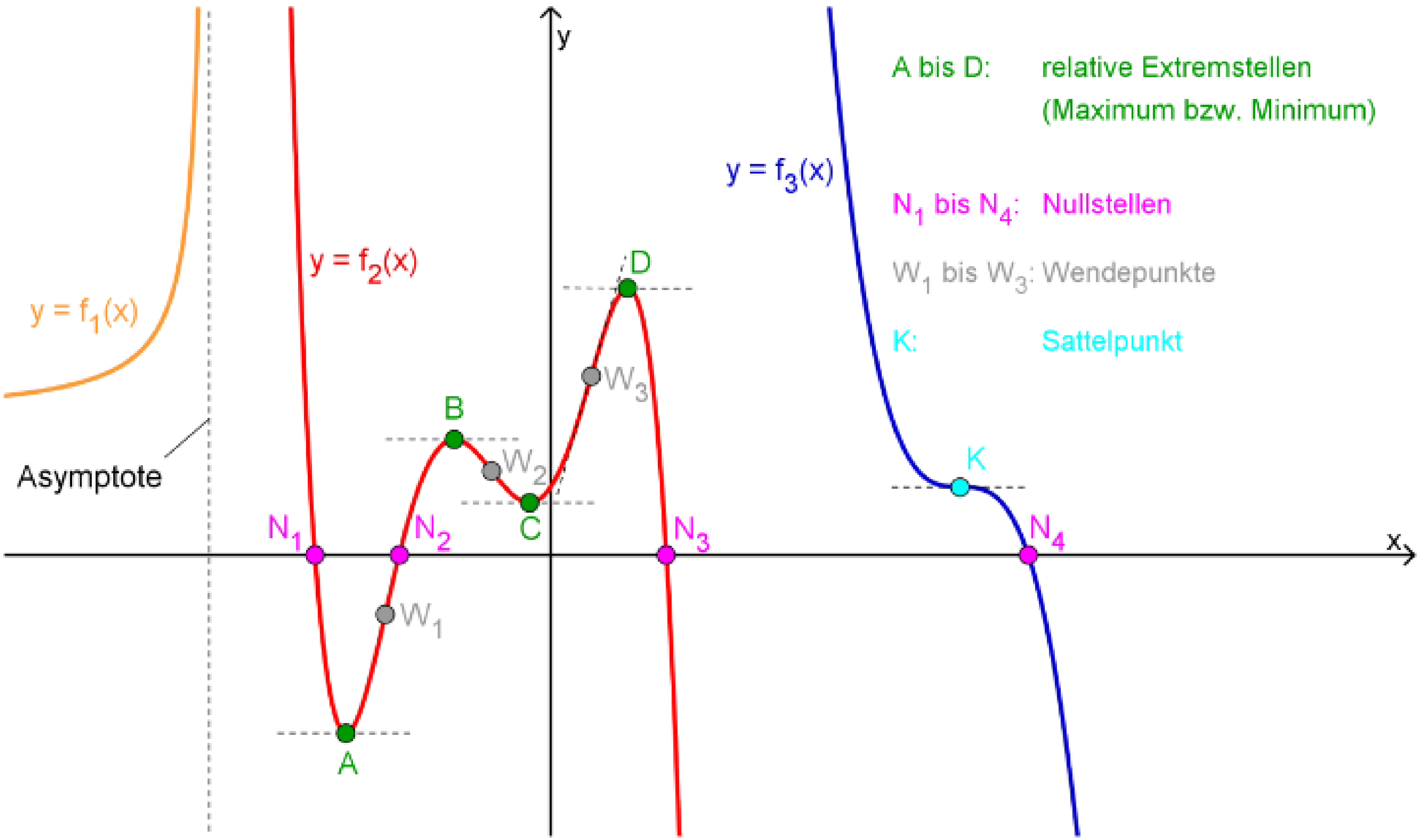


Potenzfunktion: $y = x^n$



Nichtlineare Funktionen

Charakteristische Punkte



Maximum

Punkt eines Graphen dessen benachbarte Punkte (vorher und nachher) einen kleineren y-Wert aufweisen. Die Tangente an den Graphen verläuft in diesem Punkt parallel zur x-Achse bzw. ihre Steigung ist gleich Null.

Minimum

Punkt eines Graphen dessen benachbarte Punkte einen grösseren y-Wert aufweisen. Die Tangente an den Graphen verläuft in diesem Punkt parallel zur x-Achse bzw. ihre Steigung ist gleich Null.

Wendepunkt

Punkt eines Graphen in dem sich die Kurve von der einen Seite der Tangente auf die andere Seite der Tangente wendet. Die Tangente im Wendepunkt heisst Wendetangente.

Sattelpunkt

Punkt eines Graphen bei dem die Wendetangente parallel zur x-Achse verläuft bzw. ihre Steigung gleich Null ist.

Nullstellen

Jene Stellen einer Funktion, bei denen der Graph die x-Achse schneidet bzw. wo die y-Werte gleich Null sind.

Asymptoten

Sind Geraden oder Kurven, die man als Tangenten von Funktionen im Unendlichen auffassen kann. Der Graph der Funktion nähert sich der Asymptote, erreicht sie aber nie! (asymptos ist griechisch und bedeutet «nicht zusammenfallend»)

Spezielle Funktion: exponentielles Wachstum

Definition

$$G = G_0 * a^{t/\tau}$$

$$G_0 = G \ / \ a^{t/\tau}$$

G	Größe, die exponentiell von der Zeit t abhängt
G_0	Wert der Größe G im Zeitpunkt $t = 0$
a	Wachstums- oder Abnahmefaktor, bezogen auf die Zeitspanne τ
t	Zeit
τ	Zeitspanne, auf die sich a bezieht

Beispiel

Eine Bakterienkultur mit exponentiellem Wachstum:
nach 25min, Dichte = 500, nach 45min Dichte = 1.200

τ = 20min (45-25), a = 1200/500 = 2,4

$G_0 = 500 \ / \ 2{,}4^{25/20} = \sim 167$

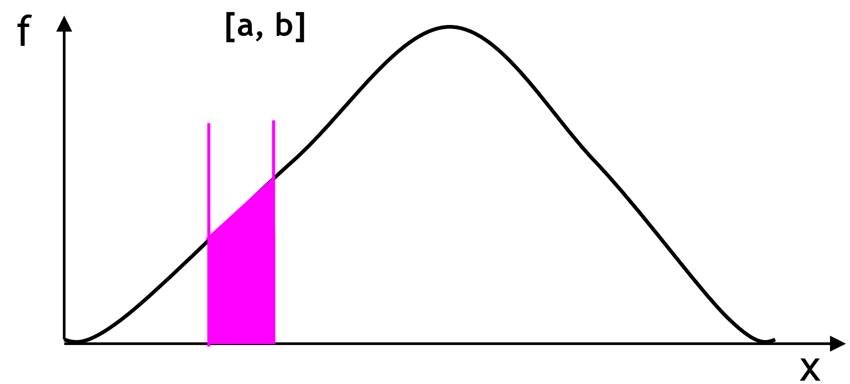
$\boxed{\text{Wachstumsfunktion } G = 167 * 2{,}4^{t/20min}}$

Quelle: Frommenwiler

Integrale

Definition

Ein **bestimmtes integral** ist definiert als die Fläche, die von dem Graphen der Funktion f auf dem Intervall [a, b] eingeschlossen wird, wobei die vertikalen Linien x = a und x = b als Begrenzung dienen.



Um den Flächeninhalt dieses Bereiches zu berechnen, unterteilt man die Fläche in (unendlich schmale) Rechtecke der Breite Δx und der Höhe f(x). Die Summe des Produkts f(x)∗Δx ist der entsprechende Flächeninhalt.

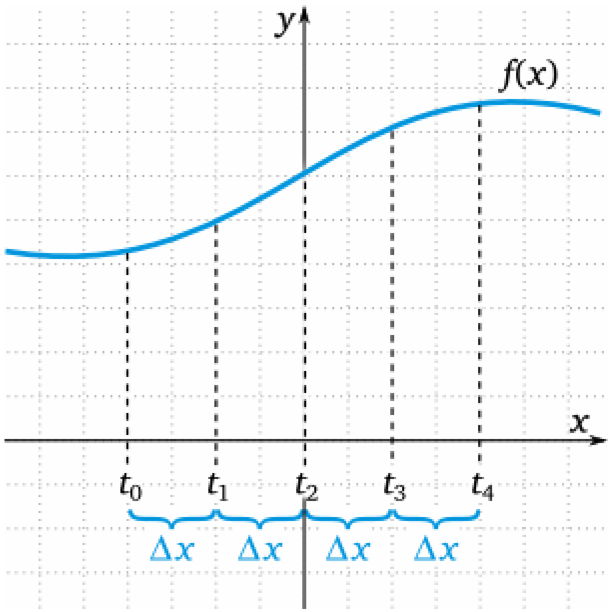
$$\int_a^b f(x) * \Delta x$$

Riemann-Integral

Definition

Das Riemann-Integral ist eine Methode zur numerischen Integration. Die Fläche unter dem Graphen wird mit Hilfe von Formen, in diesem Falle Rechtecke, berechnet.

Zerlegung der Fläche (Intervall [a,b]) in n Teilintervalle t_n.



$$\boxed{\Delta x = \frac{b - a}{n}}$$

$$t_n = a + \Delta x * n \longrightarrow t_0 = a,\ t_n = b$$

Bestimmtes Integral

Notation

[a,b]: Grenzen $\int_a^b \boxed{f(x)}\,\boxed{d}\,\boxed{x}$

Differential

Integrationsvariable

Integrand:
zu integrierende Funktion

Berechnung

$$\int_a^b f(x)dx = [\boxed{F}(x)]_a^b = \boxed{F}(b) - \boxed{F}(a)$$

gesucht: Stammfunktion F

Beispiel

$$f(x) = x^1 \longrightarrow F(x) = \frac{1}{n+1} * x^{n+1} = \frac{x^2}{2}$$

$$\int_0^1 x\,dx = F(1) - F(0) = \frac{1}{2} - 0 = \frac{1}{2}$$

Unbestimmtes Integral

= Stammfunktion plus Konstante C
(**Integrationskonstante**)

Beispiele Stammfunktionen

Konstante Funktion $\int a \, dx = a * x$

Potenzfunktion $\int x^n \, dx = \frac{1}{n+1} * x^{n+1}$

Exponentialfunktion $\int e^x \, dx = e^x$

Logarithmusfunktion $\int \ln(x) \, dx = x + x * \ln(x)$

Sinusfunktion $\int \sin x \, dx = -\cos x$

Grundbegriffe

Zufallsexperiment

beliebig wiederholbar

nach bestimmten Regeln durchgeführt

Ergebnis muss unsicher (= zufällig) sein

(Zufalls-)Ereignis (A)

Ergebnis eines Zufallsexperiments

Beispiel Würfeln

Elementarereignis

nicht in weitere Ereignisse zerlegbar

$E = \{4\}$

Ereignis

Klasse/Menge von Elementarereignissen

$A = \{1,3,5\}$

Ereignisraum

$\Omega = \{1,2,3,4,5,6\}$

Jedes Ereignis ist eine Teilmenge des Ereignisraums und besteht aus mindestens einem Elementarereignis.

Grundbegriffe

Beispiel Würfeln

Sicheres Ereignis

Menge aller Elementarereignisse des Ereignisraums (eines dieser Ereignis *muss* auftreten

$$A = \Omega = \{1,2,3,4,5,6\}$$

Komplementäre Ereignisse

alle Elementarereignisse, die nicht zum Ereignis A gehören. Die Vereinigung von A und Nicht-A führt zum sicheren Ergebnis.

$$A = \{1,3,5\} \quad \bar{A} = \{2,4,6\} \quad A \cup \bar{A} = \{1,2,3,4,5,6\}$$

Wahrscheinlichkeit P

Definition

P(A) = Wahrscheinlichkeit für das Eintreten des Ereignisses A

$$= \frac{\text{Anzahl der für A günstigen Ereignisse}}{\text{Anzahl der insgesamt möglichen Ereignisse}}$$

Laplace-Experiment:

Alle Elementarereignisse haben die gleiche Wahrscheinlichkeit.

Beispiel

Würfeln: wie hoch ist die Wahrscheinlichkeit, eine 1 oder 2 zu würfeln?

$$P(1,2) = \frac{2}{6} = \frac{1}{3}$$

Kolmogorov-Axiome

(1) Die Wahrscheinlichkeit für ein Zufallsereignis liegt zwischen Null und Eins.

$$0 \leq P(A) \leq 1$$

(2) Die Wahrscheinlichkeit für ein sicheres Ereignis is gleich 1.

$$P(A_{sicher}) = 1$$

(3) Für zwei disjunkte (trennscharfe) Ereignisse gilt:

$$P(A \cup B) = P(A) + P(B)$$

Regeln

Additionssatz zur Berechnung der Wahrscheinlichkeit von (A oder B)

$$P(A \cup B) = P(A) + P(B)$$

Multiplikationssatz zur Berechnung der Wahrscheinlichkeit von (A und B)

$$P(A \cap B) = P(A) * P(B)$$

Komplementaritätssatz zur Berechnung der Wahrscheinlichkeit des Auftretens von A

$$P(A) = 1 - P(B)$$

Gegenwahrscheinlichkeit

Kombinatorik 1

Permutation

Wie viele Möglichkeiten gibt es, verschiedene Objekte in einer Reihenfolge anzuordnen?

Beispiel: 3 Menschen, 3 Stühle

N!

3! = 3 * 2 * 1 = 6

Variation

Anordnung mit vorgegebener Reihenfolge

Beispiel: Skat, 32 Karten, nacheinander Kreuz Ass, Pik Ass, Herz Ass, Karo Ass

32 * 31 * 30 * 29 = 863.040 *mögliche Reihenfolgen der ersten 4 Karten*

$$P = \frac{1}{863.040}$$

Wahrscheinlichkeit für eine bestimmte Reihenfolge

Allgemein

$$\boxed{\frac{n!}{(n-k)!}}$$

n Anzahl Objekte

k ausgewählte Objekte

Kombinatorik 2

Kombination

zufällige Auswahl von k Objekten aus n Objekten, ohne Zurücklegen und ohne bestimmte Reihenfolge

$$\binom{n}{k} = \frac{n!}{(n-k)! * k!}$$

Wahrscheinlichkeitsverteilungen

Binomialkoeffizient

Beispiel Lotto

n= 49 Kugeln (insgesamt mögliche Ereignisse)
k= 6 richtige Kugeln (günstige Ereignisse)

$$\binom{n}{k} = \frac{n!}{(n-k)! * k!}$$

$$\binom{49}{6} = \frac{49!}{(49-6)! * 6!} = \frac{49!}{43! * 6!} = 13983816$$

$$P(6er\ im\ Lotto) = \frac{1}{13983816} = 7{,}1 * 10^{-8} = 0{,}000000071$$

Es gibt insgesamt 13983816 mögliche Kombinationen, sechs Kugeln aus 49 Kugeln zu ziehen. Nur eine davon ist die richtige Kombination.

Messwerte: Skalenniveau

<table>
<tr><th></th><th></th><th>Messniveau</th><th>Mathematische Eigenschaften der Messwerte</th><th>Beschreibung der Messwerteigenschaften</th><th>Beispiele</th></tr>
<tr><td rowspan="4">Zunahme des Informationsgehalts</td><td rowspan="2">nicht-metrische Daten</td><td>Nominalskala</td><td>$A = A \neq B$</td><td><u>Klassifikation:</u>
Die Messwerte zweier Untersuchungseinheiten sind identisch oder nicht-identisch.</td><td><u>dichotom:</u>
▪ Geschlecht (m./w.)
<u>polytom:</u>
▪ Parteien (CDU / SPD / FDP)</td></tr>
<tr><td>Ordinalskala</td><td>$A > B > C$</td><td><u>Rangordnung:</u>
Messwerte lassen sich auf einer Messdimension als kleiner / größer /gleich einordnen.</td><td><u>Präferenz- u. Urteilsdaten:</u>
Marke X gefällt mir besser, gleich gut, weniger als Marke Y.</td></tr>
<tr><td rowspan="2">metrische Daten</td><td>Intervallskala</td><td>$A > B > C$
und
$A - B = B - C$</td><td><u>Rangordnung und Abstandsbestimmung:</u>
Die Abstände zwischen Messwerten sind angebbar.</td><td>▪ Temperatur (Celsius)
▪ Geburtsjahr</td></tr>
<tr><td>Rationalskala</td><td>$A = x * B$</td><td><u>Absoluter Nullpunkt:</u>
Neben Abstandsbestimmungen können auch Messwert-verhältnisse berechnet werden.</td><td>▪ Alter
▪ Jahresumsatz
▪ Artikelumfang</td></tr>
</table>

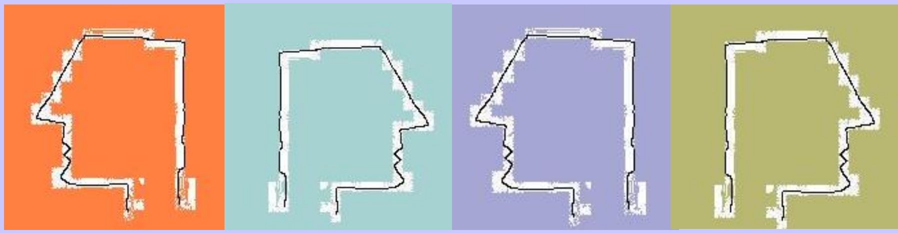
UNIVARIATE STATISTIK

Verteilung einzelner Messwerte / einzelner Variablen im Datensatz

ÜBERBLICK

BEREINIGUNG

DESKRIPTION



Häufigkeiten

Kategorie k	Strichliste	abs. Häufigkeiten	rel. Häufigkeiten (%-Werte)
bis unter 500	|	1	0,8
500 bis unter 1000	||||| ||||| ||	12	9,2
1000 bis unter 1500	||||| ||||| ||||| |||||	20	15,4
1500 bis unter 2000	||||| ||||| ||||| ||||| ||||	24	18,5
2000 bis unter 2500	||||| ||||| ||||| ||||| ||	22	16,9
2500 bis unter 3000	||||| ||||| ||||| |||	18	13,8
3000 bis unter 3500	||||| ||||| |	11	8,5
3500 bis unter 4000	||||| ||||	9	6,9
4000 und mehr	||||| ||||| |||	13	10,0
		n = 130	100

$$\boxed{p_j = \frac{f_j}{n}}$$

Häufigkeiten 2

Jetzt geht es um Radio, Fernsehen, Tageszeitungen und das Internet. Unabhängig davon, wieviel Zeit Sie für die einzelnen Medien aufwenden, möchte ich jetzt von Ihnen wissen, wie häufig Sie Fernsehen nutzen: mehrmals täglich (1), täglich, mehrmals pro Woche, einmal pro Woche, mehrmals im Monat, seltener oder nie (7).

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1 mehrmals täglich	777	17,3	17,3	17,3
	2 täglich	3045	67,7	67,7	84,9
	3 mehrmals pro Woche	460	10,2	10,2	95,1
	4 einmal pro Woche	59	1,3	1,3	96,5
	5 mehrmals im Monat	39	,9	,9	97,3
	6 seltener	54	1,2	1,2	98,5
	7 nie	66	1,5	1,5	100,0
	Gesamt	4500	100,0	100,0	

(Studie Massenkommunikation 2005)

Häufigkeiten 3

(TV-Nutzung gestern / Min.)

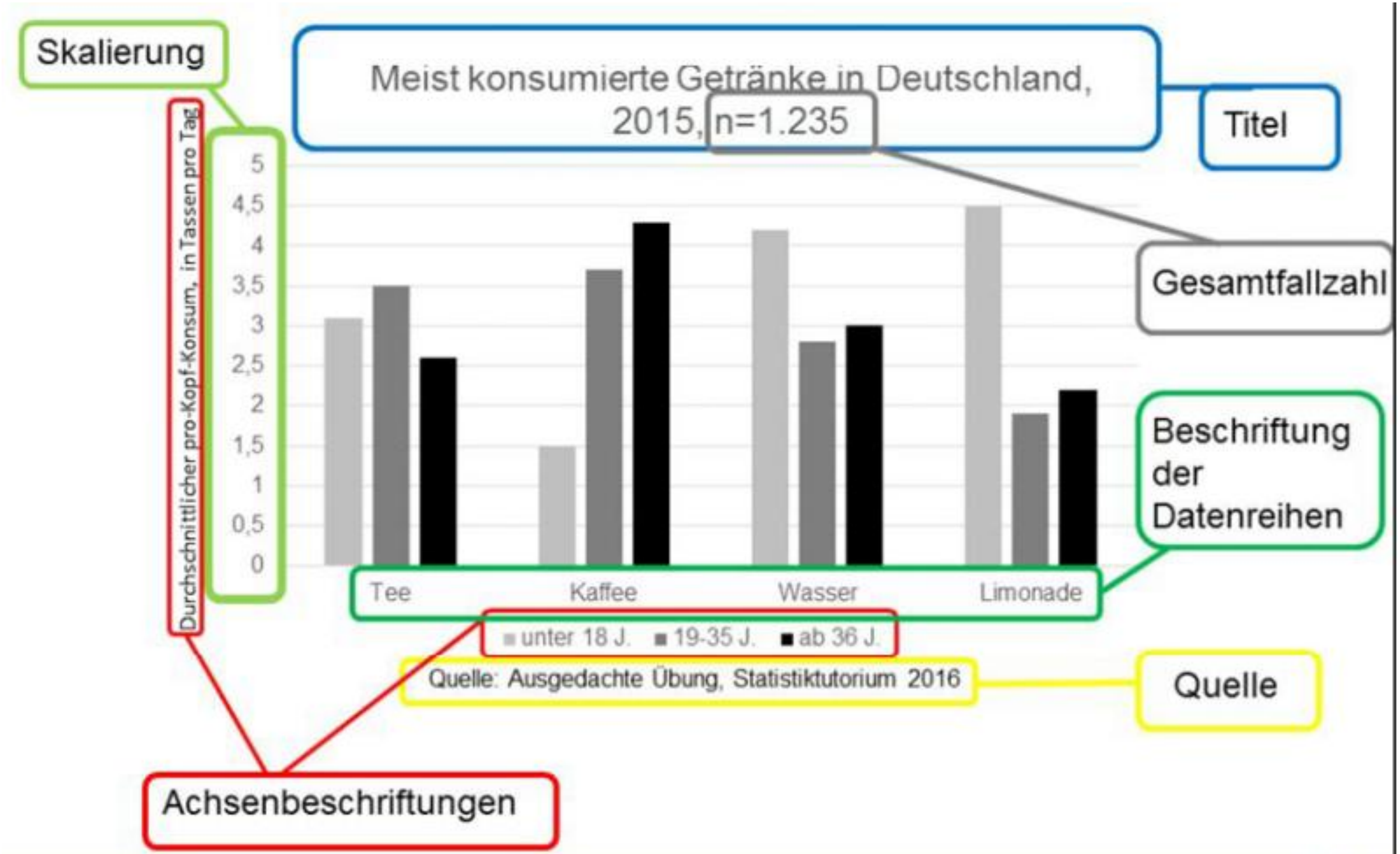
TV_min

N	Gültig	54
	Fehlend	124

TV_min

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	1	,6	1,9	1,9
	2	2	1,1	3,7	5,6
	5	3	1,7	5,6	11,1
	10	1	,6	1,9	13,0
	15	3	1,7	5,6	18,5
	20	4	2,2	7,4	25,9
	30	6	3,4	11,1	37,0
	45	1	,6	1,9	38,9
	50	1	,6	1,9	40,7
	60	13	7,3	24,1	64,8
	90	5	2,8	9,3	74,1
	100	2	1,1	3,7	77,8
	120	4	2,2	7,4	85,2
	180	2	1,1	3,7	88,9
	200	2	1,1	3,7	92,6
	240	3	1,7	5,6	98,1
	360	1	,6	1,9	100,0
	Gesamt	54	30,3	100,0	
Fehlend	0	123	69,1		
	System	1	,6		
	Gesamt	124	69,7		
Gesamt		178	100,0		

CHECKLISTE DIAGRAMME



CHECKLISTE TABELLE

Titel

Gesamtfallzahl

Stat. Maßzahl

Tabelle 7: Lieblingsfarbe nach Geschlecht* (n=166, Angaben in Prozent)

	Geschlecht		
Farbe	**m**	**w**	**Gesamt**
schwarz	5,3	14,0	**11,5**
braun	7,9	2,6	**3,8**
grün	21,1	20,2	**20,6**
weiß	7,9	7,9	**7,9**
blau	39,5	24,6	**27,9**
gelb	0,0	7,0	**5,1**
rot	18,4	23,7	**23,3**
Gesamt	**100,0**	**100,0**	**100,0**

*Geschlossene Frage: Welche der im Folgenden aufgefürten Farben ist Ihre Lieblingsfarbe?

Quelle: Umfrage zur Vorlesung Methoden II (SoSe2013)

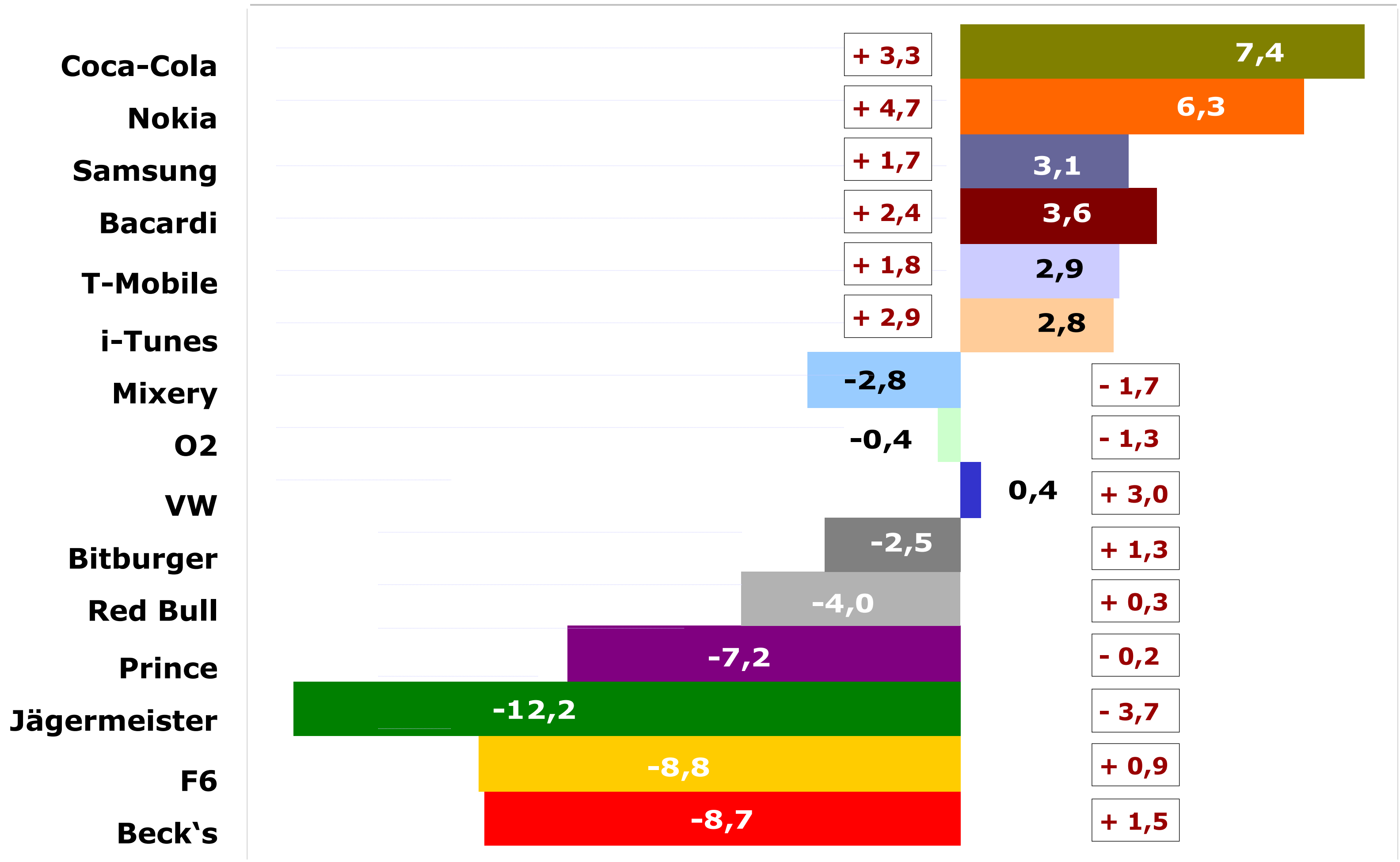
Kategorien (aussagekräftige Beschriftung)

Datenquelle

DARSTELLUNG VON BEWERTUNGSSKALEN

Wie glaubwürdig finden Sie das Engagement der folgenden Firmen und Marken im Bereich Musik?

Differenz der %-Anteile sehr glaubwürdig/glaubwürdig und unglaubwürdig/sehr unglaubwürdig



(Differenz zur 2. Welle; in 1. Welle nicht gefragt) **Basis: alle Befragten; n = 1.008; in %**

Univariate Kennwerte

Definition

„Die in einem Datensatz für ein Merkmal enthaltene Information lässt sich zu *Kenngrößen* verdichten.

Diese charakterisieren das **Zentrum** oder die Variabilität des Datensatzes. Man hat also Kenngrößen zur Beschreibung der **„mittleren“ Lage** der Elemente des Datensatzes und solche zur Charakterisierung der Streuung.“

Mittag, Hans-Joachim (2015): Statistik: Eine Einführung mit interaktiven Elementen. Berlin, Heidelberg: Springer. S. 103.

- Datenverdichtung, Reduktion von Komplexität
- Interpretationshilfen, Vergleichsgrößen
- Kommunikation der Dateneigenschaften
- Maße der zentralen Tendenz, Lagemaße: Typische Werte
- Streumaße: Heterogenität, Unterschiedlichkeit der Werte

Skalenniveau und univariate Kennwerte

	Messniveau	Eigenschaften	mögliche Aussage	Beispiele
non-metrisch	Nominalskala	klassifizierend	gleich, ungleich	Farben, Geschlecht
	Ordinalskala	Rangordnung, keine gleichen Abstände	größer, kleiner	Bewertung von Kinofilmen
metrisch	Intervallskala	gleiche Abstände	Gleichheit von Differenzen	Temperatur in Grad C
	Ratioskala	absoluter Nullpunkt	Gleichheit von Verhältnissen	TV-Nutzung in Min/Tag

Skalenniveau	Lagemaß
nominal	Modalwert
ordinal	Median, Modalwert
metrisch	Arithmetisches Mittel, Modalwert, Median

Modalwert (Modus)

mindestens Nominalskalenniveau

der Wert (Merkmalsausprägung), der innerhalb einer Datenmenge am häufigsten vorkommt

Median (Md, $\tilde{x}$)

mindestens Ordinalskalenniveau

der Wert (Merkmalsausprägung), der in der Mitte steht, wenn alle Beobachtungswerte x_i der Größe nach geordnet sind.

nicht von Extremwerten beeinflusst

Ungrade Fallzahl

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

Gerade Fallzahl

$$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2}$$

Arithmetisches Mittel (AM, $\bar{x}$)

metrisches Skalenniveau

die Summe aller Werte, geteilt durch Anzahl der Fälle

„Gleichgewichtspunkt der Verteilung“

von Extremwerten beeinflusst

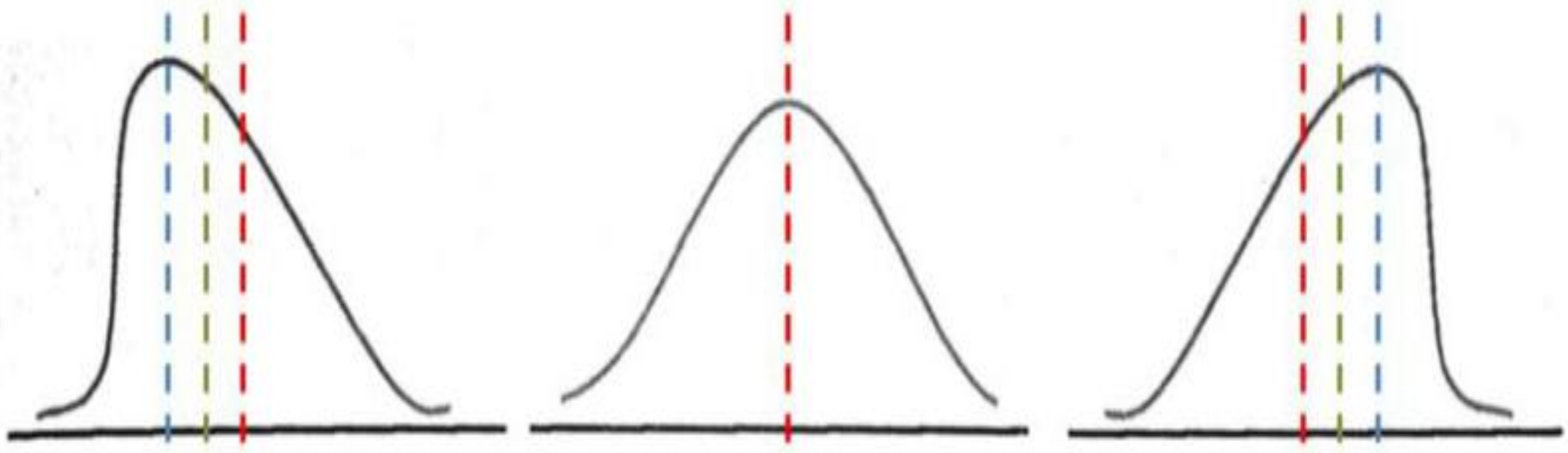
ohne Klassenbildung

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i$$

mit Klassenbildung

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i * f_i$$

Lagemaße und Verteilung



Mo. < Md. < AM

Mo. = Md. = AM

AM < Md. < Mo.

linksschief

rechtsschief

Normalverteilung

Varianz (s^2)

Summe der quadrierten Abweichungen der Einzelfälle vom Arithmetischen Mittel

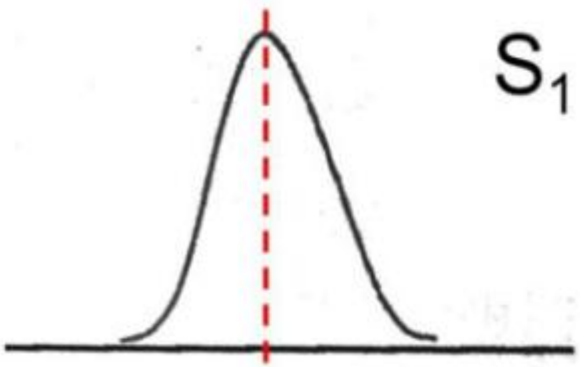
$$s^2 = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Standardabweichung (s)

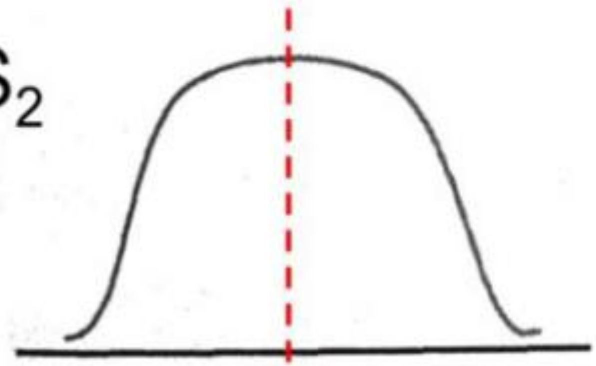
Wurzel aus der Varianz

Aussagekraft nur im Vergleich

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



$S_1 < S_2$



schmalgipflig

breitgipflig

Kleine Standardabweichung = homogene Verteilung

Große Standardabweichung = heterogene Verteilung

Diese Interpretation nur im Vergleich sinnvoll!

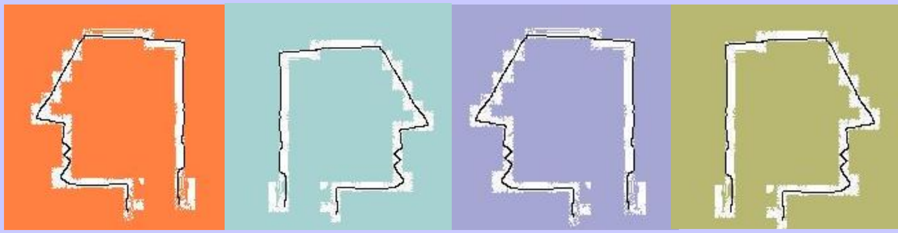
BIVARIATE STATISTIK

Zusammenhang zweier Messwerte / Variablen im Datensatz

DIFFERENZEN / UNTERSCHIEDE

ZUSAMMENHÄNGE

HYPOTHESEN-TESTS



Bivariate Häufigkeitsverteilung

Definition

„Hat man zwei diskrete Merkmale X und Y mit k bzw. m Ausprägungen, kann man die **absoluten oder relativen Häufigkeiten** für die k   m Ausprägungskombinationen tabellarisch darstellen.

Diese auch als **Kontingenztafel** bezeichnete Tabelle definiert eine bivariate Häufigkeitsverteilung.

Ein Spezialfall der Kontingenztafel ist die **Vierfeldertafel**, bei der X und Y jeweils nur zwei Ausprägungen aufweisen."

Mittag, Hans-Joachim (2015): Statistik: Eine Einführung mit interaktiven Elementen. Berlin, Heidelberg: Springer. S. 103.

Beispiel 1

Frage zur Präferenz von Filmen	Frauen (n=1.080)	Männer (n=1.090)	**Gesamt** (n=2.170)
Some like it Hot	48 %	8%	28 %
Der Sturm	3 %	7 %	5 %
Stirb langsam	10 %	40 %	25 %
Star Wars Episode IV	7 %	44 %	26 %
Anna Karenina	32 %	1 %	16 %
Gesamt	**100 %**	**100 %**	**100 %**

Konvention: Spalte = *unabhängige* (Einfluss-) Größe
Zeile = *abhängige* Größe

Beispiel 2

Frage zur Präferenz von Filmen	Frauen (n=1.080)	Männer (n=1.090)	Gesamt (n=2.170)
Some like it Hot	86 %	14%	100 %
Der Sturm	30 %	70 %	100 %
Stirb langsam	20 %	80 %	100 %
Star Wars Episode IV	14 %	88 %	100 %
Anna Karenina	97 %	3 %	100 %

→ Weniger Informationen als bei Spaltenprozentuierung

Bivariate Zusammenhangsmaße

Definition

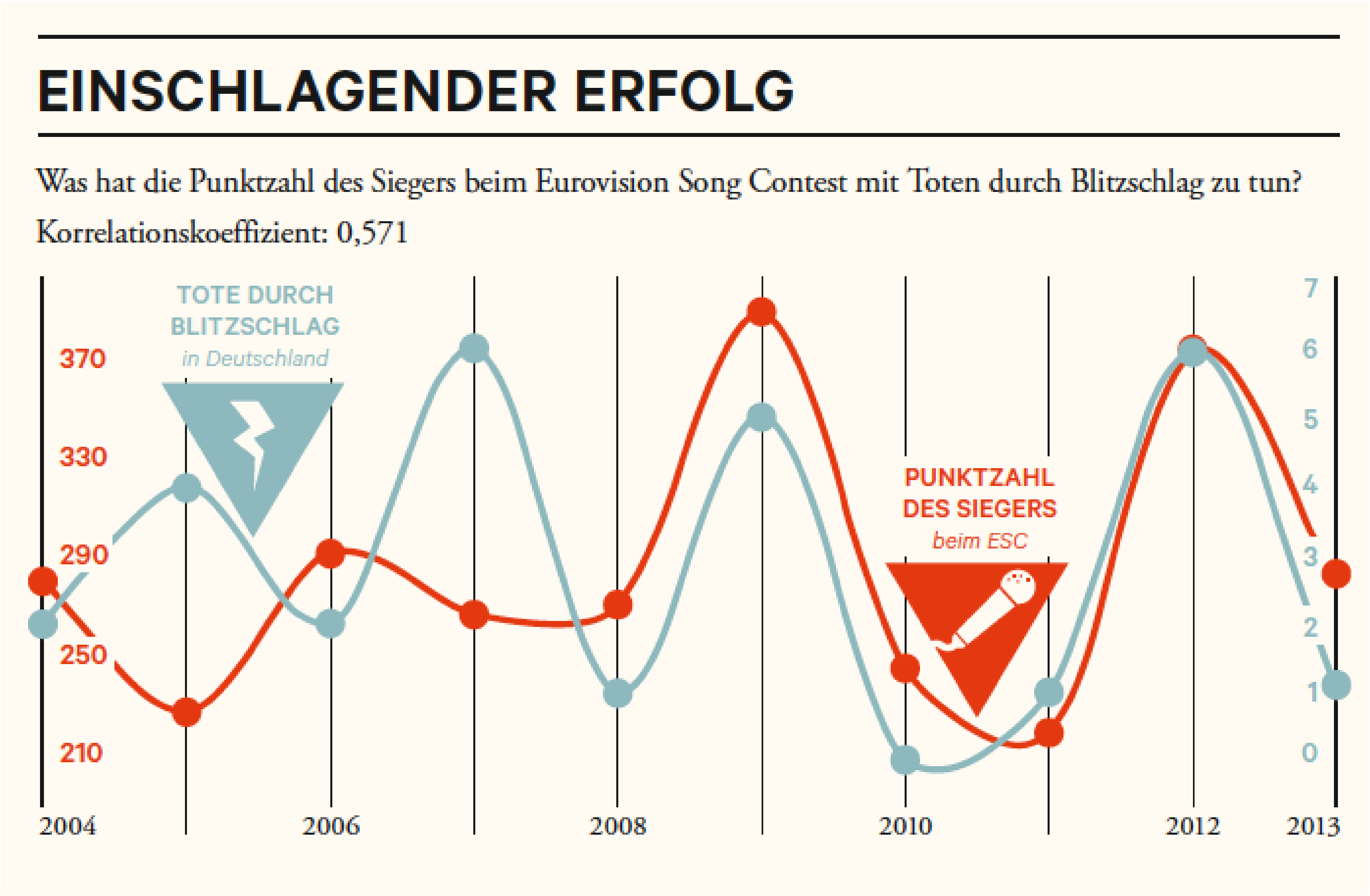
Bivariate Zusammenhangsmaße beschreiben die **gemeinsame Verteilung** zweier Variablen. Sie lassen Aussagen über Zusammenhänge und Unterschiede zu.

Mit anderen Worten: sie sind Maße für die **Koinzidenz**zweier Merkmale.

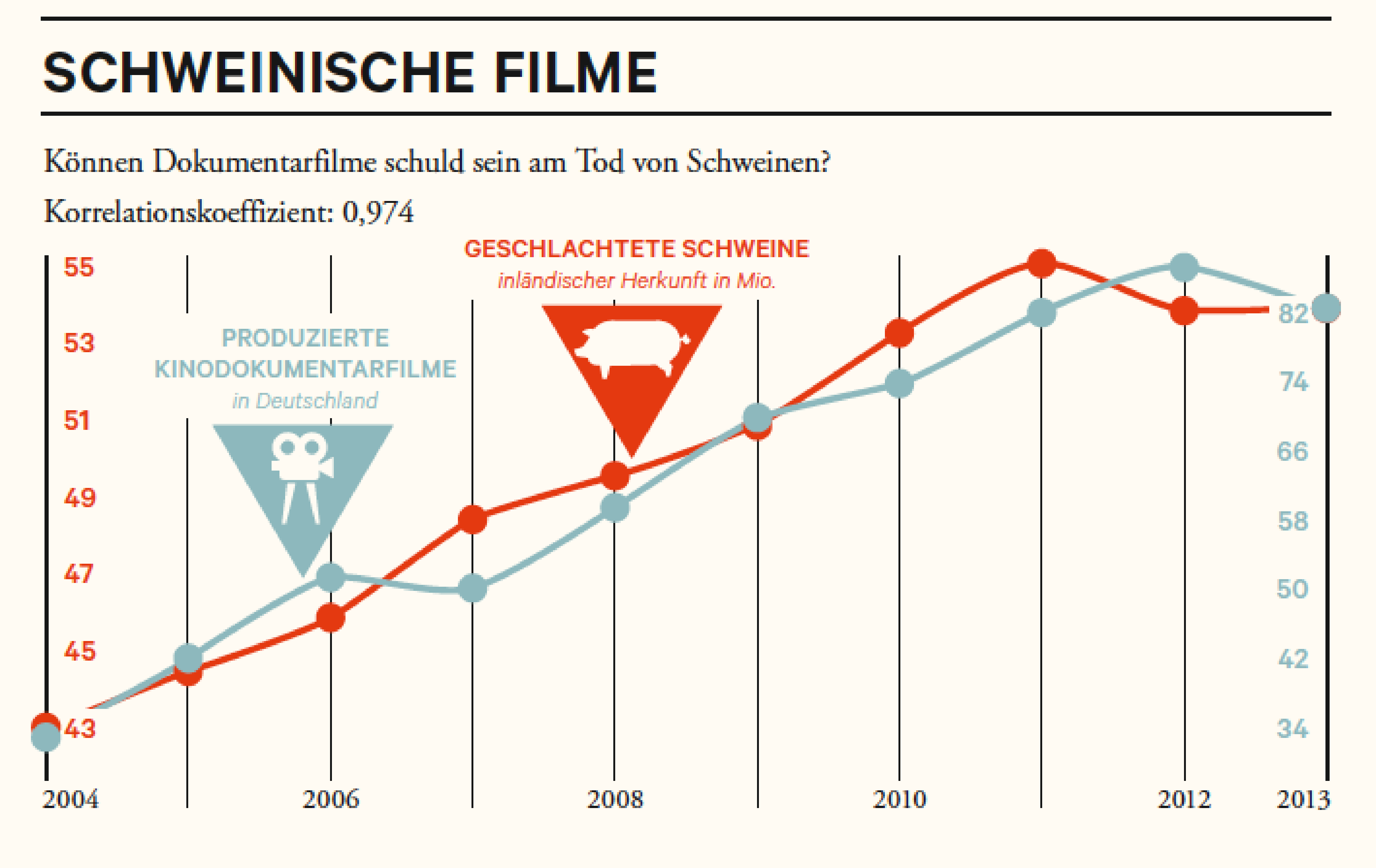
Bivariate Zusammenhangsmaße gibt es für jedes Skalenniveau.

Skalenniveau	Beispiele	Zusammenhangsmaß	Aussage
nominal	Geschlecht, Parteipräferenz	**Chi²** **Cramer's V**	Zusammenhang Stärke
ordinal	Lieblingsfilme Person A und B	**Rangkorrelationskoeffizient** **Spearman's τ** (rho)	Übereinstimmung / Stärke
metrisch	Größe, Gewicht	**Kovarianz** **Korrelationskoeffizient**	Zusammenhang je-desto / Stärke

Zusammenhänge…



Zusammenhänge…



Korrelation vs. Kausalität

Zusammenhangsmaß CHI² 1

Definition

Maßzahl für den Zusammenhang zweier nominalskalierter Variabeln.

Basis: Kreuz- bzw. **Kontingenztabelle**

Logik

Berechnung einer zweiten sog. **Indifferenztabelle** unter der Annahme, dass **kein** Zusammenhang zwischen den Werten der beiden Variablen besteht.

Das Zusammenhangsmaß **Chi2 (X^2)** ist die **Summe der Werteabweichungen** zwischen empirischer Kontingenz- und berechneter Indifferenztabelle.

Chi2 (X^2) hat einen Wertebereich von 0 (= kein Zusammenhang bis ∞ (= maximaler Zusammenhang).

Der Maximalwert ist abhängig von der Skalierung der Variablen, Tabellen unterschiedlicher Variablen lassen sich deshalb nicht ohne Weiteres vergleichen.

Zusammenhangsmaß CHI² 2

Beobachtete Häufigkeiten

Anzahl der Zellen

$$\boxed{X^2} = \sum_{i=1}^{\boxed{k}} \frac{(\boxed{f_{b(i)}} - \boxed{f_{e(i)}})^2}{\boxed{f_{e(i)}}}$$

Erwartete Häufigkeiten

Chi2-Kennwert

Vorgehensweise

Voraussetzungen

Kreuztabelle mit absoluten Zellen- und Randhäufigkeiten
nominalskalierte Variablen (u.U. auch ordinalskalierte)
Gesamtfallzahl mind. n = 60
alle Zellenwerte mind. n = 1
weniger als 20% aller Zellen mit einer Häufigkeit < n = 5

Schritte

(1) Kontingenztabelle erstellen
beobachtete Häufigkeiten (f_b) in absoluten Zahlen

(2) Indifferenztabelle berechnen
Erwartete Häufigkeiten (f_e) für alle Zellen berechnen

$$\boxed{f_e = \frac{\text{Zeilen (n)} * \text{Spalten (n)}}{\text{Gesamt (n)}}}$$

(3) Für jede Zelle die Abweichung zwischen Kontingenz- und Indifferenztabelle berechnen

$$\boxed{\frac{(f_b - f_e)^2}{f_e}}$$

(4) Aufsummieren zu **Chi2 (X^2)**

Standardisierungsmaße von CHI²

Warum?

Die Werte des Zusammenhangsmaßes Chi^2 hängen von der Anzahl n der Messwerte und der Grüße der Tabelle ab.

Chi2-Werte unterschiedlicher Tabellen können deshalb auch nicht miteinander verglichen werden.

Zur besseren Interpretation und Vergleichbarkeit stehen **standardisierte** Maße zur Verfügung:

Cramer's V (für beliebige Kreuztabellen)
Kontingenzkoeffizient C (für beliebige Kreuztabellen)
Phi (für Vierfeldertabellen)

Wertebereiche:

0 (kein Zusammenhang) $\leq$ V/C/Phi $\leq$ 1 (perfekter Zusammenhang)

Stärke des Zusammenhangs, nicht Richtung!

Beispiel Cramer‘s V

Definition

Cramer‘s V ist ein **standardisiertes** Maß, das die **Stärke** des Zusammenhangs zweier **nominalskalierter** Variablen angibt.

$$V = \sqrt{\frac{X^2}{n * (R - 1)}}$$

$$0 \leq V \leq 1$$

X^2 = Chi²-Wert

i = Anzahl der Kategorien der Zeilenvariable

j = Anzahl der Spaltenvariable

R = min (i,j) -> ist die kleinere Zahl von beiden (bei einer 3x4-Tabelle z.B. ist R = 3)

Rangkorrelationskoeffizient Spearman

Definition

Spearman‘s τ_s ist ein **standardisiertes** skalenunabhängiges Maß, das **Stärke** und **Richtung** des Zusammenhangs zweier mindestens **ordinalskalierter** Variablen angibt.

τ_s berücksichtigt die **Rangreihenfolge**, nicht deren Höhe, und ist dadurch robust gegenüber Ausreißern. es kann ab n > 5 berechnet werden.

Konstante (*don‘t ask)*

$$\tau_s = 1 - \frac{\boxed{6} * \sum_{i=1}^{n} \boxed{d_i^2}}{\boxed{n} * (n^2 - 1)}$$

quadrierte Randplatzdifferenz (d)

Rangplätze müssen vor der Berechnung auf- oder absteigend sortiert sein.

Anzahl der Ränge, nicht Fälle!

Wertebereich

-1 (perfekter negativer Zusammenhang) $\leq \tau_s \leq$ 1 (perfekter positiver Zusammenhang)

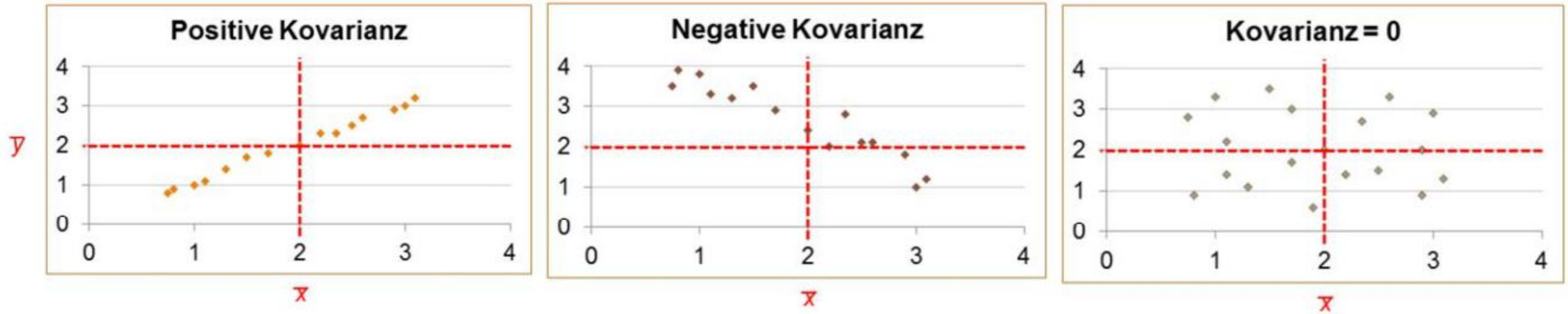
Bei τ_s = 0 sind die Variablen unabhängig voneinander.

Definition

Die Kovarianz (cov_{xy}) ist ein **nicht-standardisiertes** Zusammenhangsmaß zur Beschreibung **linearer Zusammenhänge** zwischen zwei mindestens **metrisch** skalierten Variablen X und Y.

Die Kovarianz ist das durchschnittliche Abweichungsprodukt aller Messwertepaare von ihrem jeweiligen Mittelwert.

$$\boxed{cov_{xy} = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}$$



Korrelationskoeffizient Pearson‘s r

Definition

Pearson‘s r ist ein **standardisiertes** skalenunabhängiges Maß, das die **Stärke** und **Richtung** des **linearen** Zusammenhangs zweier **metrisch** skalierter Variablen angibt.

$$r_{xy} = \frac{\boxed{\frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}}{\boxed{s_x * s_y}}$$

Kovarianz

Produkt der Standardabweichungen von X und Y

Wertebereich

-1 (perfekt negativer linearer Zusammenhang) $\leq$ $r_{x,y}$ $\leq$ 1 (perfekt positiver linearer Zusammenhang)

Bei $r_{x,y}$ = besteht kein **linearer** Zusammenhang.

Pearson's r: ein bisschen Geformel

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i - \overline{x})}{s_x}\right)\left(\frac{(y_i - \overline{y})}{s_y}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i - \overline{x})}{\sqrt{\sum \frac{(x_i - \overline{x})^2}{n}}}\right)\left(\frac{(y_i - \overline{y})}{\sqrt{\sum \frac{(y_i - \overline{y})^2}{n}}}\right)$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2)(\sum_{i=1}^{n}(y_i - \overline{y})^2)}}$$

Überblick standardisierte bivariate Zusammenhangsmaße

Y- Variable → X-Variable ↓	Nominal	Ordinal	Metrisch
nominal	Cramer's V		
ordinal		Spearman‘s Rho	
metrisch			Pearson‘s r

Wertebereich: $0 \leq V \leq 1$ $-1 \leq r_S \leq 1$ $-1 \leq r_p \leq 1$