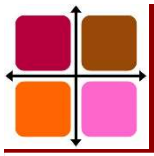


Missbrauch von Daten
- Statistik für Anfänger

Vortrag beim No Spy Day, Stuttgart

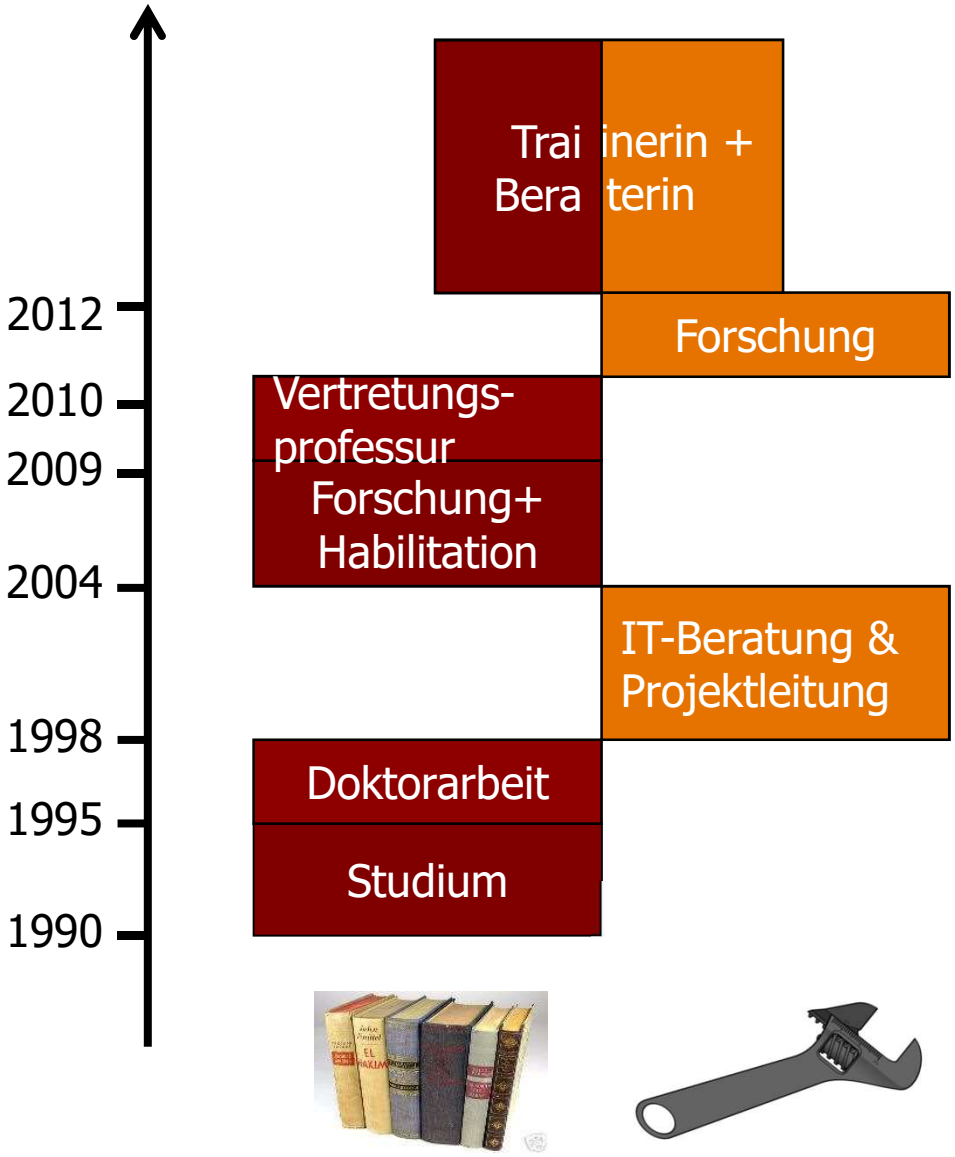
18.05.2019

Dr. habil. Andrea Herrmann

Freie Software Engineering Trainerin und Beraterin

D-70372 Stuttgart, herrmann@herrmann-ehrlich.de

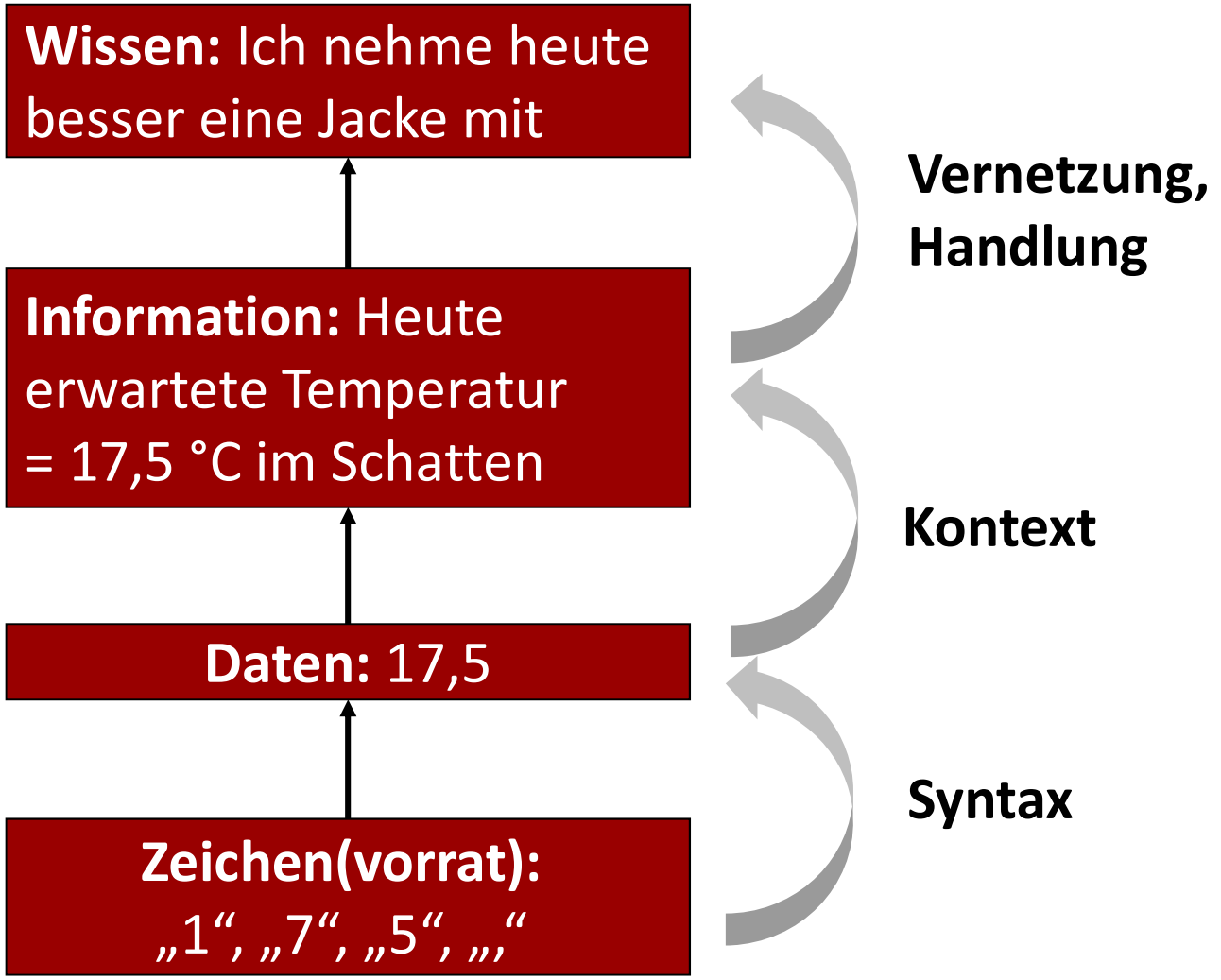
Dr. habil. Andrea Herrmann



Übersicht

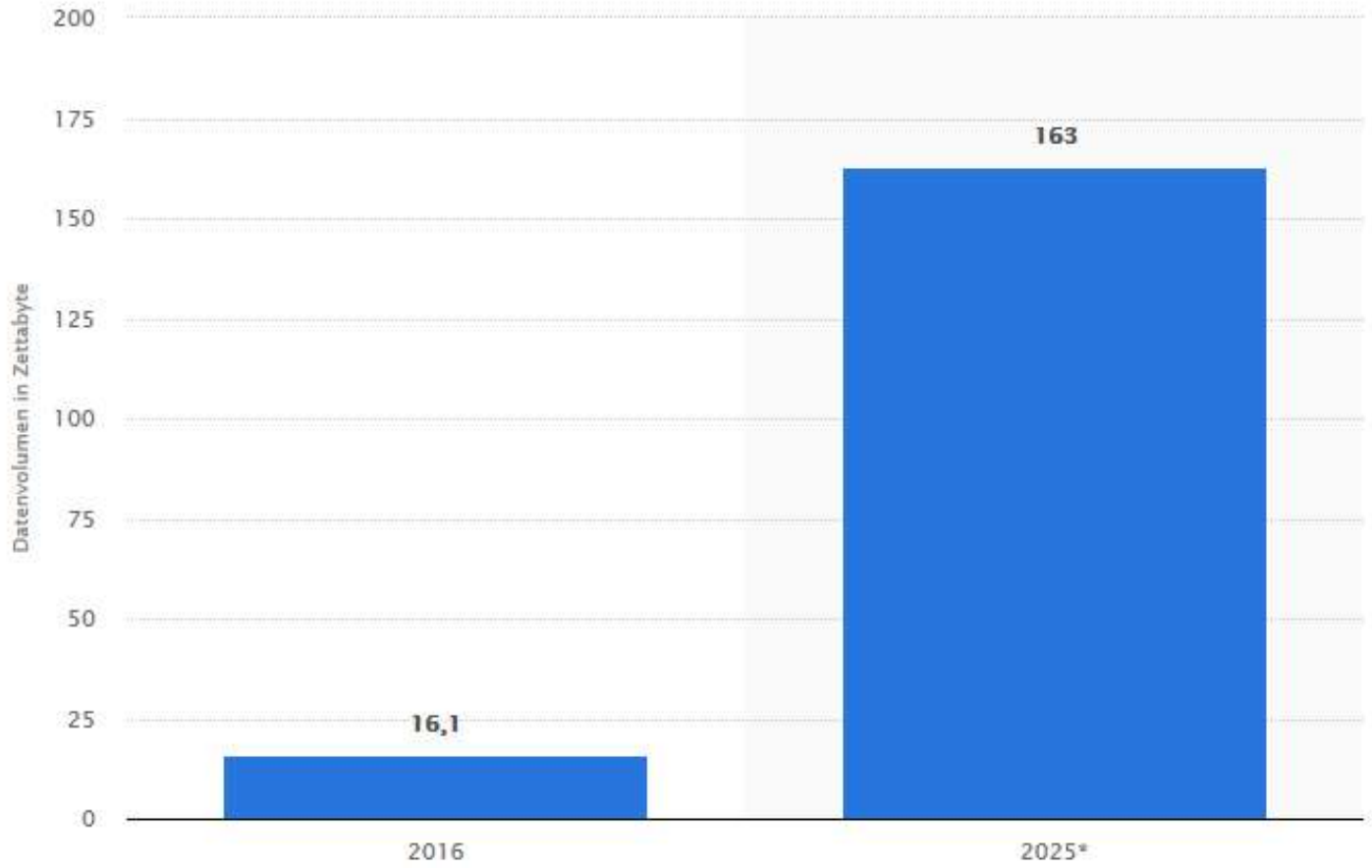
1. Daten, Information, Wissen
2. Logik und Fehlschlüsse
3. Korrelationen: Große Schuhe machen reich
4. Ausblick

Zeichen – Daten – Information - Wissen



Datenmengen

jährlich generierte digitale Datenmenge weltweit in Zettabyte = 10^{21} Byte = 1 Trillion GB



https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/

Woher stammen all diese Daten?

Vom Benutzer eingetippt

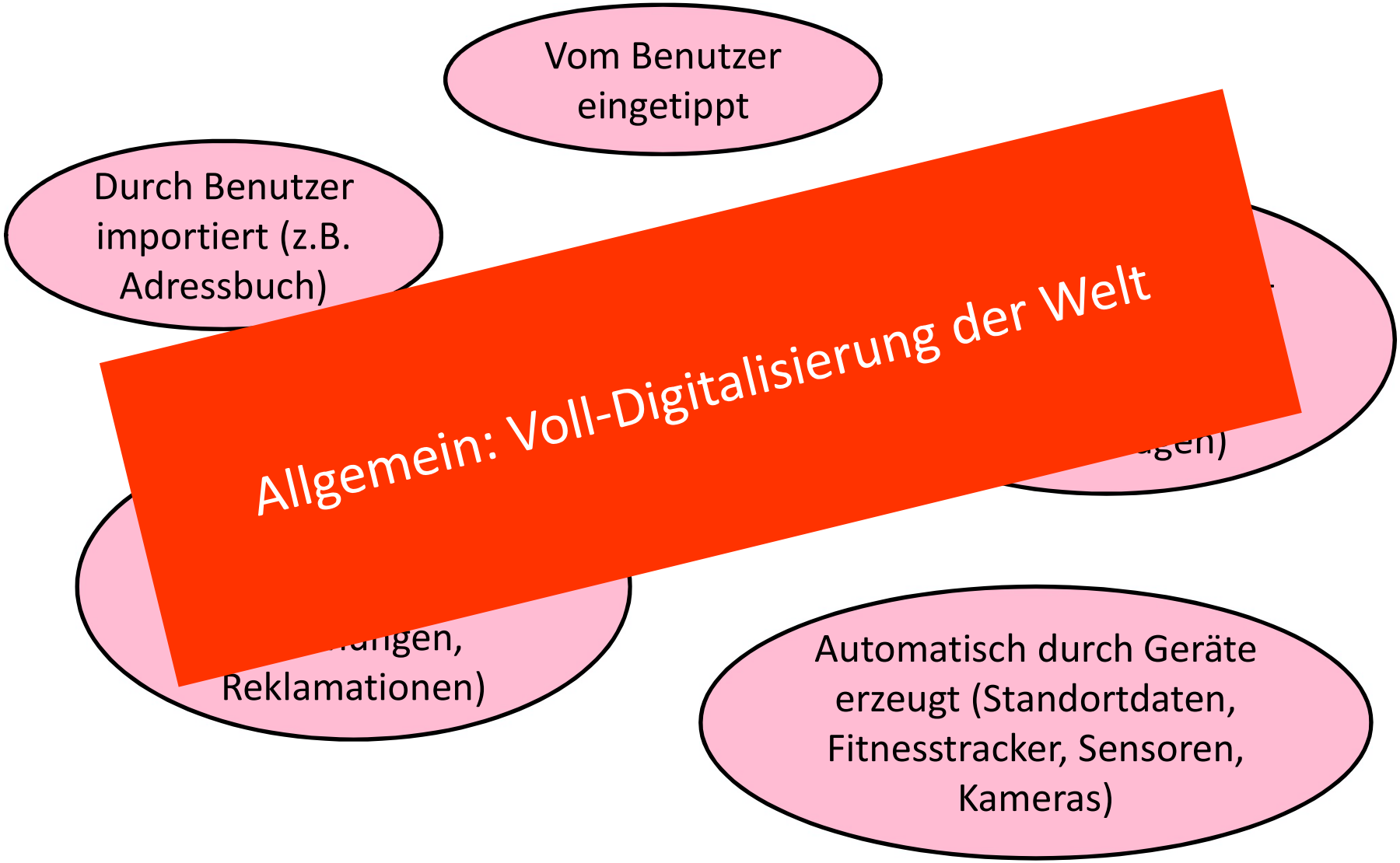
Durch Benutzer importiert (z.B. Adressbuch)

Durch Benutzer anderweitig erzeugt (Fotos, Videos, Sprachnachrichten, Suchanfragen)

Entstehen bei Geschäftstätigkeiten (Bestellungen, Rechnungen, Reklamationen)

Automatisch durch Geräte erzeugt (Standortdaten, Fitnesstracker, Sensoren, Kameras)

Woher stammen all diese Daten?



Ziel von Big Data

- Aus Daten Wissen machen!

Ziel von Big Data

- Aus Daten Wissen machen!
- Von vorhandenen Daten auf nicht Bekanntes schließen (Daten wie Alter, Geschlecht, Postleitzahl -> Kaufwünsche, Kaufkraft)
- Interpolation -> Trends und Zukunftsprognosen -> lohnende Investitionen

Ziel von Big Data

- Aus Daten Wissen machen!
- Dadurch entstehen Handlungen wie:
  - Mailverteiler für zielgerichtete Werbung für ein Produkt erstellen
  - Epidemien voraussagen -> Regierung kauft vermehrt Impfstoffe
  - Gerechtigkeit oder Stigmatisierung von Personen (Kreditsperre, Social Score)
  - Verdächtige Personen identifizieren und überwachen (Predictive Policing)

Problem: Statistischer Analphabetismus

„Es hat sich ein statistischer Analphabetismus verbreitet, der in Deutschland sogar hip ist. […]

Oft sind gar nicht die Statistiken an sich falsch, sondern sie sind falsch interpretiert oder die Ergebnisse falsch dargestellt.“

Prof. Thomas Bauer, Big Data: Chance oder Risiko? Interview für die RUB News, 2016
https://news.rub.de/wissenschaft/2016-09-28-interview-big-data-chance-oder-risiko

Übersicht

1. Daten, Information, Wissen
2. Logik und Fehlschlüsse
3. Korrelationen: Große Schuhe machen reich
4. Künstliche Dummheiten

Formen von Daten und Informationen

Daten

- Quantitativ (17,5°C)
- Qualitativ

Informationen

- Aussagen (A ist wahr; oder: Wenn A, dann B)
- Korrelationen (je stärker A, umso stärker B)
- Bedingte Wahrscheinlichkeit (wenn A, dann mit 90%-Wahrscheinlichkeit auch B)

Aussagen-Logik

- Aussagen können wahr oder falsch sein, in der Fuzzy-Logik auch teilweise wahr
- A → B: Aus Aussage A = „wahr“ folgt immer Aussage B = „wahr“
- Beispiel:
  - A = Die Kühlschranktür ist offen
  - B = Das Licht im Kühlschrank ist an

	B = „wahr“	**B = „falsch“ bzw. ¬B**
A = „wahr“	Richtig, kann ich sehen und testen	Es liegt ein Defekt vor
A = „falsch“ bzw. ¬A	Das wäre ein Defekt, aber kann ich nicht prüfen	Wäre richtig, aber ich kann es nicht überprüfen

Logik und Fehlschlüsse

- Ich weiß: **A -> B** (Immer wenn die Kühlschranktür offen ist, dann ist auch das Licht an.)
- Fehlschluss: **B -> A** (Immer wenn das Licht an ist, ist die Kühlschranktür offen.)
- Fehlschluss: **¬A -> ¬B** (Immer wenn die Tür zu ist, ist auch das Licht aus.)

	B = „wahr“	B = „falsch“ bzw. ¬B
A = „wahr“	Richtig, kann ich sehen und testen	Es liegt ein Defekt vor
A = „falsch“ bzw. ¬A	Das wäre ein Defekt, aber kann ich nicht prüfen	Wäre richtig, aber ich kann es nicht überprüfen

Achtung: Kausale Zusammenhänge sind selten

- **Kausaler Zusammenhang**: A verursacht B (A -> B)
- Häufiger: **bedingte Wahrscheinlichkeit oder Korrelation**

Beispiel Rasterfahndung

- Wenn A und B und C und D, dann E (Person ist verdächtig)
- Einfache Datenbankabfrage: SELECT * FROM Personen WHERE (bezahlt Miete in bar=‚wahr‘) AND (Kindergeld = ‚falsch‘) AND (Geschlecht=‚m‘) AND ((Alter >= 18) AND (Alter <= 40))
- Die Kriterien werden nie alle Terroristen vollständig identifizieren, sondern optimieren die Balance zwischen Aufwand und Effektivität

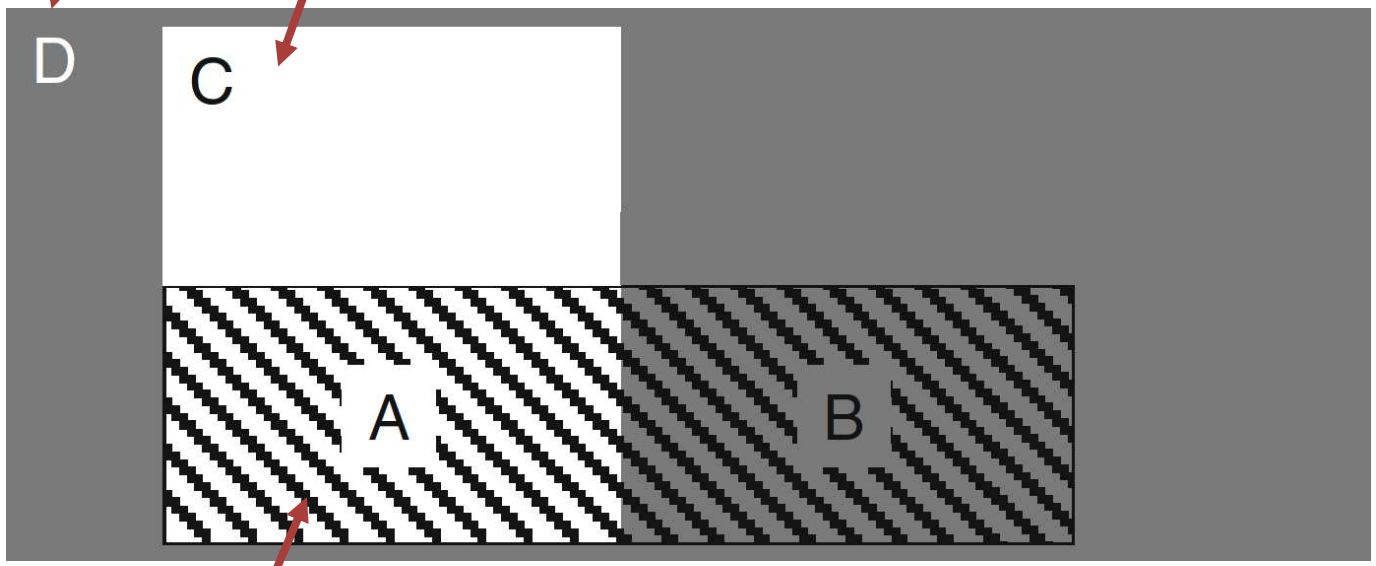
Qualität der Suchkriterien: recall und precision

Recall = A / (A + C)

Precision = A / (A + B)

D Suchuniversum

C relevante Elemente, die nicht gefunden wurden



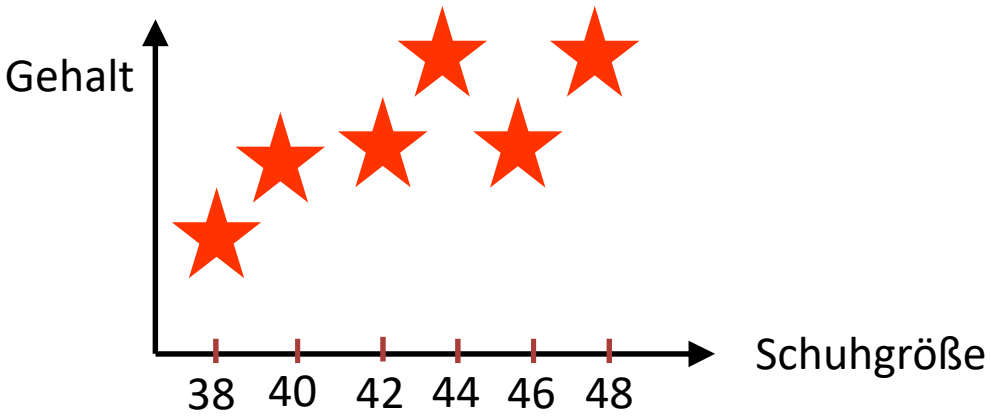
(A+B) Suchergebnisse, davon A relevant, B irrelevant

Übersicht

1. Daten, Information, Wissen
2. Logik und Fehlschlüsse
3. Korrelationen: Große Schuhe machen reich
4. Ausblick

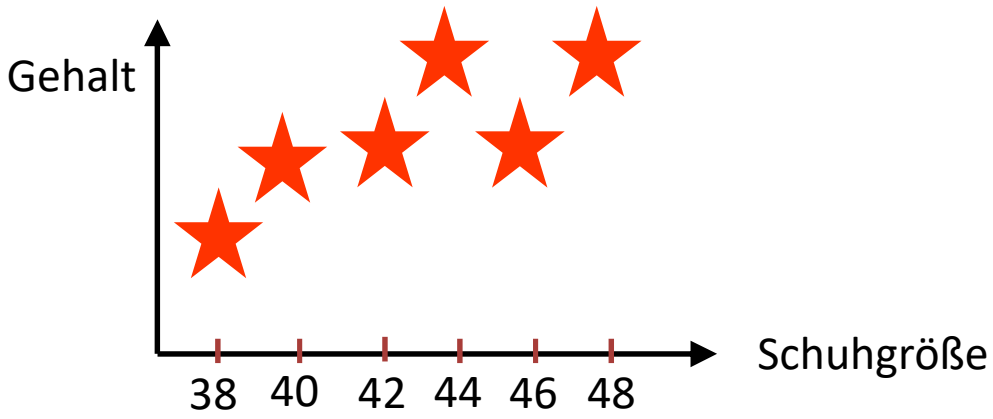
Menschen mit großen Schuhen verdienen mehr

Daten:

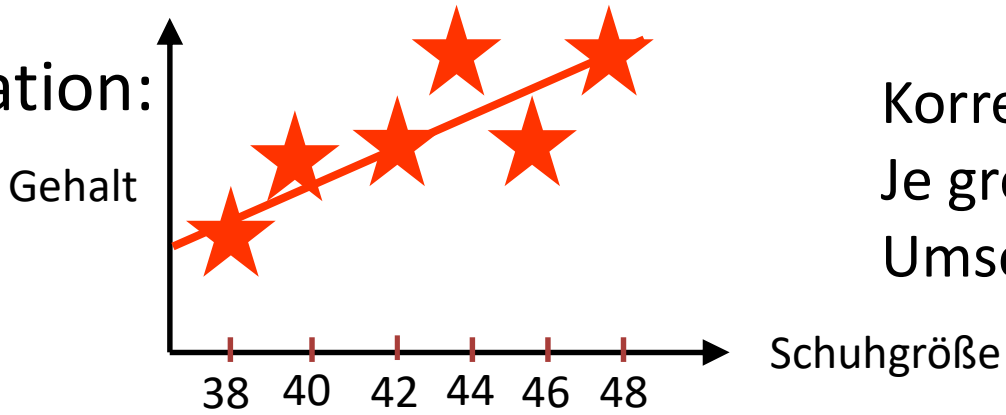


Menschen mit großen Schuhen verdienen mehr

Daten:



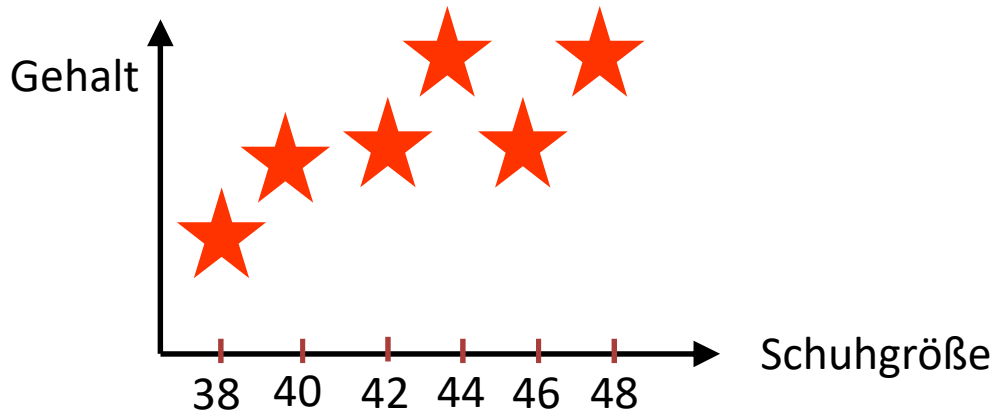
Information:



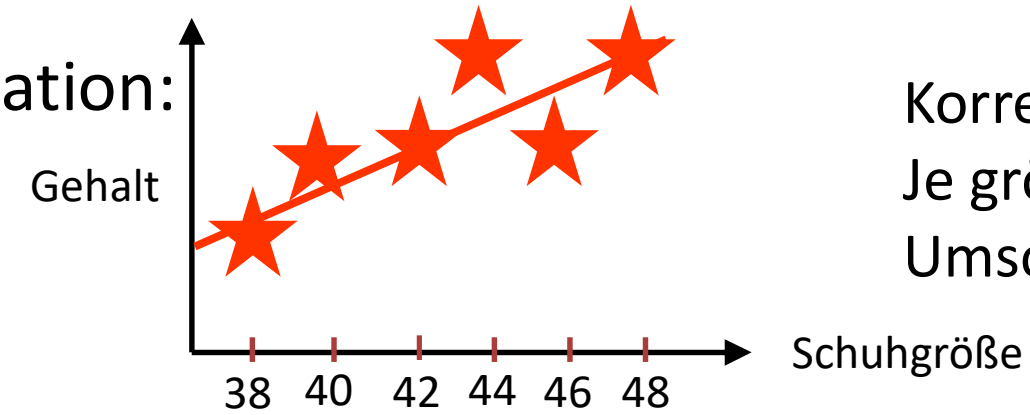
Korrelation:
Je größer die Schuhe,
Umso mehr Gehalt.

Menschen mit großen Schuhen verdienen mehr

Daten:



Information:



Korrelation:
Je größer die Schuhe,
Umso mehr Gehalt.

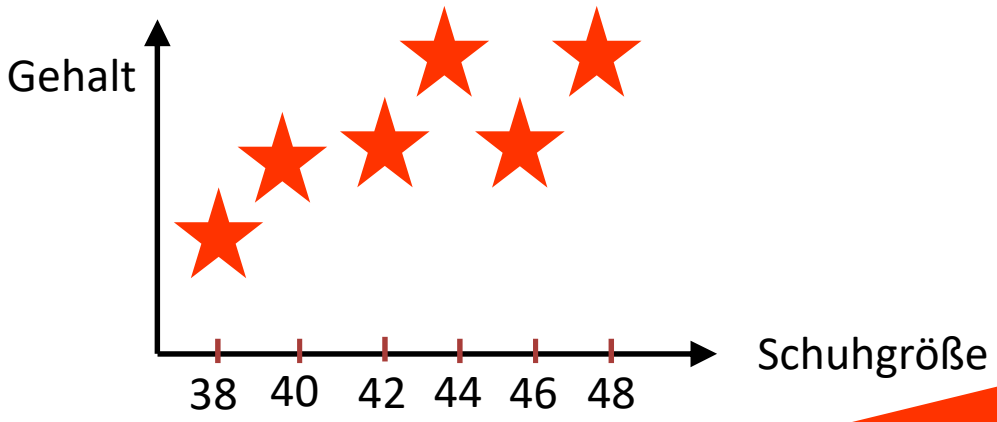
Wissen: Mitarbeiter: Wenn ich größere Schuhe anziehe, bekomme ich eine Gehaltserhöhung!
Arbeitgeber: Mitarbeiter mit kleinen Schuhen kosten weniger!

Logik versus Korrelation

- In der **Logik** geht es um Aussagen und Regeln („Immer wenn“), zwangsläufig verursacht durch
  - Gesetzmäßigkeiten (z.B. Naturgesetz)
  - Ursache-Wirkung (z.B. der Kühlschrank enthält einen Mechanismus, der beim Öffnen das Licht einschaltet)
- **Korrelationen** ermitteln Häufigkeiten von Zusammentreffen, verursacht durch
  - Vorlieben, Lebensstil, Empfehlungen, Moden,...
  - Z.B. Schwangere Frauen kaufen eine bestimmte Hautcreme mit höherer Wahrscheinlichkeit als nichtschwangere Frauen

Menschen mit großen Schuhen verdienen mehr

Daten:

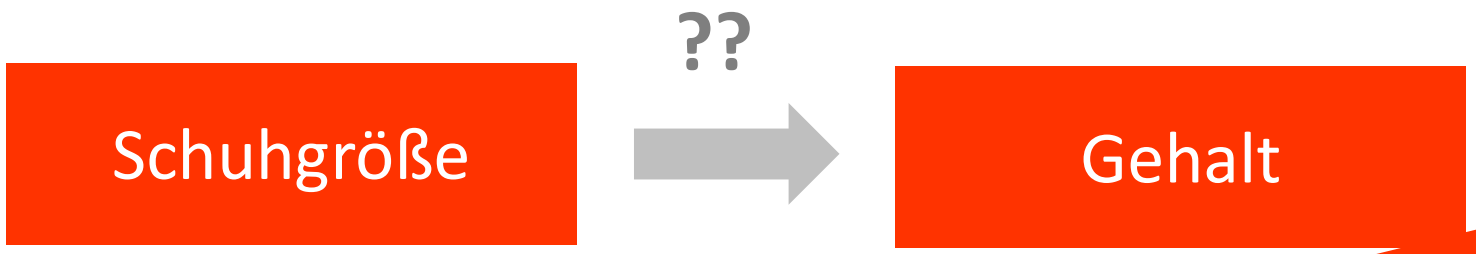


Information:

Denkfehler:
Korrelation ≠ Ursache-Wirkungs-Beziehung

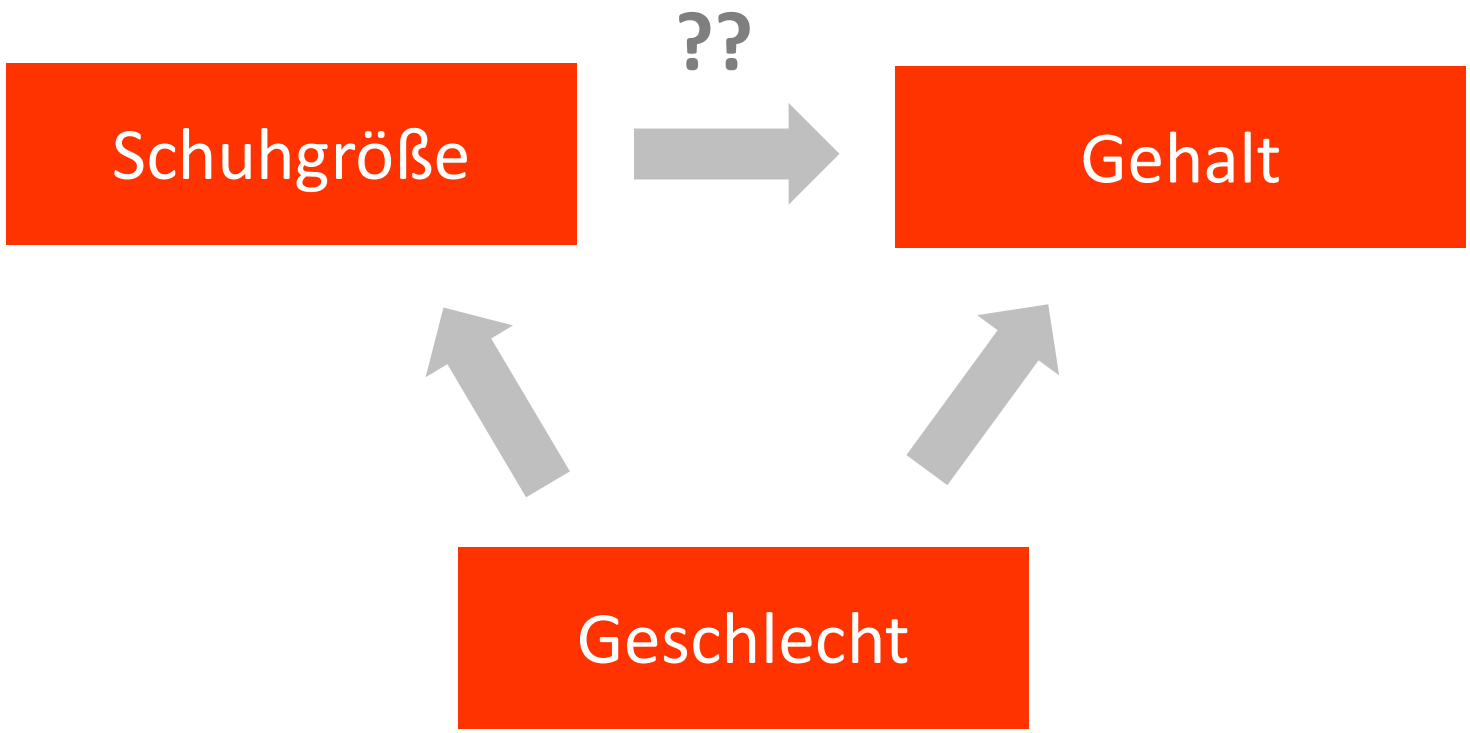
Wisse ich größere Schuhe anziehe,
ne ich eine Gehaltserhöhung!
Arbeitgeber: Leute mit kleinen Schuhen kosten weniger!

Menschen mit großen Schuhen verdienen mehr



Denkfehler:
Korrelation ≠ Ursache-Wirkungs-Beziehung
Scheinkorrelation: dritte Variable = Geschlecht

Menschen mit großen Schuhen verdienen mehr



Schokolade macht intelligent!

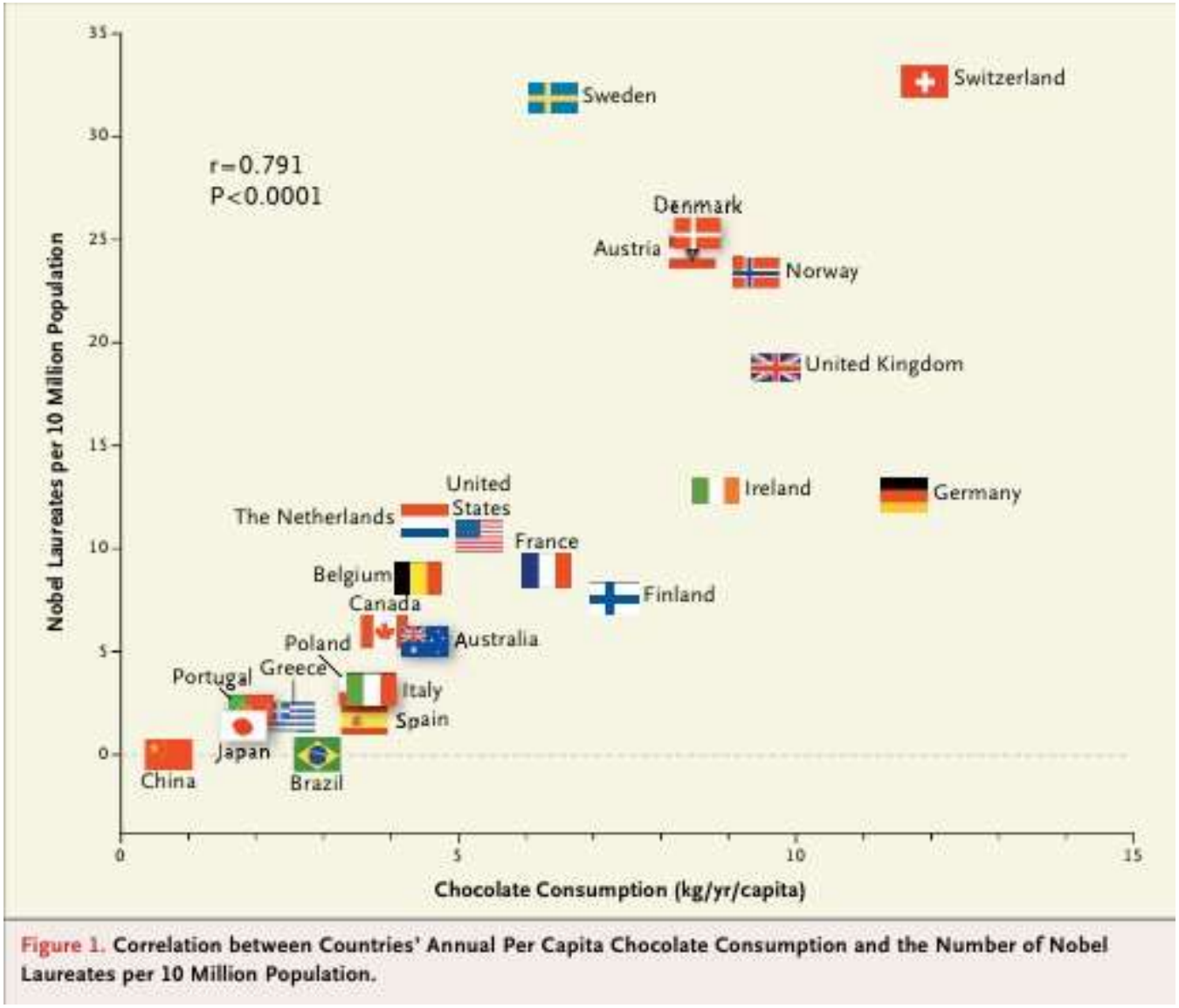
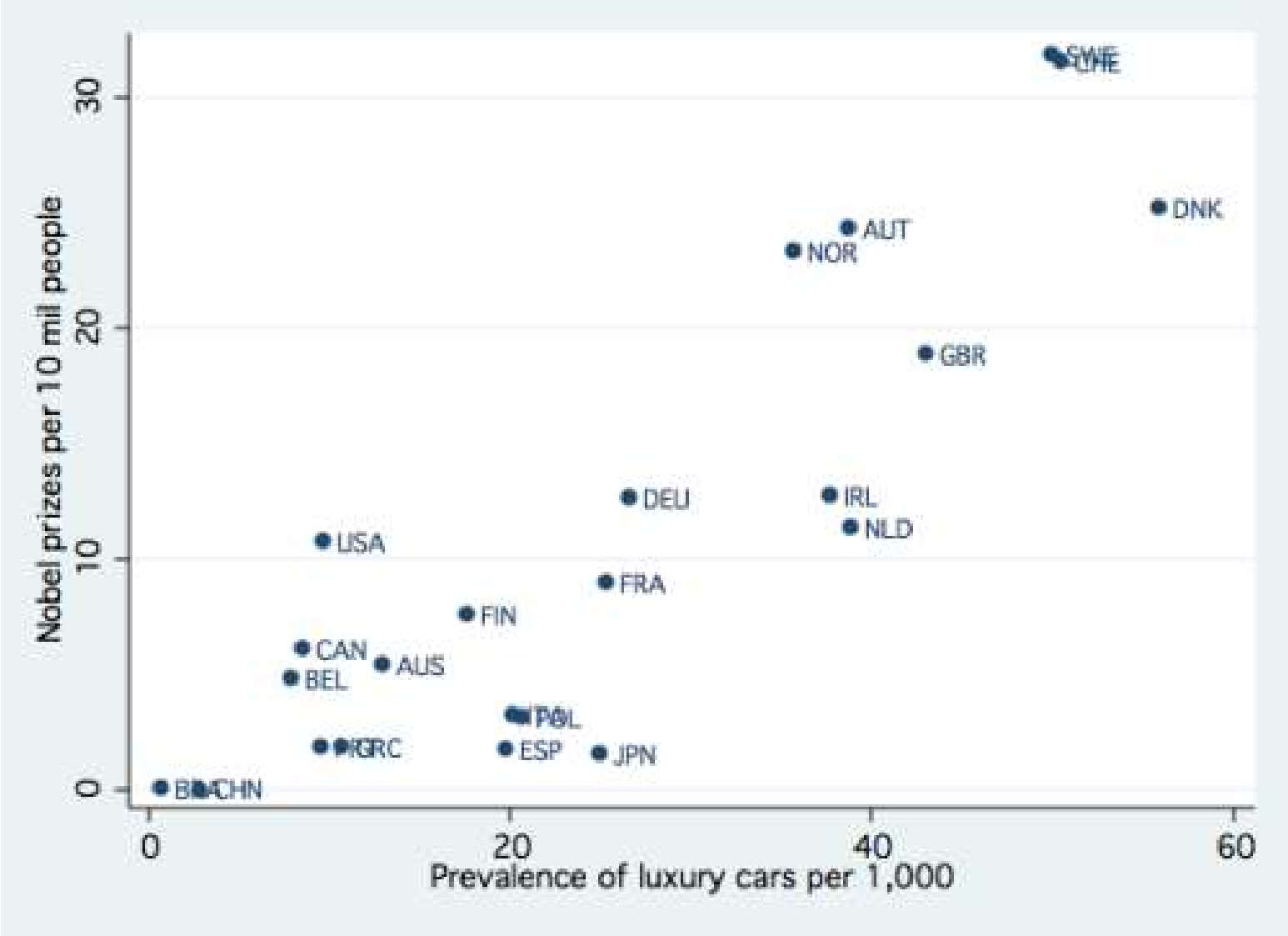


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

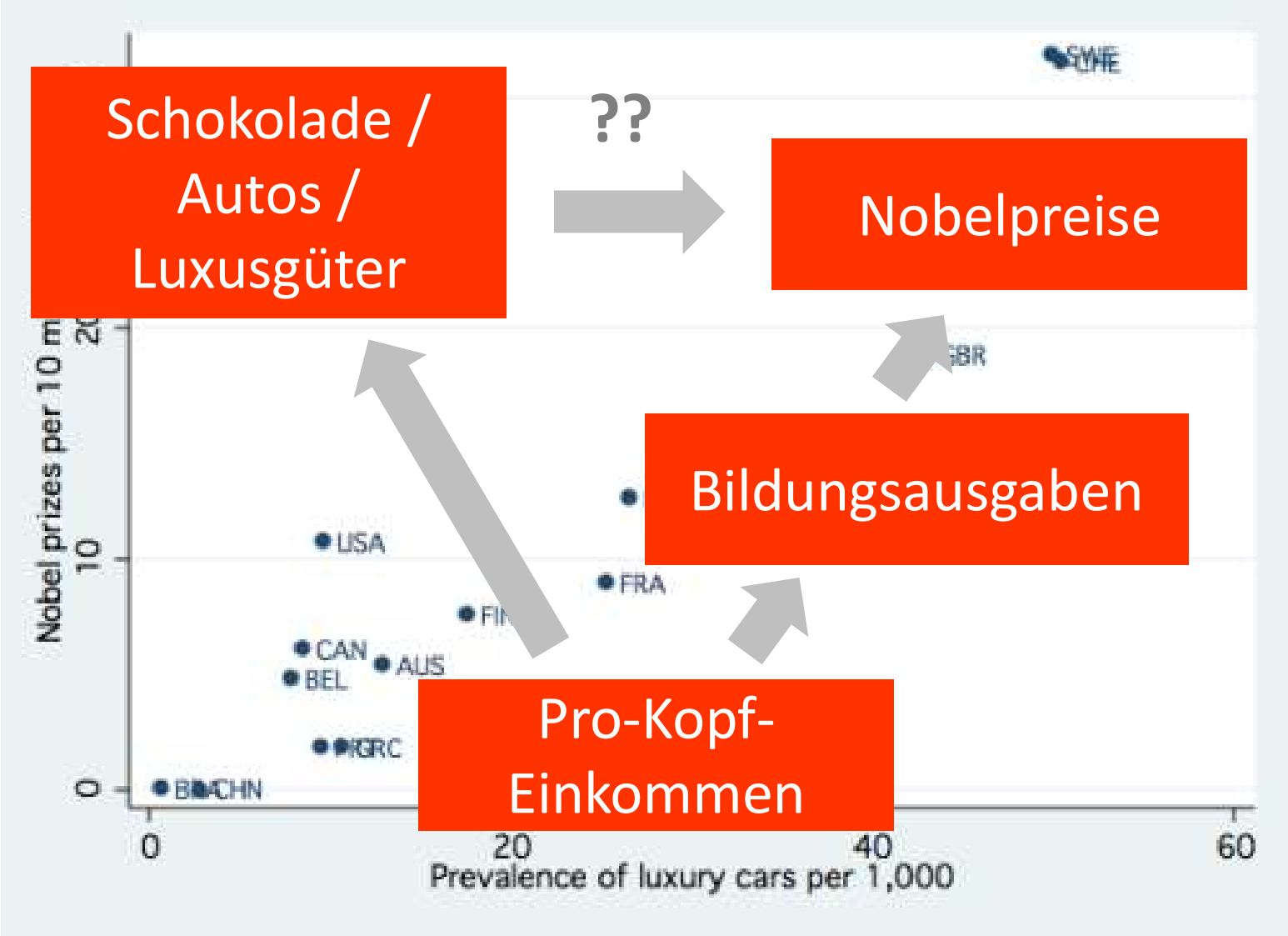
Messerli, Franz H. (2012). *Chocolate Consumption, Cognitive Function, and Nobel Laureates*. New England Journal of Medicine, 367:16, 1562-1564

Teure Autos machen intelligent??



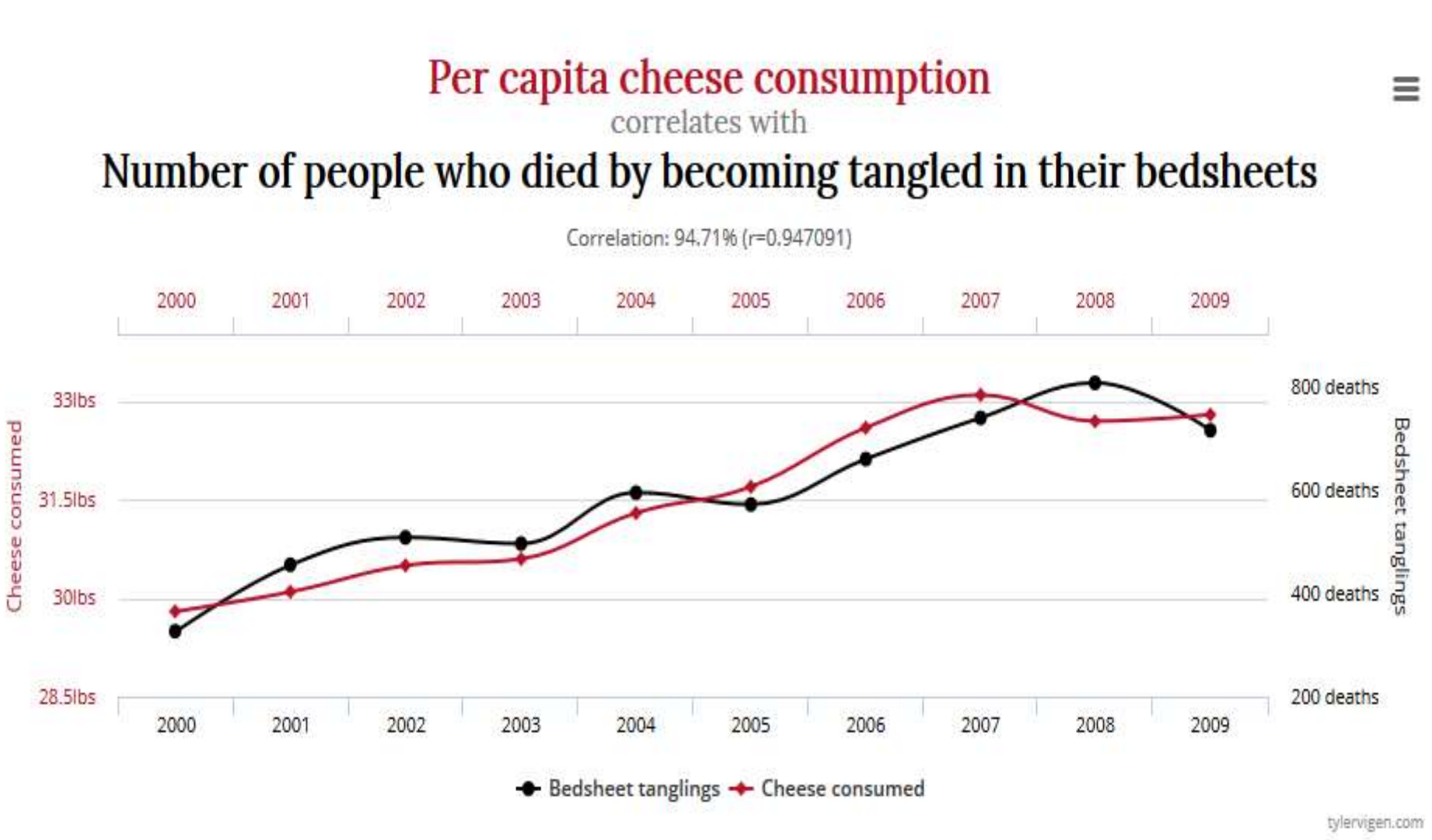
https://epianalysis.wordpress.com/2012/11/19/chocolate/

Teure Autos machen intelligent??



https://epianalysis.wordpress.com/2012/11/19/chocolate/

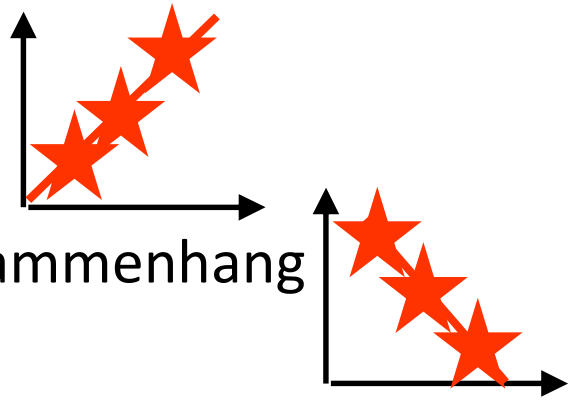
Zufällige Korrelationen sind möglich!



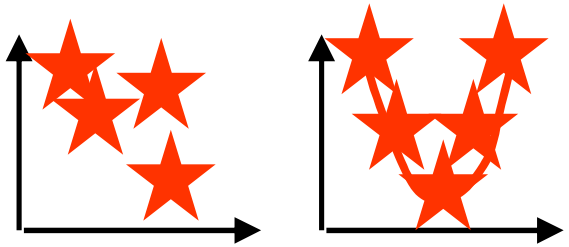
http://www.tylervigen.com/spurious-correlations

Pearson-Korrelationsfaktor

- Misst die Stärke eines linearen Zusammenhangs zwischen zwei Merkmalen
- Wertebereich: $-1 \leq r \leq 1$:
  - $r = 1$ → linearer Zusammenhang
  - $r = -1$ → gegenläufiger linearer Zusammenhang

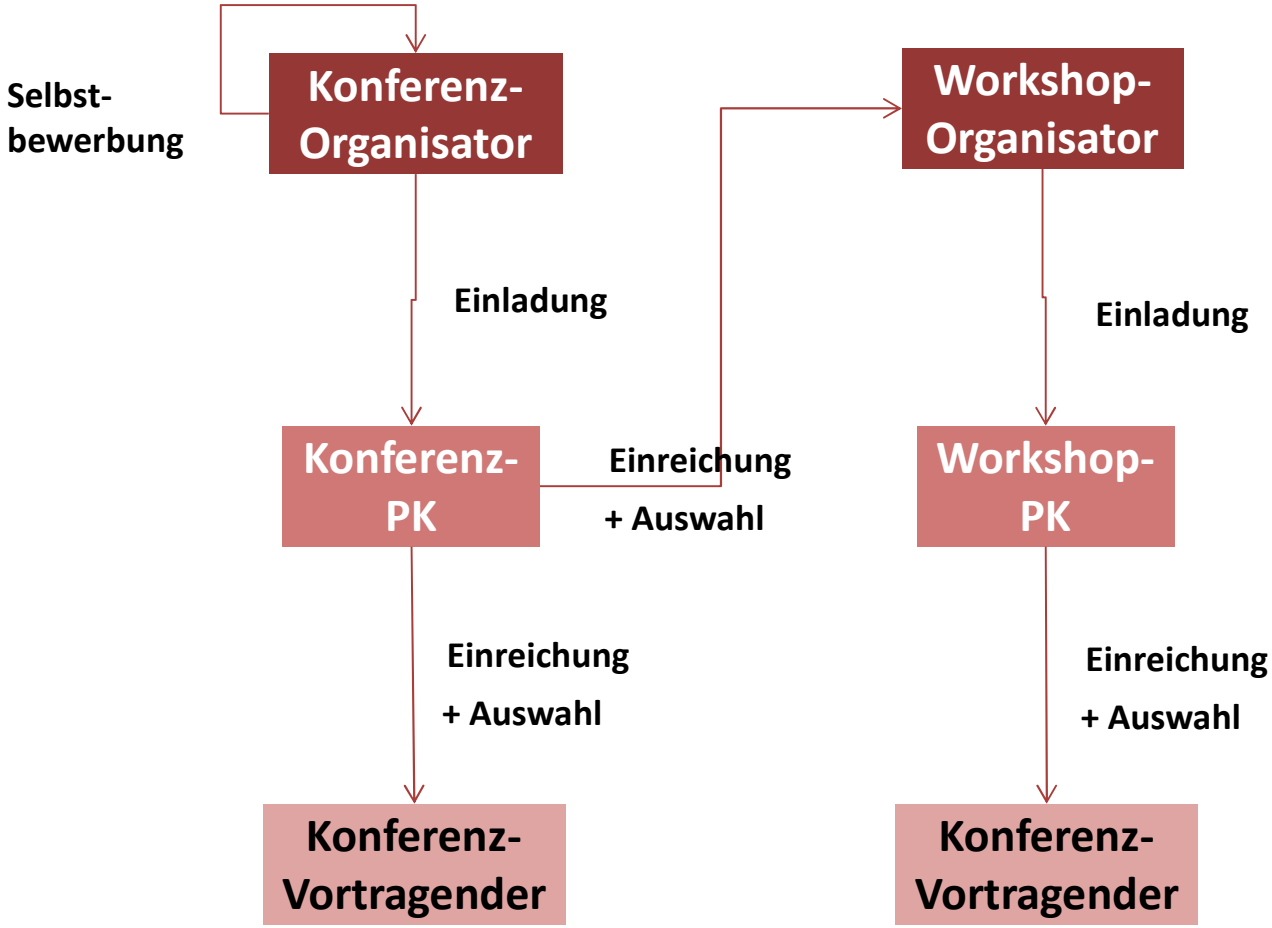


  - $1 < r < 0$ → Werte lassen sich mehr oder weniger gut durch eine Gerade annähern
  - $r = 0$ → kein linearer Zusammenhang zwischen den Variablen (nichtlineare Abhängigkeiten können aber bestehen)



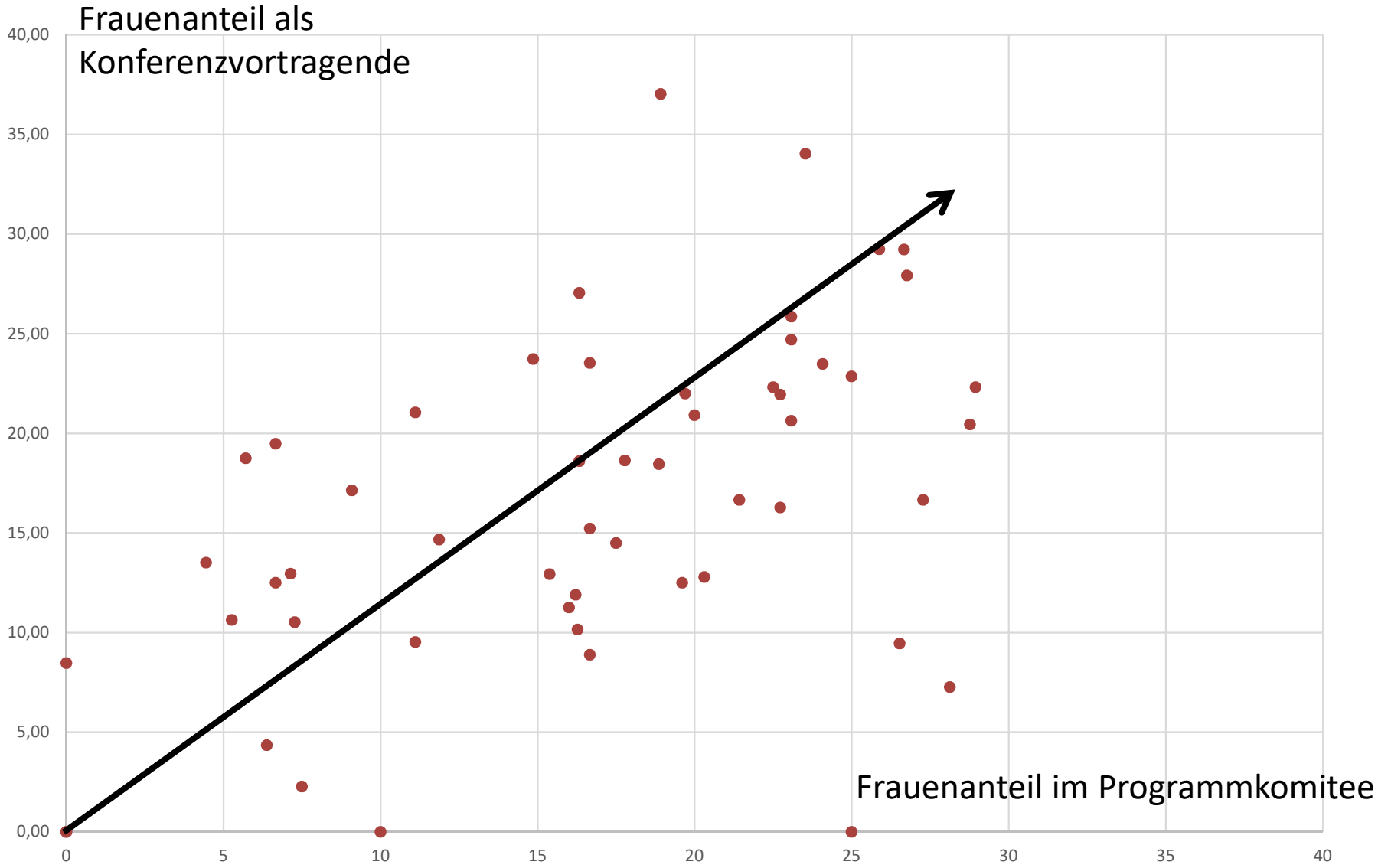
Beispiel: Korrelationsanalyse

Frage: Fördern Frauen Frauen?



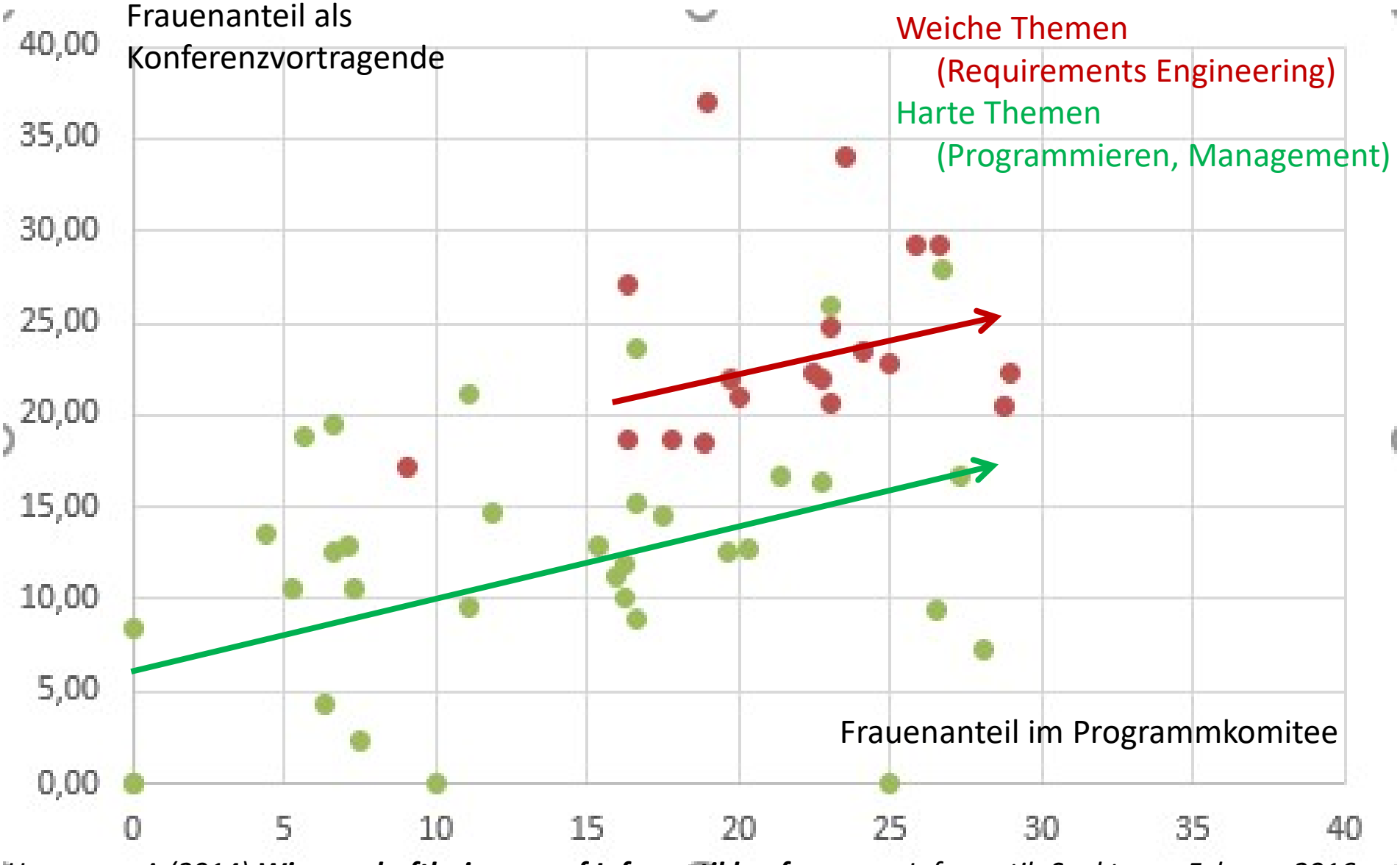
PK = Programmkomitee

Fördern Frauen Frauen?



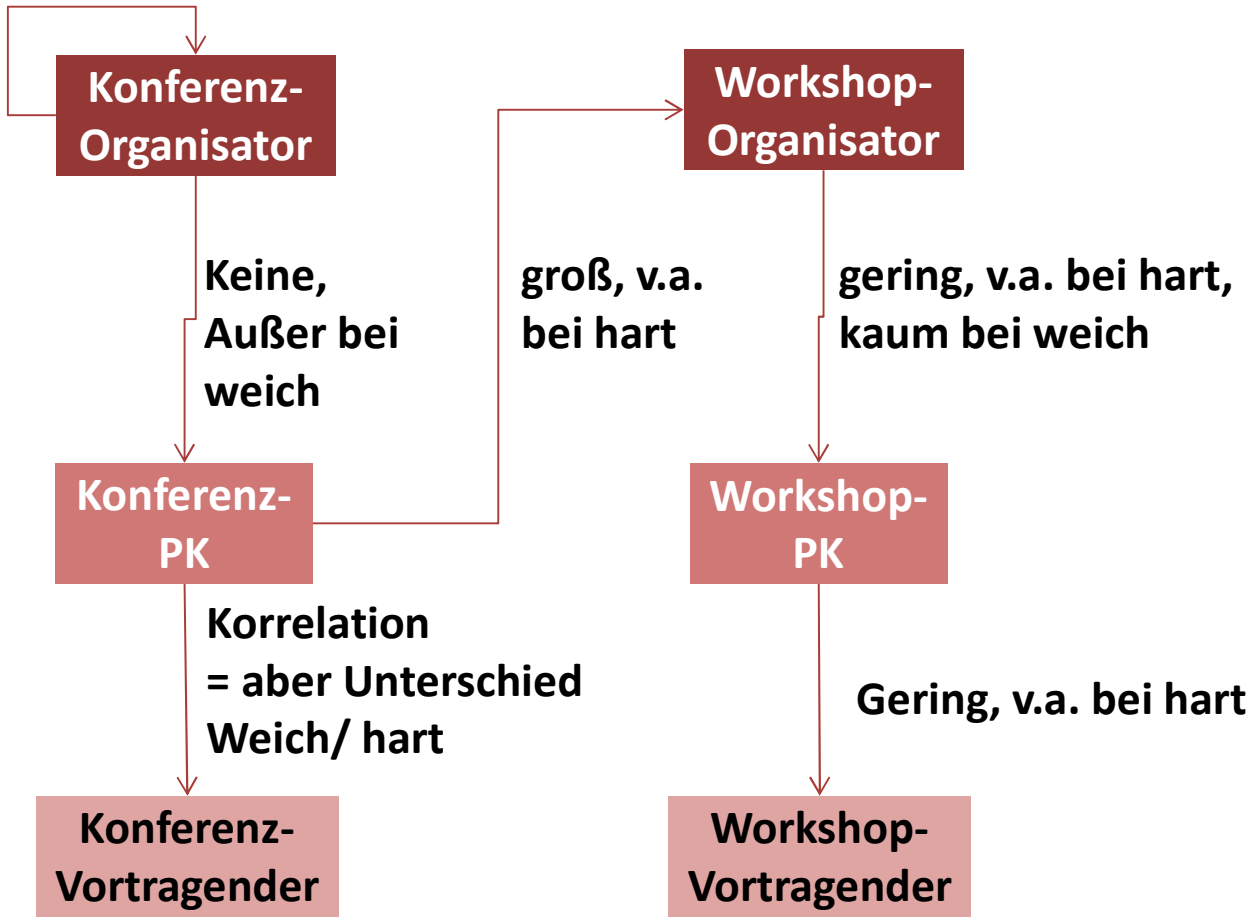
Herrmann A (2014) ***Wissenschaftlerinnen auf Informatikkonferenzen****. Informatik-Spektrum, Februar 2016, Volume 39, Issue 1, S. 38-56, DOI 10.1007/s00287-014-0839-8*

Fördern Frauen Frauen?



Herrmann A (2014) ***Wissenschaftlerinnen auf Informatikkonferenzen****. Informatik-Spektrum, Februar 2016, Volume 39, Issue 1, S. 38-56, DOI 10.1007/s00287-014-0839-8*

Korrelationen qualitativ



Übersicht

1. Daten, Information, Wissen
2. Logik und Fehlschlüsse
3. Korrelationen: Große Schuhe machen reich
4. Ausblick

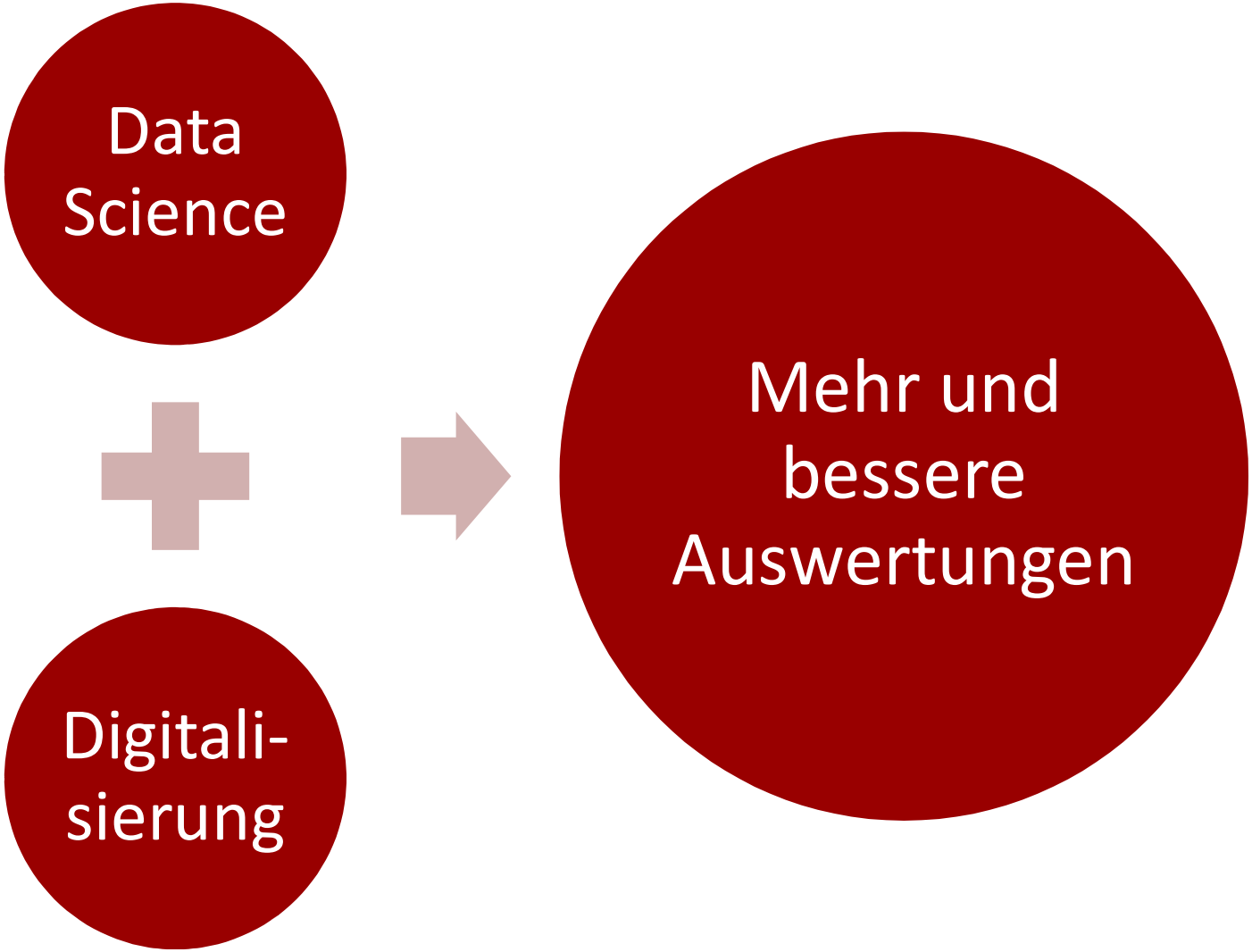
Zusammenfassung

- Der Weg von Daten zu Wissen ist mit Unsicherheiten und Fallen gepflastert.
- Daten ≠ Realität
- Korrelation ≠ Erklärung oder Ursache-Wirkung
- Gefahren: Scheinkorrelationen und Fehlschlüsse
- Ungenauigkeiten sind je nach Anwendungsbereich mehr oder weniger akzeptabel
- Daten korrekt auszuwerten verlangt Kompetenz und Disziplin

Fazit

- Traue nur der Statistik, die du selbst erstellt hast.
- A fool with a tool is a faster fool.
- Korrelationen nicht zu ernst nehmen
- Mehr Daten (Big Data) -> bessere Auswertungen?

Ausblick


Data Science
Digitali-sierung
Mehr und bessere Auswertungen

Literatur: Unstatistiken

- Gerd Gigerenzer: Das Einmaleins der Skepsis - Über den richtigen Umgang mit Zahlen und Risiken. Berliner Taschenbuchverlag, 3. Auflage 2007
- Thomas Bauer, Gerd Gigerenzer, Walter Krämer: Warum dick nicht doof macht und Genmais nicht tötet – Über Risiken und Nebenwirkungen der Unstatistik. Campus, 2014
- aktuelle Unstatistiken monatlich hier: http://www.rwi-essen.de/unstatistik/
- Tyler Vigen: Spurious Correlations http://www.tylervigen.com/spurious-correlations
- Wolfgang Walla: Wie man sich durch statistische Grafiken täuschen lässt. Statistisches Landesamt Baden-Württemberg, 2008, https://www.destatis.de/GPStatistik/servlets/MCRFileNodeServlet/BWMonografie_derivate_00000082/8020_08001.pdf
- Waltraud Kahle: Lügt Statistik? Über den sorglosen Umgang mit Daten http://www.math.uni-magdeburg.de/~wkahle/luegtstat.pdf