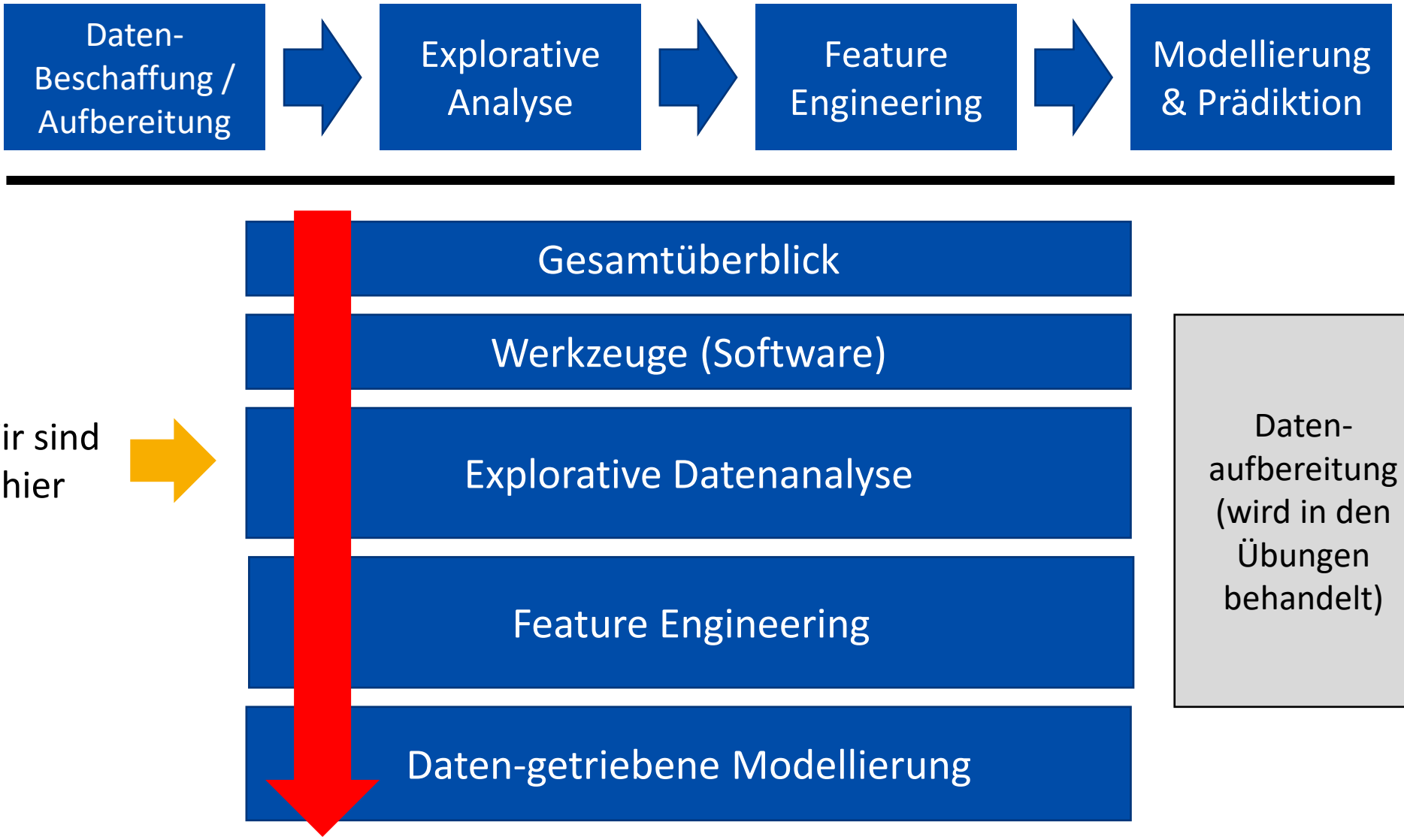
Einführung in Data Science

Unser Plan für heute:

1. Wiederholung

2. Multivariate explorative Analyse – Teil 2
   - Zusammenhangsmaße:
     Spearman Korrlationskoeffizient,
     Mutual Information
   - Interpretationsfehler

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

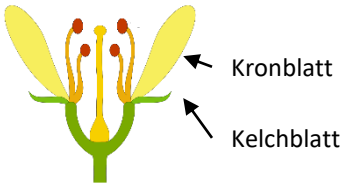
M‘variate Explorative Analyse | Visualisierung

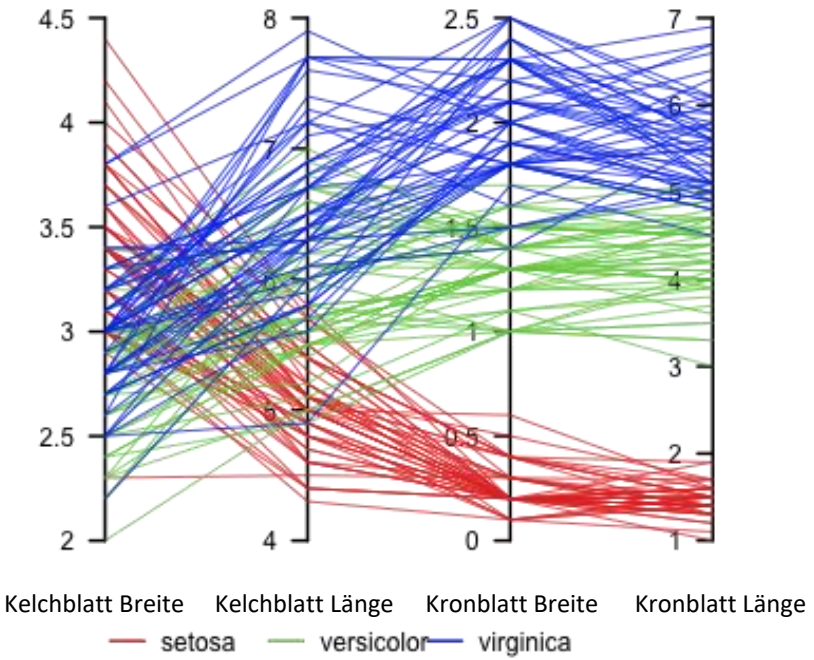
Parallel Coordinate Plots (PCPs, auch genannt: ||-Koordinaten)

- Technik zur Visualisierung multivariater Daten
  (Henry Gannett, 19. Jahrhundert; Inselberg (1985) Wegman (1990))
- hilfreich bei Exploration höherdimensionaler Räume
- Koordinaten werden auf parallelen Achsen dargestellt

Beispiel: 4-dimensionaler Raum

- Schwertlilien-Datensatz:
  50 Pflanzen mit 4 Merkmale
  (Features)
- 3 Schwertlilienarten:
  Setosa, Versicolor, Virginia





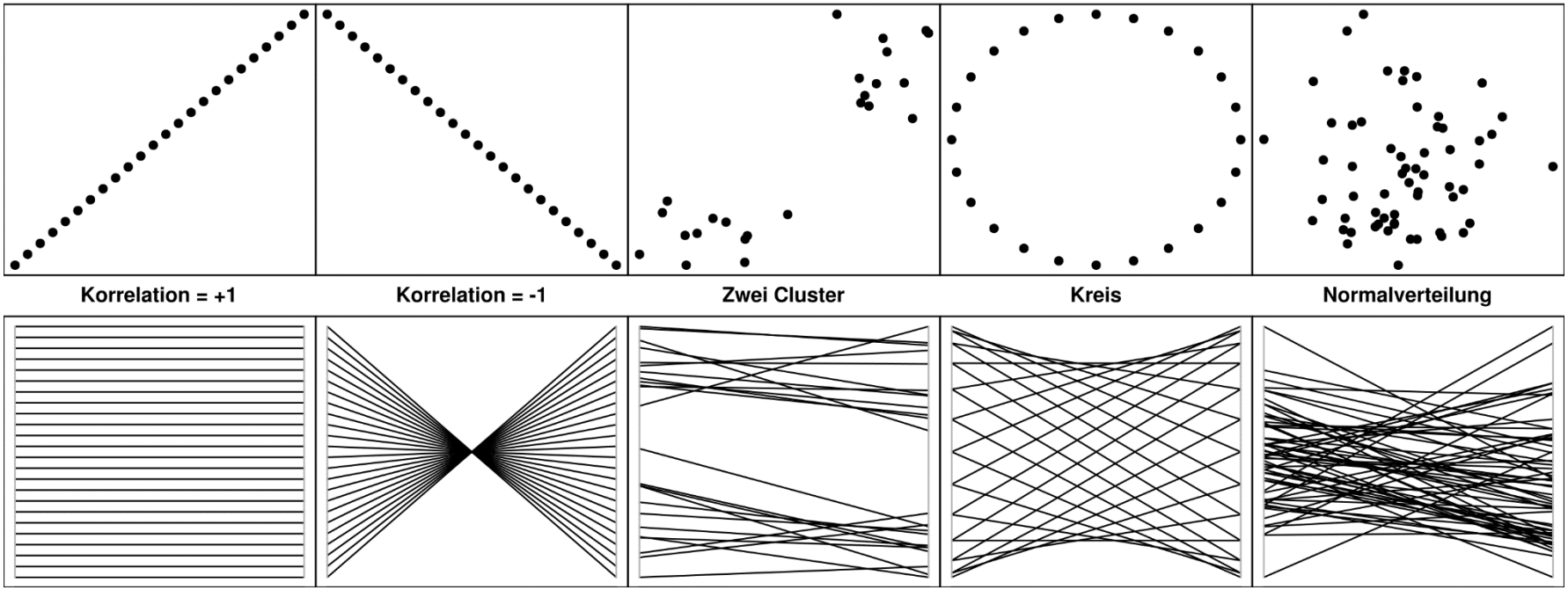
A. Inselberg, Vis. Comp. 1, 69-91 (1985); E.J. Wegman, J. Am. Stat. Assoc. 85, 664-675 (1990)

M‘variate Explorative Analyse | Visualisierung

Parallel Coordinate Plots

Höherdimensionale Strukturen werden in PCPs erkennbar:

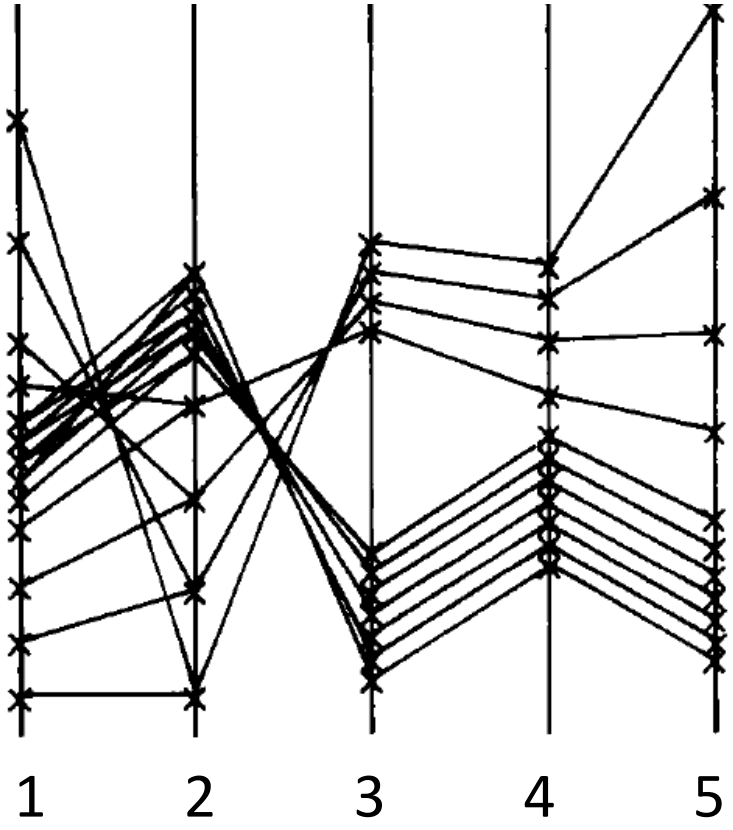
Streudiagramme



Parallel Coordinate Plots

M‘variate Explorative Analyse | Visualisierung

Parallel Coordinate Plots



Frage
Welche Strukturen erkennen Sie im dargestellten PCP?

Frage
Welche Faktoren beeinflussen die Visualisierung des PCPs?

Quelle: E.J. Wegman, J. Am. Stat. Assoc. 85, 664-675 (1990)

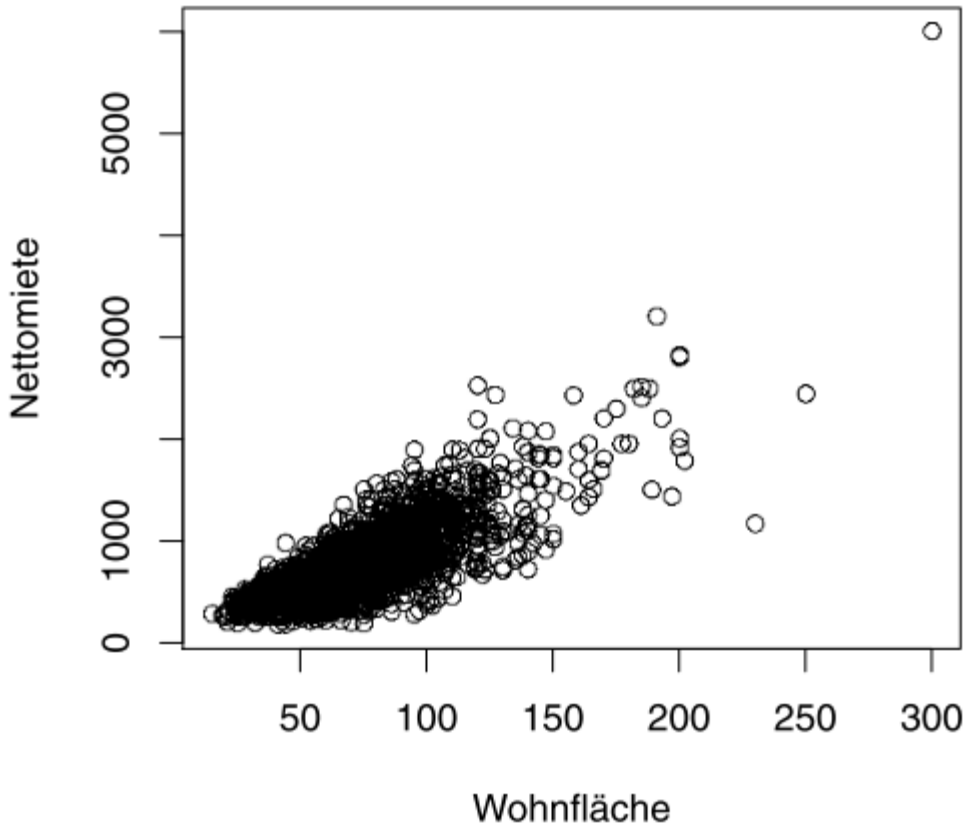
M‘variate Explorative Analyse | Zusammenhangsmaße

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)

Spearmans Korrelationskoeffizient

Spearmans Korrelationskoeffizient

- charakterisiert Stärke monotoner Zusammenhänge

Frage: Was ist ein monotoner Zusammenhang?

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:

$$rg(x_i) = i$$

Beispiel:

x_i	2.17	8.00	1.09	2.01
$rg\,(x_i)$	3	4	1	2

Wir übersetzen ebenfalls die Werte y_i in ihre Ränge $rg(y_i)$.

Spearmans Korrelationskoeffizient

Umgang mit identischen Werten („Ties“, „Bindungen“)

- Bei identischen Werten: Rangvergabe nicht mehr eindeutig.
- Stattdessen: Bilden des Durchschnittsrangs
  D.h.: identischen Werten wird als Rang das arithmetische Mittel der infrage kommenden Ränge zugewiesen

Beispiel

x_i	1.09	2.17	2.17	2.17	3.02	4.5
Zwischenschritt: Zuweisung von Rängen (im Falle von Ties: zufällig)	1	3	2	4	5	6
Ermittlung des Durchschnittsrangs für Ties		$(2+3+4)/3 = 3$				
$rg(x_i)$	1	3	3	3	5	6

Spearmans Korrelationskoeffizient

Spearmans Korrelationskoeffizient

- entspricht Pearsons Korrelationskoeffizient, angewandt auf die Rangpaare

$$r_{SP} = \frac{\sum(rg\,(x_i) - \bar{rg}_X)(rg\,(y_i) - \bar{rg}_Y)}{\sqrt{\sum(rg\,(x_i) - \bar{rg}_X)^2 \sum(rg\,(y_i) - \bar{rg}_Y)^2}}$$

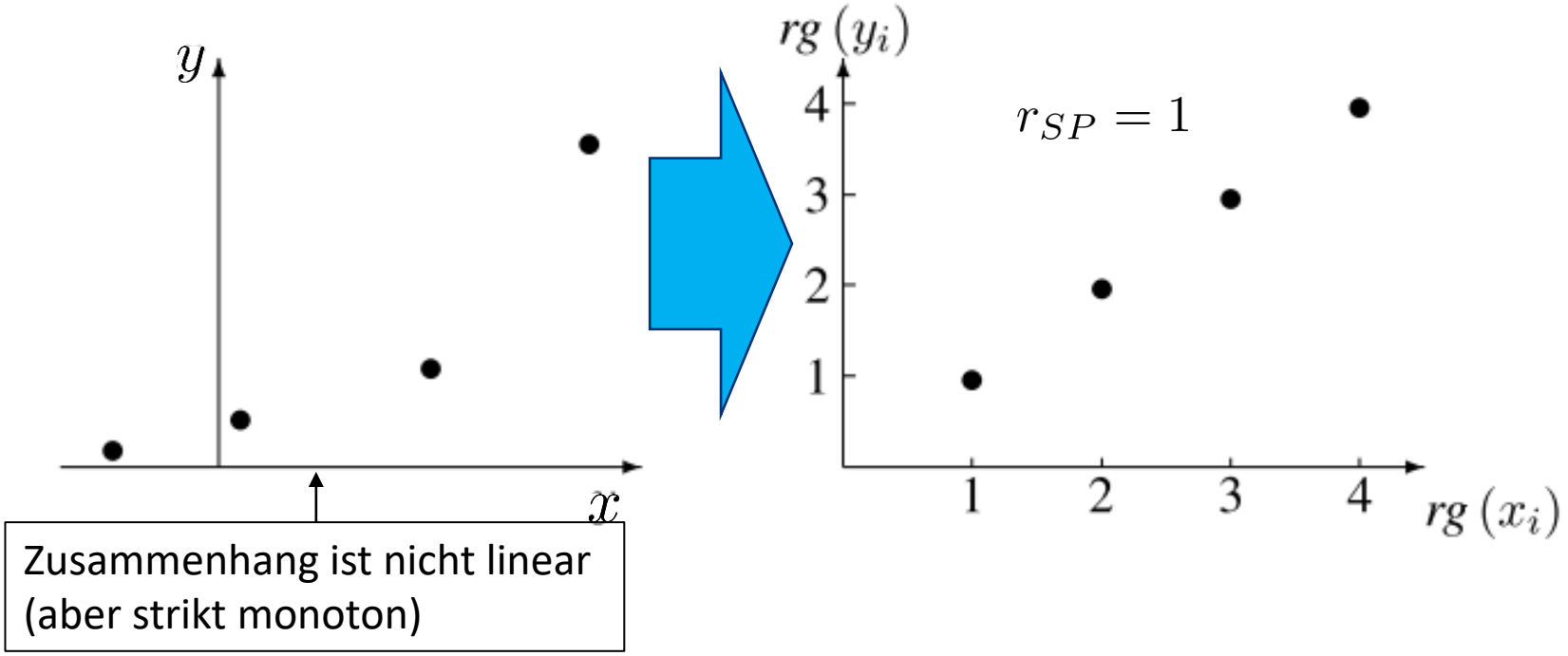
mit folgenden Mittelwerte der Ränge:

$$\bar{rg}_X = \frac{1}{n}\sum_{i=1}^{n} rg\,(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2$$

$$\bar{rg}_Y = \frac{1}{n}\sum_{i=1}^{n} rg\,(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2$$

Spearmans Korrelationskoeffizient

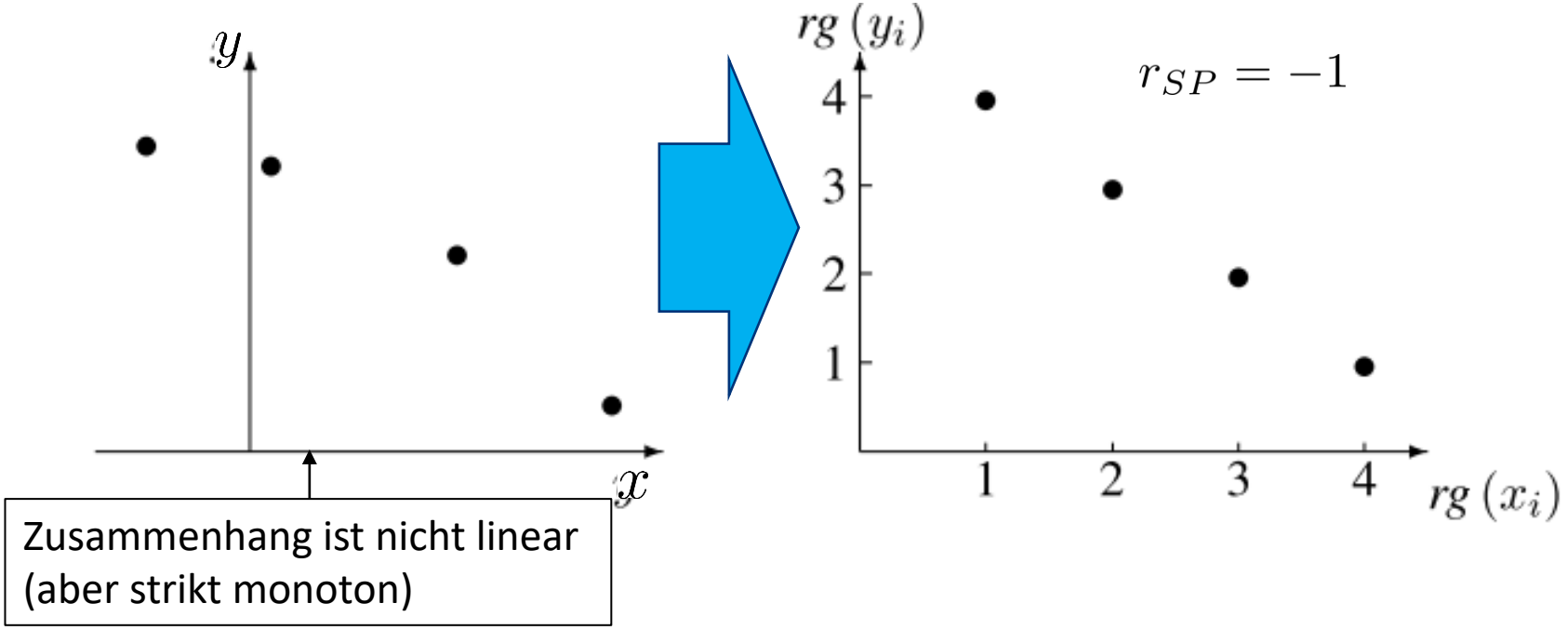
- $-1 \leq r_{SP} \leq 1$
- $r_{SP} = 1$ sofern Rangpaare $(rg(x_i), rg(y_i))$ auf einer Geraden positiver Steigung liegen



Spearmans Korrelationskoeffizient misst Stärke des monotonen Zusammenhangs zwischen x und y.

Spearmans Korrelationskoeffizient

- $-1 \leq r_{SP} \leq 1$
- $r_{SP} = -1$ sofern Rangpaare $(rg(x_i), rg(y_i))$ auf einer Geraden negativer Steigung liegen



Spearmans Korrelationskoeffizient misst Stärke des monotonen Zusammenhangs zwischen x und y.

Korrelationskoeffizienten: Pearson vs Spearman

Pearsons Korrelationskoeffizient vs Spearmans Korrelationskoeffizient

- charakterisiert lineare Zusammenhänge
- nutzbar für Merkmale (Features) auf *metrischem Skalenniveau*,
  [d.h. Merkmale unterstützen Operationen wie ≠, =, >, <, +, -, *, /]

  Beispiele: Nettomiete, Alter

- charakterisiert monotone Zusammenhänge
- nutzbar für Merkmale (Features) auf metrischem oder *ordinalem Skalenniveau*,
  [ordinale Merkmale unterstützen Operationen wie ≠, =, >, <,]

  Beispiele: Ränge, Schwierigkeitsgrade von Skipisten (blau, rot, schwarz).

Invarianzen von Korrelationskoeffizienten

Invarianzen

1. Invarianz des Absolutbetrags unter linearen Transformationen:
   (betrifft: Pearsons und Spearmans Korrelationskoeffizienten)

$$\tilde{X} = a_X X + b_X$$
$$\tilde{Y} = a_Y Y + b_Y$$

linear transformierte Features (Merkmale)

Es gilt: $$r_{\tilde{X}\tilde{Y}} = \frac{\sum[a_X x_i + b_X - (a_X \bar{x} + b_X)][a_Y y_i + b_Y - (a_Y \bar{y} + b_Y)]}{\sqrt{\sum[a_X x_i + b_X - (a_X \bar{x} + b_X)]^2 \sum[a_Y y_i + b_Y - (a_Y \bar{y} + b_Y)]^2}}$$

$$= \frac{a_X a_Y \sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[a_X^2 \sum[x_i - \bar{x}]^2][a_Y^2 \sum[y_i - \bar{y}]^2]}} = \frac{a_X a_Y}{|a_X||a_Y|} r_{XY}$$

Es folgt: $$|r_{\tilde{X}\tilde{Y}}| = |r_{XY}|$$

Maßstabsunabhängigkeit des Pearson bzw. Spearman Korrelationskoeffizienten

Invarianzen von Korrelationskoeffizienten

Genauer: $r_{\tilde{X}\tilde{Y}} = \frac{a_X a_Y}{|a_X||a_Y|} r_{XY}$

Der Betrag von *r* ist invariant unter linearer Transformation.
Das Vorzeichen wird bestimmt durch die Vorzeichen der Koeffizienten a_X, a_Y.

2. Invarianz unter streng monotonen Transformationen
(betrifft: Spearmans Korrelationskoeffizient)

$$\tilde{X} = g(X)$$
$$\tilde{Y} = h(Y)$$

mit streng monoton wachsenden oder fallenden Funktionen *g* und *h*

Frage: Wie verhält sich das Vorzeichen von r_{SP} in Abhängigkeit davon, ob *g* bzw *h* streng monoton wachsen oder fallen?

Invarianzen von Korrelationskoeffizienten

3. Invarianz unter Vertauschung der Merkmale *X, Y*
(betrifft: Pearsons und Spearmans Korrelationskoeffizienten)
ergibt sich direkt aus den Definitionen der Korrelationskoeffizienten:

$$r_{XY} = r_{YX} \qquad \text{bzw.} \qquad r_{SP}(X,Y) = r_{SP}(Y,X)$$

→ Merkmale *X* und *Y* sind gleichberechtigt.

→ Korrelationskoeffizienten charakterisieren nur Stärke eines möglichen Zusammenhangs zwischen X und Y und nicht die Richtung!

Beispiel: Daten zu Körpergrößen von Vätern und Söhnen.
Befund: positiver Korrelationskoeffizient.

Korrelation ist ein Maß für die Stärke eines Zusammenhangs zwischen *X* und *Y*. **Die Richtung der Wirkung (sofern existent) wird nicht erfasst**.

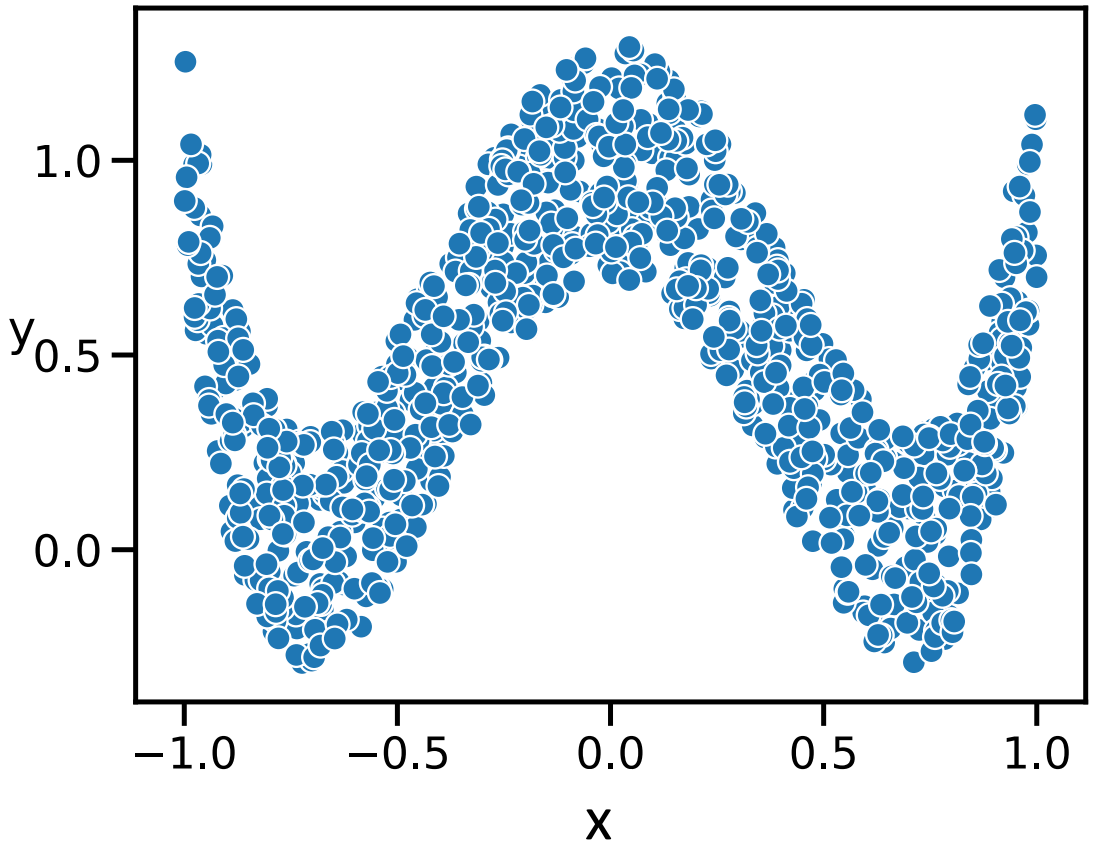
Korrelationskoeffizienten | Pearson- und Spearman

Fragen F

1. Sind x und y in der rechten Abbildung unabhängige Variablen?

2. Welche Werte werden Sie für Pearsons und Spearmans Korrelations-koeffizienten erhalten?

3. Was ist die Ursache für die Diskrepanz zwischen Ihren Beobachtungen in (1) und (2)?



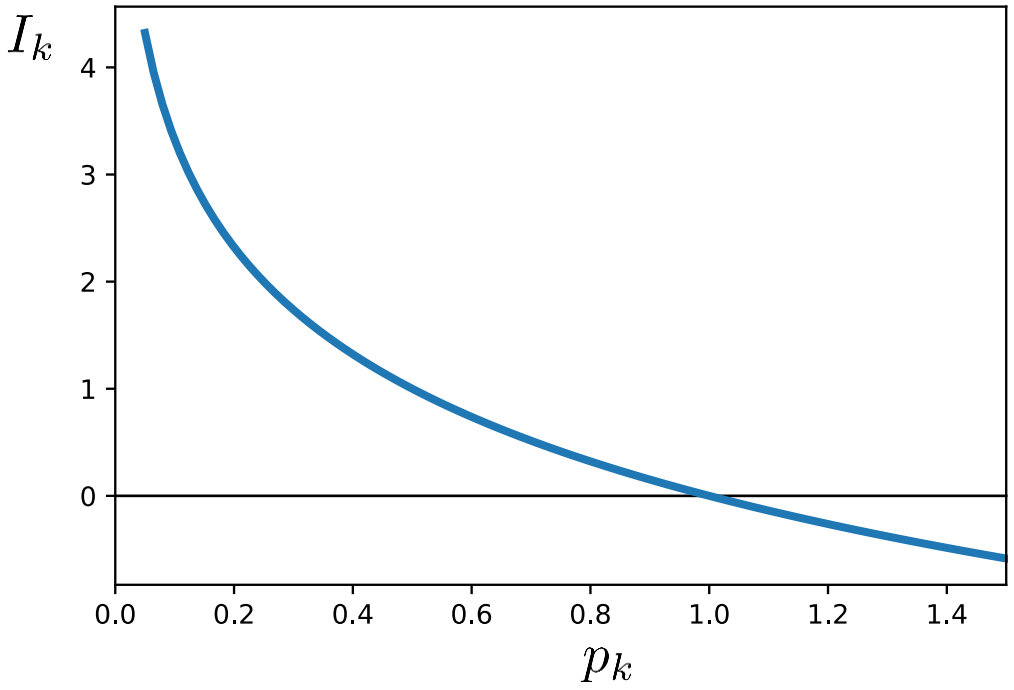
Mutual Information | Informationsgehalt

Mutual Information basiert auf Konzepten aus der Informationstheorie.

Definition (Informationsgehalt)
Informationsgehalt I_k für das Eintreffen eines Ereignisses k mit Wahrscheinlichkeit p_k sei:

$$I_k := -\log_2 p_k$$

Maßeinheit: bit



Bemerkungen

- Je seltener ein Ereignis k, desto größer sein Informationsgehalt.
- Logarithmus erleichtert das Rechnen mit Informationsgehalten.
- $I_k \geq 0$, da $p_k \in [0, 1]$
- Wahl einer anderen Basis des Logarithmus verändert nur die Einheit, in der „Informationsgehalt“ gemessen wird:

$$\log_2 p_k = \frac{\log_{10} p_k}{\log_{10} 2} \approx 3.32 \log_{10} p_k$$

(Maßeinheit in der Basis 10: ban)

Mutual Information

Wir betrachten ein Zufallsexperiment mit einer diskreten Zufallsvariablen *X,* die die möglichen diskreten Werte (Ereignisse, Ausgänge) x_i mit Wahrscheinlichkeiten p_i annimmt.

Definition (Entropie)[1]
Der mittlere Informationsgehalt eines Ereignisses (Ausgangs) eines Zufallsexperiments mit Zufallsvariable X heißt **Entropie H(X).**

$$H(X) := \mathbb{E}[I] = -\sum_{k=1}^{C} p_k \log_2 p_k \qquad \text{mit} \quad 0 \log_2 0 = 0$$

Erwartungswert

Mittelung gewichtet bzgl der Wahrscheinlichkeiten der Ereignisse (bzw. der Klassen)

(entsprechend dem Grenzwert: $\lim_{x \to 0} x \log_2 x = 0$)

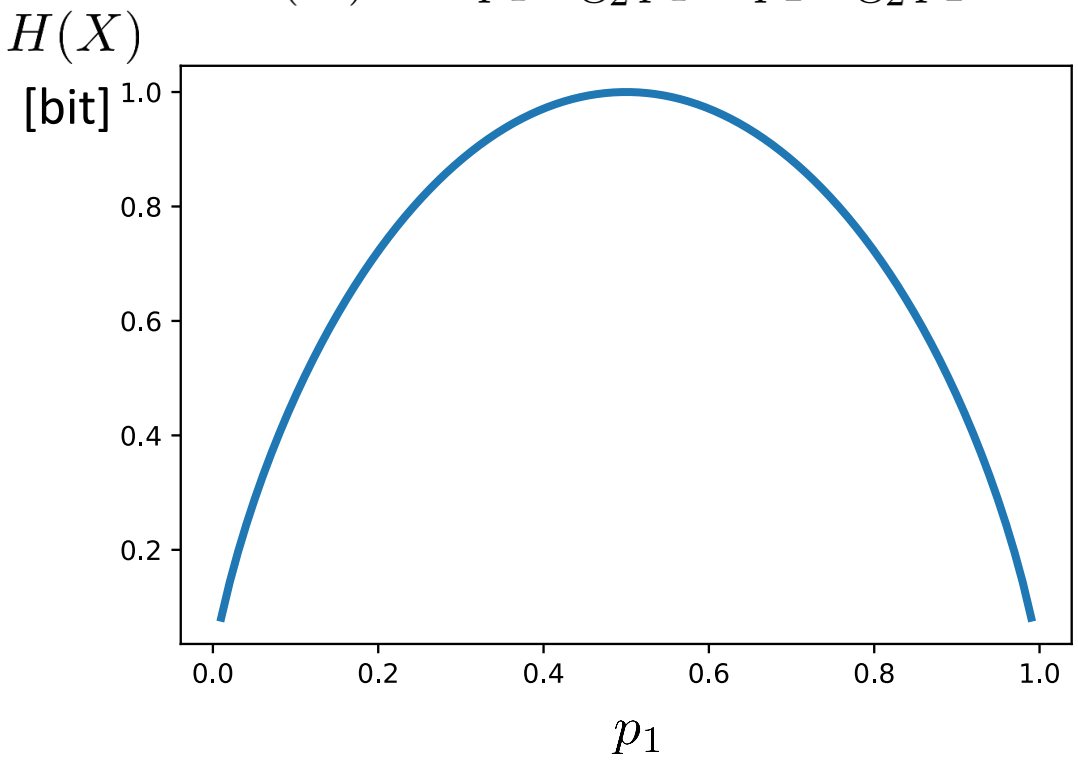
1) bekannt unter den Namen „Shannon-Entropie“, „Gibbs-Boltzmann Entropie“

Mutual Information | Entropie

Beispiel – Entropie für diskrete Variable mit zwei Ausgängen

- 2 Klassen: C=2
- p_1 : Wahrscheinlichkeit für Ereignis (Ausgang) 1
- $p_2 = 1 - p_1$: Wahrscheinlichkeit für Ereignis (Ausgang) 2

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$



Mutual Information | Bedingte Entropie

Wir betrachten zwei Zufallsexperimente mit den diskreten Zufallsvariablen *X* und *Y,* die die möglichen Werte x_i und y_j mit Wahrscheinlichkeiten $p(x_i)$ bzw. $p(y_j)$ annehmen können.

- $p(x_i, y_j)$ bezeichne die gemeinsame Wahrscheinlichkeit (Verbundwahrscheinlichkeit) für das gemeinsame Auftreten von x_i, y_j
- $p(x_i|y_j) = \frac{p(x_i,y_j)}{p(y_j)}$ bezeichne die bedingte Wahrscheinlichkeit für x_i unter der Bedingung, dass y_j vorgegeben ist.

Entropie von X unter der Bedingung des Auftretens eines Wertes y_j:

$$H(X|Y = y_j) = -\sum_i p(x_i|y_j) \log_2 p(x_i|y_j)$$

Mutual Information | Bedingte Entropie

Entropie von X unter der Bedingung des Auftretens eines Wertes y_j:

$$H(X|Y = y_j) = -\sum_i p(x_i|y_j) \log_2 p(x_i|y_j) \qquad (1)$$

Definition (bedingte Entropie)
Der mittlere Informationsgehalt eines Ergebnisses einer Zufallsvariablen *X* unter der Bedingung, dass der Wert einer Zufallsvariablen *Y* bekannt ist, heißt **bedingte Entropie** *H(X|Y)*.

$$H(X|Y) = \sum_j p(y_j) H(X|Y = y_j)$$

Dies ist die gewichtete Summe des Ausdrucks (1) für jeden möglichen Wert von *Y*.

$$= -\sum_{i,j} p(y_j) p(x_i|y_j) \log_2 p(x_i|y_j)$$

$$= -\sum_{i,j} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(y_j)}$$

weil $p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)}$

Mutual Information

Definition (Mutual Information)
Die Abnahme des mittleren Informationsgehalts eines Ergebnisses der Zufallsvariablen *X* durch Kenntnis des Ergebnisses einer Zufallsvariablen *Y* heißt *Mutual Information*.

$$\mathrm{I}(X;Y) = H(X) - H(X|Y)$$

Auf Deutsch
Die Mutual Information misst, wieviel Information Y über X offenbart.

Mutual Information ist auch bekannt unter der Bezeichnung *Transinformation* oder *Information Gain* (im Kontext von Bäumen als Lernmodelle).

Mutual Information

Eigenschaft 1: Symmetrie

$$
\begin{aligned}
\mathrm{I}(X;Y) &= H(X) - H(X|Y) \\
&= -\sum_i p(x_i) \log_2(p(x_i)) + \sum_{i,j} p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(y_j)} \right) \\
&= -\sum_i \left(\sum_j p(x_i, y_j) \right) \log_2(p(x_i)) + \sum_{i,j} p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(y_j)} \right) \\
&= -\sum_{i,j} p(x_i, y_j) \log_2(p(x_i)) + \sum_{i,j} p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(y_j)} \right) \\
&= \sum_{i,j} p(x_i, y_j) \left(\log_2 \left(\frac{p(x_i, y_j)}{p(y_j)} \right) - \log_2(p(x_i)) \right) \\
&= \sum_{i,j} p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(x_i) p(y_j)} \right) \\
&= \mathrm{I}(Y;X)
\end{aligned}
$$

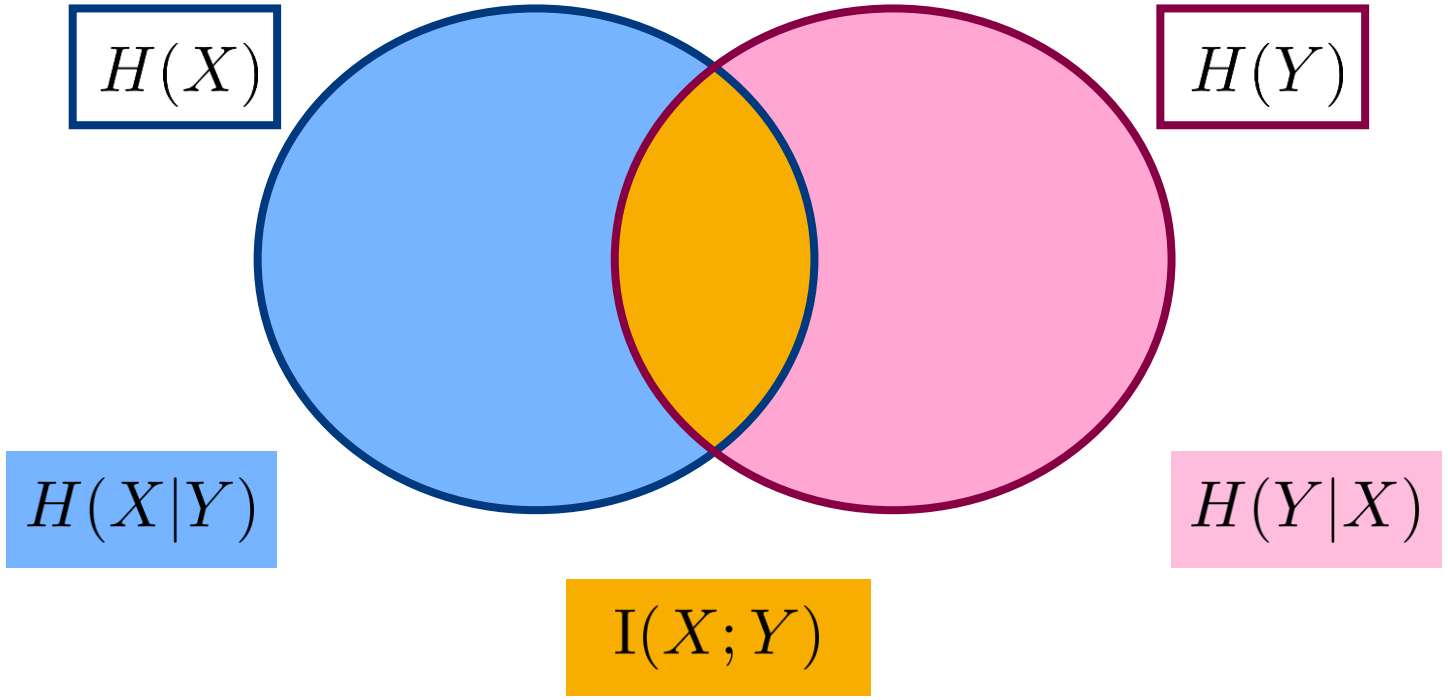
symmetrisch unter Vertauschung von X und Y

Eigenschaft 2: Nicht-Negativität[1]: $\mathrm{I}(X;Y) \geq 0$

1) beweisen wir hier nicht (lässt sich über die Jensensche Ungleichung zeigen)

Mutual Information

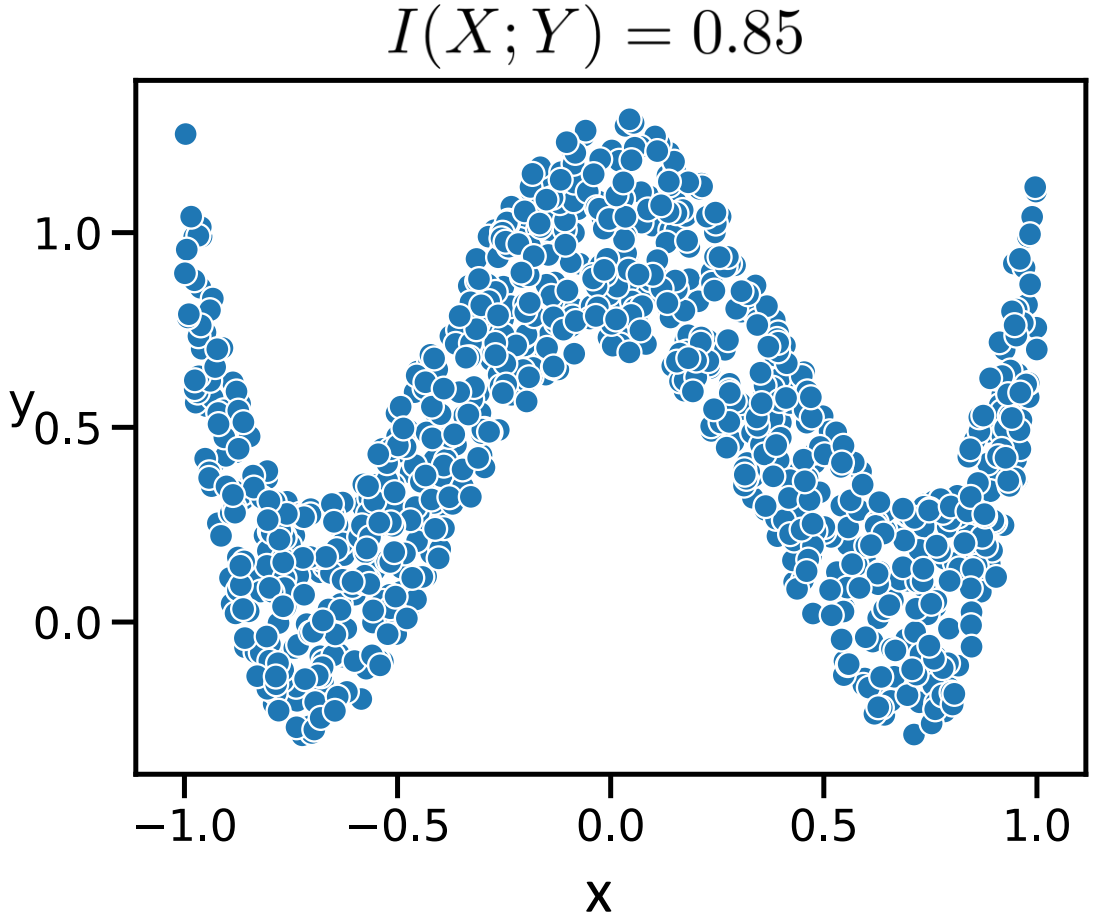
Venn-Diagramm (Mengendiagramm)
zur Veranschaulichung der Zusammenhänge zwischen Entropien H(X) und H(Y), bedingten Entropien H(X|Y), H(Y|X) und Mutual Information I(X;Y)



Mutual Information

Mutual Information …

- kann nichtlineare Zusammenhänge charakterisieren
- ist schwer zu schätzen bei kleinen Stichproben (→ robuste Schätzer wurden in den letzten 20 Jahren entwickelt)



Mutual Information | Nebenbemerkung

Einer der bekanntesten Schätzer für Mutual Information wurde am Forschungszentrum Jülich entwickelt:

PHYSICAL REVIEW E **69**, 066138 (2004)

Estimating mutual information

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger
John-von-Neumann Institute for Computing, Forschungszentrum Jülich, D-52425 Jülich, Germany
(Received 28 May 2003; published 23 June 2004)

We present two classes of improved estimators for mutual information $M(X,Y)$, from samples of random points distributed according to some joint probability density $\mu(x,y)$. In contrast to conventional estimators based on binnings, they are based on entropy estimates from k-nearest neighbor distances. This means that they are data efficient (with $k=1$ we resolve structures down to the smallest possible scales), adaptive (the resolution is higher where data are more numerous), and have minimal bias. Indeed, the bias of the underlying entropy

M‘variate Explorative Analyse

Korrelation und Kausalität

- hohe Werte von Zusammenhangsmaßen *können hindeuten auf* kausale Zusammenhänge, diese aber nicht begründen
- *Kontrollierte Experimente* können zur Entdeckung von kausalen Zusammenhängen genutzt werden.
  Typisch: Merkmal *X* wird im Experiment verändert, und man beobachtet die daraus resultierenden oder nicht resultierenden Änderungen von Merkmal *Y*.
- Experimente lassen sich aber in vielen Fällen nicht realisieren.
  → Entwicklung von Maßen zur Charakterisierung der Richtung eines Zusammenhangs (z.B. *Granger Causality*)

Zusammenhangsmaße | Interpretationsfehler

Fehlerquellen bei der Beurteilung von Zusammenhangsmaßen

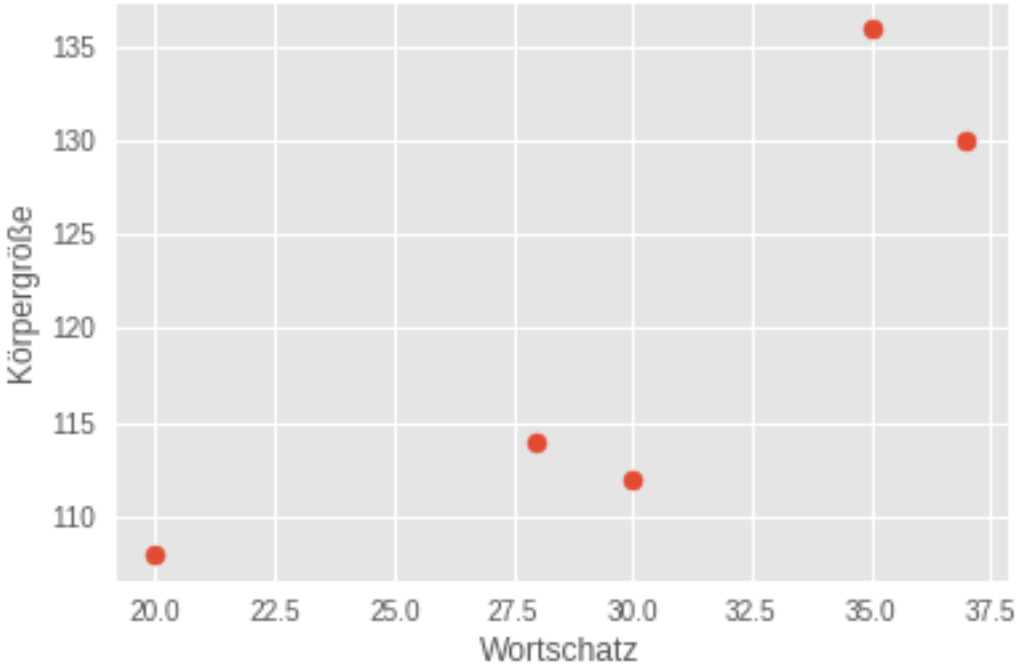
- wesentlicher Merkmale (Features) nicht berücksichtigen

F

1. **Scheinkausalität** (*spurious correlation*)
   (meist unscharf als *Scheinkorrelation* bezeichnet)

Beispiel (fiktiv):

- Körpergrößen (*Y*) von Kindern und ihr Wortschatz (*X*, in einem Aufsatz) wurden erfasst
- Korrelation (Pearson):
  $r_{XY} = 0.86$



Frage: Welchen Kausalzusammenhang vermuten Sie zwischen *X* und *Y*?

Zusammenhangsmaße | Interpretationsfehler

Zusammenhangsmaße | Interpretationsfehler

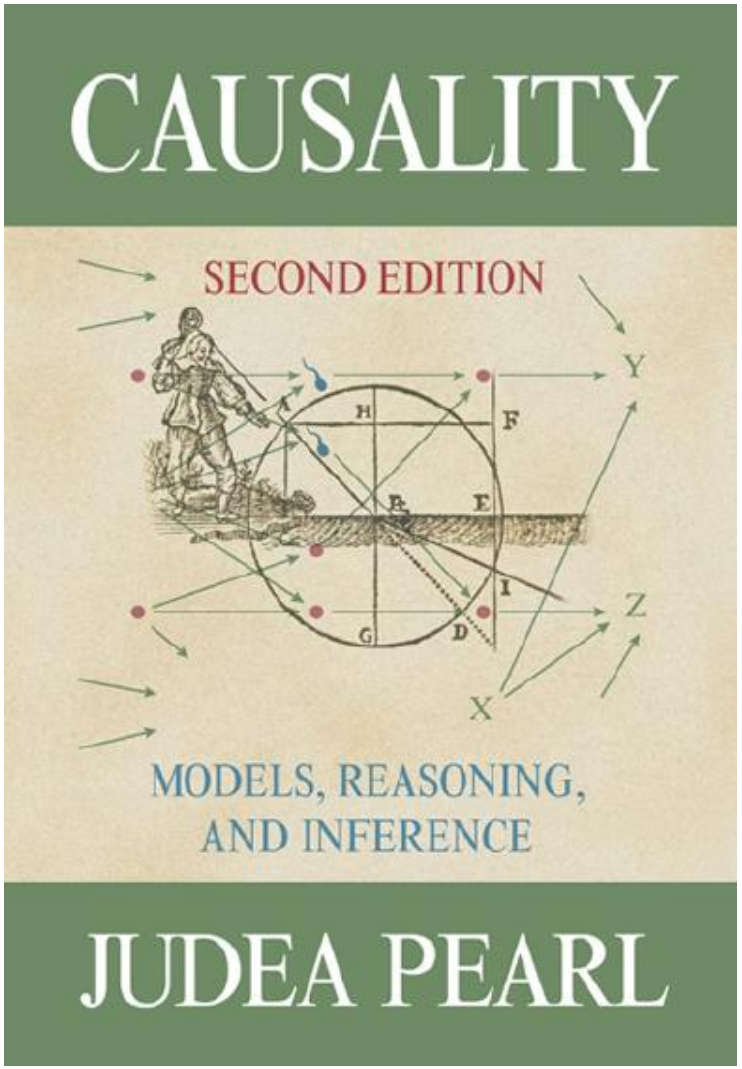
Beispiel: Jäger-Beute Beziehungen in komplexen Ökosystemen

Zusammenhangsmaße | Exkurs

Identifikation von Kausalitätsnetzen (*graphical models*)

- Netzwerke statistischer Unabhängigkeit (d.h. relevant sind *fehlende* Verbindungen)
- Entwicklung vieler Methoden zur Identifikation von direkter/indirekter bzw. Schein- vs echter Korrelation
- aktueller Forschungsgegenstand (siehe Buch von J. Pearl, 2009)

Unterhaltsames Buch (ab Doktorandenniveau)
J. Pearl, Prof. an der UCLA, Informatiker, KI-Forscher

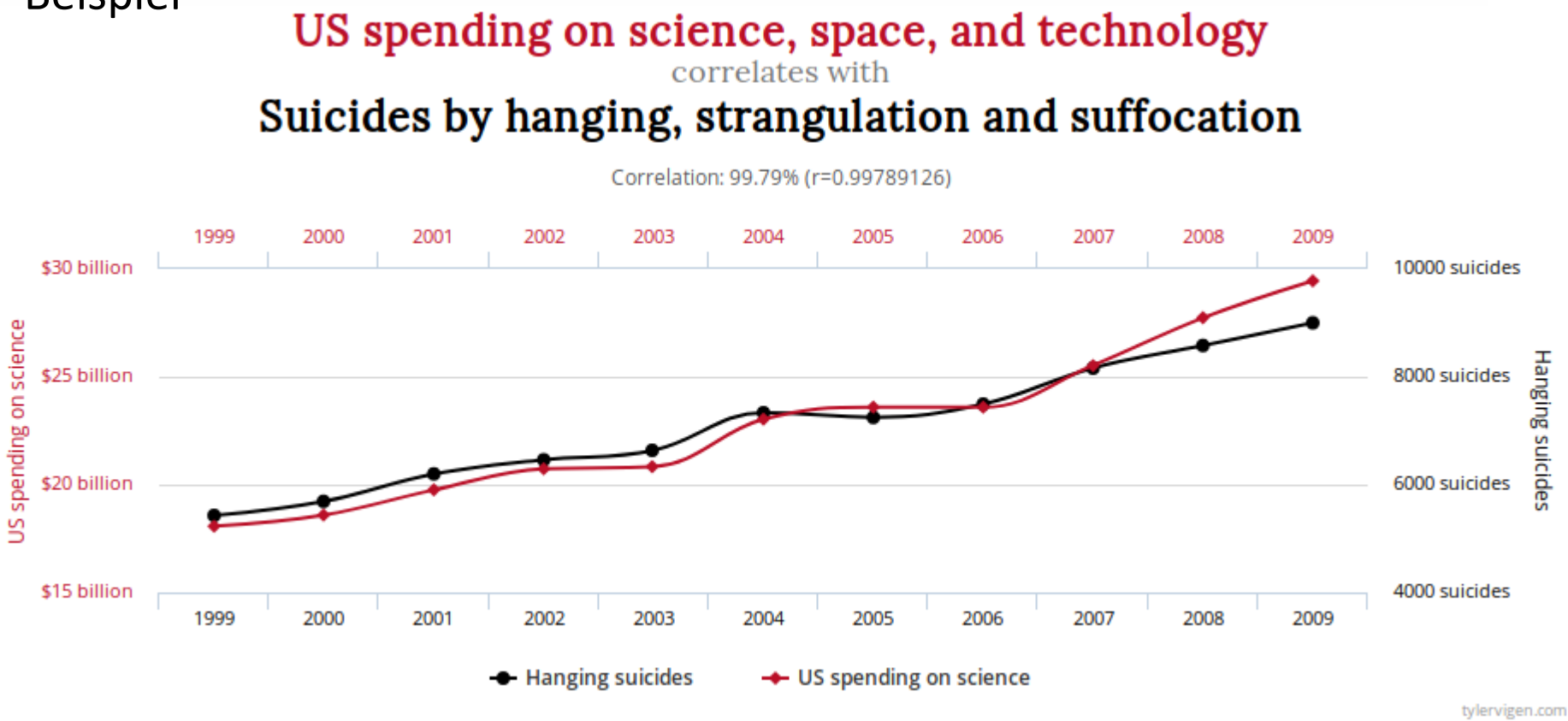


Pearl, J. Springer 2009

Zusammenhangsmaße | Interpretationsfehler

c) Zufällige Korrelation

Beispiel

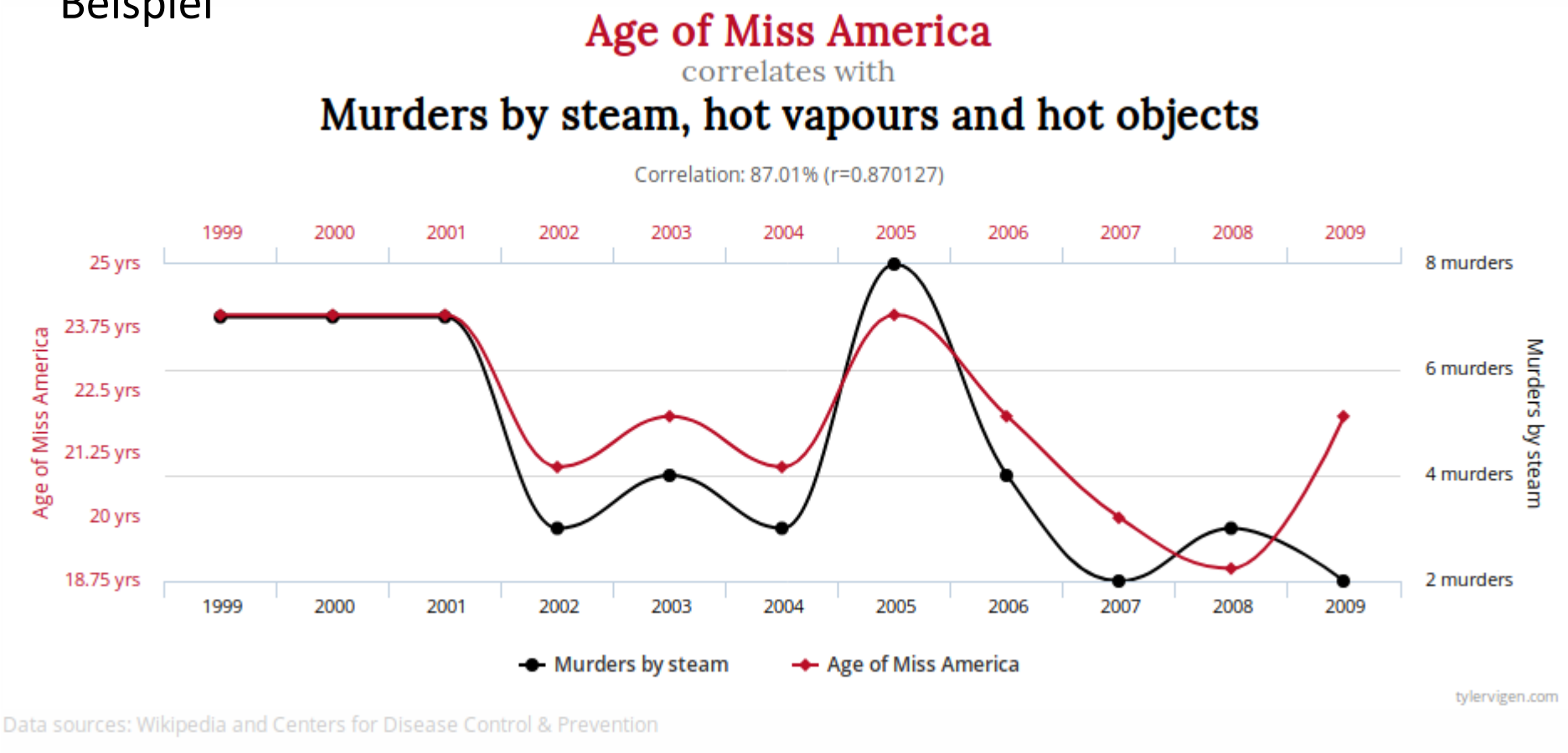


Quelle: https://www.tylervigen.com/spurious-correlations

Zusammenhangsmaße | Interpretationsfehler

c) Zufällige Korrelation

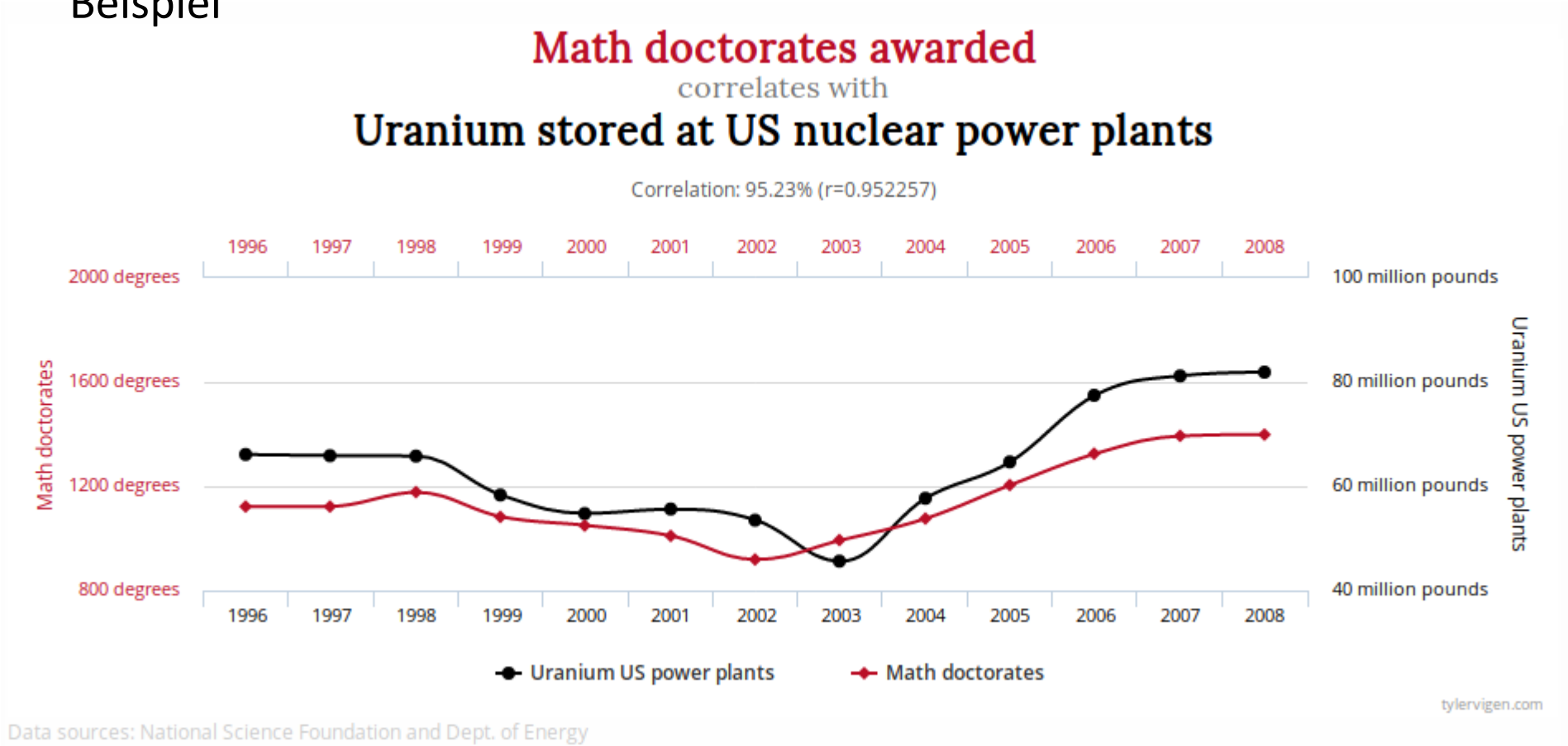
Beispiel



Zusammenhangsmaße | Interpretationsfehler

c) Zufällige Korrelation

Beispiel



Zusammenhangsmaße | Interpretationsfehler

c) Zufällige Korrelation

Tyler Vigen (ehemaliger Doktorand der Harvard Law School) generierte die vorherigen gezeigten Abbildungen zum Spaß auf folgende Weise[1]:

1. Anlegen eines Datensatzes mit vielen Merkmalen (darunter z.B. „Age of Miss America“)
2. Berechnung von Korrelationskoeffizienten für alle Merkmalspaare.

„data dredging“

> “The correlations are also strong because very few points are being compared. Instead of comparing just ten years, we should ideally be looking at hundreds of points of comparison. Correlations are an important part of scientific analysis, but they can be misleading if used incorrectly. Even the charts are designed to be subtly deceptive. The data on the Y-axes doesn’t always start at zero, which makes the graphs appear to line up much better than they otherwise would. The data points are real and mathematically placed, but they are displayed in a very specific way.”[1]

1) Quelle: Vigen T, *Spurious Correlations*, Hachette 2015

Zusammenhangsmaße | Interpretationsfehler

2. Verdeckte Korrelation
(kann auftreten, wenn eine Population in Teilpopulationen zerfällt)

Beispiel (fiktiv):

- Dosierung eines Medikaments vs Therapieerfolg
- Patientenpopulation zerfällt in zwei Teilpopulationen (leicht und schwer erkrankte)
- In jeder Teilpopulation steigt Therapieerfolg mit Dosierung (→ positive Korrelation)
- Über beide Teilpopulationen ist der Korrelationskoeffizient aber negativ (Verdeckung der positiven Korrelation der Teilpopulationen)