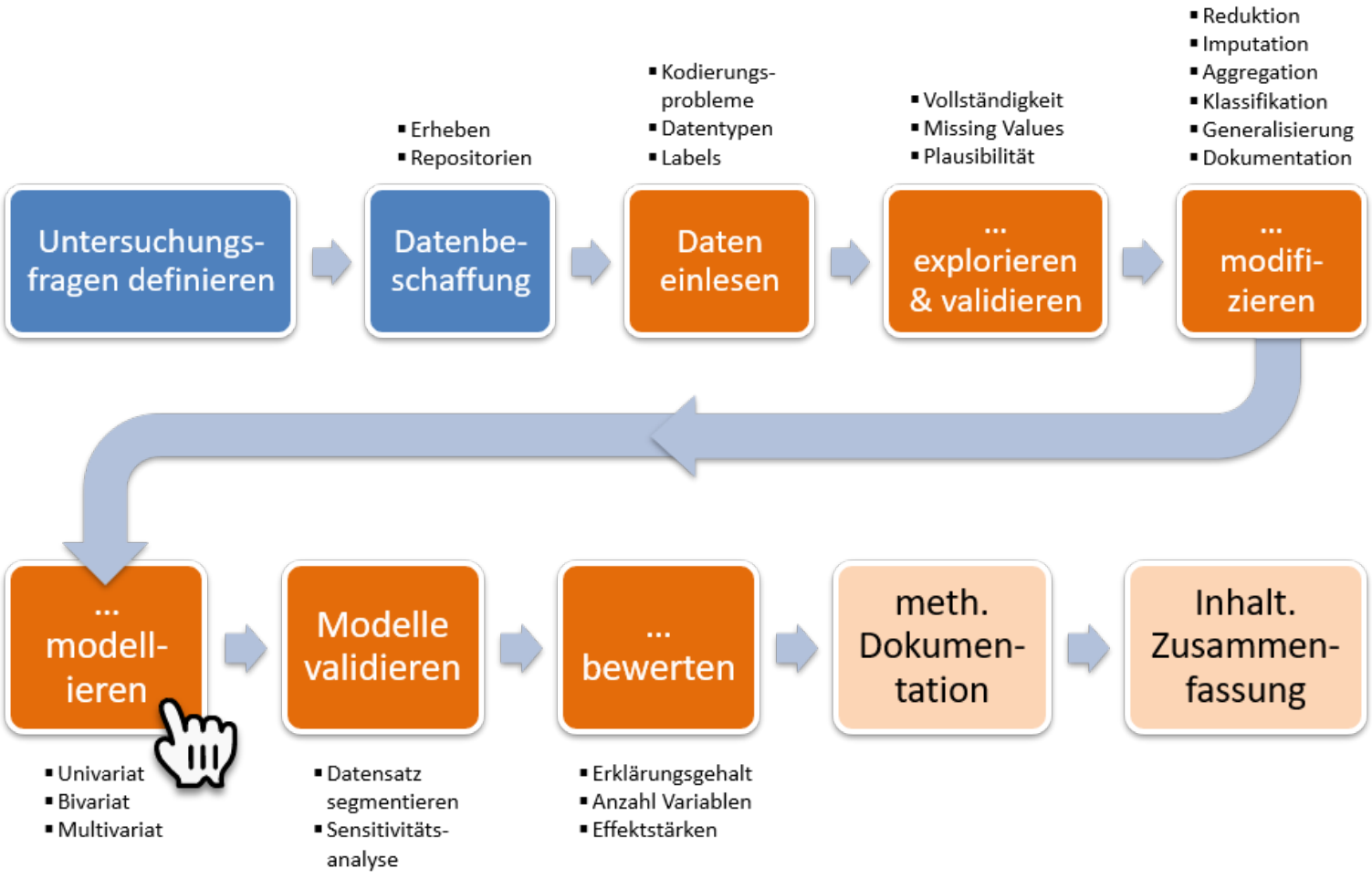
universität innsbruck
Institut für Geographie

Clusteranalyse - Grundlegendes

716408 | How 2 do Things with even more Numbers

KMH
WS 21-22 (updated: 2021-12-16)

Wo wir gerade stehen



Clusteranalyse?

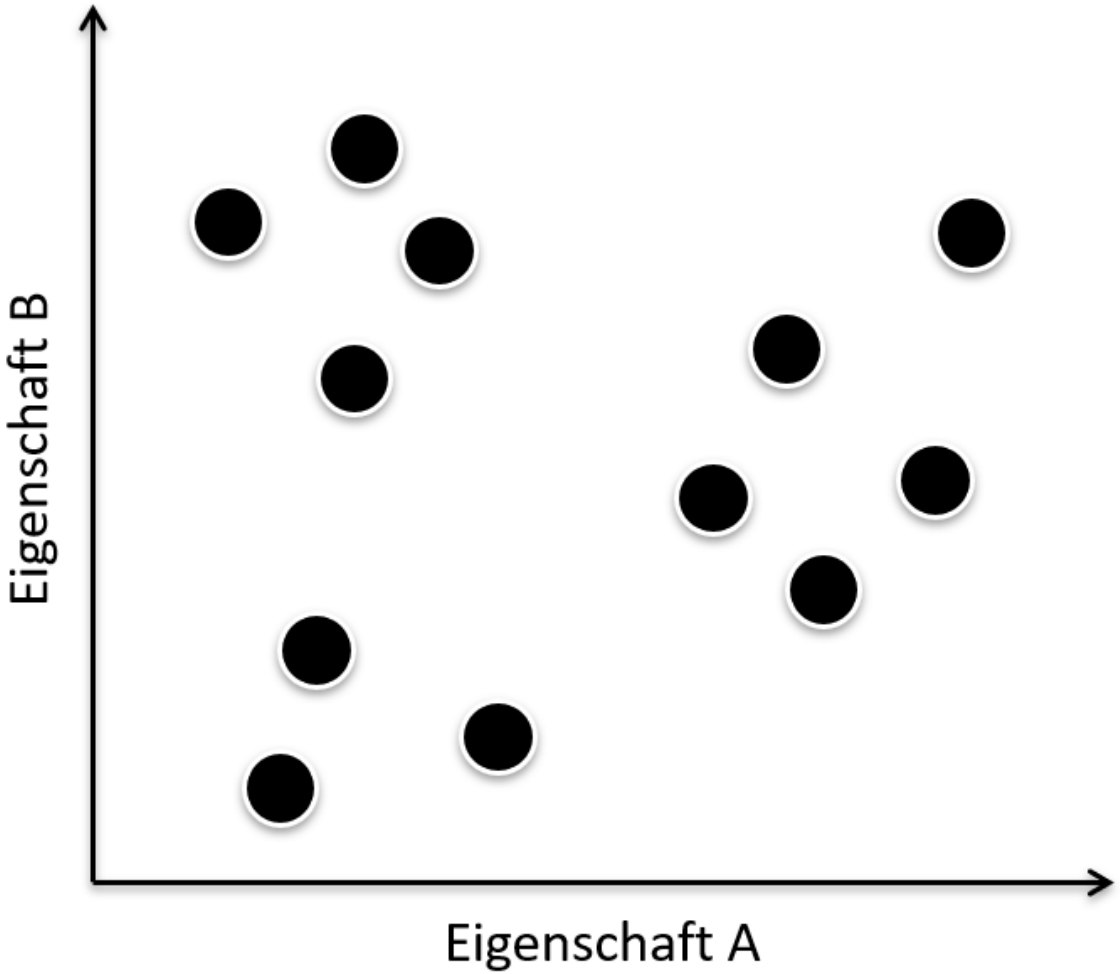
Warum und wozu Clusteranalyse?

strukturentdeckende Verfahren:
↪ "ähnliche*" Merkmalsträger in Gruppen zusammenfassen

Einige **geographische Beispiele**:

- um Aktivitäten zur Anpassung an den Klimawandel inhaltlich zu sortieren (z.B. Grüneis et al. 2018);
- um Verhaltenstypen zur Vorsorge gegenüber Naturgefahren zu unterscheiden (z.B. Posch et al. 2019);
- oder um Typen ländlicher Räume zu unterscheiden (z.B. Höferl et al. 2007).

@ Ähnlichkeit



Zielsetzung beim Bilden der Gruppen

A. **hohe Intracluster-Homogenität**: Ähnliches in einen Cluster

B. **hohen Intercluster-Heterogenität klar**:
Cluster klar unterschiedlich

Wie geht das?

Ablauf einer Clusterung

Zwei zentrale Arbeitsschritte:

1. Zuerst: Ähnlichkeit bzw. Distanz zwischen den zu gruppierenden Merkmalsträgern ermitteln
    - ↪ **Proximitätsmaße**.
2. Danach: Anhand eines **Gruppierungsverfahrens** gruppierbare (= die "ähnlichsten") Merkmalsträger zusammenfassen
    - ↪ **Fusionsalgorithmen**

📚 Mittlerweile gibt es Vielzahl clusteranalytischer Verfahren, die sich hinsichtlich dieser beiden Arbeitsschritte unterscheiden (vgl. Backhaus et al. 2017:438ff.).

Welche Proximitätsmaße gibt es?

Ein kurzer Überblick auf Proximitätsmaße

- **Ziel:** Bestimmung der Ähnlichkeit von Merkmalsträgern
  - Ähnlichkeit = Indikator der je Merkmalsträger über all seine Merkmale hinweg ermittelt und aggregiert wird.
- **Operationalisierung** von "Ähnlichkeit":
  - Ähnlichkeitsmaße wie zB Korrelationen
  - Distanzmaße wie zB euklidische Distanz

Ausgewählte Ähnlichkeitsmaße

abhänging vom Skalenniveau:

metrische Merkmale	nominale Merkmale	binäre (0/1) Merkmale
Kosinus	Transformation in binäre Variable	Würfelmaß (Dice- oder Czekanowski-Koeffizient)
Pearson-Korrelation		Jaccard-Koeffizient
		M-Koeffizient (einfache Übereinstimmung)
		Kulczynski-Koeffizient
		Rogers und Tanimoto
		Russel & Rao (RR) Koeffizient

Ausgewählte Distanzmaße

abhänging vom Skalenniveau:

metrische Merkmale	nominale Merkmale	binäre (0/1) Merkmale
(Quad.) Euklidische Distanz	Chi-Quadrat-Maß	Binäre Euklidische Distanz
Minkowski Metrik	Phi-Quadrat-Maß	Lance-Williams-Maß
Block Metrik		Binäre Form-Differenz
Tschebyscheff Metrik		Größendifferenz
		Varianz
		Rogers und Tanimoto
		Russel & Rao (RR) Koeffizient

Exkurs: ... and there is more

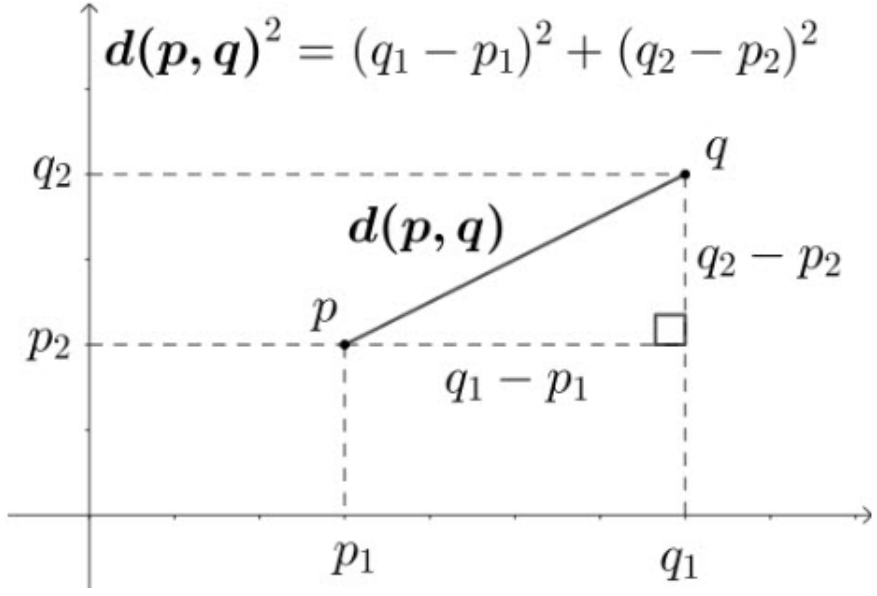
Wie immer: Packages bieten natürlich noch mehr Ähnlichkeits- und Distanzmaße

```
library(philentropy)
getDistMethods()
```

```
##  [1] "euclidean"         "manhattan"         "minkowski"
##  [4] "chebyshev"         "sorensen"          "gower"
##  [7] "soergel"           "kulczynski_d"      "canberra"
## [10] "lorentzian"        "intersection"      "non-intersection"
## [13] "wavehedges"        "czekanowski"       "motyka"
## [16] "kulczynski_s"      "tanimoto"          "ruzicka"
## [19] "inner_product"     "harmonic_mean"     "cosine"
## [22] "hassebrook"        "jaccard"           "dice"
## [25] "fidelity"          "bhattacharyya"     "hellinger"
## [28] "matusita"          "squared_chord"     "squared_euclidean"
## [31] "pearson"           "neyman"            "squared_chi"
## [34] "prob_symm"         "divergence"        "clark"
## [37] "additive_symm"     "kullback-leibler"  "jeffreys"
## [40] "k_divergence"      "topsoe"            "jensen-shannon"
## [43] "jensen_difference" "taneja"            "kumar-johnson"
## [46] "avg"
```

Wie entscheidet man sich für ein Ähnlichkeitsmaß?

- 👉 **Skalennievau** entscheidet
- danach: situationsabhängig
- Ein (klassisches) Beispiel: **Euklidische Distanz** bei metrischen Variablen



(Kmhkmh, Wikikmedia, CC BY)

Exkurs: Euklidische Distanz

Für einen n-dimensionalen (n = Anzahl der Variablen) Fall kann die euklidische Distanz wie folgt ermittelt werden:

$$d_{(p,q)} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Zur Betonung von Unterschieden in den so ermittelten Distanzen:

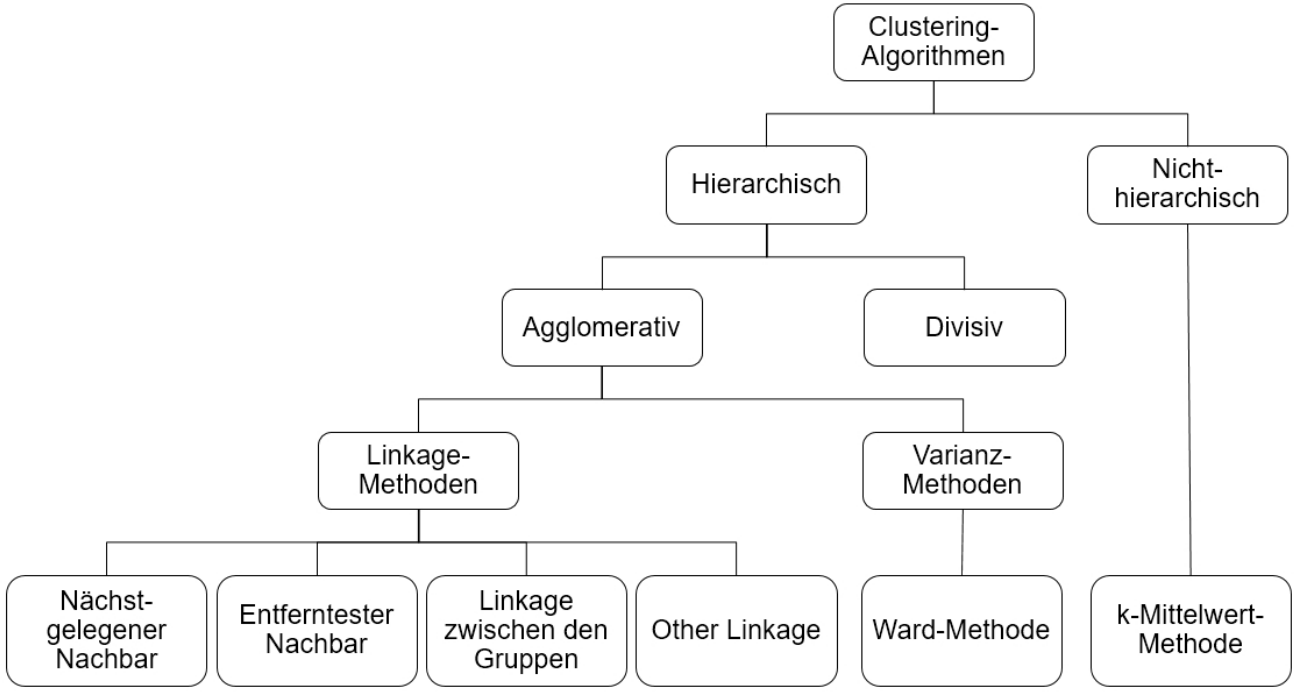
quadrierte euklidische Distanz $d^2_{(p,q)}$

↪ Betonung der Unterschiedlichkeiten von Merkmalsträgern

Welche Fusionsalgorithmen gibt es?

Ein kurzer Überblick auf Gruppierungsverfahren

Ziel:
Merkmalsträger mit geringer Proximität (hohe Ähnlichkeit bzw. geringe Distanz) in Gruppen zusammenfassen bzw. Grundgesamtheit in solche Gruppen zerteilen



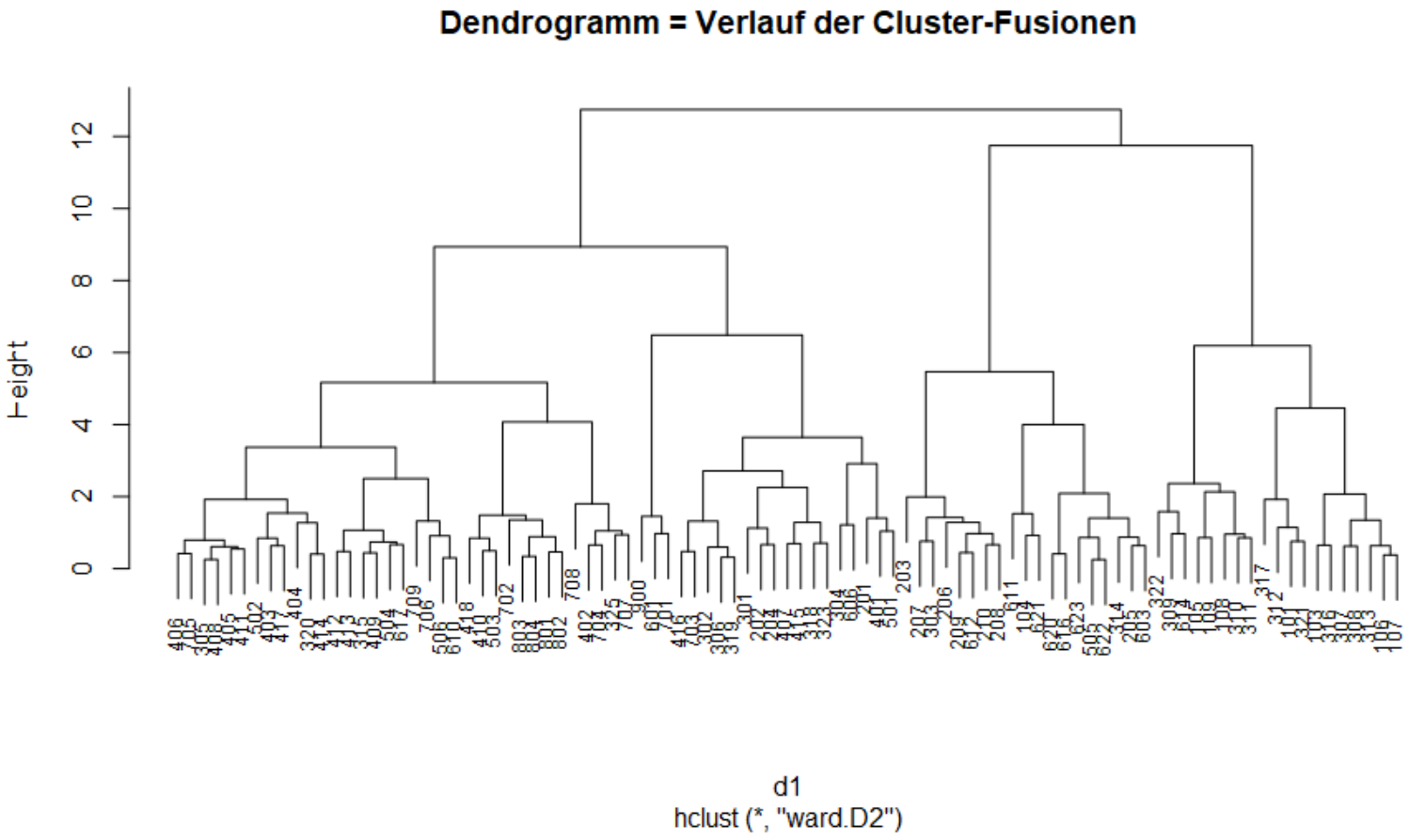
(Universität Zürich)

Der Klassiker:

Hierarchisch agglomerative Verfahren

- Beginn: jeder Merkmalsträger = eigener Cluster
- Ermittlung der Proximität
- Fusion der zwei "ähnlichsten" Cluster
- Ermittlung der Proximität
- Fusion ...
- ...
- Endergebnis: 1 Supercluster (= umfasst alle Merkmalsträger)

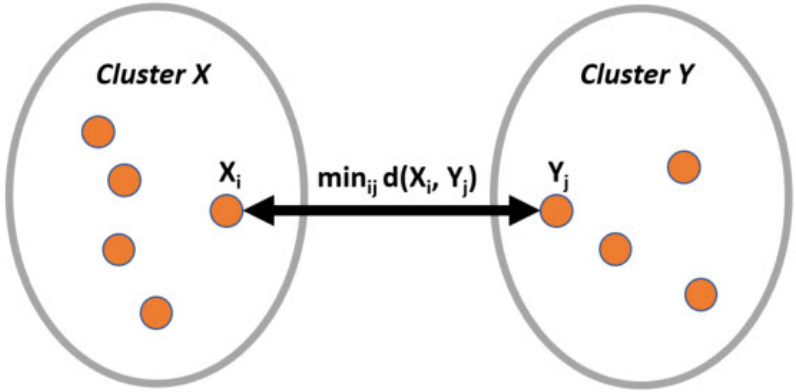
Der Klassiker graphisch gedacht:



Fusionsalgorithmus ≠ Fusionsalgorithmus

Das **Single Linkage (aka. "Nächstgelegener Nachbar") Verfahren**:

Bei diesem Verfahren werden die beiden Cluster am ähnlichsten bzw. nähesten zueinander eingestuft, deren Merkmalsträger die geringste Distanz zueinander aufweisen:



- Effekt 1: **Ausreißer** werden sichtbar
- Effekt 2: "Kettenbildung" bedingt wenige, dafür große Cluster

Fusionsalgorithmus ≠ Fusionsalgorithmus

Die **Ward Methode**:

Dieses Verfahren fokussiert nicht auf die Distanz von Clusterelementen zueinander, sondern auf die Varianz der Cluster. Ward definiert Varianz dabei als die Summe der quadrierten Abweichungen ("ESS - Error Sum of Squares") der Merkmalsträger in einem Cluster zum Cluster-Mittelwert.

- ↪ jene zwei Cluster fusionieren, deren Fusion die **Varianz über alle Cluster** am wenigsten erhöht
- **Effekt:** gleich große Cluster