M. Thrun, AG Datenbionik bei Prof. Ultsch

Quality Measures of Projections

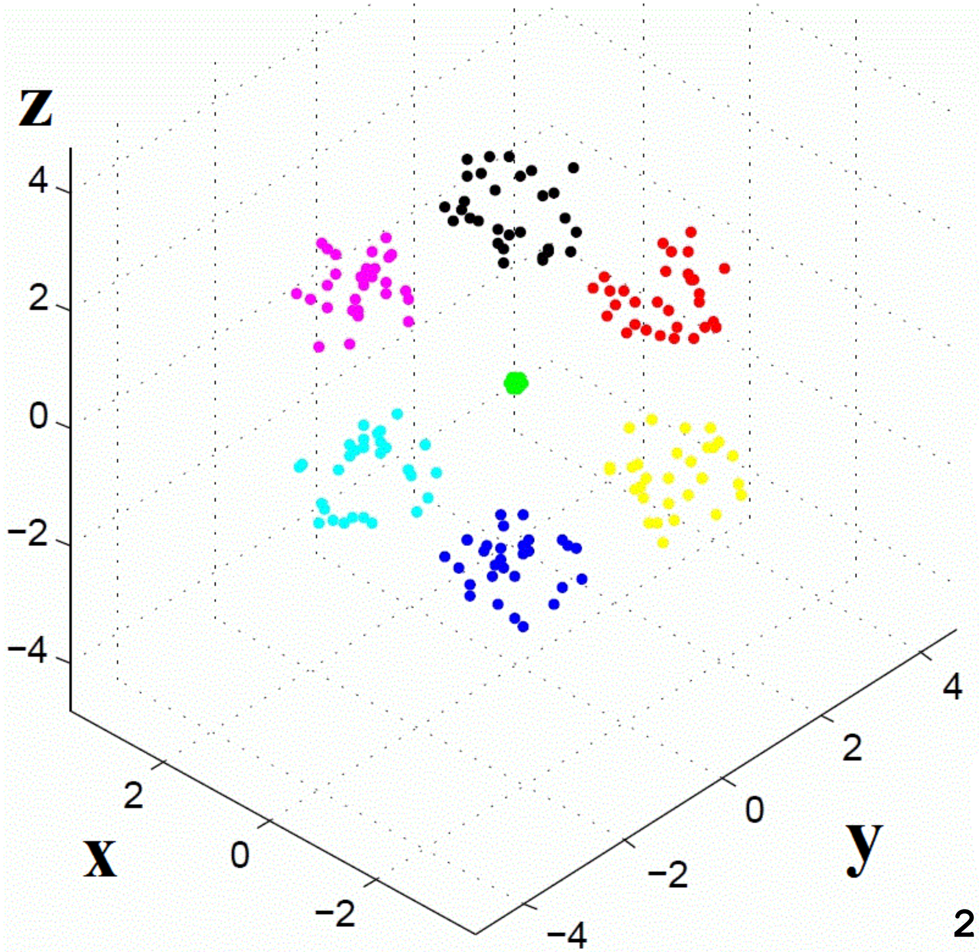
Chapter 6 der Dissertation

Philipps Universität Marburg

Discovering patterns in data

- Searching for similarities
- Clustering: process of finding groups of similar objects in high-dimensional data
  - e.g. labelling data points with colors
- Example: N=3 dimensions in Hepta
  - Equidistant clusters
  - One Cluster has higher density
  - Small distances in each cluster

Hepta data set

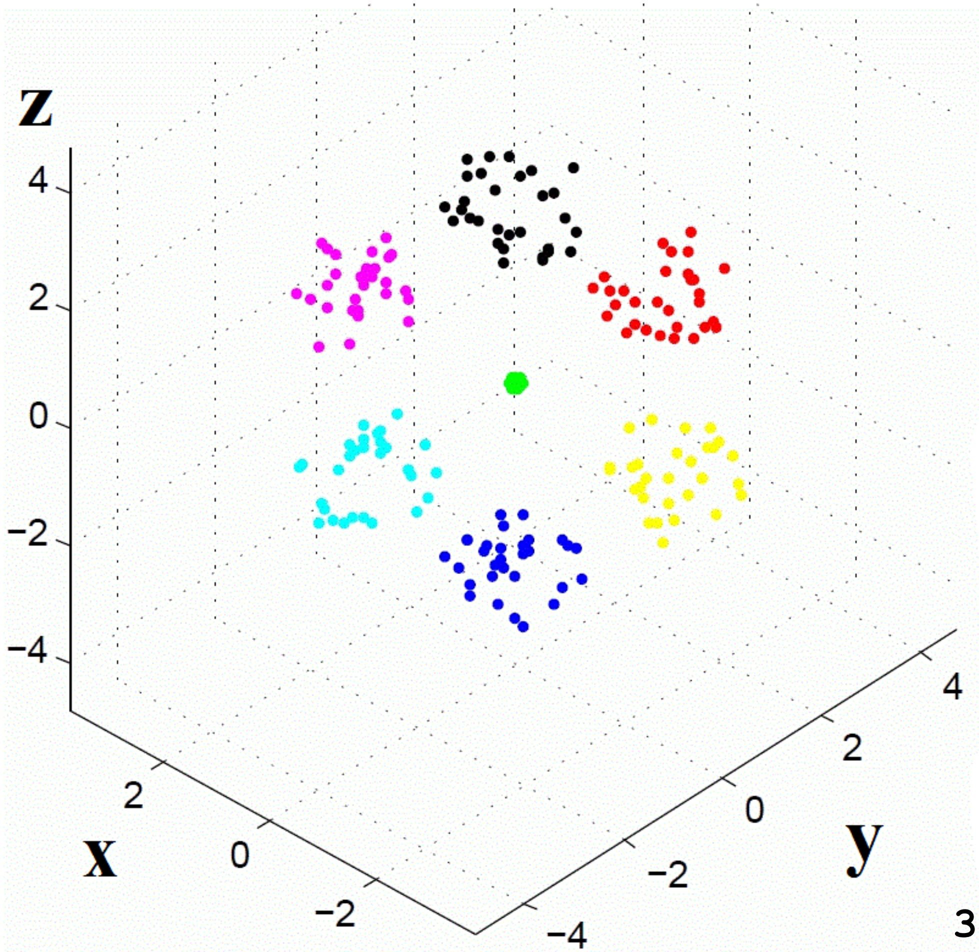


Discovering patterns in data

- Searching for similarities
- Clustering: process of finding groups of similar objects in high-dimensional data
- Problem: high-dimensional: N>>3

How are we able to find similarities?

Hepta data set



Finding similarities

1. Use a clustering algorithm
   - Every algorithm has a geometric model for a cluster
   - How to define a cluster? -> application-specific

2. Dimensionality Reduction (DR)
   - Type I: *manifold learning,*
   - Type II: ***projections*** into two dimensions

Dimensionality Reduction (DR)

- Type I: *manifold learning,* see [Lee, Verleyson 2007]
  - "manifold learning methods are not necessarily good for[…] visualization […] since they have been designed to find a manifold, not compress it into a lower dimensionality"[Venna et al., 2010, p. 452]
  - they do not outperform the classical principal component analysis (PCA) in real world tasks [L. J. van der Maaten et al., 2009],

- Type II: *projections* into two dimensions
  - a scatter plot of a projection method (mostly PCA) still remains state-of-the-art for cluster analysis (e.g. [Everitt et al., 2001, pp. 31-32; Hennig et al., 2015, pp. 119-120, 683-684; Mirkin, 2005, p. 25; Ritter, 2014, p. 223]

DR mit Projektionen

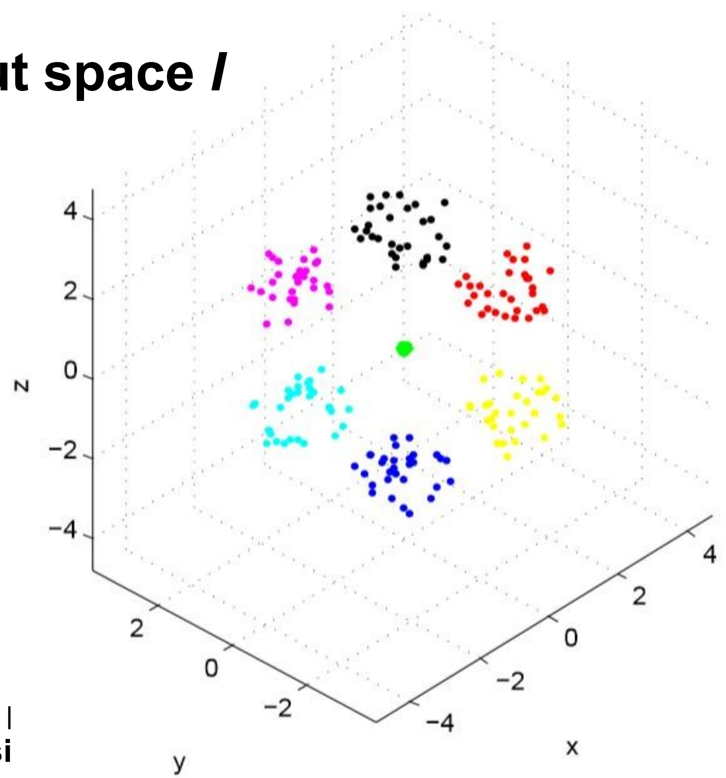
- Hochdimensionale Daten $\mathbb{R}^d, d > 4$, nicht vollständig darstellbar

=> Projektion von hochdimensionalen Daten in 2 Dimensionen

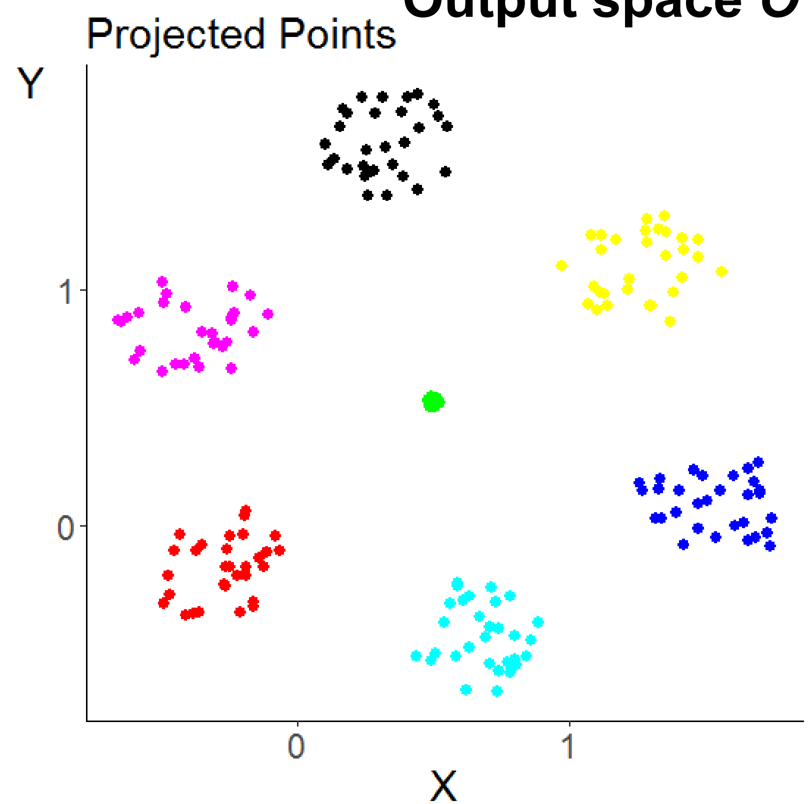
- Projektion soll der Erkennung von „Ähnlichkeit“ dienen

  B/ Aus Projektion Anzahl an Clustern schätzen

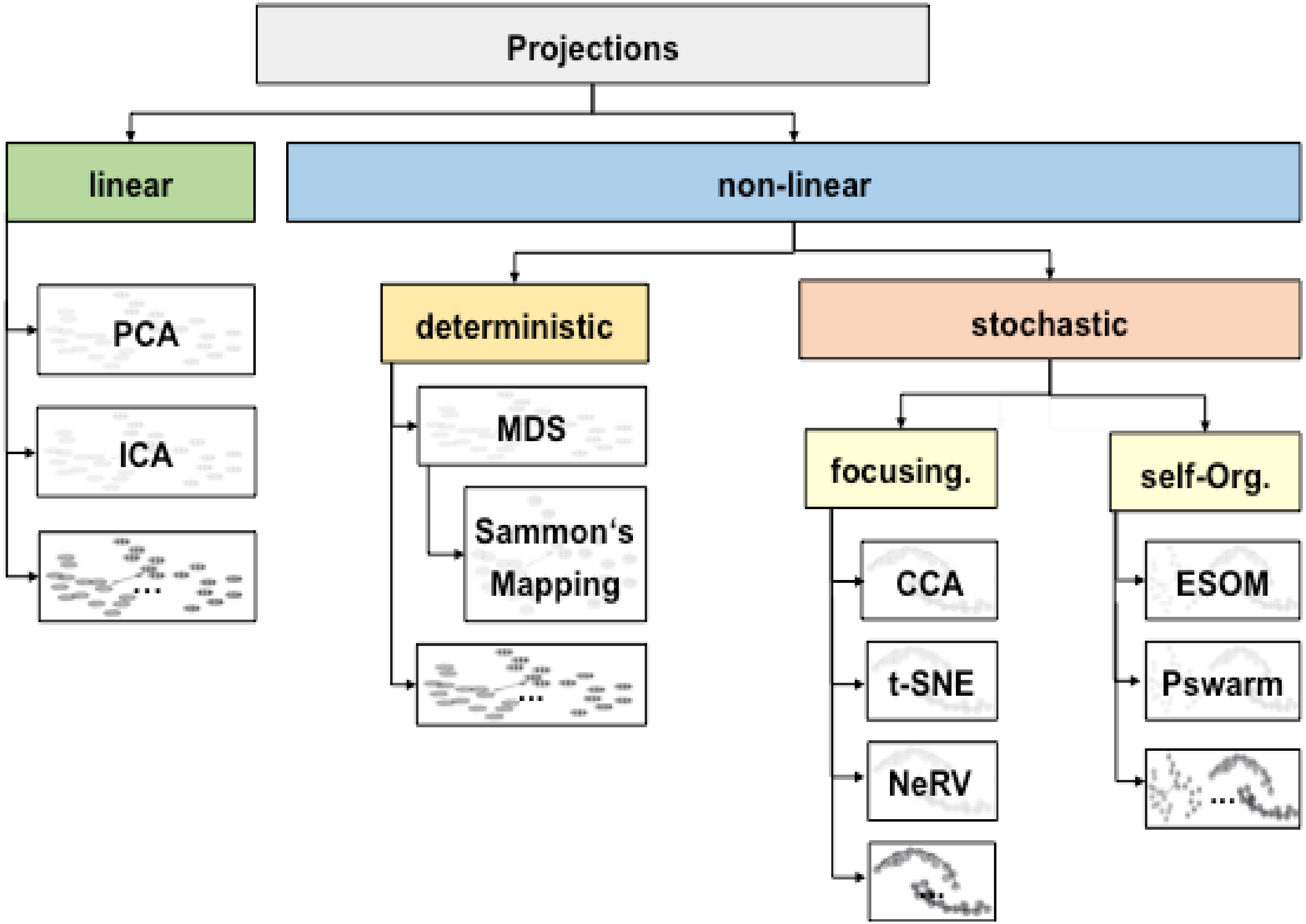
Input space *I*



Output space *O*



Typische Projektionsverfahren



Praktische Probleme

Data -> Projection -> Cluster Analysis ->Visual Verification

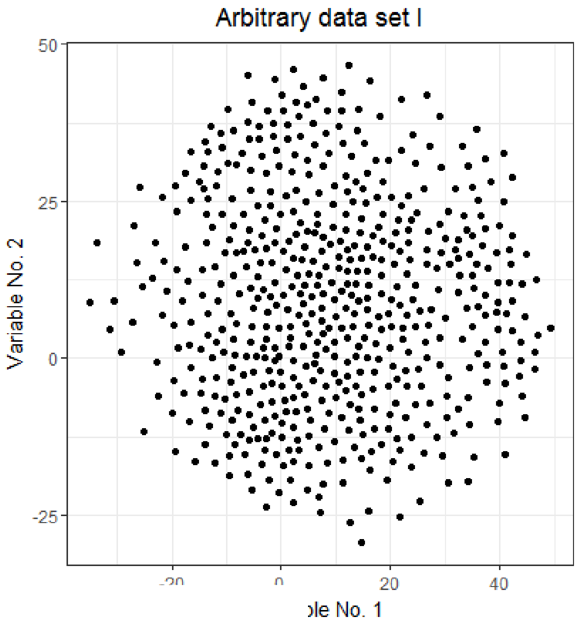
1. **Grundsätzliche Fehler durch DR**
2. Stochastische Projektionsverfahren haben zufällige „Fehler“ abhängig vom Durchlauf (Versuch)
3. Falsches Projektionsverfahren ausgewählt

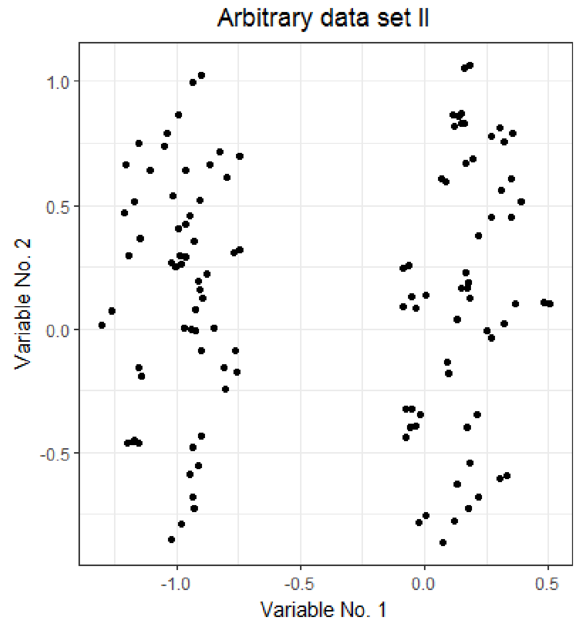
Wie misst man die Qualität eines Projektionsverfahrens?

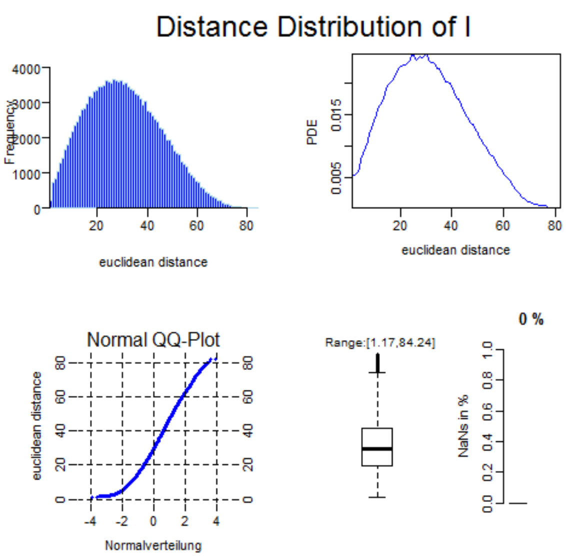
Datenverteilung im Hochdimensionaler Raum

Fall 1: Kontinuierlich

Fall 2: Diskontinuierlich

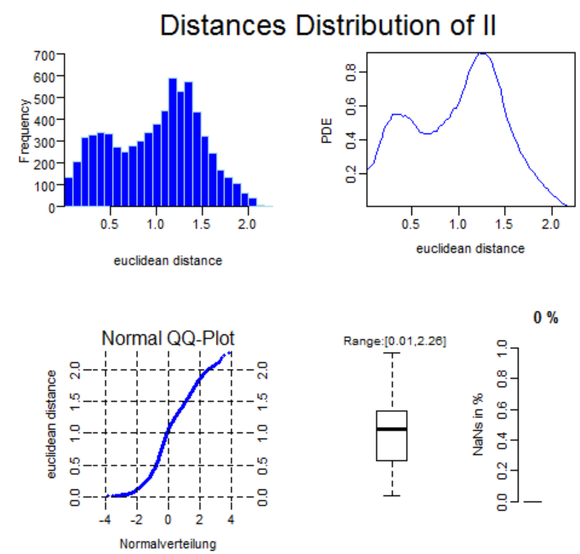






Input space

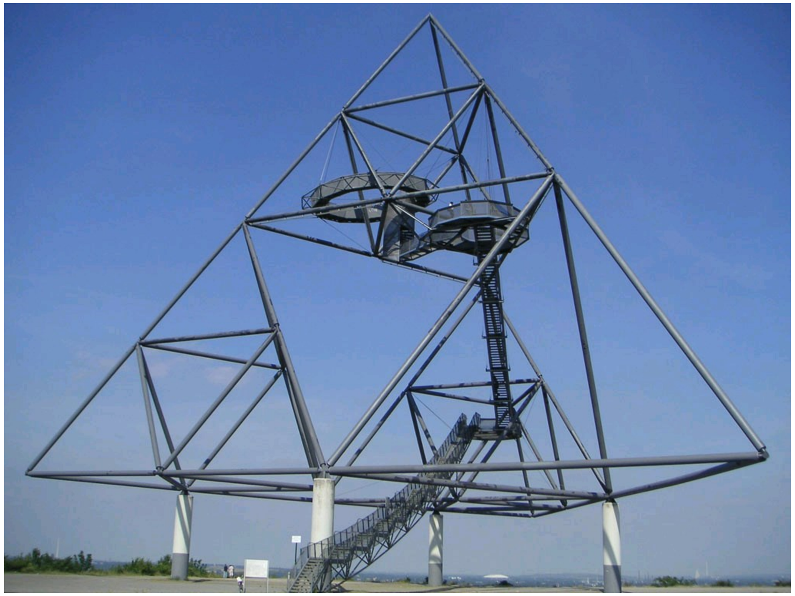
Distances: $\boldsymbol{D(l,j)}$

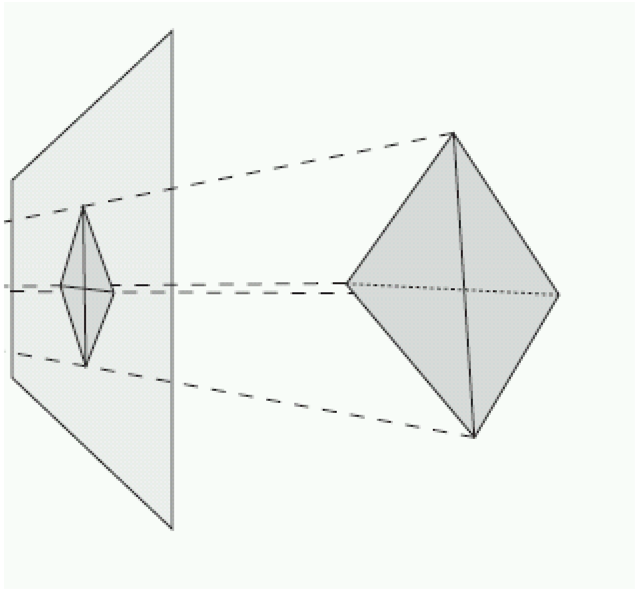


Fall 1: Kontinuierlich

- By limiting the Output space to two dimensions, low dimensional similarities $\boldsymbol{d}(\boldsymbol{l},\boldsymbol{j})$ do not represent high-dimensional distances $\boldsymbol{D}(\boldsymbol{l},\boldsymbol{j})$ coercively

  => Two kind of errors: **BPE** and **FPE** [Ultsch, Herrmann 2005]

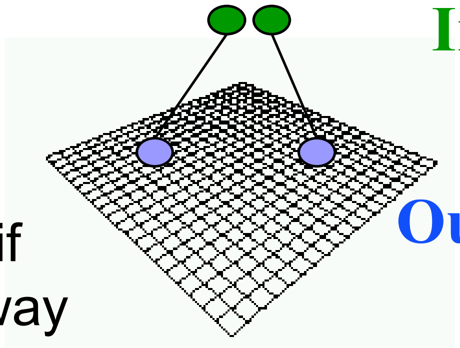




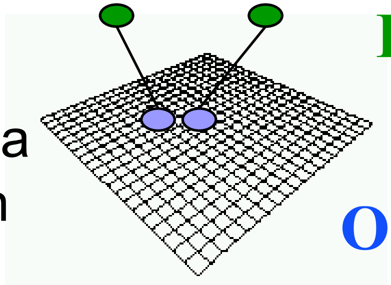
BPE vs FPE

- Best Case: close proximity data points stay in a close proximity and remote data points stay in remote positions

- Let's assume a pair of similar high dimensional data points $(l_I, j_I) \in I$:



  - Forward projection errors (**FPE**) which occur if similar data points in *I* are mapped onto faraway points $(l, j\ ) \in O$

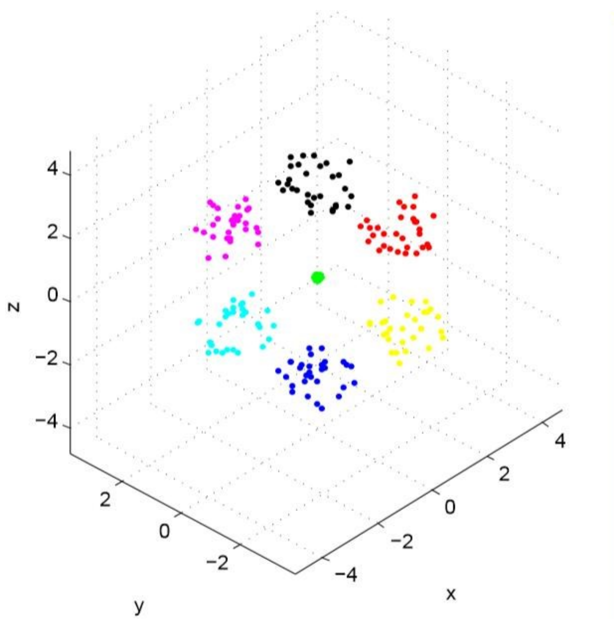


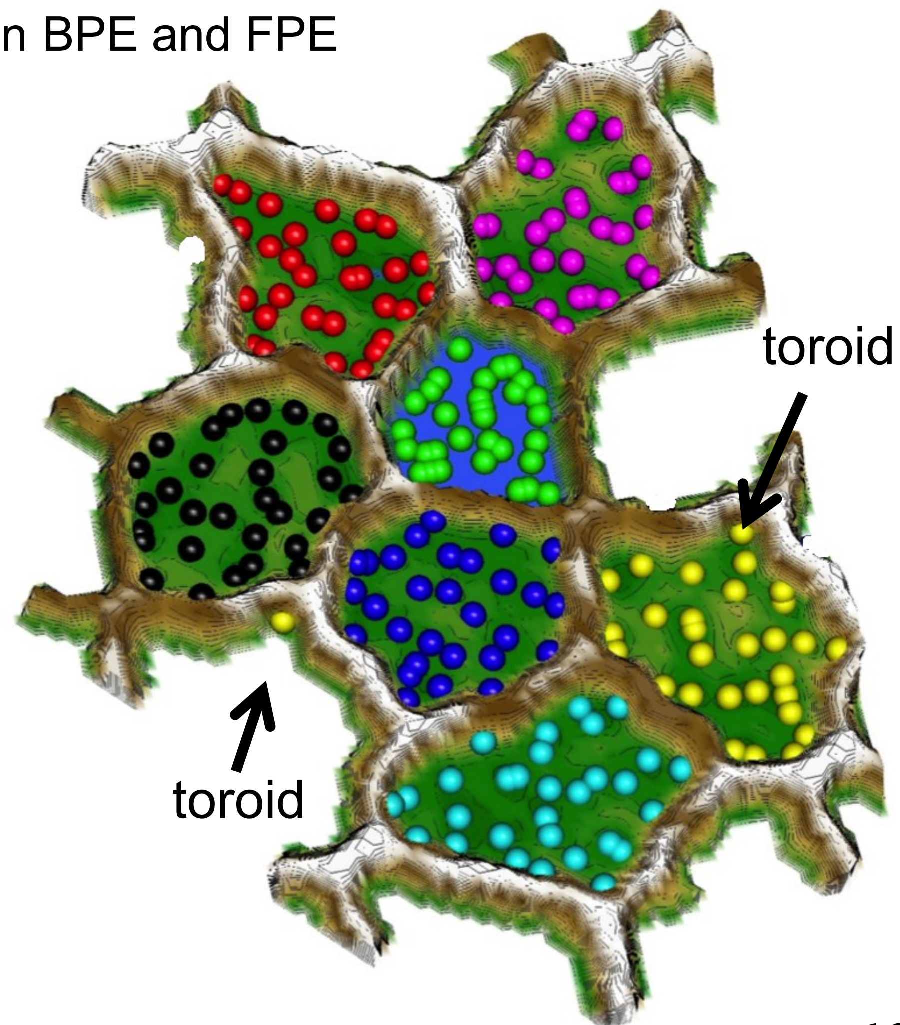
  - Backward projection errors (**BPE**) occurs if a pair of close neighboring positions $(l, j\ ) \in O$ is a representation of a pair of distant data points in *I*

Fehlerbehandlung, Fall 1

- Information Retrieval with Precision and Recall
  (1-BPE and 1-FPE) -> Simulated annealing -> NeRV
  - Problematik: Gewichtung zwischen BPE and FPE

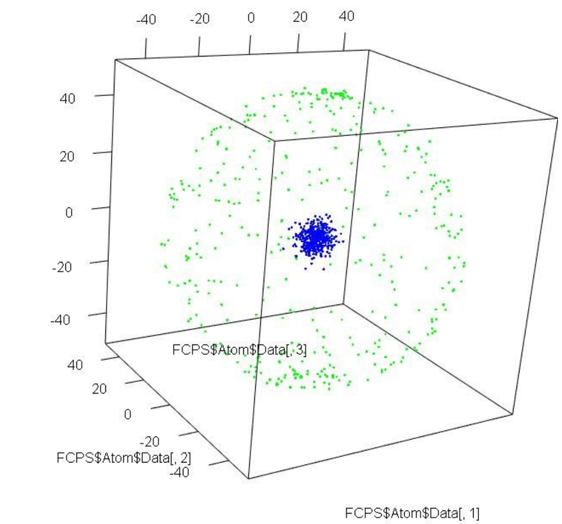
- Wir: Generalisierte Umatrix





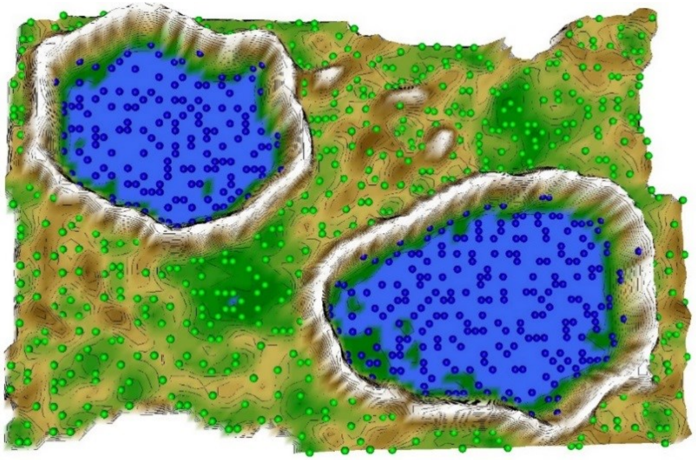
Fall 2: Diskontinuierlich:

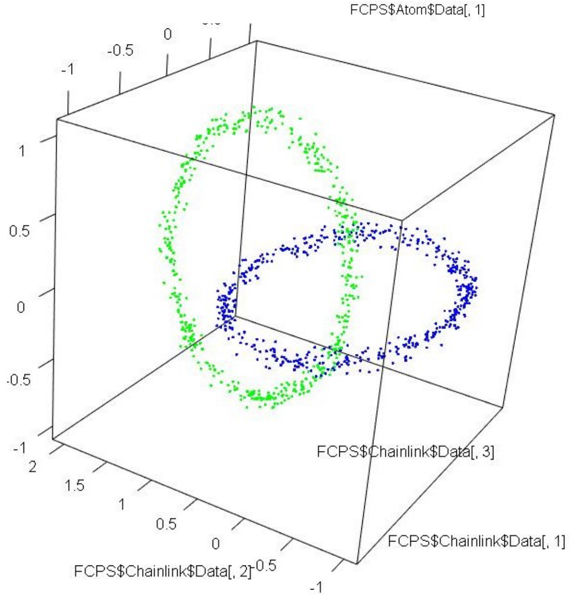
- dritte Fehlerart: **„low structure preservation“**



ESOM

Umatrix

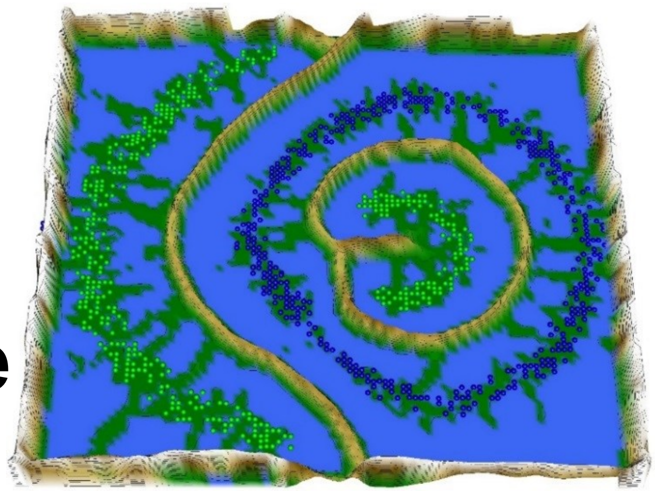




CCA

+generalisierte

Umatrix



Assesement of quality measures

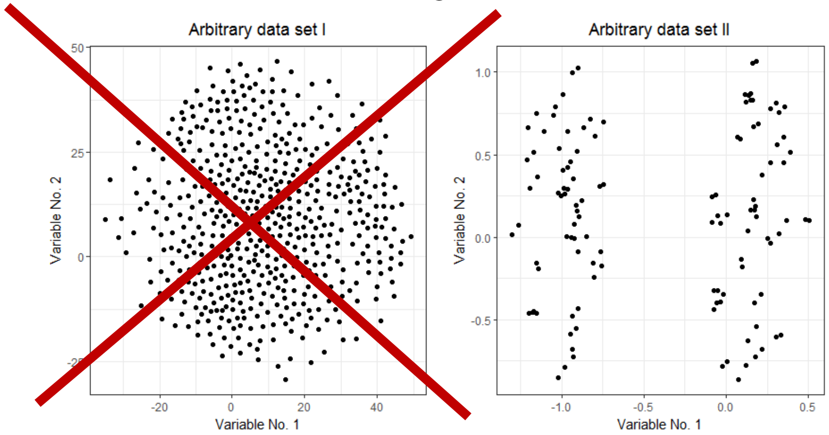
- Goal: Easily understandable quality measures

Is the visualisation by a projection method a good representation of the submanifold?

- good representation = structure preserving projection method

Structure Preservation: Describe the quality

- Structure: Pattern characterized by discontinuity



1. **Compact Structures**
   - the arrangement of **all** given points in space specified by a distance is compared

-> Distance Based

2. **Connected Structures**
   - Local neighborhoods are compared

-> Quality of local proximity

How to describe the quality?

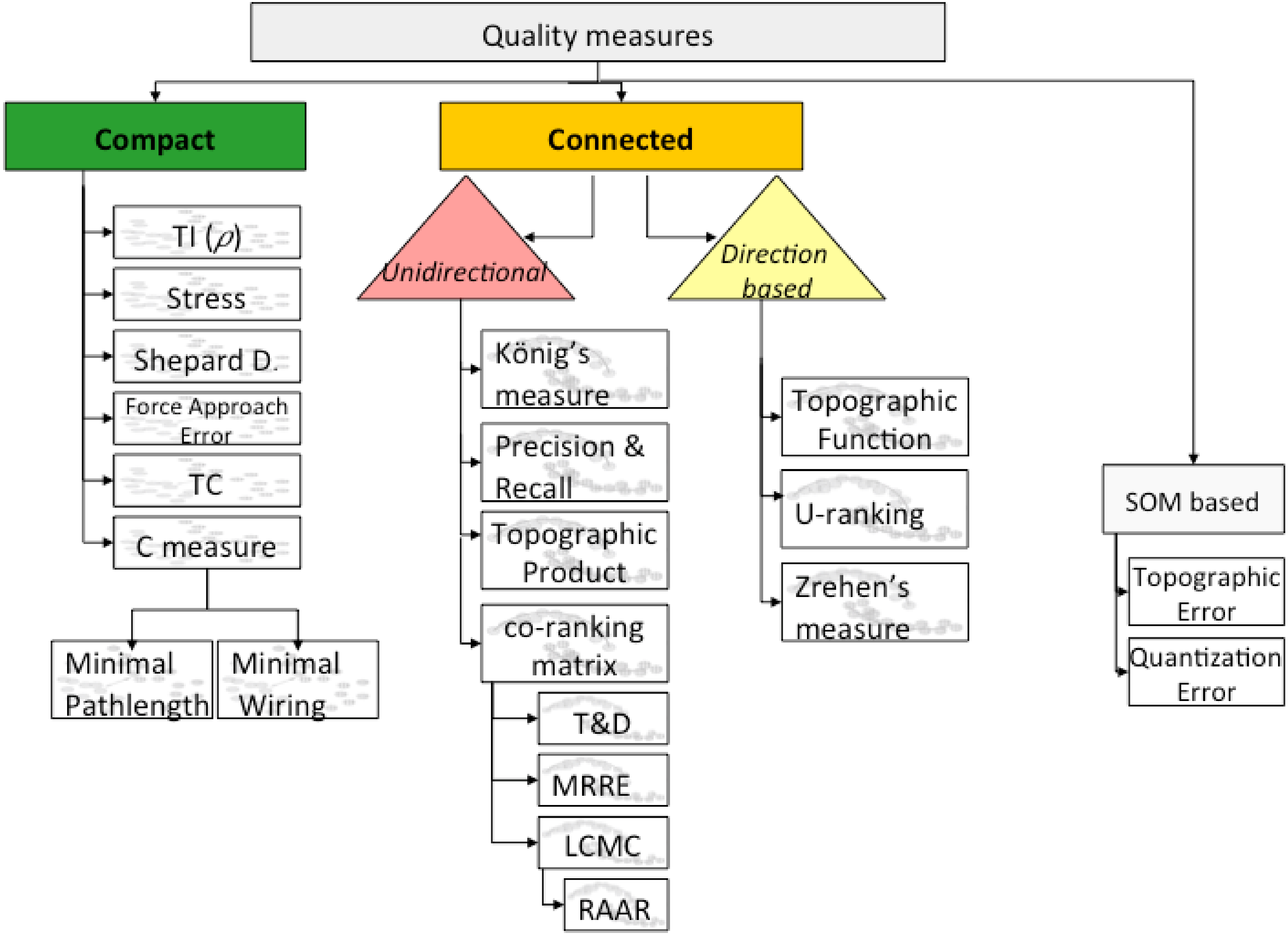
3. Classification Requiring (supervised)

I. Condition: Classification of data known (Input I)

II. Condition: Classification through the projection computable (Output II)

=>Measure quality by comparing both classifications

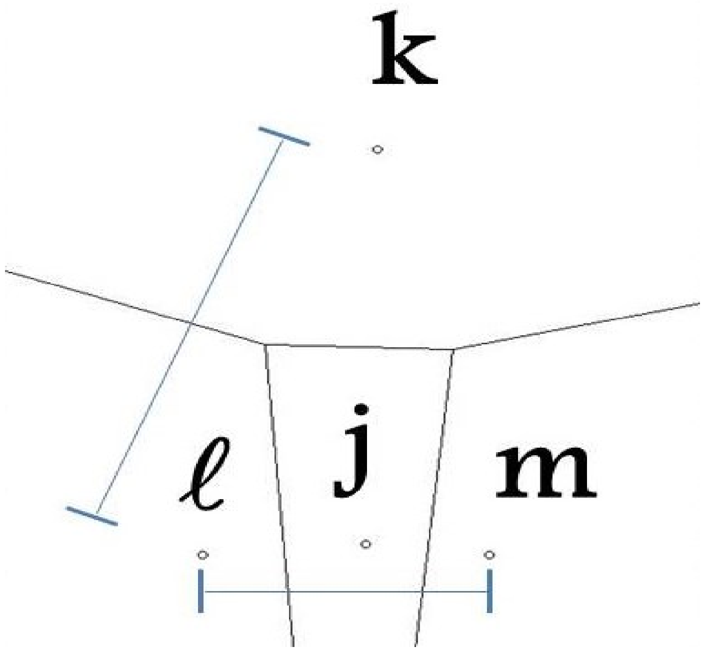
- *Not the topic here*



Quality measures
Compact
Connected
TI (ρ)
Stress
Shepard D.
Force Approach Error
TC
C measure
Minimal Pathlength
Minimal Wiring
Unidirectional
Direction based
König's measure
Precision & Recall
Topographic Product
co-ranking matrix
T&D
MRRE
LCMC
RAAR
Topographic Function
U-ranking
Zrehen's measure
SOM based
Topographic Error
Quantization Error

Unidirectional based versus direction based

- **Consider the following Voronoi Cells:**



(l,k) always neighbors in Delaunay Graph

but **dependend on knn** maybe neighbors in KNN-graph

(l,j) Always neighbors in Delaunay Graph and in KNN-graph

(l,m) Never neighbors in Delaunay Graph (or Gabriel Graph)

but **dependend on knn** maybe neighbors in KNN-graph

Connected quality measures

- Quality based on graph theory

=> Pros and Cons of the quality measurement are the Pros and Cons of the specific graph

- quality measuement F(I,O) is a function

⇒ Consideration of functional profile

- Hope: Only Cluster relevant Distances are considered

Example Trustworthiness & Continuity

$$\boldsymbol{T}(knn) = 1 - \frac{1}{N(knn)} * \sum_{j,} \sum_{l \in \mathrm{H(knn,O\backslash I)}} R(j,l) - \sum_{l \in \mathrm{H(knn,O\backslash I)}} l \qquad (9)$$

$$\boldsymbol{C}(knn) = 1 - \frac{1}{N(knn)} * \sum_{j,} \sum_{l \in \mathrm{H(knn,I\backslash O)}} r(j,l) - \sum_{l \in \mathrm{H(knn,I\backslash O)}} l \qquad (10)$$

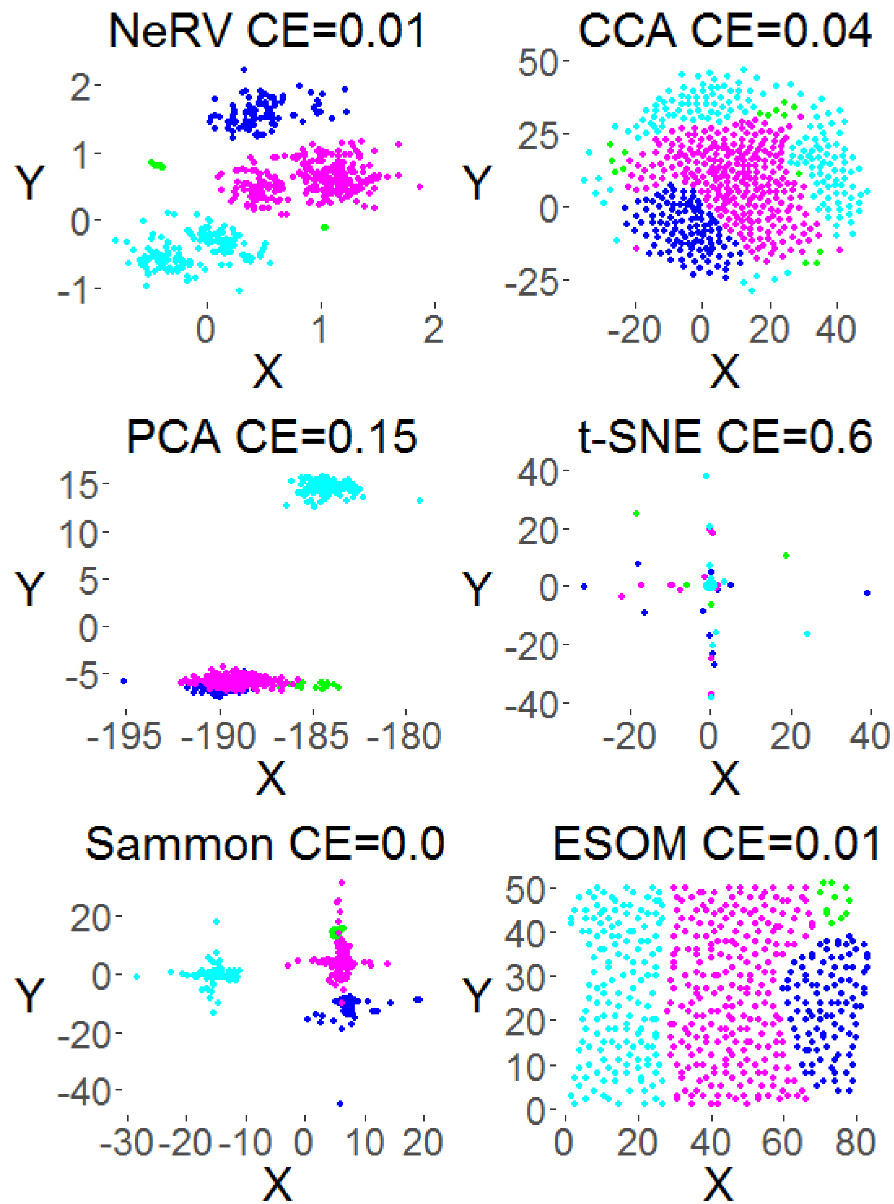
- Sort distances d(x,y) and assign consequitive numbers
  -> **Rang** $\boldsymbol{r}(\boldsymbol{j},\boldsymbol{l}) \in O$ and R$(\boldsymbol{j},\boldsymbol{l}) \in I$
- **Let for each point j, the points $l \in \mathrm{H_j(knn, O\backslash I)}$ be in the neighborhood of the Output space O, but not in the k nearest neighborhood (knn) of the Input space around the point j**
- The size of a Set in the neighborhood H is often defined by *knn*, and is a subset of I or O. We use the short Notation defined by
  $$\mathrm{H}_{knn}(x_j) \subset \mathrm{I} := \mathrm{H(knn, I)} \qquad (3)$$
- ideal arrangement in the neighborhood: $\sum_{l \in \mathrm{H(knn,I)}} l$

Practical Example: Leukemia

- 554 patients (points) with prior diagnosis: healthy, AML,APL,CLL

->prior classification (colors of pints)

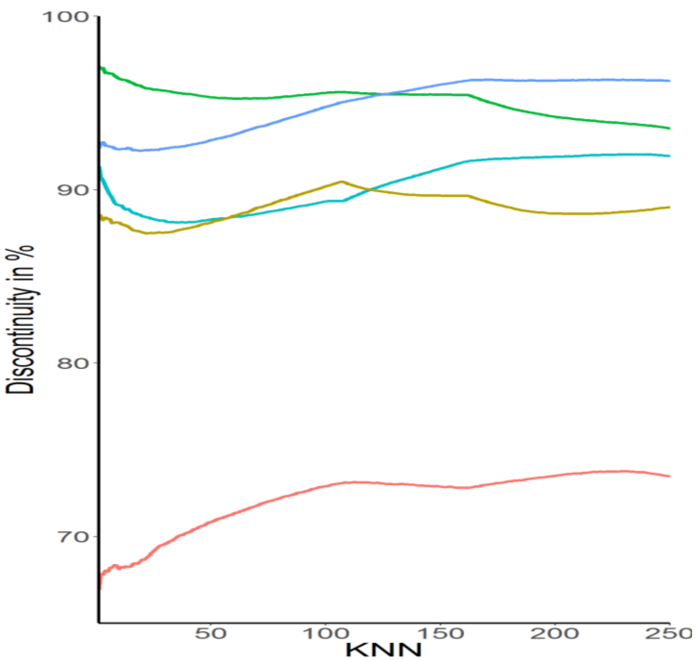
- ~8000 Genes -> ~8000 Dimensions

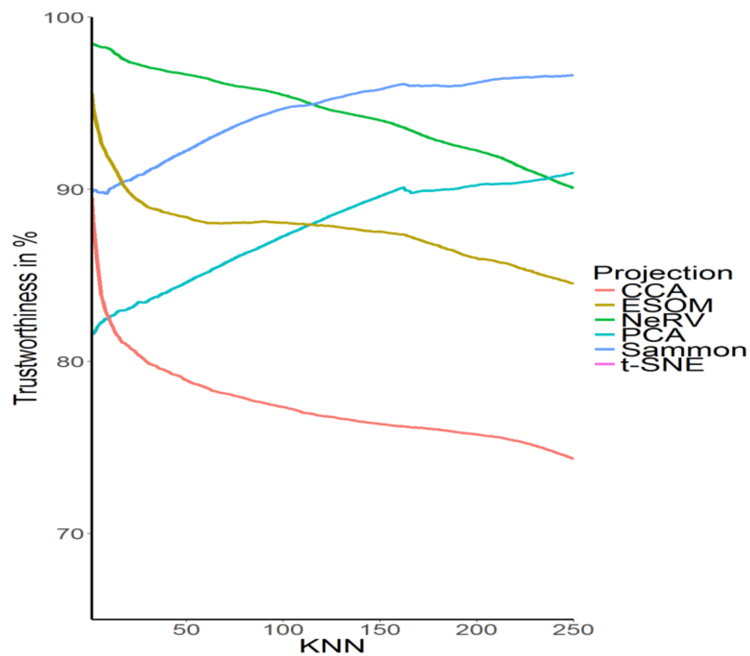


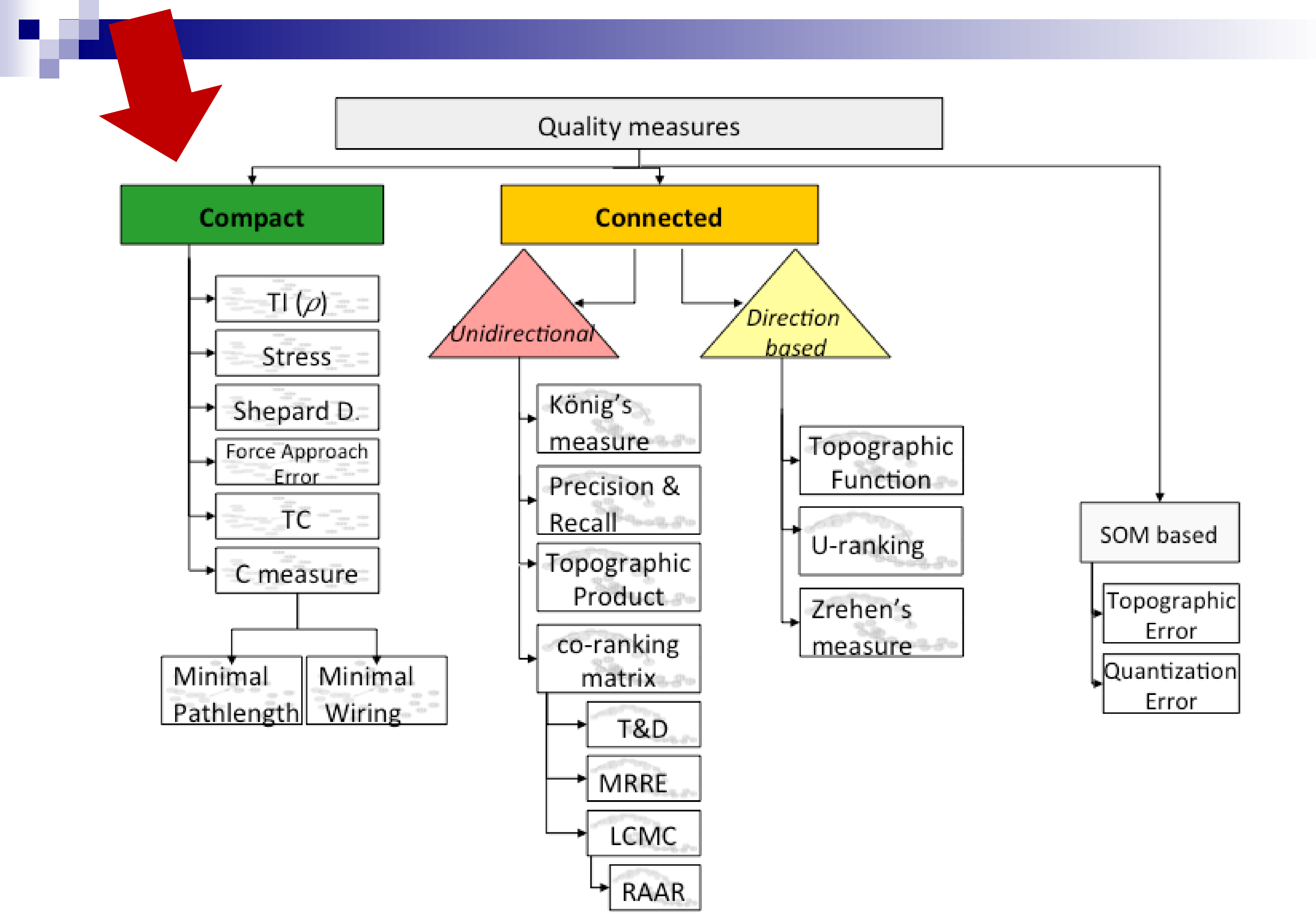
- Only ESOM is structure preserving!

Trustworthiness and Discontinuity

- Difficult to interpret
- Maybe NeRV and Sammon are the best
- Weiter oben ist beser
- Wo aufhören mit k fuer kNN
- Was ist wenn die Kurven sich schneiden?





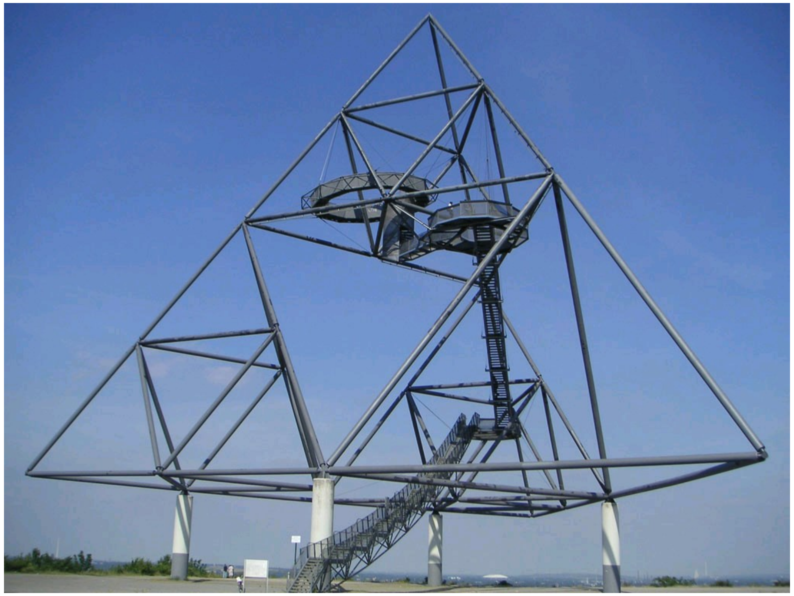


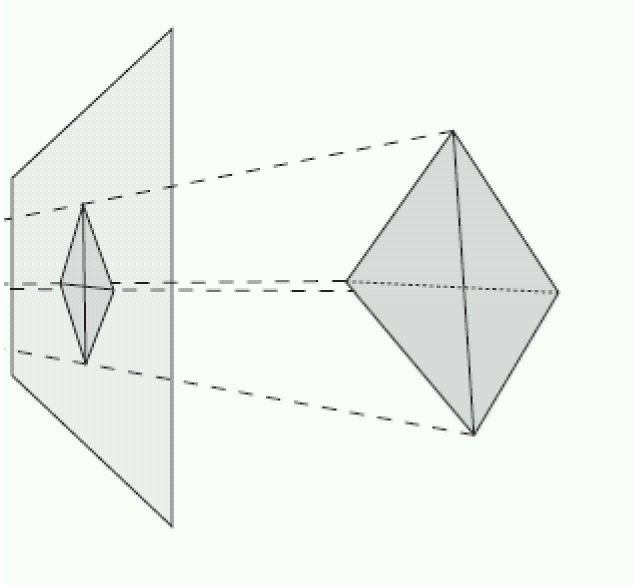
Quality measures
Compact
Connected
TI (ρ)
Stress
Shepard D.
Force Approach Error
TC
C measure
Minimal Pathlength
Minimal Wiring
Unidirectional
Direction based
König's measure
Precision & Recall
Topographic Product
co-ranking matrix
T&D
MRRE
LCMC
RAAR
Topographic Function
U-ranking
Zrehen's measure
SOM based
Topographic Error
Quantization Error

Distance based quality measures

- Distance D(I,j) and d(I,j) for all/some points are compared
- Ranks of distances
  - E.g.: Correlations like Spearmans Rho
- BUT: Preservation of all distance or even rank of distances is not possible!

Examples:





Example: C Measure

$$\sum_l \sum_j D(l,j) * d(l,j)$$

- C ist das Produkt aller Distanzen der Gewichtsvektoren und aller Kartendistanzen für alle Neuronen
- C wird maximal, wenn die Rangfolge der Abstände in Eingabe- und Kartenraum übereinstimmt.

- Sub categories, e.g. $\sum_l \sum_j D(l,j) * s(l,j)$, where $s(k,j)$ defines the k nearest neighbors with knn=1

Analyse C-measure

Man könnte das C-Mass einfach normieren

Summe(sort(Dij) * sort(kij)) ist das Maximum =100%

Summe(sort(Dij) * antisort(kij)) ist das Minimum

Dann könnte man C in % angeben

Minimal Pathlength and Minimal Wiring

Two C-Variants: Minimal Pathlength and Minimal Wiring

Number (5) presents the definition of the Minimal Pathlength [Durbin/Mitchison 1990] and (6) the definition of the Minimal Wiring [Mitchison 1995]

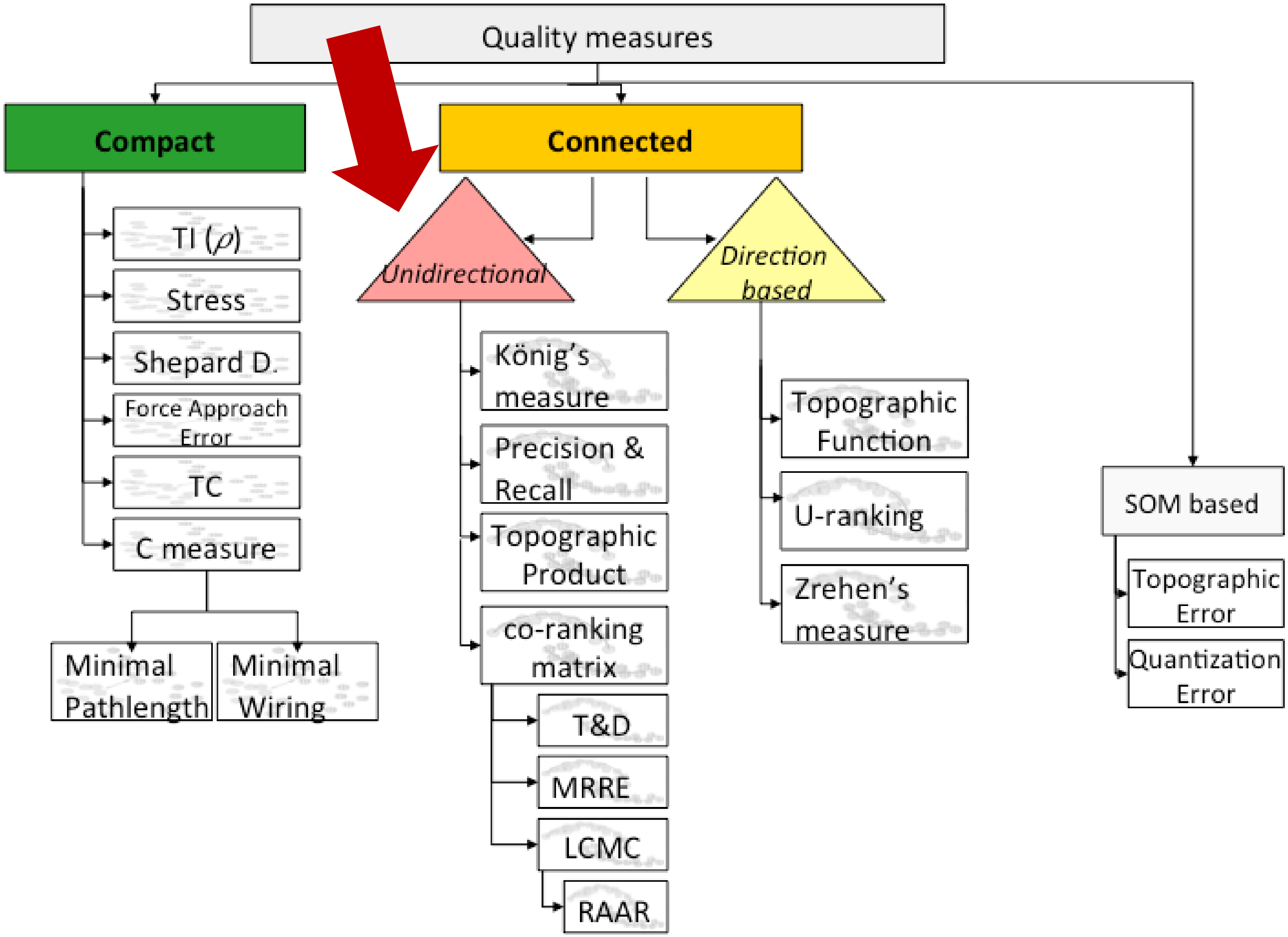
$$F = \sum_{j,l} D(j,l) \cdot s(j,l) \qquad (5)$$

$$F = \sum_{j,l} \mathrm{d}(\mathrm{j},\mathrm{l}) \cdot s(j,l) \qquad (6)$$

where $s(k, j)$ defines the k nearest neighbors. Thus, it is analogical to the KNN graph, e.g. [Brito et al. 1997];

$$s(j,l) = \begin{cases} 1: & j \in H(knn = 1, IorO) \\ 0: & otherwise \end{cases}$$

where $H(knn = 1)$ defines a set of the nearest space neighbors within the Input space I in (5) and within the Output space O in (6). So the measurement is a mixture of Euclidean graph and KNN graph with


Quality measures
Compact
Connected
Unidirectional
Direction based
TI (ρ)
Stress
Shepard D.
Force Approach Error
TC
C measure
Minimal Pathlength
Minimal Wiring
König's measure
Precision & Recall
Topographic Product
co-ranking matrix
T&D
MRRE
LCMC
RAAR
Topographic Function
U-ranking
Zrehen's measure
SOM based
Topographic Error
Quantization Error

Unidirectional quality measures (QMs)

- Based on KNN-Graphs

1. One „right“ k is chosen, e.g. Königs Measure, LCMC
2. Functional profile by calculation a lot of k‘s, e.g. MRRE, T&D

LCMC

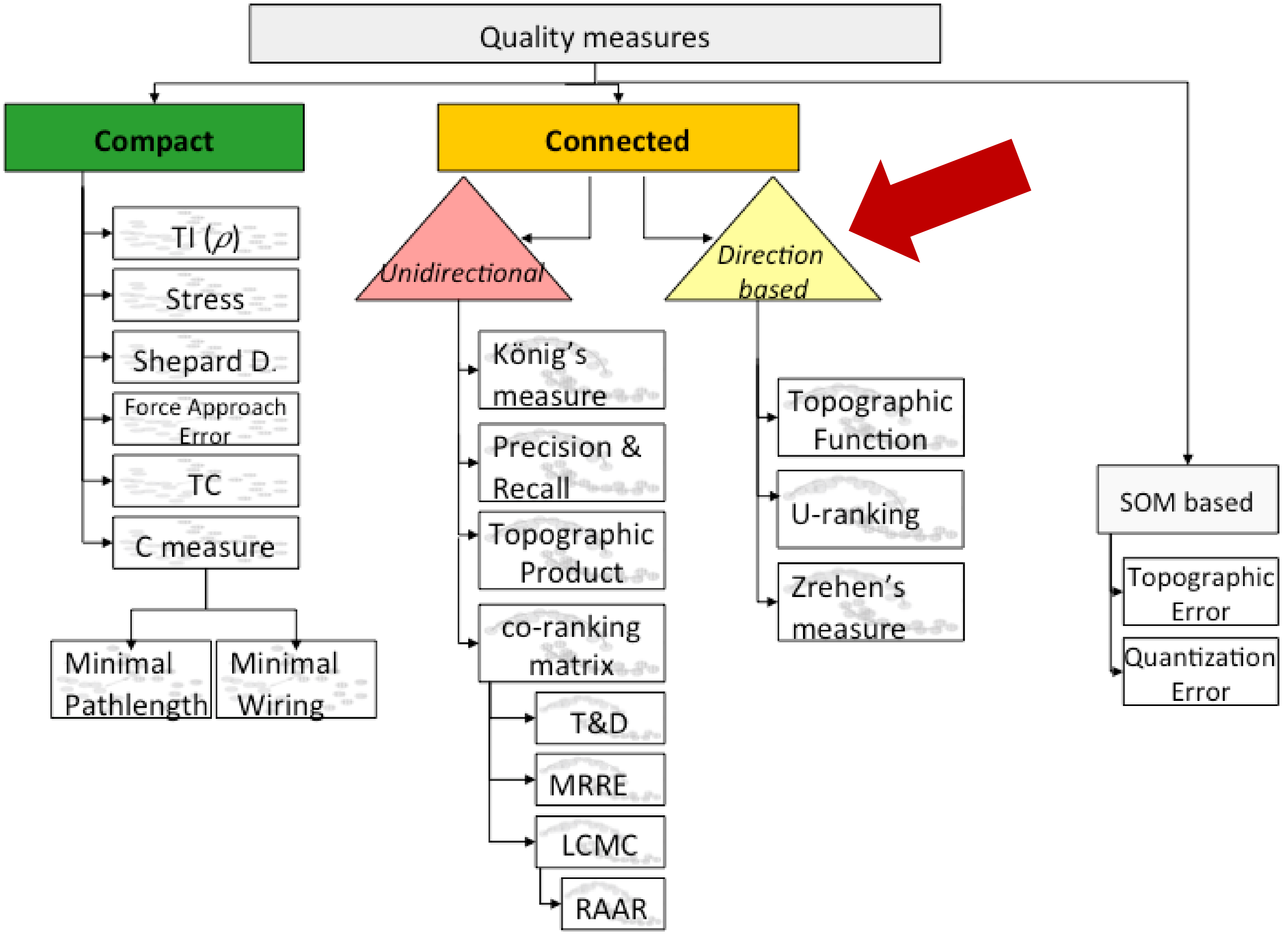
LCMC is defined as the average size of an overlap of k nearest neighborhoods in I and O [Lisha Chen/Buja 2009]:

$$\mathrm{A}(j) = \left| H(knn, I) \ \cap H(knn, O) \ \right|, \qquad \overline{A_{knn}} = \frac{1}{N} \sum_{j=1}^{N} A(j) \qquad (11)$$

- For each $x_j \in I$ and $\mathrm{w}_j \in O$ there is a set of points in the neighborhood $H(knn, I)$ and $H(knn, O)$, which are calculated with a given knn of an KNN-graph. The overlap is measured pointwise as in (7)

$$F(knn) = \frac{1}{knn} \overline{A_{knn}} - \frac{knn}{N-1} \qquad (12)$$

- The mean $\overline{A_{knn}}$ is normalized with *knn*, because it is the upper bound of $\overline{A_{knn}}$ and adjusted by modelling a hypergeometric distribution with *knn* defectives out of N-1 items and *knn* draws.
- In contrast to T&D and Mean Relative Rank Error, LCMC accounts for things that go well


Quality measures
Compact
Connected
TI (ρ)
Stress
Shepard D.
Force Approach Error
TC
C measure
Minimal Pathlength
Minimal Wiring
Unidirectional
Direction based
König's measure
Precision & Recall
Topographic Product
co-ranking matrix
T&D
MRRE
LCMC
RAAR
Topographic Function
U-ranking
Zrehen's measure
SOM based
Topographic Error
Quantization Error

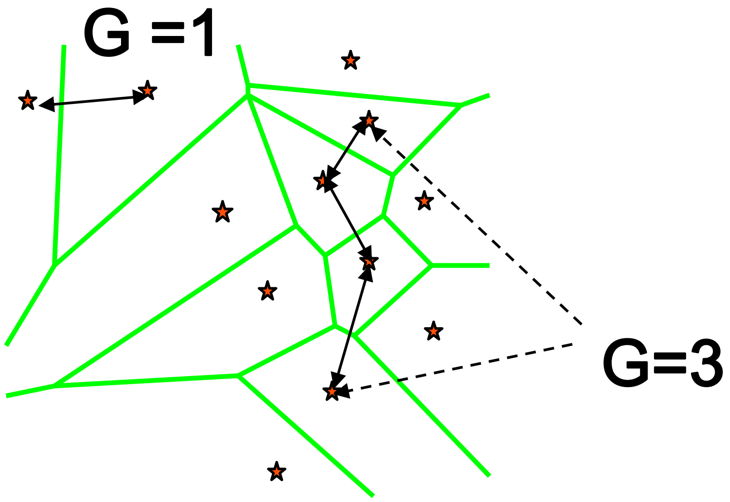
Direction based quality measures (QMs)

- Based on Delaunay Graphs
- Based on Gabriel Graph: Zrehen
- Quality measurements resulting in one value

Distances in a Graph

$$\phi(j,h) = \#\{\forall l \in I: g(l,j,\mathcal{D}) > h \wedge G(l,j,\mathcal{D}) = 1\}, \qquad h > 0$$
$$\phi(j,h) = \#\{\forall l \in I: g(l,j,\mathcal{D}) = 1 \wedge G(l,j,\mathcal{D}) > |h|\}, h < 0$$

(18): Delaunay Path Distance between fixed point j and all points l with g$(l,j,\mathcal{D})$>h in Output O, where the Delaunay Cells of (l,j) are neighbors in Input ($G(l,j,\mathcal{D})$=1)



Example: Topographic Function

- TF quantifies the identity of the Delaunay graphs in I and O

$$F(h) = \frac{1}{N} \sum_{j=1, j \in I}^{N} \phi(j,h) \qquad h \neq 0 \;\; (17)$$

$$\phi(j,h) = \#\{\forall l \in I: g(l,j,\mathcal{D}) > h \wedge G(l,j,\mathcal{D}) = 1\}, \qquad h > 0$$

$$\phi(j,h) = \#\{\forall l \in I: g(l,j,\mathcal{D}) = 1 \wedge G(l,j,\mathcal{D}) > |h|\}, h < 0$$

- The shortest path in the Delaunay graph $\mathcal{D}$ of the Input space between the data points $(l,j) \in I$ is $G(l,j,\mathcal{D}) =$ and between projected points$g(l,j,\mathcal{D})$
- The Delaunay graph's distances $G(l,j,\mathcal{D})$ and $g(l,j,\mathcal{D})$ are equivalent to the number of Voronoi cells between the two points.

Example Topographic Function II

- **h=1: Delaunay Graphs in I and O are the same**
- **If *h* is greater than zero, (x_j, x_l) are neighbors in the Input space and if *h* is smaller than zero $\mathrm{w}_j, \mathrm{w}_l$ are neighbors in the Output space.**
- **"*Small values of h indicate that there are only local dimensional conflicts, whereas large values indicate the global character of a dimensional conflict*" [Villmann et al. 1997]. Therefore, [Bauer et al. 1999] proposed the simplified equation (20):**

$$F(1) + F(-1) \qquad (20)$$

- **h equals zero if and only when two points are neighbors in Input space and Output space, thus the overlap of Voronoi neighbors in I and O is required.**

Example Topological Correlation

The shortest path in the Delaunay graph of the Input space between the data points $(x_j, x_l) \in I$ is $Del(j,l)$ and between projected points $(\mathrm{w}_j, \mathrm{w}_l) \in O$ is $del(j,l)$.

$x = \frac{1}{\frac{N(N-1)}{2}} \sum_{l=2}^{N} \sum_{j=1}^{l-1} \quad Del(l,j)$, y analog mit del(l,j)

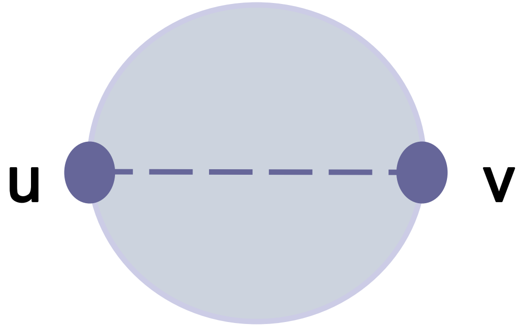
Pearsons Correlationcoefficient

TC=1/N* $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

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

Zrehens Mass

Idee: man nehme 2 im Ausgaberaum unmittelbar benachbarte Neuronen u und v und die zugehörigen Datenvektoren (Gewichtsvektoren) u und v im Eingaberaum



m Umkreis um u-v soll kein weiterer Punkt des Datenraums vorkommen

(Gabriel Graph-Bedingung:) Empty Ball condition / Umkreisbedingung

Pros and Cons: compact QMs

- Pro:
  - Only one value describes the quality
  - Range of values is specified and meaningfull

- Cons:
  - Measurements based on correlations describe linear relationsships between D(l,j) and d(l,j)
  - Outliers/Extremes are overweighted
  - Preservation of the whole arrangement of points is measured

    => Structure preservation is not considered

Pros and Cons: unidirectional based

- Pro:
  - Focusing with local neighborhoods
  - KNN-Graphs are easily computable in R^n
  - BPE and FPE considering by two different functions F(I,O)
- Cons:
  - Structure preservation is only sometimes considered
  - The right k for KNN-graph is unknown
  - Functional profile is abstract and for different projection methods not easily comparable

Pros and Cons: direction based

- Pro:
  - Focusing with local neighborhoods
  - Considering BPE, FPE, Gaps
  - One value => different projection methods comparable

- Cons:
  - Graphs are difficult to compute for R^n
  - Range of the value not always specified
  - Do the quality measures really show structure preserving submanifolds?

Zusammenfassung

- Keine Projektion ist perfekt, kann sie auch nicht sein
- Für Clusterhafte (Daten mit „Lücken) ist die Bestimmung eines Fehlermaßes für Projektionen ungelöst
- Fehlermaße lassen sich anhand der in ihnen getroffen Vornahmen gruppieren
-
- Unser Ansatz: Sichtbarmachung von Fehlern
- Generalisierte Umatrix =Umatrix für beliebige Projektion nach $\mathbb{R}^2$ zeigt die Fehler

Thank you for listening