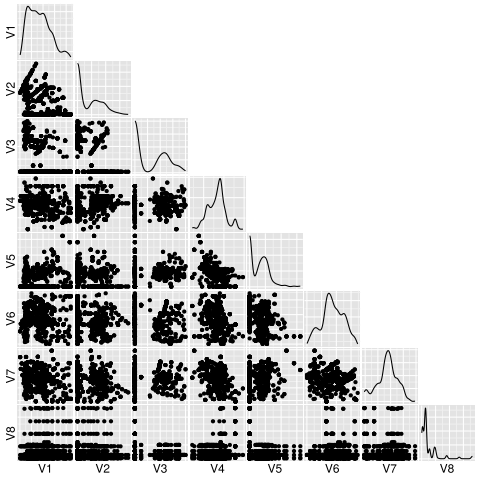
Analysis of quality measures for dimensionality reduction

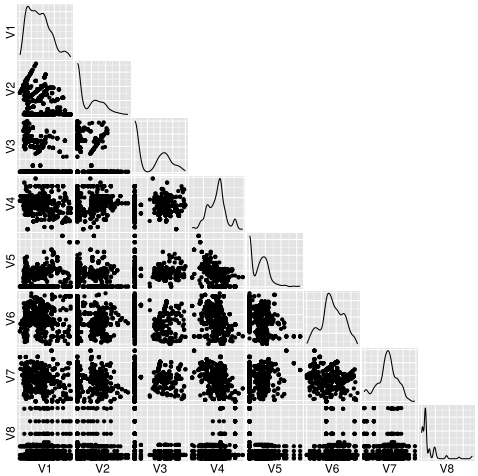
Bastian Rieck Heike Leitte

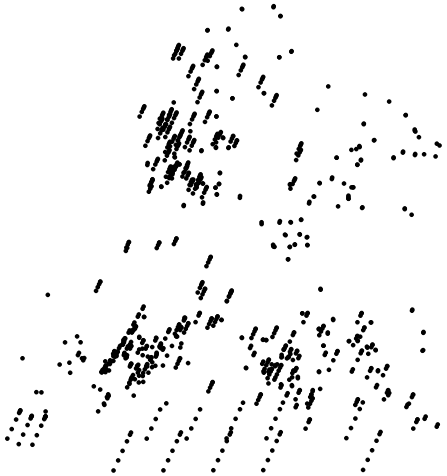
Interdisciplinary Center for Scientific Computing
Heidelberg University

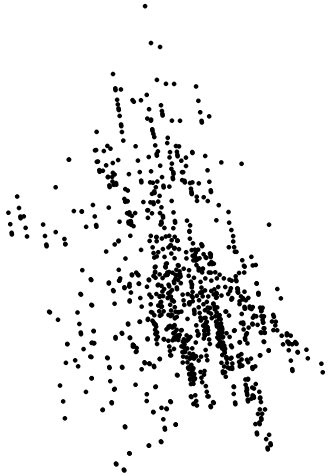
Motivation



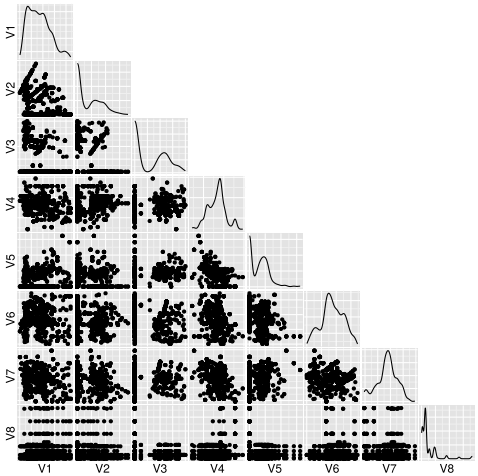
Motivation

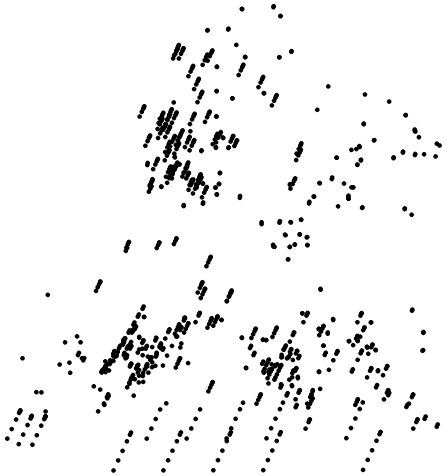


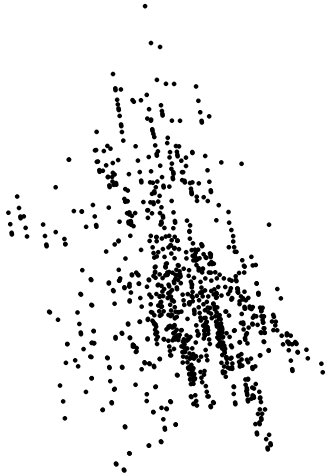




Motivation





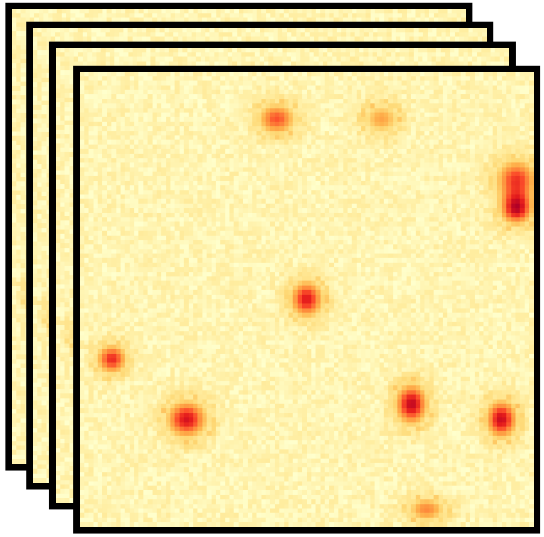


$$Q_{\text{Stress}}(x_i) = \sqrt{\sum_{j=1}^{n} \left(d_{ij} - \delta_{ij}\right)^2 / \sum_{j=1}^{n} \delta_{ij}^2}$$

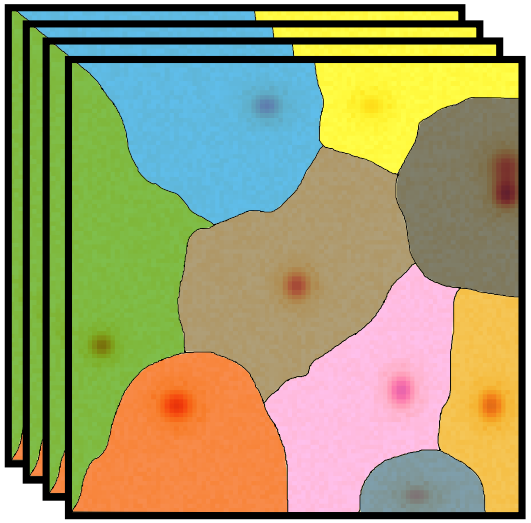
Q_{Stress}, $Q_{\text{Residual variance}}$, Q_{RMSE}, $Q_{\text{Spearman}} \cdots$

- Is more than one aspect of the data preserved by a method?
- How to compare different quality measures with each other?
- Do multiple quality measures *agree* on the data set?

Our method

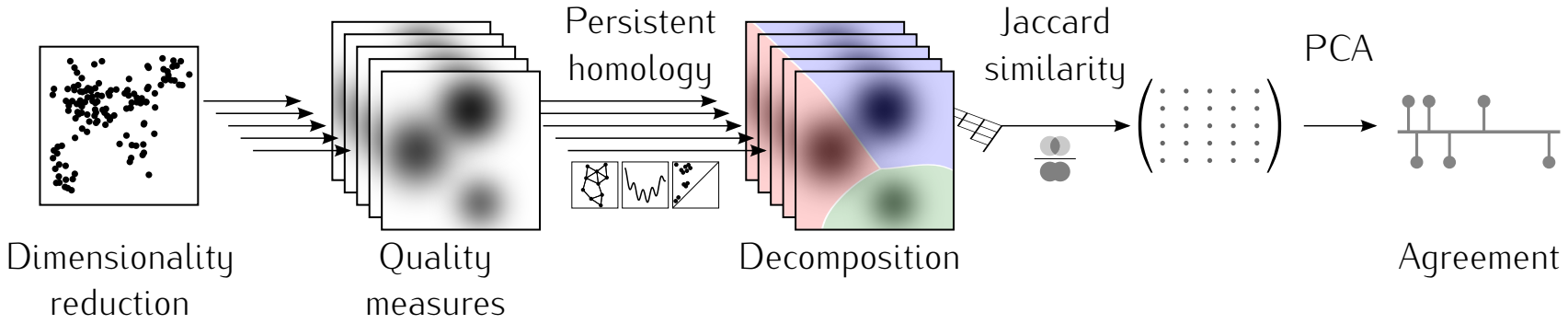


Quality measures



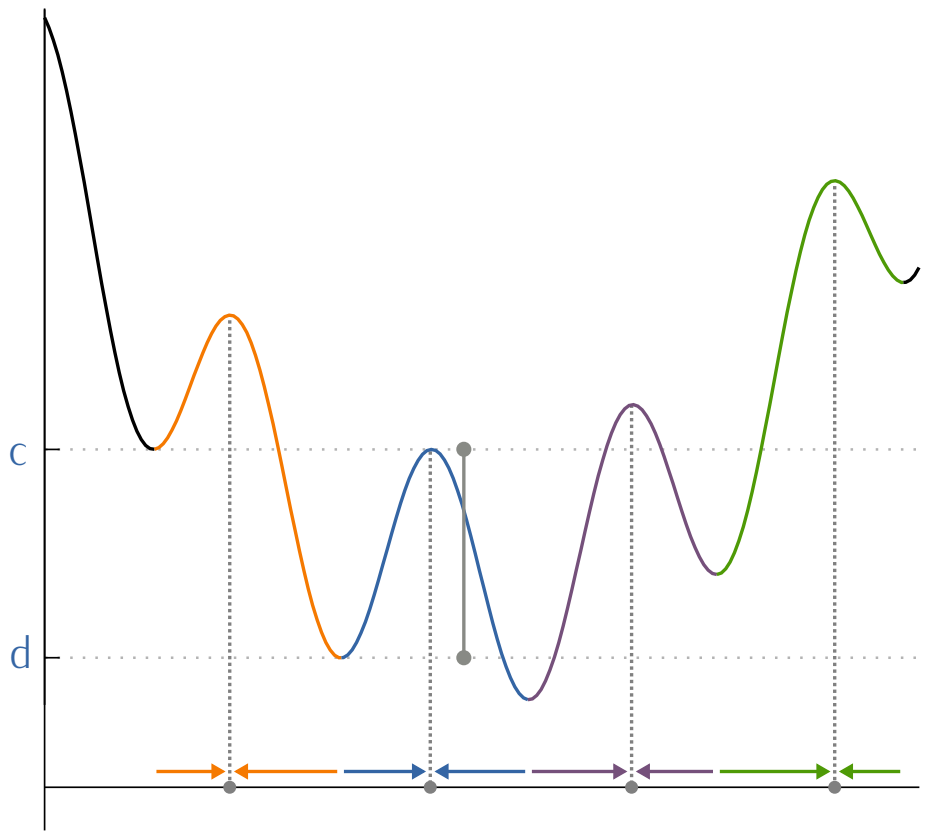
Decompositions

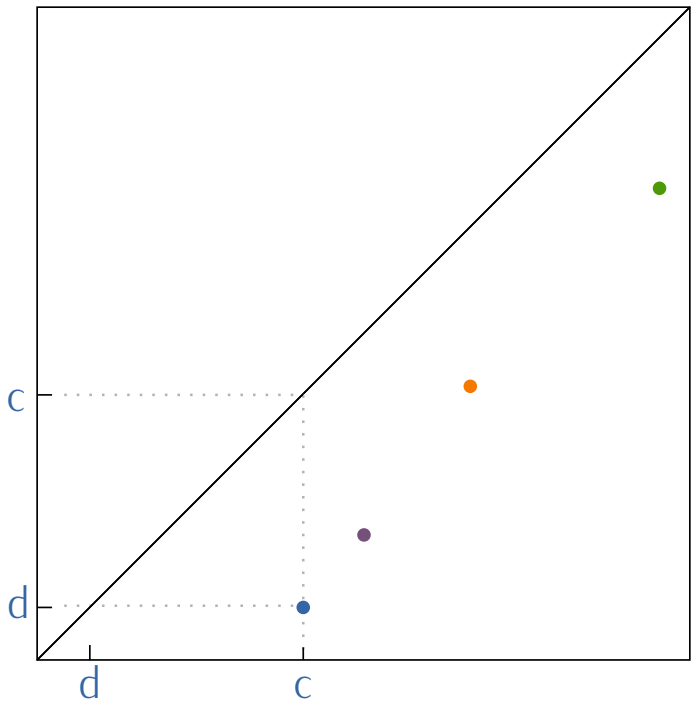
Workflow


Persistent homology
Jaccard similarity
PCA
Dimensionality reduction
Quality measures
Decomposition
Agreement

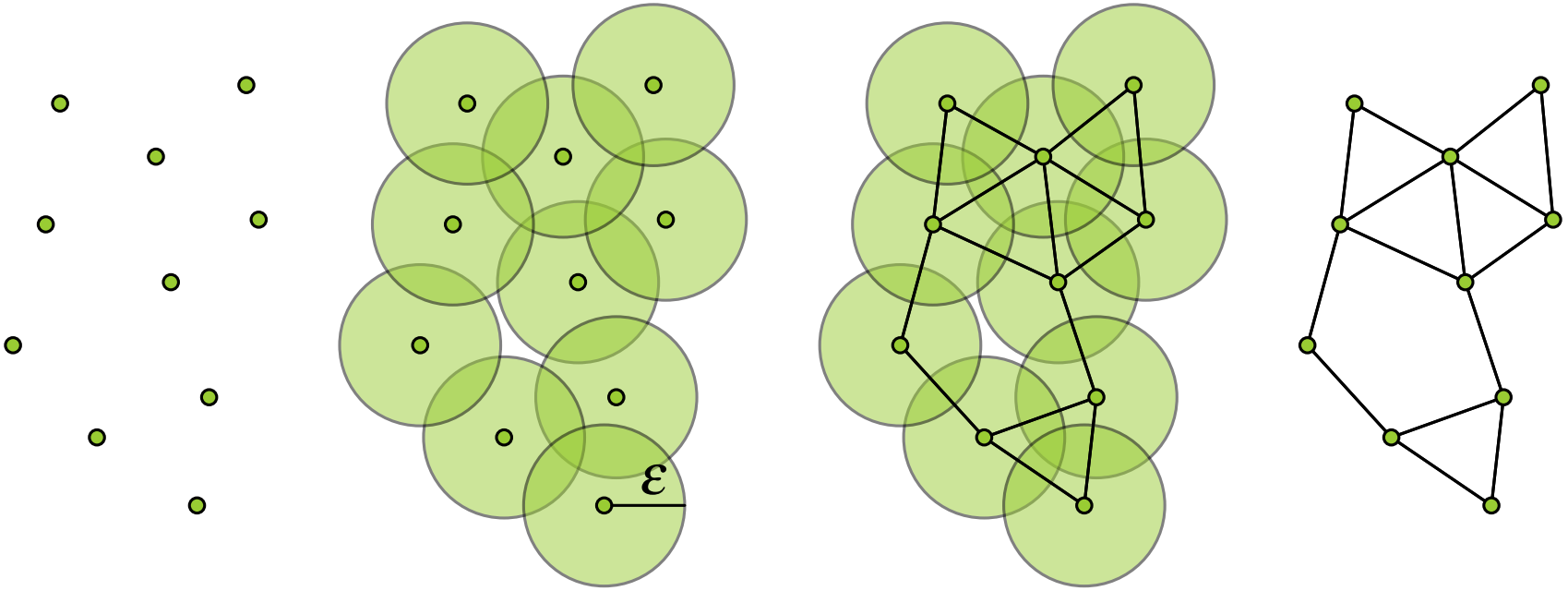
Persistent homology

Superlevel set filtration





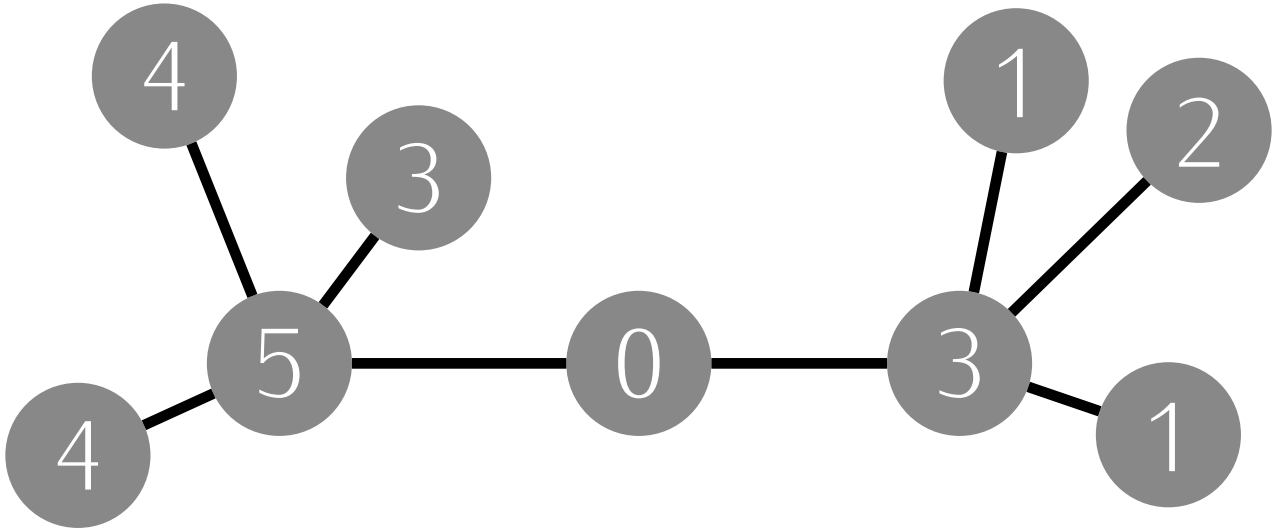
High-dimensional data



- Rips graph calculation to obtain neighbourhoods.
- Distance threshold estimation controls coarseness of approximation.

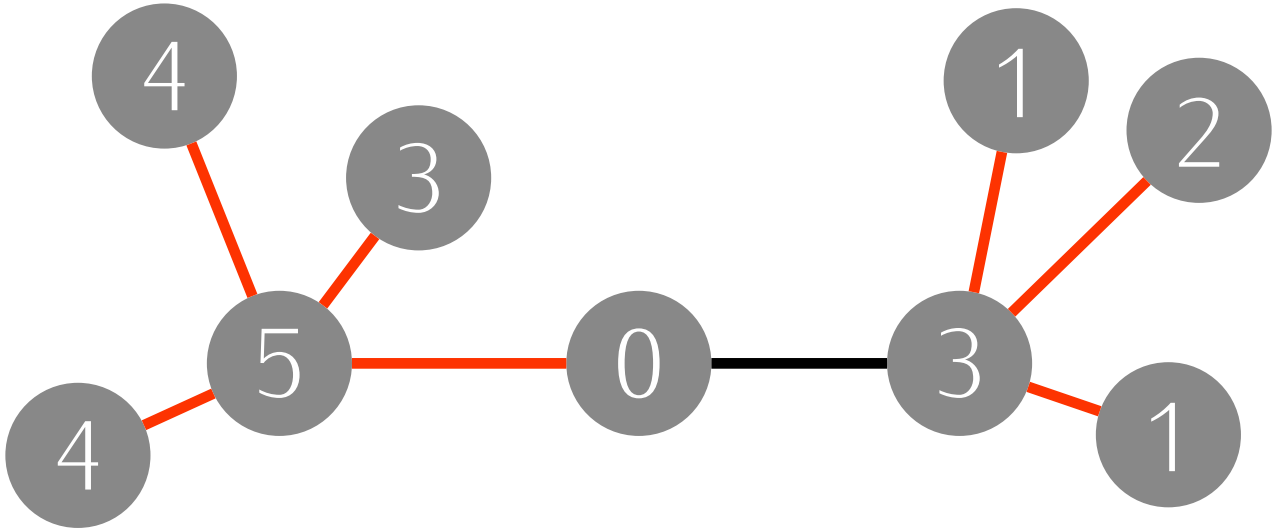
High-dimensional data

Obtaining a decomposition



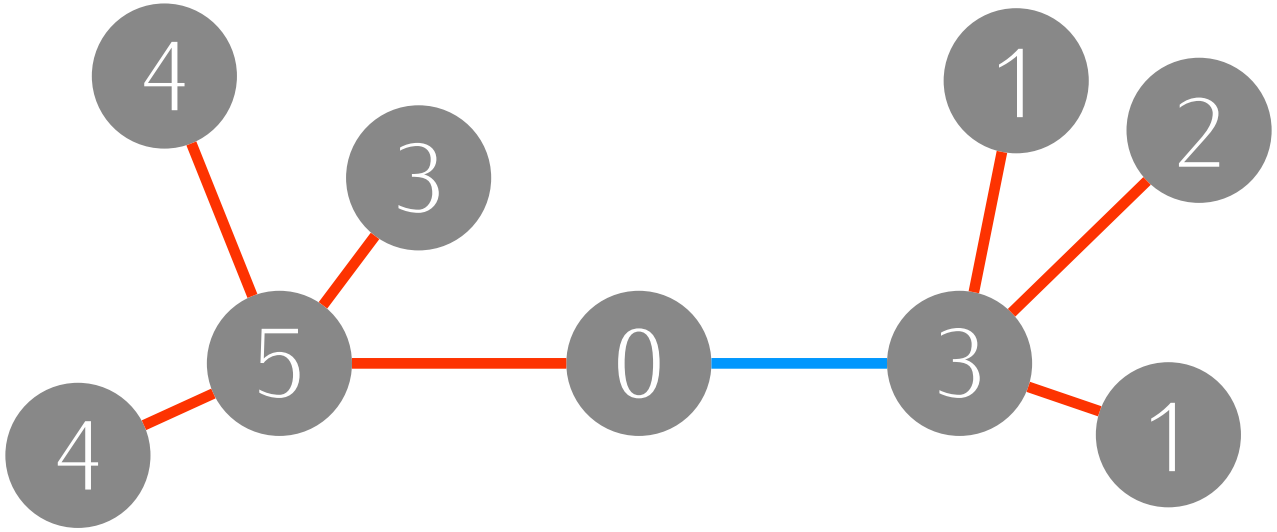
Example

Peak-seeking



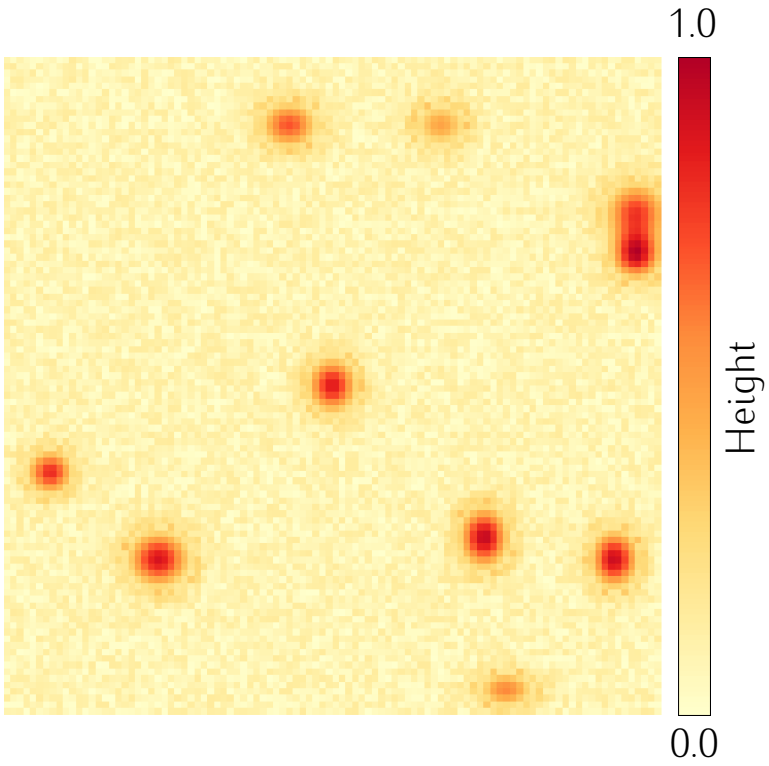
Example

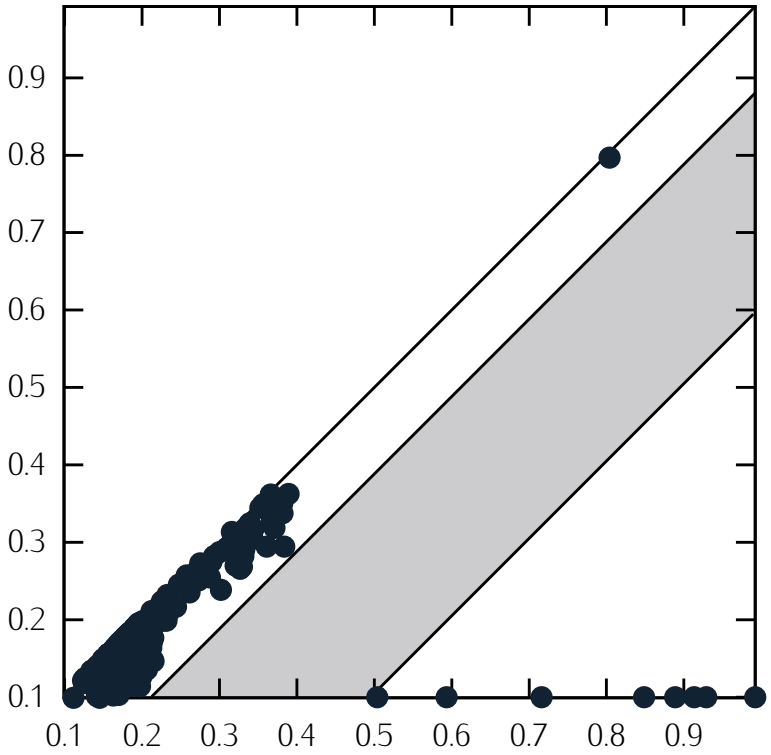
Merging



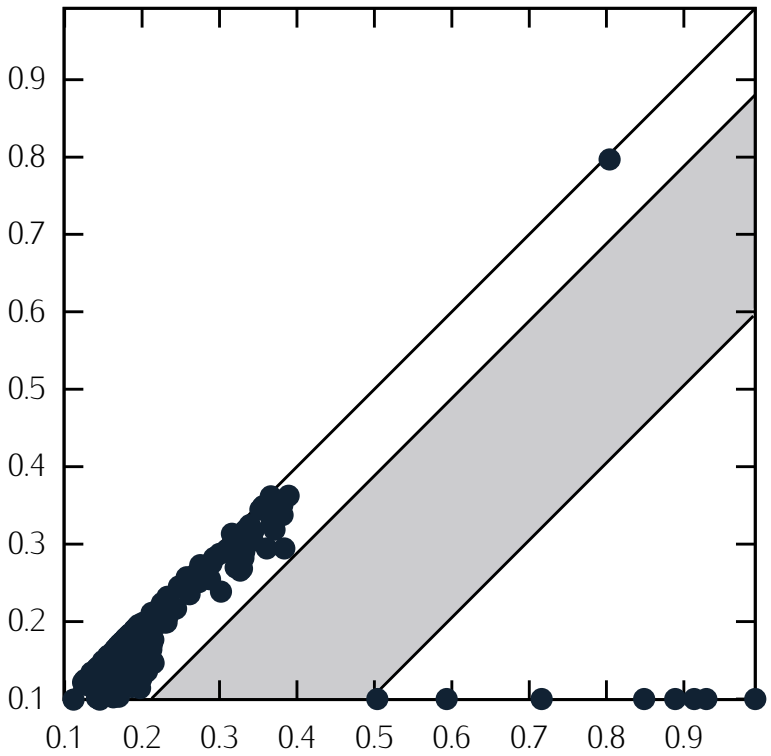
Merge threshold estimation

Chazal et al.: If data is sampled sufficiently dense, “relevant” peaks and topological noise are well-separated.

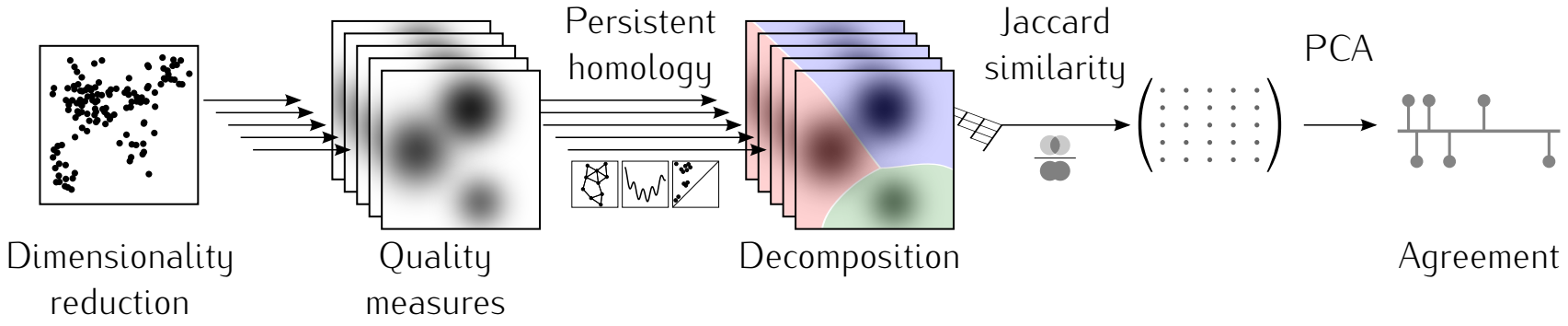




In practice



- Relate size of largest empty region to average size of empty regions.
- If ratio is large enough, consider region to be significant.

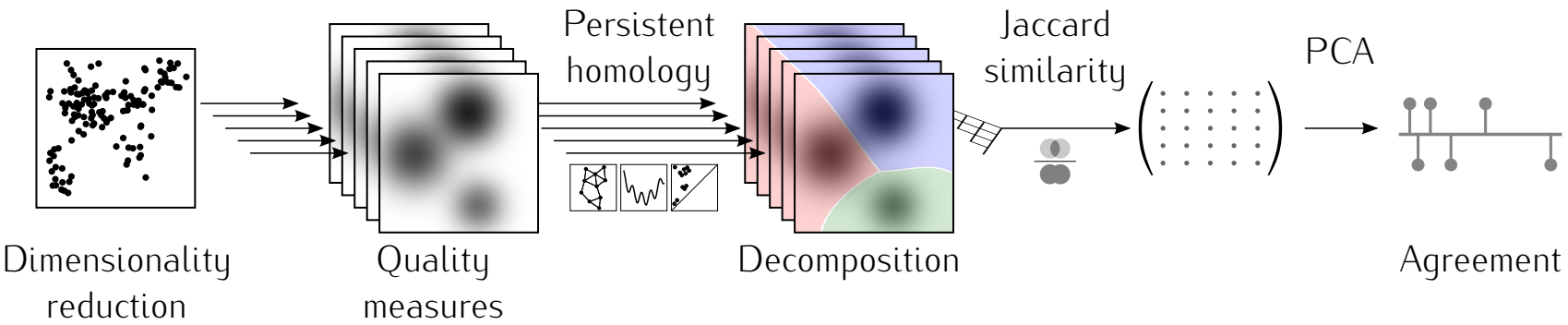

Persistent homology
Jaccard similarity
PCA
Dimensionality reduction
Quality measures
Decomposition
Agreement

Decomposition comparison

Jaccard index

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Visualization



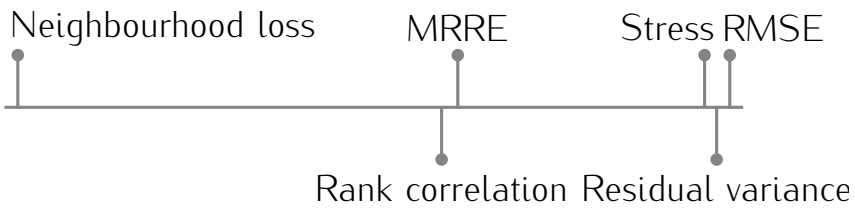
Encode local costs in scatterplot (pointwise).

Compressive strength

- 1030 cement mixtures with 8-dimensional feature vectors.
- Linear substructures.

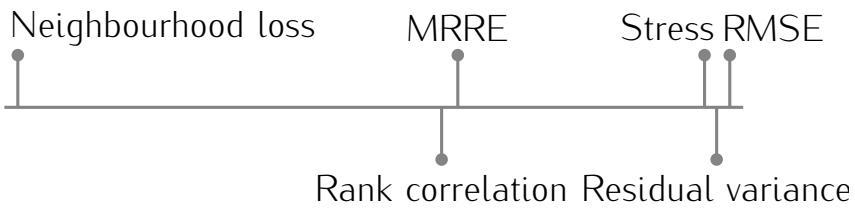
Compressive strength

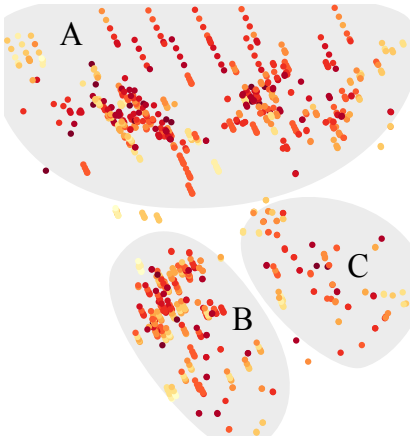
- 1030 cement mixtures with 8-dimensional feature vectors.
- Linear substructures.



Compressive strength

- 1030 cement mixtures with 8-dimensional feature vectors.
- Linear substructures.

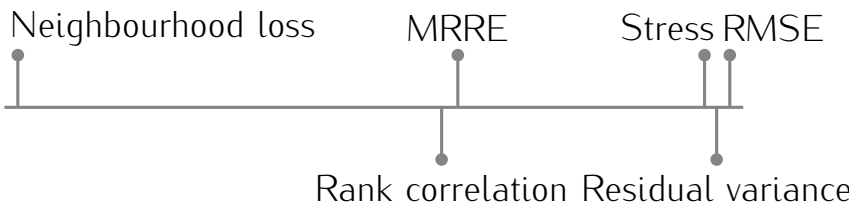


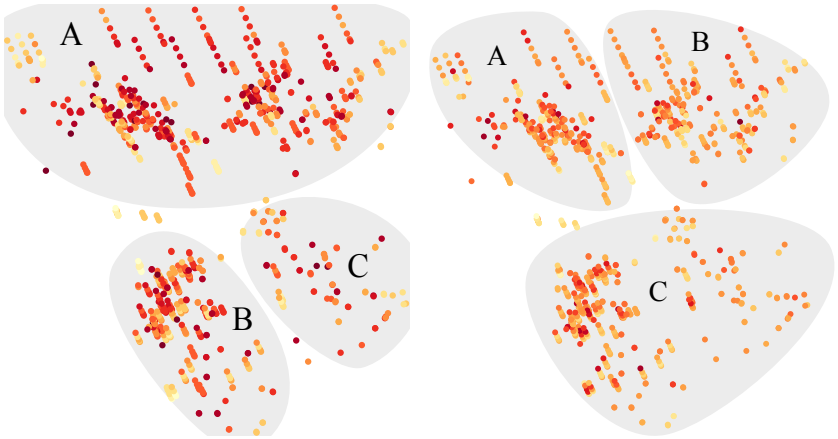


Nb'hood loss

Compressive strength

- 1030 cement mixtures with 8-dimensional feature vectors.
- Linear substructures.

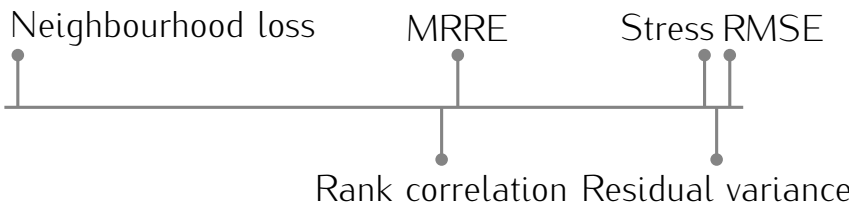


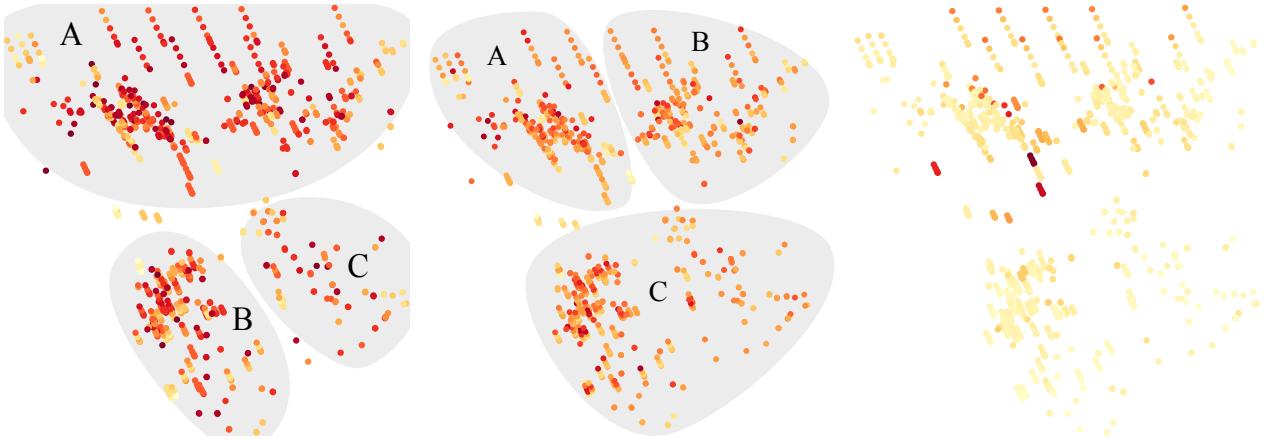


Nb'hood loss MRRE

Compressive strength

- 1030 cement mixtures with 8-dimensional feature vectors.
- Linear substructures.

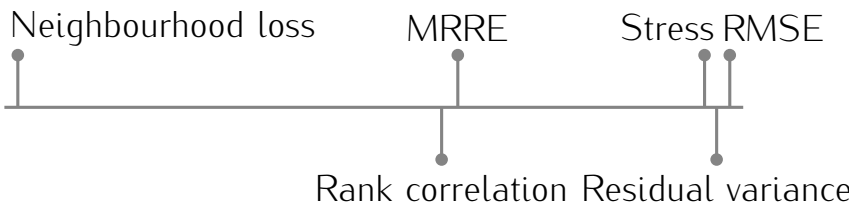


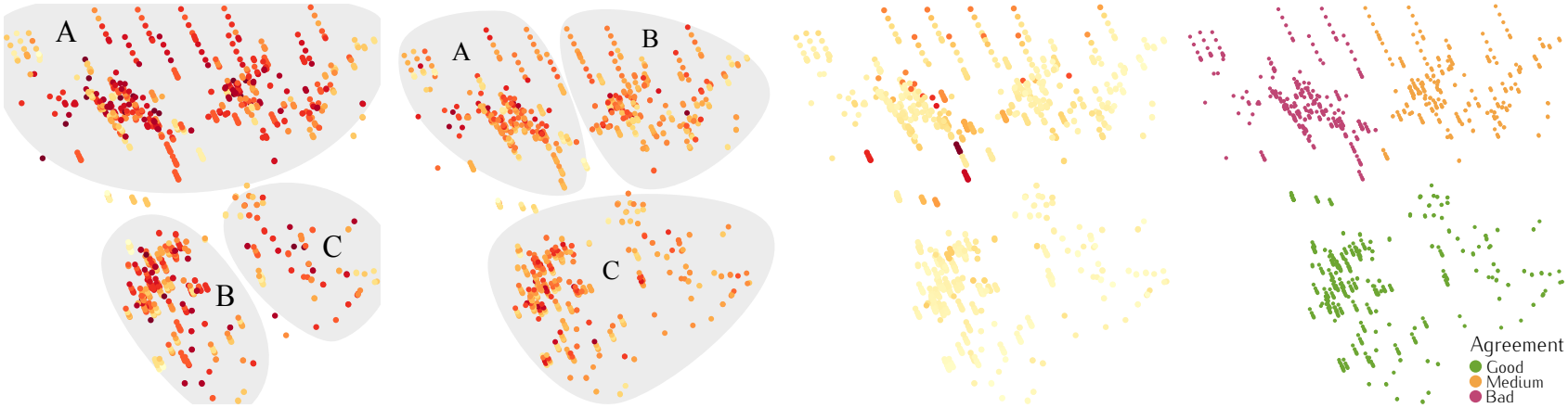


Nb'hood loss MRRE RMSE

Compressive strength

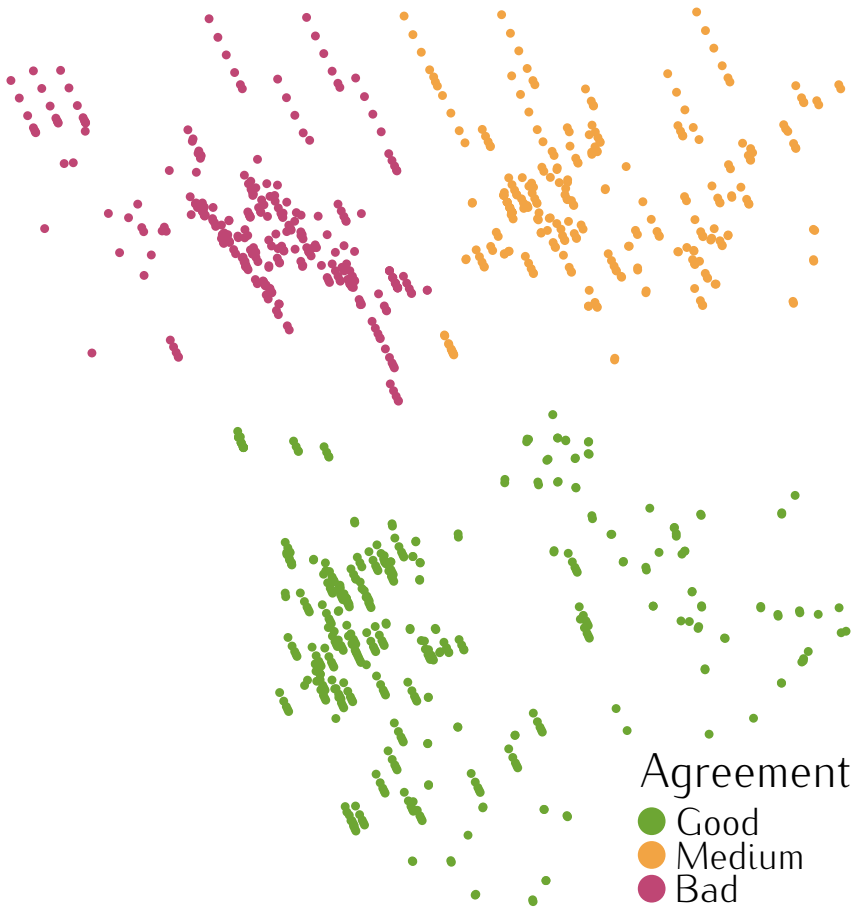
- 1030 cement mixtures with 8-dimensional feature vectors.
- Linear substructures.





Nb'hood loss MRRE RMSE Local costs

Compressive strength



- Upper region: Rank-based and distance-based measures disagree
- Misrepresentation possible

Handwritten digits

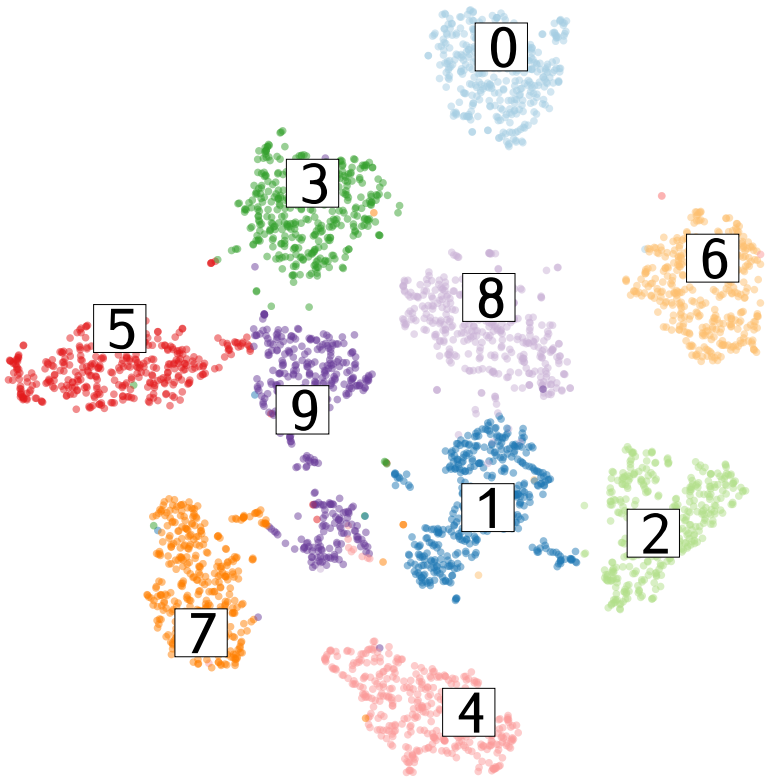
Comparing linear and non-linear dimensionality reduction

- 5620 instances of 64-dimensional feature vectors.
- Handwritten digits of different writers.

Handwritten digits

Comparing linear and non-linear dimensionality reduction

- 5620 instances of 64-dimensional feature vectors.
- Handwritten digits of different writers.

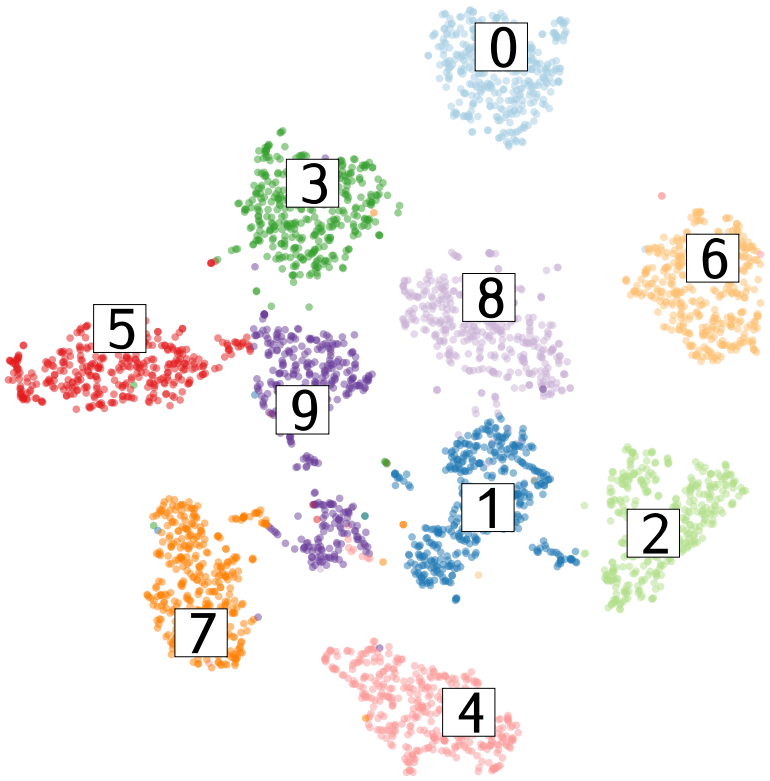


t-SNE

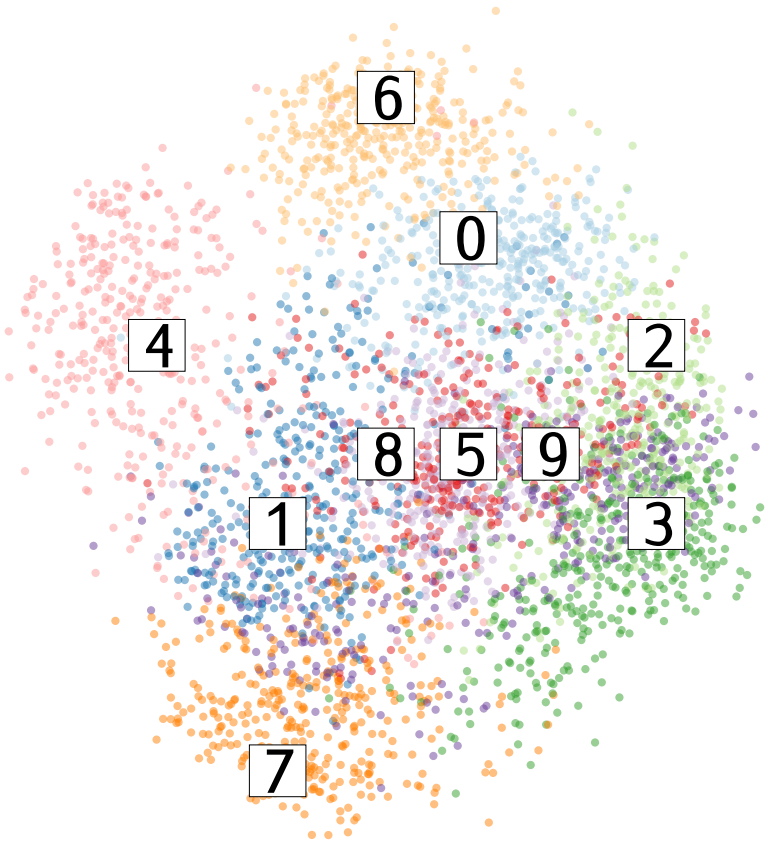
Handwritten digits

Comparing linear and non-linear dimensionality reduction

- 5620 instances of 64-dimensional feature vectors.
- Handwritten digits of different writers.

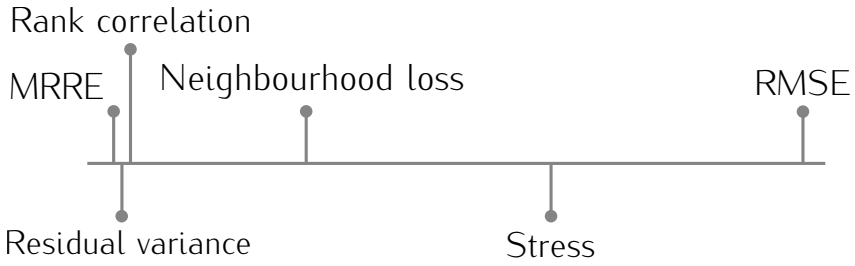


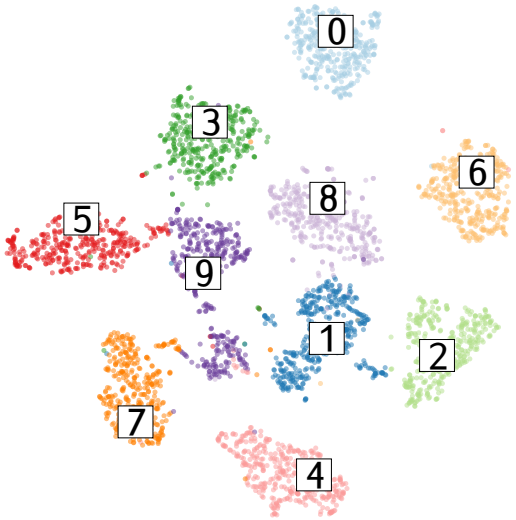
t-SNE

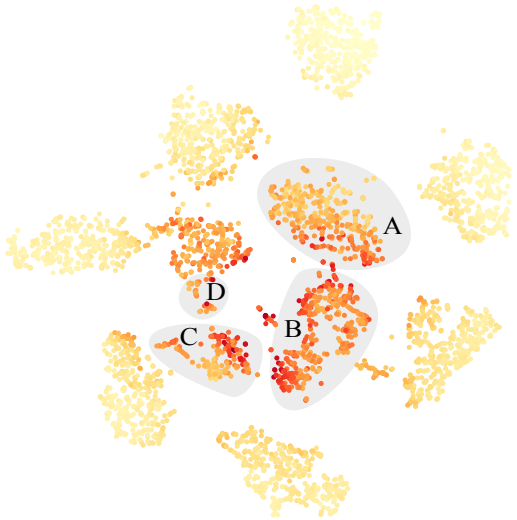


PCA

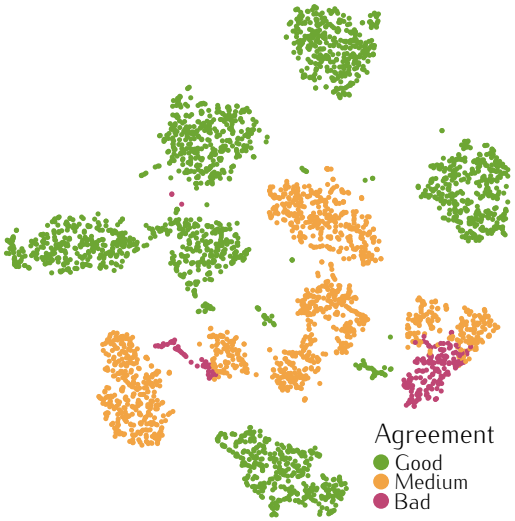
t-SNE





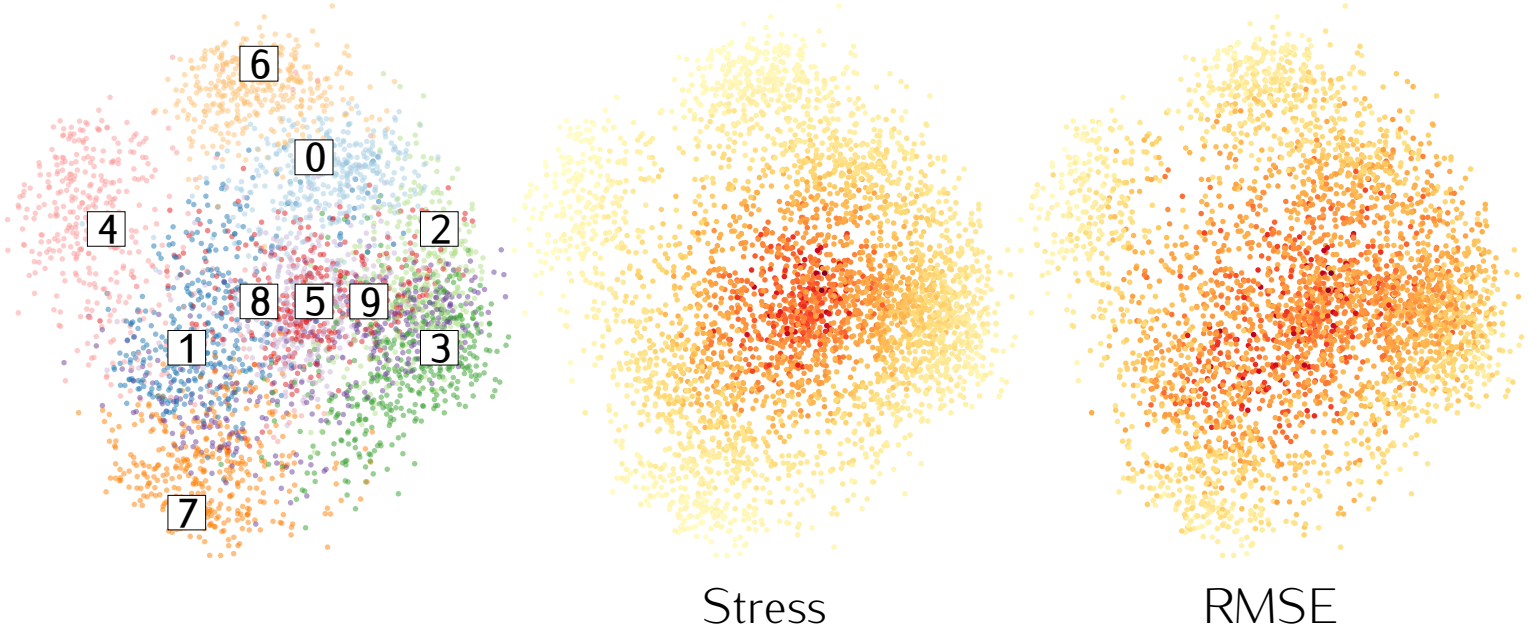


Stress



Local costs

PCA



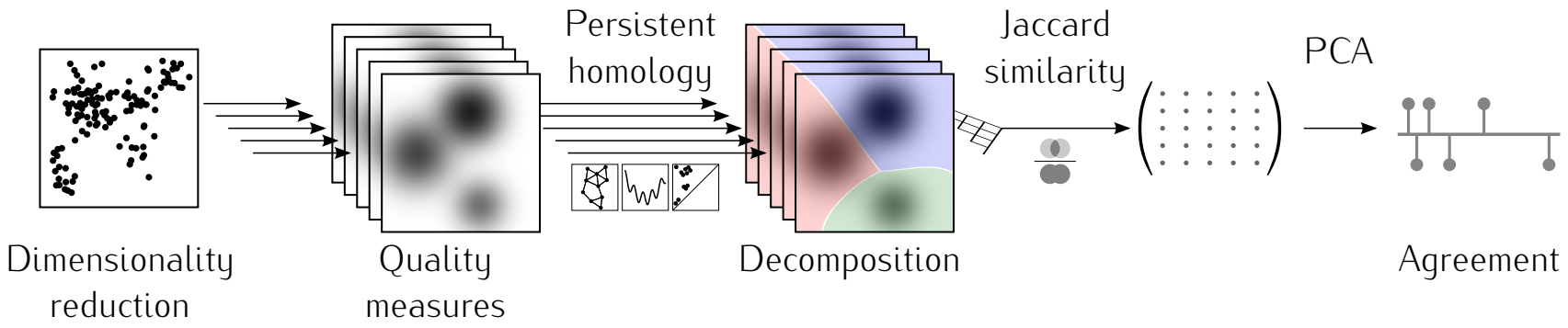
- Quality measures highlight a single persistent peak in the data.
- Peak is centred around the region of digits 5, 8, 9.

Choosing between PCA and t-SNE

- PCA: Quality measures are “equally bad”
- t-SNE: Pronounced differences in groups

- Here: Can compare e.g. rank correlations to see that t-SNE performs better
- In general: Different ranges

Conclusion



- Use persistent homology to analyse the behaviour of quality measures on embeddings.
- Clustering & automatic merge threshold selection.
- Judge agreement of quality measures for dimensionality reduction.