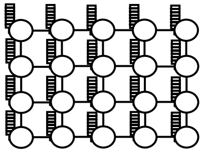
SOM Visualisierungen

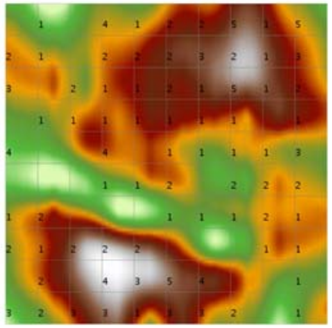
VU Selbst-Organisierende Systeme
Andreas Rauber
http://www.ifs.tuwien.ac.at/~andi

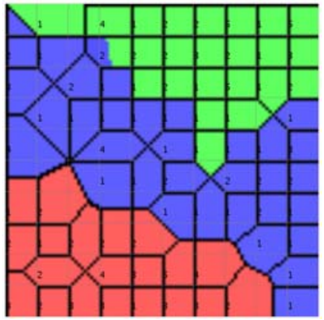
Outline

- SOM Basics

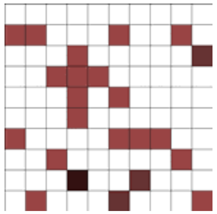


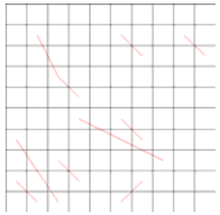
- Visualisierungen



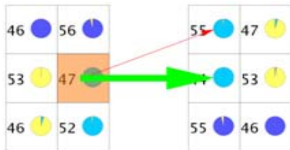


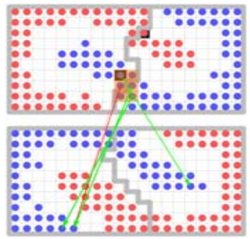
- SOM Qualitätsmaße



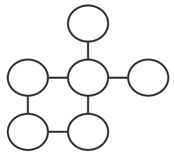


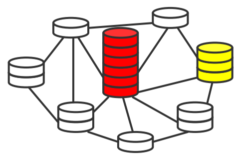
- SOM Comparison



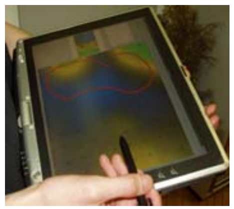


- Verwandte Verfahren & Architekturen





- Applikationen



SOM Visualisierungen

- SOM nur Basis für weitere Analyse und Verwendung
- Visualisierung: meist Karte als Basis
- verschiedene Arten der Einteilung:
  - Verwendete Information:
    - Verwendung nur des Codebooks
    - Verwendung von Codebook und Daten (mit/ohne Klasseninfo)
  - Art der Darstellung
    - Einfärbung / Hintergund
    - Überlagerbare Information
  - Art der Information:
    - Datenanalyse: Dichte, Topologie, Klassen, Distanzen
    - Qualitätsanalyse der Karte
- Wichtig ist Kombination von Visualisierungen zur Interpretation der Information, welche die SOM bietet

Visualizations

- Juha Vesanto. SOM-based Data Visualization Methods. Intelligent Data Analysis 3(2):111-126. Elsevier Science.1999

Outline

- Überblick Visualisierungsarten
- Visualisierung der SOM
  - Codebook-Projektion
  - Adaptive Coordinates
- Visualisierungen auf der SOM
  - Textuelle Informationen
  - Dichte
  - Distanzen
  - Klasseninfo
  - Attribute
  - Clustering der SOM

Visualisierungsarten

Überblick Visualisierungsmöglichkeiten

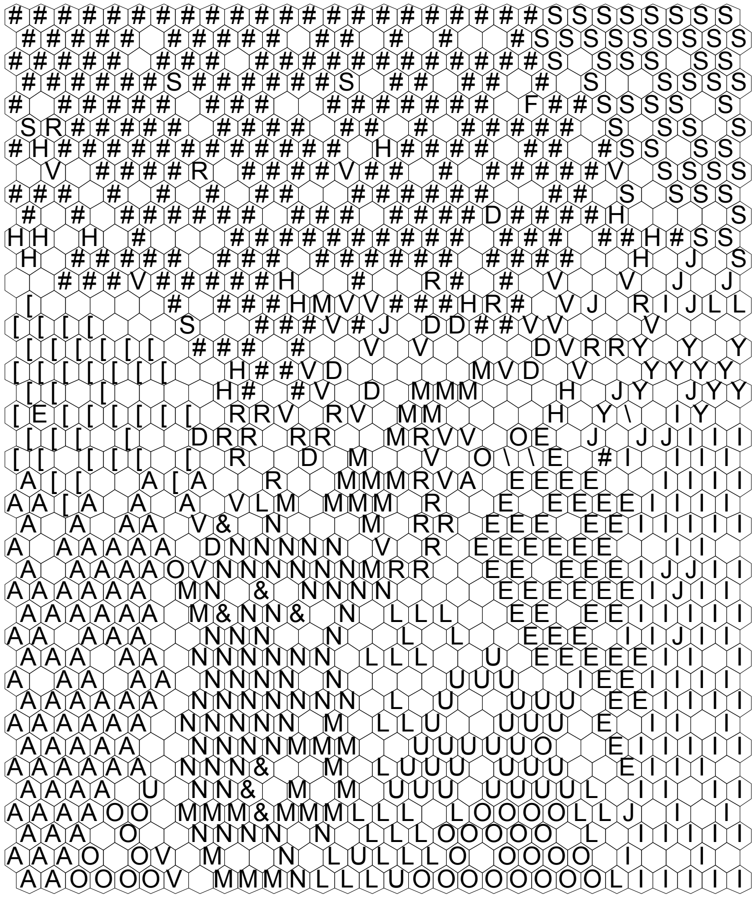
- Zahlreiche Möglichkeiten, Information auf den Knoten abzubilden
  - textuelle /numerische Information
  - (farbliche) Symbole, Metaphergrafiken
  - Einfärben der Zellen
  - Einfärben durch Interpolation über Knoten
  - Verbindungslinien, Graphen als Overlay
  - sonstige (z.B. 3D Welten)
- Dies kann zur Visualisierung unterschiedlicher Informationen genutzt werden
  - Datenvektoren
  - Klassen
  - Qualitätsmaße

Visualisierungsarten

- Textuelle Information auf den Knoten:
  - Anzahl der Vektoren
  - Namen der Vektoren
  - Klassenlabels
  - Qualitätsmaße
- als Overlay auf Einfärbungen kombinierbar

Visualisierungsarten

Labels (only most frequent label)



Klassen-IDs auf Units

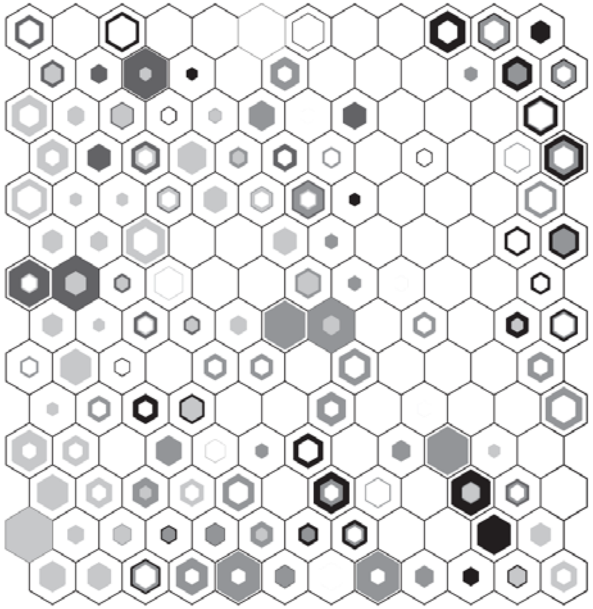
	1		4	1	2	2	5	1	5
2	1		2	2	2	3	2	1	3
3		2	1	1	2	1	5	1	2
	1	1	1	1	1	1	1		1
4			4		1	1	1	1	3
			1	1	2		2	2	2
1	2				1	1	1	2	1
2	1	2	2	2				1	1
	2		4	3	5	4			1
3	2	3	3	1	3	3	2		1

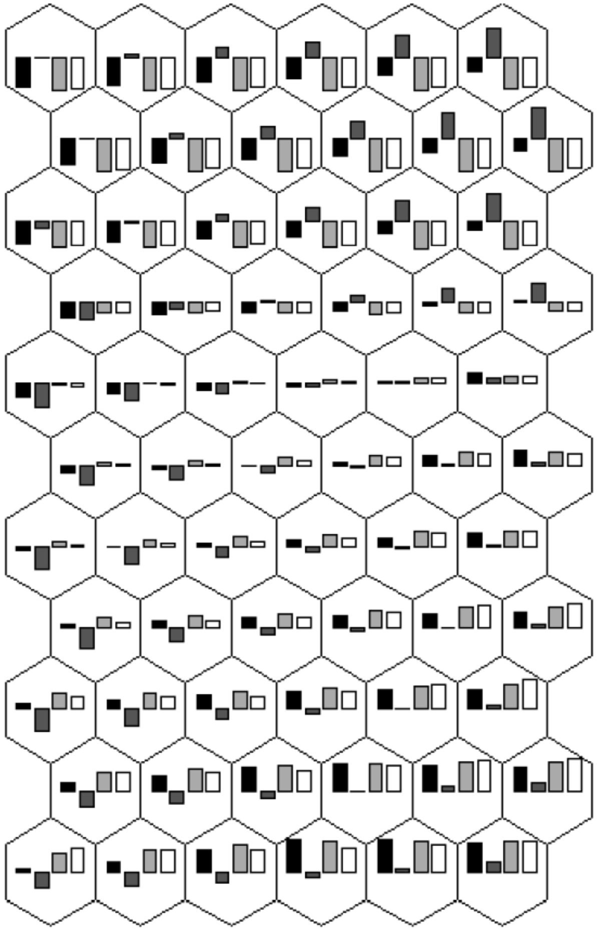
Anzahl Vektoren auf Units

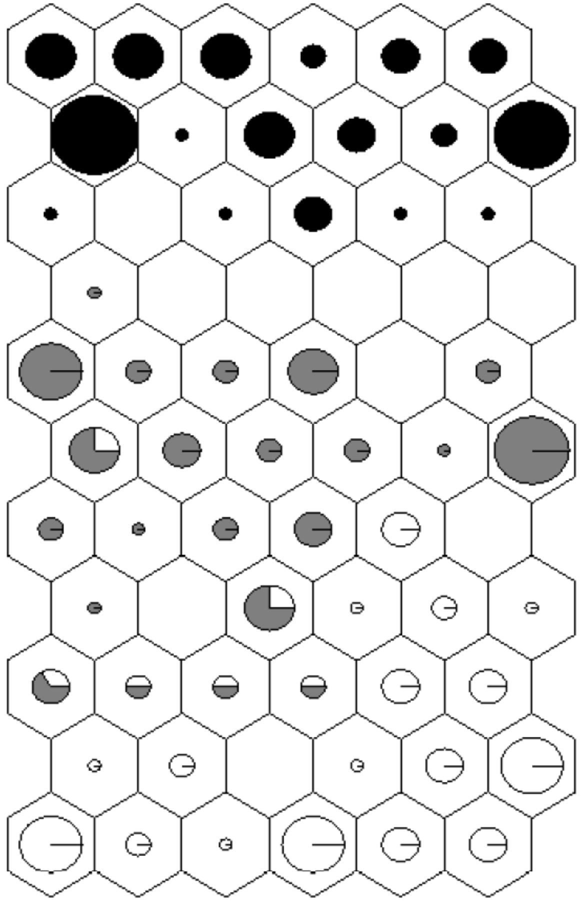
Visualisierungsarten

- Symbole auf Knoten
- Information aus den Gewichtsvektoren oder Datenvektoren wird farblich oder durch Symbole dargestellt
- Diagrame oder Symbole (z.B. Chernoff Faces)
- Beispiele:
  - Klassenverteilung
  - Attributverteilungen

Visualisierungsarten



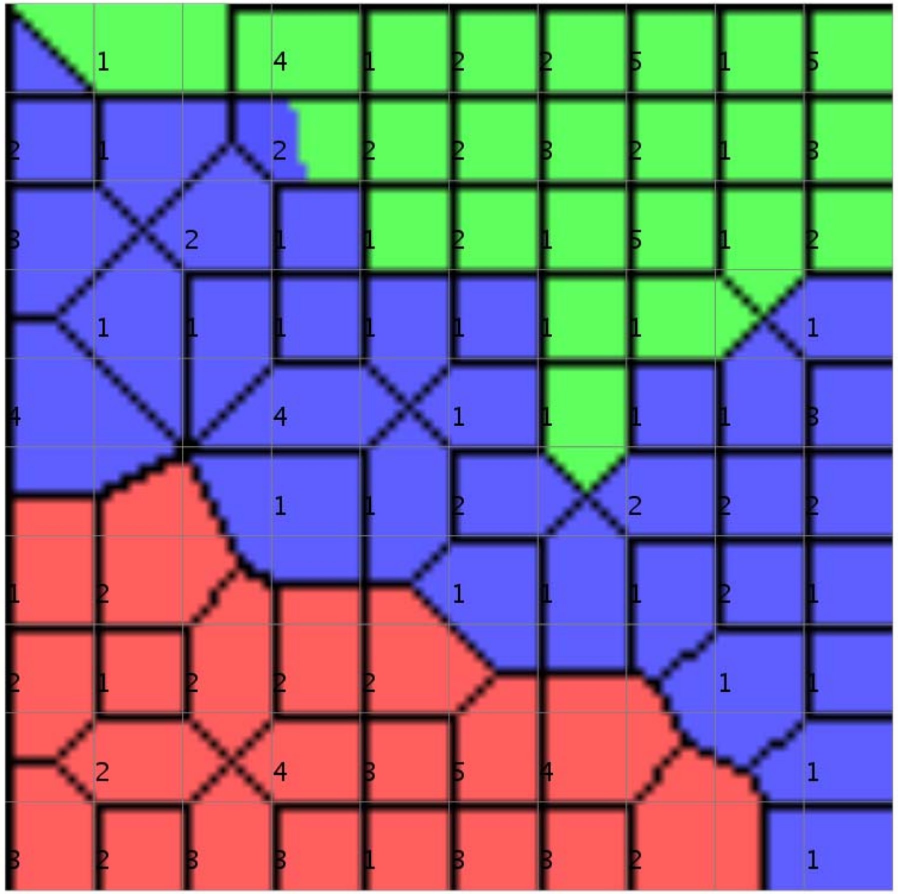


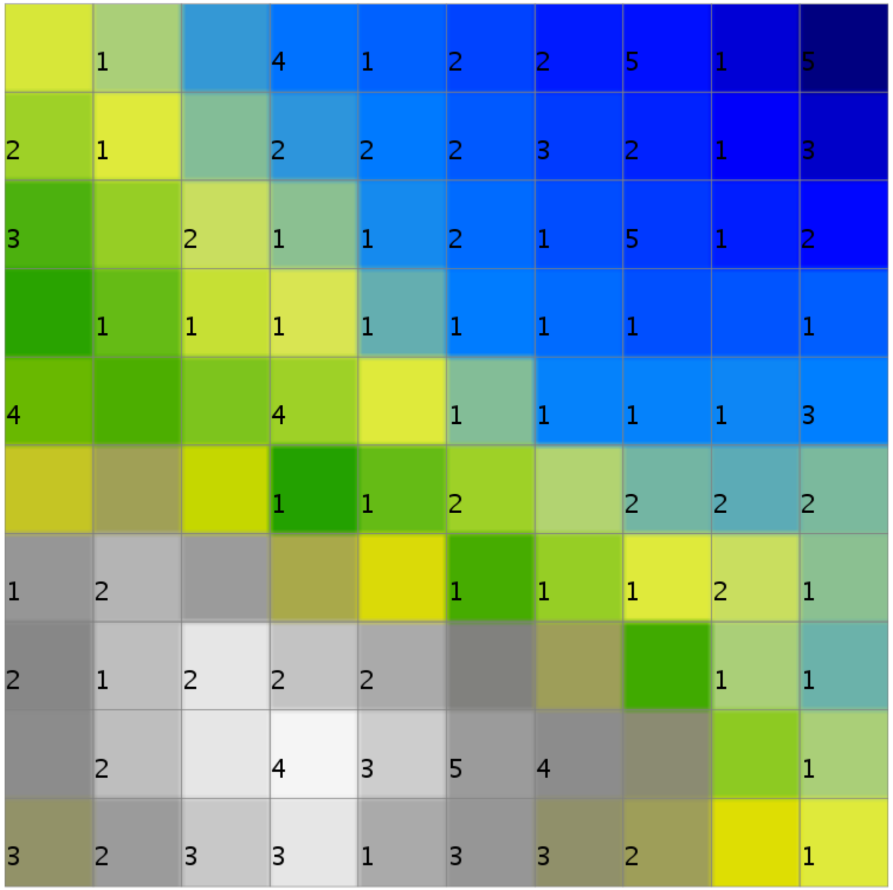


Visualisierungsarten

- Einfärben von Knoten
- Numerische Information als Basis zur Einfärbung
- Unterschiedliche Farbskalen haben entscheidende Auswirkung auf Interpretation
  - Farbwahl
  - Gradienten
  - Invertierung
- 2 Arten
  - Einfärben der Zellen
  - Numerische Werte als Stützpunkte, Interpolation
- “Hintergrundbild” für Text und Graphen

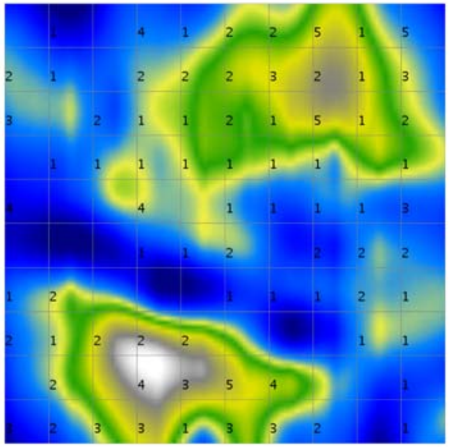
Visualisierungsarten

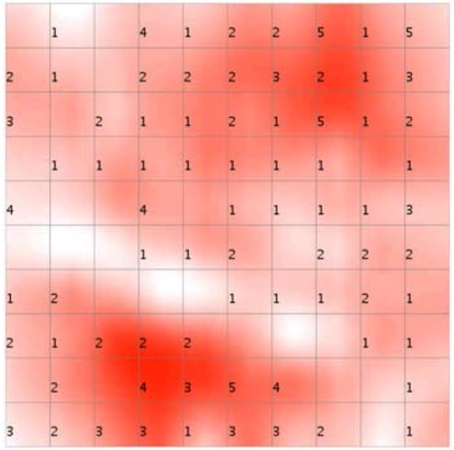


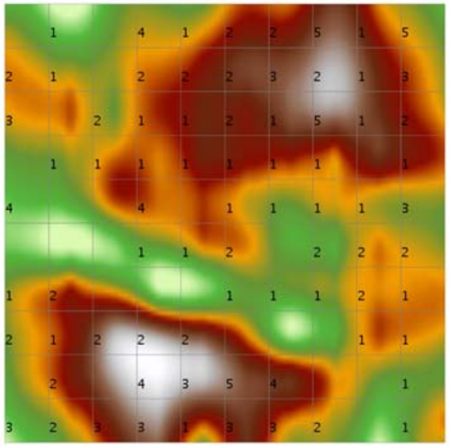


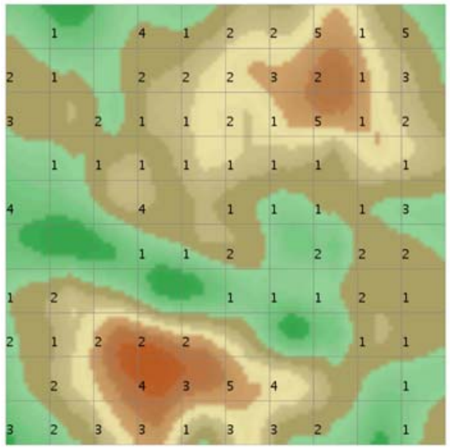
Visualisierungsarten

- Interpolation über Knoten: Farbskalen



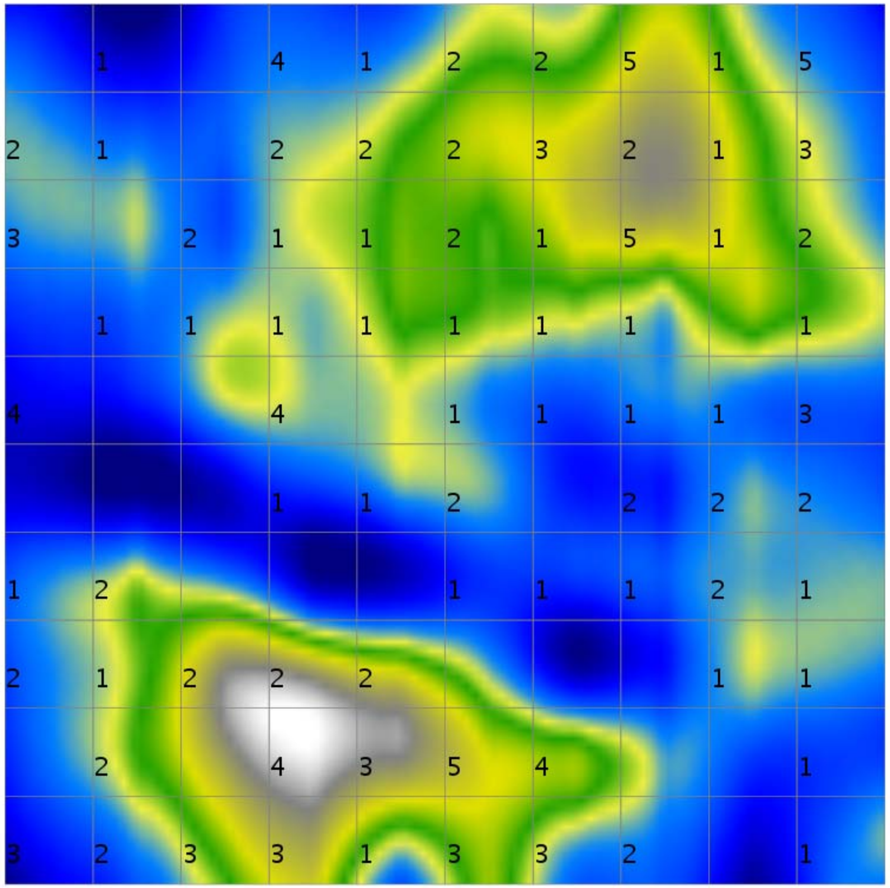


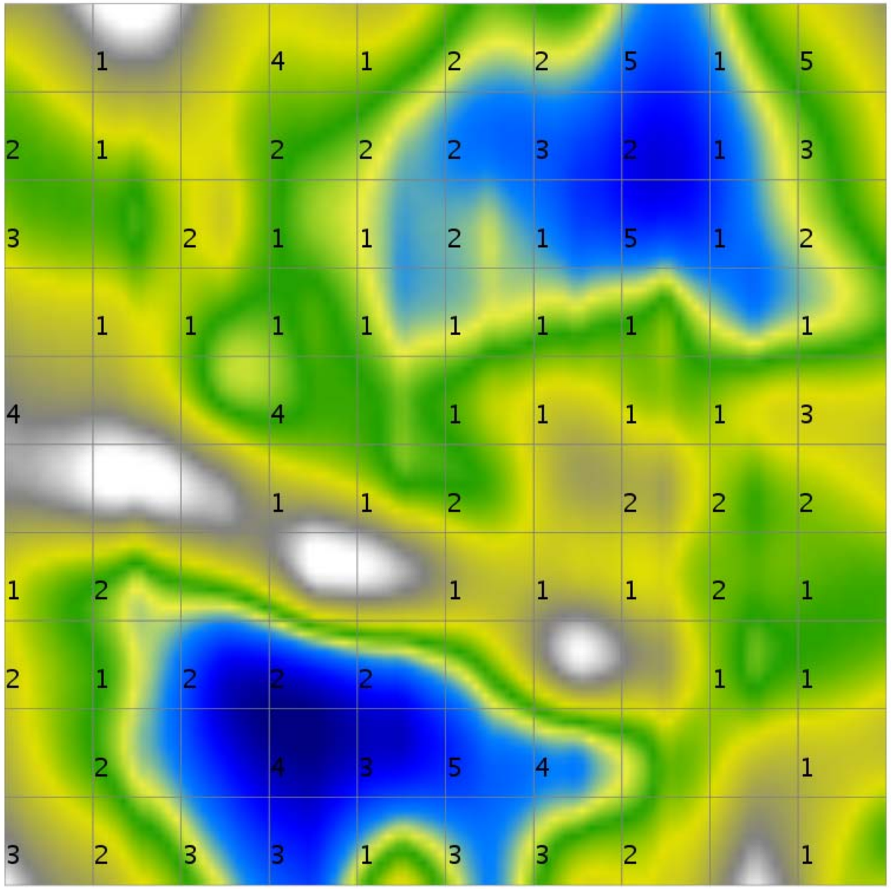




Visualisierungsarten

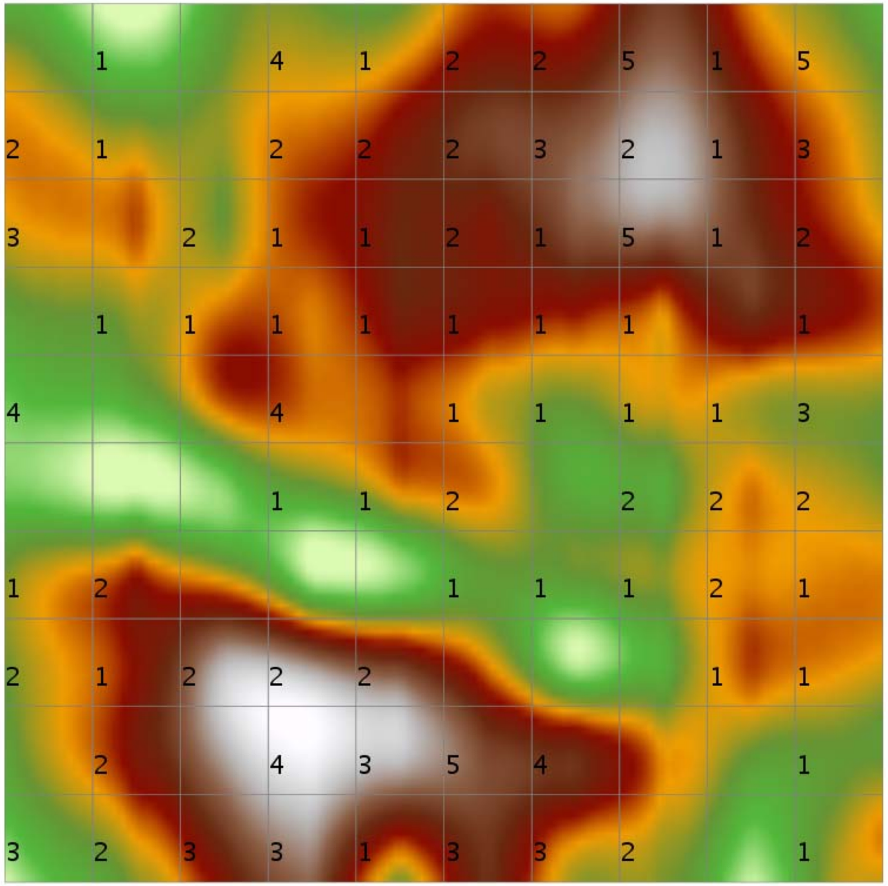
- Interpolation über Knoten: Invertierung

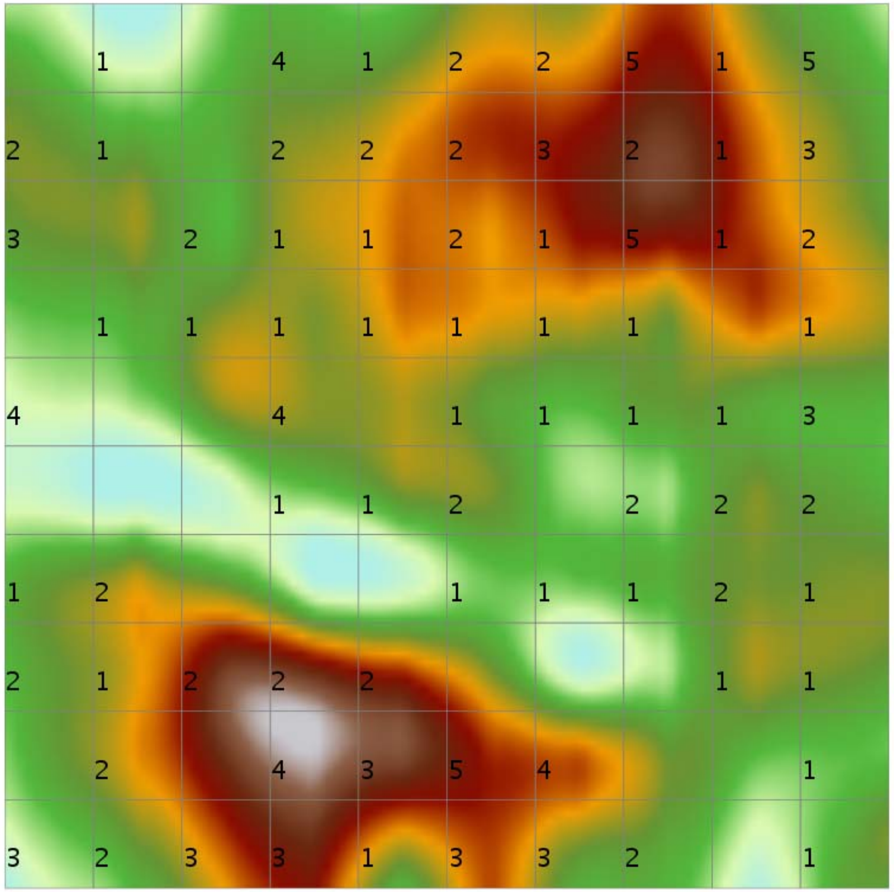




Visualisierungsarten

- Interpolation über Knoten: Farbskalen - Skalierung

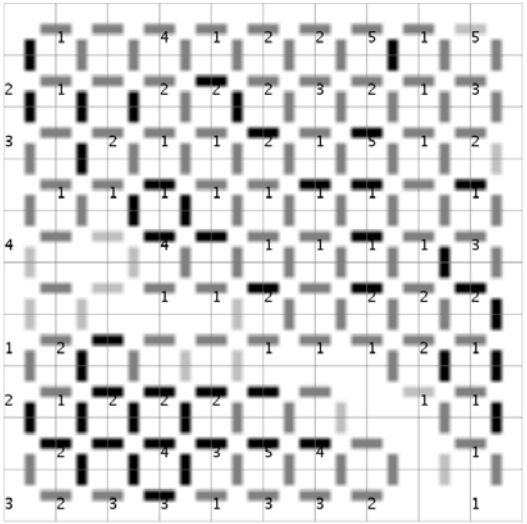
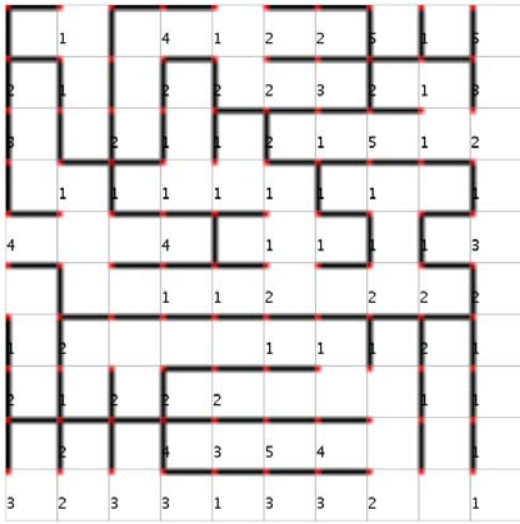
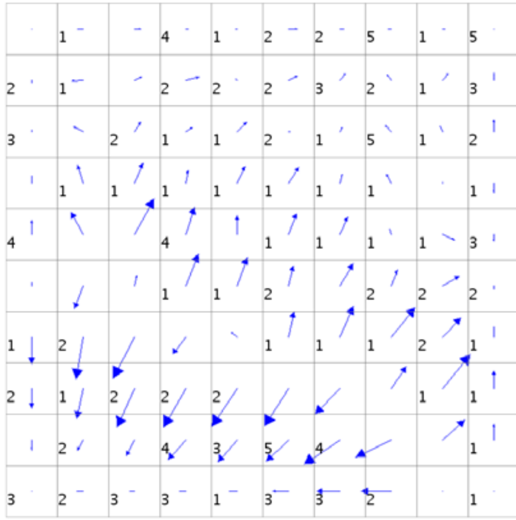
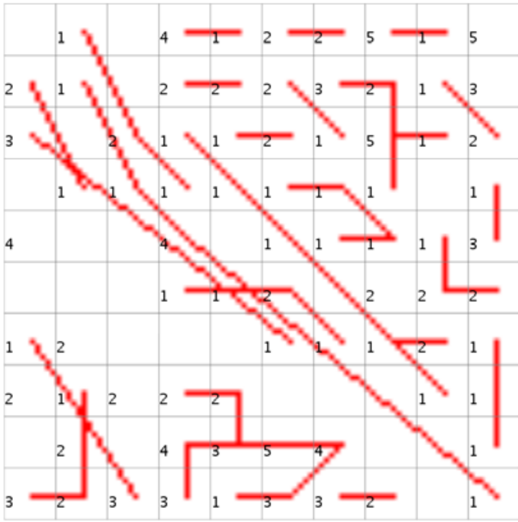
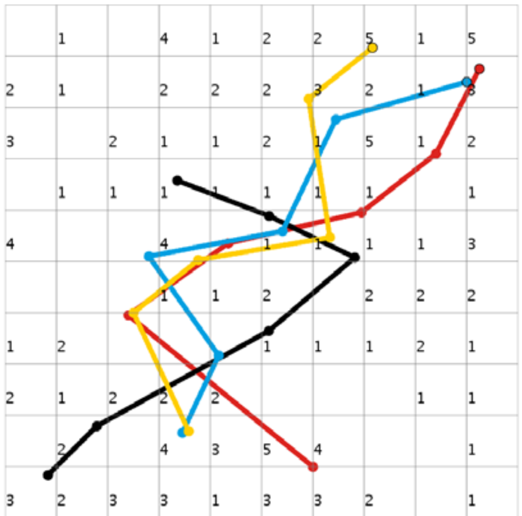




Visualisierungsarten

- Verbindungslinien / Graphen auf Karte
- Einzeichnen von Strukturen und Verbindungen
- Arten der Darstellung
  - Richtung der Linien
  - Farbe
  - Dicke
  - Form
- Als Overlay auf Einfärbungen kombinierbar

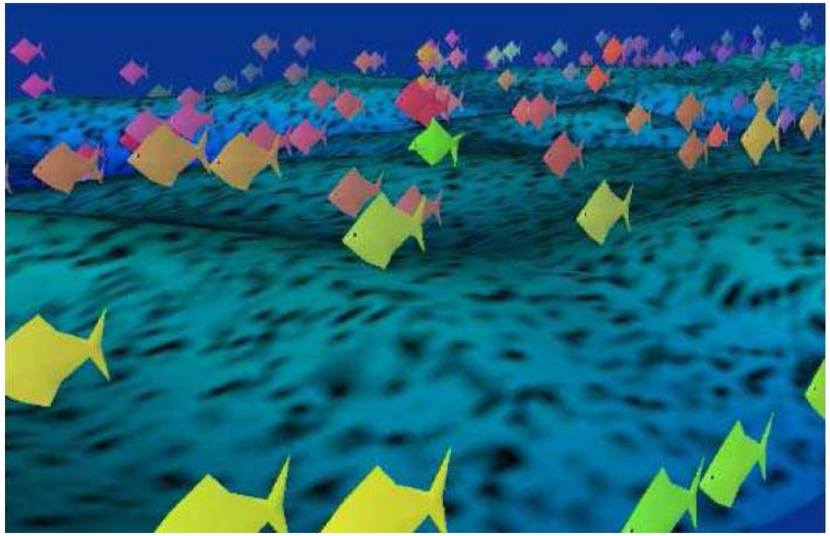
Visualisierungsarten

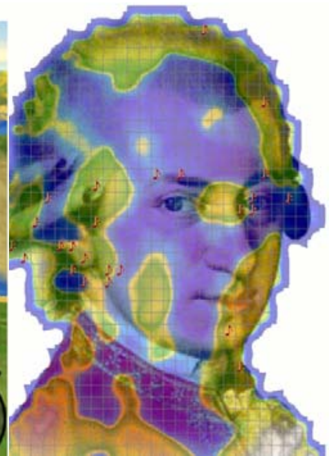






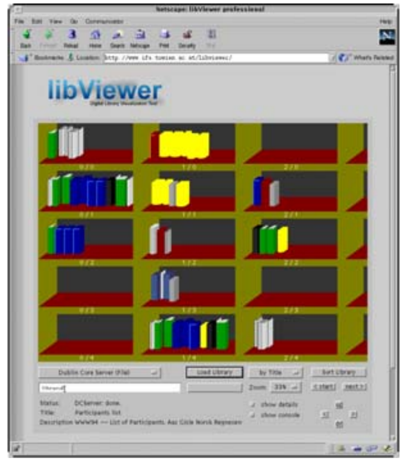
Visualisierungsarten

- Zahlreiche weitere Formen
- SOM als Organisationsbasis für Visualisierungen
- SOM schafft Datenraum, darauf andere Visualisierungen
  - 3-D Welten
  - Metapher Grafiken
  - Basis für Interpolationen
- Domänenspezifische Darstellungen

Visualisierungsarten







Visualisierungsarten

- Zusammenfassung
- Zahlreiche Visualisierungsmöglichkeiten
  - Projektion des Codebook
  - SOM als Basis
    - textuelle /numerische Information
    - (farbliche) Symbole, Metaphergrafiken
    - Einfärben der Zellen
    - Einfärben durch Interpolation über Knoten
    - Verbindungslinien, Graphen als Overlay
    - sonstige (z.B. 3D Welten)
- Basis für
  - Analyse der SOM
  - Verwendung der SOM
- Welche Informationen können abgebildet werden?

Outline

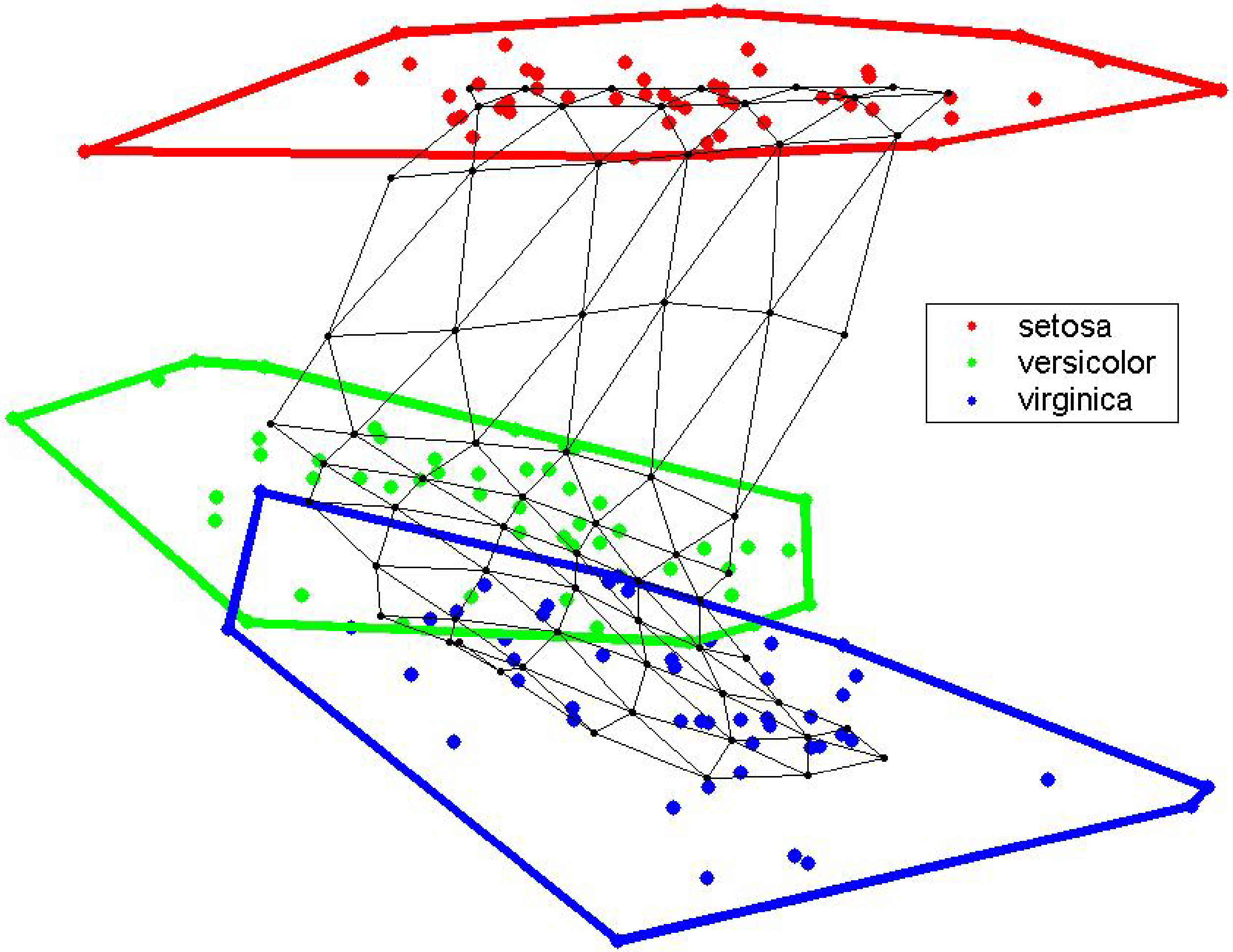
- Überblick Visualisierungsarten
- Visualisierung der SOM
  - Codebook-Projektion
  - Adaptive Coordinates
- Visualisierungen auf der SOM
  - Textuelle Informationen
  - Dichte
  - Distanzen
  - Klasseninformation
  - Attribute
  - Clustering der SOM

Codebook-Projektion

Projektion des SOM Codebooks

- Abbildung des Verhaltens der Daten und Gewichtsvektoren
- nicht Abbildung auf Karte!
- Codebook hat nur im 2-dim Inputraum eine 2D-Position
- Daten in 2D-Raum projizieren
  (z.B: PCA, Sammons Mapping)
- Verzerrte Visualisierung der Beziehungen zwischen Daten und Gewichtsvektoren
- Gemeinsame Projektion von Daten und Codebook

Codebook-Projektion

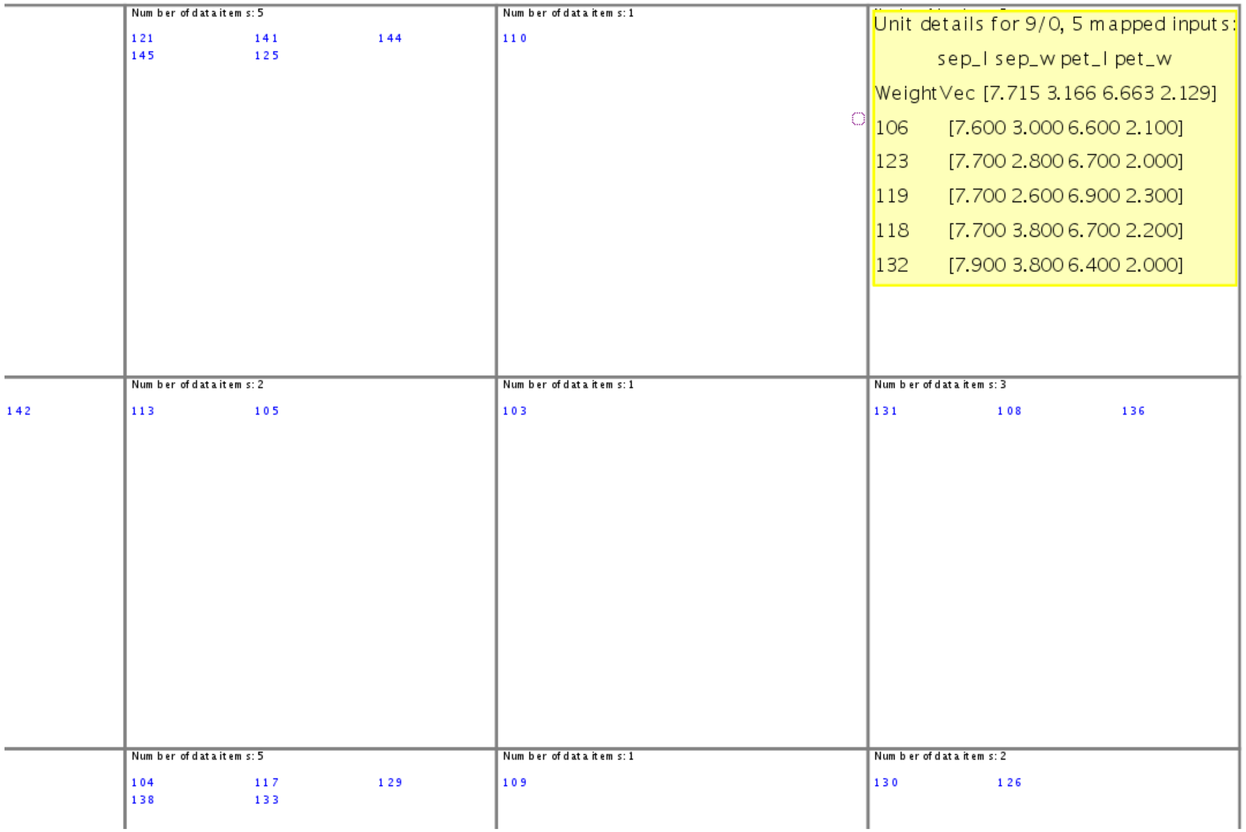


Outline

- Überblick Visualisierungsarten
- Visualisierung der SOM
  - Codebook-Projektion
  - Adaptive Coordinates
- Visualisierungen auf der SOM
  - Textuelle Informationen
  - Dichte
  - Distanzen
  - Klasseninformation
  - Attribute
  - Clustering der SOM

Text-Informationen auf der SOM

- Vektor-Infos
  - Vector IDs, Abbildungsabstände, Fehlermaße, Klassen,…



Outline

- Überblick Visualisierungsarten
- Visualisierung der SOM
  - Codebook-Projektion
  - Adaptive Coordinates
- Visualisierungen auf der SOM
  - Textuelle Informationen
  - Dichte
  - Distanzen
  - Klasseninformation
  - Attribute
  - Clustering der SOM

Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
  - Hit Histogramm
  - Activity Histogram
  - Smoothed Data Histograms
  - P-Matrix
  - Sky Metaphor
  - Neighborhood Graphs
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM

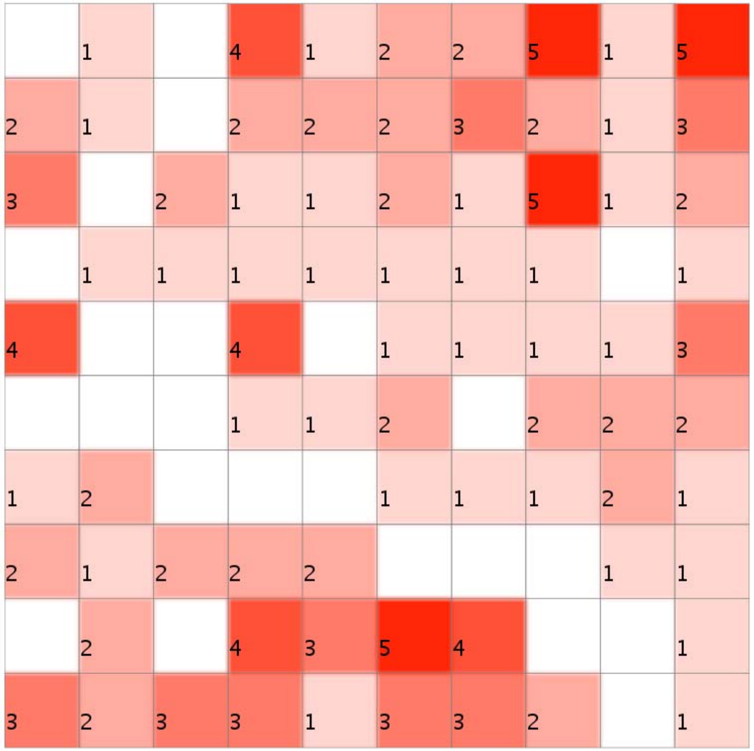
Dichteverteilungen auf der Karte

- Dichte: Verteilung der Datenpunkte auf der Karte
- jedes Sample wird seiner Position auf der Karte zugewiesen
- leere Felder = „interpolierende Units“, d.h. Übergangsregionen zwischen dicht besiedelten Gebieten
- Magnification Factor
- verschiedene Arten
  - Textuell: Anzahl der Vektoren
  - Hit Histogram: Farbe, patches
  - Smoothed Data Histograms
  - P-Matrix

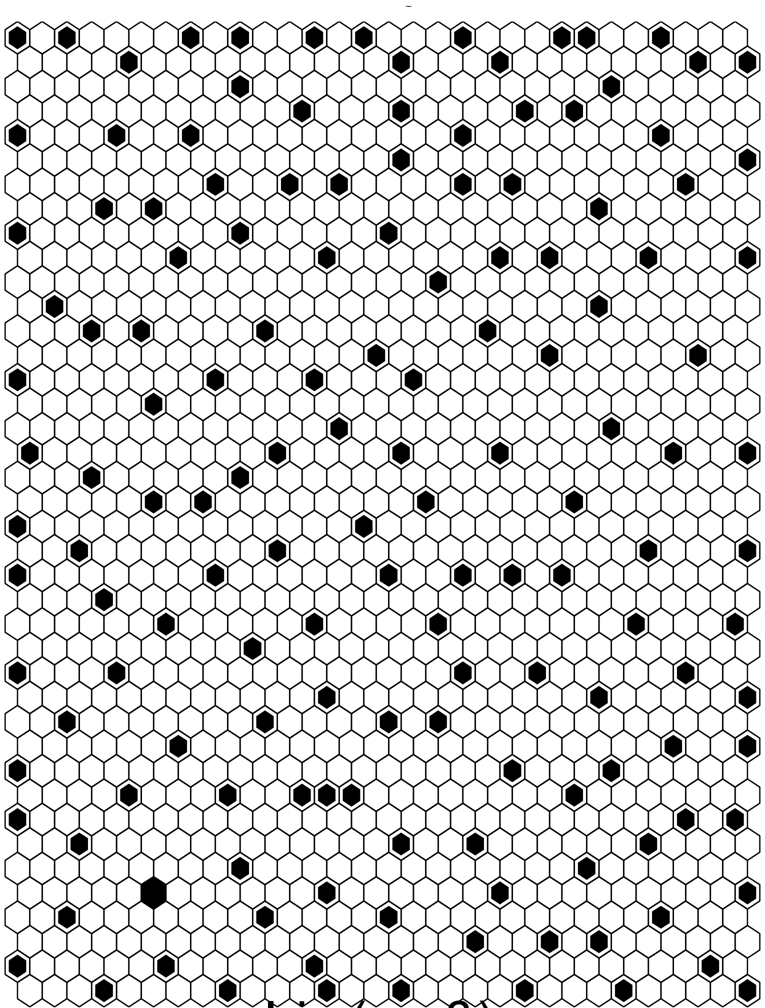
Dichte – Hit Histogramm

- Vektor-Infos
  - Anzahl der Vektoren pro Unit: Hit histogram

	1		4	1	2	2	5	1	5
2	1		2	2	2	3	2	1	3
3		2	1	1	2	1	5	1	2
	1	1	1	1	1	1	1		1
4			4		1	1	1	1	3
			1	1	2		2	2	2
1	2				1	1	1	2	1
2	1	2	2	2				1	1
	2		4	3	5	4			1
3	2	3	3	1	3	3	2		1



Dichte – Hit Histogramm



Iris (groß)

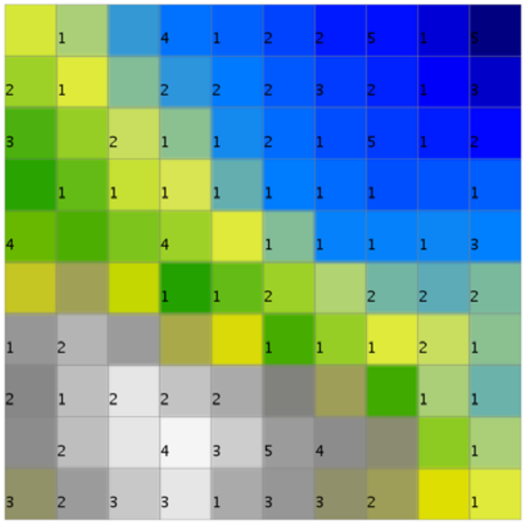
Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
  - Hit Histogramm
  - Activity Histogram
  - Smoothed Data Histograms
  - P-Matrix
  - Sky Metaphor
  - Neighborhood Graphs
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM

Activity Histogram

- Datenpunkt wird auf best-matching unit abgebildet
- Wie passend wäre die second-best unit gewesen?
- Activity Histogram pro Datenpunkt
- Visualisierung der Distanz für einzelne Datenpunkte zu allen Gewichtsvektoren



Outline

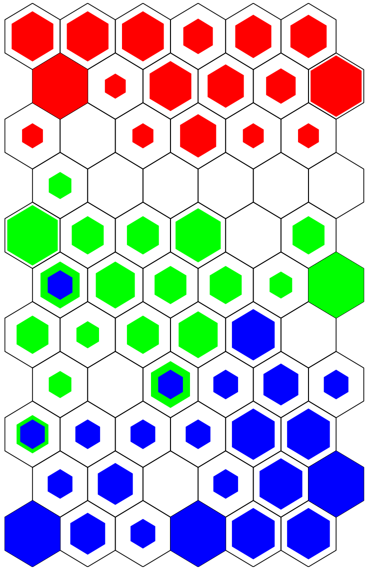
Visualisierungen der SOM

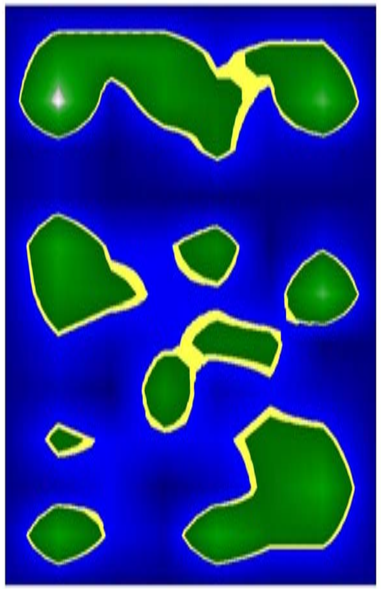
- Textuelle Informationen
- Dichte
  - Hit Histogramm
  - Activity Histogram
  - Smoothed Data Histograms
  - P-Matrix
  - Sky Metaphor
  - Neighborhood Graphs
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM

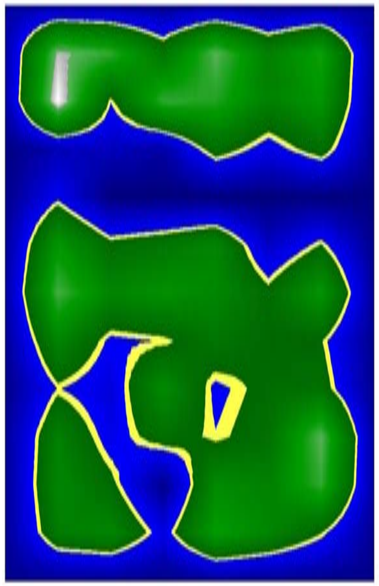
Dichte: Smoothed Data Histogram

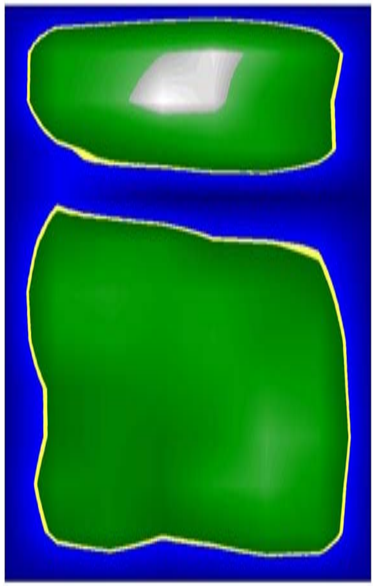
- Erweiterung von Hit Histograms
- jeder Datenpunkt wird nicht nur zu einer Map Unit gerechnet, sondern auf n weitere units abgebildet
- 3 verschiedene Varianten
  - auf allen n units voll gezählt
  - auf 1. unit 1, dann 1/2, 1/3, ...
  - Gewichtung abhängig zur Abbildungsdistanz
- erzeugt Verschmierungseffekt
- Parameter n: Anzahl der n-best matching units: Granularität
- E. Pampalk, A. Rauber, D. Merkl: **Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps**. In: Proceedings of the Intl.Conf on Artificial Neural Networks (ICANN 2002), pp.871-876.

Dichte: Smoothed Data Histogram





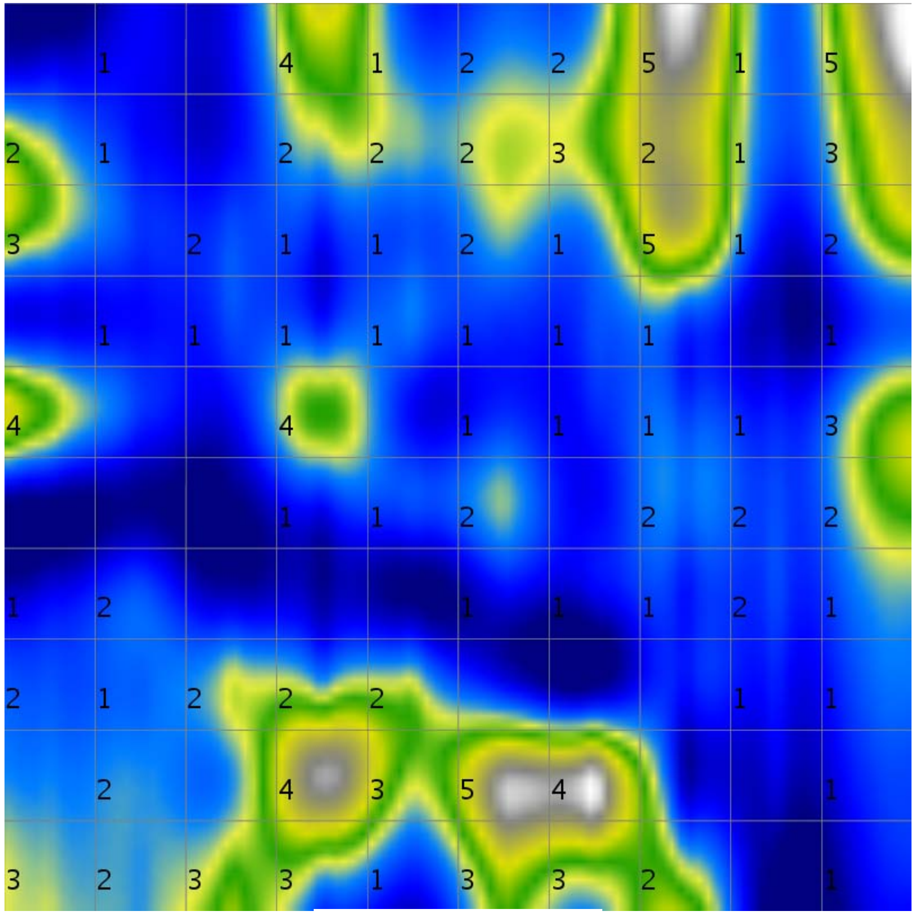




Hit-histogram n = 1 n = 3 n = 10

Dichte: Smoothed Data Histogram

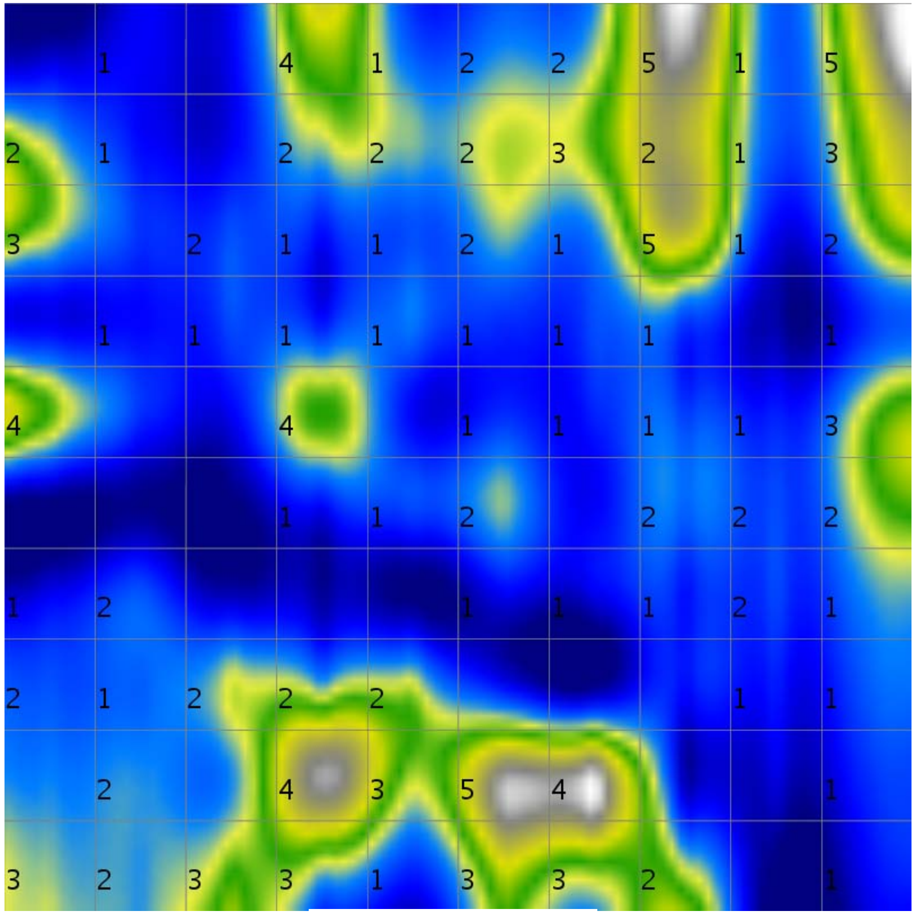
- Iris Dataset



n = 2

Dichte: Smoothed Data Histogram

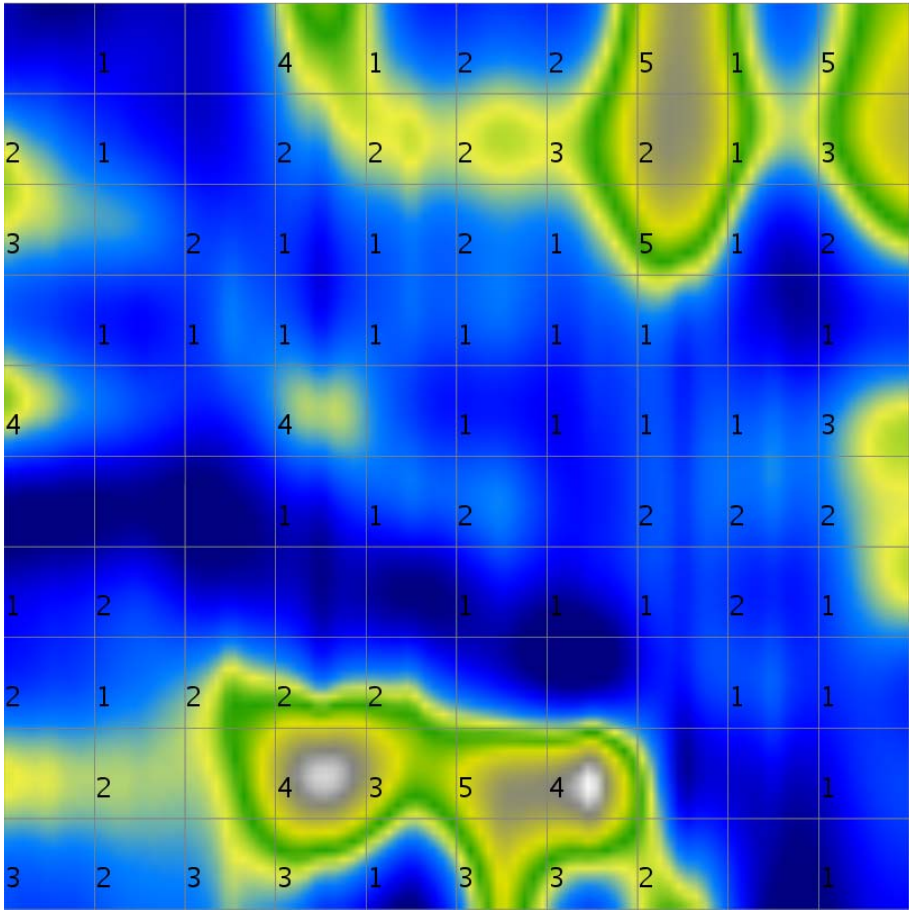
- Iris Dataset



n = 2

Dichte: Smoothed Data Histogram

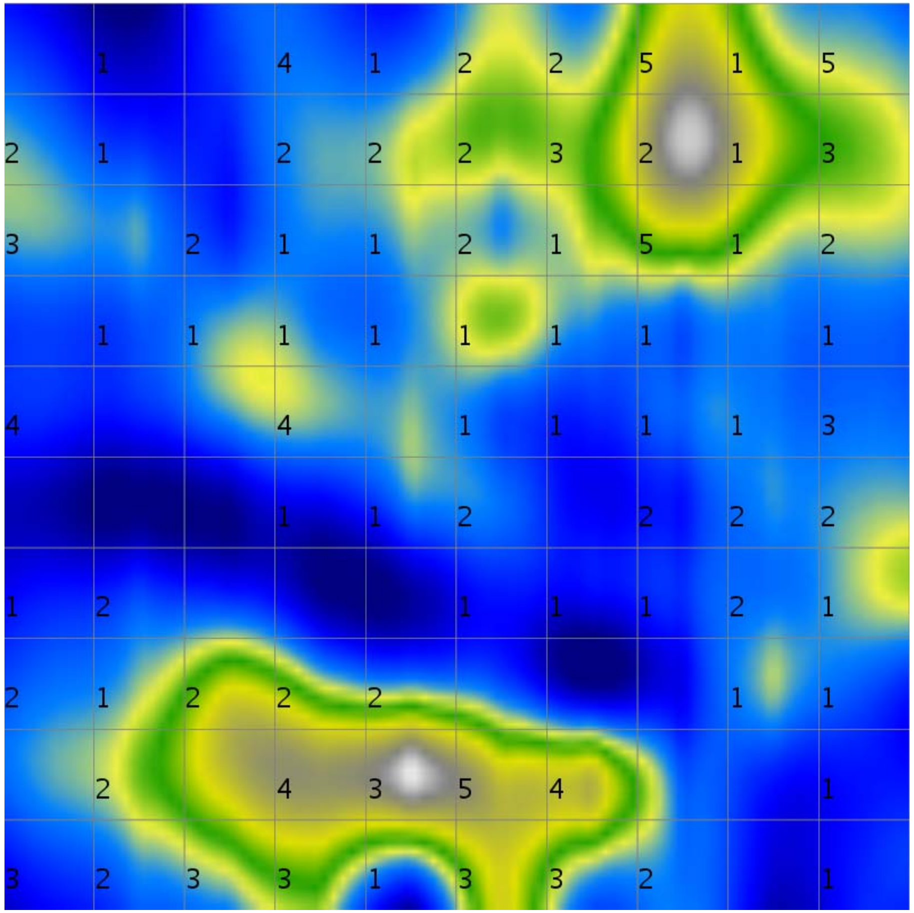
- Iris Dataset



n = 3

Dichte: Smoothed Data Histogram

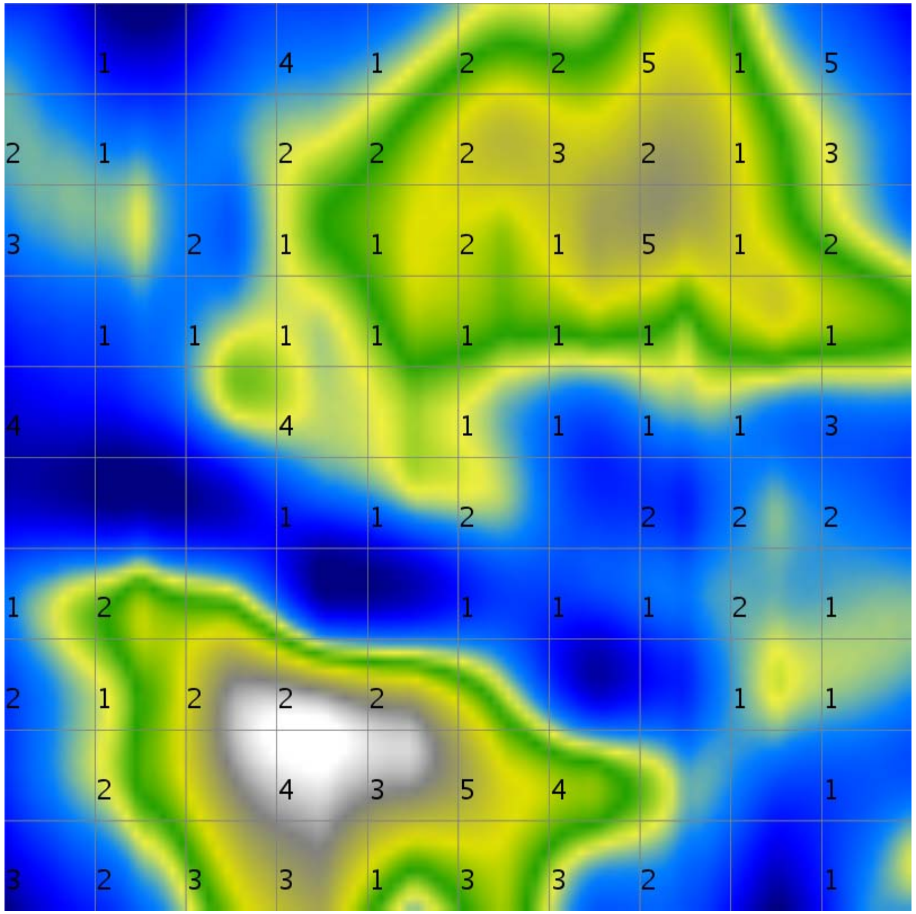
- Iris Dataset



n = 8

Dichte: Smoothed Data Histogram

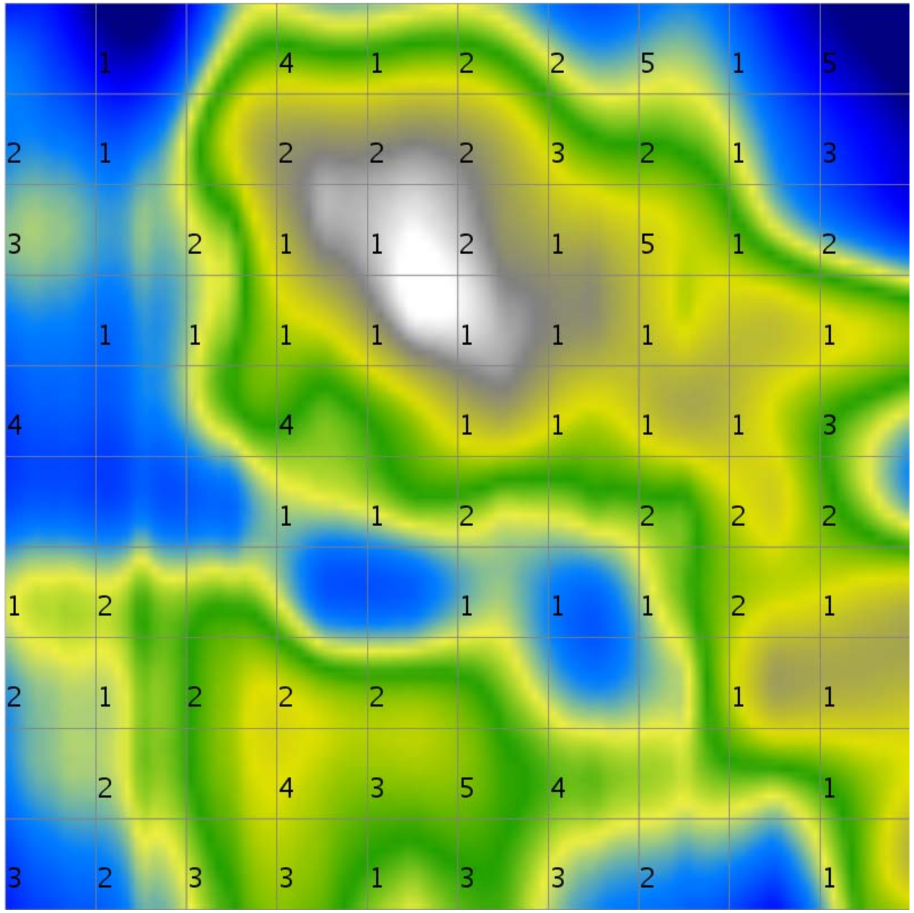
- Iris Dataset



n = 20

Dichte: Smoothed Data Histogram

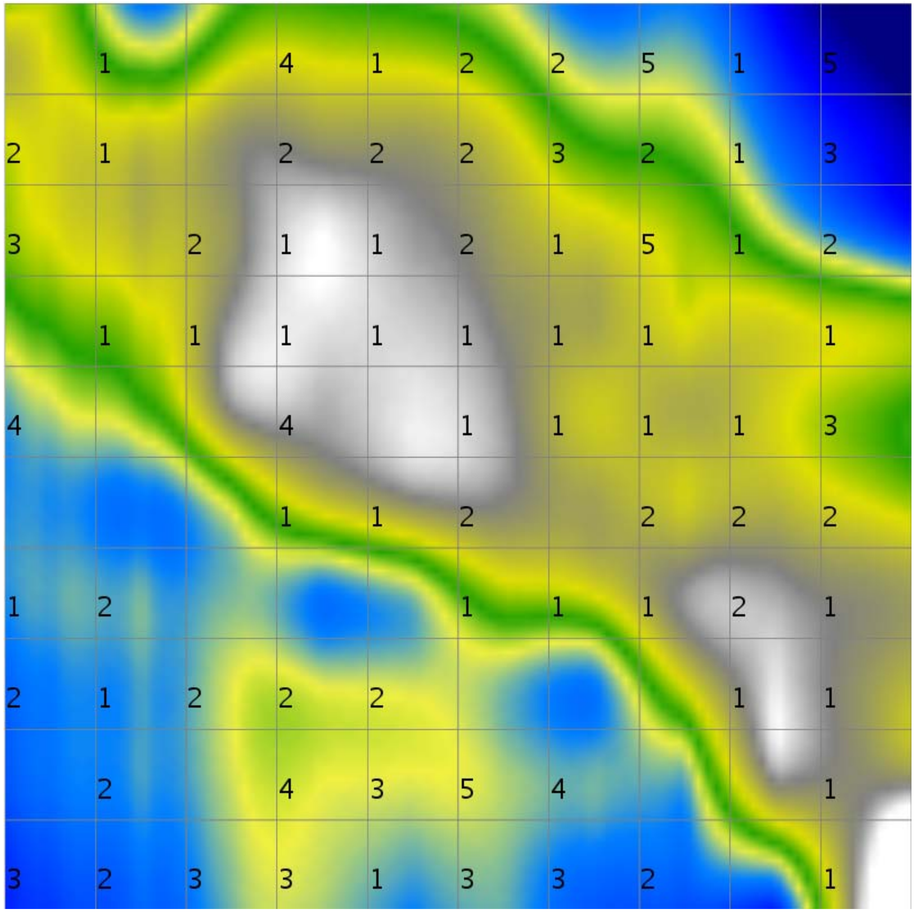
- Iris Dataset



n = 50

Dichte: Smoothed Data Histogram

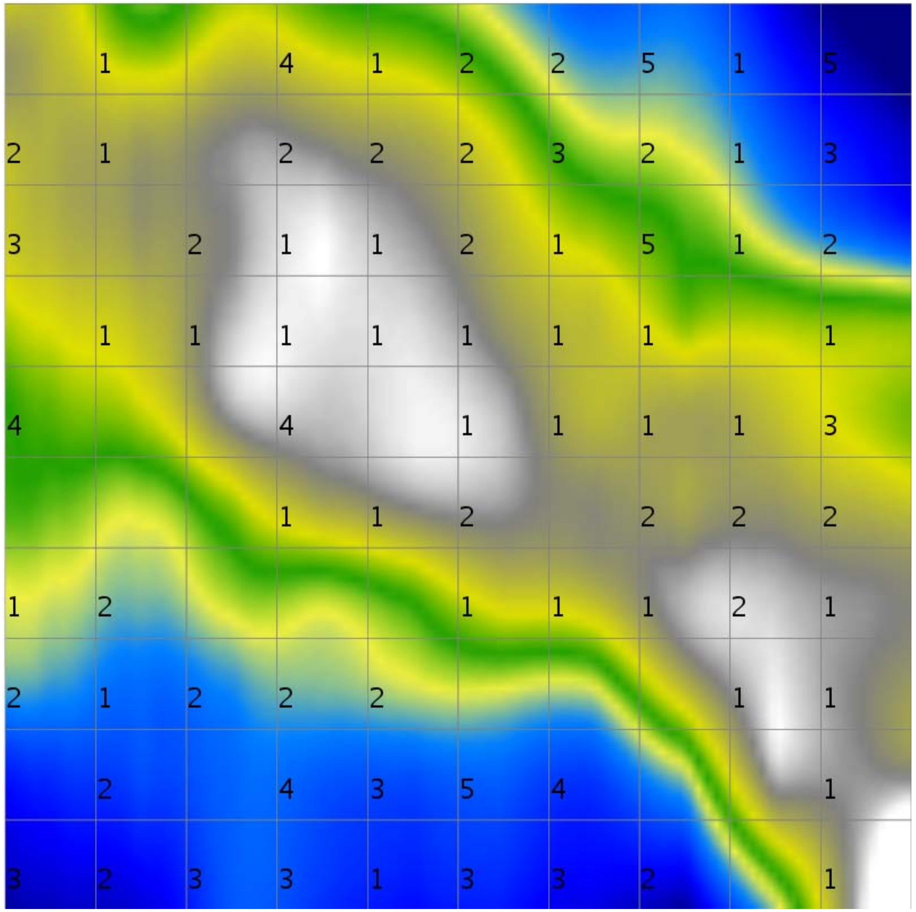
- Iris Dataset



n = 70

Dichte: Smoothed Data Histogram

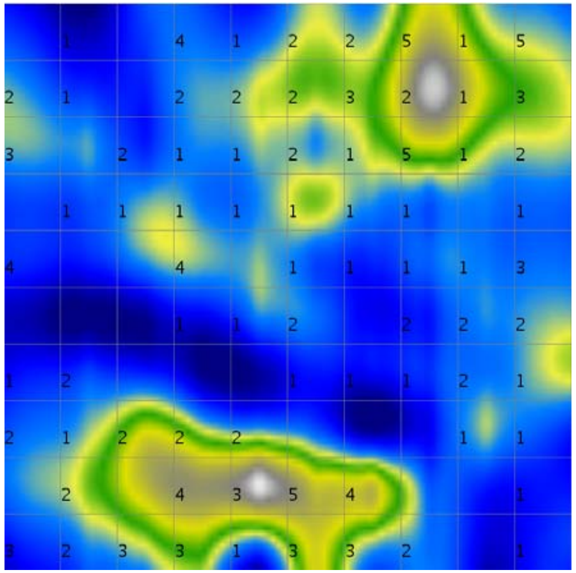
- Iris Dataset



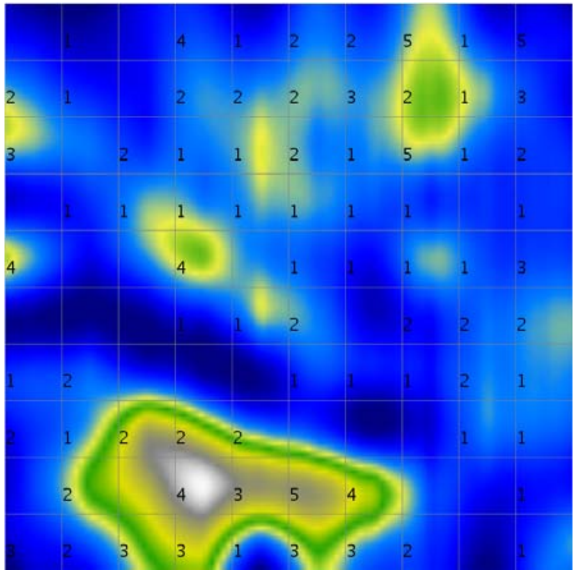
n = 90

Dichte: Smoothed Data Histogram

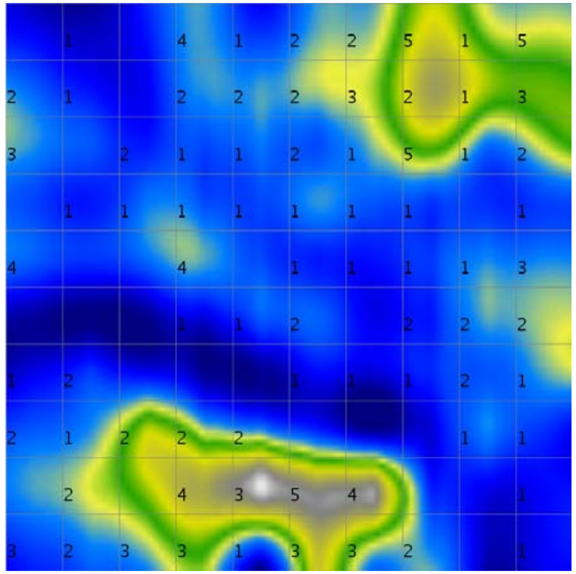
- Iris Dataset, SDN, n=8



SDH



SDH-weighted



SDH-weighted, normalized

Dichte: Smoothed Data Histogram

Fragen

- was passiert bei großen Smoothing-Faktoren? Warum?
- Unterschied zwischen normalen SDH, gewichteten SDH – wann werden die Visualisierungen unterschiedlich sein? Wo?
- Welche Aussagen können getroffen werden, wenn die Visualisierungen unterschiedlich sind?

Outline

Visualisierungen der SOM

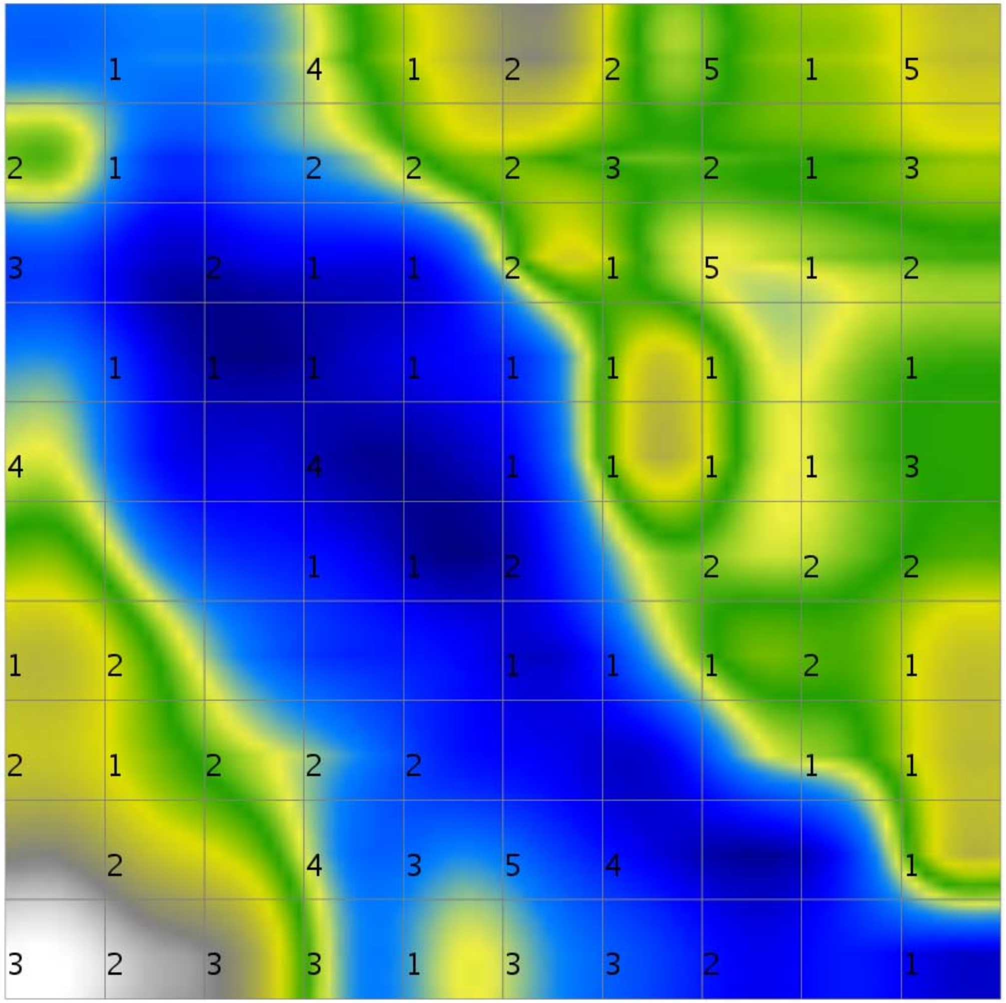
- Textuelle Informationen
- Dichte
  - Hit Histogramm
  - Activity Histogram
  - Smoothed Data Histograms
  - P-Matrix
  - Sky Metaphor
  - Neighborhood Graphs
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM

Dichte: P-Matrix

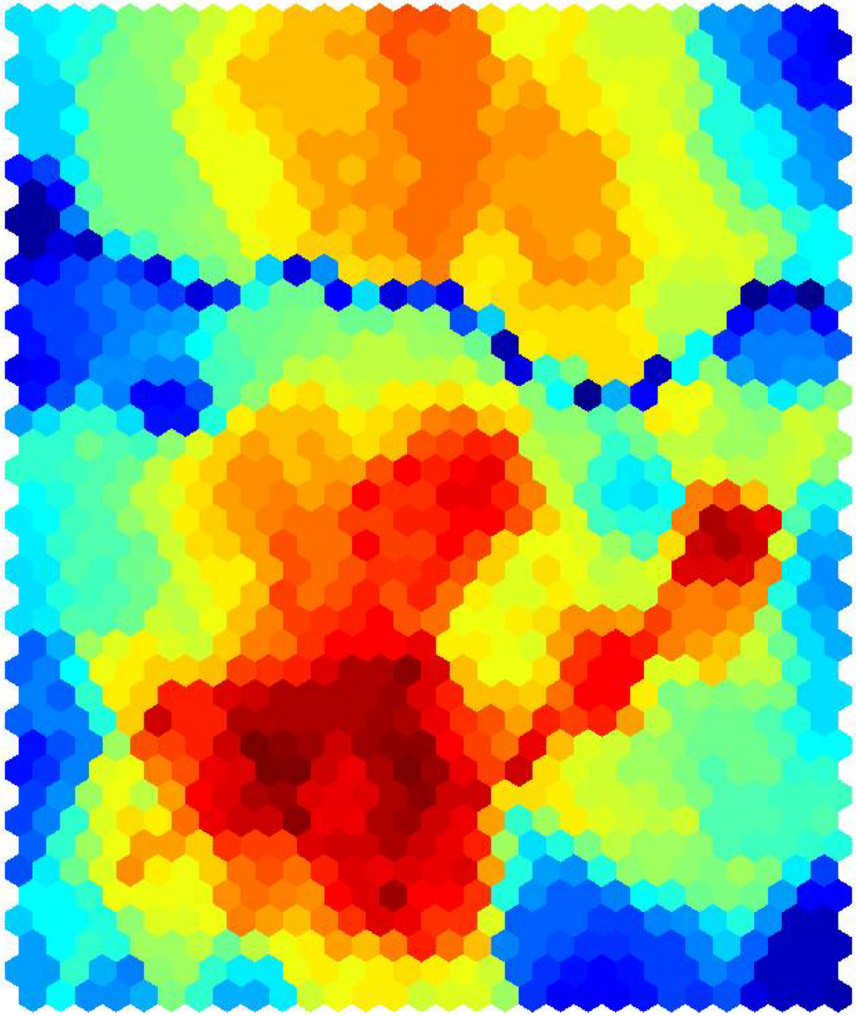
- Pareto-Matrix
- Hypersphere um jeden Gewichtsvektor im Datenraum
- Ermittlung der Anzahl der Datenpunkt in der Hypersphere ergibt Schätzung der Dichte um diese Unit
- gut geeignet für Karten mit großer Anzahl von units im Verhältnis zu Datenpunkten (Emergent SOMs)
- Ultsch, A.: Maps for the Visualization of High-Dimensional Spaces. In: Proceedings of the 2003 Workshop on Self-Organizing Maps (WSOM), Kyushu, Japan, 2003. pp.91-100.

Dichte: P-Matrix

- Iris-Datenset, P-Matrix



Visualisierung der Dichte



Iris (groß)
P-Matrix

Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
  - Hit Histogramm
  - Activity Histogram
  - Smoothed Data Histograms
  - P-Matrix
  - Sky Metaphor
  - Neighborhood Graphs
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM

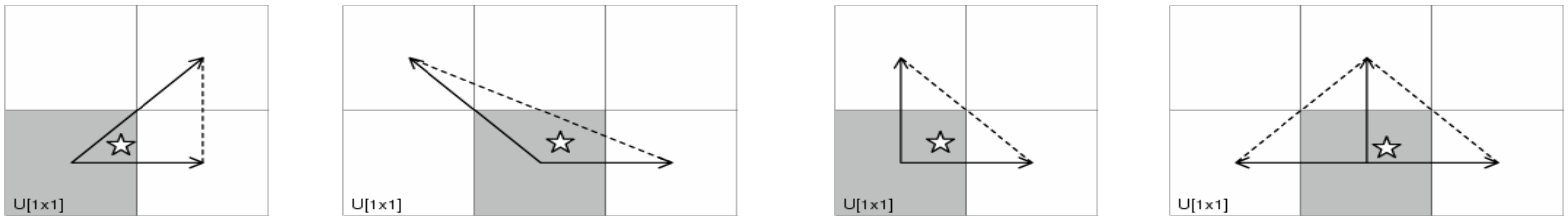
Dichte: Sky Metaphor

- Conventionally, data items mapped “on unit”
- Sky: display them on their “exact” position within the unit, not just in the centre
- Visualise similarity of an input with other inputs
  - within the same
  - across the neighbouring units
- Triangulation
- Khalid Latif and Rudolf Mayer. **Sky-Metaphor Visualisation for Self-Organising Maps.** In Proceedings of the 7th International Conference on Knowledge Management (I-KNOW'07), Graz, Austria, September 5 - 7 2007.

Dichte: Sky Metaphor

- Apply pull force to determine exact position
  - Relative to the distance of the input to BMU
  - Inverse proportional to the distance of the input to the other units

$$\mathbf{F}_i \quad \propto \quad \frac{d(x, U_1)}{d(x, U_i)} \quad \text{for} \quad i > 1$$



Dichte: Sky Metaphor

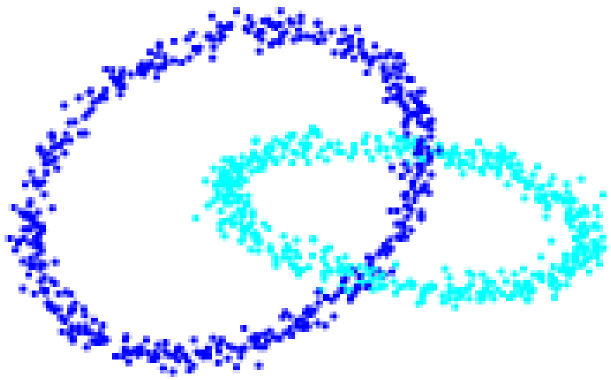
- X and y coordinates of the exact position p of input x on unit U_1calculated as

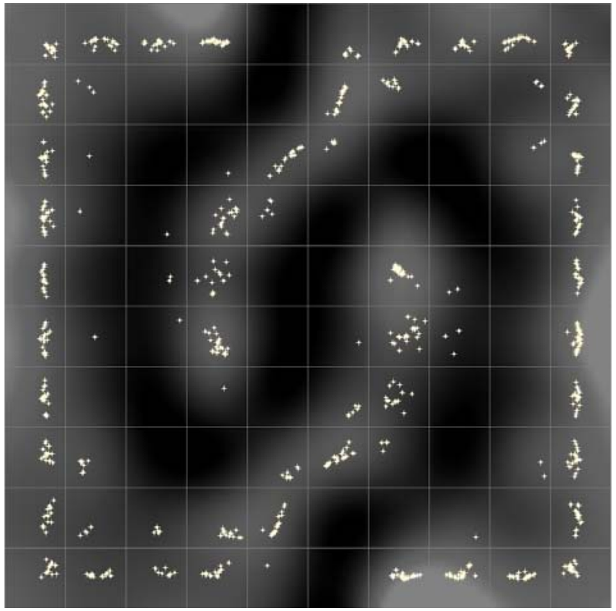
$$\mathbf{p}_{<x,y>} = \langle \lambda * \sum_{i=2}^{k} \mathbf{F}_i * \frac{1}{U_{i<x>} - U_{1<x>}}, \\ \lambda * \sum_{i=2}^{k} \mathbf{F}_i * \frac{1}{U_{i<y>} - U_{1<y>}} \rangle$$

- k: index over the 2 or 3 nearest units U_2 ..U_4
- λ: grid-constant to reconcile the displacement according to the display co-ordinates

Dichte: Sky Metaphor

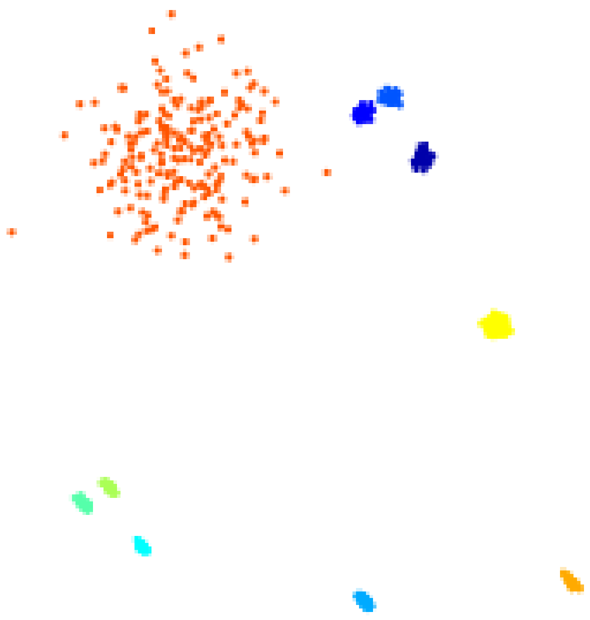
- Chain-link data set

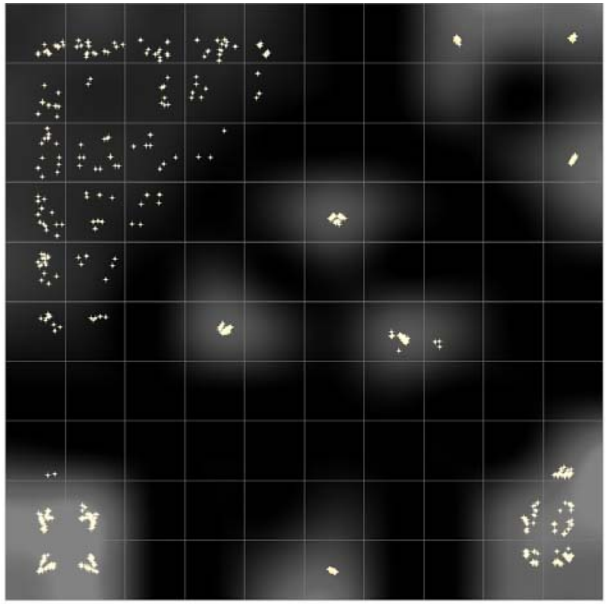




Dichte: Sky Metaphor

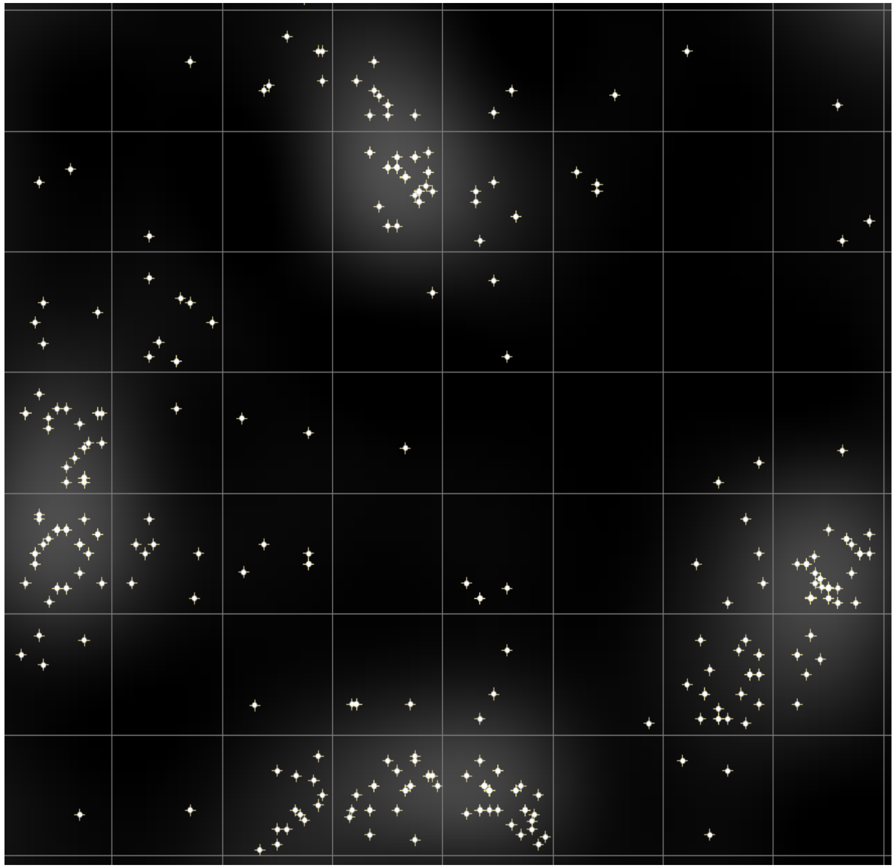
- 10-dimensional Gaussian distributions





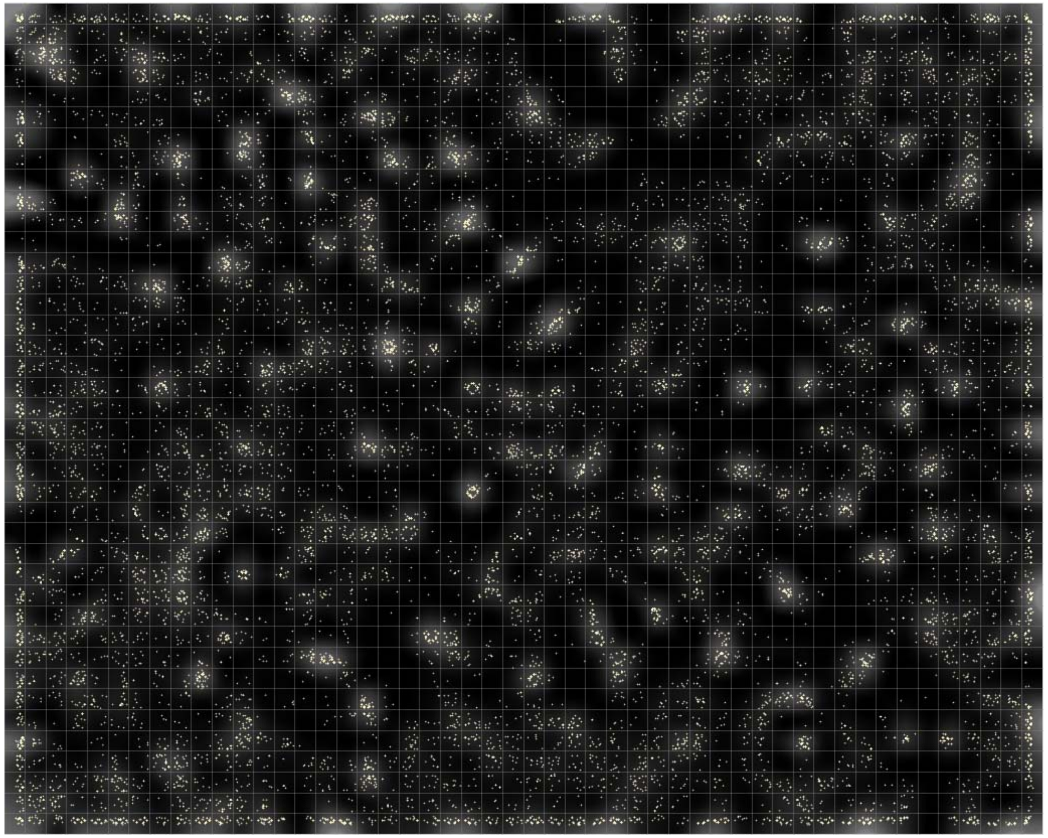
Dichte: Sky Metaphor

- 20 newsgroups data set



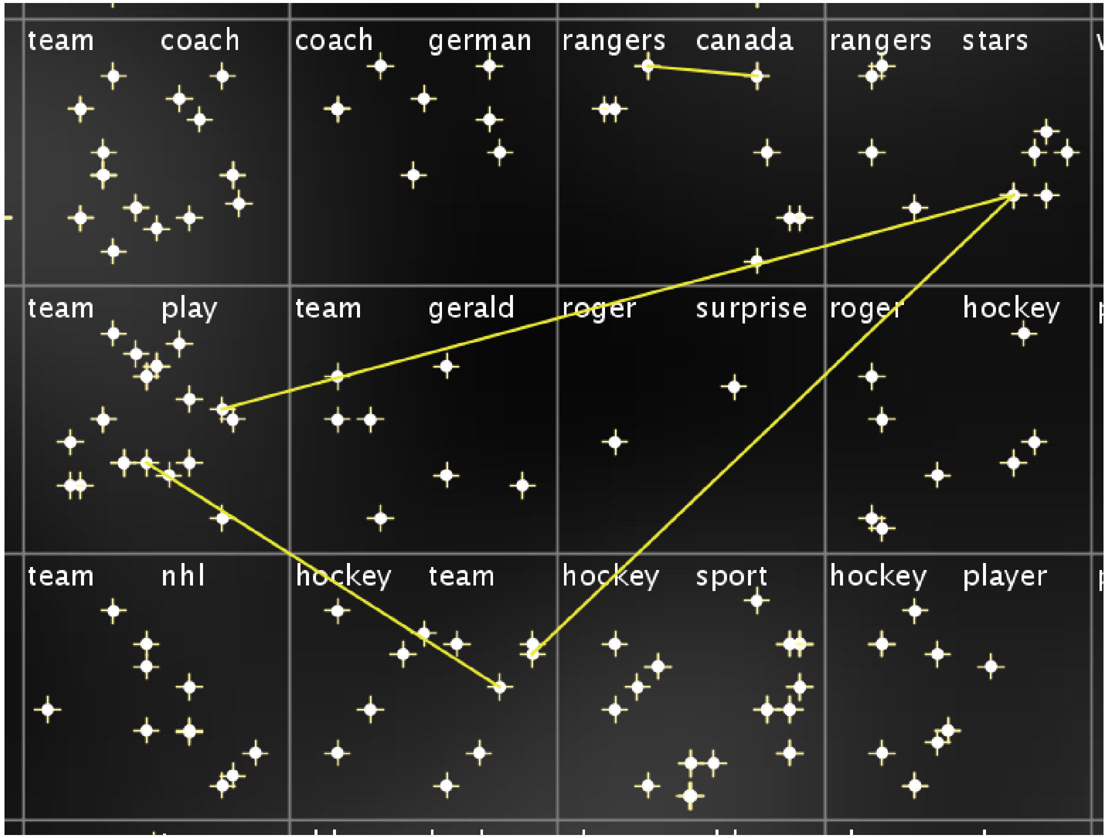
Dichte: Sky Metaphor

- 20 newsgroups data set



Sky Metaphor Visualisation

- 20 newsgroups data sets – semantic links



Sky Metaphor Visualisation

- 20 newsgroups data sets – semantic links

Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
  - Hit Histogramm
  - Activity Histogram
  - Smoothed Data Histograms
  - P-Matrix
  - Sky Metaphor
  - Neighborhood Graphs
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM

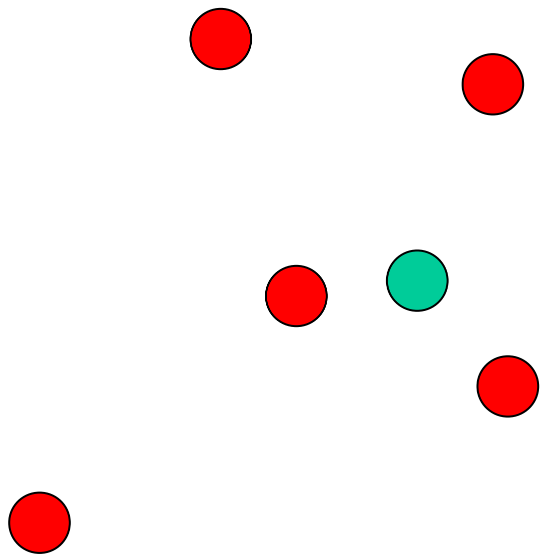
Dichte: Neighbourhood Graphs

- show, which areas of the map are close to each other based on density in input space
- how well the topology is preserved
- the location of dense or sparse regions
- Georg Pölzlbauer, Andreas Rauber, and Michael Dittenbach. **Graph projection techniques for self-organizing maps.** In Michel Verleysen, editor, Proceedings of the European Symposium on Artificial Neural Networks (ESANN'05), pages 533-538, Bruges, Belgium, April 27-29 2005. d-side publications.

Neighbourhood Graphs

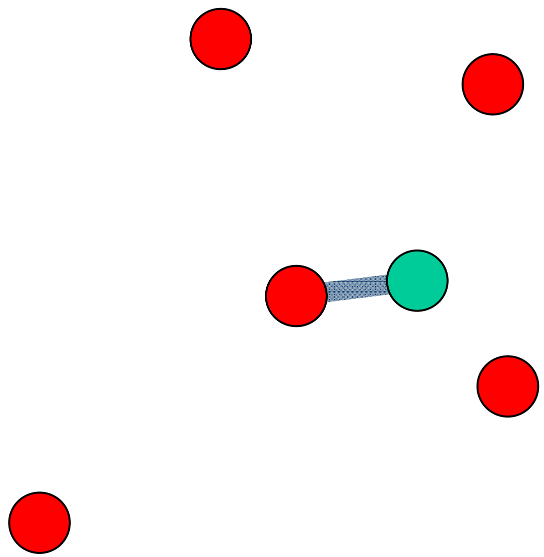
- similar to P-matrix, but for pairs of units
- different levels of granularity interactive analysis
- 2 approaches:
  - knn - based distances
  - radius-based distances

Neighbourhood Graphs: k-n



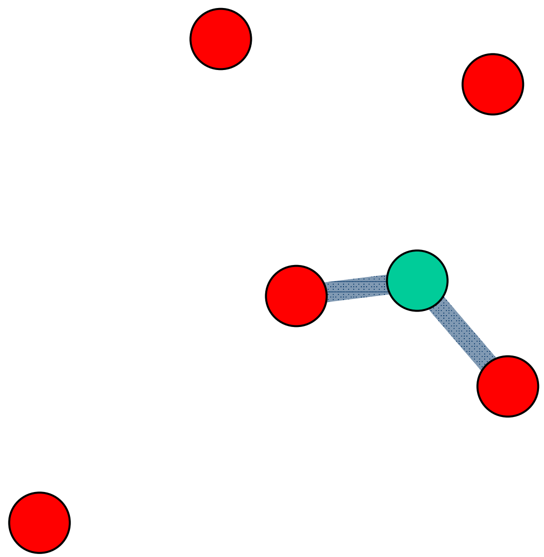
KNN-based distances in input space

Neighbourhood Graphs: k-n



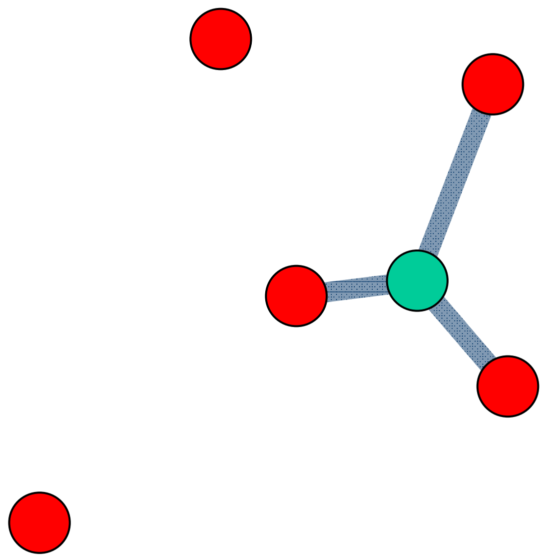
1-nearest neighbor

Neighbourhood Graphs: k-n



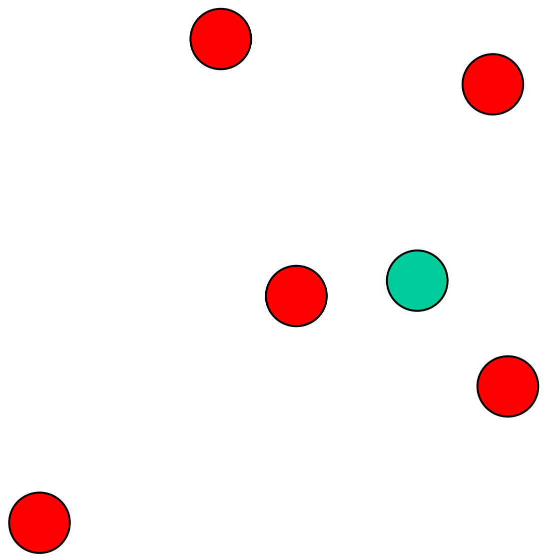
2-neares neighbors

Neighbourhood Graphs: k-n



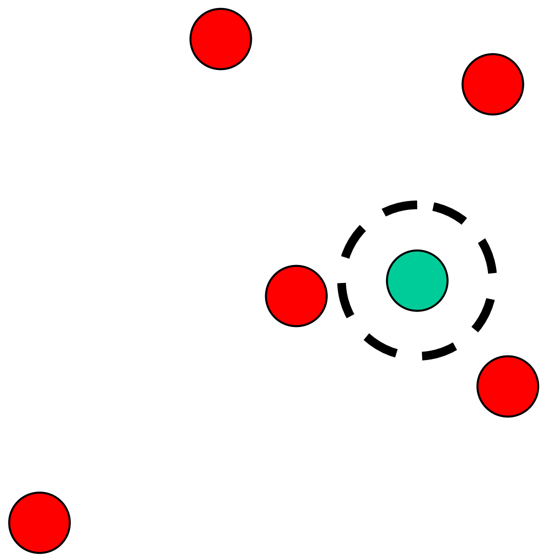
3-nearest neighbors

Neighbourhood Graphs: Radius



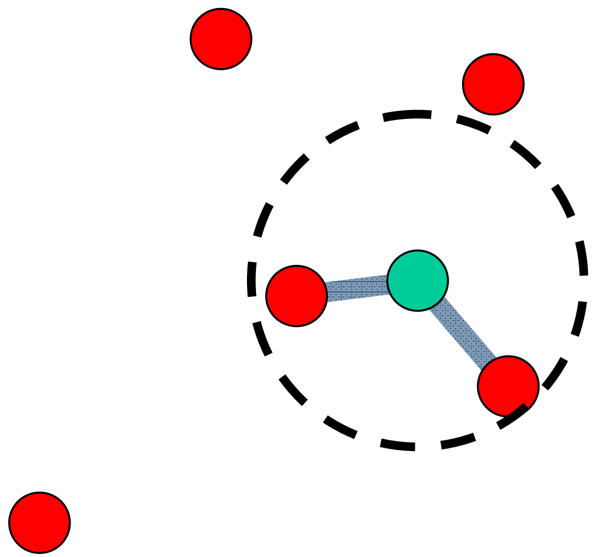
Radius-based distances in input space

Neighbourhood Graphs: Radius



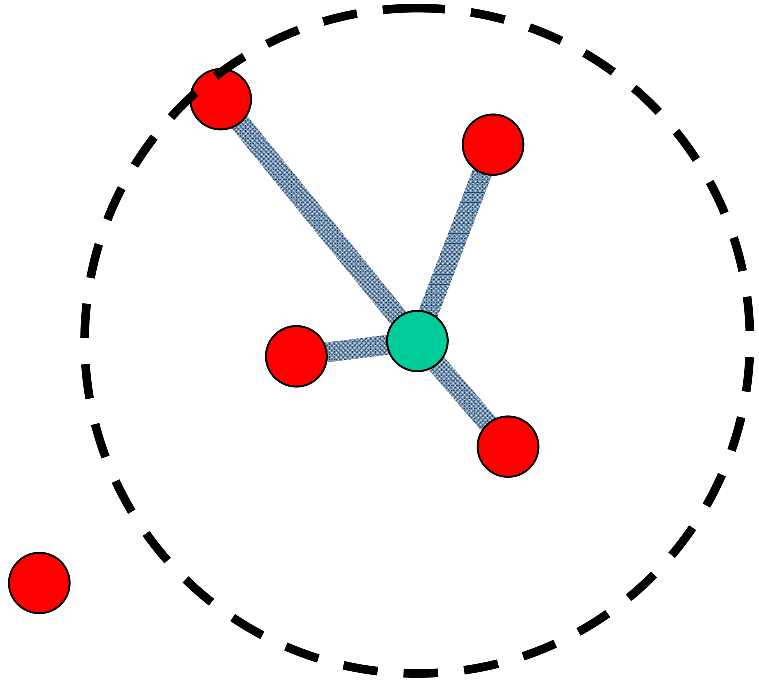
Radius: 1.0

Neighbourhood Graphs: Radius



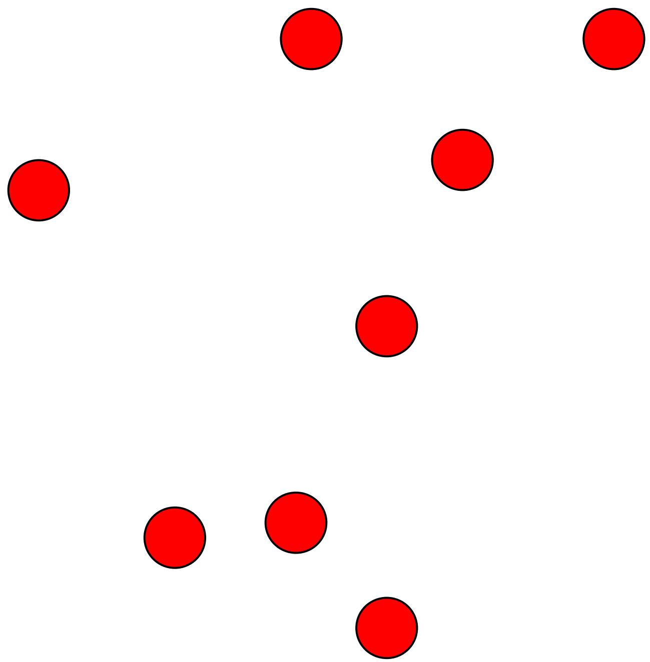
Radius: 2.0

Neighbourhood Graphs: Radius

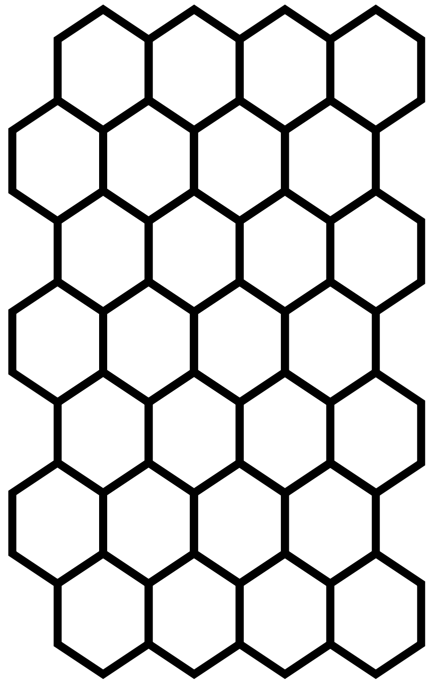


Radius: 3.0

Neighbourhood Graphs: Projection

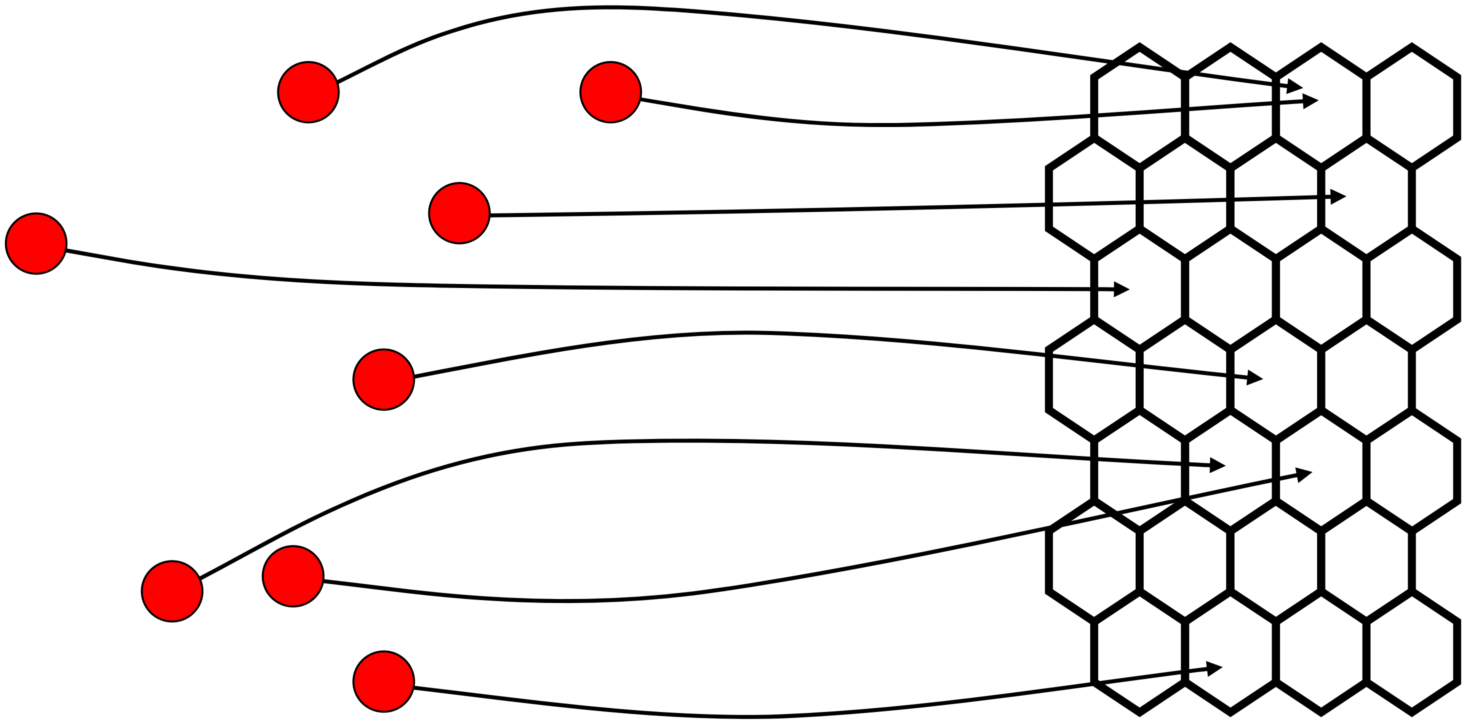


Input Space



Map Lattice

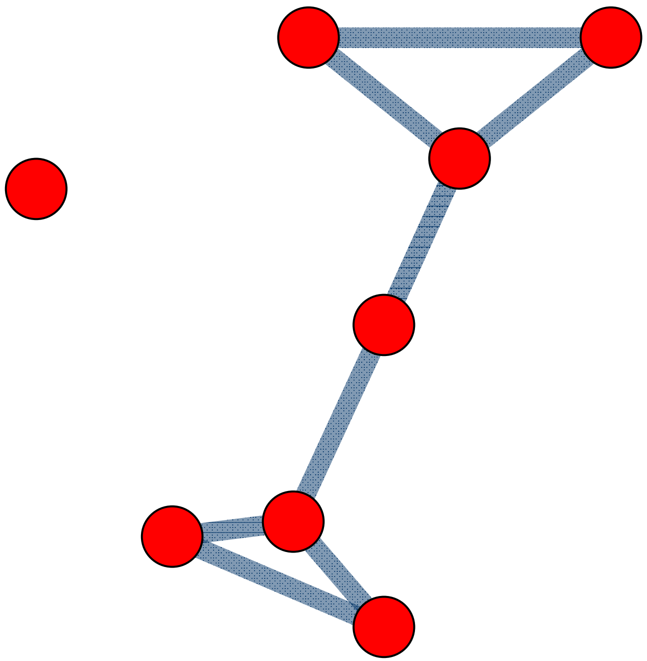
Neighbourhood Graphs: Projection



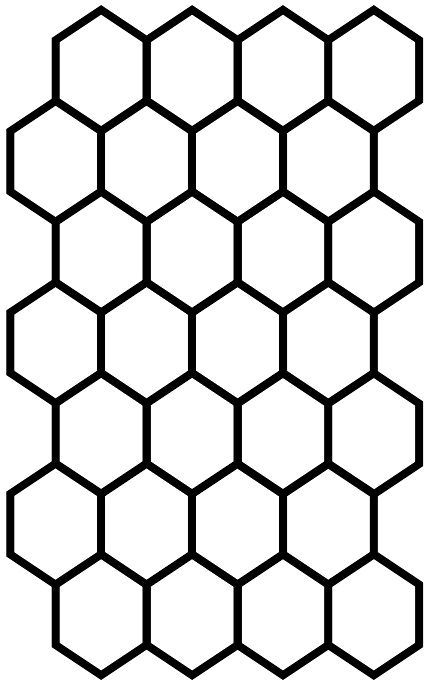
Input Space

Map Lattice

Neighbourhood Graphs: Projection

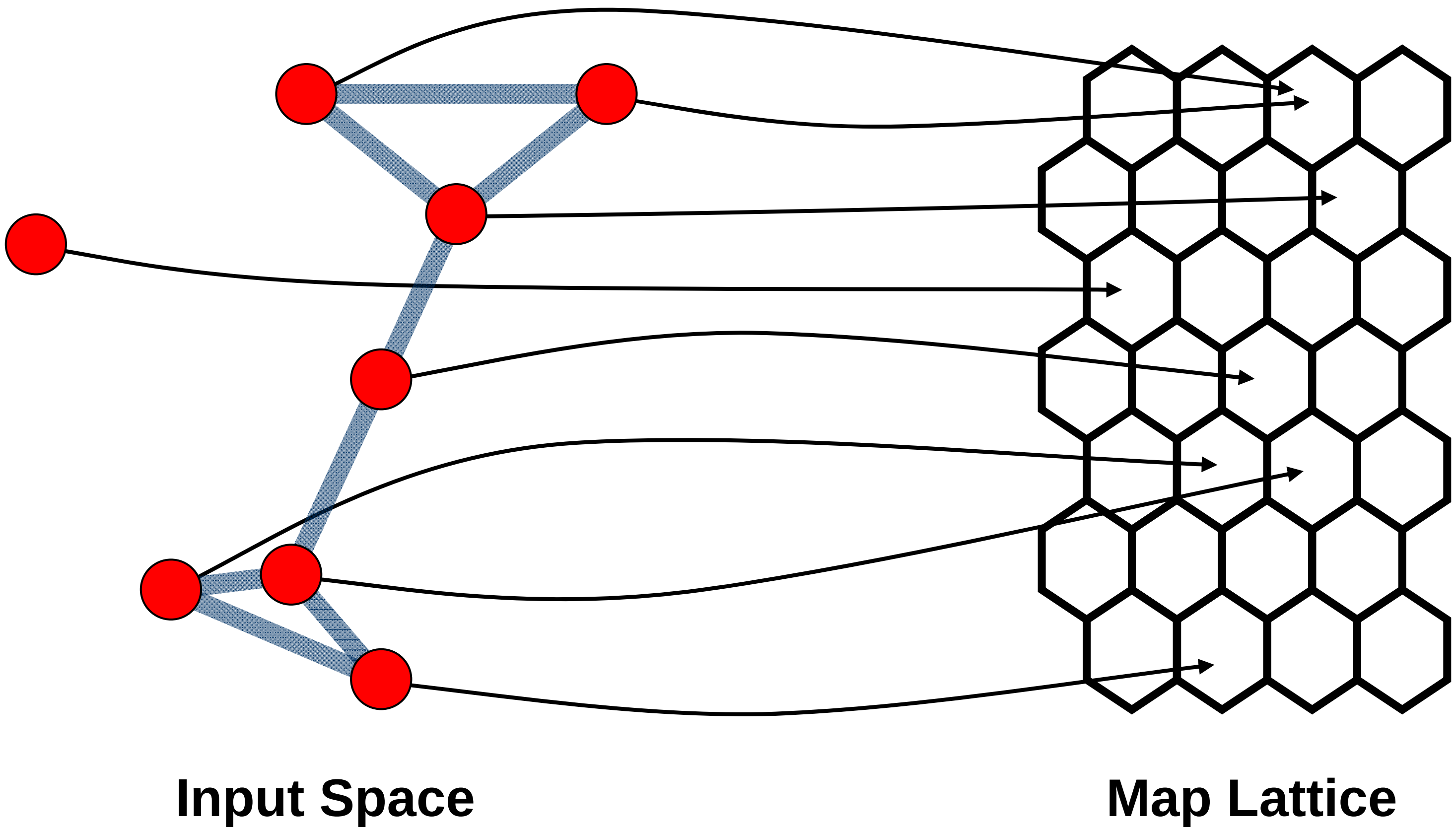


Input Space

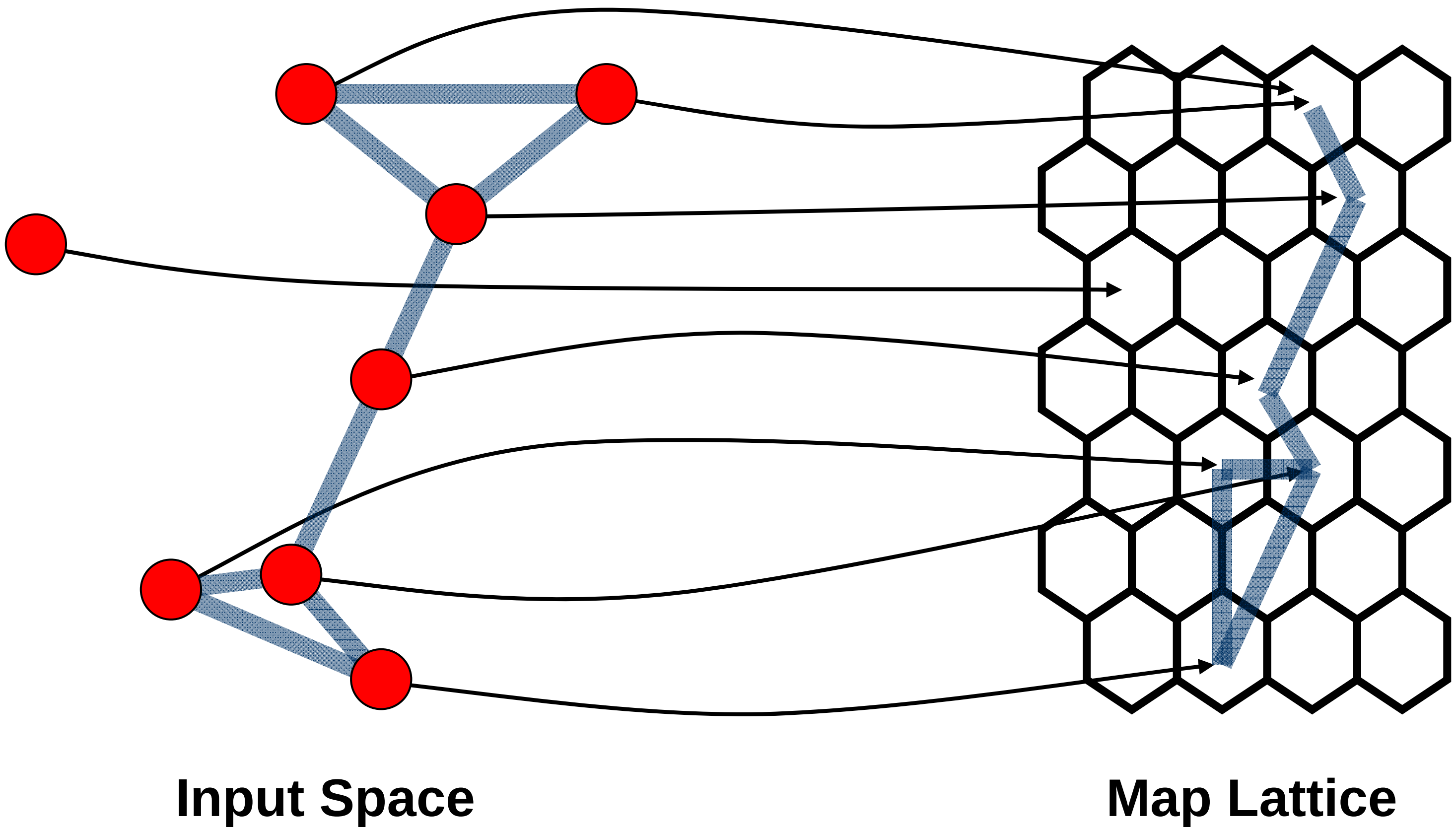


Map Lattice

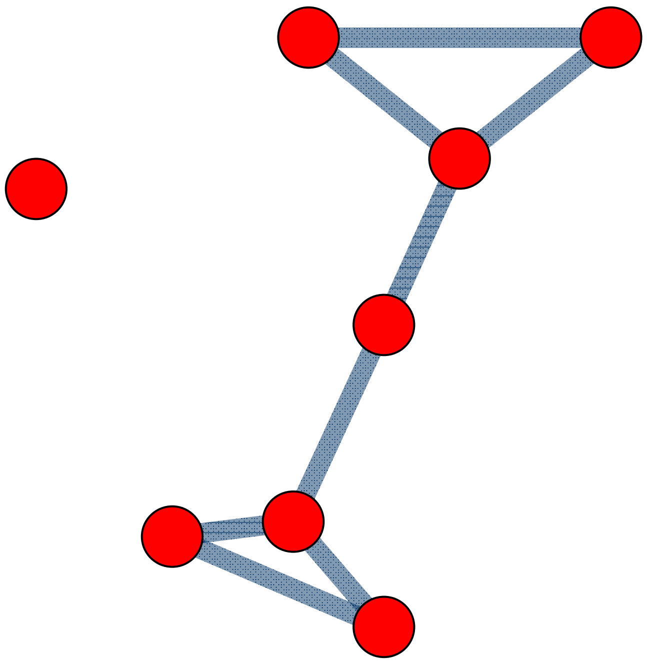
Neighbourhood Graphs: Projection



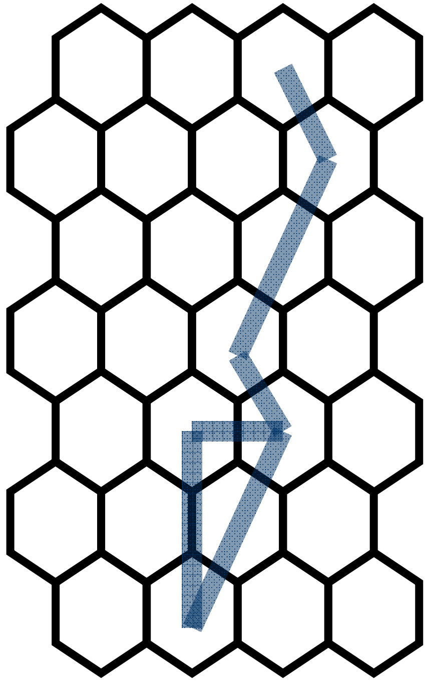
Neighbourhood Graphs: Projection



Neighbourhood Graphs: Projection

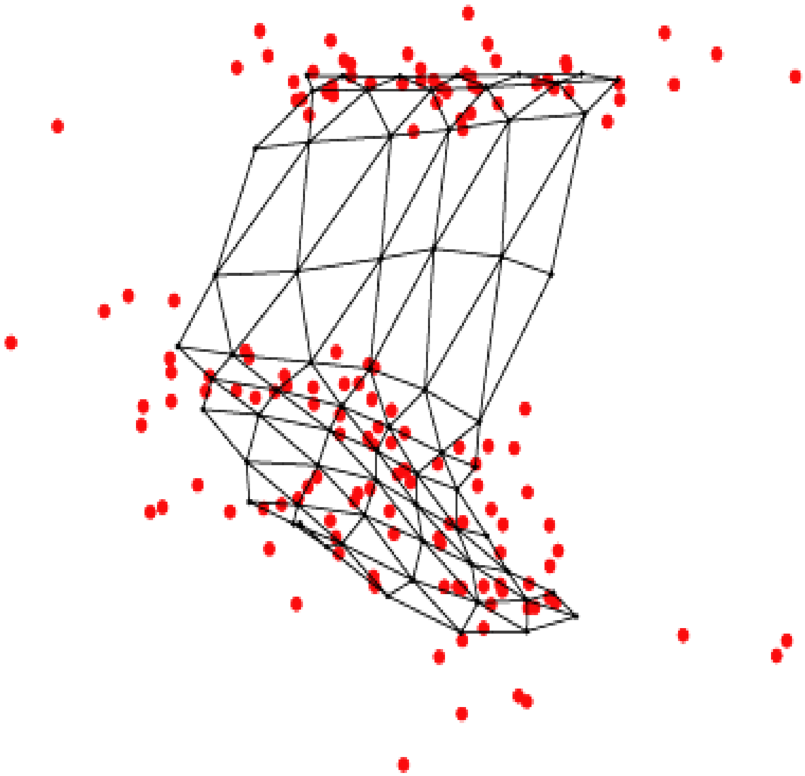


Input Space



Map Lattice

Example: Iris Data Set

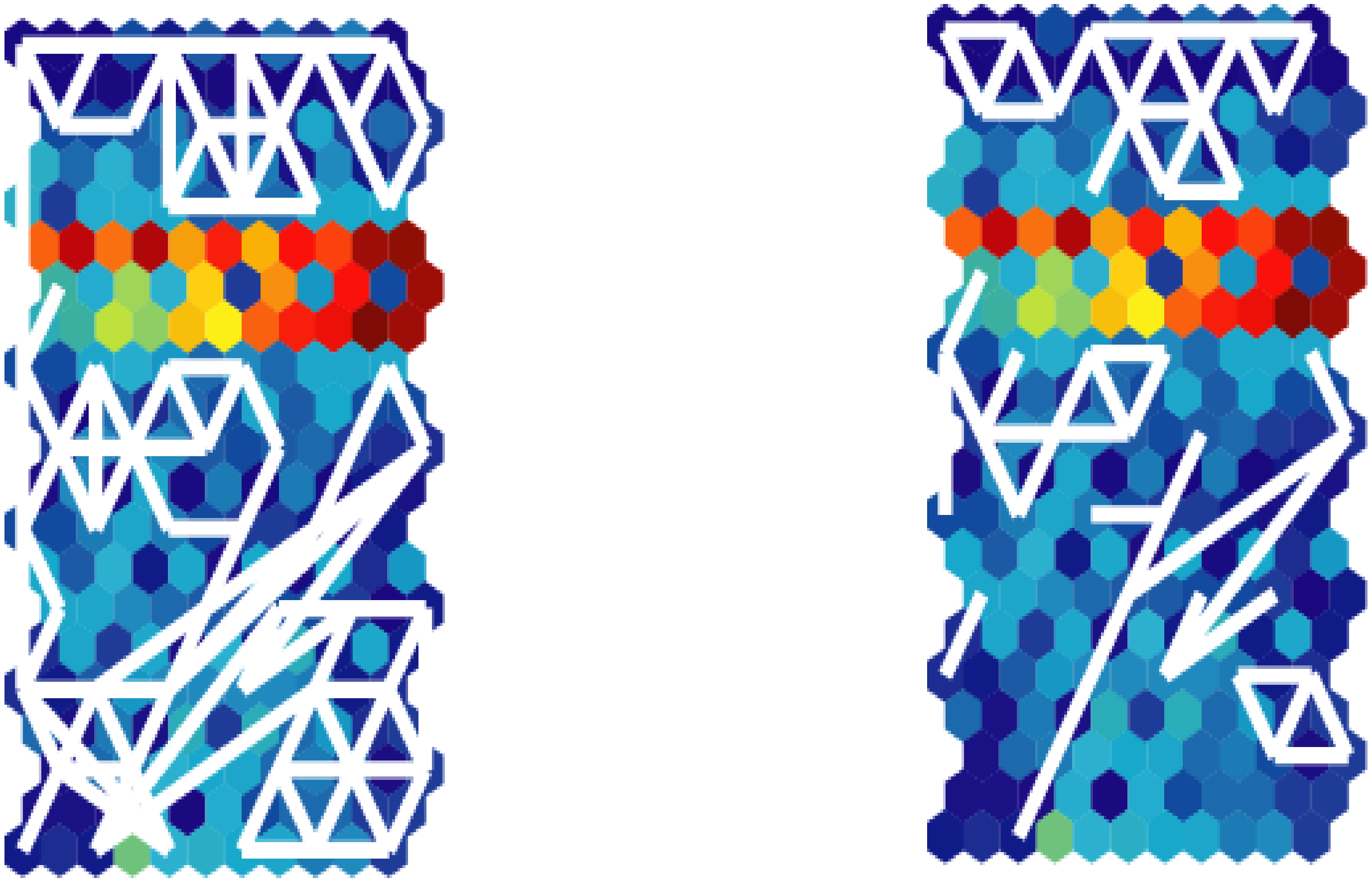


Neighborhood Graphs: Iris

1-Nearest Neighbors

Radius: 0.2

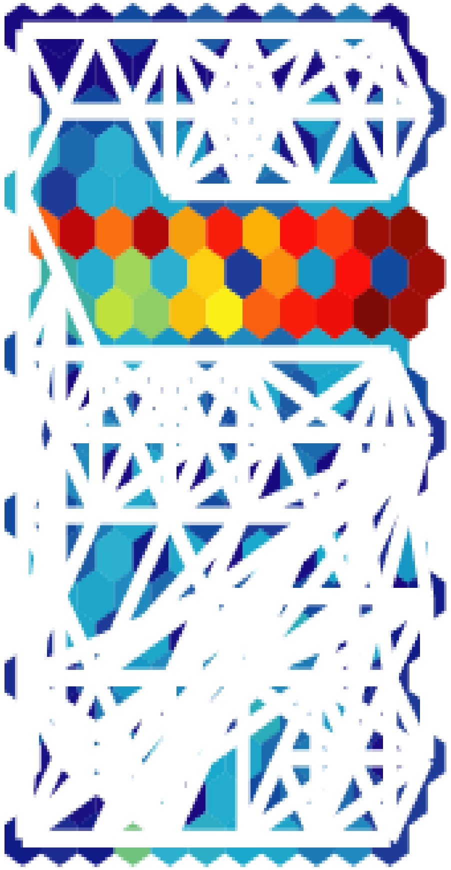
Neighborhood Graphs: Iris



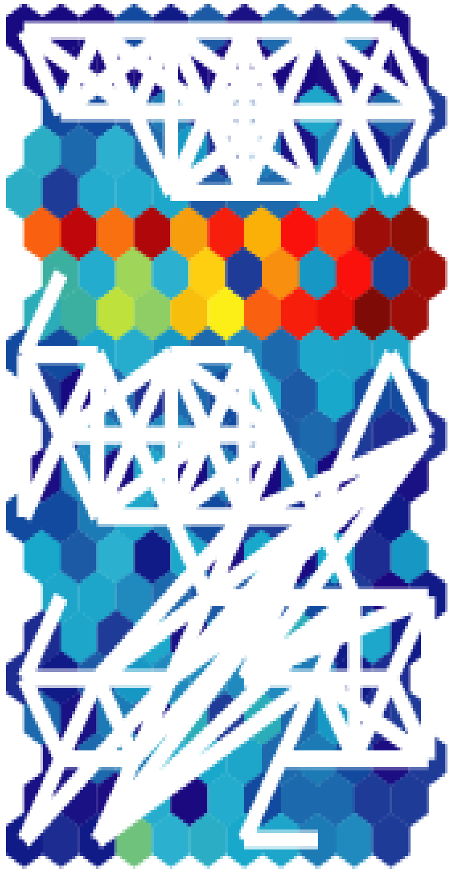
3-Nearest Neighbors

Radius: 0.4

Neighborhood Graphs: Iris

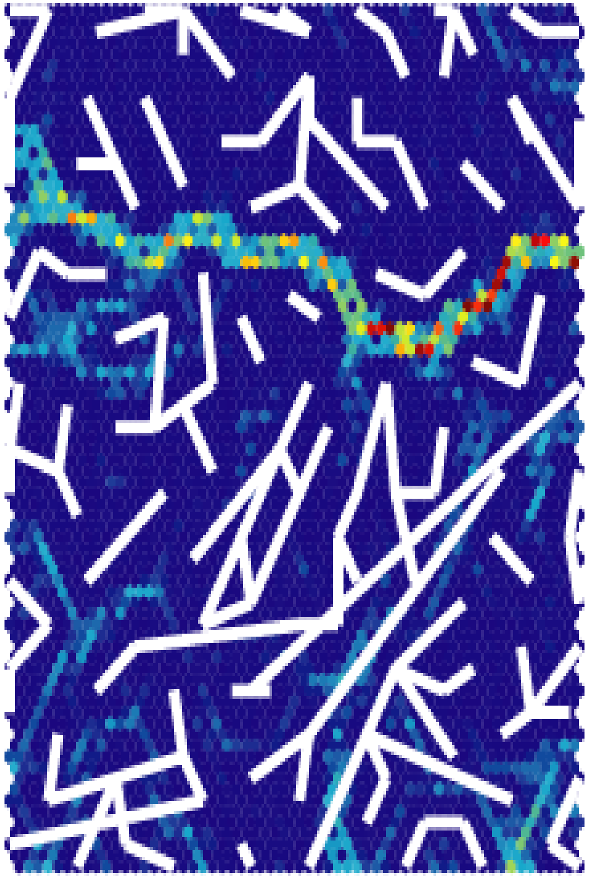


10-Nearest Neighbors

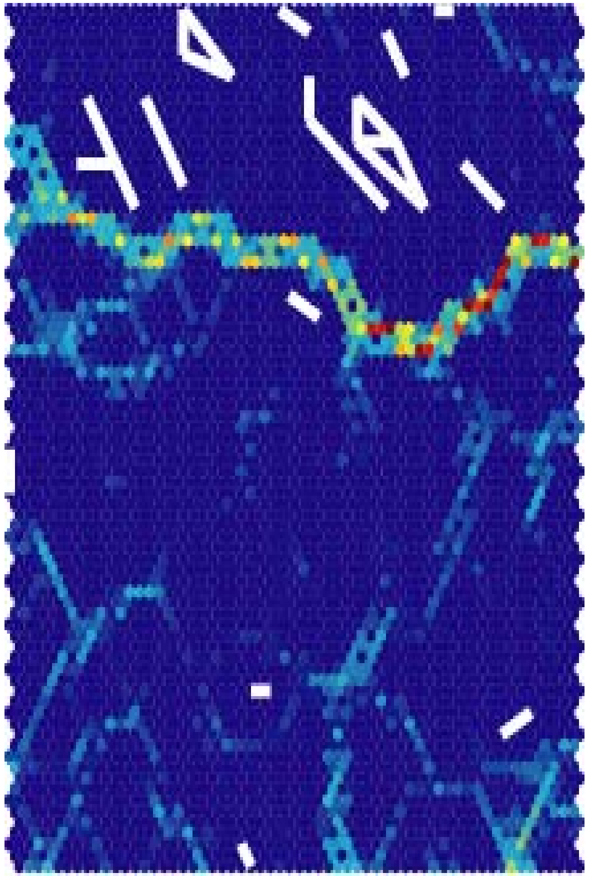


Radius: 0.6

Neighborhood Graphs: Iris

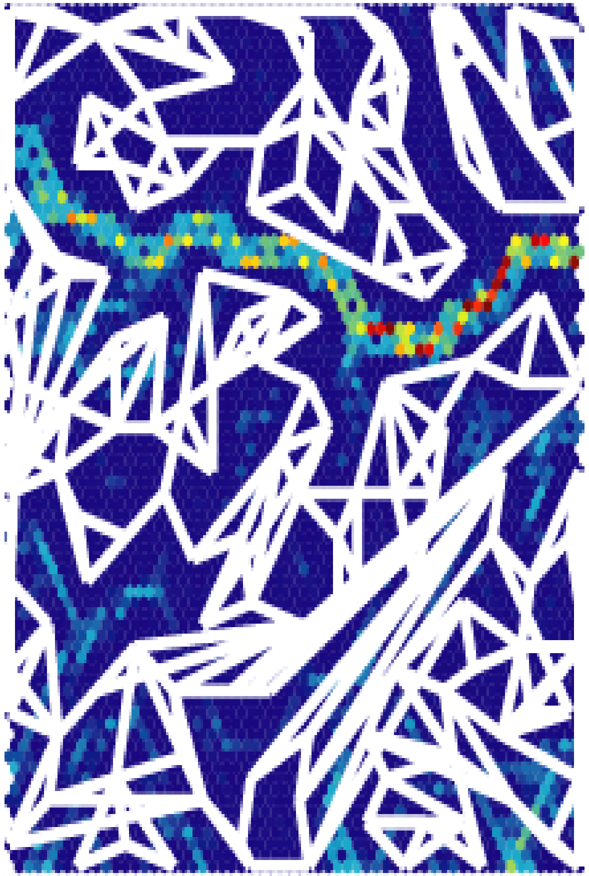


1-Nearest Neighbors

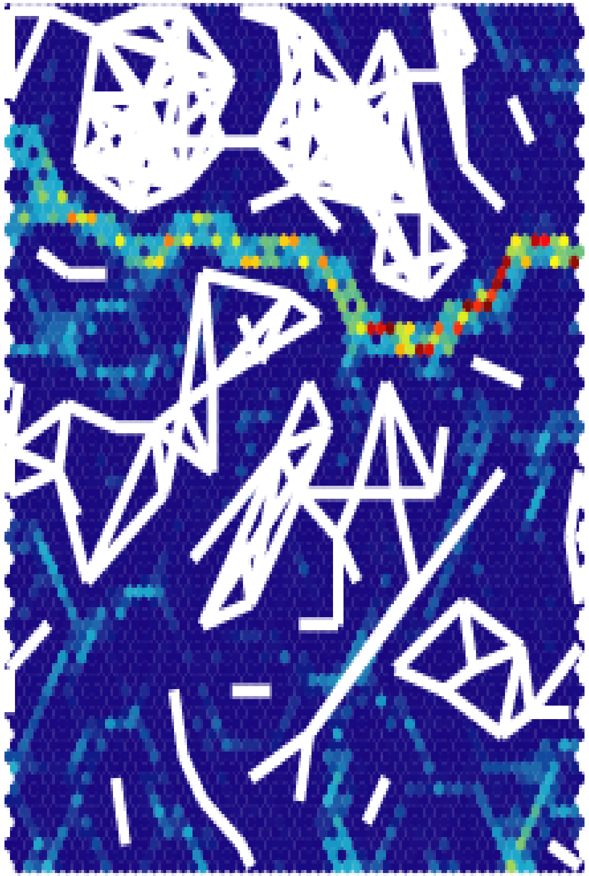


Radius: 0.2

Neighborhood Graphs: Iris

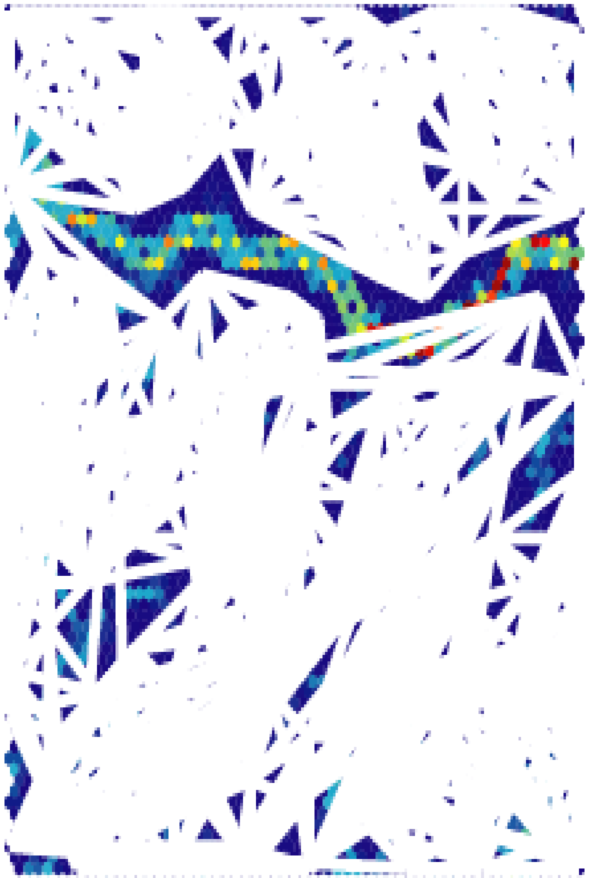


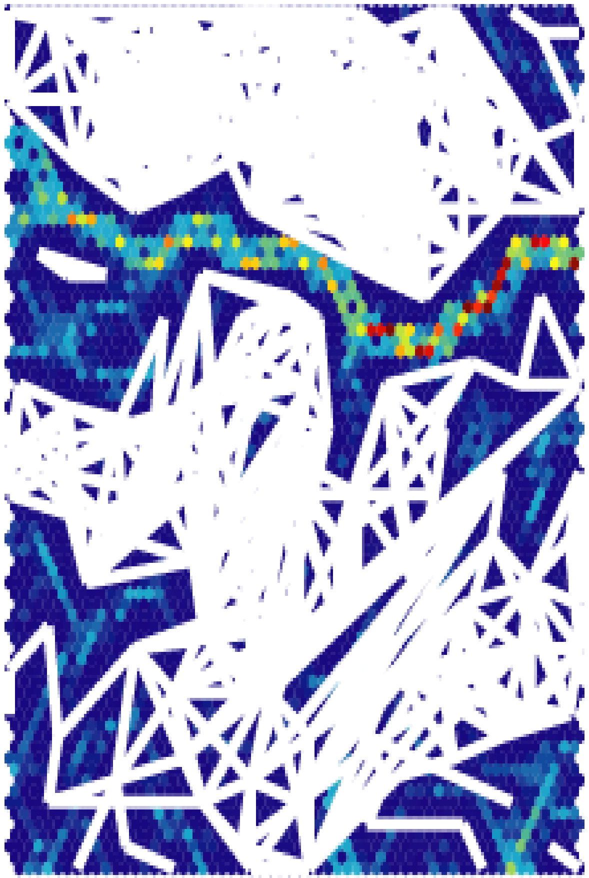
3-Nearest Neighbors



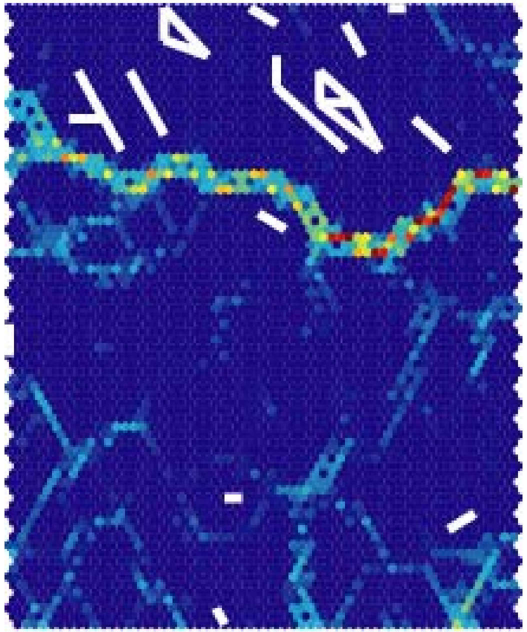
Radius: 0.4

Neighborhood Graphs: Iris

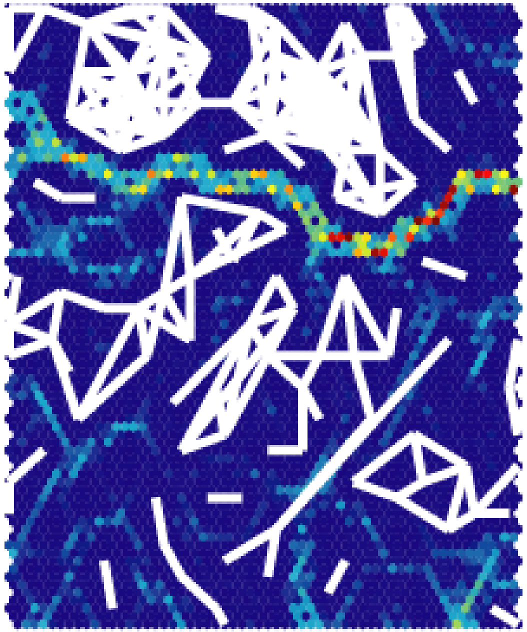

10-Nearest Neighbors


Radius: 0.6

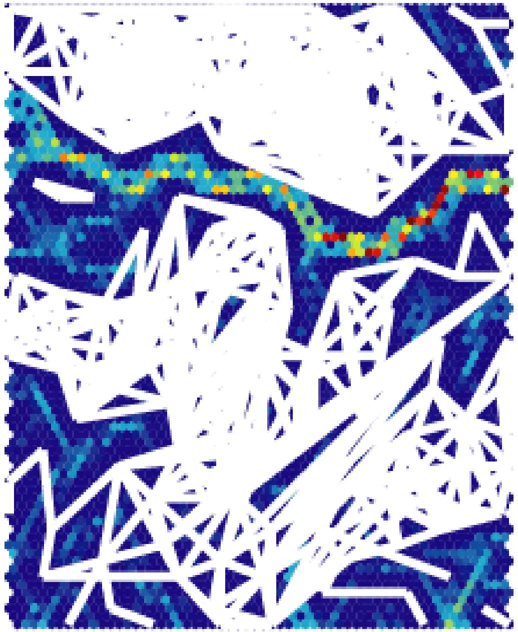
Neighborhood Graphs: Iris



Radius: 0.2



Radius: 0.4



Radius: 0.6

Neighbourhood Graphs

- Show relationship between input & map space
- Radius method: focus on density
  - for clusters of different size
- Nearest neighbors: focus on cluster size
  - for clusters with different densities
- High parameter values:
  show clusters and their connections
- Long lines: **topology violation**
  (quality indicator)

Neighborhood Graphs

Fragen

- Was ist der Unterschied zwischen der Radius und der nn-Methode zur Ermittlung der Nachbarschaft?
- Wann werden die beiden Methoden unterschiedliche Ergebnisse liefern? Welche Aussagen können abgeleitet werden, wenn sich die Ergebnisse unterscheiden?
- Wann hat nn ein dichtes Liniennetz, wann die Radius-Methode?
- Was sagen lange Linien aus, die quer über die Karte gehen?

Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
  - Hit Histogramm
  - Activity Histogram
  - Smoothed Data Histograms
  - P-Matrix
  - Sky Metaphor
  - Neighborhood Graphs
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM

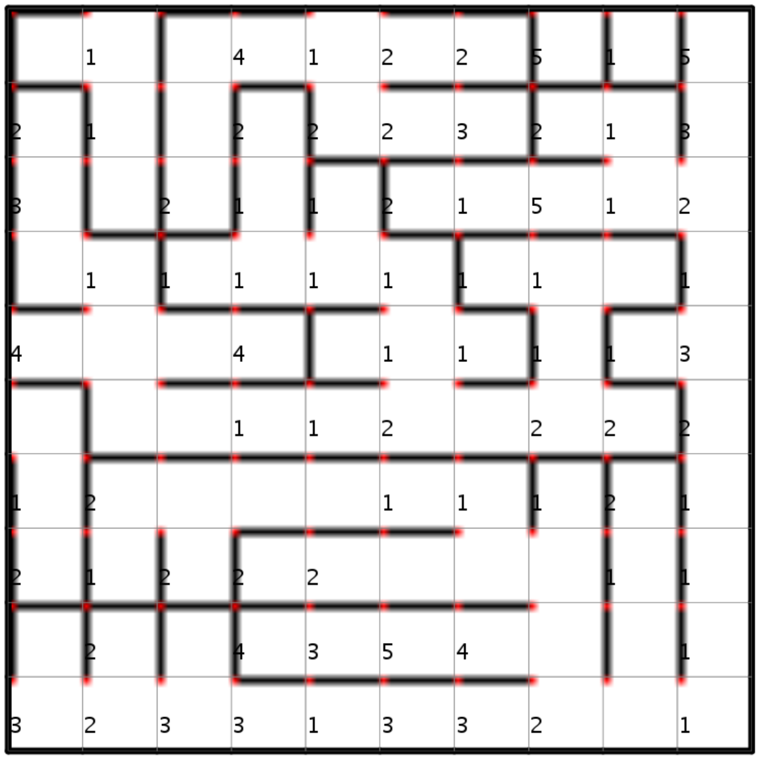
Outline

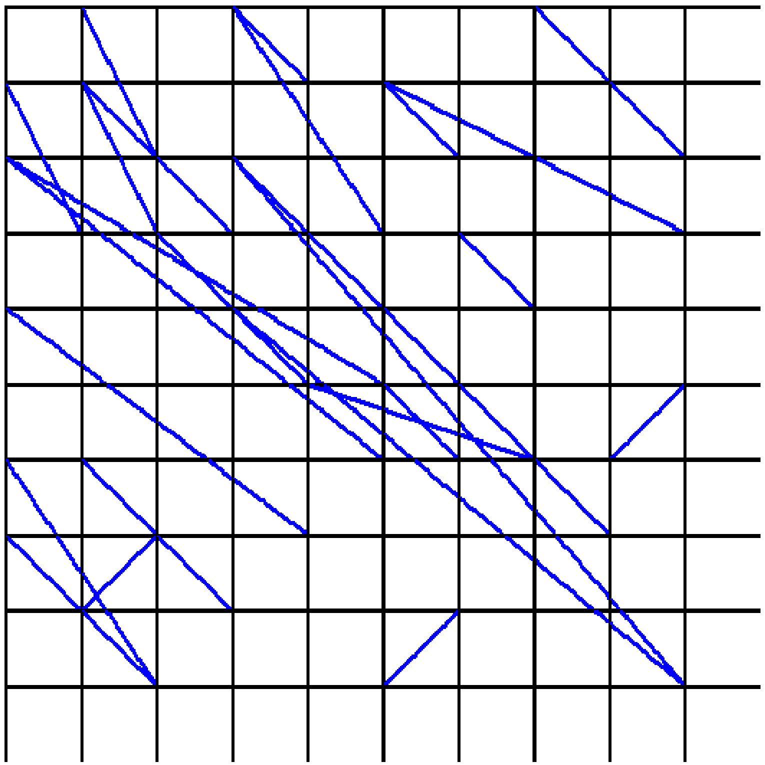
Visualisierungen der SOM

- Textuelle Informationen
- Dichte
- Distanzen
  - Minimum Spanning Trees
  - Cluster Connections (CC)
  - D-Matrix
  - U-Matrix
  - U* Matrix: U-Matrix + P-Matrix
  - Vectorfields: Flow / Borderline
- Klasseninformation
- Attribute
- Clustering der SOM

Distanzen: MSPT

- Minimum Spanning Tree
  - MSPT der Vektoren
  - MSPT der Units





Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
- Distanzen
  - Minimum Spanning Trees
  - Cluster Connections (CC)
  - D-Matrix
  - U-Matrix
  - U* Matrix: U-Matrix + P-Matrix
  - Vectorfields: Flow / Borderline
- Klasseninformation
- Attribute
- Clustering der SOM

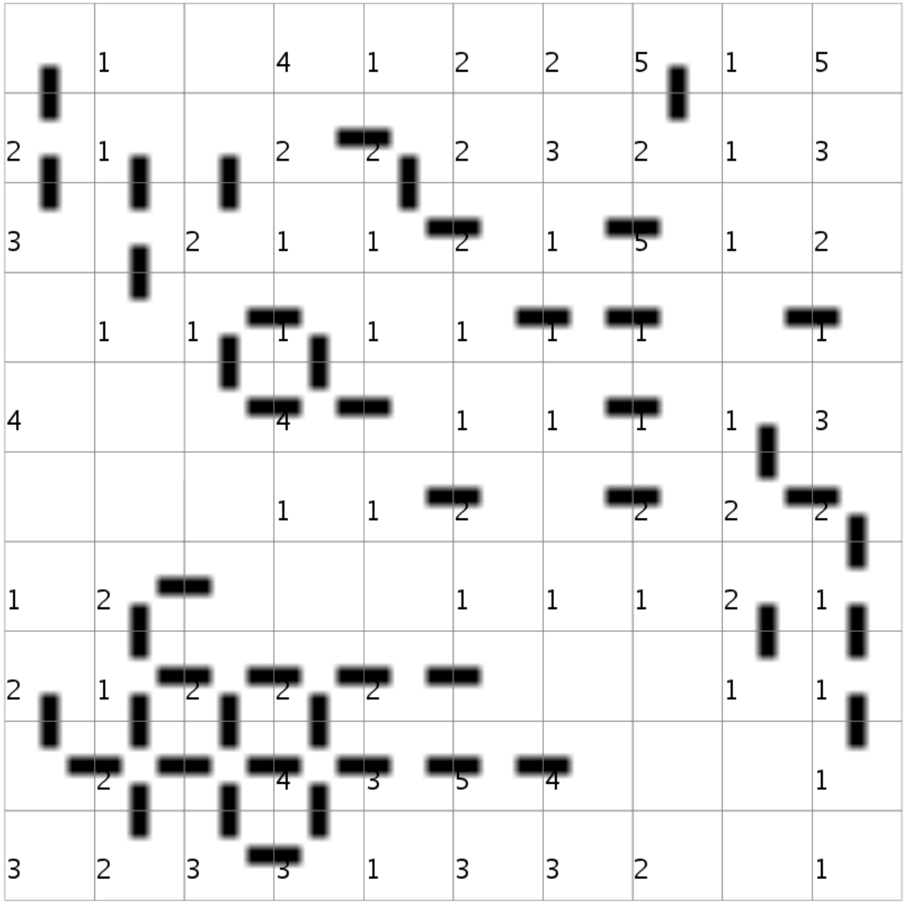
Distanzen

- Berechnung der Abstände zwischen den Gewichtsvektoren
  - kleine Abstände: ähnliche Datenpunkte
  - große Abstände: Clustergrenzen
- verschiedene Verfahren
  - Cluster Connections (CC)
  - D-Matrix
  - U-Matrix
  - U* Matrix: U-Matrix + P-Matrix
  - Vectorfields: Flow / Borderline

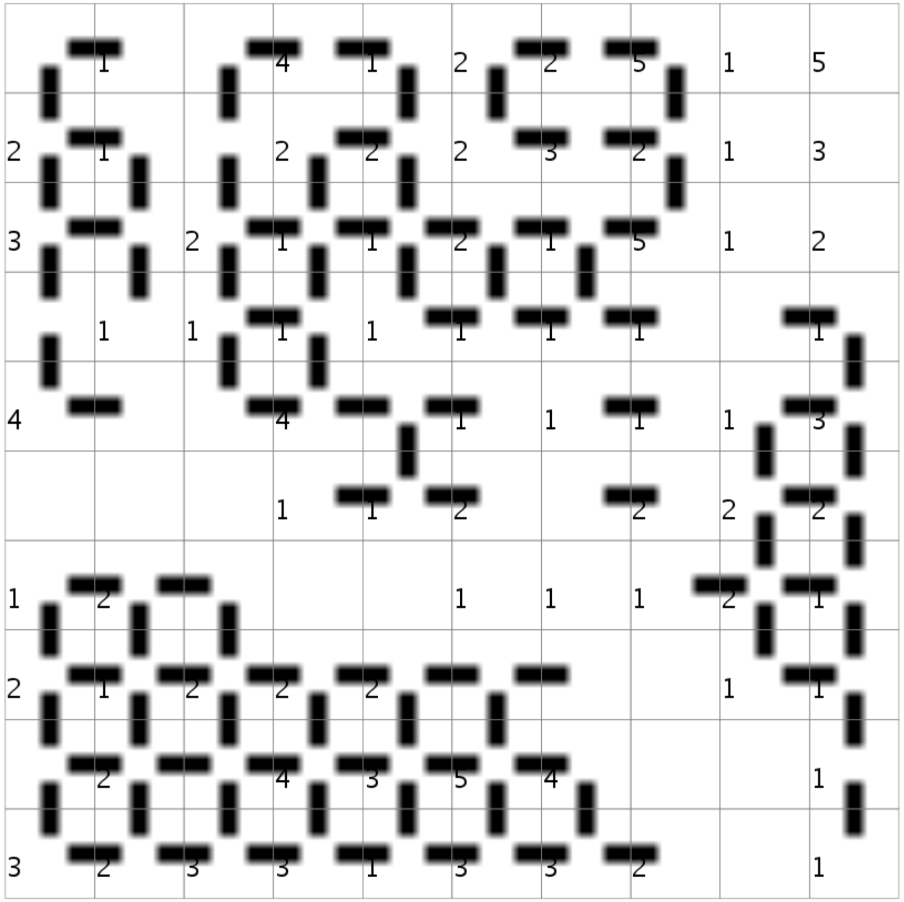
Distanzen: CC

- Cluster Connections (CC)
- Berechnung der Distanz zwischenbenachbarten Gewichtsvektoren im Datenraum
- Einzeichnen von Verbindungen im Ausgaberaum
- Schwellwerte, Graph
- erlaubt interaktive Ermittlung der Clusterstruktur
- Dieter Merkl, Andreas Rauber: Cluster Connections -- A visualization technique to reveal cluster boundaries in self-organizing maps. In: Proc 9th Italian Workshop on Neural Nets (WIRN97), Vietri sul Mare, Italy, Springer, 1997.

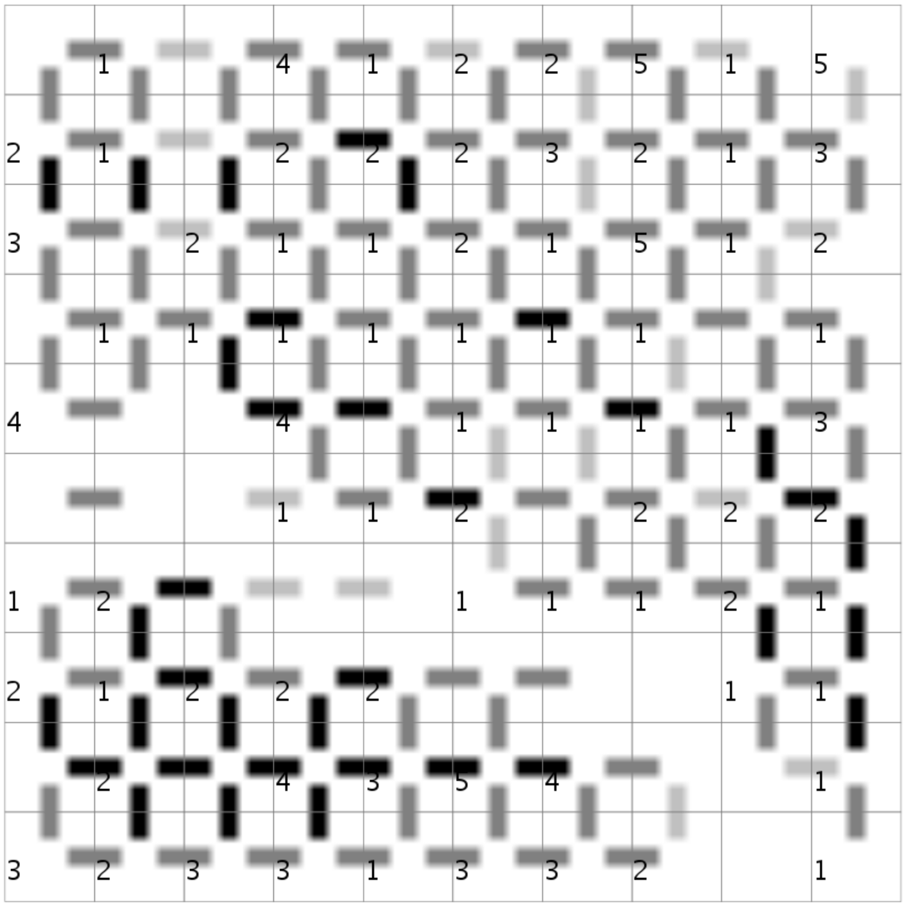
Distanzen: CC



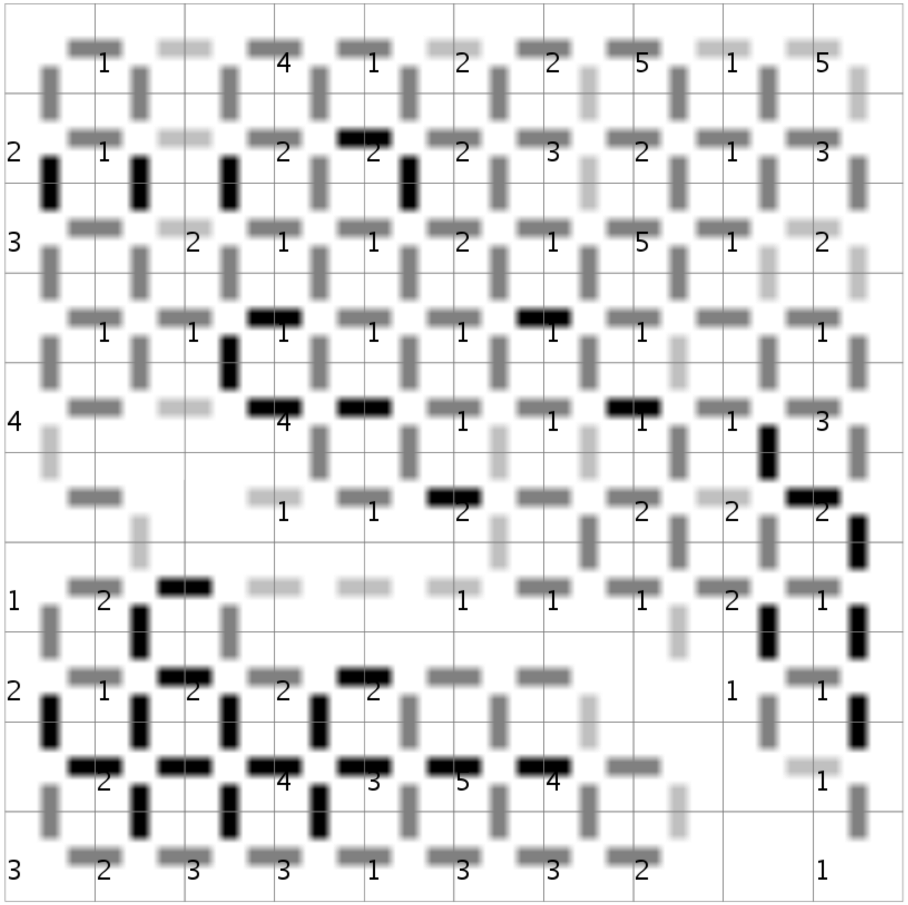
Distanzen: CC



Distanzen: CC



Distanzen: CC



Outline

Visualisierungen der SOM

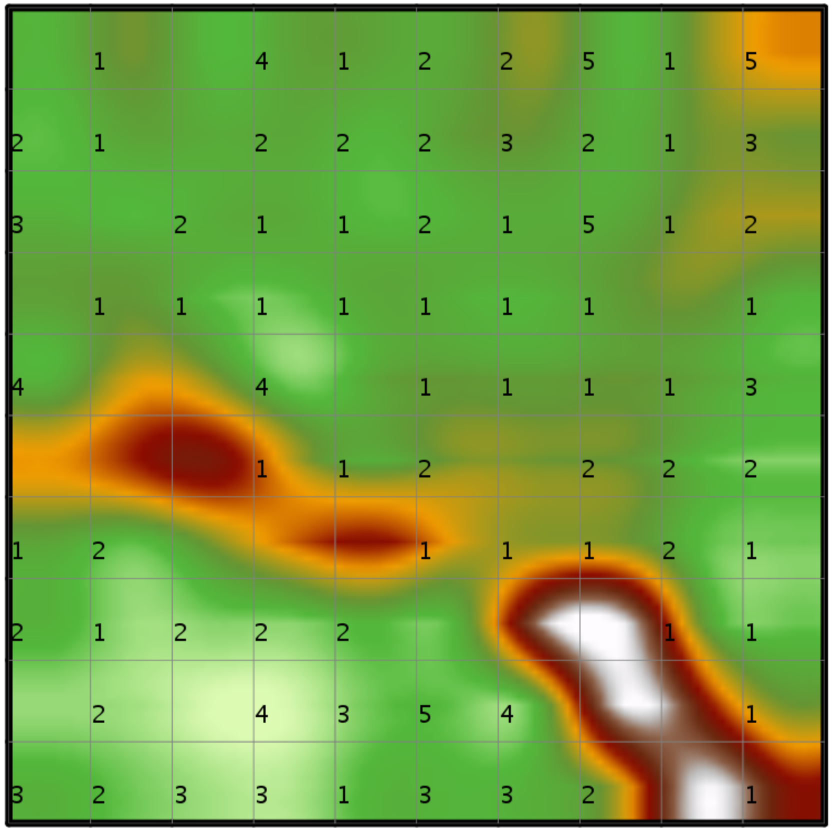
- Textuelle Informationen
- Dichte
- Distanzen
  - Minimum Spanning Trees
  - Cluster Connections (CC)
  - D-Matrix
  - U-Matrix
  - U* Matrix: U-Matrix + P-Matrix
  - Vectorfields: Flow / Borderline
- Klasseninformation
- Attribute
- Clustering der SOM

Distanzen: D/U-Matrix

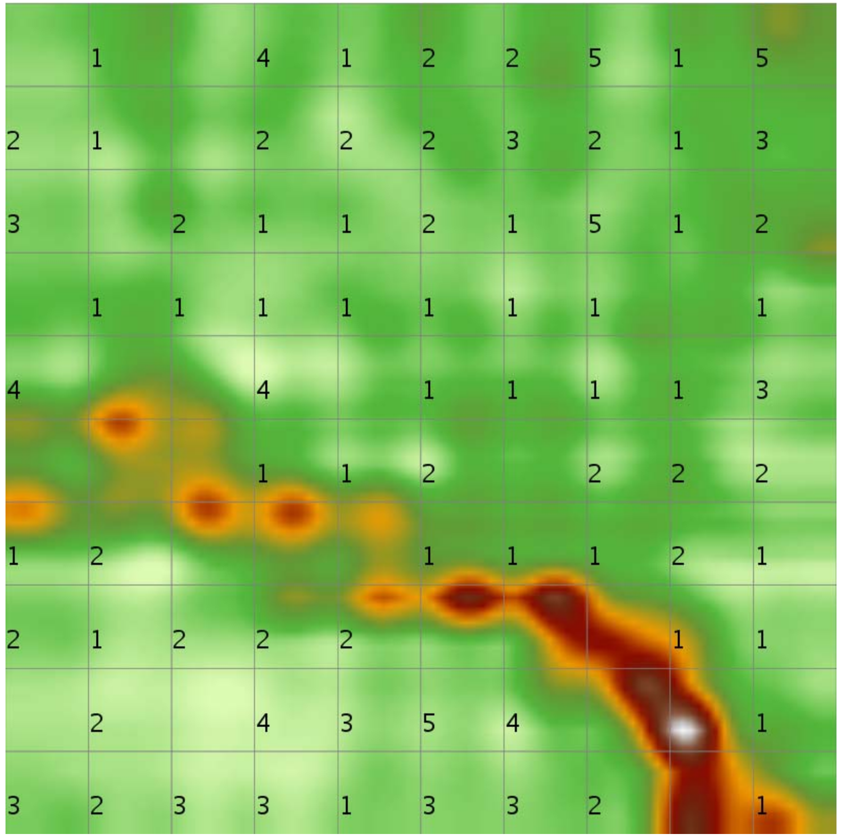
- berechnet aus Distanz zwischen benachbarten Units
- D-Matrix: durchschnittliche Distanz der Unit zu Nachbarn
- U-Matrix: Interpolationsfelder mit Distanz zu jedem Nachbarn – feinere Auflösung
- veranschaulicht Clusterstruktur
- „Berge“ (hohe Werte): Clustergrenzen
- „Täler“ (niedrige Werte): kohärente, ähnliche Regionen
- 2D oder 3D Ansicht
- A. Ultsch.: Self-organizing neural networks for visualization and classification. In *Information and Classification. Concepts, Methods and Application*. Springer Verlag, 1993.

Distanzen: D/U-Matrix

- Iris Dataset:



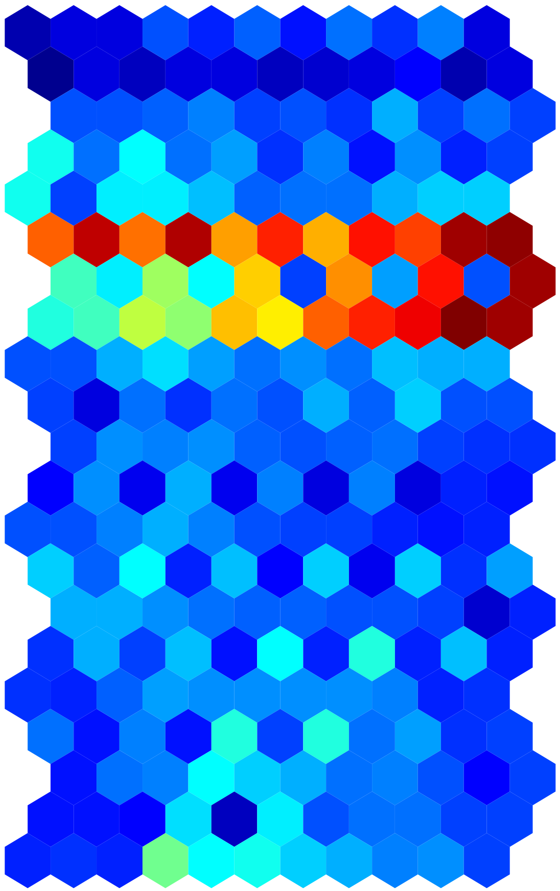
D-Matrix

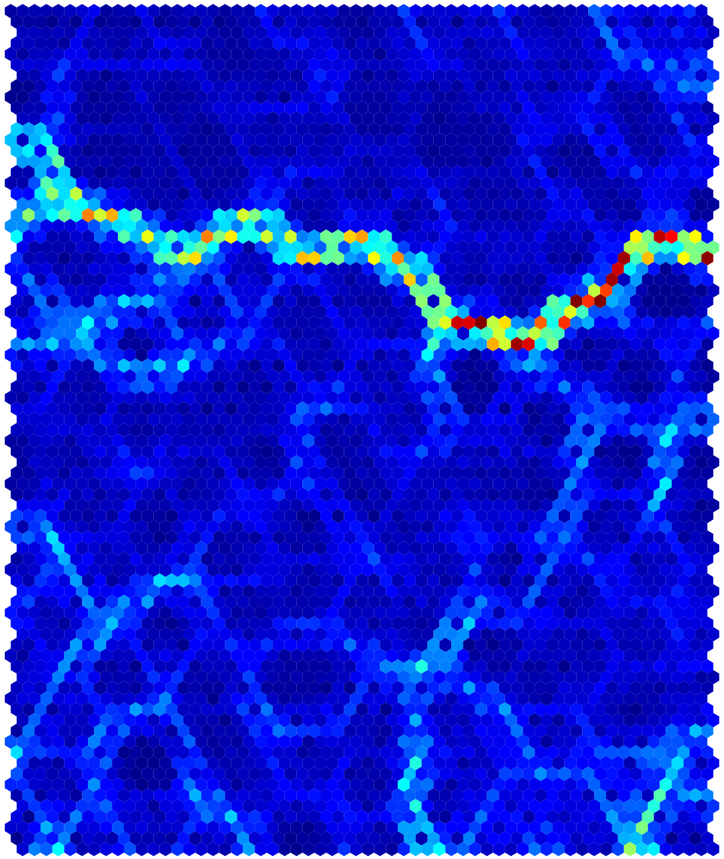


U-Matrix

Distanzen: U-Matrix

- Iris Dataset: kleine / große SOM





Outline

Visualisierungen der SOM

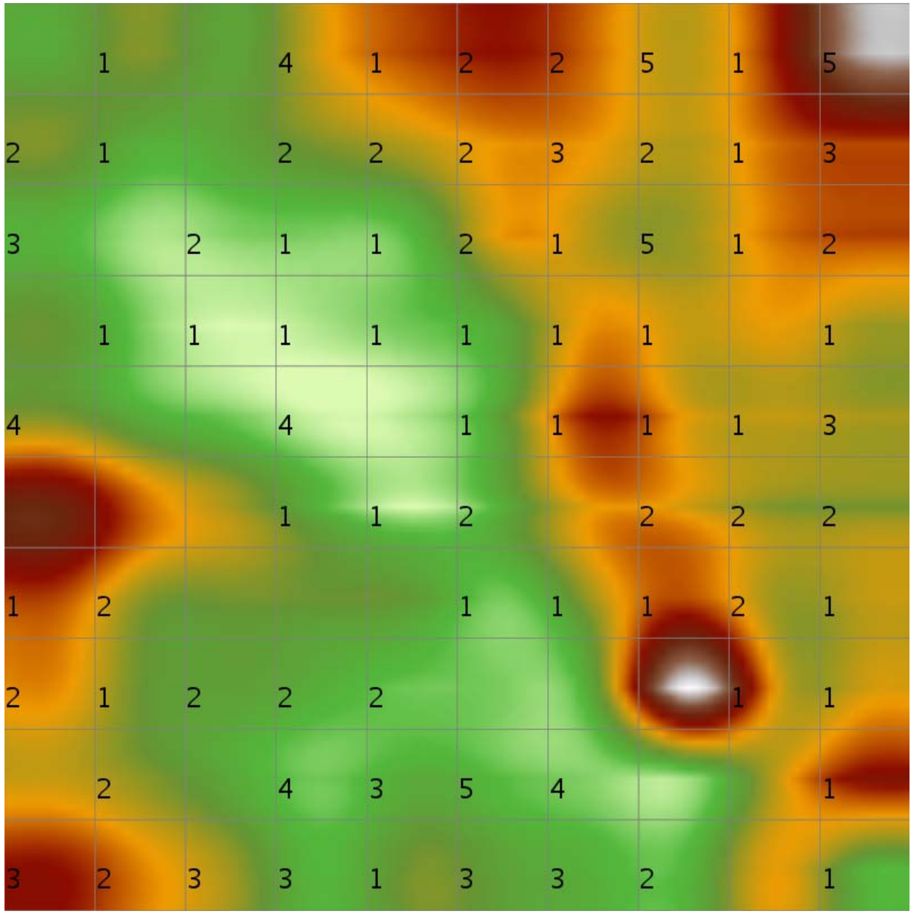
- Textuelle Informationen
- Dichte
- Distanzen
  - Minimum Spanning Trees
  - Cluster Connections (CC)
  - D-Matrix
  - U-Matrix
  - U* Matrix: U-Matrix + P-Matrix
  - Vectorfields: Flow / Borderline
- Klasseninformation
- Attribute
- Clustering der SOM

Distanzen + Dichte: U*

- P-Matrix: zeigt Dichte der Daten auf Nodes
- U-Matrix zeigt Distanzen/Clustergrenzen
- U*-Matrix: Kombination von U-Matrix und P-Matrix
- Ultsch A. U*-Matrix: A Tool to Visualize Cluster in High-Dimensional Data. Tech. Report, Dept. of Mathematics and Computer Science, University of Marburg.

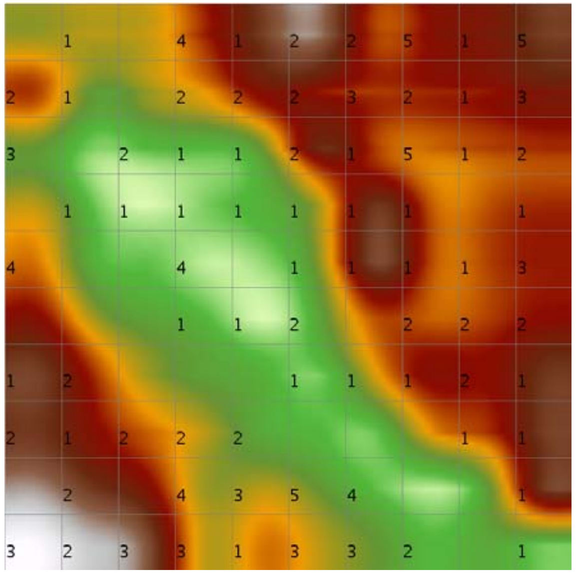
Distanzen + Dichte: U*

- Iris Dataset:

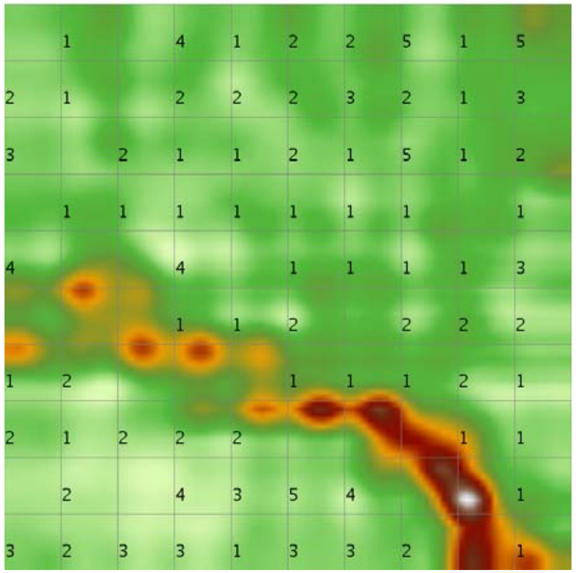


Distanzen + Dichte: U*

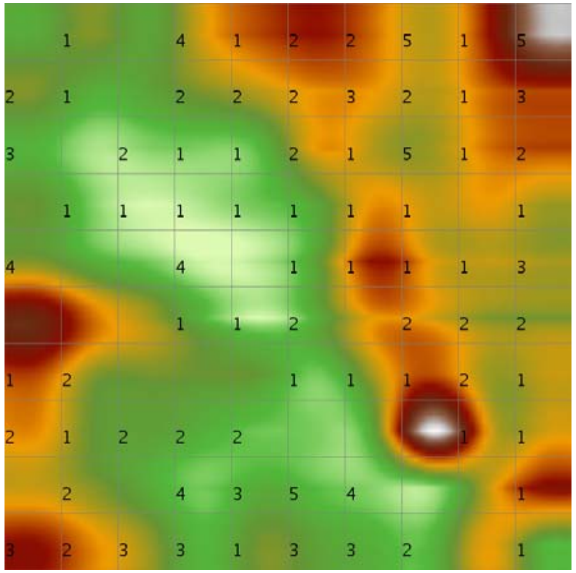
- Iris Dataset:



P-Matrix



U-Matrix



U*-Matrix

Distanzen + Dichte: U*

Fragen

- Welche Aussagen lassen sich aus der P-Matrix ableiten, welche aus der U-Matrix
- Welchen Sinn macht die U* Matrix – wann ist sie besonders wichtig?

Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
- Distanzen
  - Minimum Spanning Trees
  - Cluster Connections (CC)
  - D-Matrix
  - U-Matrix
  - U* Matrix: U-Matrix + P-Matrix
  - Vectorfields: Flow / Borderline
- Klasseninformation
- Attribute
- Clustering der SOM

Distanzen: SOM Vector Fields

Vector field graphs show

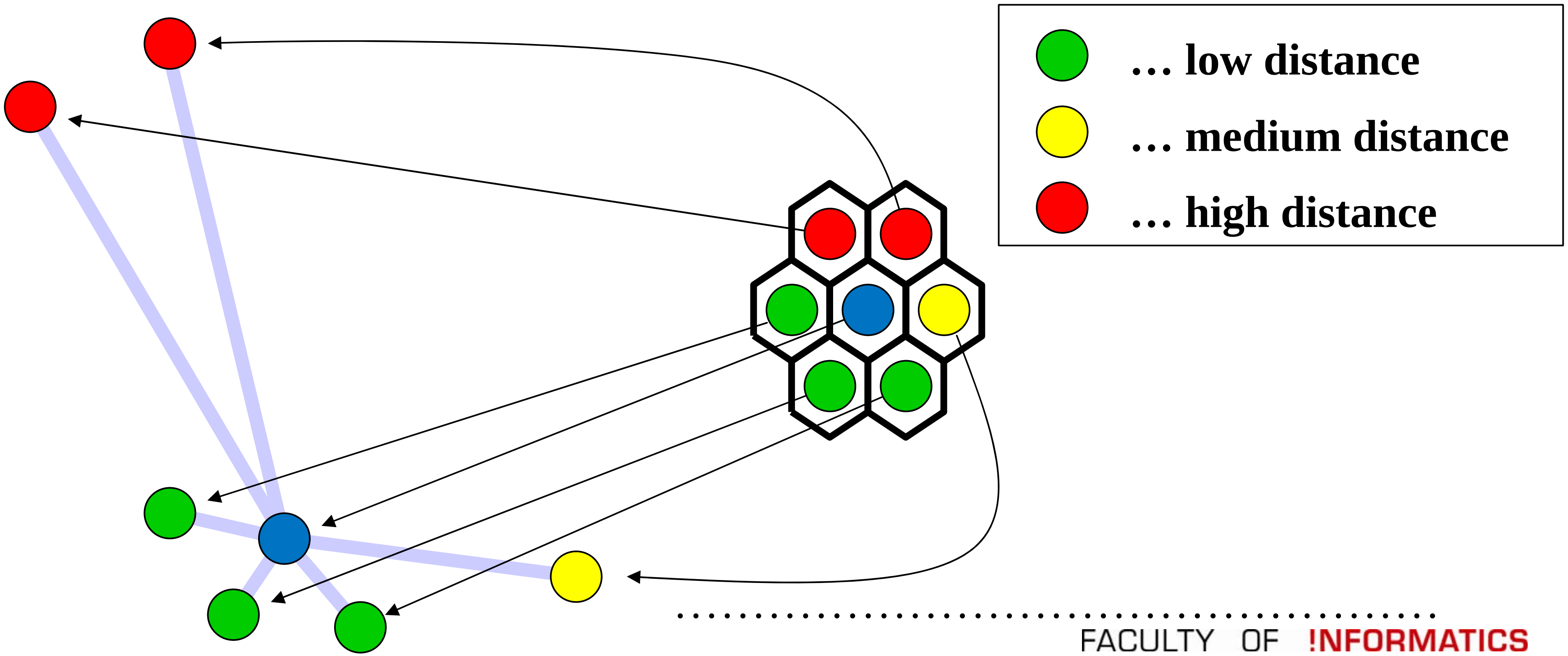
- which areas of the map are close to each other
- relationships of groups of attributes
- Georg Pölzlbauer, Andreas Rauber, and Michael Dittenbach. **A vector field visualization technique for self-organizing maps** In Tu Bao Ho, David Cheung, Huan Li, editors, Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05), pages 399-409, Hanoi, Vietnam, May 18-20 2005. Springer-Verlag.
- Georg Pölzlbauer, Michael Dittenbach, Andreas Rauber. **Advanced visualization of Self-Organizing Maps with vector fields**. Neural Networks, 19(6-7):911-922, July-August 2006

SOM Vector Fields

- similar to U-matrix, but for pairs of units
- based solely on units' weight vectors
- different levels of granularity (interactive analysis)
- optimized for engineering disciplines

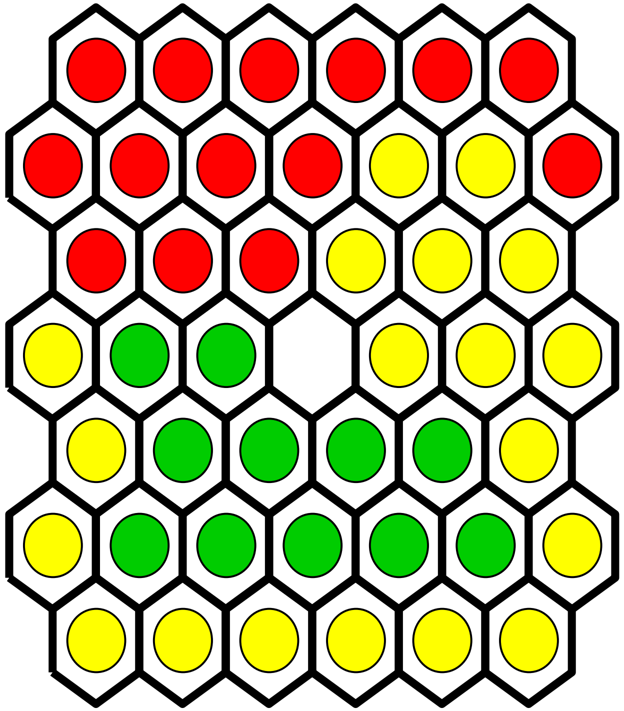
Vector Field Principles

- Vector field representation
- Vectors pointing to cluster centers
- Smoother version of U-Matrix



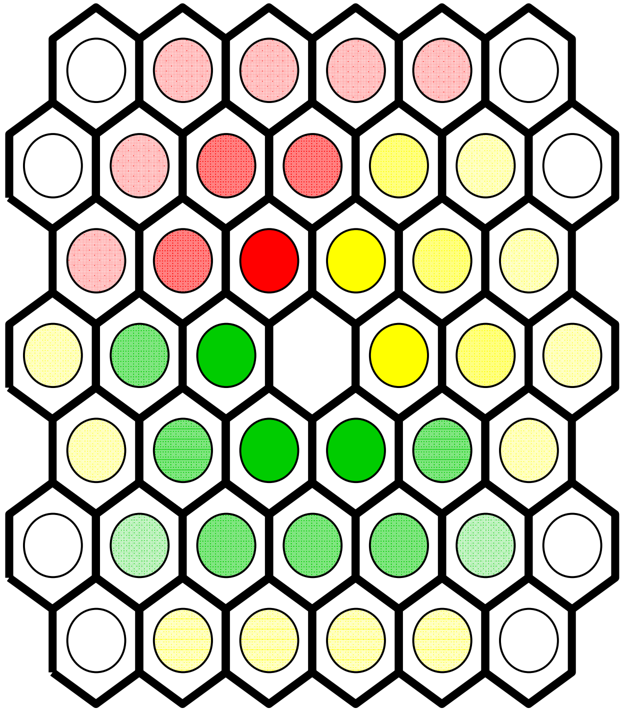
Flow Computation

1. calculate distances



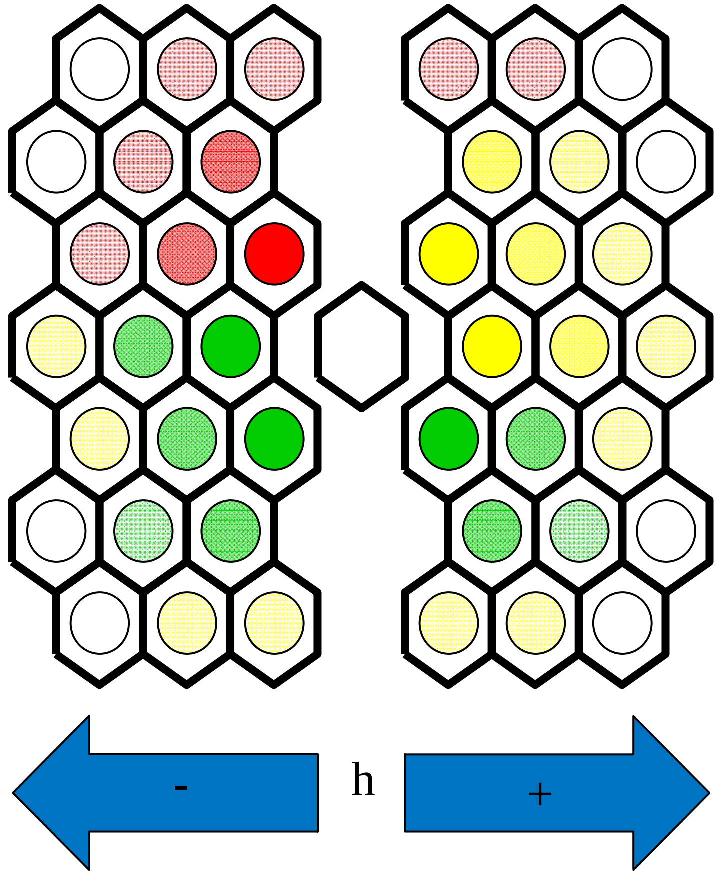
Flow Computation

1. calculate distances
2. weight with kernel function (different levels of granularity)



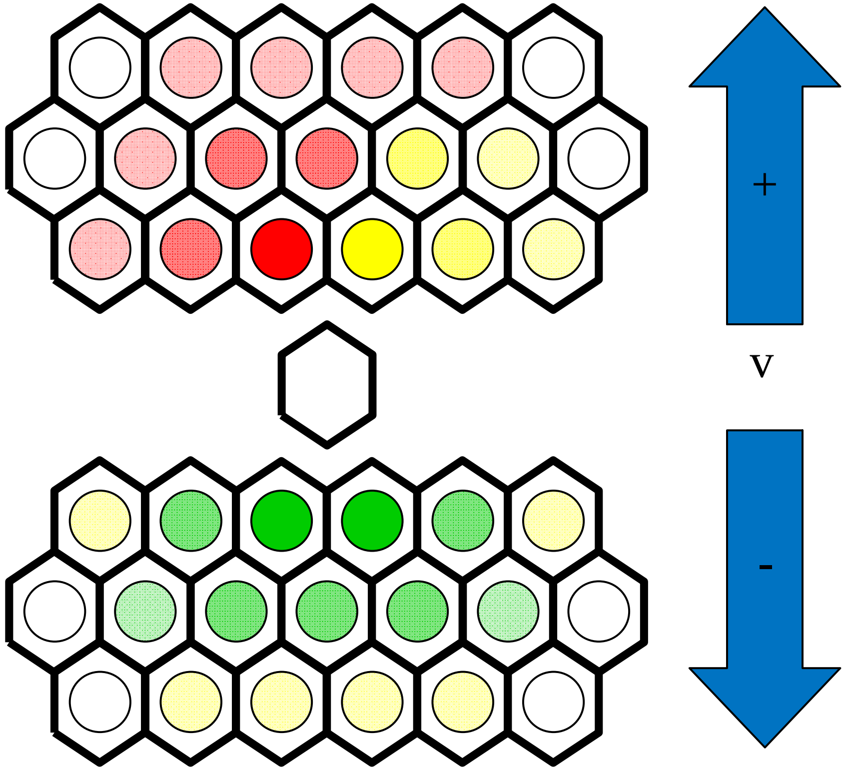
Flow Computation

1. calculate distances
2. weight with kernel function
3. divide into positive and negative h/v directions



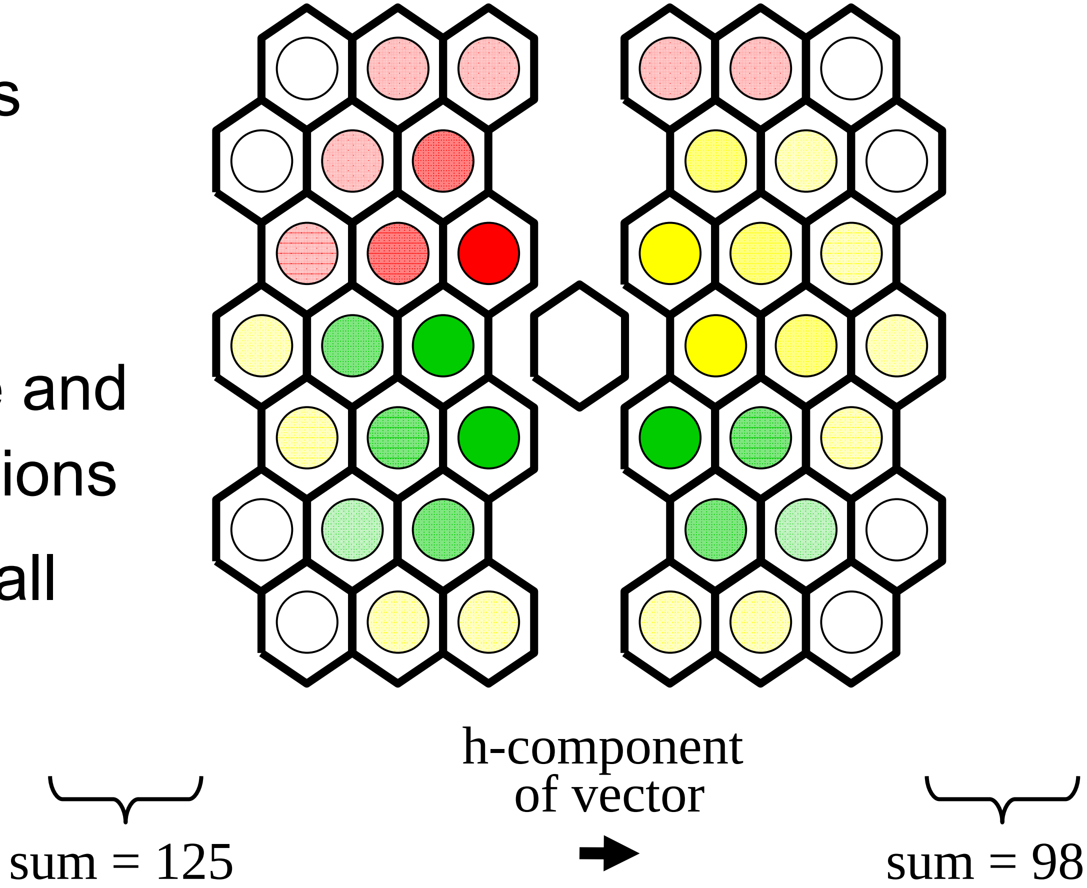
Flow Computation

1. calculate distances
2. weight with kernel function
3. divide into positive and negative h/v directions



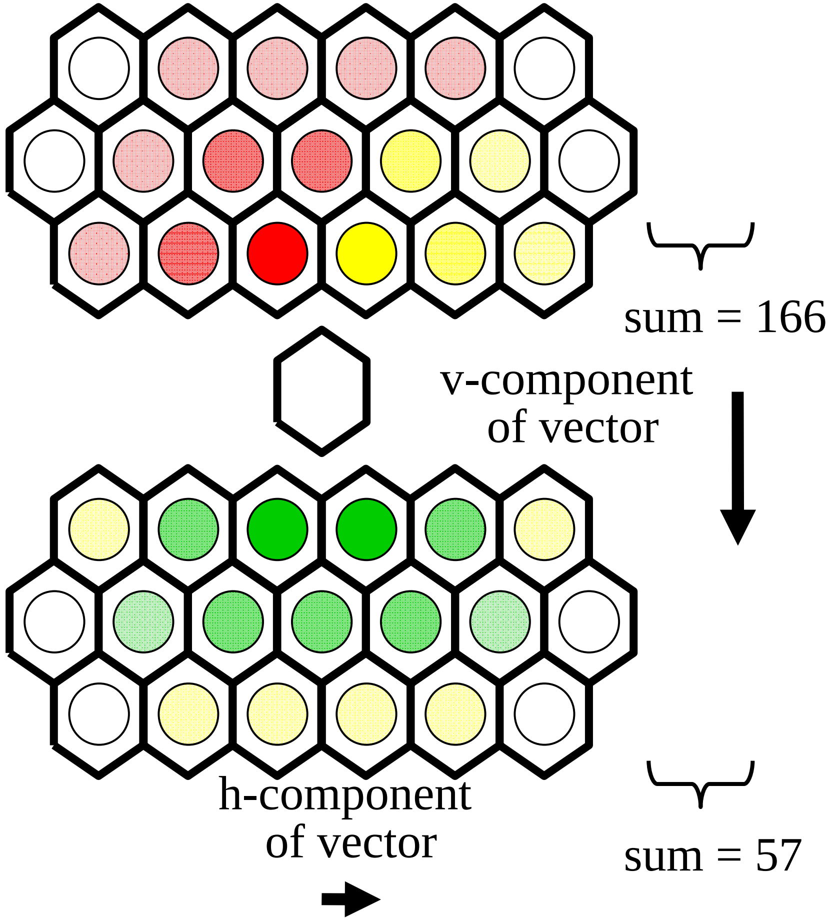
Flow Computation

1. calculate distances
2. weight with kernel function
3. divide into positive and negative h/v directions
4. calculate sums of all contributions



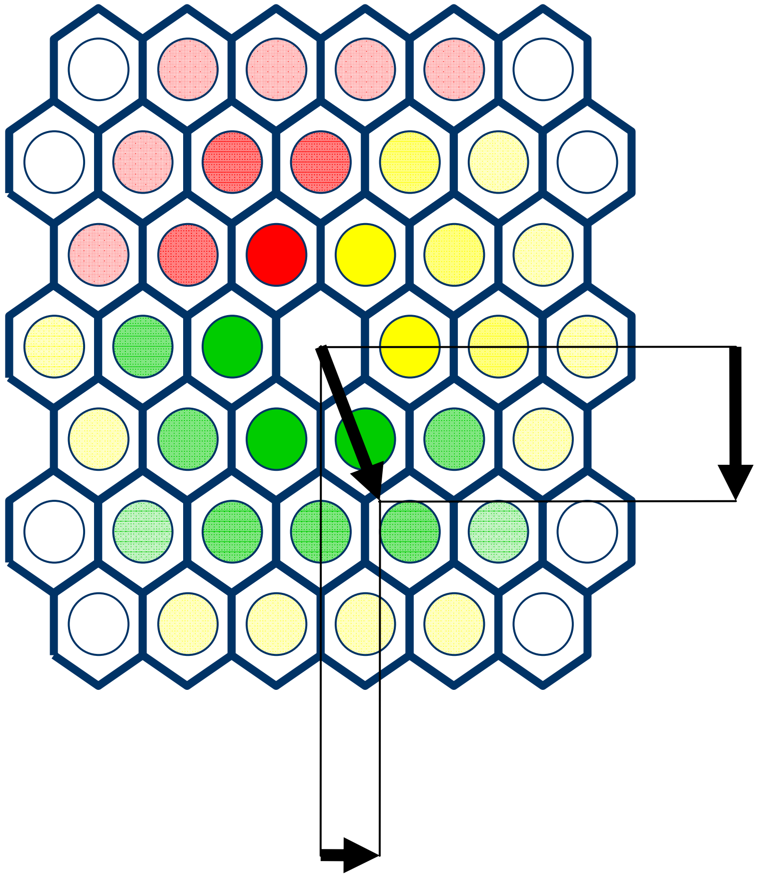
Flow Computation

1. calculate distances
2. weight with kernel function
3. divide into positive and negative h/v directions
4. calculate sums of all contributions



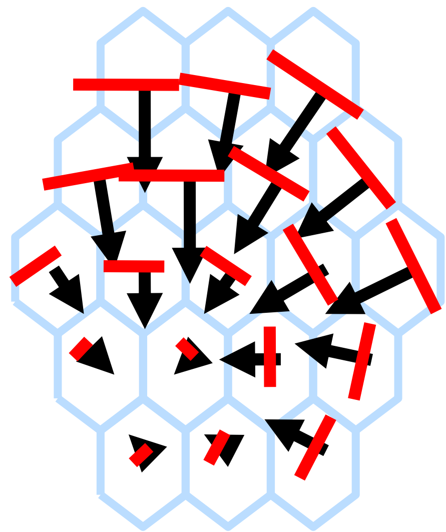
Flow Computation

1. calculate distances
2. weight with kernel function
3. divide into positive and negative h/v directions
4. calculate sums of all contributions
5. aggregate h/v components



Flow vs. Borderlines

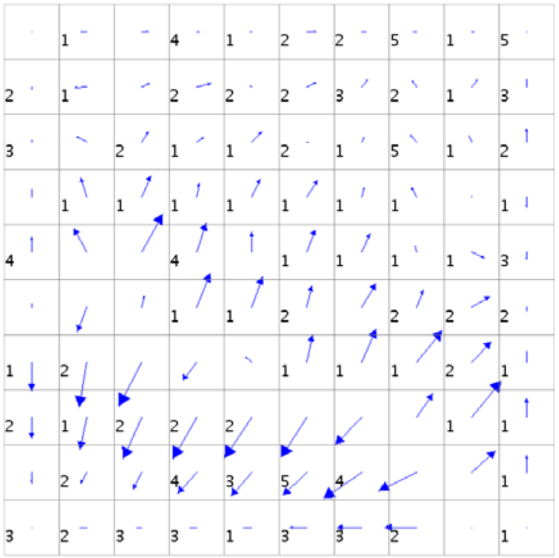
- Flow
  - arrows point to cluster centers
  - length shows intensity



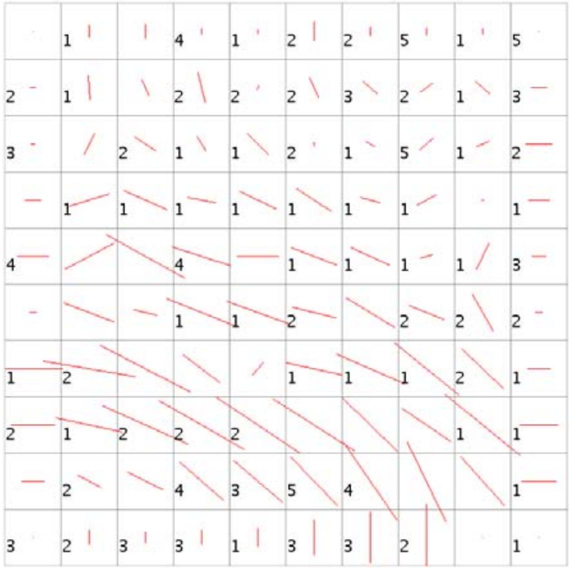
- Borderlines
  - dual representation
  - rotate arrows by 90 degrees
  - show cluster boundaries

Flow/Borderlines

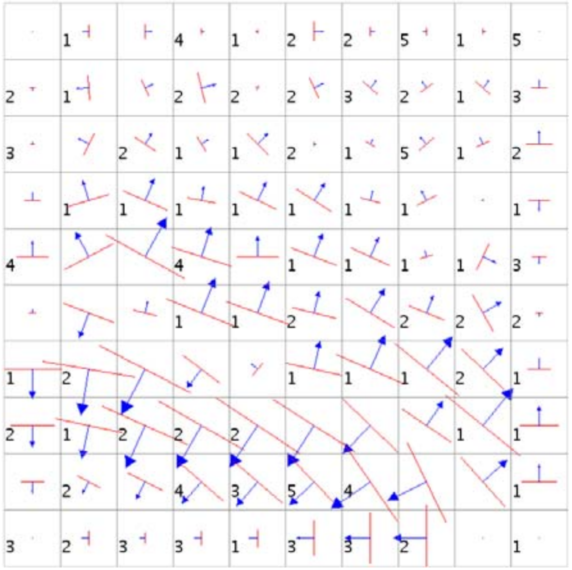
- Iris Dataset:



Flow

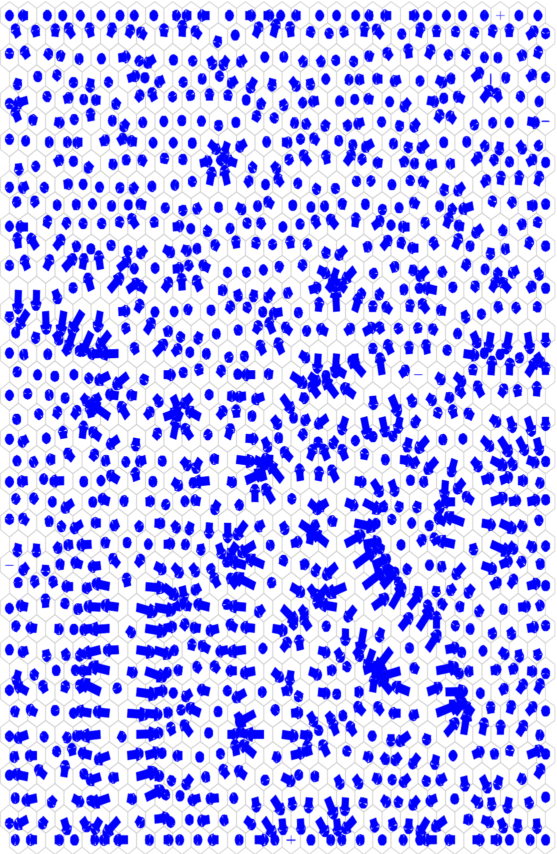


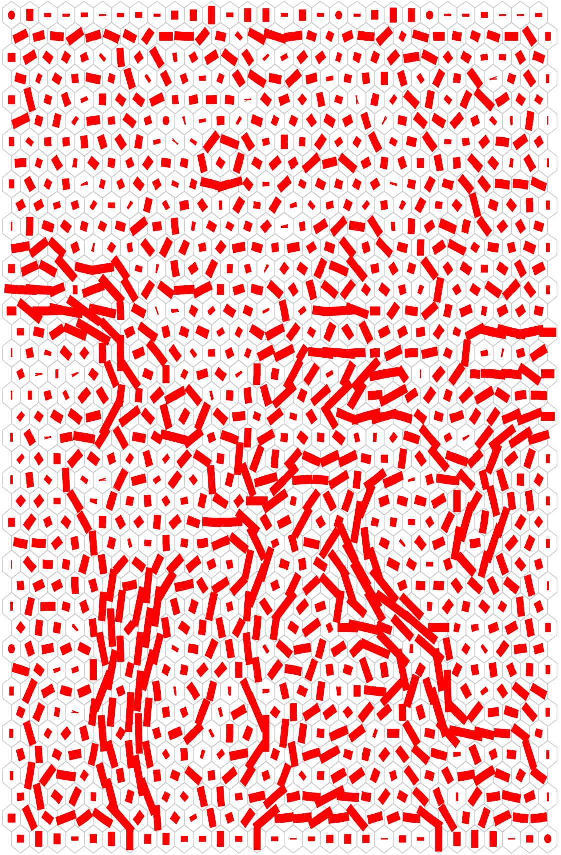
Borderlines



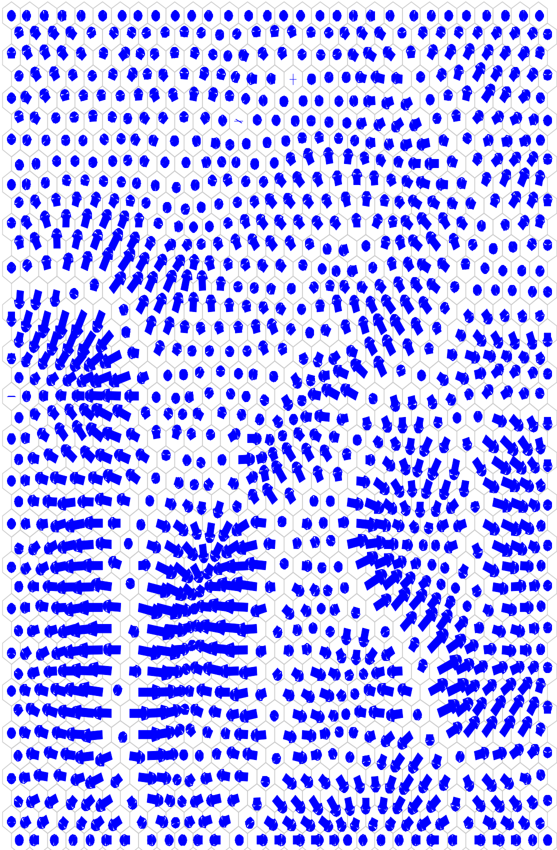
Flow+Borderlines

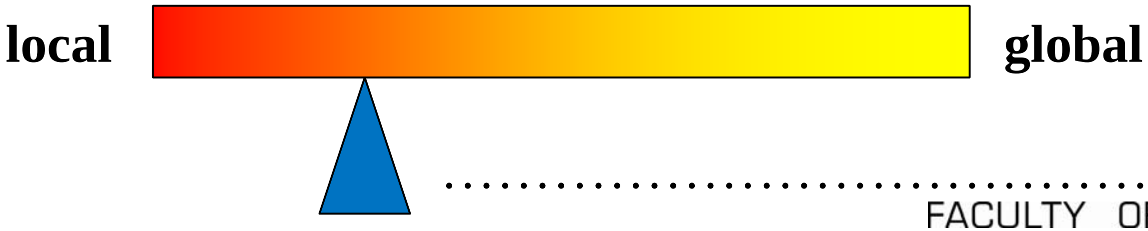
Flow/Borderline: sigma = 1



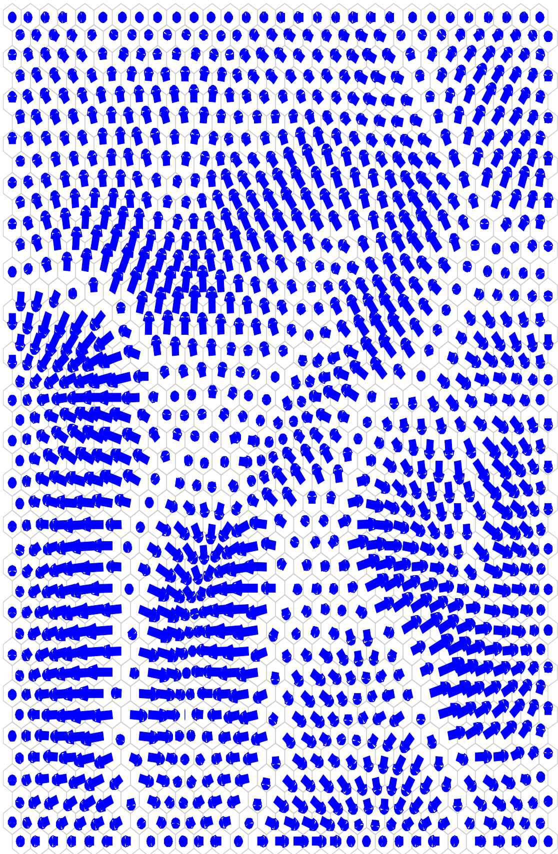


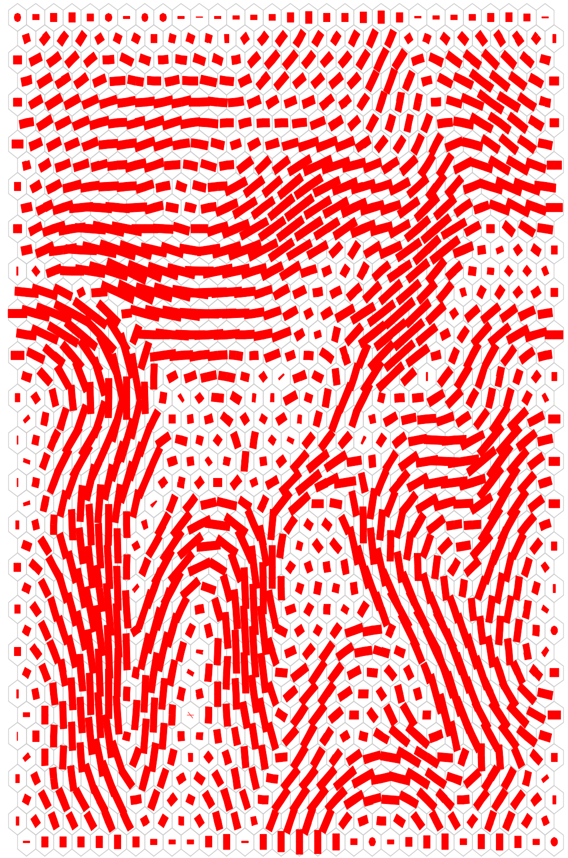
Flow/Borderline: sigma = 3

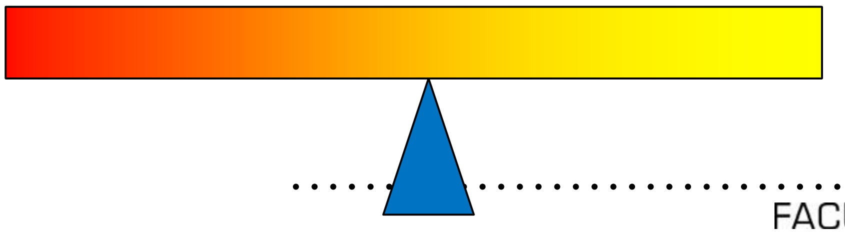




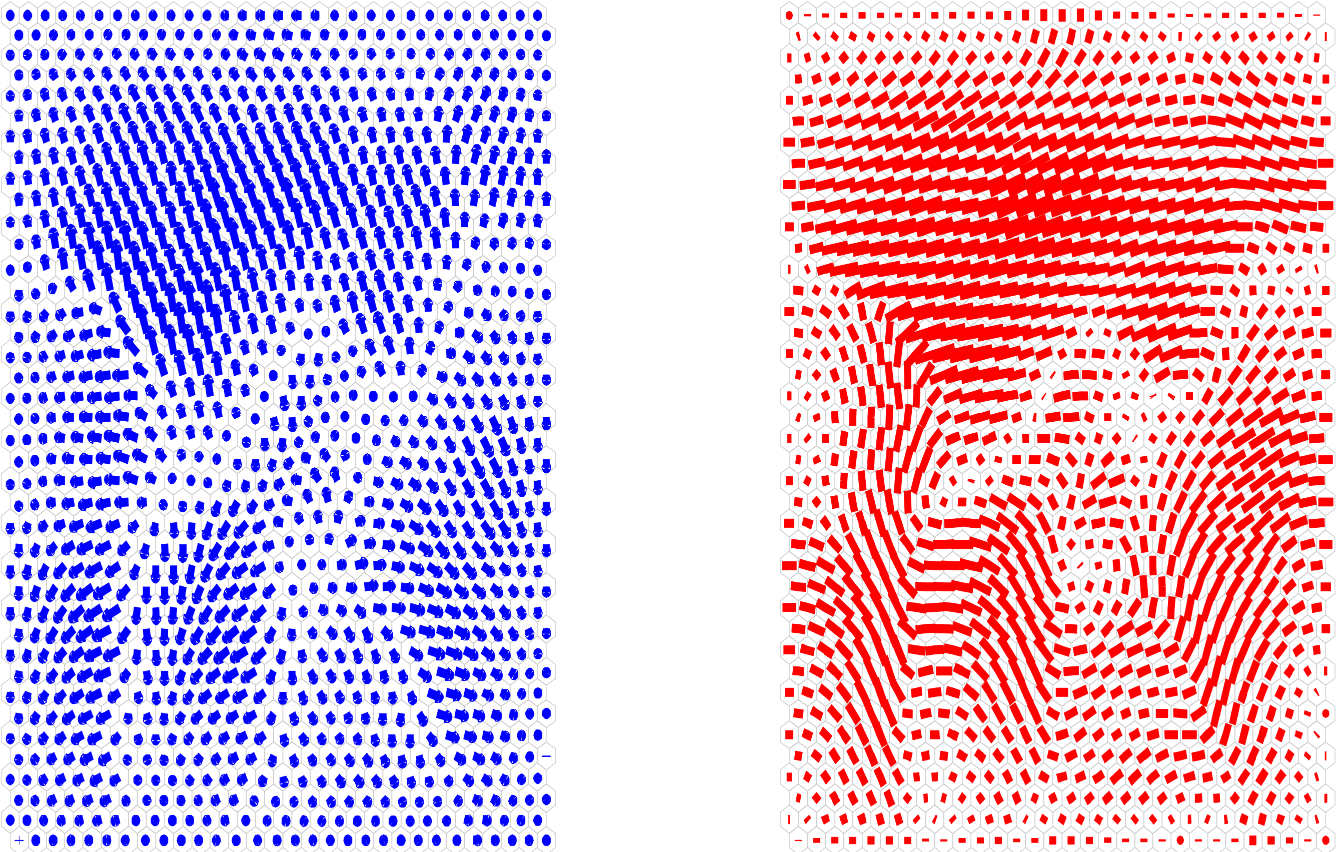
Flow/Borderline: sigma = 5

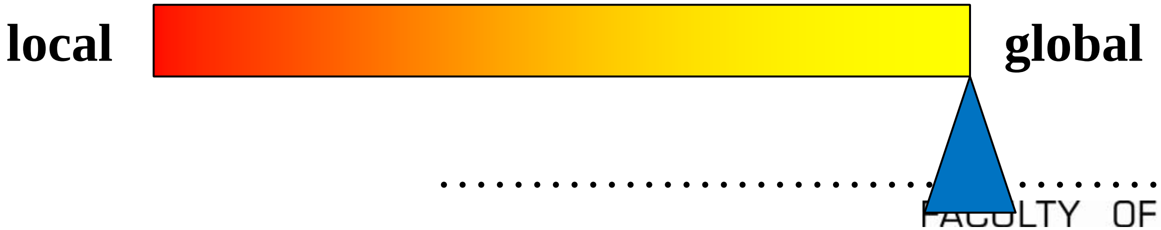




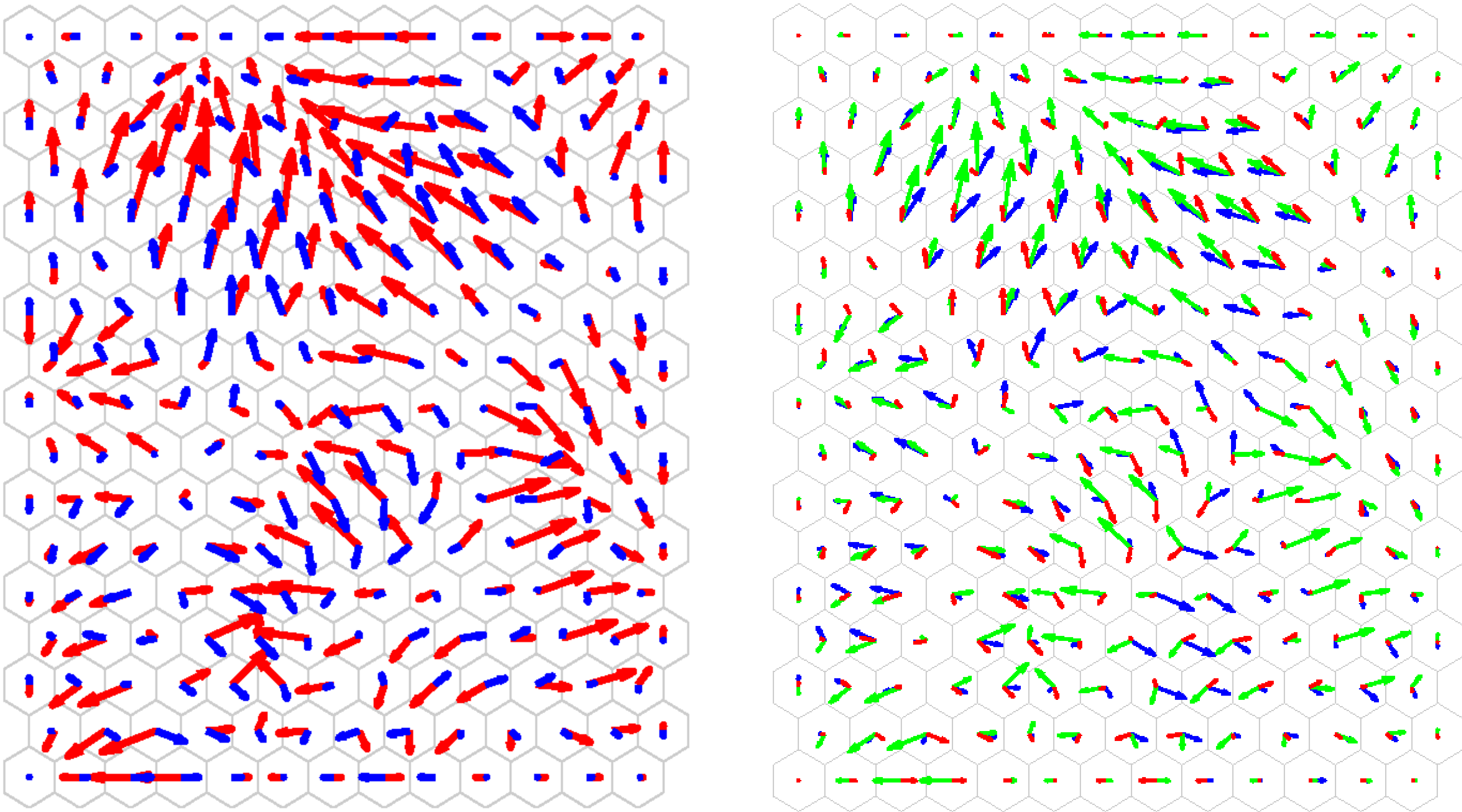
local  **global**

Flow/Borderline: sigma = 15





Flow: Groups of Attributes



Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
- Distanzen
  - Minimum Spanning Trees
  - Cluster Connections (CC)
  - D-Matrix
  - U-Matrix
  - U* Matrix: U-Matrix + P-Matrix
  - Vectorfields: Flow / Borderline
- Klasseninformation
- Attribute
- Clustering der SOM

Outline

Visualisierungen der SOM

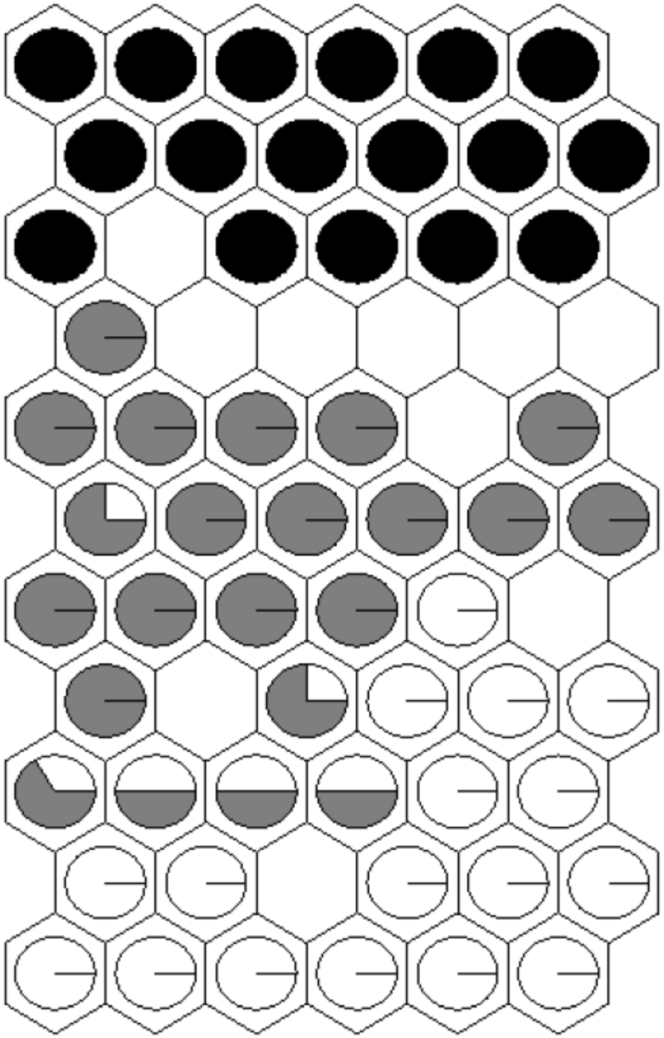
- Textuelle Informationen
- Dichte
- Distanzen
- Klasseninformation
  - Pie Charts / Patches
  - Class Coloring:
    - Chessboard
    - Color Filling with Attractor
- Attribute
- Clustering der SOM

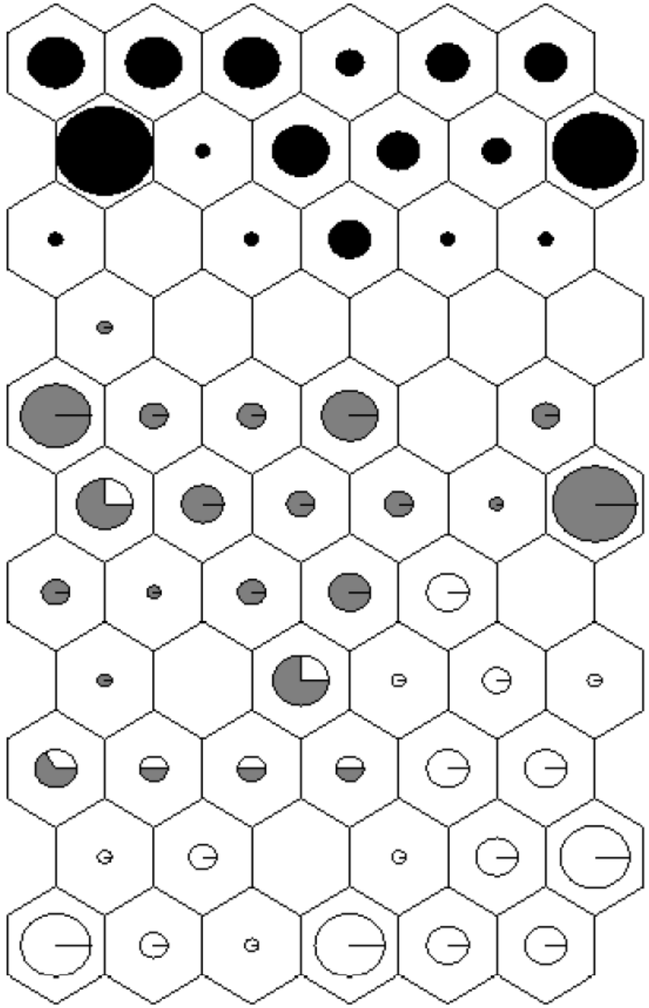
Klassenverteilung

- Klasseninformation: Verteilung der Klassen auf die Knoten
- Standard-Ansätze:
  - diskrete Behandlung der Knoten
  - Pie Charts: mit Größenänderung / ohne Größenänderung
  - überlappende Patches
- Class-coloring, Chessboard-Visualisierung
  - Einfärben der gesamten Karte
  - bessere Visuelle Eindrücke der Übergänge und Durchmischungen

Klassenverteilung

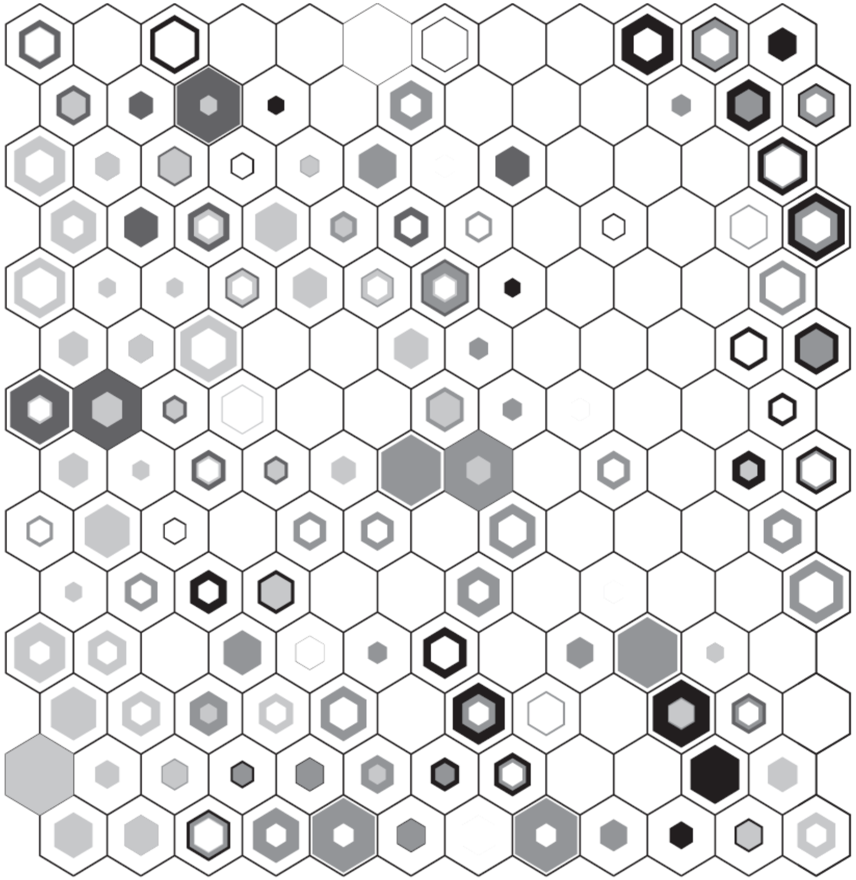
- Pie Charts: ohne / mit Hithistogramm

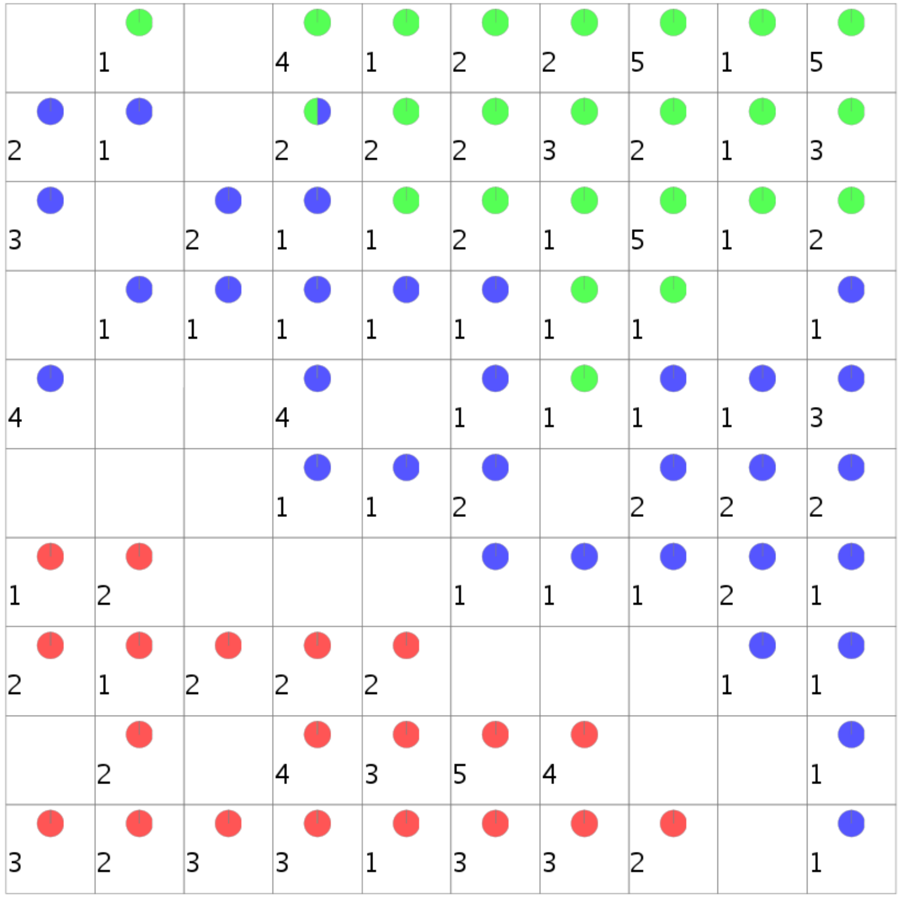




Klassenverteilung

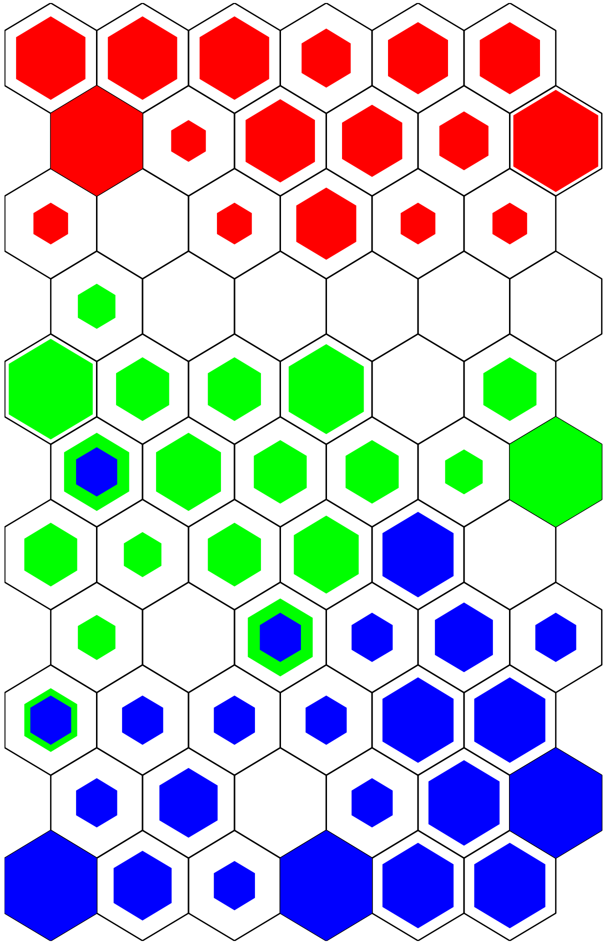
- Patches mit Hit Histogramm, Pie-Charts



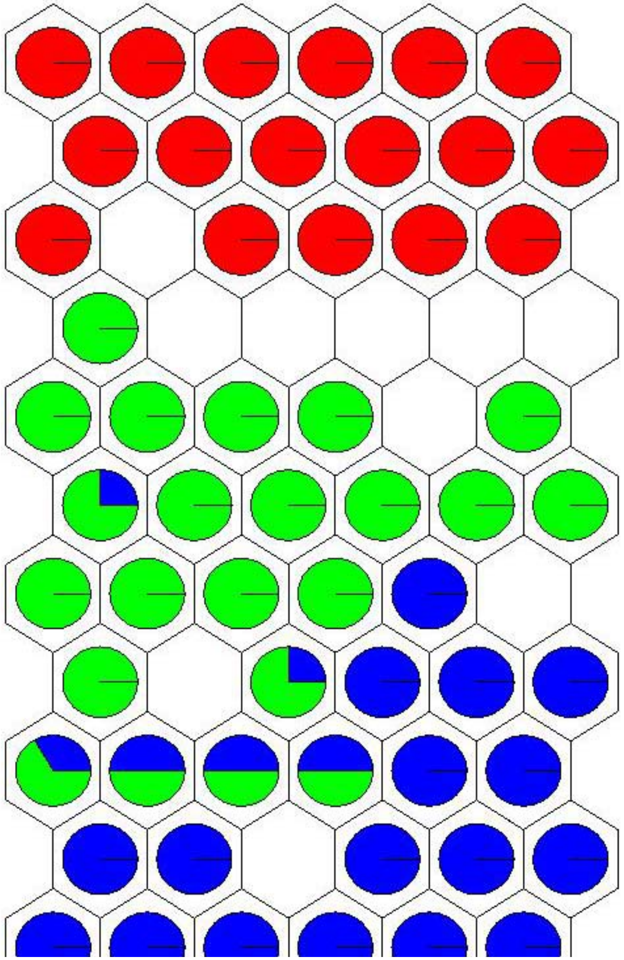


Klassenverteilung

- Patches mit Hit Histogramm, Pie-Charts



Iris (klein),
3 Klassen

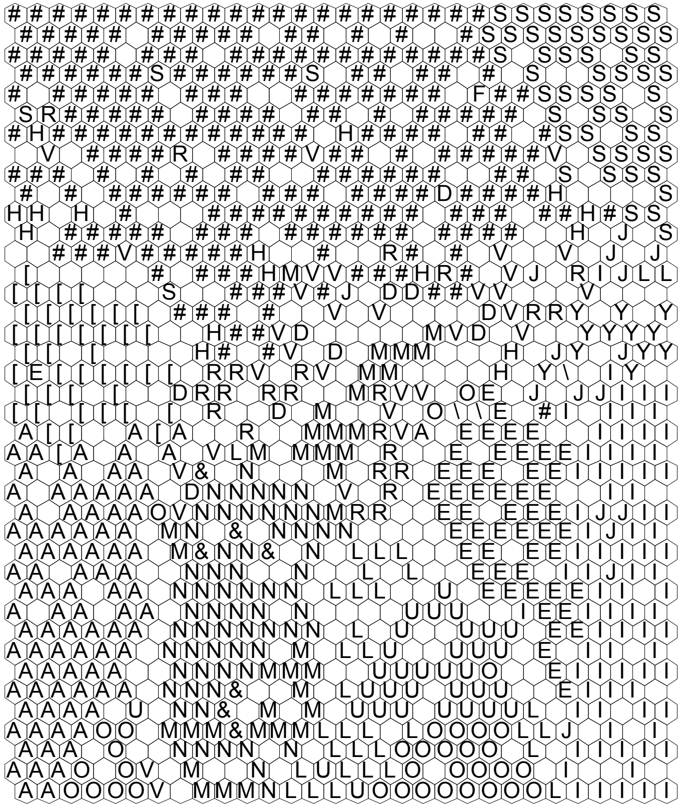


Iris (klein),
3 Klassen, Pie-Charts

Klassenverteilung

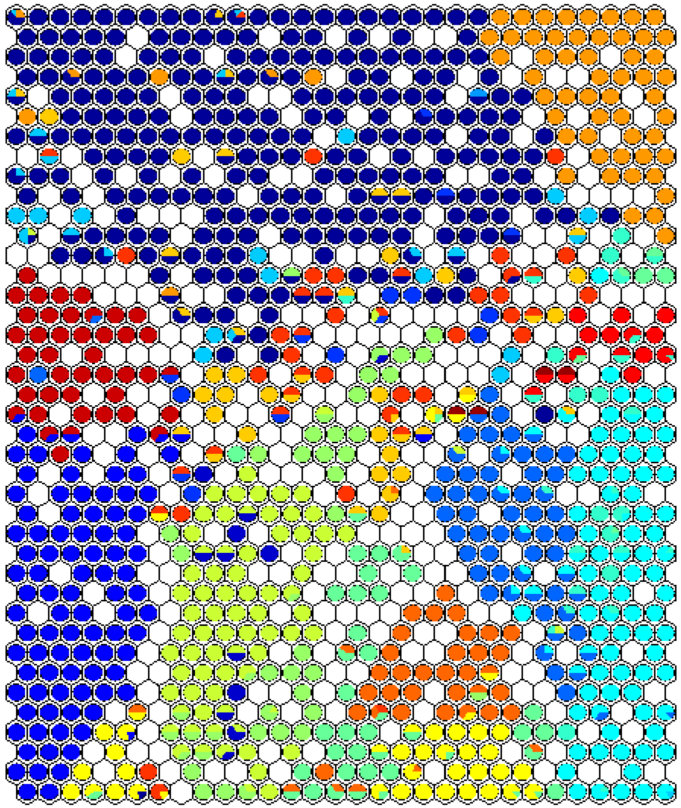
- Textuelle Class-IDs, Pie-Charts

Labels (only most frequent label)



class IDs

Pie Chart of Categoriy (unscaled)



pie charts

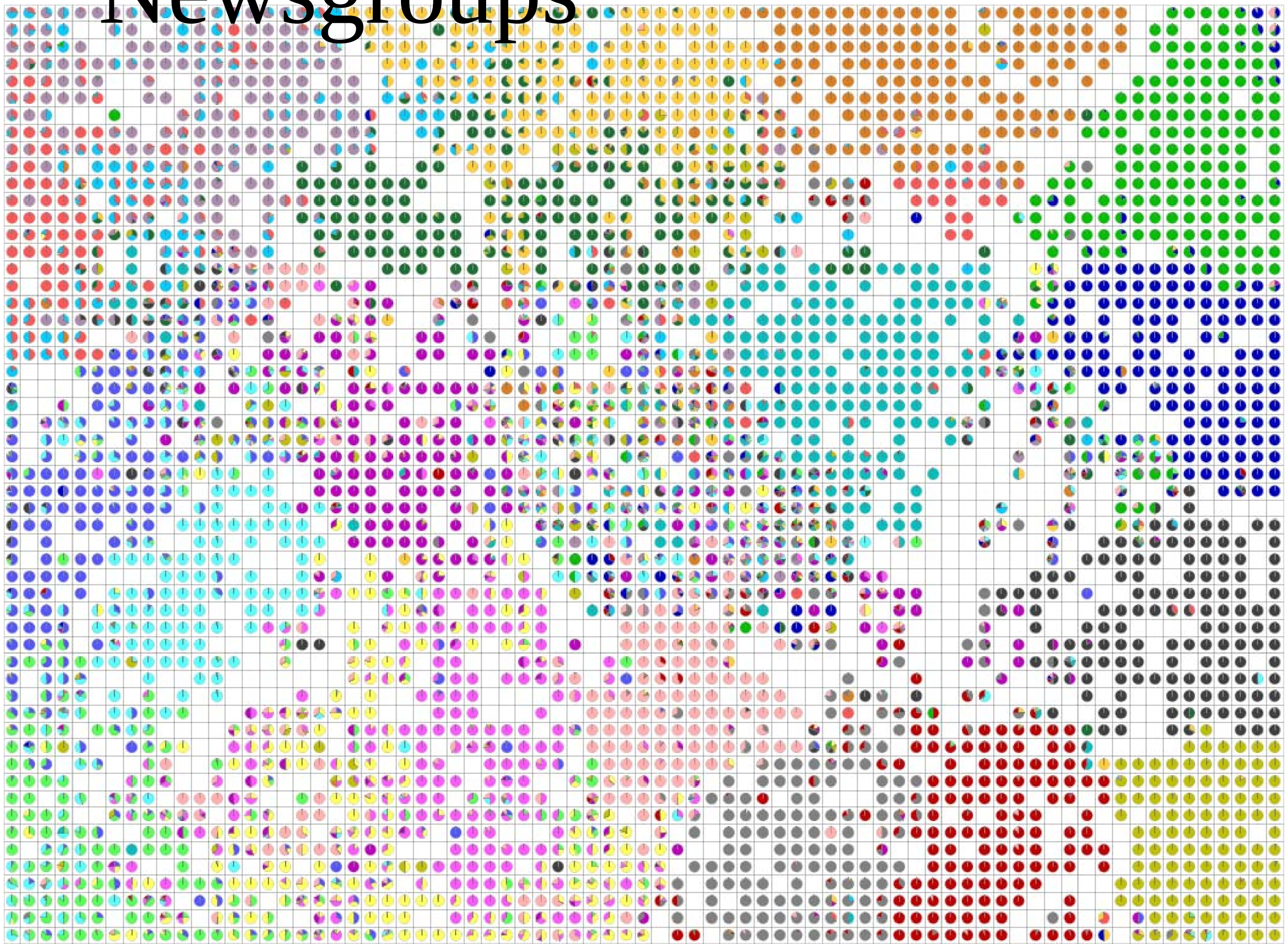
Klassenverteilung

- 20-Newsgroups Benchmark Datenset
- Je 1000 Postings pro newsgroup
- Hierarchie von Newsgroups

- Full-term indexing
- Wörter auf Wortstämme reduziert

alt.atheism	
comp.graphics	
comp.os.ms-windows.misc	
comp.sys.ibm.pc.hardware	
comp.sys.mac.hardware	
comp.windows.x	
misc.forsale	
rec.autos	
rec.motorcycles	
rec.sport.baseball	
rec.sport.hockey	
sci.crypt	
sci.electronics	
sci.med	
sci.space	
soc.religion.christian	
talk.politics.guns	
talk.politics.mideast	
talk.politics.misc	
talk.religion.misc	

Klassenverteilung: 20 Newsgroups



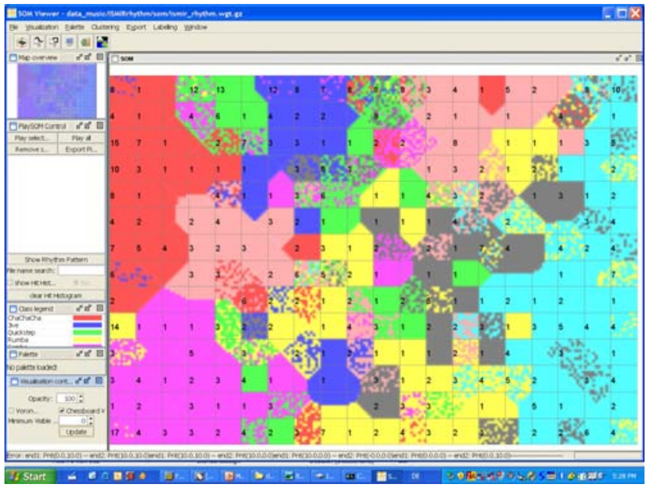
Outline

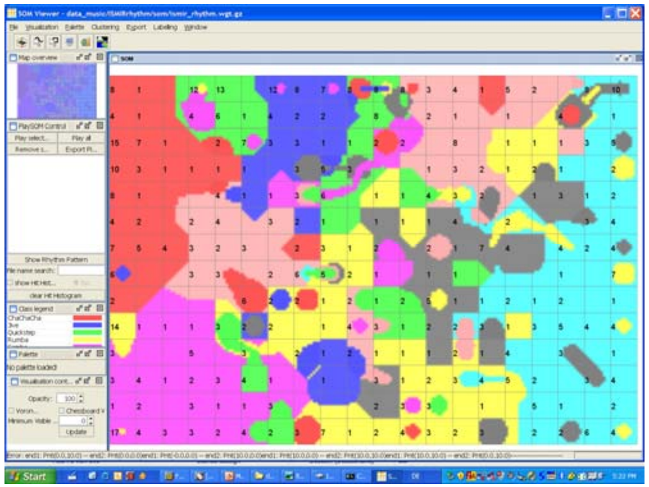
Visualisierungen der SOM

- Textuelle Informationen
- Dichte
- Distanzen
- Klasseninformation
  - Pie Charts / Patches
  - Class Coloring:
    - Chessboard
    - Color Filling with Attractor
- Attribute
- Clustering der SOM

Klassenverteilung: Class Coloring

- Color SOM with class information
- Similar to pie chart representation
- 2 visualization types:
  - chessboard visualization
  - color flooding with attractor





Klassenverteilung: Class Coloring

- Taha Abdel-Aziz: **Coloring of the SOM based on Class Labels**. Master Thesis, Department of Software Technology and Interactive Systems, Vienna University of Technology, October 2006.
- Rudolf Mayer, Taha Abdel Aziz, and Andreas Rauber. **Visualising Class Distribution on Self-Organising Maps (accepted for publication)**.
  In Proceedings of the International Conference on Artificial Neural Networks (ICANN'07), Porto, Portugal, September 9 - 13 2007. Springer Verlag.

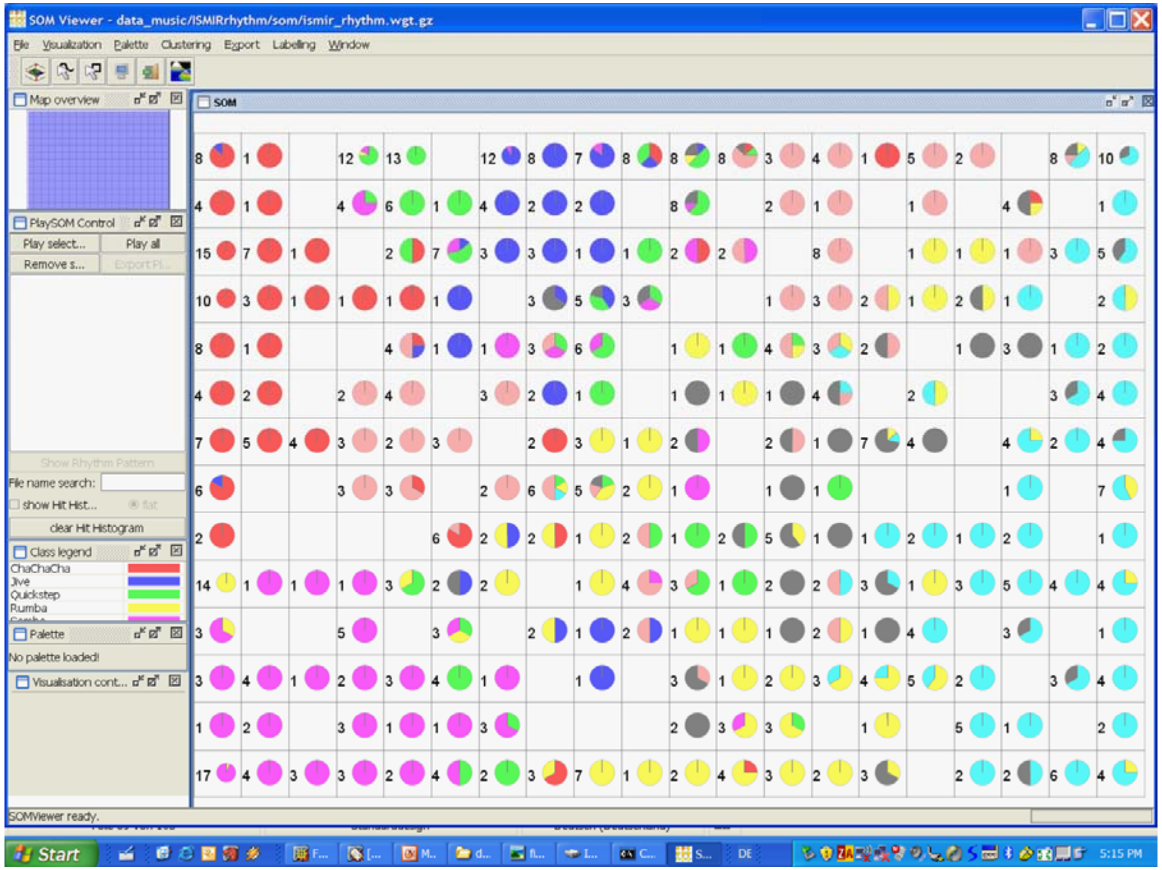
Class Coloring: Attractor Flooding

- Chessboard Visualization:
  - Voronoi Tesselation
  - color voronoi cells with patches accoridng to percentual share of class
  - opt.: set frequency threshold for small classes

Class Coloring: Chessboard

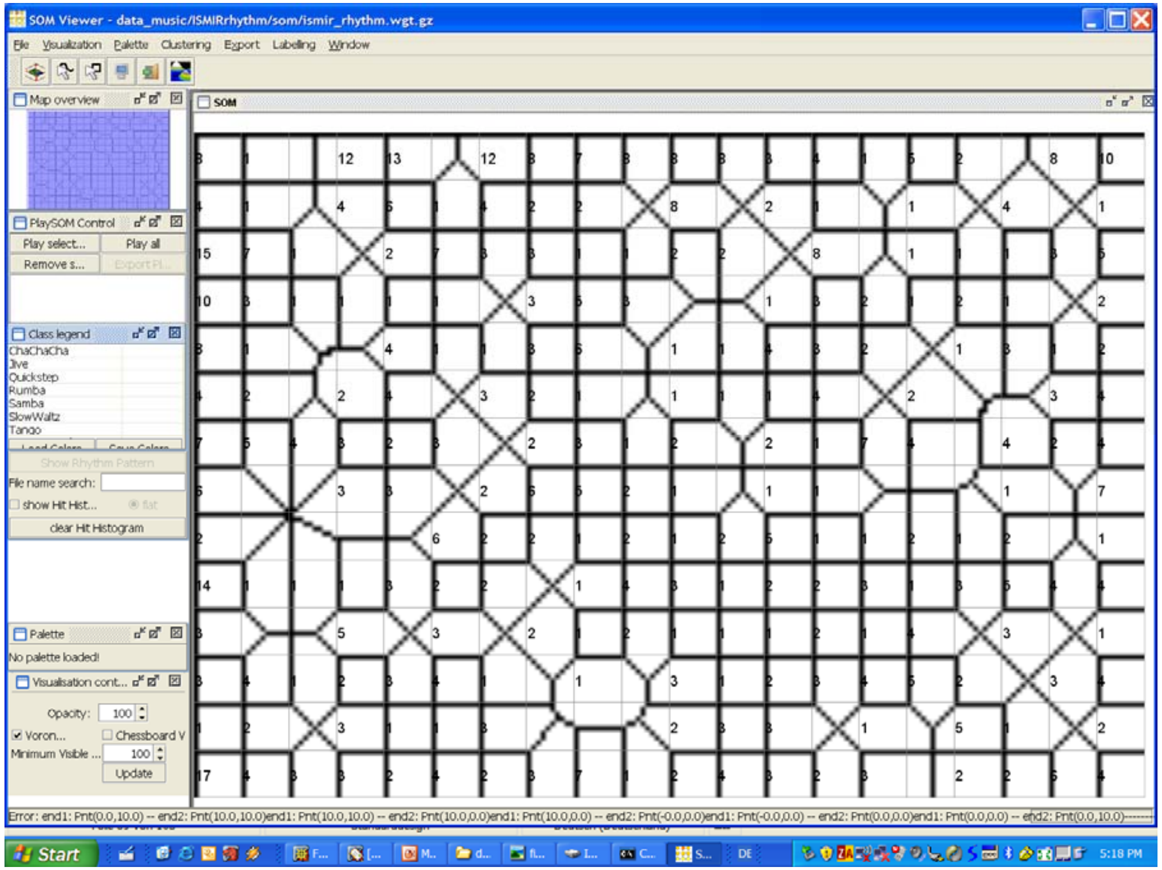
Chessboard:

- Step 1: Class information pie charts



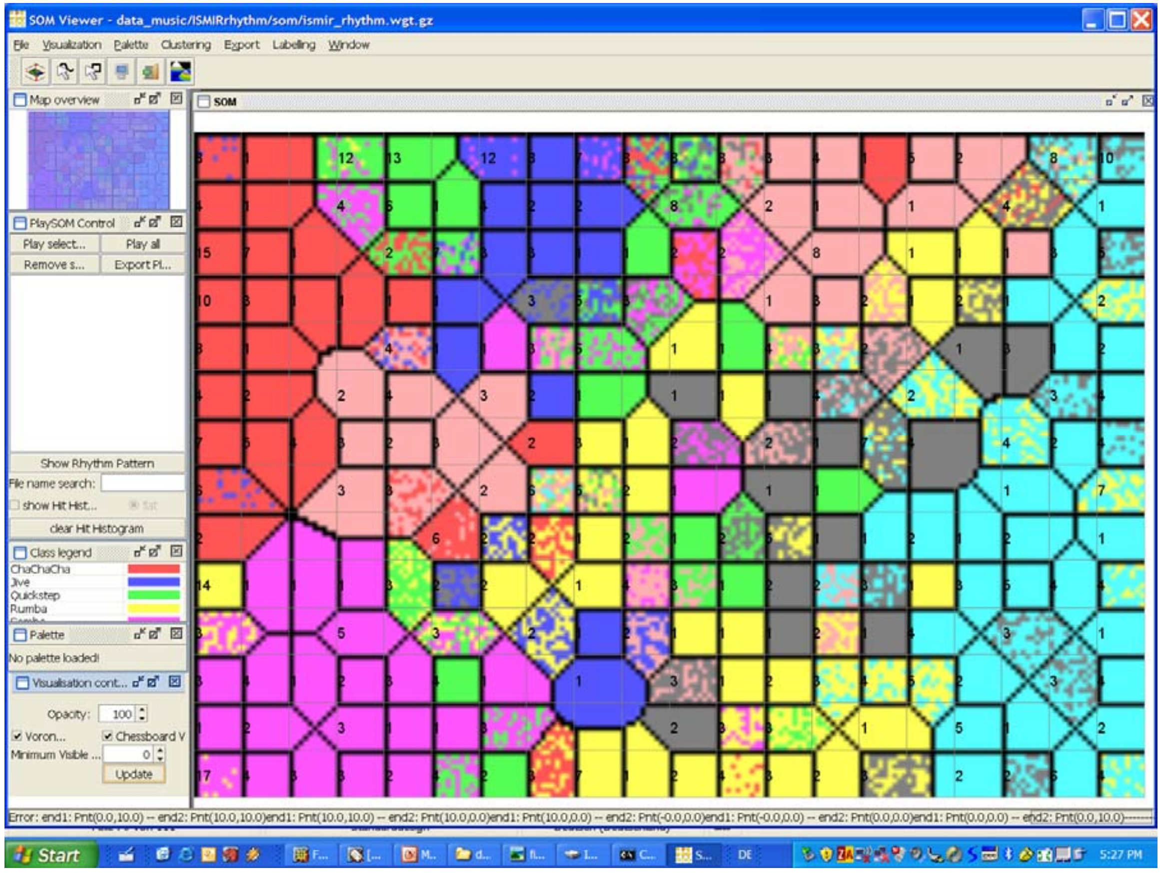
Class Coloring: Chessboard

- Step 2: Voronoi Tesselation



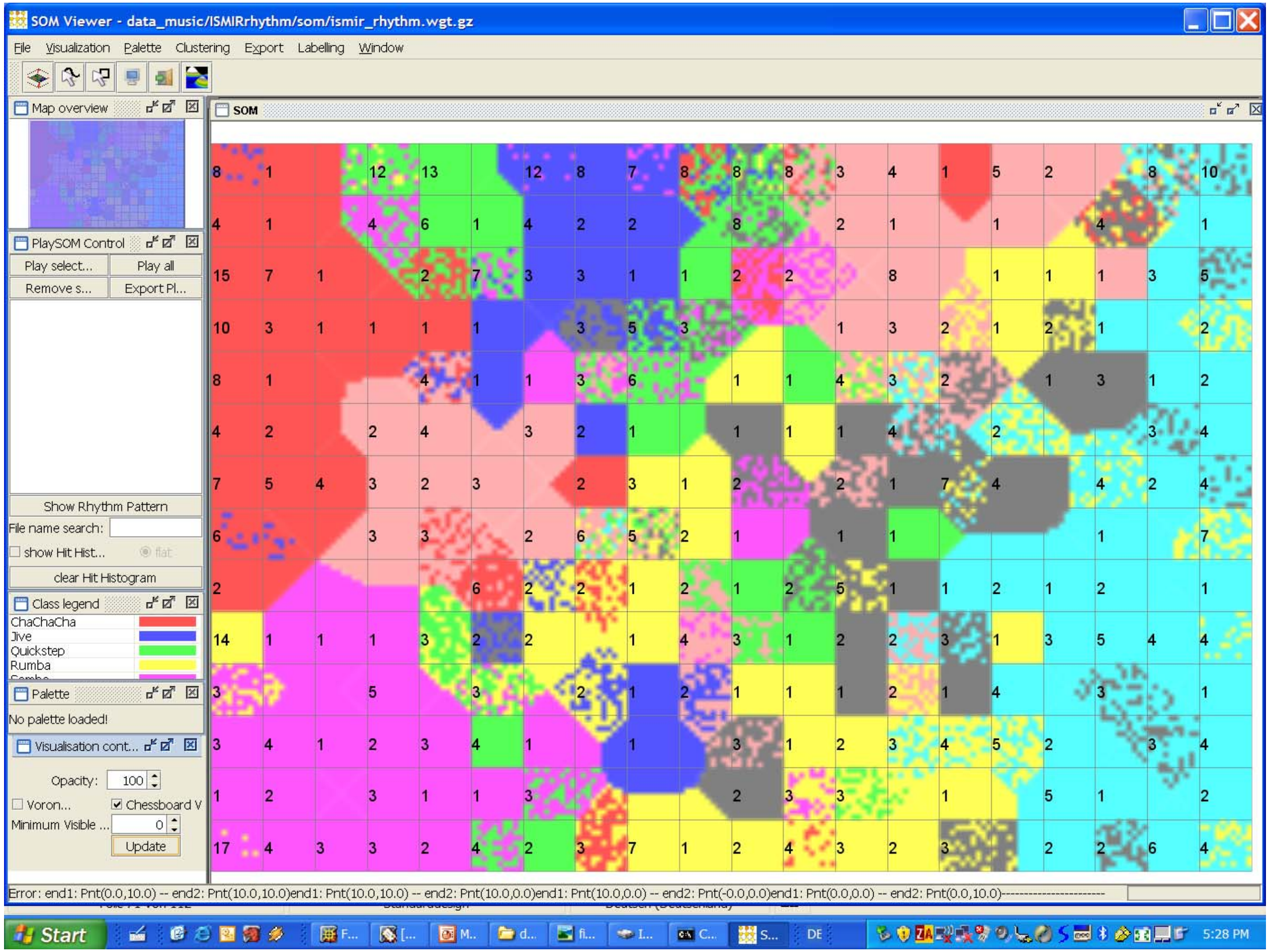
Class Coloring: Chessboard

- Step 3: Chessboard-style pixel coloring according to class frequency



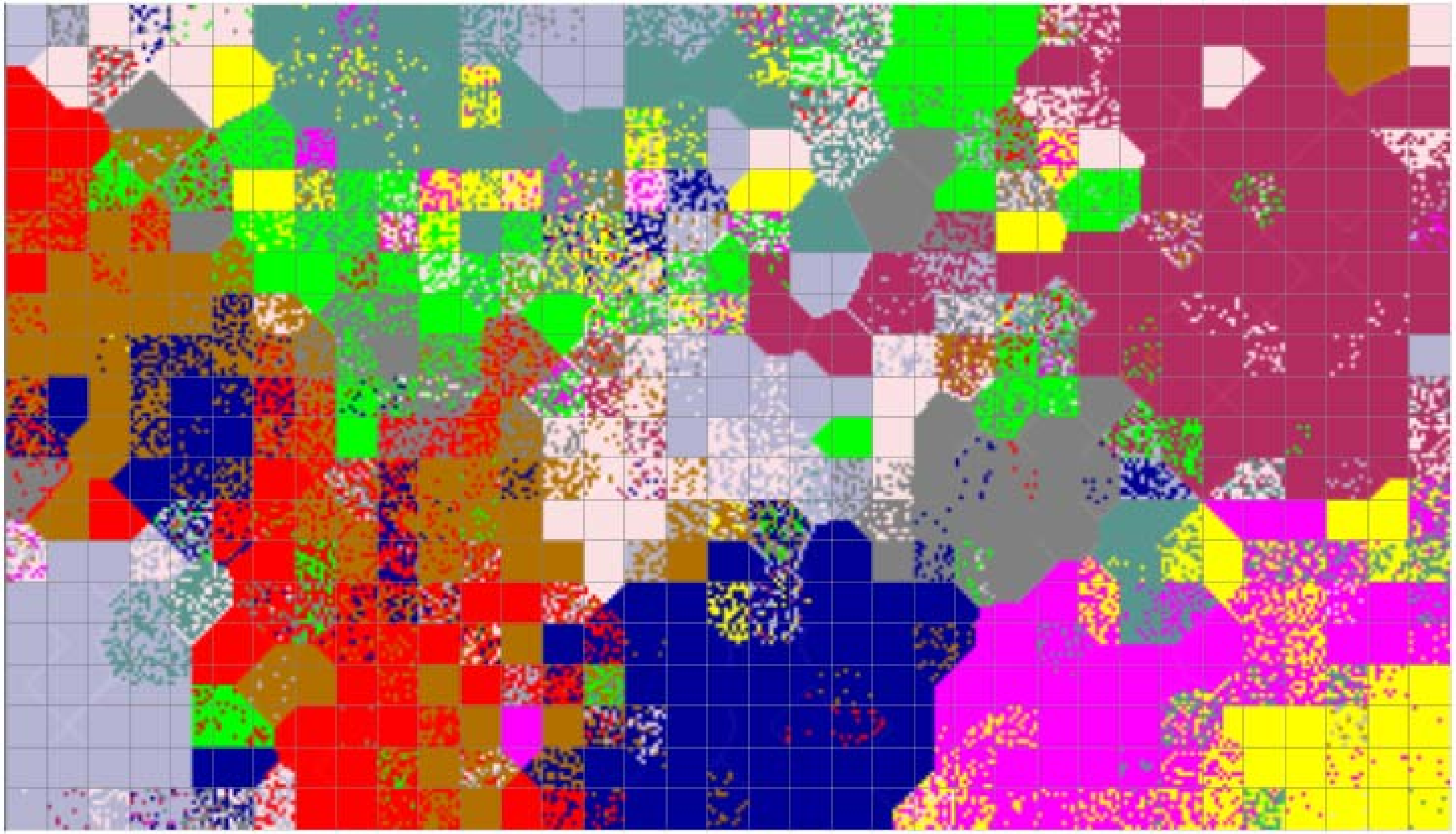
Class Coloring: Chessboard

- final:



Class Coloring: Chessboard

Chessboard visualization of Banksearch Data SOM



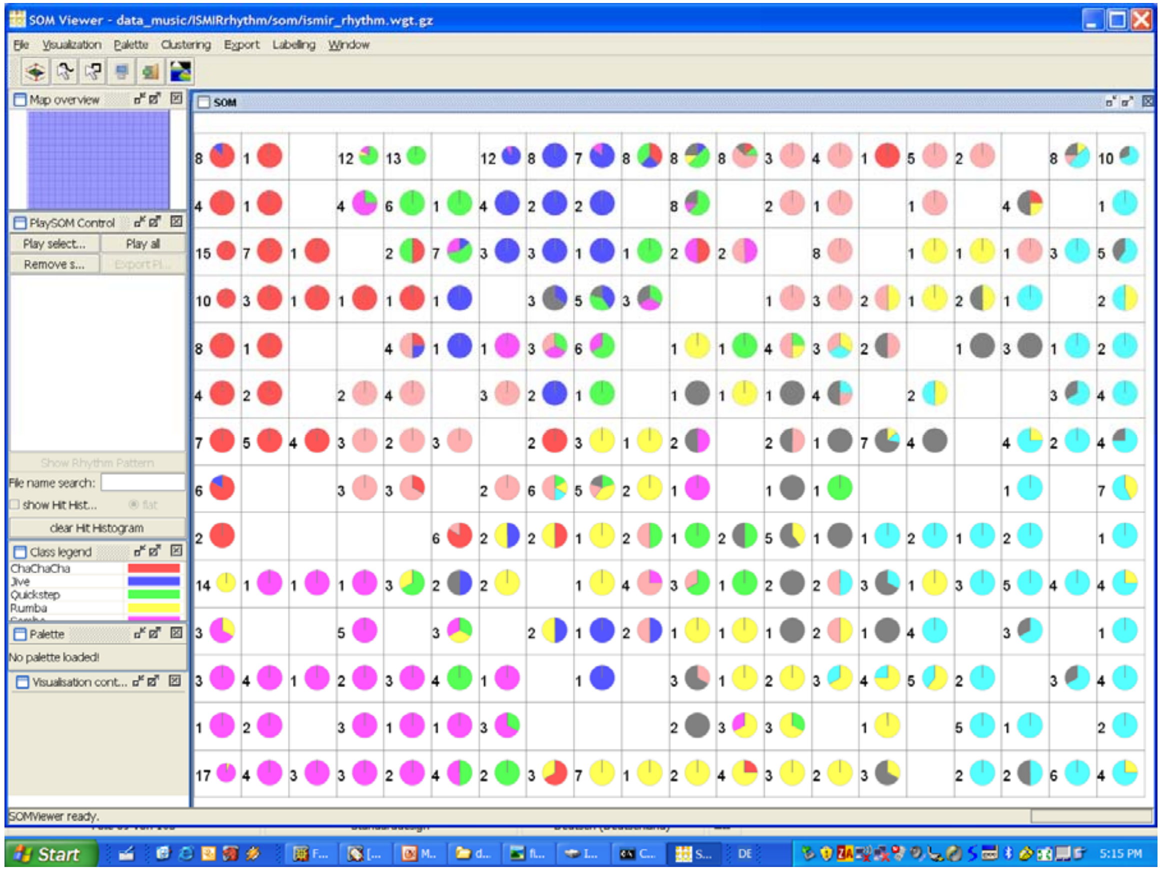
Class Coloring: Attractor Flooding

- Attractor Flooding:
  - Voronoi Tesselation
  - Fill with dominant class color
  - Identify neighboring class distributions
  - Identify attractors
  - Flood-fill style coloring along attractors according to frequency
  - opt.: set frequency threshold for small classes

Class Coloring: Attractor Flooding

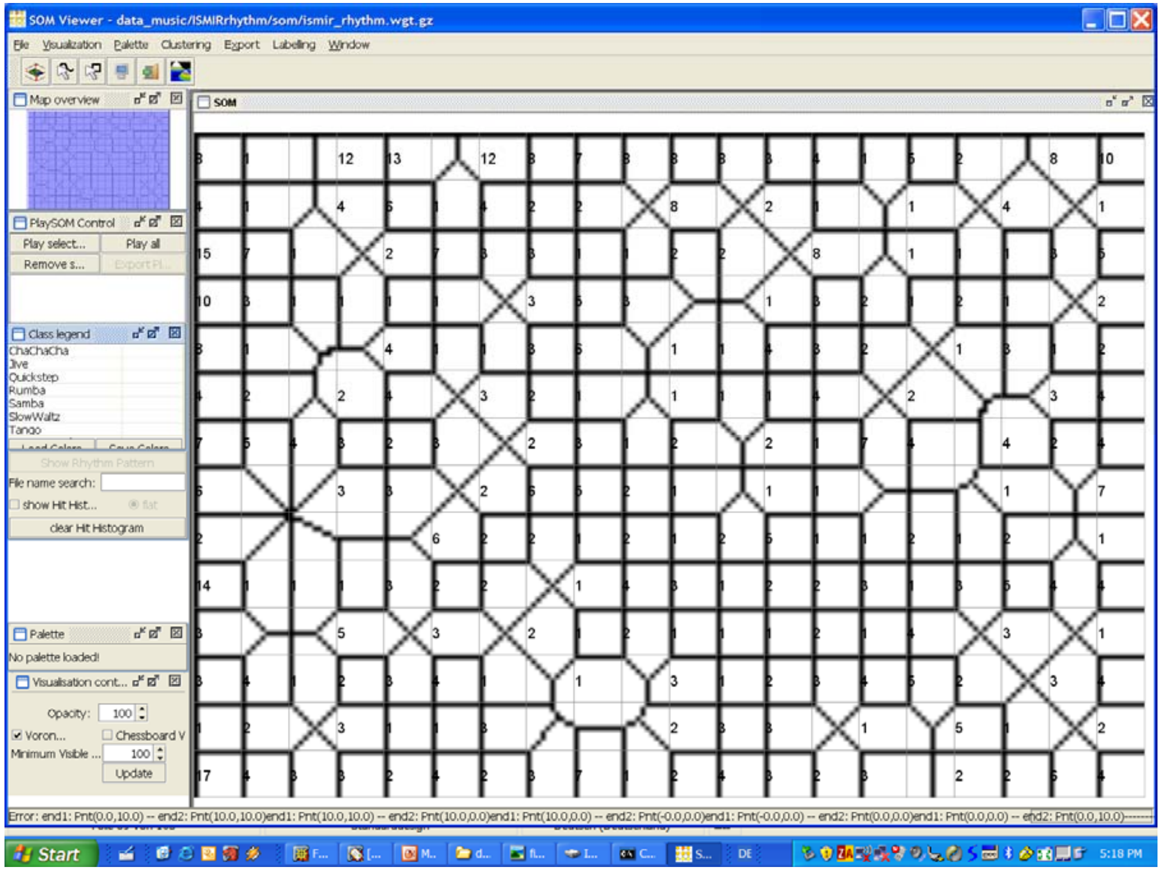
Attractor Flooding:

- Step 1: Class information pie charts



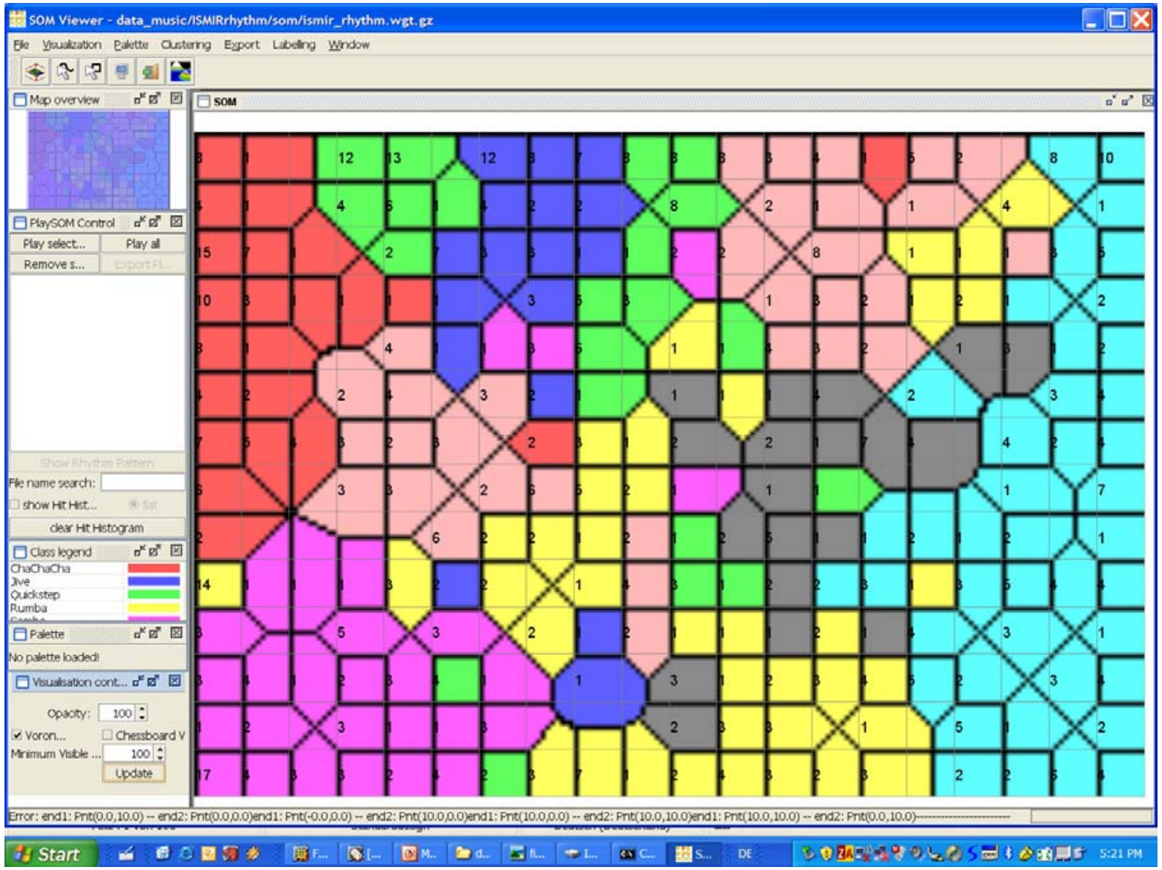
Class Coloring: Attractor Flooding

- Step 2: Voronoi Tesselation



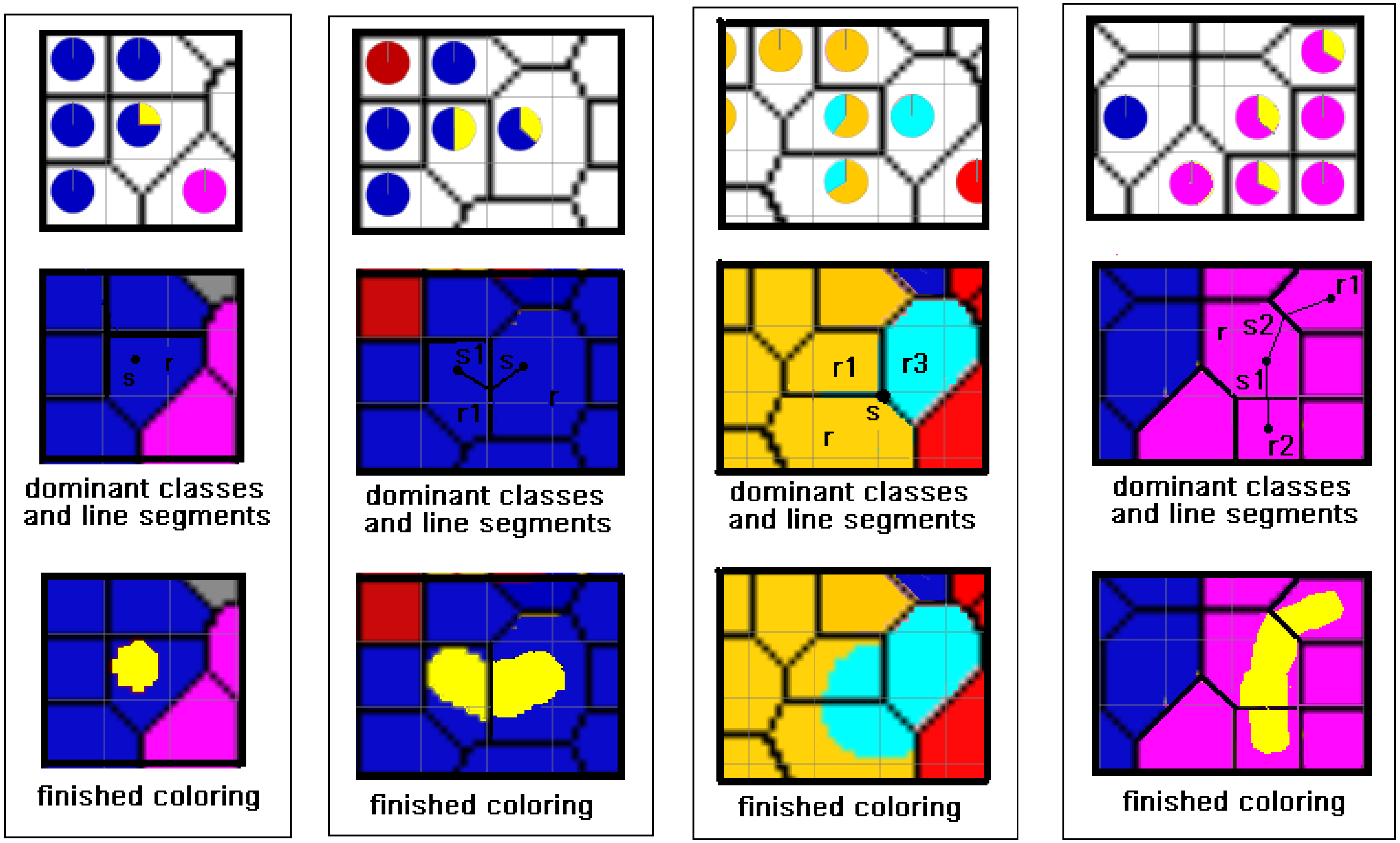
Class Coloring: Attractor Flooding

- Step 3: fill with dominant class color



Class Coloring: Attractor Flooding

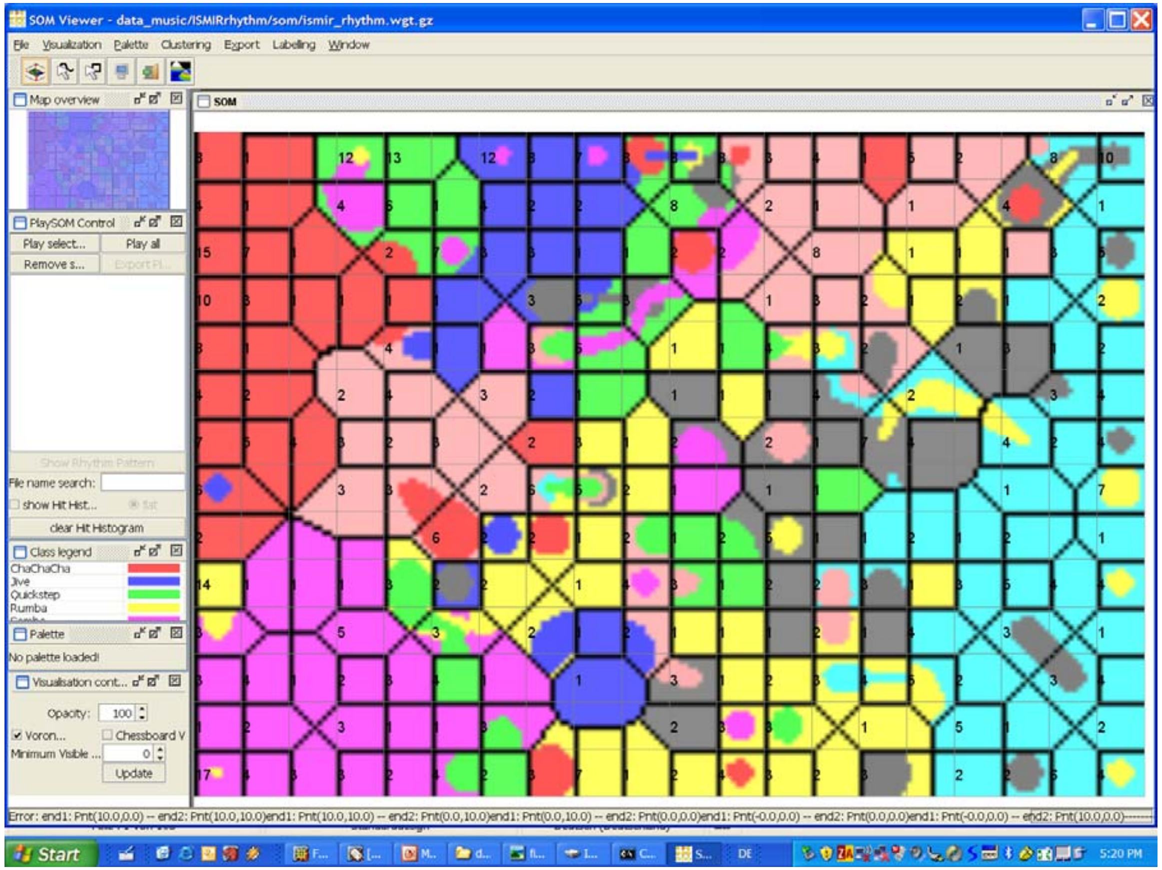
- Step 4: identify regions and attractors



(A) (B) (C) (D)

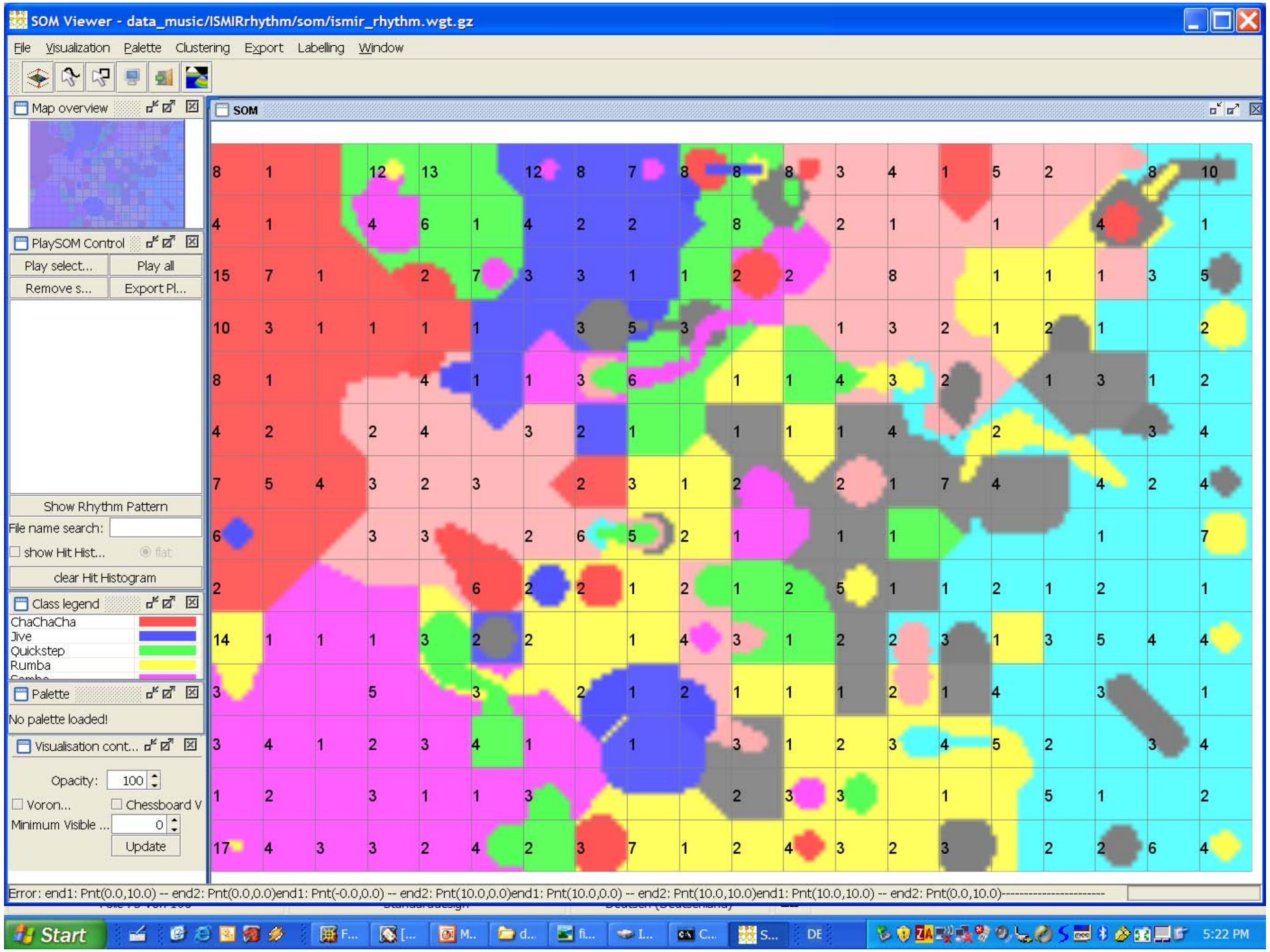
Class Coloring: Attractor Flooding

- Step 5: flood-fill along attractors according to class frequency



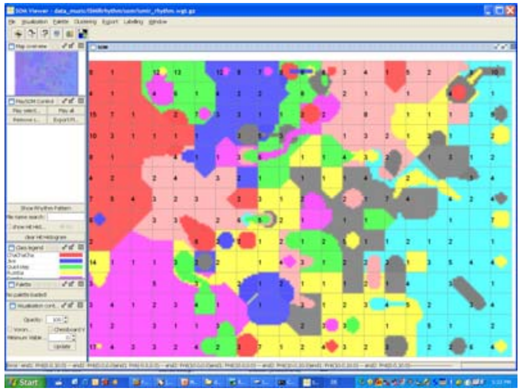
Class Coloring: Attractor Flooding

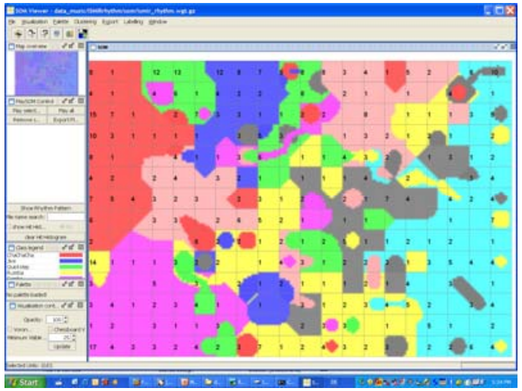
- final:

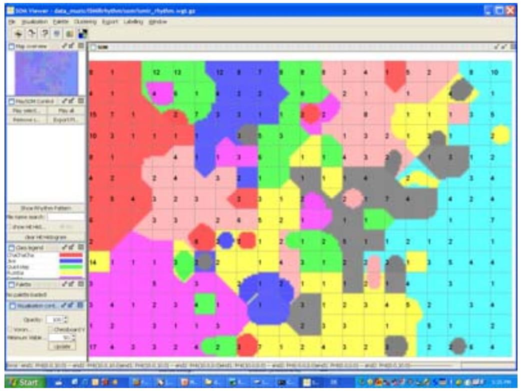


Class Coloring: Attractor Flooding

- Class frequency thresholding:





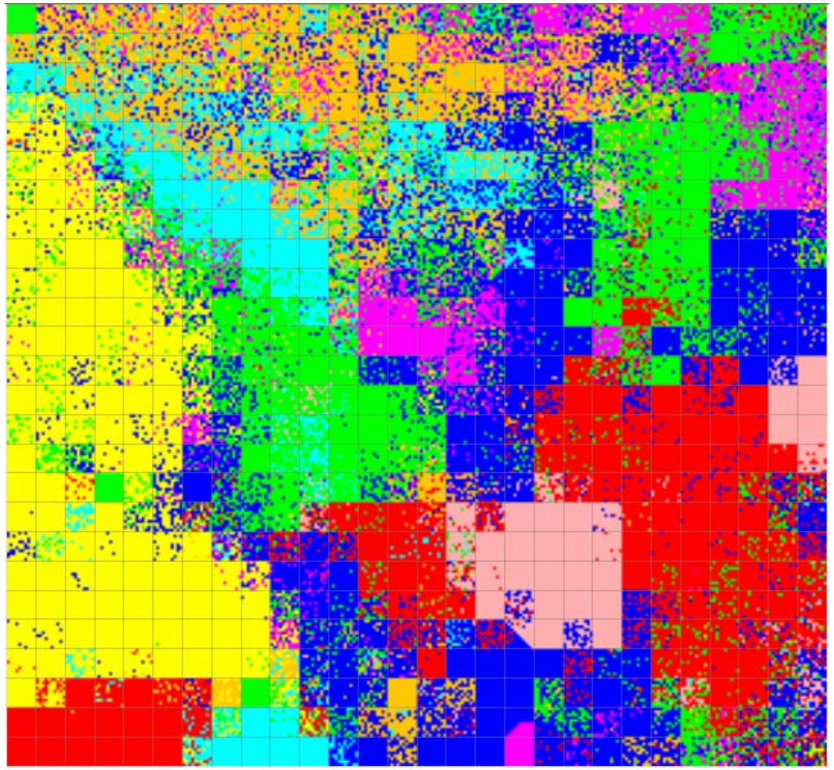


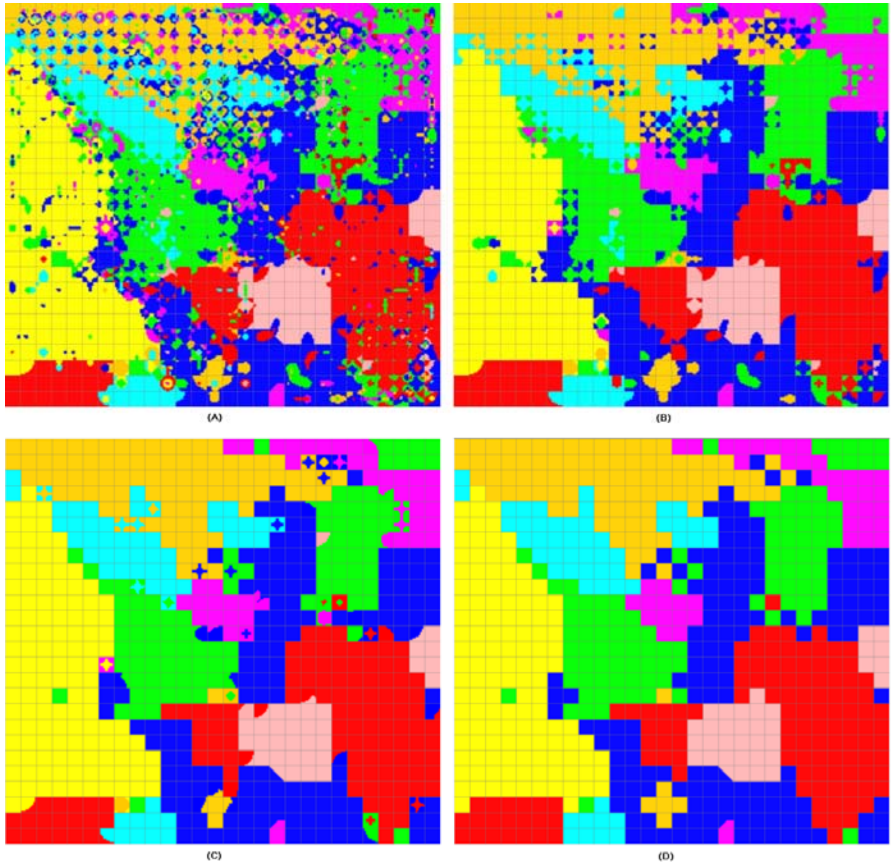
100% 50% 25%

Class Coloring

Radio Search data set:

chessboard and attractor flooding





Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
- Distanzen
- Klasseninformation
- Attribute
  - Component Planes
  - Clustering of Component Planes
  - Metro Maps
  - Vectorfields: grouped Flow
- Clustering der SOM

Attributes

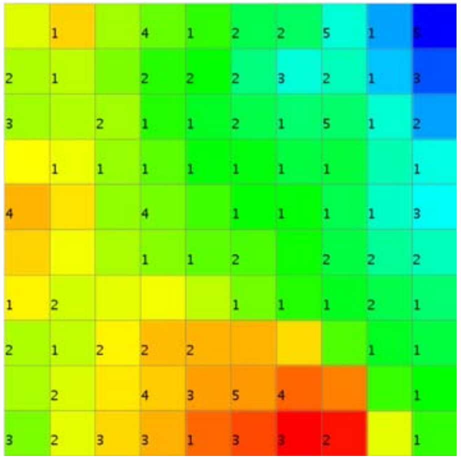
- Analysis of individual attributes or groups of attributes
- Distribution of attribute values
- Correlation between attribute values
- different visualizations
  - component planes
  - clustering of component planes
  - metro maps
  - (component-based flow)

Attributes: Component Planes

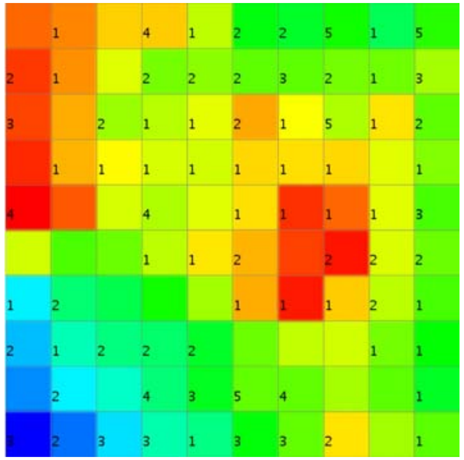
- Component Planes
- „Horizontal slice“: color each unit according to the value of a given attribute
- Analyze regularity of distribution:
  - clear gradient
  - islands with high/low value
  - quasi-random, no structure
  - analyze correlations

Attributes: Component Planes

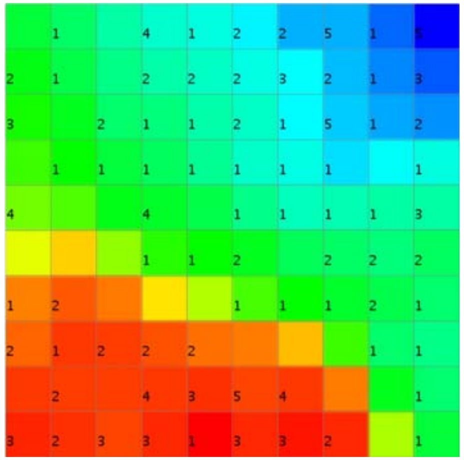
- Iris Dataset – Component Planes



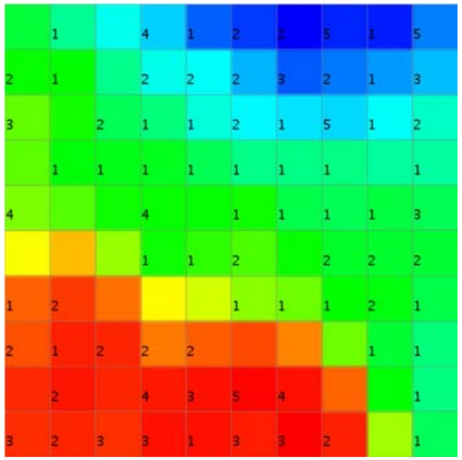
sepal length



sepal width



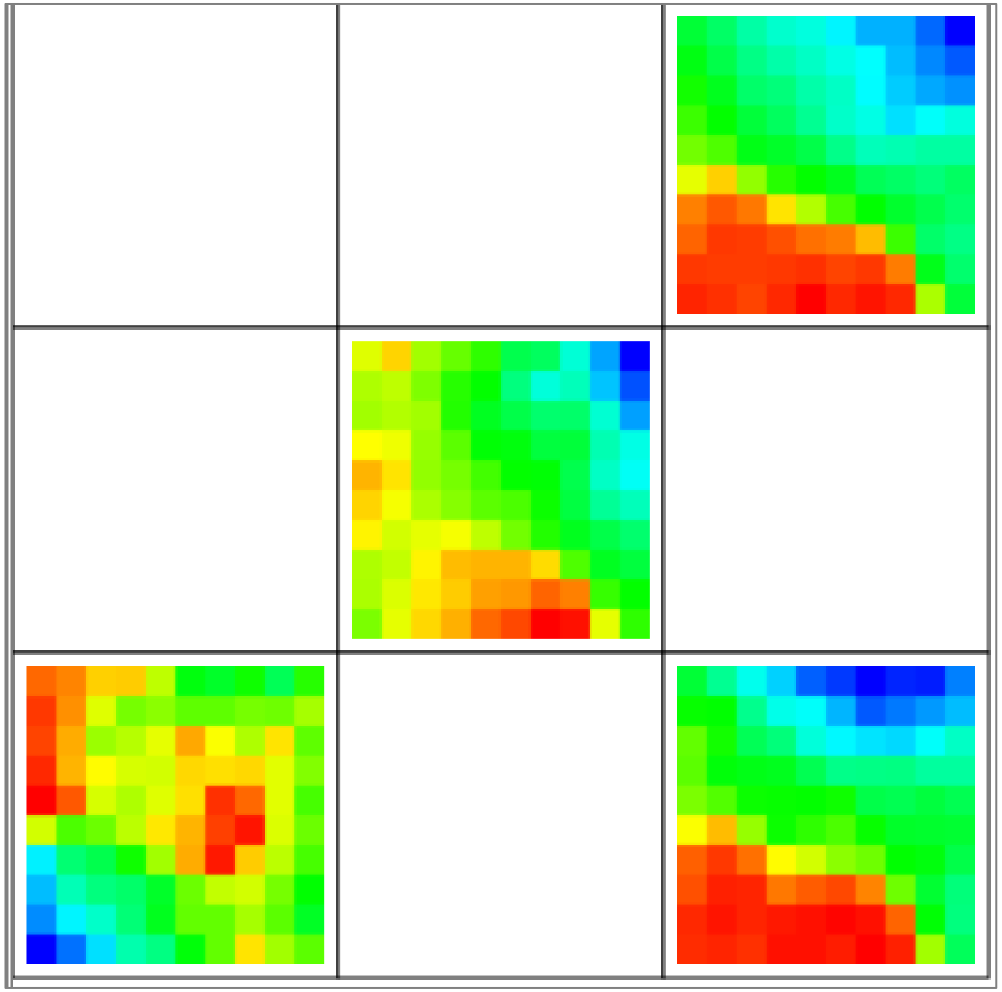
petal length



petal width

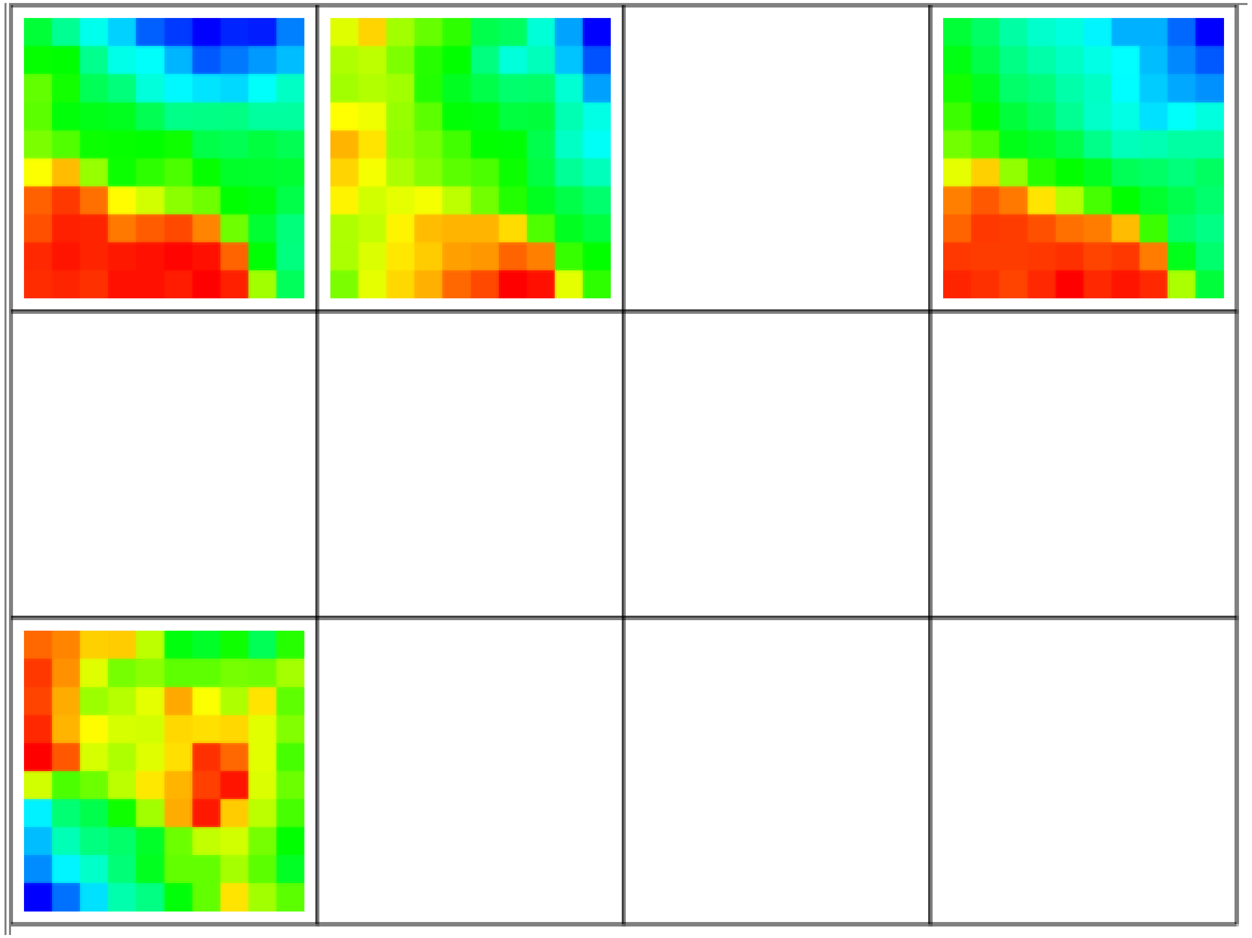
Attributes: Component Planes

- Iris Dataset – Clustered Component Planes



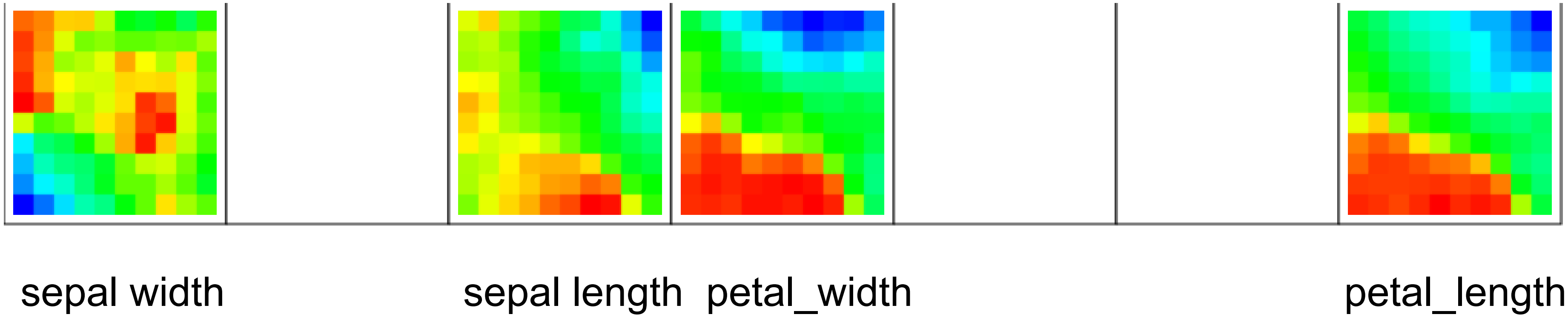
Attributes: Component Planes

- Iris Dataset – Clustered Component Planes

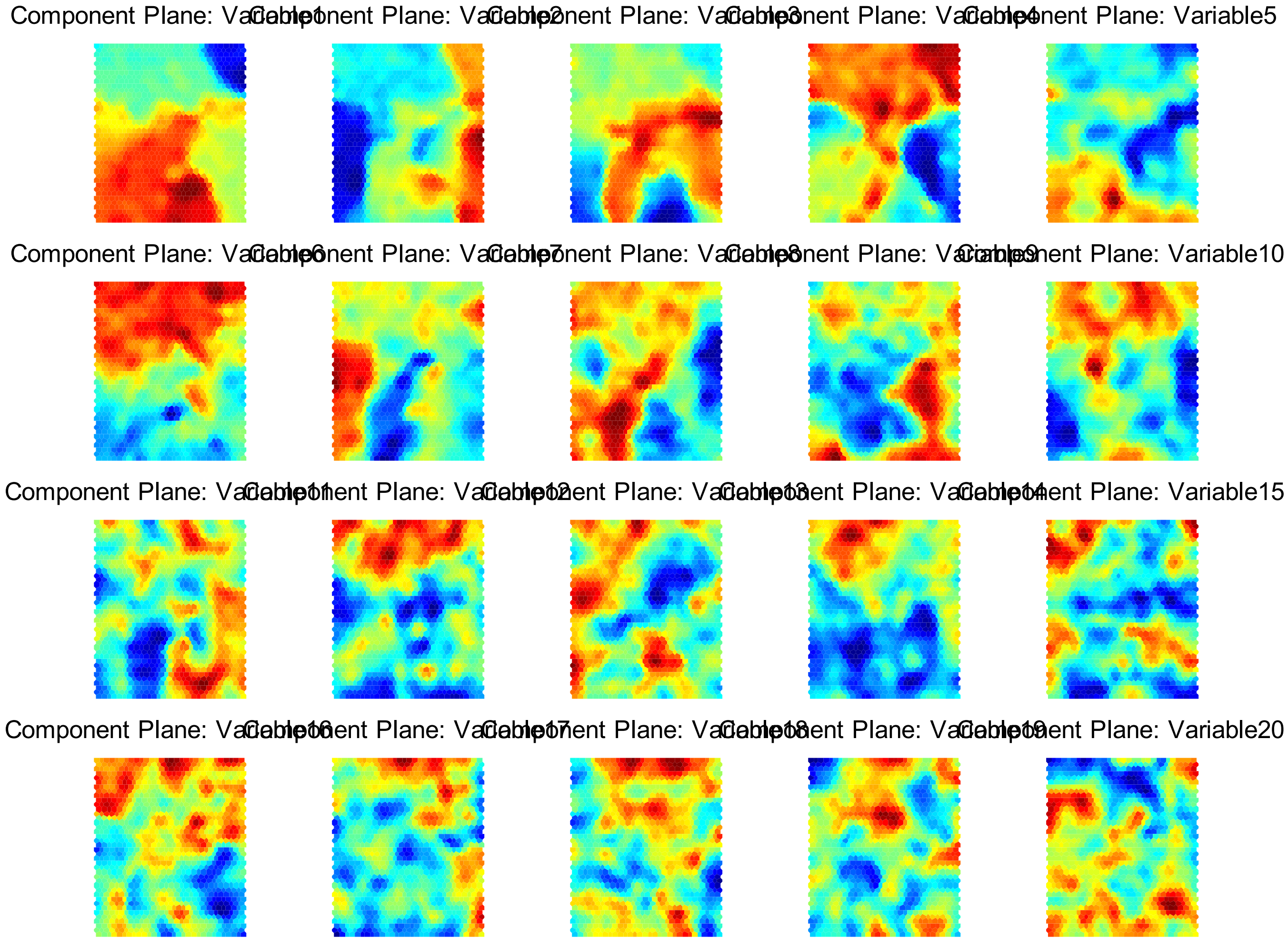


Attributes: Component Planes

- Iris Dataset – Clustered Component Planes
  - linear 7x1 SOM



Attributes: Component Planes



Outline

Visualisierungen der SOM

- Textuelle Informationen
- Dichte
- Distanzen
- Klasseninformation
- Attribute
  - Component Planes
  - Clustering of Component Planes
  - Metro Maps
  - Vectorfields: grouped Flow
- Clustering der SOM

Attributes: MetroMaps

- Component Planes provide overview of attribute value distribution across SOM
- Multiple images (1 per attribute)
- Difficult to comprehend
- Hard to understand correlations between attributes
- MetroMaps
  - concept of skewed distances
  - simplified structure
  - aggregation of correlated component planes
  - overlay to any colored SOM visualization

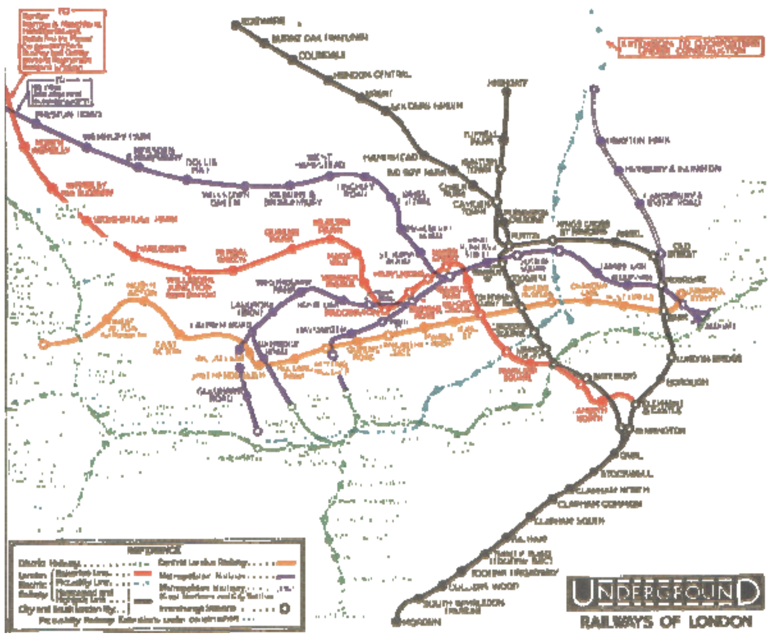
Attributes: MetroMaps

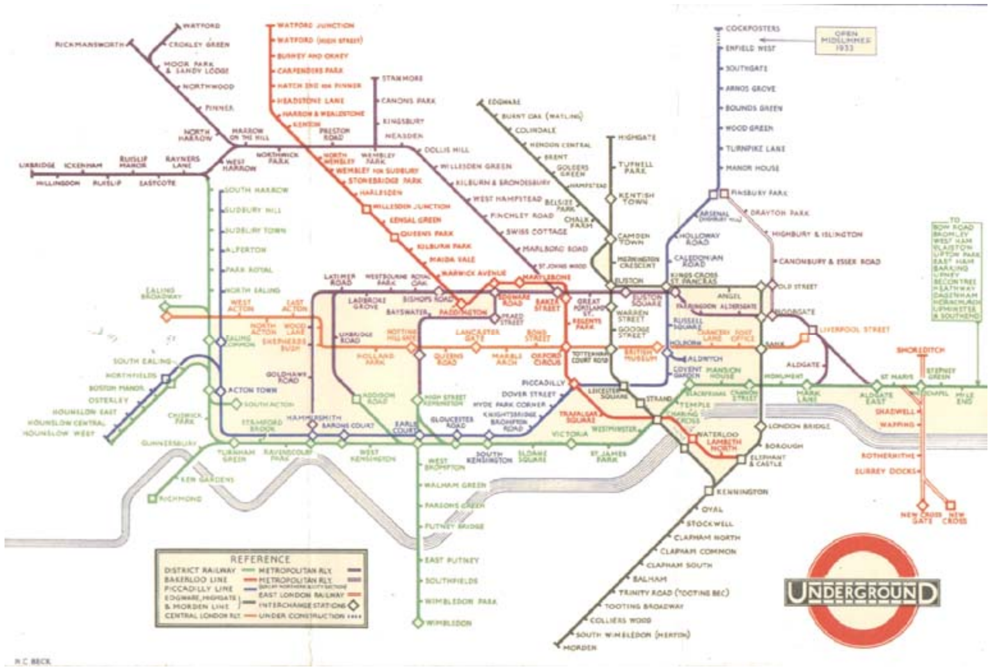
- Robert Neumayer, Rudolf Mayer, Georg Pölzlbauer, and Andreas Rauber. The metro visualisation of component planes for self-organising maps. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'07)*, Orlando, FL, USA, August 12-17 2007. IEEE Computer Society.
- Robert Neumayer, Rudolf Mayer, and Andreas Rauber. Component selection for the metro visualisation of the SOM. In *Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM'07)*, Bielefeld, Germany, September 3-6 2007.

Attributes: MetroMaps

- London Metro Map

1932 | 1933

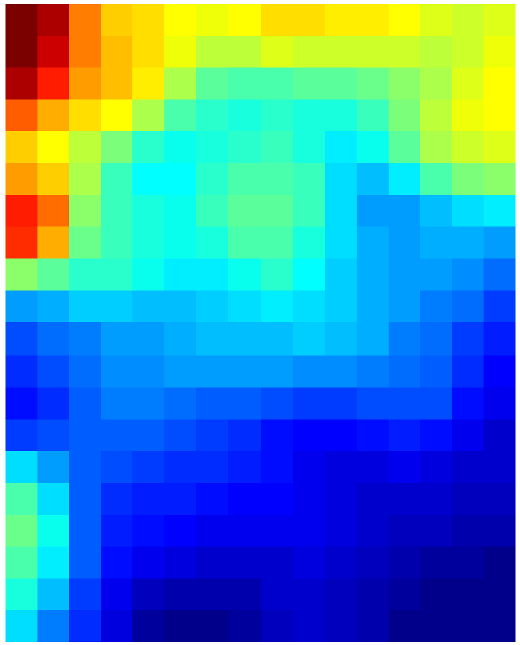




Attributes: MetroMaps

Component Planes

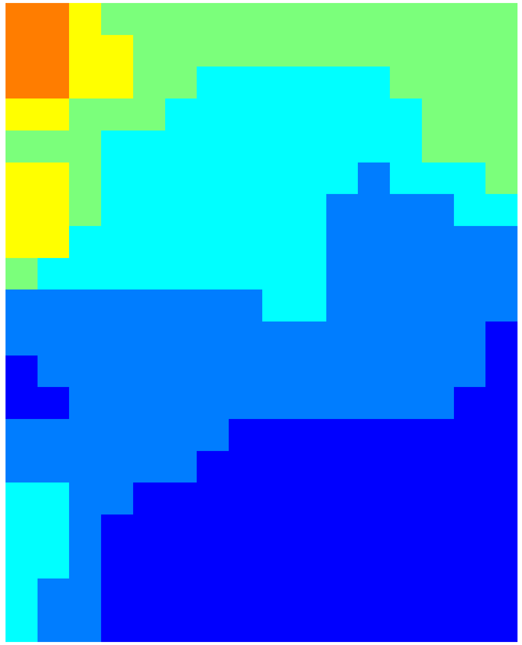
- Single component plane: visualization of 1 model vector component across map
- continuous gradients



Attributes: MetroMaps

Discretization

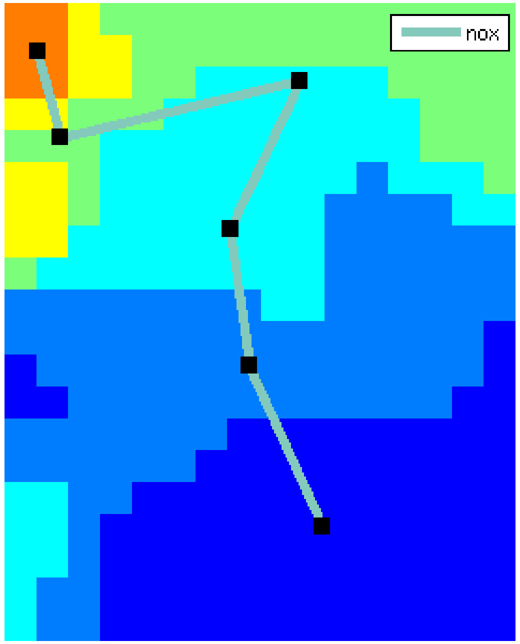
- Discretization of values
- Binning into n bins



Attributes: MetroMaps

Component Lines

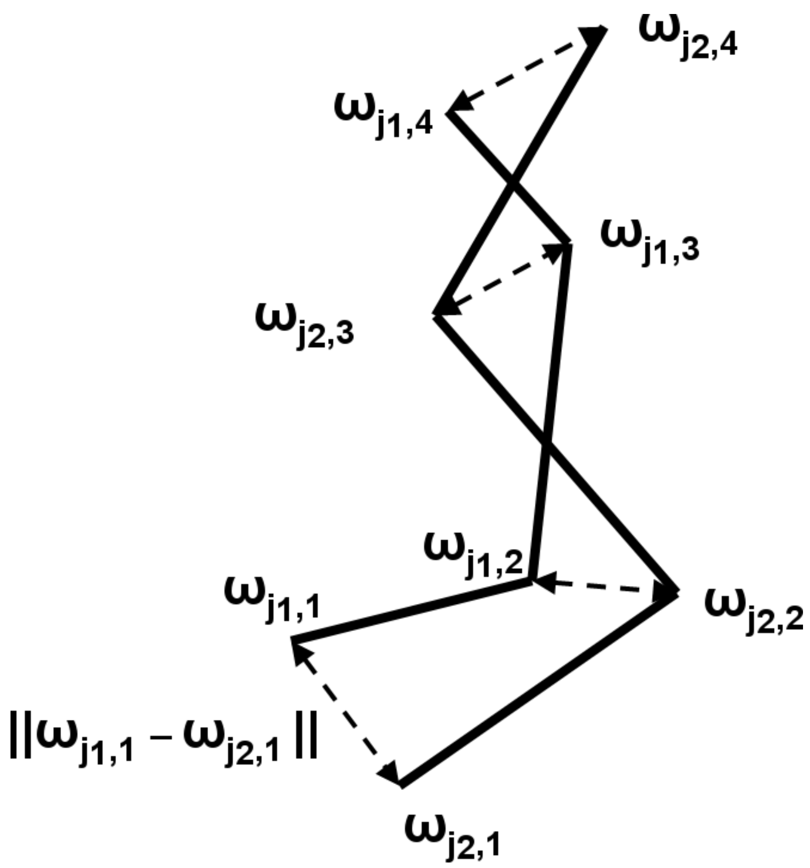
- Computation of centers of gravity
- interconnecting lines
- revealing the gradient of single components



Attributes: MetroMaps

Aggregation

- Calculate distance between component lines
- based on minimum pairwise distances
- Cluster component lines
- Visualization of aggregated subset



Attributes: MetroMaps

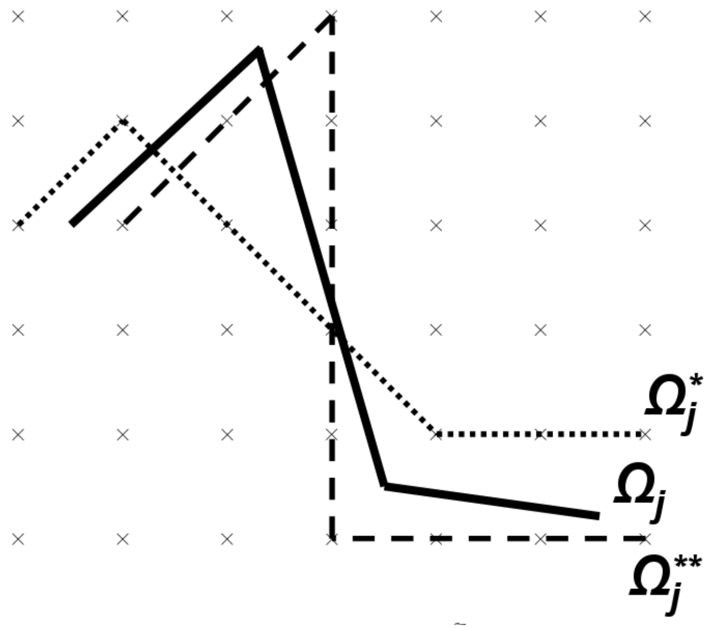
Selection

- use only components with consistent scattering
- ratio of number of bins divided by number of closed regions
- select component lines that have a high ration, i.e. little to no scattering /local minima

Attributes: MetroMaps

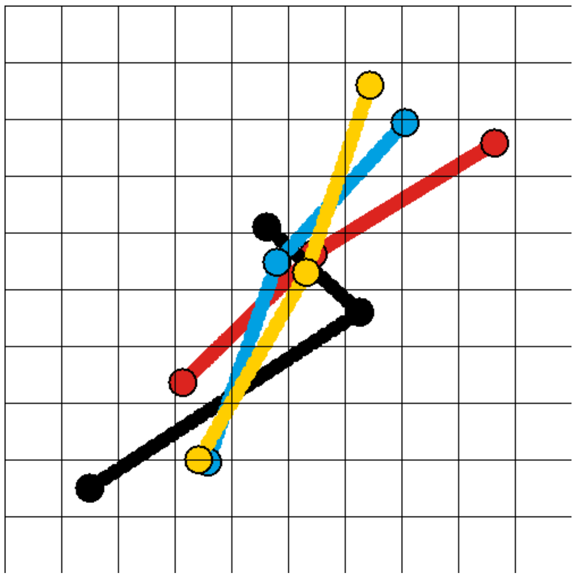
Snapping

- Visual enhancement
- Snapping component lines onto SOM grid
- Allowing only horizontal, vertical, and 45 degree lines
- Clearer structure

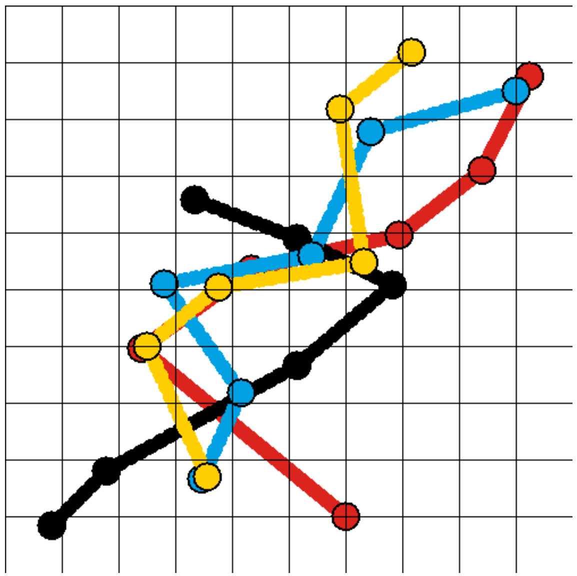


Attributes: MetroMaps

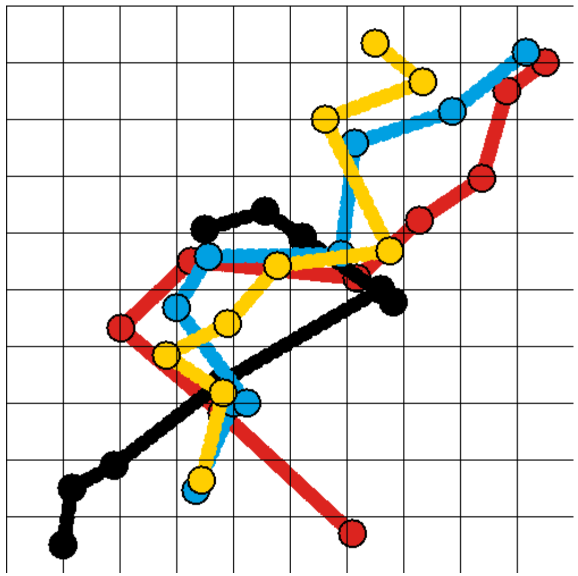
- Iris Data Set
  - Metro Maps, unsnapped



3 bins



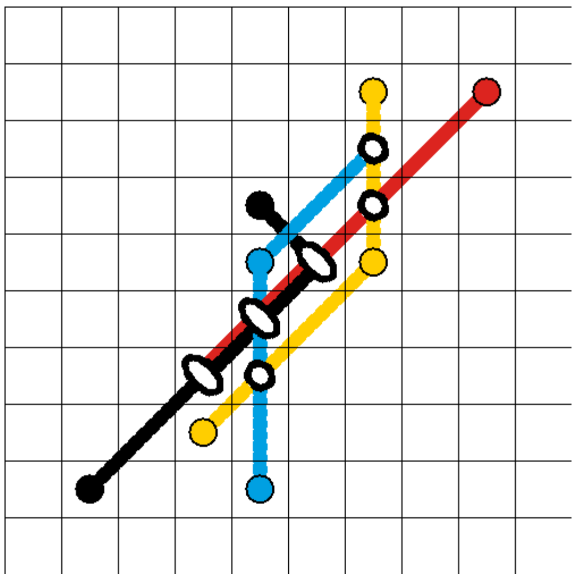
6 bins



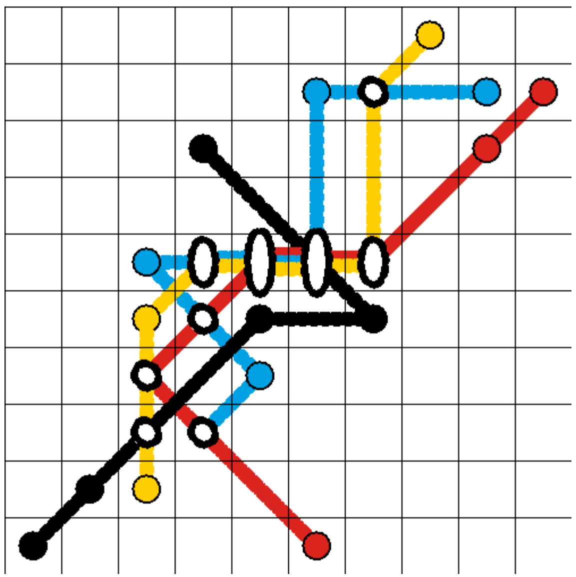
9 bins

Attributes: MetroMaps

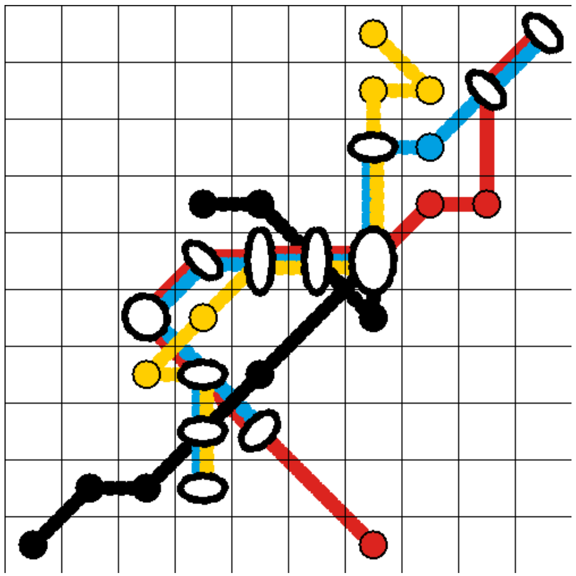
- Iris Data Set
  - Metro Maps, snapped



3 bins



6 bins



9 bins

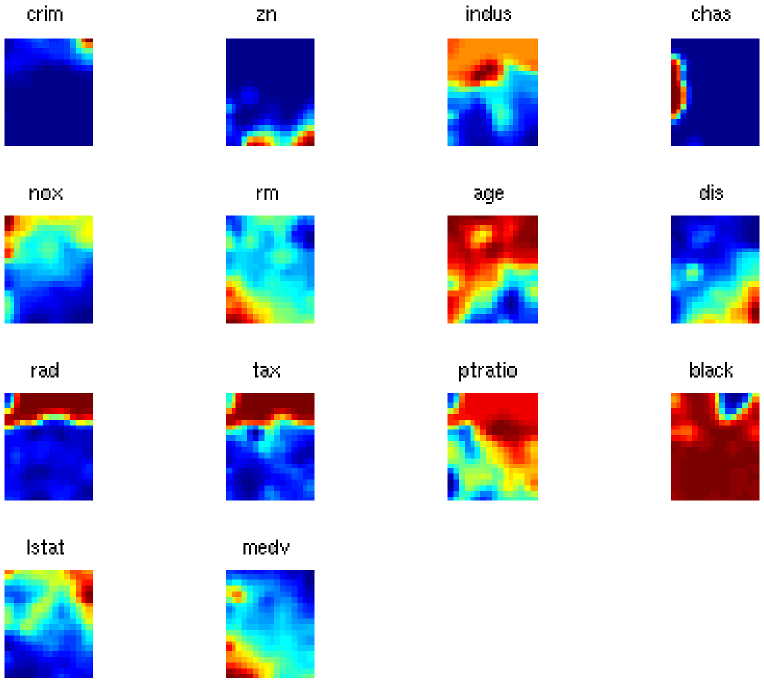
Attributes: MetroMaps

- Boston Housing Data
  - UCI Machine Learning Repository
  - describes households in Boston
  - 506 instances
  - 14 attributes
  - 20x16 SOM
  - discretization based on 6 bins
  - U-matrix as background visualization

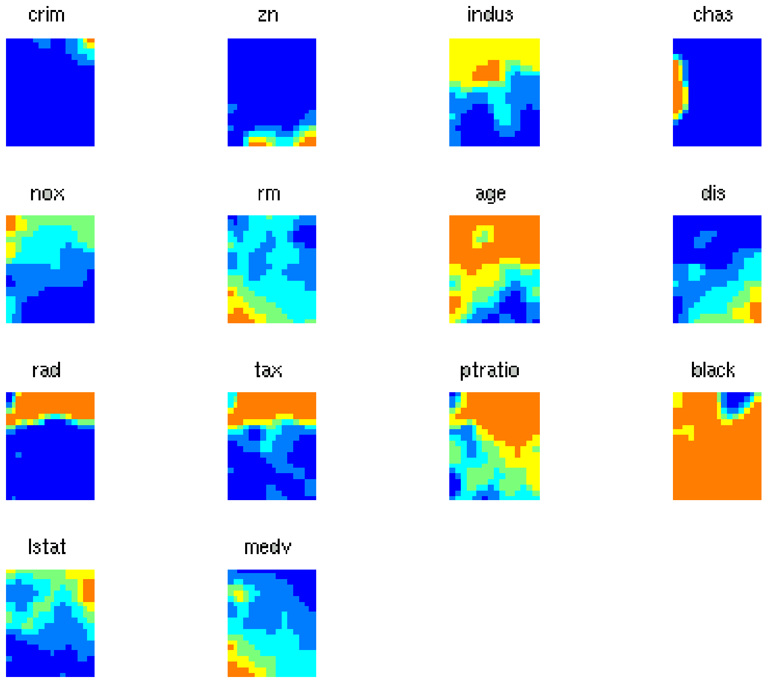
Attributes: MetroMaps

- Boston Housing - Discretization

component planes

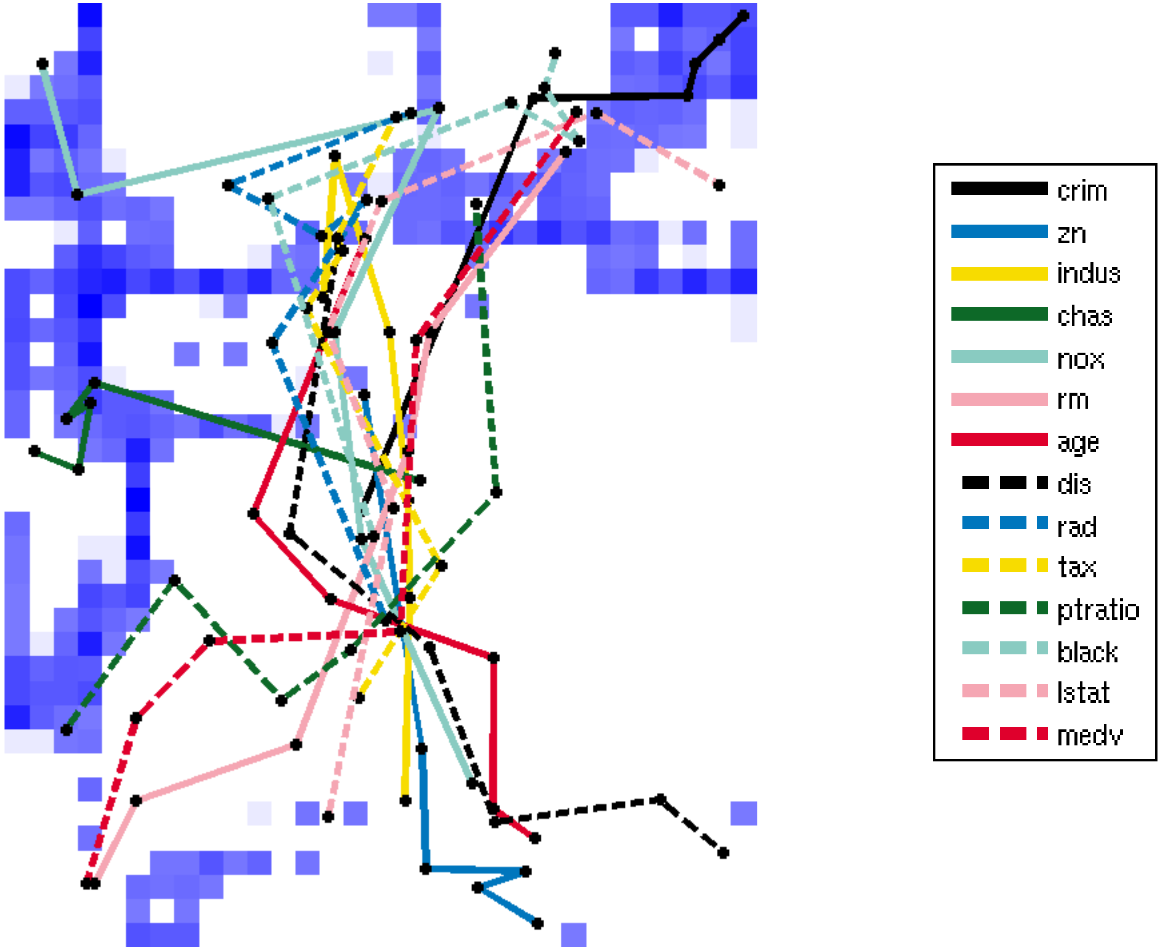


binned



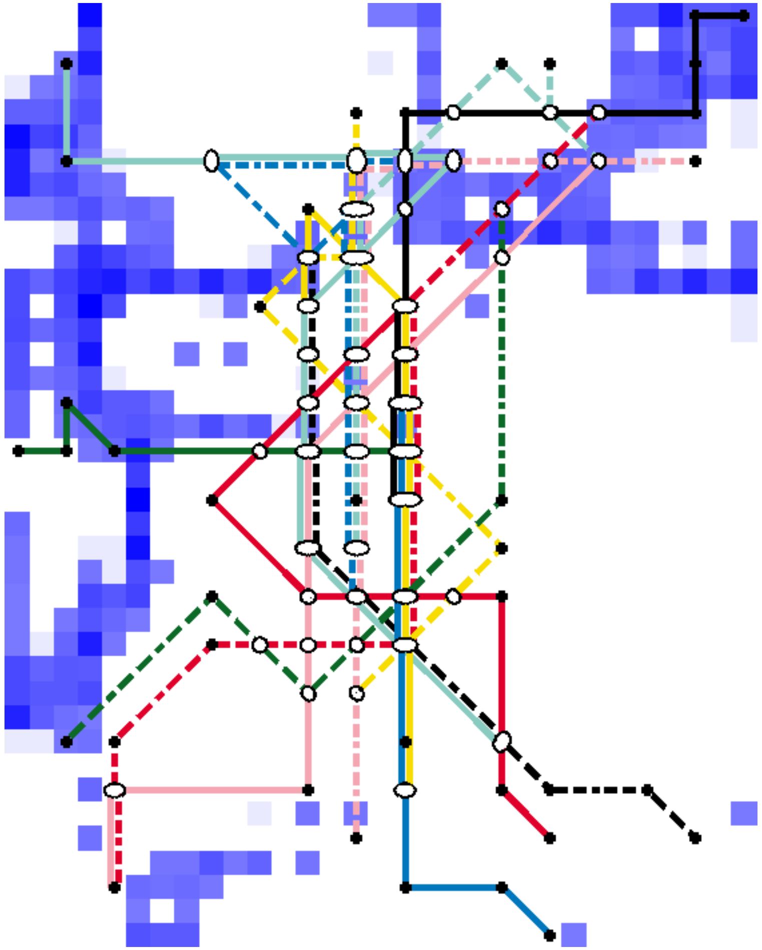
Attributes: MetroMaps

- Boston Housing - Component Lines



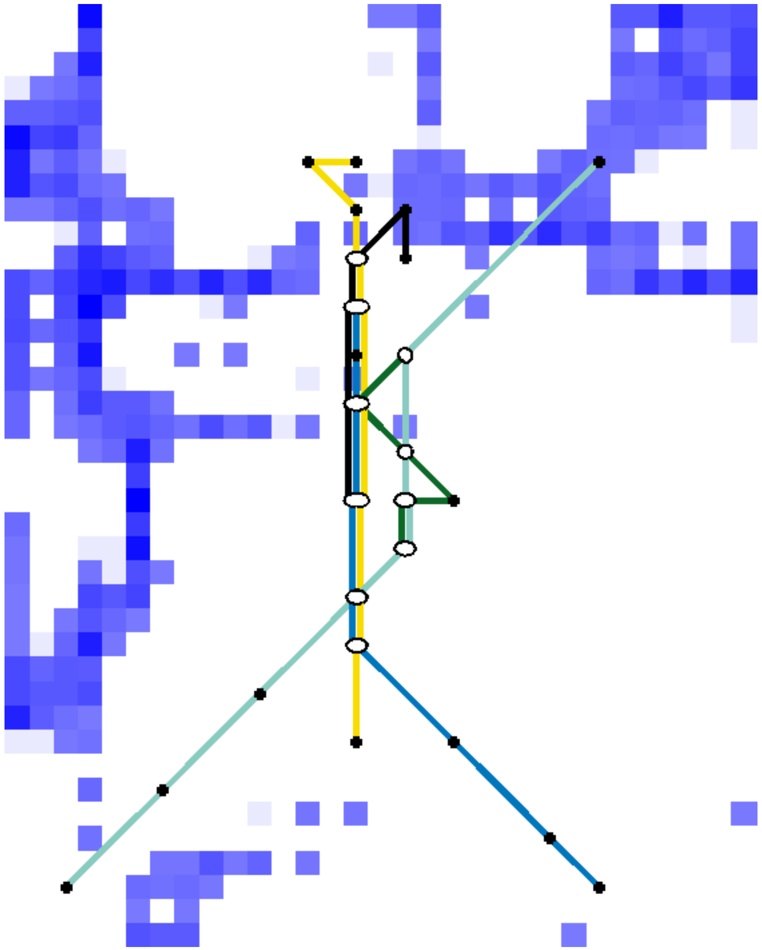
Attributes: MetroMaps

- Boston Housing - Snapped



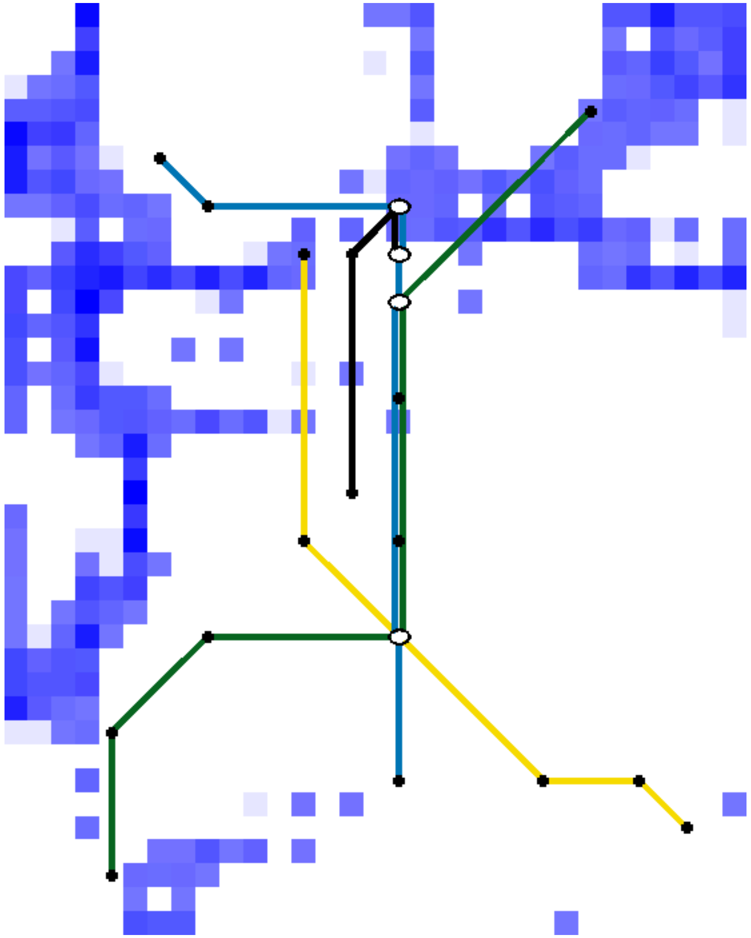
Attributes: MetroMaps

- Boston Housing - Aggregated



Attributes: MetroMaps

- Boston Housing - Selected



Outline

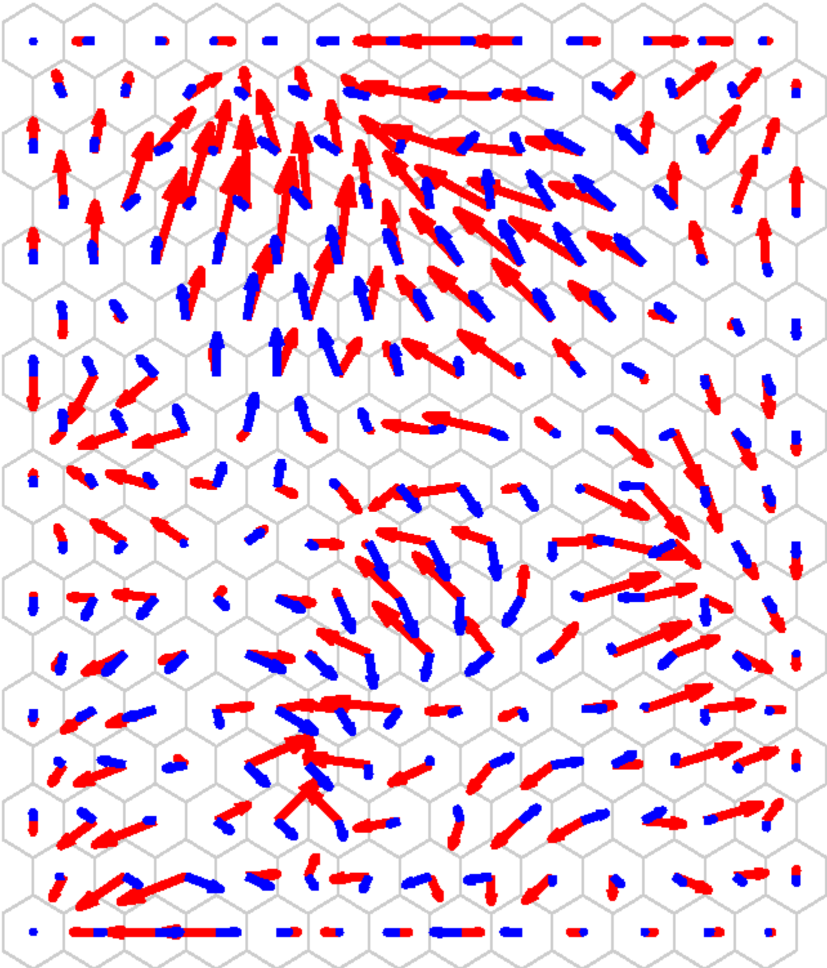
Visualisierungen der SOM

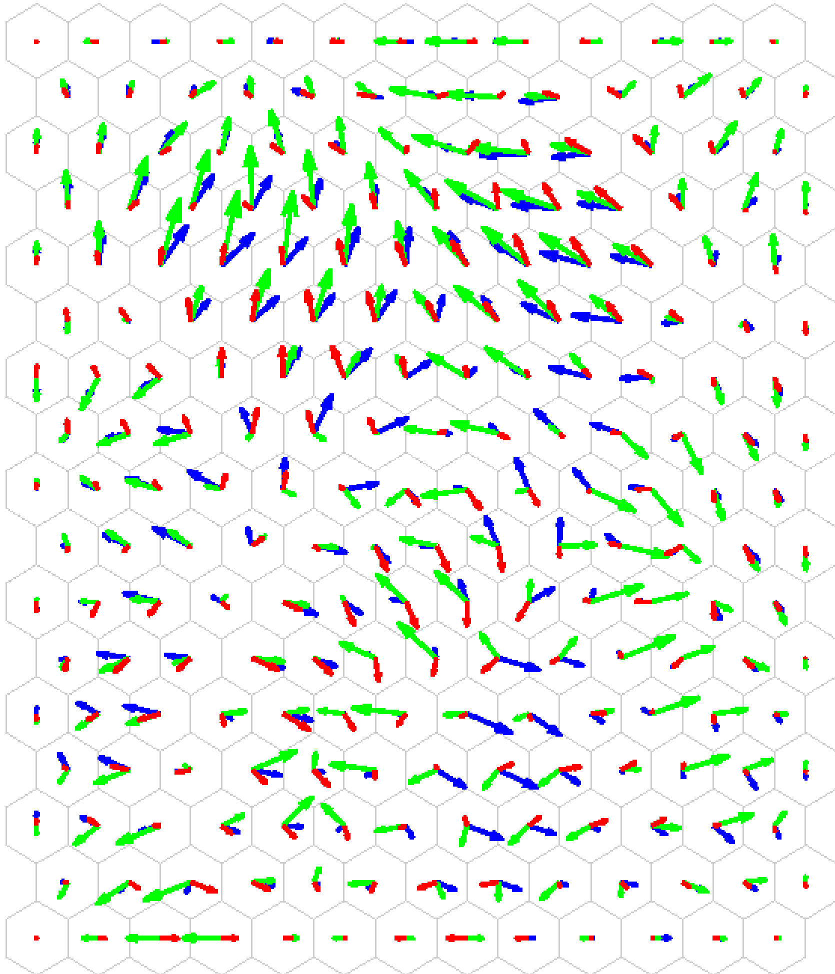
- Textuelle Informationen
- Dichte
- Distanzen
- Klasseninformation
- Attribute
  - Component Planes
  - Clustering of Component Planes
  - Metro Maps
  - Vectorfields: grouped Flow
- Clustering der SOM

Flow: Groups of Attributes

- Flow-based visualization for groups of attributes
- Attributes may be grouped by
  - clustering: data correlation
  - semantics: source, type of information
- Up to 3 groups can be meaningfully interpretable

Flow: Groups of Attributes





Outline

Visualisierungen der SOM

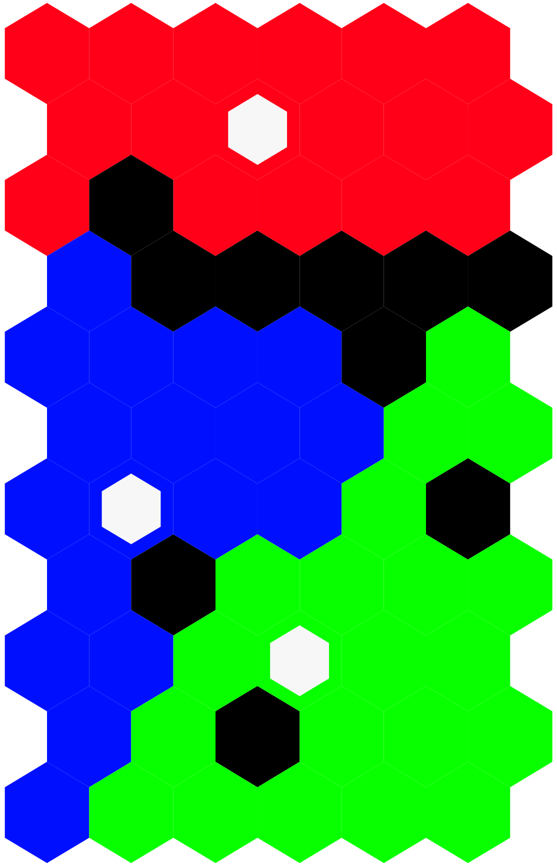
- Textuelle Informationen
- Dichte
- Distanzen
- Klasseninformation
- Attribute
- Clustering der SOM
  - flach: k-means
  - hierarchisch: single/complete linkage, WARD,…

Clustering the SOM

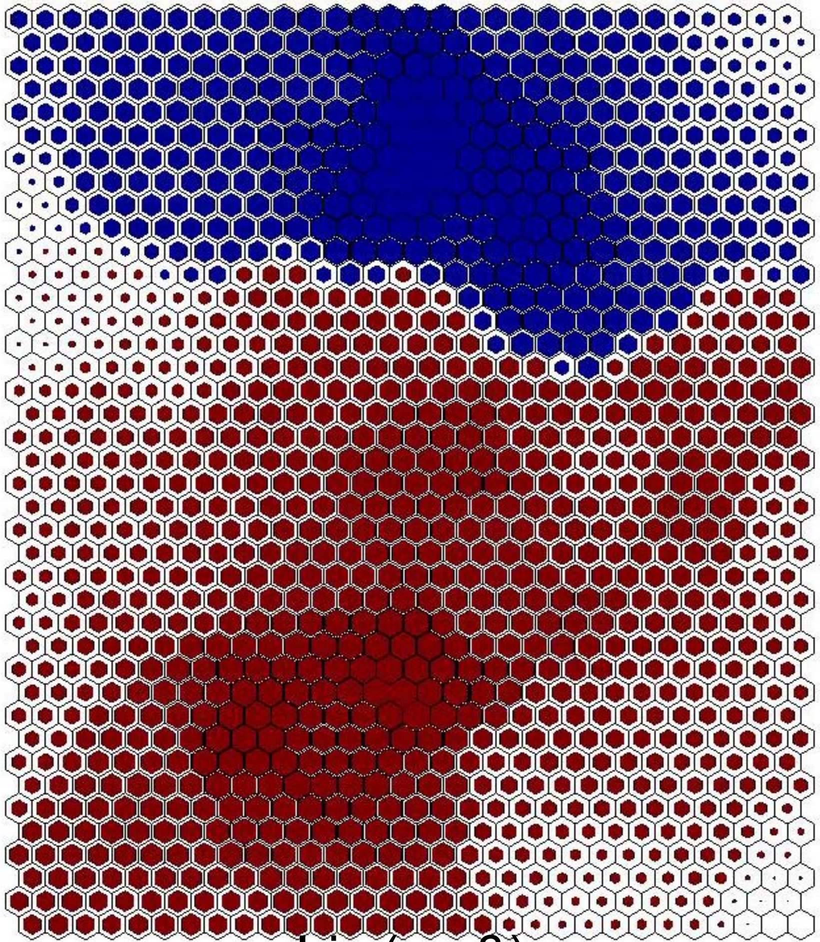
- Clustering: Unterteilung der Karte in Regionen
- k-means: erzeugt *k* Cluster
- Berechnung ähnlich SOM
- nicht deterministisch: Ergebnis bei gleicher Eingabe (=SOM) nicht unbedingt gleich

- Juha Vesanto, Esa Alhoniemi:Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks 11(3):586-600. 2000. IEEE.
- Angela Roiger: **Analyzing, Labeling, and Interacting with SOMs for Knowledge Management**. Master Thesis, Department of Software Technology and Interactive Systems, Vienna University of Technology, March 2007.

Clustering: k-means



Iris (klein)
k = 3

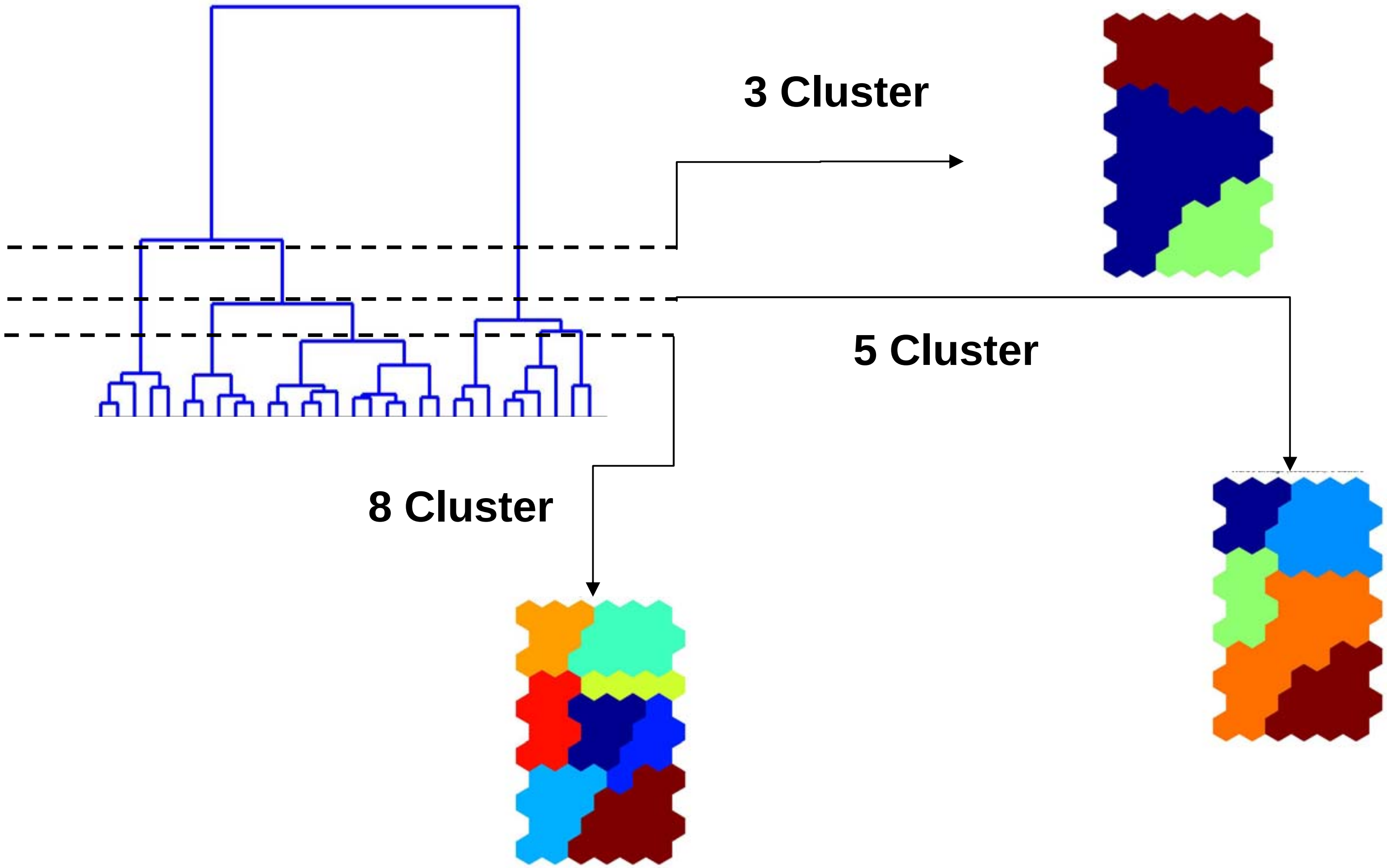


Iris (groß)
k = 2

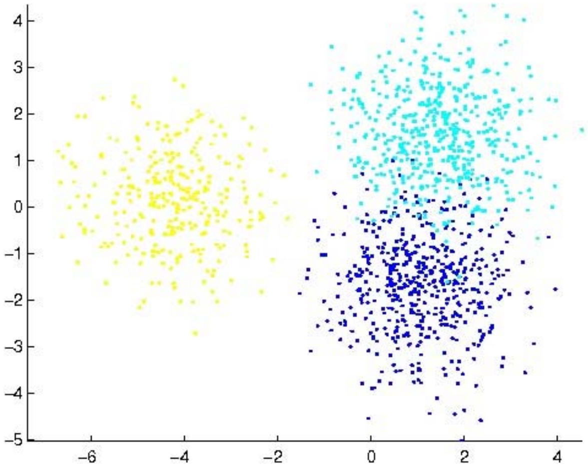
Clustering: Hierarchisch

- Baumstruktur, baut Hierarchie von Unterteilungen auf
- jeweils 2 Cluster werden zusammengefaßt, ergibt nächste Ebene
- Hierarchie kann als Dendrogramm visualisiert werden
- Verschiedene Ansätze
  - single linkage
  - complete linkage
  - WARDs clustering

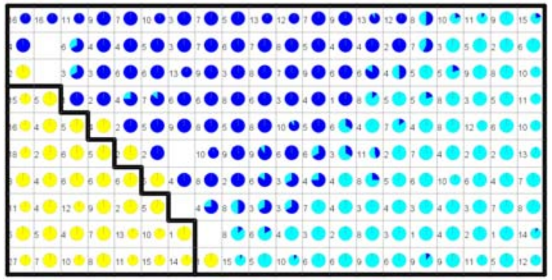
Hierarchisches Clustering



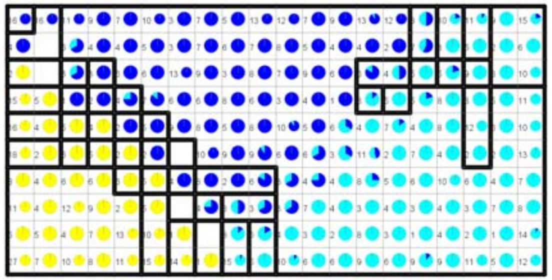
Clustering



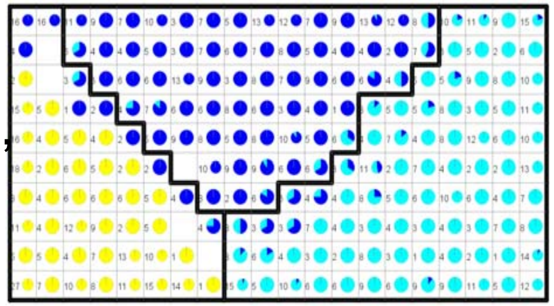
single-linkage,
2 clusters



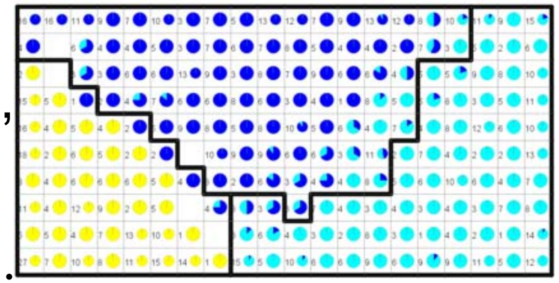
single-linkage,
40 clusters



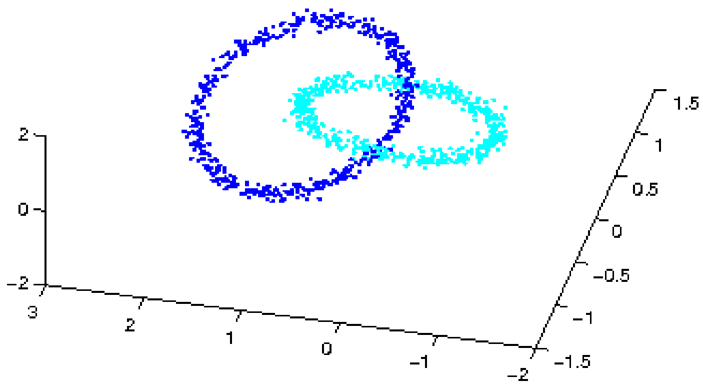
Ward's clustering,
3 clusters



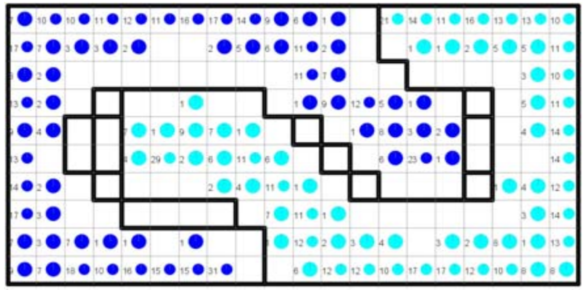
complete-linkage,
3 clusters



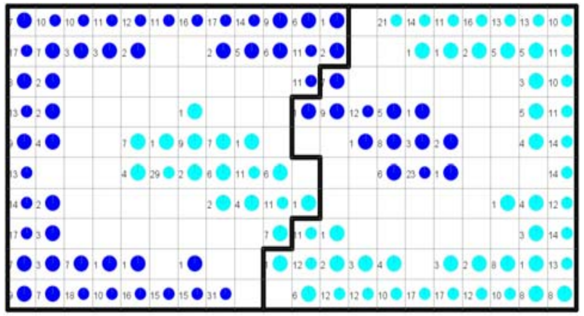
Clustering



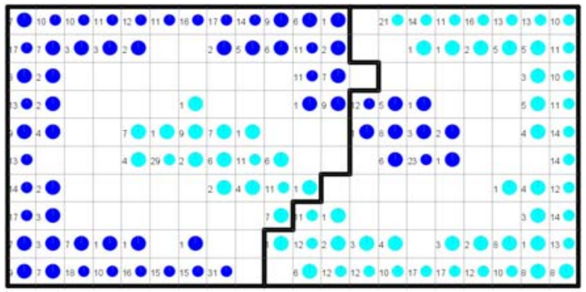
single-linkage,
11 clusters



Ward clustering,
2 clusters



complete-linkage,
2 clusters



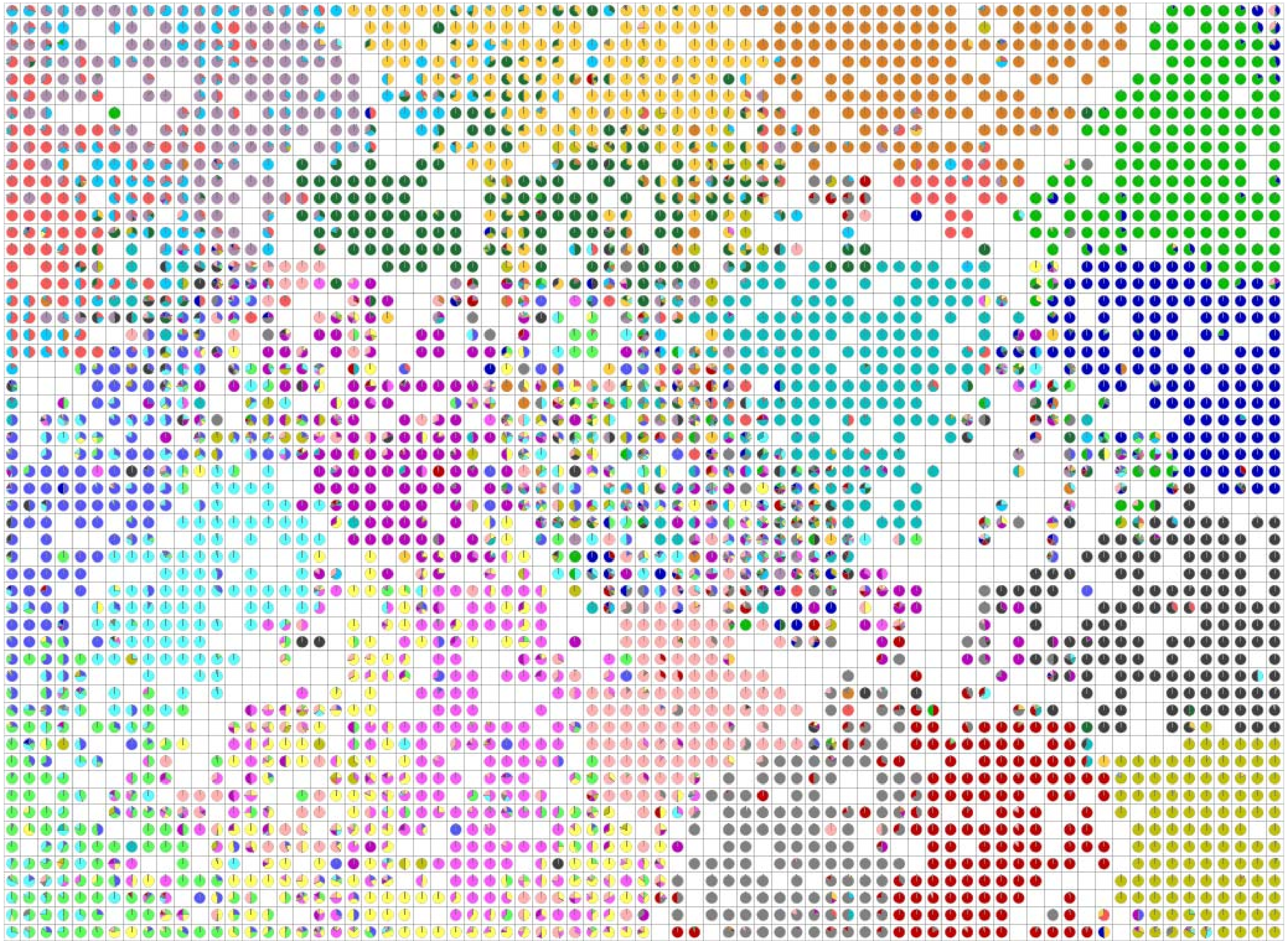
Clustering

- Beispiel: 20 Newsgroups
- Benchmark Datenset
- Je 1000 Postings pro Newsgroup
- Hierarchie von Newsgroups

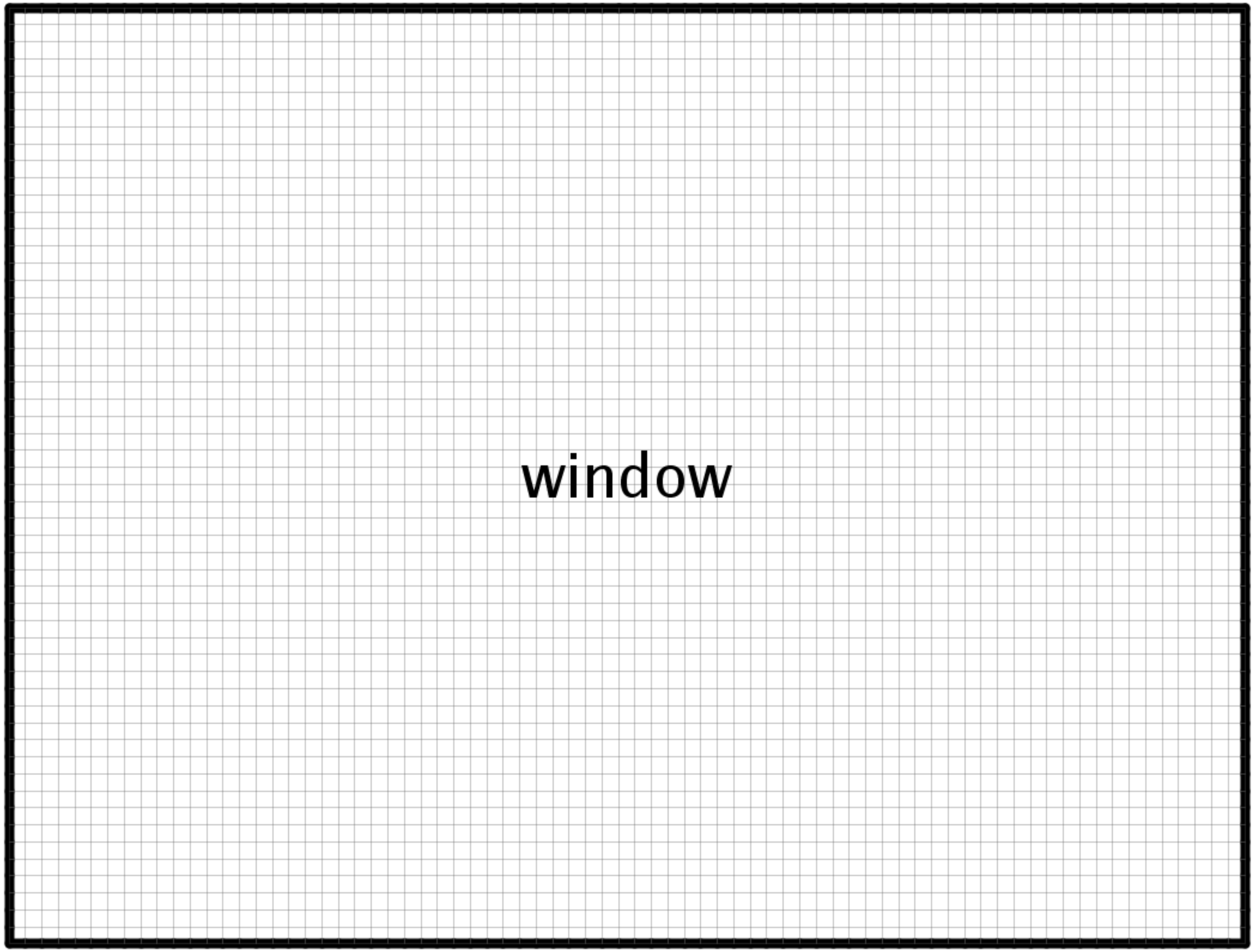
- Full-term indexing
- Wörter auf Wortstämme reduziert

alt.atheism
comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x
misc.forsale
rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey
sci.crypt
sci.electronics
sci.med
sci.space
soc.religion.christian
talk.politics.guns
talk.politics.mideast
talk.politics.misc
talk.religion.misc

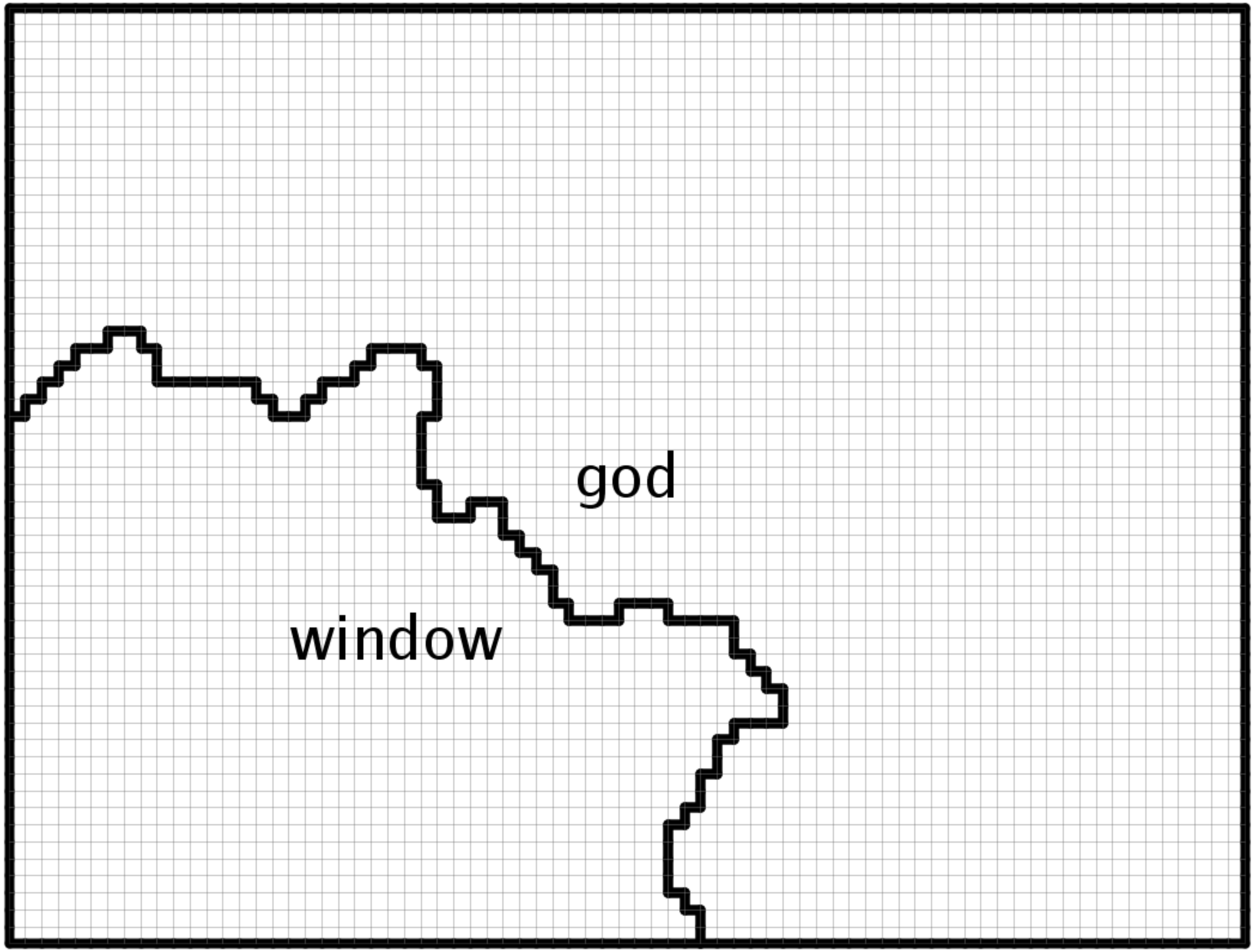
Beispiel: 20 Newsgroups



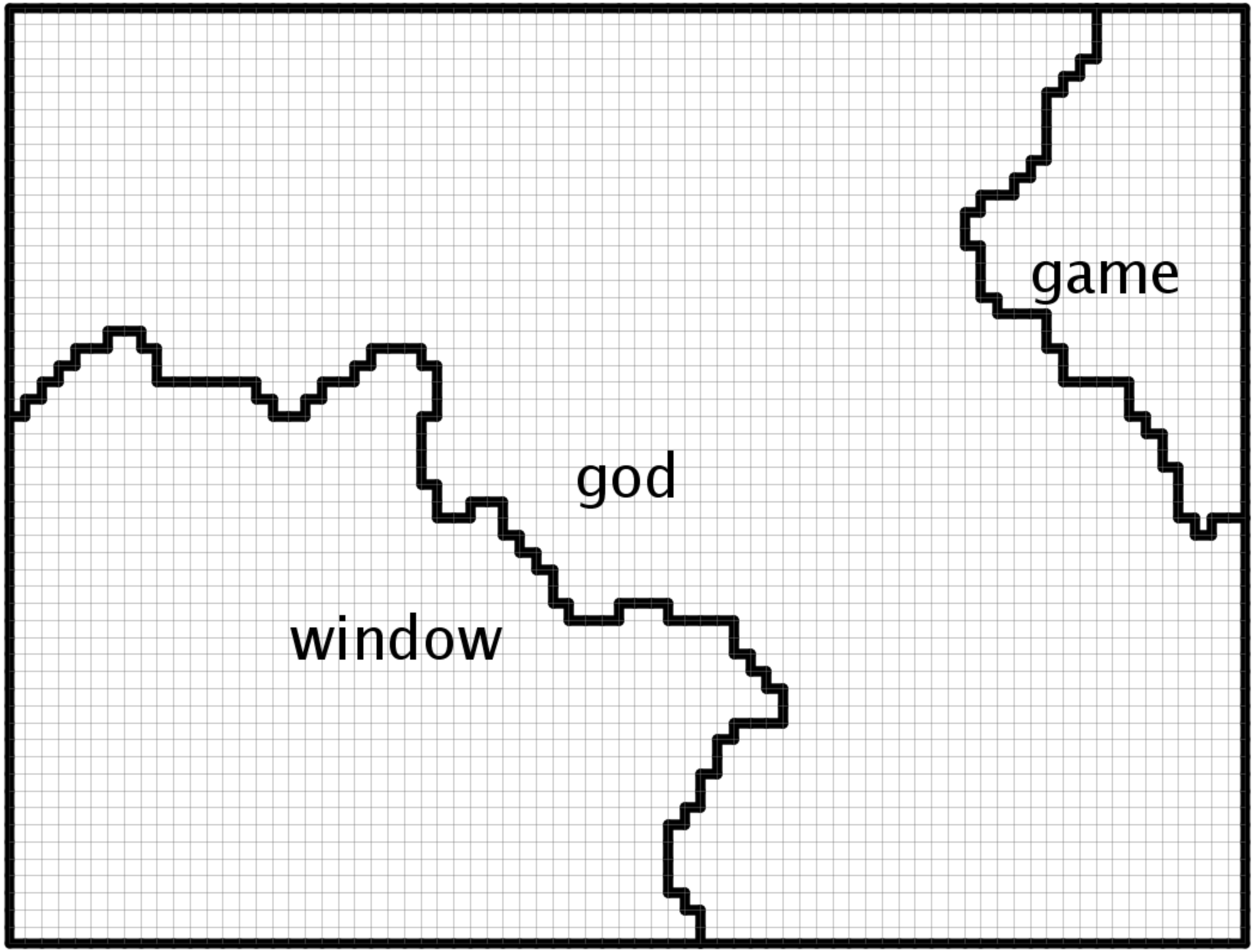
20 Newsgroups: Ward+Labels



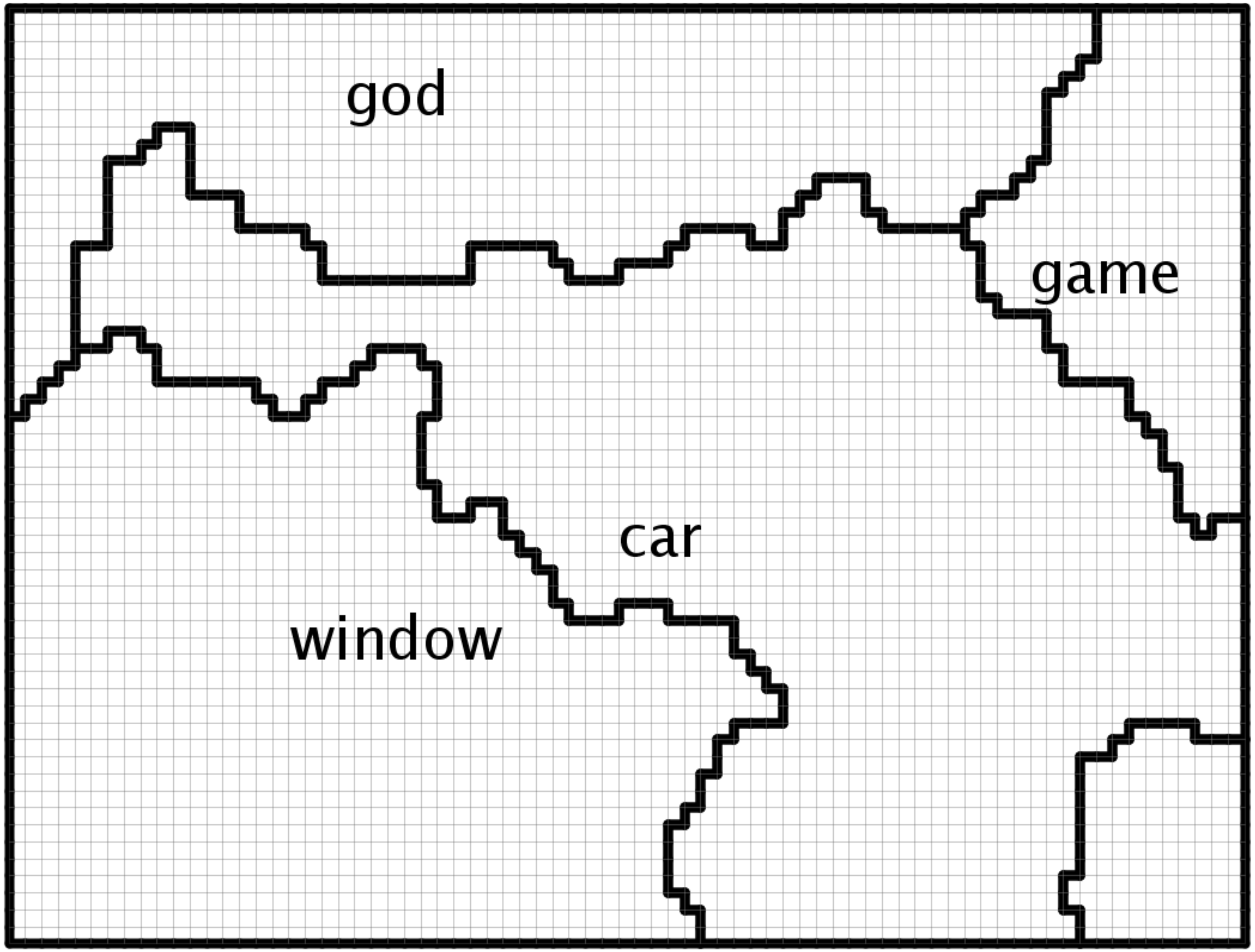
20 Newsgroups: Ward+Labels



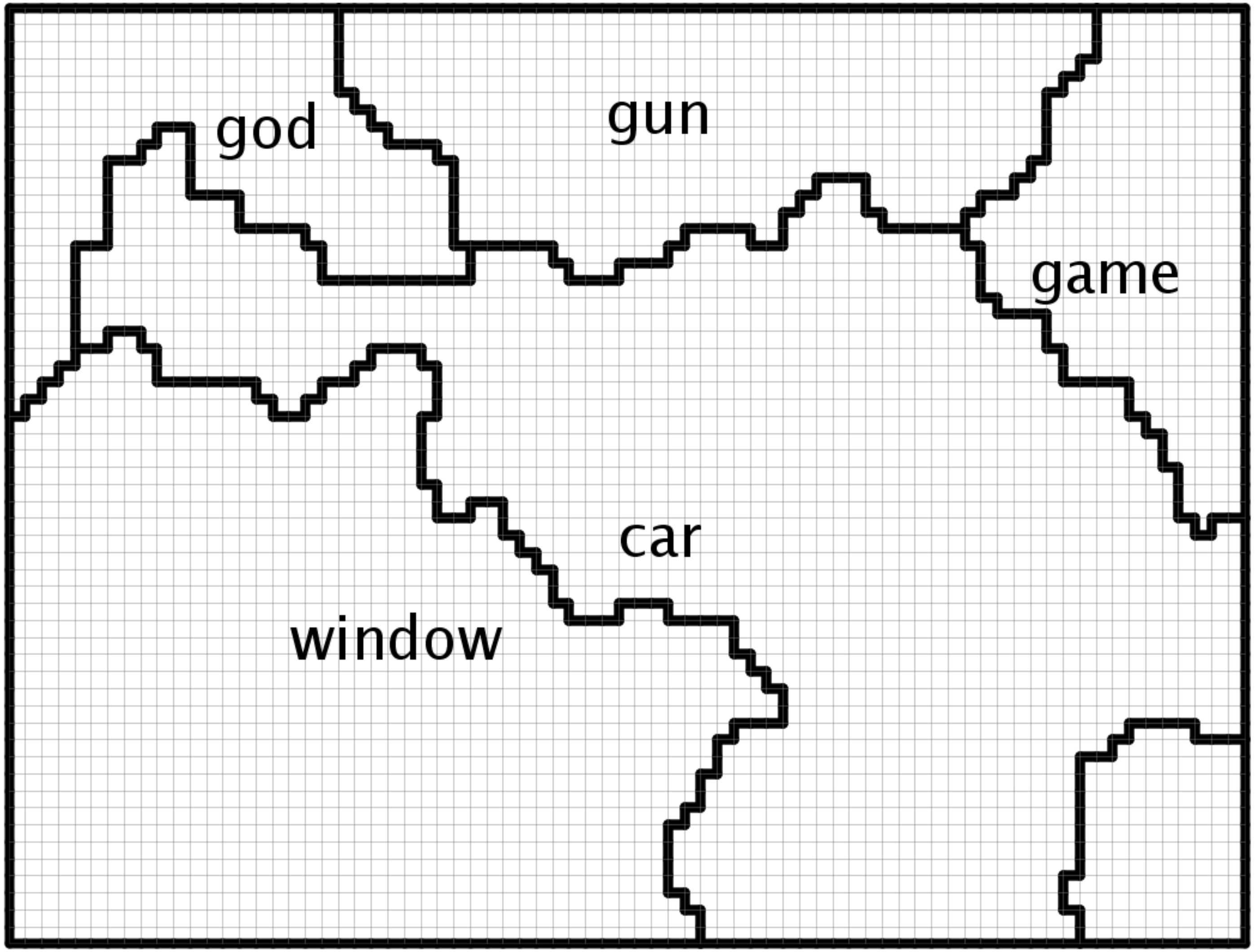
20 Newsgroups: Ward+Labels



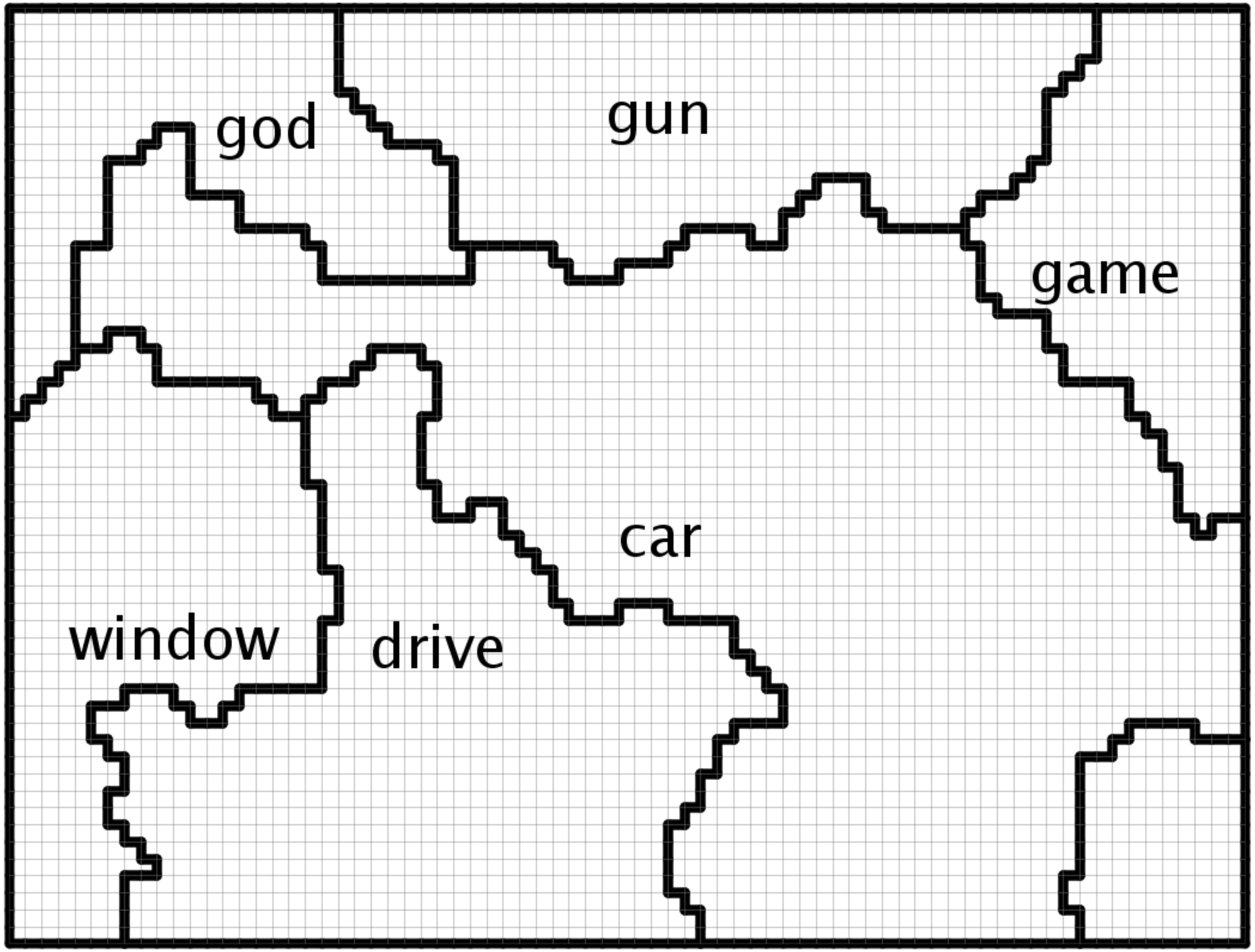
20 Newsgroups: Ward+Labels



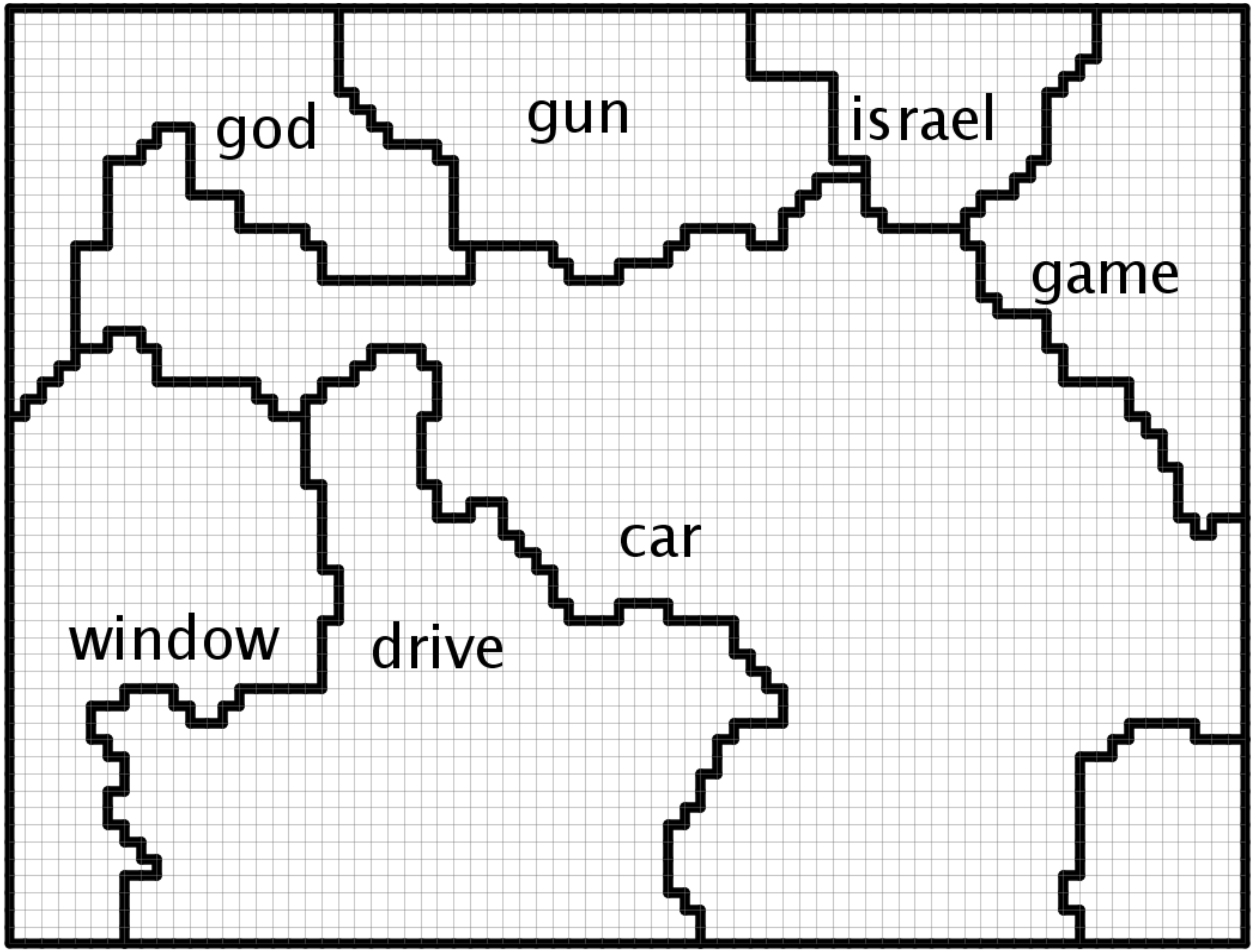
20 Newsgroups: Ward+Labels



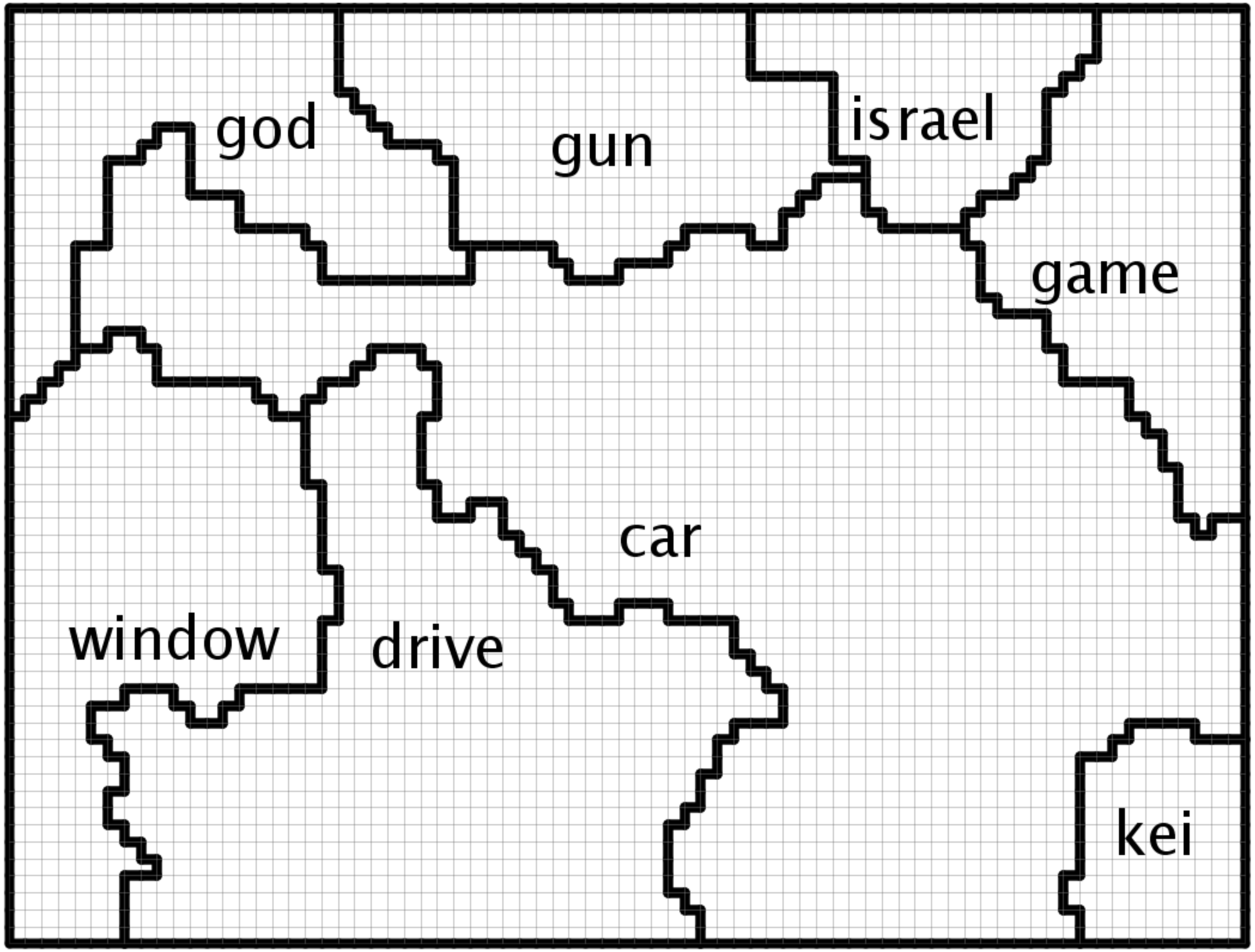
20 Newsgroups: Ward+Labels



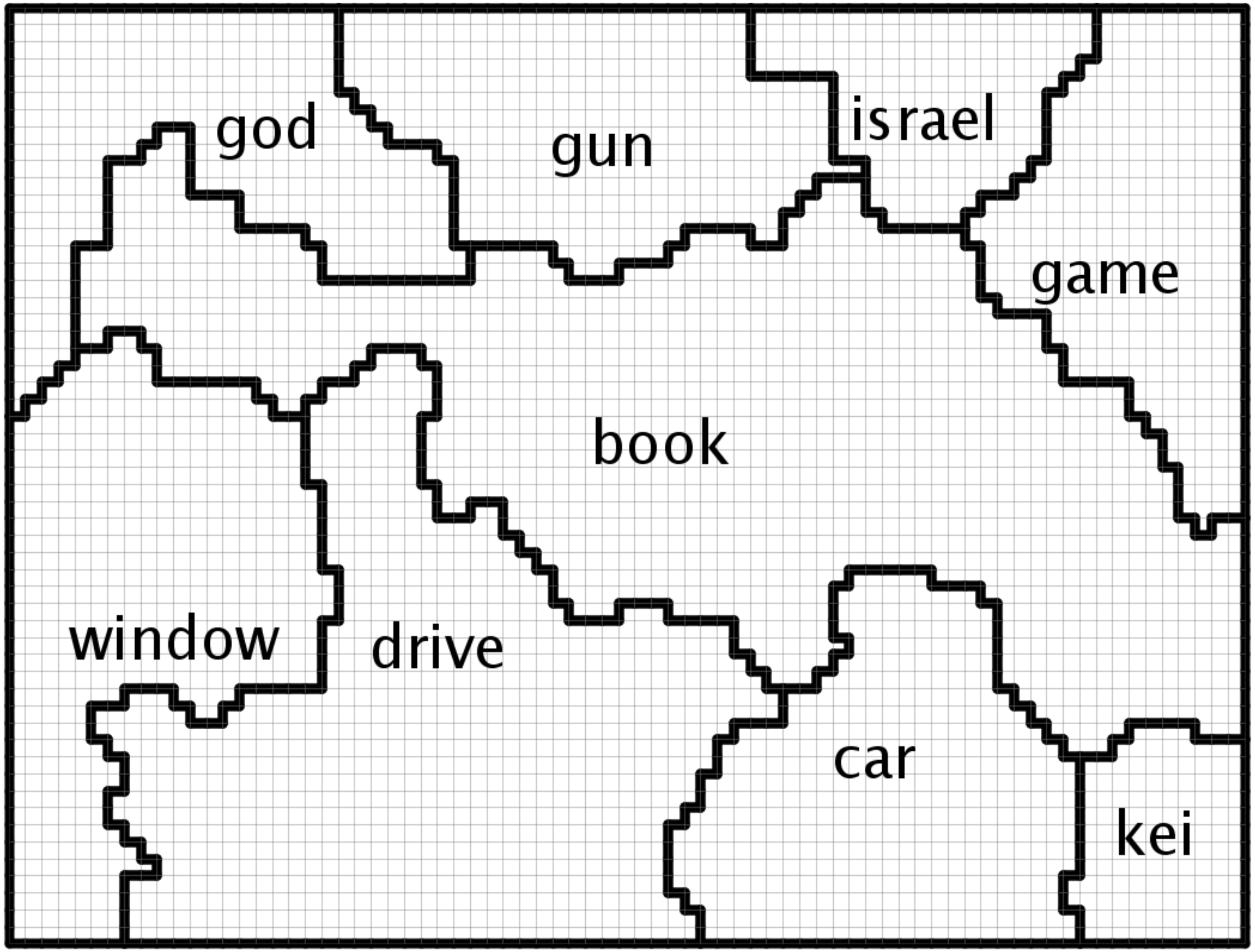
20 Newsgroups: Ward+Labels



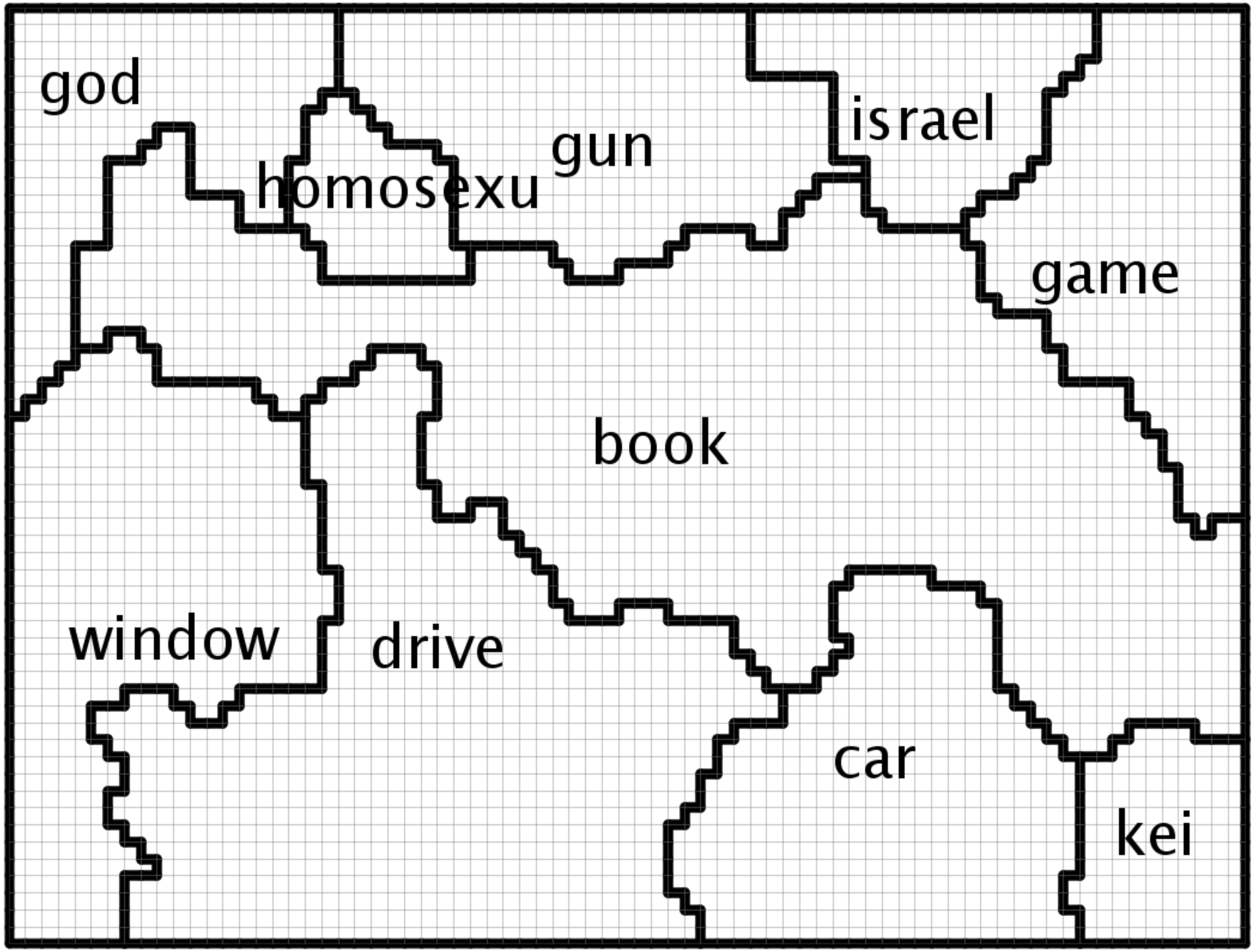
20 Newsgroups: Ward+Labels



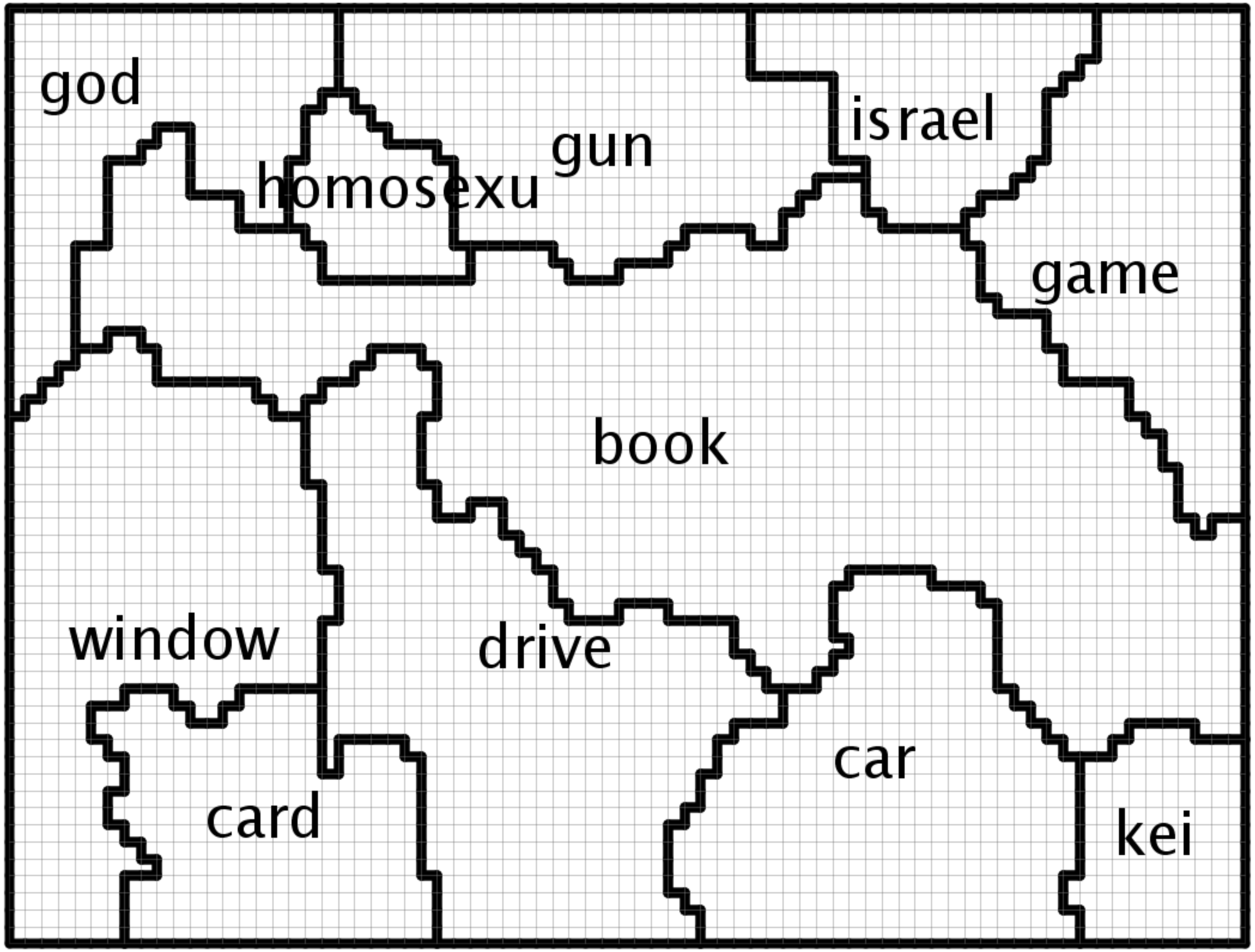
20 Newsgroups: Ward+Labels



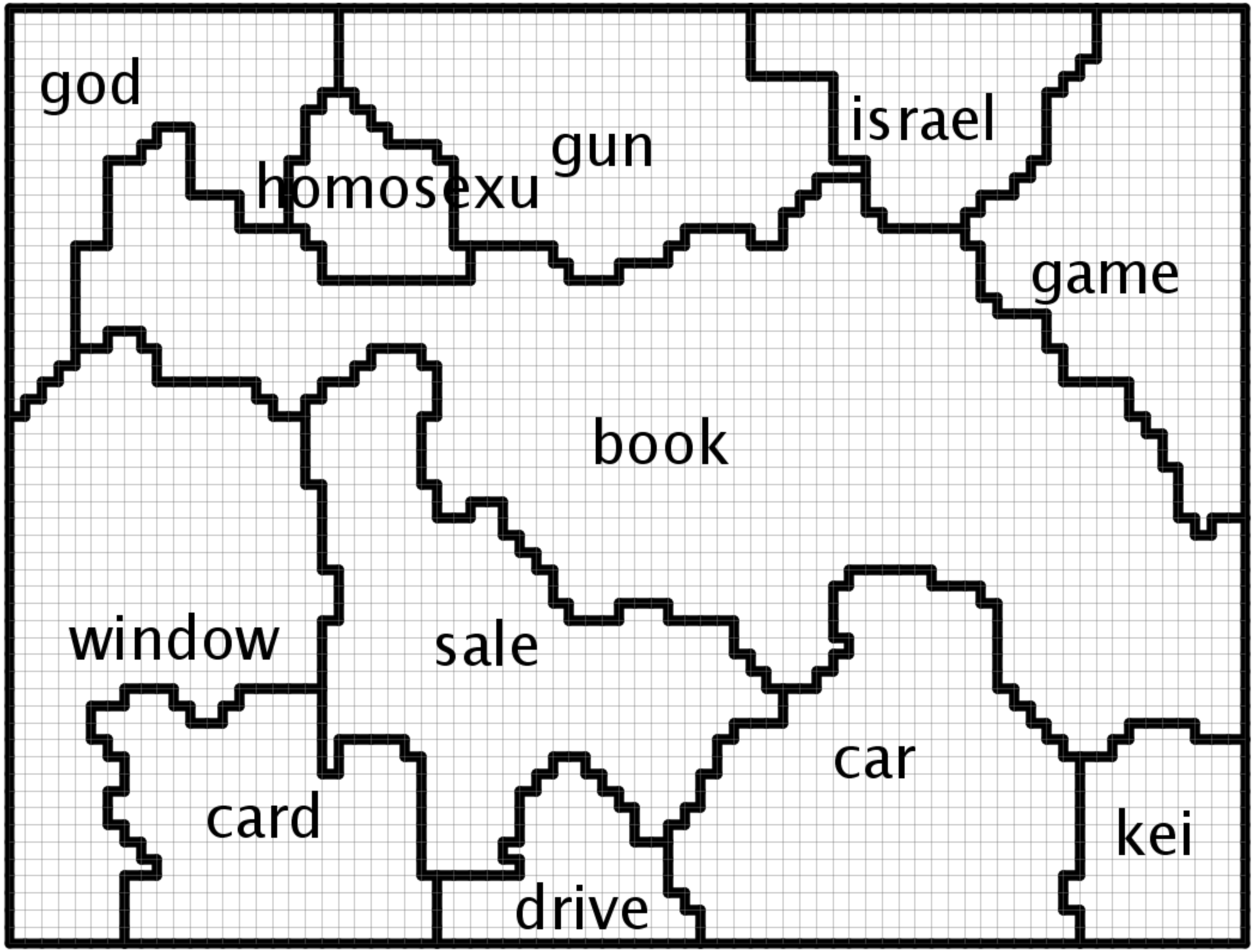
20 Newsgroups: Ward+Labels



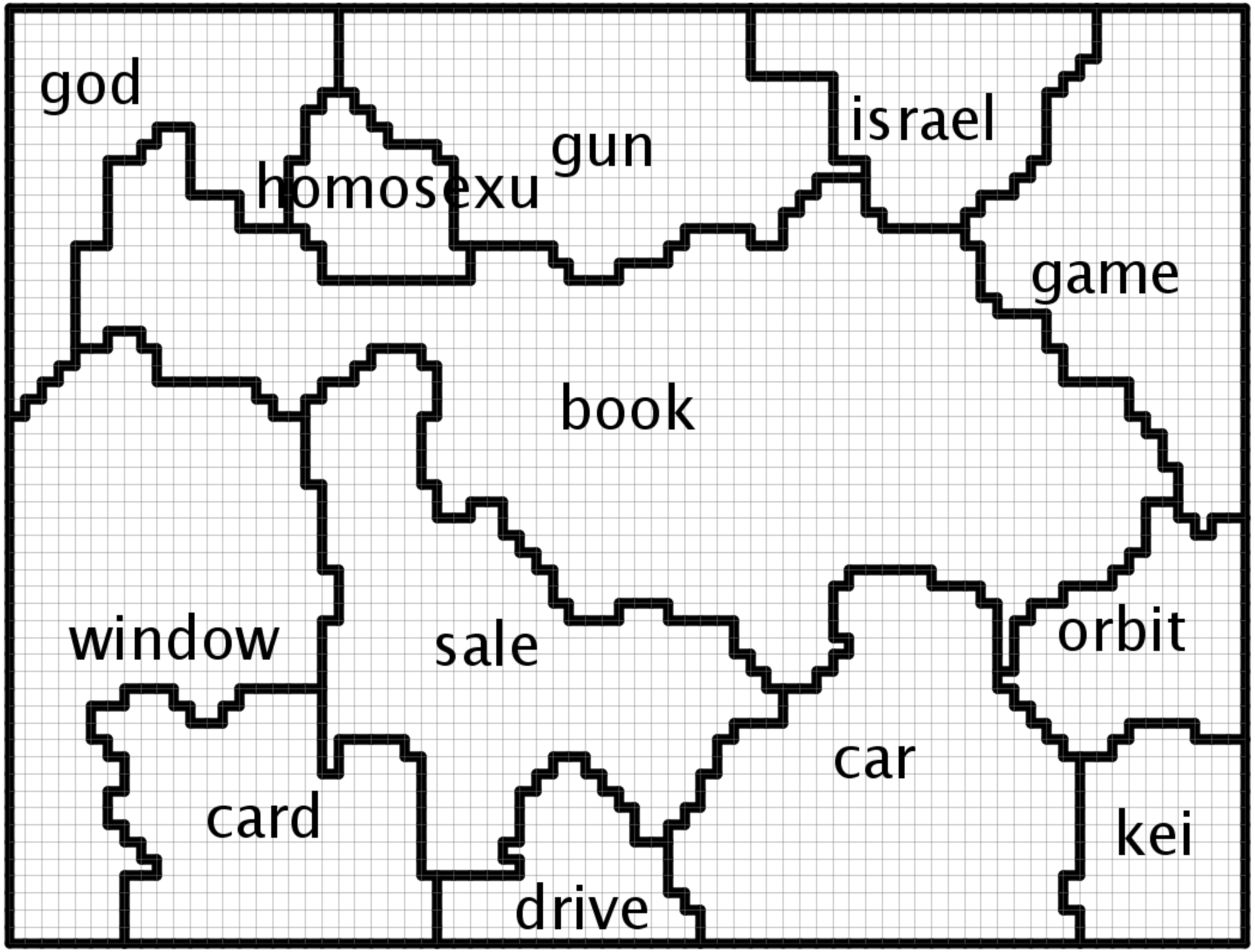
20 Newsgroups: Ward+Labels



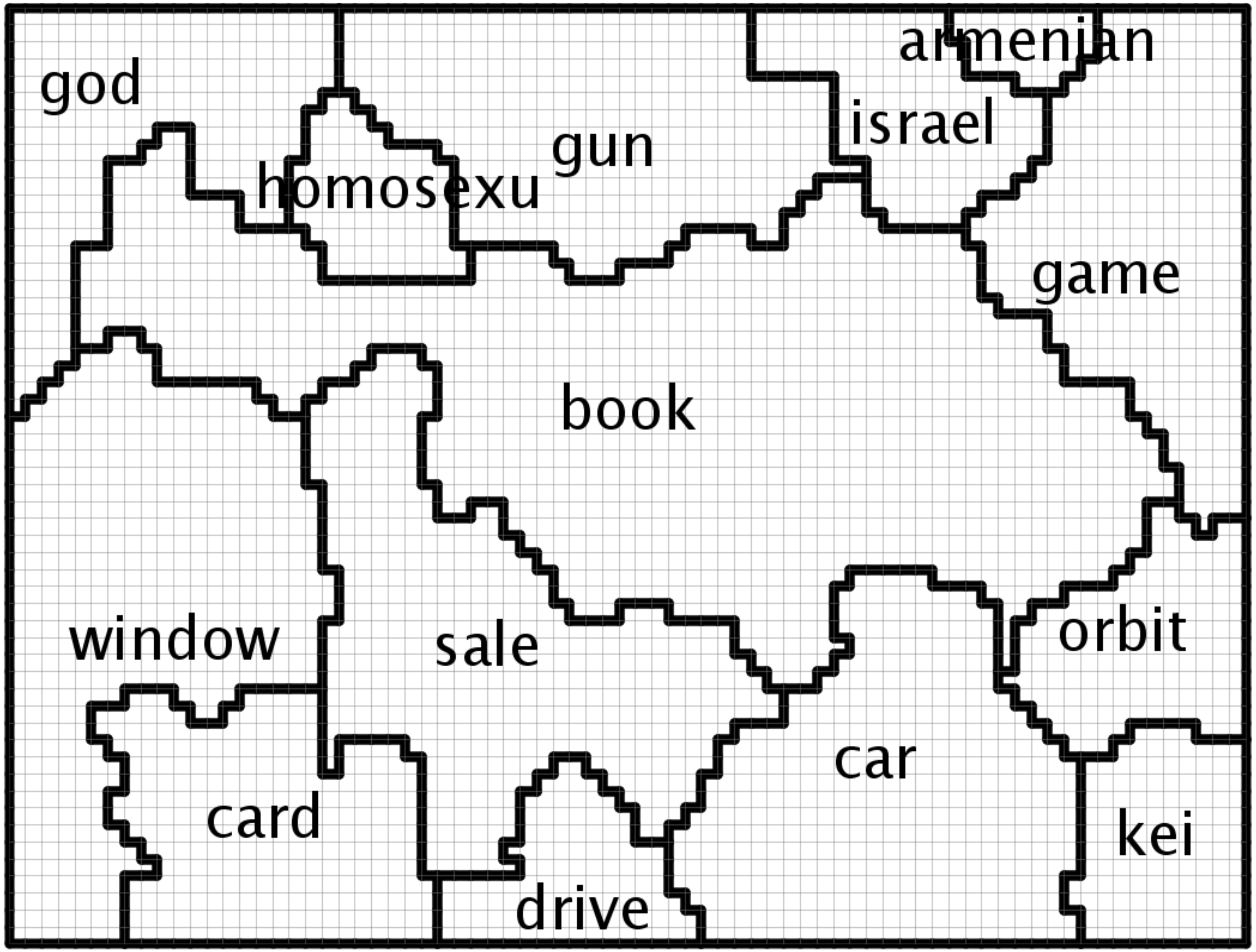
20 Newsgroups: Ward+Labels



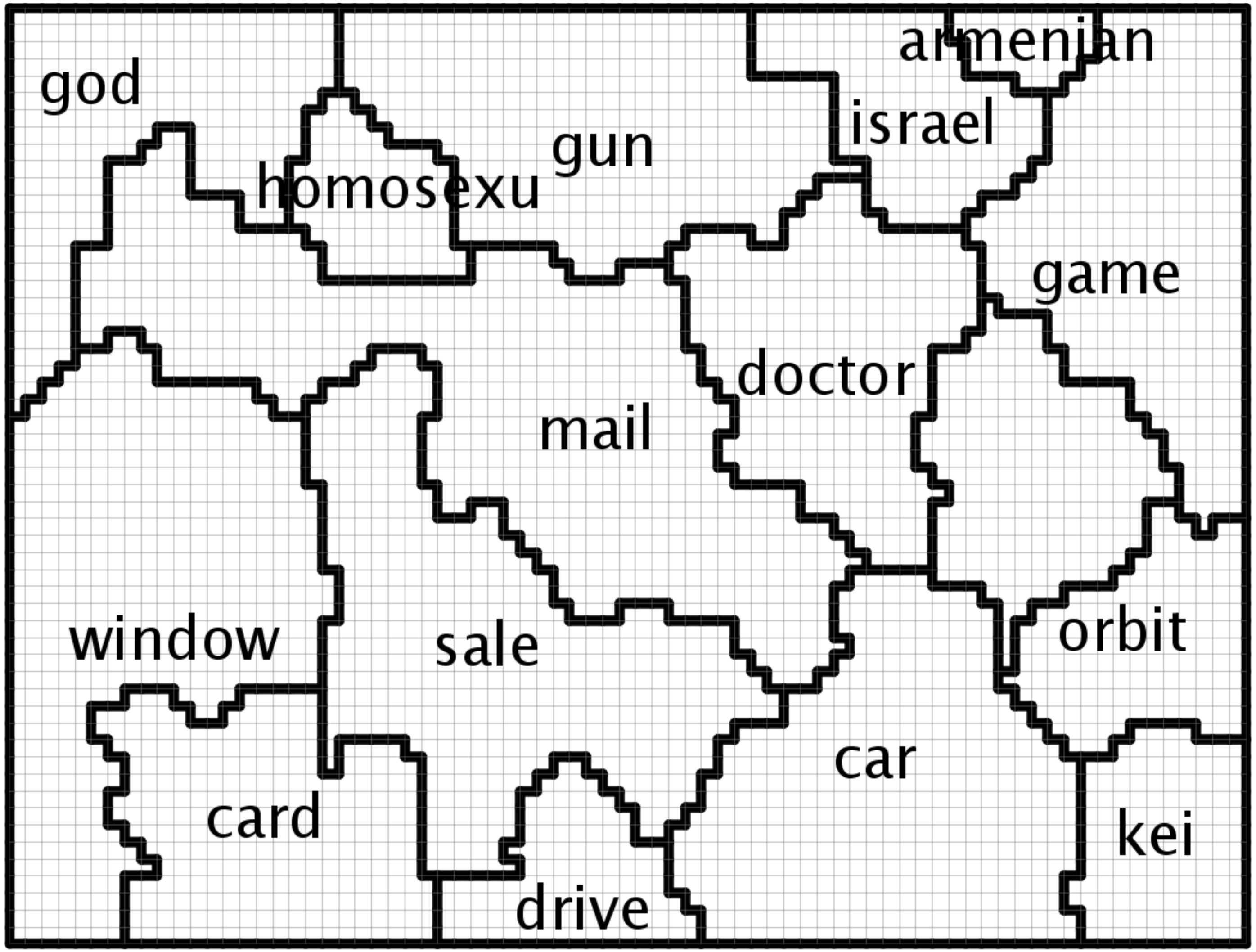
20 Newsgroups: Ward+Labels



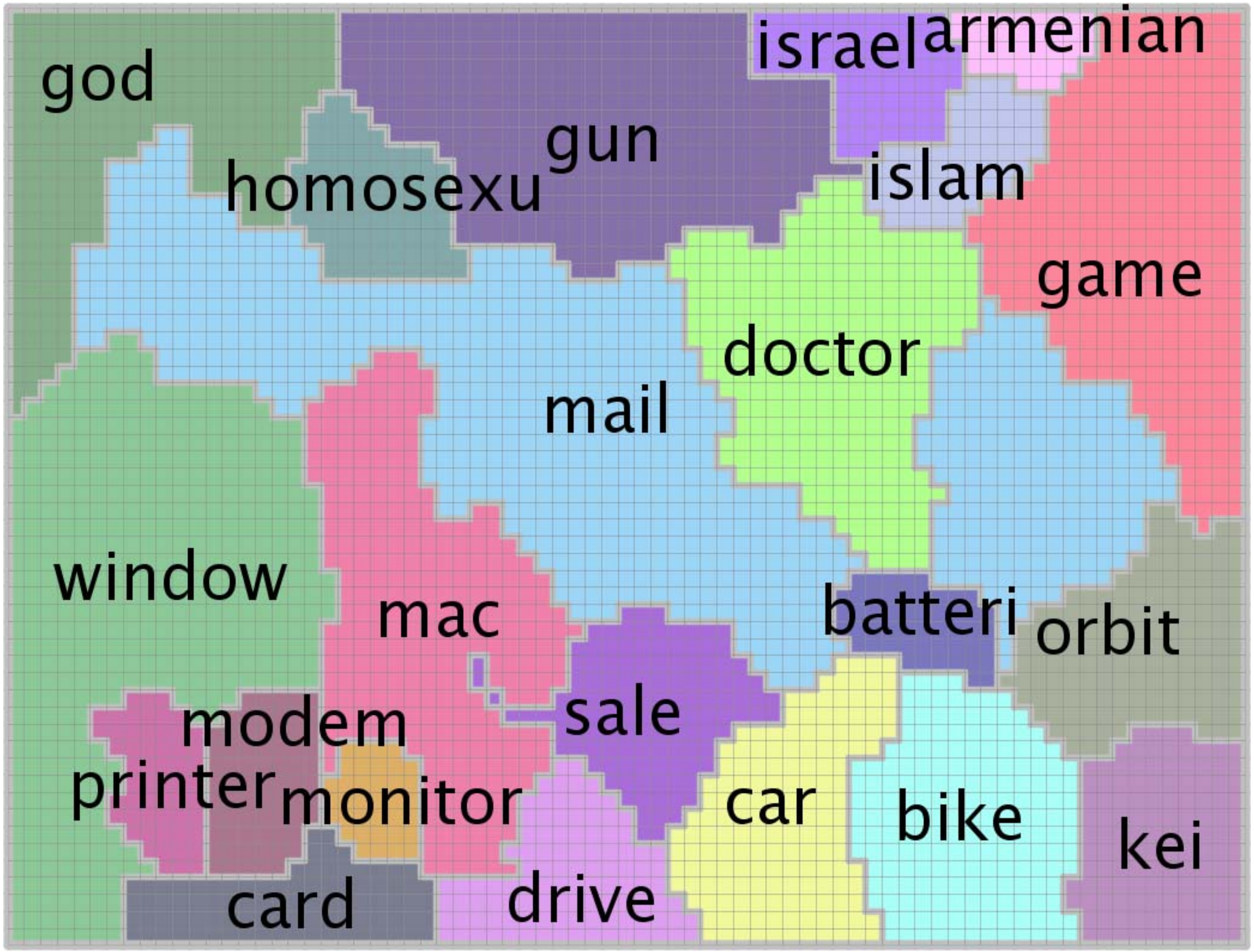
20 Newsgroups: Ward+Labels



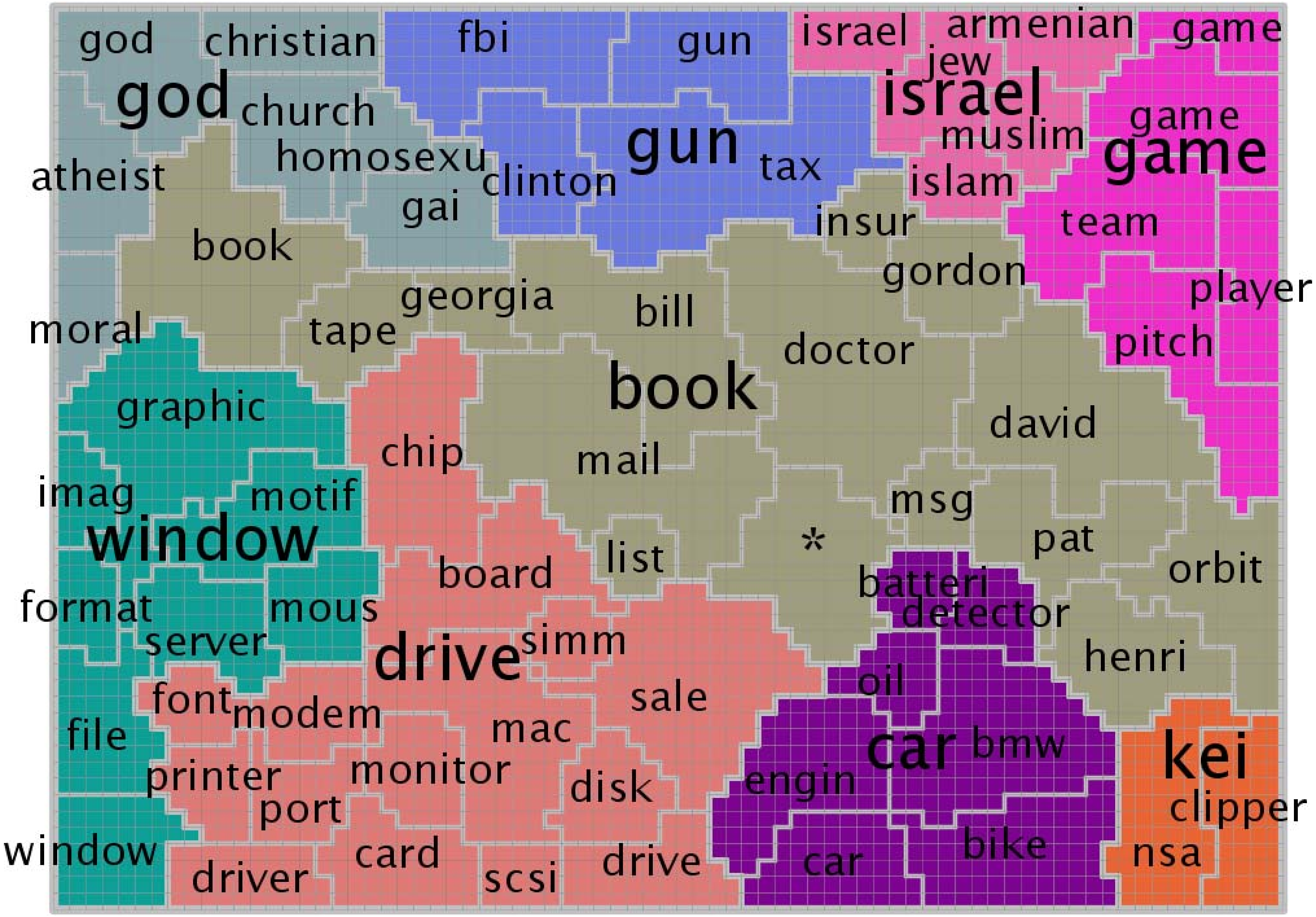
20 Newsgroups: Ward+Labels



20 Newsgroups: Ward+Labels

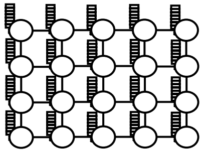


20 Newsgroups: Ward+Labels

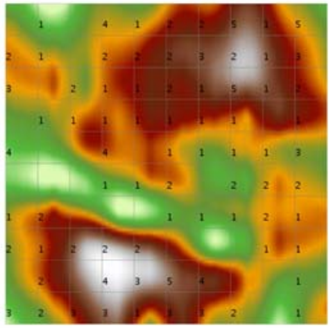


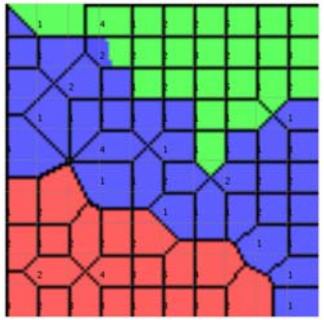
Outline

- SOM Basics

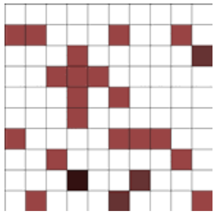


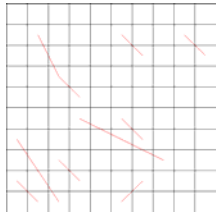
- Visualisierungen



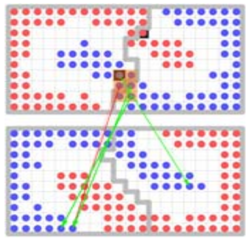


- SOM Qualitätsmaße

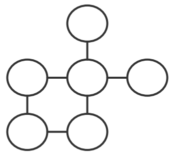


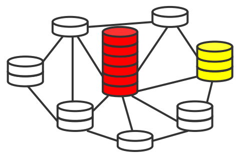


- SOM Comparison



- Verwandte Verfahren & Architekturen





- Applikationen