M.C. Thrun

Cluster Analysis of the World Gross-Domestic Product Based on the Emergent Self-Organization of a Swarm

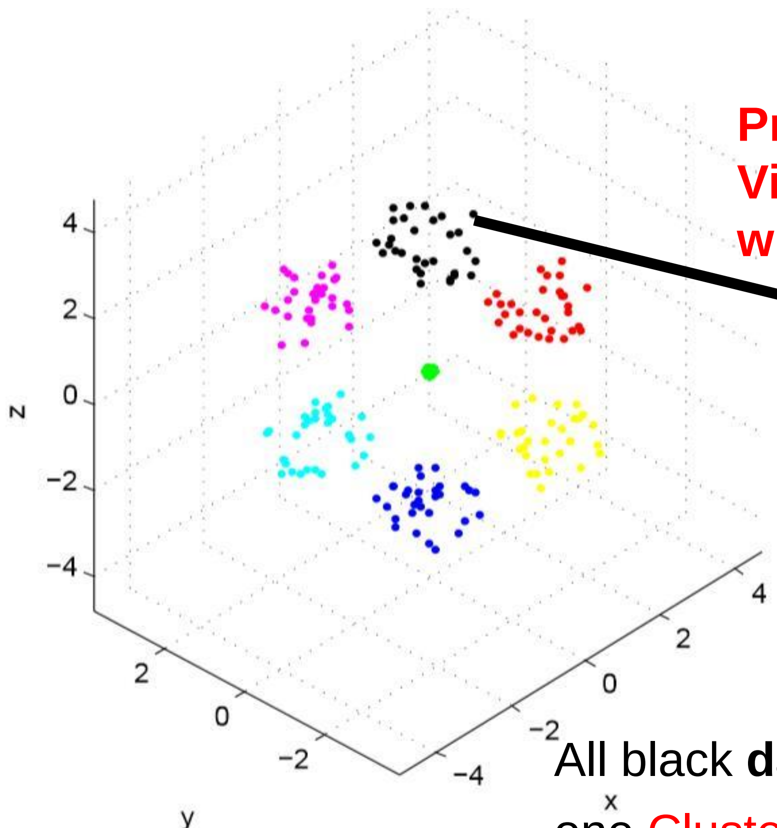
Feel free to contact me through www.deepbionics.org

Motivation

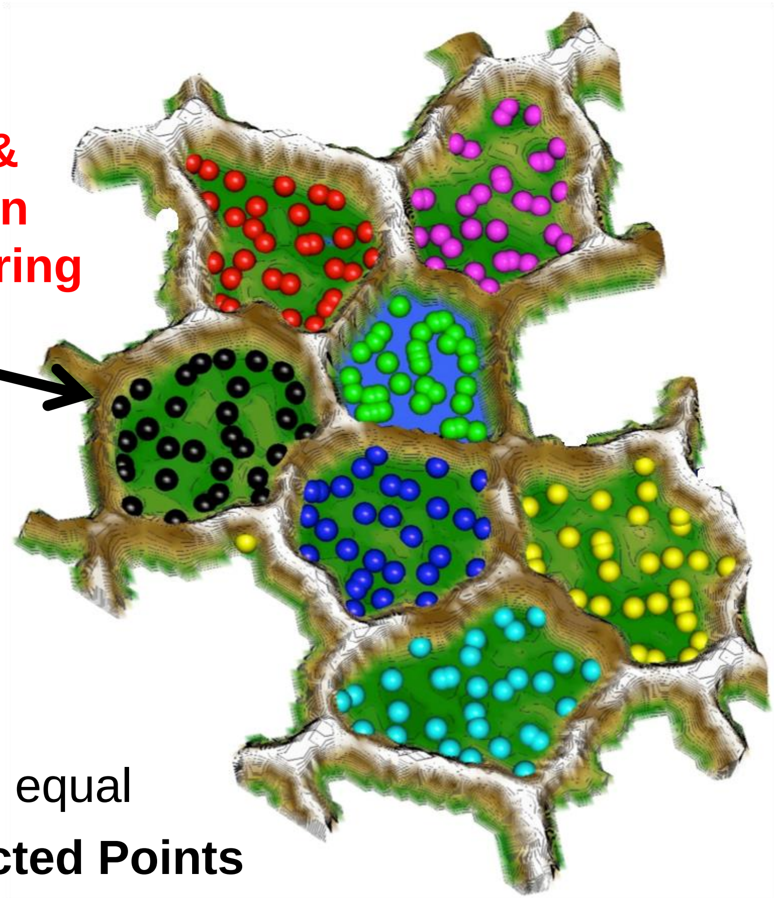
Problem: Separate data into similar groups -> **Clustering**

Goal: detect meaningful cluster structures

Solution: Project high-dimensional data in two dimensions and visualize as 3D landscape



Projection & Visualization with Clustering



All black **data points** equal one Cluster of **Projected Points**

Background

- Clusters can be of arbitrary shapes (structures) (1)
  - No generally accepted definition of clusters exists in the literature (2)
  - Number of clusters difficult to estimate
- Projection methods and clustering methods are separatly used

Implicit assumptions about structures of data are made by

- Clustering criterions (3)
- Projection methods (besides ESOM) (4)
- Quality measures (QMs) for projection methods (4)
- Quality assessments for clustering methods in the case of unknown class labeles (4),(5)

What happens if the structures in the data are unknown?

(1) [Jain/Dubes, 1988]; (2) [Hennig et al., 2015, p. 705]; (3) [Duda et al., 2001; Everitt et al., 2001; Handl et al., 2005; Theodoridis/Koutroumbas, 2009; Ultsch/Lötsch, 2016];

(4) [Thrun, 2017], (5) [Handl et al., 2005]

Challenges and Questions

- 27 Clustering algorithms on 15 datasets confirm hypothesis, see

  http://www.deepbionics.org/Projects/ClusteringAlgorithms.html

- Does the structure defined by a cluster algorithm lead to consistent insights

- How can a cluster analysis be performed on a data set of unknown structures without prior assumptions?

- Maybe the default parameter settings were incorrect?
  - How to choose the right parameter setting of more elaborated methods (e.g. t-SNE, Spectral Clustering)

=> Search for alternative concepts in literature

Concepts for Databionic Swarm

- Swarm Intelligence (1)
- Self-Organization
  - Swarms (2)
  - Self-organizing map (SOM), (3)
- Bionics (4)
  - Used in prior works (5)
- Applied Game Theory (6)
- Applied Emergence (6)

“the application of biological methods and systems found in nature”

(1) [Beni 1989]; (2) [Bonabeau/Dorigo et al., 1999]; (3) [Ultsch 1992];
(4) [Deneubourg 1991, Reynolds, 1987]; (5) [Herrmann 2007];
(6) [Thrun, 2018]: https://www.springer.com/la/book/9783658205393

What is Swarm Intelligence (SI)?

- In general the definition of intelligence is controversial (1) and complex (2)
- In the context of swarms, behavior and intelligence are used synonymously
- **Collective behavior** generically denotes any behavior of agents in a system having *more than one agent* (3)
- Five principles of swarm behavior (4)

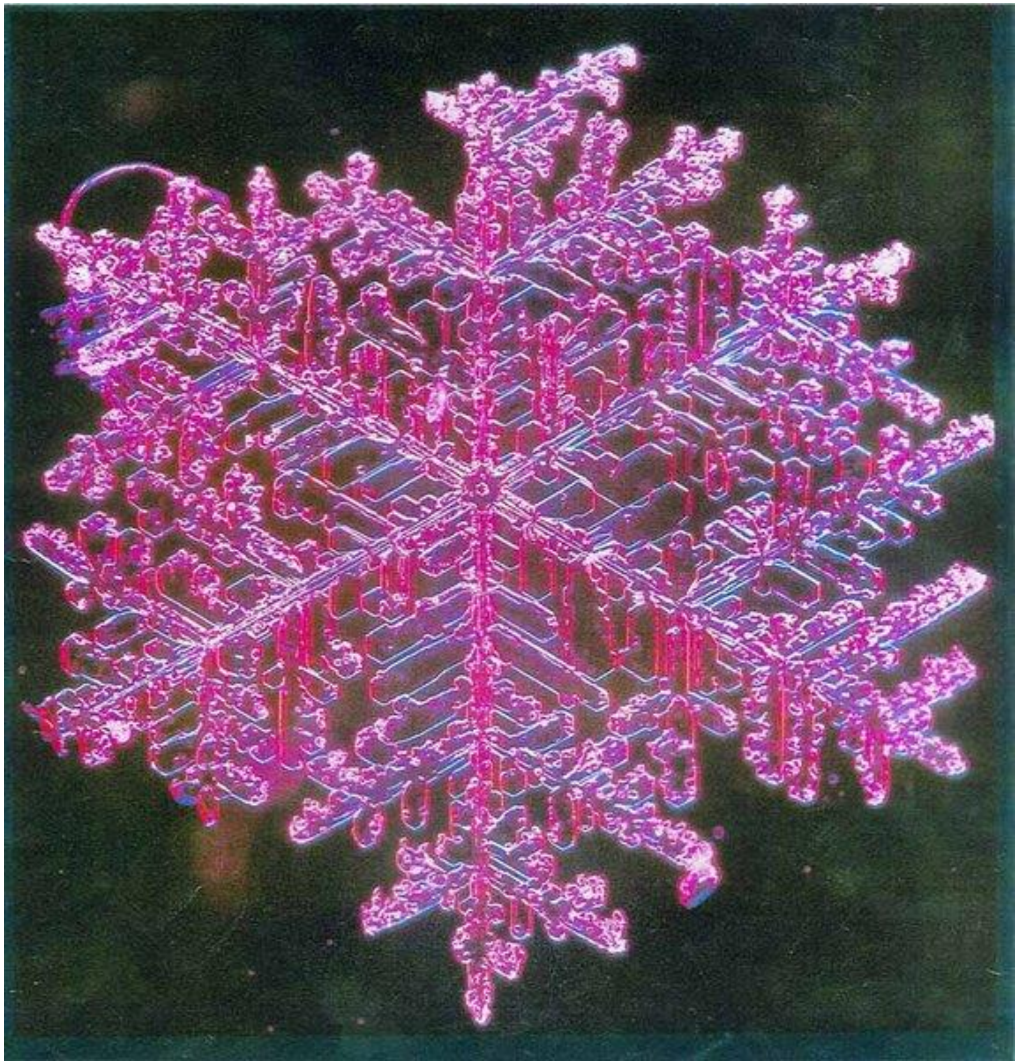
(1) [Legg/Hutter, 2007]; (2) [Zhong, 2010]; (3) [Cao et al., 1997]; (4) [Grosan et al., 2006; Reynolds, 1987].

Self-organization (SO)

Self-organization is defined by spontaneous pattern formation by a system itself, without responsibility of any determinate inside agent.

- Four basic ingredients for SO in a swarm (1)
- A swarm using SO should have more than 100 agents (2)

Ice flake



(1) [Bonabeau et al., 1999]; (2) [Beni, 2004]

Emergence

- An ability of an system
- The arising of novel and coherent properties during the process of self-organization (1)
- Four factors lead to emergence in swarm

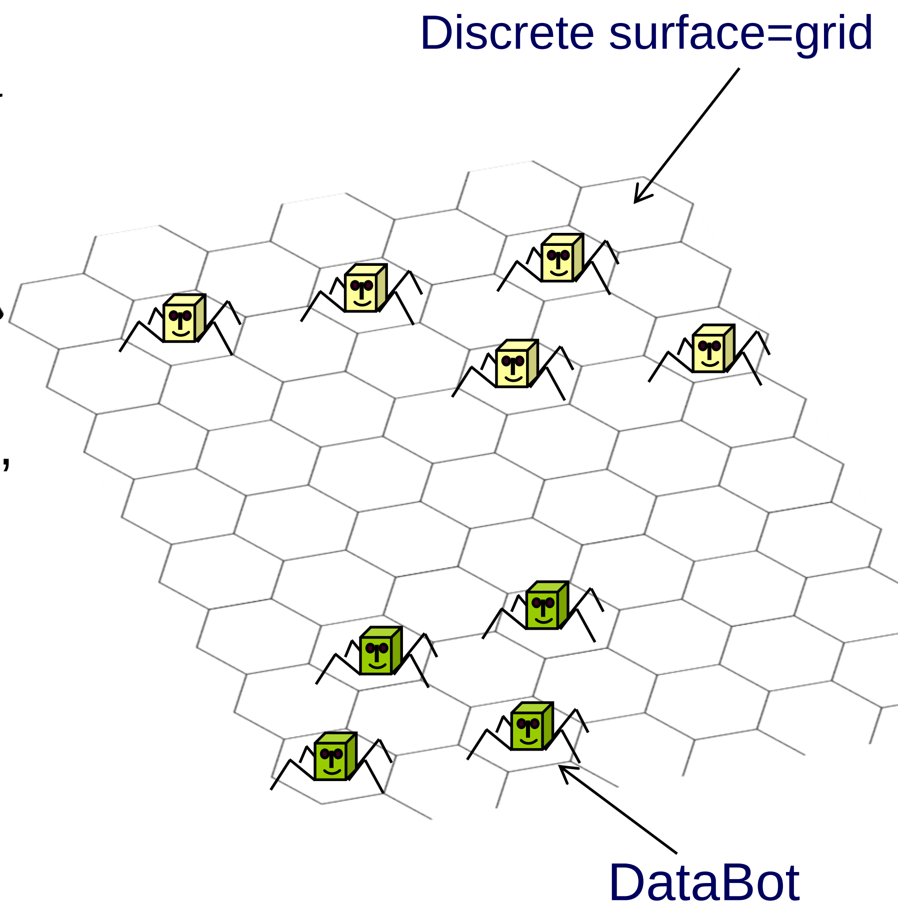
Example: One H_2O molecule

-> Wetness of Water

(1) [Goldstein 1999, Ultsch 1999, 2007]

Bionics of Databionic Swarm (DBS)

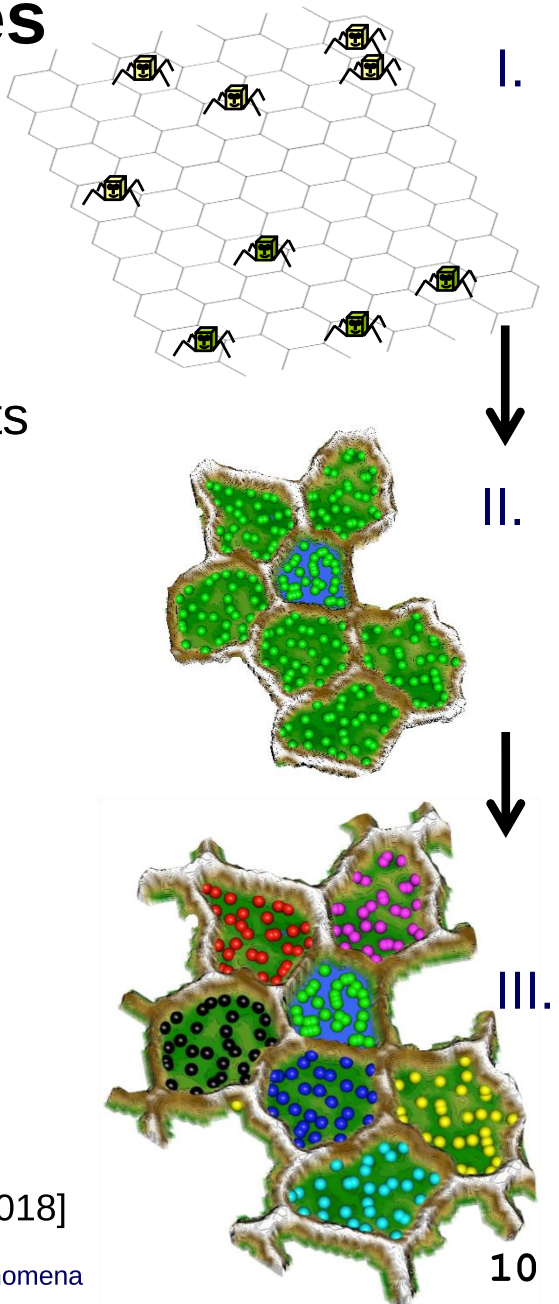
- Observation of Ants
  - Living in and moving on a flat toroidal surface, wearing one data point per agent (4)
  - Communication: Scent (3)
  - Smelling the surroundings of ones place (1)
  - DataBot moves to a free positions, if it *prefers* the scent of the new position (2)
  - Preference through an application of Game theory (5)



(1)[Herrmann 2007]; (2) [c.f. Schelling 1971];

(3) [Herrmann/Ultsch 2008a]; (4) [Ultsch 1999], (5) [Thrun, 2018]

DBS - Three Interchangeable Modules



I. Projection with Pswarm (3)
  - Application of Self-Organization, Swarm Intelligence, Bionics and game theory, parameter-free

II. Visualization: topographic map with hypsometric tints
  - Emergence through a unsupervised artificial neural network, parameter-free

  -> Results in 3D landscape generation and 3D printing (1)

III. Semi-automated Clustering on the visualization
  - U-matrix is approximation of Abstract Umatrix (2)

    which is based on Voronoi cells

    -> Shortest Paths (Djikstra) of Delaunay graph of projected points weighted with high-dimensional Distances used for a hierarchical clustering approach (3)

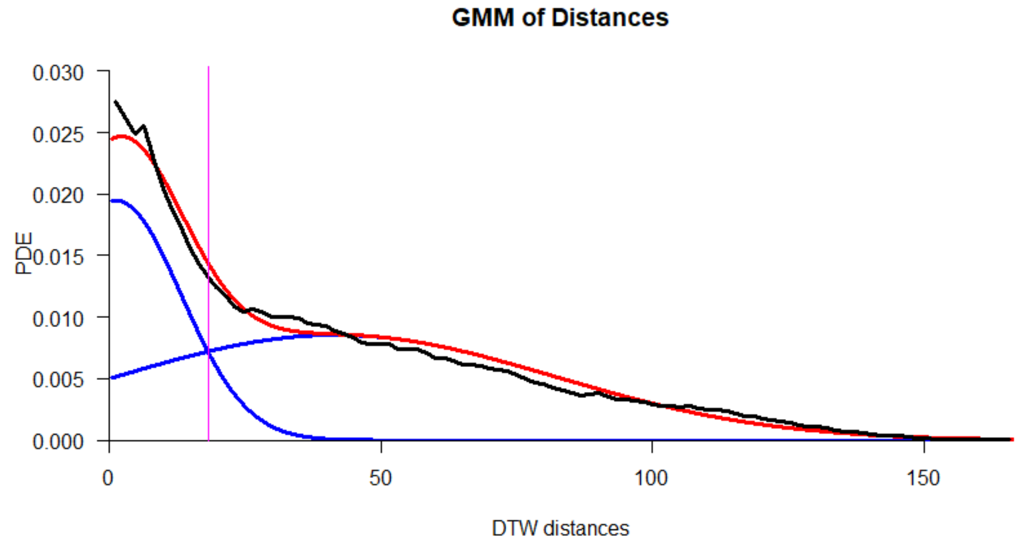
(1) [Thrun et al., 2016, Ultsch/Thrun 2017]; (2) [Lötsch/Ultsch 2014]; (3) [Thrun, 2018] https://cran.r-project.org/web/packages/AdaptGauss/index.html

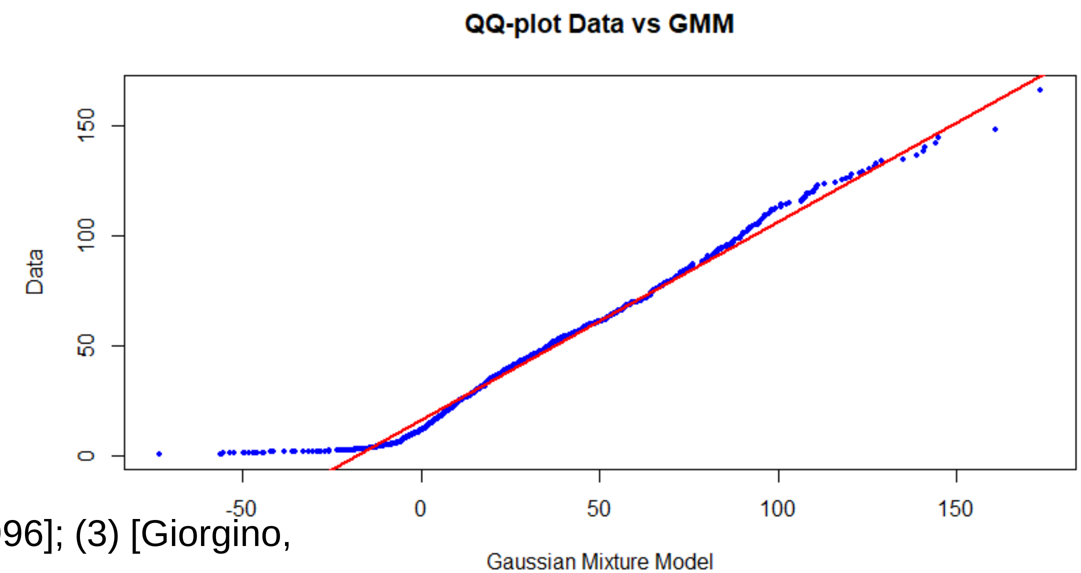
World Gross Domestic Product (GDP)

- PPP-converted GDP per capita (1) for the years from 1970 to 2010
- World GDP timeseries of 160 coutries was logarithmized
- For the distances D(l, j) the dynamic time warping (DTW) distances (2) were calculated using the R package dtw (3)
  - GMM has two modes

-> Distances can be clearly separated in larger inter-cluster distances and smaller intra cluster distances

-> Clear distance structure

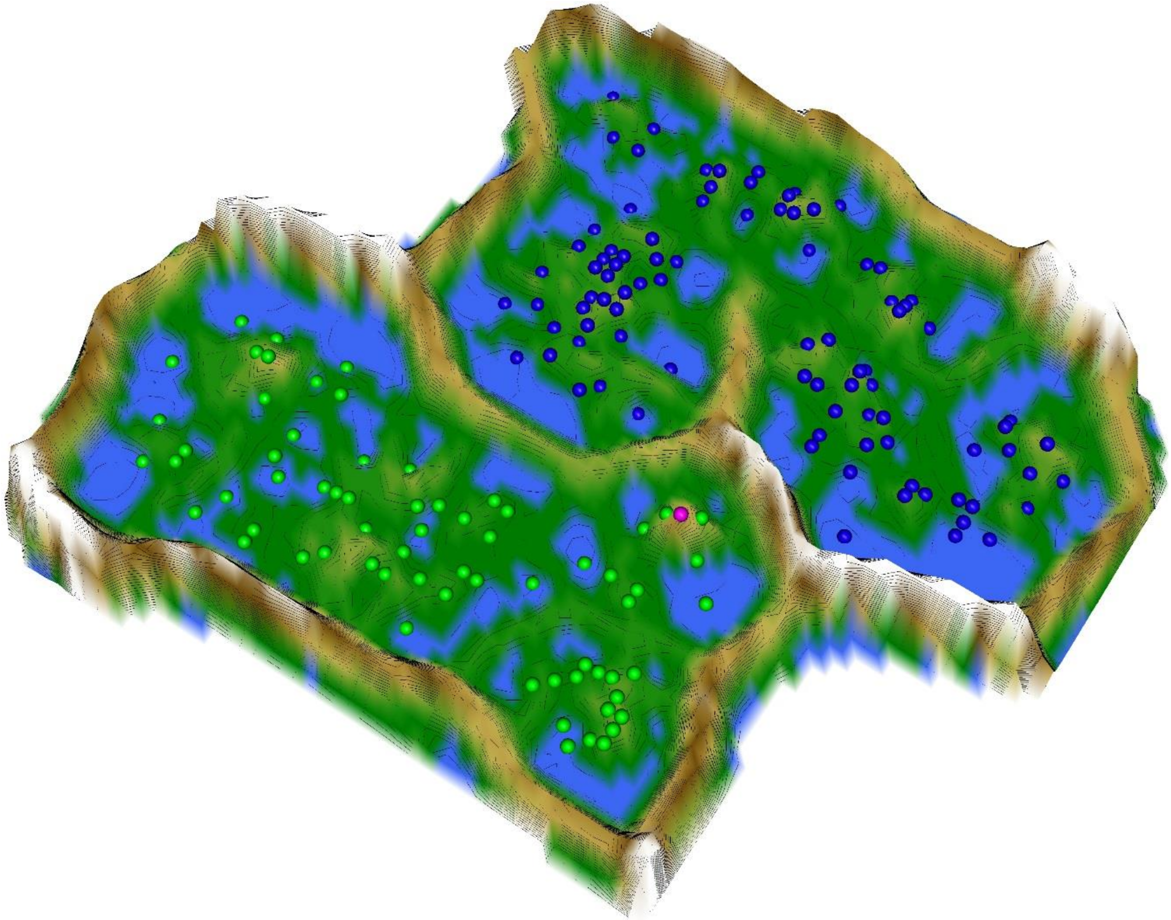




(1) [Leister, 2016]; (2) [Bernad & Clifford, 1996]; (3) [Giorgino, 2009]

Apply DBS to World GDP I

- Borders of the grid are cyclically connected with a periodicity
- Here we cut-out an Island
- Every Point symbolizes a country
- High-dimensional distances are visualized of the low dimensional projected points
- If the mountain is high the distances are large
- If the valley is low the distances are small
- Hypsometric tints: colors are height dependent



https://www.springer.com/la/book/9783658205393

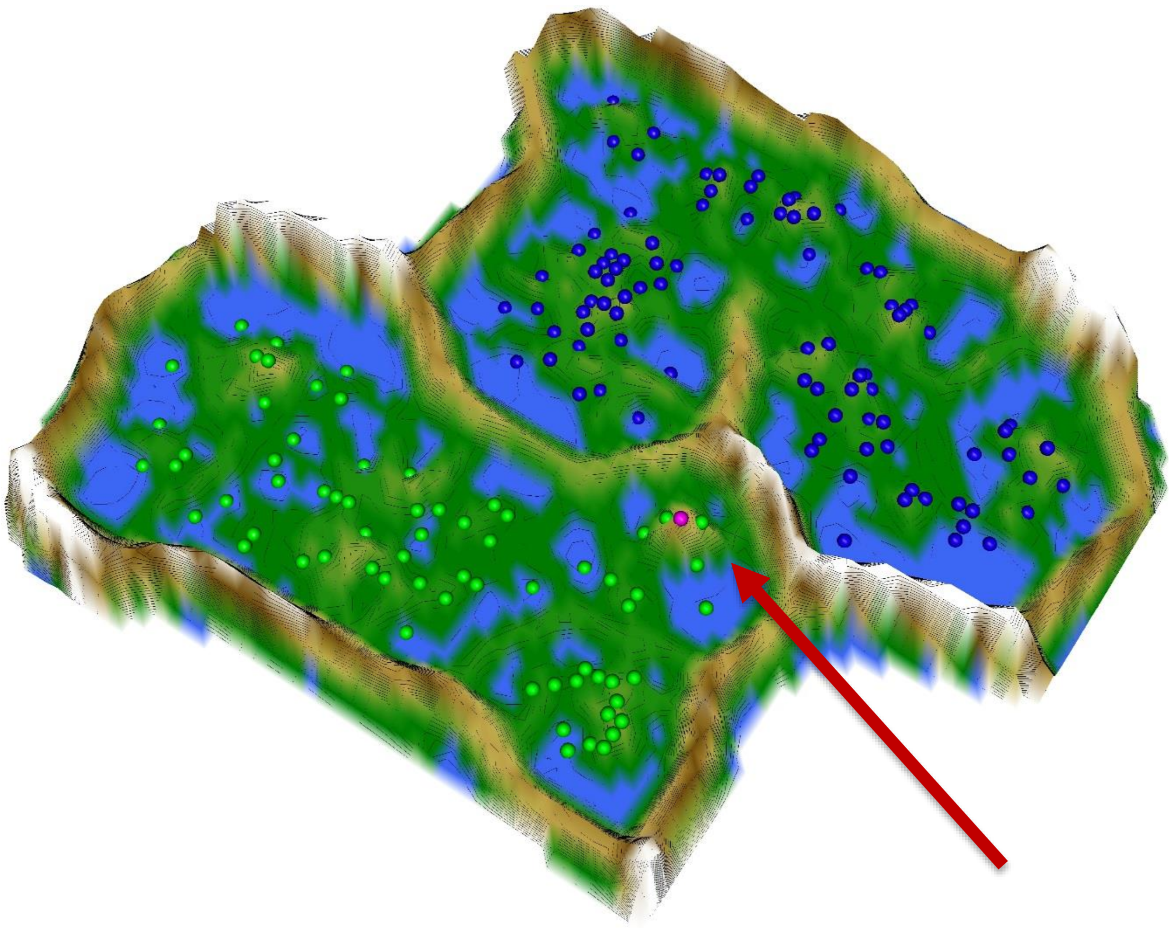
https://cran.r-project.org/web/packages/DatabionicSwarm/index.html

Apply DBS to World GDP II

Topographic Map:

- Valleys and basins indicate clusters
- Watersheds of hills and mountains indicate borderlines of clusters

⇒ **Number of Clusters is number of valleys!**



World GDP:

- High-Dimensional Disonctinuites are still visible through hills and valleys
- One outlier

https://www.springer.com/la/book/9783658205393

https://cran.r-project.org/web/packages/DatabionicSwarm/index.html

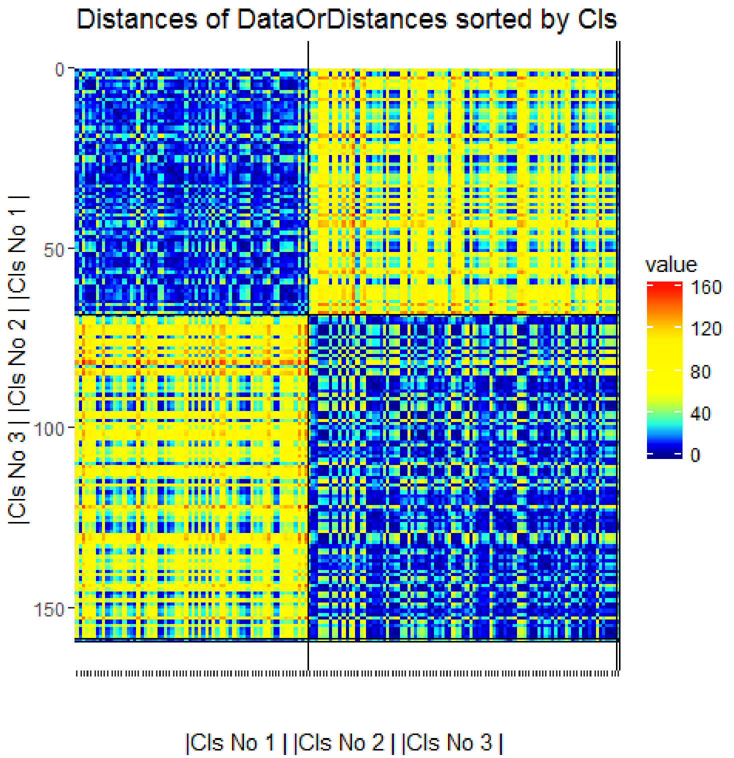
External Verification of Cluster Homogeneity

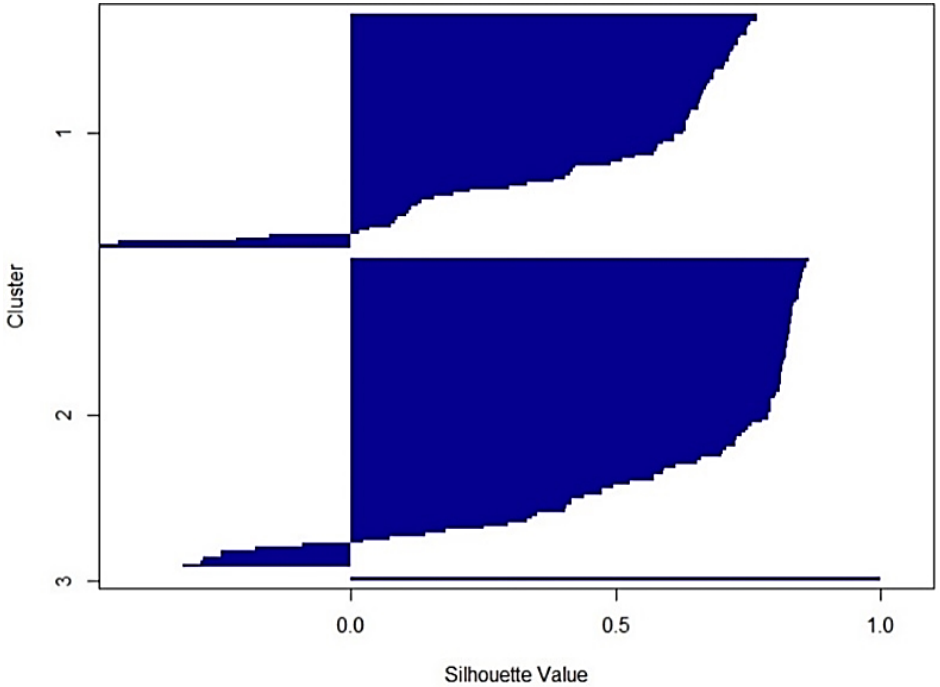
- We are searching for similar countries

-> Heatmap shows intra vers inter-cluster distances (cd)

indicates that intra-cd are small and inter-cd large

-> Silhouette plot indicates approx. spherical cluster structures

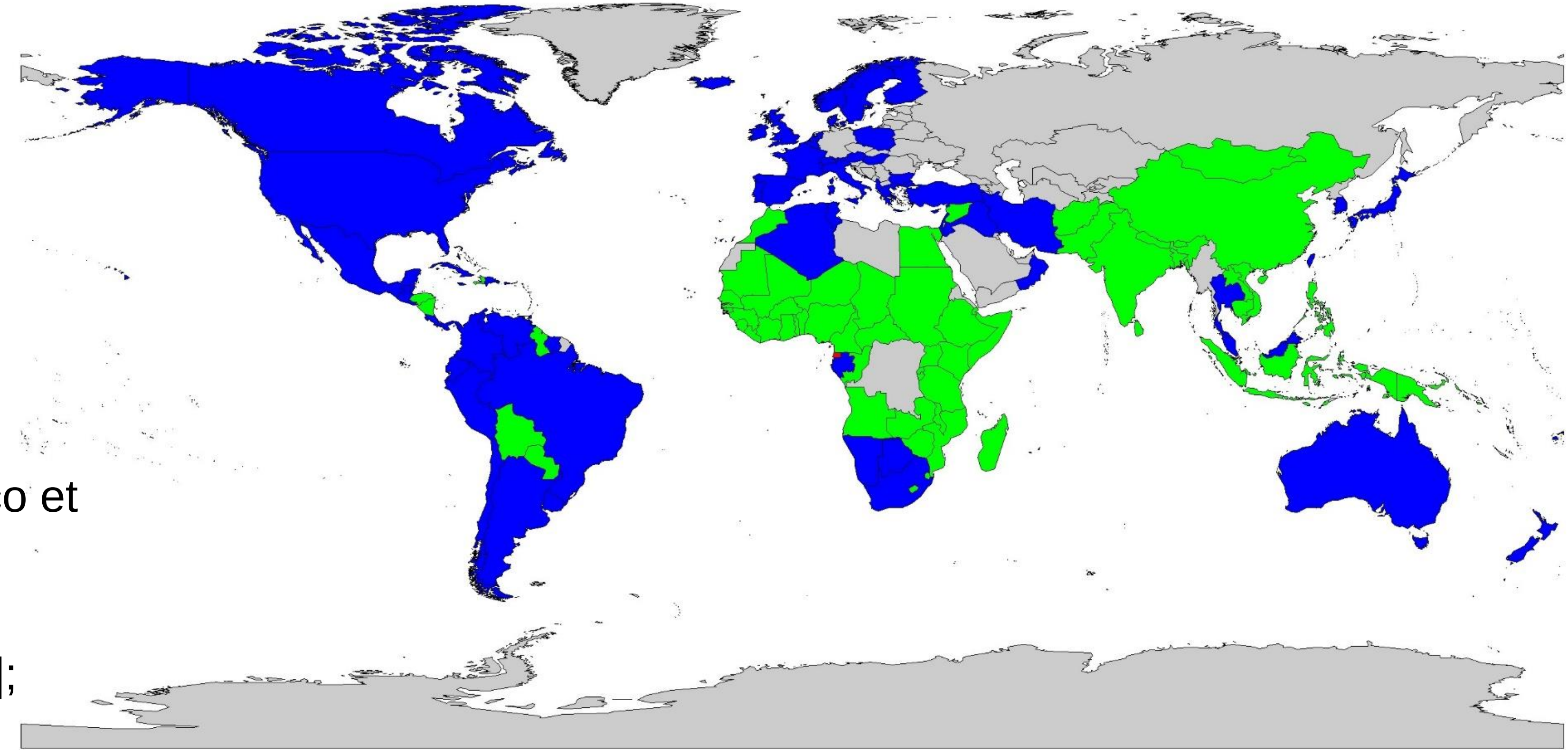




https://cran.r-project.org/web/packages/DataVisualizations/index.html

Geographical distribution

- Colors of countries in blue and green as in clustering
- Coherent geographical distribution of countries by clustering
- Priory cluster analysis on GDP datasets was performed for
  - Latin American countries (1) and European countries (2)



(1) [Redelico et al., 2009];

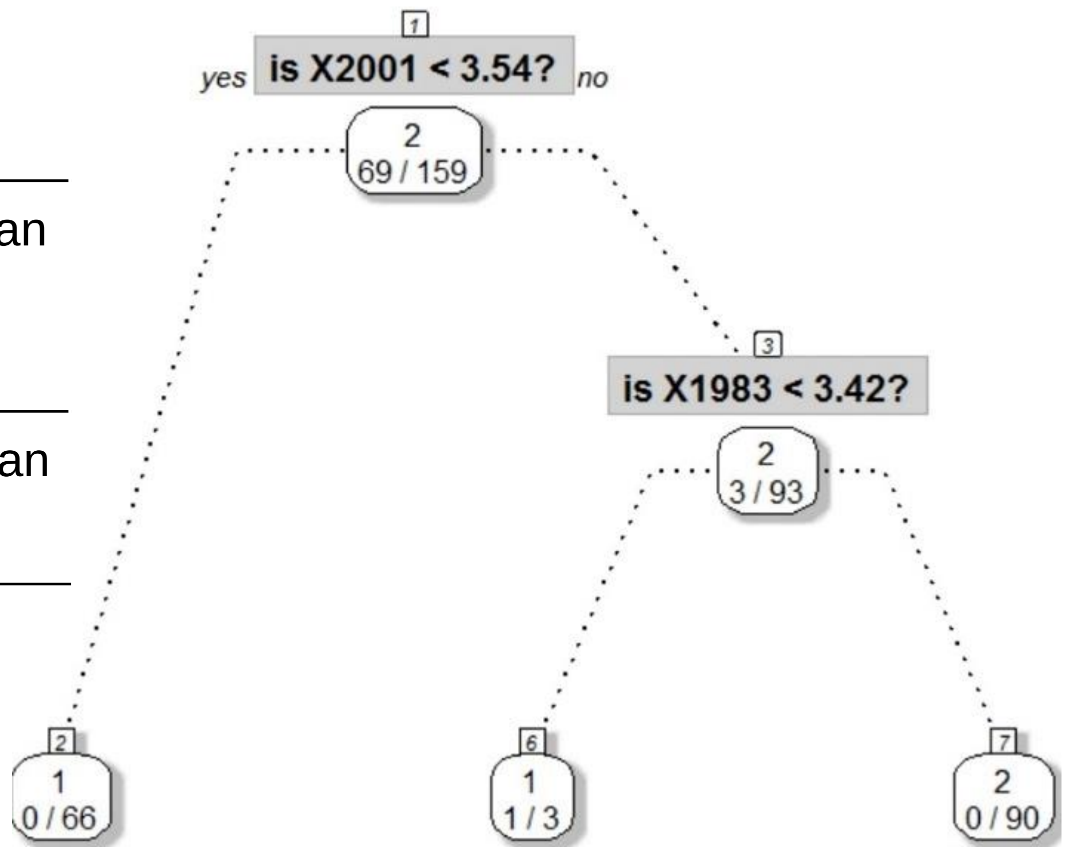
(2) [Gallo & Ertur, 2003];

https://cran.r-project.org/web/packages/DataVisualizations/index.html

Do the clusters make sense?

- Explaining the cluster structure by CART
- Simple Rules can be extractedout of the tree.

Rule No.	DBS Cluster No.	No. of Nations	Rules for the year 2001
R1	1	66	GDP lower than 3469 U
R2	2	93	GDP higher than 3469 U



Discussion of Results

- First cluster consists mostly of African and Asian countries
- Second clusters of industrialized countries of predominantly Europe and America
- Outliers
  - Equatorial Guinea (in DBS)
  - Incorrectly classified countries Egypt and Micronesia (in CART)

-> GDP is sensitive by economic shocks (e.g., oil-price), number and the change of inhabitants

- Economic achievement of 157 countries was profoundly affected in the year 2001
  - Could be the crashing of airplanes into the World Trade Center

  -> World economy was experiencing its first synchronized global recession in a quarter-century (1)

(1) [Makinen, 2002, p. 17]

Databionic swarm

- DBS is a flexible and robust clustering framework
  - Three interchangeable modules
  - Swarm-based technique combining swarm intelligence, self-organization and emergence
  - Combined with a human-understandable visualization technique
- Parameter-free in projection and visualization

- Clustering/absence of clusters is verified by visualization (1)
- Number of cluster can be estimated by visualization

- Detects meaningful structure in the data
  - Emerging structures lead to new, unknown but useful knowledge in data

(1) https://www.springer.com/la/book/9783658205393

https://cran.r-project.org/web/packages/DatabionicSwarm/index.html

Drawbacks

- Stochastic projection (trial depending)

⇒ These problems exist also in most other common methods

- Realm of Big Data is not discussed here due to time complexity of Pswarm

If prior knowledge of the data set to be analyzed is available, then a projection method that is appropriately chosen with regard to the structures that should be preserved can outperform Pswarm

Outlook/Further Research

- If prior knowledge available

  => Projection based clustering (1)

  => Or use appropriate clustering method with
  topographic map for verification

- DBS Algorithm will be parallelized in future to be used Cloud Computing

(1) [Thrun/Ultsch, 2017b]
https://cran.r-project.org/web/packages/ProjectionBasedClustering/index.html

.

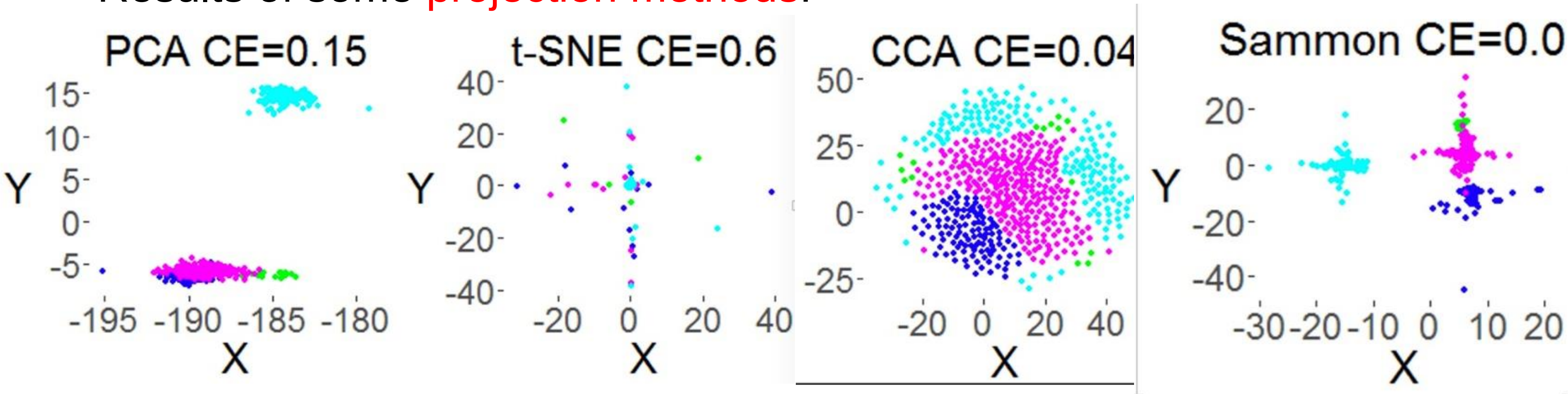
Thank you for listening.

Any questions?

Example: Leukemia data set

- 7747 gene expressions of 554 subjects (1)
- Prior classification is made availible by domain expert
- Data structures of this high-dimensional data set is unknown

Results of some projection methods:



Results of some clustering methods:

Algorithm	Ward	SL	kMeans	MoG	PAM	Spectral
Accuracy in %	100	80.1	76.53	Not Computable	78.3	59.0

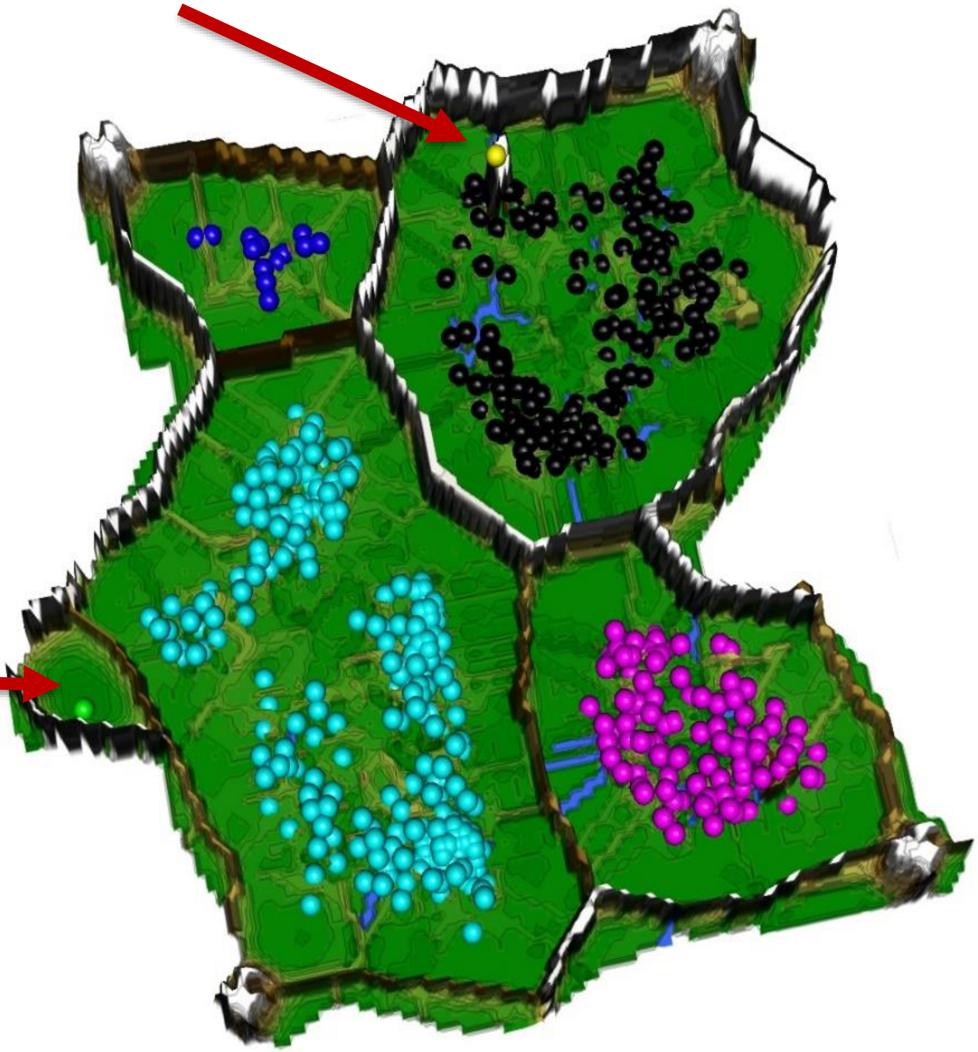
https://cran.r-project.org/web/packages/ProjectionBasedClustering/index.html

(1) [Haferlach et al., 2010]

Apply DBS to Leukemia data set

- ~7500 Dimensions visualized in 3 Dimensions
- High-Dimensional Disonctinuites still visible thorugh hills and valleys
- Types of leukemia diagnoses and healthy patients visible in different colors of projected points
- Accurarcy of 99.6%
- Two outliers

-> Possible problem in diagnosis



Key Idea's for DBS

- Clustering method

-> Should be based on visualization of the projection

- □ Visualization based on Umatrix for ESOM (1)

  => Allows to estimates number of clusters

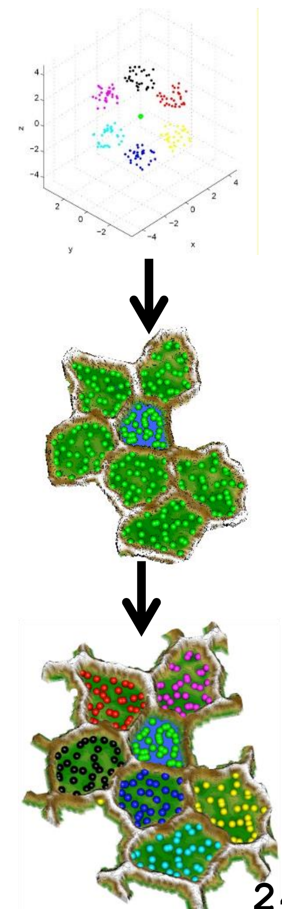
- Projection method
  - □ Annealing scheme with the decreasing neighborhood radius for DataBots (2)
  - □ No objective function, no parameters

Apply

I. *Physics:*
  - □ Given problem -> First search for symmetry
  - □ Solution results in parameter reduction and neighborhood definition

II. *Game Theory* (3)
  - □ Requirements for ***preference*** (using I.)
  - □ Annealing scheme (without parameters)

Goal: separate data points into similar groups("colors")



(1) [Ultsch 2003]; (2) [Kämpf/Ultsch, 2006]; (3) [Nash 1951]

Game Theory in Detail

- [Neumann/Morgenstern, 1953, p. 84]: A game is defined as a scenario with n players i=1, …, n in which each player makes a choice
- Let a game *G* be defined by *n* players associated with *n* non-empty sets $\Pi_1, \dots, \Pi_n$, where every set Π_i represent all choices made by player i; then, the pay-off function is defined as

$$p = (p_1, \dots, p_n): \Pi_1 \times \dots \times \Pi_n \to \mathbb{R}^n$$

- Mixed strategies may include the five main principles of collective behavior

Mixed strategies

- In a game with *n* players, let the k choices of player i be defined by a set $\Pi_i = \{\pi_1^i, \dots \pi_\alpha^i, \dots, \pi_k^i\}$, where π_α^i indicates the i^{th} player's α^{th} choice,

  then, a mixed strategy $s_j(i) \in S_i$ for player i is defined by

$$s_j(i) = \sum_{\alpha=1}^{k(i)} c_\alpha(i) \pi_\alpha(i)$$

where

$$\sum_{\alpha=1}^{k(i)} c_\alpha(i) = 1 \text{ and all } c_\alpha(i) \geq 0.$$

„Every Finite Game Has an Equilibrium Point" (1)

Let $t_j(i) \in S_i$ be the mixed strategy that maximizes the payoff for player i; then, the Nash equilibrium is defined as

$$p_i\Big(s(1),\ldots,s(i-1),t_j(i),s(i+1),\ldots,s(n)\Big) = \max_{t_j(i)\in S_i} p_i\Big(s(1),\ldots,s(n)\Big) \qquad (2)$$

if and only if this equation holds for every i [Nash, 1951]

- For a weak Nash equilibrium multiple mixed strategies for the same person that result in the same maximal payoff p_i,
- For a strong Nash equilibrium, even a coalition of players cannot further increase their payoffs by simultaneously changing their strategies
- A Nash equilibrium is not necessarily unique

(1) [Nash, 1951, p 288]

SI for Unsupervised Machine Learning

1. Particle Swarm Optimization (PSO)
   - Bionics: Bird flocking -> agents communicate directly
   - Normally applied as a population-based search algorithm [Rana et al., 2011]
   - Rule-based classification models, e.g. AntMiner, or as an optimizer within other learning algorithms
2. Ant Colony Optimization (ACO)
   - Agents communicate through stigmergy
   - Applied to the task of sorting [Martens et al., 2011]
   - **Ant Based Clustering (ABC)**
3. Artificial Behavior based on DataBots

- And some special cases like
  - Prey model [Stephens/Krebs, 1986], [Giraldo et al., 2011]

Four basic ingredients for SO in a swarm

1. *Positive feedback*
   - promotes a creation of convenient structures and helps to stabilize them
2. *Negative feedback*
   - promotes a creation of convenient structures and helps to stabilize them
3. *Amplification of fluctuations*
   - *Fluctuations* defined as errors, random movements and task switching
4. *Multiple interactions*.
   1. For swarm behavior to emerge, multiple interactions are required

[Bonabeau et al., 1999]

Five principles of swarm behavior

1. *Homogeneity,*
   1. every agent has the same behavior model
2. *Locality*
   1. the motion of each agent is only influenced by its nearest neighbors;
3. *Velocity matching*
   1. every agent attempts to match the velocity of nearby flock mates
4. *Collision Avoidance*
   1. every agent avoids collisions with nearby agents
5. *Flock Centering*
   1. the agents attempt to stay close to the neighboring agents

[Grosan et al., 2006]

Factors leading to Emergence

- The three factors leading to emergence in swarms are
  1. Randomness
     - Uses a source of random numbers in its calculations (non-determinism) [Ultsch, 2007].
  2. Temporal and structural unpredictability

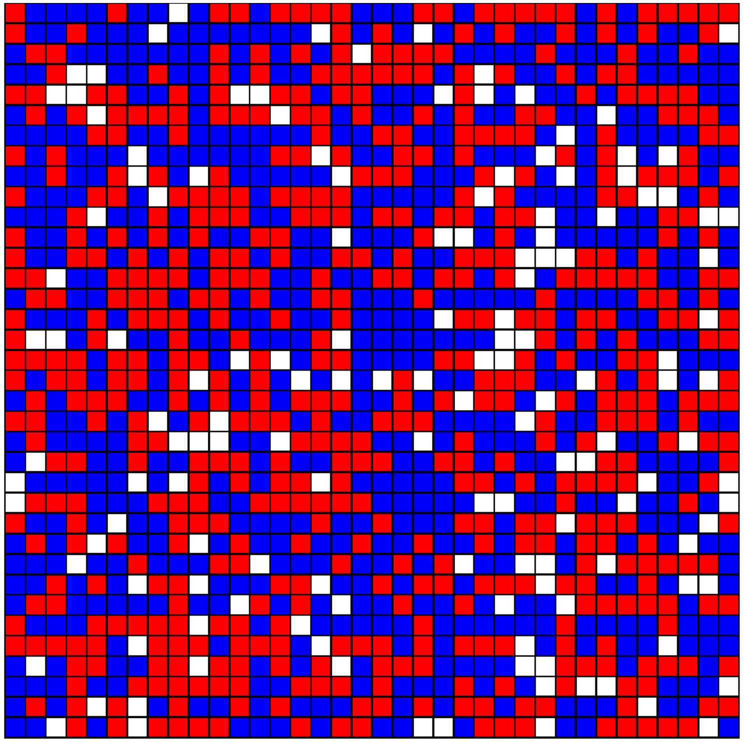
     => No objective function.
  3. Multiple non-linear interactions among **many** agents
     - Many elementary processes are required
     - Nonlinearity means that adding or removing interactions among agents or any agents themselves results in behavior that is linearly unpredictable
       - e.g. Adding/Removing DataBots
- For applications, the existence of emergence is irrelevant. Even if emergent phenomena can be causally explained, they can still be used in the future.

Example for Emergence: Schelling‘s Model

- *The Schelling model consists of a grid of square patches.*
- *Agents are located on this landscape, initially at random, with no more than one on* any patch.

[Schelling, 1969, 1971]

Agents: red or blue squares



Free positions: white squares

Schelling‘s Modell – Live Example

- Mild preference of the agent’s own color, results in segregation
  - Each agent has a tolerance parameter. Green agents are “happy” when the ratio of greens to reds in its **Moore neighborhood** (the eight immediately adjacent cells or patches) is more than its tolerance
  - Unhappy agents are allowed to move

http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/

Settings:

Similar 66

Red/Blue: 50/50

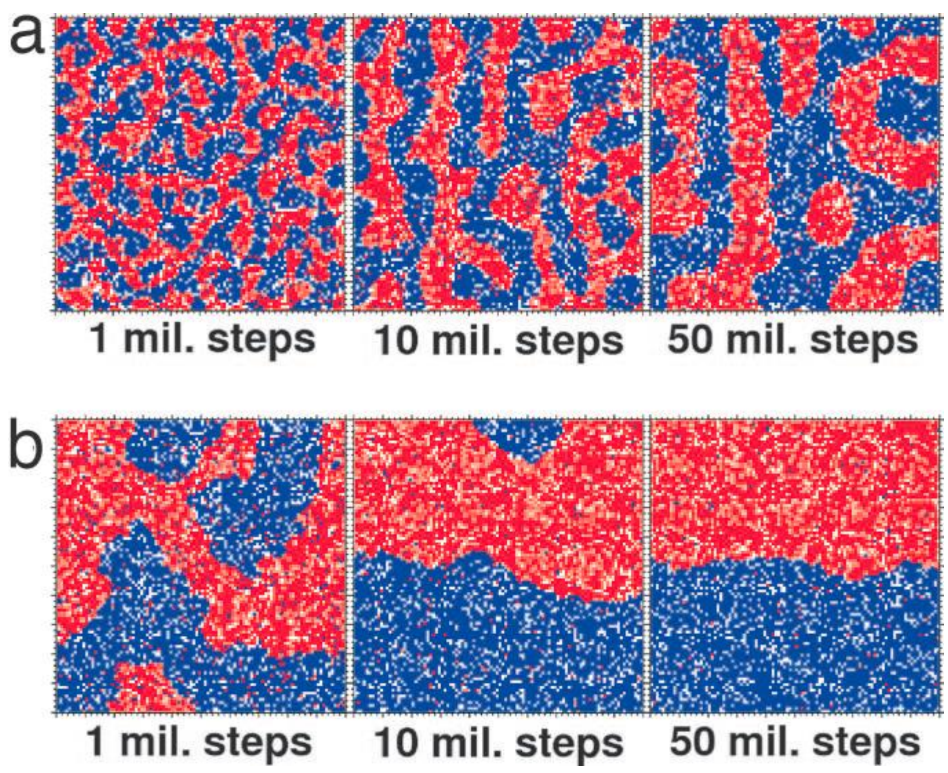
Empty: 10

Size 36x36

Delay 11ms

Schelling‘s Modell

- Behaviour Explanation of emergence of Ghetos:
  - Initially integrated communities changed to full segregation
  - Even if the people’s happiness rules expressed only a mild preference for having neighbors of their own type

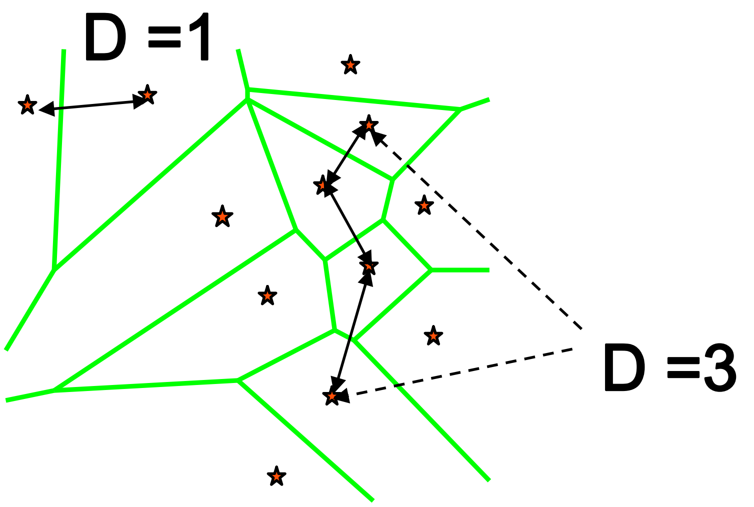


Abstract U-distances

- Delaunay Graph: Graph of Voronoi cells
  - A region corresponds to each BestMatch consisting of all points closer to that BestMatch than to any other
- Delaunay Path: Number of edges between two BestMatches
- Djikstra Shortest Paths of Delaunay graph weighted with highdimensional Distances

Reason:

- U-matrix is approximation of Abstract Umatrix (1) which is based on Voronoi cells
- Height of Voronoi-Borders=Euclidean High-dimensional distance of data



(1) [Lötsch/Ultsch 2014]

Why a swarm based approach for cluster analysis?

1. No objective function

⇒ possibility for emergence

2. Redundant decentralized algorithm

⇒ New data can be added inclemently

⇒ Swarm techniques are known for their properties of flexibility and robustness (1)

3. Alternative to neural networks

(1) [Bonabeau/Meyer, 2001; Şahin, 2004]