Clustering High Dimensional Data

Roland Winkler (roland.winkler@dlr.de), Frank Klawonn, Rudolf Kruse

July 11, 2011

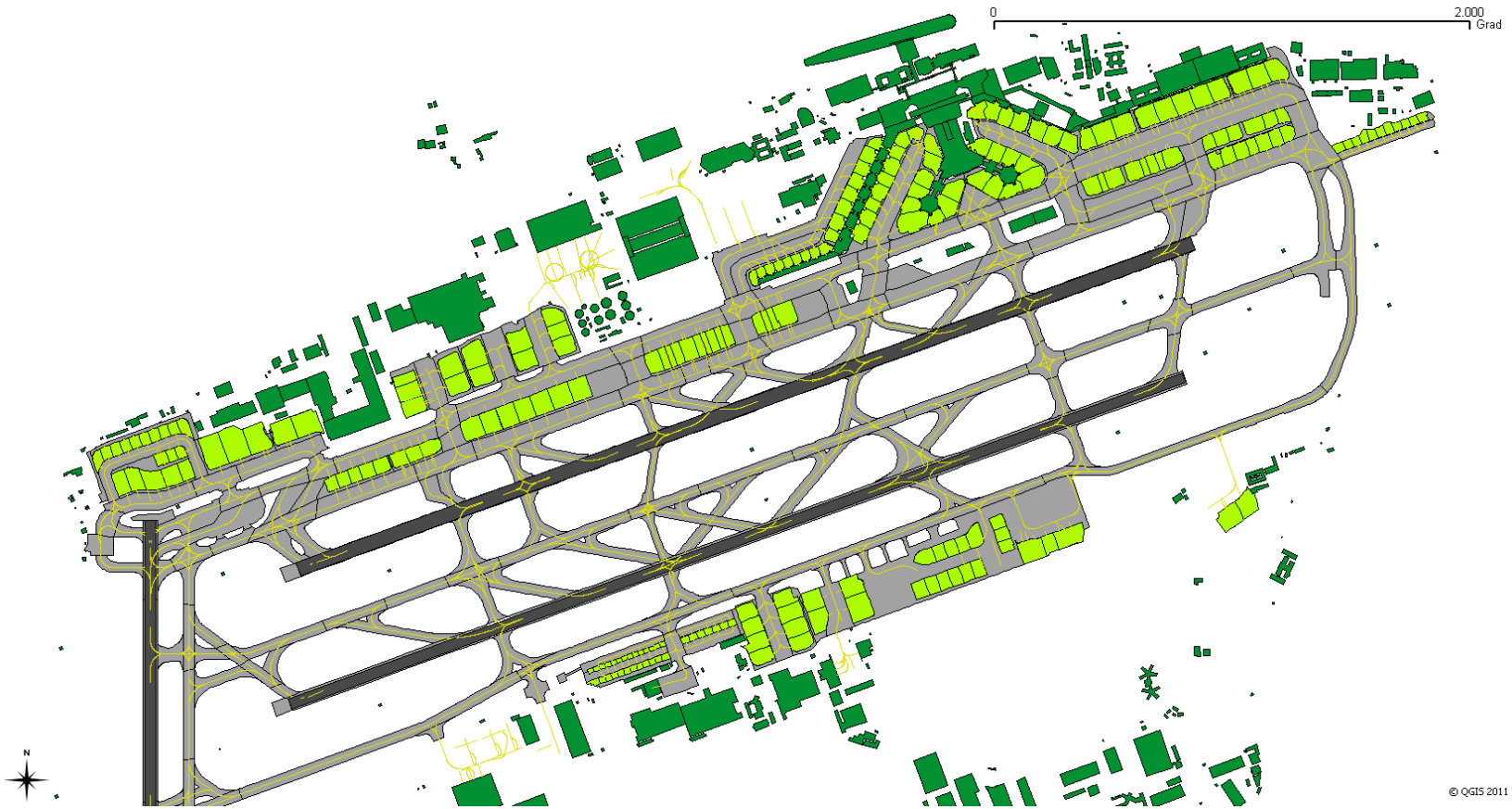
German Aerospace Center Braunschweig

Institute of Flight Guidance

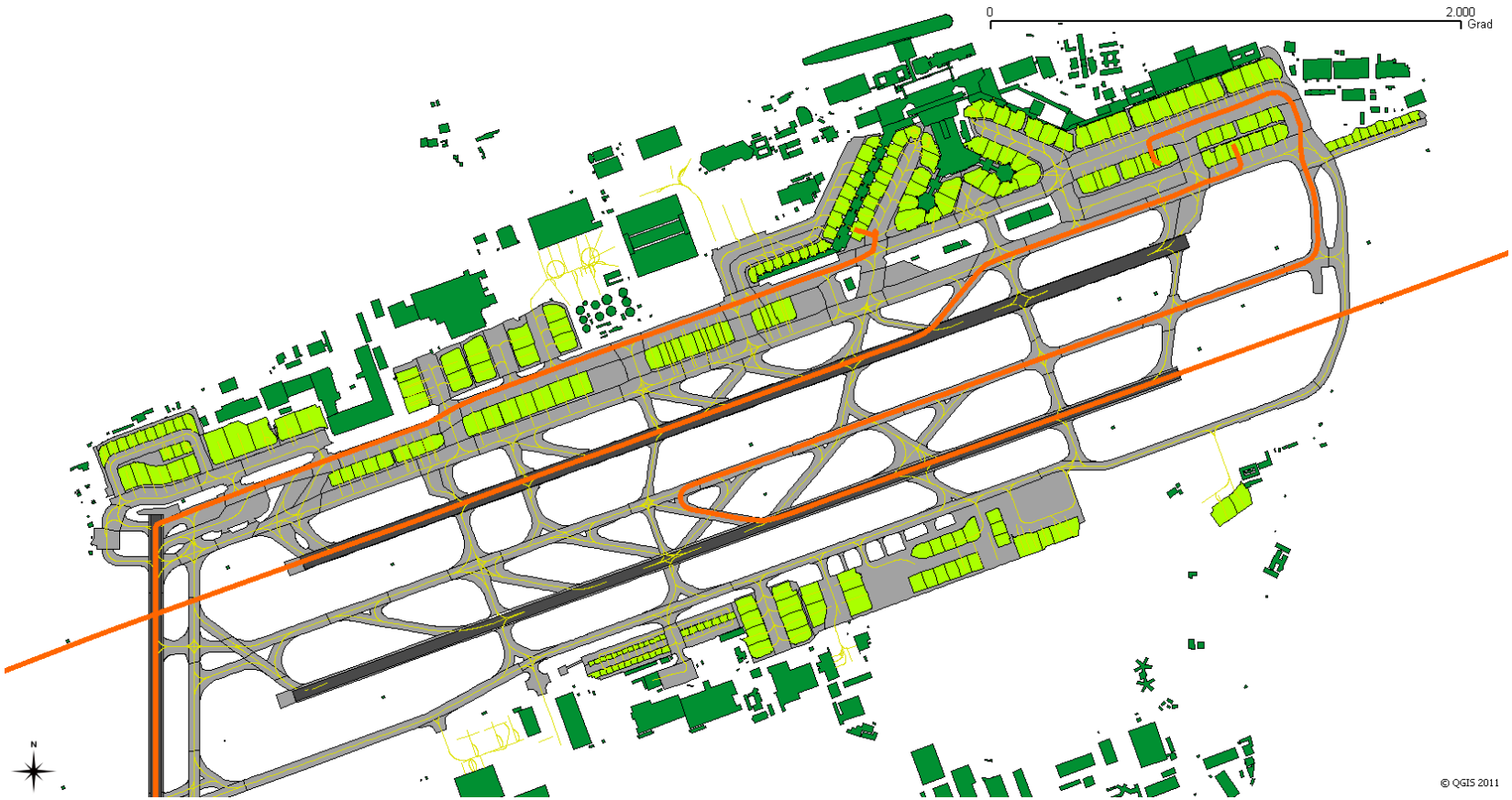
DLR Deutsches Zentrum für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

Current Section

S.O.D.A.: Frankfurt Airport

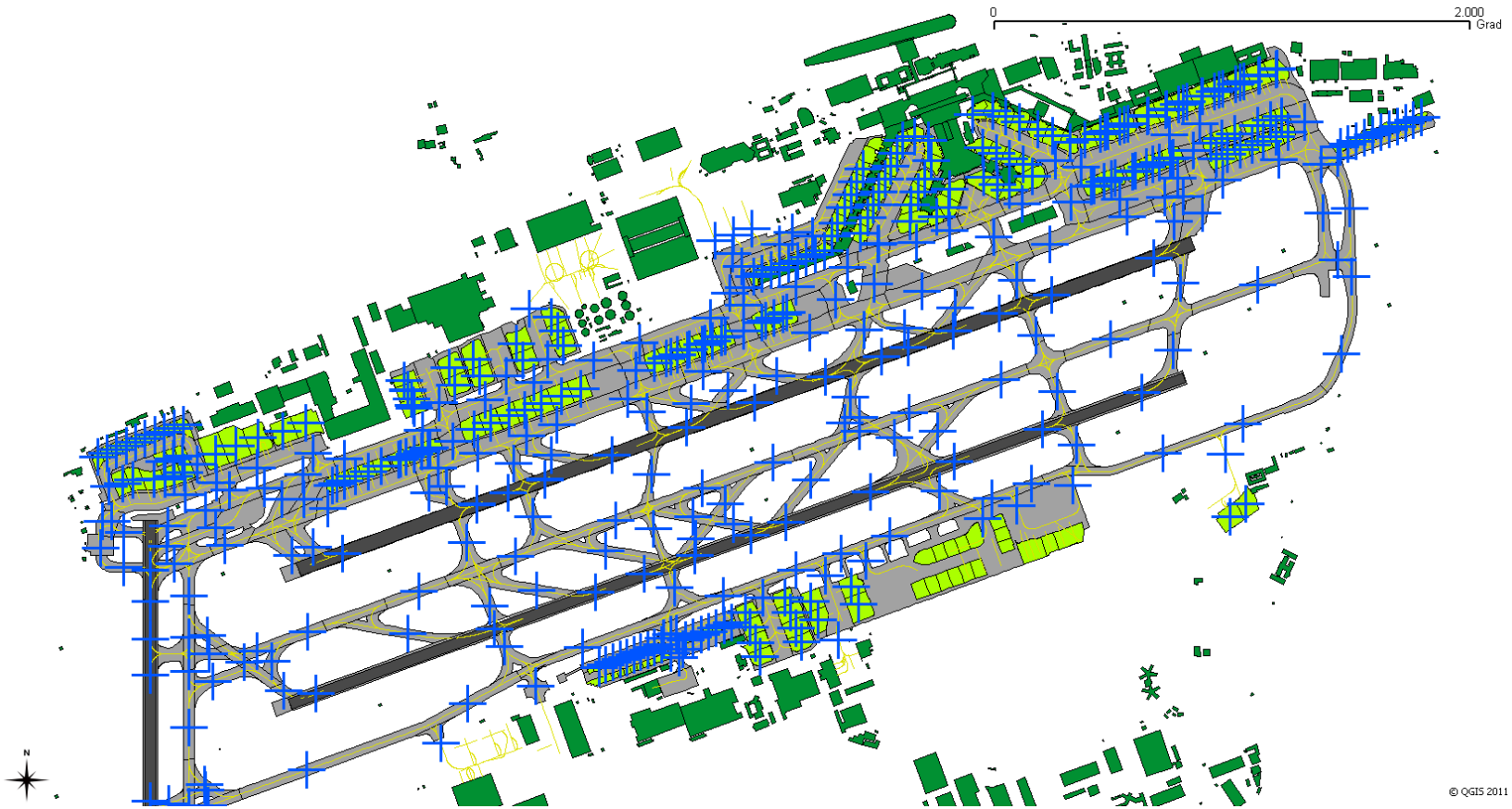


S.O.D.A.: 3 Aircraft transitions



S.O.D.A.: 10000 Aircraft transitions

S.O.D.A.: Characteristic Positions



Current Section

Challenges with high dim data sets in clustering

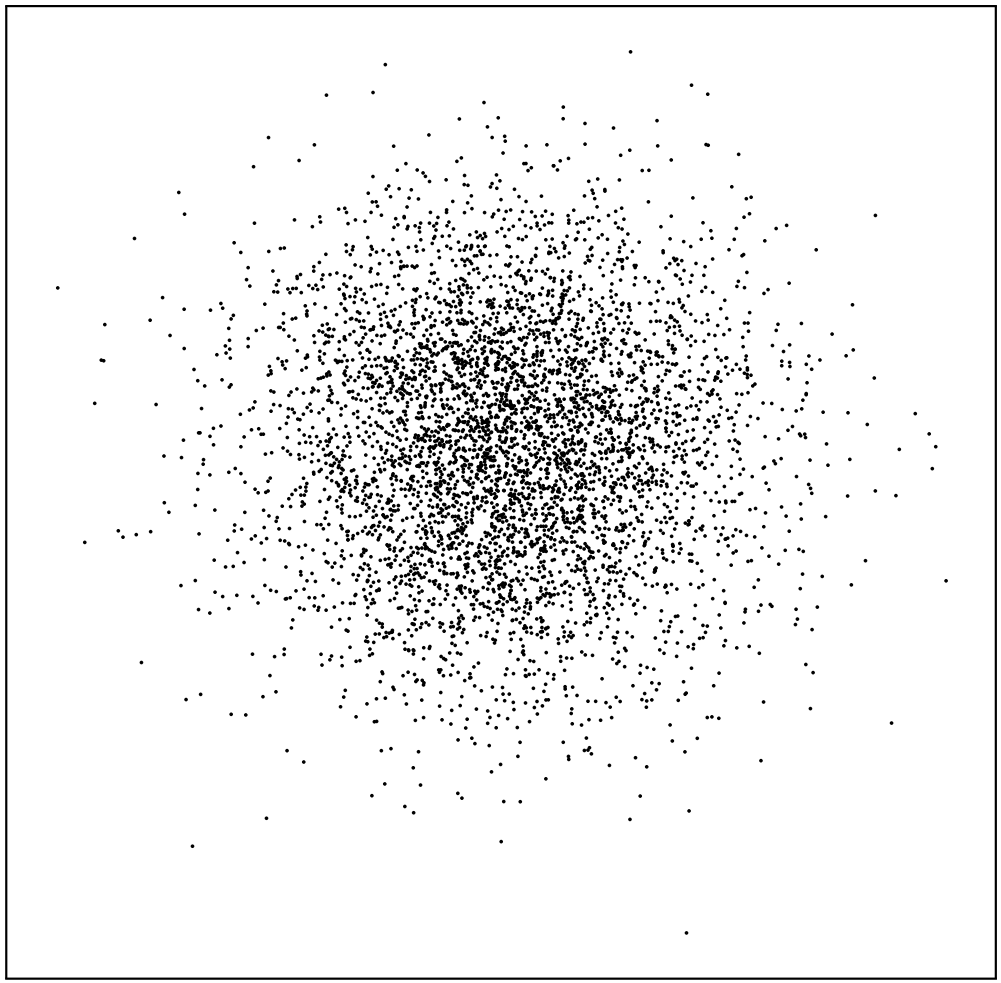
- Huge space that is very thin populated (for comparison: the m-dimensional hypercube has 2^m corners)
- The intrinsic dimensionality might be lower and form a complex geometry
- Dimension reduction is not necessarily helpful
- Occurrence of hubs (data objects that are part of the k-NN of a large portion of other data objects)
- **Distance concentration**

Distance concentration

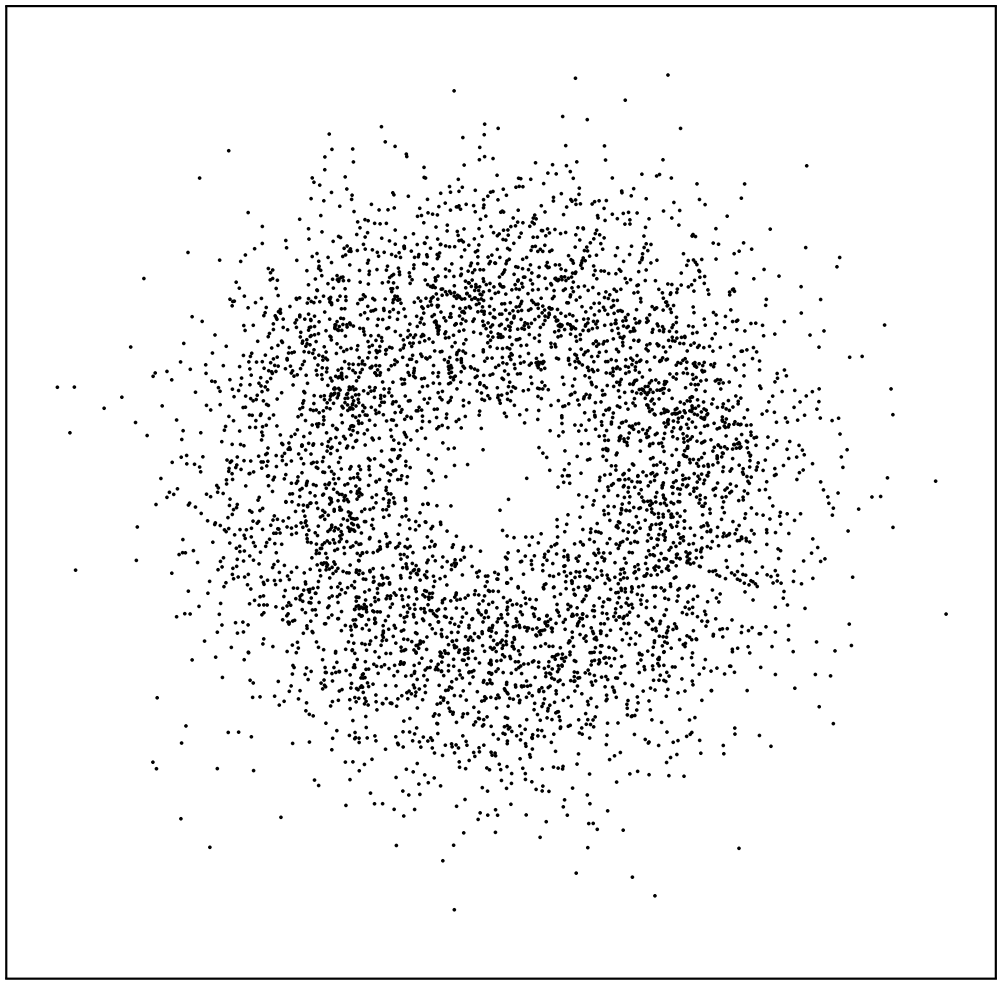
- Data set: $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^m$ with n data objects
- Query point: $Q \in \mathbb{R}^m$
- Metric: $\|\cdot\| : \mathbb{R}^m \longrightarrow \mathbb{R}$
- Set of distance: $D_Q(X) = \{\|x_i - Q\| \ : \ x_i \in X\}$
- Sample mean: $\widehat{E}(D_Q(X))$ and sample variance $\widehat{V}(D_Q(X))$
- Relative variance of distances: $\frac{\widehat{V}(D_Q(X))}{\widehat{E}(D_Q(X))^2} = \widehat{RV}(D_Q(X)))$

$$\lim_{m\to\infty} \widehat{RV}(D_Q(X)) = 0 \Longleftrightarrow \lim_{m\to\infty} P\left((1+\varepsilon)\cdot \min(D_Q(X)) > \max(D_Q(X))\right) = 1 \tag{1}$$

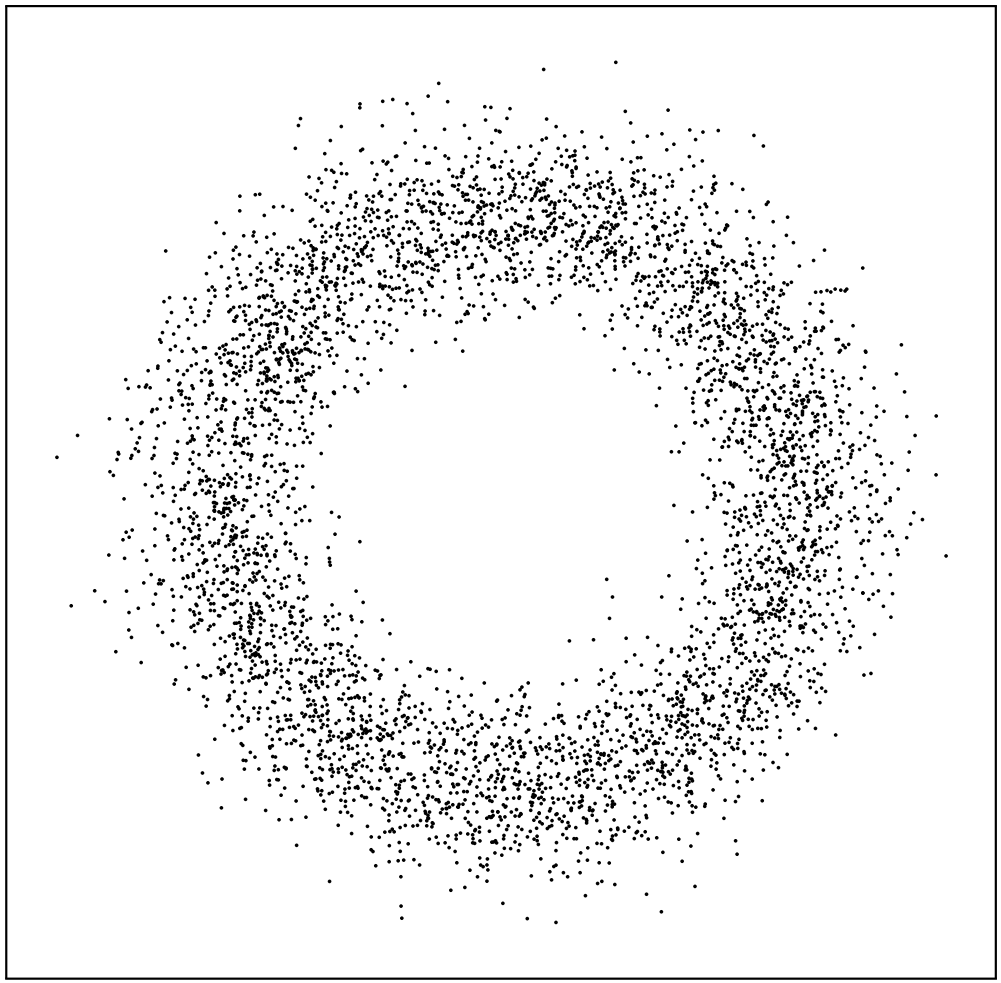
2 dimensions normal distribution



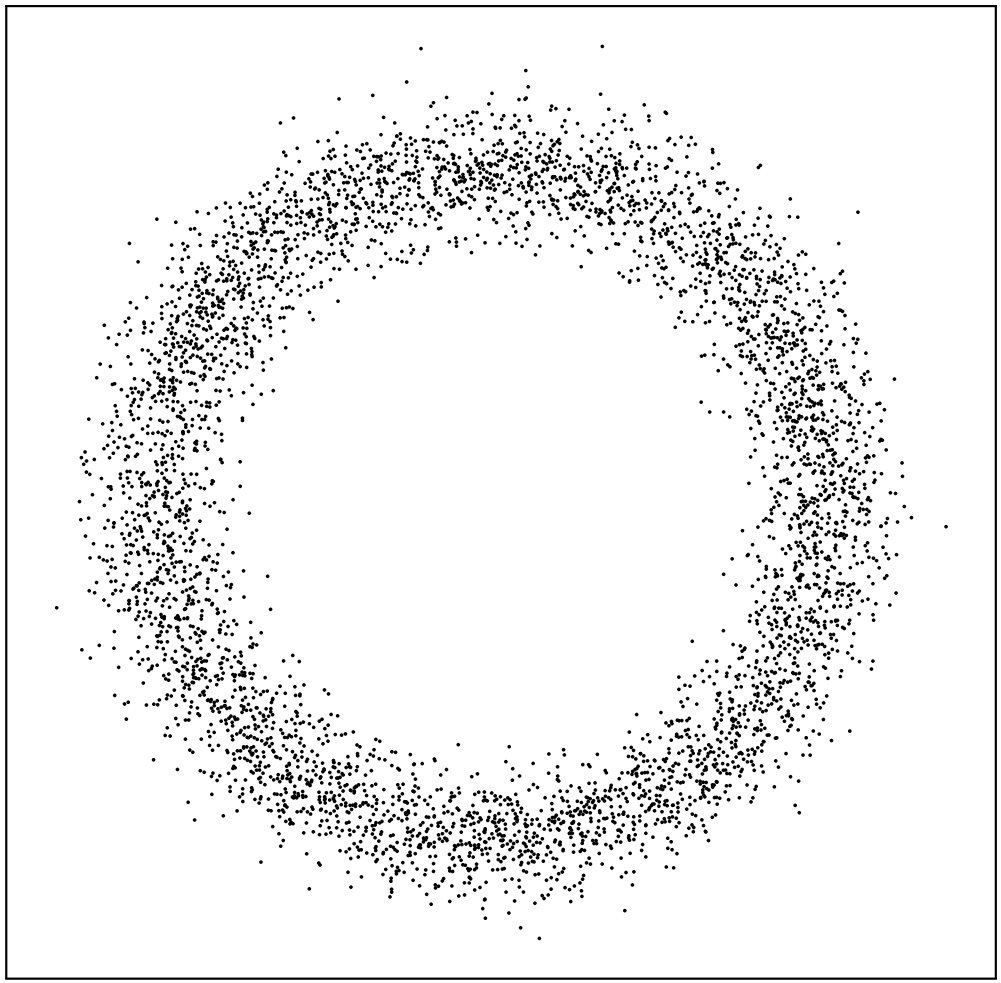
5 dimensions normal distribution



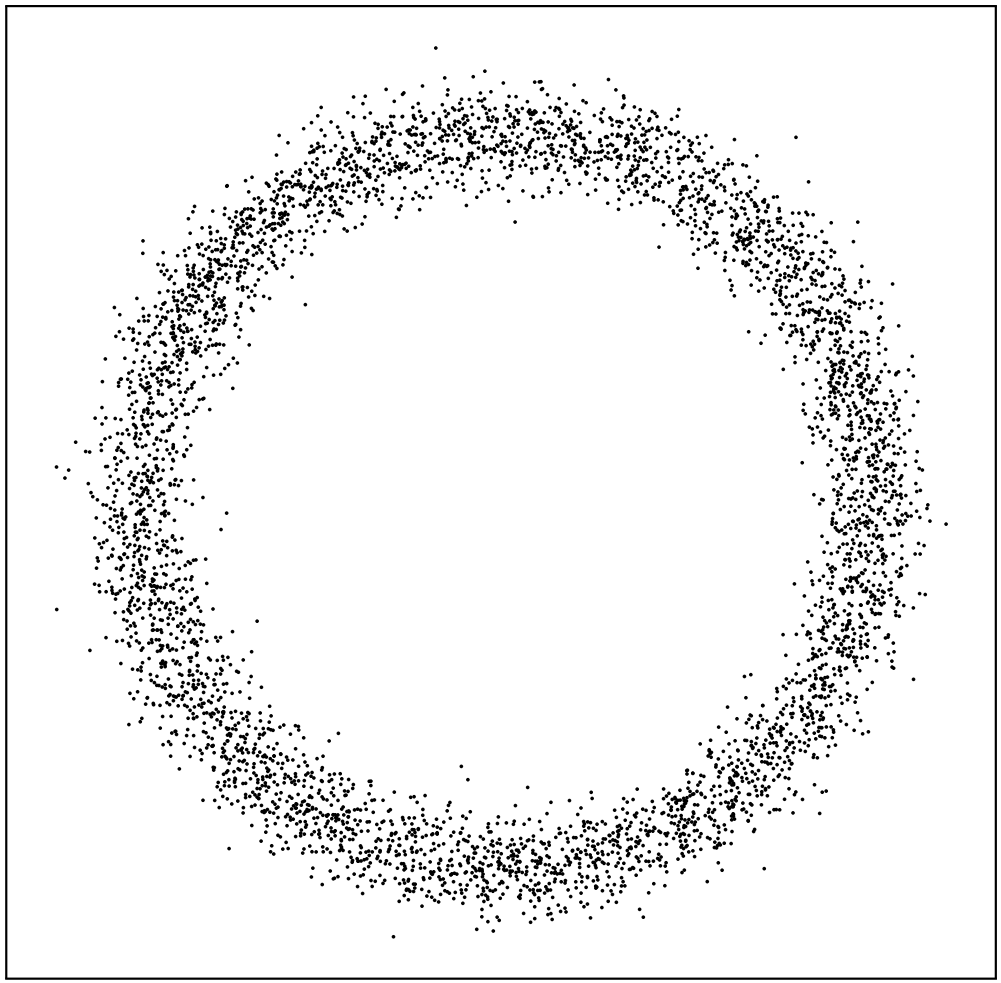
20 dimensions normal distribution



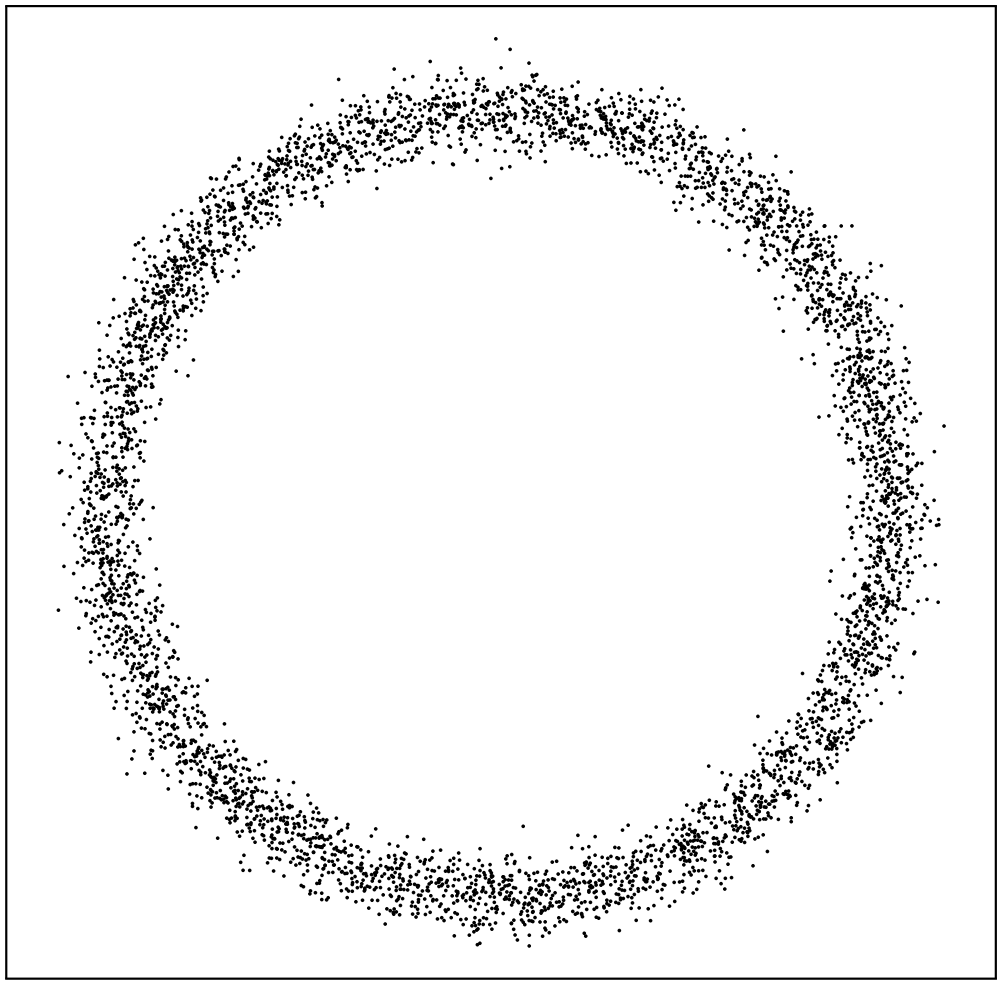
50 dimensions normal distribution



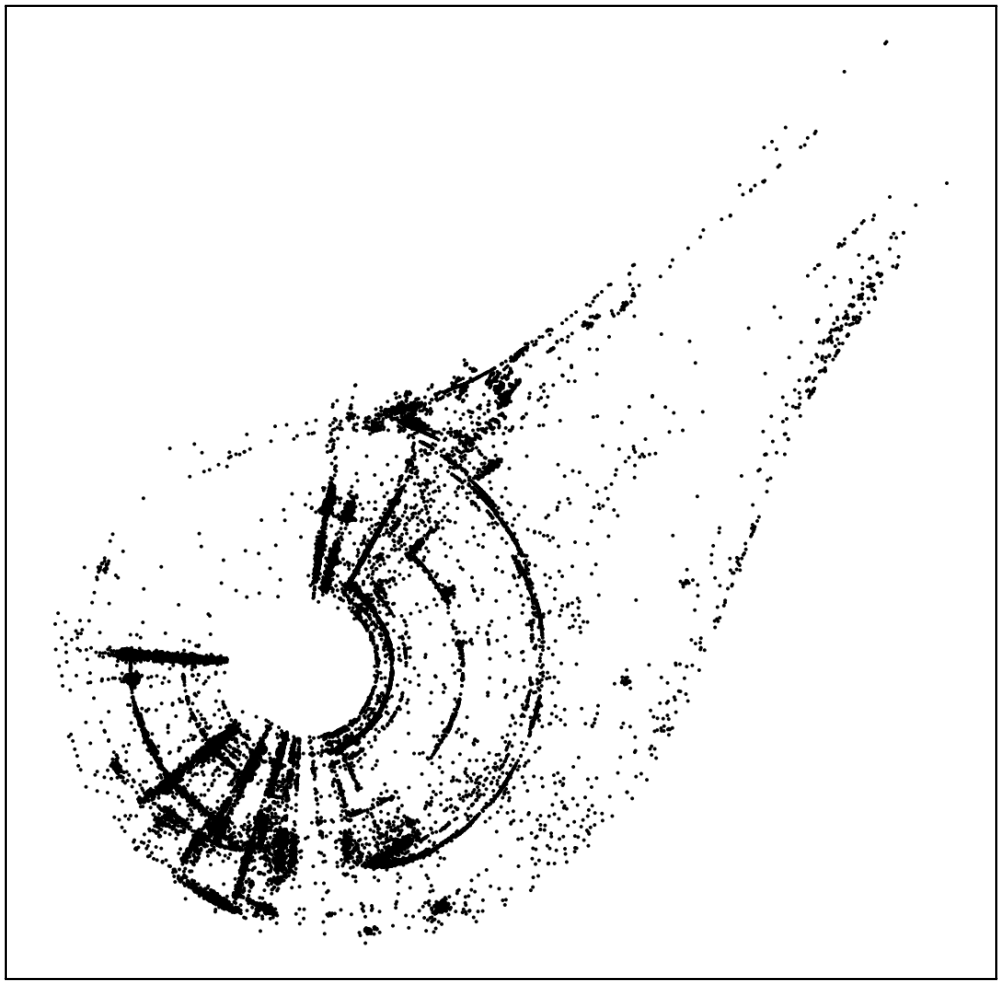
100 dimensions normal distribution



200 dimensions normal distribution



S.O.D.A.: Characteristic vector data set

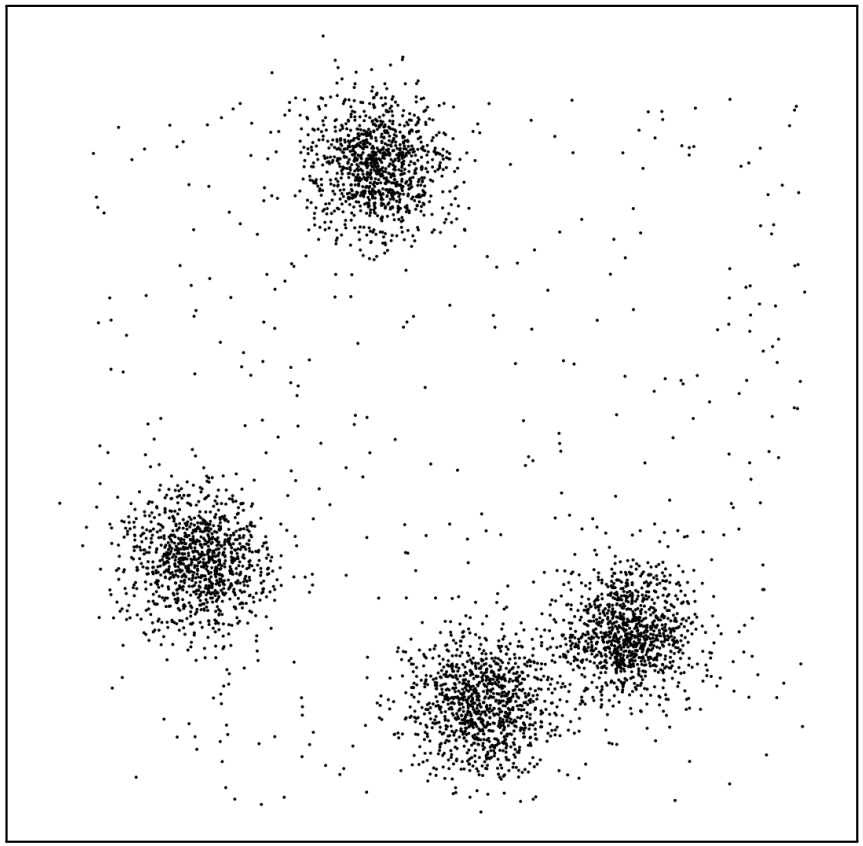


Current Section

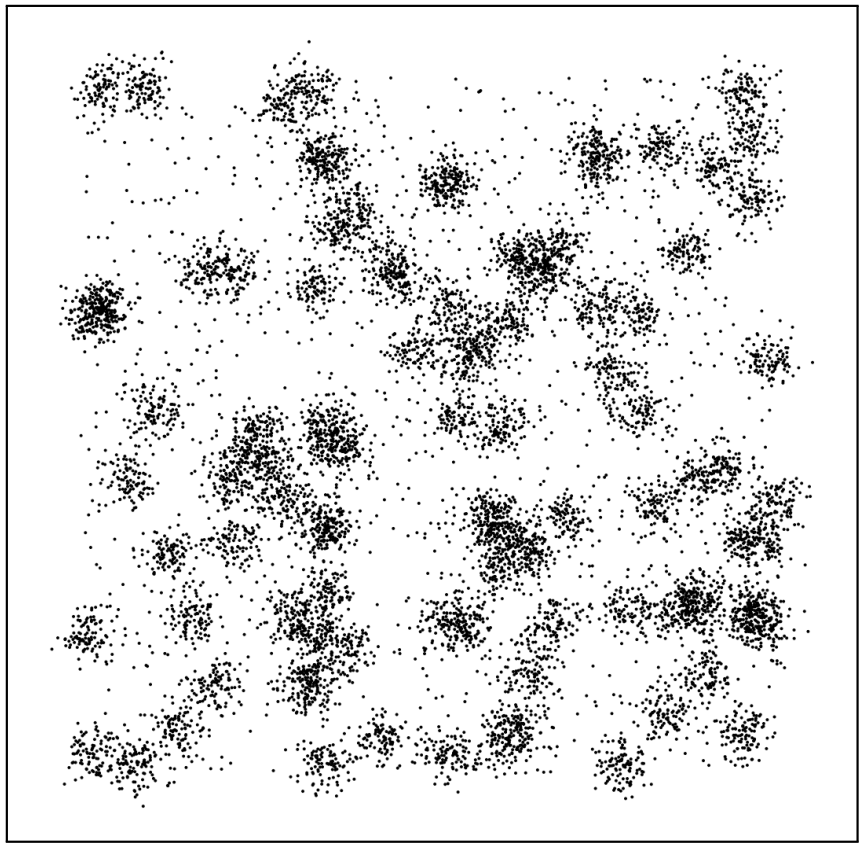
Objective function for HCM, FCM and PFCM

- Data set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^m$ with n data objects
- Prototype set: $Y = \{y_1, \ldots, y_c\} \subset \mathbb{R}^m$
- Distance function: $d_{ij} = d(y_i, x_j) = \|y_i - x_j\|$
- Membership values: $u_{ij} \in [0, 1]$
- Fuzzifier function: $f_{\text{fuzzy}} : [0, 1] \to [0, 1]$
- Objective function $\sum_{i=1}^{c} \sum_{j=1}^{n} f_{\text{fuzzy}}(u_{ij}) d_{ij}^2$

Test Data set



$m = 2$

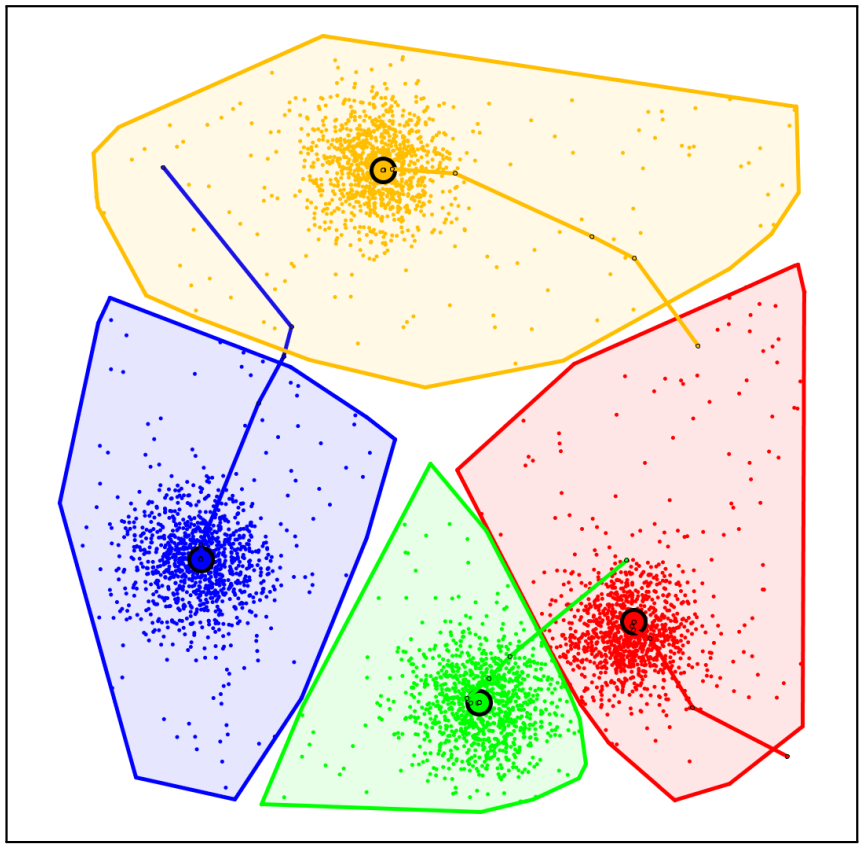


$m = 50$

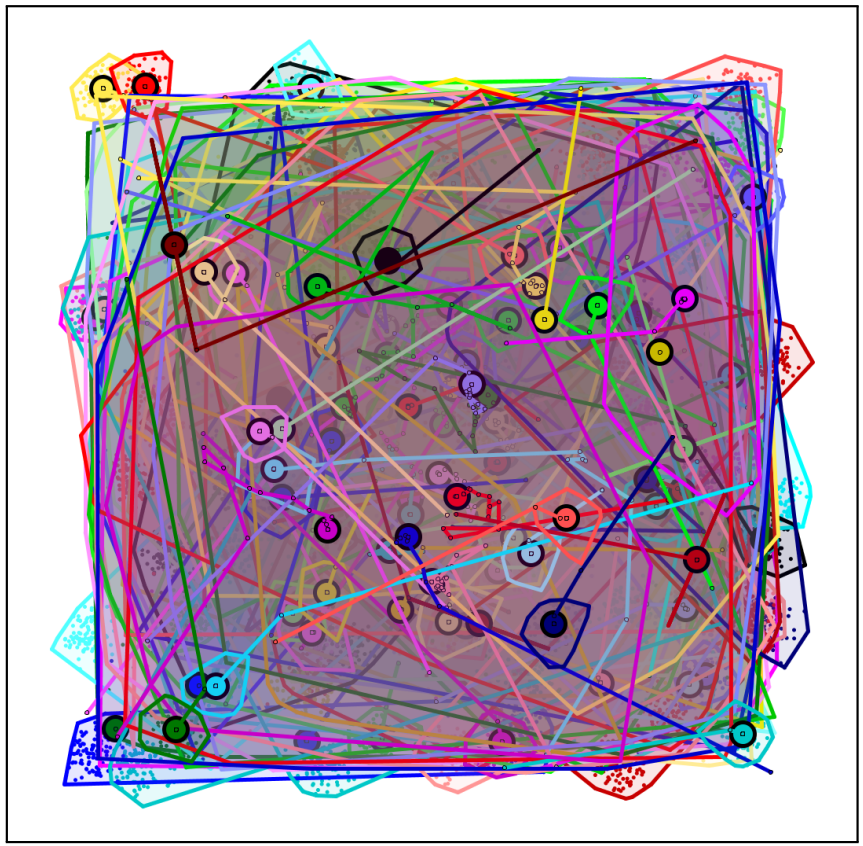
HCM

$f_{\text{fuzzy}}(u) = u$

found clusters: 47



$m = 2$

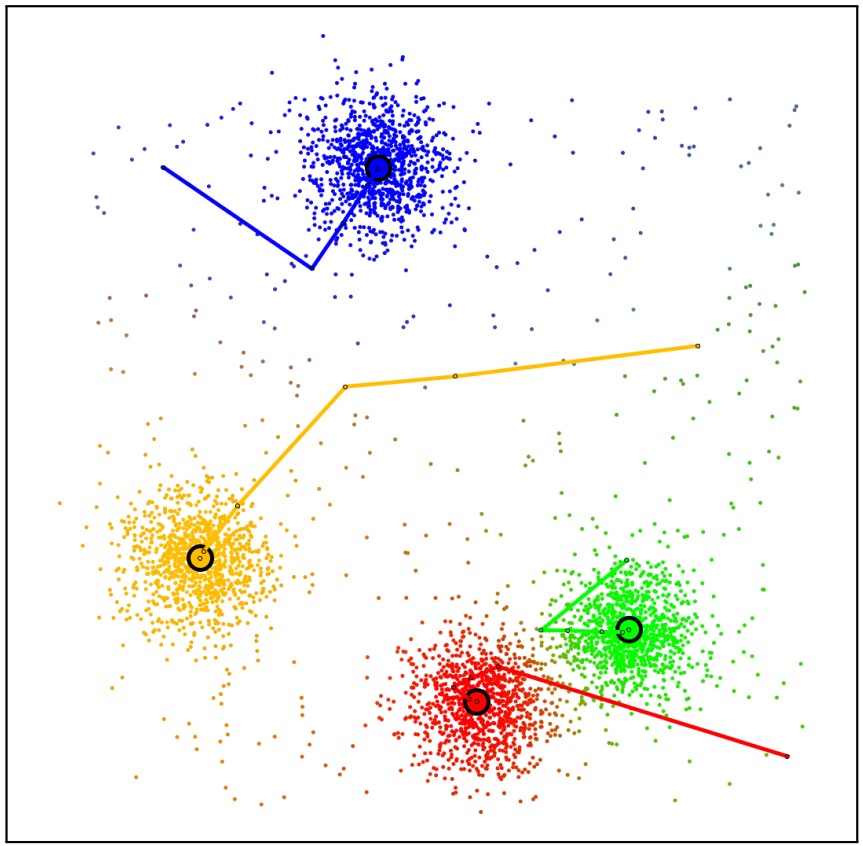


$m = 50$

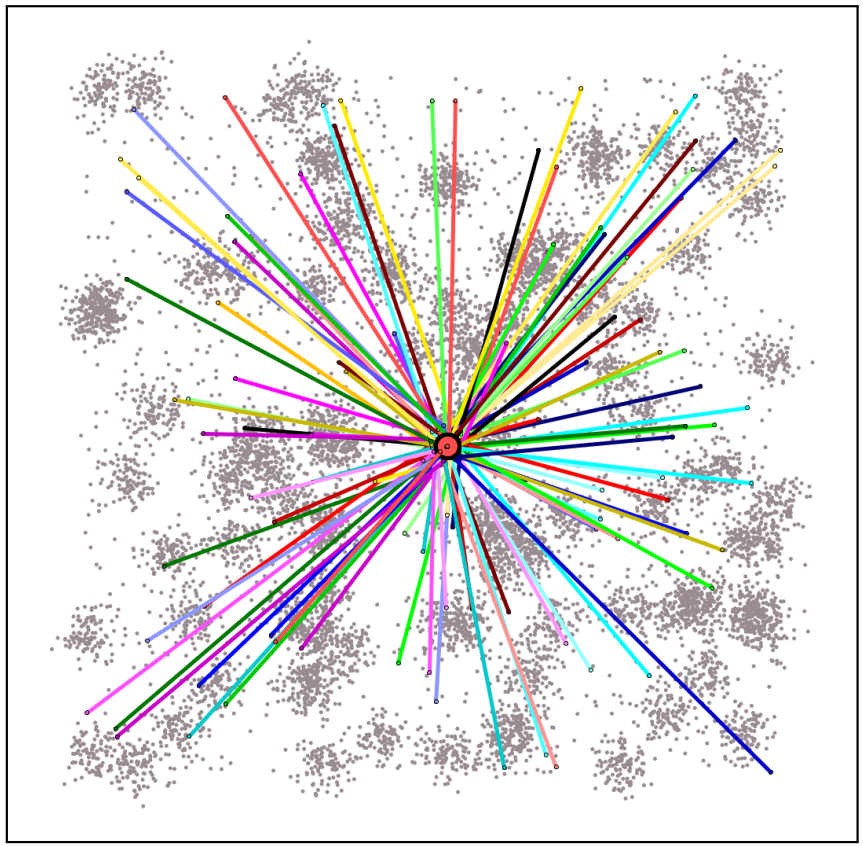
FCM

$f_{\text{fuzzy}}(u) = u^{\omega}$ (here: $\omega = 2$)

found clusters: 0



$m = 2$



$m = 50$

FCM with distance concentration

- Suppose $d(y_i, x_j) = d_{ij} \approx d^*$
- All membership values tend to become equal
- All prototypes tend to become equal

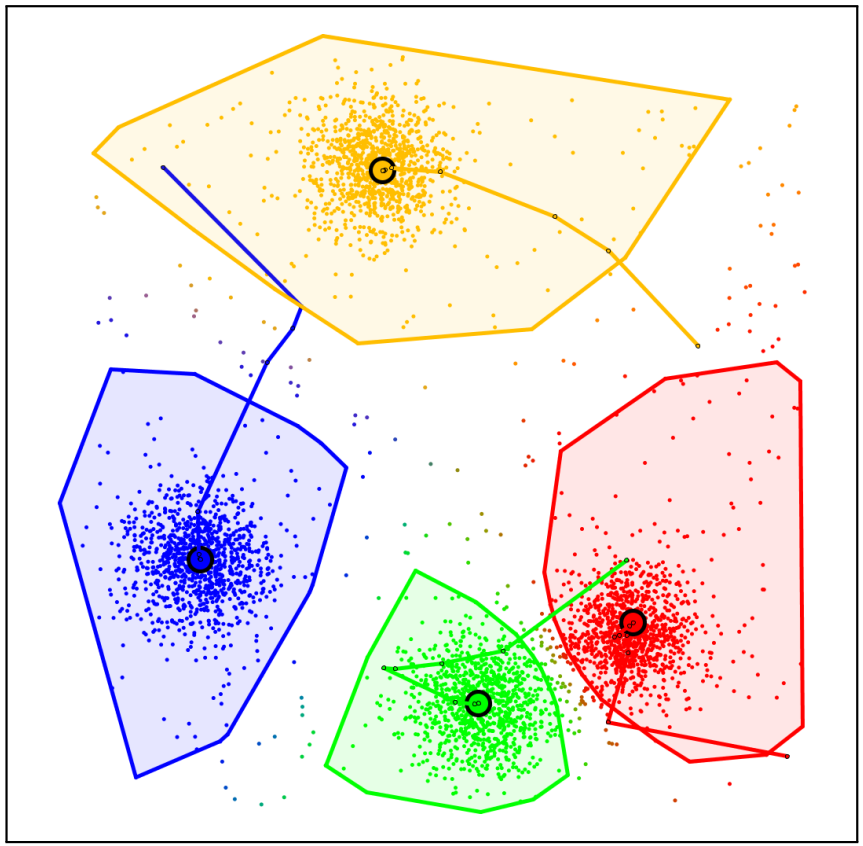
$$u_{ij} \approx \frac{\left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}} = \frac{\left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}}{c \cdot \left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}} = \frac{1}{c}$$

$$y_i \approx \frac{\sum_{j=1}^{n}\left(\frac{1}{c}\right)^{\omega} x_j}{\sum_{j=1}^{n}\left(\frac{1}{c}\right)^{\omega}} = \frac{\sum_{j=1}^{n}\left(\frac{1}{c}\right)^{\omega} x_j}{n \cdot \left(\frac{1}{c}\right)^{\omega}} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

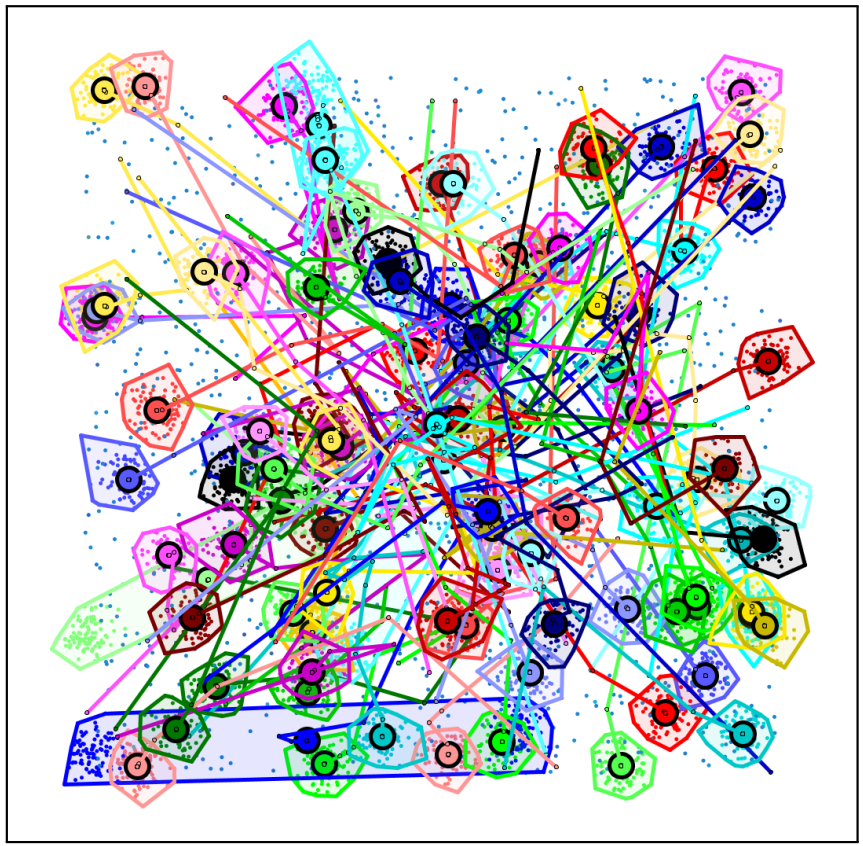
PFCM

$f_{\text{fuzzy}}(u) = \alpha \cdot u + (1 - \alpha)u^2$

found clusters: 92



$m = 2$

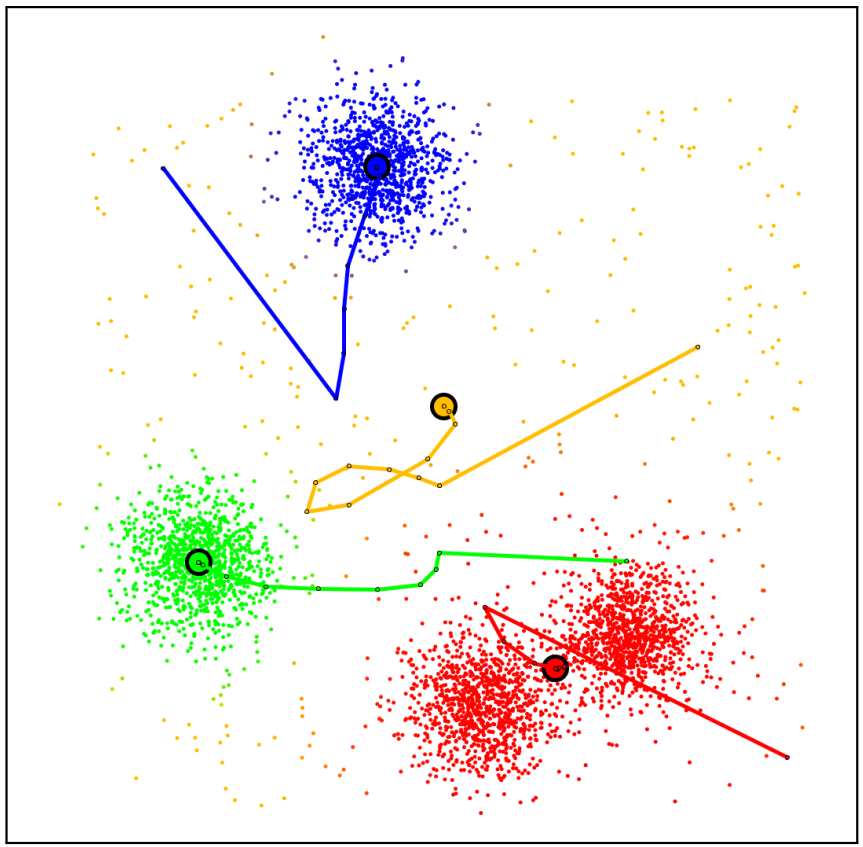


$m = 50$

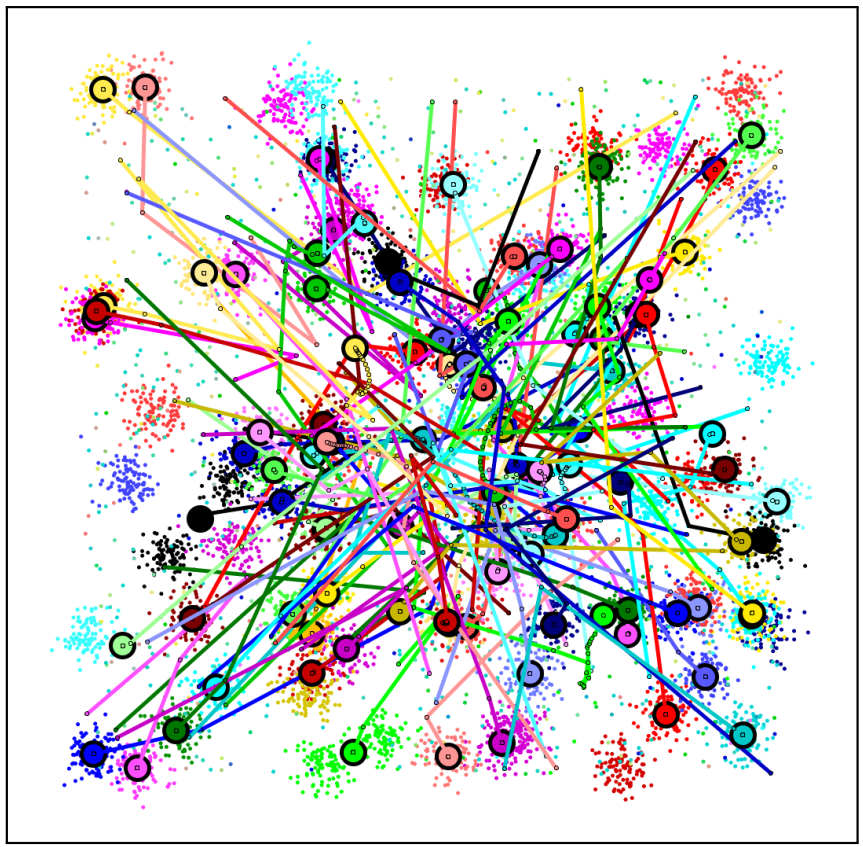
EM Gaussian Mixture Model

Covariance matrix = δE_m

found clusters: 68



$m = 2$



$m = 50$

Current Section

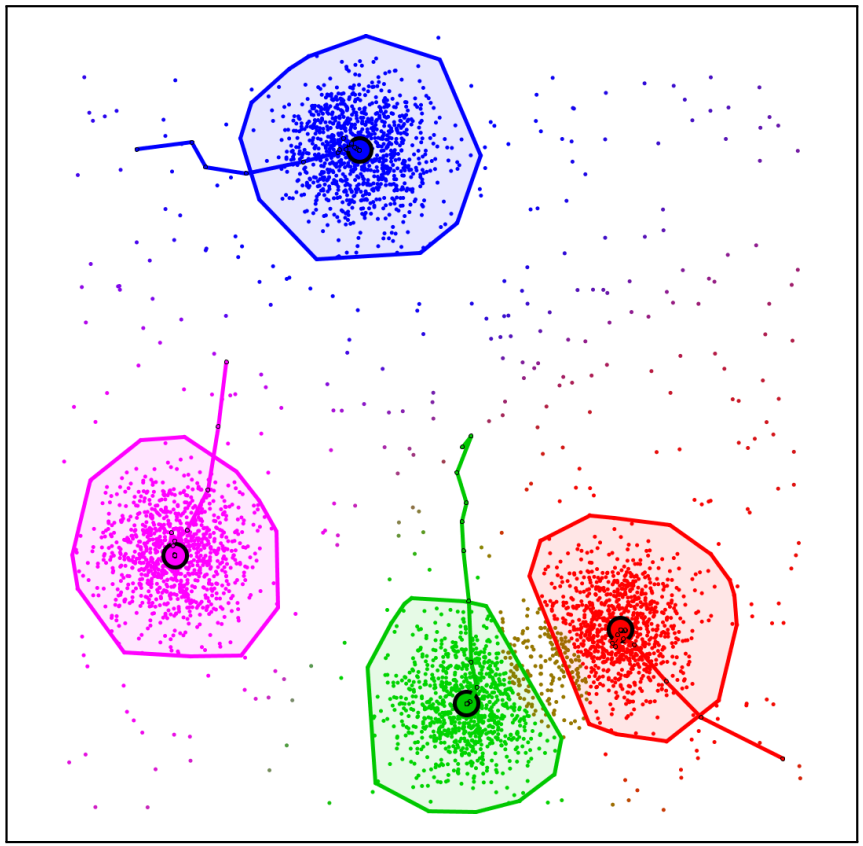
Alternative Distance Function

- $D_Q(X) = \{\|x - Q\| : x \in X\}$
- $a > 0$
- $\delta = \max\left(\widehat{E}(D_Q(X)) - a\sqrt{\widehat{V}(D_Q(X))}, 0\right)$
- $d_{ij} = d_\delta(y_i, x_j) = \max\left(\|y_i - x_j\| - \delta, 0\right)$ with $Q = y_i$
- $a = 3$ implies: for less than 10% of data objects, $d_\delta(y_i, x_j) = 0$
- Objective function $\sum_{i=1}^{c}\sum_{j=1}^{n} f_{\text{fuzzy}}(u_{ij}) d_\delta^2(y_i, x_j)$ does not work well with d_δ
- Use d_δ only for calculating Membership values

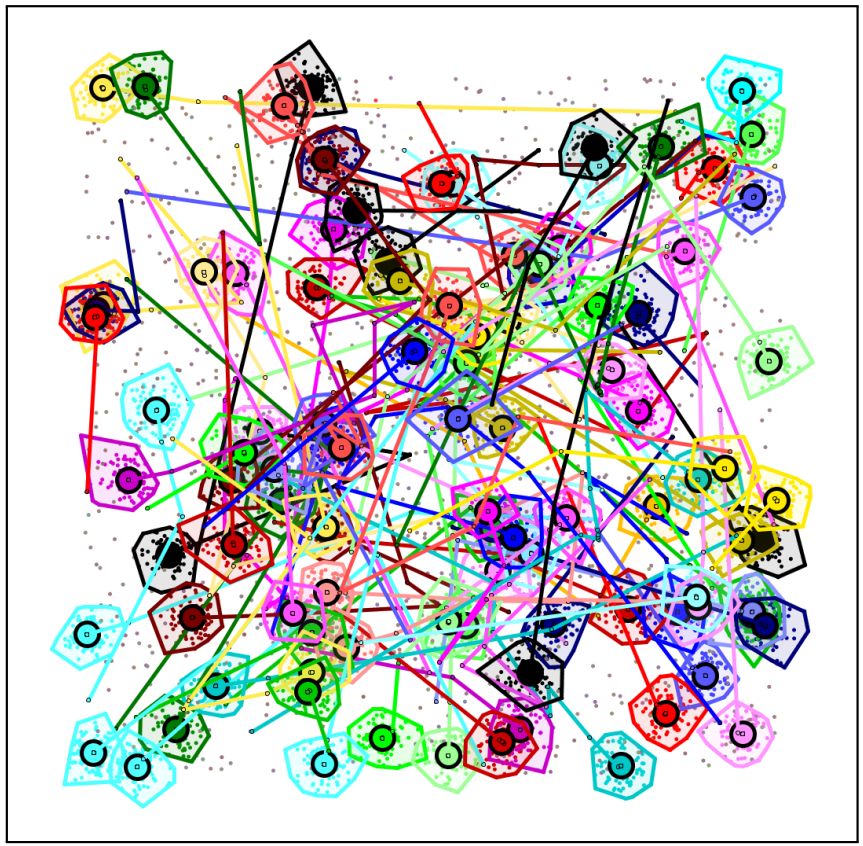
FCM with Alternative Distance Function Example

$f_{\text{fuzzy}}(u) = u^{\omega}$, $d = d_{\delta}$

found clusters: 99



$m = 2$



$m = 50$

Thank you for your attention

Bibliography

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999).
When is nearest neighbor meaningful?
In *Database Theory - ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg.

Durrant, R. J. and Kabán, A. (2008).
When is 'nearest neighbour' meaningful: A converse theorem and implications.
Journal of Complexity, 25(4):385 – 397.

Hsu, C.-M. and Chen, M.-S. (2009).
On the design and applicability of distance functions in high-dimensional data space.
Knowledge and Data Engineering, IEEE Transactions on, 21(4):523 –536.

Kabán, A. (2011).
Non-parametric detection of meaningless distances in high dimensional data.
Statistics and Computing.
status: accepted.

S.O.D.A.

- S.O.D.A (Surveillance Data Analysis System) is developed by the Fraport AG, the organization operating the international airport in Frankfurt
- It is based on the open source data base system PostGre SQL, the visualizations are done using PostGIS
- DLR (German Aerospace Center) Braunschweig and Fraport are working as partners in the S.O.D.A. project to increase data usage and quality

Distance concentration

- $F^{(m)}$, $m \in \mathbb{N}$ be a sequence of random variables
- $X^{(m)} = \{x_1^{(m)}, \ldots, x_n^{(m)}\} \sim F^{(m)}$ be a sample of n data objects
- $Q^{(m)} \in dom(F^{(m)})$ a query point
- $\| \cdot \| : dom(F^{(m)}) \longrightarrow \mathbb{R}$ be a distance function and $p > 0$
- $D^{(m)} = D_{Q^{(m)}}(X^{(m)}) = \{\|x_i^{(m)} - Q^{(m)}\|^p \ : \ x_i^{(m)} \in X^{(m)}\}$
- $\widehat{E}(D^{(m)})$ and $\widehat{V}(D^{(m)})$ be finite, $\widehat{E}(D^{(m)}) > 0$
- $\frac{\widehat{V}(D^{(m)})}{\widehat{E}(D^{(m)})^2} = \widehat{RV}(D^{(m)}))$

$$\lim_{m\to\infty} \widehat{RV}(D^{(m)}) = 0 \Longleftrightarrow \lim_{m\to\infty} P\left((1+\varepsilon) \cdot \min(D^{(m)}) > \max(D^{(m)})\right) = 1 \qquad (2)$$

Distance concentration

- Let $\|\cdot\|$ be one of the L_p norms and in all cases $Q^{(m)} \sim F^{(m)}$
- When are the conditions in (1) true?
  - $F^{(m)}$ is i.i.d. in all dimensions (see the following example)
  - $F^{(m)}$ is correlations in all dimensions: $U_1, \ldots, U_m$ so that $U_k \sim U[0, \sqrt{k}]$, define $F_1^{(m)} = U_1$ $F_k^{(m)} = U_k + \frac{1}{2} F_{k-1}^{(m)}$
  - $F^{(m)}$ is the uniform distribution on a m-dimensional hyperqube surface: $F^{(m)} = U([0,1]^m \setminus (0,1)^m)$
  - $F^{(m)}$ with $\sigma_m \to 0$: define $F_k^{(m)} \sim N(0, \frac{1}{k})$, $k = 1 \ldots m$
- Open question: What are the satisfying conditions for distance concentration?