Clustering with Repulsive Prototypes

R. Winkler, F. Rehm, R. Kruse

roland.winkler@dlr.de,
frank.rehm@dlr.de,
kruse@iws.cs.uni-magdeburg.de

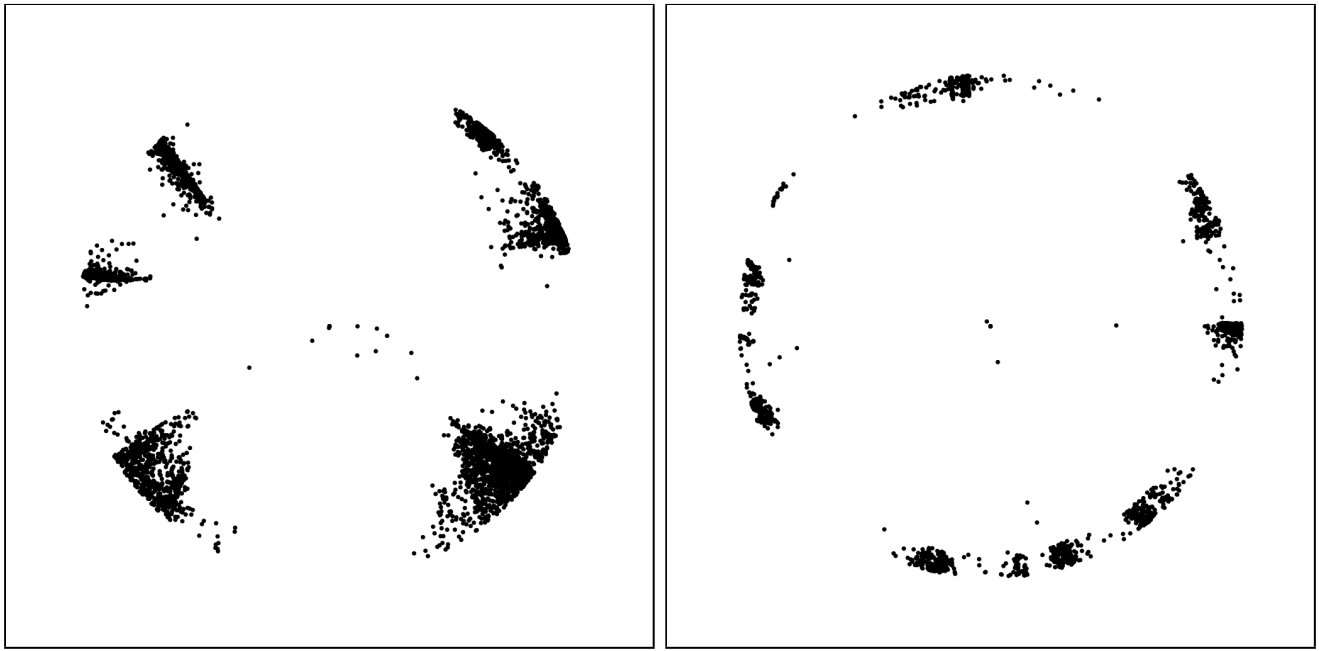
Contents

1. Introduction

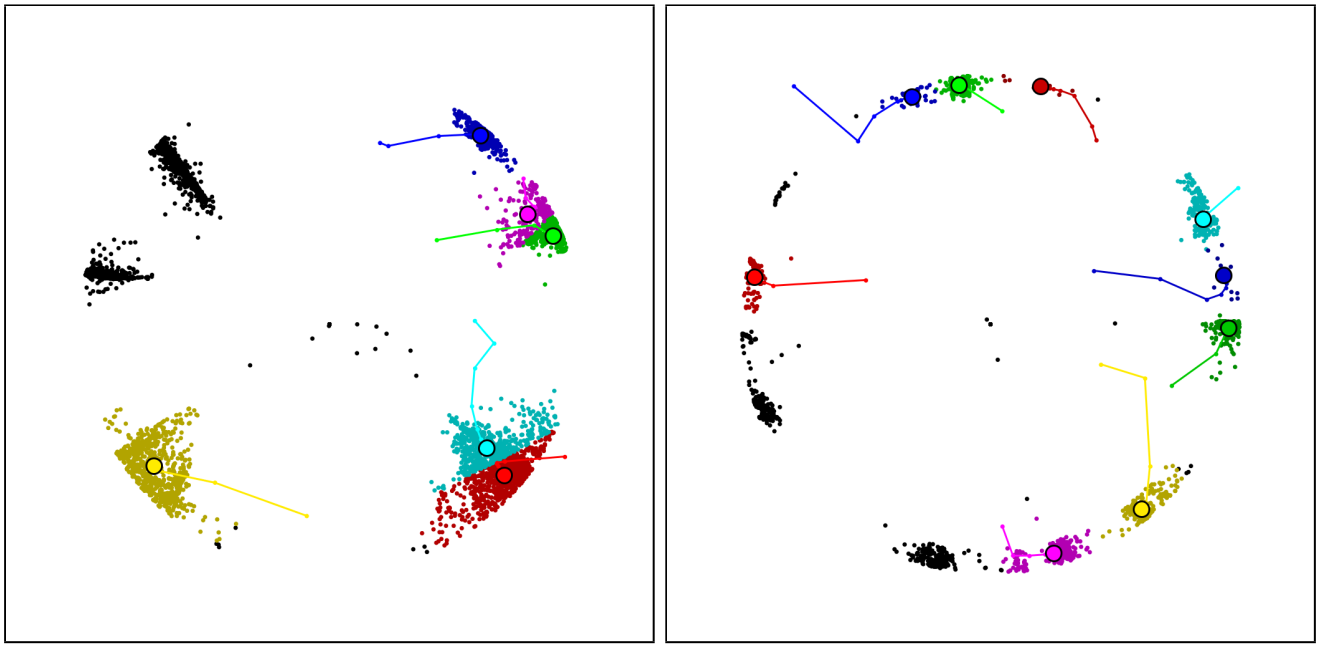
Initialization of Clustering Algorithms

- **The focus in this presentation lies on the following two problems in prototype based clustering:**
  - **find the number of clusters in a dataset**
  - **find a good initialization for the prototypes**
- **Repulsive prototypes is an extension for Noise Clustering**
- **With repulsive prototypes, it is possible to use additional information about a dataset, to solve the above mentioned problems**
- **The following information is required:**
  - **volume expansion of clusters (noise distance)**
  - **minimal distance between prototypes**

Example Datasets



Example Datasets with Noise Clustering



2. Update Function for Repulsive Prototypes

Fuzzy c-Means Update Rules

- **Objective function:**

$$L(X,U,B,\Lambda) = \sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}^{\omega} d_{ij}^2 + \sum_{j=1}^{n} \lambda_j \left(1 - \sum_{i=1}^{m} u_{ij}\right)$$

- **Update rules:**

$$u'_{ij} = \frac{(d_{ij})^{\frac{2}{1-\omega}}}{\sum_{k=1}^{m}\left((d_{kj})^{\frac{2}{1-\omega}}\right)} \qquad \beta'_i = \frac{\sum_{j=1}^{n}(u_{ij})^{\omega}\cdot x_j}{\sum_{j=1}^{n}(u_{ij})^{\omega}}$$

- **Notation:**

m: **number of Clusters**

n: **number of datapoints**

i,k: **prototype indices**

j: **datapoint index**

β: **prototype**

x: **datapoint**

ω: **fuzzyfier**

λ: **lagrange variable**

u_{ij}: **partition matrix element**

d_{ij}: **distance β_i to x_j**

Repulsive Extension

- **Repulsive prototypes are an extension for Noise Clustering**
- **Update rule of repulsive prototypes:**

$$\beta_i' = \frac{\sum_{j=1}^{n} (u_{ij})^{\omega} \cdot x_j}{\sum_{j=1}^{n} (u_{ij})^{\omega}} + c \cdot \sum_{\substack{k=1 \\ k \neq i}}^{m} \left(\frac{\beta_i - \beta_k}{\|\beta_i - \beta_k\|} \cdot \frac{\sum_{j=1}^{n} u_{kj}}{\sum_{j=1}^{n} (u_{ij} + u_{kj})} \cdot \varphi(d(\beta_i, \beta_k)) \right)$$

- **Notation:**

m: **number of clusters**

n: **number of datapoints**

i, k: **prototype indices**

j: **datapoint index**

β: **prototype**

x: **datapoint**

ω: **fuzzyfier**

u_{ij}: **partition matrix element**

d_{ij}: **distance β_i to x_j**

c: **distance β_i to x_j**

φ: **repulsion function**

Repulsion Function

$$\varphi(x) = \frac{1}{1 + e^{a \cdot (x - \sigma)}}$$

$$\varphi(\sigma) = 0.5$$

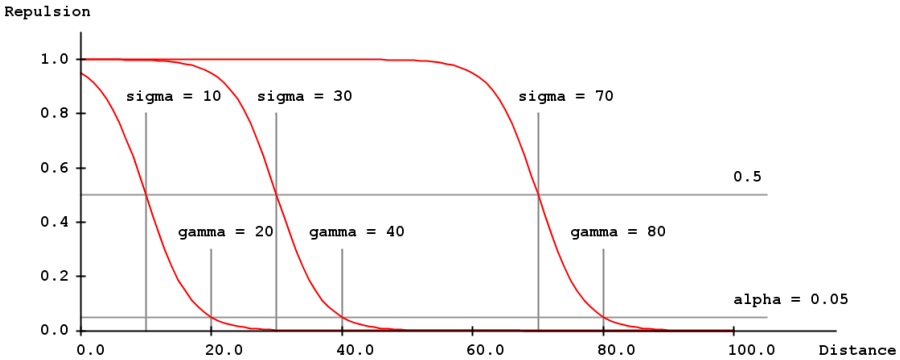
$$\varphi(\gamma) = \alpha$$

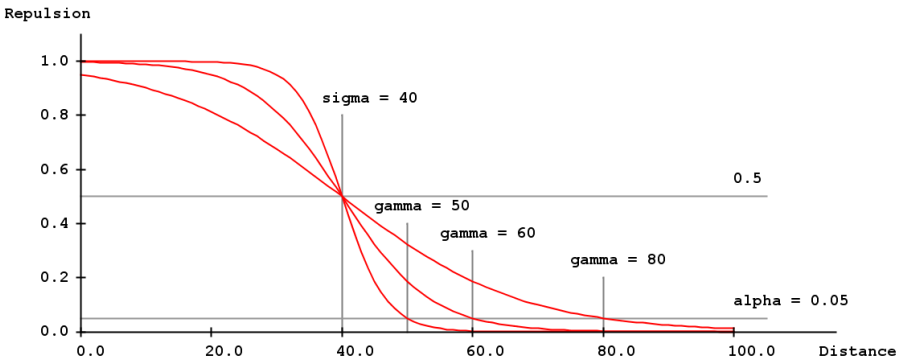
$$\Rightarrow a = \frac{\ln(\frac{1}{\alpha} - 1)}{\gamma - \sigma}$$

$$\alpha < 0.5$$

$$\sigma > 0$$

$$\gamma > \sigma$$



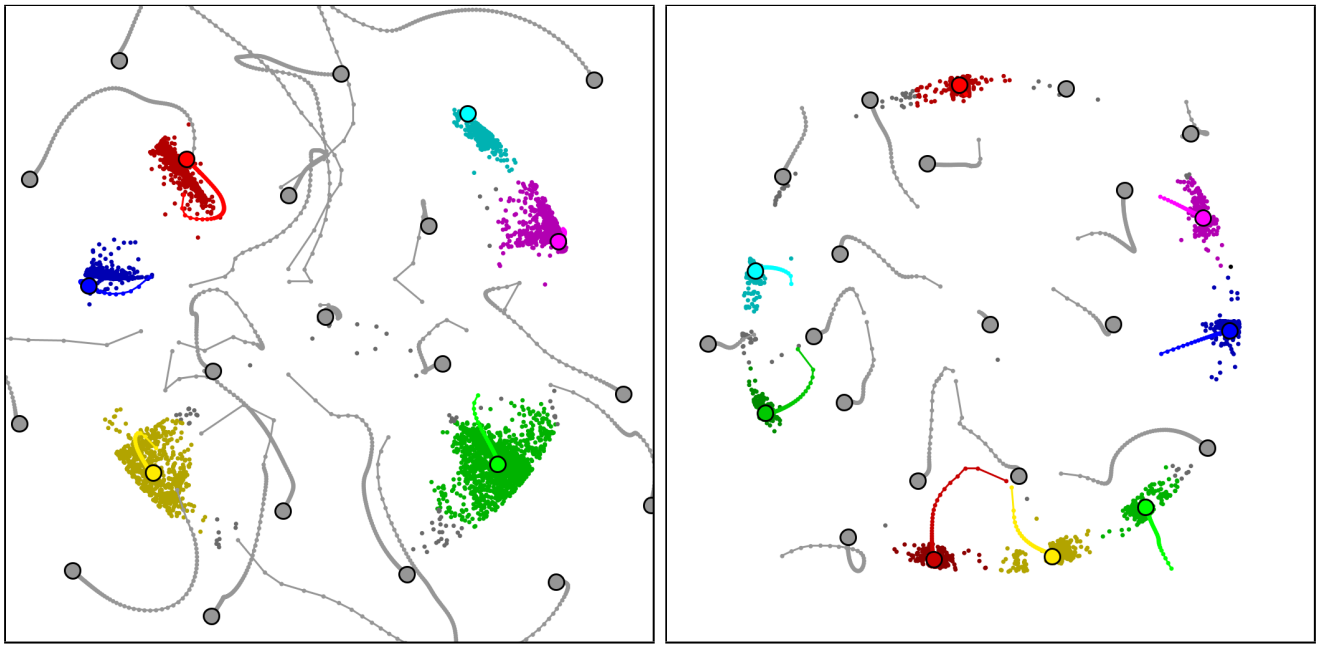


Clustering Parameter

- **Important parameter for clustering with repulsive prototypes:**
  - **Parameter of the repulsion function φ: σ and γ**
  - **Upper bound for the number of clusters $\hat{m}$ (so there are more prototypes than clusters)**
- **A test is used to differenciate between 'full' and 'empty' prototypes. For example: the prototype β_i is full, if $\sum_{j=1}^{n} u_{ij} > T \in \mathbb{R}^+$ and empty otherwise**
- **A roughly estimation of the above mentioned parameter is sufficient for a successful clustering**
- **It is possible to estimate σ and γ using the noise distance ν, for example: $\sigma = a \cdot \nu$ and $\gamma = b \cdot \nu$ with $a, b \in \mathbb{R}^+$ and $a < b$**

3. Experimental Results

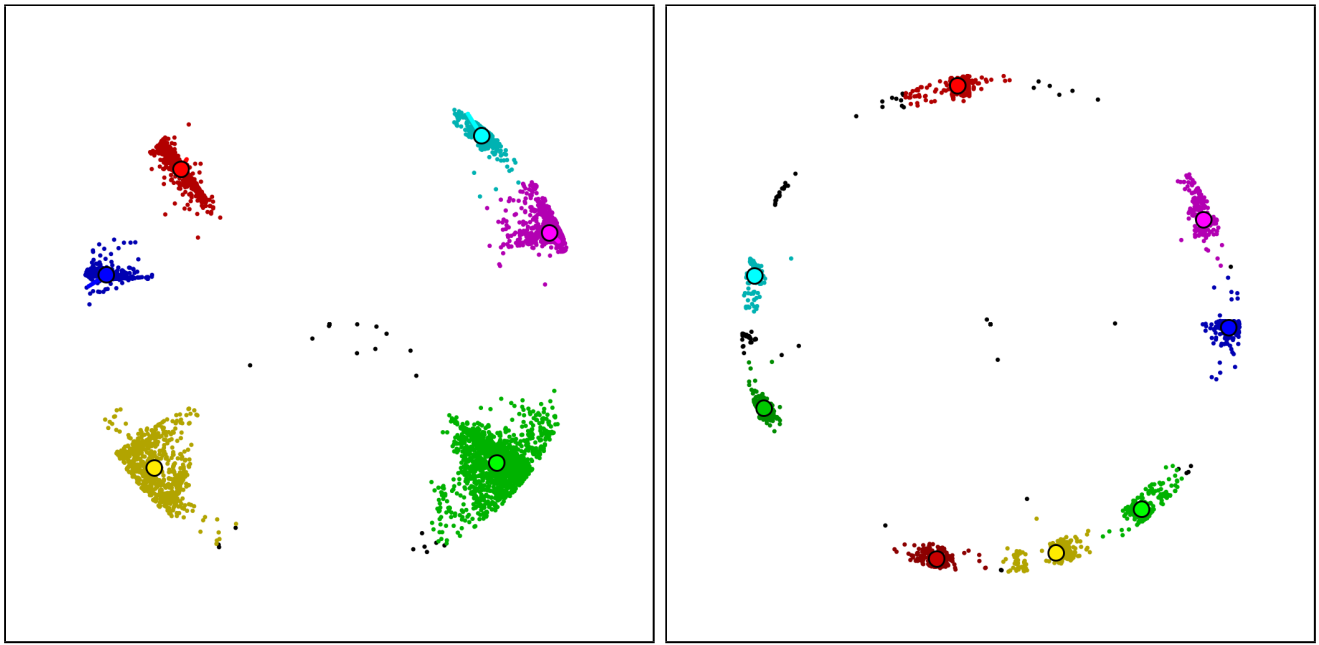
Example Dataset with Repulsive Initialization



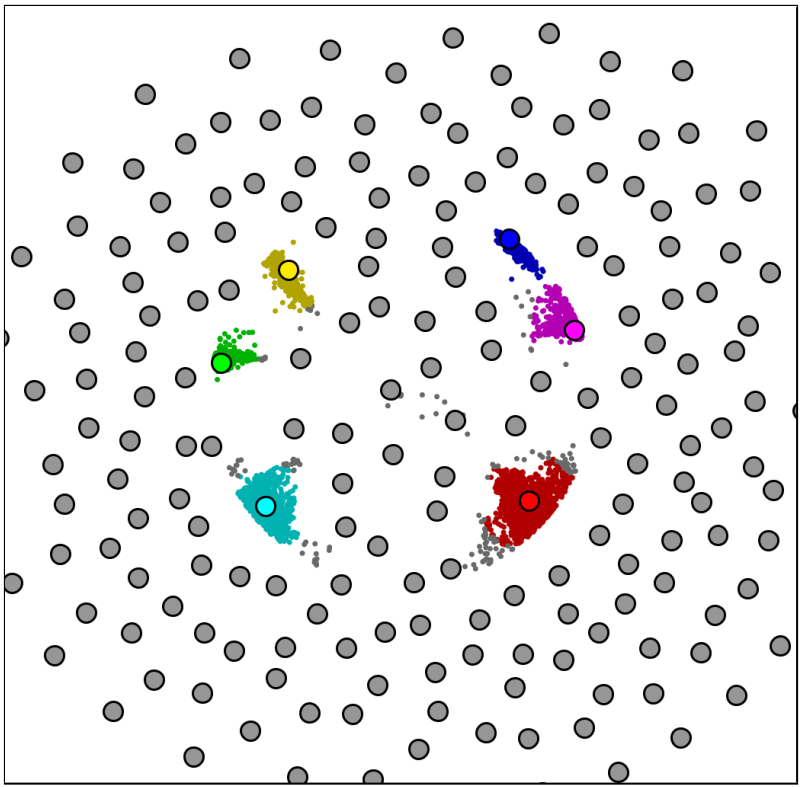
left $\nu = 0.2, \quad \sigma = 0.9 \cdot \nu, \quad \gamma = 1.8 \cdot \nu, \quad \alpha = 0.05$

right $\nu = 0.1, \quad \sigma = 0.9 \cdot \nu, \quad \gamma = 1.8 \cdot \nu, \quad \alpha = 0.05$

Noise Clustering initialized with full Prototypes

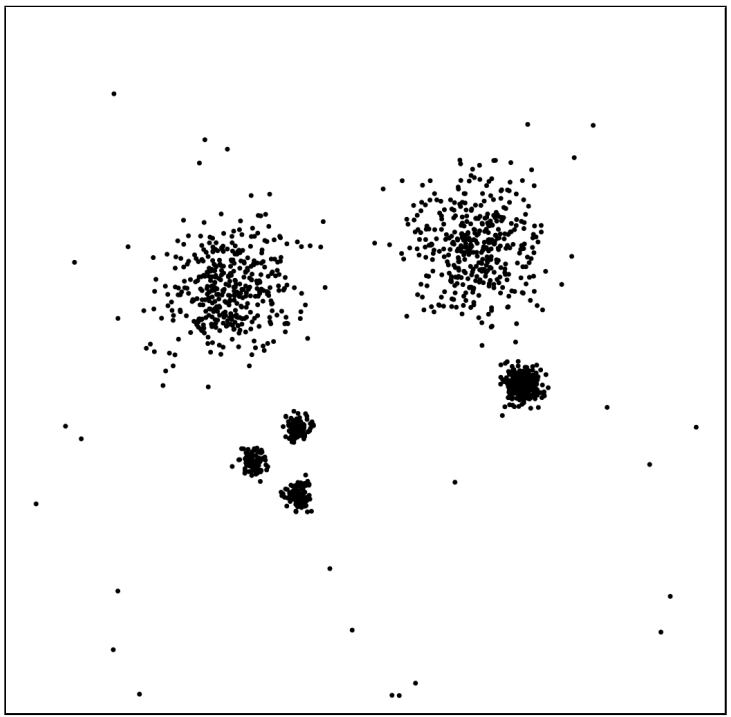


Example Dataset with many Prototypes

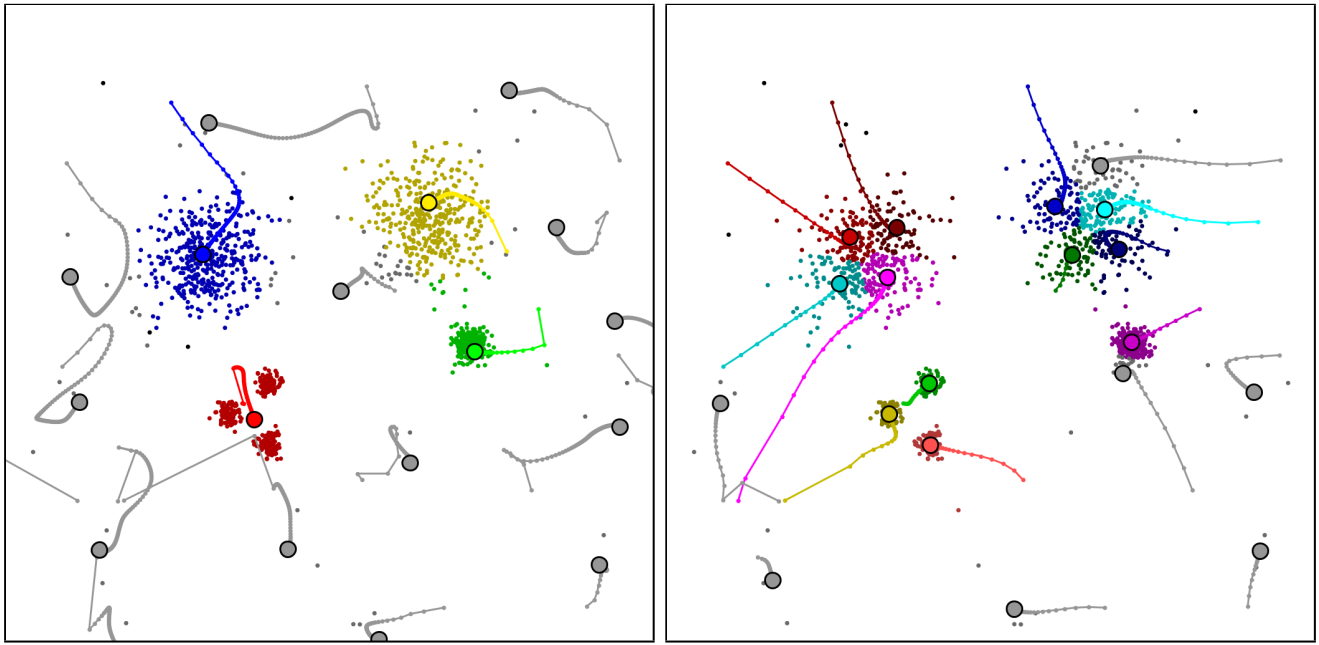


Problematic Dataset

- At least one vast cluster in large distance to other clusters
- At least 2, well seperated clusters in close proximity to each other



Repulsive Prototypes on the Problematic Dataset



4. Conclusions

Conclusions

- **It is possible use knowledge such as cluster size and distance between clusters to estimate the number and position of clusters in a dataset**

Future Work

- **Test repulsive prototypes on high dimensional datasets, measure the success with several different measurements and compare the results with other clustering algorithms**
- **Apply repulsive prototypes with various other, prototype based clustering algorithms**
- **Find an objective function for clustering with repulsive prototypes, or proove that it does not exist**
- **An open source implementation**