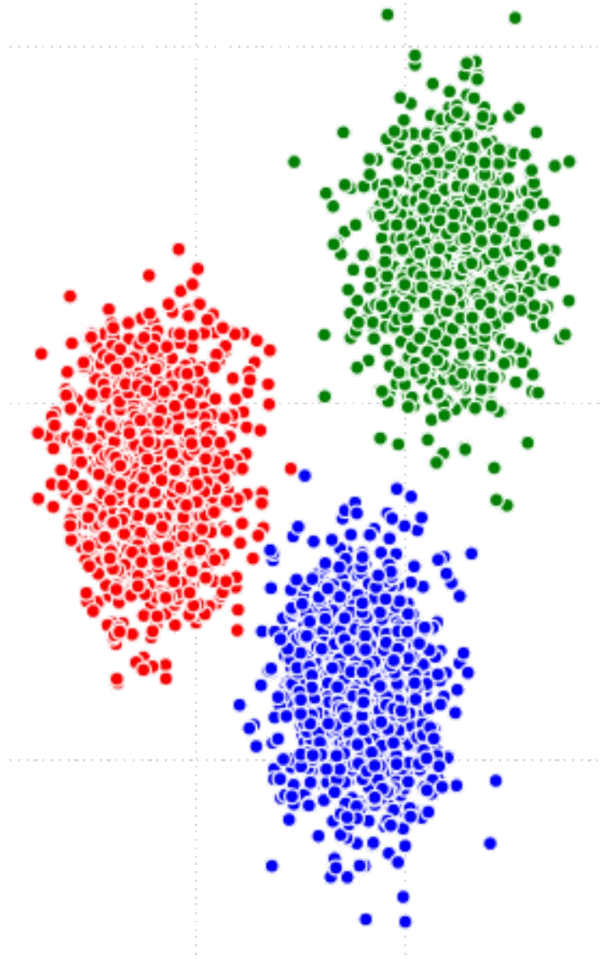# Clustering of complex datasets
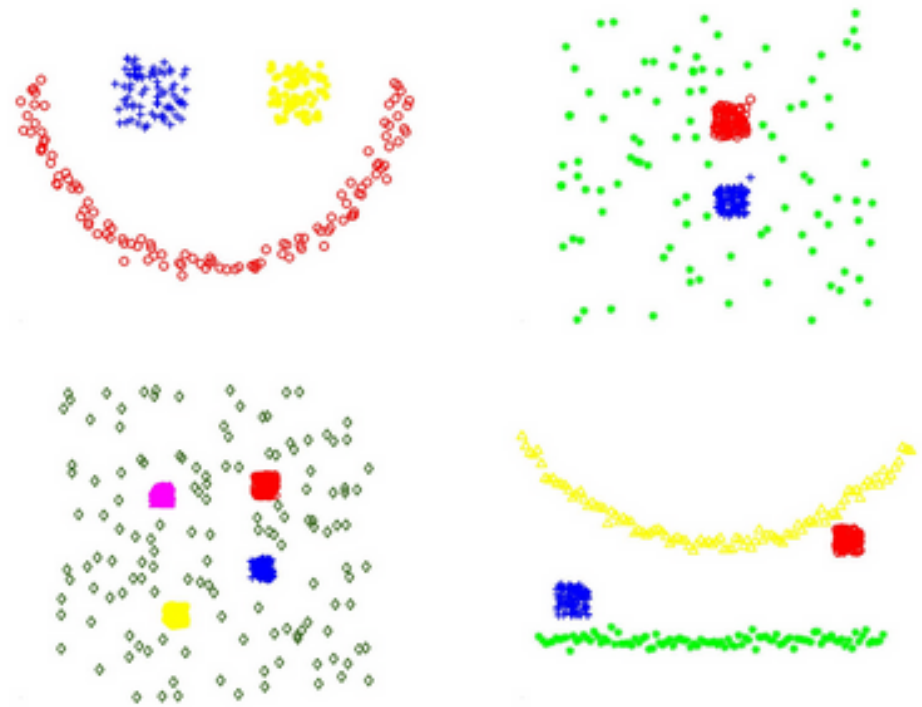
Milestone Presentation

Ahmed Tidjani, Fabio Buso, Paul Velthuis & Zahin Azher
Service-centric Networking | Tu Berlin | May 11th, 2016

Technische
Universität
Berlin

TELEKOM
INNOVATION
LABORATORIES

Service-centric
NETWORKING

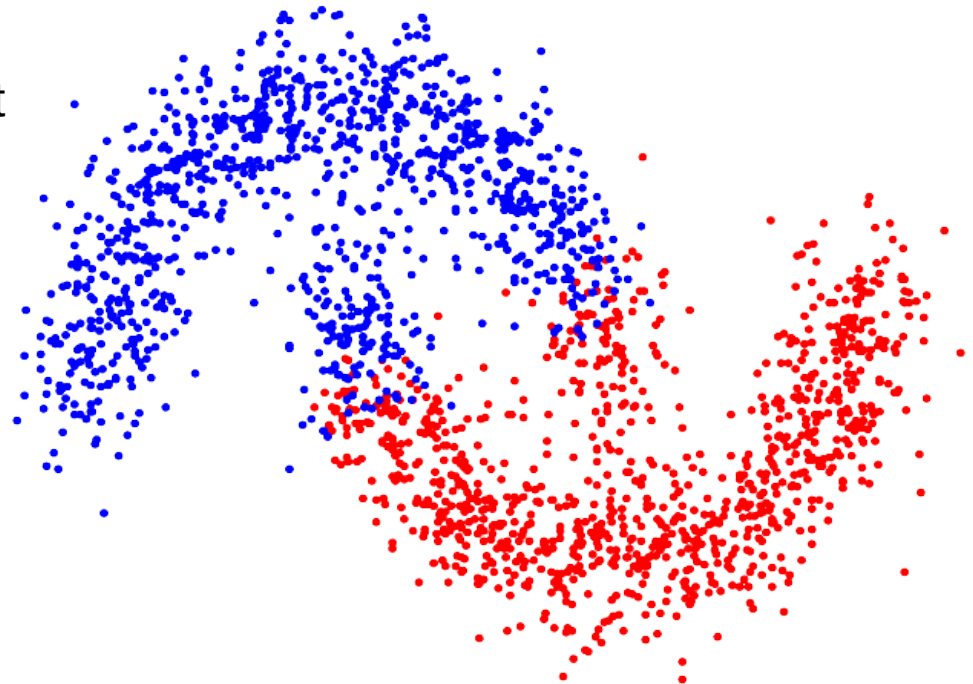# Complex Datasets

Normal Dataset

Complex Datasets

Clustering

Grouping a set of objects in a way that similar objects are in the same cluster.

Types:
- Connectivity Based (Hierarchical)
- Centroid Based (kMeans)
- Density Based (OPTICS)
- Self-Tuning Spectral Clustering

Kmeans Clustering
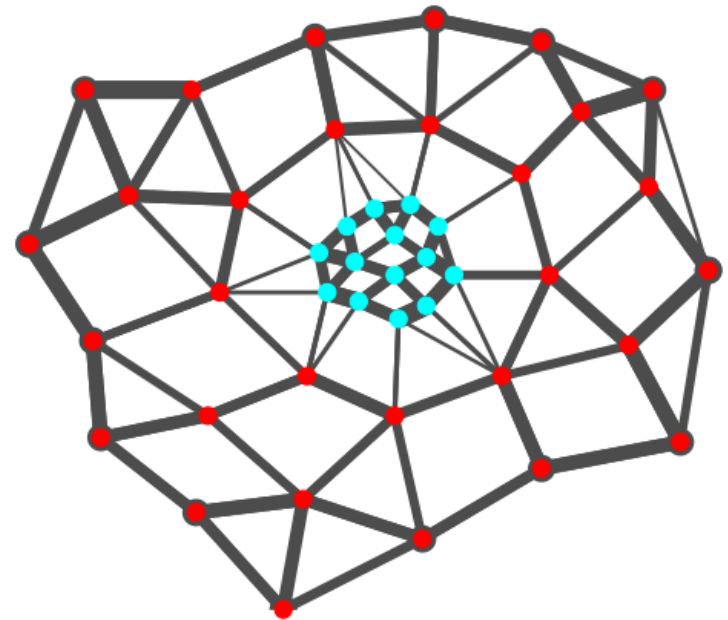
Self-tuning Spectral Clustering[2]

Requires:
- Max number of Clusters

Builds the affinity matrix based on the relations between objects in the dataset.

Operate a dimensionality reduction based the eigenvectors of the matrix.

Exploits the eigenvectors structure to determine the optimal number of clusters.

Building of the Affinity Matrix:

$$\hat{A}_{ij} = e^{\left(\frac{-d^2\left(s_i, s_j\right)}{\sigma_i \sigma_j}\right)}$$

$$\sigma_i = d\left(s_i, s_K\right), K = 7$$

Self-tuning Spectral Clustering[2] - contd

Implementations already available in: C++ and Matlab

Objective:
- Investigate the construction of the affinity matrix
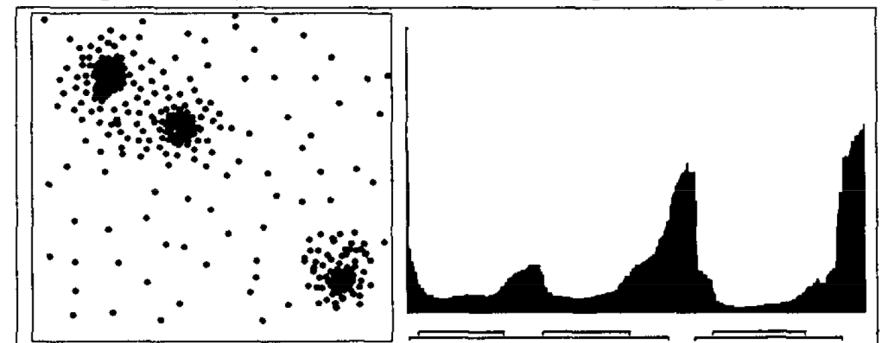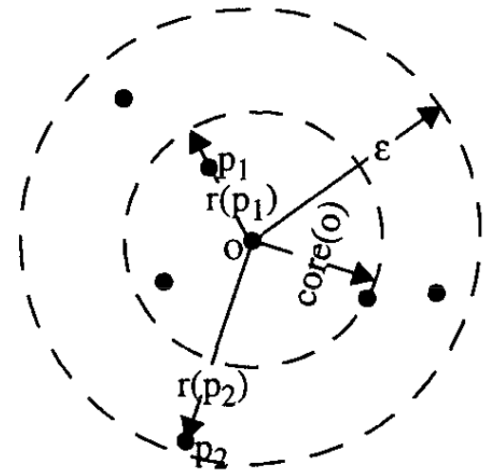- Comparison with OPTICS
- Performance Analysis

OPTICS [1]

Density Based

Requires:
- Min # Points in a Cluster (MinPts),
- Max Neighborhood radius (ε),
- Degree of Steepness (ξ).

Produces Reachability Plot

Automatic cluster extraction from the plot

OPTICS[1] - contd

Implementations already available in: Java, Python, R and Matlab.

Objective:
- Comparison of the different Implementations
- Focus on the automatic cluster extraction
- Performance Analysis

Problem:
- Not all implementations extract automatically the clusters

Visualization Tool

Website to run the experiments and visualize the results.

Possibility to set the parameters and see the result in real time.

Plot with the Javascript library D3.js

Backend in Python to execute the different algorithms.
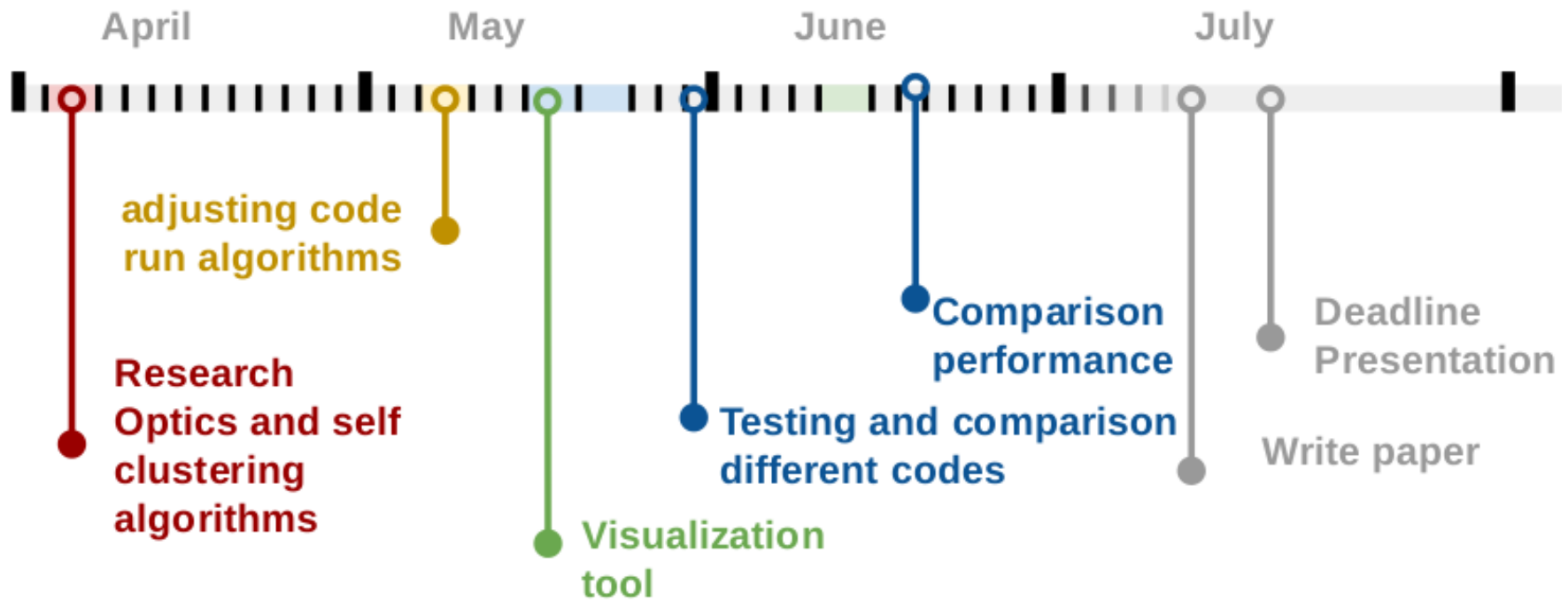
Load Dataset

OPTICS Result

Min Points
ε
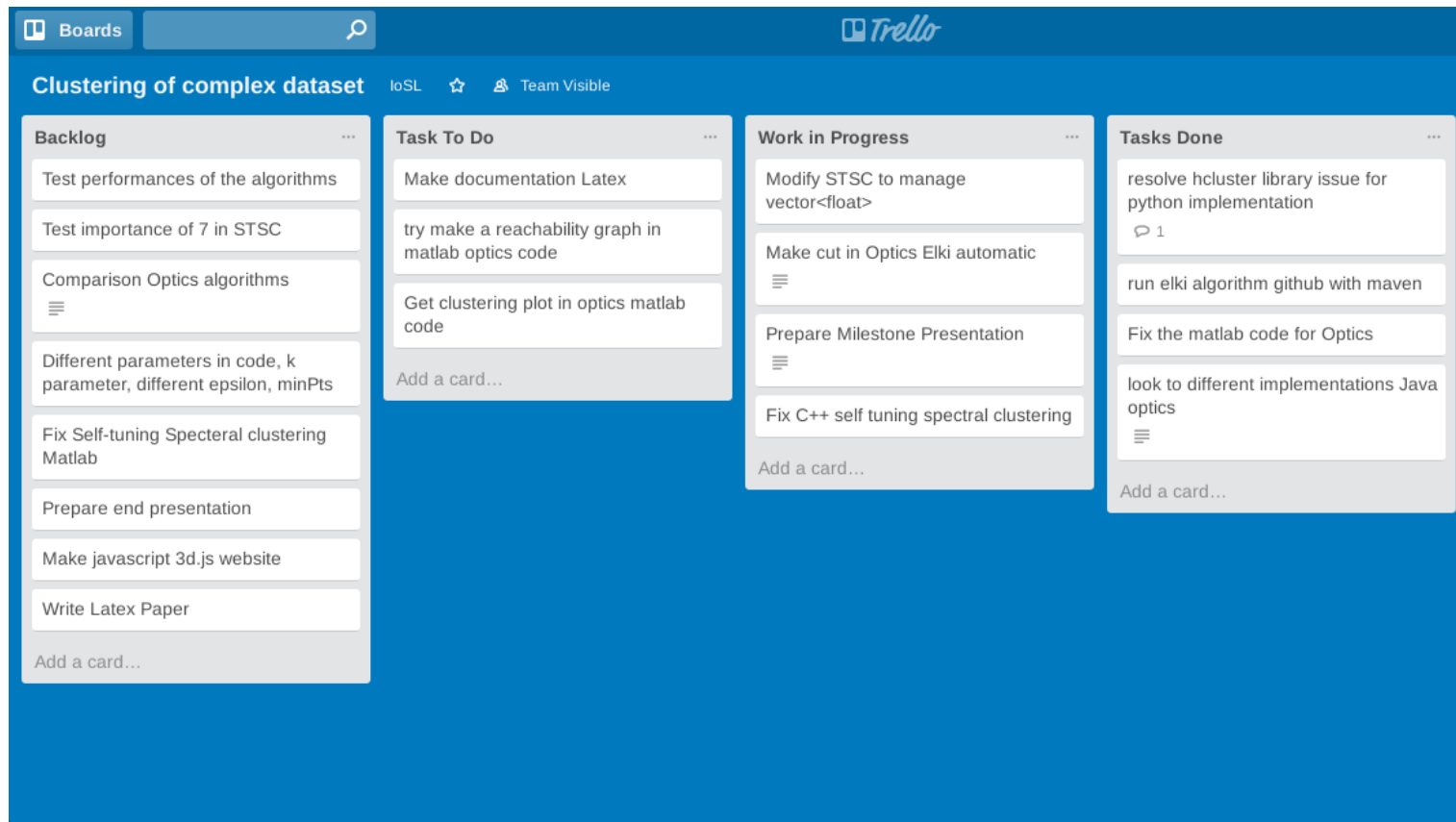Stepness Degree

Self Tuning Spectral Clustering Result
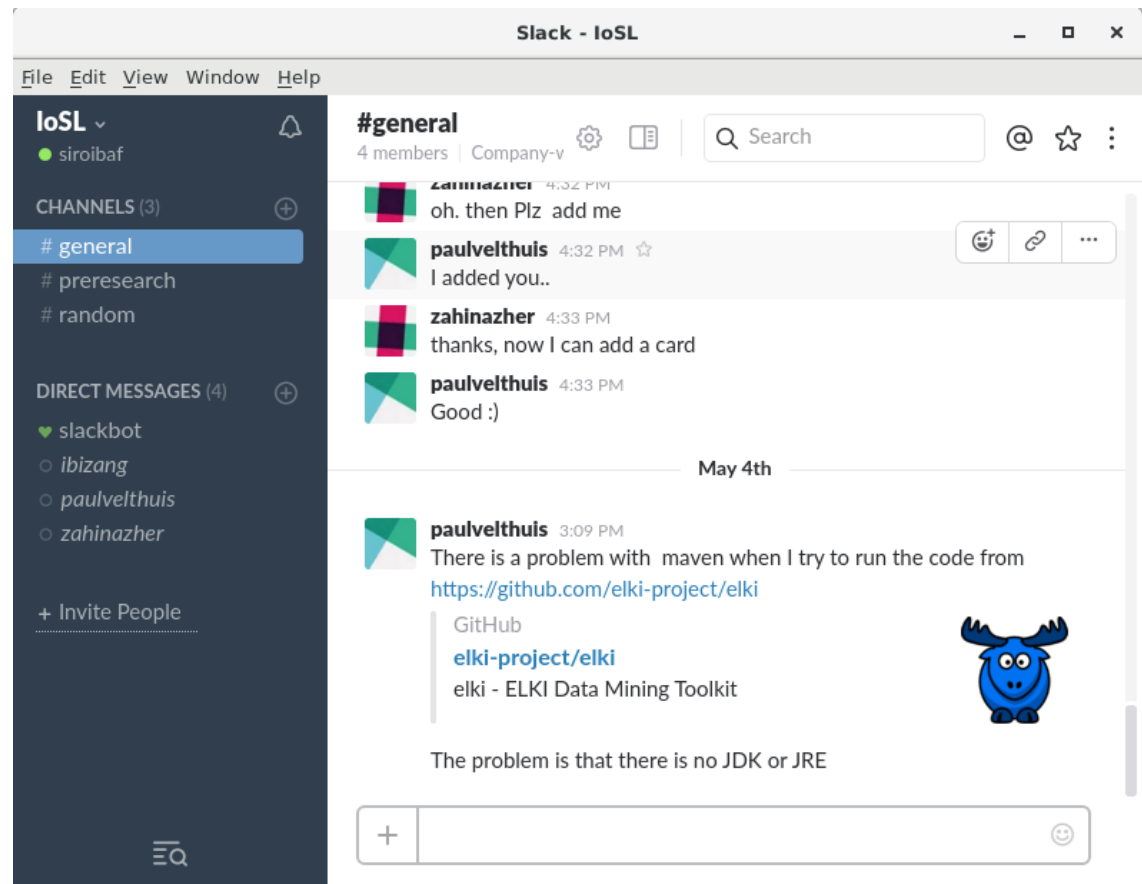
K

ORGANIZATION

Timeline

April

May

June

July

adjusting code
run algorithms

**Research
Optics and self
clustering
algorithms**

**Visualization
tool**

**Testing and comparison
different codes**

**Comparison
performance**

Deadline
Presentation

Write paper

Tools

Trello

Clustering of Complex Datasets
Ahmed Tidjani, Fabio Buso, Paul Velthuis & Zahin Azher

Tools

Trello

Slack

Tools

Trello

Slack

GitLab (Wiki)

THANK YOU FOR YOUR ATTENTION
QUESTIONS?

References

[1] Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999, June). OPTICS: ordering points to identify the clustering structure. In ACM Sigmod Record (Vol. 28, No. 2, pp. 49-60). ACM.

[2] Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. InAdvances in neural information processing systems (pp. 1601-1608).

Clustering of Complex Datasets
Ahmed Tidjani, Fabio Buso, Paul Velthuis & Zahin Azher