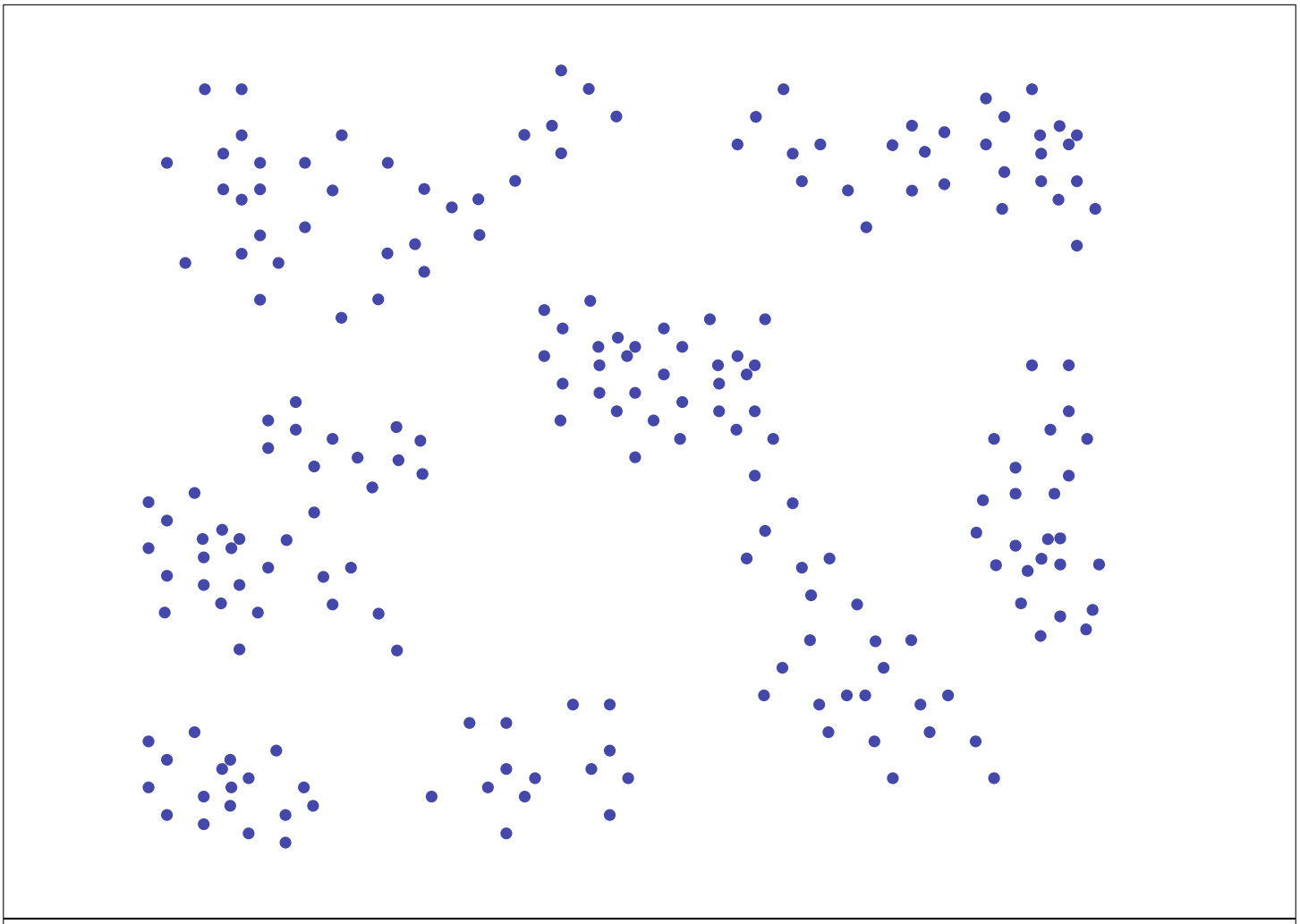
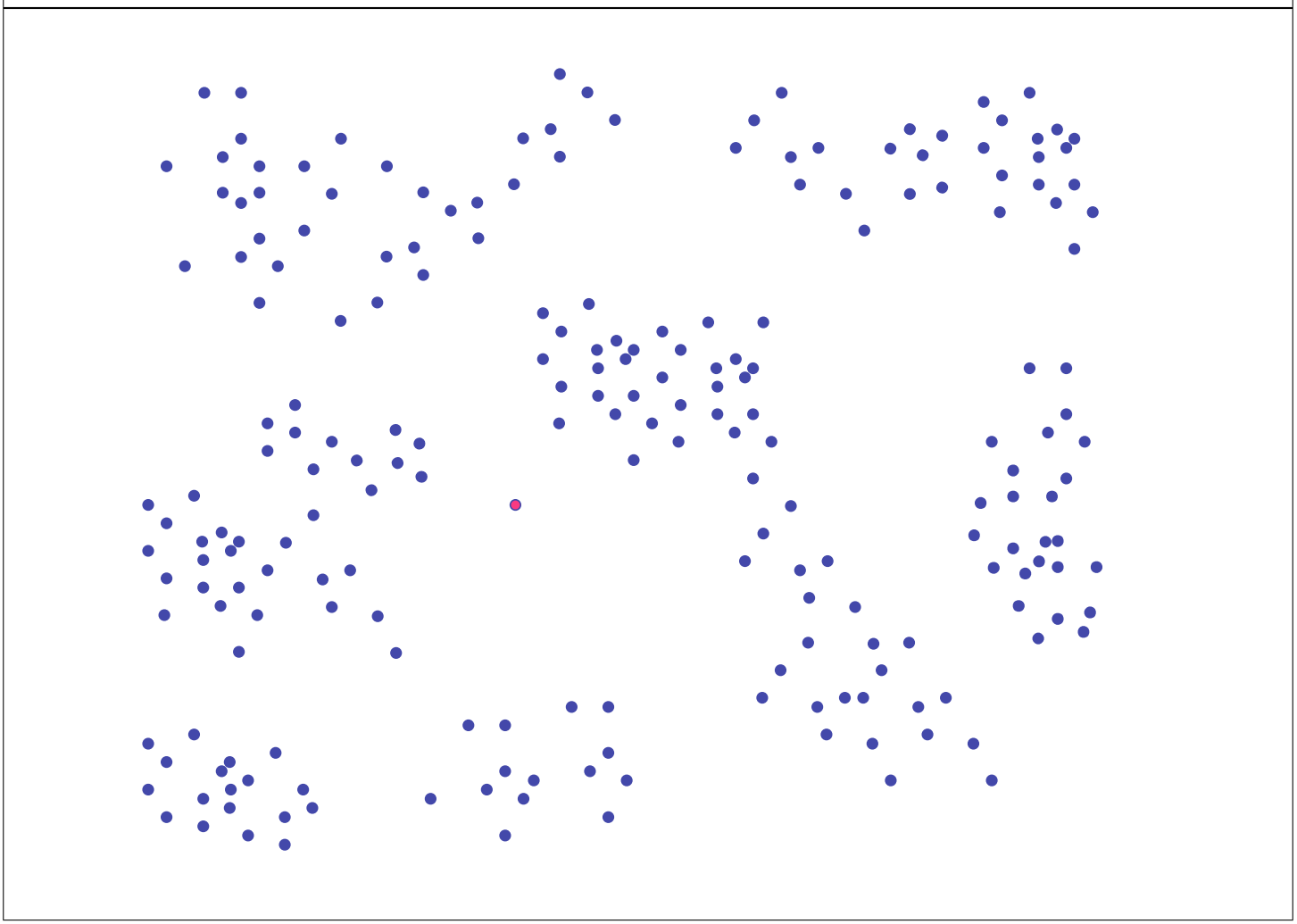
The Music of the (p) Spheres

Alessandro Panconesi, La Sapienza of Rome

Joint work with: Prabakhar Raghavan, Mauro Sozio, Alessandro Tiberi, Eli Upfal

Nearest Neighbours



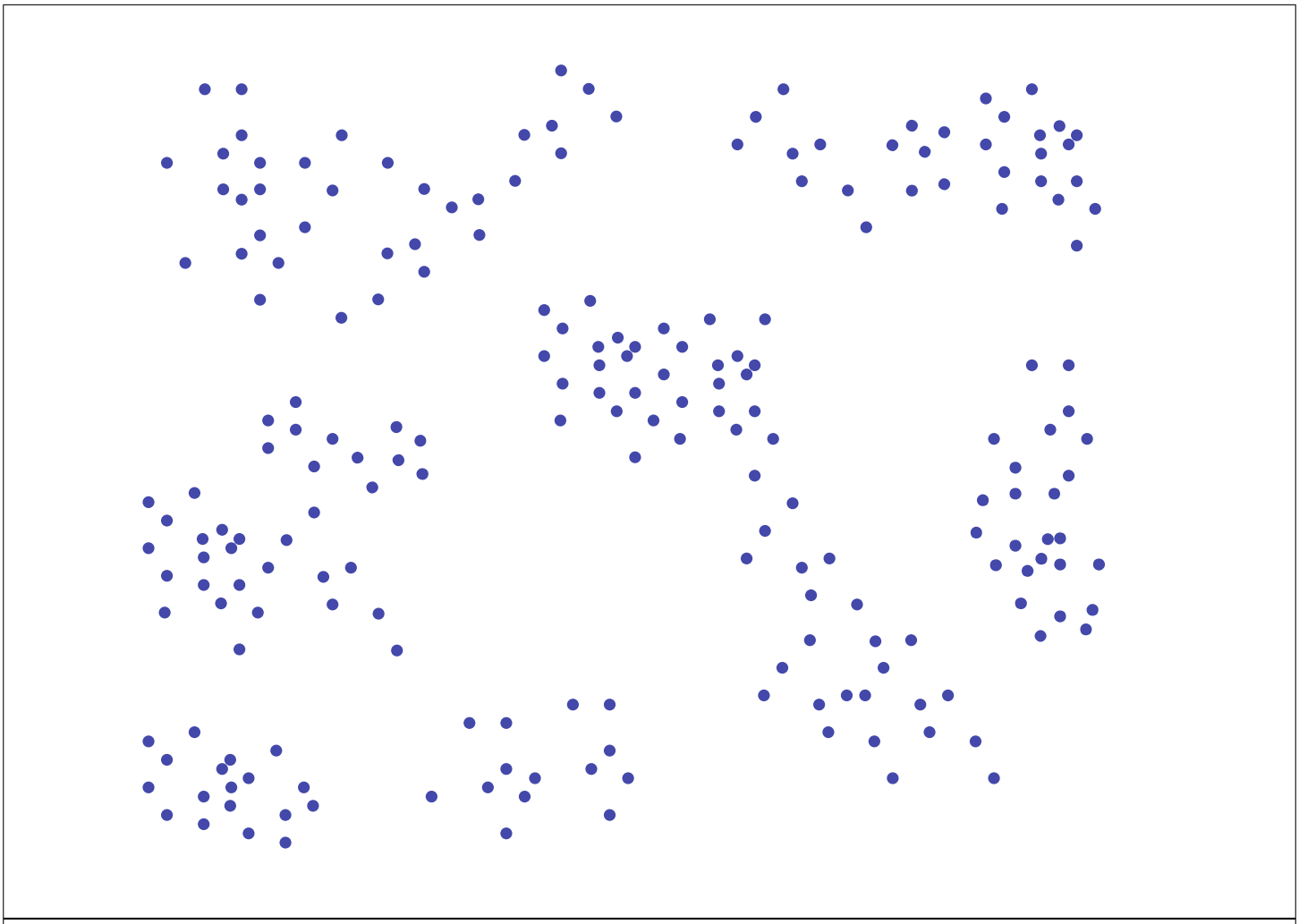


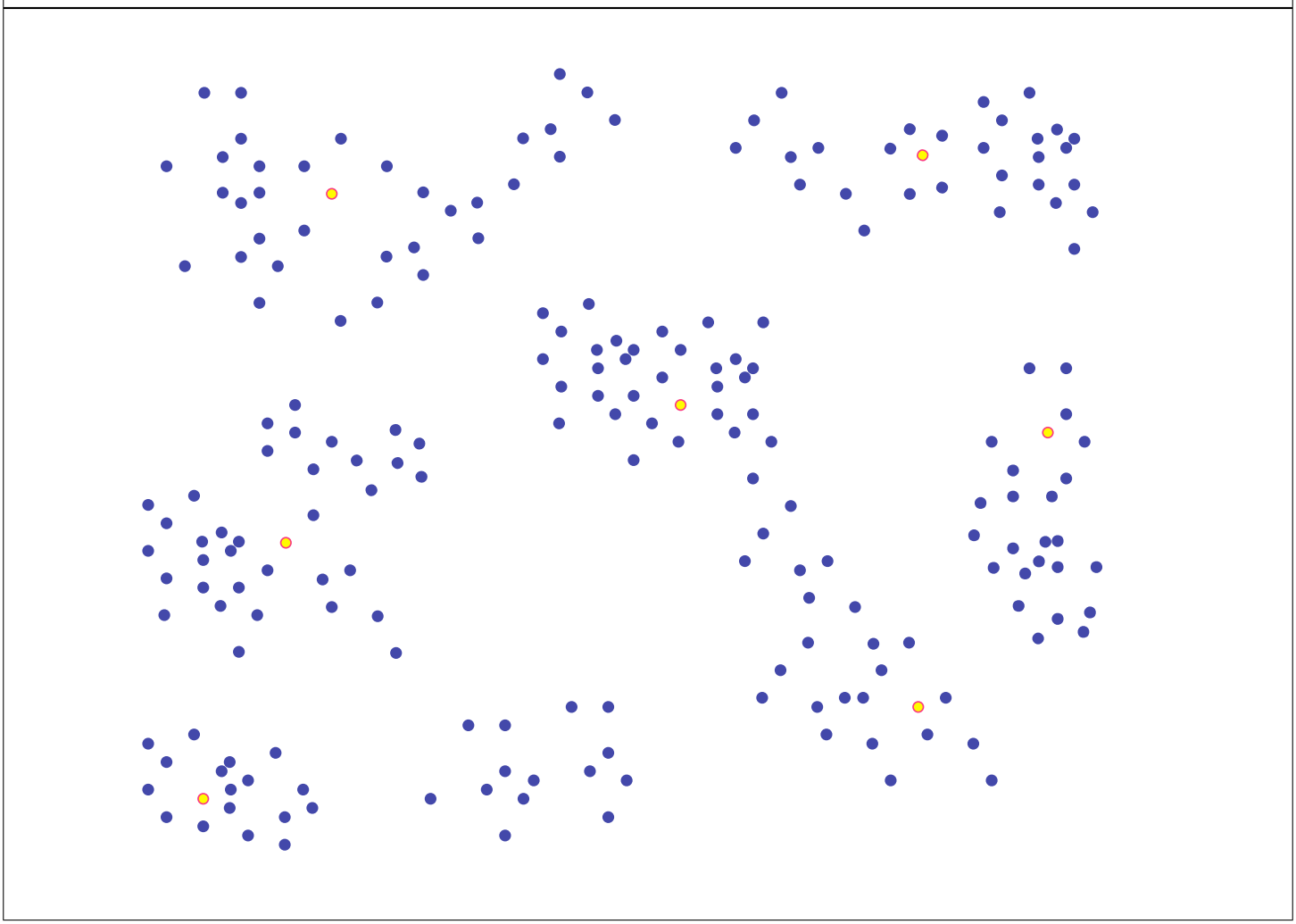
Nearest Neighbour

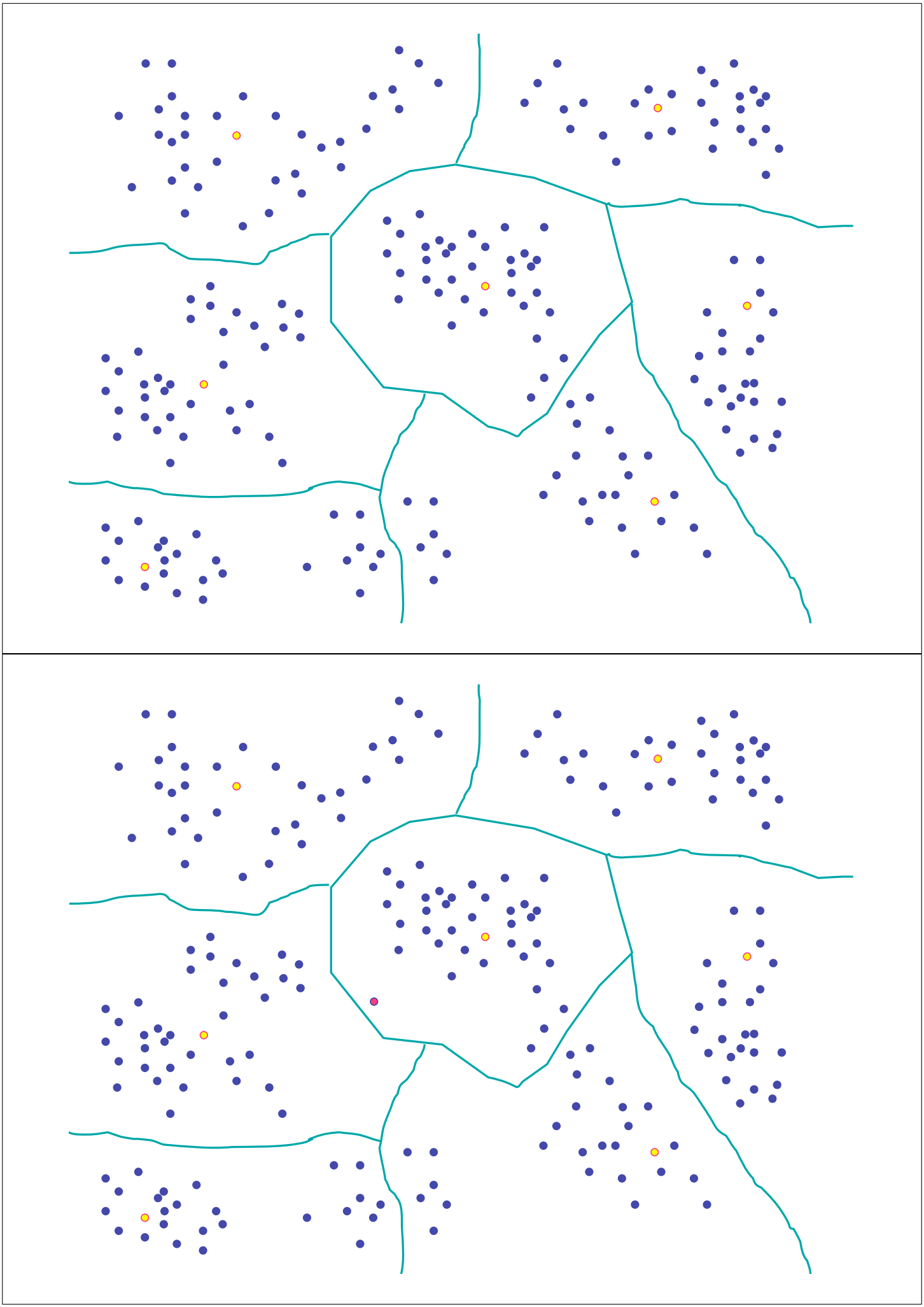
Euclidean distance

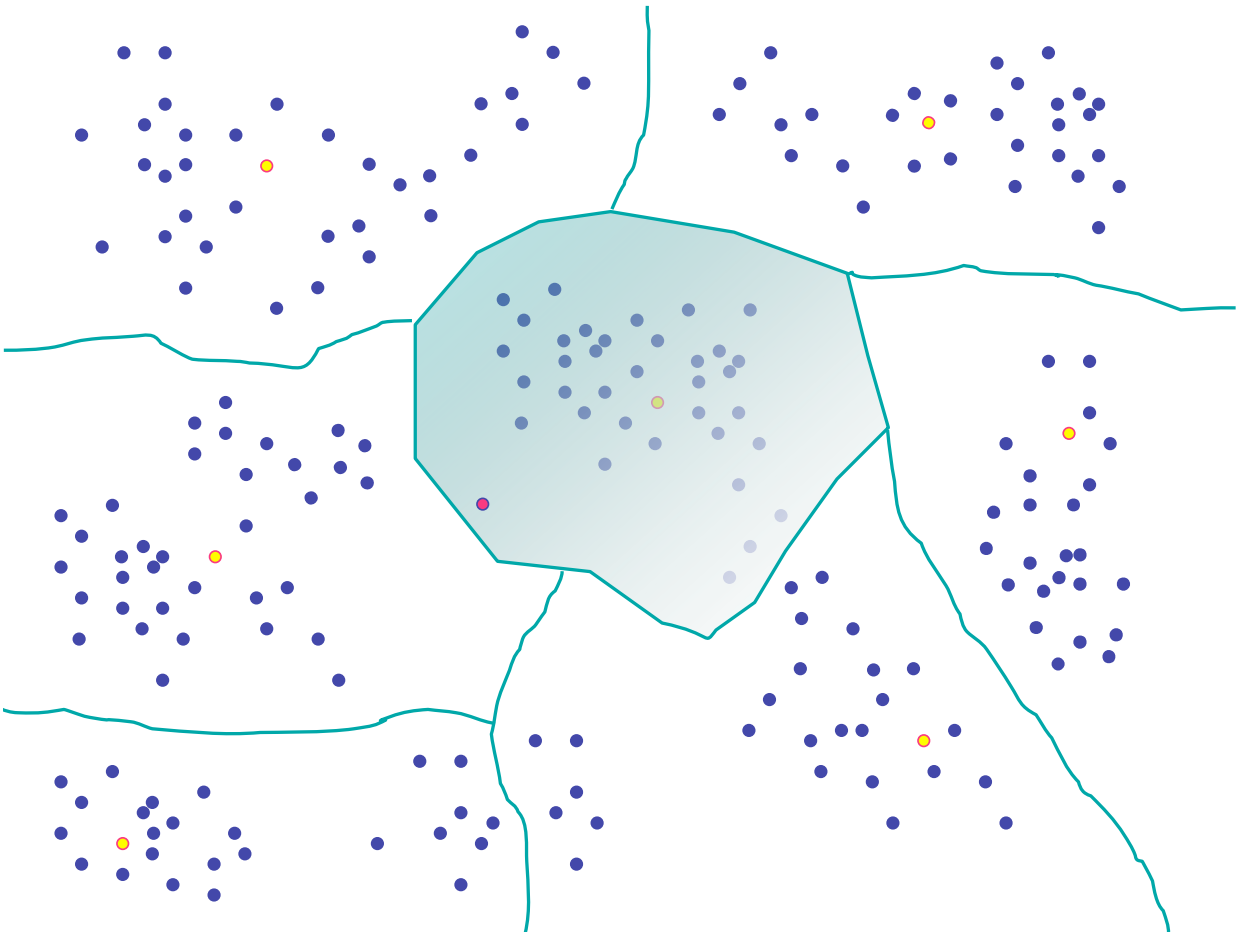
Cosine similarity

Random Clustering





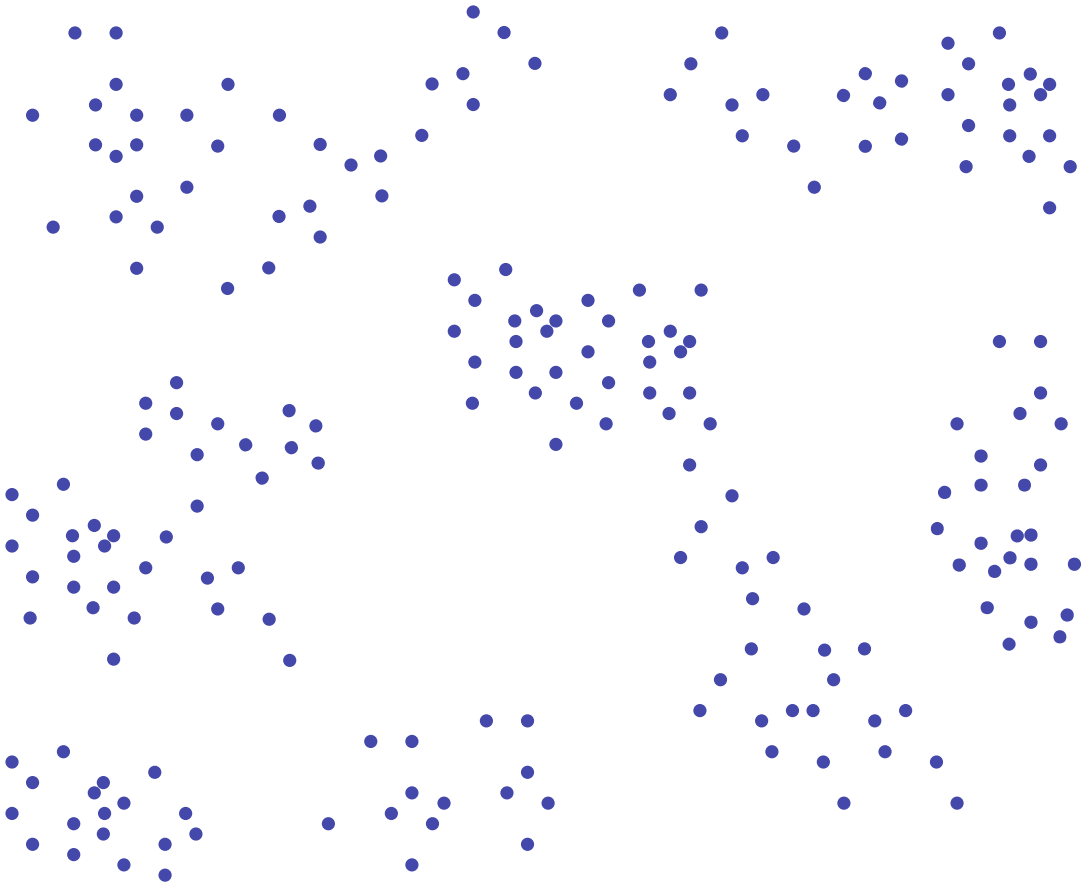


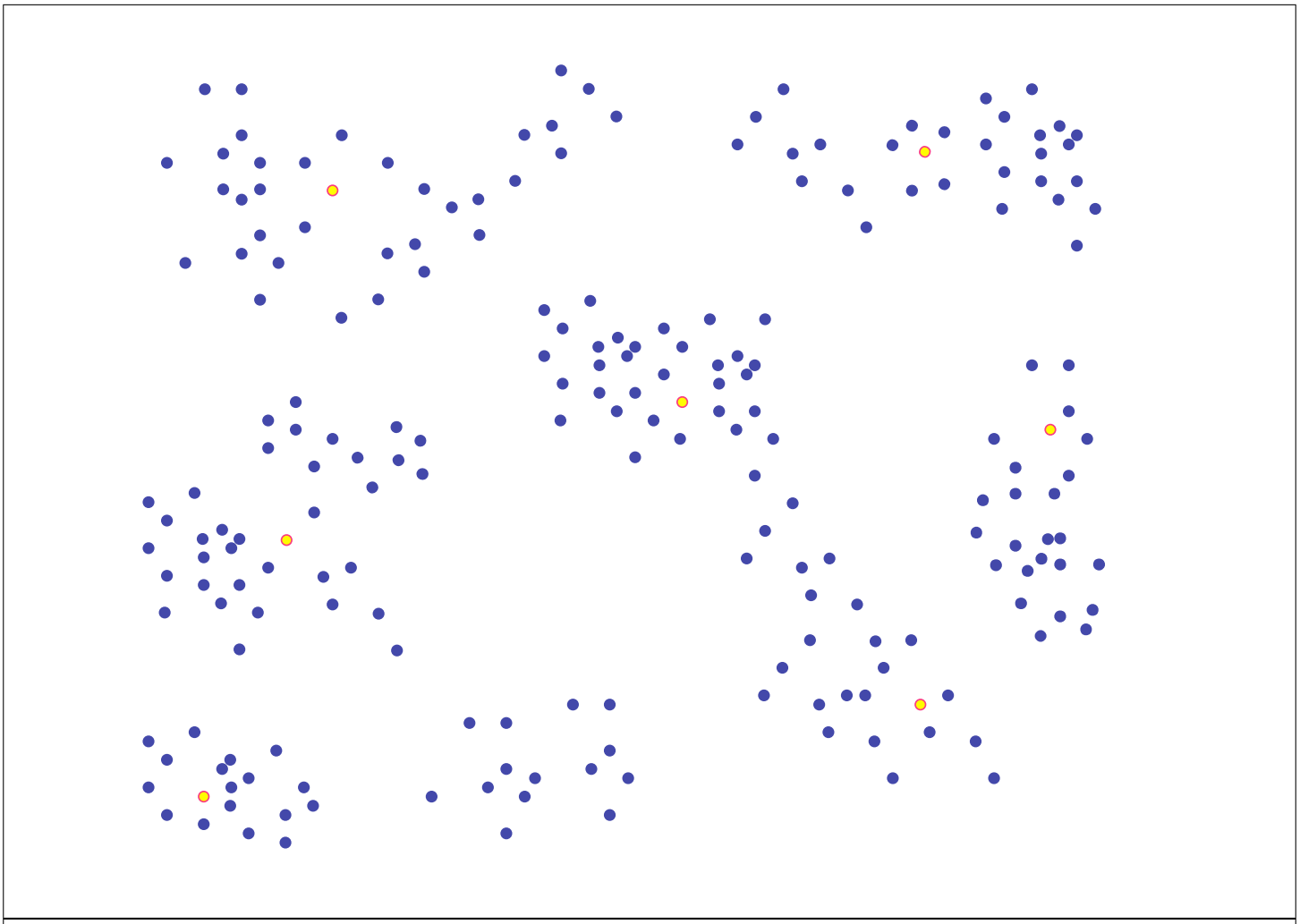


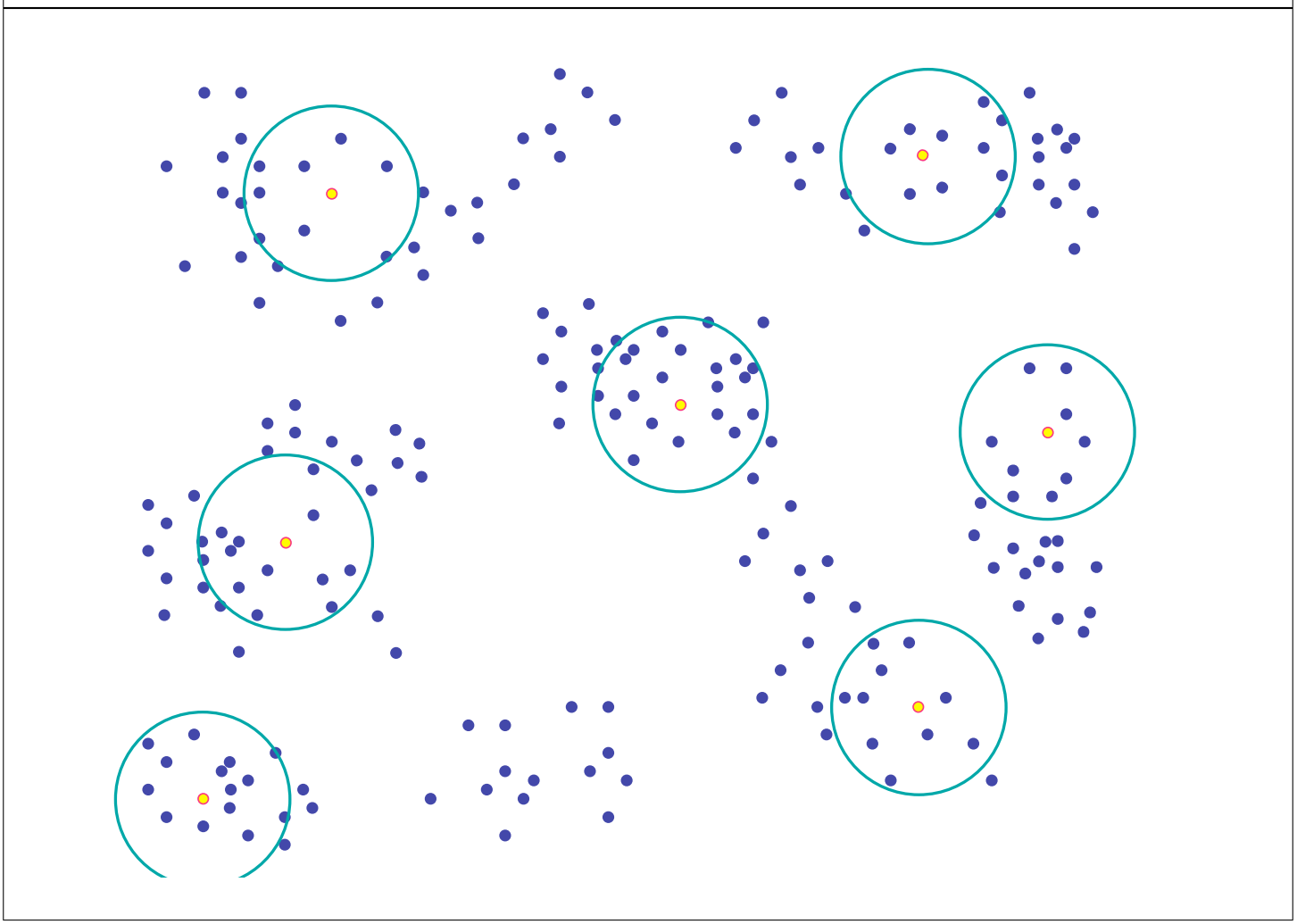
P-spheres

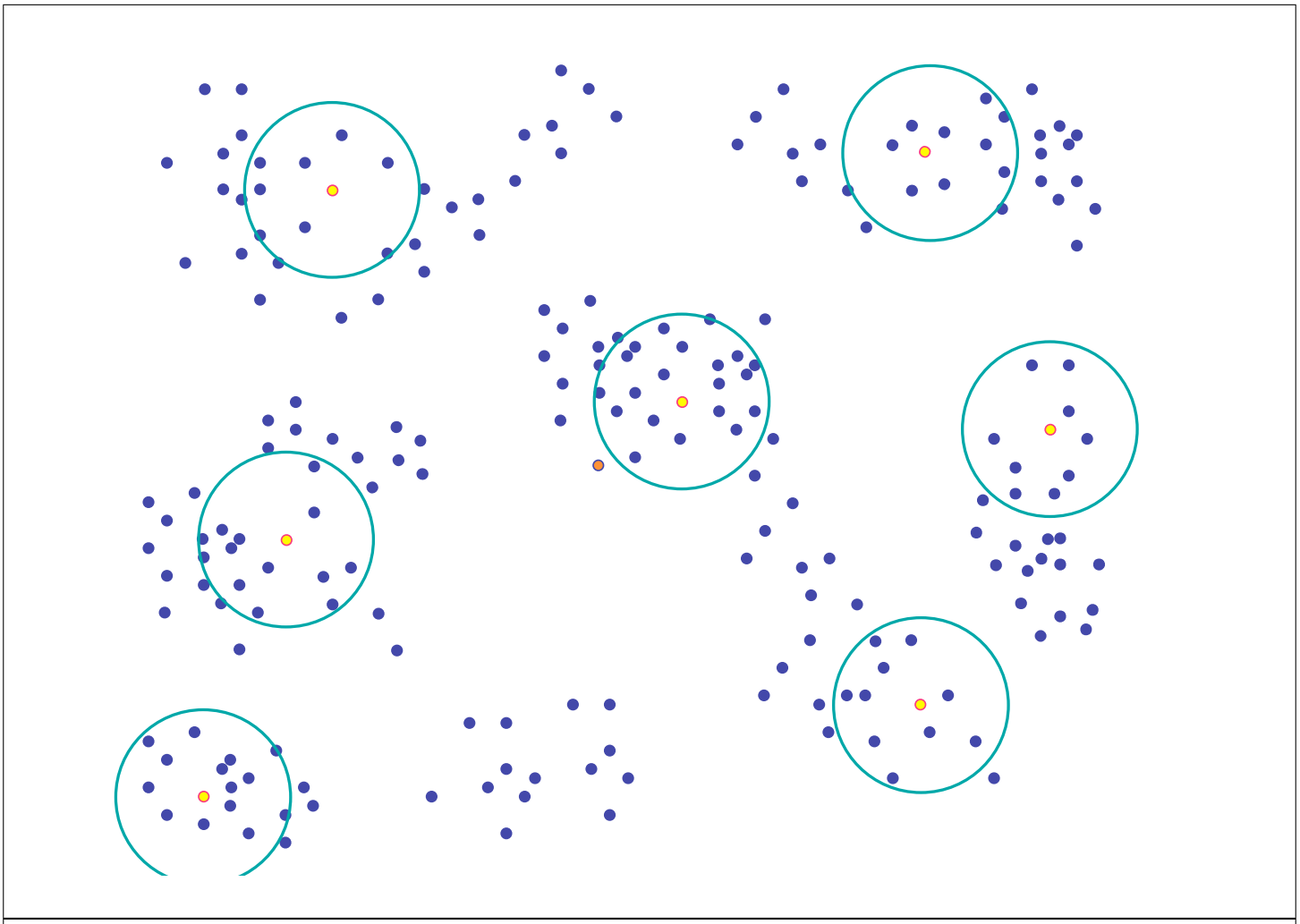
Once upon a time

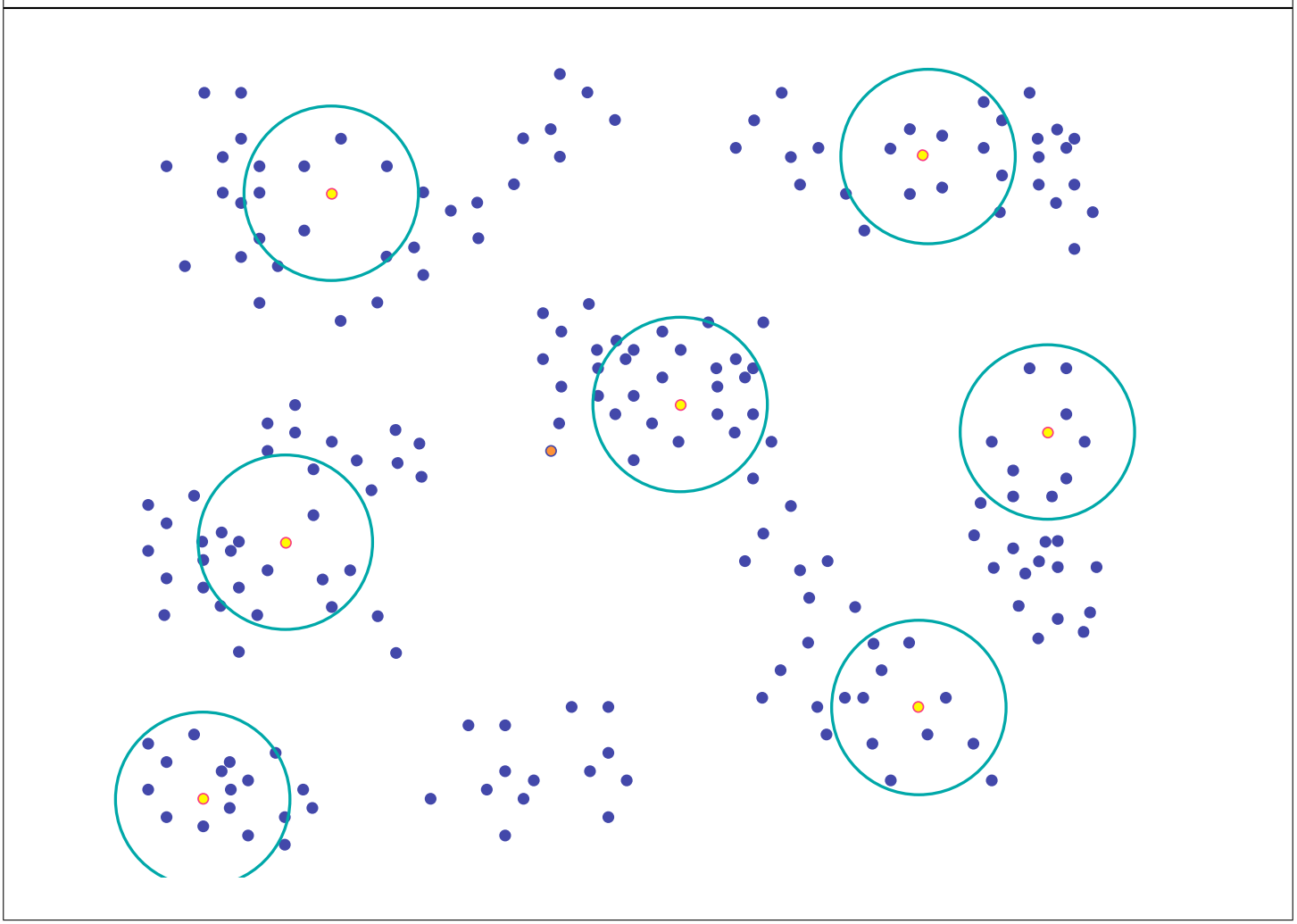
- **Musica universalis** or **music of the spheres** is a medieval philosophical concept that regards the proportions in the movements of the celestial bodies - the Sun, Moon and planets - as a form of musica, the medieval Latin name for music. This music was not thought of as an audible sound, but simply as a mathematical concept. The Greek philosopher Pythagoras was frequently credited with originating the concept, which stemmed from his semi-mystical, semi-mathematical philosophy and its associated system of numerology of Pythagoreanism. Some Surat Shabda Yoga, Satgurus considered the music of the spheres to be a term synonymous with the Shabda or the Audible Life Stream in that tradition, because they considered Pythagoras to be a Satguru as well.

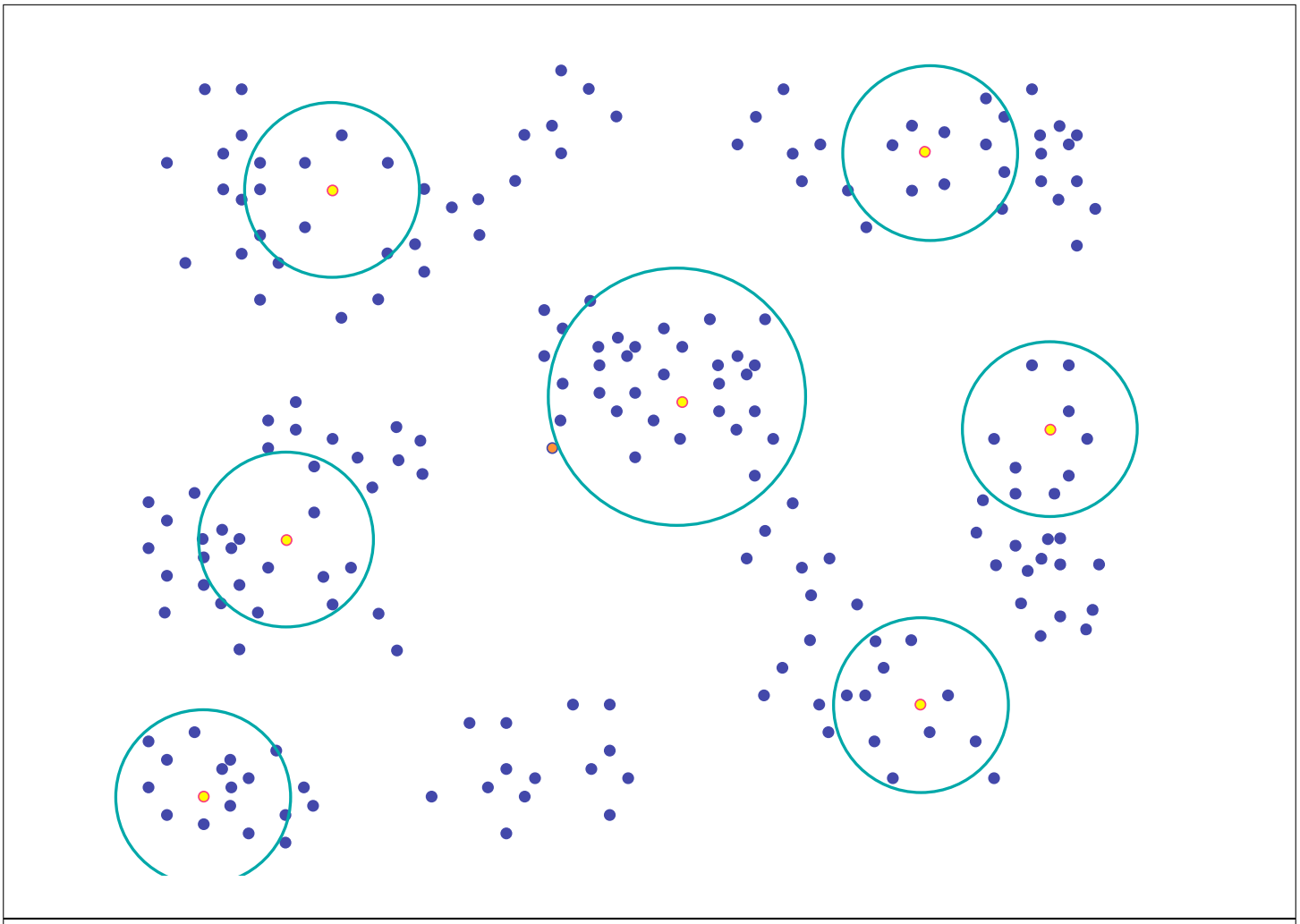


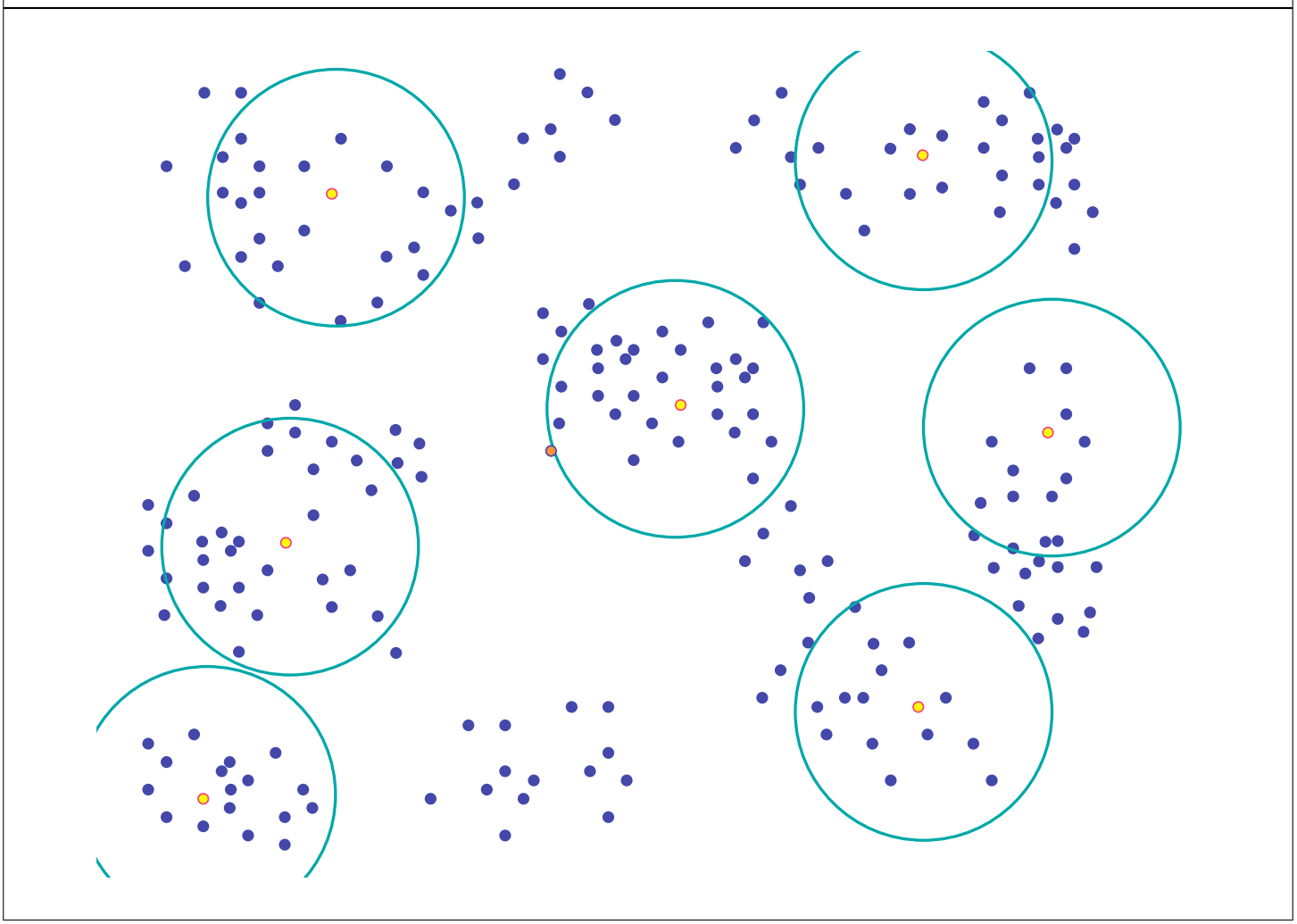






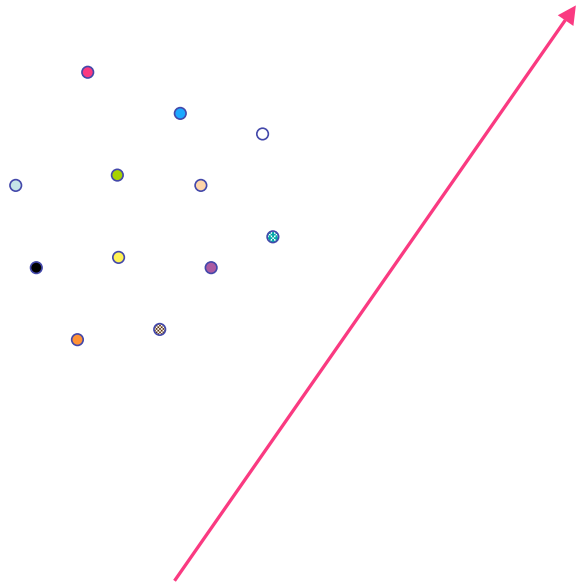




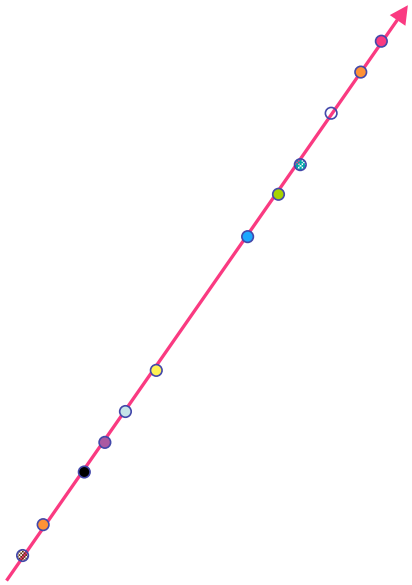


Rank Aggregation

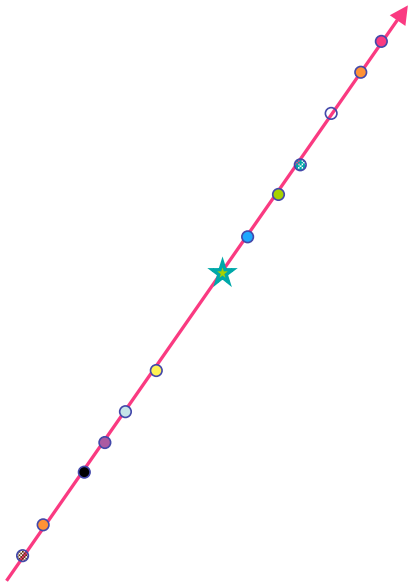
Rank Aggregation



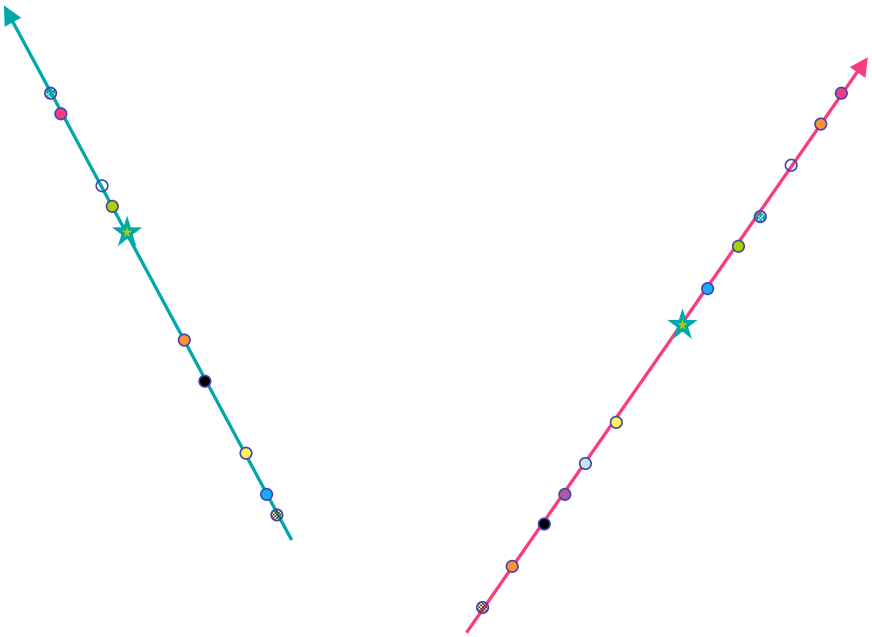
Project docs on a random line



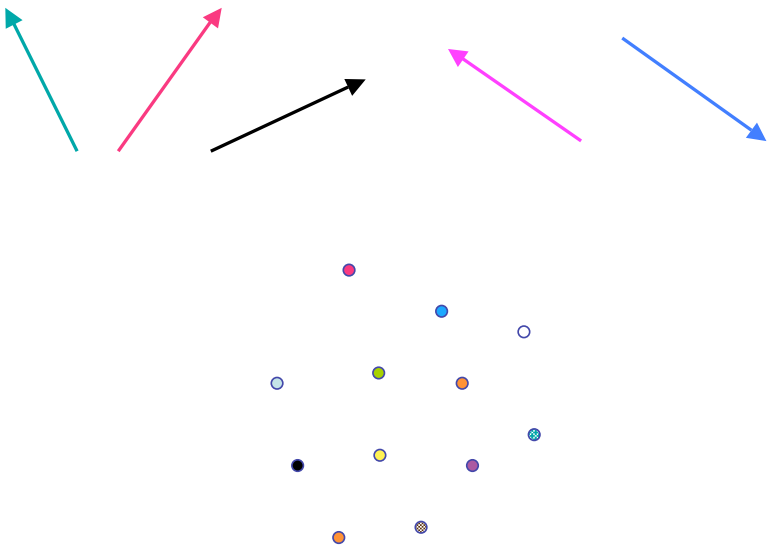
Project query: closest doc gets one vote



Repeat with a set of random lines

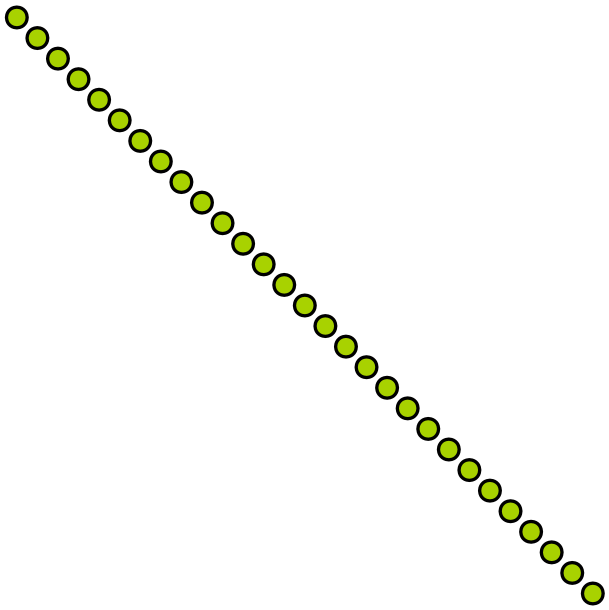


Elections

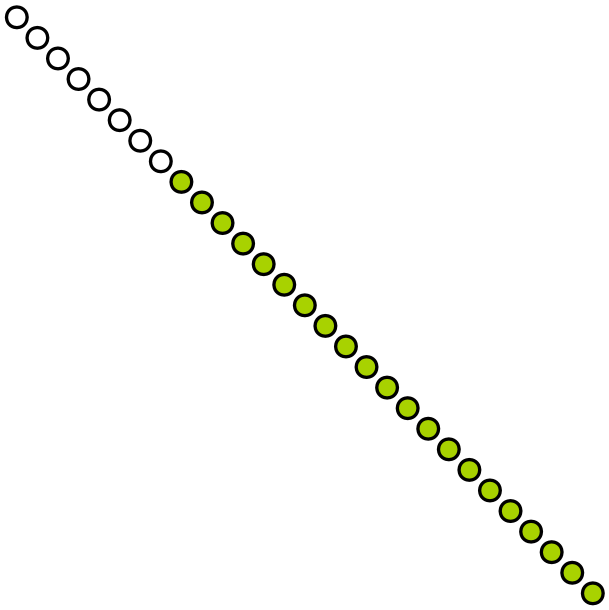


Evaluation

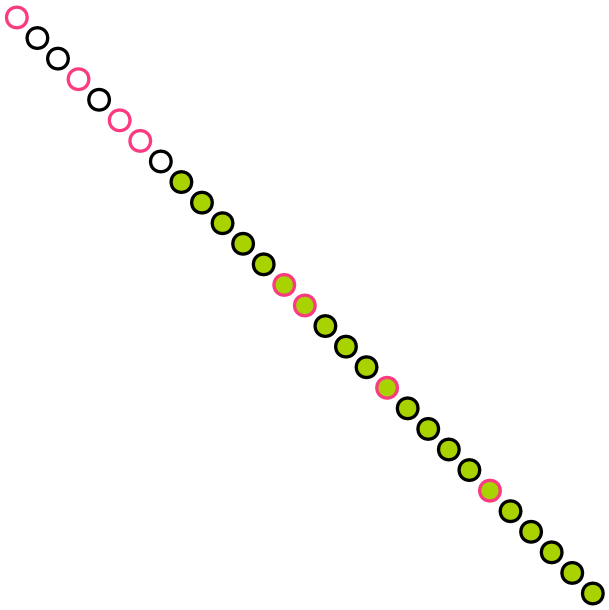
Competitive Recall

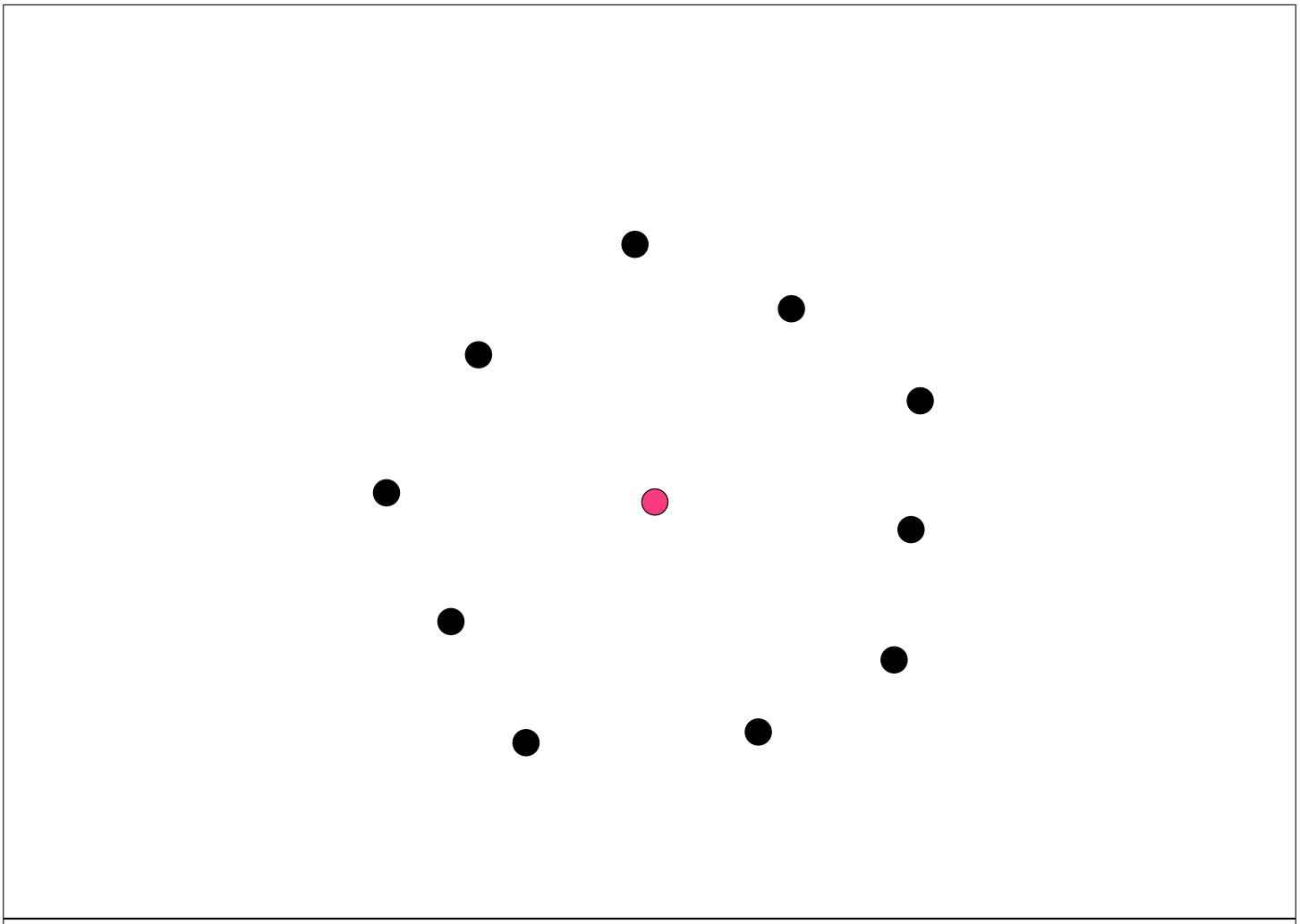


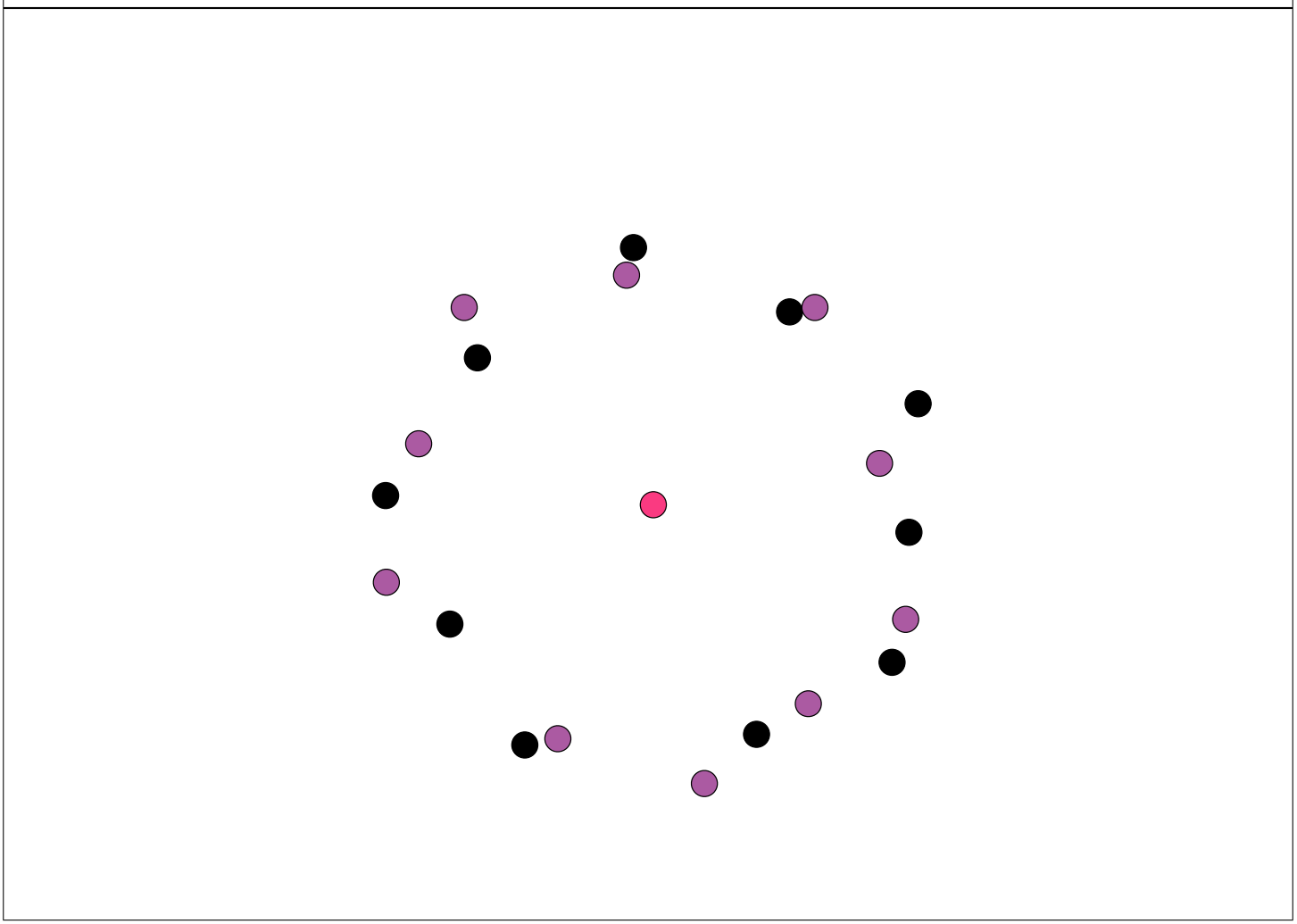
Competitive Recall



Competitive Recall







Competitive Similarity

$$AD(S,q) := \sum_{x \in D} x \bullet q \;/\; |D|$$

Average distance of q from S

Competitive Similarity

$$CS(S,q) := AD(S,q) \;/\; AD(Best,q)$$

We consider a normalised version of this..

The dataset

- *100,000 docs from CiteSeer*
- *#dimensions = 400,000*
- *Normalized to unit vectors*
- *Words were stemmed and stopwords removed*

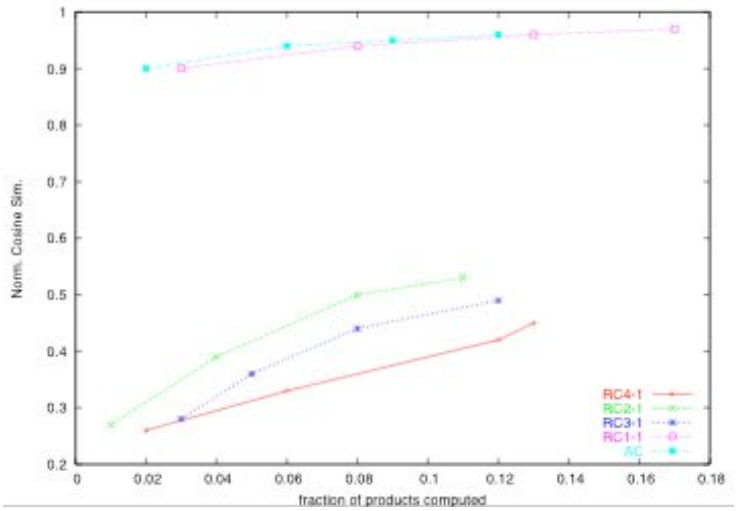
What we measure

- *Quality, ie recall and (normalized) competitive similarity*
- *Computational effort (machine independent)*
- We tried long and short queries

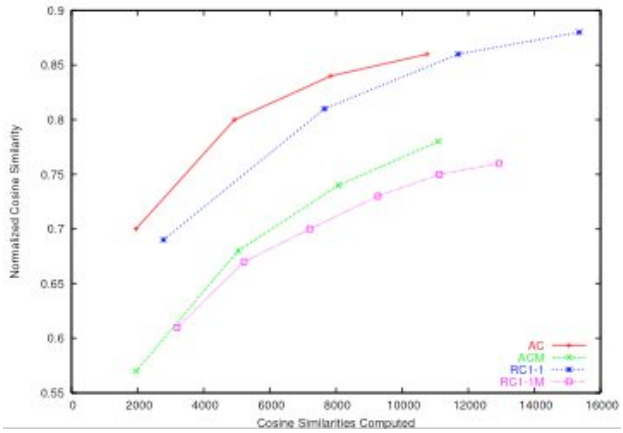
Points of Interest

To the best of our knowledge, this is the 1st empirical study of p-spheres and rank aggregation for text data. These data are characterized by huge number of dimensions and very sparse vectors

Random Clustering: 1 level of recursion is best

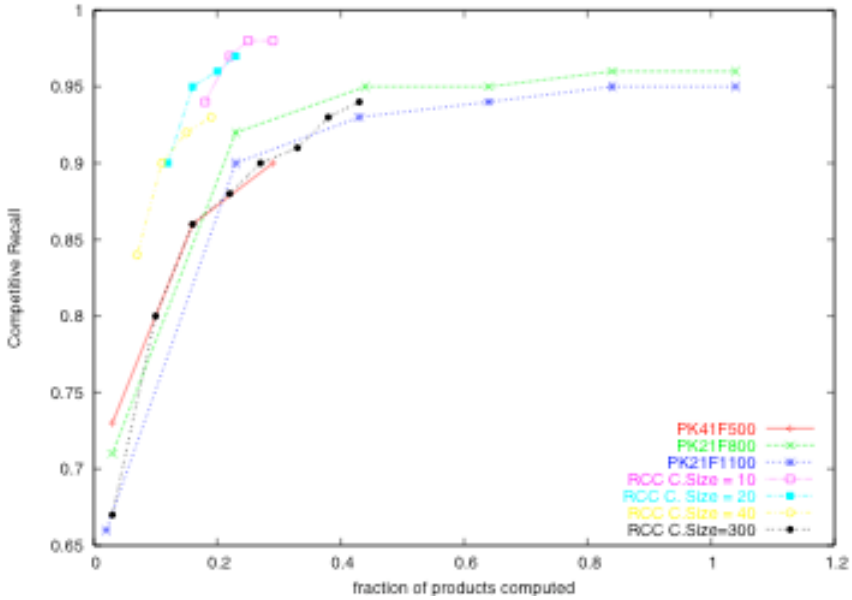


Random Clustering: Centroids are best

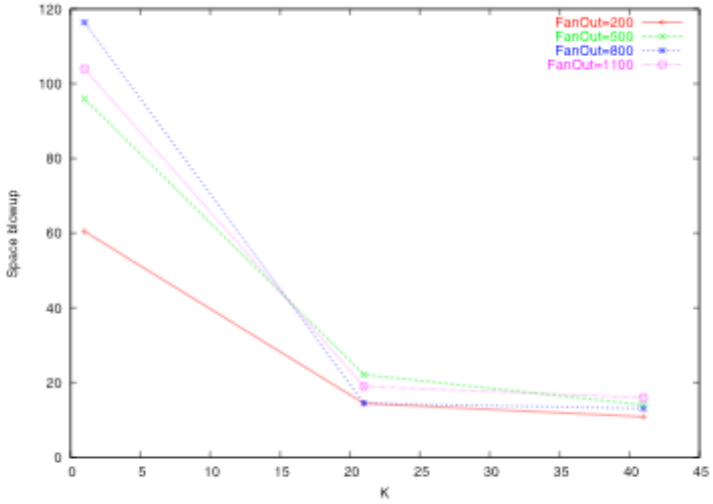


P-Spheres vs Random Clustering

Quality vs Computational Effort



Space:the Final Frontier



The Bottomline

Random Clustering

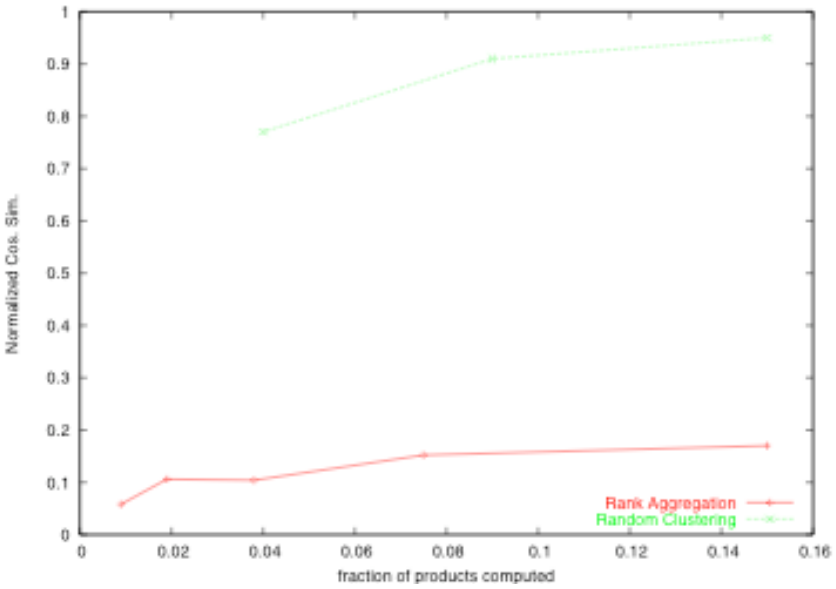
- Space is optimal
- Full coverage of corpus
- Significantly better quality for same effort
- No a-priory knowledge required
- Extremely simple

P-spheres

- Significant space blow-up
- Partial coverage of corpus
- Significantly lower quality for same effort
- A-priory knowledge of query distribution required

Rank Aggregation

Rank Aggregation vs Random Clustering



Possible Explanation

Very sparse vectors, huge dimension (we had to bring the #dimensions from 400,000 to 60,000)

It works well for dense vectors

Sanity check: Approximation of Euclidean distance for our data same as that reported in original paper

Future Directions

- Simple randomized clustering seems to be ***remarkably*** effective
- We have an ***extremely cool*** generative model I did not have time to tell you about with which we can prove amazing things. To do: see how well it fits the data
- We are trying to improve random sampling by combining it with Rank Aggregation
- ..and to augment it with Pagerank

A Challenge

Does anybody know the origin of the word *Yahoo!* ?