Fast Clustering of Flow Cytometry Data via Adaptive Mean Shift

Suchismit Mahapatra

Why Clustering in Flow Cytometry data ?

- Has been used with subset lymphocyte cell populations in multicolor immunophenotyping assays.
- The results showed high agreement with those obtained for well-studied subset populations i.e. CD4, CD8, etc.
- Can provide researchers with new insight into the high-dimensional flow cytometry data.
- Enable development of new methods for automated flow cytometry data analysis.

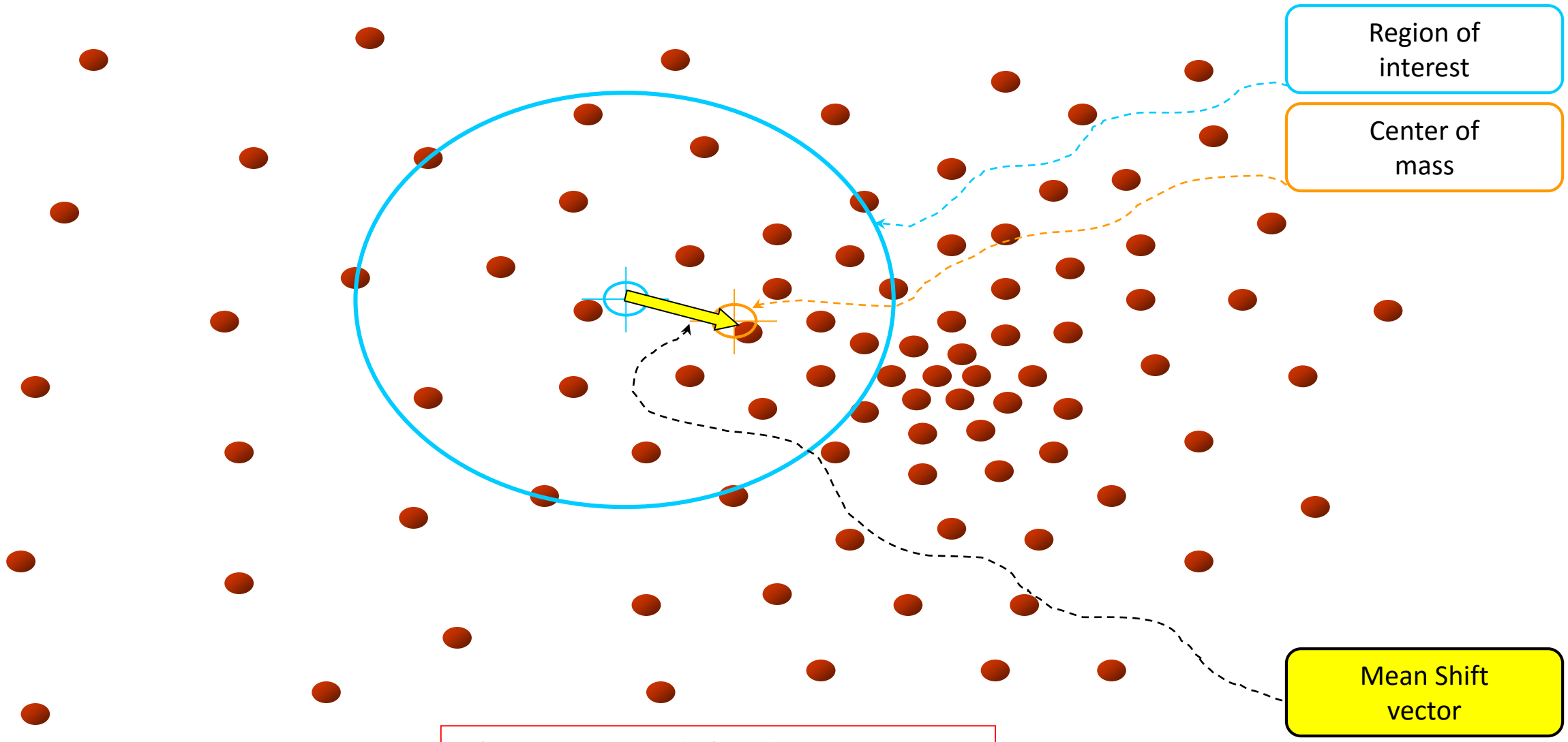
Clustering algorithms

- Parametric approaches rely upon a priori knowledge of the number of clusters and regarding the shape of the clusters.
- In general, they result in simpler models and are cheaper to compute but perform poorly when the underlying data distribution does not match assumptions made.
- Non-parametric techniques make no such assumptions, however they tend to be computationally expensive. Robust performance wise.

Mean Shift

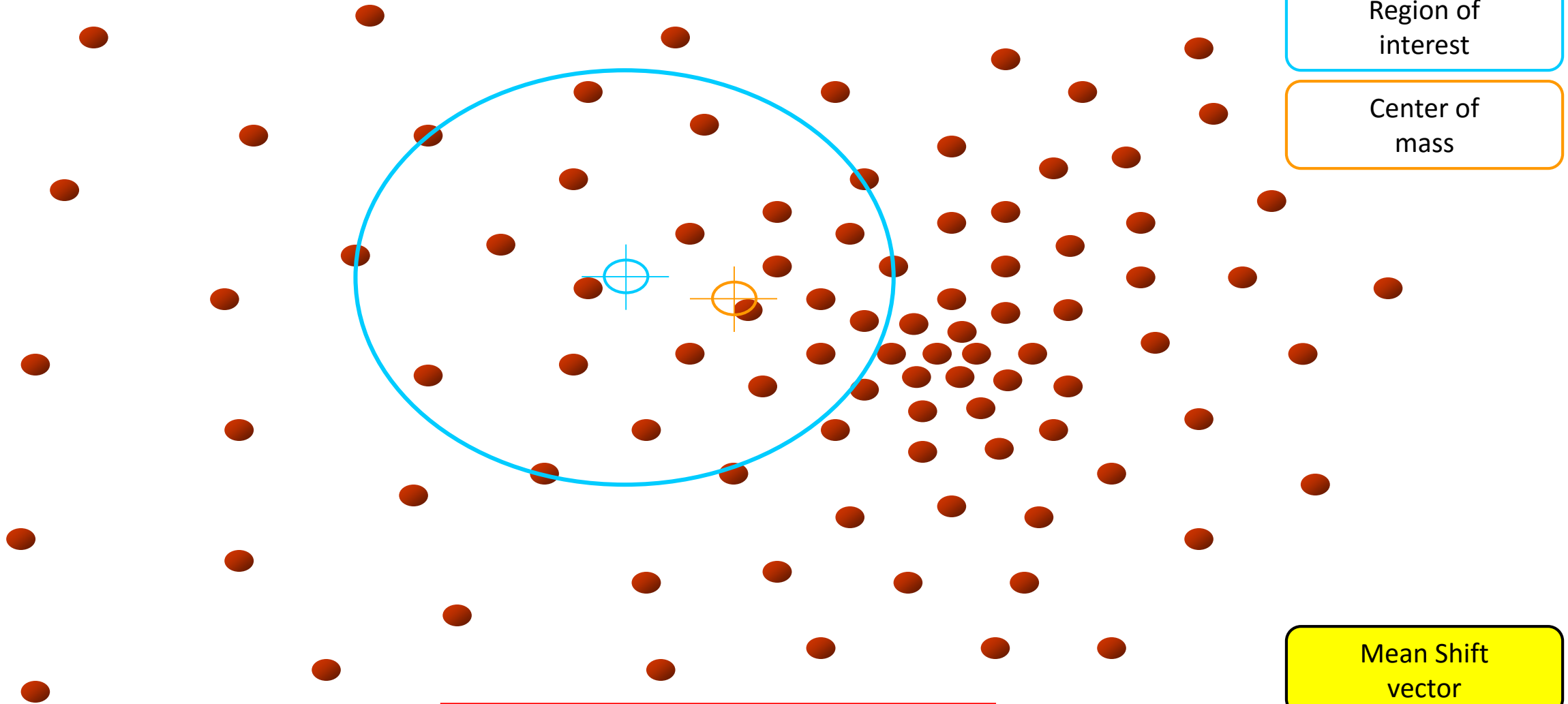
- Belongs to the family of non-parametric density estimation based approaches.
- Given a set of data samples, it tries to find the regions of maximum density i.e. modes of the distribution.
- Iterative.
- Gradient based.

Mean Shift - Intuition



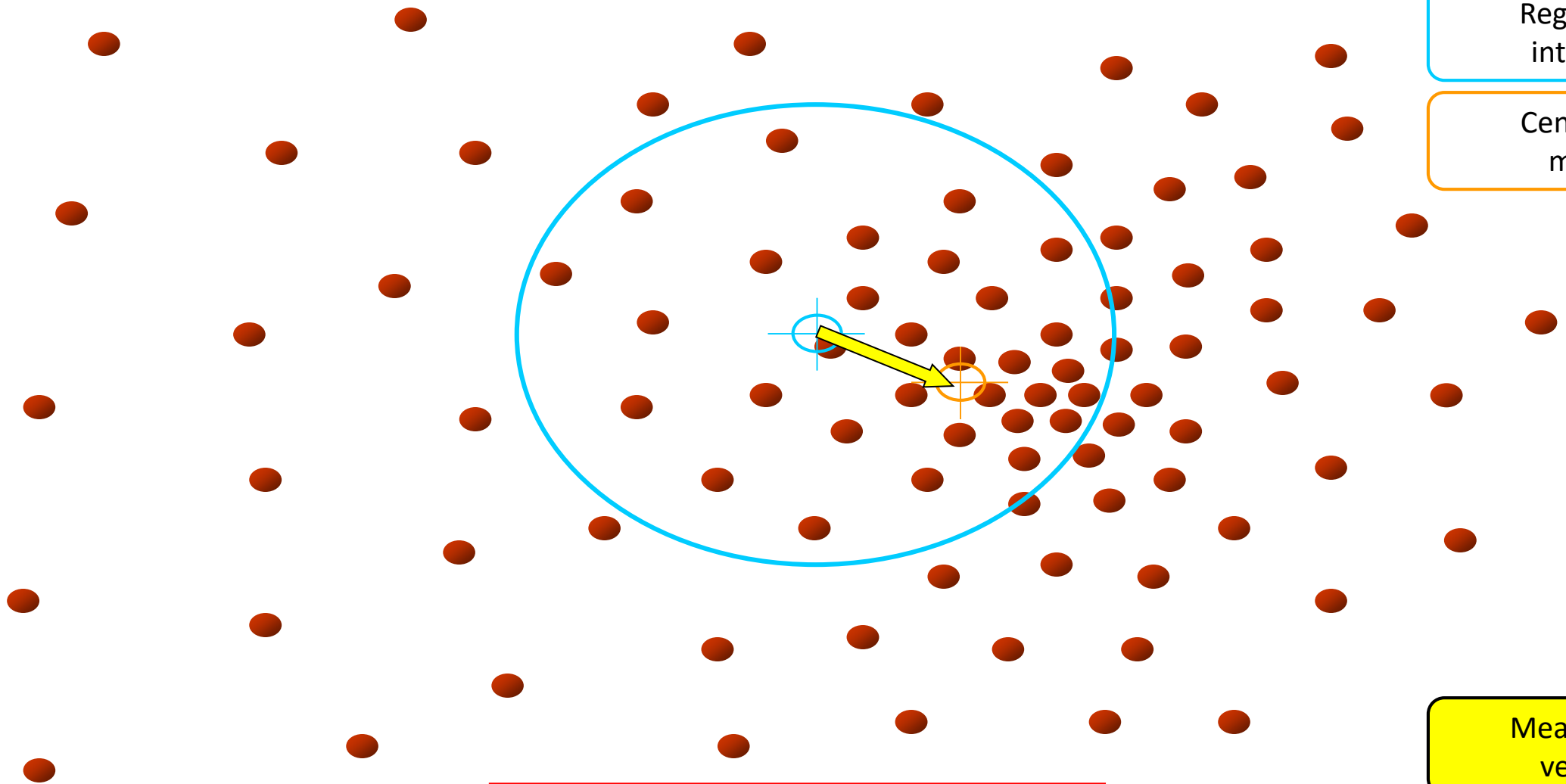
<u>Objective</u> : Find the densest region

Mean Shift - Intuition



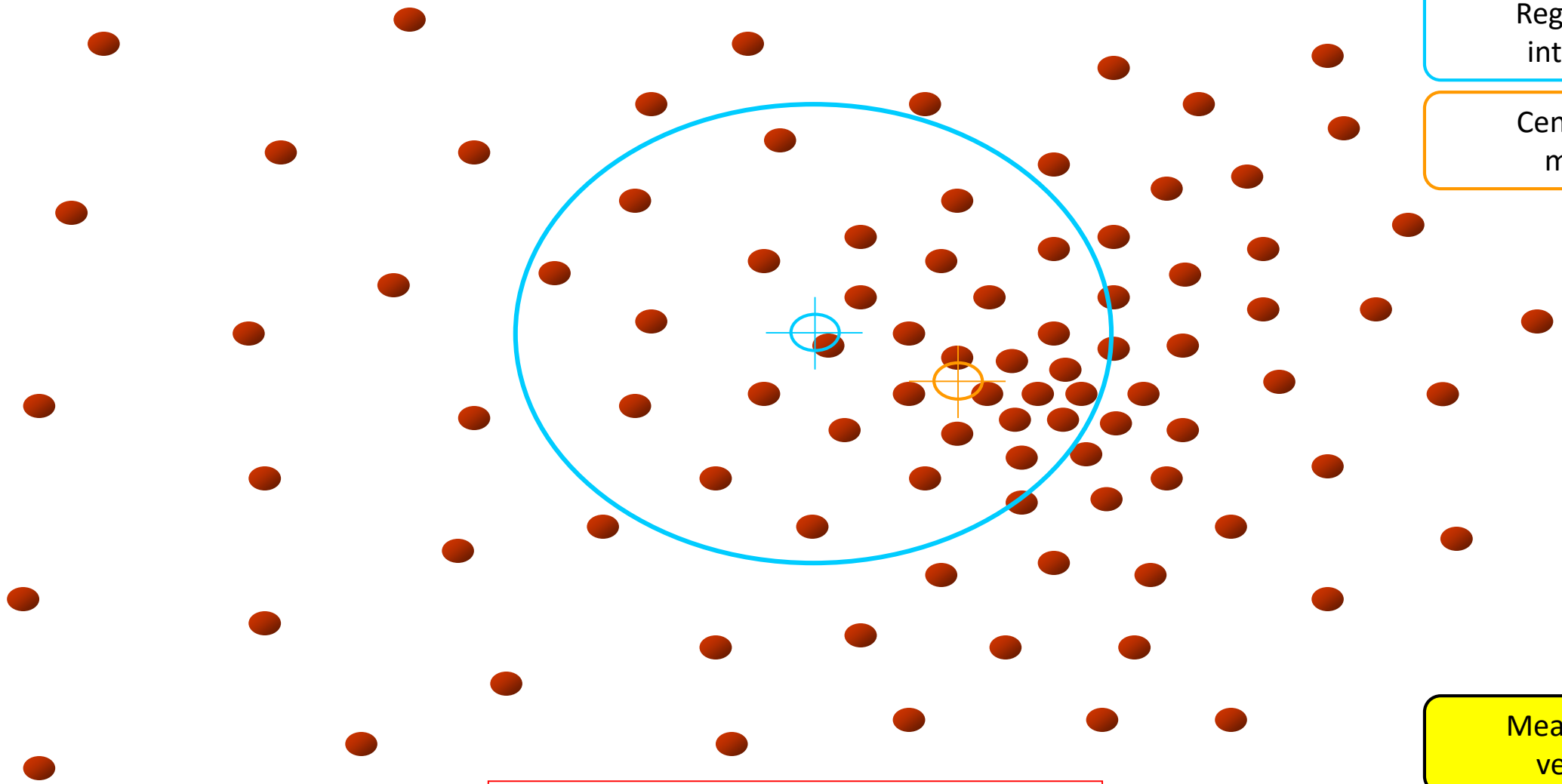
<u>Objective</u> : Find the densest region

Mean Shift - Intuition



<u>Objective</u> : Find the densest region

Mean Shift - Intuition



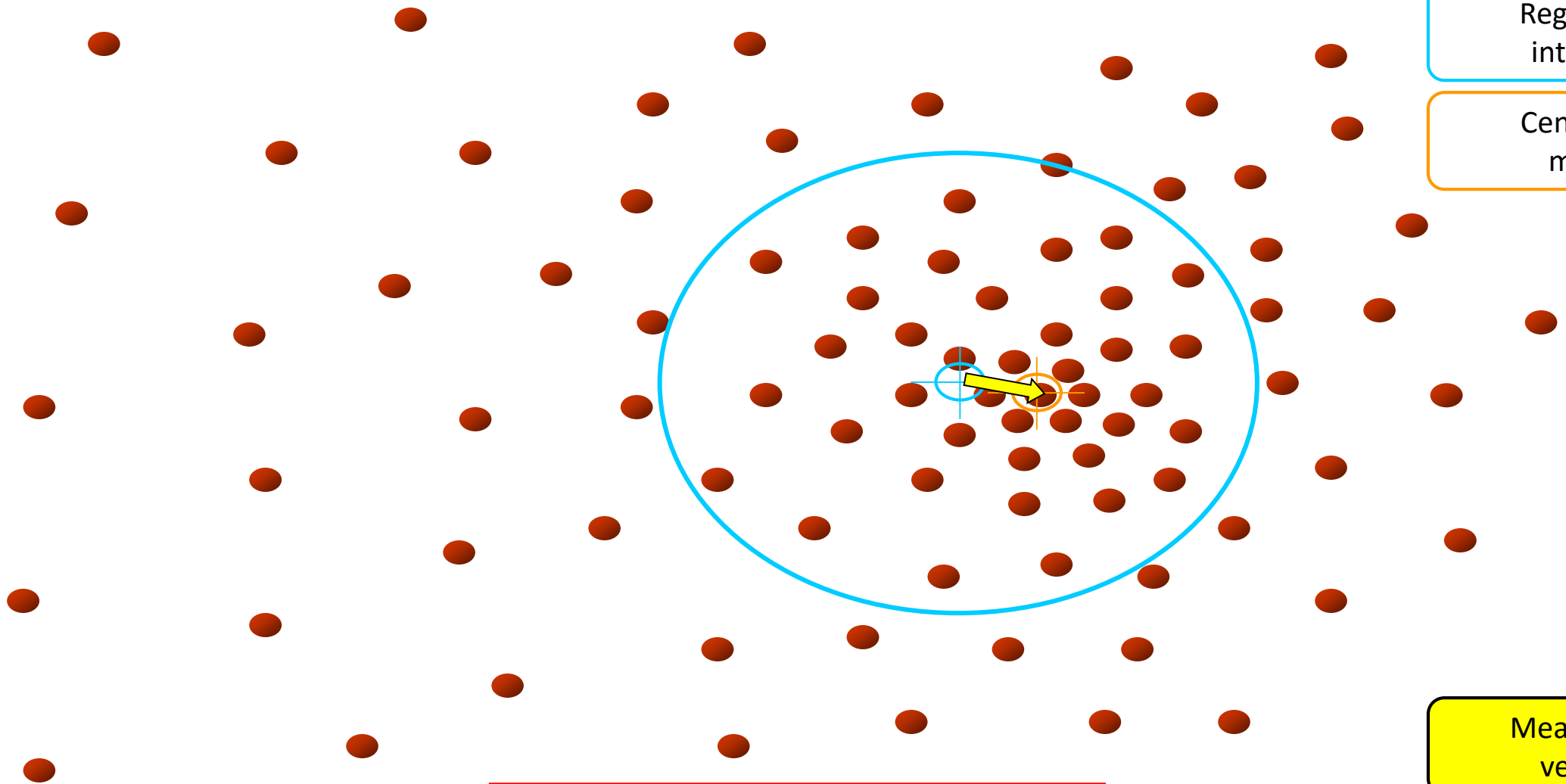
Region of interest

Center of mass

Mean Shift vector

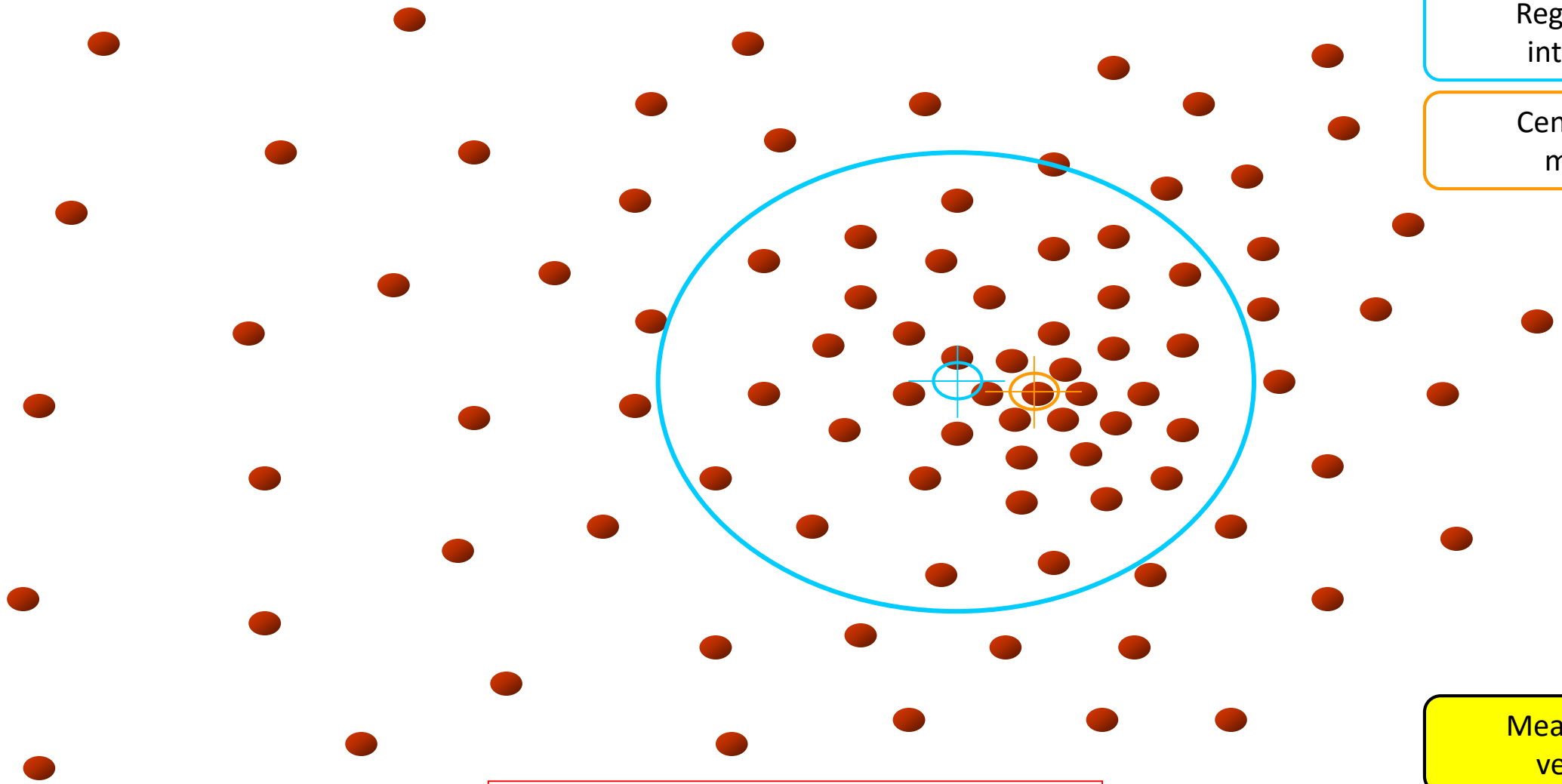
<u>Objective</u> : Find the densest region

Mean Shift - Intuition



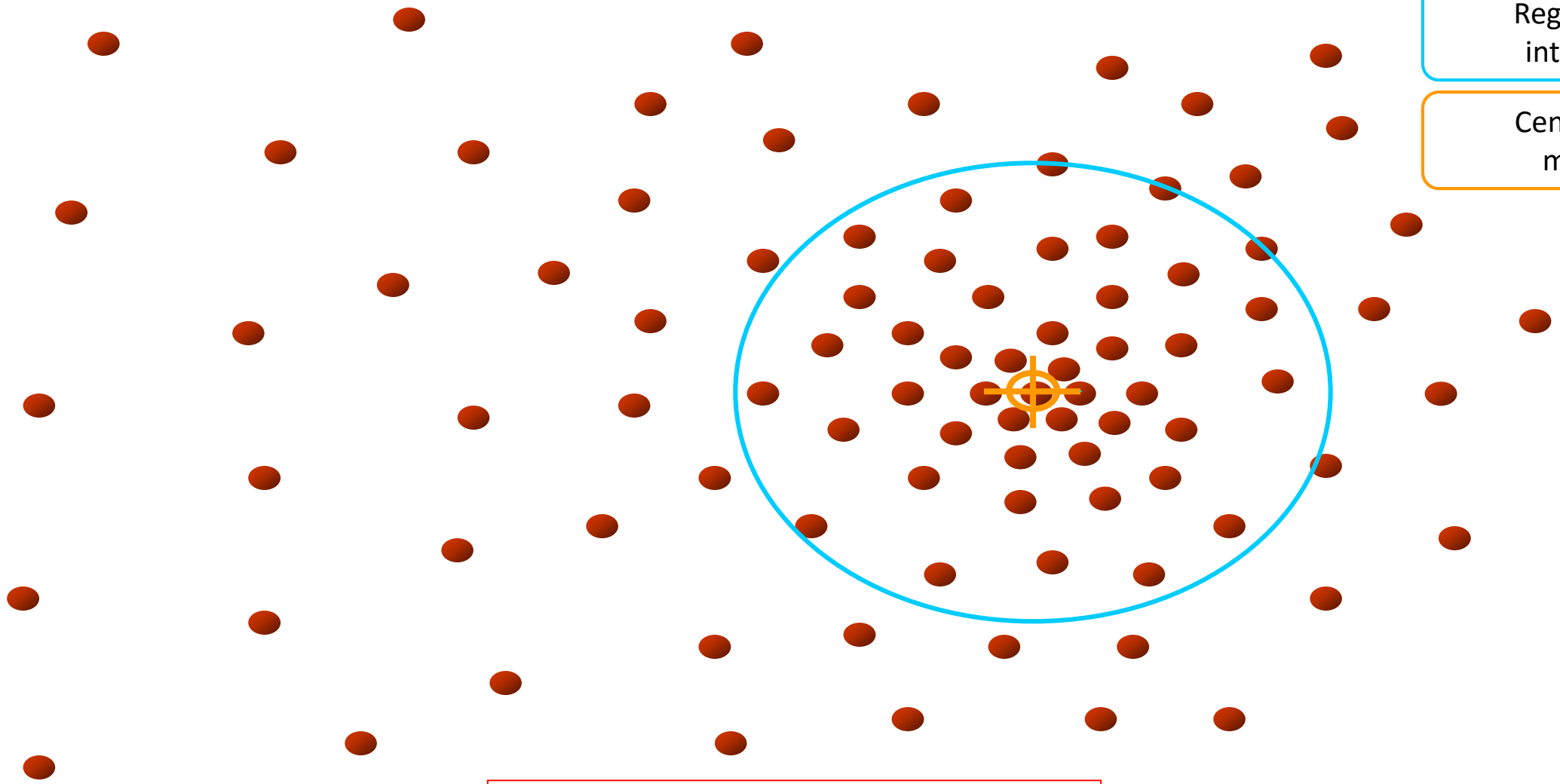
<u>Objective</u> : Find the densest region

Mean Shift - Intuition



<u>Objective</u> : Find the densest region

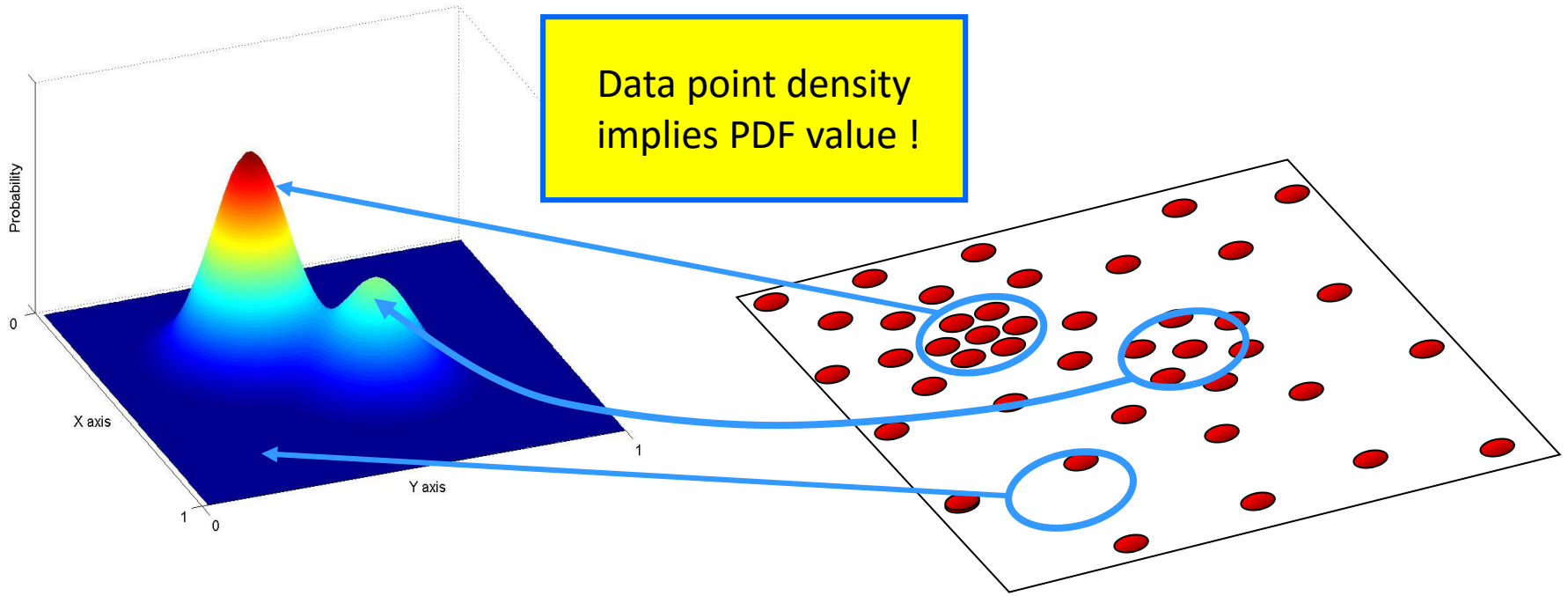
Mean Shift - Intuition



<u>Objective</u> : Find the densest region

Why search for densest regions ?

<u>Assumption</u> : The data points are actually samples from an underlying PDF



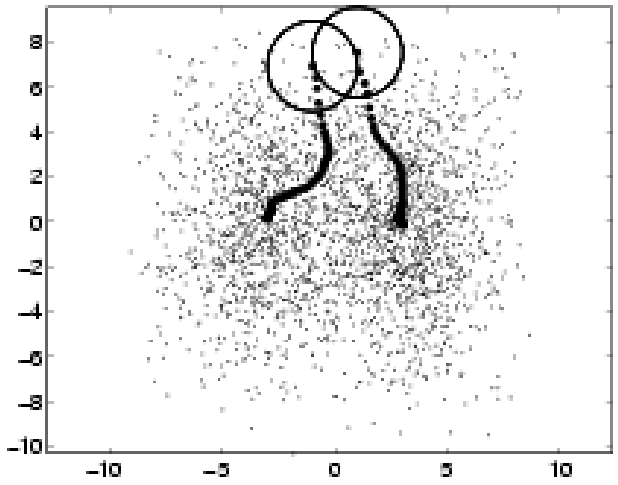
Assumed Underlying PDF

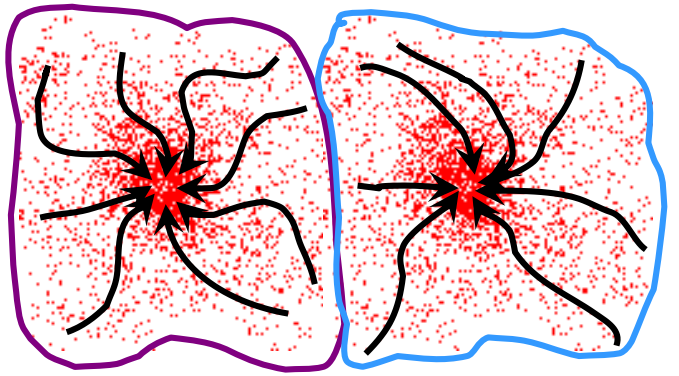
Real Data Samples

Mean Shift in clustering

Cluster : All data points in the ***attraction basin*** of a mode

Attraction basin : the region for which all trajectories lead to the same mode





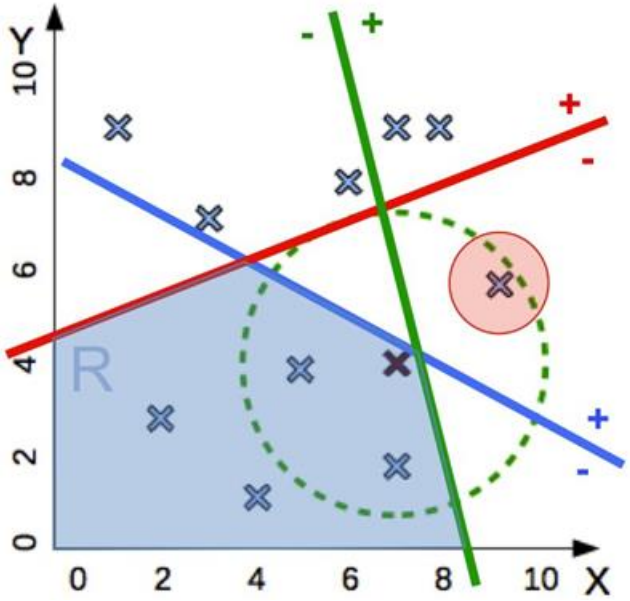
Tessellations of the feature space, containing the basins of attraction, which are the regions for which all trajectories lead to the same mode.

Locality Sensitive Hashing

- Used for computing fast, efficient Nearest Neighbor searches.
- Belongs to the class of randomized algorithms, which do not guarantee an exact answer but return the correct answer or one close to it with a high probability guarantee.
- Random hyper-planes h_1 ... h_K
  - Feature space sliced into 2^K partitions.
  - Compare query point with only $\mathbb{E}(N/2^K)$ points.

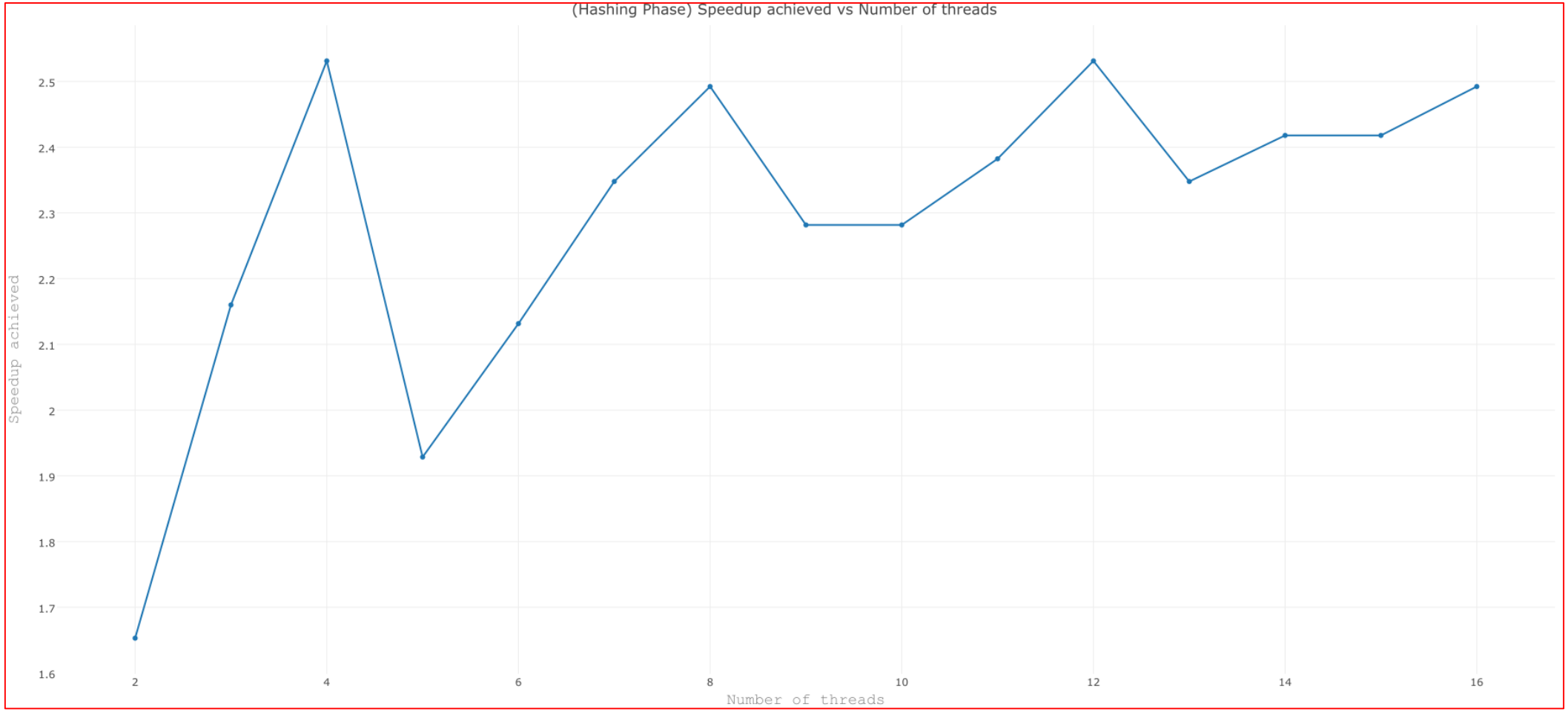
Locality Sensitive Hashing

- Inexact: missed neighbors
  - Repeat with $\mathcal{L}$ sets of $h_1 \ldots h_{\mathcal{K}}$

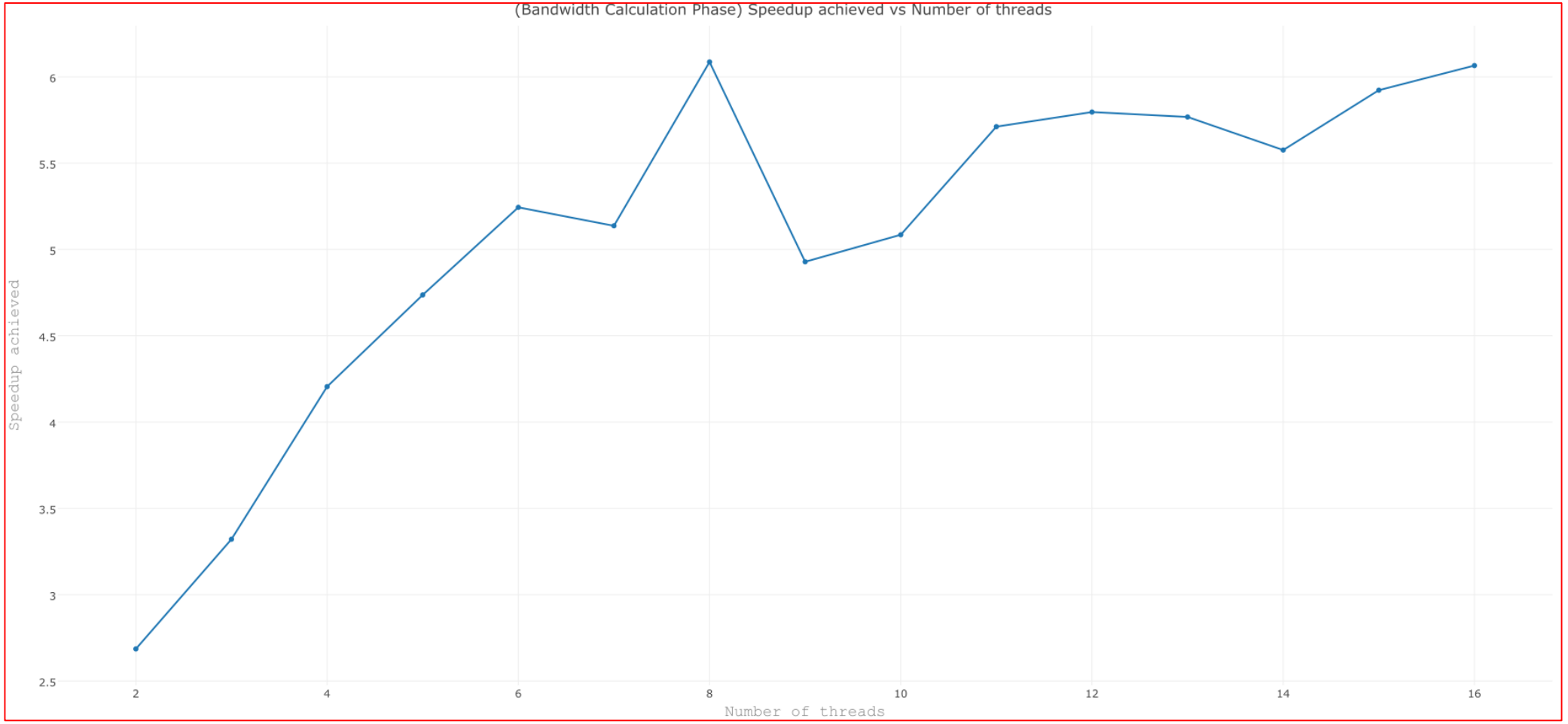


- Intuitively map $\mathbb{R}^D \Longrightarrow \mathbb{R}^{\mathcal{K}}$
- Higher value of $\mathcal{K}$ means a more faithful representation.
- $\mathcal{L}$ introduces redundancy to cover for inexactness.

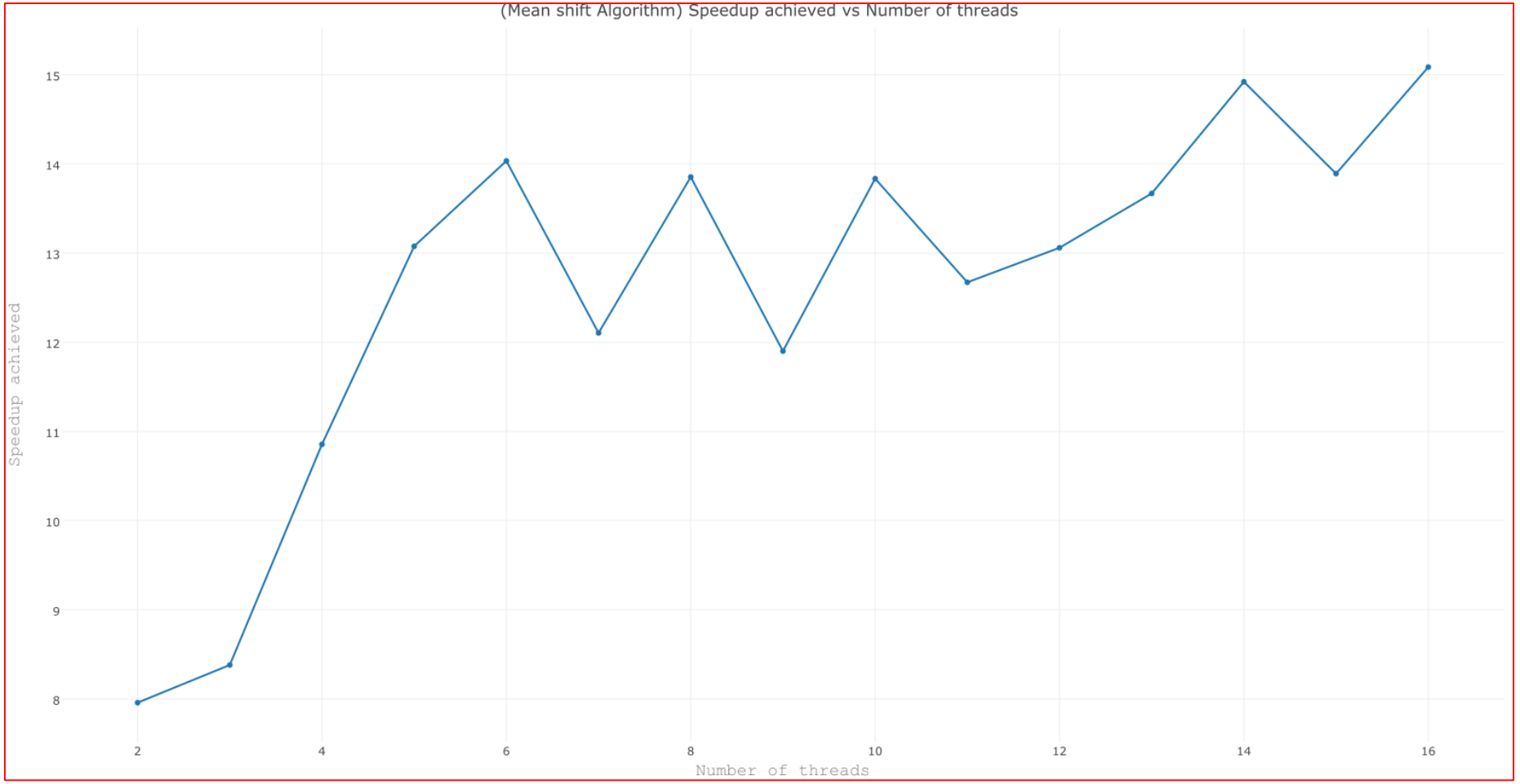
Results – Speed up



Results – Speed up



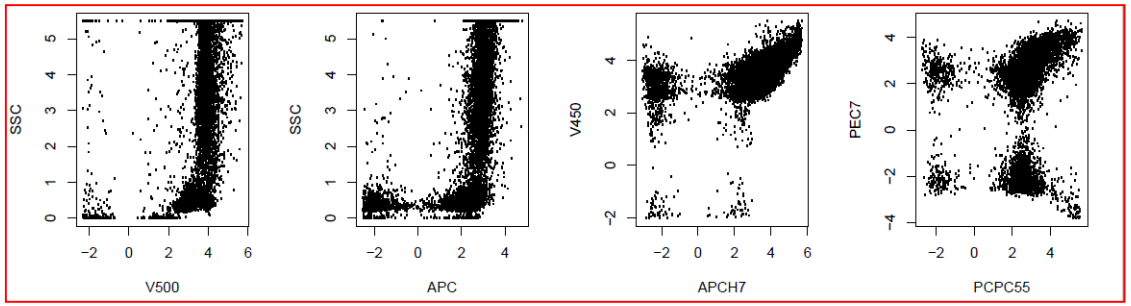
Results – Speed up

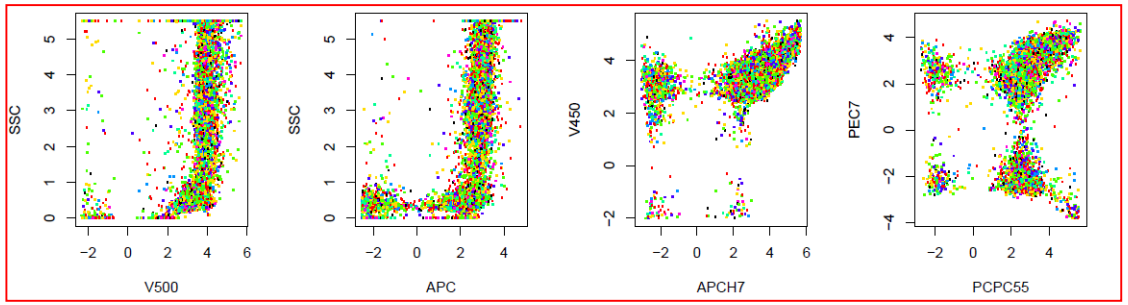


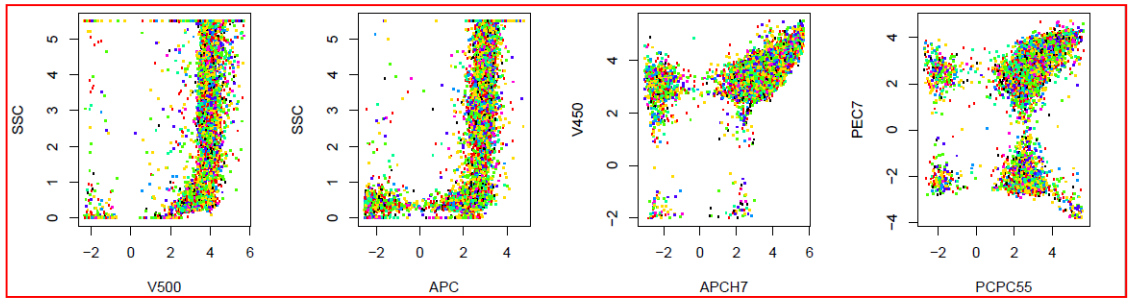
A “Not-so-novel” approach

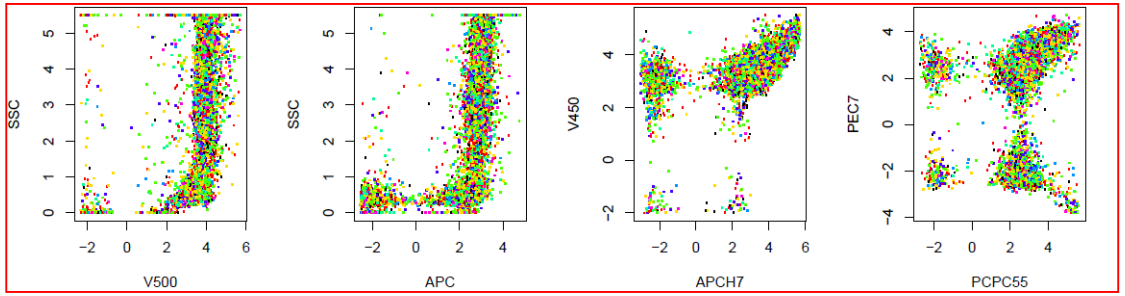
- Experimented with running Mean Shift on only a random fraction of the data points and compared quality/speed.
- Large datasets which took forever to execute previously could now be analyzed.
- Observed quality remained more or less the same.
- With the pure speedup achieved, this significantly helped matters.

Results – Clustering quality



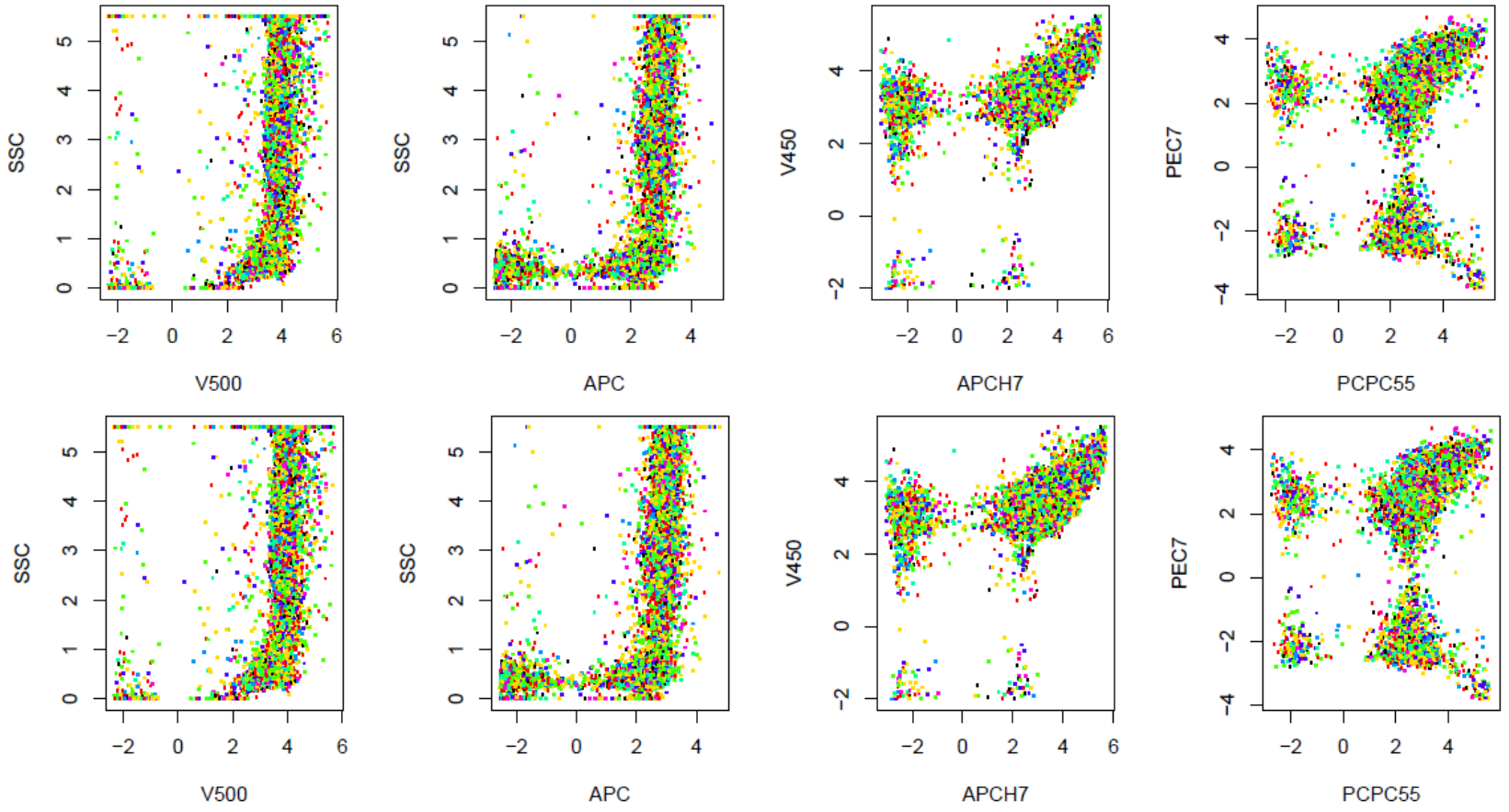






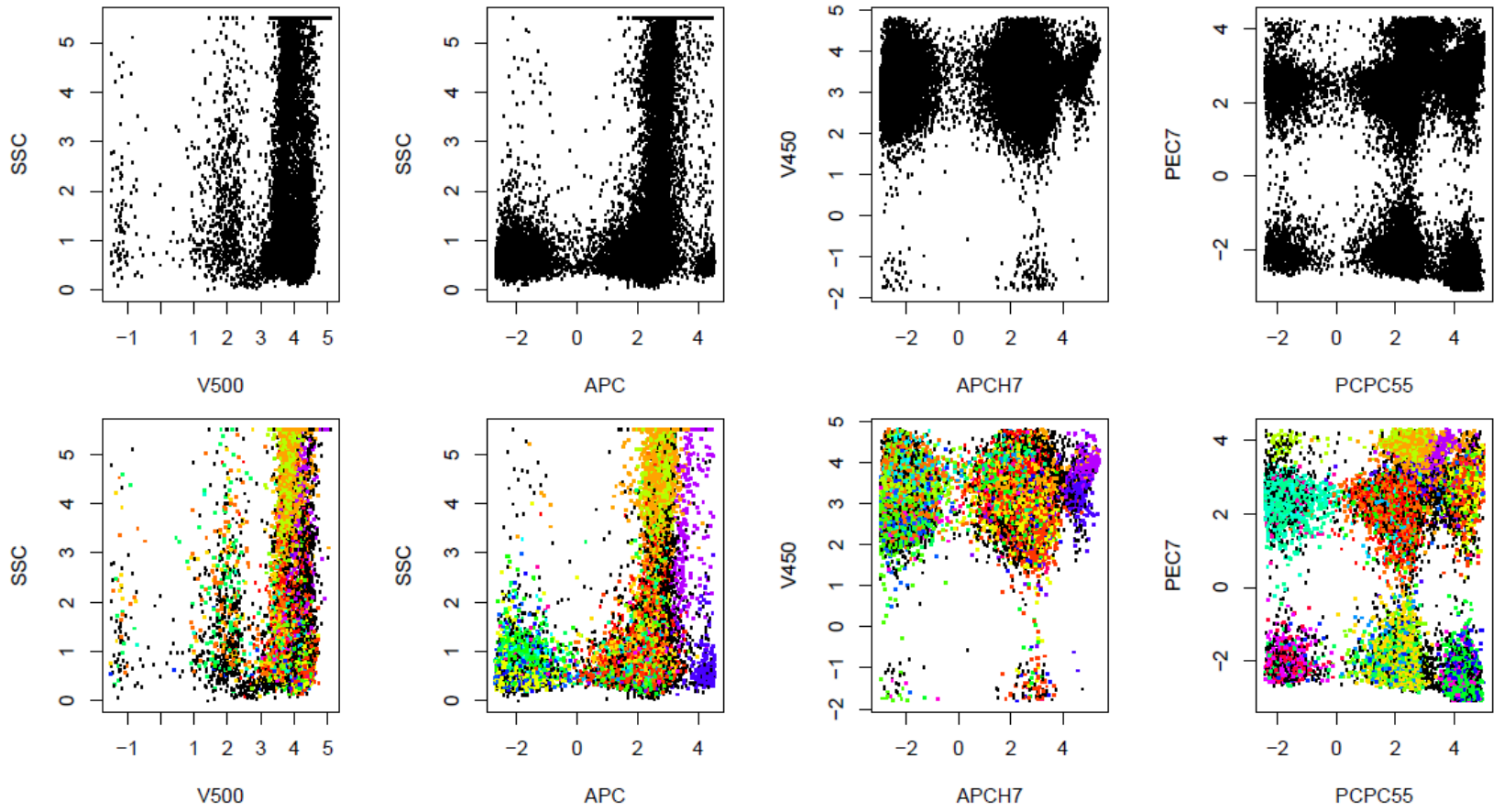
Top left: Original data. Bottom left: Using 100% of data.
Top right: Using 50% of data. Bottom right: Using 10% of data.

Results (Clustering quality)



Top: FAMS. Bottom: Parallel FAMS

Results - Clustering quality



Top: Original dataset. Bottom: Parallel FAMS

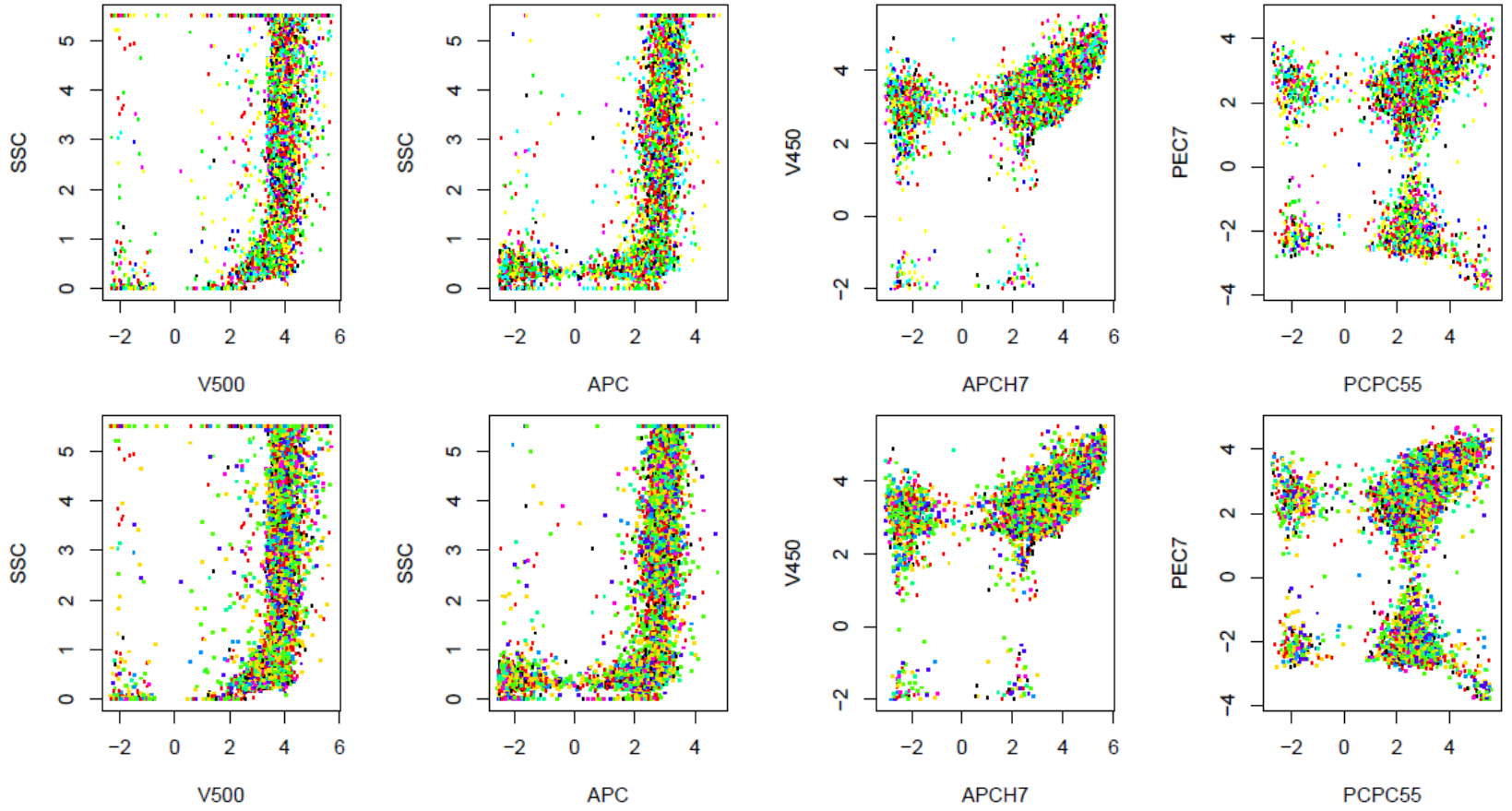
Remarks

- Parallelization of algorithms is in general difficult but “doable”.
- Most approaches either failed or rejected for better ones.
- Multiple strategies to achieve goal i.e. shared memory, critical sections, local buffers.
- Local buffers worked best in this scenario.
- Practical experience and experimentation helps.
- Understanding the problem at an intuitive level allows you to find avenues to take advantage of.

Conclusion

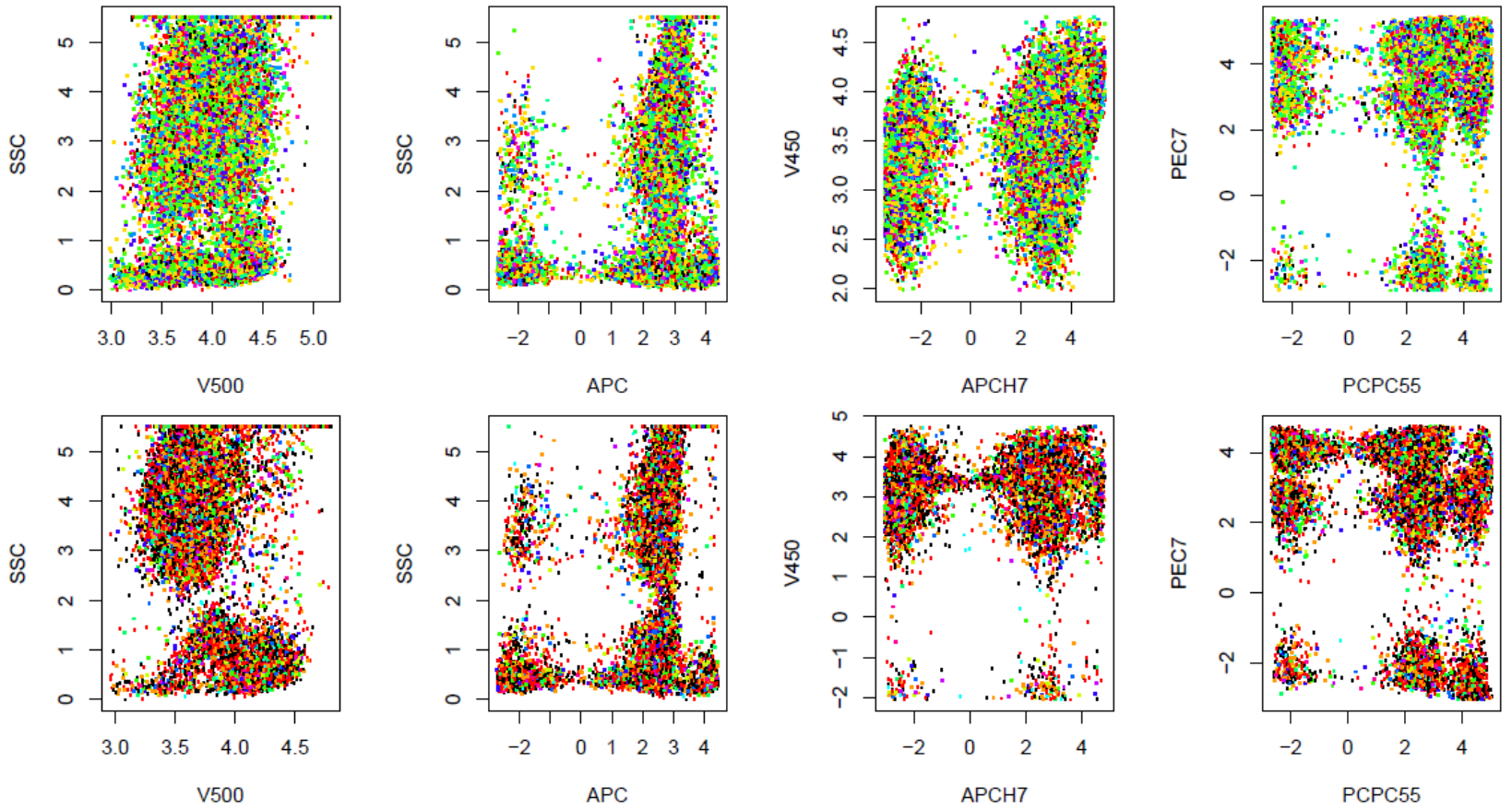
- Possible to use machine learning as well as parallelization in health sciences domain.
- However some preprocessing might be needed.
- Amount of data available is plenty as well as in different views.
- Plenty of opportunities to leverage on this.
- In presence of noise, simple i.e. linear models should be preferred not to overfit, however many of the problems are non-linear in nature.

Future work – Different mode sizes



Top: Using max modes = 100.
Bottom: Using max modes = 1000
Can gain insight into subpopulations ? Do these sub-clusters have significance ?

Future work – Signatures ?



Top: Acute cancer individual. Bottom: Healthy individual
Can we use modes or their collection as a signature to determine test cases?

Thank you !!