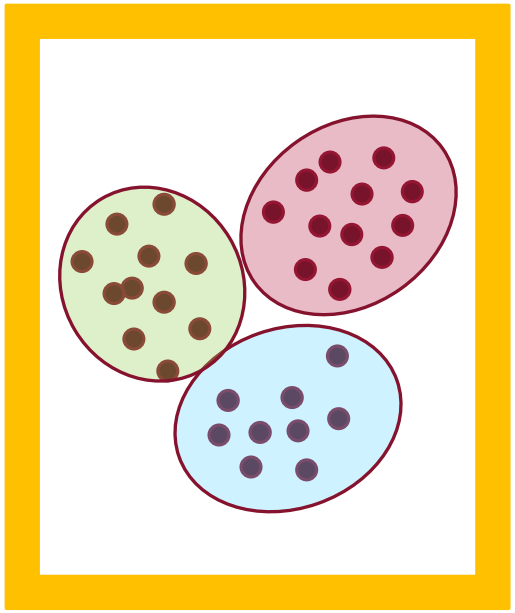
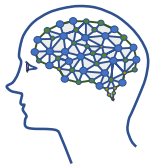


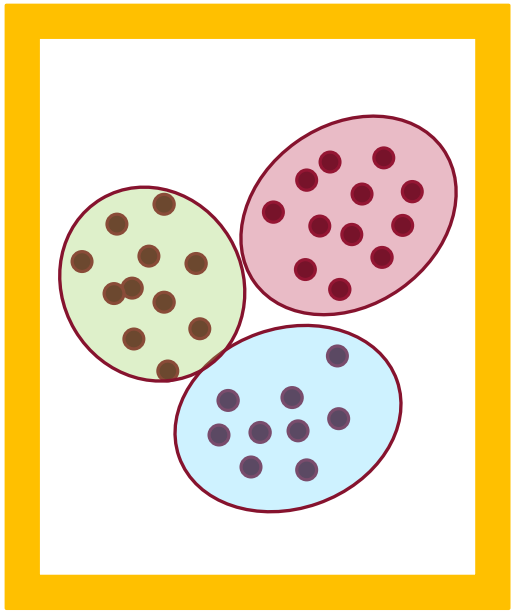
Unsupervised Learning and Visualization

Looking for Patterns in Unlabelled Data

Plan for the Day



- Data Clustering
  - K-Means
  - Hierarchical Clustering
- Data Visualization
  - Linear Methods: PCA
  - Non-linear Methods: MDS, ISOMAP, LLE
  - Non-linear Method: SNE, t-SNE
- Plans for Next Week

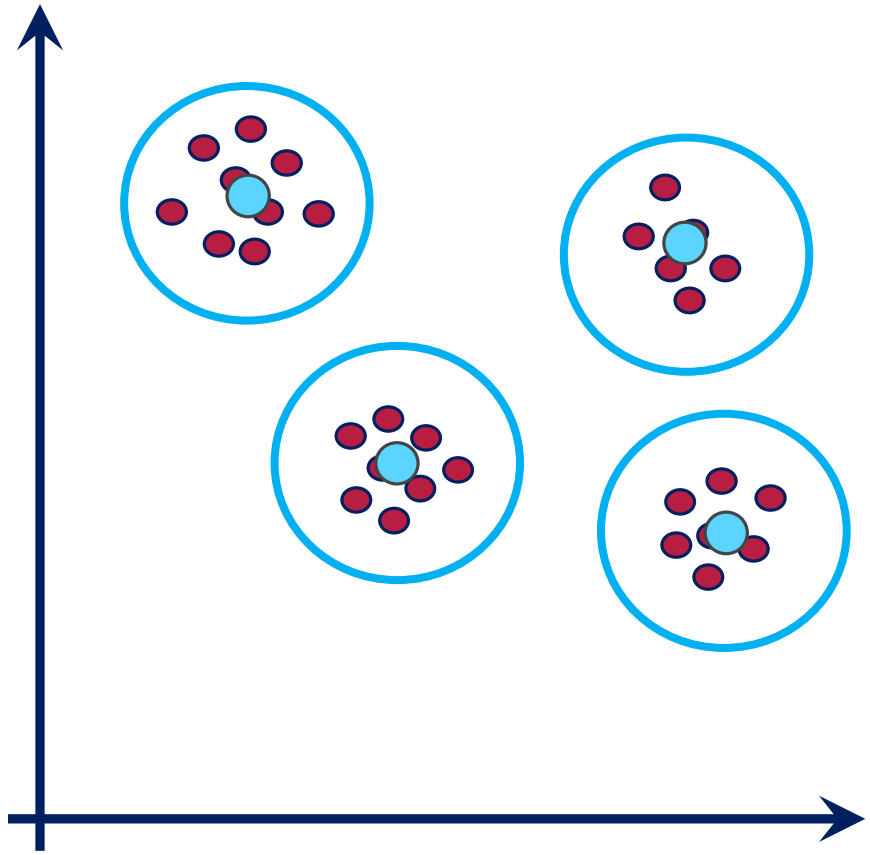


Data Clustering

Discovering Patterns in Data

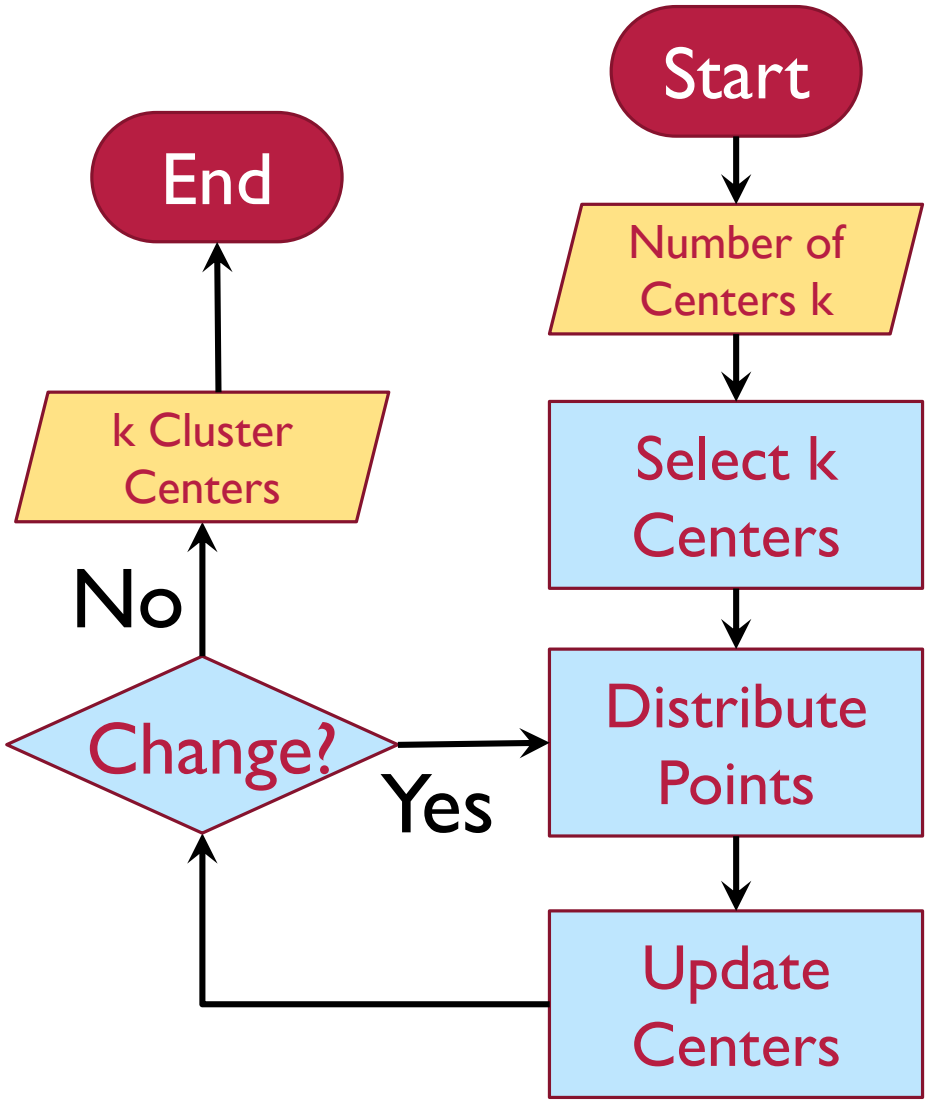
K-Means

- You are given N points
- How do we find k clusters?
  - What if we know the cluster centers?
- How do we find the cluster centers?
  - What if we know the k clusters?

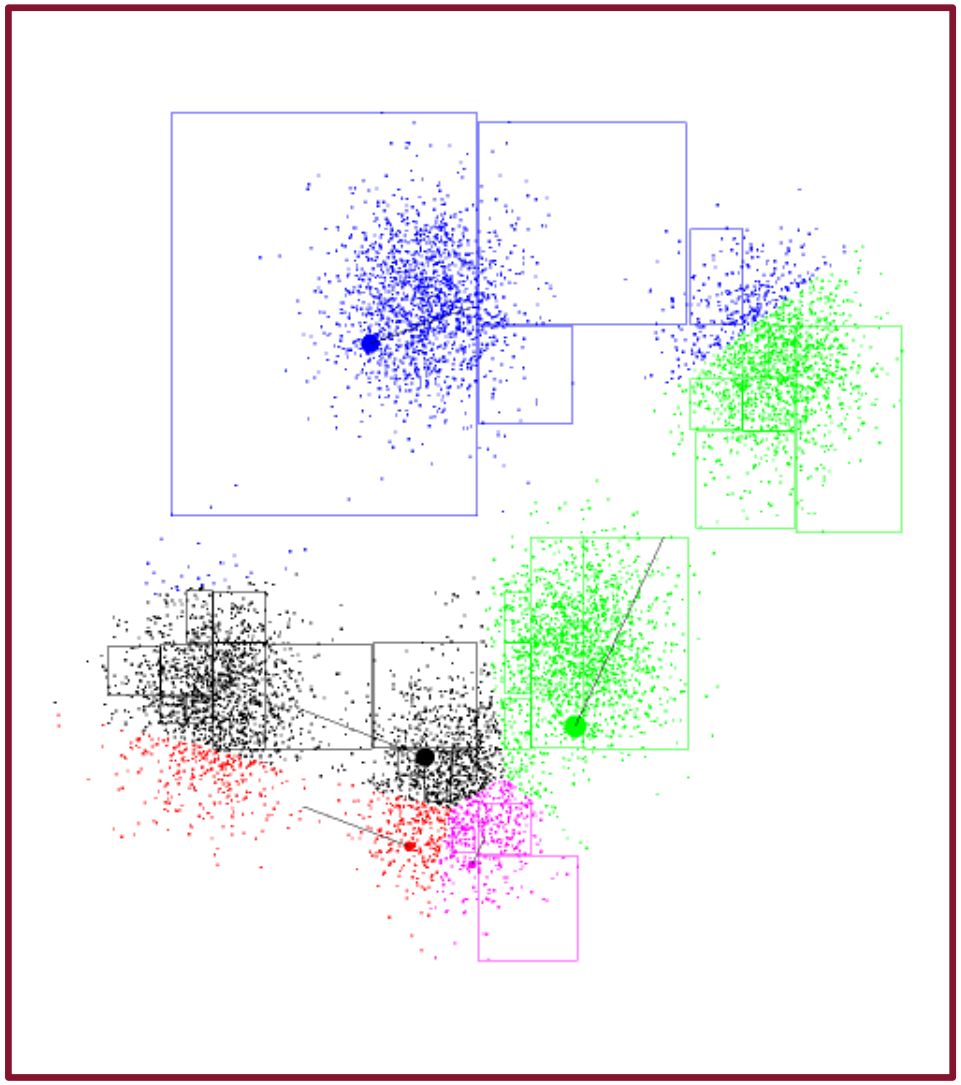


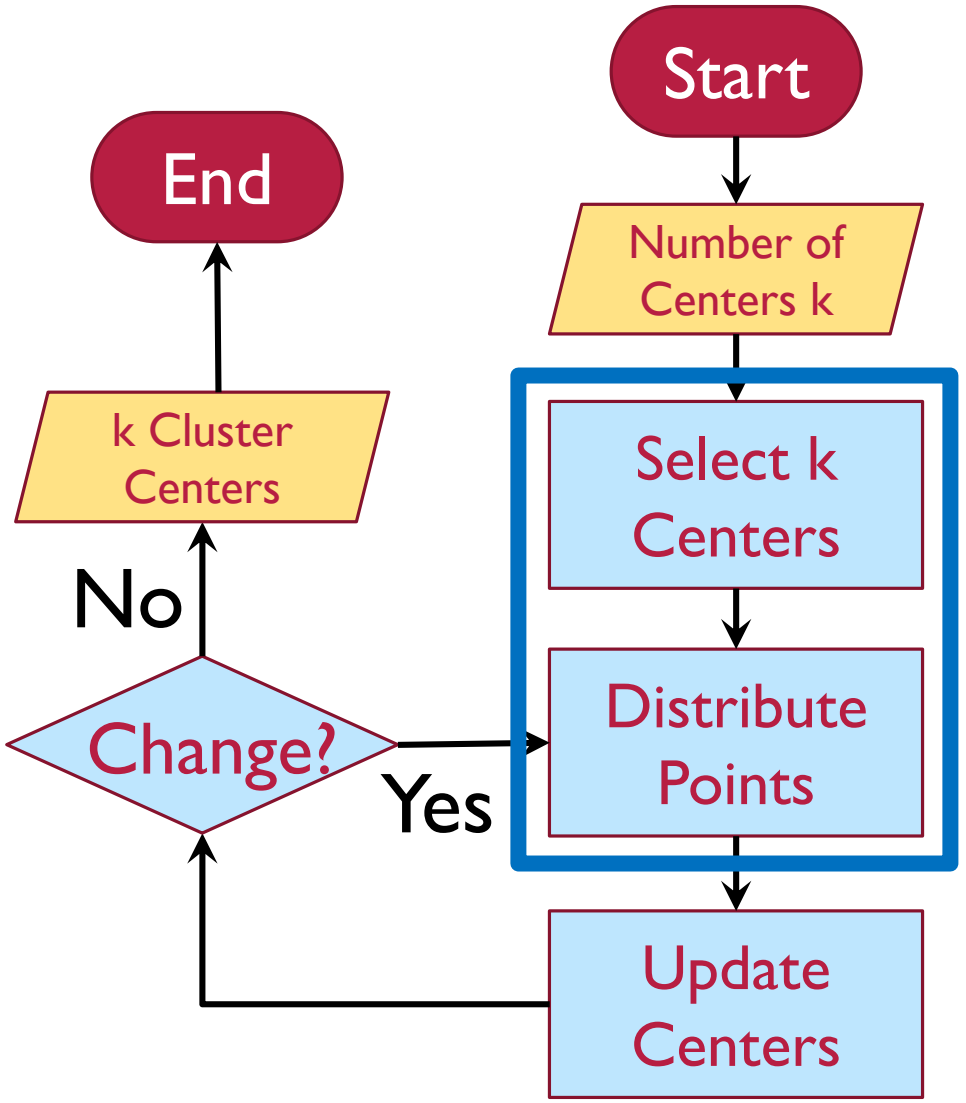
K-Means

1. Input: k (number of clusters)
2. Randomly select k centers
   3. Distribute Points
   4. Update Centers
   5. Repeat 3,4 till convergence
6. Output: Cluster centers

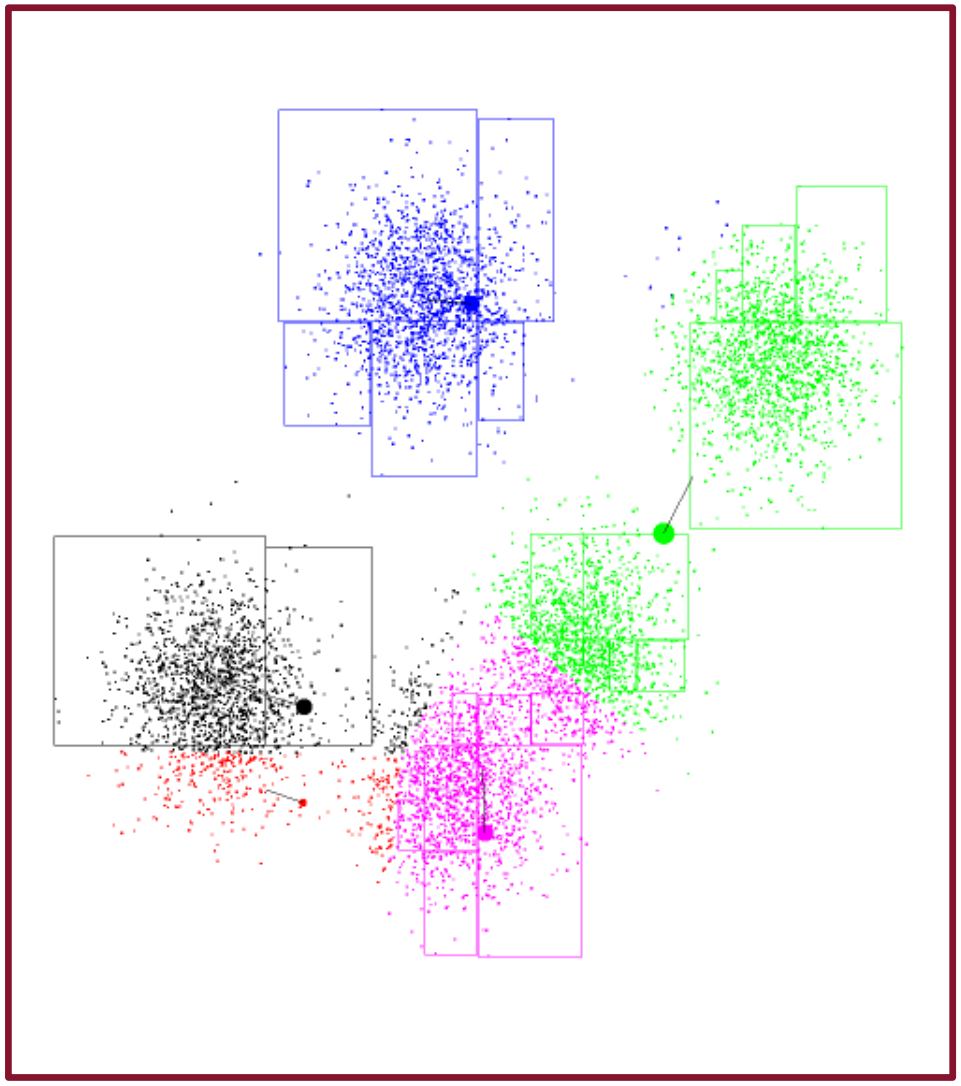


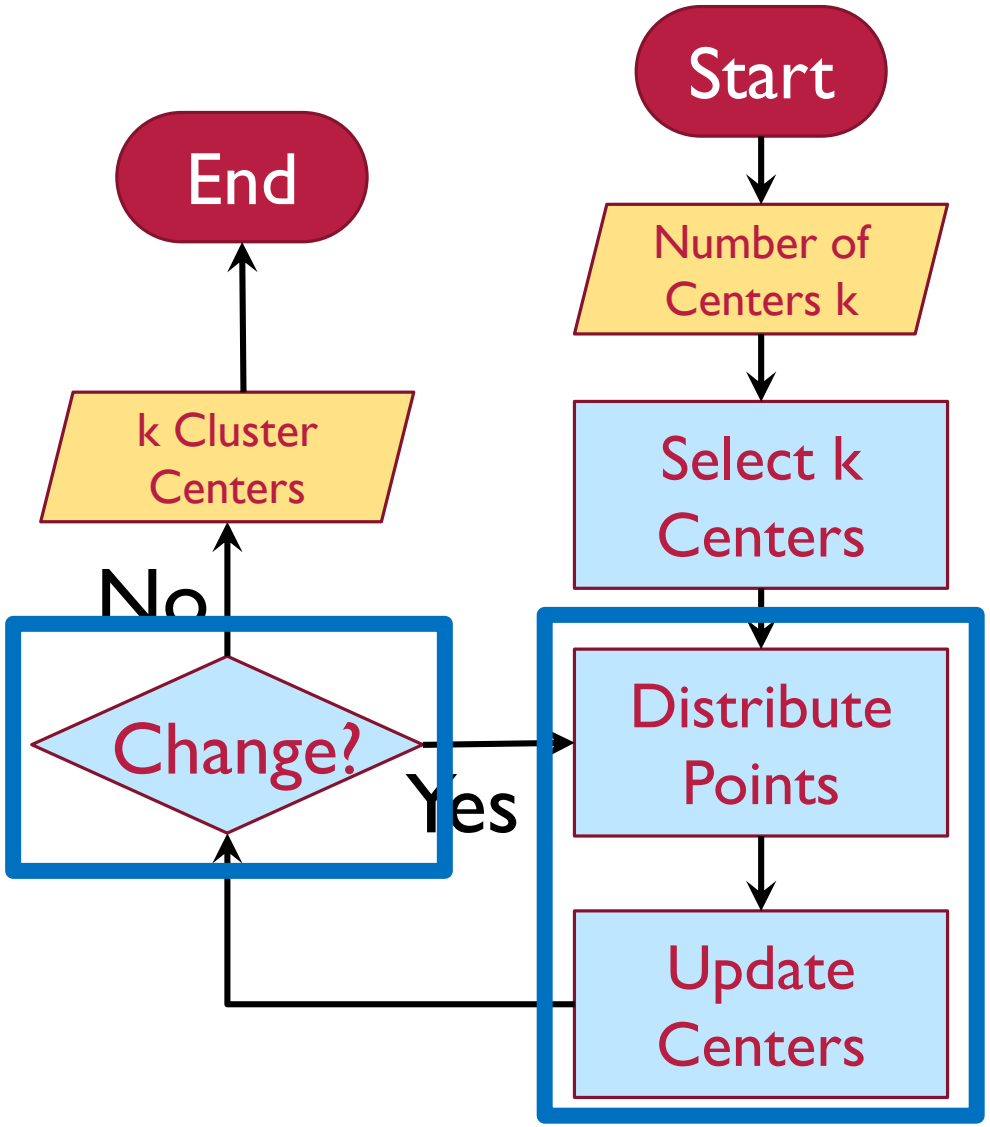
K-Means



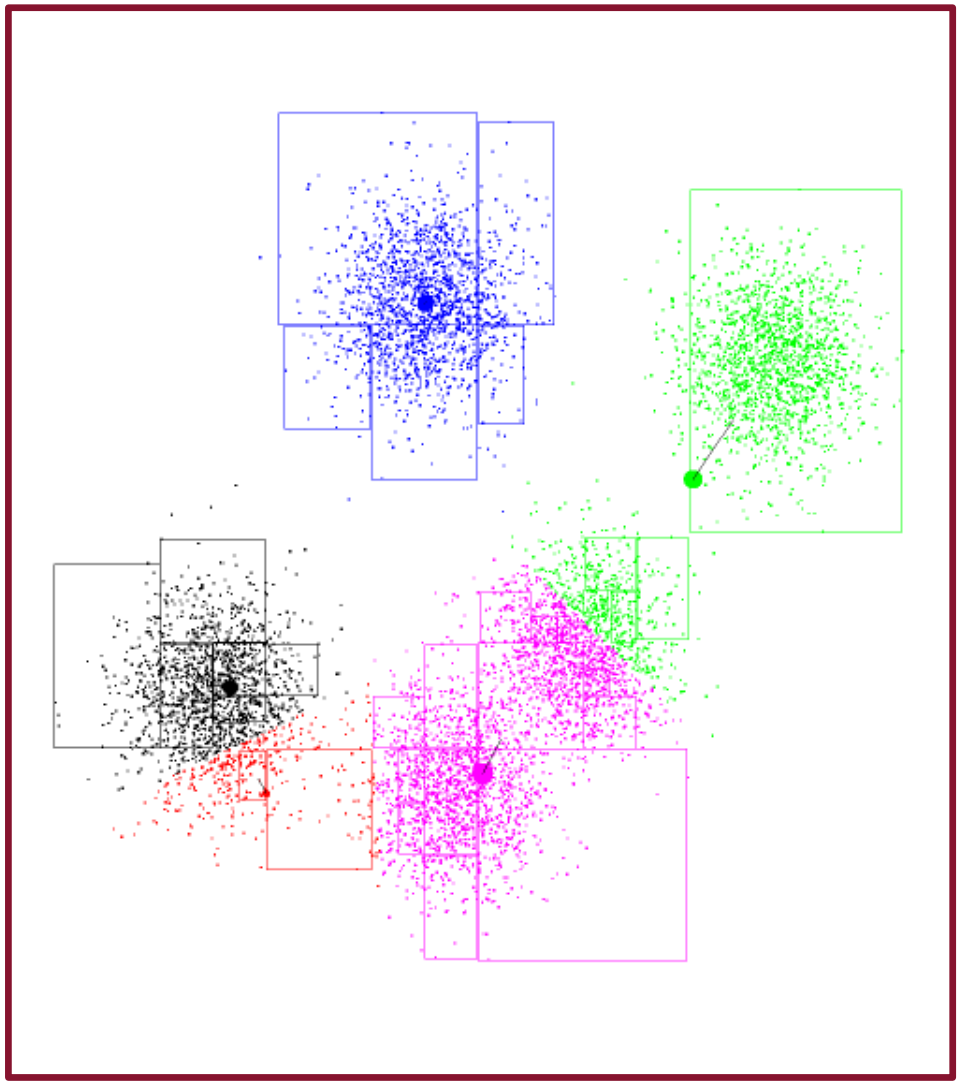

Start
Number of Centers k
Select k Centers
Distribute Points
Update Centers
Change?
Yes
No
k Cluster Centers
End

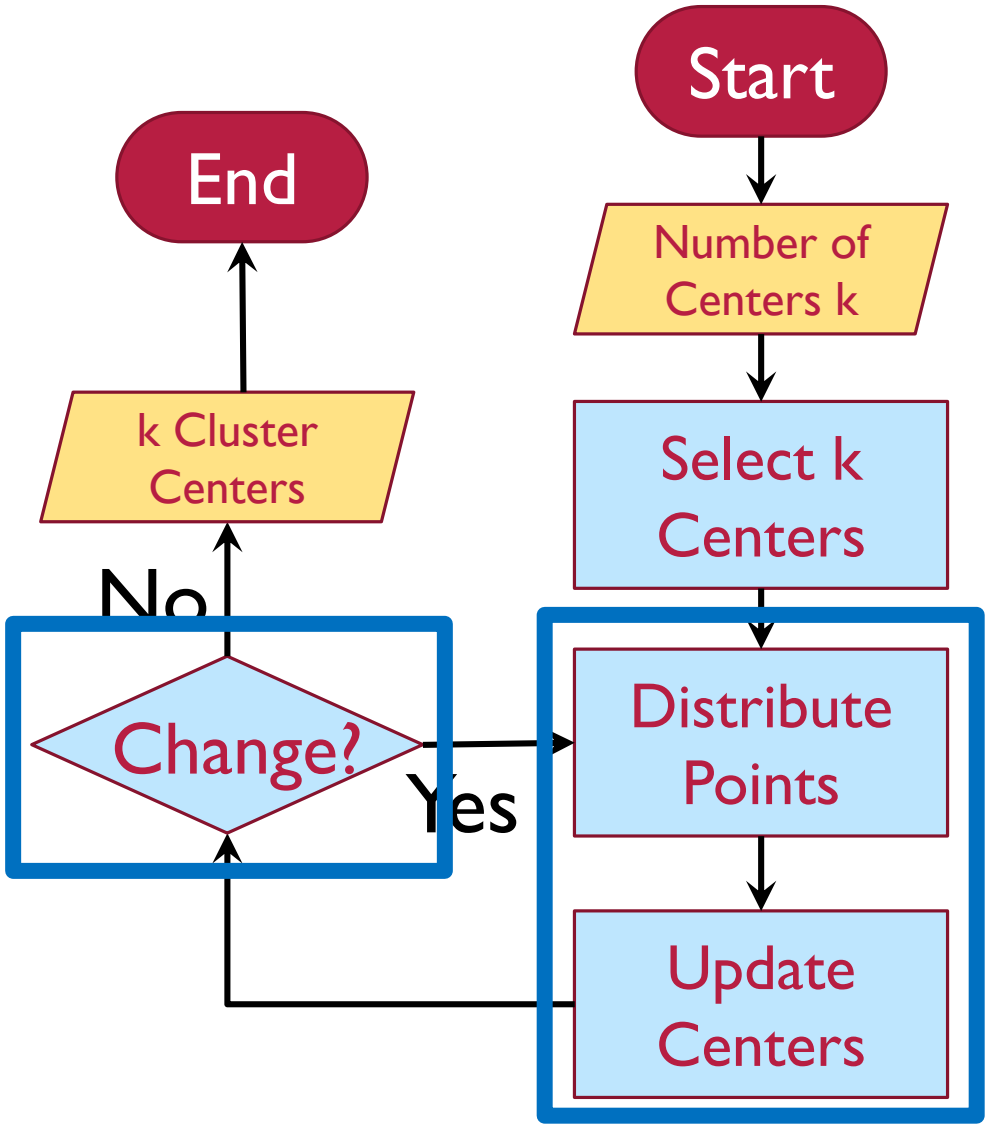
K-Means: Update 1



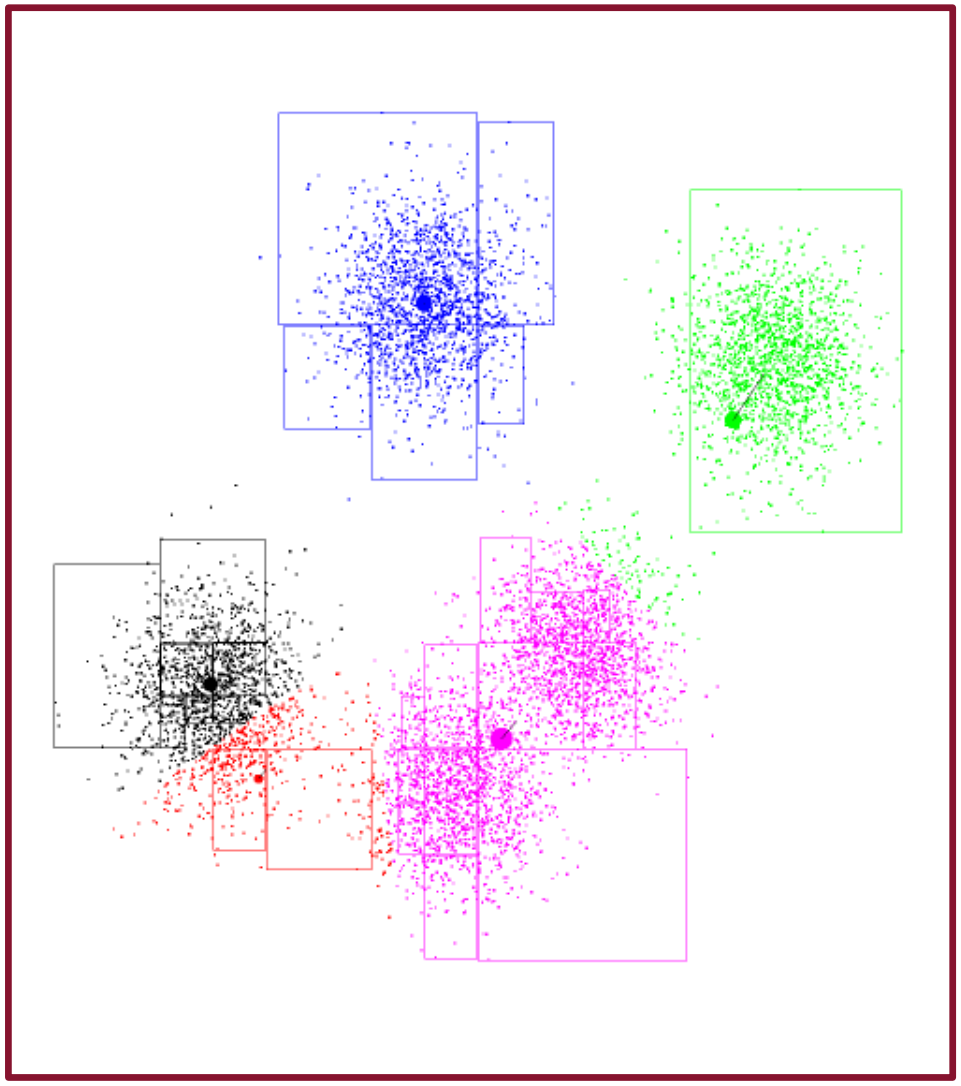


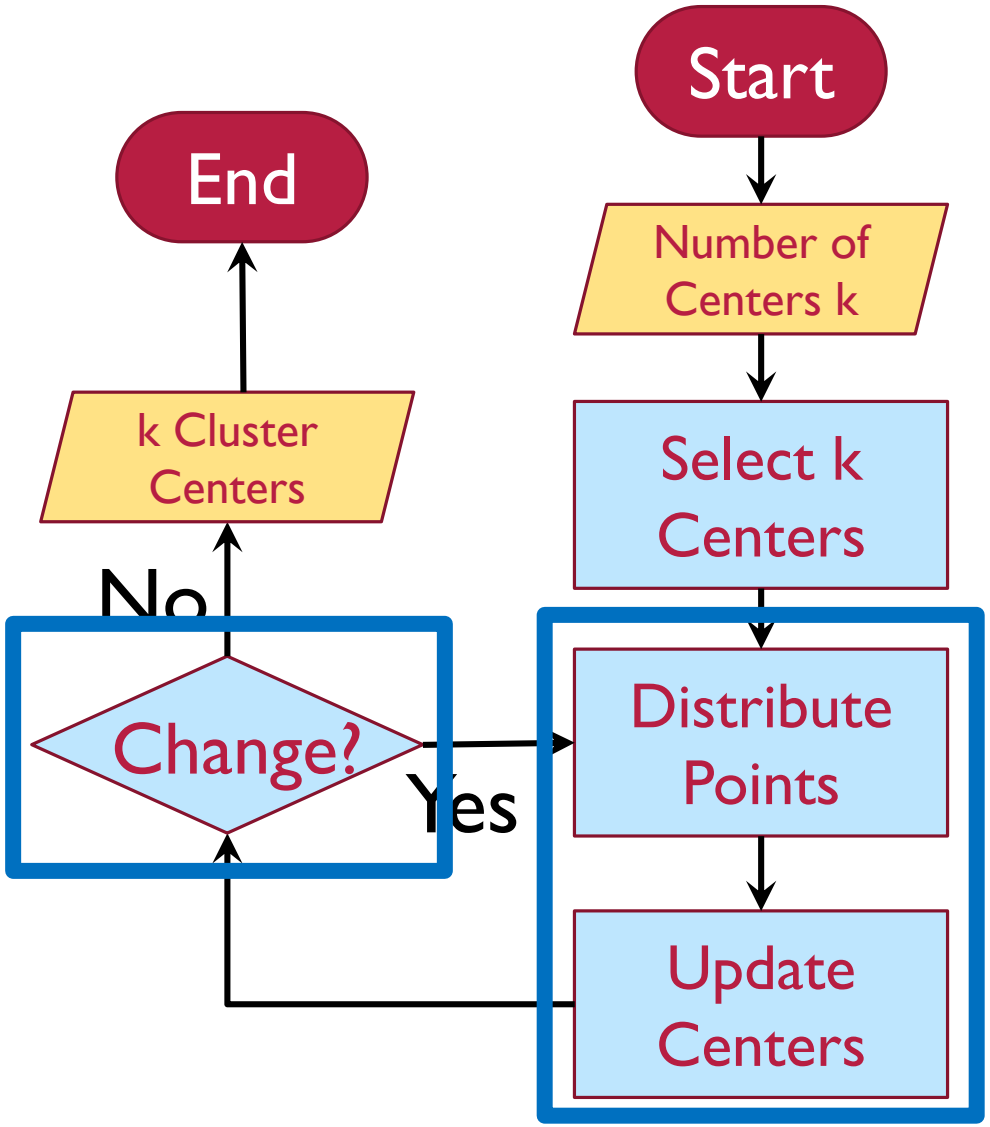
K-Means: Update 2



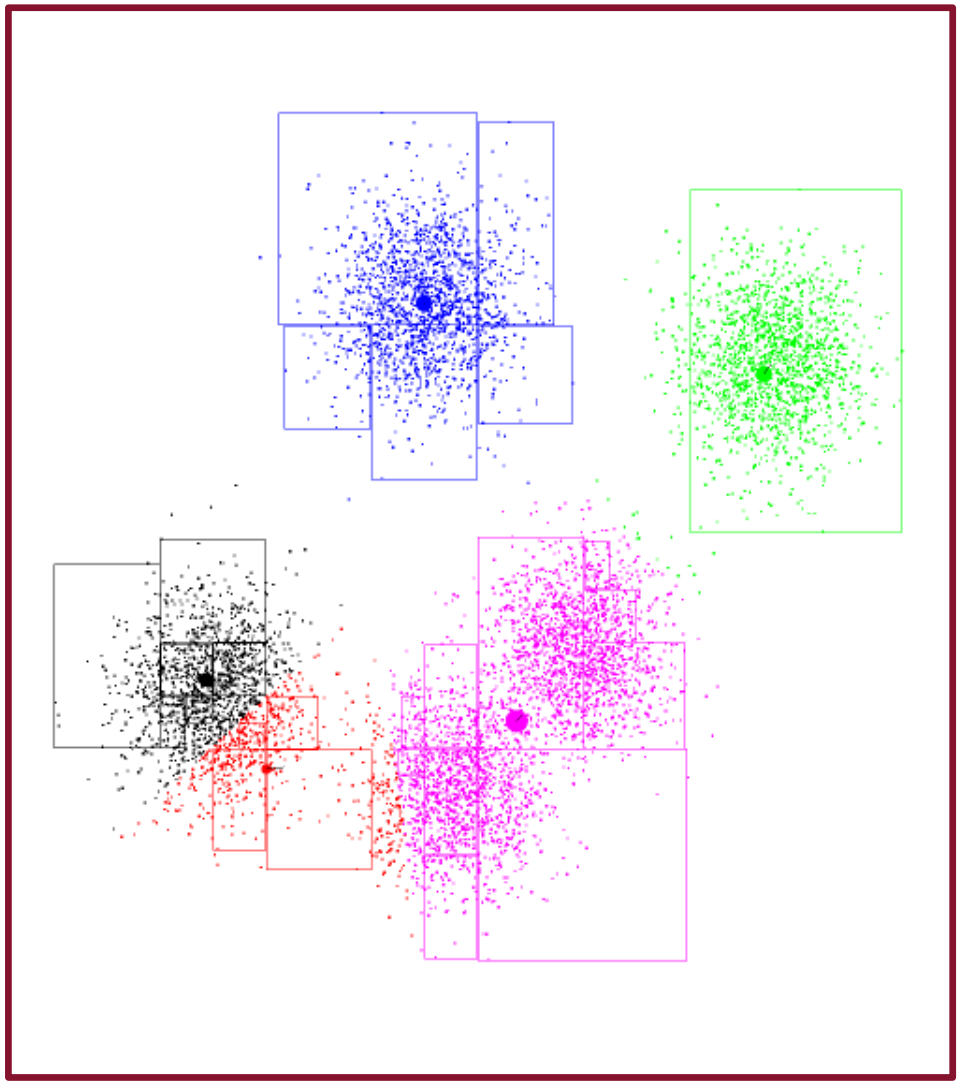


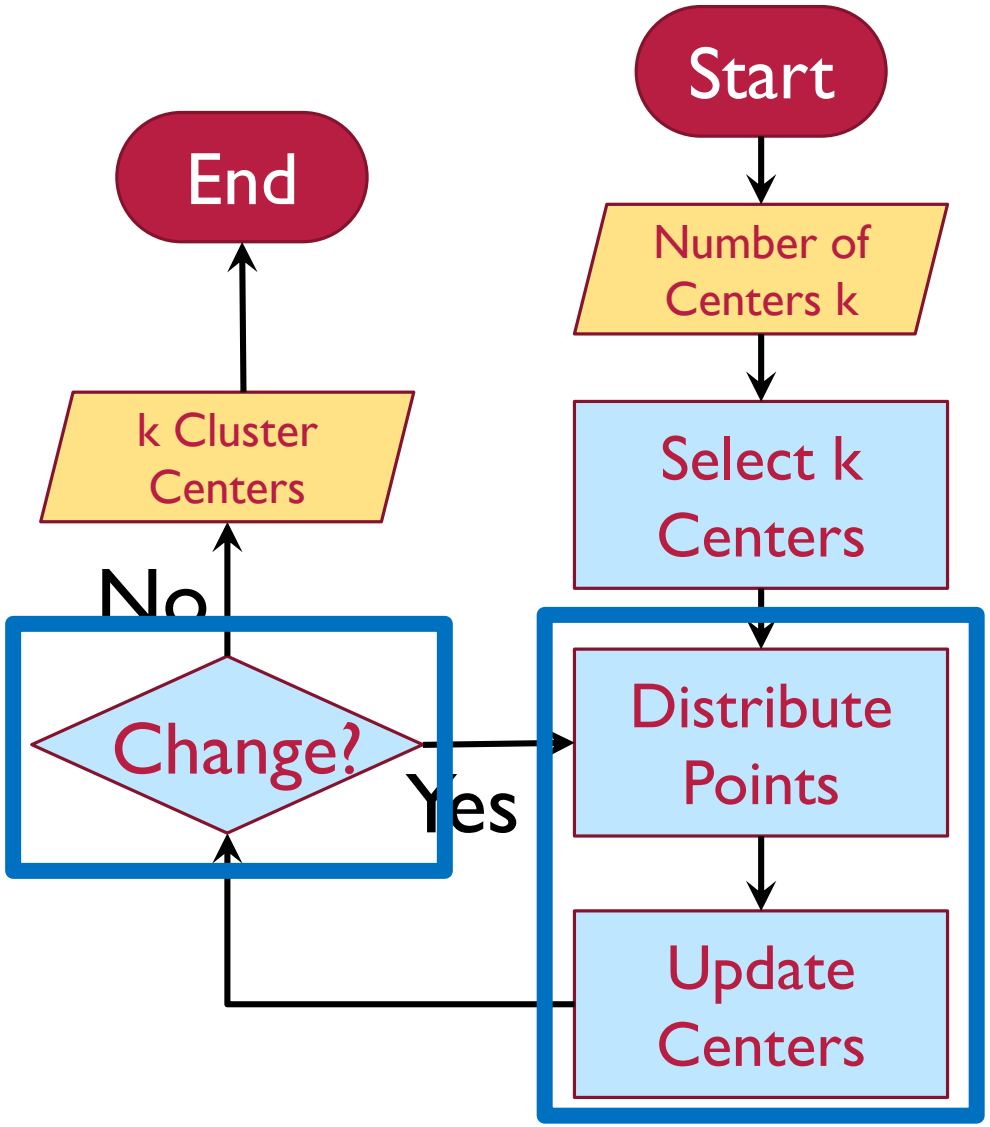
K-Means: Update 3



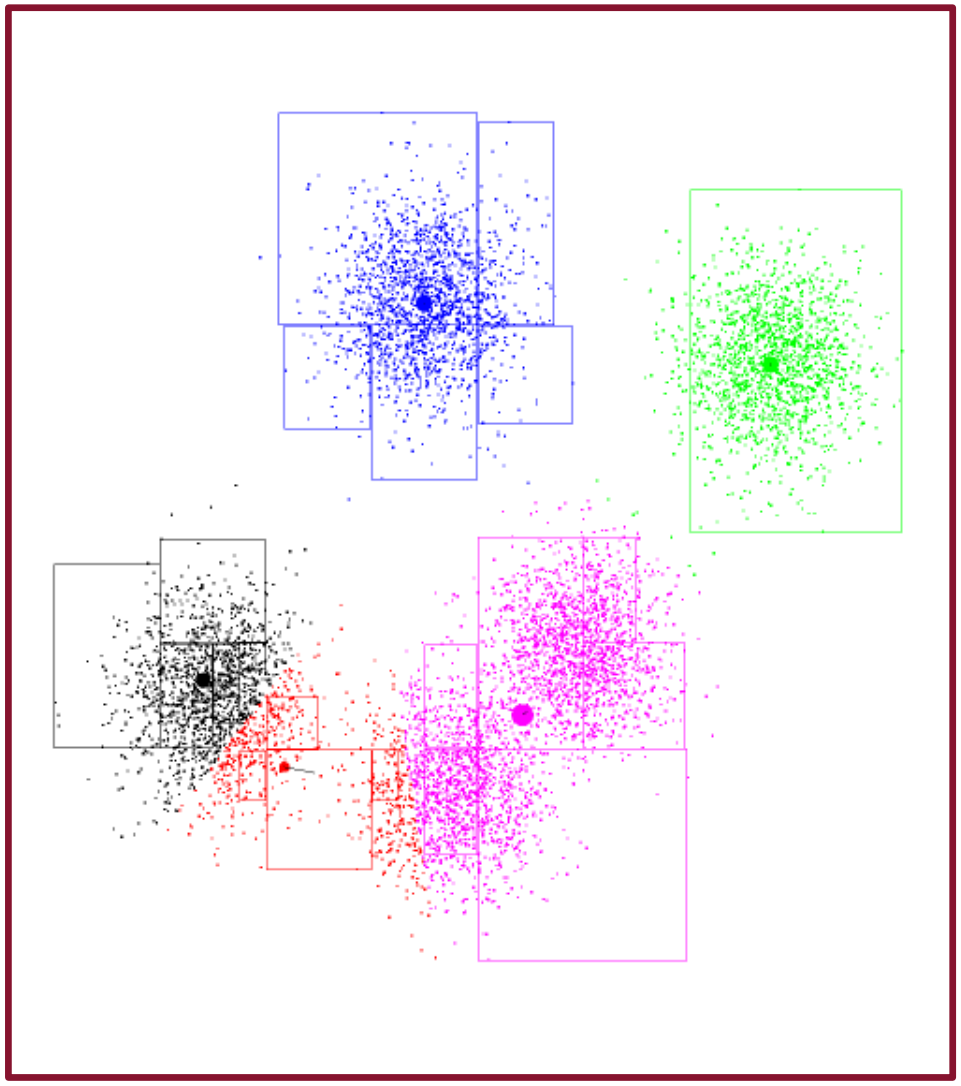


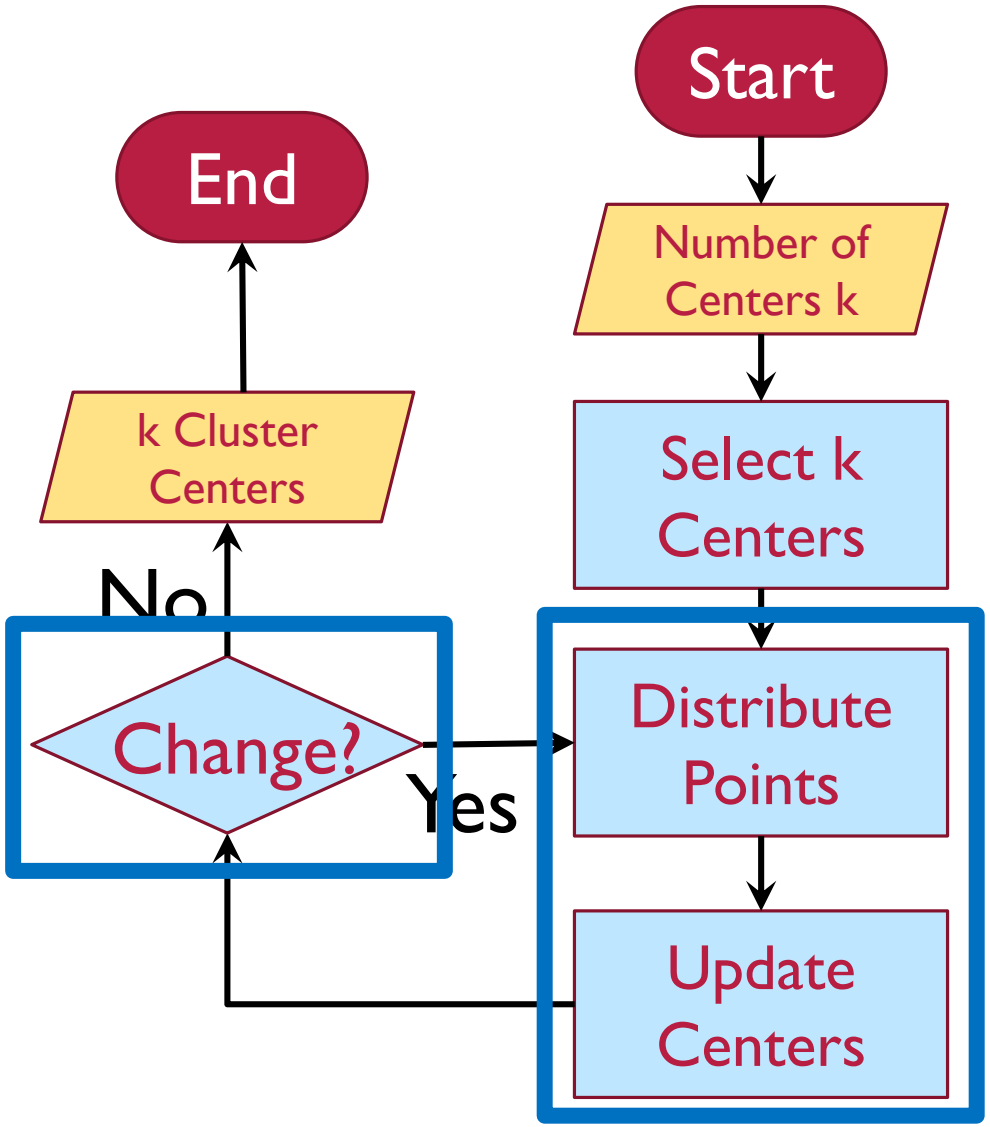
K-Means: Update 4



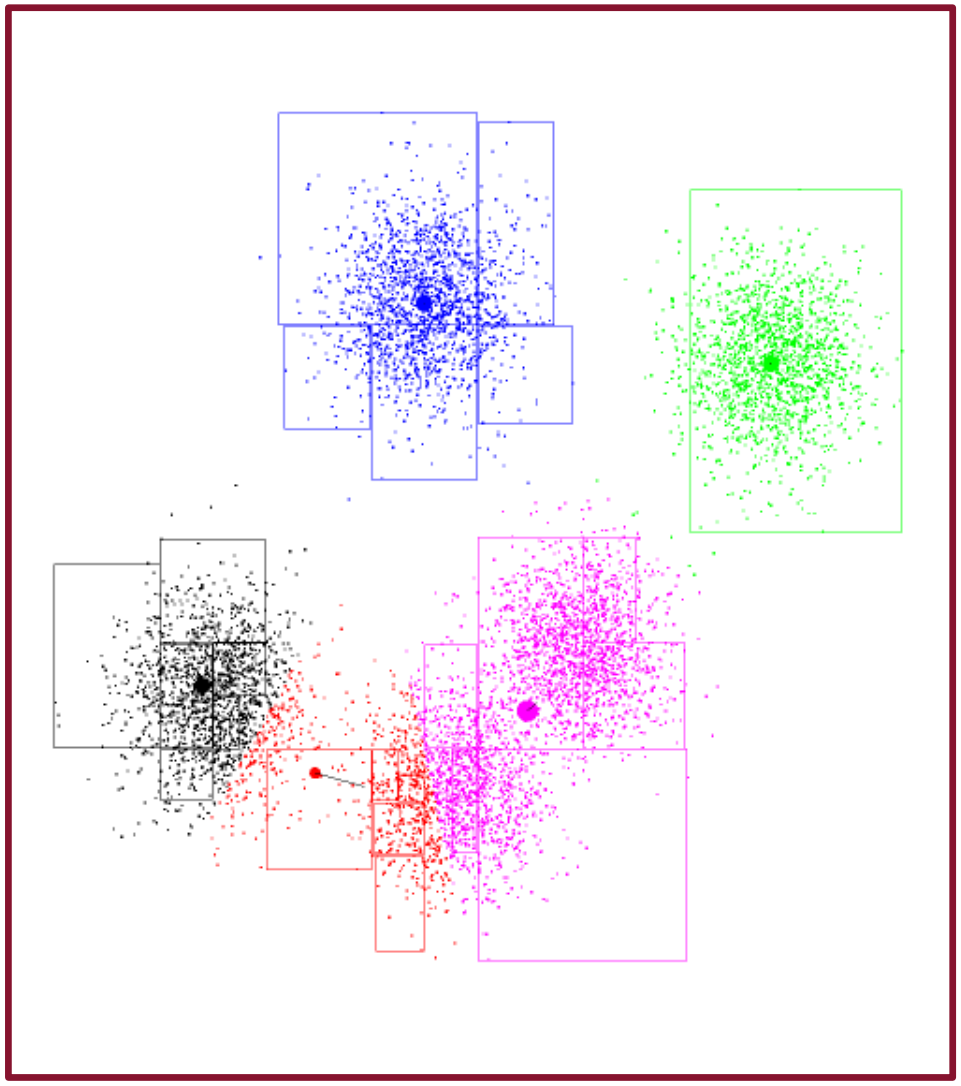


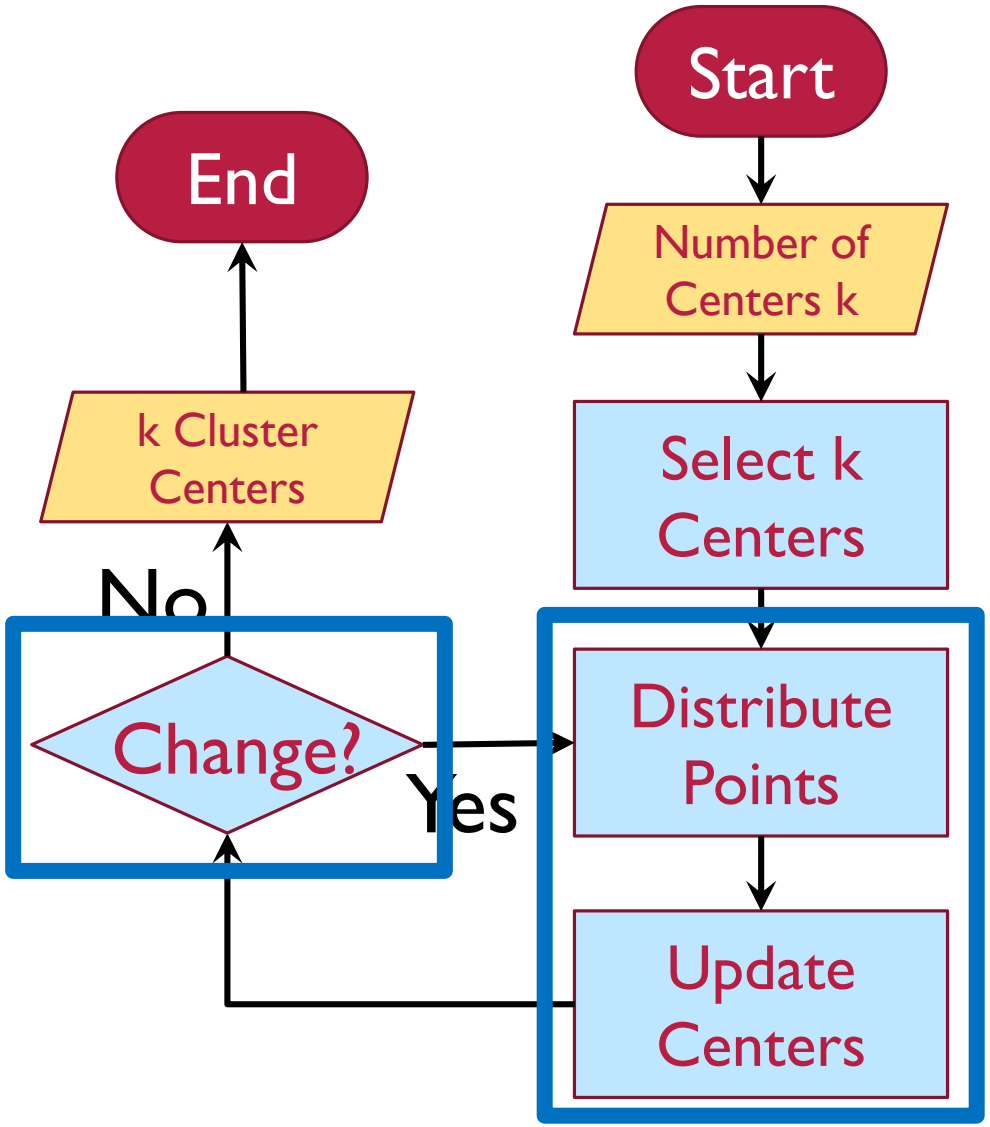
K-Means: Update 5



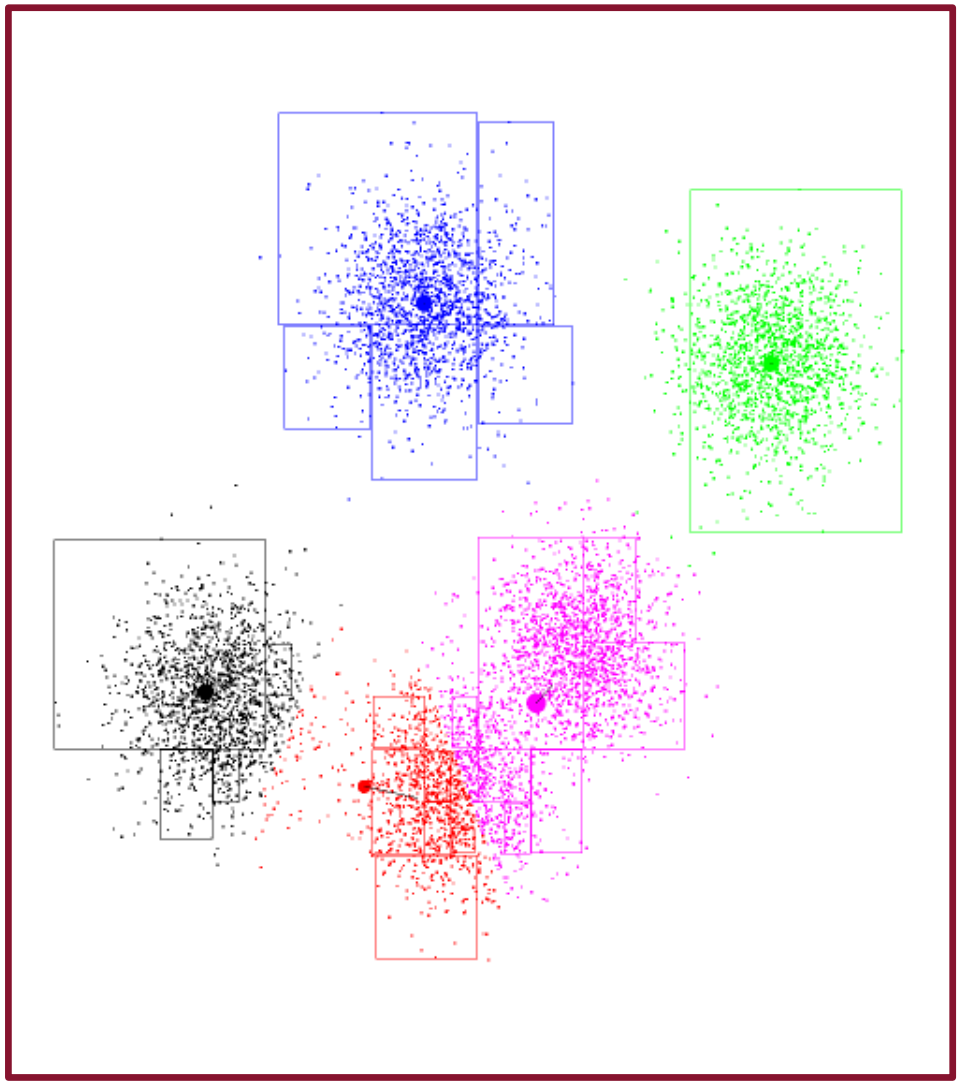


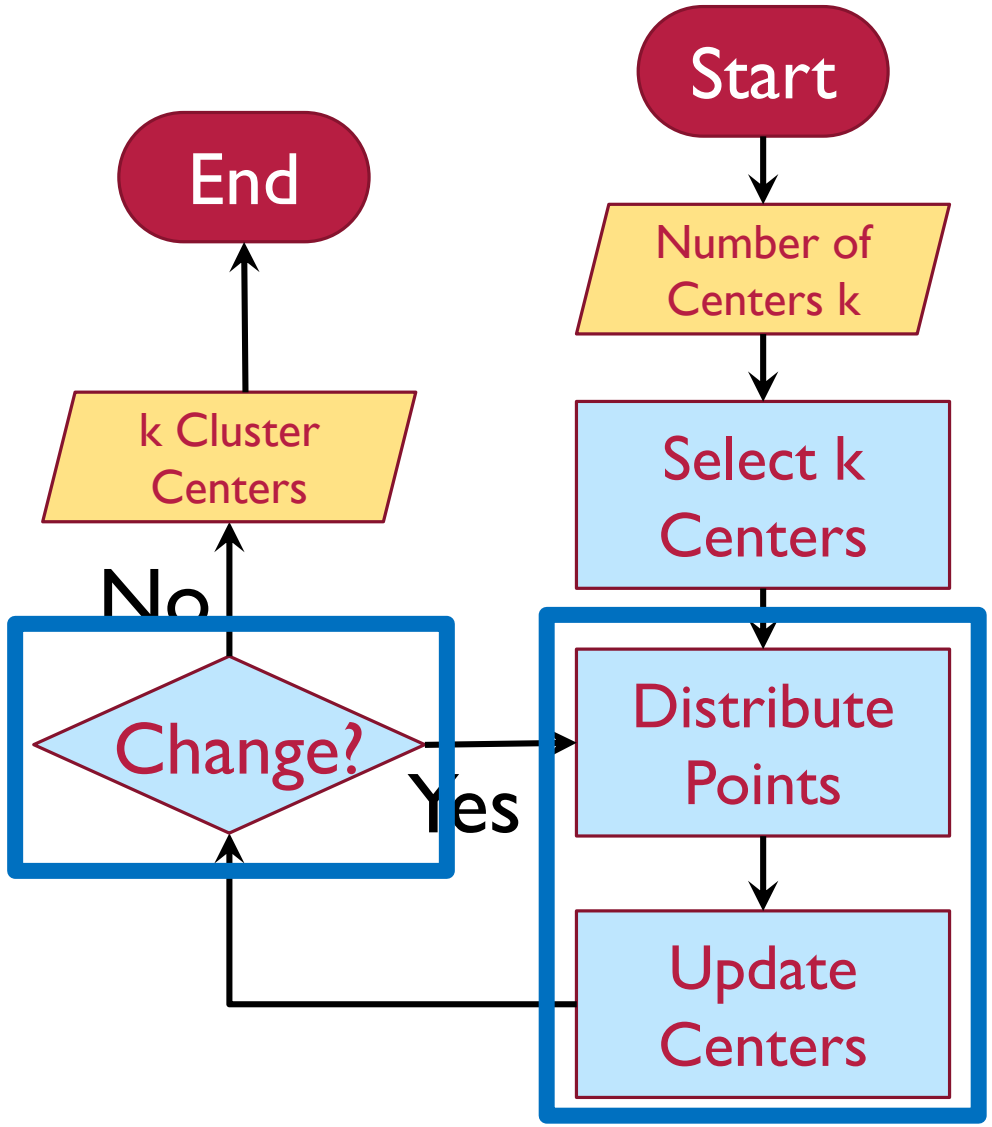
K-Means: Update 6



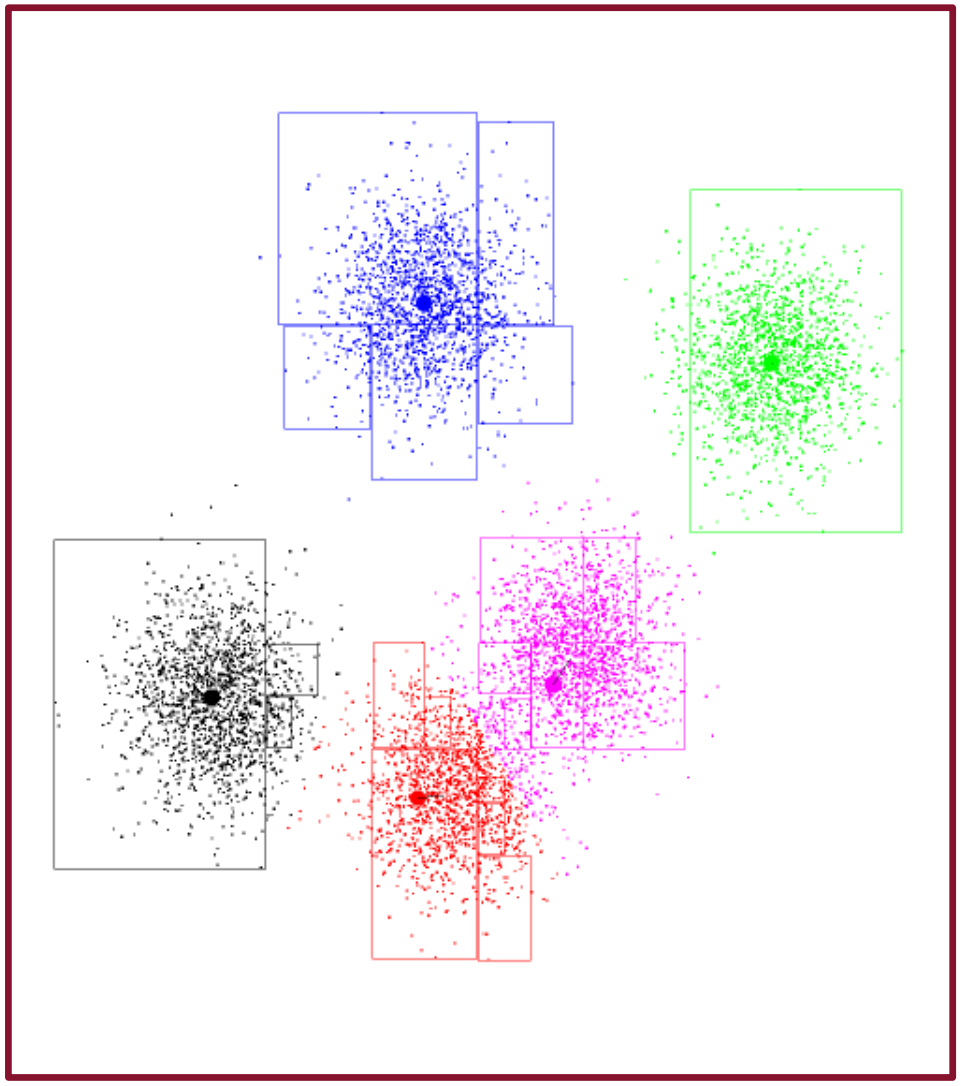


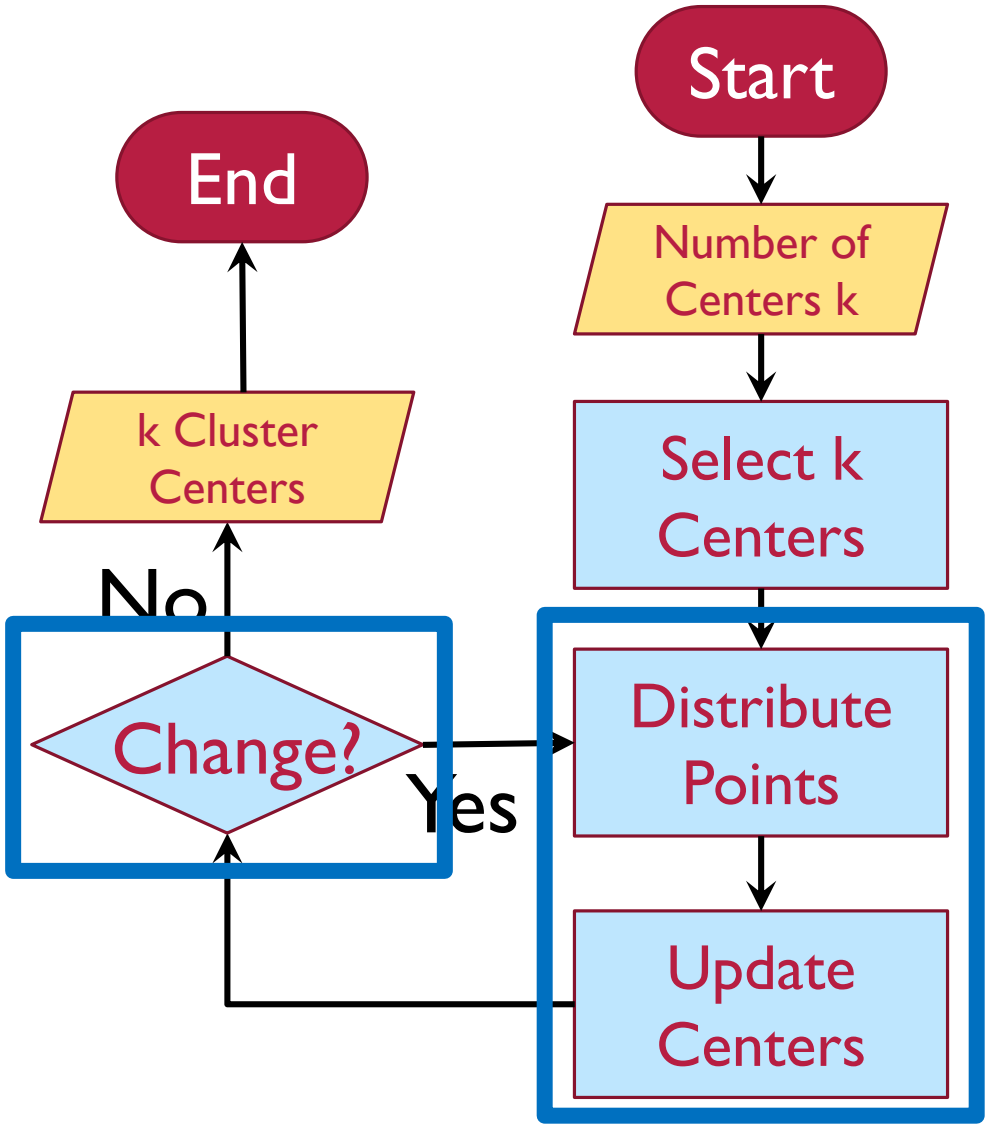
K-Means: Update 7



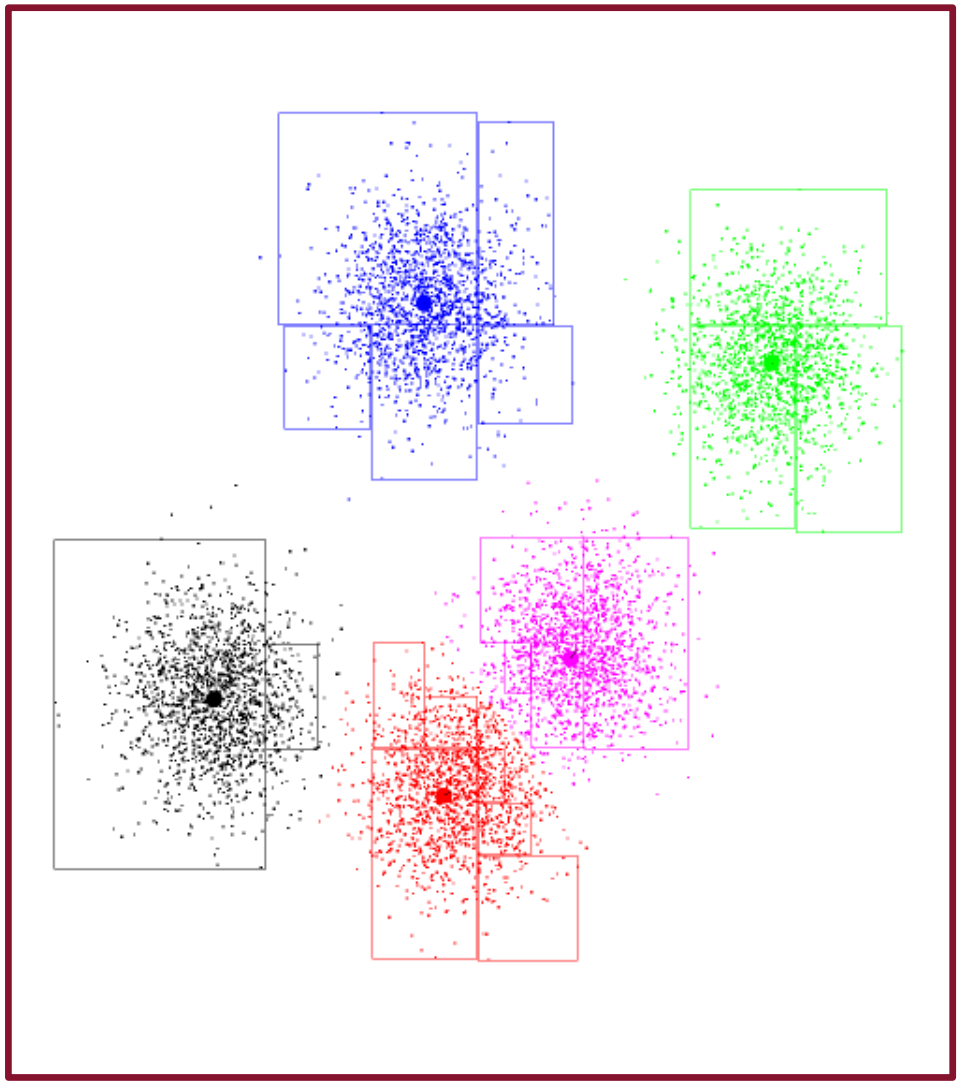


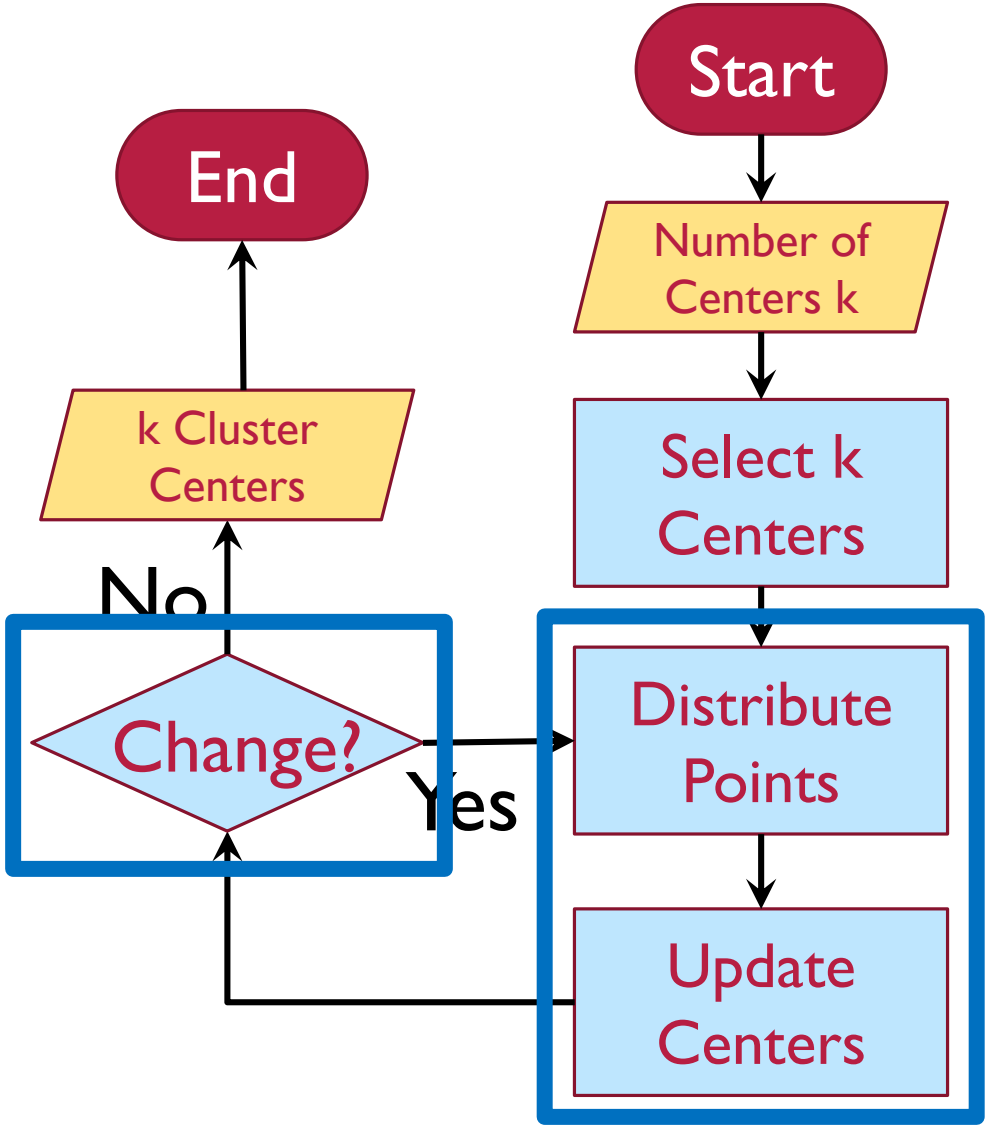
K-Means: Update 8





K-Means: Update 9





K-Means Demo

Let us try it for real

Questions?

Hierarchical Clustering

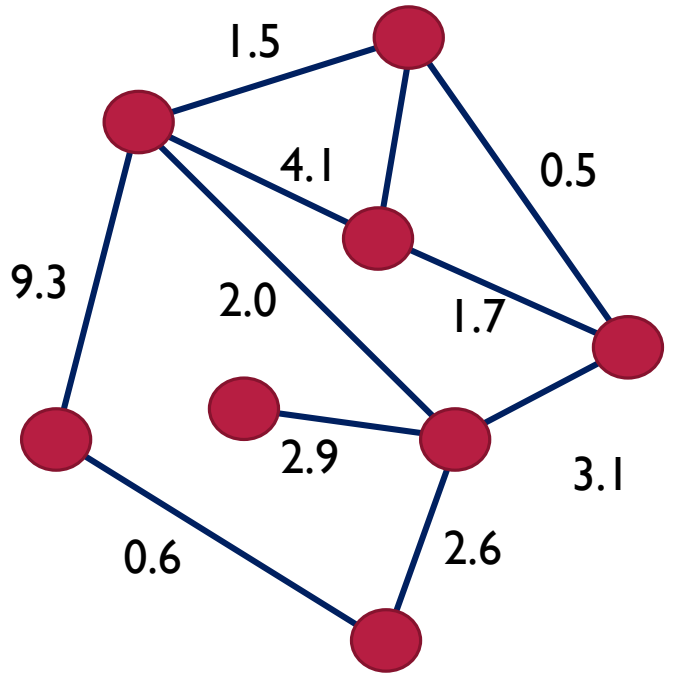
The World is not Flat

Hierarchical Clustering

- Data in the world is not flat
  - Animals, Trees, Birds, Fish, Rocks
    - Types of Animals, Species, Sub-species, Types of Rocks, …
- Can we recover the hierarchical structure from clustering
- Agglomorative vs. Divisive
  - Bottom-up vs. Top-down

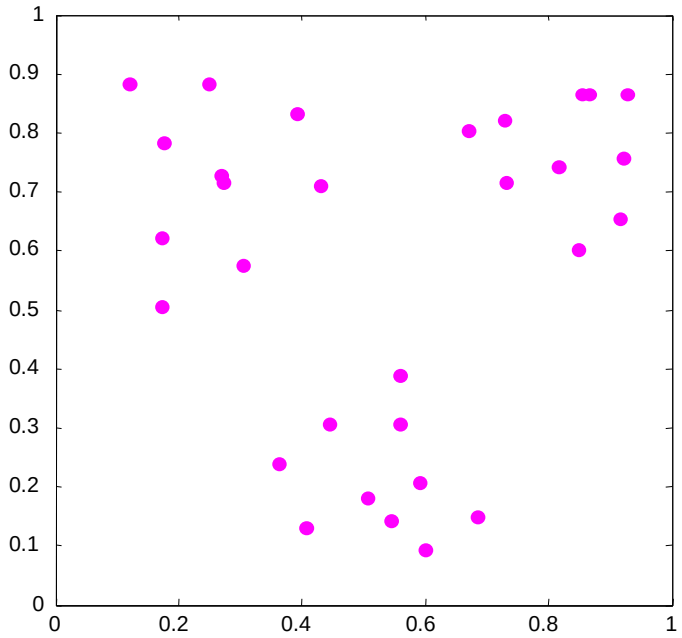
Graph Theory: A short review

- A graph is a collection of
  - Vertices, and
  - Edges
- Edges can have weights on them
- Tree is a special kind of graph
  - No cycles in the graph

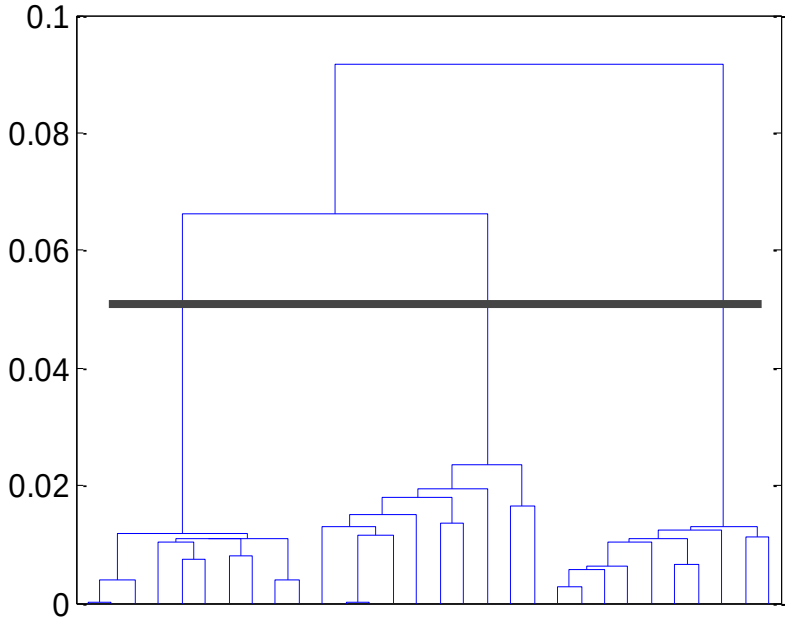


Single-Link Algorithm

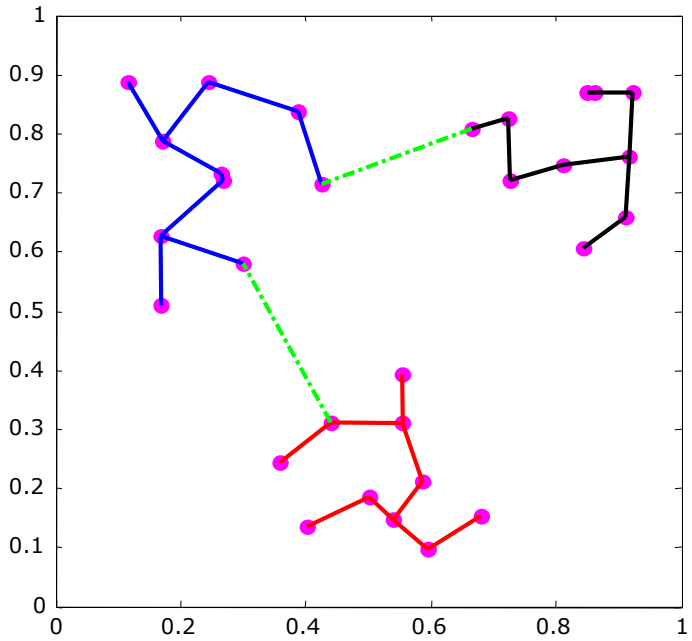
- Form a hierarchy for the data points (dendrogram), which can be used to partition the data
- The "closest" data points are joined to form a cluster at each step
- Closely related to the minimum spanning tree-based clustering



Data



Dendrogram



Single-link, k=3

User's Dilemma!

- Which similarity measure and which features to use?
- How many clusters?
- Which is the "best" clustering method?
- Are the individual clusters and the partition valid?
- How to choose algorithmic parameters?

Data Clustering: Jain and Dubes.

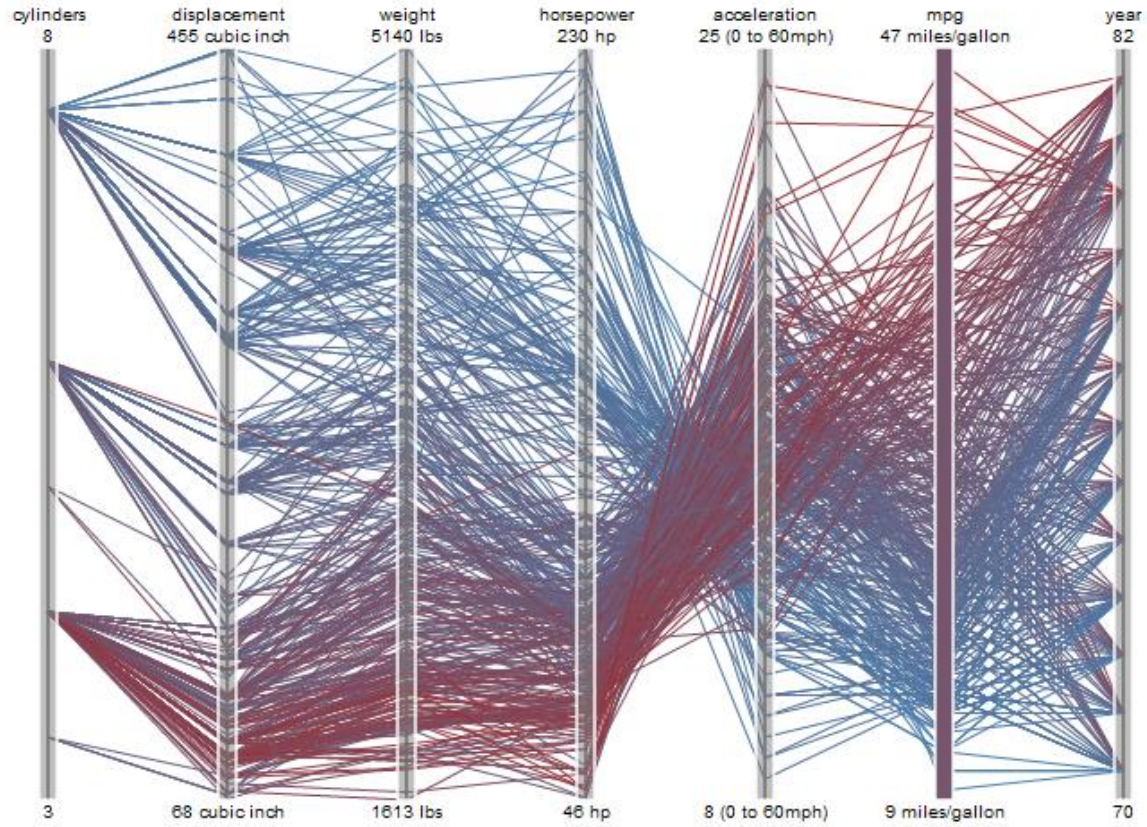
Questions?

Data Visualization

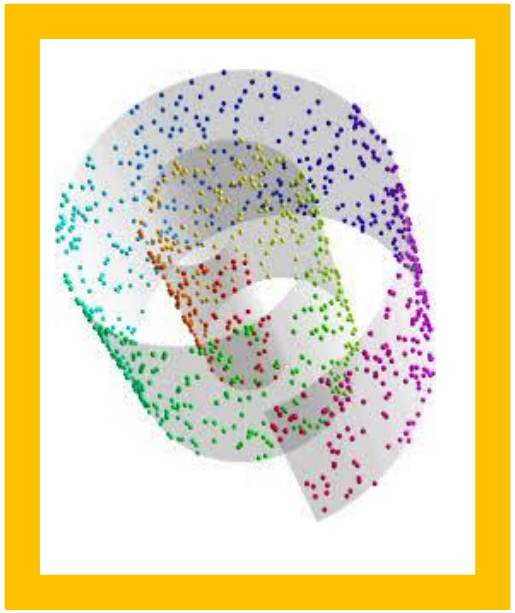
When Data is High-Dimensional

Why Data Visualization

- How do we look at HD data?
  - Dimensionality Reduction
  - Other methods
    - Have limitations
- Difference between DR and Visualization goals?



Questions?



Dimensionality Reduction

World is not always Linear

Dimensionality Reduction: Assumptions

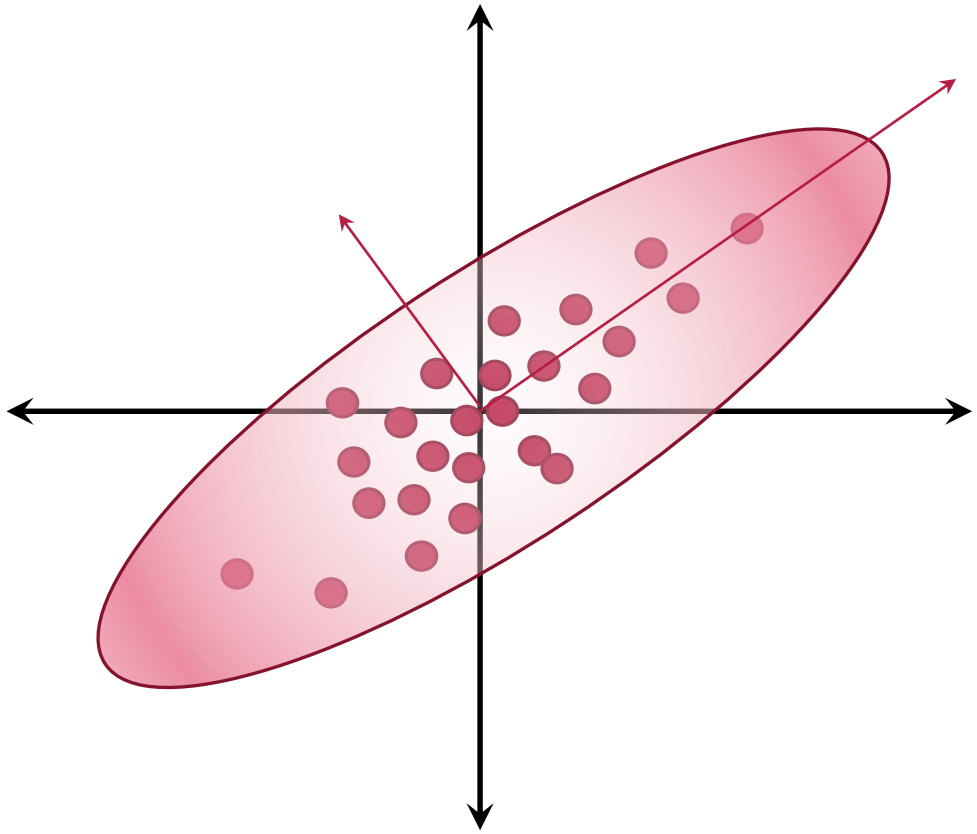
- High-dimensional data often lives in a lower dimensional manifold (sub-space)
- We can represent the data points well by using just their lower dimensional co-ordinates
- The lower dimensional data will capture the distribution of (pair-wise distances between) points in high dimensions
- If the manifold is a linear sub-space, we can use PCA.

PCA and Covariance Matrix

- Going from 2D to 1D
- Which feature to select?
  - This may be any feature (or vector in any direction)
- Two view points:
  - Maximize variance
  - Minimize error
- Solution to both happens to be:

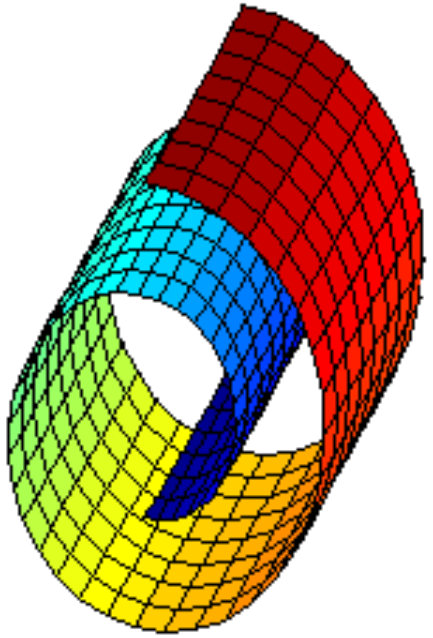
$$\mathbf{Ax} = \lambda\mathbf{x}$$

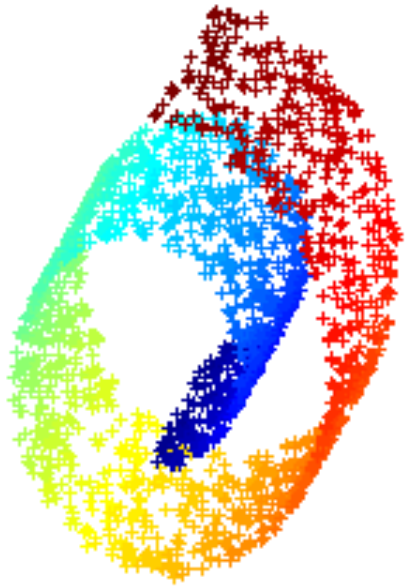
- What is A? How do we solve this?



PCA: Fundamental Shortcoming

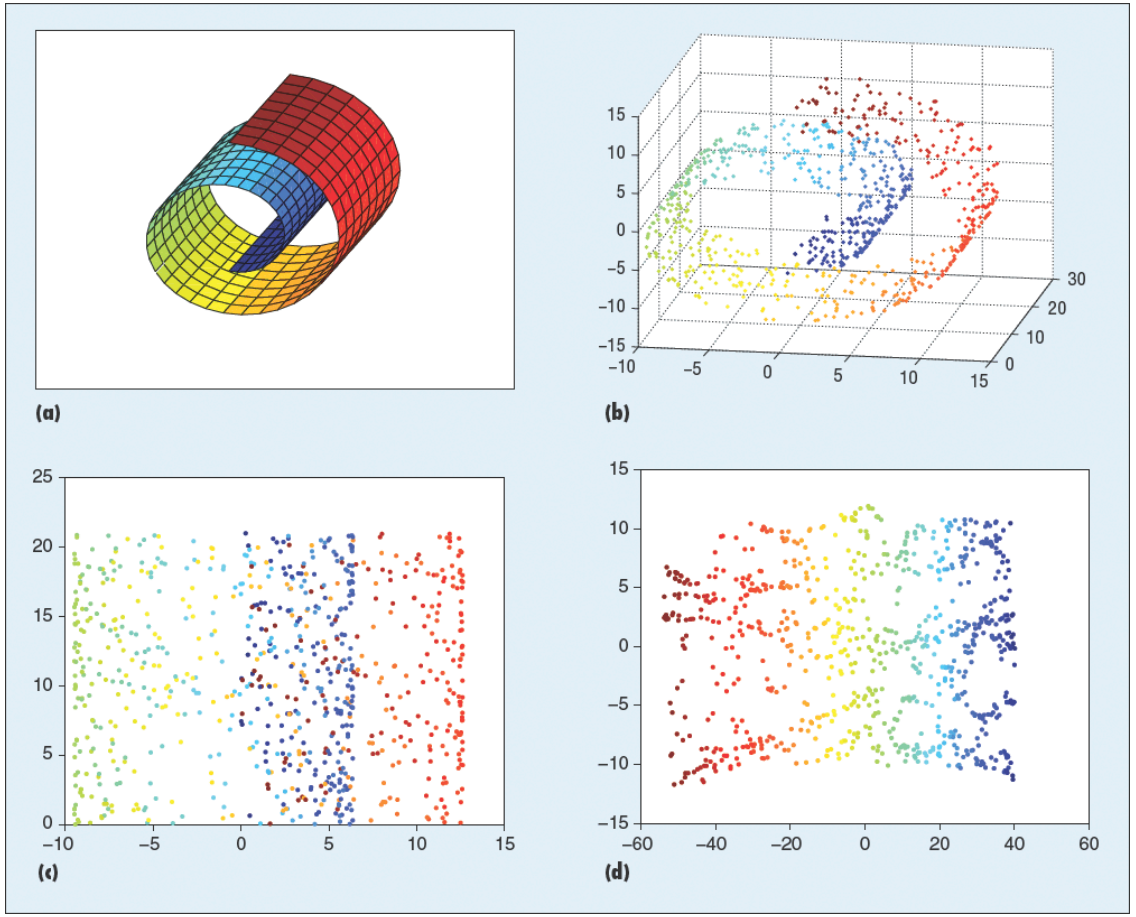
- World is often non-linear
- Consider the Swiss-Roll dataset
  - What would PCA give?
  - What do we want?





What do we really want?

- Find a lower-dimensional representation such that:
  - Distances in LD ≅ Distances in HD
  - Closer distances are more important
- Unrolling the swiss roll
- Do not insist on being able to get HD back from LD
  - Using for visualization



MDS: Multidimensional Scaling

- Find a representation that best preserves pair-wise distances
- Mathematically:
- Start with random vectors in LD
- Update the locations so as to minimize the cost function
  - Gradient descent

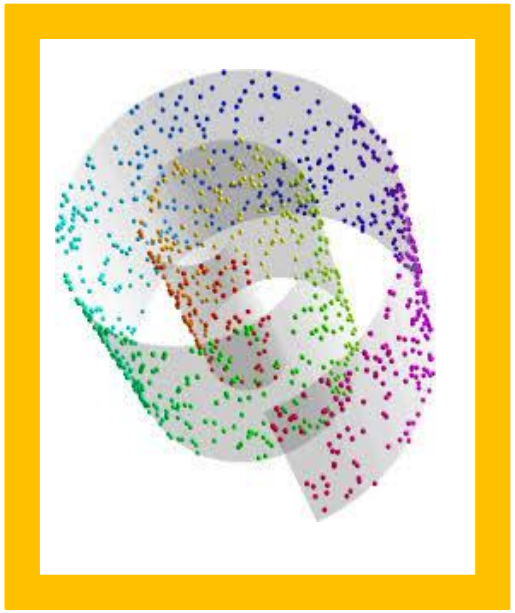
$$Cost = \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2$$

$$d_{ij} = \| x_i - x_j \|^2$$

$$\hat{d}_{ij} = \| y_i - y_j \|^2$$

Can get stuck in local minima

Still Linear Can be PCA



Non-Linear Dimensionality Reduction

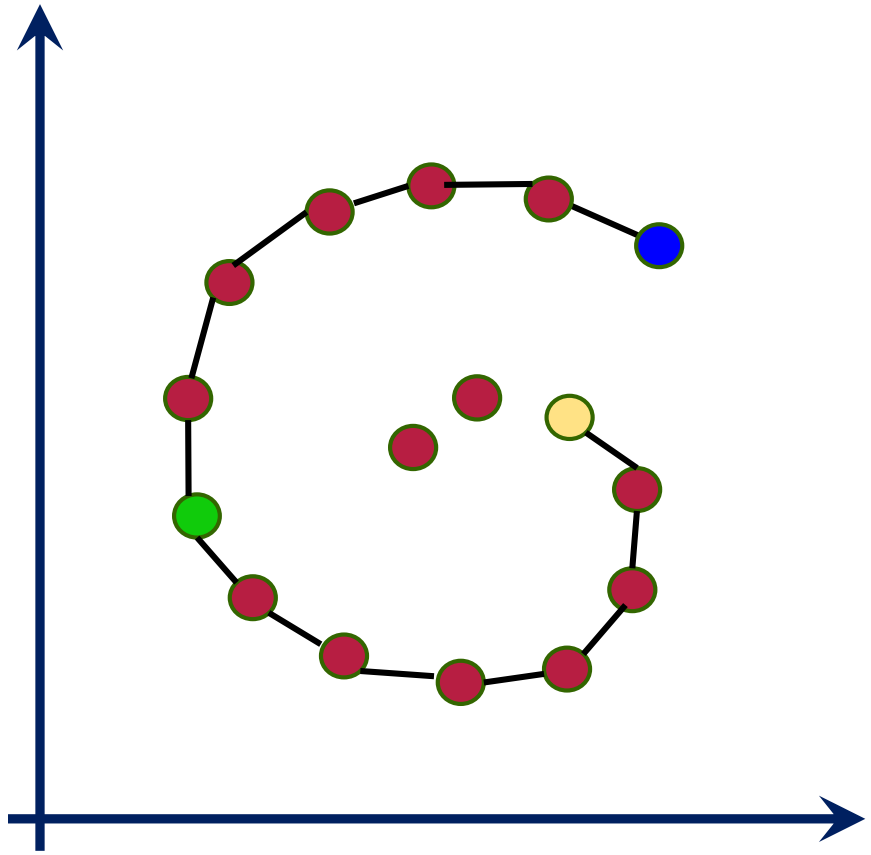
ISOMAP and LLE

Dimensionality Reduction: Approaches

- Global Approach: All distances in HD are equally important and should be captured in the LD representation
  - Is this PCA?
- Local Approach: Only smaller distances in HD are meaningful/reliable/ interesting to us
  - Could also weigh smaller and larger distances differently

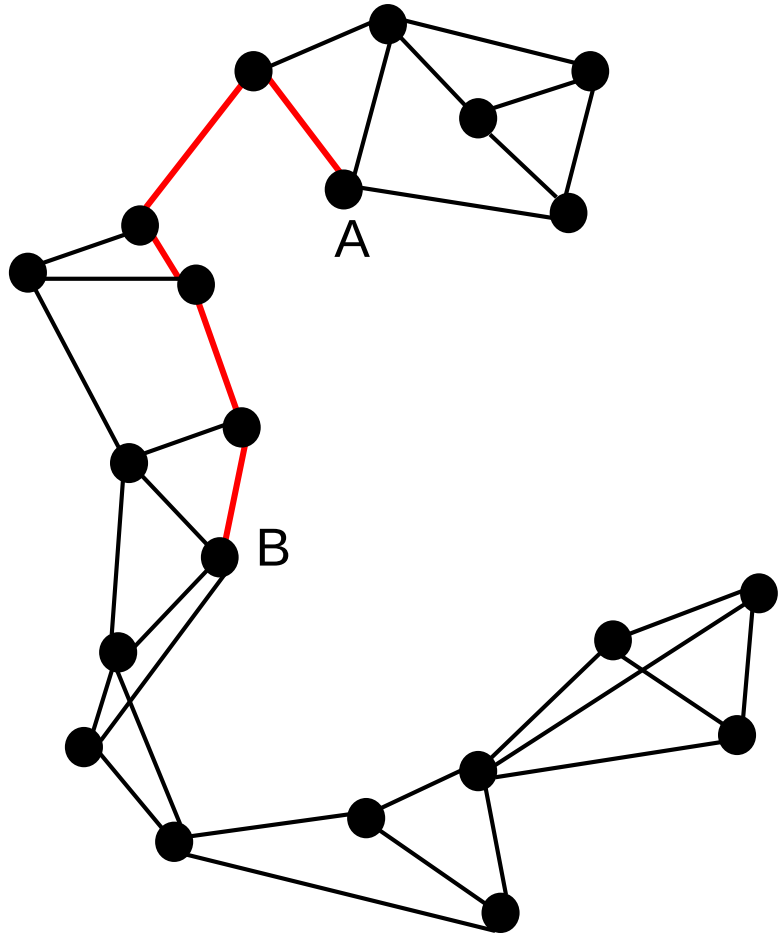
ISOMAP

- d(●,○) > d(●,●)
- Is Euclidean metric the right distance metric?
- How to robustly measure distances along the manifold?
- If we can do that, the global optimization is efficient (a variant of PCA)

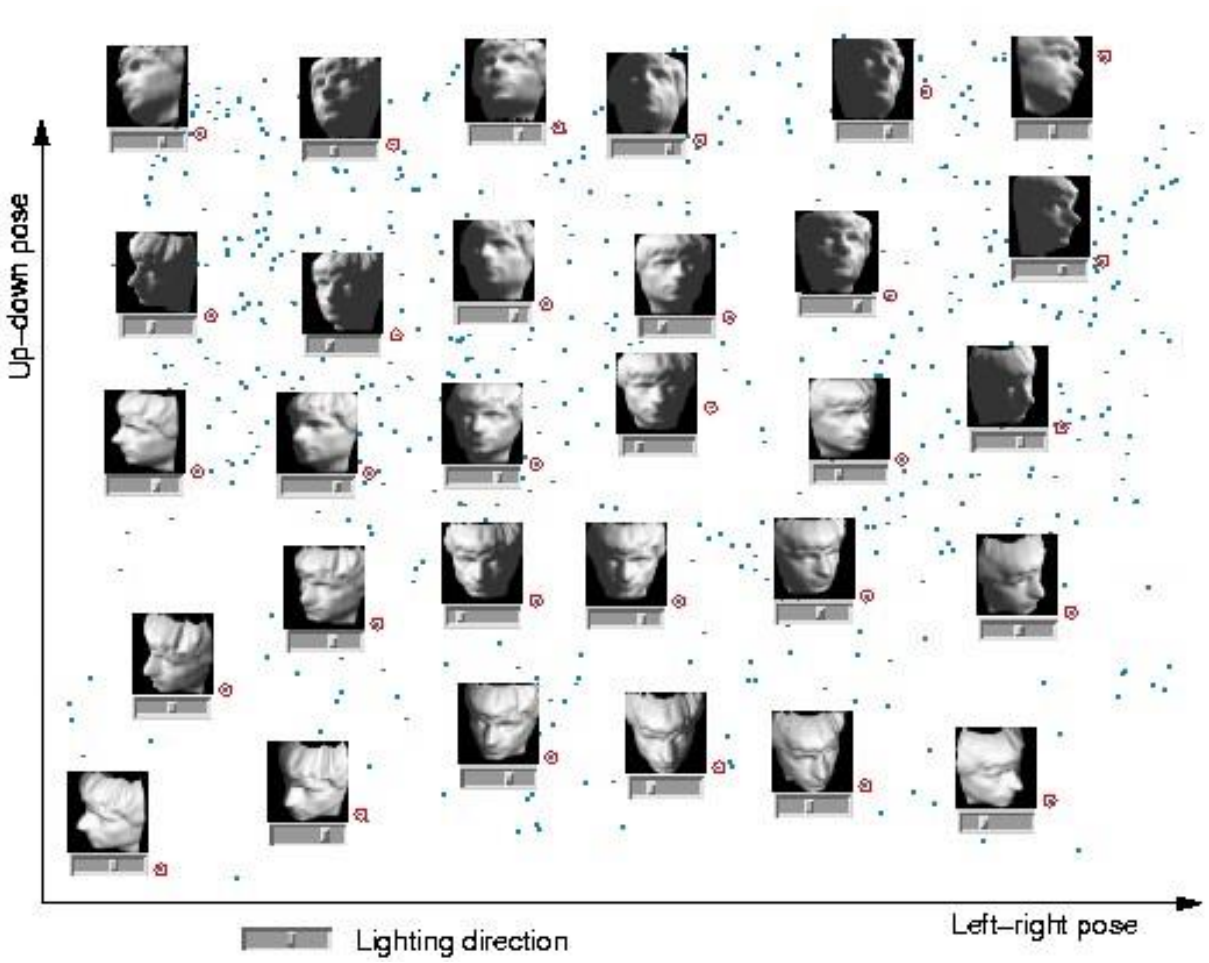


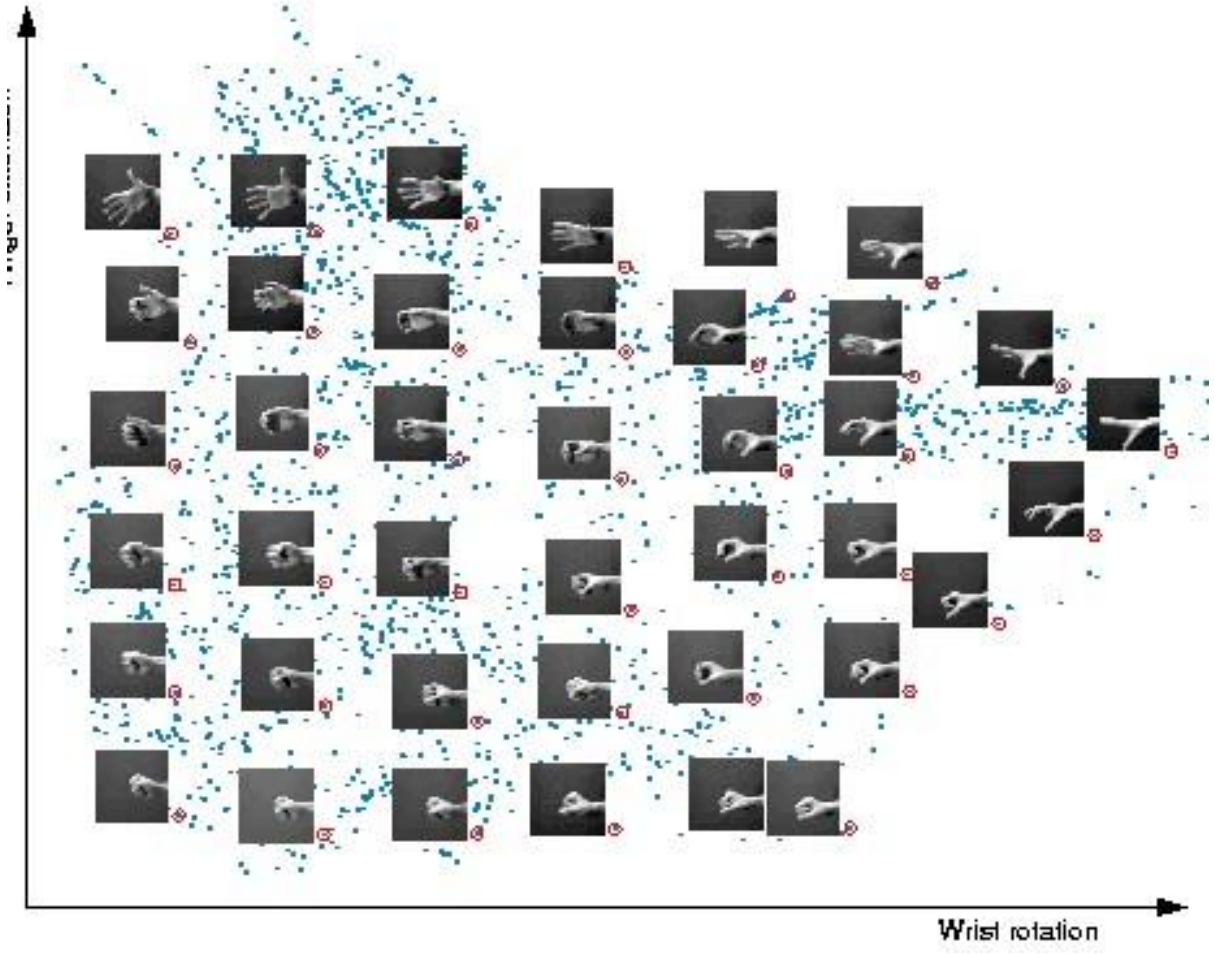
ISOMAP

- How does ISOMAP measure the MD?
- Connect each data point to its K nearest neighbors in the high-dimensional space.
- Link weights: True Euclidean distances.
- MD(A,B) = ShortestPath (A,B) in this *neighborhood graph*.
- Compute the low-dimensional embedding as in Metric MDS.

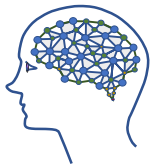


ISOMAP Results





LLE: Locally Linear Embedding



- Idea: Preserve the structure of local neighborhood
- Approach:
  - Represent each point as a weighted combination of its Neighbors in HD. Remember the w_{ij}s.
  - Find a LD representation that minimize the representation error:

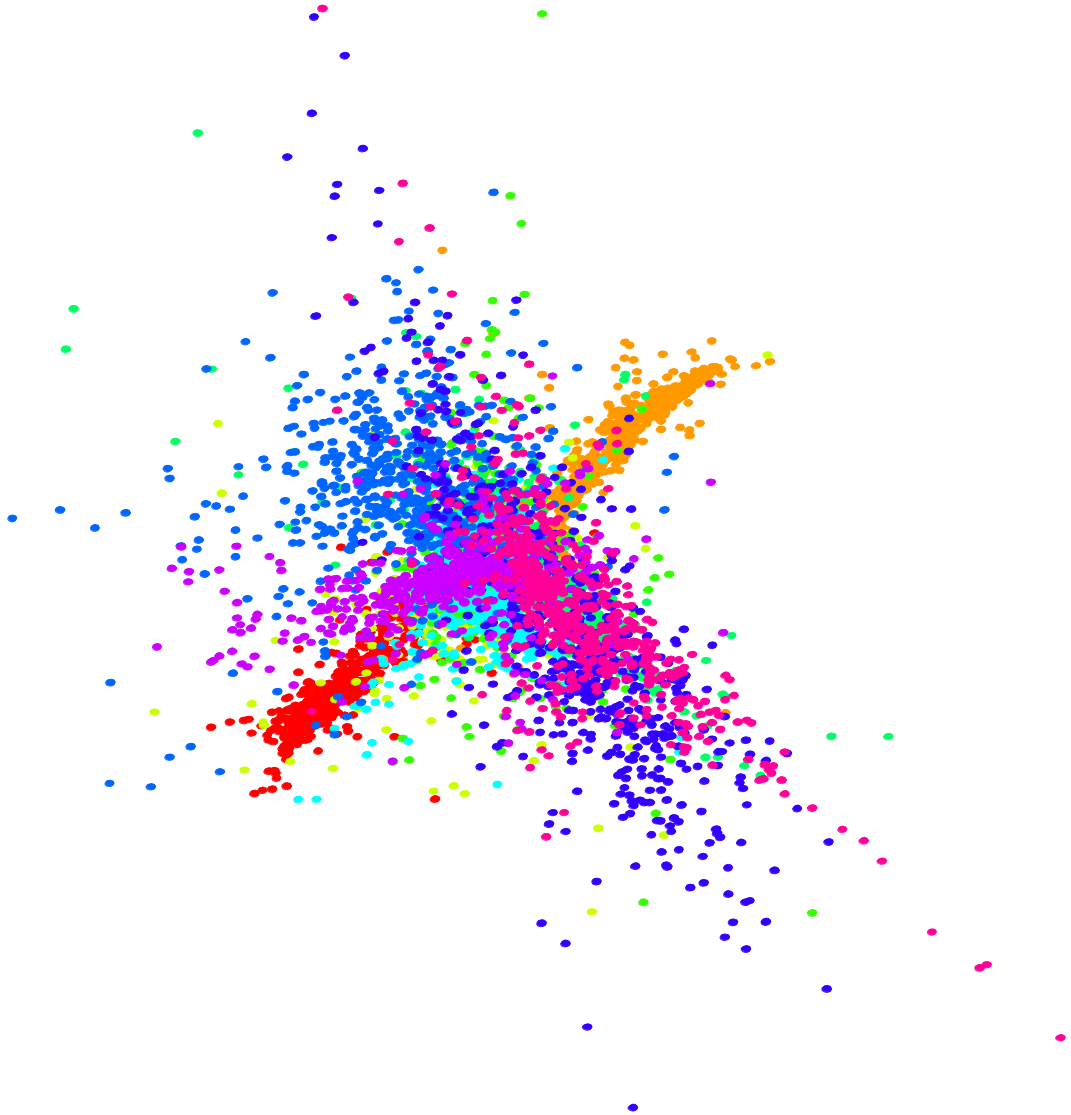
$$\mathbf{x}_i \approx \sum_j w_{ij}\mathbf{x}_j$$

$$Cost = \sum_i \| \mathbf{y}_i - \sum_{j \varepsilon N(i)} w_{ij}\mathbf{y}_j \|^2$$

  - Constraint: ys have unit variance in each dim.

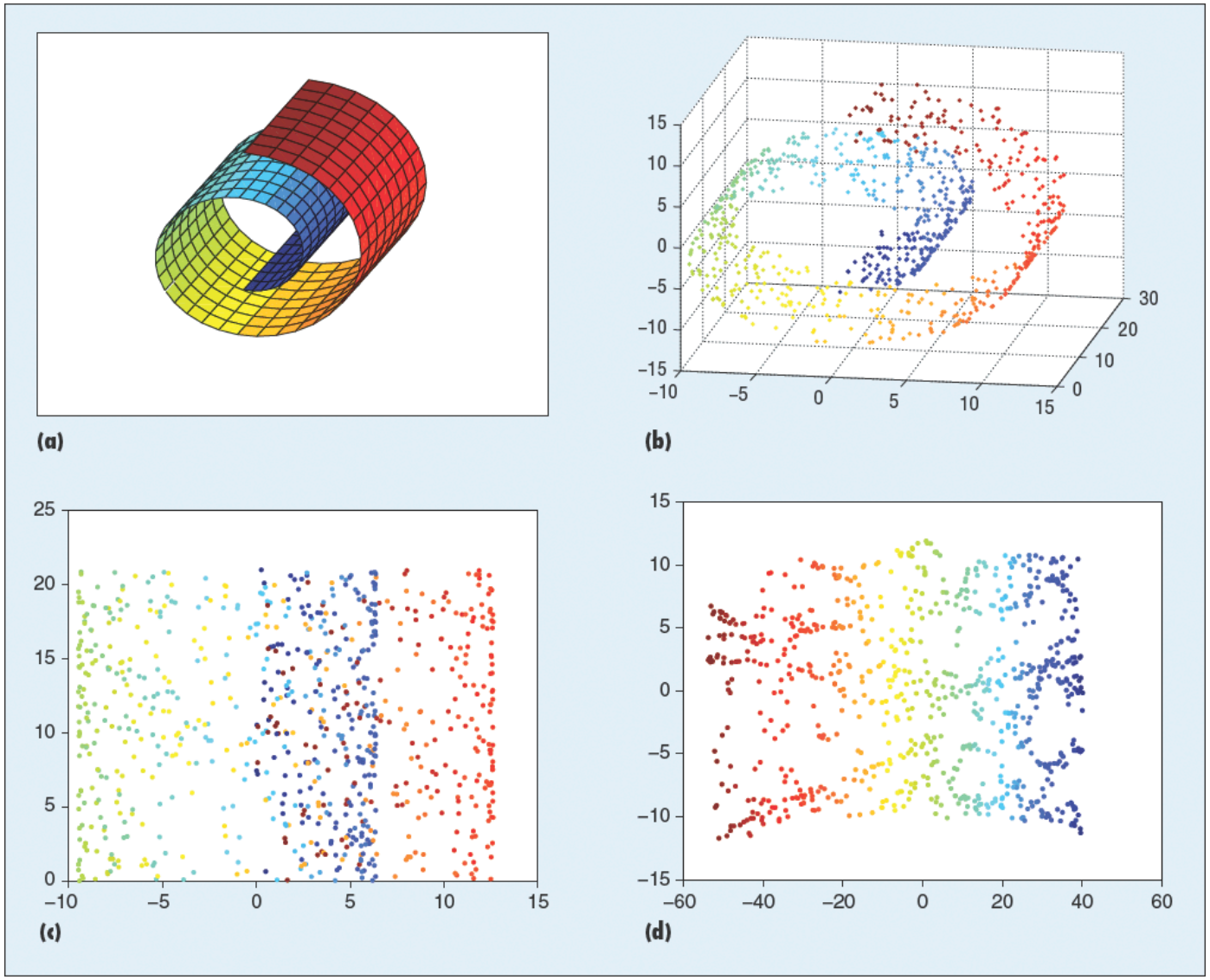
Typical LLE Embedding

- Most of the data is close to the center of the space.
- A few points are far from the center to satisfy the unit variance constraint.

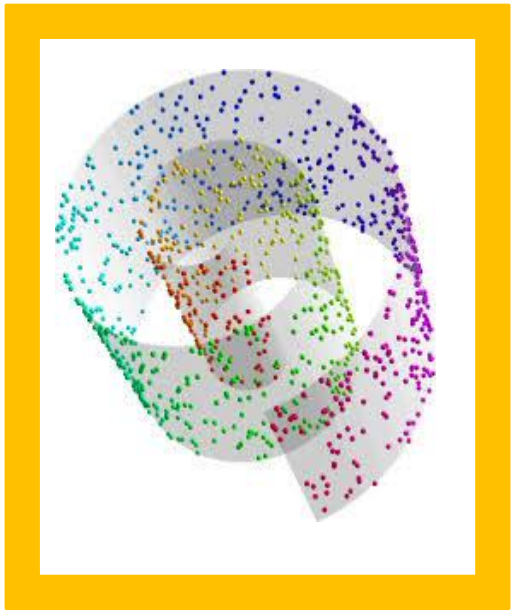


What about the Swiss Roll?

- The bottom two are ISOMAP and LLE



Questions?

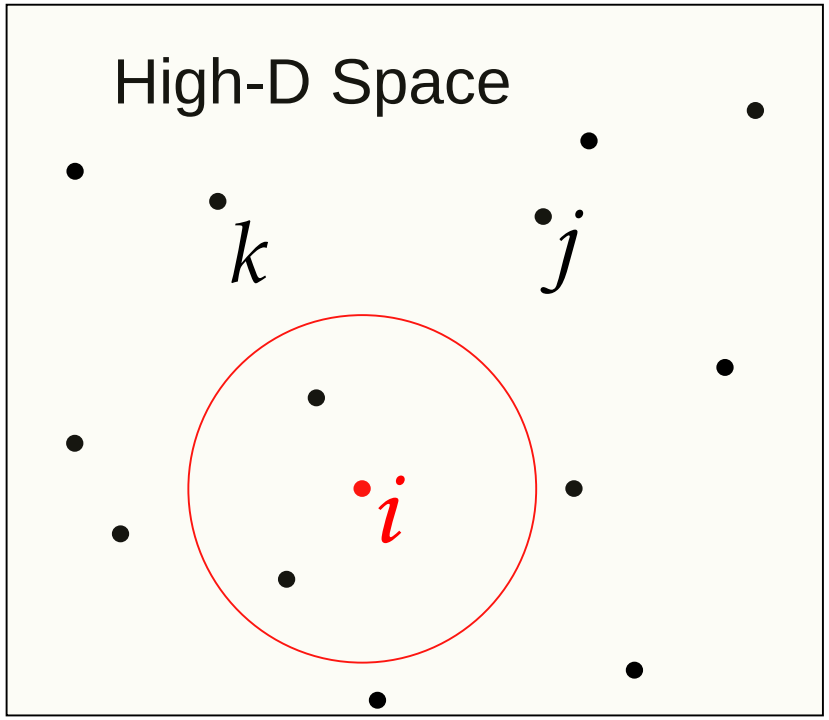


t-SNE

Improving Visualization

SNE: A Probabilistic Embedding

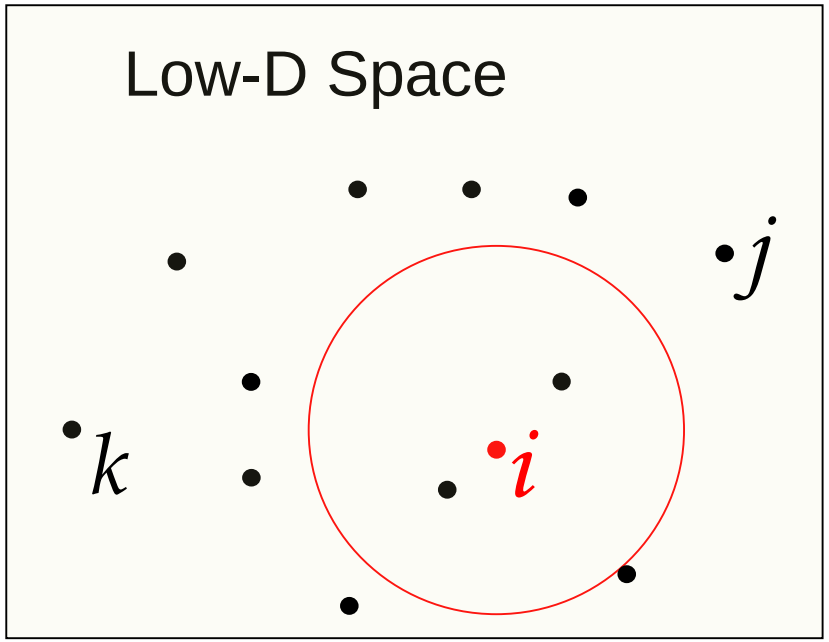
- For point j, there is a probability of it being called a neighbour of i.
- The probability is a function of the distance between i and j in HD.
- We end up with a matrix of probabilities.
- Each point is then represented as a probability distribution over all other points: A row of the above matrix



$$p_{j|i} = \frac{e^{-d_{ij}^2 / 2\sigma_i^2}}{\sum_k e^{-d_{ik}^2 / 2\sigma_i^2}}$$

AIML

- Give each data point a location in the LD space.
- Evaluate this representation by seeing how well the LD probabilities model the HD ones.



$$q_{j|i} = \frac{e^{-d_{ij}^2}}{\sum_k e^{-d_{ik}^2}}$$

Computing the LD Embedding

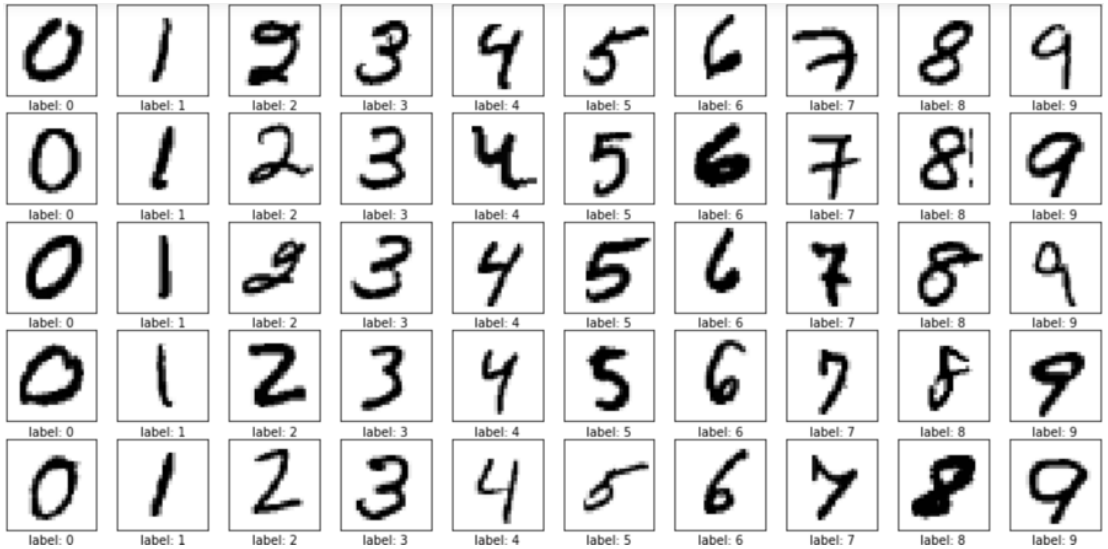
$$Cost = \sum_i KL(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

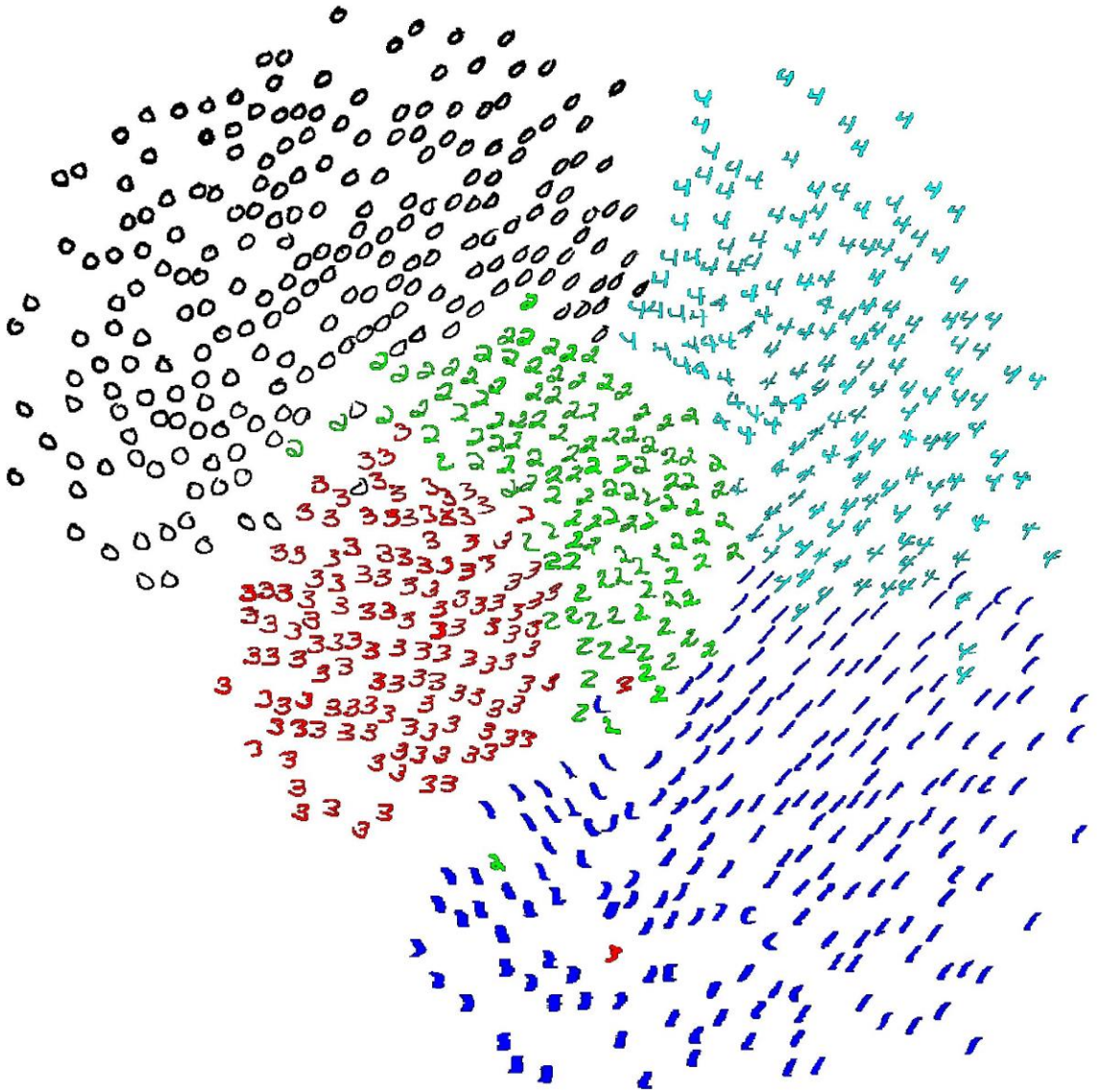
- For points where p_{ij} is large and q_{ij} is small we lose a lot.
  - Nearby points in high-D really want to be nearby in low-D
- For points where q_{ij} is large and pij is small we lose a little because we waste some of the probability mass in the Q_i distribution.
  - Widely separated points in HD have a mild preference for being widely separated in LD.

SNE on MNIST

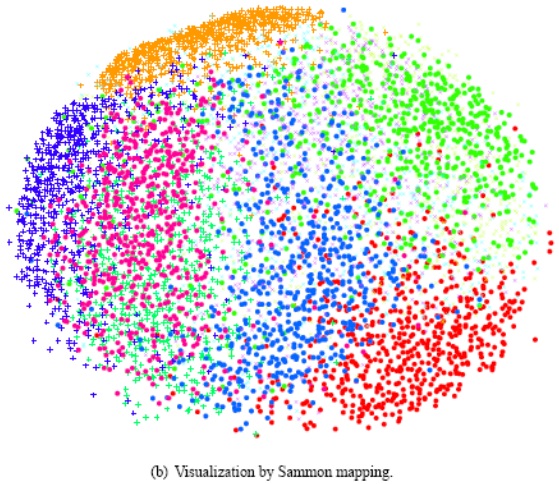
MNIST Handwritten digits dataset

- 28x28 binary images
- Large variations in writing



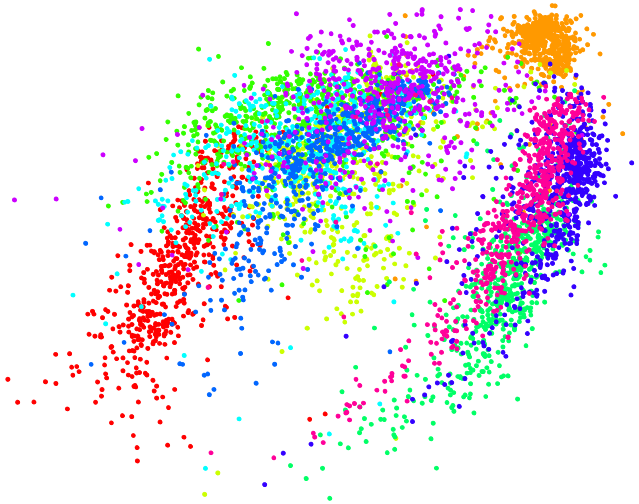


MDS, ISOMAP and LLE on MNIST

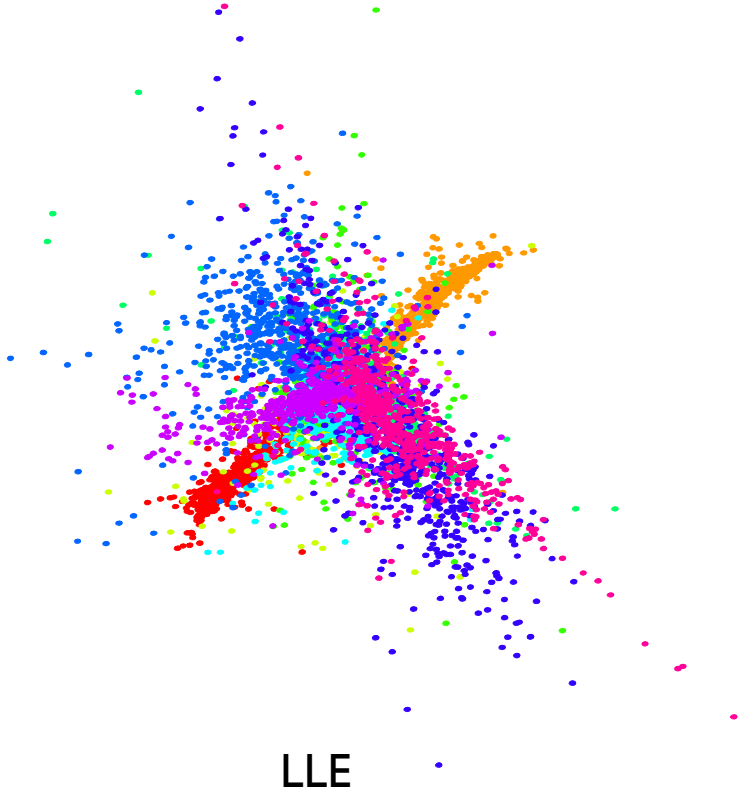


(b) Visualization by Sammon mapping.

MDS



ISOMAP

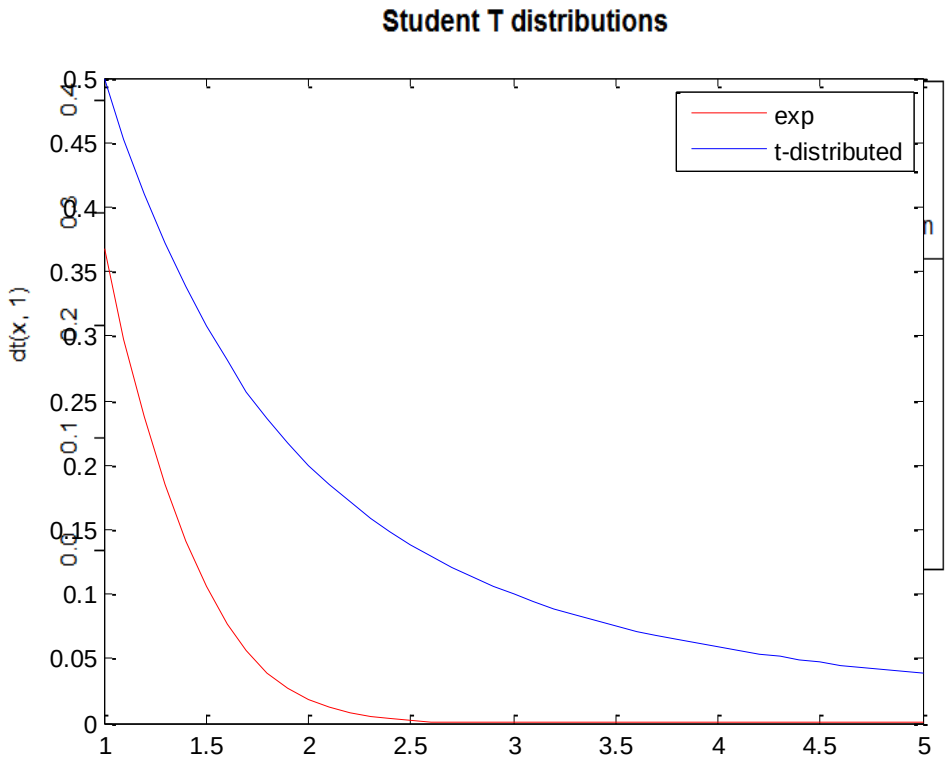


LLE

t-SNE: SNE With a heavy tail

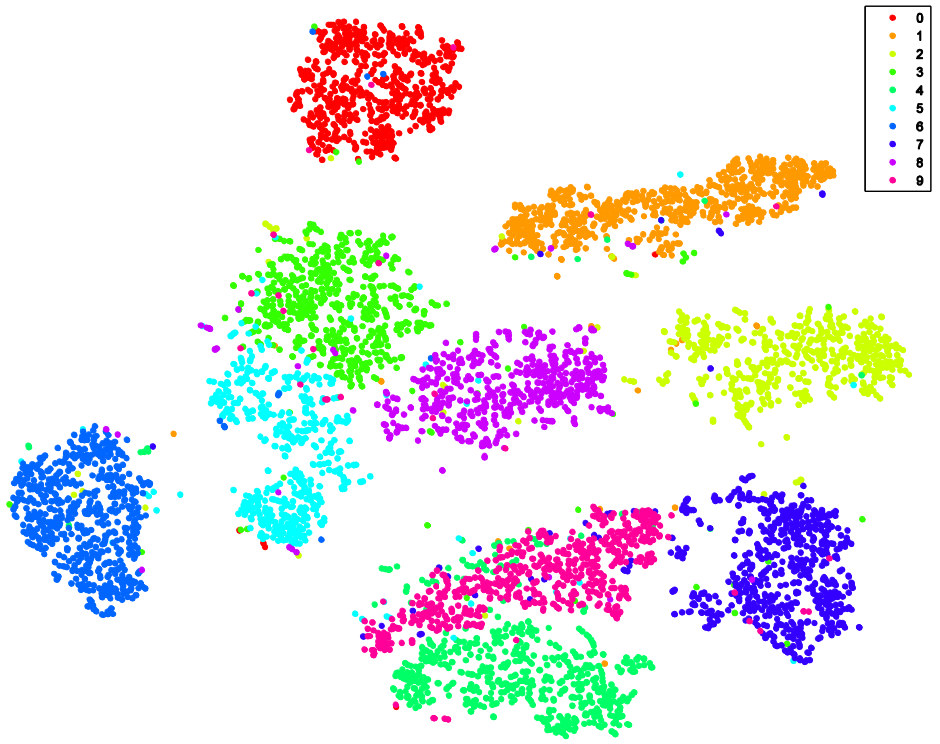
- Instead of using gaussian for LD, use a t-distribution
- Allows for farther points moving away and gaps to appear between densities.
- Can be efficiently computed

$$q_{ij} \propto \frac{1}{1+d_{ij}^2}$$



$$\exp(-x^2) \Rightarrow \frac{1}{1+x^2}$$

t-SNE on MNIST; Summary



- Classes are much better separated
- Note that the method is unsupervised!!!
- Efficient approximations exist
- Most popular LD visualization at the moment.

Questions?

Labs and Action Workshop

What happens next week