Clustering

CURE Algorithm

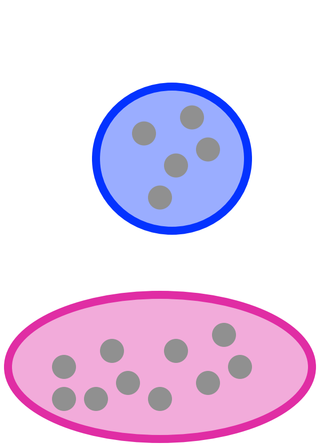
Mining of Massive Datasets

Leskovec, Rajaraman, and Ullman

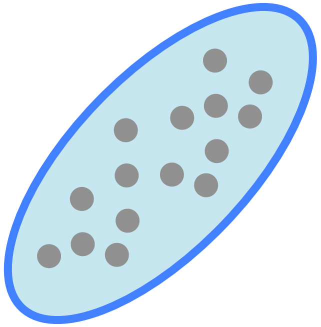
Stanford University

Limitations of BFR Algorithm

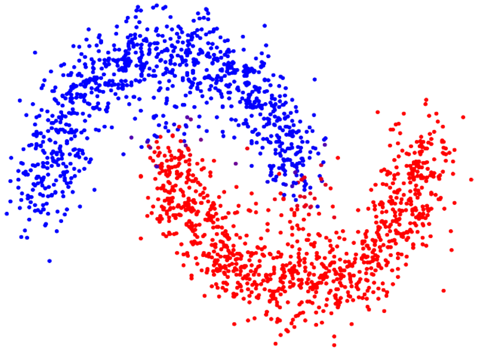
- Makes strong assumptions:
  (1) Clusters normally distributed in each dimension
  (2) Axes are fixed – ellipses at an angle are **not** OK



OK



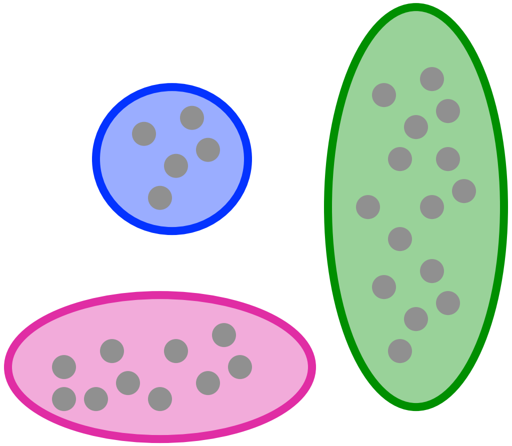
Not OK



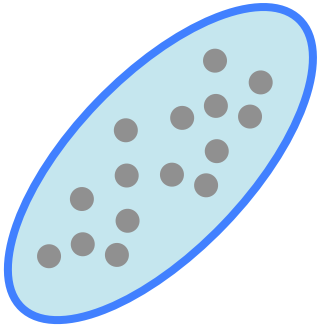
Not OK

CURE Algorithm

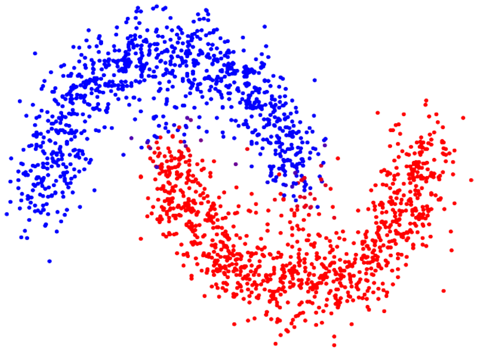
- **CURE (Clustering Using REpresentatives):**
  - Assumes a Euclidean distance
  - Allows clusters to assume any shape
  - **Uses a collection of representative points to represent clusters**



OK

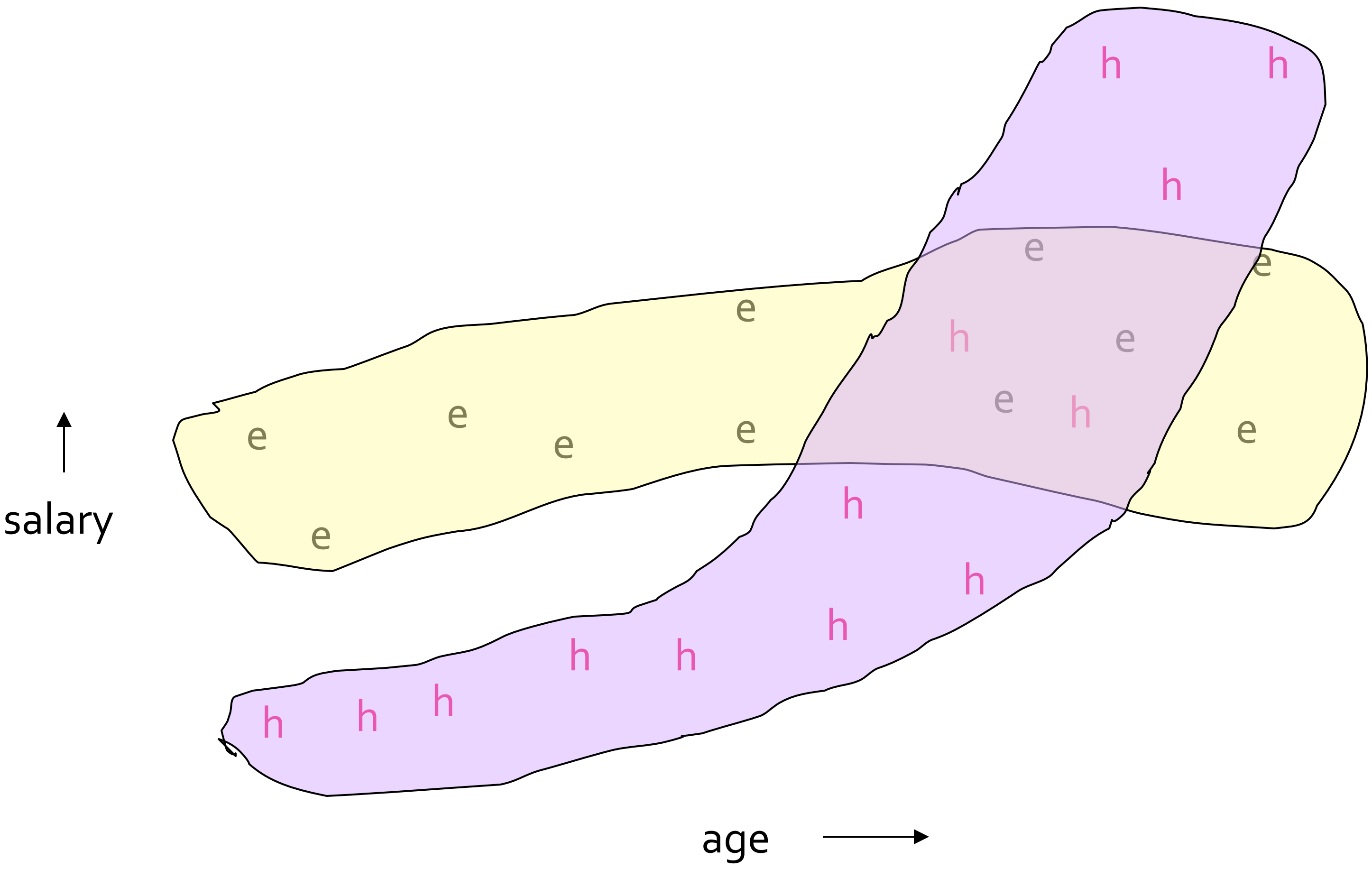


OK



OK

Example: Stanford Salaries

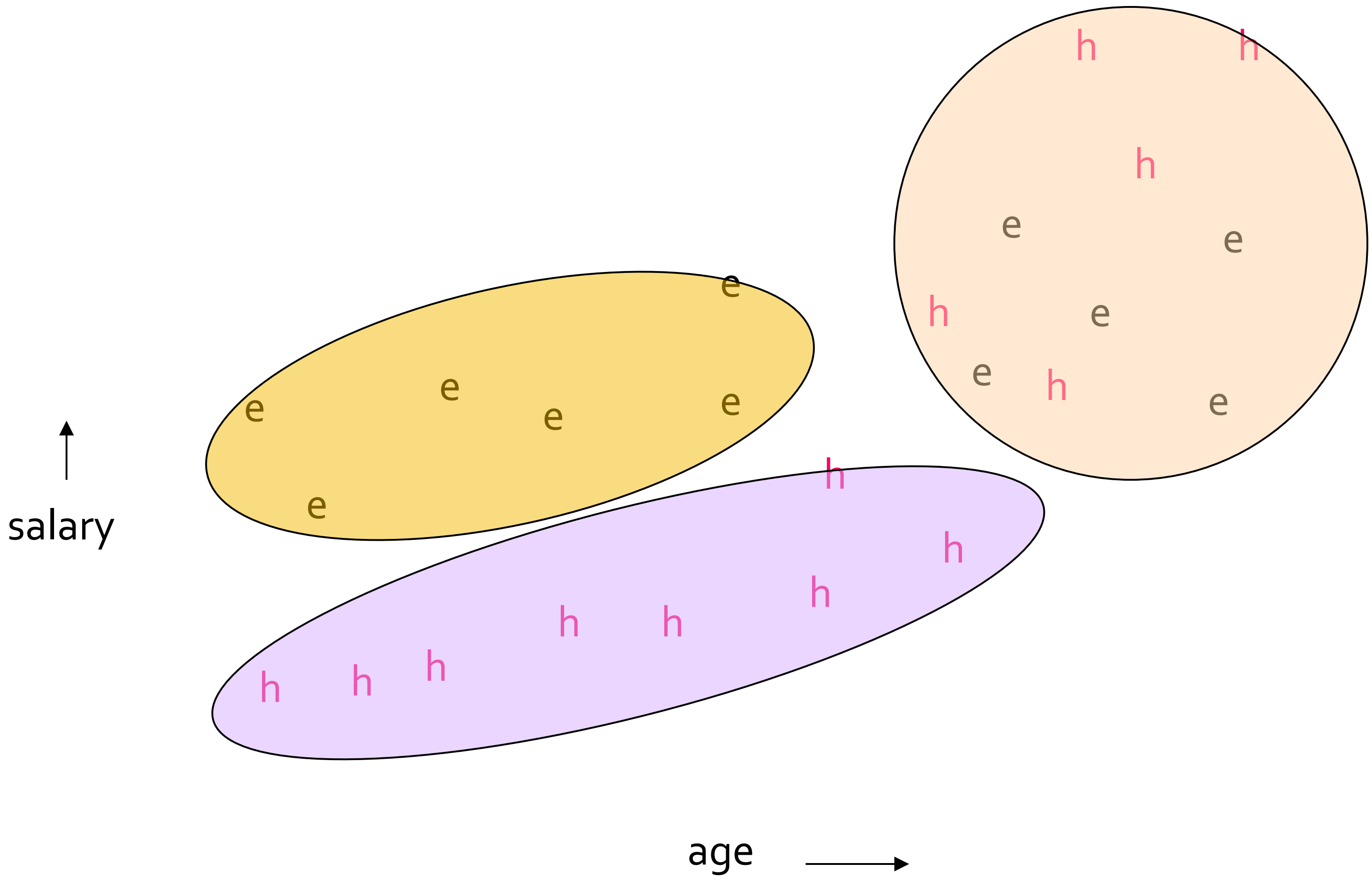


Starting CURE

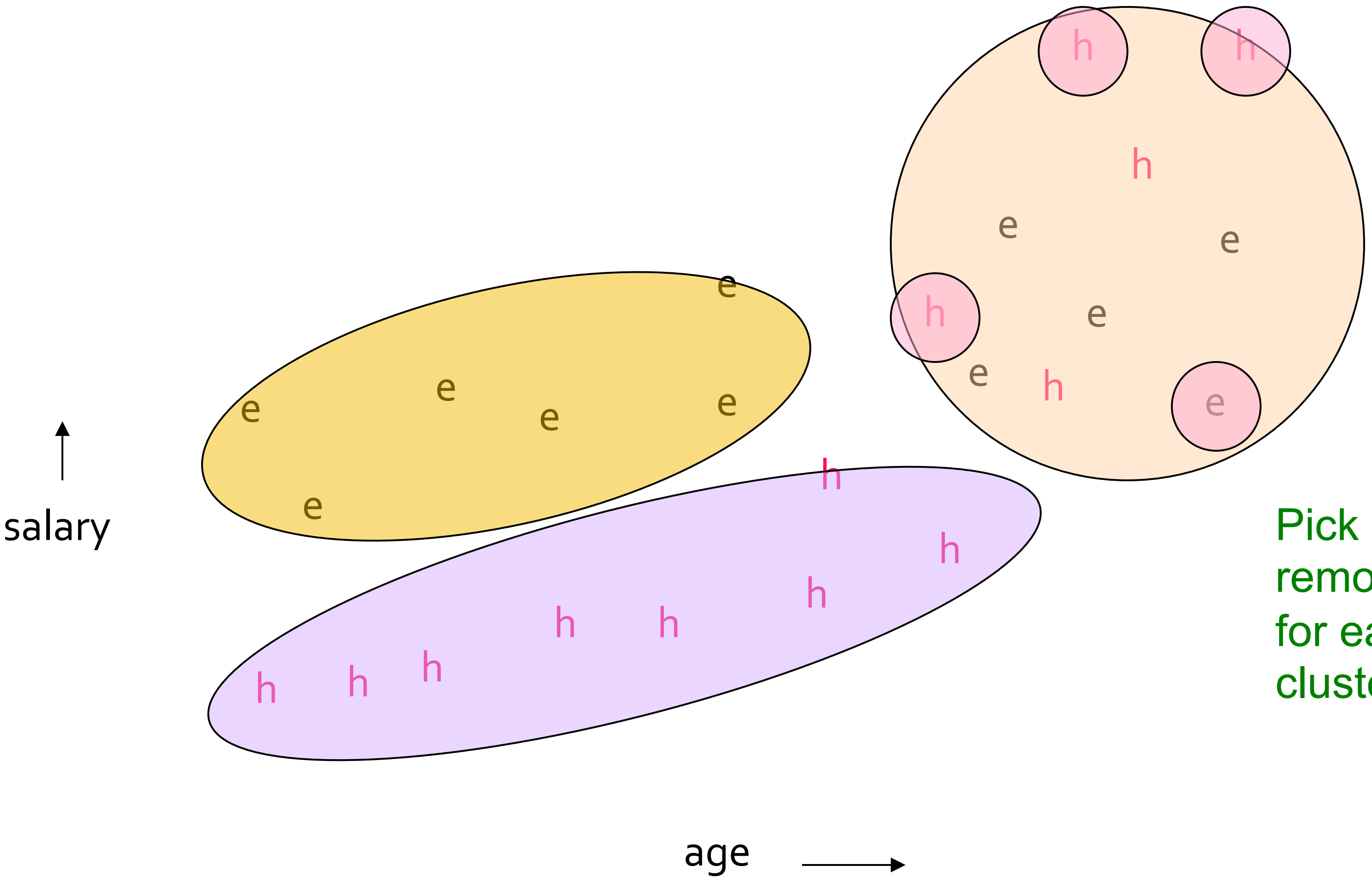
Pass 1 of 2:

- Pick a random sample of points that fit in main memory
- Cluster sample points hierarchically to create the initial clusters
- **Pick representative points:**
  - For each cluster, pick *k* (e.g., 4) representative points, as dispersed as possible
  - Move each representative point a fixed fraction (e.g., 20%) toward the centroid of the cluster

Example: Initial Clusters

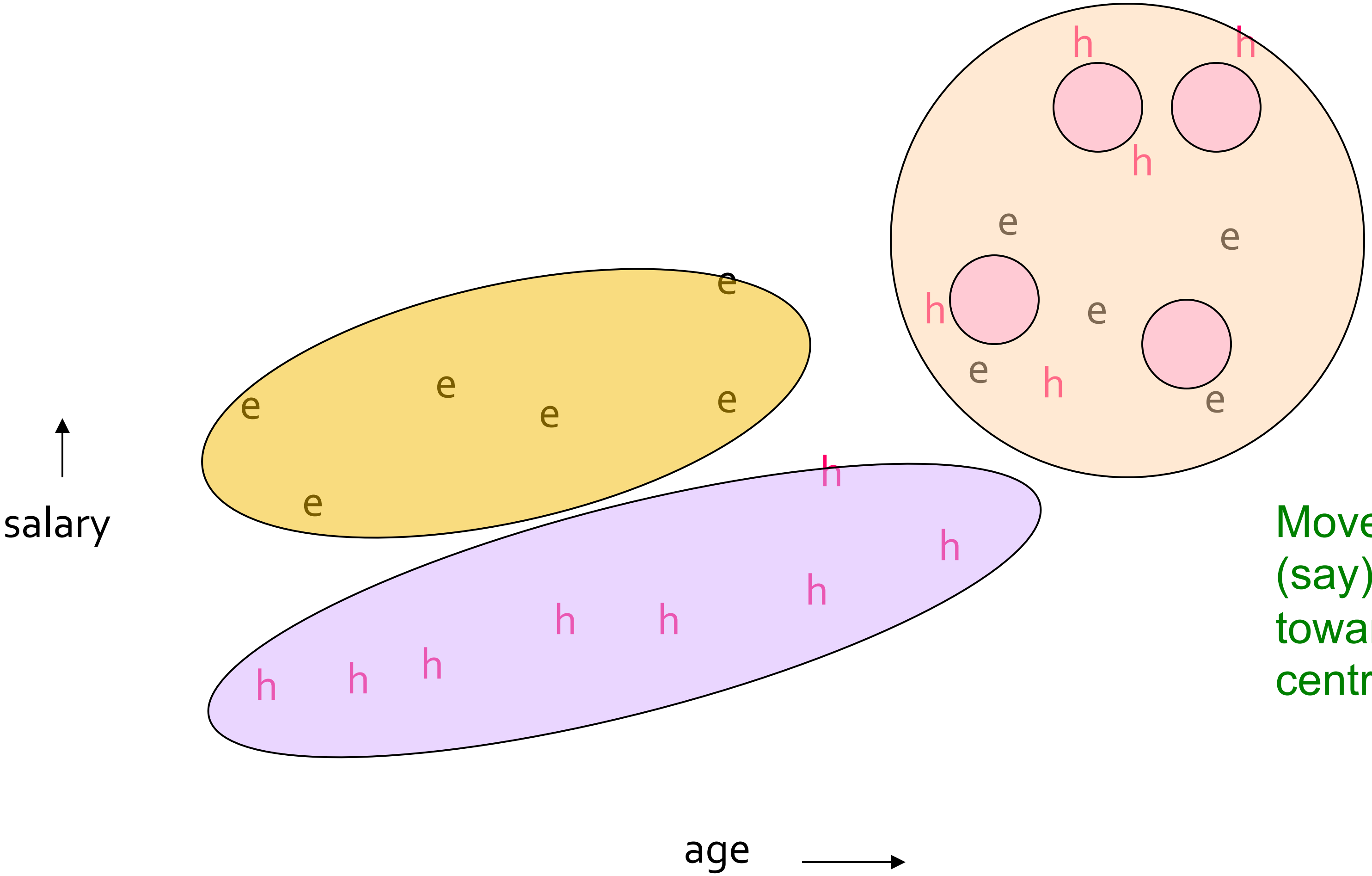


Example: Pick Dispersed Points



Pick (say) 4 remote points for each cluster.

Example: Pick Dispersed Points



Move points (say) 20% toward the centroid.

Finishing CURE

Pass 2 of 2:

- Now, rescan the whole dataset and visit each point ***p*** in the data set

- **Place it in the "closest cluster"**
  - Normal definition of "closest": that cluster with the closest (to *p*) among all the representative points of all the clusters

- And that's it!

Summary

- **Clustering:** Given a **set of points**, with a notion of **distance** between points, **group the points** into some number of ***clusters***
- **Algorithms:**
  - Agglomerative **hierarchical clustering**
    - Centroid and clustroid
  - ***k*-means**
  - **BFR**
  - **CURE**