



Clustering Algorithms

- Clustering algorithms have been designed to handle very large datasets
- E.g. the **Birch algorithm**
 - Main idea: use an in-memory R-tree to store points that are being clustered
 - Insert points one at a time into the R-tree, merging a new point with an existing cluster if it is less than some δ distance away
 - If there are more leaf nodes than fit in memory, merge existing clusters that are close to each other
 - At the end of first pass we get a large number of clusters at the leaves of the R-tree
 - Merge clusters to reduce the number of clusters