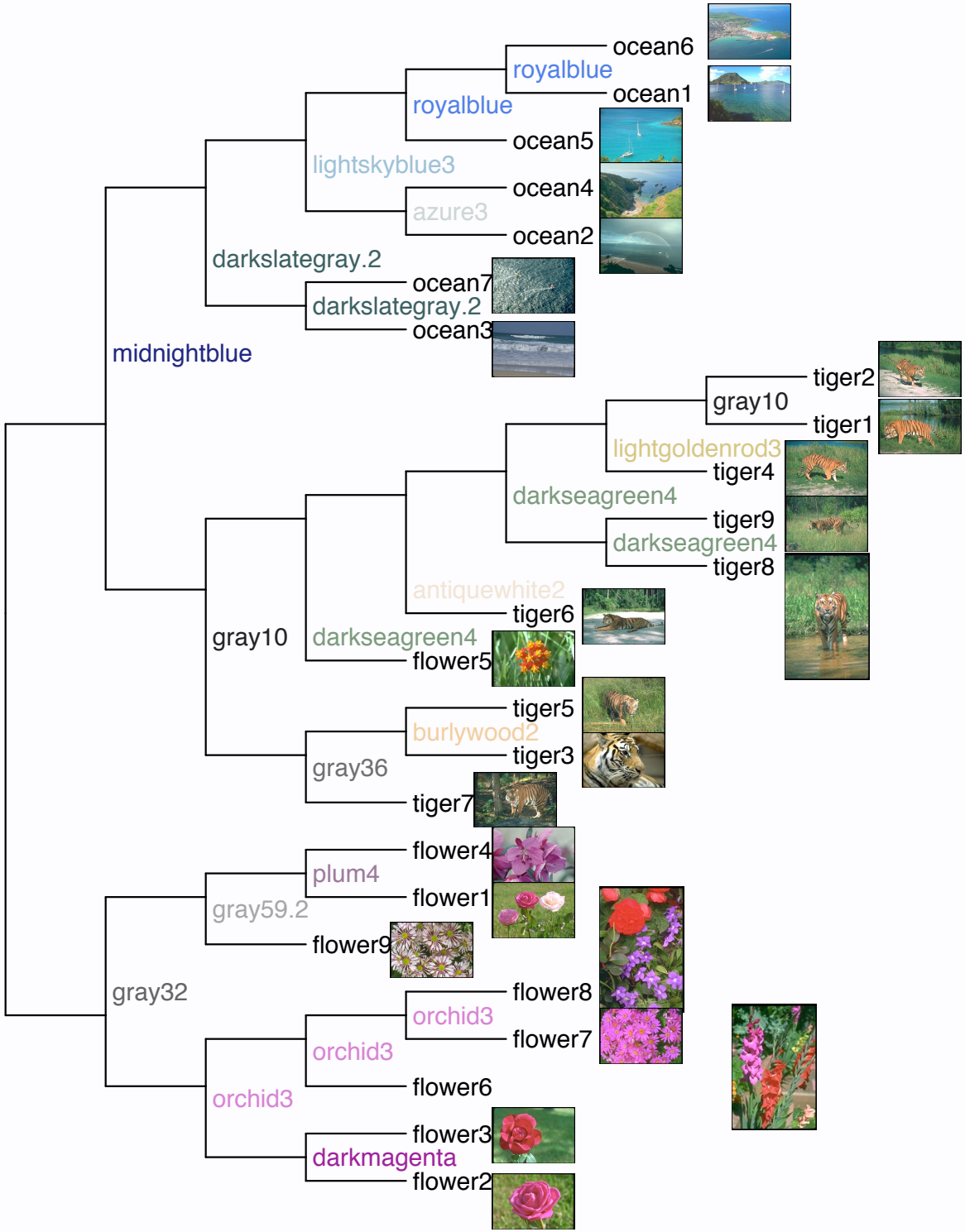
Yet More Clustering

36-350: Data Mining
19 September 2008

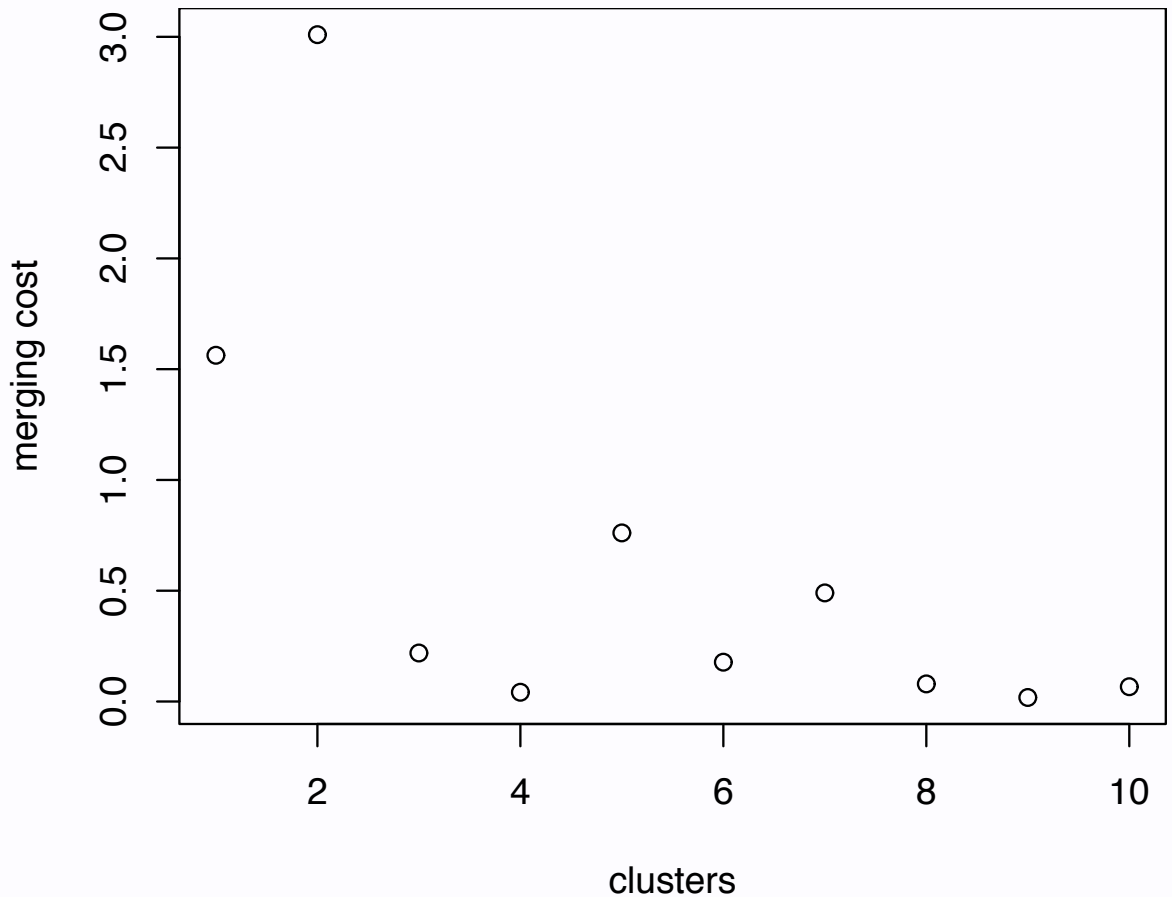
Last time...

- K-means: divide into k clusters to reduce within-cluser variance*cluster size
- Ward's method: start with each point in own cluster, cluster the clusters



Ward's method applied to the images from earlier: ocean, tigers, flowers

Jump in merging cost suggests 3 clusters - almost exactly right ones, too (but thinks flower5 is a tiger)



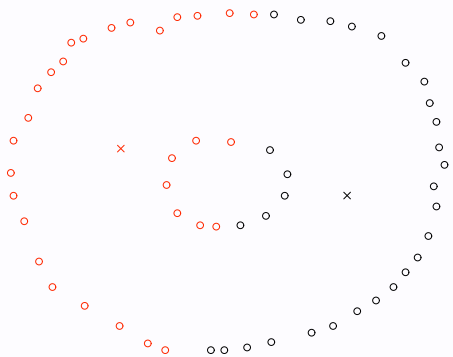
Merging cost vs. # of clusters

Rule of thumb: stop when merging costs go way up

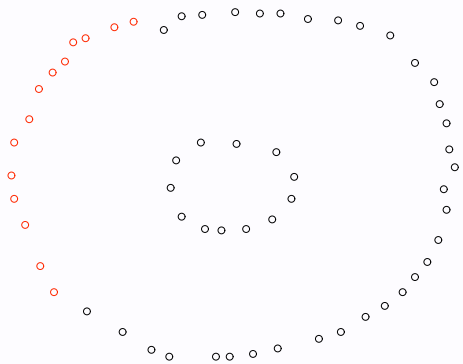
Here: 3 clusters (or 6 or 8...)

Minimizing the mean distance from the cluster center tends to make spheres, which can be silly

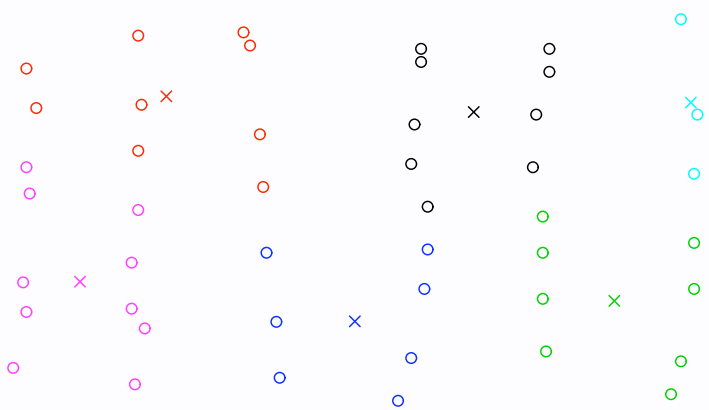
k-Means

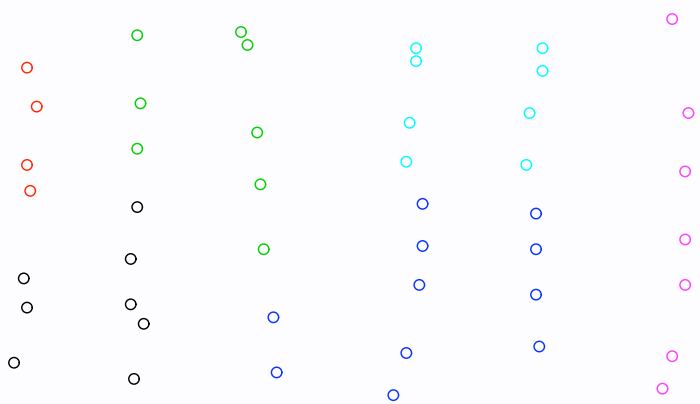


Ward's



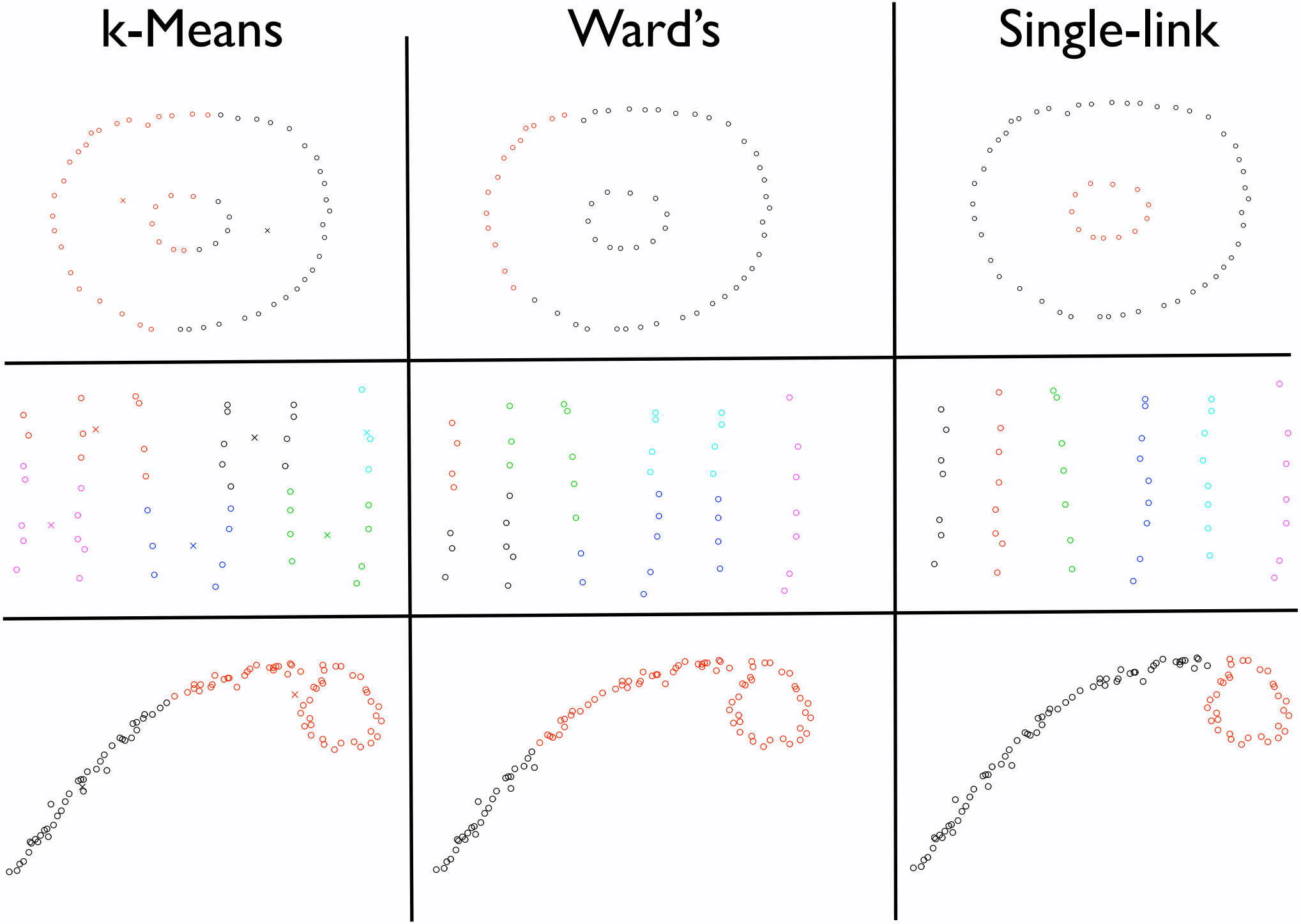
note how Ward's is less balanced



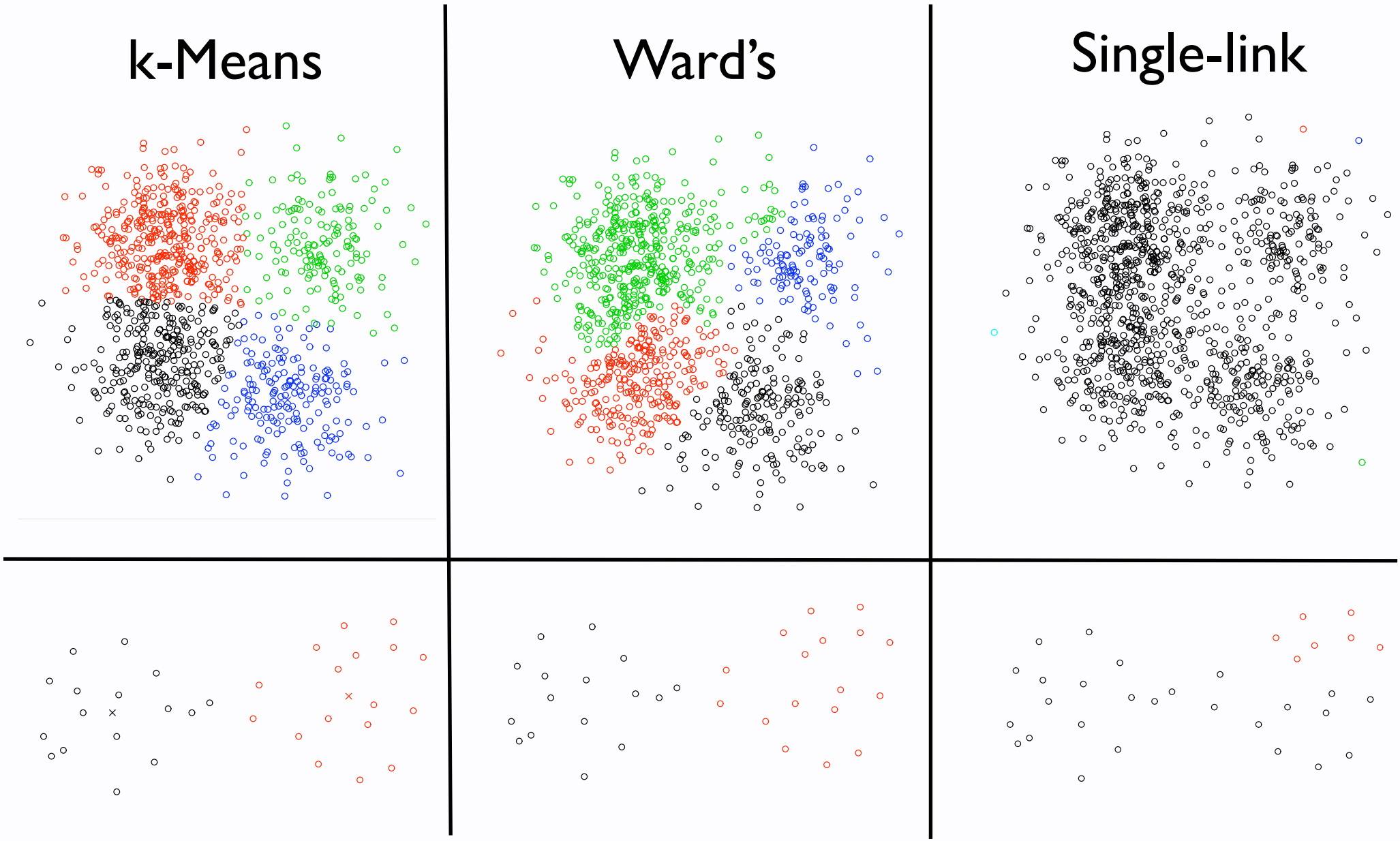


Single-link clustering

1. Start with every point in its own cluster
2. Calculate gaps between every pair of clusters = distance between 2 closest points in each cluster
3. Merge clusters with smallest gap


k-Means
Ward's
Single-link

Examples where single-link doesn't work so well



How many clusters?

- Can always improve sum-of-squares by adding more clusters
- Can generally improve *any* criterion by adding more clusters
- It seems silly to say that each point is in its own cluster

Heuristics

- Merging cost: if reducing the number of clusters gives a big hit in performance, stop
  - Why?
  - What's a big hit?
- Add a cost-per-cluster
  - Why that cost?

Missing from the heuristics

- Clusters are good if the data really do fall into different categories with different characteristics; if not, not
- Summarizing the training data isn't what we want!

Real criteria

- External validity: Does knowing the cluster predict variables *other than* the ones used to determine cluster membership?
  - If so, is it really the cluster, or one of those variables?
- Generalization to new data

Generalization

- The model *generalizes* if it performs about as well on new data from the same source as it did on the training data
- This notion really only applies to predictive models
- Clustering tools we've seen hardly give us predictions

Faking prediction

- Look at sum of squares with old cluster centers on new data
- Look at tree structure on new data: leaves will be different, but how much of the tree shape changes? Does the merging-cost graph look the same?
- How much do things change with a little new data added in to the old?

Cross-validation

- Randomly divide the data into training and testing sets, say 90/10
- Fit the model on the training set and evaluate on the testing set
- Repeat several times, say 10
- Average the results

Reification

- Treating your idea as an independently-existing thing (*res*)
- Sometimes a good idea (bacteria), sometimes not (zodiac sign)
- Overwhelming temptation with clustering
  - especially once you add names

To play with

- Go to http://yawyl.claritas.com/ and figure out what they are doing
- Should you believe it?