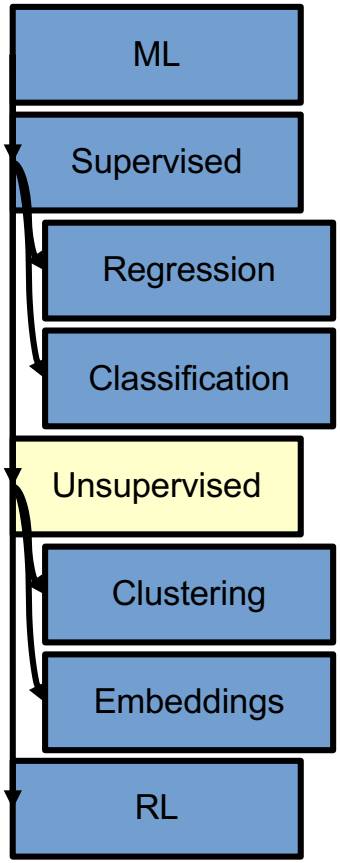
CS181: Introduction to Machine Learning

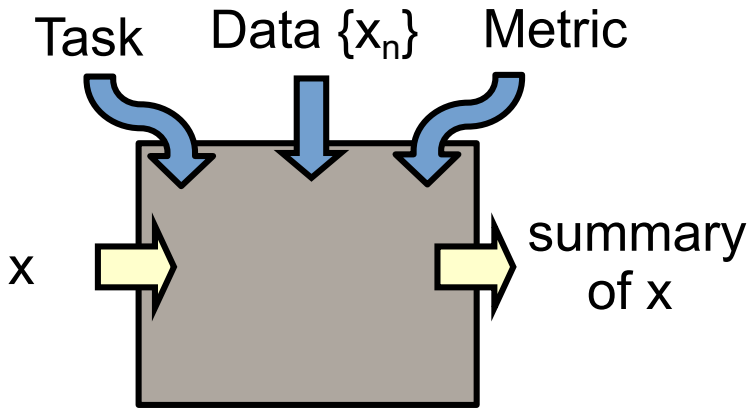
Lecture 13 (Clustering)

Spring 2021

Finale Doshi-Velez and David C. Parkes
Harvard Computer Science

Unsupervised Learning

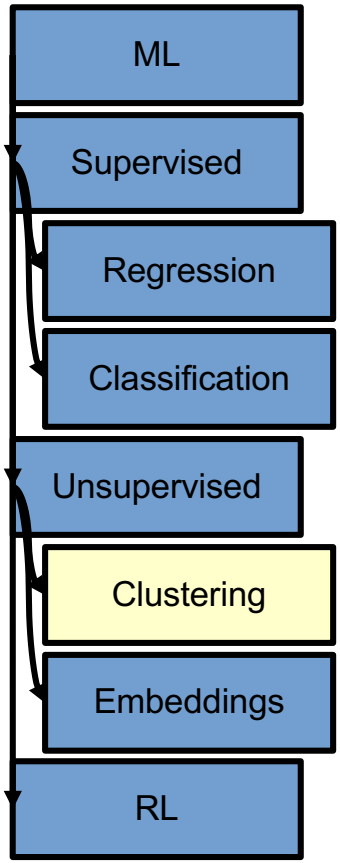


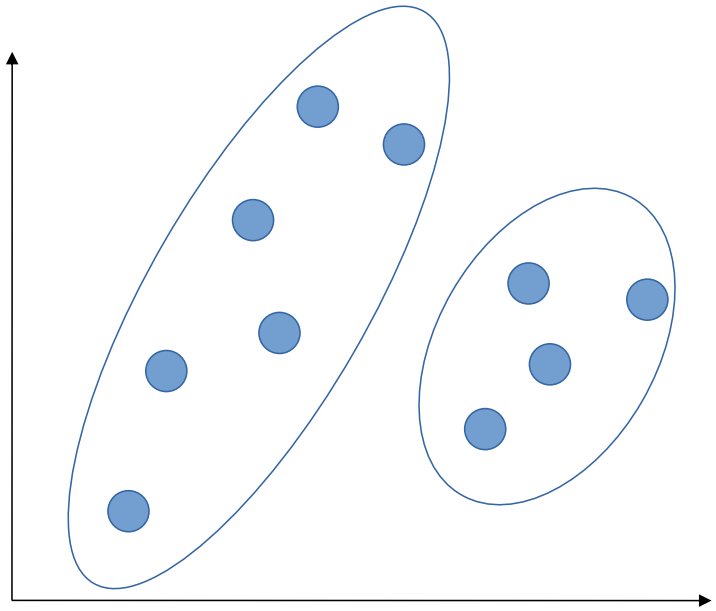


Data D = {x_1, x_2, … x_N}

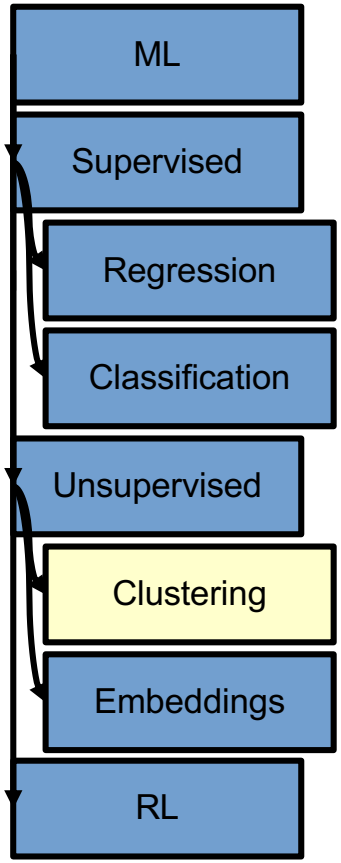
Typical goals: understand, summarize, identify concepts

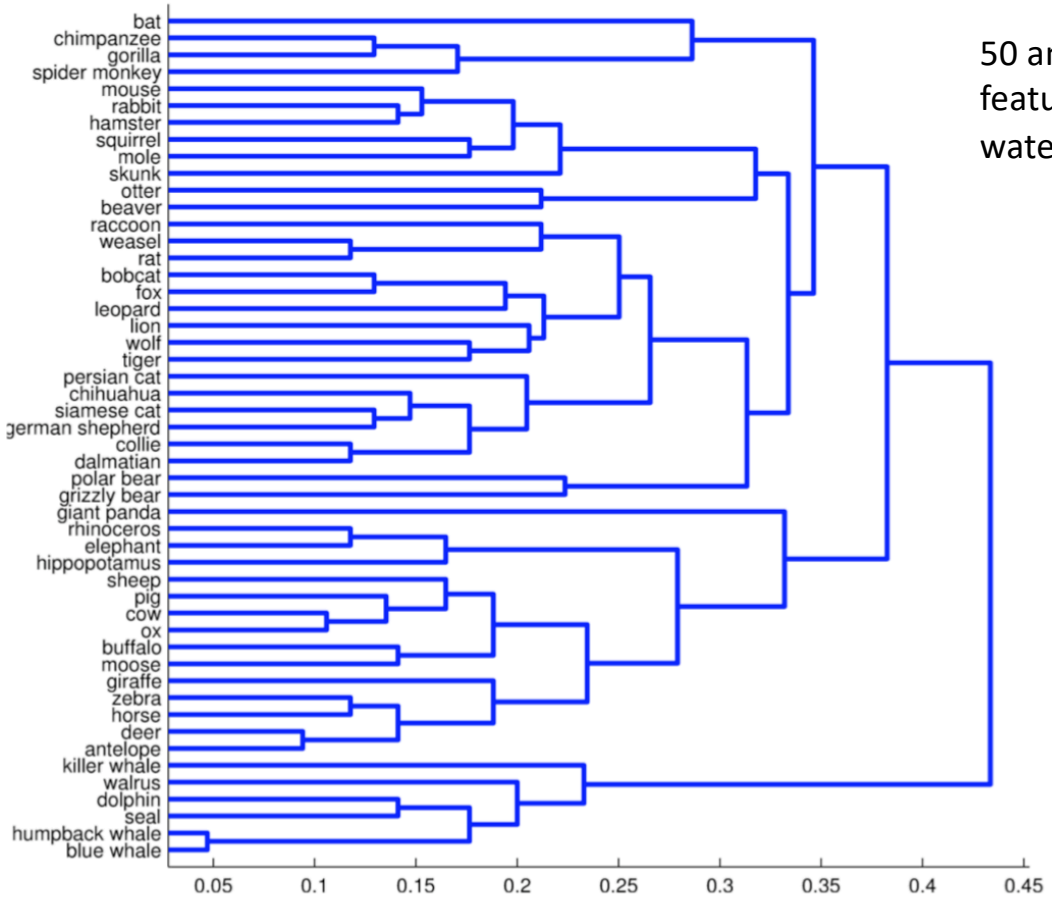
Unsupervised Learning: Clustering





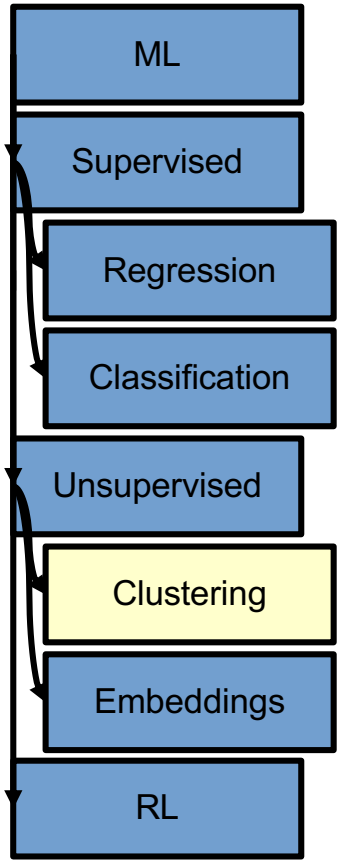
Unsupervised Learning: Clustering

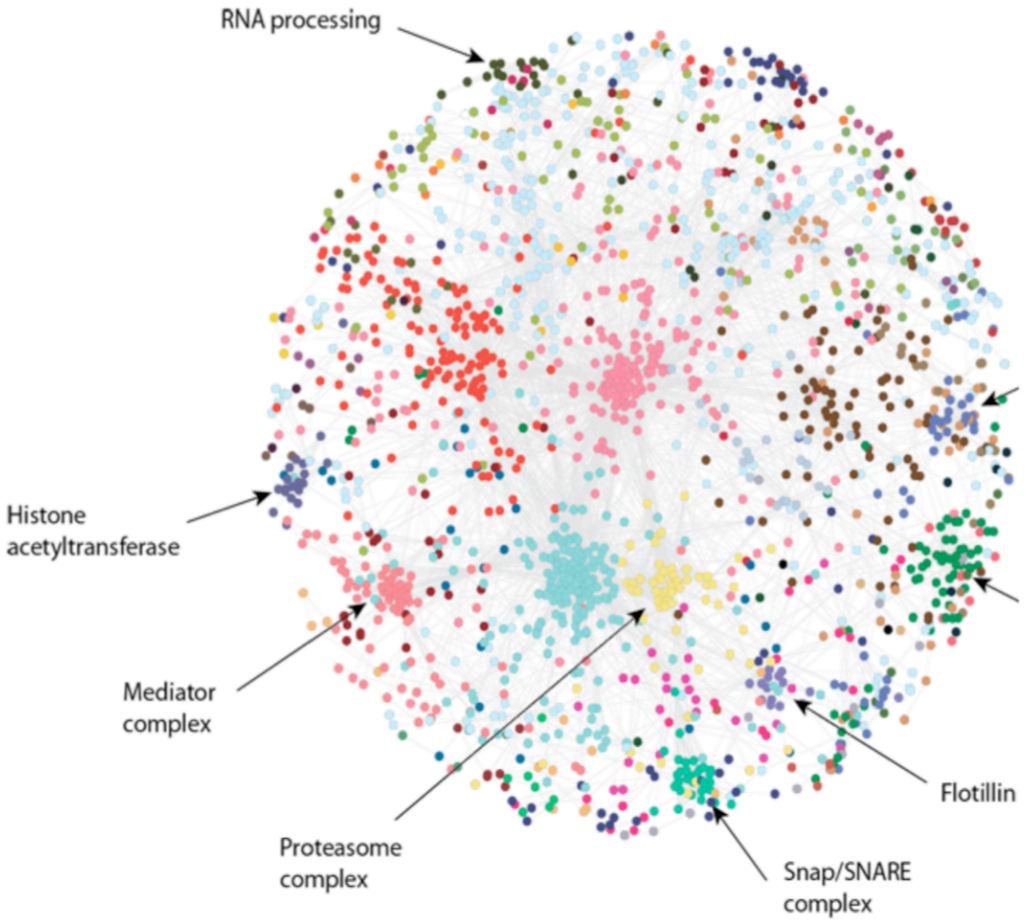




50 animals, 85 binary features (e.g., long neck, water, smelly)

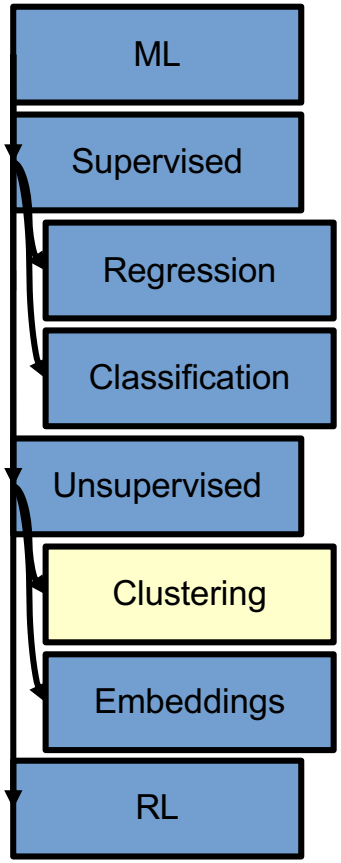
Unsupervised Learning: Clustering

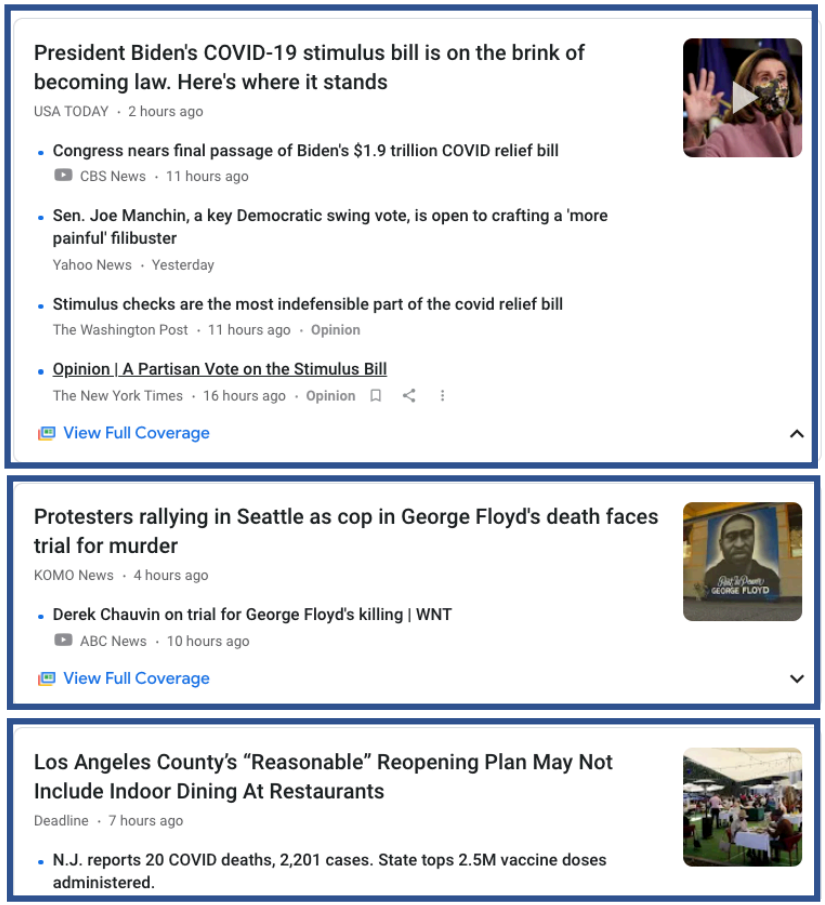




Understanding gene regulation

Unsupervised Learning: Clustering

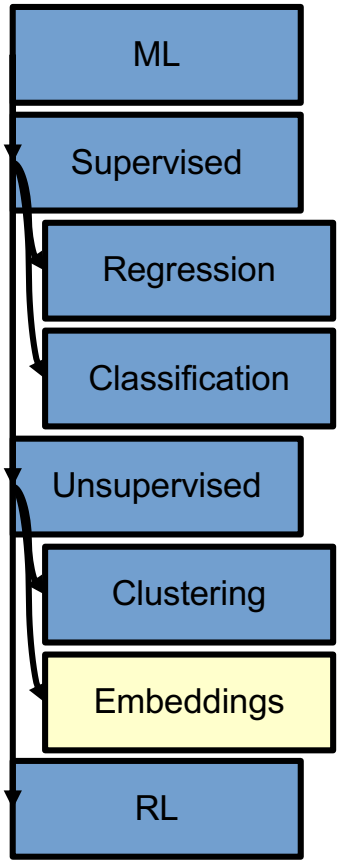


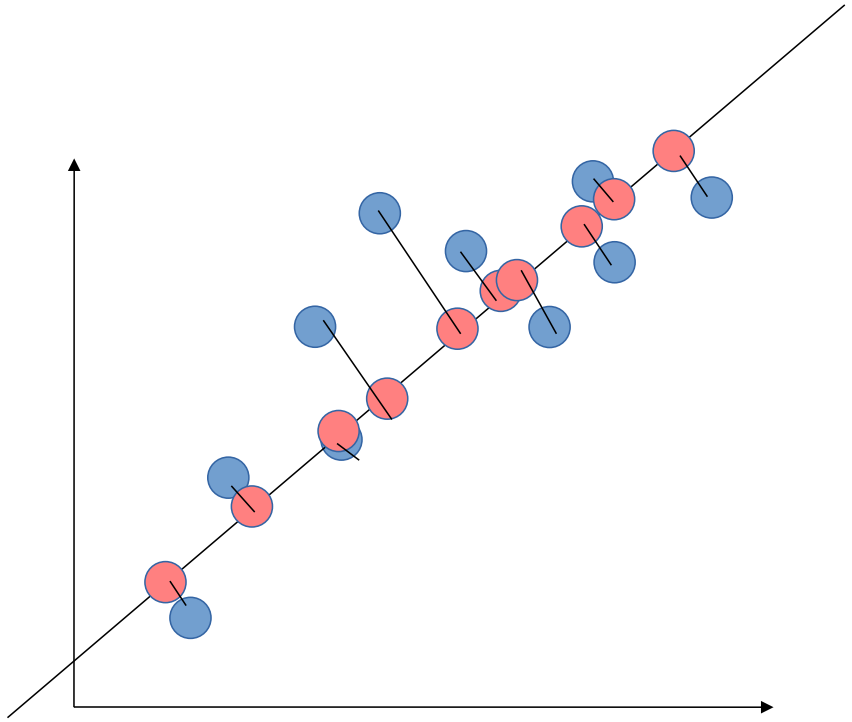


https://news.google.com/topstories?hl=en-US&gl=US&ceid=US:en

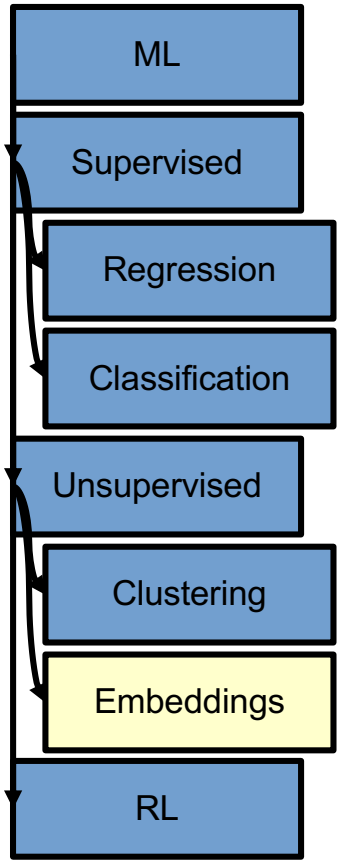
Organizing news stories

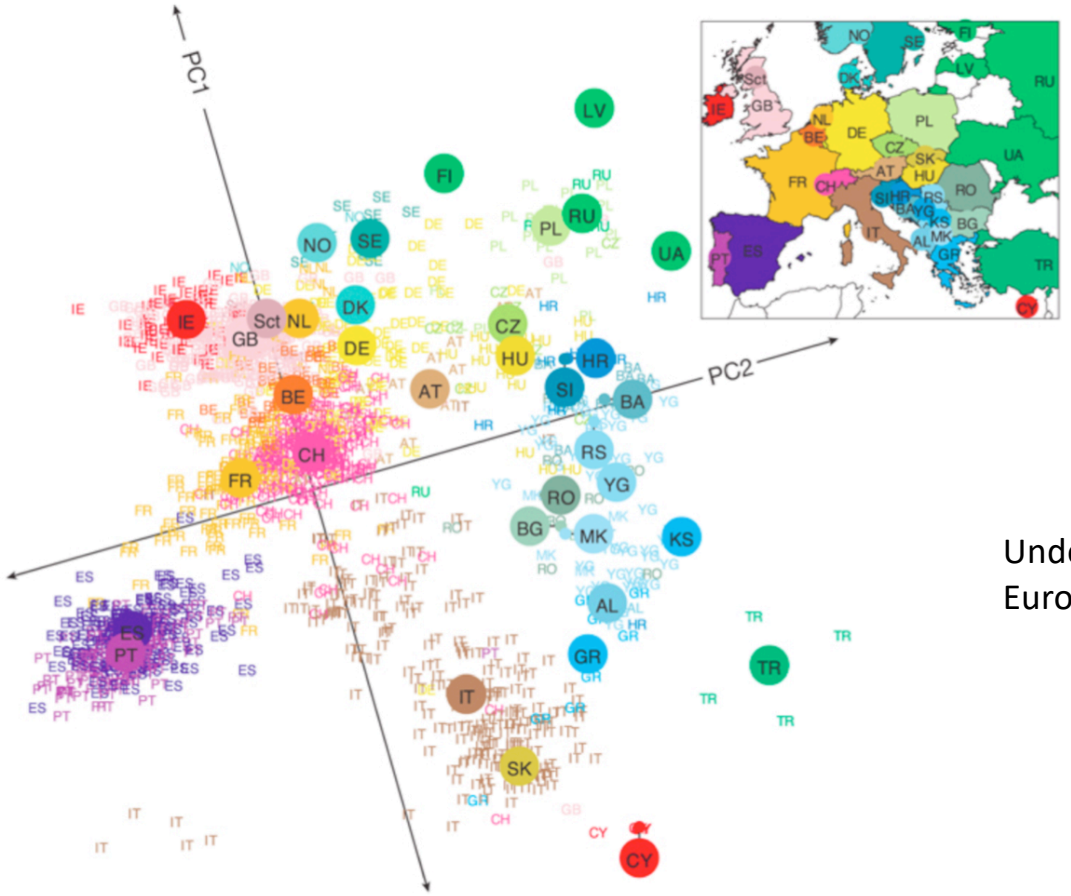
Unsupervised Learning: Embedding





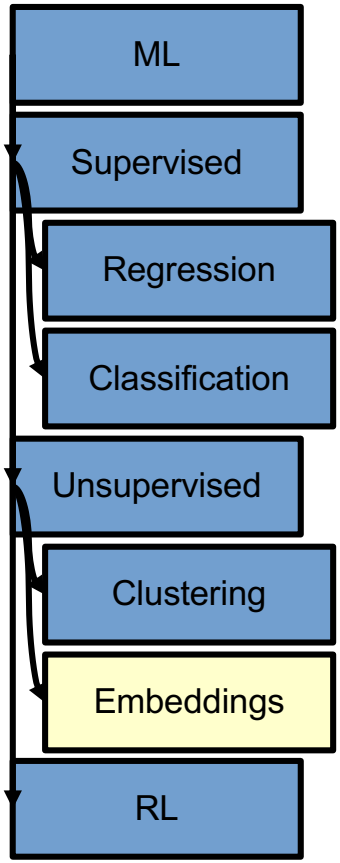
Unsupervised Learning: Embedding

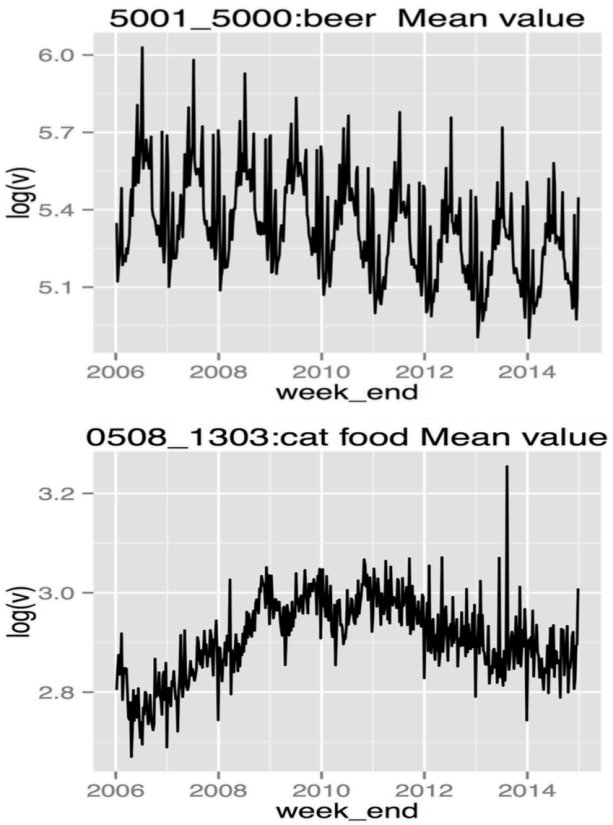




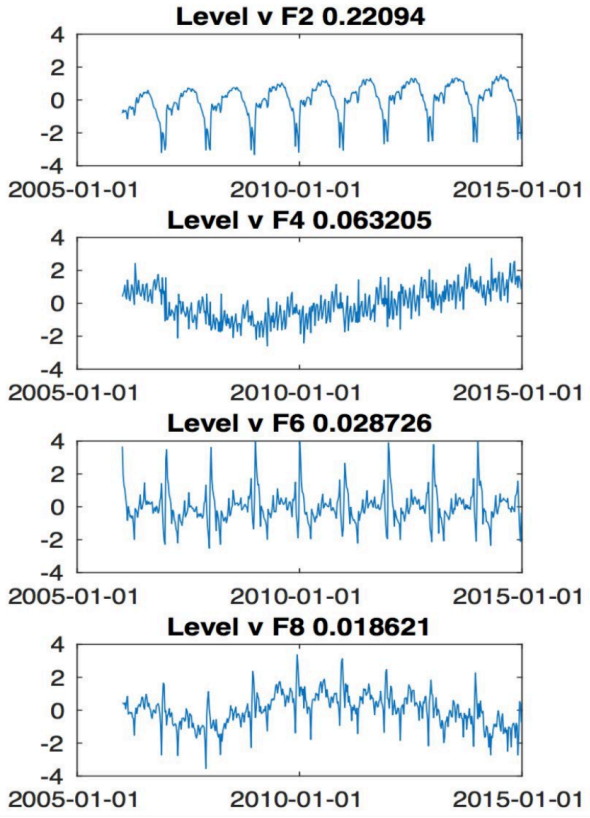
Understanding the European gene pool

Unsupervised Learning: Embedding



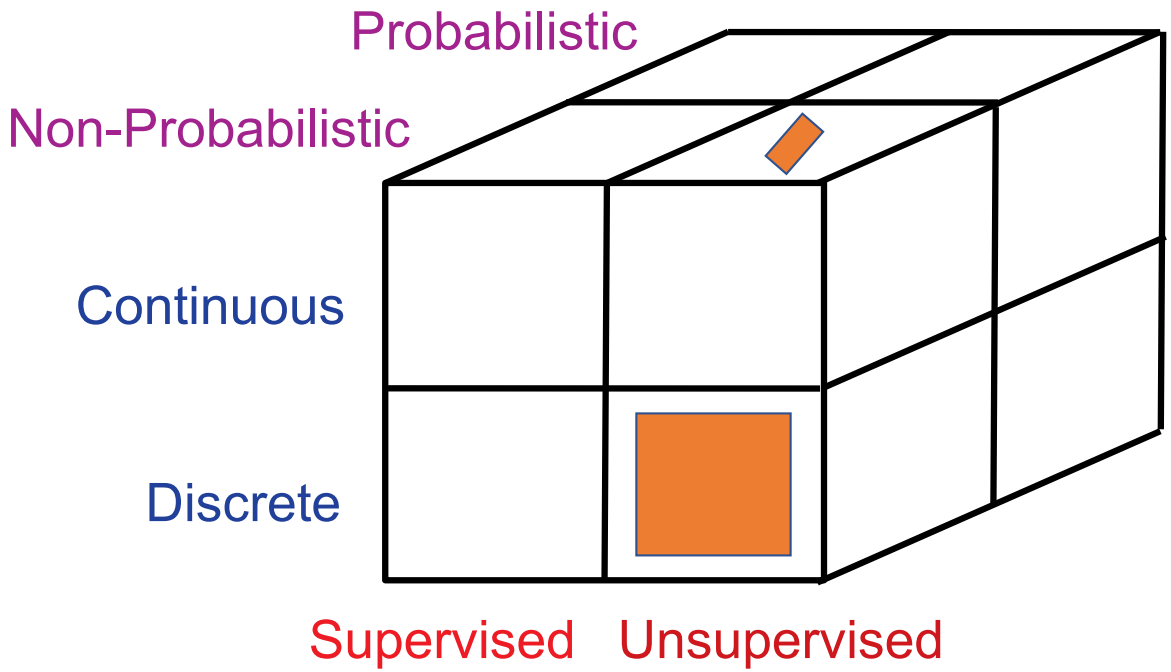


(S. Ng 2016)



Nielsen scanner data, 1000 products, 8 years

Through PCA we see the effect of the 08-09 financial crisis on demand

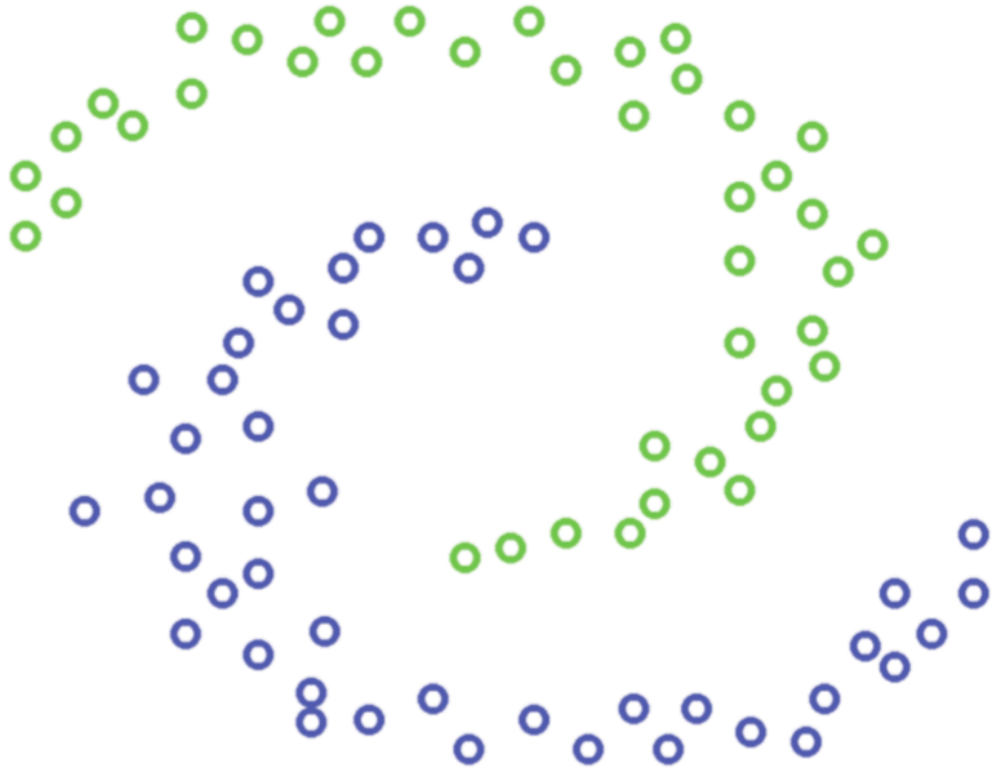


+ graphical models, reinforcement learning

Today: Clustering

How Would you Cluster these Points?

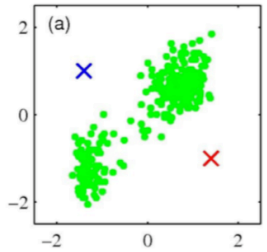
How Would you Cluster these Points?



K-Means on “Old Faithful” Geyser Eruptions

- Yellowstone National Park, Wyoming
- 272 data points
- Features
  - Duration of current eruption (x_{n2})
  - Duration of next eruption (x_{n1})
  - Standardized: $x_{nj} = (x_{nj}-\boldsymbol{\mu}_j)/\boldsymbol{\sigma}_j$, where $\boldsymbol{\mu}_j$ is mean of feature j, $\boldsymbol{\sigma}_j$ is standard deviation of feature j

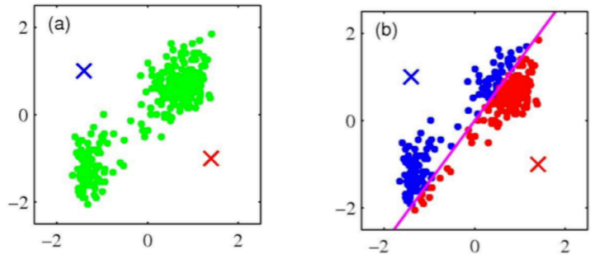
K-Means on “Old Faithful” Geyser Eruptions



- Duration of next eruption (x_1)
- Duration of current eruption (x_2)

(Bishop)

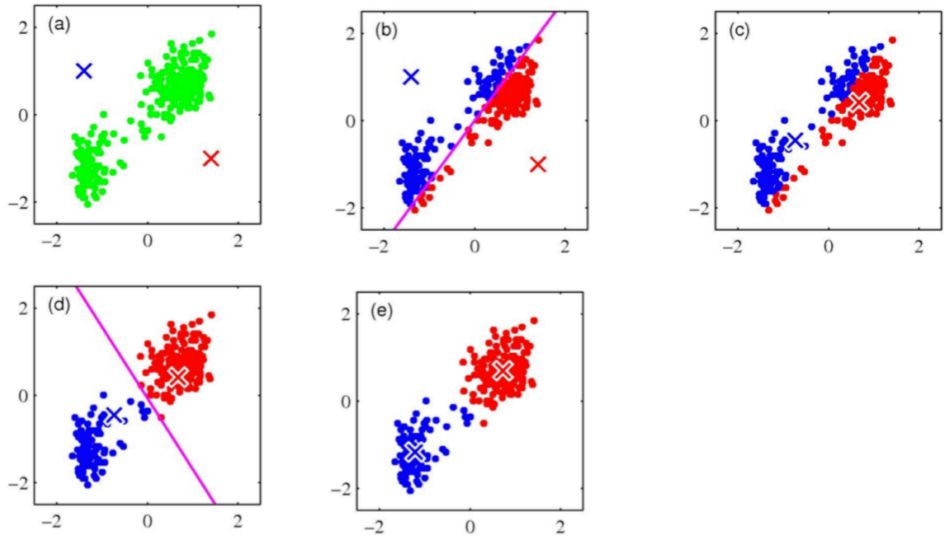
K-Means on “Old Faithful” Geyser Eruptions



- Duration of next eruption (x_1)
- Duration of current eruption (x_2)

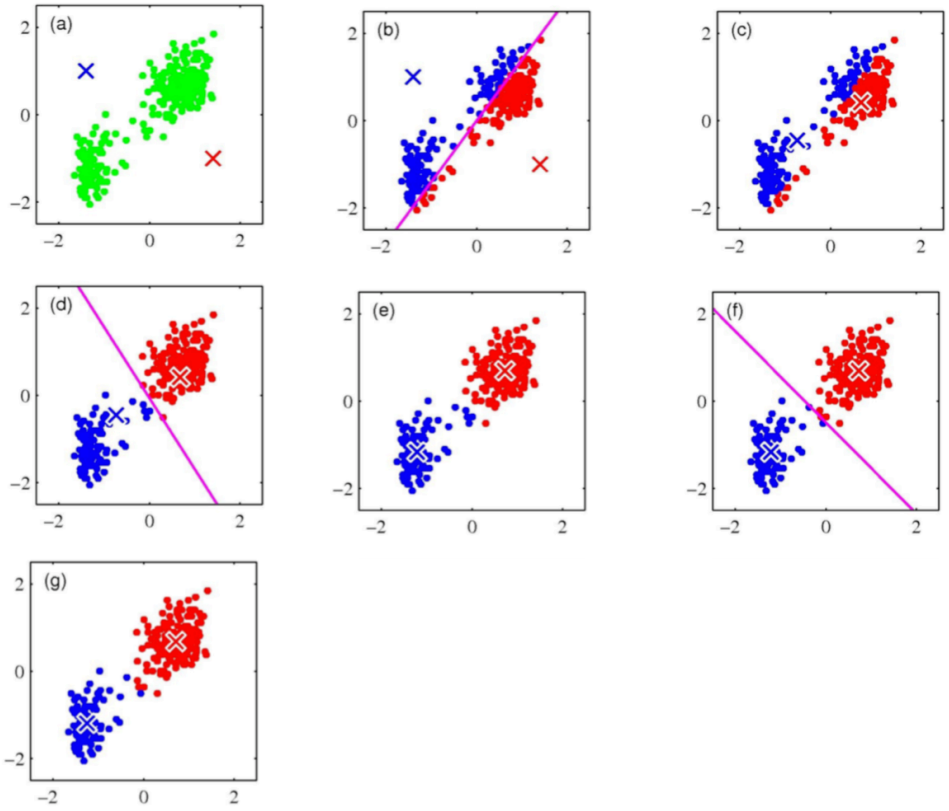
(Bishop)

K-Means on “Old Faithful” Geyser Eruptions



- Duration of next eruption (x_1)
- Duration of current eruption (x_2)

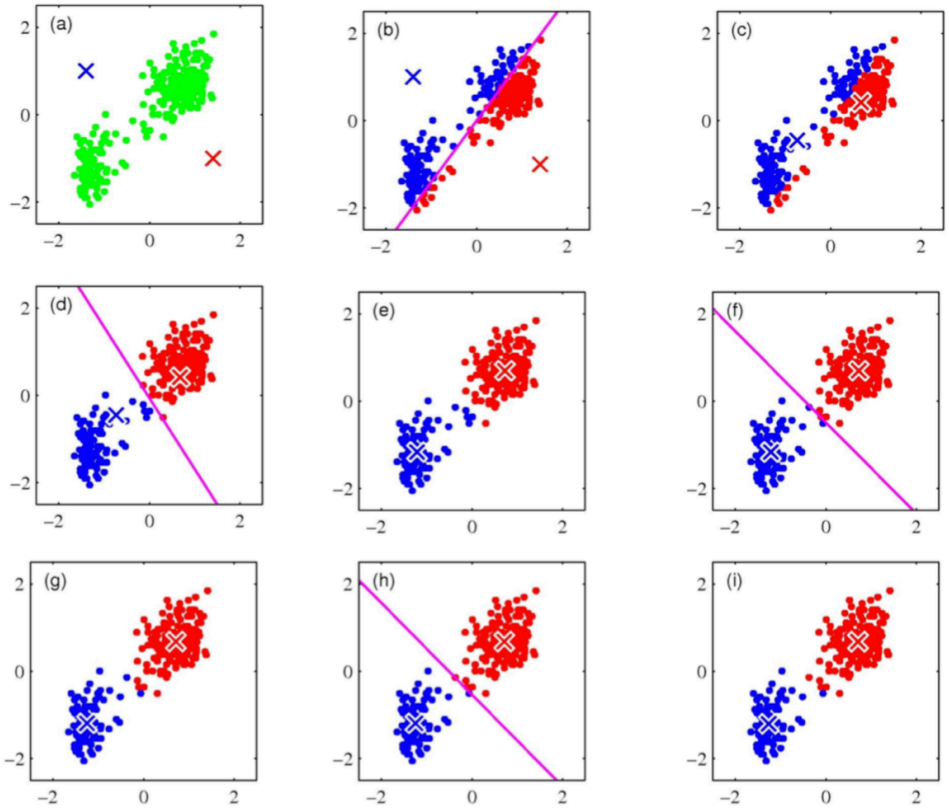
K-Means on “Old Faithful” Geyser Eruptions



- Duration of next eruption (x_1)
- Duration of current eruption (x_2)

(Bishop)

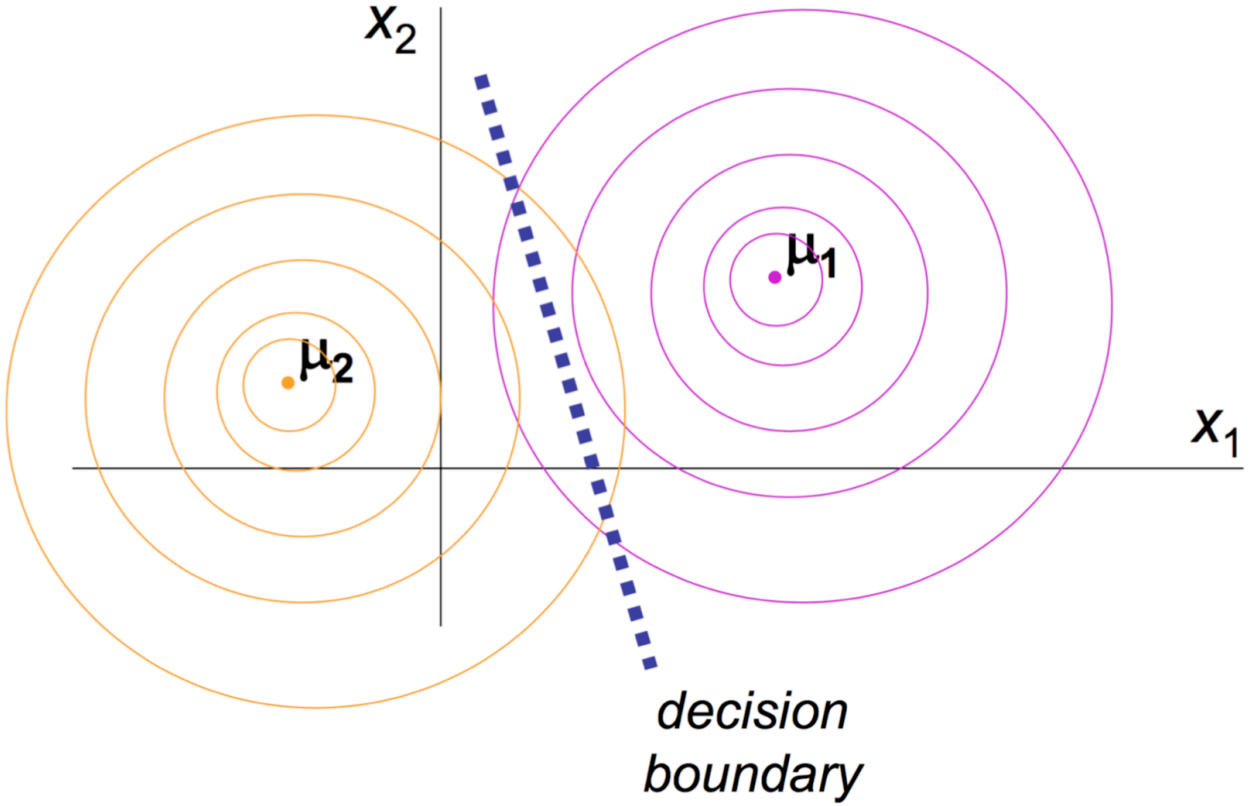
K-Means on “Old Faithful” Geyser Eruptions



- Duration of next eruption (x_1)
- Duration of current eruption (x_2)

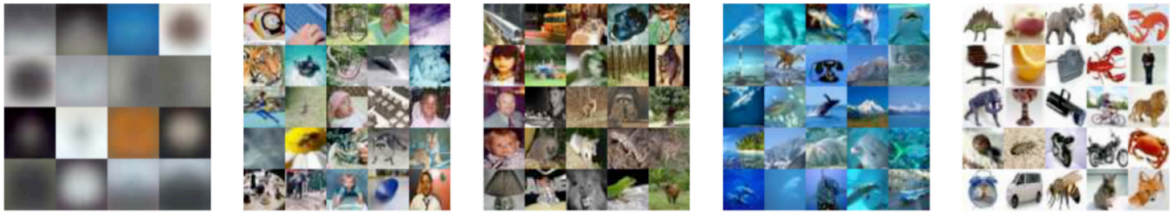
(Bishop)

K-Means: Linear Decision Boundaries

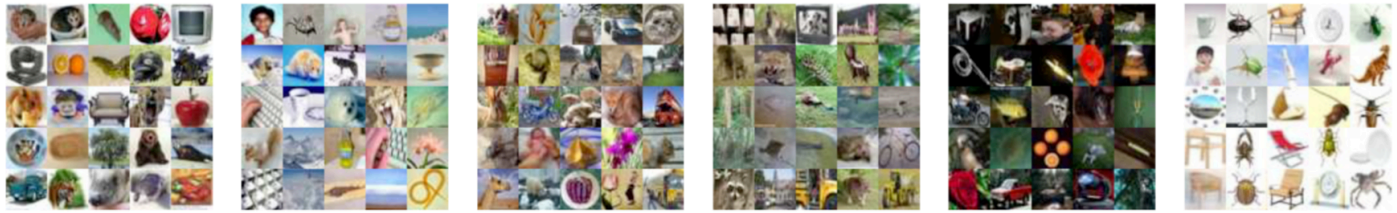


(Bishop)

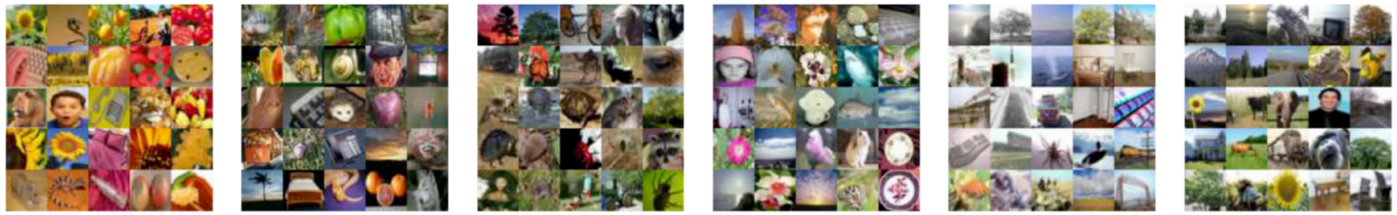
K-Means: CIFAR-100



(a) Cluster Centers (b) Cluster 1 (c) Cluster 2 (d) Cluster 3 (e) Cluster 4



(f) Cluster 5 (g) Cluster 6 (h) Cluster 7 (i) Cluster 8 (j) Cluster 9 (k) Cluster 10



(l) Cluster 11 (m) Cluster 12 (n) Cluster 13 (o) Cluster 14 (p) Cluster 15 (q) Cluster 16

50,000 images
32x32x3 (RGB)
K=16
Clusters pick up on low frequency color variations

(Adams)

K-Medoids: CIFAR-100

(a) Cluster Centers (b) Cluster 1 (c) Cluster 2 (d) Cluster 3 (e) Cluster 4

(f) Cluster 5 (g) Cluster 6 (h) Cluster 7 (i) Cluster 8 (j) Cluster 9 (k) Cluster 10

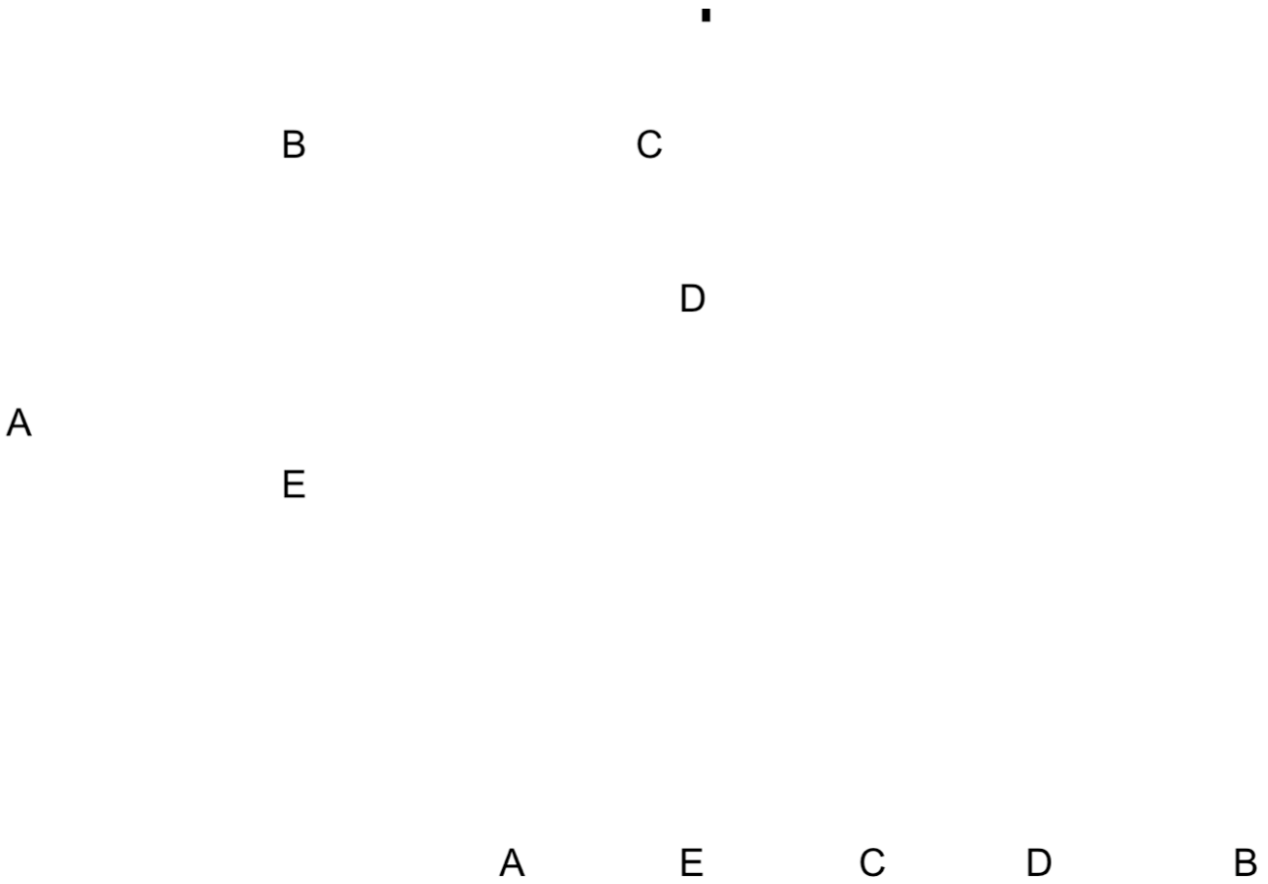
(l) Cluster 11 (m) Cluster 12 (n) Cluster 13 (o) Cluster 14 (p) Cluster 15 (q) Cluster 16

50,000 images
32x32x3 (RGB)
K=16

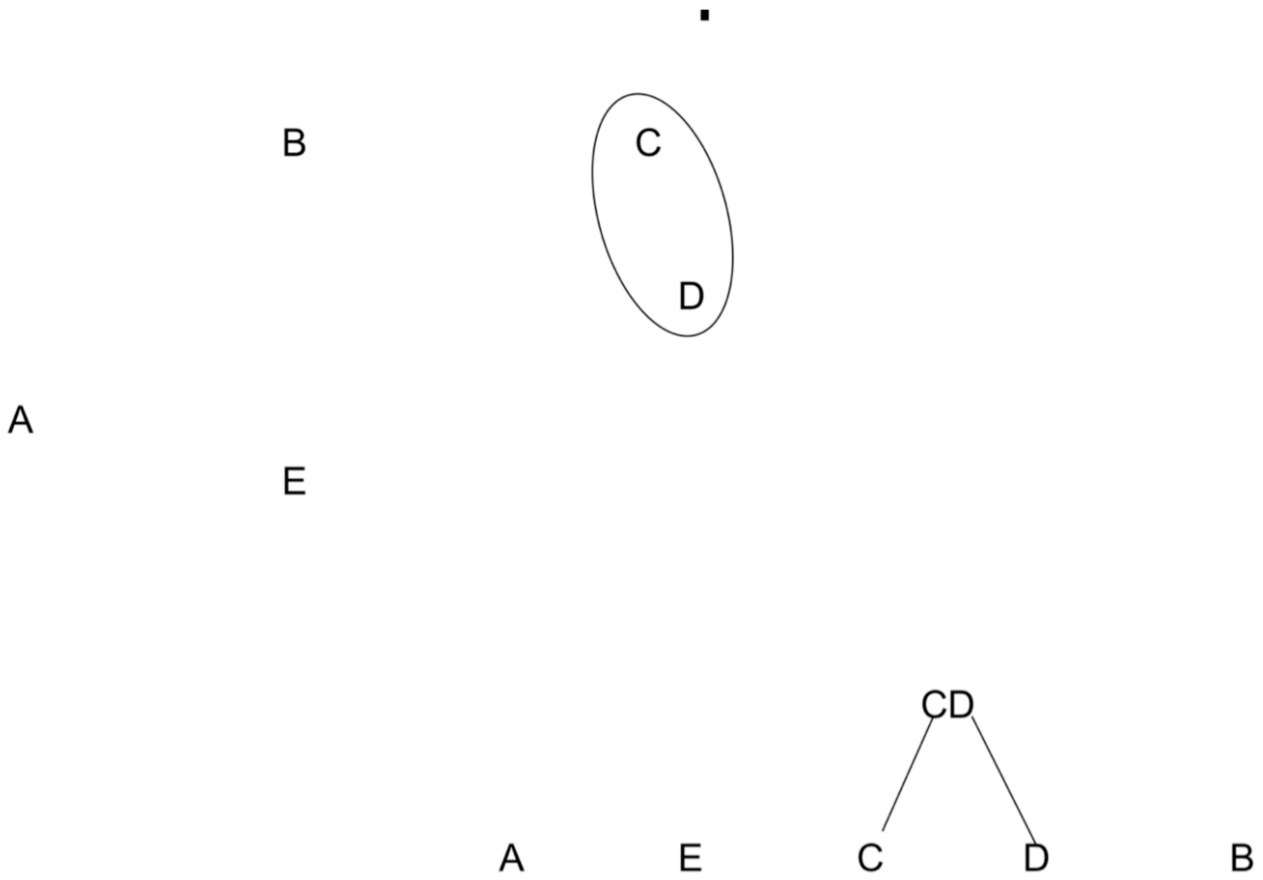
Now use actual examples as prototypes

(Adams)

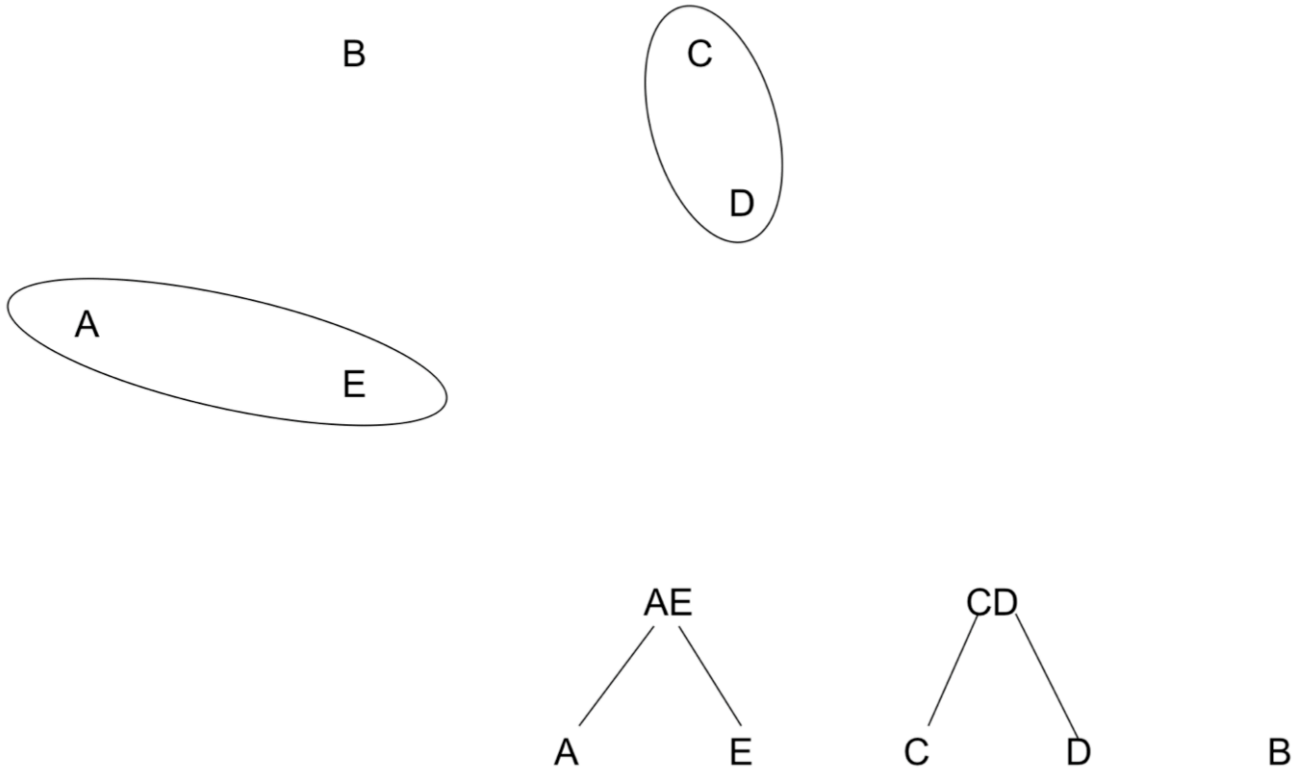
HAC (1 of 5)



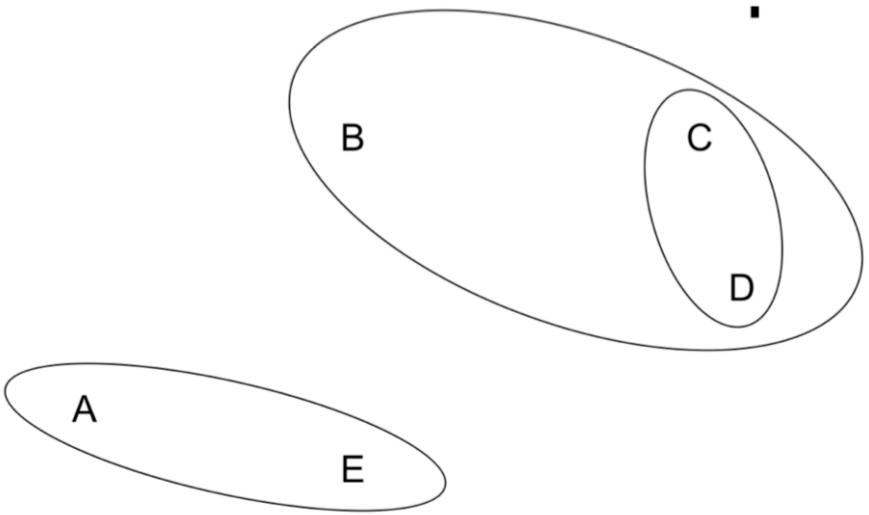
HAC (2 of 5)

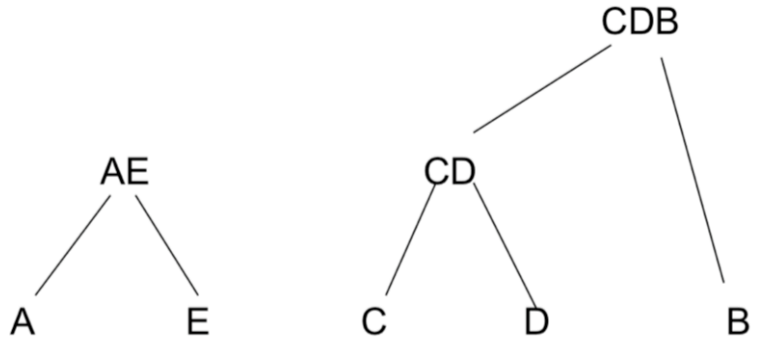


HAC (3 of 5)

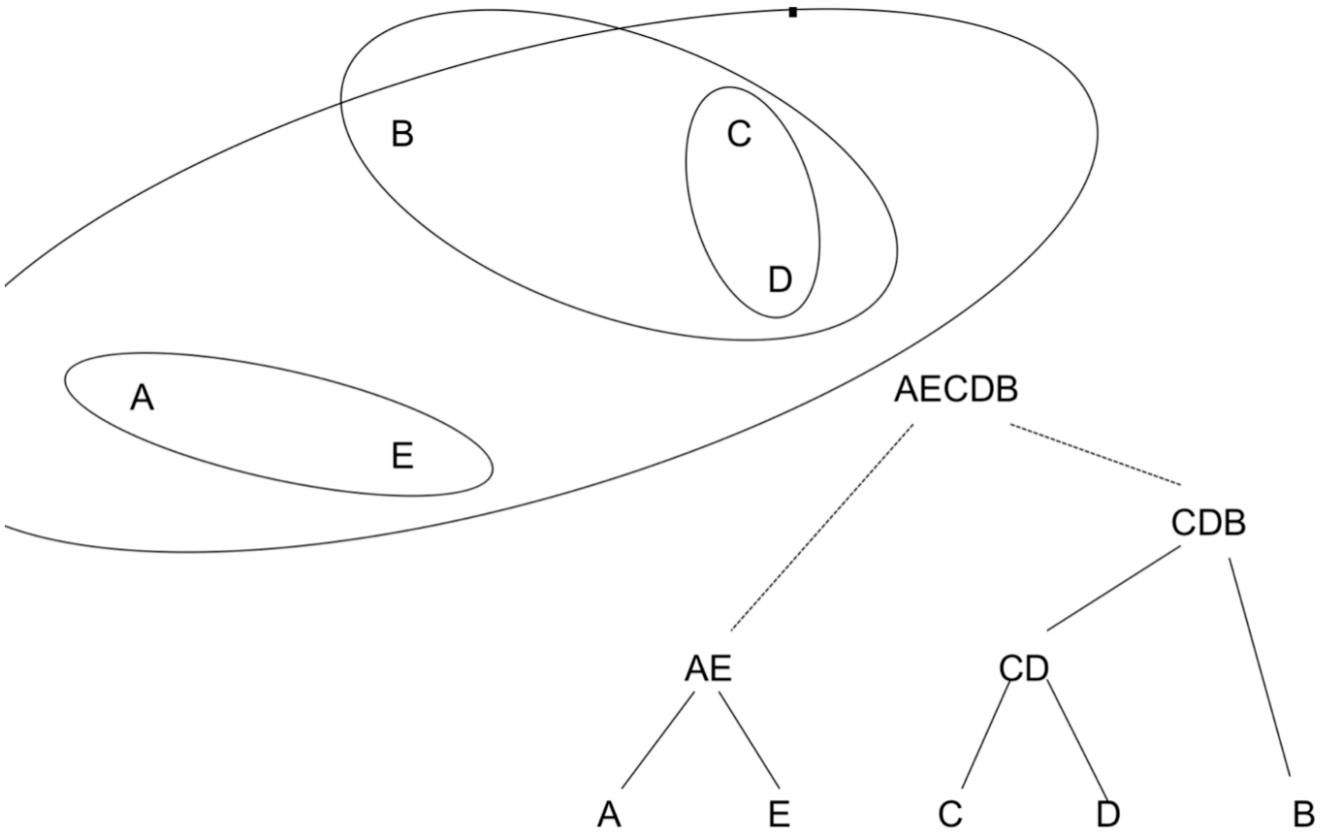


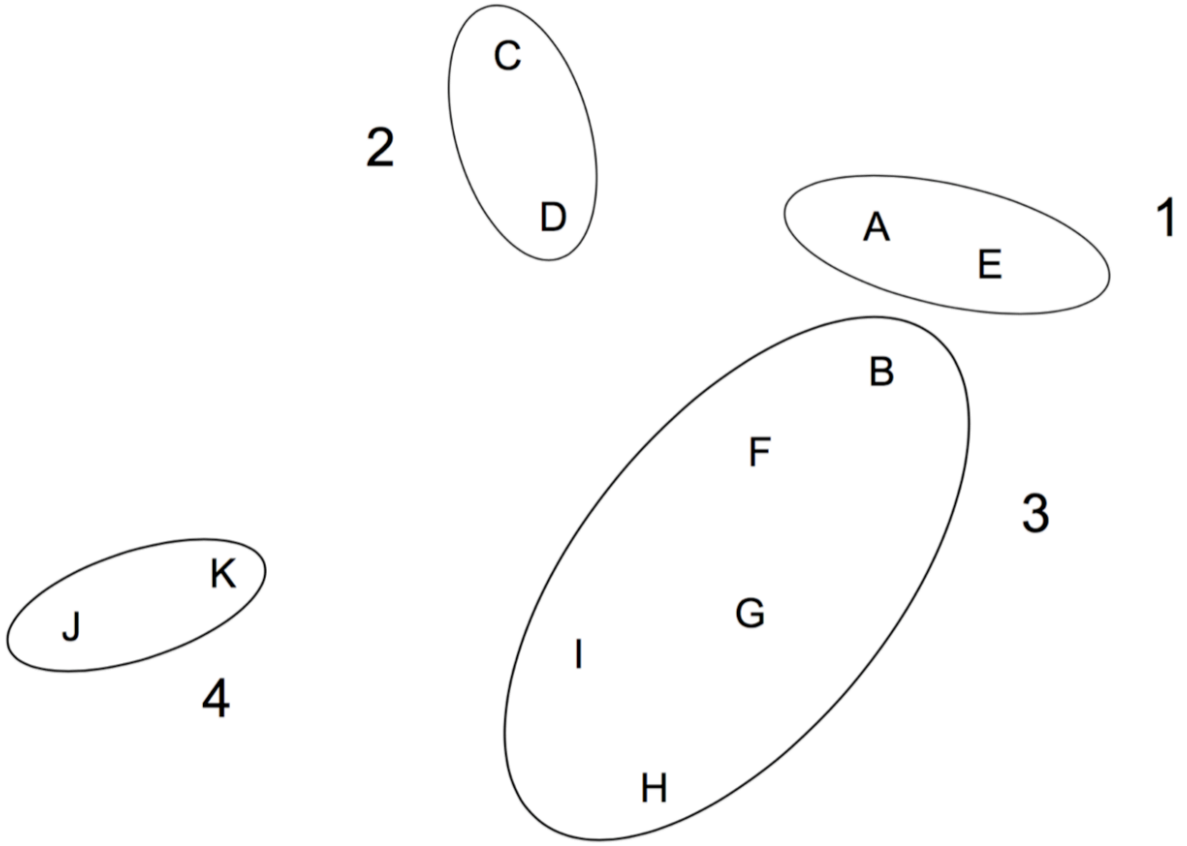
HAC (4 of 5)



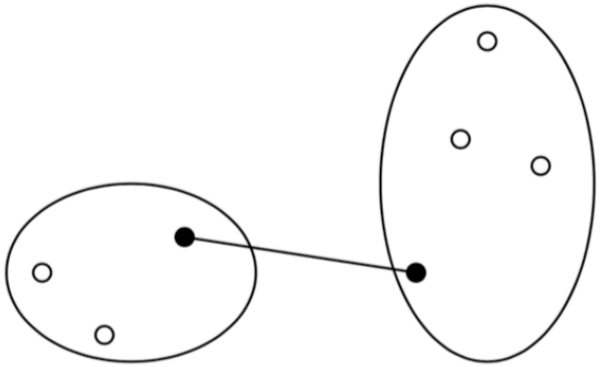


HAC (5 of 5)

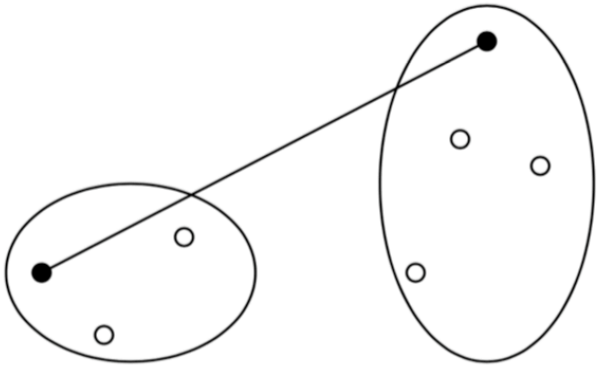




What is closest to cluster {A,E}?



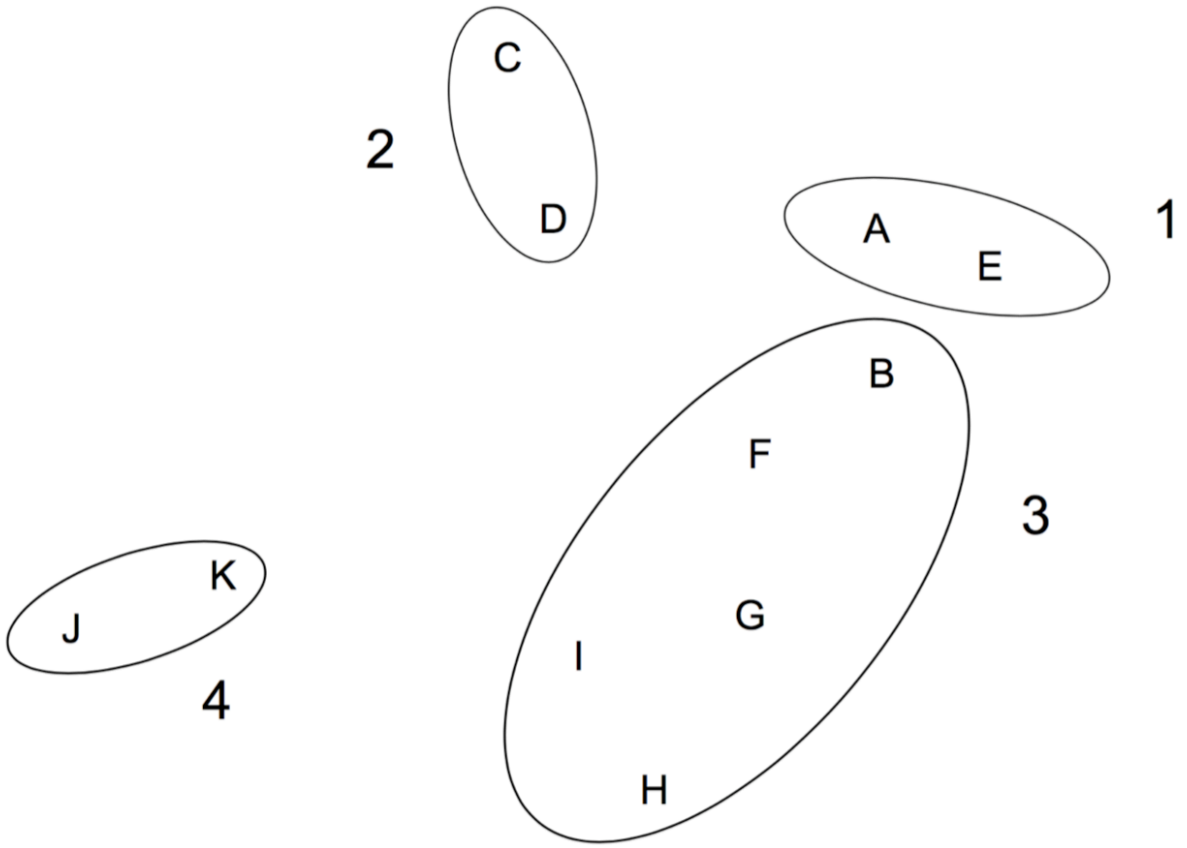
(a) min distance



(b) max distance

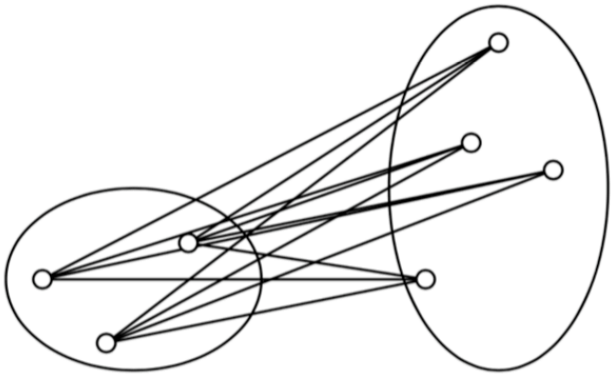
$$d_{\mathsf{min}}(G, G') = \min_{\mathbf{x} \in G, \mathbf{x}' \in G'} ||\mathbf{x} - \mathbf{x}'||$$

$$d_{\mathsf{max}}(G, G') = \max_{\mathbf{x} \in G, \mathbf{x}' \in G'} ||\mathbf{x} - \mathbf{x}'||$$

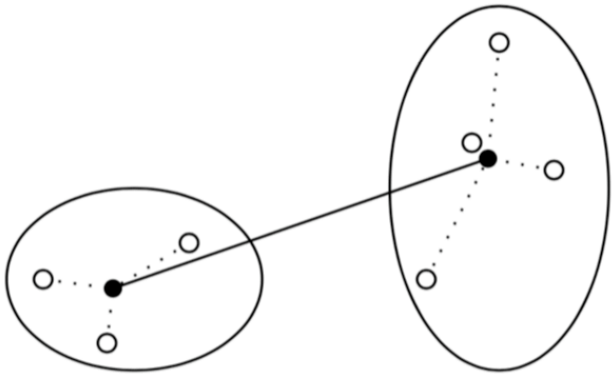


What is closest to cluster {A,E}?

“min” linkage
“max” linkage



(c) average distance

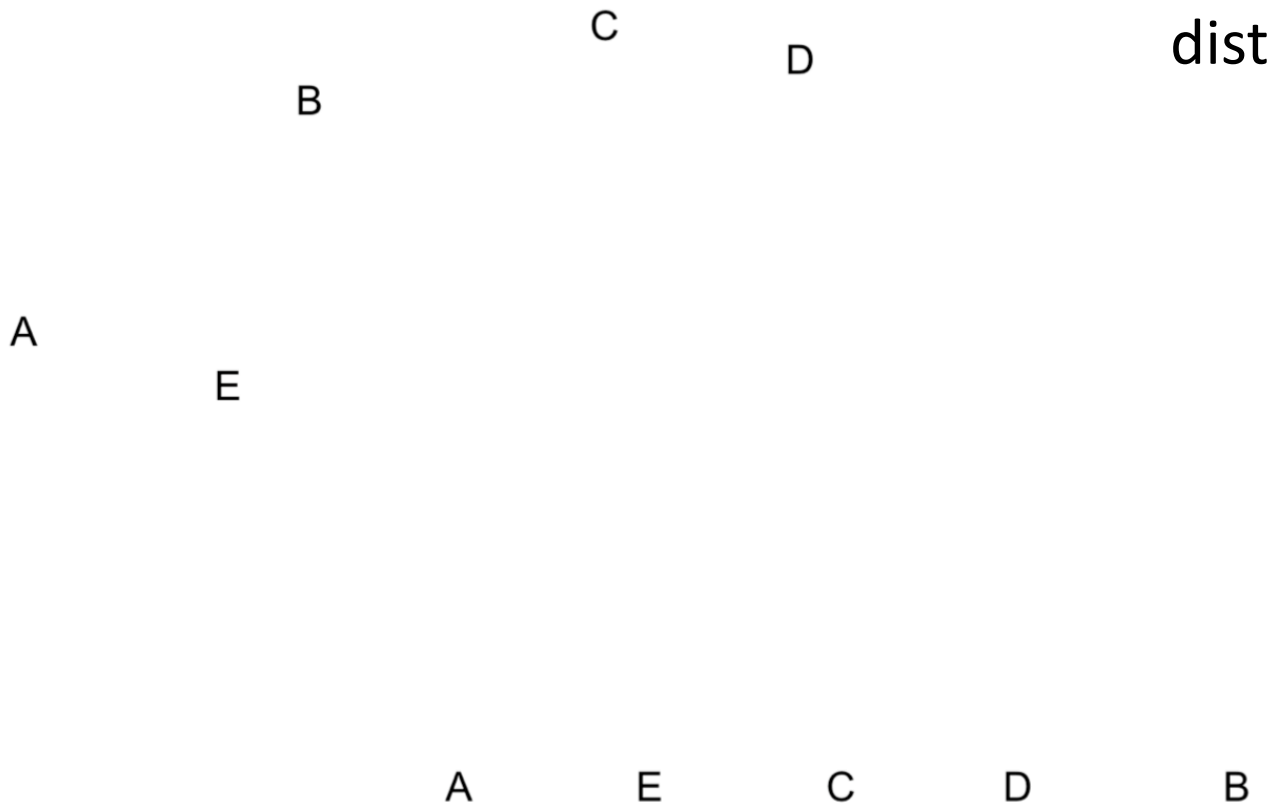


(d) centroid distance

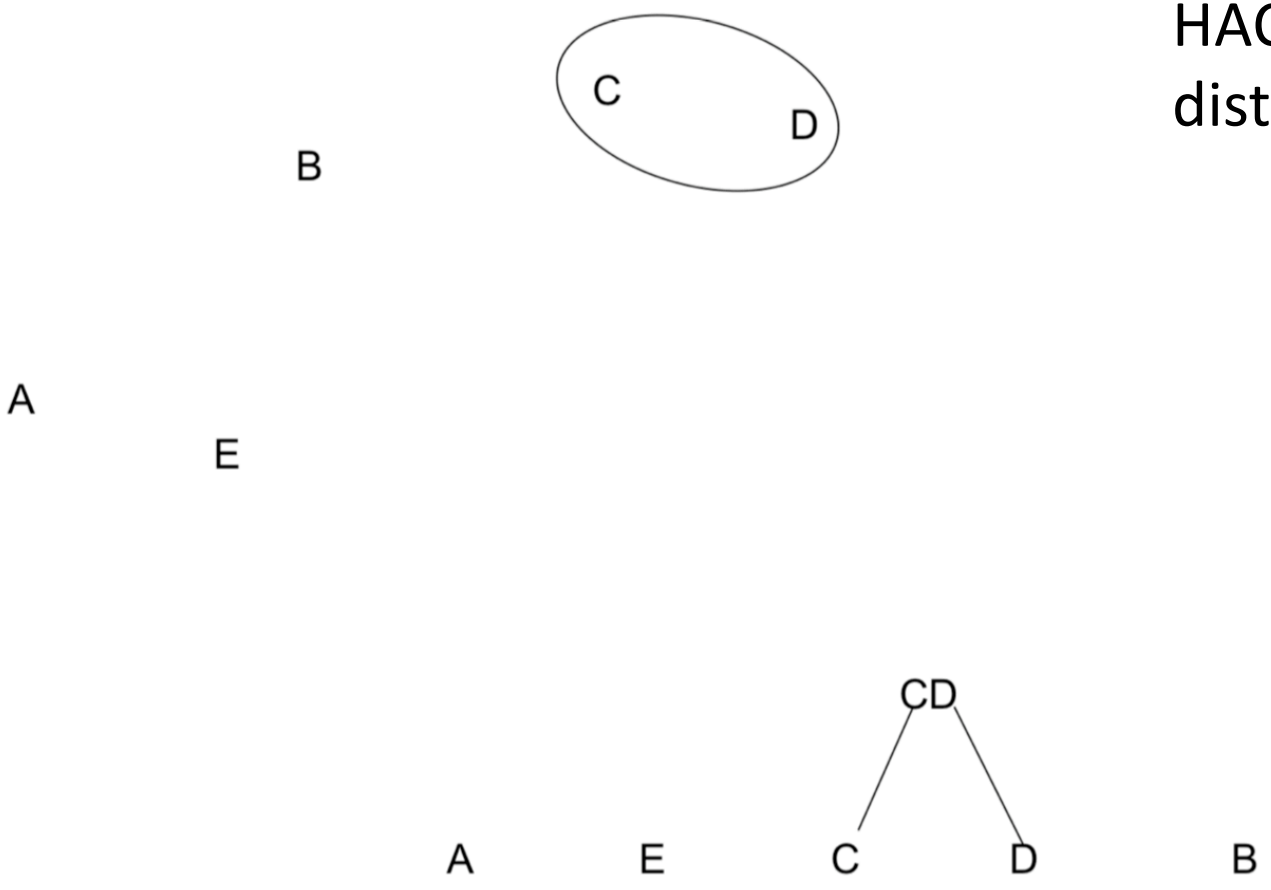
$$d_{\text{avg}}(G, G') = \frac{1}{|G||G'|} \sum_{\mathbf{x} \in G, \mathbf{x}' \in G'} ||\mathbf{x} - \mathbf{x}'||$$

$$d_{\text{centroid}}(G, G') = ||\boldsymbol{\mu}_G - \boldsymbol{\mu}'_G||$$

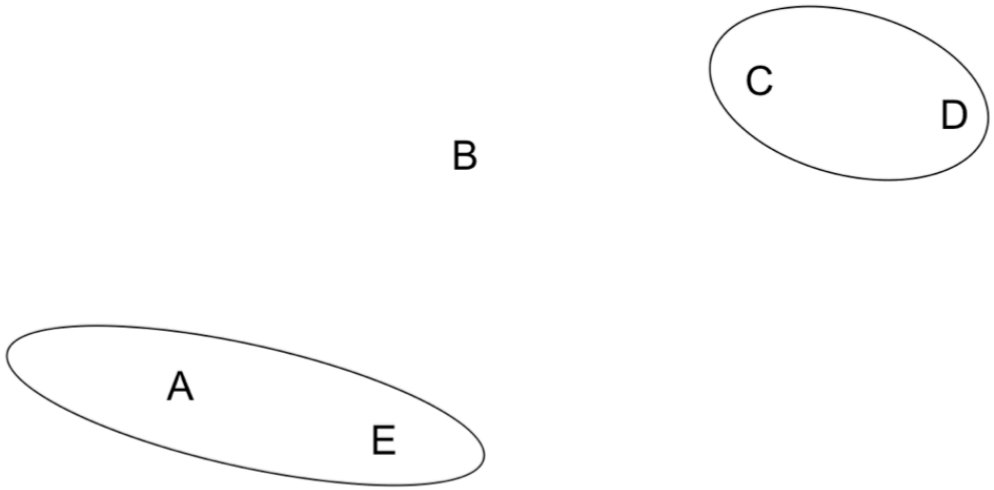
HAC with min distance (1 of 6)

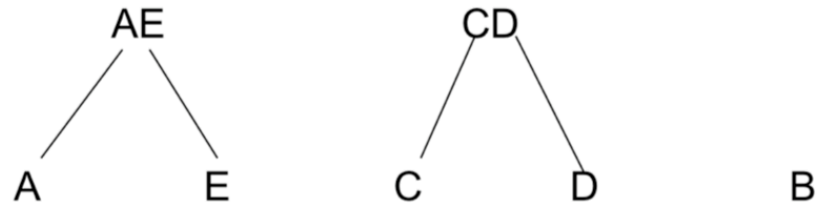


HAC with min distance (2 of 6)

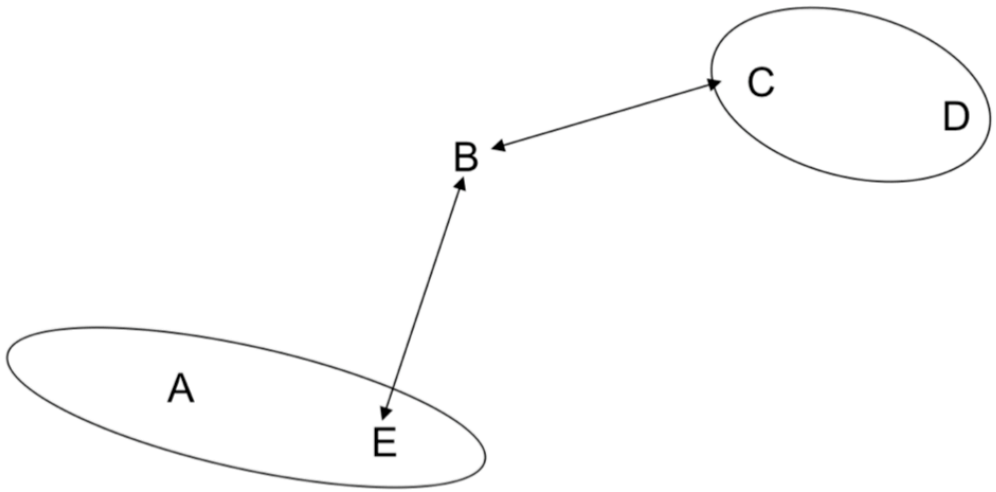


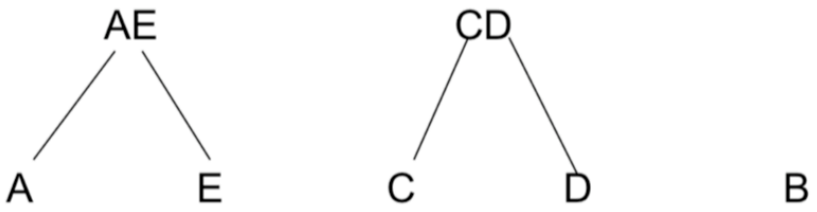
HAC with min distance (3 of 6)



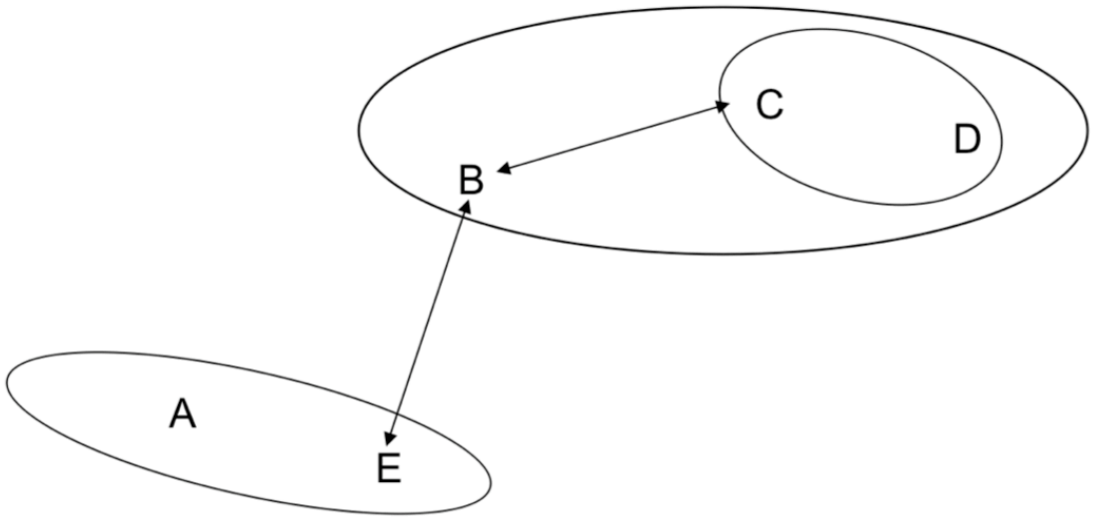


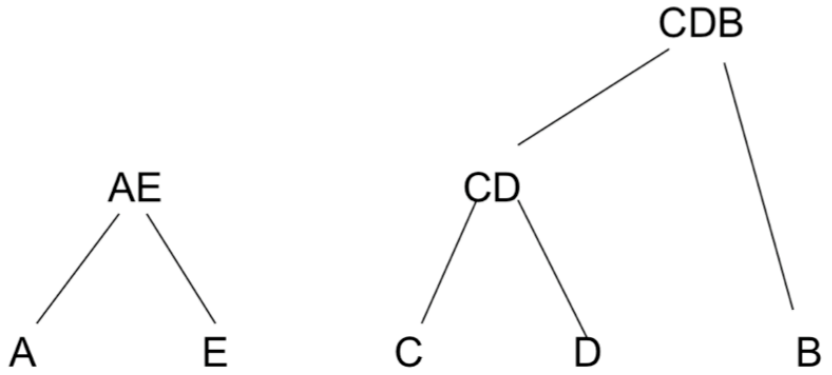
HAC with min distance (4 of 6)



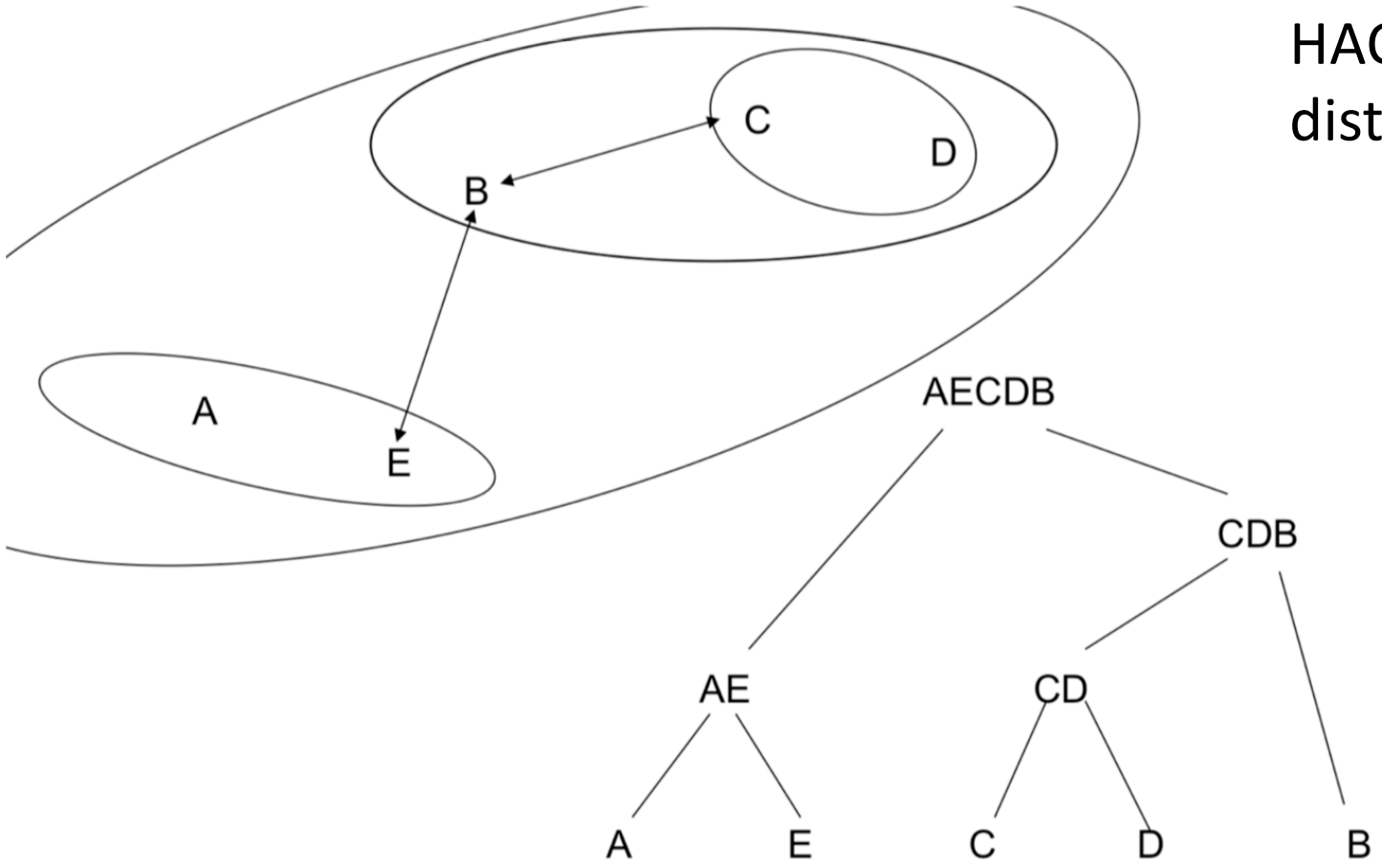


HAC with min distance (5 of 6)

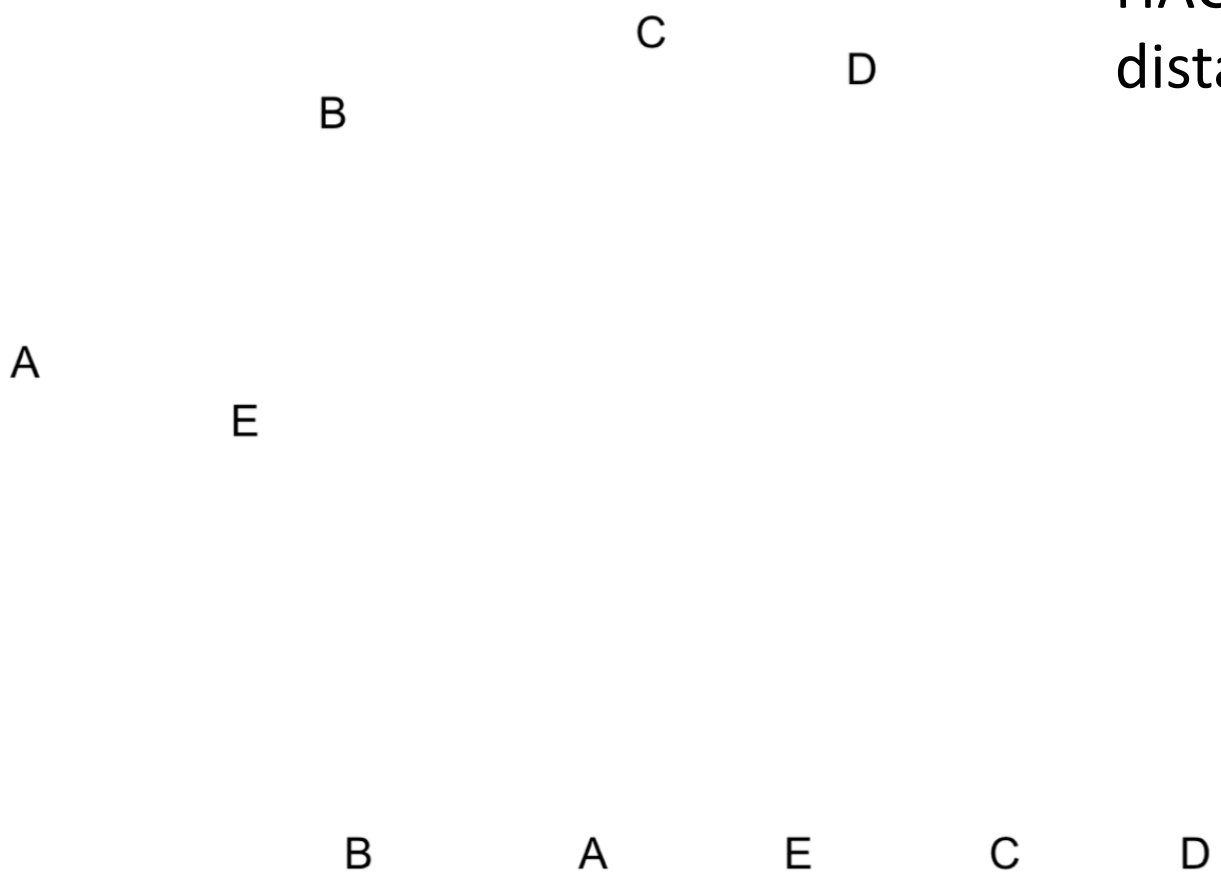




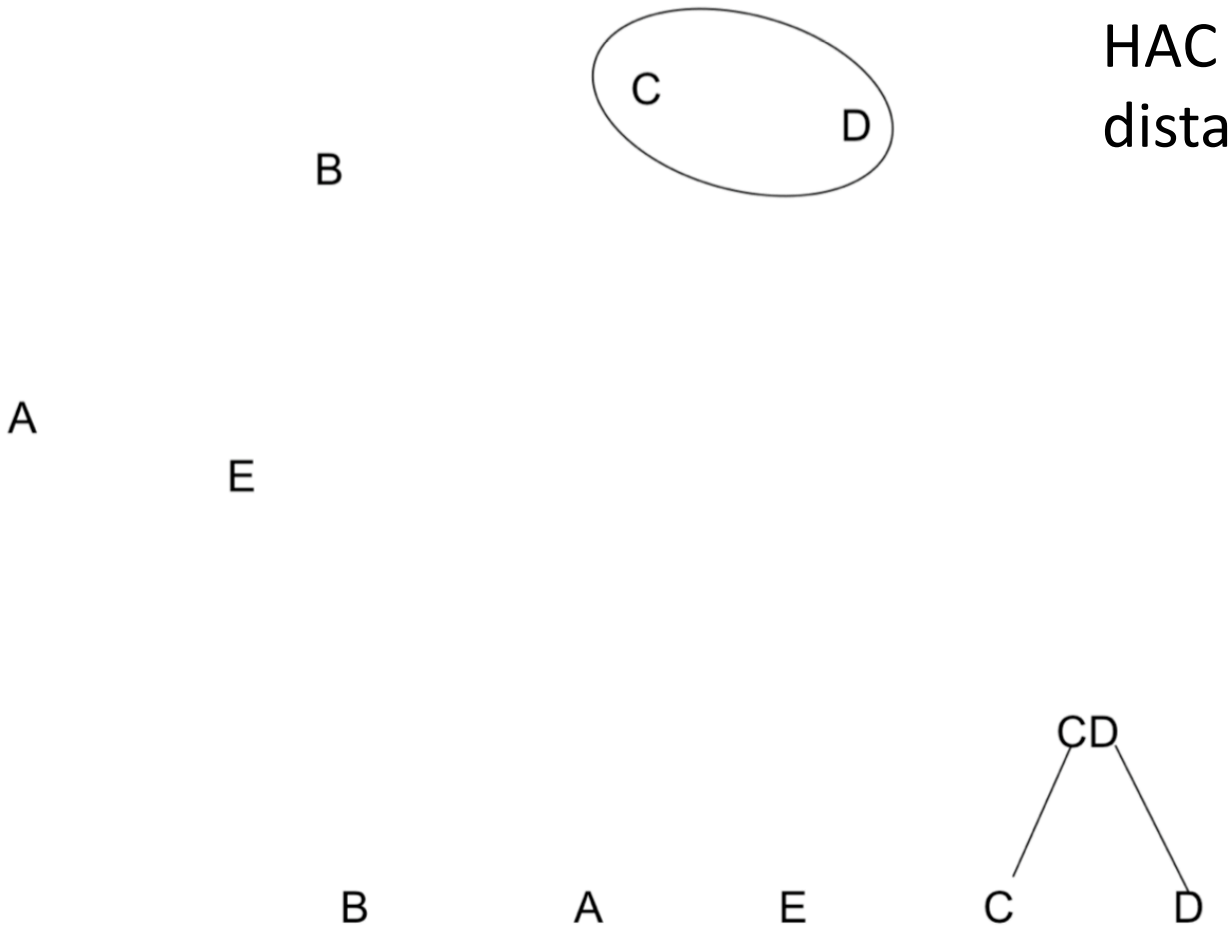
HAC with min distance (6 of 6)



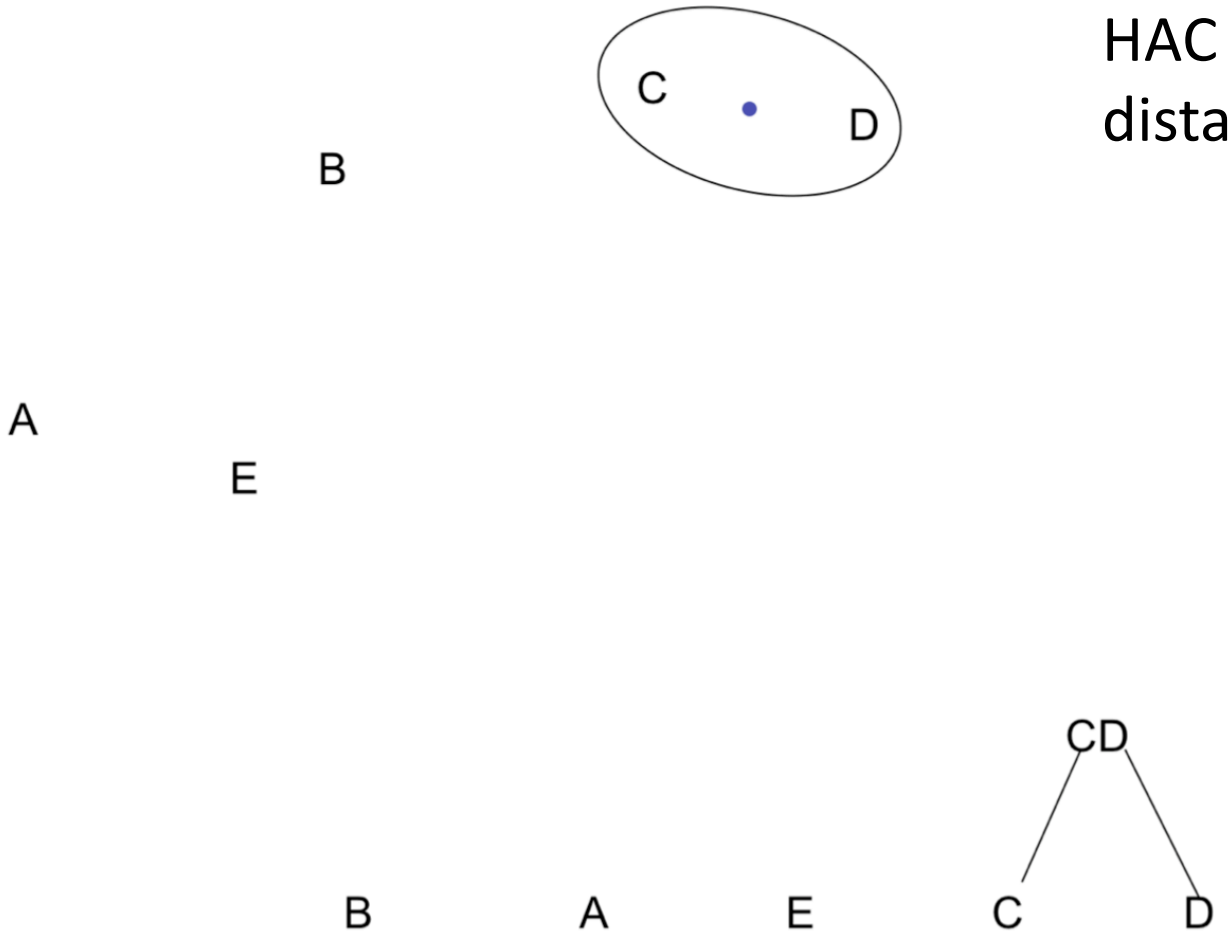
HAC with centroid distance (1 of 7)



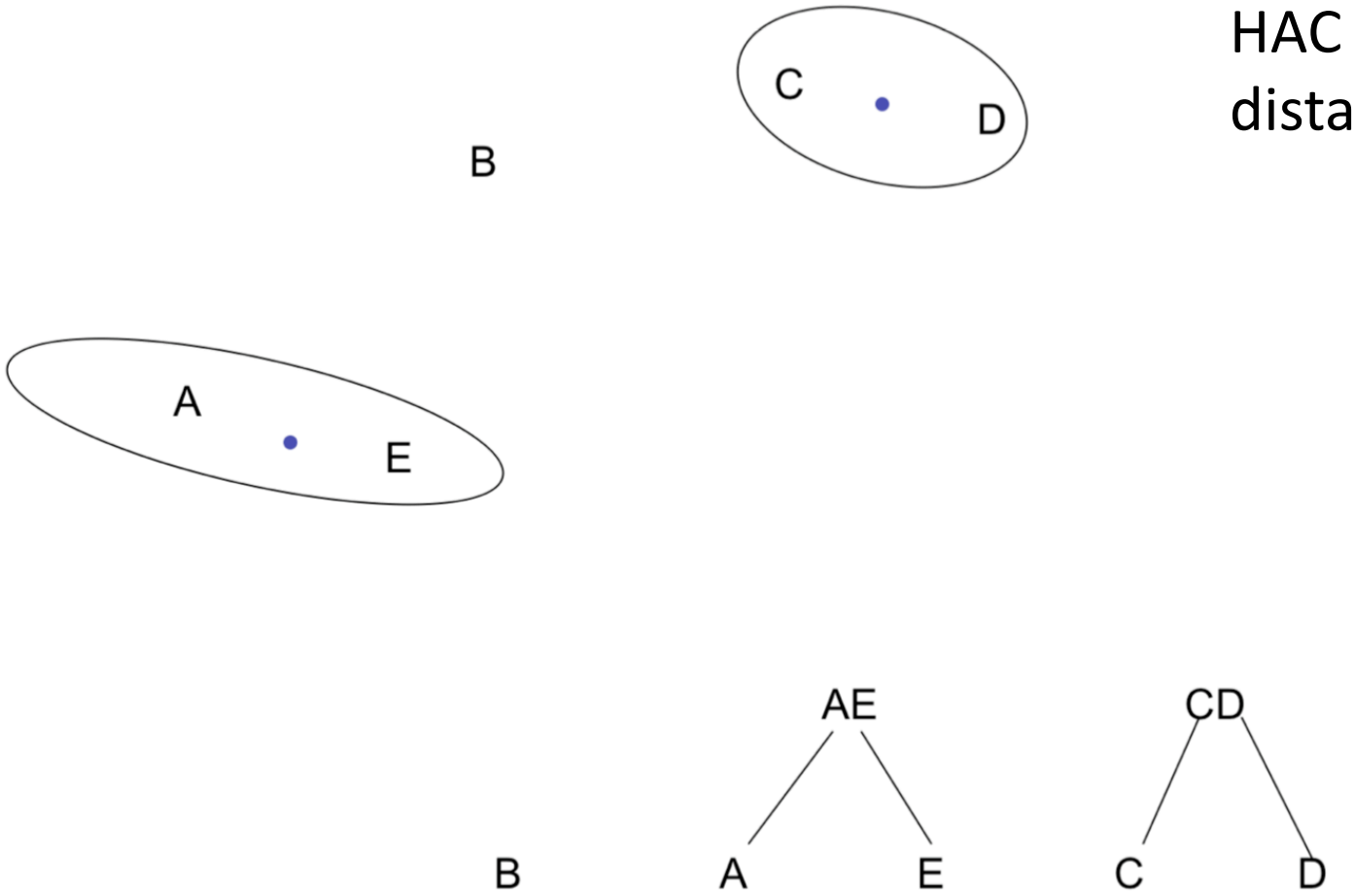
HAC with centroid distance (2 of 7)



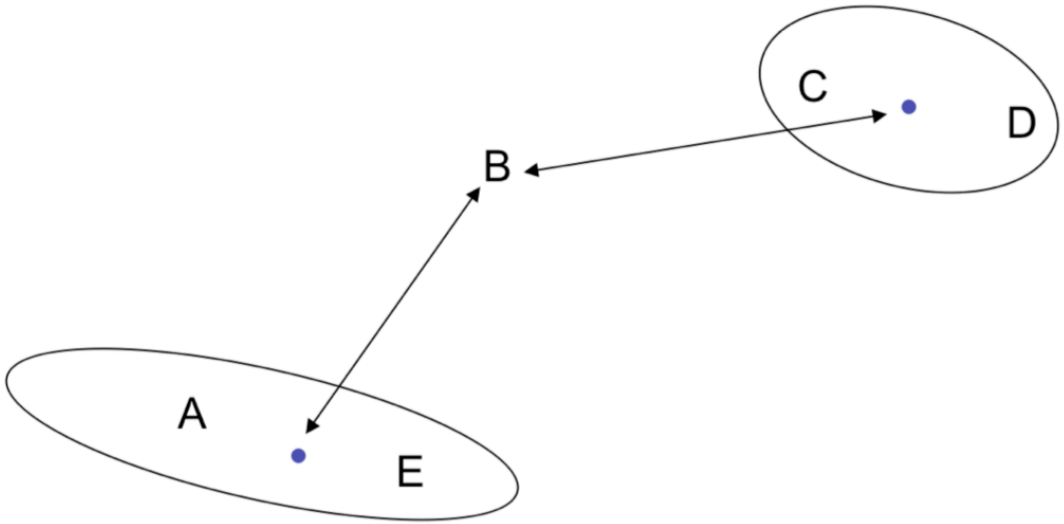
HAC with centroid distance (3 of 7)

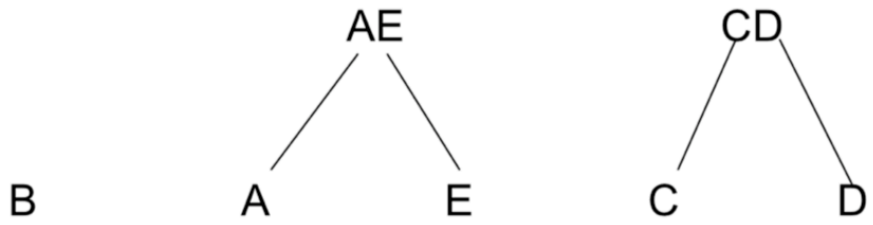


HAC with centroid distance (4 of 7)

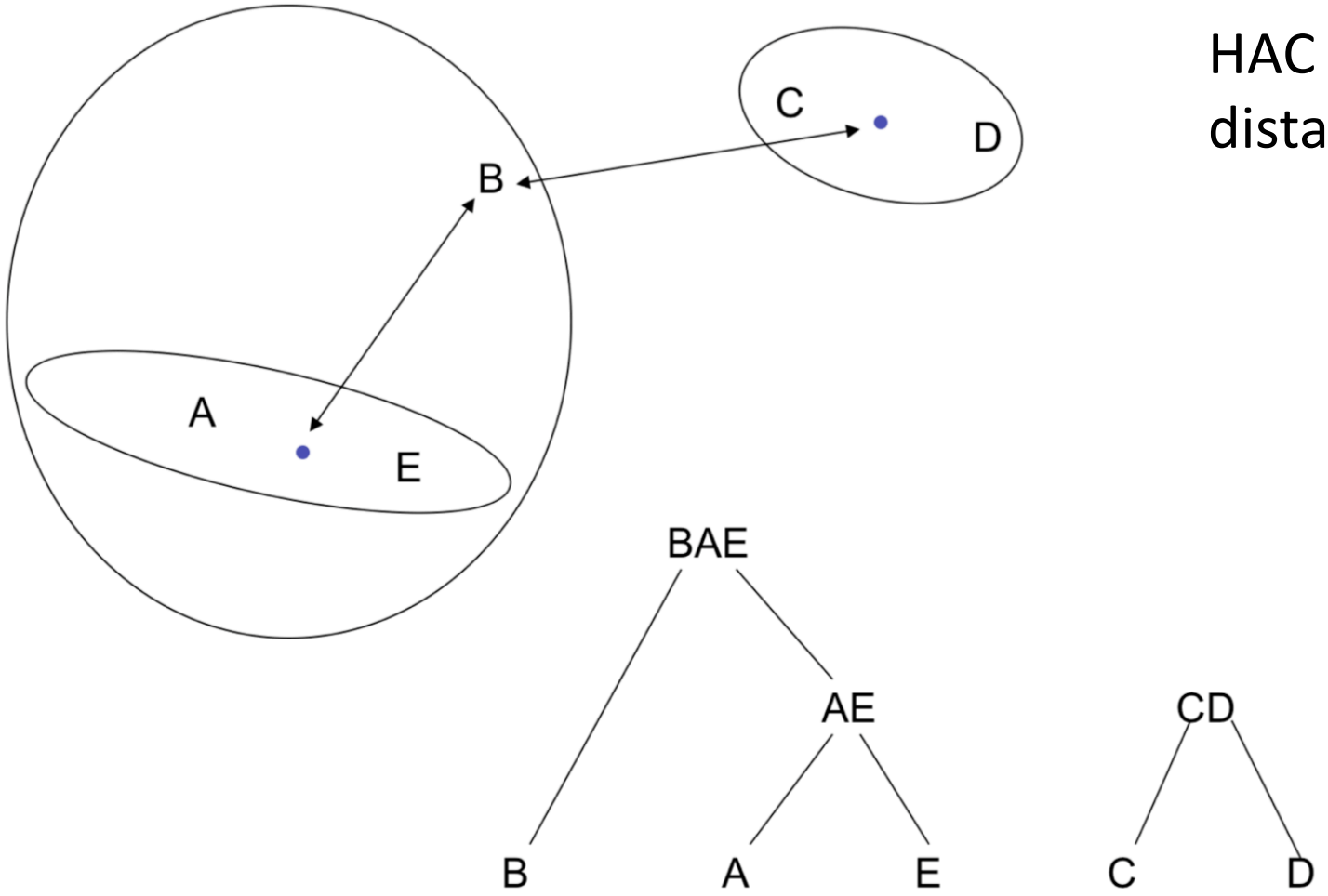


HAC with centroid distance (5 of 7)

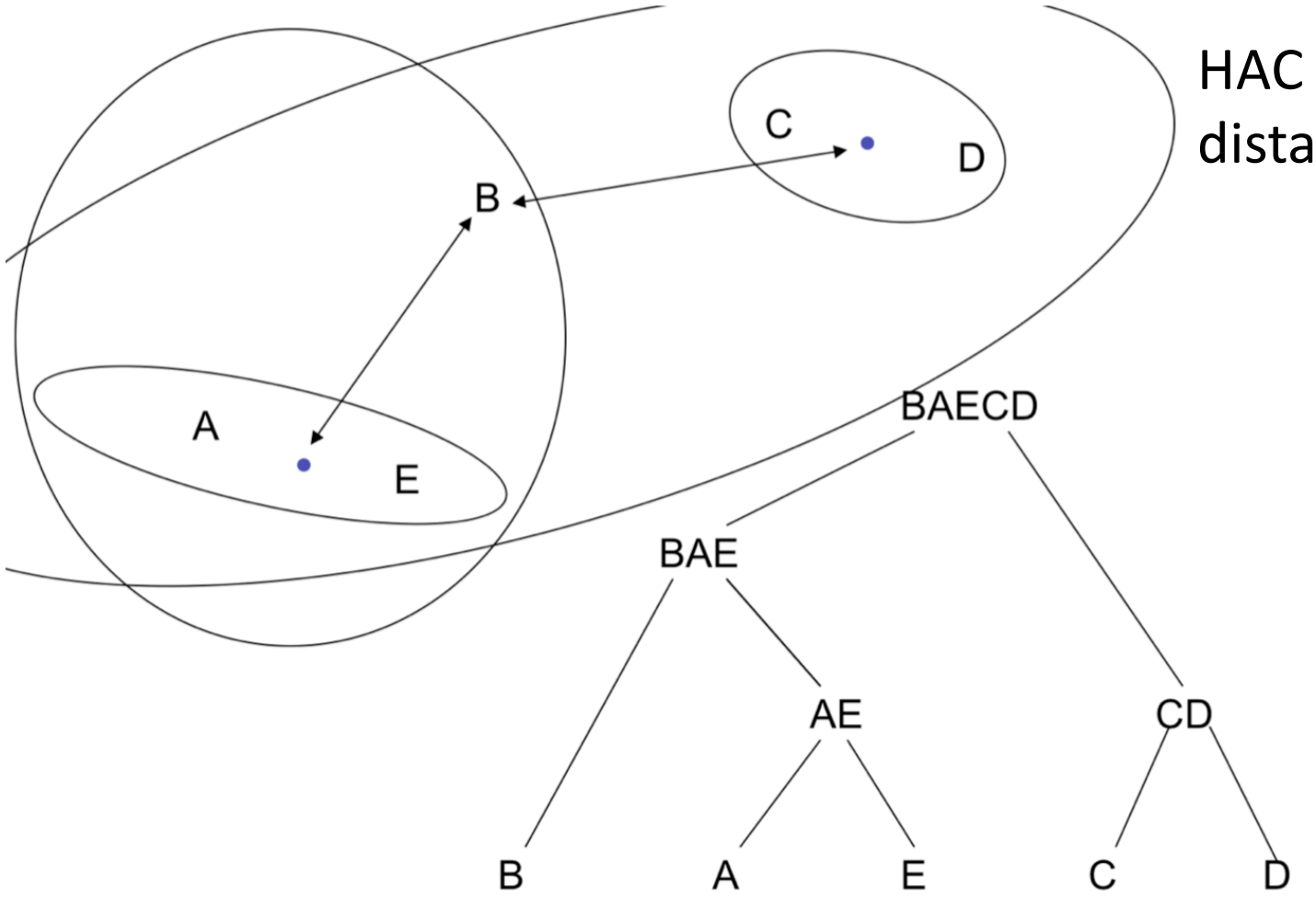




HAC with centroid distance (6 of 7)



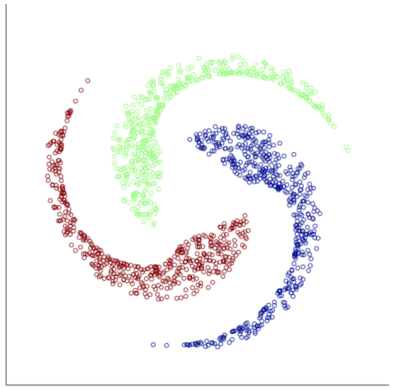
HAC with centroid distance (7 of 7)



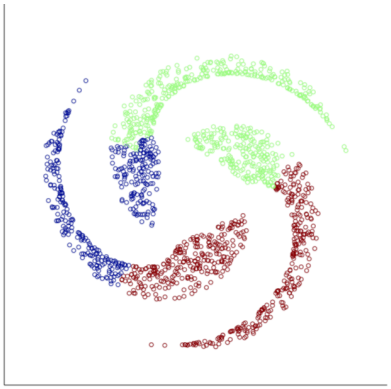
Comparing HAC Group Distance criteria

- Which of “min” and “max” linkages will tend to merge large clusters with each other?
  - A: **Min**. Large clusters more likely to have a pair of examples that are close
- Which of “min” and “max” will tend to have a “chaining effect” and lead to “long, stringy” clusters?
  - A: **Min**. Only one distance has to be small to merge
- Which of “min” and “max” will tend to prefer compact clusters?
  - A: **Max**. All distances have to be small to merge
- The “average” and “centroid” linkages are compromises, allowing some elongation but also preferring some compactness

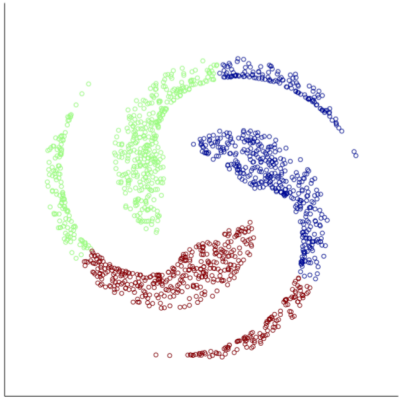
HAC applied to Pinwheel



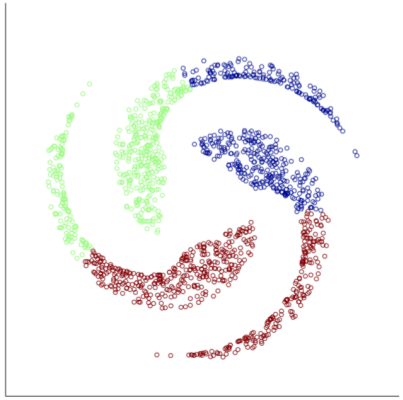
(a) min distance



(b) max distance

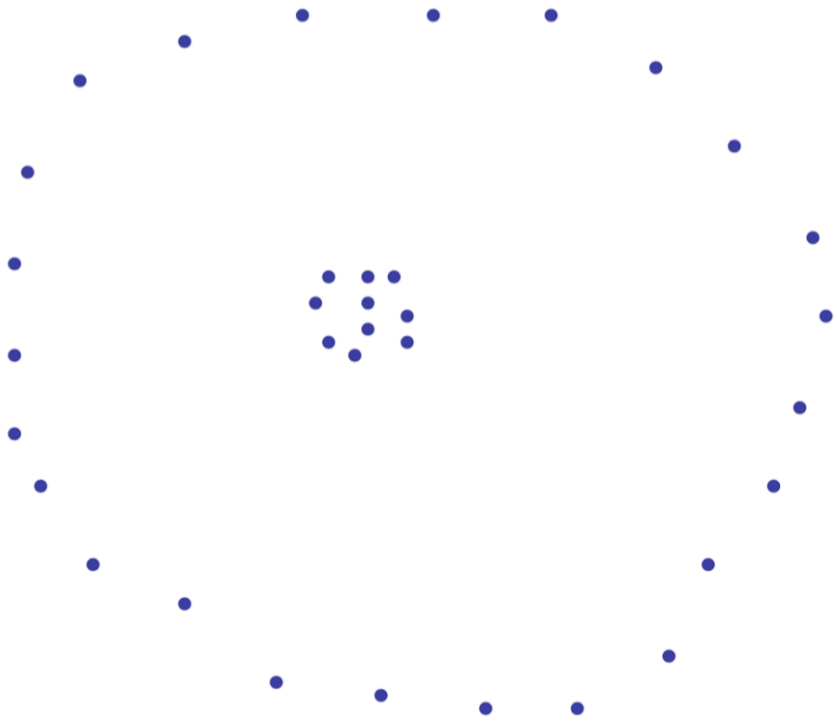


(c) average distance



(d) centroid distance

Comprehension Question

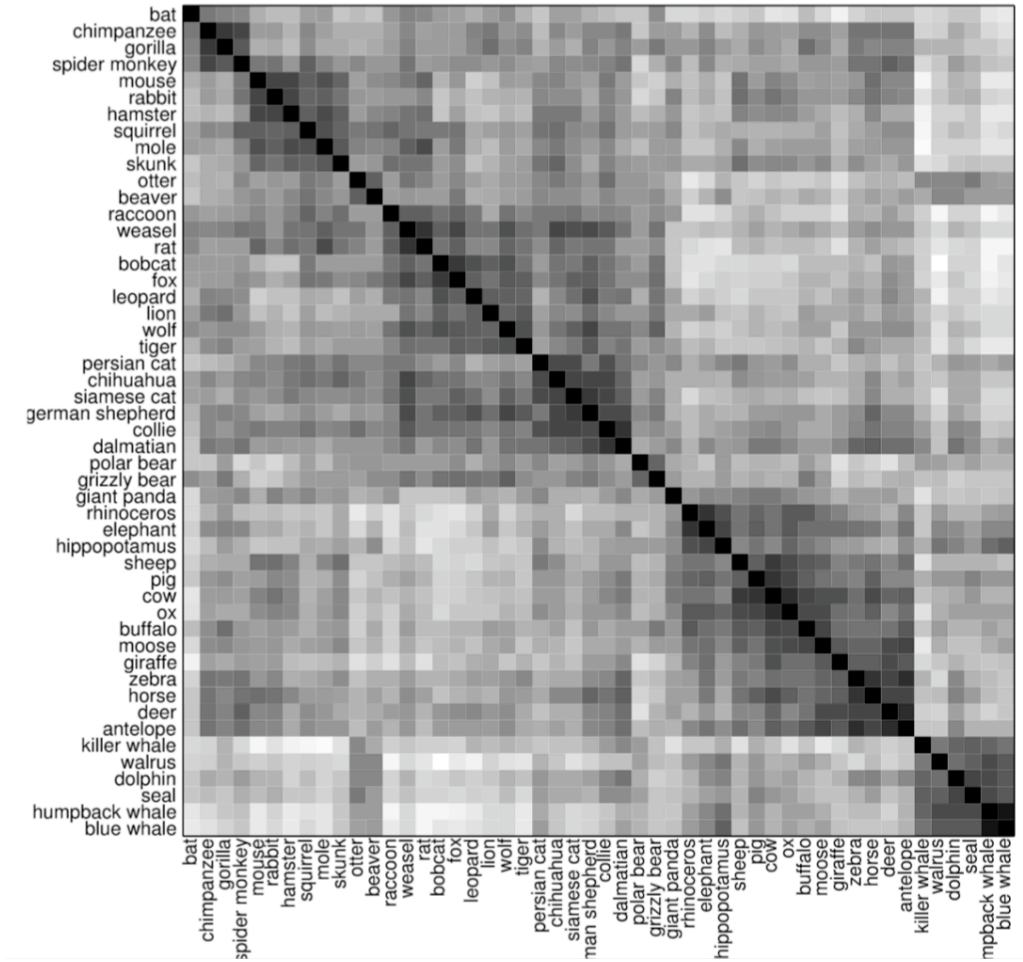


What will “min” do here?

What will “max” do here?

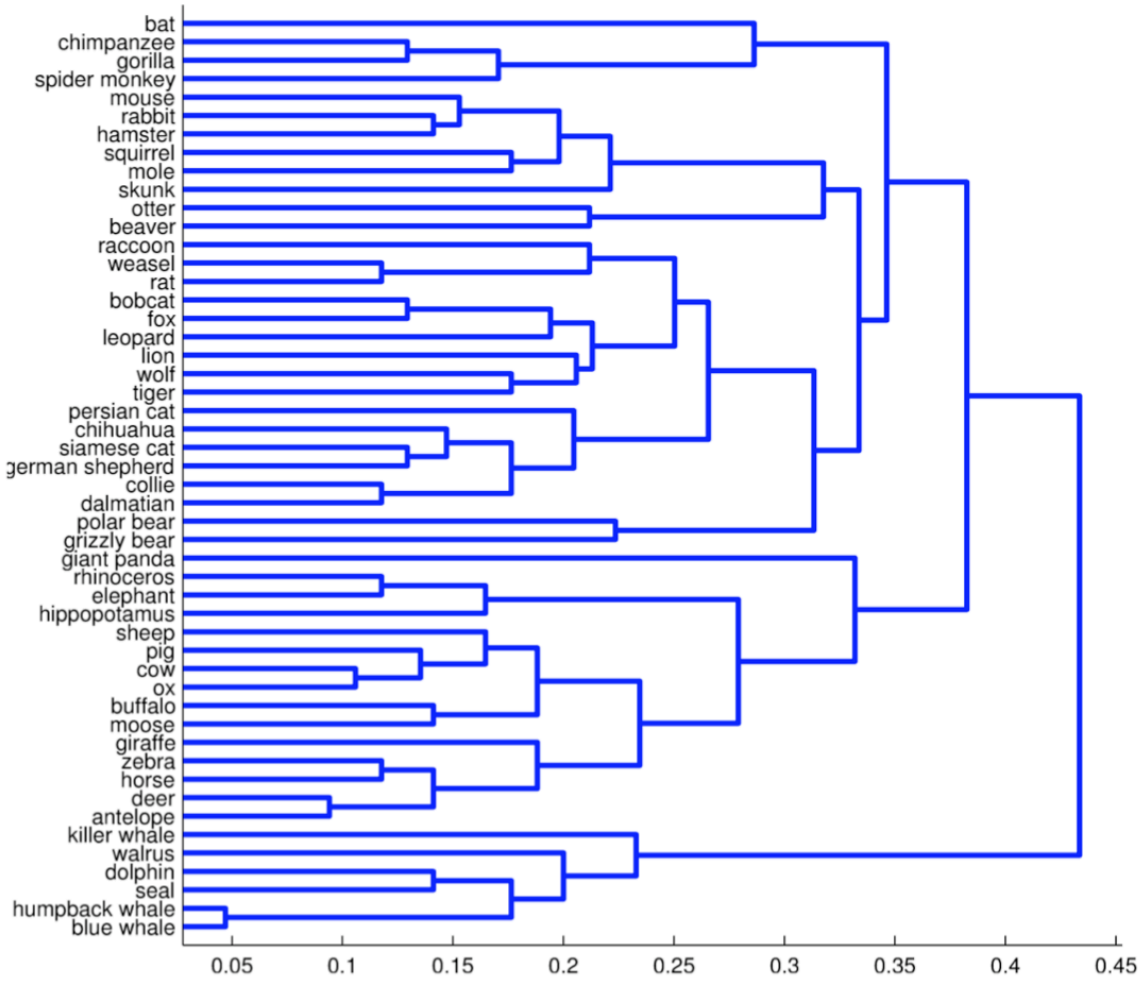
Example: Animal Clustering

Example: Animal Clustering



Step 1:
Compute the pairwise distances (Hamming distances)

Example: Animal Clustering

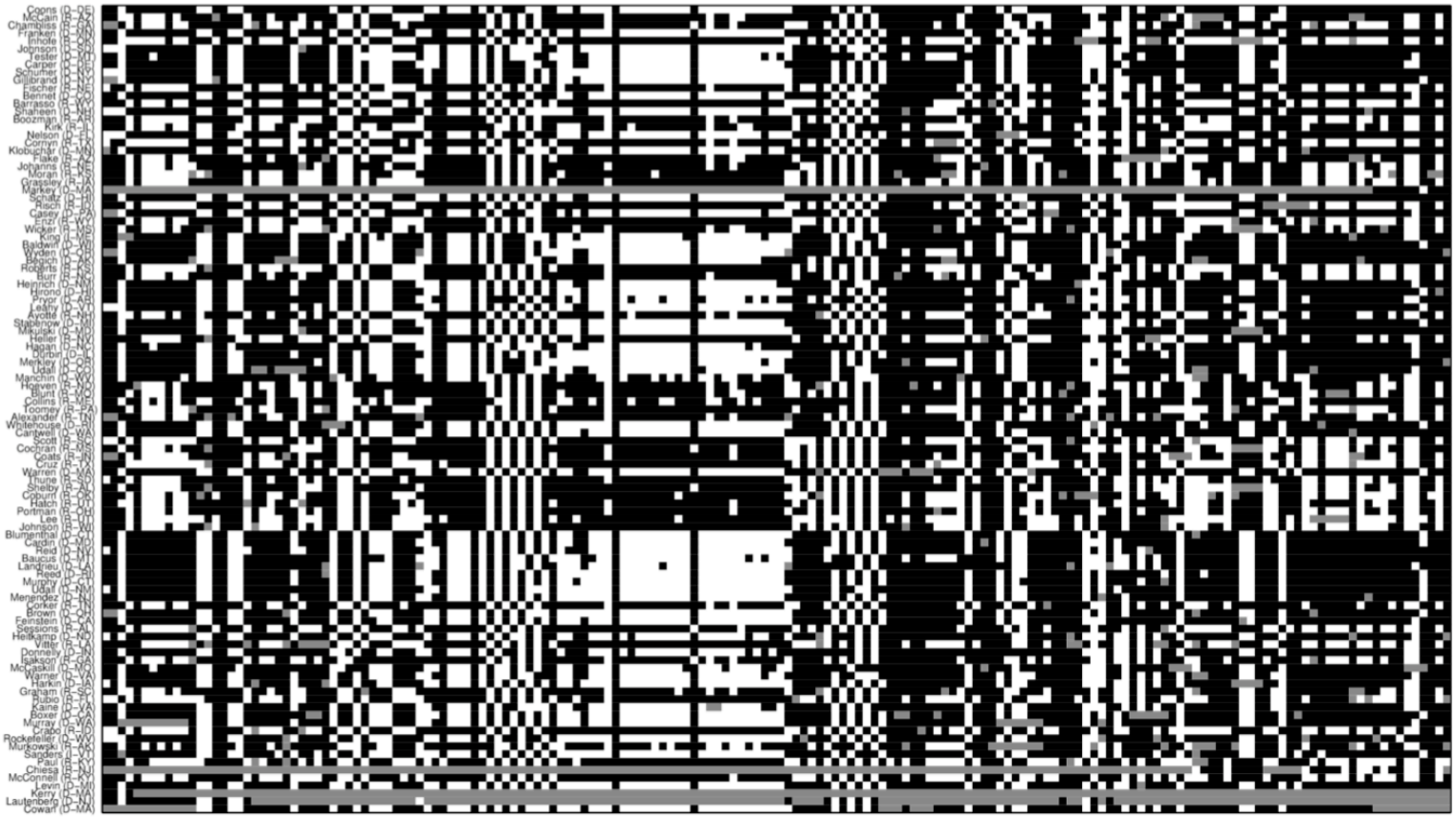


Step 2: Apply HAC clustering with “average” linkage

Result illustrated through “dendrogram”

Shows groups that were merged. The x-distance provides the distance between groups when merged.

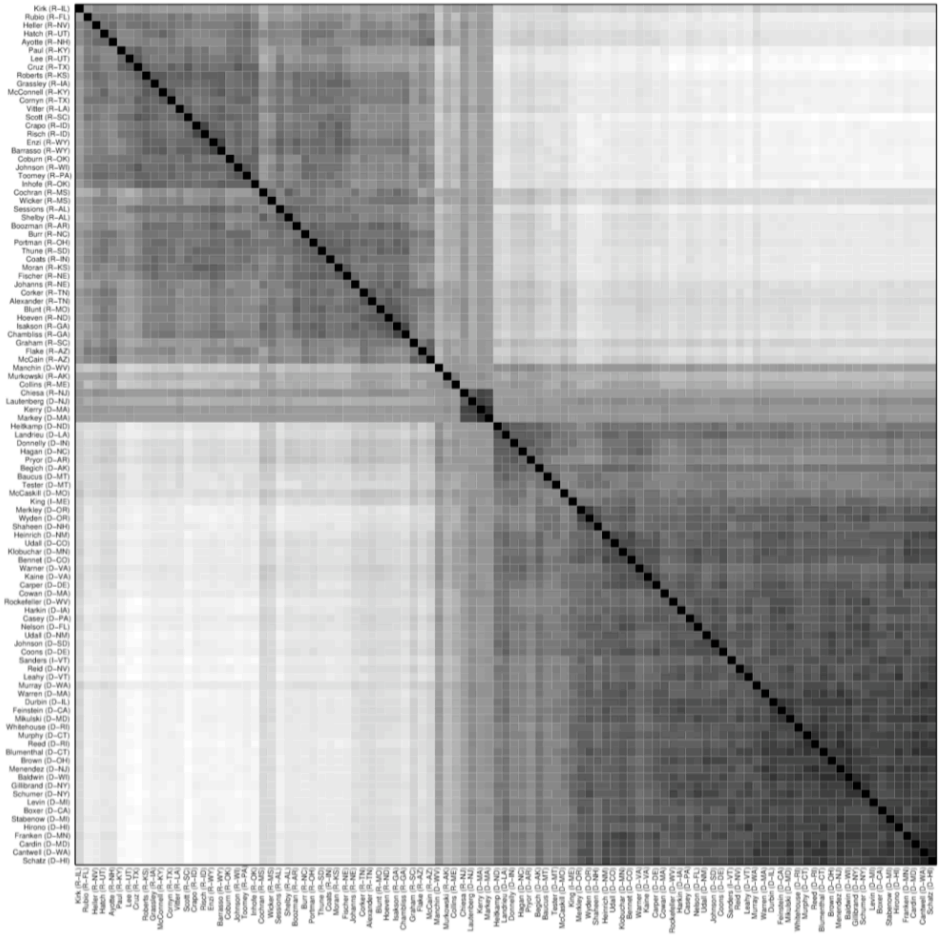
Example: Senators in 113th US Congress



(a) Senators and Votes

Features: votes on 172 bills

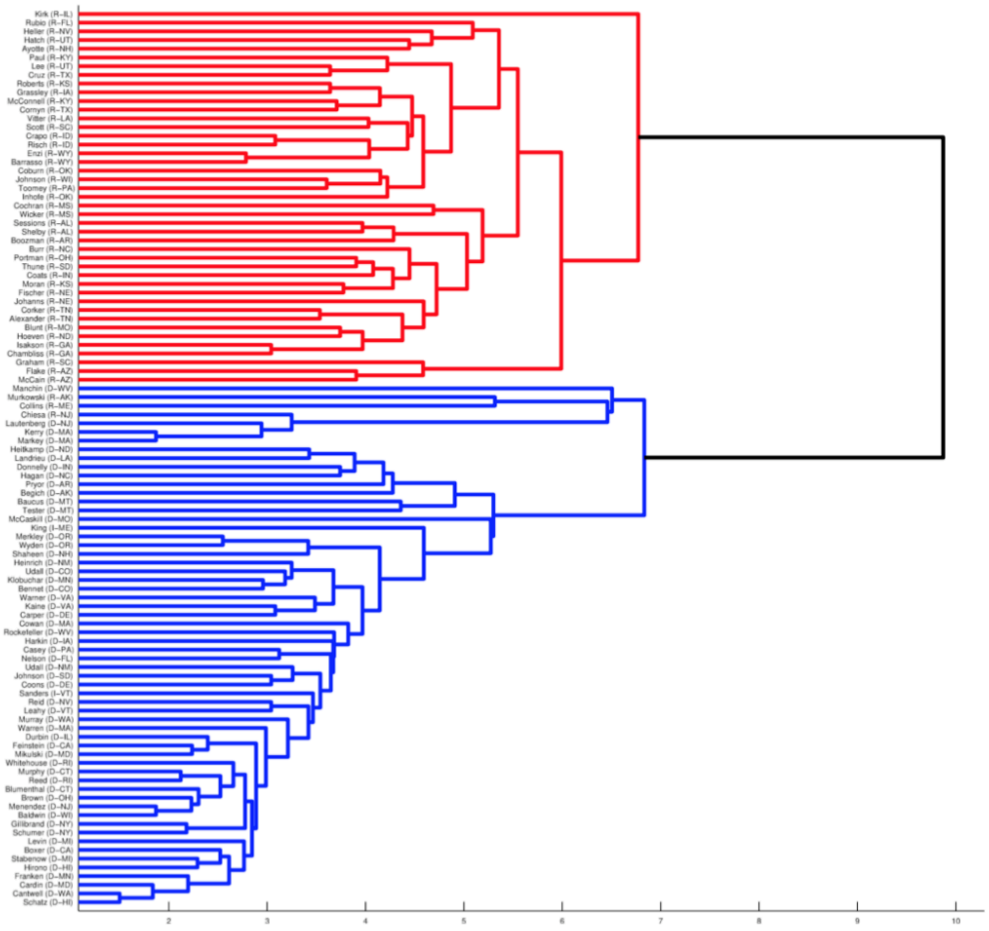
Example: Senators in 113th US Congress



Step 1: Computer pairwise distances between voting record (L2 norm)

Darker is smaller distance
Ordered by similarity

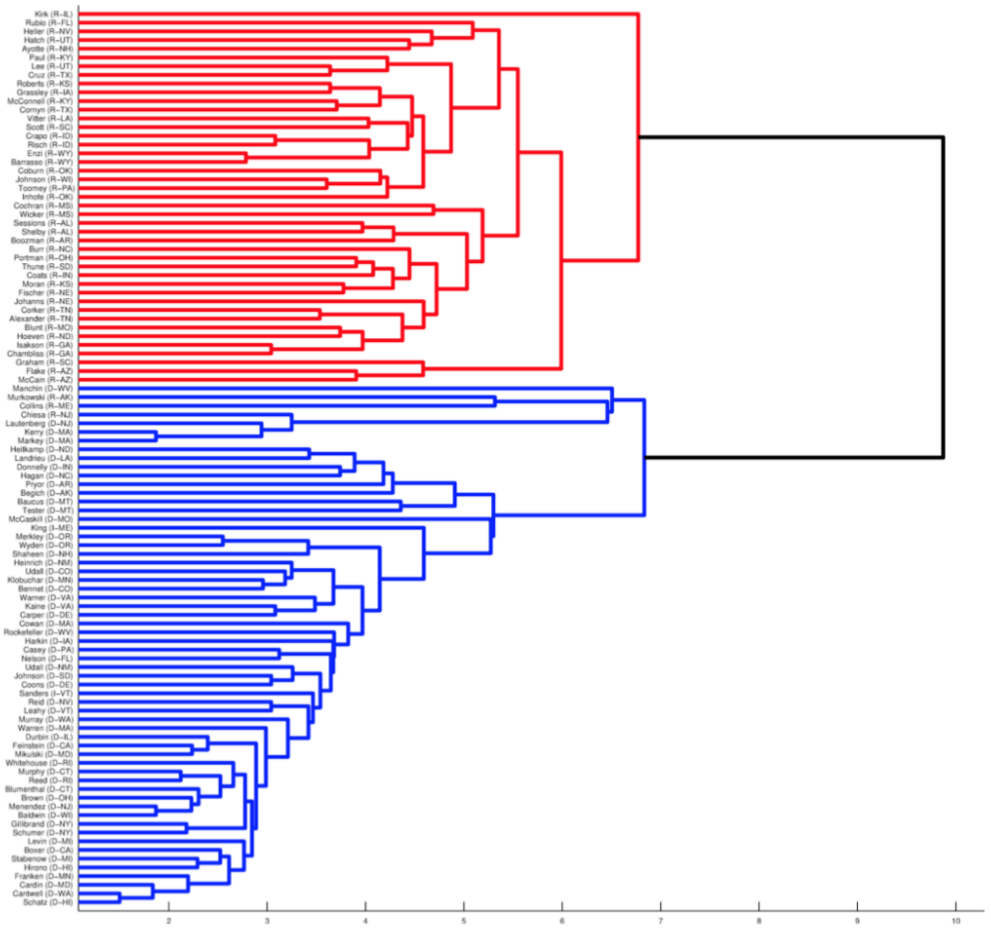
Example: Senators in 113th US Congress

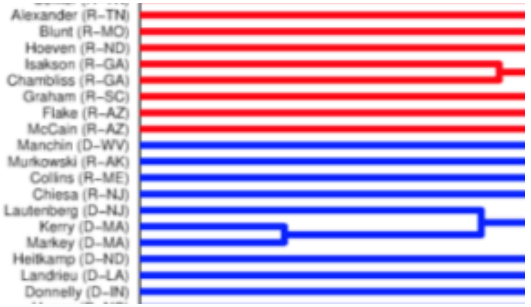


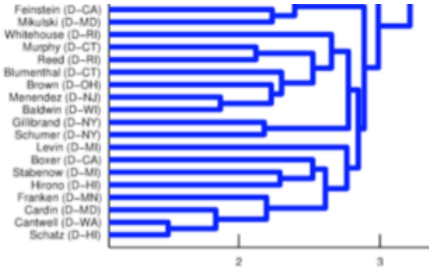
Step 2: HAC average

Dendrogram, showing the "top 2" clusters

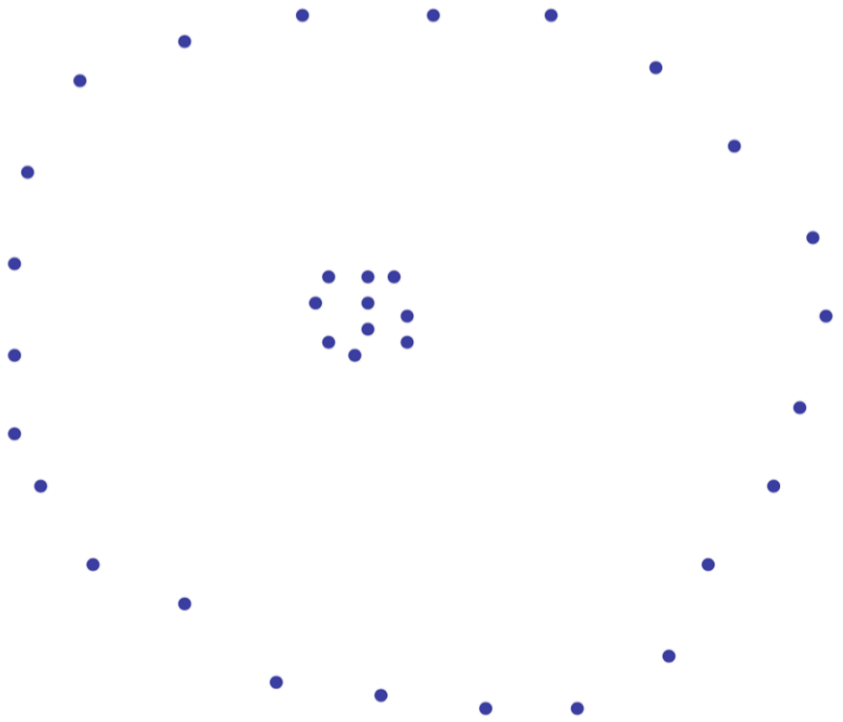
Example: Senators in 113th US Congress





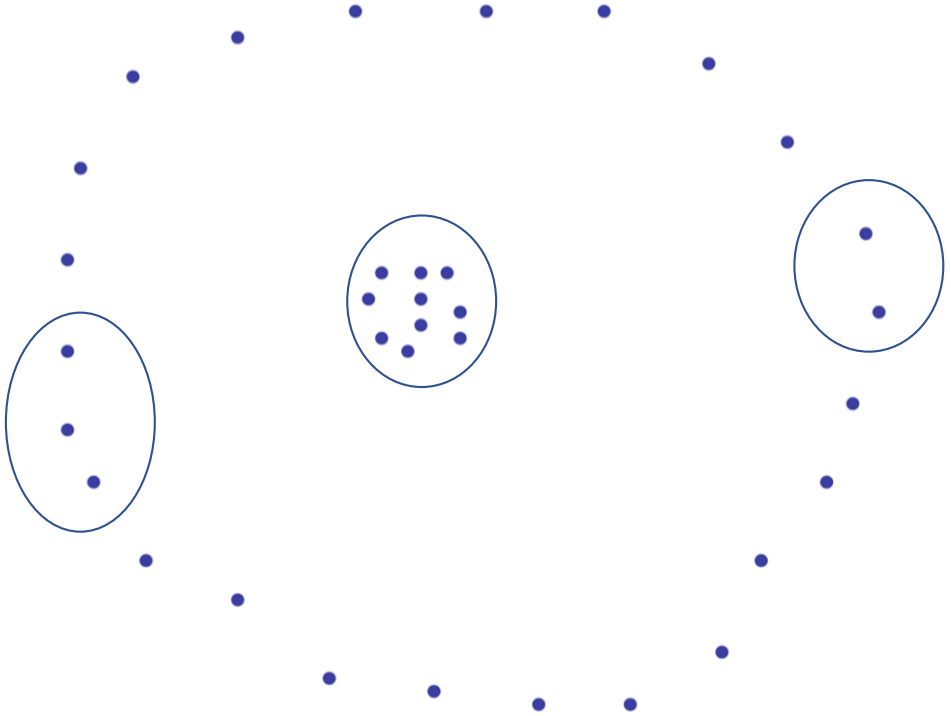


Comprehension Question



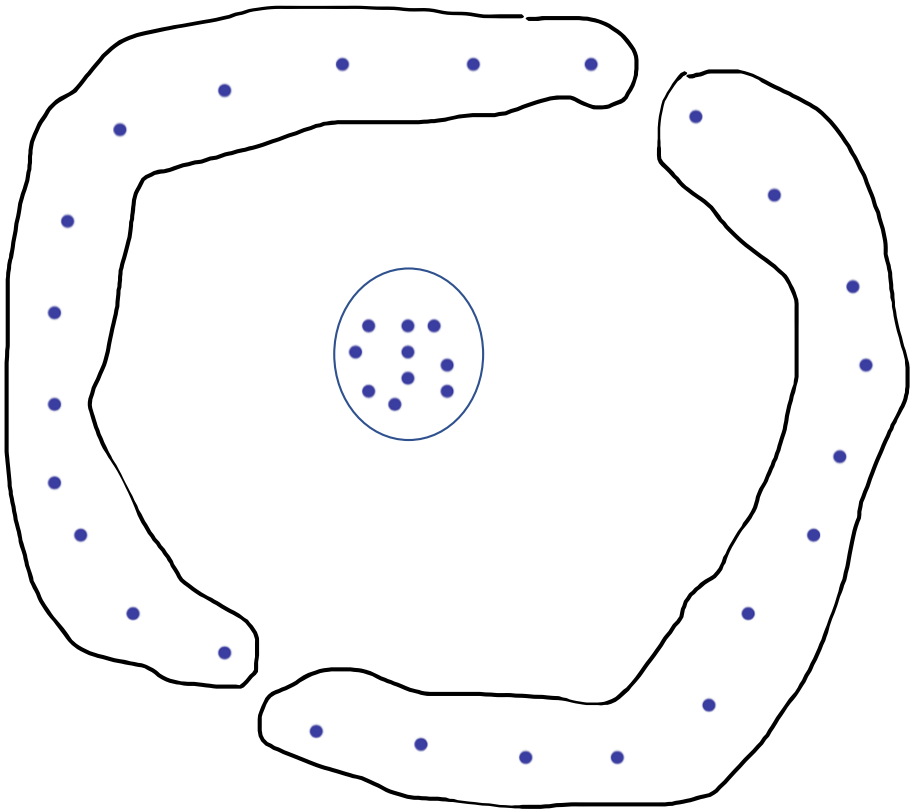
What will “min” do here?

Comprehension Question



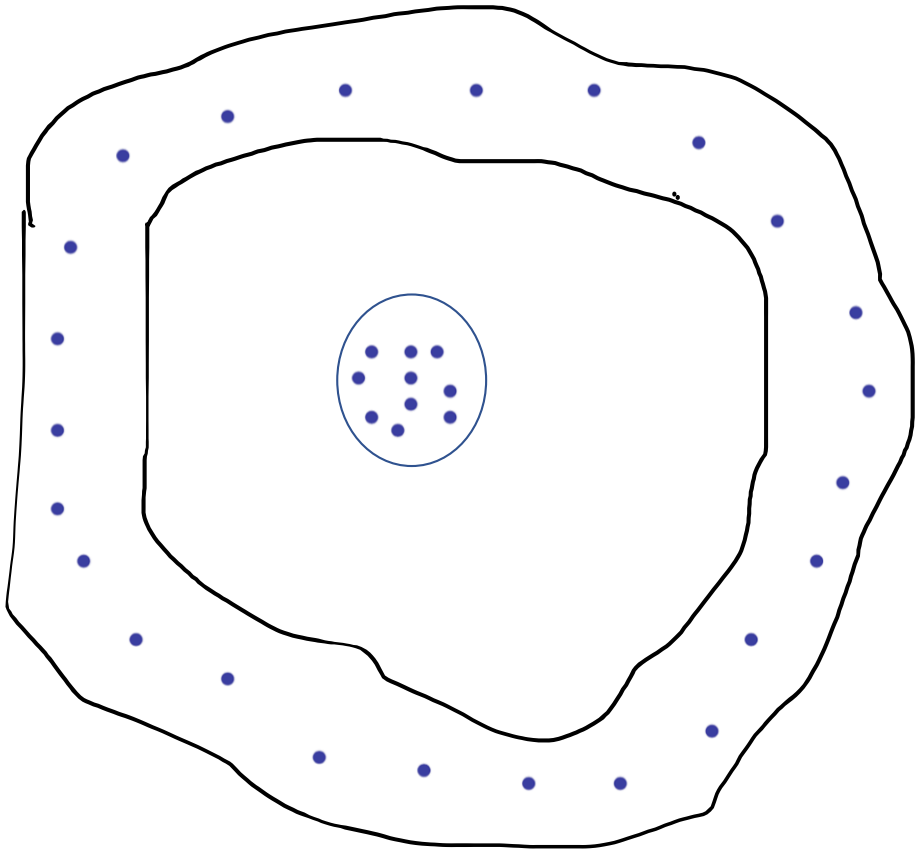
What will “min” do here?

Comprehension Question



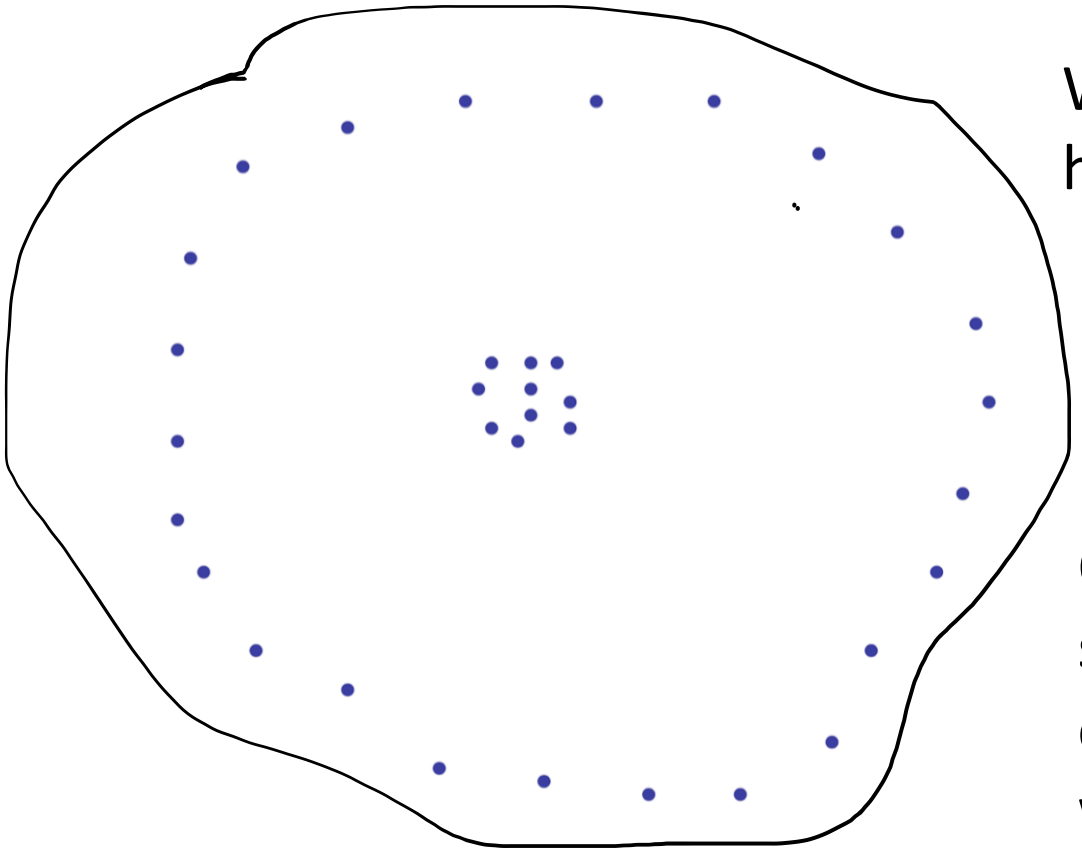
What will “min” do here?

Comprehension Question



What will “min” do here?

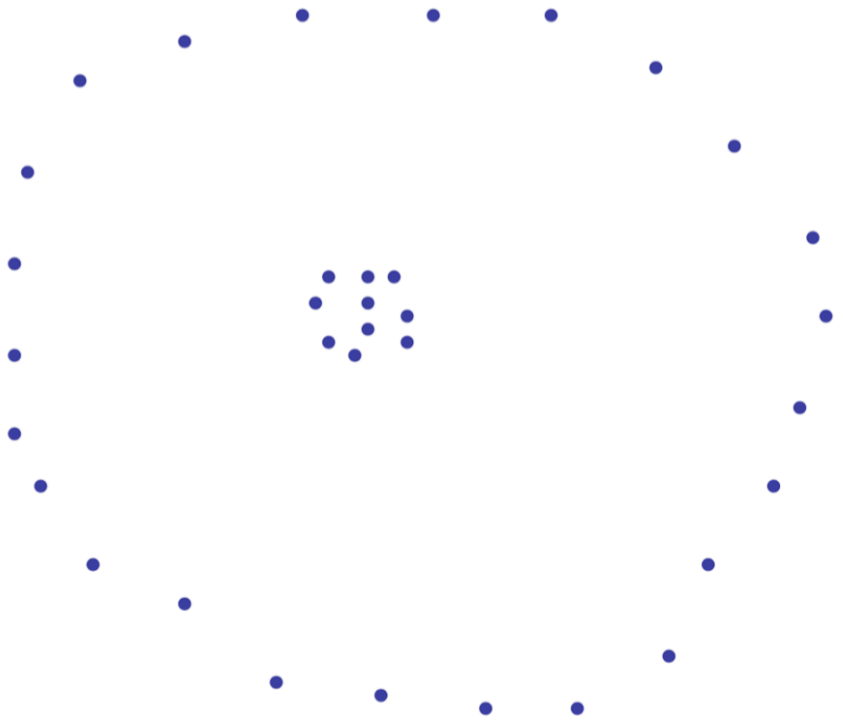
Comprehension Question



What will “min” do here?

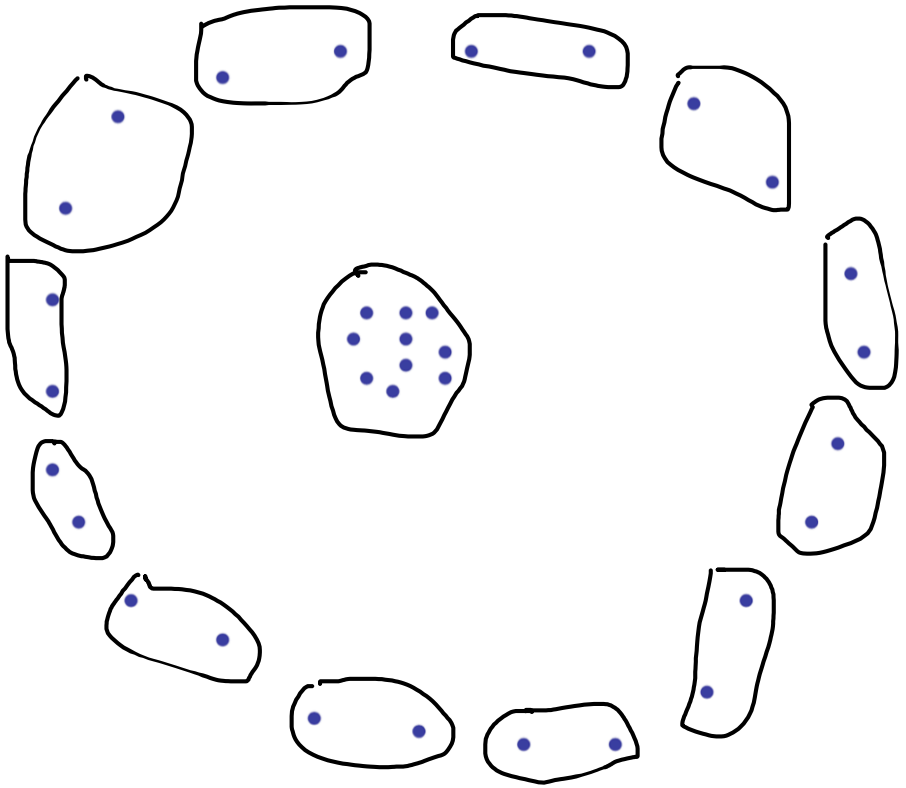
Only as last step would combine with center

Comprehension Question

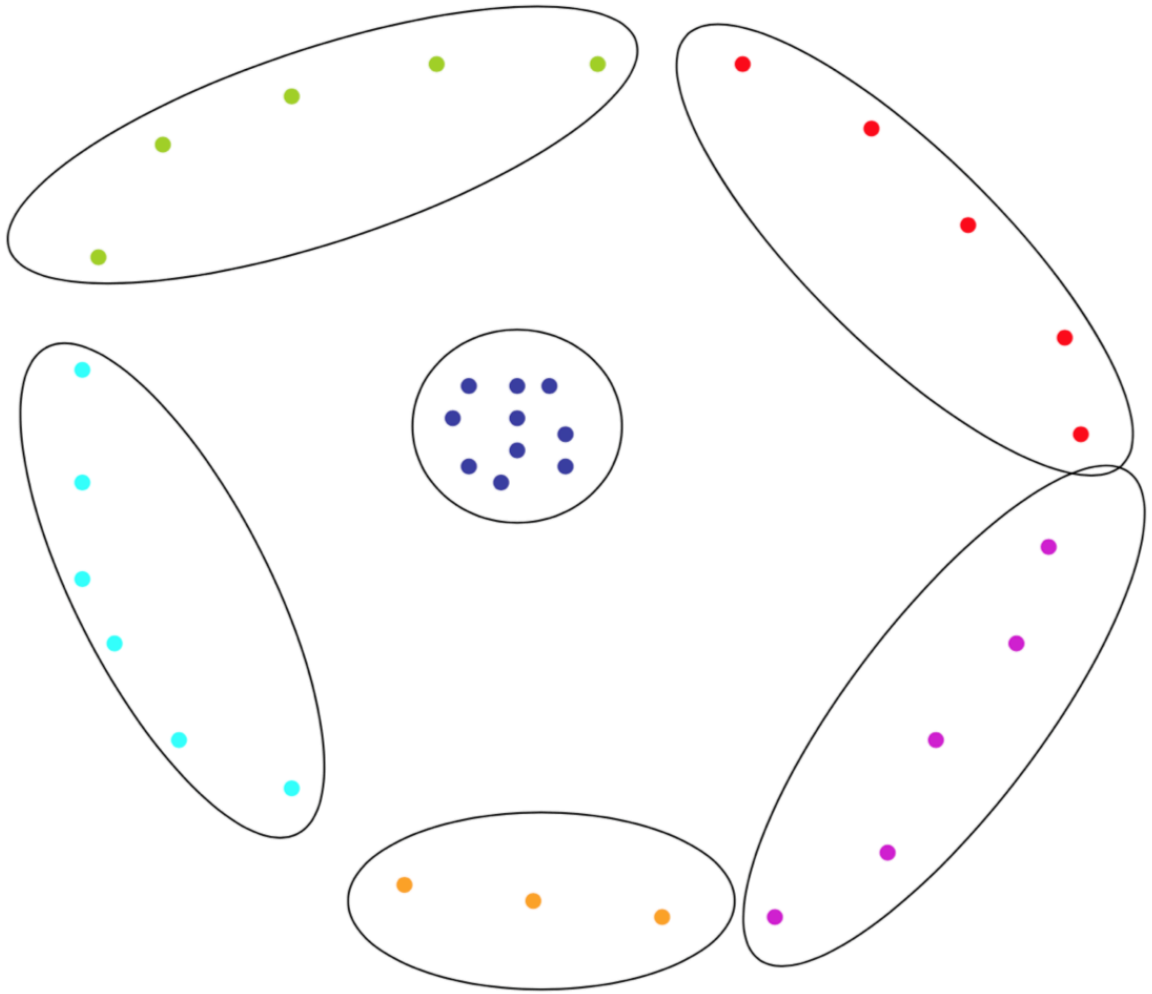


What will “max” do here?

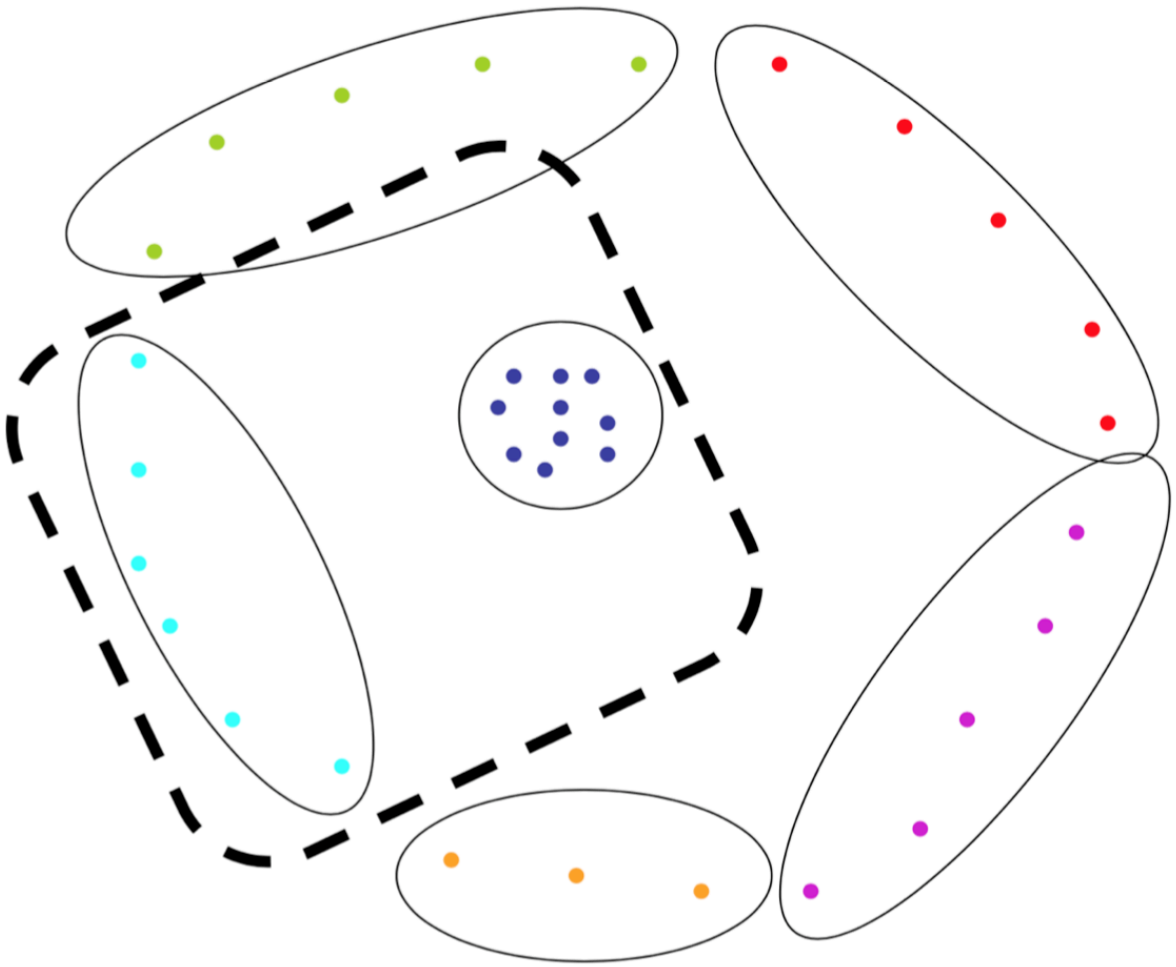
Comprehension Question



What will “max” do here?

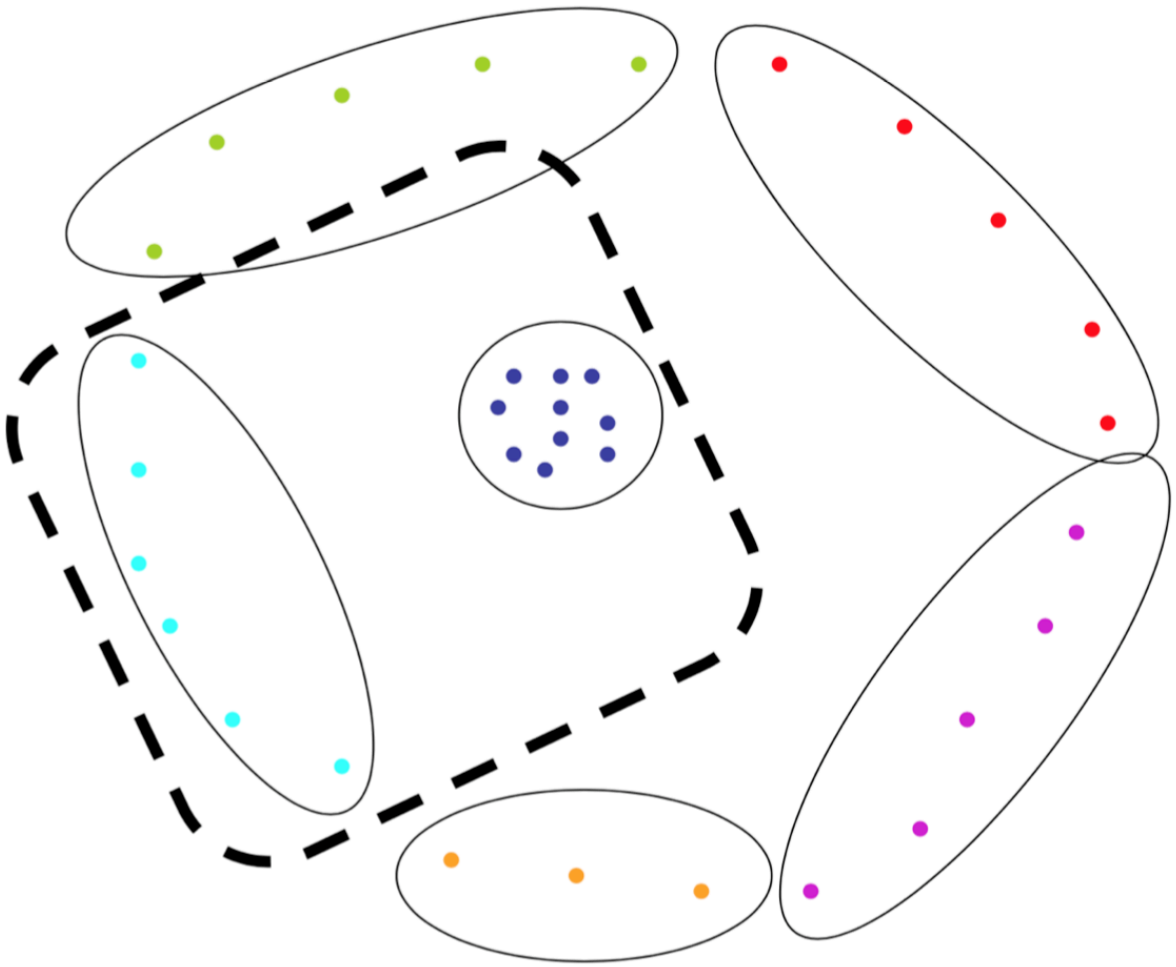


"max" linkage will get to this point



"max" linkage will get to this point

Next: combine outside clusters with center



"max" linkage will get to this point

Next: combine outside clusters with center

Eventually merge in additional outside clusters