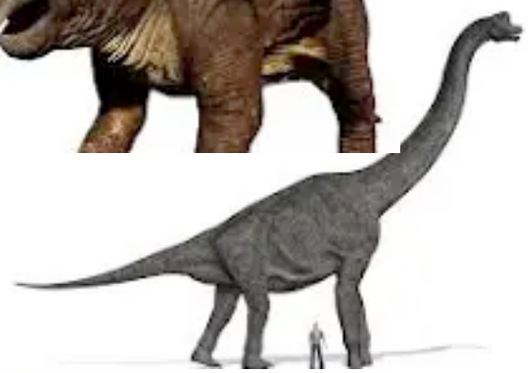
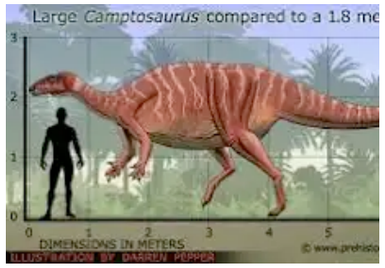
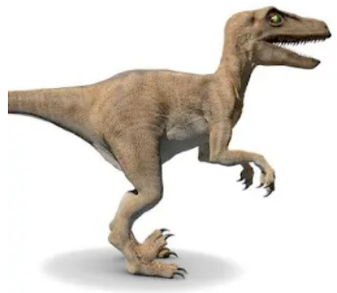



6'
3'
0'


Large Camptosaurus compared to a 1.8 me
DIMENSIONS IN METERS
ILLUSTRATION BY DARREN PEPPER



Hierarchical Agglomerative Clustering

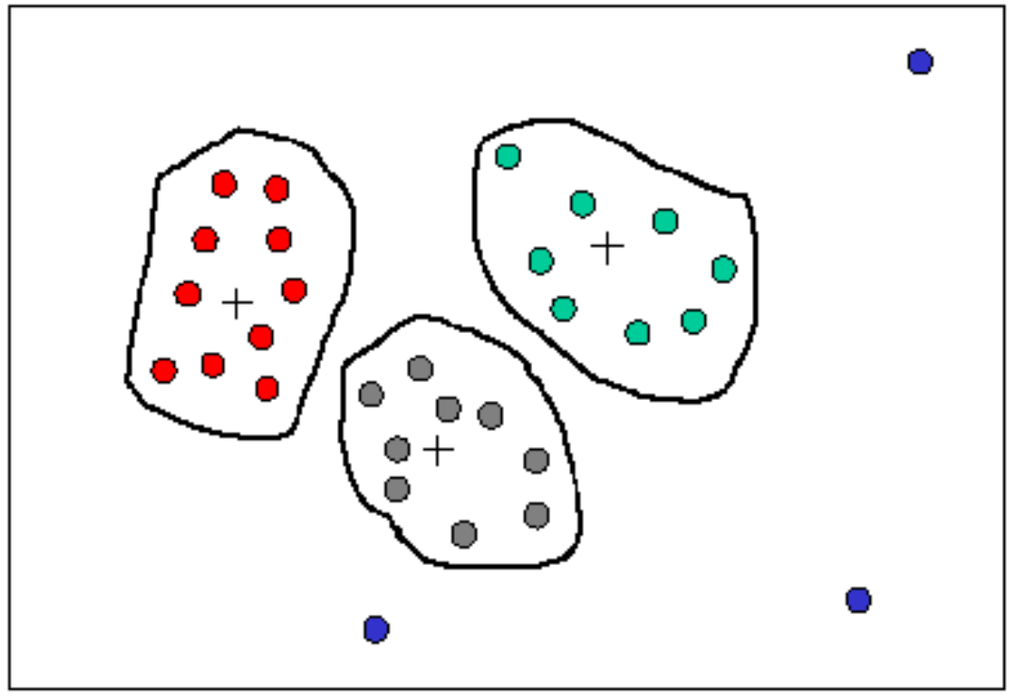
CS 337

What is Clustering?

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called **clusters**

Helps users understand the natural grouping or structure in a data set

- **Cluster:**
  - a collection of data objects that are “similar” to one another
  - and "different" from other clusters

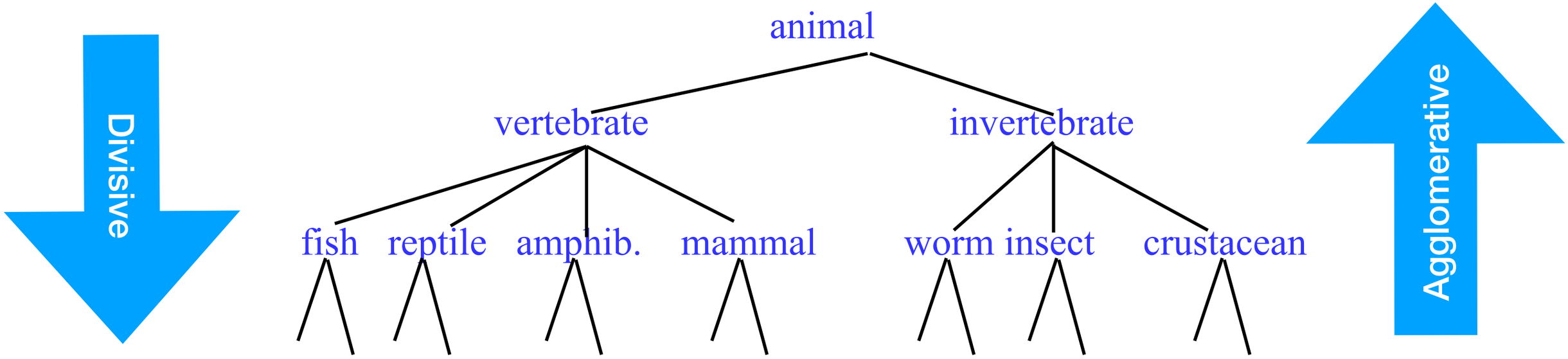


Clustering Methods

- *Agglomerative* (*bottom-up*)
  - iteratively combine clusters to form larger and larger clusters.
- *Divisive* (*partitional, top-down*) separate all examples into clusters. Separate clusters into subclusters.
  - Decision Trees
    - But they require labelled examples
- *K-Means* start with K "centers", assign data to nearest. Compute new centers. Repeat.
  - This week's lab!!!

Hierarchical Clustering

- Builds a tree-based hierarchical taxonomy (*dendrogram*) from a set of examples.



Hierarchical Agglomerative Clustering (HAC)

- Assumes a *similarity function* for determining the similarity of two instances.
- Repeatedly join the two clusters that are most similar until there is only one cluster.
- The history of merging forms a binary tree or hierarchy (dendrogram).

HAC Algorithm

Start with all instances in their own cluster.
Until there is only one cluster:
    Among the current clusters, determine the two clusters, c_i and c_j, that are most similar.
Replace c_i and c_j with a single cluster $c_i \cup c_j$

Cluster Similarity

- Assume a similarity function that determines the similarity of two instances: *sim*(*x*,*y*).
- How to compute similarity of two clusters each possibly containing multiple instances?
  - Single Link: Similarity of two most similar members.
  - Complete Link: Similarity of two least similar members.
  - Centroid: Average similarity between members.

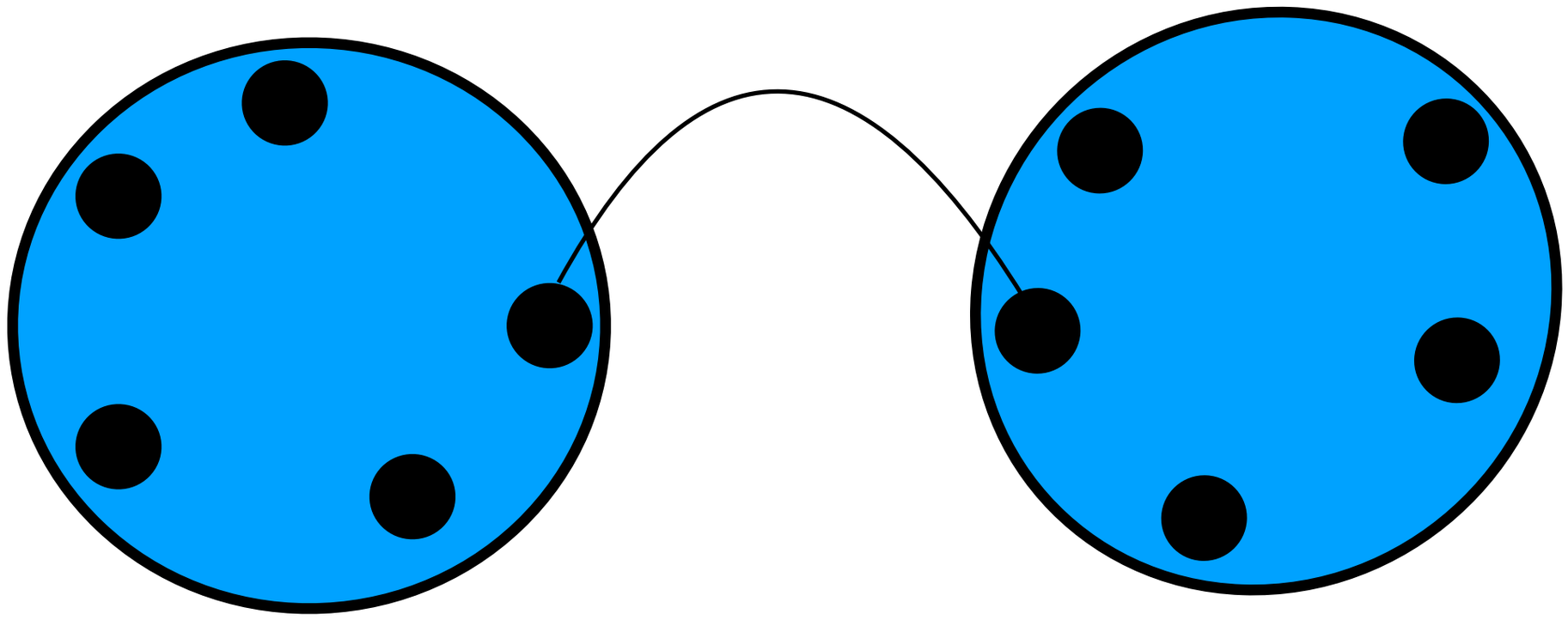
– Monotonically increasing!
  – Desirable but not true of every technique
  – Yes: Single Link, Complete Link
  – No (possibly): Centroid

Single Link Agglomerative Clustering

- Use maximum similarity of pairs:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Can result in "straggly" (long and thin) clusters due to *chaining effect*.

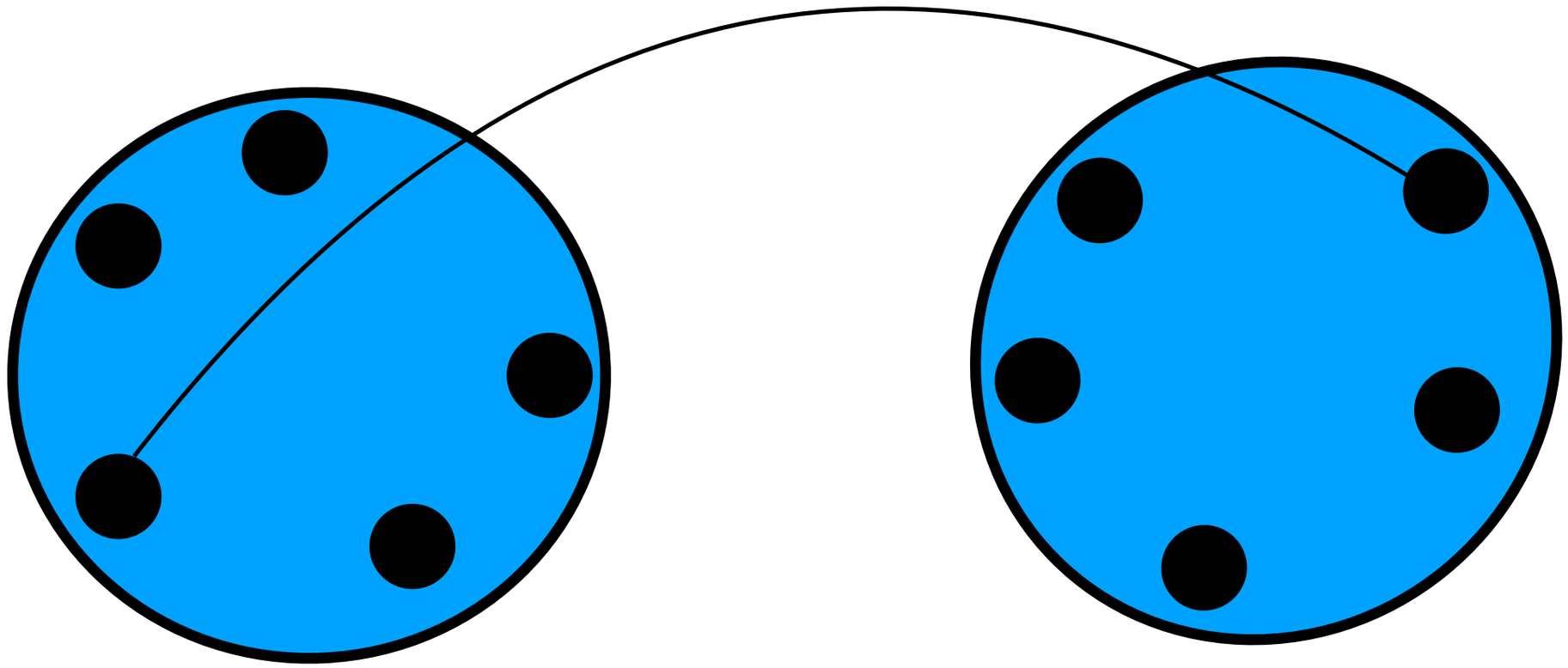


Complete Link Agglomerative Clustering

- Use minimum similarity of pairs:

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- Makes more "tight," spherical clusters that are typically preferable.



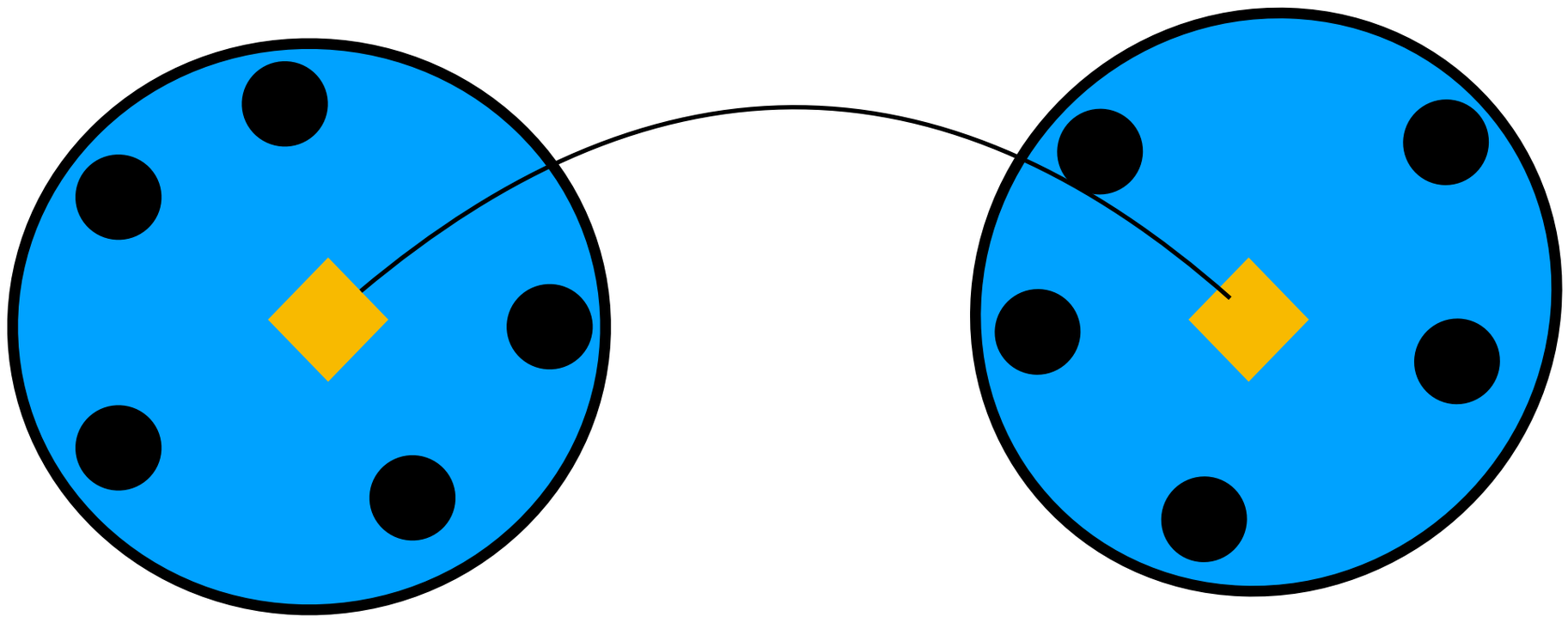
What is a "centroid"

- The average of the points in a cluster.
  - $c_z = (1/N_z)\ sum_{i=0..Nz}(d_i)$
    - where c_z is the centroid of cluster Z
    - N_z is number of docs in cluster
    - d_i is a document in cluster Z
  - "sum" does each dimension independently

	Dim1	Dim2	Dim3
Doc1	1	4	7
Doc2	17	20	-5
Doc3	10	10	10
Doc4	2	4	0
Doc5	5	2	3
C_z	7.0	8.0	3.0

Centroid Agglomerative Clustering

- distance to other clusters is the distance between centroids
  - Math: the distance from a document to a cluster centroid is average of distance from that document to each document in the cluster.
  - Every time you change a cluster, you have to recompute distance to every other cluster



Computational Complexity

- In the first iteration, all HAC methods need to compute similarity of all pairs of n individual instances which is $O(n^2)$.
- In each of the subsequent $n-2$ merging iterations, must compute the distance between the most recently created cluster and all other existing clusters.
  - Why n-2?
- Overall $O(n^3)$
  - can get $O(n^2)$ for single link

Stopping HAC

- HAC will continue until there is a single cluster
- Often this does not make sense
  - Cutoff clustering at a specified distance
  - Cutoff clustering when you get a big jump in distance
  - Cutoff when you have a set number of clusters

Distance Dendrograms -- Visualizing HAC

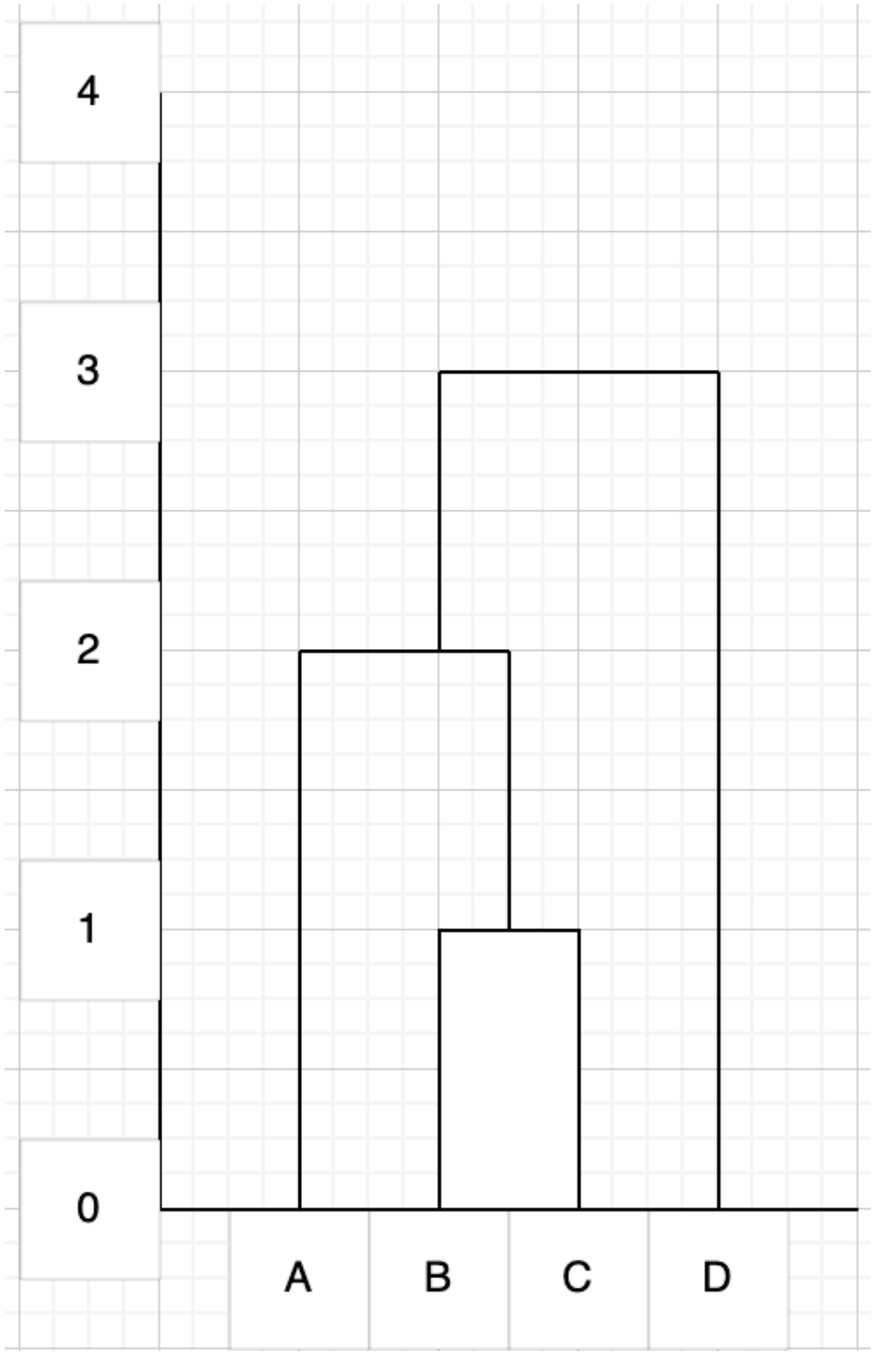
- Create a graph in which y axis is distance between clusters.
- Given monotonically increasing distance get a nice looking graphic
  - Each time you merge a cluster
    - draw two vertical lines from last cluster distance to new cluster distance
    - draw a horizontal line connecting vertical lines at distance between merged clusters
- clusters all start at 0.

Dendrograms --Example

- Suppose 4 items and using single link distance. Distances are in the table

	A	B	C	D
A	--	2	4	6
B		--	1	7
C			--	3

- Step 1: Group together B,C with cost 1
- Step 2: Group A, (B,C) with cost 2
- Step 3: Group D, (A,B,C) with cost 3



Exercise

	a	b	c	d	e	f	g
a	-	-	-	-	-	-	-
b	87	-	-	-	-	-	-
c	92	91	-	-	-	-	-
d	74	22	60	-	-	-	-
e	24	21	28	17	-	-	-
f	6	31	90	17	8	-	-
g	5	3	78	8	56	2	-
h	54	47	15	65	61	80	47

Suppose you have 7 items (a-g) with distances between items in table above.

Show the dendrogram for single link and complete-link clustering for the above data

Clustering Issues

- Suppose you have a feature for which distance is not defined (or at least not well-defined)
- Features on different ranges
- Features with same range but different variance

Normalization

- Ensure features have same mean and variance
- Why?

```
Algorithm
Given D -- an NxM array where N is the number of independent data
items and M is the dimensionality of the items
Return -- an NxM array in which the original date has been transforme
to have 0 mean and unit variance
Normalize(D):
   let R = an NxM array, initially 0
   for m in 0..(M-1):
      let av = average of the N items on feature m
      let sd = standard deviation of N items on feature m
      for n in 0..(N-1):
          R[n][m] = (D[n][m]-av)/sd
    return R
```

Similarity Measure

- A similarity measure is a function that computes the *degree of similarity* between two vectors.

- Using a similarity measure between two vectors:
  - It is possible give a total ordering of the distances between a target vector
  - a set of vectors
  - It is possible to enforce a certain threshold so that the size of the retrieved set can be controlled.

Similarity Measure
Inner Product

- Similarity between vectors for the document $\boldsymbol{d}_i$ and query $\boldsymbol{q}$ can be computed as the vector inner product (a.k.a. dot product):

$$\text{sim}(\boldsymbol{d}_j, \boldsymbol{q}) = \boldsymbol{d}_j \bullet \boldsymbol{q} = \sum_{i=1}^{t} w_{ij} w_{iq}$$

  where w_{ij} is the weight of term i in document j and w_{iq} is the weight of term i in the query

- For binary vectors, the inner product is the number of matched query terms in the document (size of intersection).
- For weighted term vectors, it is the sum of the products of the weights of the matched terms.

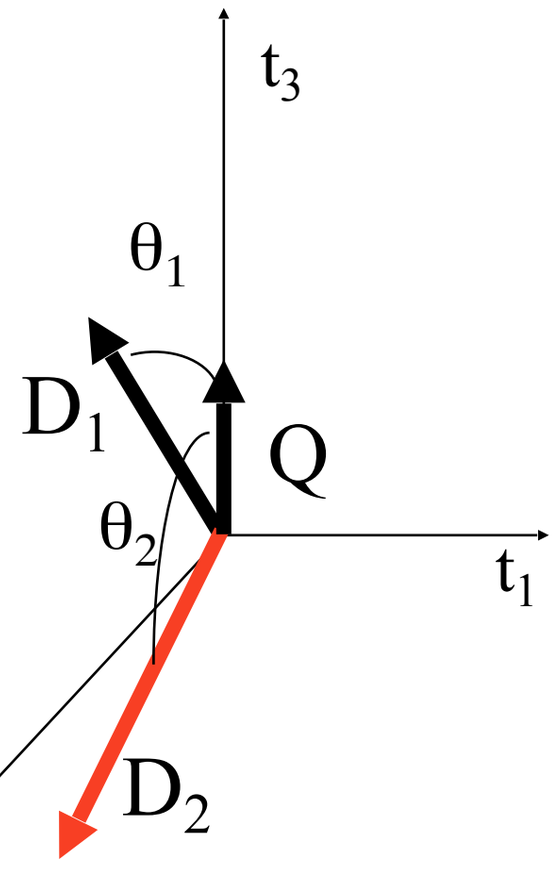
Properties of Inner Product

- The inner product is unbounded.
- Works best when features are either binary or binary-like
  - 0 indicates absence / false
  - Requires that features are strictly non-negative

Cosine Similarity Measure

- Cosine similarity measures the cosine of the angle between two vectors.
- Inner product normalized by the vector lengths.

$$\text{CosSim}(\mathbf{d}_j, \mathbf{q}) = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$



Suppose that VS has only 3 terms (dimensions) A, B, and C, corresponding to t_1, t_2 and t_3 in graph

$D_1 = (2, 3, 5)$
$D_2 = (3, 7, 1)$
$Q = (0, 0, 2)$

$\text{CosSim}(D_1, Q) = 10 / \sqrt{(4+9+25)(0+0+4)} = 0.81$
$\text{CosSim}(D_2, Q) = 2 / \sqrt{(9+49+1)(0+0+4)} = 0.13$

D_1 is 6 times better than D_2 using cosine similarity but only 5 times better using inner product.

"Norms"

- Properties of norms:
  - Non-negativity: It should always be non-negative.
  - Definiteness: It is zero if and only if the vector is zero, i.e., zero vector.
  - Triangle inequality: The norm of a sum of two vectors is no more than the sum of their norms.
  - Homogeneity: Multiplying a vector by a scalar multiplies the norm of the vector by the absolute value of the scalar.
- Given 2 vectors X and Y
  - Euclidian Norm (2-Norm)
    - sqrt(sum of squares of X_i-Y_i)
  - 1-Norm (Manhattan Distance)
  - Infinity-Norm
    - conceptually similar to complete link
  - P-Norm