

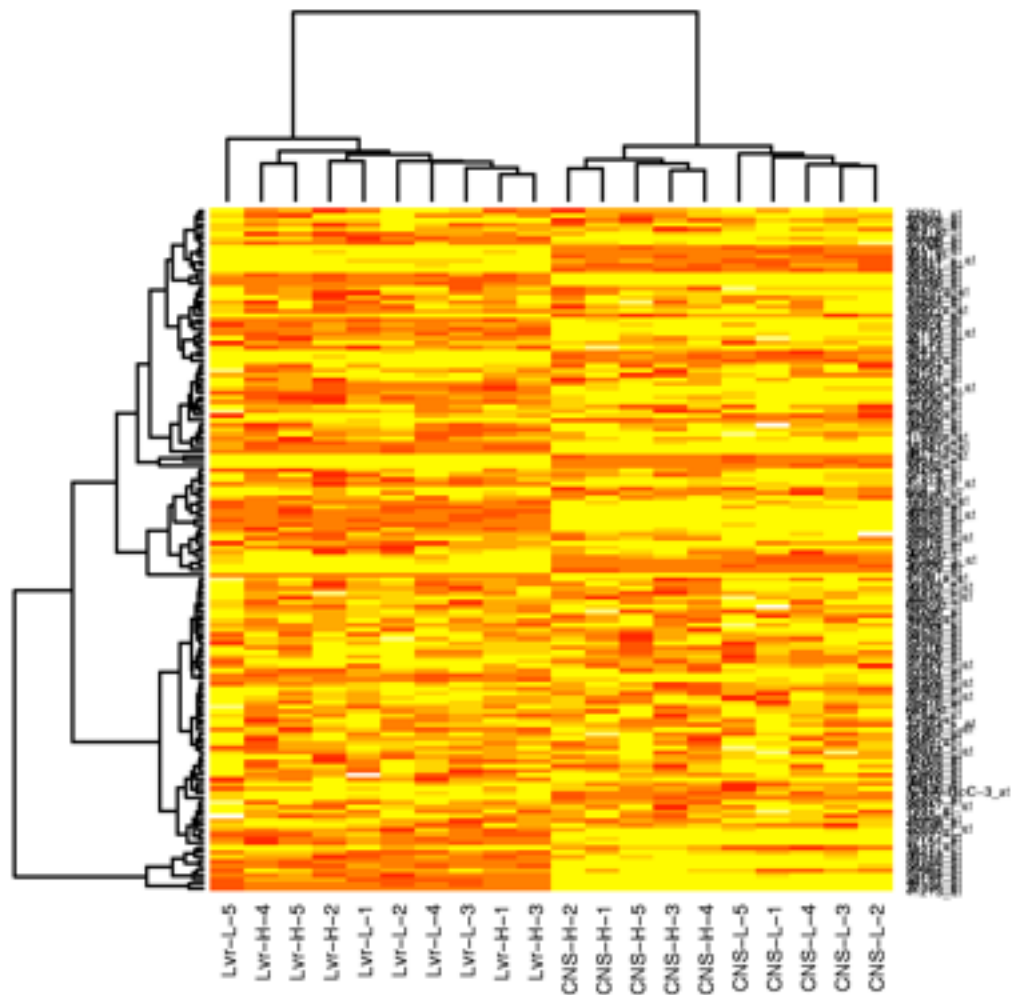
Clustering Gene Expression

CMSC423 Spring 2014
Héctor Corrada Bravo

Outline

- **K-means (and K-medoids) clustering**
- **Model-Based clustering (soft K-means)**
 - **EM algorithm**

Heatmaps



Inspection of the genes differentially expressed between MLL and ALL is instructive (Fig. 1). Many genes underexpressed in MLL have a function in early B-cell development. These include genes expressed in early B cells^{14,15}, *MME*, *CD24*, *CD22*

Heatmap showing the distribution of 1000 genes across 1000 samples, categorized into ALL and MLL groups. The color scale ranges from -3.0 (blue) to 3.0 (red).

Gene list (from top to bottom):

- CD7-2
- ALDH3B1
- CD7-3
- ADAM27A
- CD7-4
- CD7-5
- CD7-6
- CD7-7
- CD7-8
- CD7-9
- CD7-10
- CD7-11
- CD7-12
- CD7-13
- CD7-14
- CD7-15
- CD7-16
- CD7-17
- CD7-18
- CD7-19
- CD7-20
- CD7-21
- CD7-22
- CD7-23
- CD7-24
- CD7-25
- CD7-26
- CD7-27
- CD7-28
- CD7-29
- CD7-30
- CD7-31
- CD7-32
- CD7-33
- CD7-34
- CD7-35
- CD7-36
- CD7-37
- CD7-38
- CD7-39
- CD7-40
- CD7-41
- CD7-42
- CD7-43
- CD7-44
- CD7-45
- CD7-46
- CD7-47
- CD7-48
- CD7-49
- CD7-50
- CD7-51
- CD7-52
- CD7-53
- CD7-54
- CD7-55
- CD7-56
- CD7-57
- CD7-58
- CD7-59
- CD7-60
- CD7-61
- CD7-62
- CD7-63
- CD7-64
- CD7-65
- CD7-66
- CD7-67
- CD7-68
- CD7-69
- CD7-70
- CD7-71
- CD7-72
- CD7-73
- CD7-74
- CD7-75
- CD7-76
- CD7-77
- CD7-78
- CD7-79
- CD7-80
- CD7-81
- CD7-82
- CD7-83
- CD7-84
- CD7-85
- CD7-86
- CD7-87
- CD7-88
- CD7-89
- CD7-90
- CD7-91
- CD7-92
- CD7-93
- CD7-94
- CD7-95
- CD7-96
- CD7-97
- CD7-98
- CD7-99
- CD7-100

exactly the germinal centre phenotype *in vitro*, as determined by the failure of a variety of activation conditions to induce the expression of BCL-6 protein, a highly specific marker for germinal centre B

signature of germinal centre B cells was reproduced virtually unchanged in FL, supporting the view that this lymphoma arises from this stage of B-cell differentiation (Fig. 2).

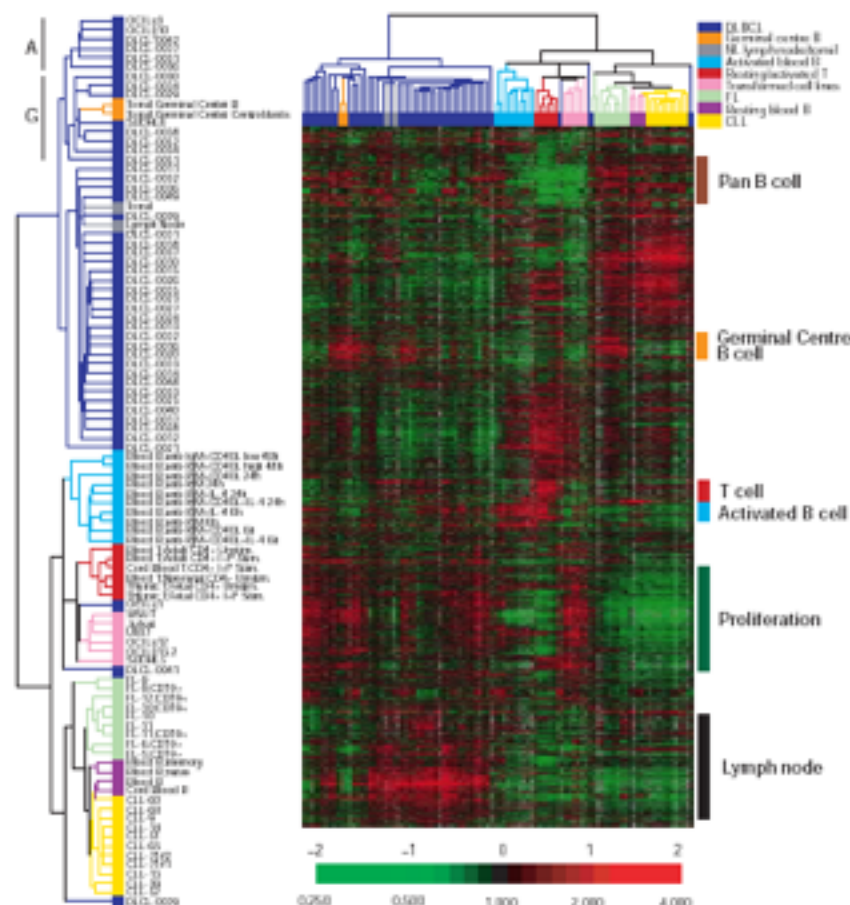
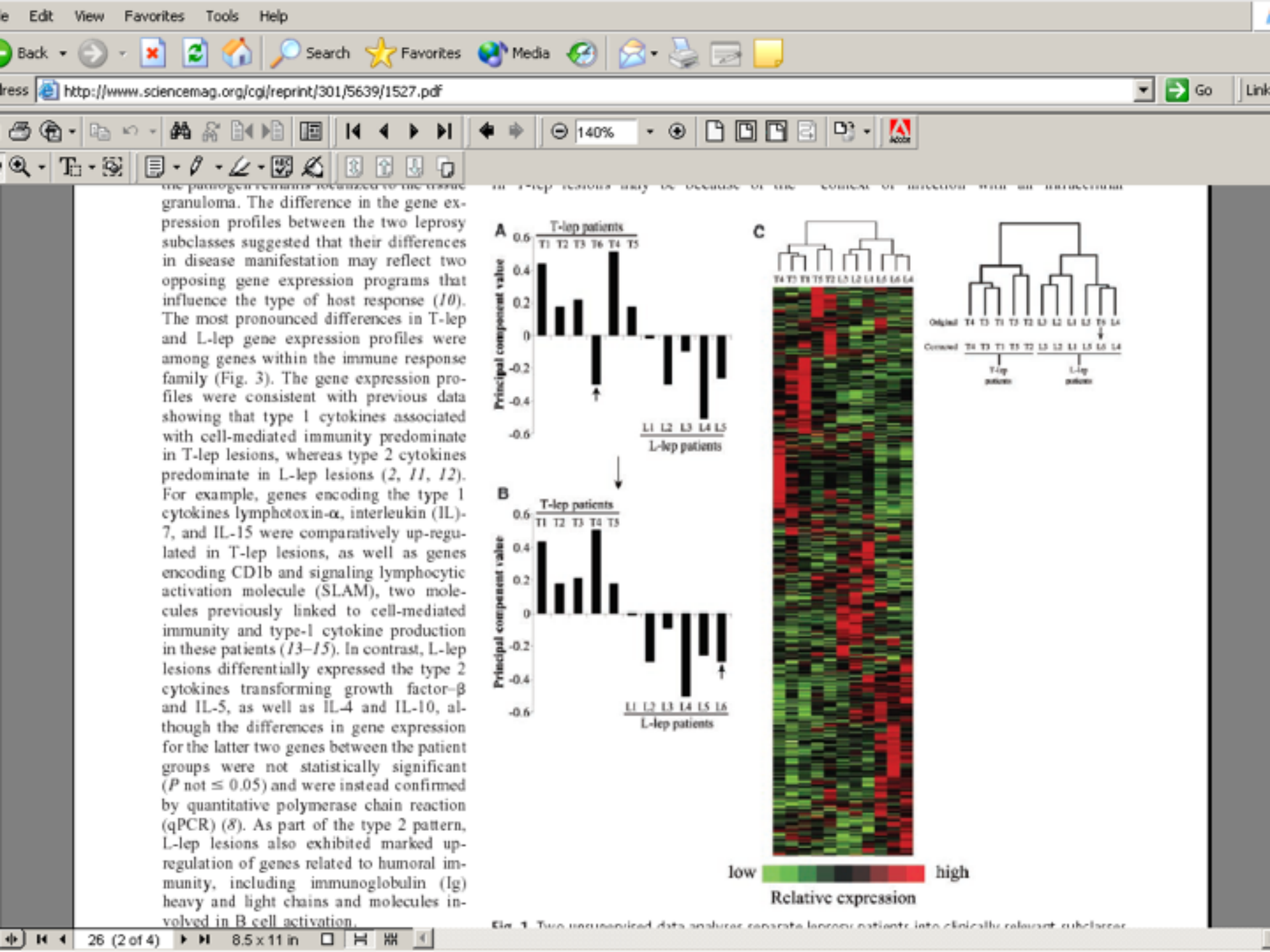
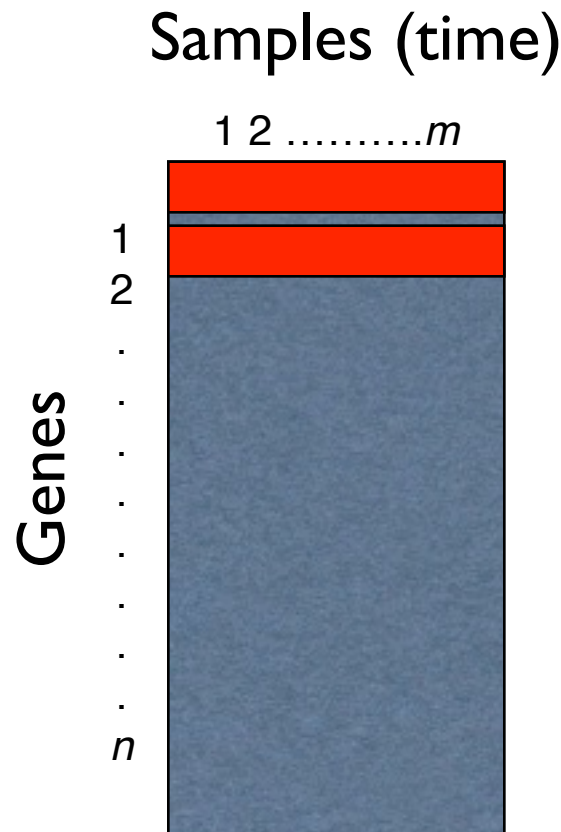


Figure 1 Hierarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression from 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of the relatedness of gene expression in each sample. The dendrogram is colour coded according to the category of mRNA sample studied (see

hybridization of fluorescent cDNA probes prepared from each experimental mRNA samples to a reference mRNA sample. These ratios are a measure of relative gene expression in each experimental sample and were depicted according to the colour scale shown at the bottom. As indicated, the scale extends from fluorescence ratios of 0.25 to 4 (-2 to +2 in log base 2 units). Grey indicates missing or excluded data. See



Measurements



Measurement is
expression of
gene i at time j

DATA MATRIX

Points

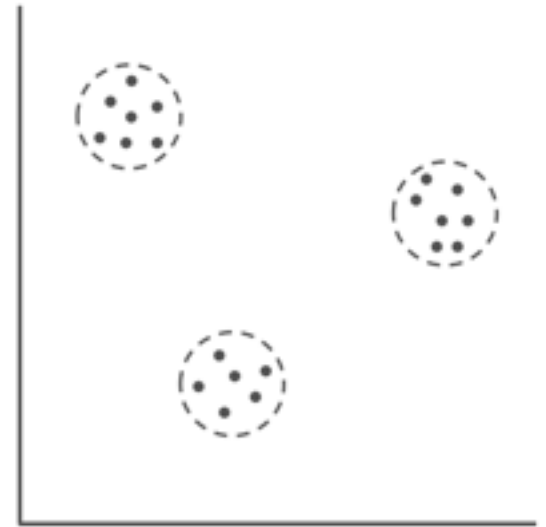
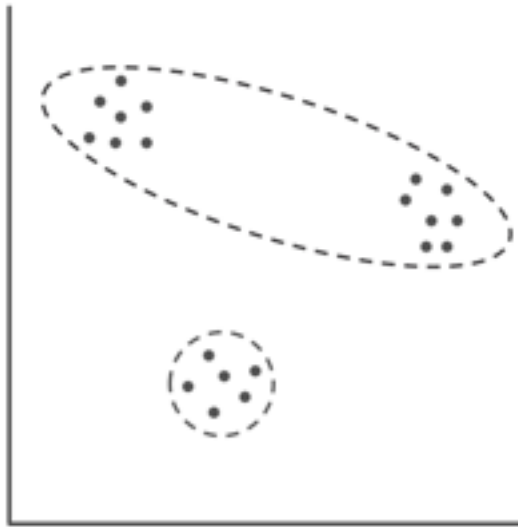
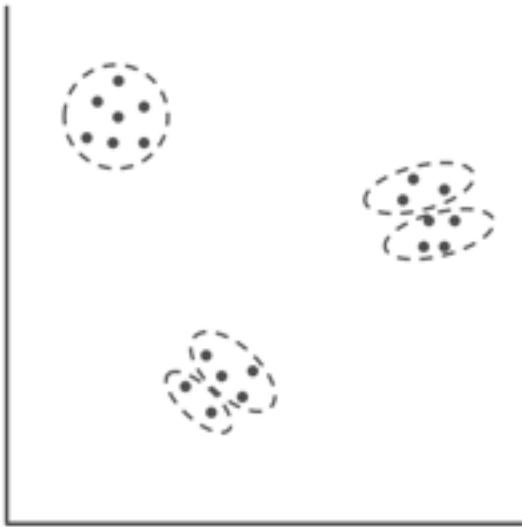
- **Gene1 = $(E_{11}, E_{12}, \dots, E_{1N})'$**
- **Gene2 = $(E_{21}, E_{22}, \dots, E_{2N})'$**
- **Sample1 = $(E_{11}, E_{21}, \dots, E_{G1})'$**
- **Sample2 = $(E_{12}, E_{22}, \dots, E_{G2})'$**
- **E_{gi} = expression gene g , sample i**

Distance

- **Clustering organizes things that are *close* into groups**
- **What does it mean for two genes to be close?**
- **What does it mean for two samples to be close?**
- **Once we know this, how do we define groups?**

Clustering

- *Separation and homogeneity*

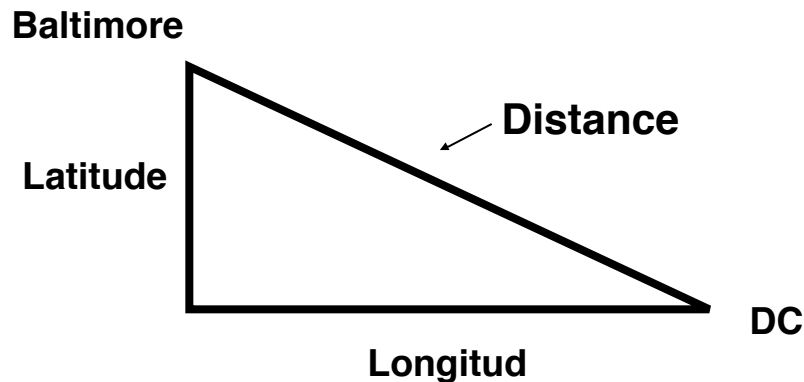


Clustering Problem

- *Partition a set of expression vectors into clusters*
- Input: An $n \times m$ gene expression matrix E
- Output: Clusters of the n expression vectors from E satisfying the conditions of homogeneity and separation

Most Famous Distance

- **Euclidean distance**
 - Example distance between gene 1 and 2:
 - Sqrt of Sum of $(E_{1i} - E_{2i})^2, i=1, \dots, N$
- **When N is 2, this is distance as we know it:**



When N is 20,000 you have to think abstractly

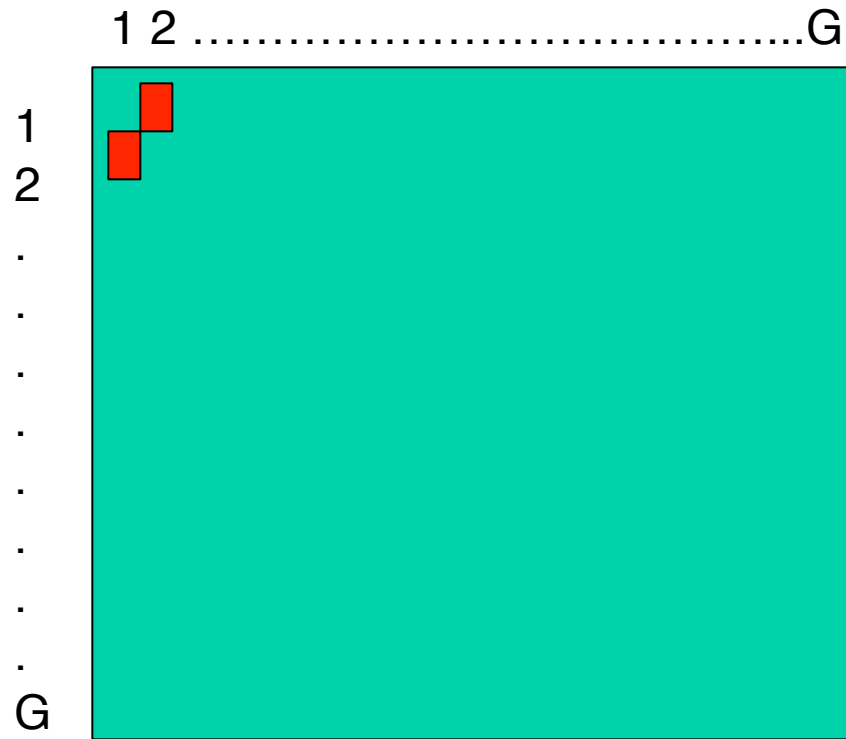
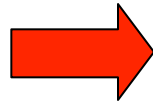
Similarity

- Instead of distance, clustering can use *similarity*
- If we standardize points then Euclidean distance is equivalent to using absolute value of correlation as a similarity index
- Other examples:
 - Spearman correlation
 - Categorical measures

The similarity/distance matrices

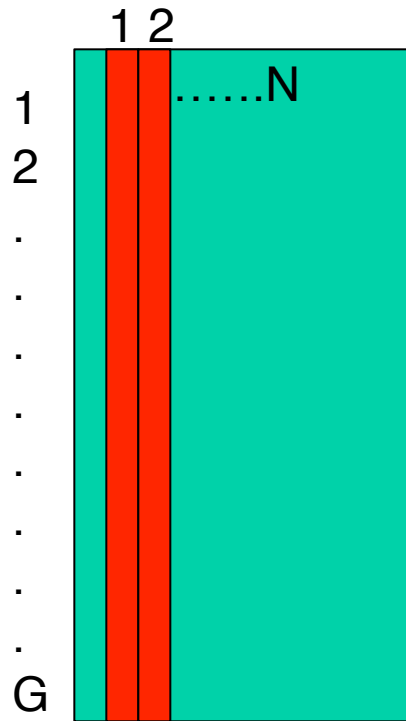


DATA MATRIX

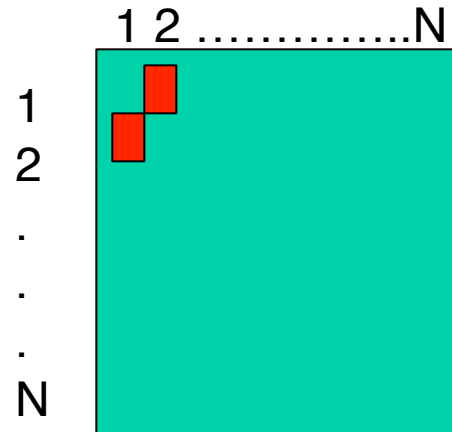
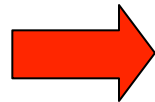


GENE SIMILARITY MATRIX

The similarity/distance matrices



DATA MATRIX



SAMPLE SIMILARITY MATRIX

K-center Clustering Problem

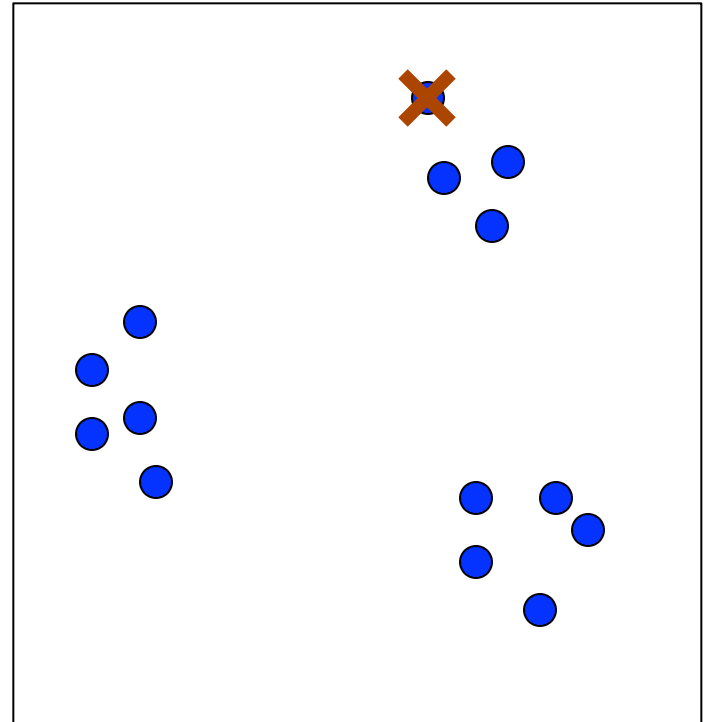
- *Given a set of data points, find k centers minimizing the maximum distance between these data points and centers*
- **Input:** A set of points $Data$ and an integer k
- **Output:** A set X of k centers that minimizes $MaxDistance(Data, X)$ over all possible choices of X .

Furthest Traversal

- Choose some point in Data as center
- While more centers needed:
 - Select the point *farthest* from current centers as next center

Furthest Traversal

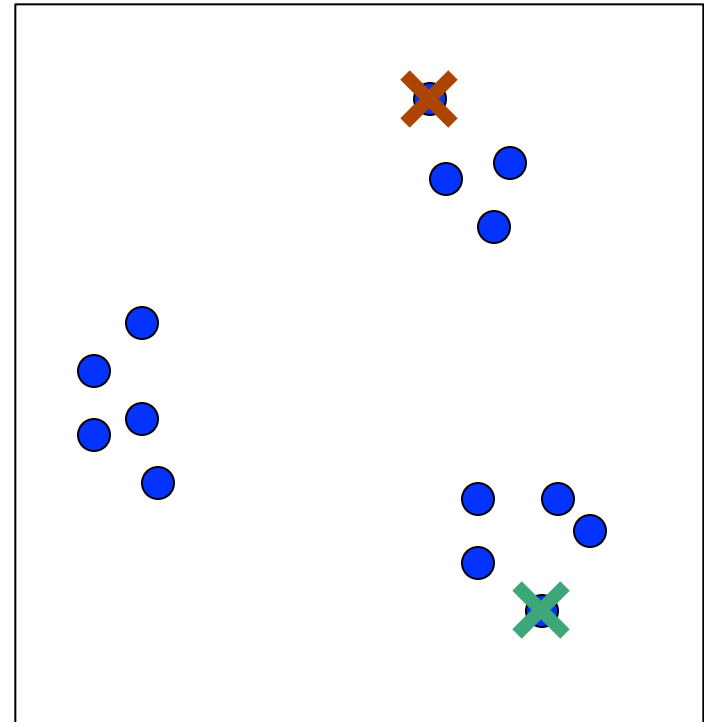
- **Choose 1 point as center**
- **This is arbitrary**



Iteration = 0

Furthest Traversal

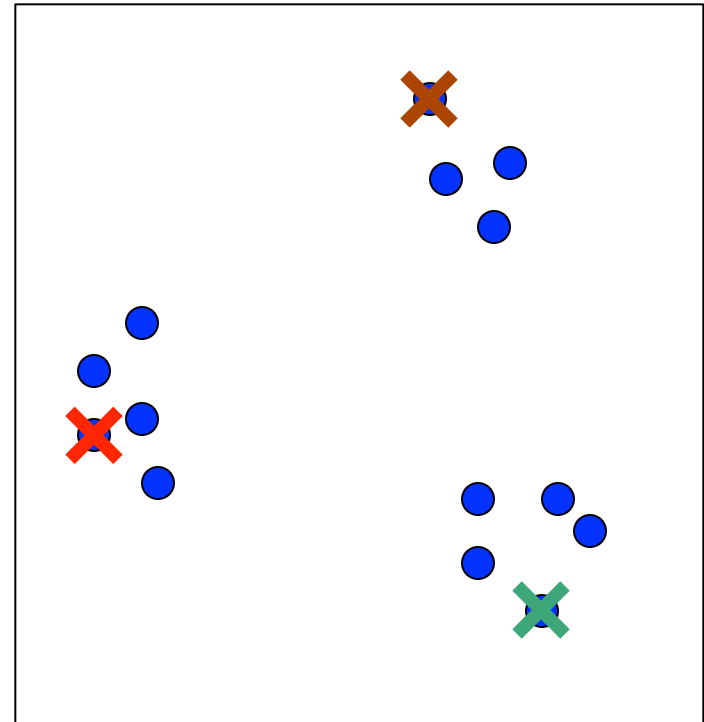
- Choose point farthest from current set of centers as next center



Iteration = 1

Furthest Traversal

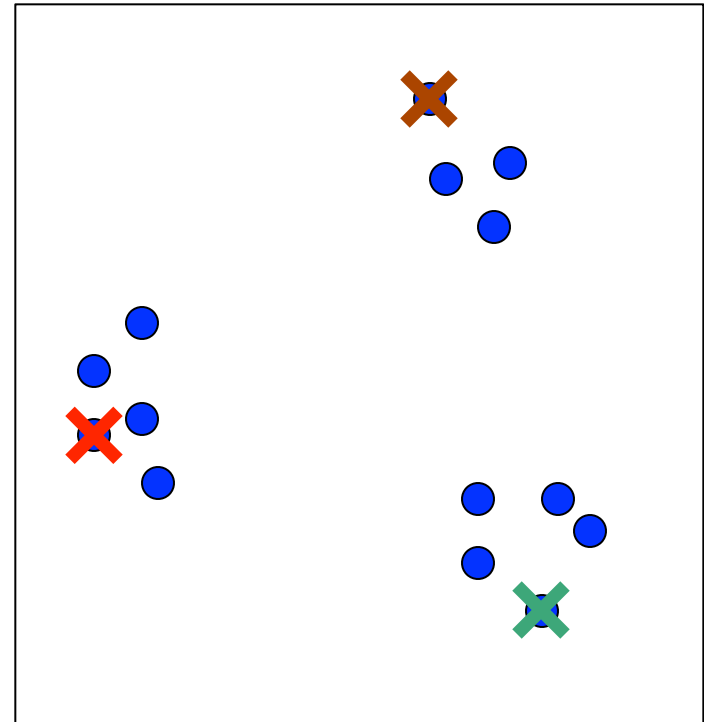
- Choose point farthest from current set of centers as next center



Iteration = 2

Furthest Traversal

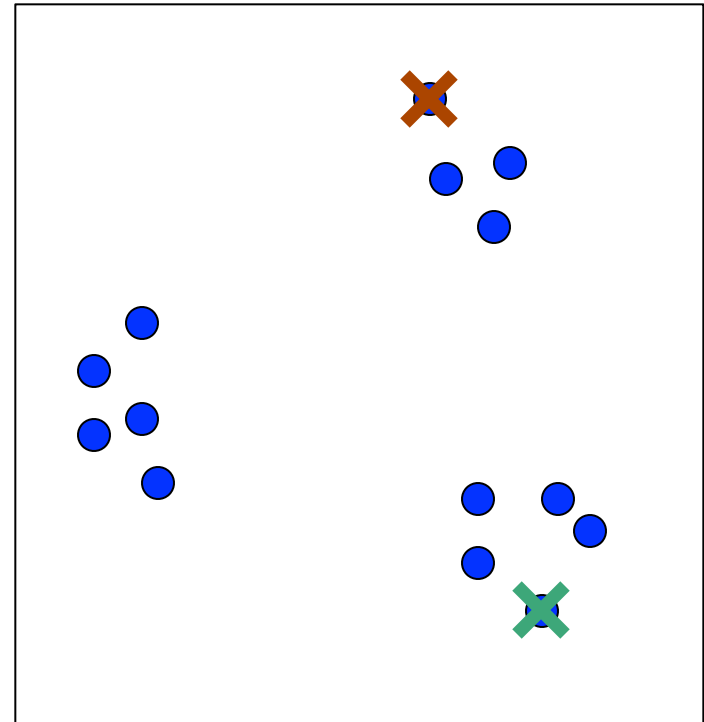
- Choose point farthest from current set of centers as next center



Iteration = 2

Furthest Traversal

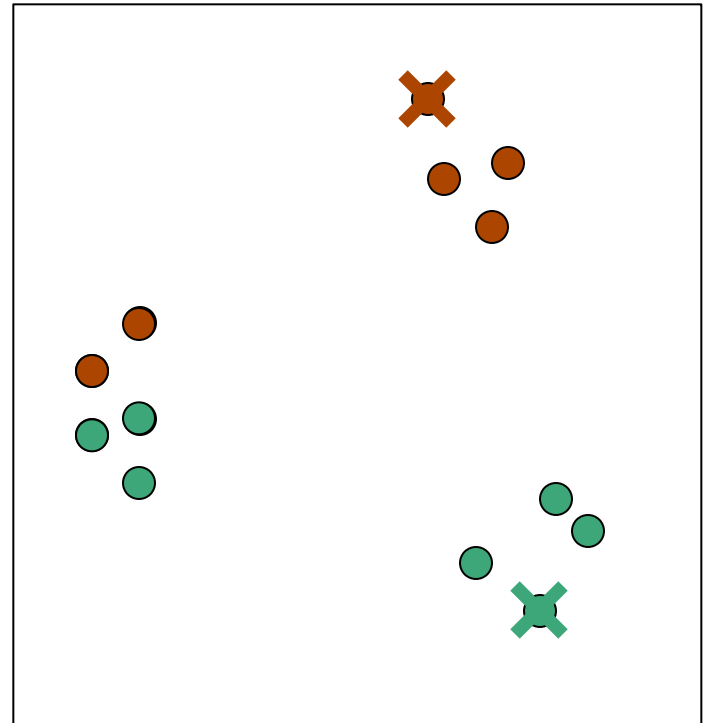
- Choose point farthest from current set of centers as next center



Iteration = 2

Furthest traversal

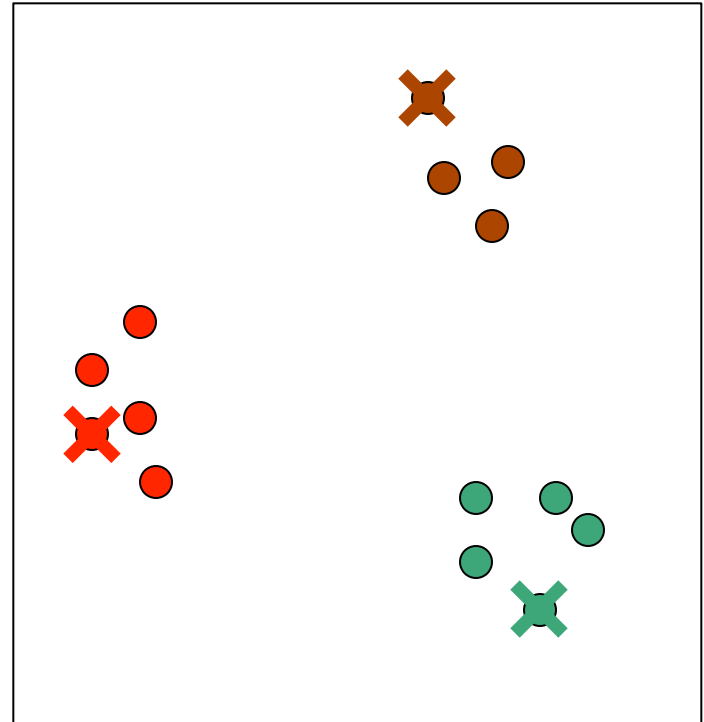
- Assign each point to its nearest center



Iteration = 2

Furthest traversal

- Assign each point to its nearest center



Iteration = 2

Analysis

- Running time?
- The ‘how good is it theorem’:
 - Let X_{opt} be an optimal set of centers
 - what does that mean?
 - and let X_{ft} be the solution given by *furthest traversal*
 - *then*
$$\text{MaxDistance}(\text{Data}, X_{ft}) \leq 2 * \text{MaxDistance}(\text{Data}, X_{opt})$$
- Why is this not a good algorithm to use?

Better center choice

- Instead of using data points themselves as centers
- We can do better by choosing centers that *are not in the dataset*

Distortion

$$\text{Distortion}(\text{Data}, X) = \frac{1}{n} \sum_{y \in \text{Data}} d(y, X)^2$$

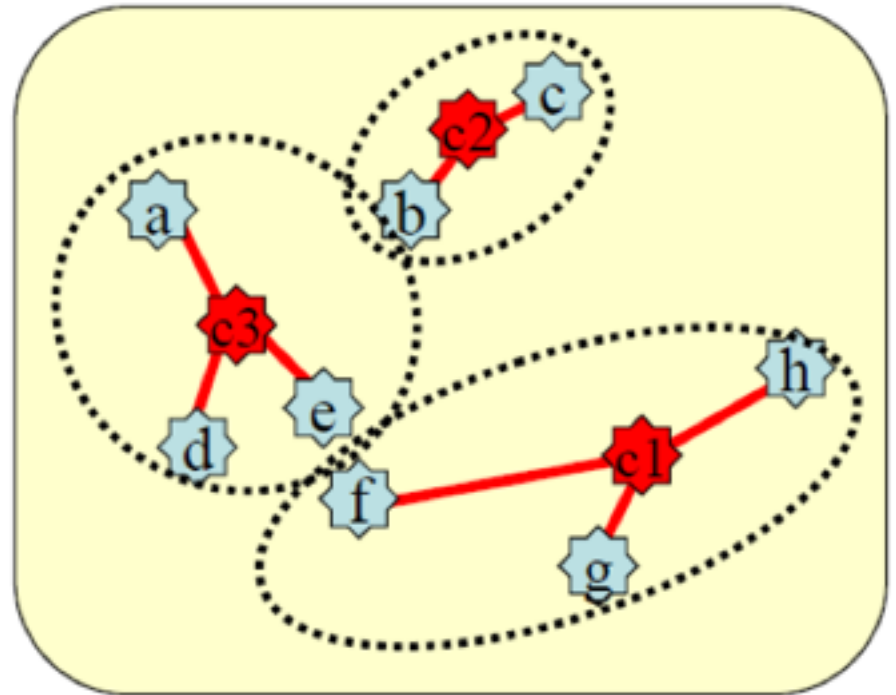
Center of gravity:

Given a set of points, what is the *center* that minimizes *distortion*?

Construct the *center* by taking the *mean* of each coordinate

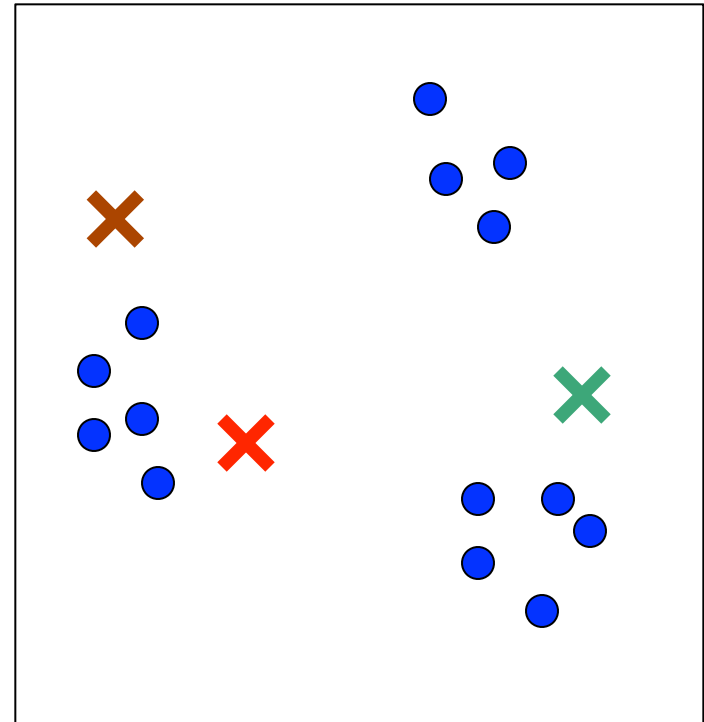
K-means problem:

minimize Distortion
instead of MaxDistance



K-means

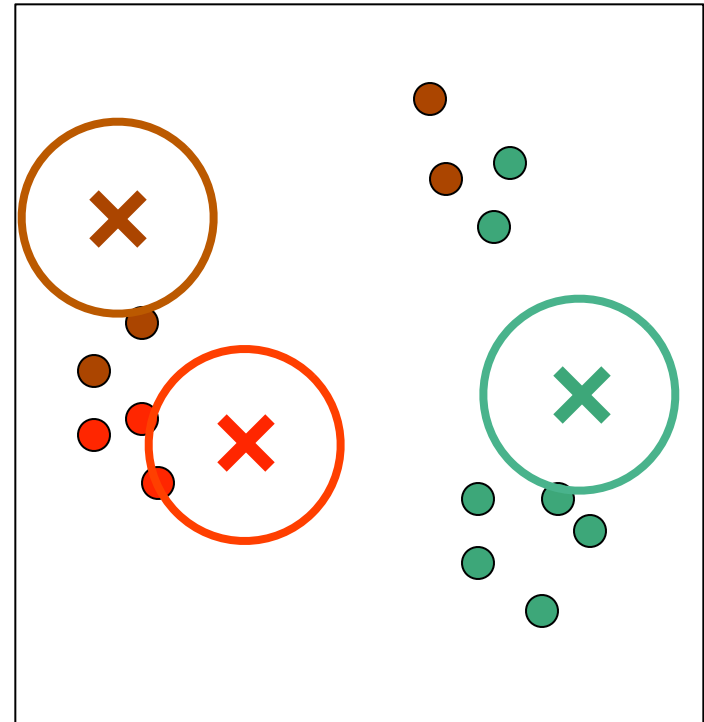
- Choose K *centroids*
- These are starting values that the user picks.
- There are some data driven ways to do it



Iteration = 0

K-means

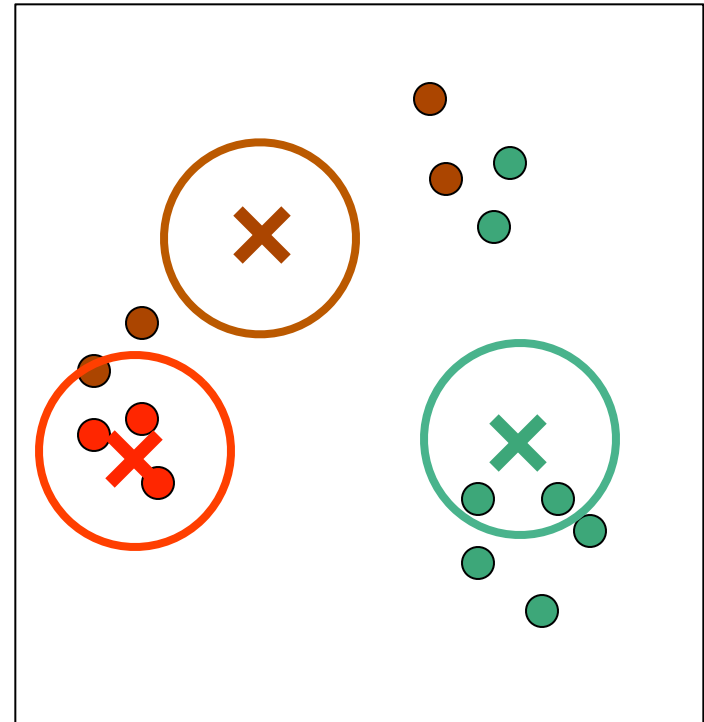
- Make first *partition* by finding the closest centroid for each point
- This is where distance is used



Iteration = 1

K-means

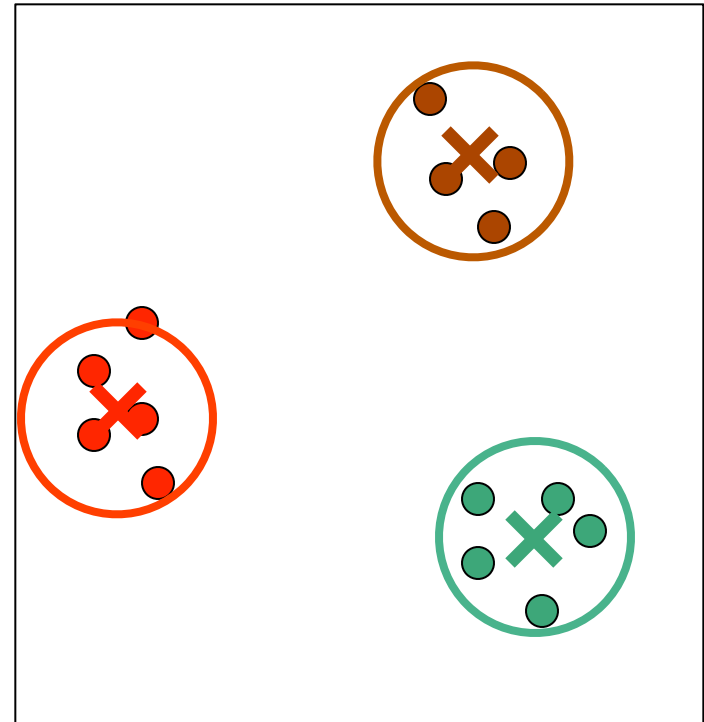
- Now re-compute the centroids by taking the *middle* of each cluster



Iteration = 2

K-means

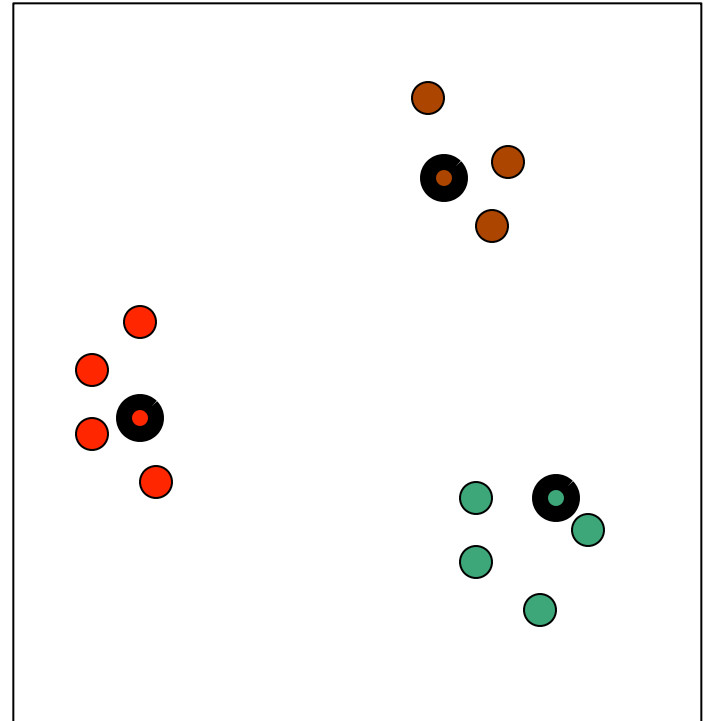
- Repeat until the centroids stop moving or until you get tired of waiting



Iteration = 3

K-medoids

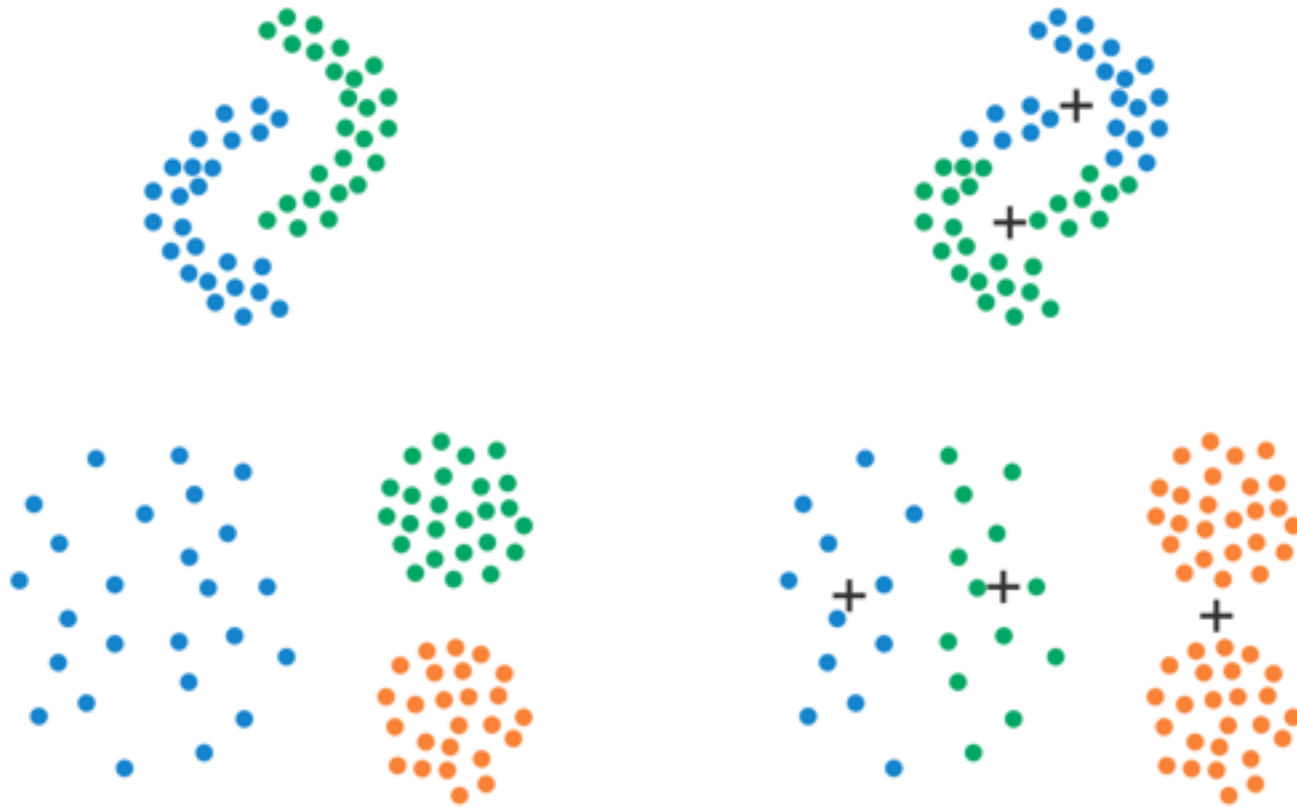
- A little different
- **Centroid:** The average of the samples within a cluster
- **Medoid:** The “representative object” within a cluster.
- Initializing requires choosing medoids at random.



K-means Limitations

- **Final results depend on starting values**
- **How do we choose K? There are methods but not much theory saying what is best.**

K-means limitations

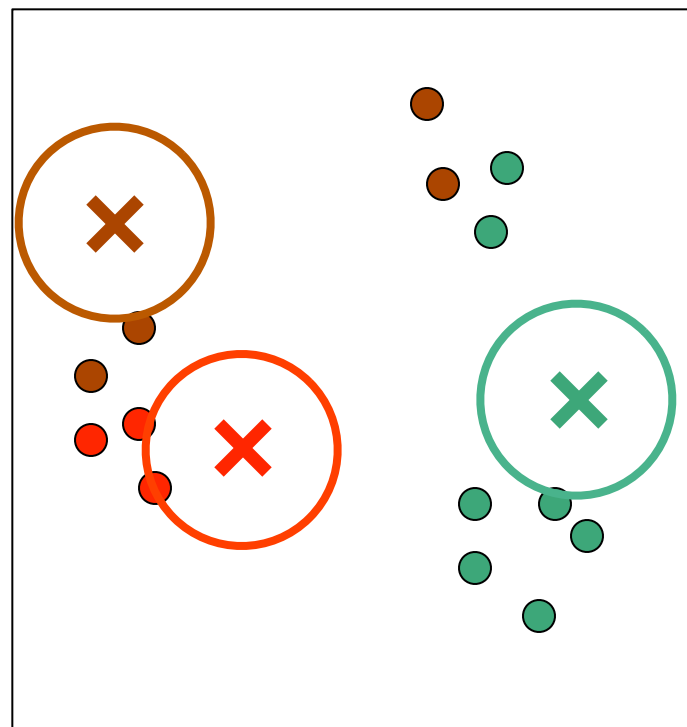


Analysis

- Does it converge?

Fuzzy K-means Clustering

- No *partitions* now
- Assumption:
 - *What we care to estimate are the centers, not the partitions*
 - *So, let's use all points to estimate each center, but weigh them by how likely they belong to that cluster*

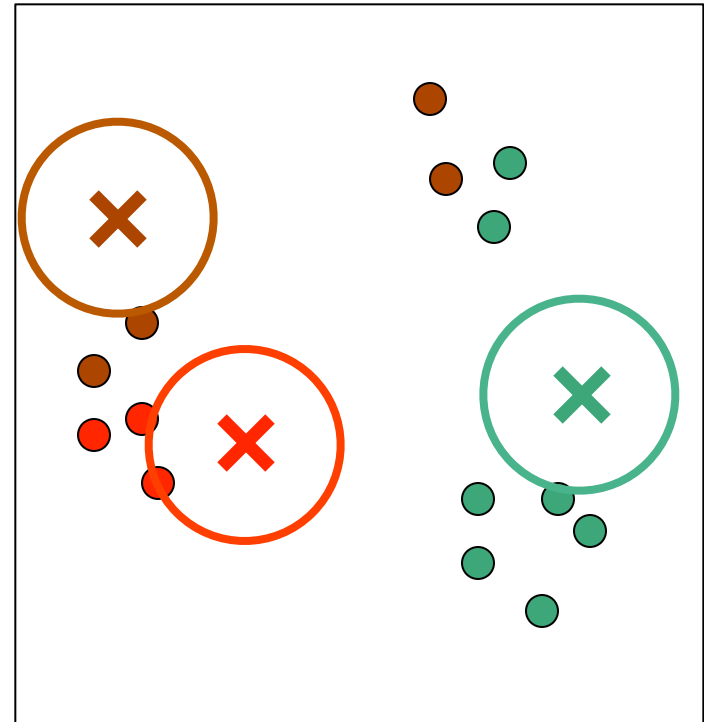


Iteration = 1

Model-Based Clustering

- No *partitions* now
- Points can be assigned to clusters with a *probability*

$$P(cl(\text{DataPoint}) = k \mid X) = \frac{f_k(\text{DataPoint})}{\sum_l f_l(\text{DataPoint})}$$



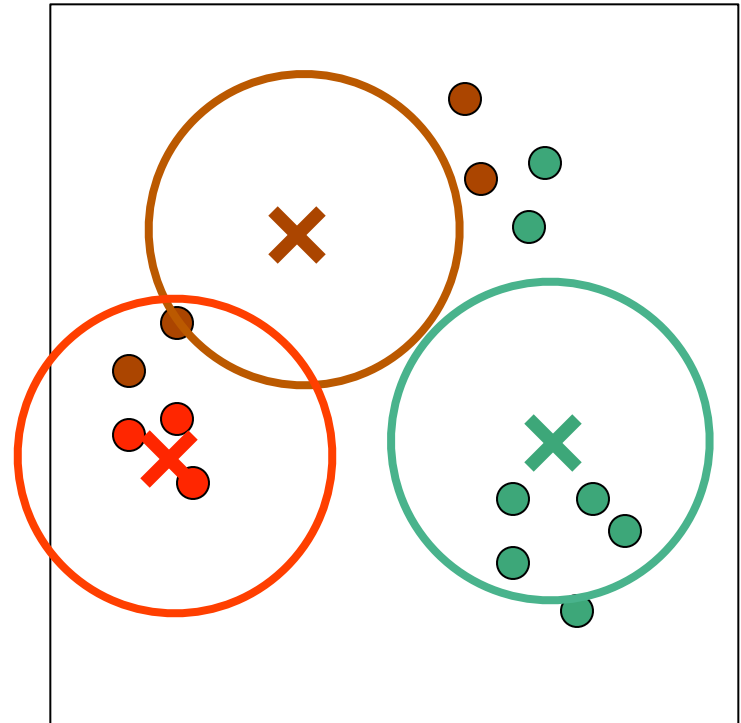
Iteration = 1

Model-Based Clustering

- Now re-compute the centers by taking the *weighted mean* of each cluster

$$\hat{\mu}_k = \frac{\sum_i z_{ik} x_i}{\sum_i z_{ik}}$$

$$z_{ik} = P(cl(x_i) = k | \Theta)$$



Iteration = 2

Final Thoughts

- **The most overused statistical method in gene expression analysis**
- **Gives us pretty pictures with patterns**
- **But, pretty picture tends to be pretty unstable.**
- **Many different ways to perform clustering**
- **Tend to be sensitive to small changes in the data**