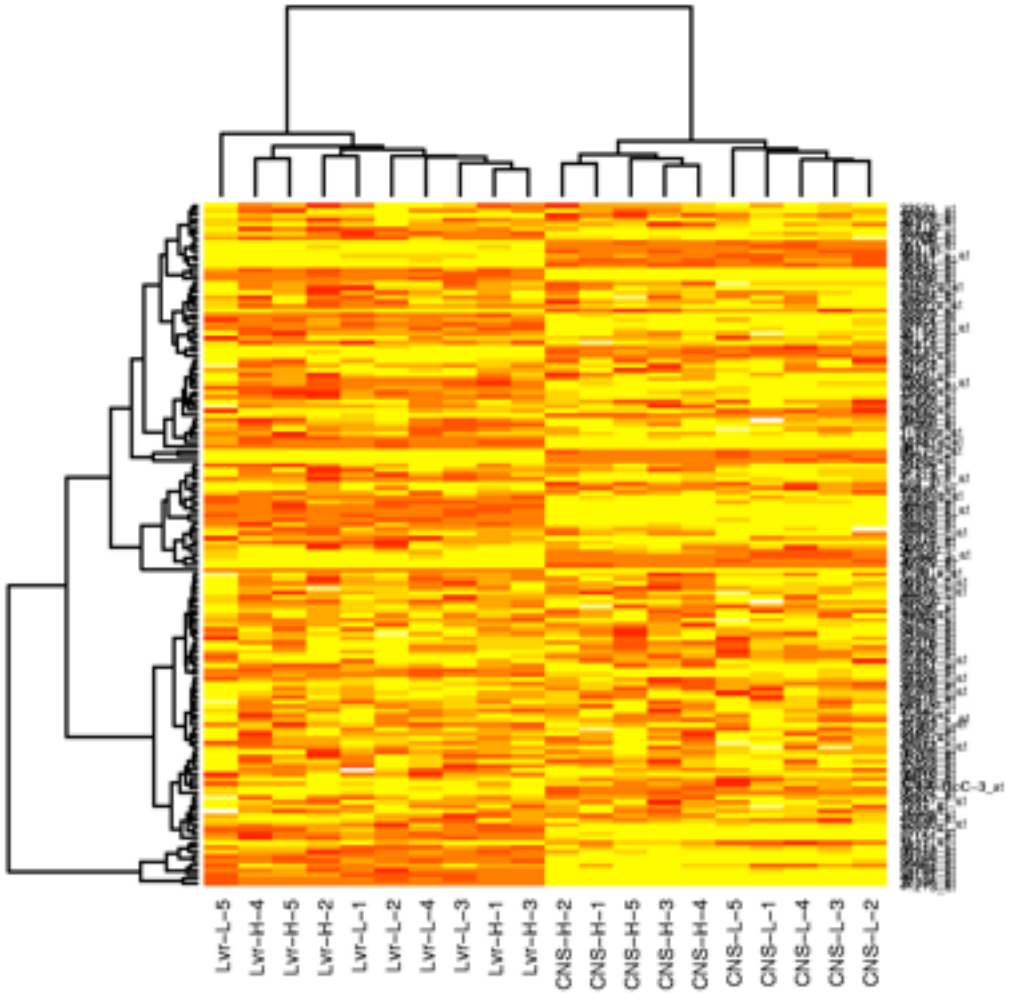
Clustering Gene Expression

CMSC423 Spring 2014
Héctor Corrada Bravo

Outline

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

Heatmaps


LVI–L–5
LVI–H–4
LVI–H–5
LVI–H–2
LVI–L–1
LVI–L–2
LVI–L–4
LVI–L–3
LVI–H–1
LVI–H–3
CNS–H–2
CNS–H–1
CNS–H–5
CNS–H–3
CNS–H–4
CNS–L–5
CNS–L–1
CNS–L–4
CNS–L–3
CNS–L–2

Edit View Favorites Tools Help
Back Search Favorites Media
http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v30/n1/full/ng765.html&filetype=pdf
Go
120%

© 2002 Nature Publishing Group http

against those from individuals diagnosed with conventional B-precursor ALL that lack this translocation. Initially, we collected samples from 20 individuals with conventional childhood ALL (denoted ALL), 10 of which had a *TEL/AML1* translocation. In addition, we collected samples from 17 individuals affected with the *MLL* translocation (denoted MLL). Details of the affected individuals and expression data are available online (Methods).

First, we determined whether there were genes among the 12,600 tested whose expression pattern correlated with the presence of an *MLL* translocation. We sorted the genes by their degree of correlation with the MLL/ALL distinction (Fig. 1) and used permutation testing to assess the statistical significance of the observed differences in gene expression[13]. For the 37 samples tested, roughly 1,000 genes are underexpressed in MLL as compared with conventional ALL, and about 200 genes are relatively highly expressed (data not shown). Thus, MLL shows a gene expression profile markedly different from that of conventional ALL.

MLL shows multilineage gene expression

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*

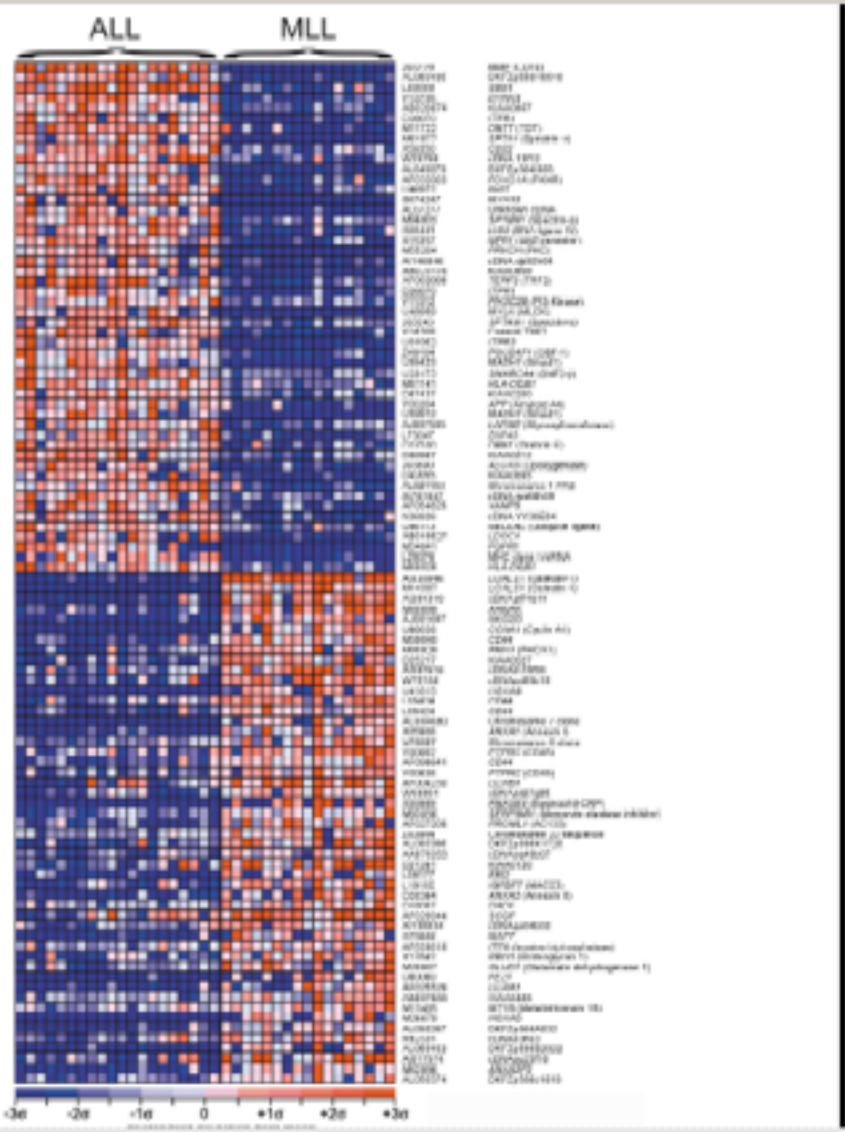


Fig. 1 Genes that distinguish ALL from MLL. The 100 genes that are most highly correlated with the class distinction are shown. Each column represents a leukemia sample, and each row represents an individual gene. Expression levels are normalized for each gene, where the mean is 0, expression levels greater than the mean are shown in red and levels less than the mean are in blue. Increasing distance from the mean is represented by increasing color intensity. The top 50 genes are relatively underexpressed and the bottom 50 genes relatively overexpressed in MLL. Gene accession numbers and the gene symbols or DNA sequence names are labeled on the right. Individual samples are arranged such that column 1 corresponds to ALL patient 1, column 2 corresponds to ALL patient 2, and so on. Information about the samples along with the top 200 genes that make the ALL/MLL distinction and their accession numbers can be found on our web site

2 of 7 8.5 x 11 in

Edit View Favorites Tools Help
Back Search Favorites Media
http://astor.som.jhmi.edu/~gp/688/alizadeh2000.pdf
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).

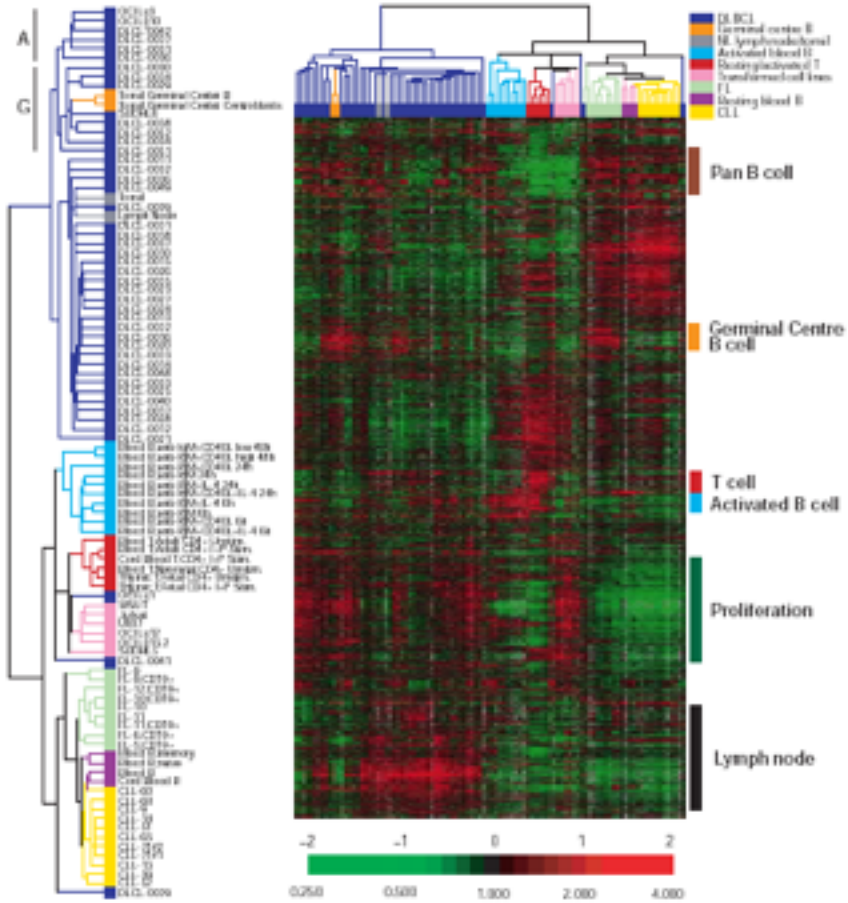


Figure 1 Hierarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression from 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of the relatedness of gene 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

3 of 9 8.26 x 10.88 in

Edit View Favorites Tools Help
Back Search Favorites Media
http://www.sciencemag.org/cgi/reprint/301/5639/1527.pdf
Go
140%

granuloma. The difference in the gene expression profiles between the two leprosy subclasses suggested that their differences in disease manifestation may reflect two opposing gene expression programs that influence the type of host response (*10*). The most pronounced differences in T-lep and L-lep gene expression profiles were among genes within the immune response family (Fig. 3). The gene expression profiles were consistent with previous data showing that type 1 cytokines associated with cell-mediated immunity predominate in T-lep lesions, whereas type 2 cytokines predominate in L-lep lesions (*2*, *11*, *12*). For example, genes encoding the type 1 cytokines lymphotoxin-α, interleukin (IL)-7, and IL-15 were comparatively up-regulated in T-lep lesions, as well as genes encoding CD1b and signaling lymphocytic activation molecule (SLAM), two molecules previously linked to cell-mediated immunity and type-1 cytokine production in these patients (*13*–*15*). In contrast, L-lep lesions differentially expressed the type 2 cytokines transforming growth factor–β and IL-5, as well as IL-4 and IL-10, although the differences in gene expression for the latter two genes between the patient groups were not statistically significant (P not ≤ 0.05) and were instead confirmed by quantitative polymerase chain reaction (qPCR) (*8*). As part of the type 2 pattern, L-lep lesions also exhibited marked up-regulation of genes related to humoral immunity, including immunoglobulin (Ig) heavy and light chains and molecules involved in B cell activation.

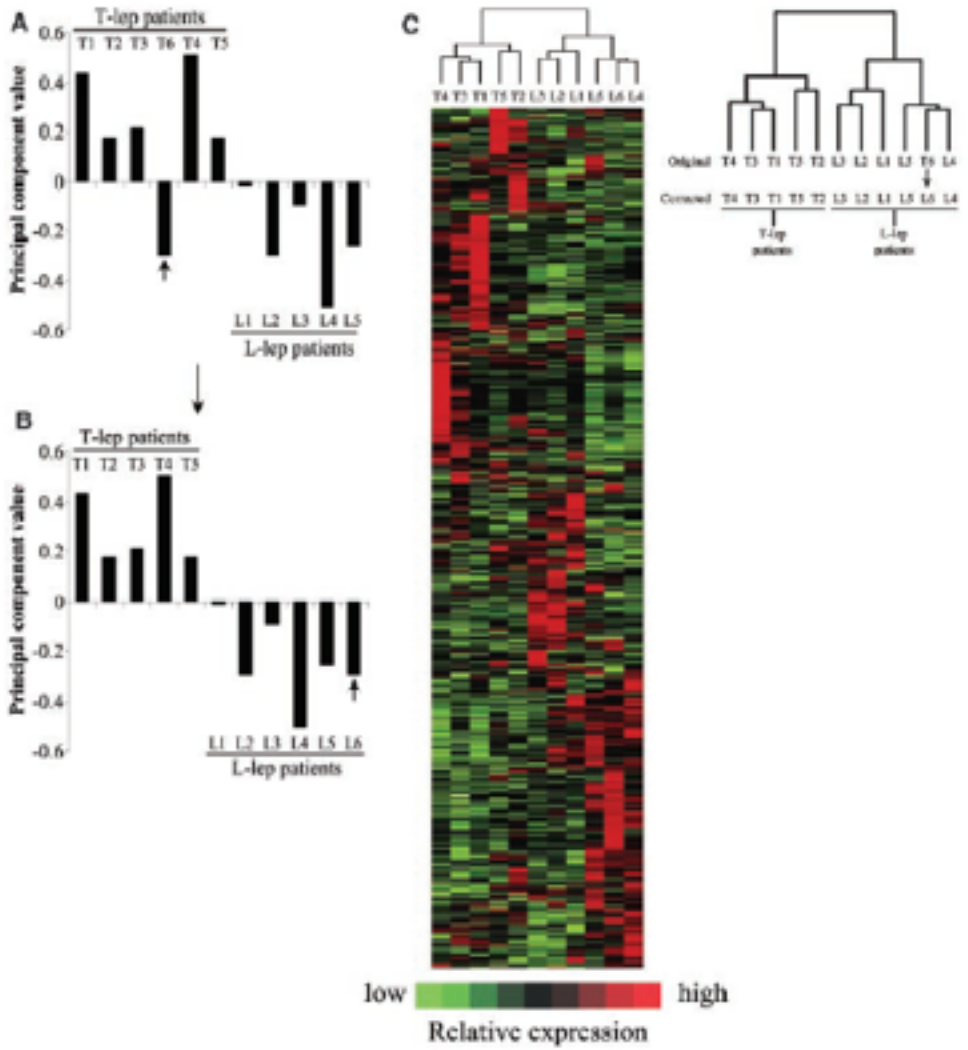
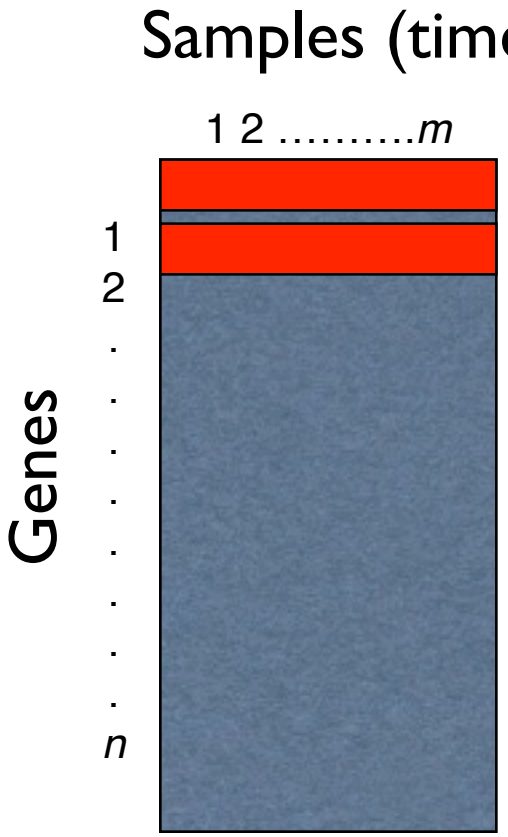


Fig. 1. Two unsupervised data analyses separate leprosy patients into clinically relevant subclasses

26 (2 of 4) 8.5 x 11 in

Measurements



DATA MATRIX

Measurement is expression of gene *i* at time *j*

Points

- Gene1= $(E_{11}, E_{12}, \ldots, E_{1N})'$
- Gene2= $(E_{21}, E_{22}, \ldots, E_{2N})'$

- Sample1= $(E_{11}, E_{21}, \ldots, E_{G1})'$
- Sample2= $(E_{12}, E_{22}, \ldots, E_{G2})'$

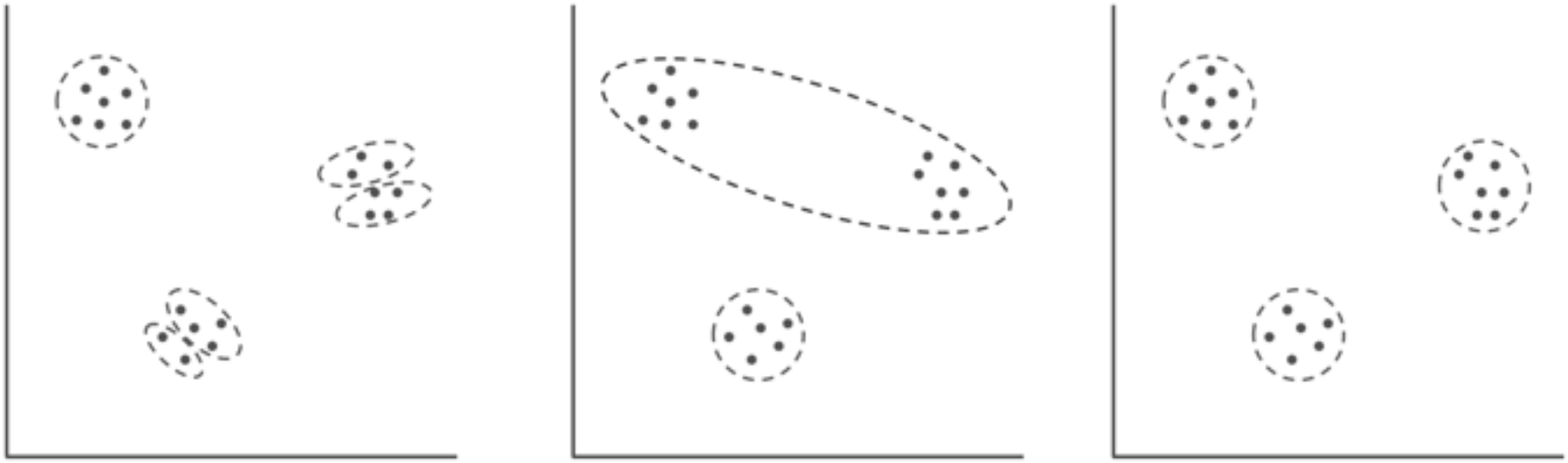
- E_{gi}=expression gene g, sample i

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*



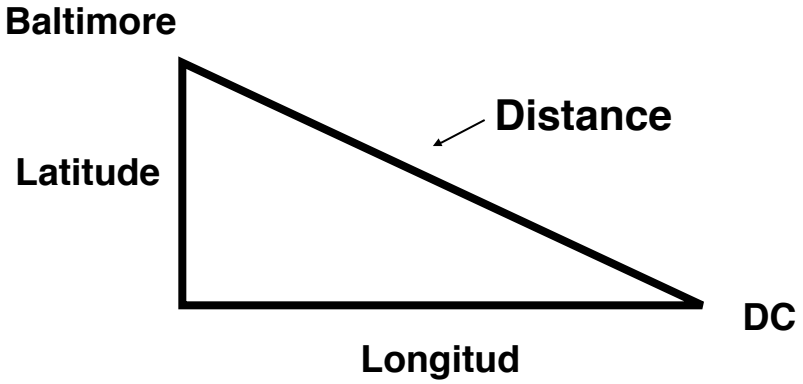
Clustering Problem

- *Partition a set of expression vectors into* ***clusters***

- Input: An *n x 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 $(E_{1i}-E_{2i})^2$, $i=1,\ldots,N$**
- **When *N* is 2, this is distance as we know it:**

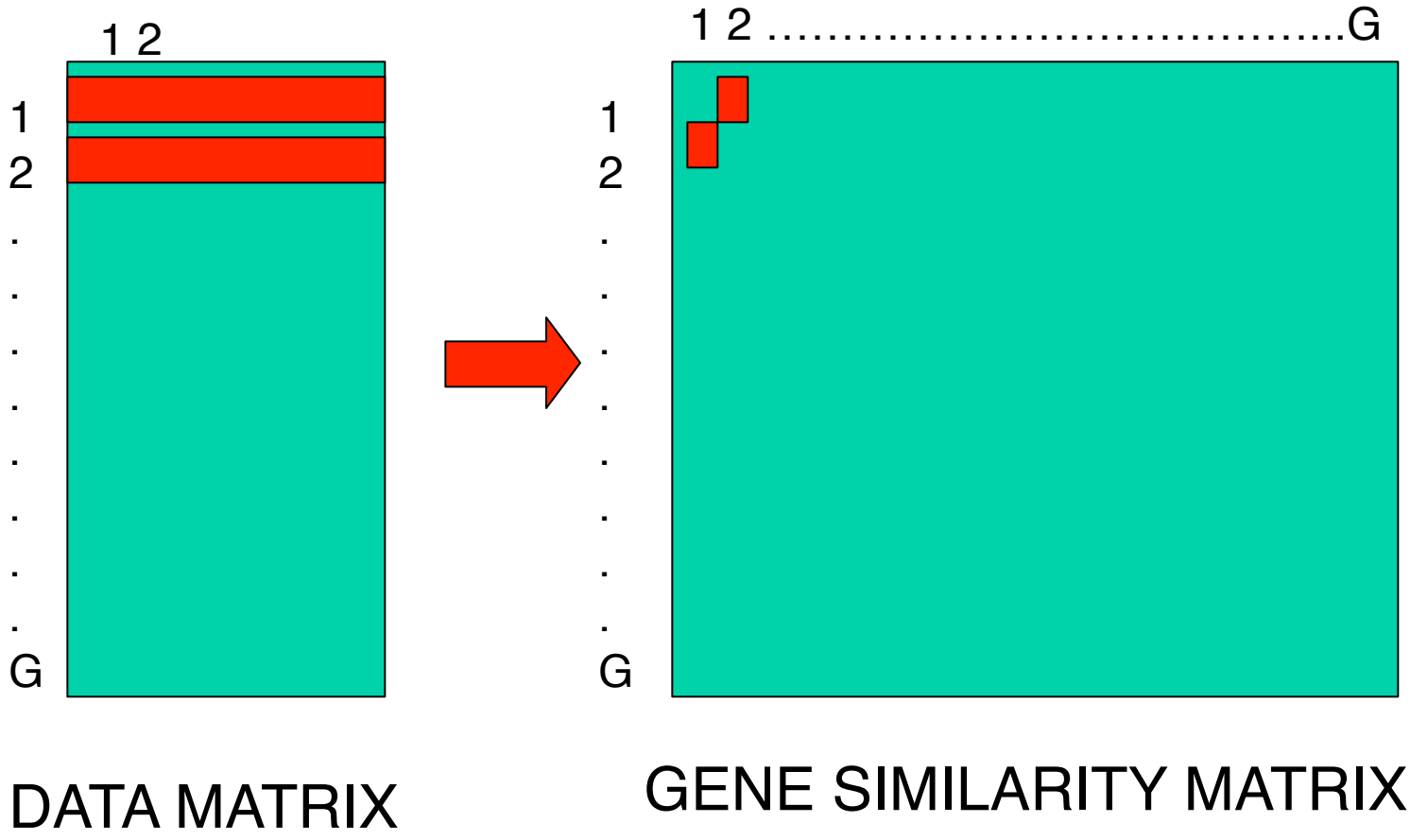


When *N* is 20,000 you have to think abstractly

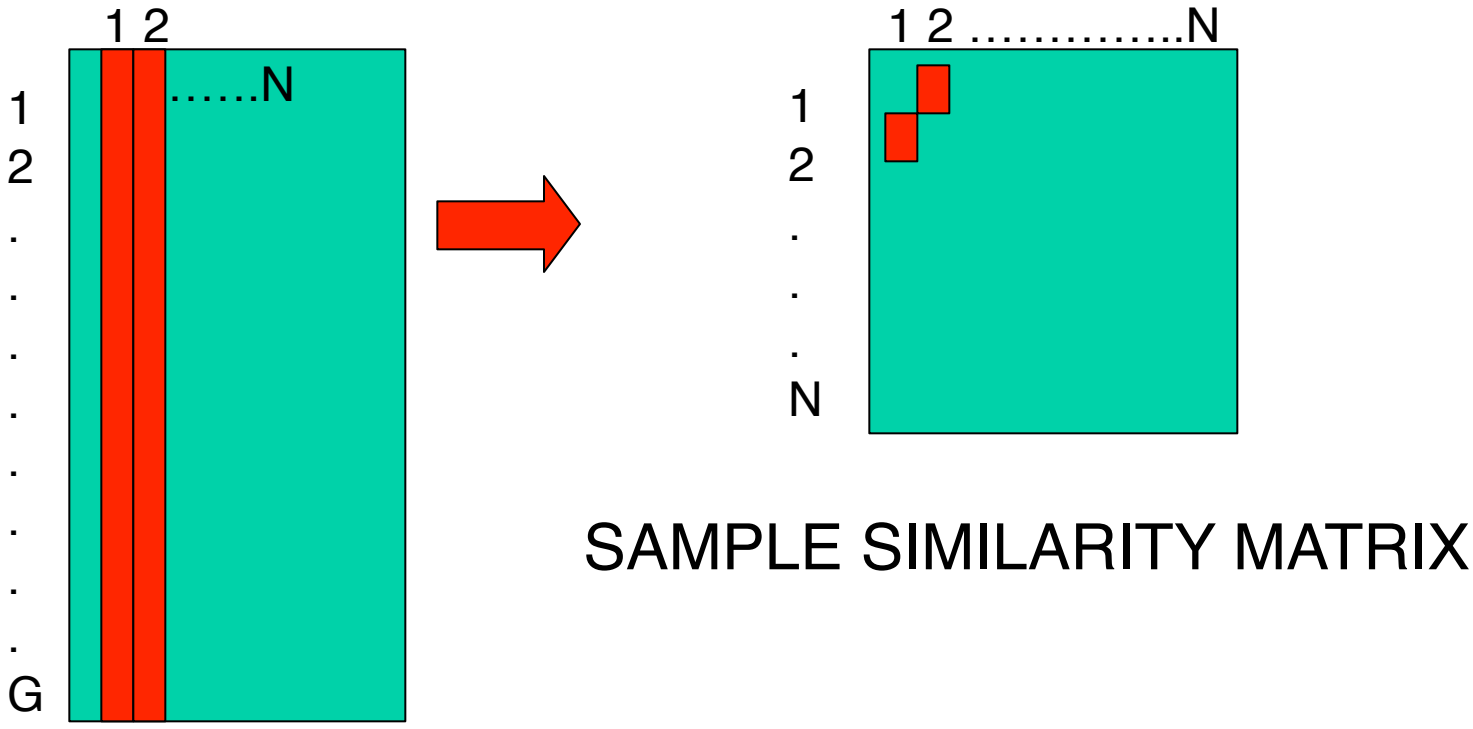
Similarity

- **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



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*

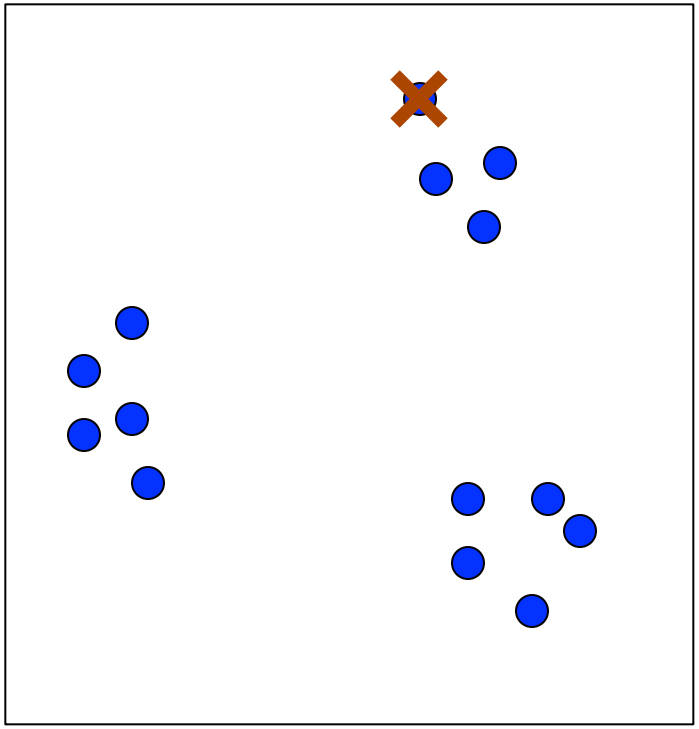
- **Input:** A set of points *Data* and an integer *k*
- **Output:** A set *X* of *k* centers that minimizes *MaxDistance(Data,X)* over all possible choices of *X*.

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

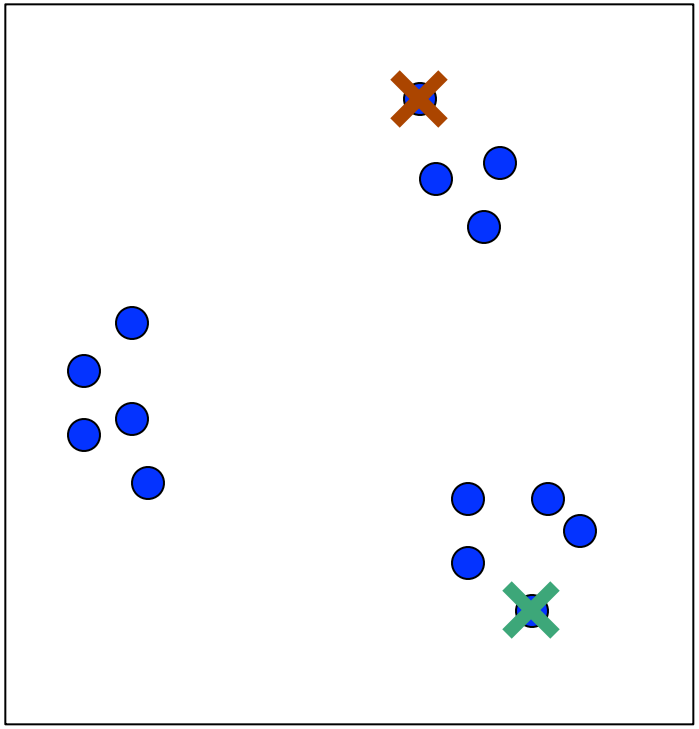
- **Choose 1 point as center**
- **This is arbitrary**



Iteration = 0

Furthest Traversal

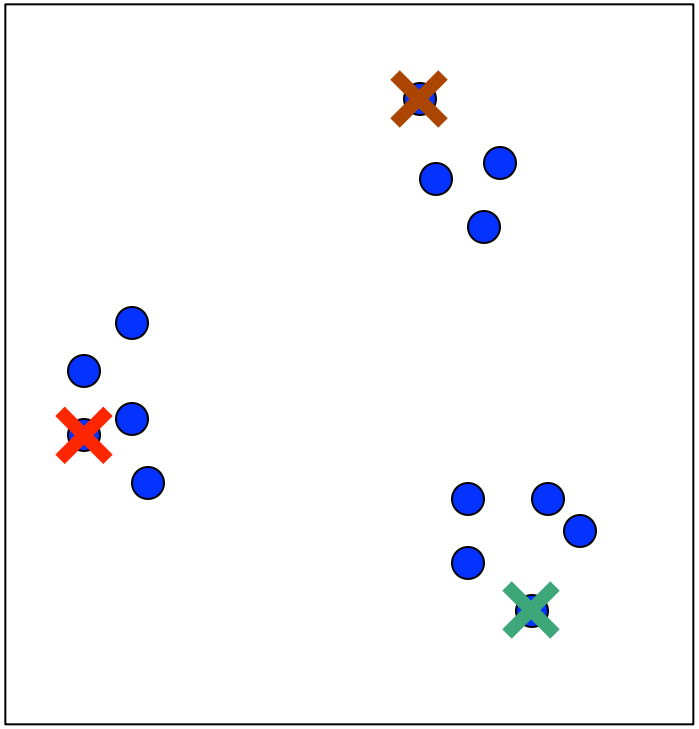
- **Choose point farthest from current set of centers as next center**



Iteration = 1

Furthest Traversal

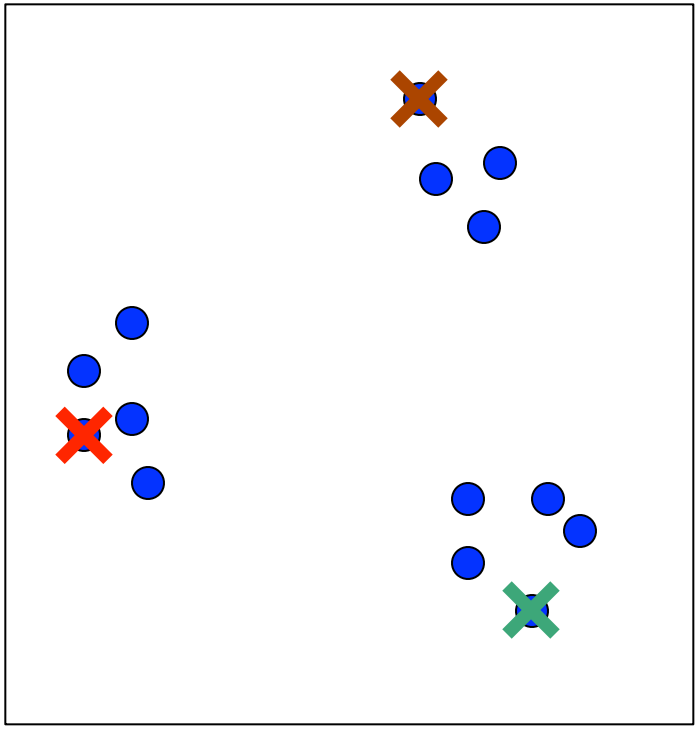
- **Choose point farthest from current set of centers as next center**



Iteration = 2

Furthest Traversal

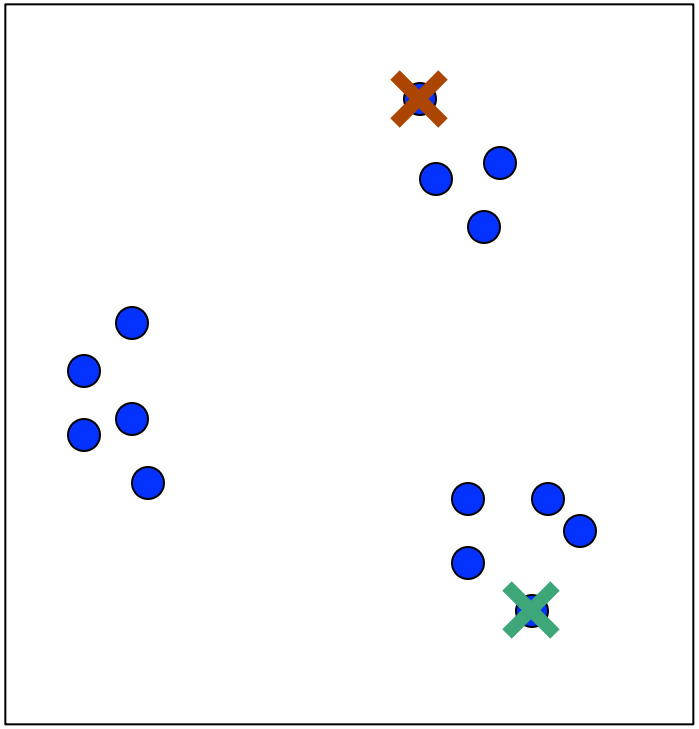
- **Choose point farthest from current set of centers as next center**



Iteration = 2

Furthest Traversal

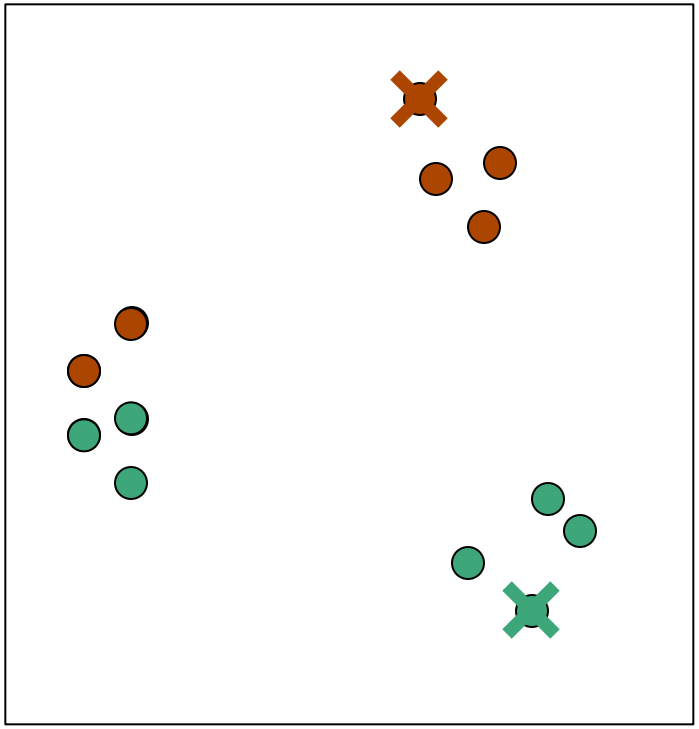
- **Choose point farthest from current set of centers as next center**



Iteration = 2

Furthest traversal

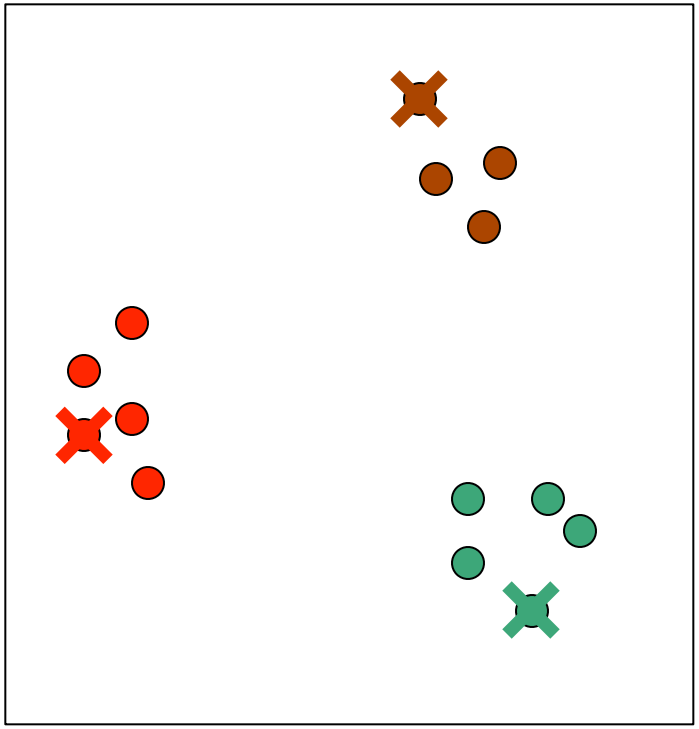
- **Assign each point to its nearest center**



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***

$$\text{MaxDistance}(Data, X_{ft}) \leq 2 * \text{MaxDistance}(Data, X_{opt})$$

- **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

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

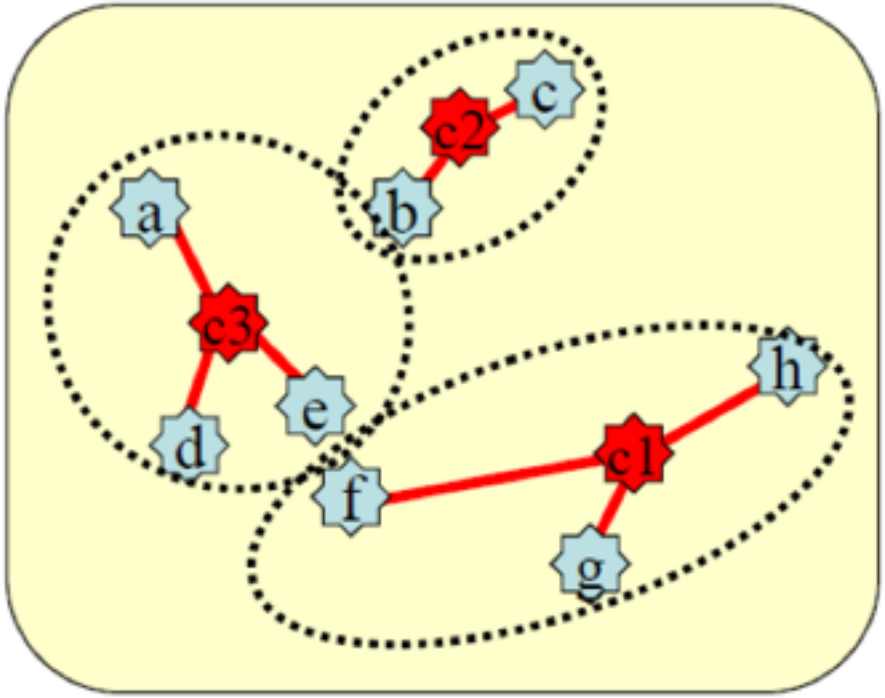
Center of gravity:

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

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

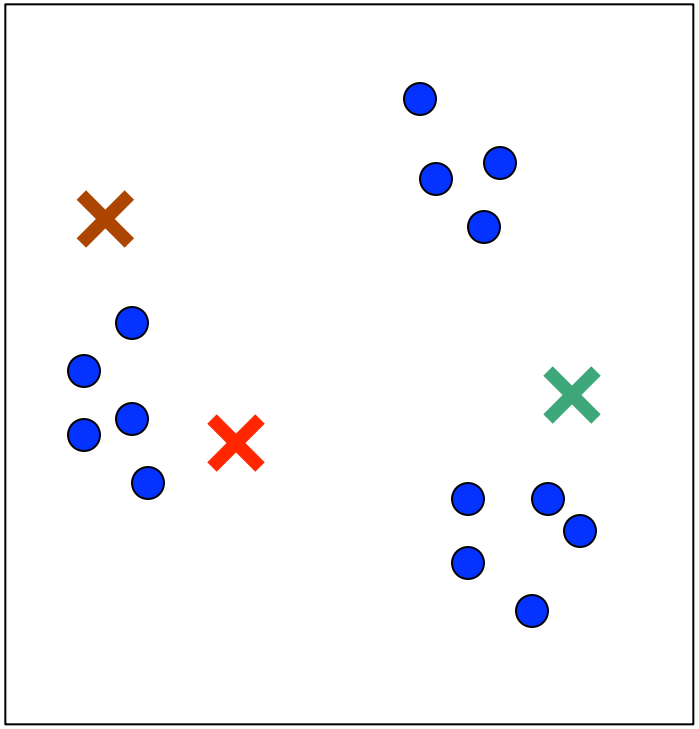
K-means problem:

minimize Distortion instead of MaxDistance



K-means

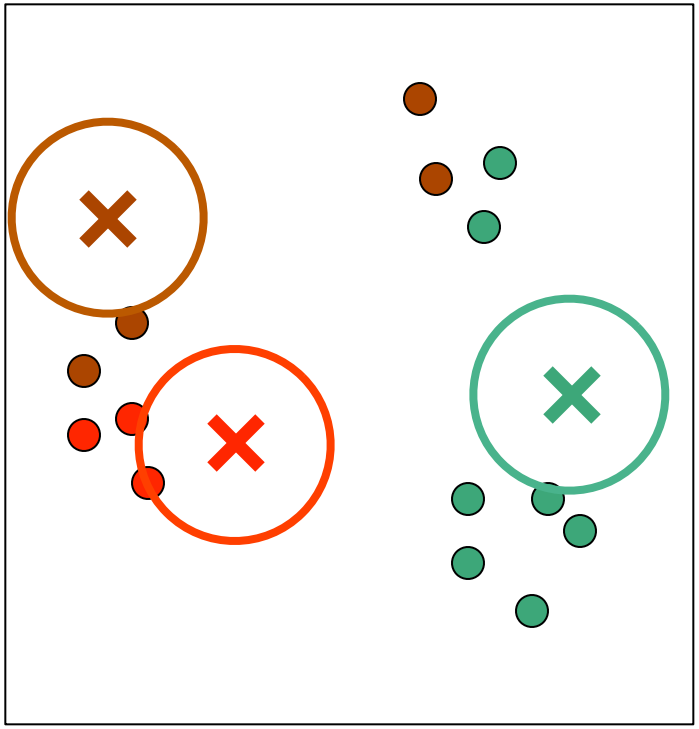
- Choose K *centroids*
- These are starting values that the user picks.
- There are some data driven ways to do it



Iteration = 0

K-means

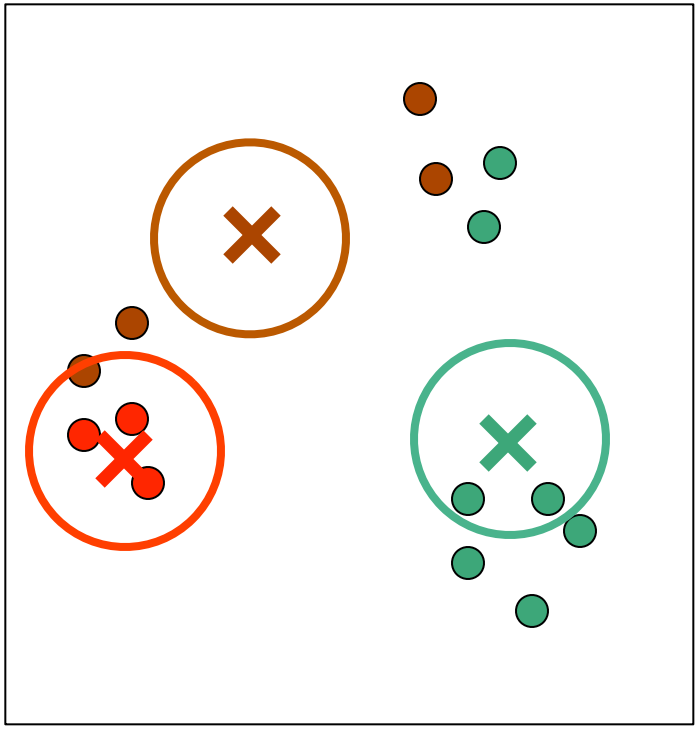
- **Make first *partition* by finding the closest centroid for each point**
- **This is where distance is used**



Iteration = 1

K-means

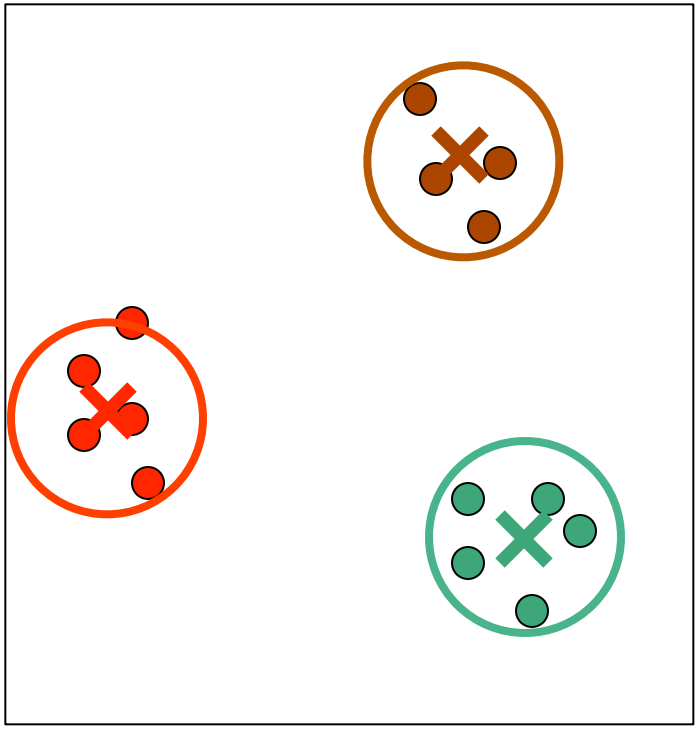
- **Now re-compute the centroids by taking the *middle* of each cluster**



Iteration = 2

K-means

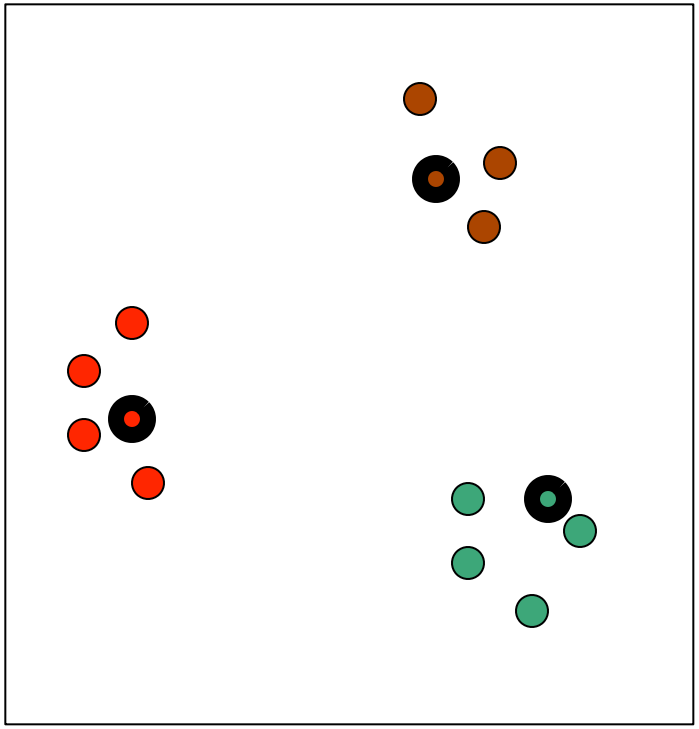
- **Repeat until the centroids stop moving or until you get tired of waiting**



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 chose K? There are methods but not much theory saying what is best.**

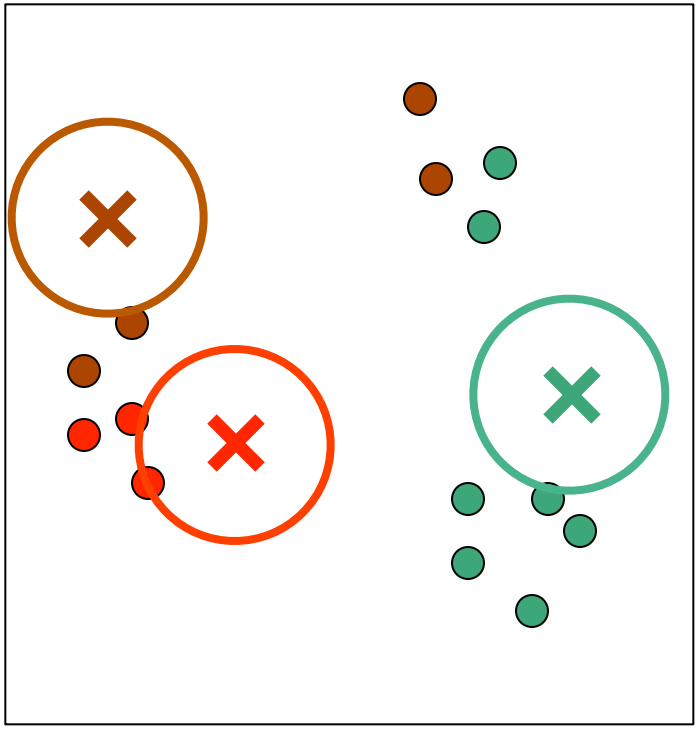
K-means limitations

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 <u>how likely</u> they belong to that cluster*

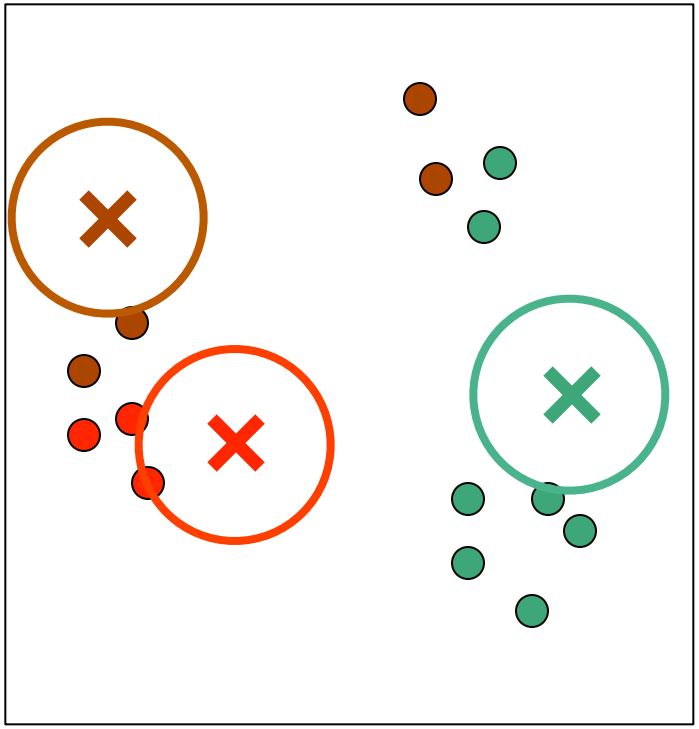


Iteration = 1

Model-Based Clustering

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

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



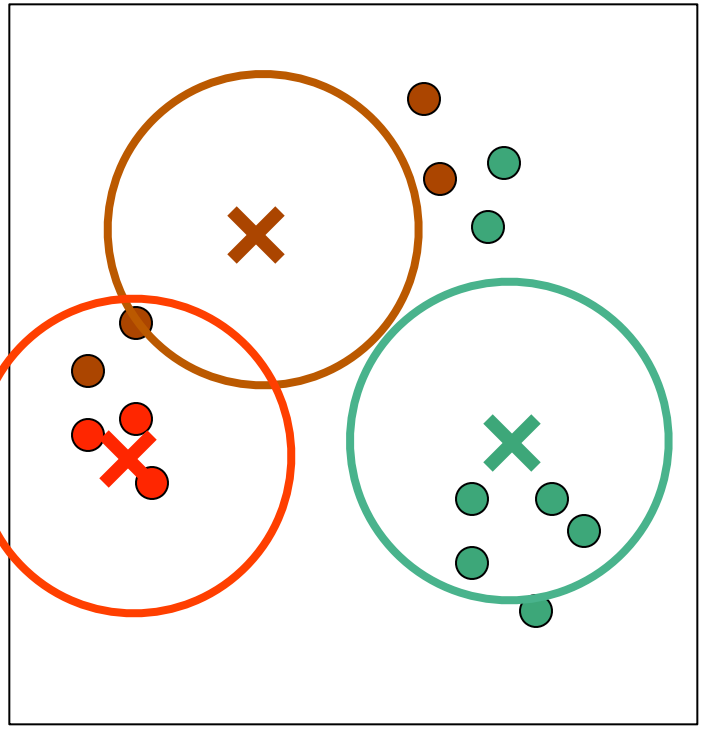
Iteration = 1

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