

Clustering

1

Clustering Aim

Data Preparation

Distance Metrics

2

Hierarchical Clustering

K-means Clustering

3

Principal Component Analysis

Practical Consideration

- <https://lgatto.github.io/IntroMachineLearningWithR/unsupervised-learning.html>
 - (in particular 4-4.5)

- Exploratory data analysis for large data
- Summarize major trends
- What is similar to what, what behaves like what?

- High Throughput analysis
 - 384 or 1536 well plates
- 'Omics
 - Metabolomics
 - 100s-1000s of metabolites at once
 - Transcriptomics
 - 10,000s of transcripts at once
 - Single cell transcriptomics
 - 10,000s of transcripts x 1000s of cells at once
 - Meta genomics
 - 100s-1000s of species at once
- ...

Box plot won't quite cut it for large data...

- 2D – tabular format
 - All samples x All measured units
- You also might consider...
 - Filtering
 - Very low values
 - Values that don't change
 - e.g. filter out vector like ``c(0, 0, 0, 0, 0, 0)``
 - Log transform (if your data are near log normal)
 - e.g. transcript expression data
 - Scaling to Z-score
 - Subtract mean
 - Divide by standard deviation

- Relative / pattern

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Pearson's r
 - Pearson's distance: $1 - r$
- Coefficient of determination (r^2)
 - coefficient of determination distance ($1 - r^2$)
- Spearman's ρ (as above, but with ranks, not raw x)
 - Spearman's distance: $1 - \rho$

- Absolute

- Manhattan
- Euclidean

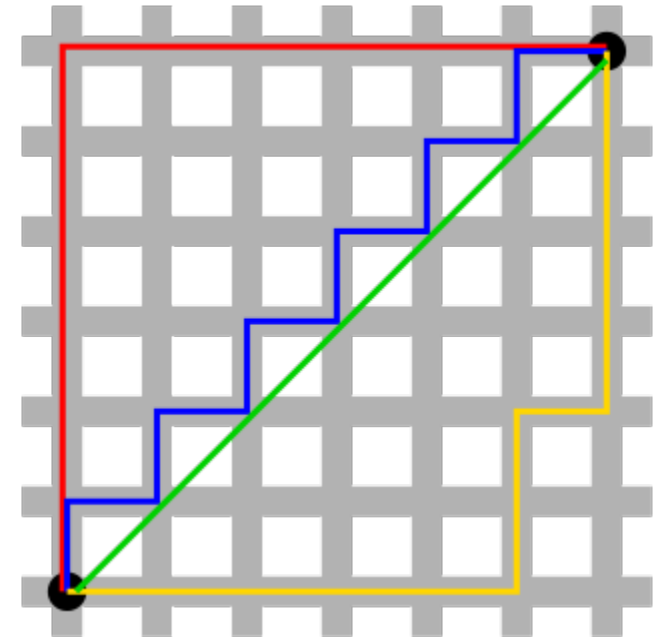
- Manhattan
$$d(p, q) = \sum_{i=1}^n |p_i - q_i|$$

- Euclidean
$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

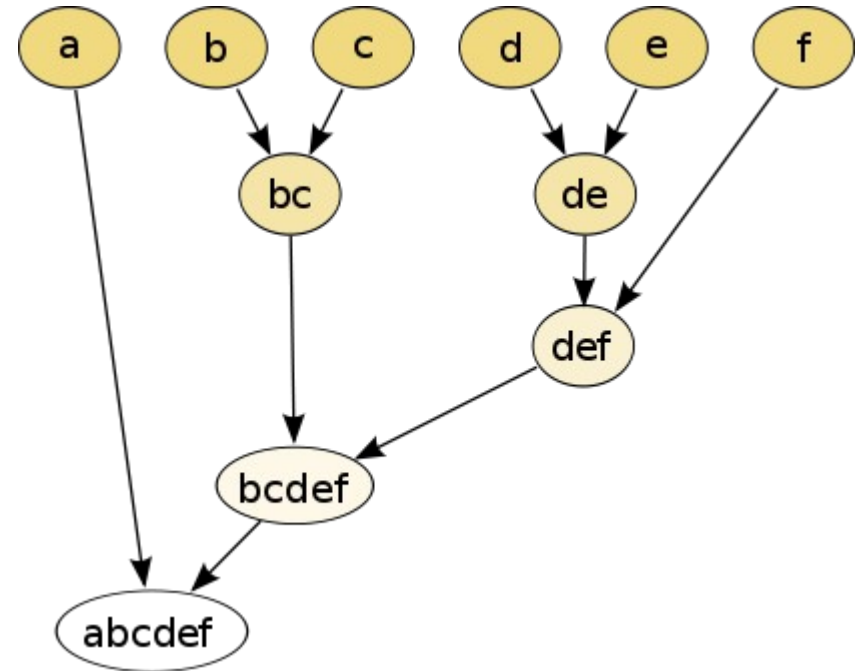
Where where p and q are vectors.

E.g. of the abundance of a particular gene in different samples

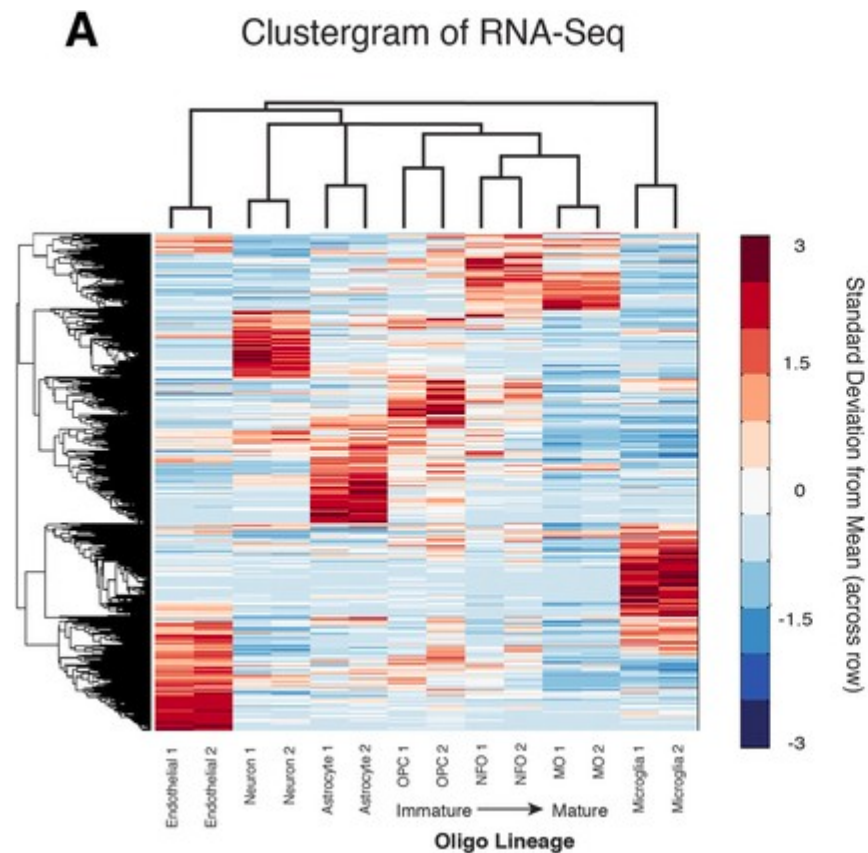
- Manhattan distance is also called "Taxi-cab" distance (red, blue, yellow).
- Euclidean distance is the diagonal (green)



- Always combine the next two items (or clusters) with the smallest distance between them
- Stop combining when only one cluster remains

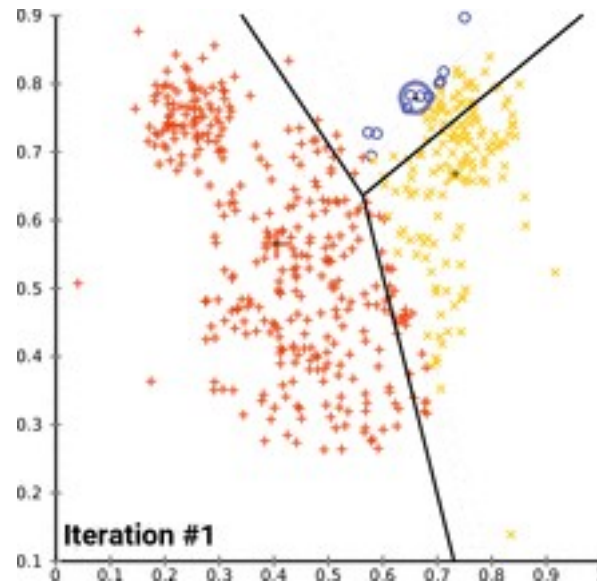


Gene expression in neurons

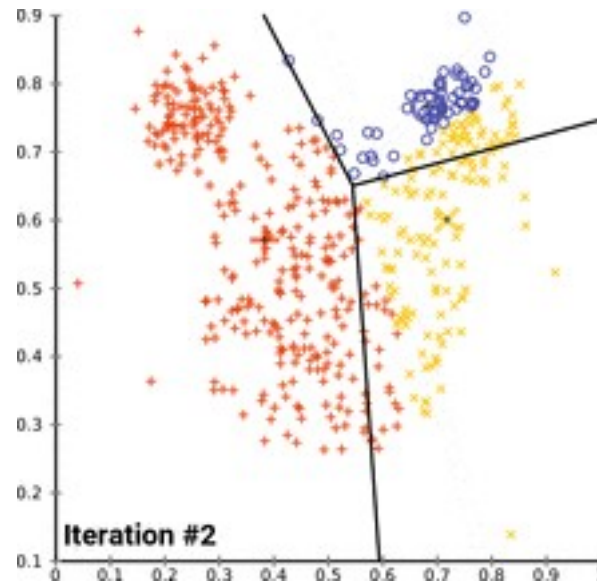


- Choose k random "centers"
- Assign every item to the cluster with the closest center
- Change the center to be the average of the new cluster
- Re-assign any genes that are now closer to a different center
- Repeat until convergence

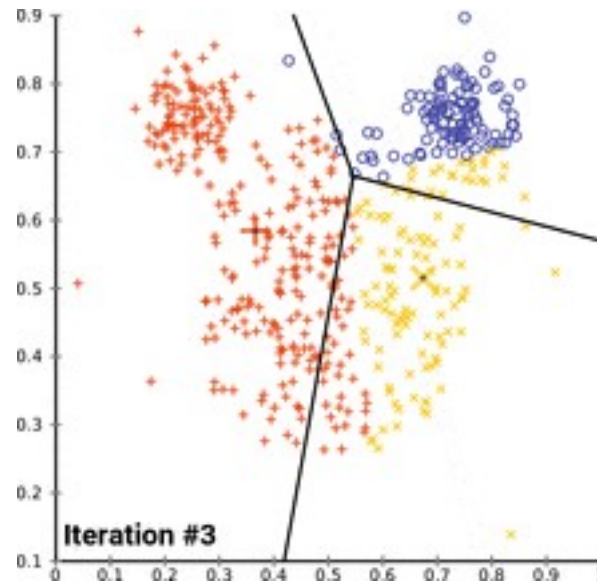
K-means clustering



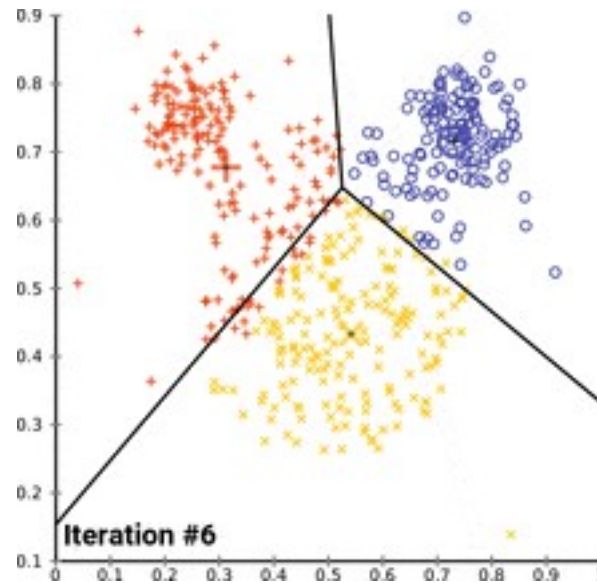
K-means clustering



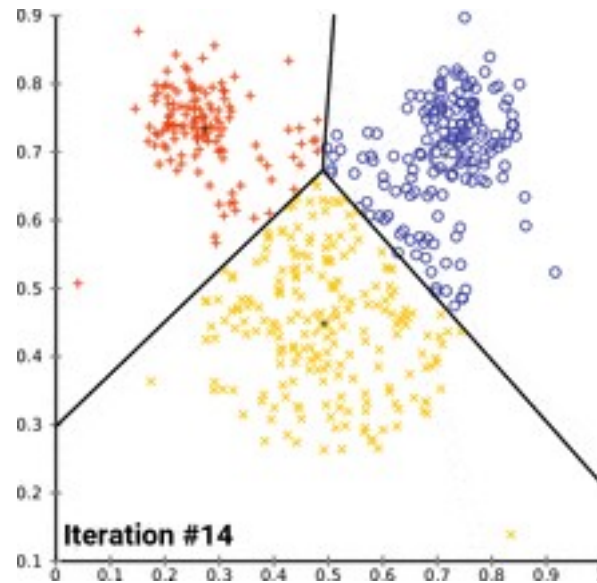
K-means clustering



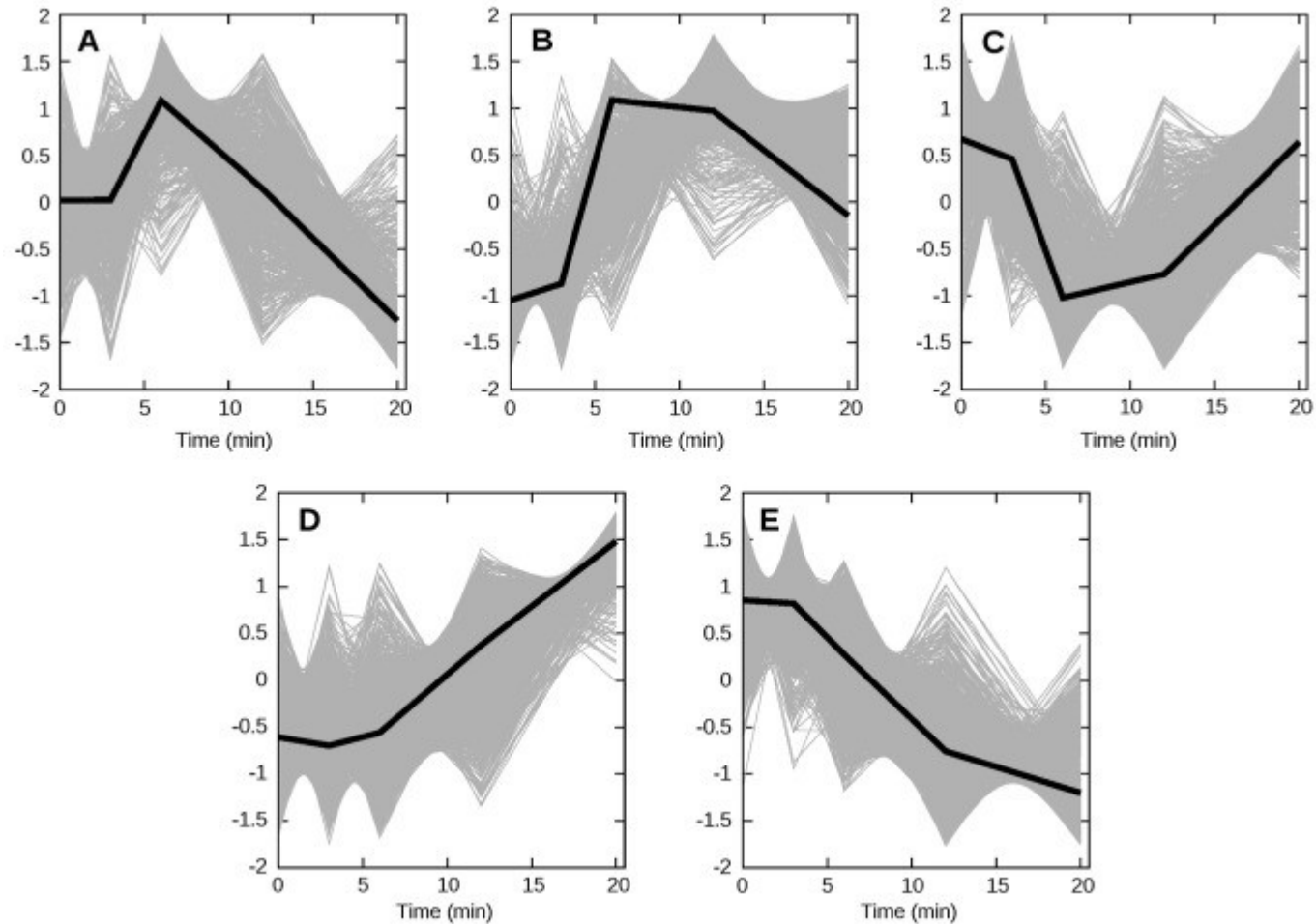
K-means clustering



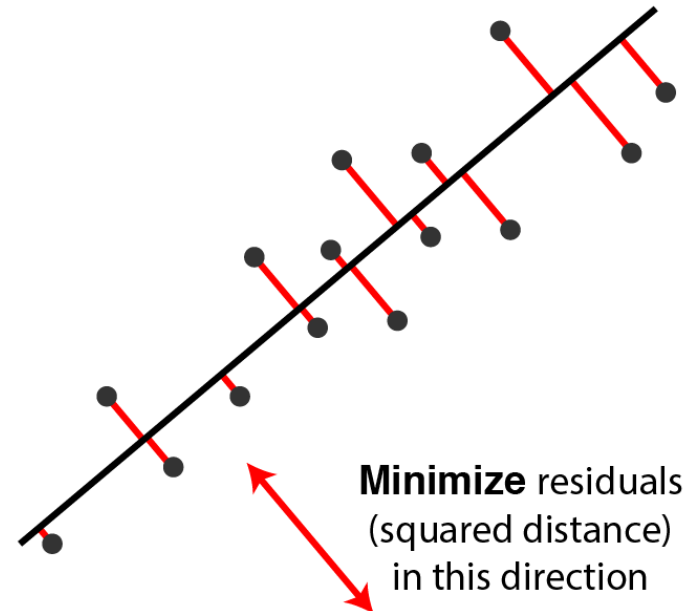
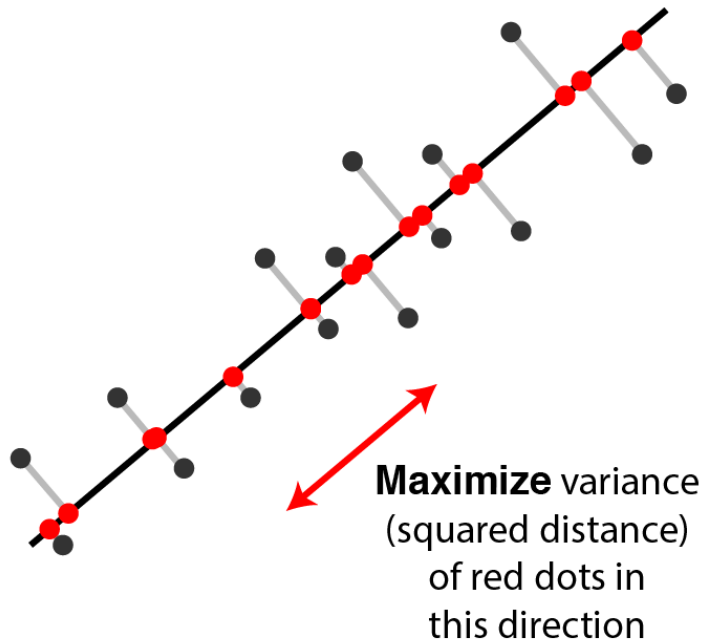
K-means clustering



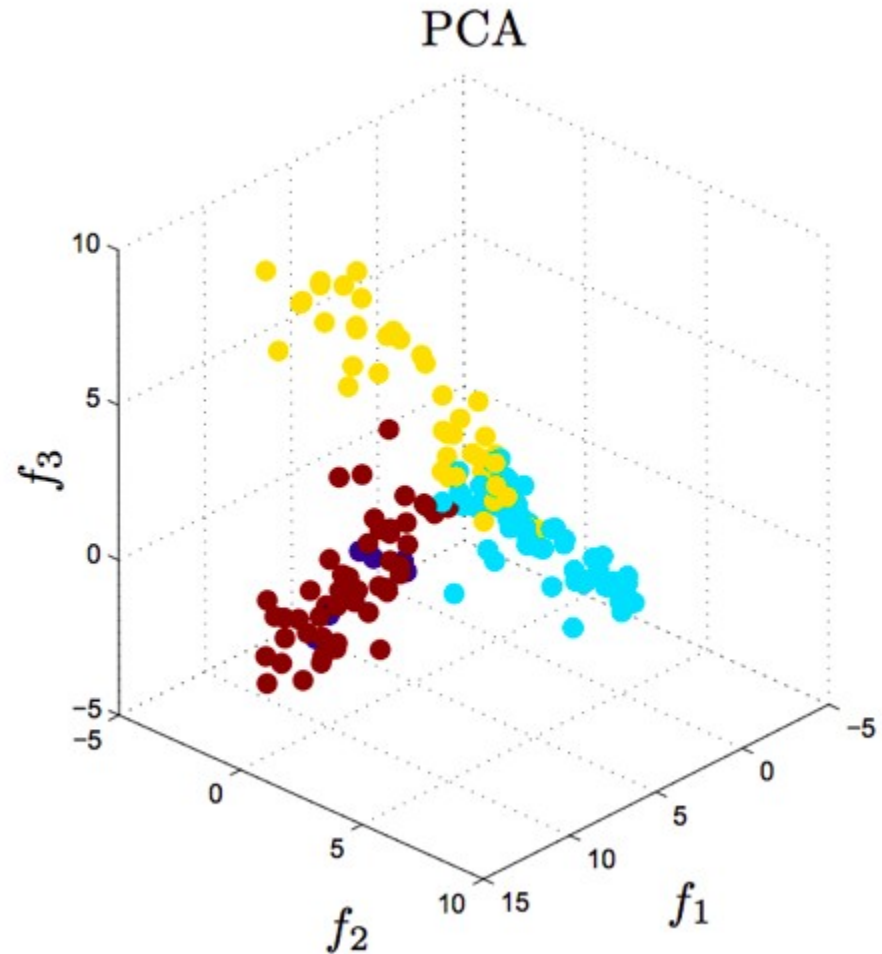
Transcriptional response of yeast to oxidative stress



- Find low dimensional projection of data that captures maximal amount of variance



- Identify largest signals in the data
- Quality control
 - Do replicates and similar samples cluster together

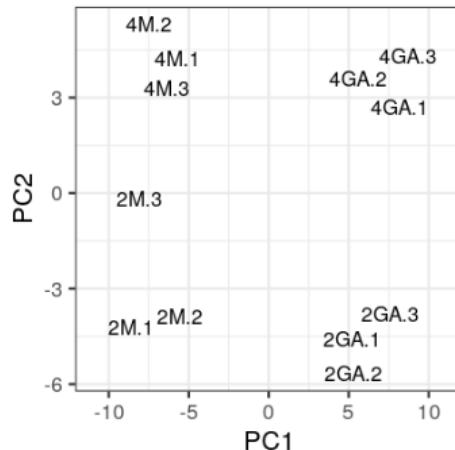


- Hierarchical clustering
 - One can "cut" the dendrogram to define clusters
 - One gets mix of large and small clusters and outlying items
- K-means
 - It's necessary to choose k , the number of centers
 - Biases towards evenly sized clusters
 - Fast
- PCA
 - Very helpful in visualization / understanding
 - Doesn't directly define clusters

- Next Up → We are done!

- Aim of clustering (summarizing big-data, unsupervised learning)
- Typical data prep with filtering, maybe transforming, and scaling
- Distance metrics (1 – correlation metrics, absolute distance metrics)
- Types of clustering
 - Hierarchical, K-means, PCA
 - Basic idea of how, but more importantly how to interpret results.
- Example Question

19. Take a look at the following result of a clustering analysis on the transcriptome (10,000+ transcripts) of samples treated for 2 or 4 days with a mock (M) control or gibberellic acid (GA) treatment.



- (b) (2 points) Which of the following **can** you confidently conclude from the analysis?
- ☐ Sample 2M.3 is definitely an outlier
 - ☐ The most prominent pattern in the data is differences between treatments (M & GA)
 - ☐ Gene expression is higher after 4 days than 2.
 - ☐ The most prominent pattern in the data is differences between 2 and 4 days.