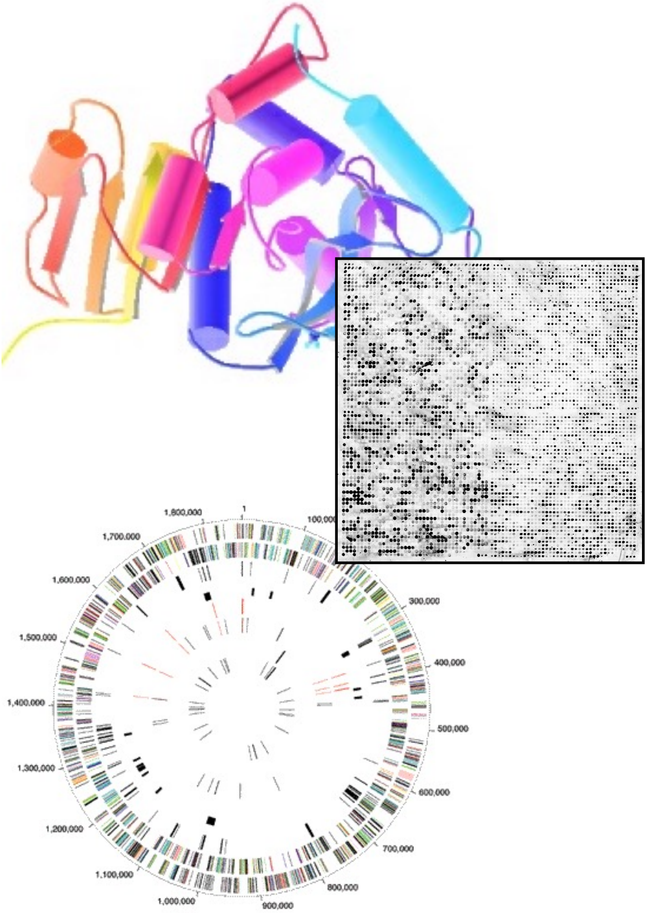
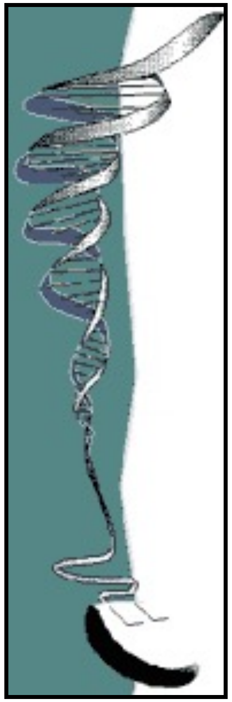
Biomed. Data Science:

# Unsupervised Datamining A: General Clustering





Mark Gerstein, Yale University

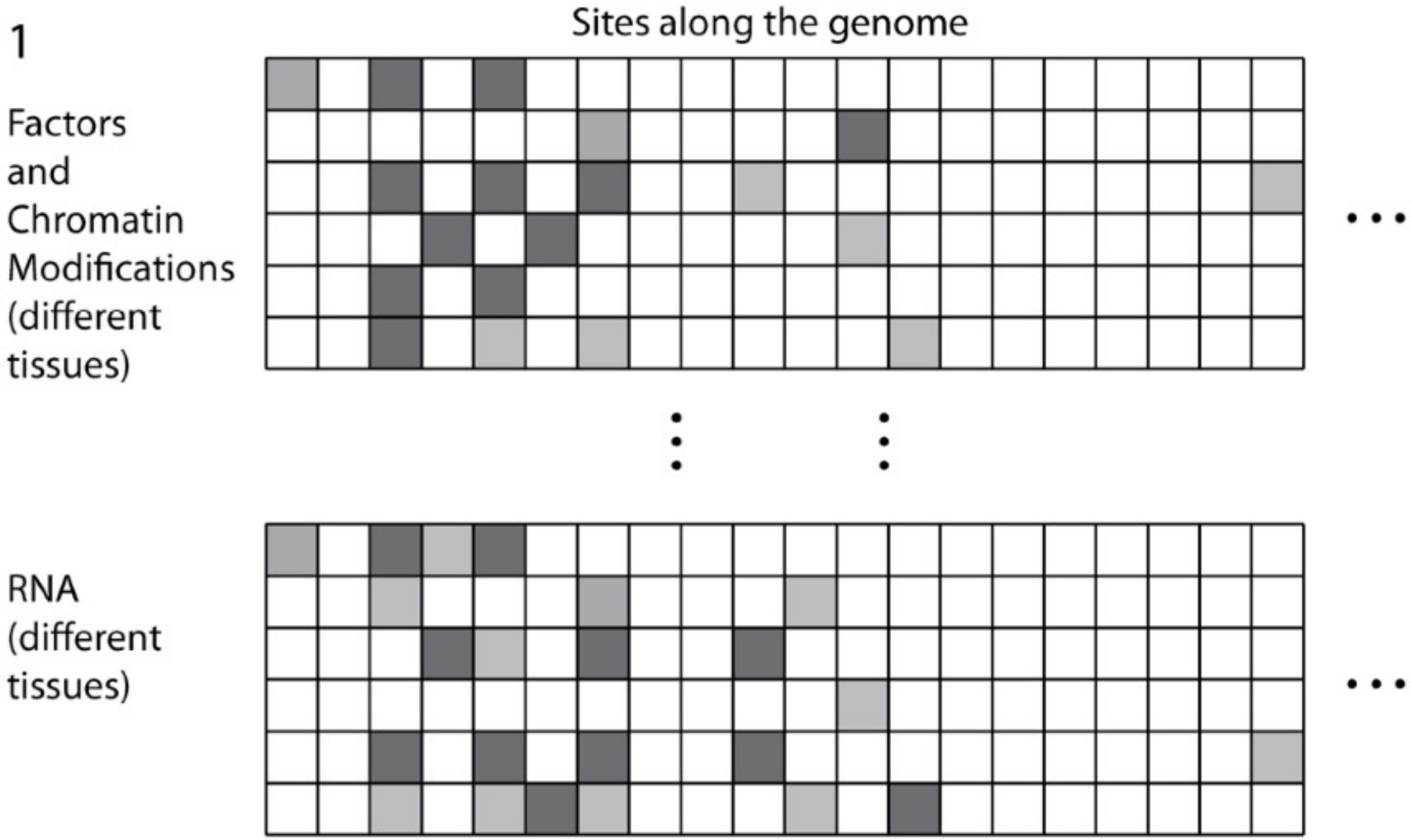
gersteinlab.org/courses/452

(last edit in spring ’21, pack #9a, final)

# Unsupervised Mining

## Columns & Rows
## of the Data Matrix

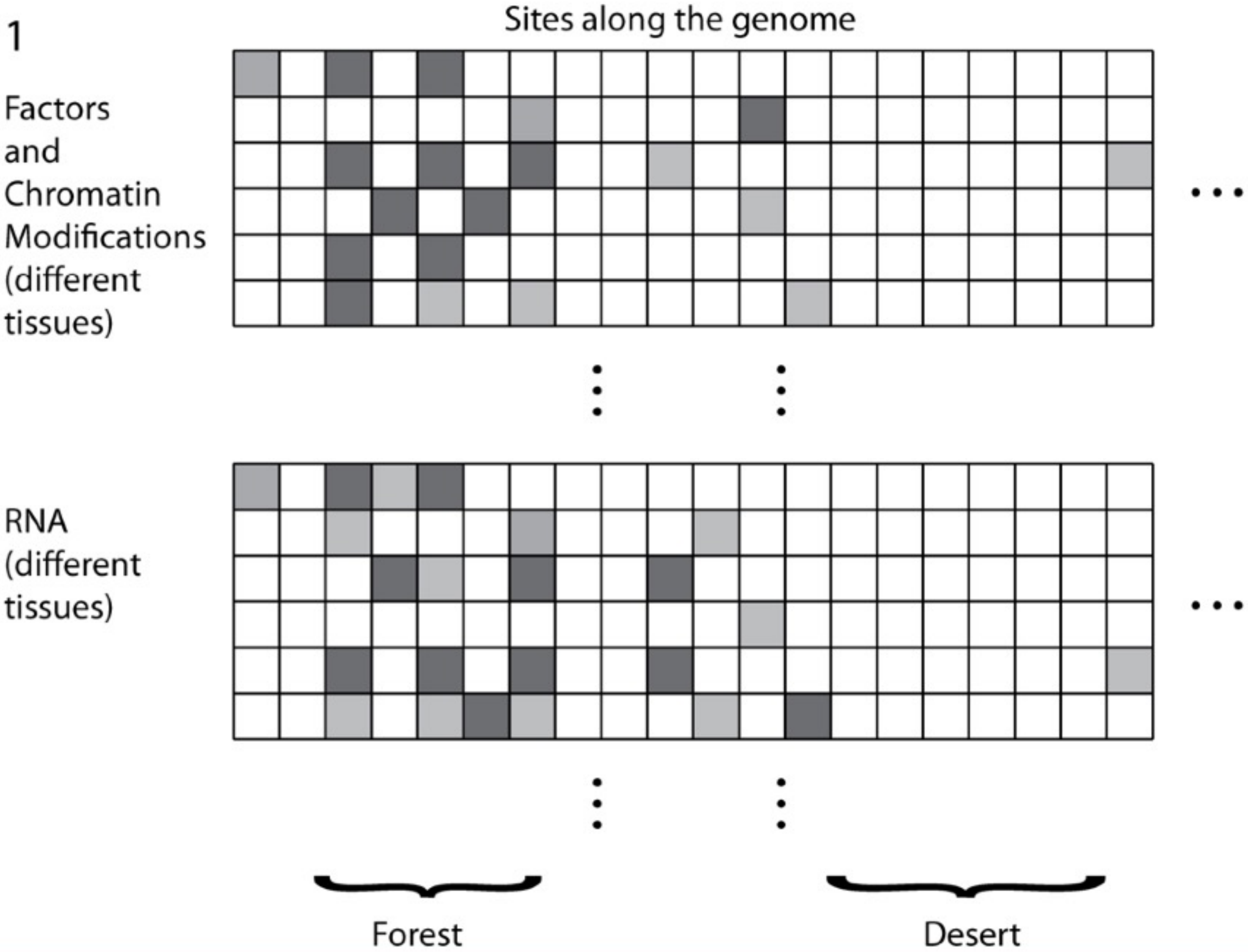
# Structure of Genomic Features Matrix



# Unsupervised Mining

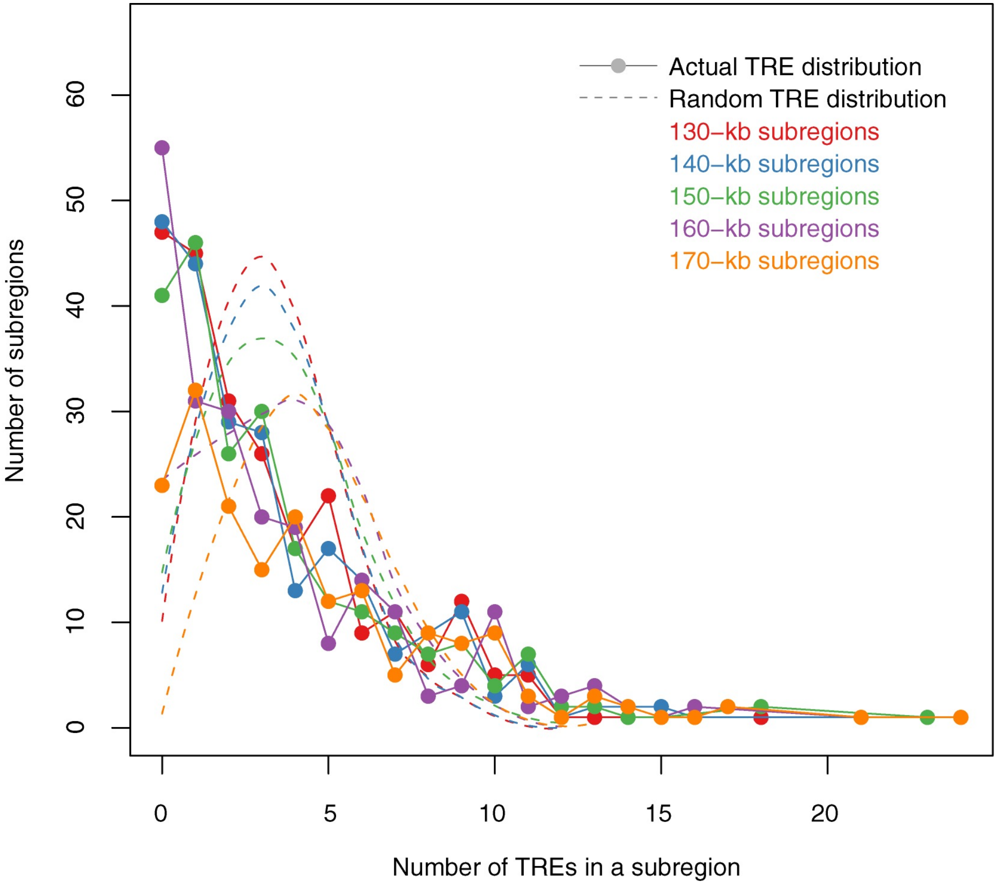
- Simple overlaps & enriched regions
- Clustering rows & columns (networks)
- PCA/SVD (theory + appl.)
- Biplot
- RCA
- CCA
- tSNE
- LDA
- (Variational Autoencoders)

# Genomic Features Matrix: Deserts & Forests



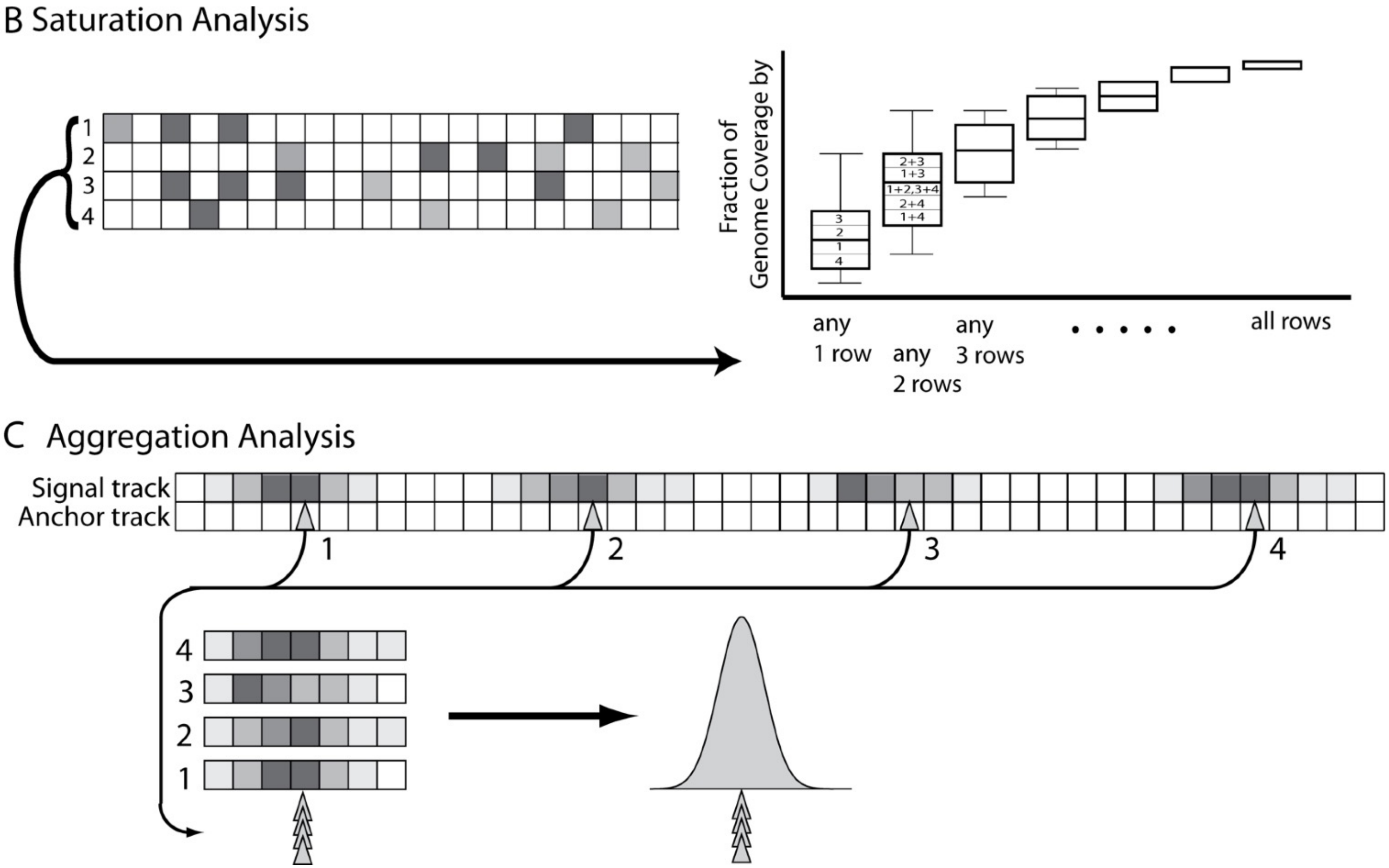
# Non-random distribution of TREs

- TREs are not evenly distributed throughout the encode regions ($P < 2.2 \times 10^{-16}$ ).
- The actual TRE distribution is power-law.
- The null distribution is ‘Poissonesque.’
- Many genomic subregions with extreme numbers of TREs.



Zhang et al. (2007) Gen. Res.

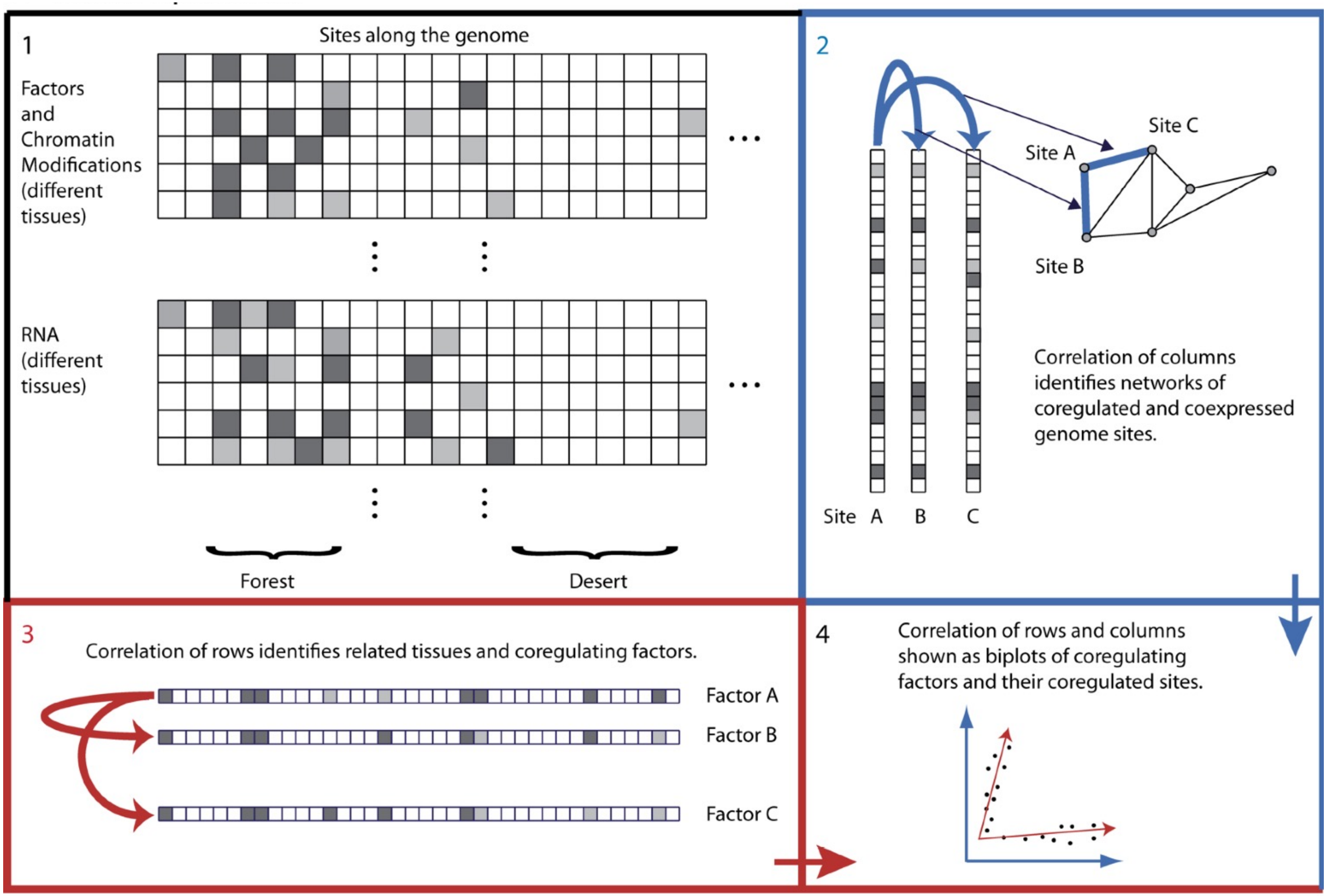
# Aggregation & Saturation



[*Nat. Rev. Genet.* (2010) 11: 559]

# Expression Clustering

# Correlating Rows & Columns



Correlation of columns identifies networks of coregulated and coexpressed genome sites.

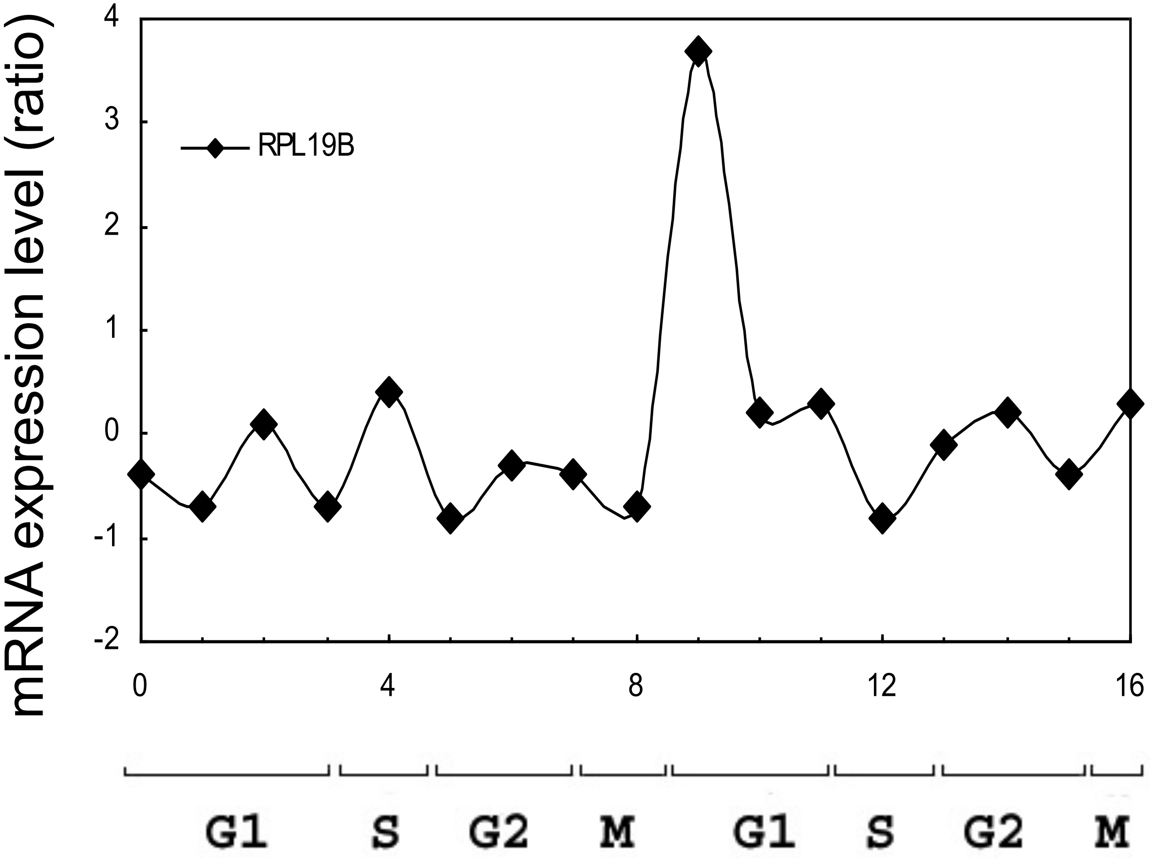
Correlation of rows identifies related tissues and coregulating factors.

Correlation of rows and columns shown as biplots of coregulating factors and their coregulated sites.

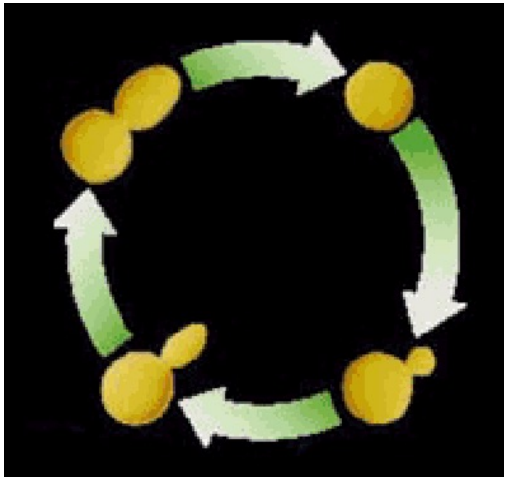
[*Nat. Rev. Genet.* (2010) 11: 559]

# Clustering the yeast cell cycle to uncover interacting proteins

[Brown, Davis]

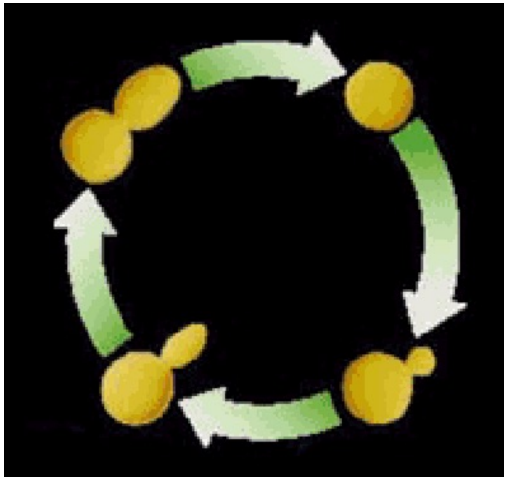


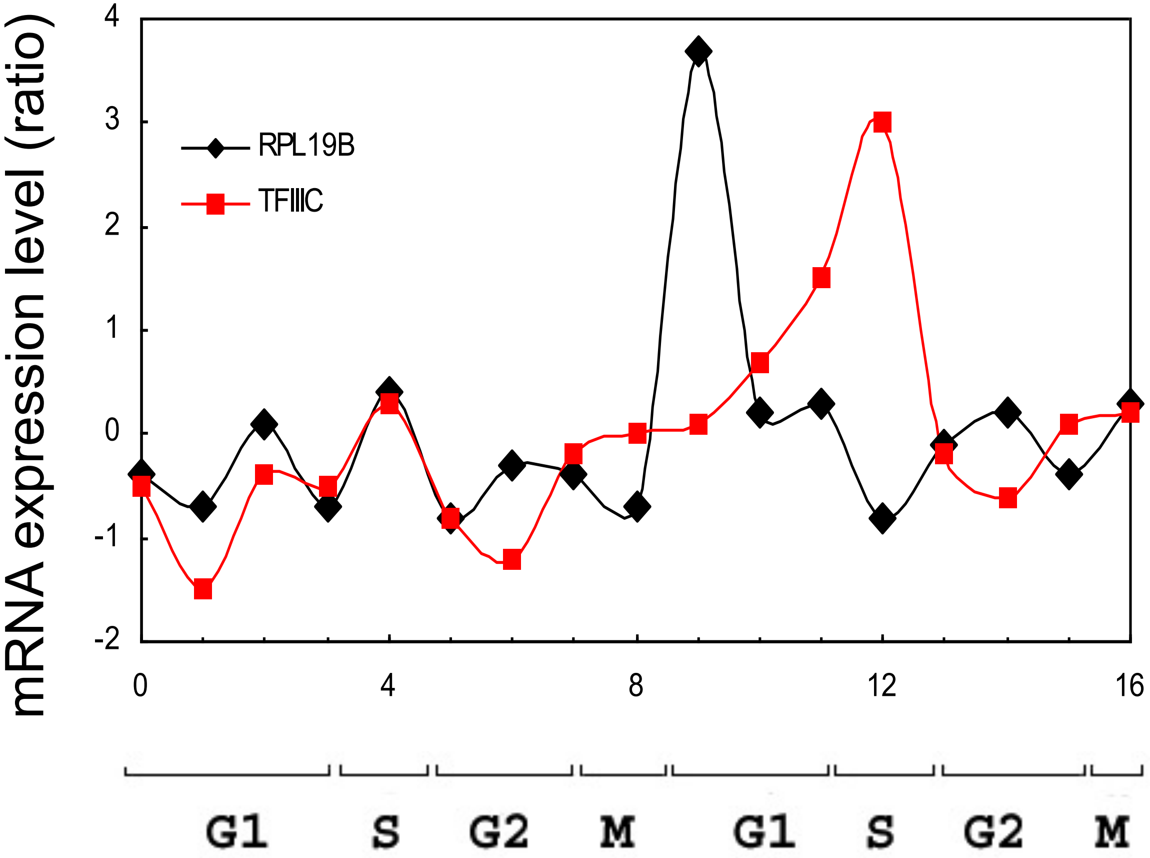
Time->



Microarray timecourse of 1 ribosomal protein

# Clustering the yeast cell cycle to uncover interacting proteins



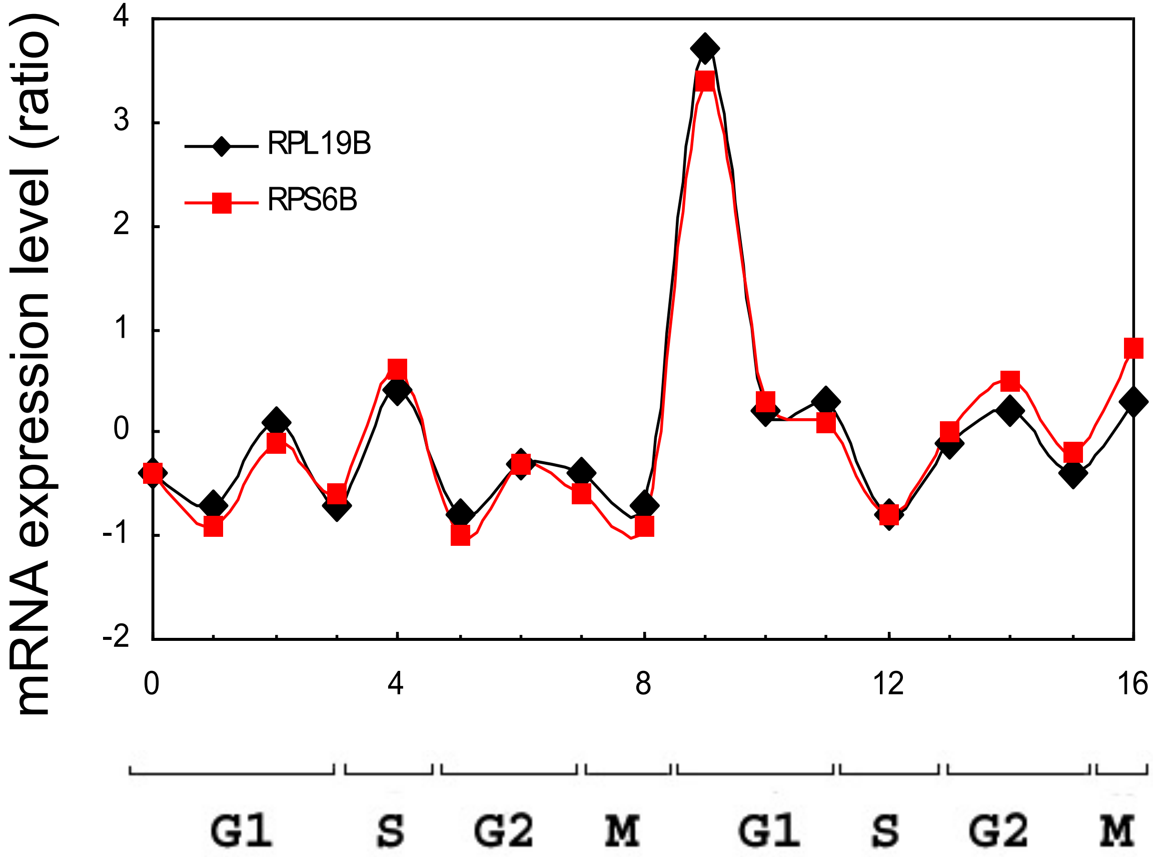


Time->

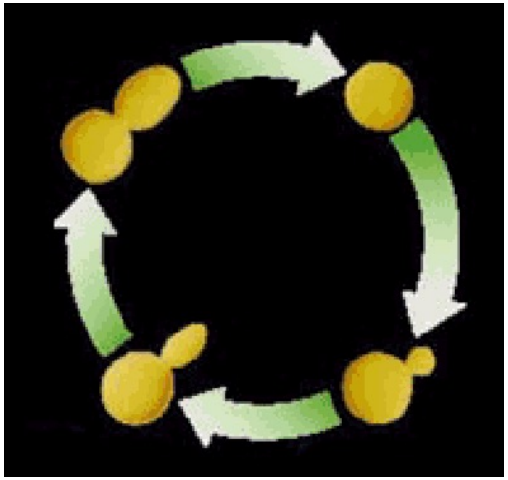
Random relationship from ~18M

# Clustering the yeast cell cycle to uncover interacting proteins

[Botstein; Church, Vidal]

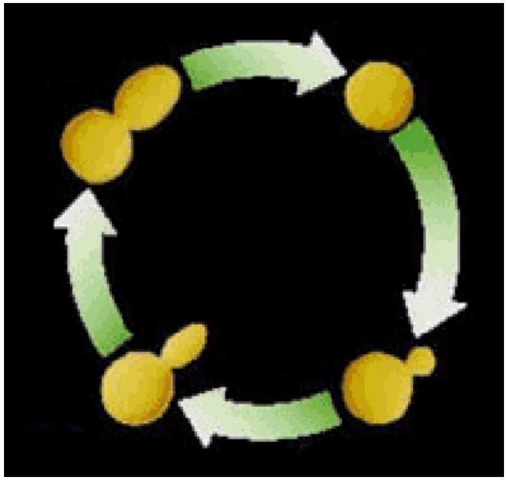


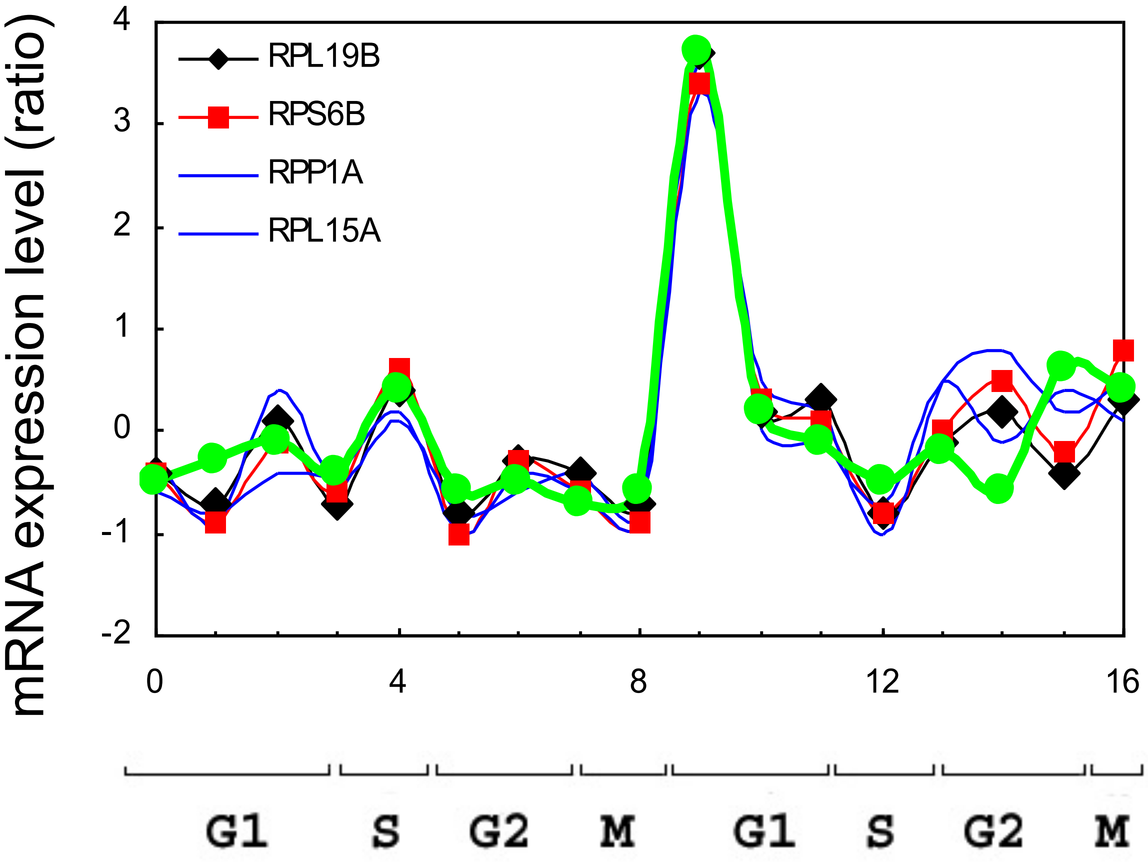
Time->



Close relationship from 18M
(2 Interacting Ribosomal Proteins)

# Clustering the yeast cell cycle to uncover interacting proteins



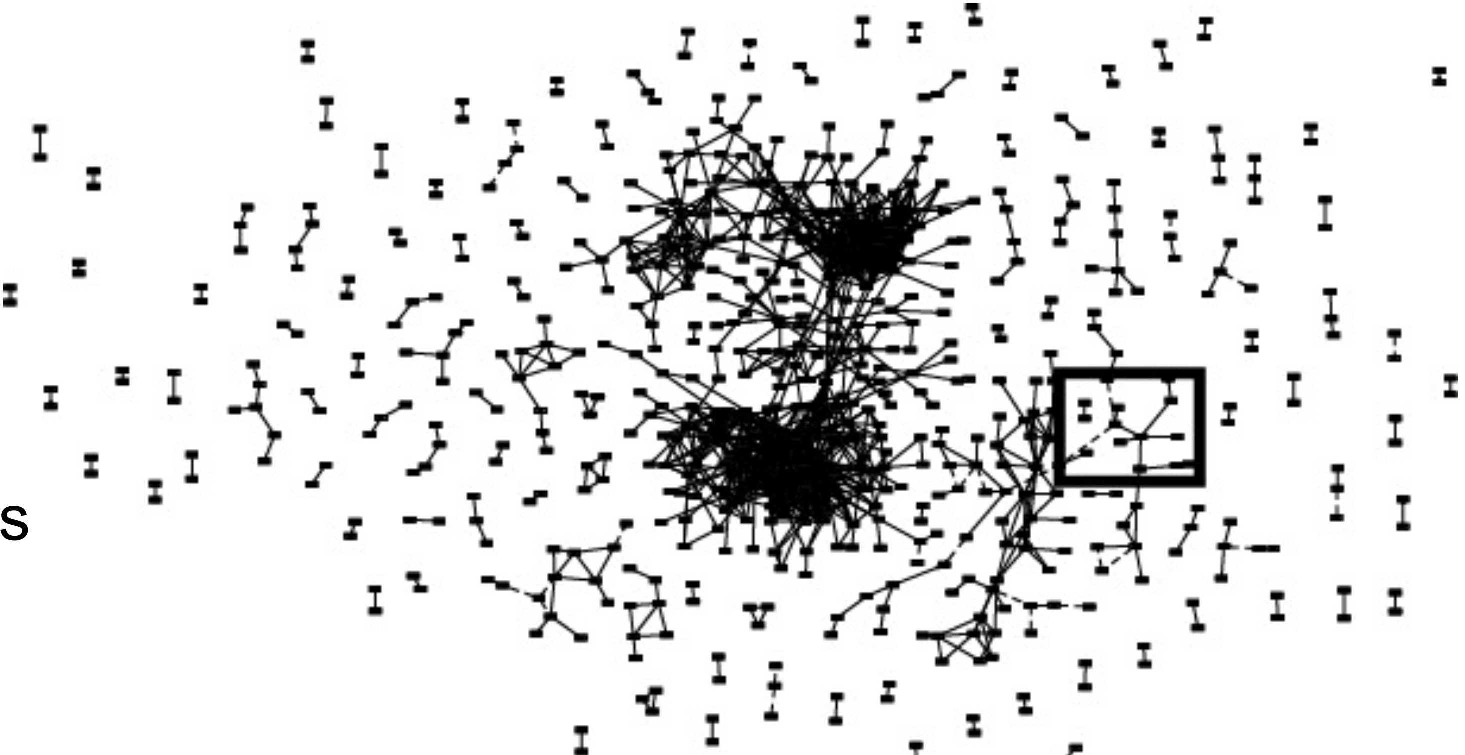


Time->

Predict Functional Interaction of Unknown Member of Cluster

●—●—●

# Global Network of Relationships



**~470K** significant relationships from **~18M** possible

# Unsupervised Mining

## General Thoughts on Clustering

# Current Clustering Methods

- Connectivity-based
- Centroid-based
- Distribution-based
- Density-based
- Community Detection

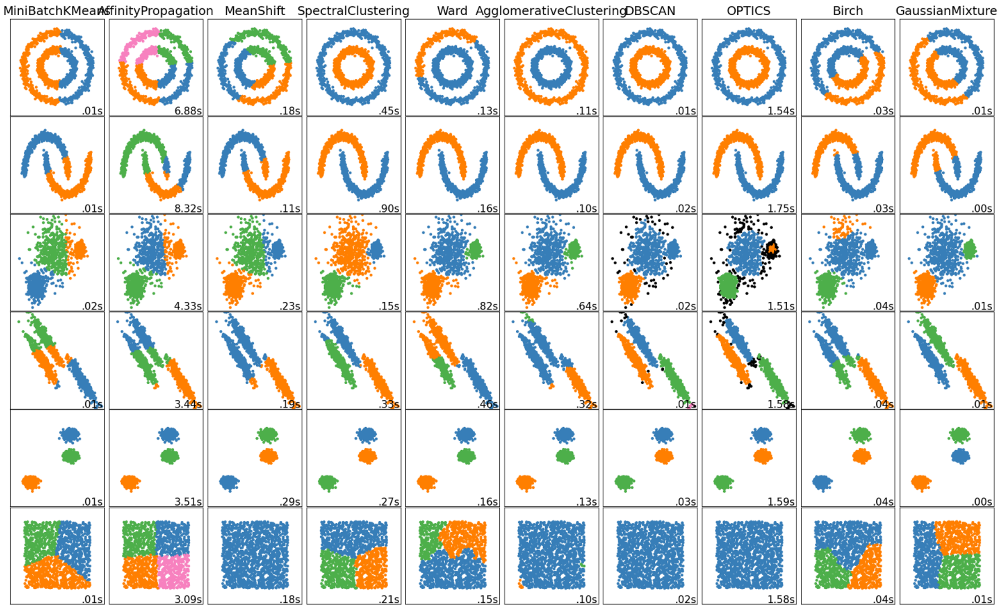


Image reference: https://scikit-learn.org/stable/modules/clustering.html

# Centroid-based Methods

- Optimizes a center vector to find data clusters
- Clusters data into a Voronoi diagram, which is interpretable
- Assumes a spherical shape for the clusters centered around the center vector
- E.g. K-means clustering
  - Heavily parameterized by K
  - Optimized by Lloyd's algorithm

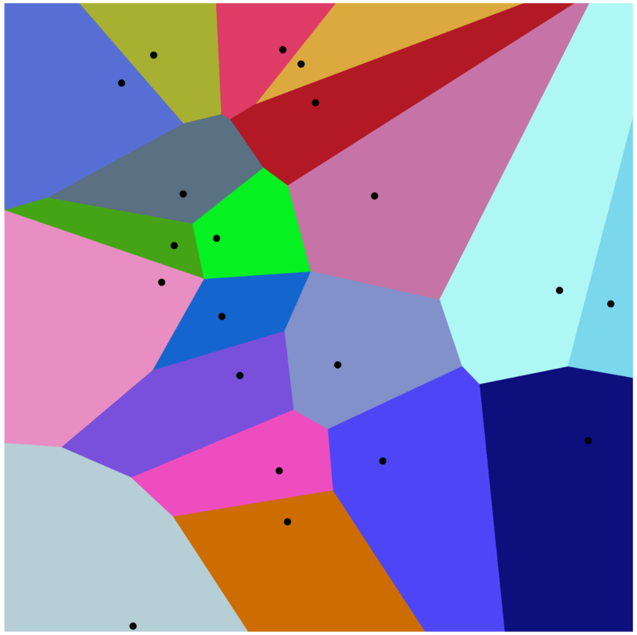
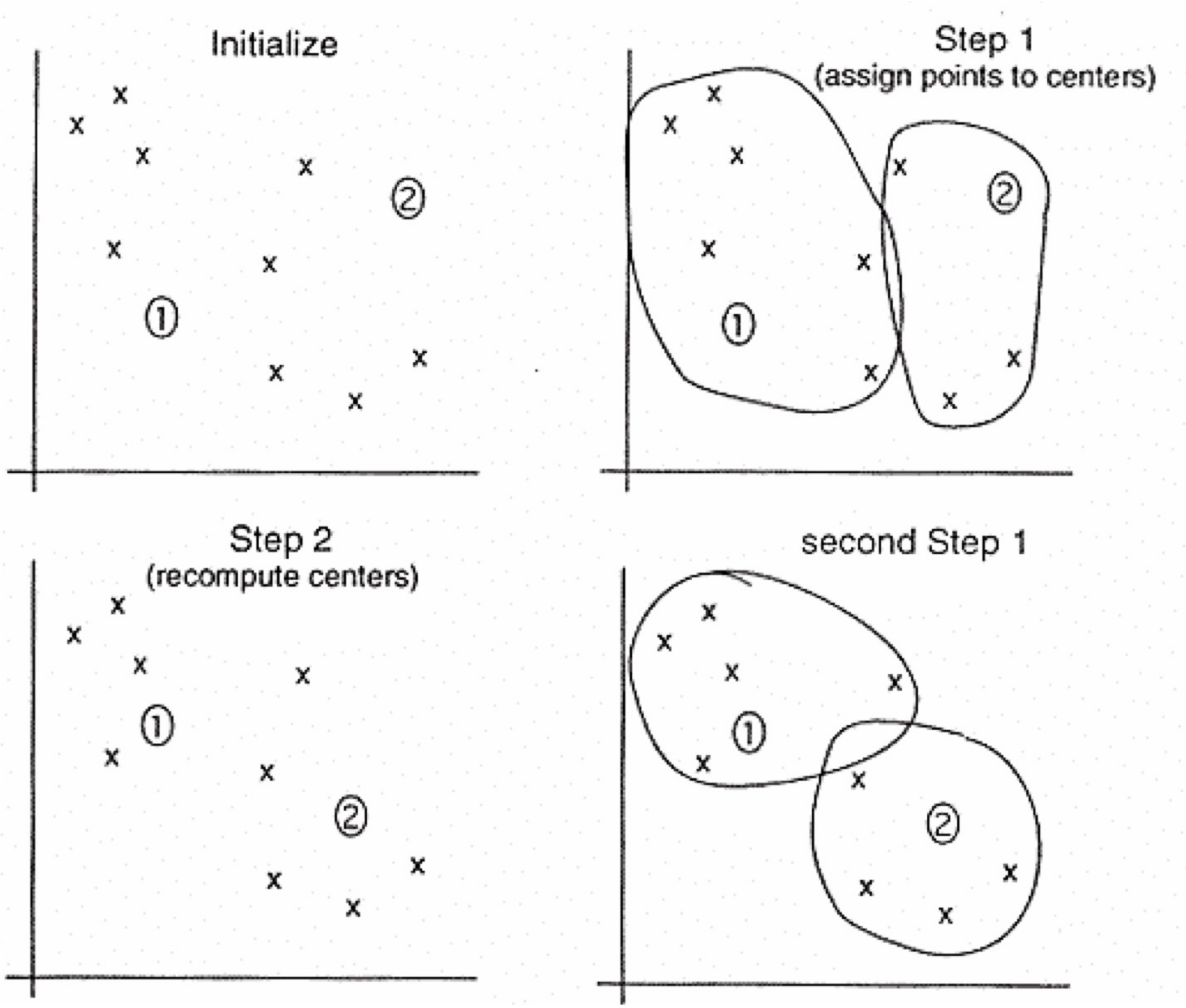


Image reference:
https://upload.wikimedia.org/wikipedia/commons/thumb/5/54/Euclidean_Voronoi_diagram.svg/1200px-Euclidean_Voronoi_diagram.svg.png

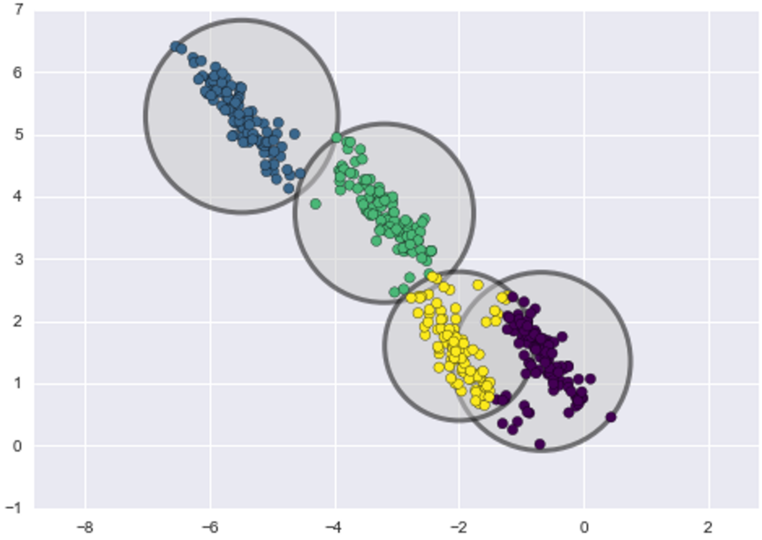
# K-means



1) Pick ten (i.e. k?) random points as putative cluster centers.
2) Group the points to be clustered by the center to which they are closest.
3) Then take the mean of each group and repeat, with the means now at the cluster center.
4) Stop when the centers stop moving.

# Distribution-based Methods

- Clusters are defined as samples from certain distributions
- Assumes the shape and number of distributions
- E.g. Gaussian Mixture Model Clustering
  - Can easily overfit by increasing the number of distributions

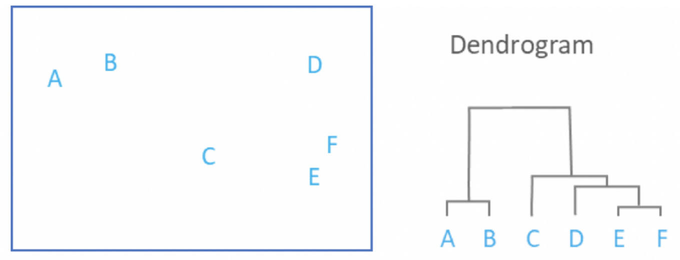


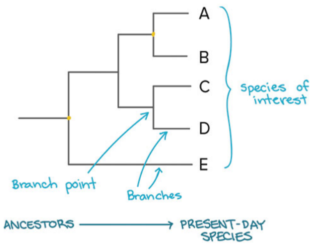
- LDA & tSNE (coming later)

Image reference: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

# Connectivity-based Methods

- Each data point start in their own cluster
- Iteratively merge clusters together based on some evaluation of distance to form a hierarchical structure
- Can be represented by a dendrogram (data point on one axis while tracking merge history on another axis)
- No definitive cut off, but can be used to trace developmental pseudo-time
- E.g. Hierarchical clustering





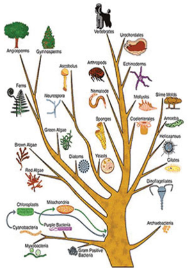
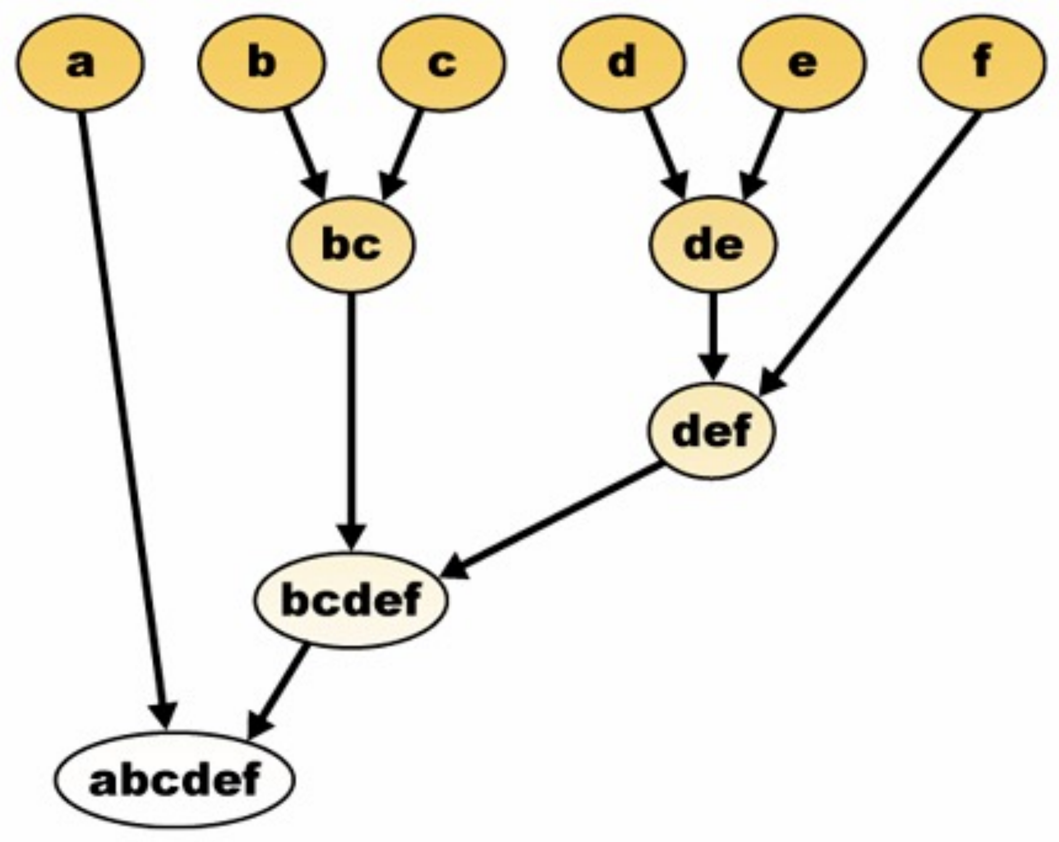


Image references:
https://46gyn61z4i0t1u1pnq2bbk2e-wpengine.netdna-ssl.com/wp-content/uploads/2018/03/Screen-Shot-2018-03-28-at-11.48.48-am.png
https://microbenotes.com/how-to-construct-a-phylogenetic-tree/

# Agglomerative Clustering

- Single or multi-link
- Threshold for connection?

- Mult-seq trees (earlier)



http://commons.wikimedia.org/wiki/File:Hierarchical_clustering_diagram.png

# Density-based Methods

- Utilize sparse regions and reachability to define clusters
- Assumes some range parameter
- E.g. DBSCAN
- Pro: Fast - O(nlogn) runtime
- Con: Some data points will not be assigned a cluster (undefined) because they are unreachable

- Edge detection

- MSB (earlier)

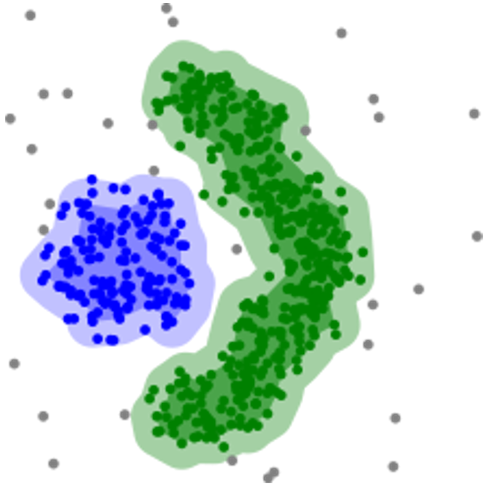


Image reference: https://upload.wikimedia.org/wikipedia/commons/thumb/0/05/DBSCAN-density-data.svg/440px-DBSCAN-density-data.svg.png