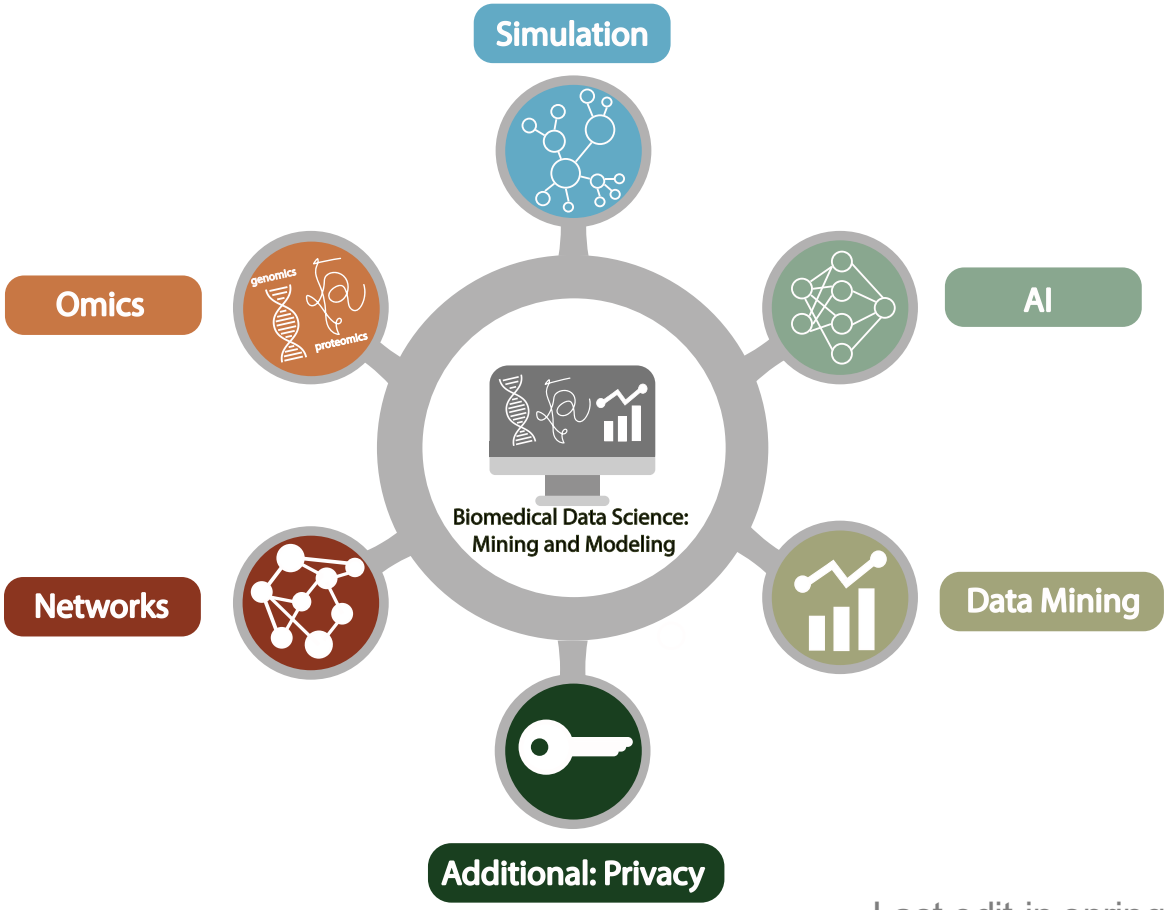
Biomedical Data Science (GersteinLab.org/courses/452)

Unsupervised Mining – General Clustering (25m9a)



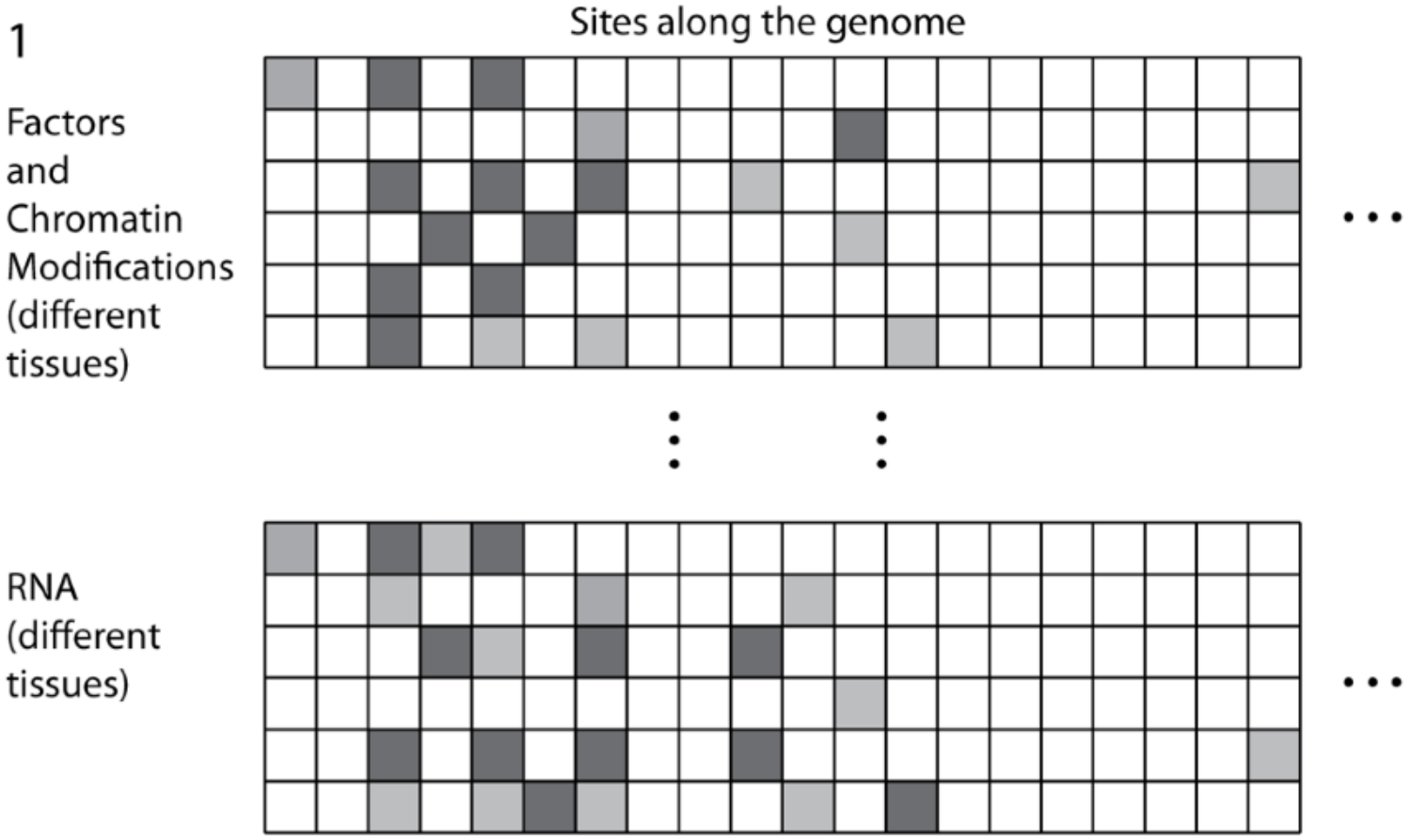
Mark Gerstein
Yale U.

Last edit in spring '25. Adding in semi-supervised clustering + more on K-means + ISLR reference. Similar to 2022's 22m9a & 2021's M9a [which has a video].

Unsupervised Mining

Columns & Rows
of the Data Matrix

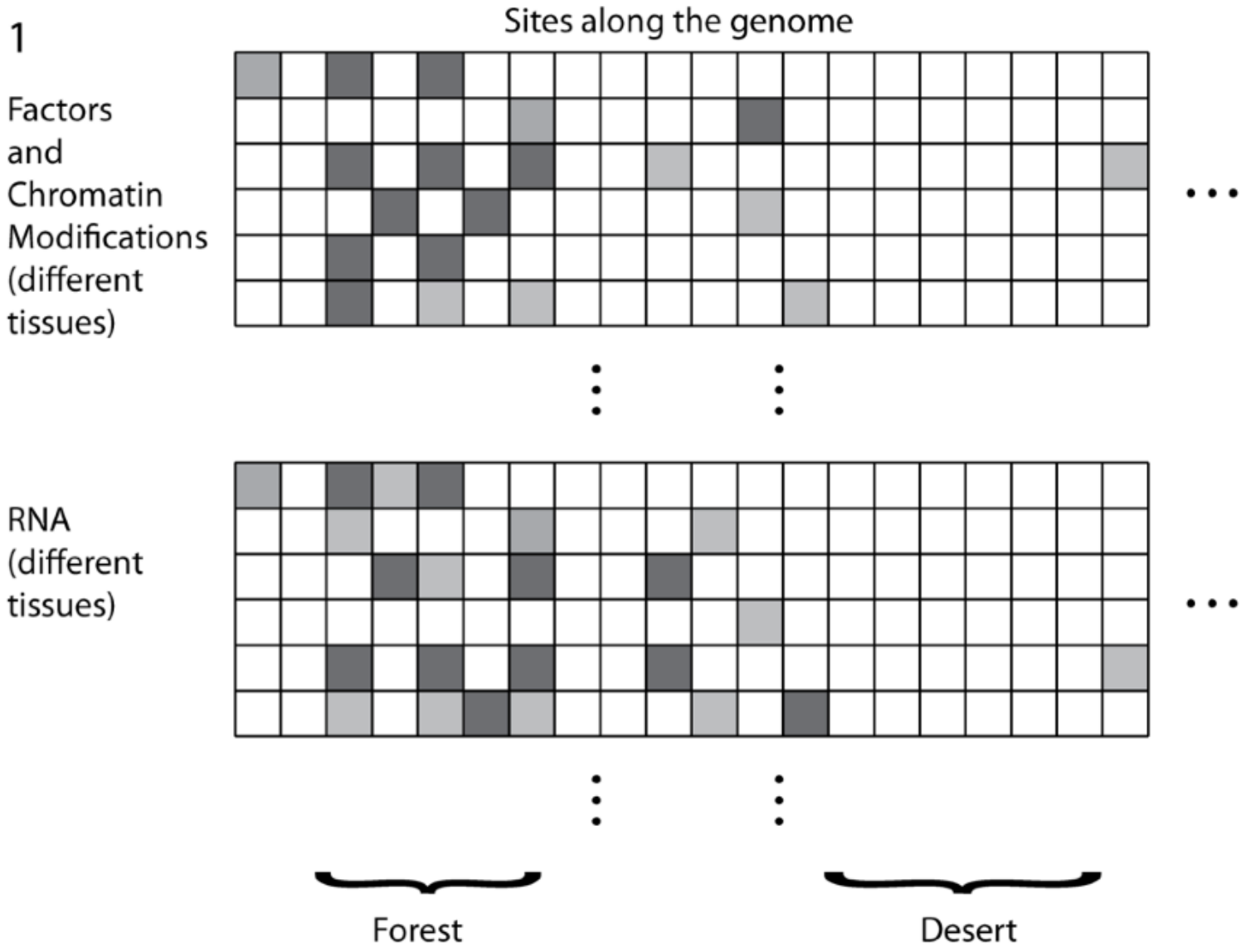
Structure of Genomic Features Matrix



Unsupervised Mining

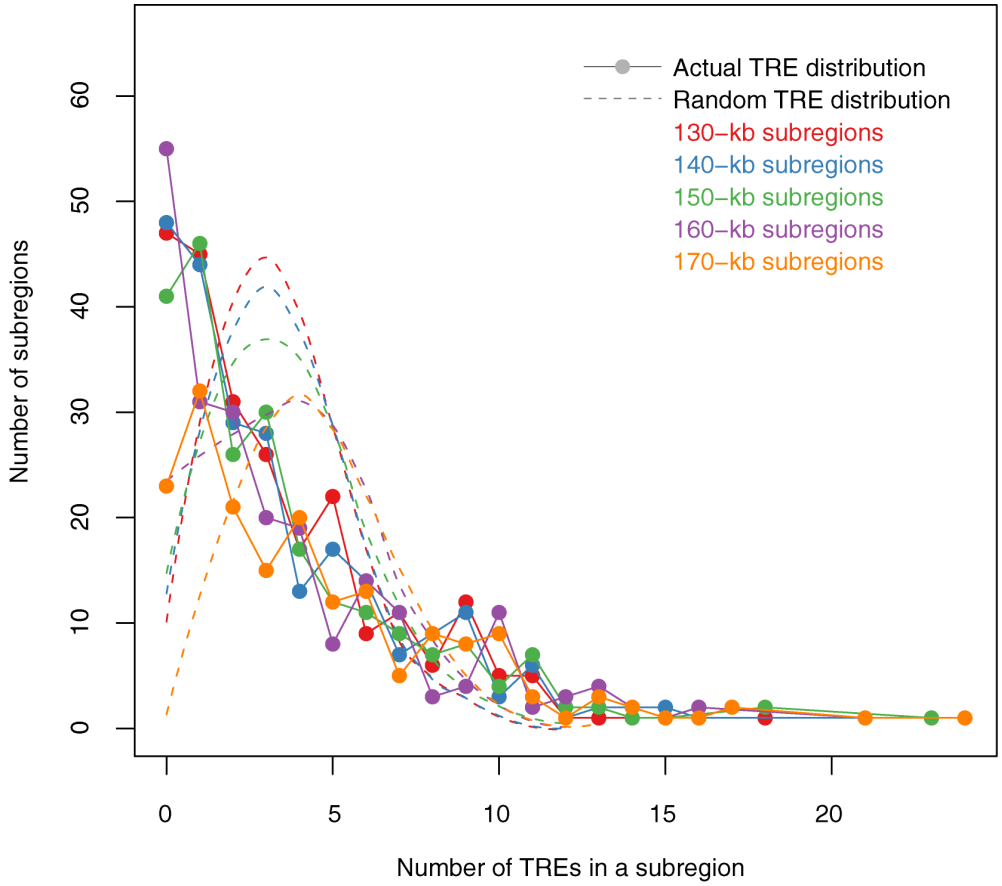
- Simple overlaps & enriched regions
- Clustering rows & columns (networks)
- PCA/SVD (theory + appl.)
- Biplot
- RCA
- CCA
- tSNE
- LDA
- (Embedding from Variational Autoencoders)

Genomic Features Matrix: Deserts & Forests



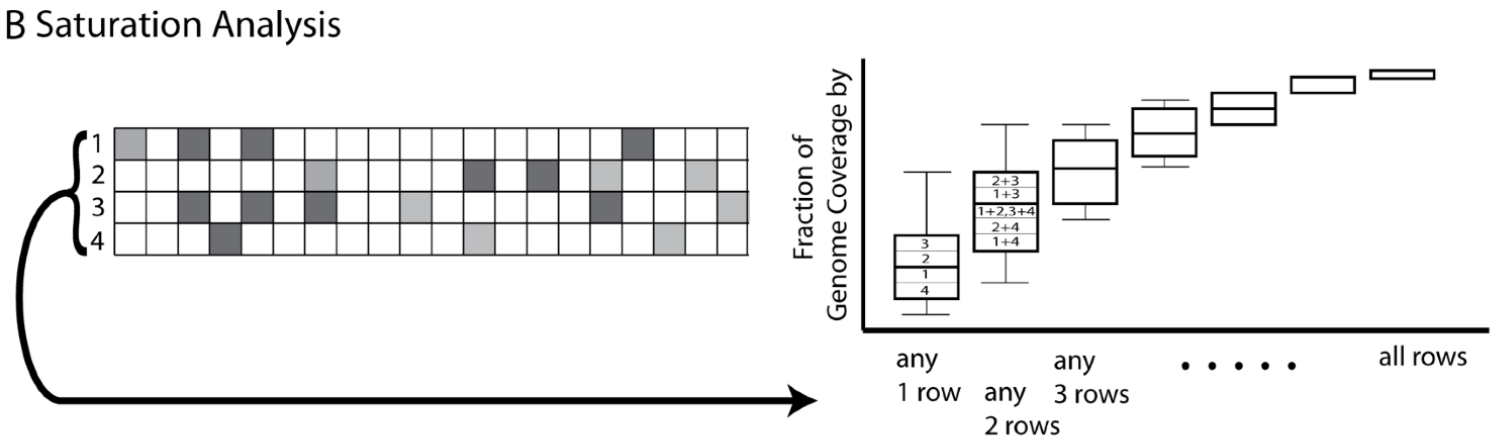
Modelling Distribution of Genomic Elements & Looking for Outliers

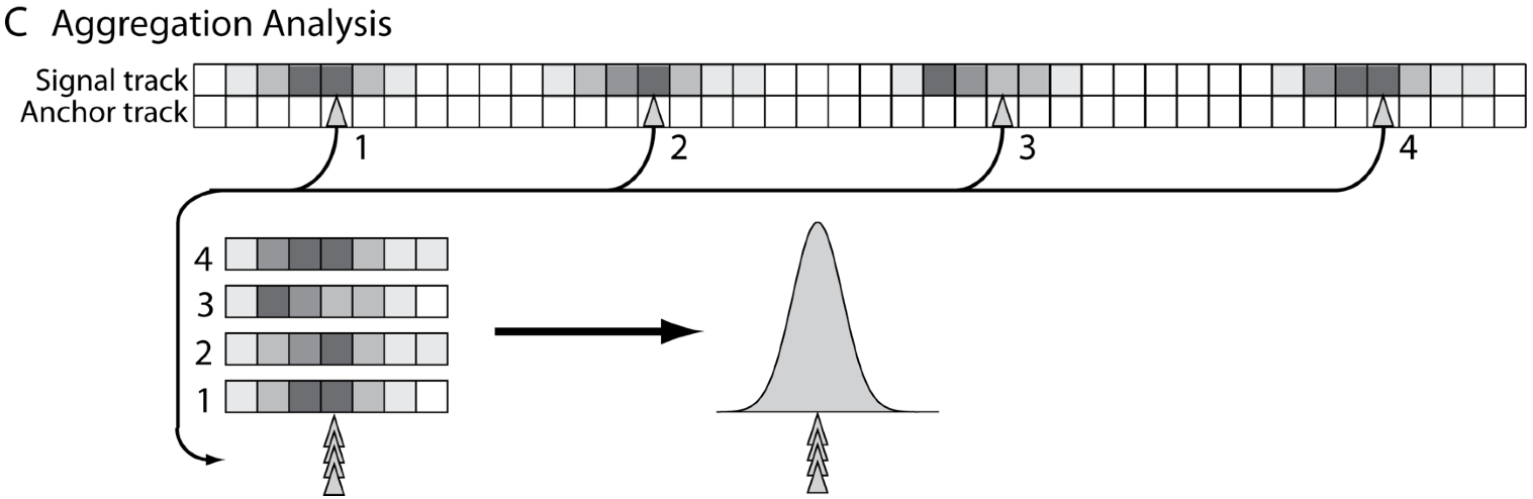
- TREs (Genomic Elements) are not evenly distributed throughout the genome
- The actual TRE distribution is power-law.
- The null distribution is ‘Poissonesque.’
- Many genomic subregions with extreme numbers of TREs.



Zhang et al. (2007) Gen. Res.

Aggregation & Saturation

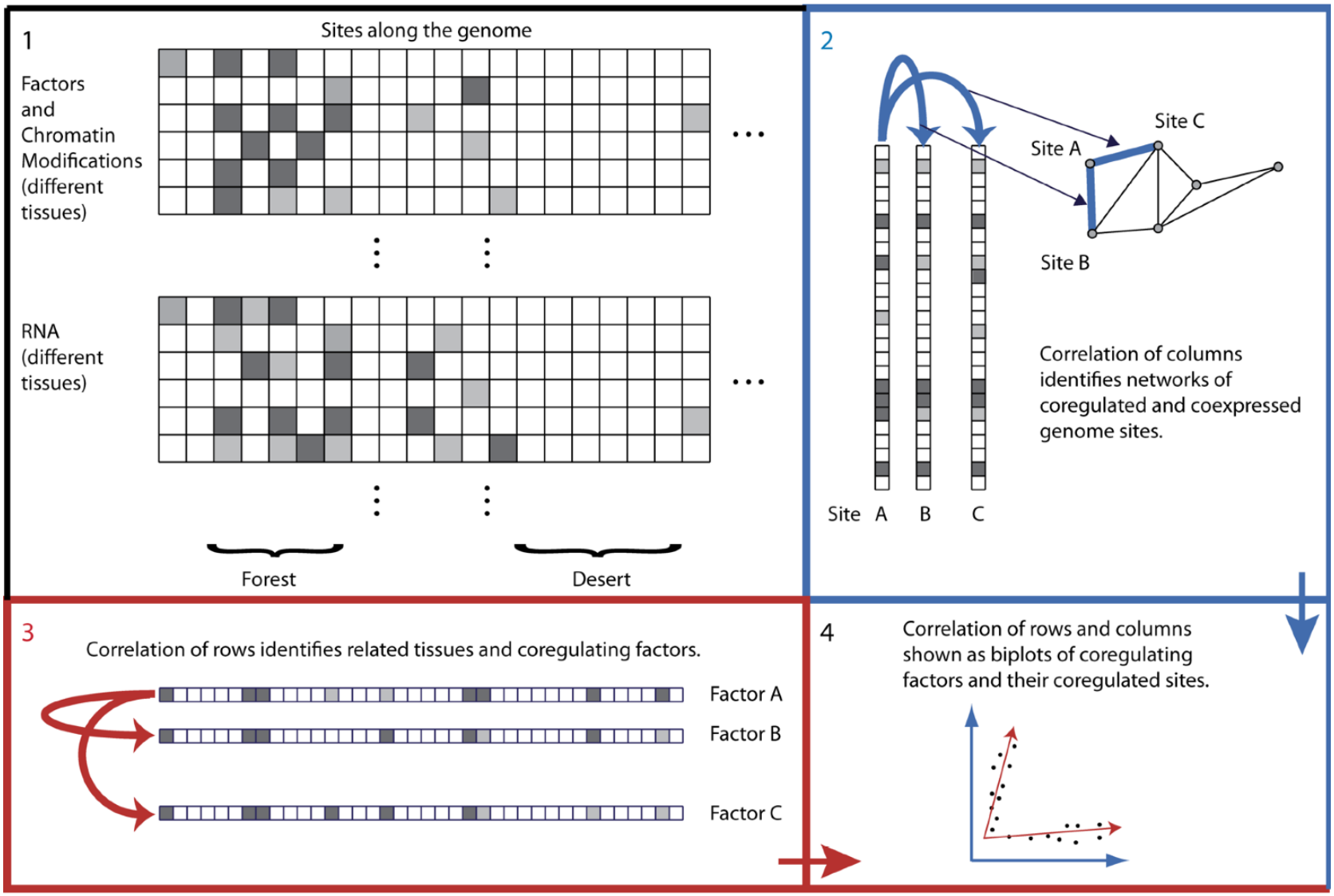




[*Nat. Rev. Genet.* (2010) 11: 559]

Expression Clustering

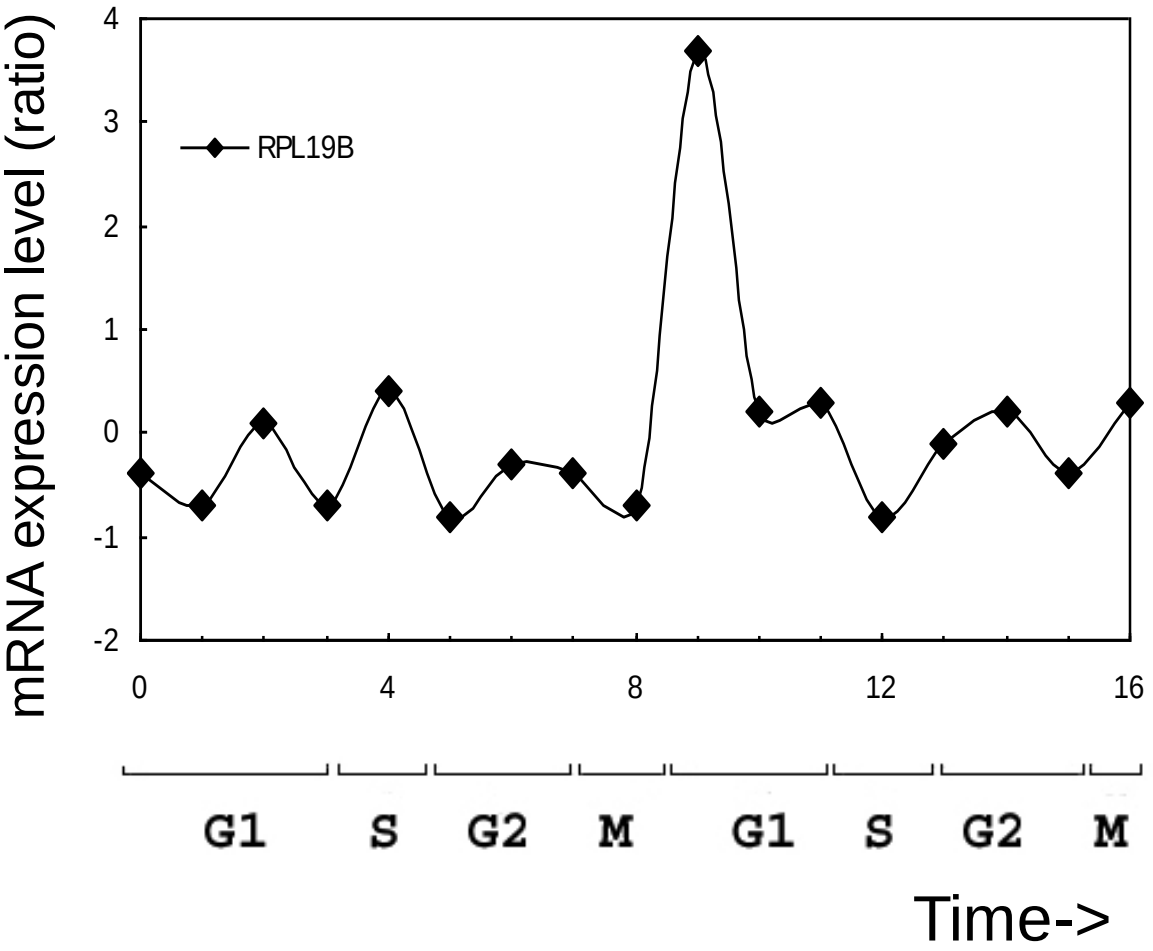
Correlating Rows & Columns

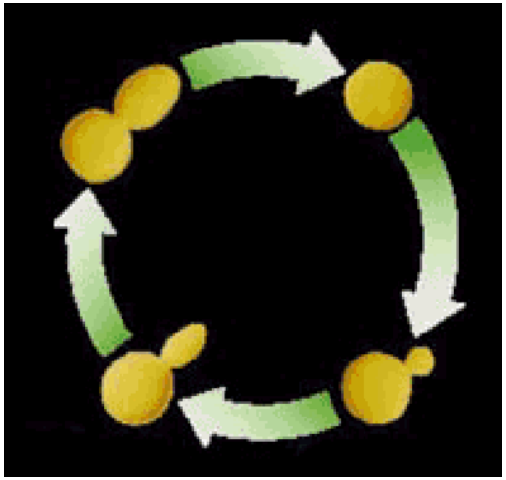


[*Nat. Rev. Genet.* (2010) 11: 559]

Clustering the yeast cell cycle to uncover interacting proteins

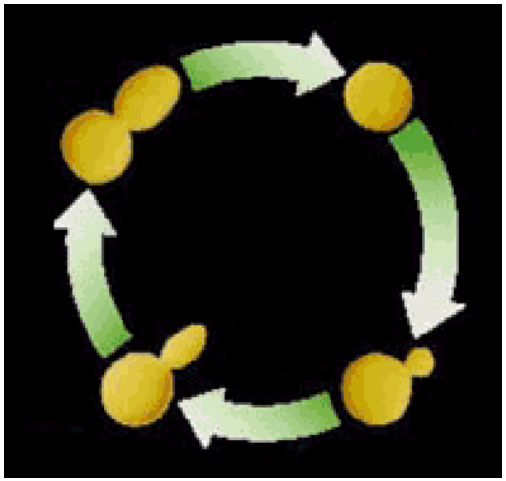
[Brown, Davis]

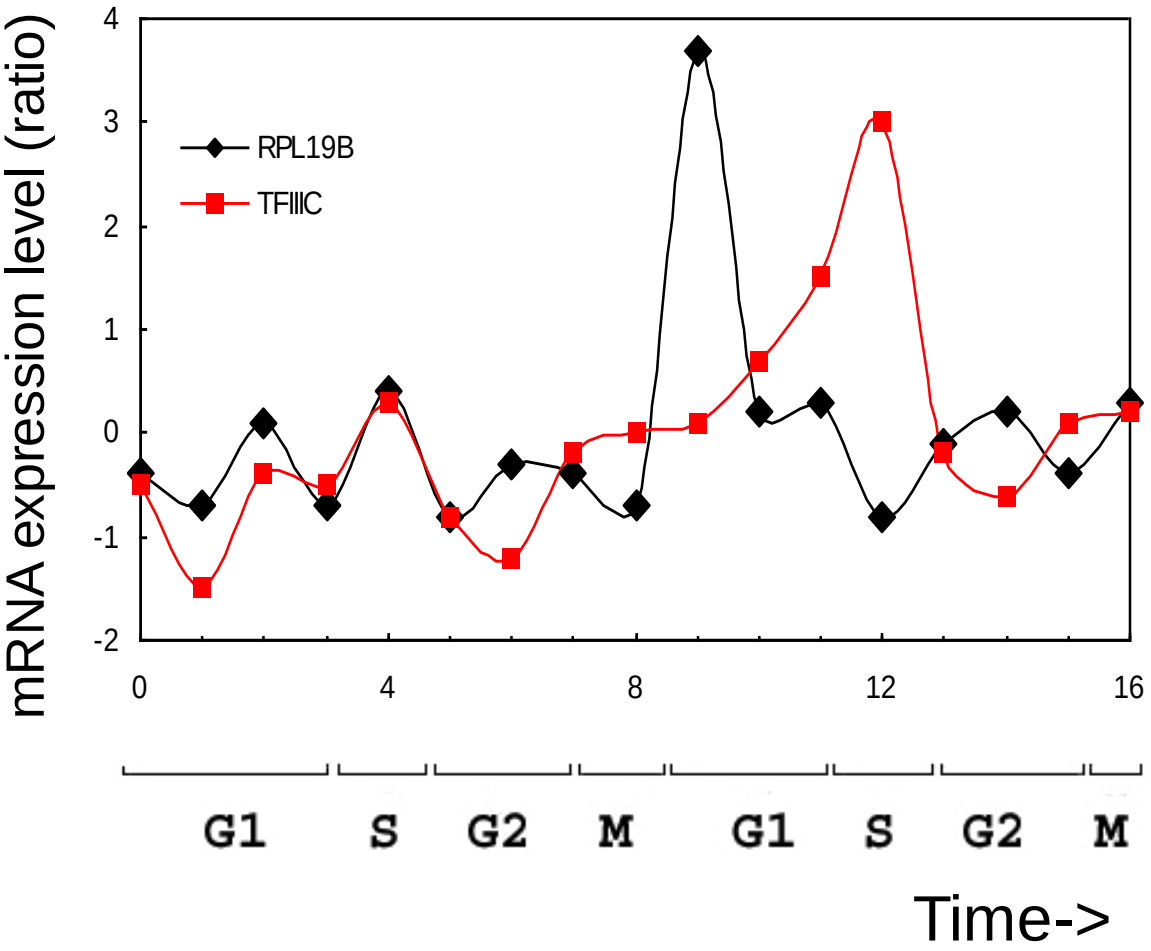




Microarray timecourse of
1 ribosomal protein

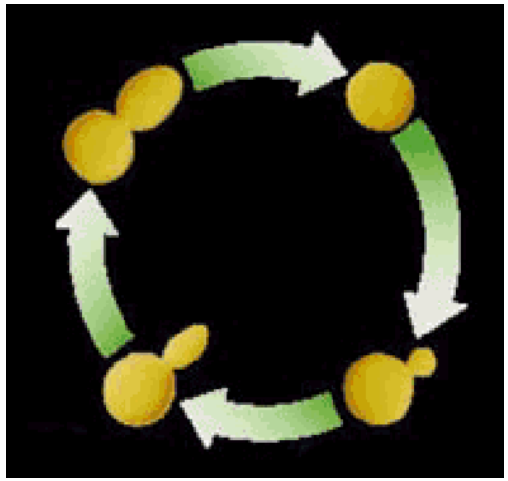
Clustering the yeast cell cycle to uncover interacting proteins

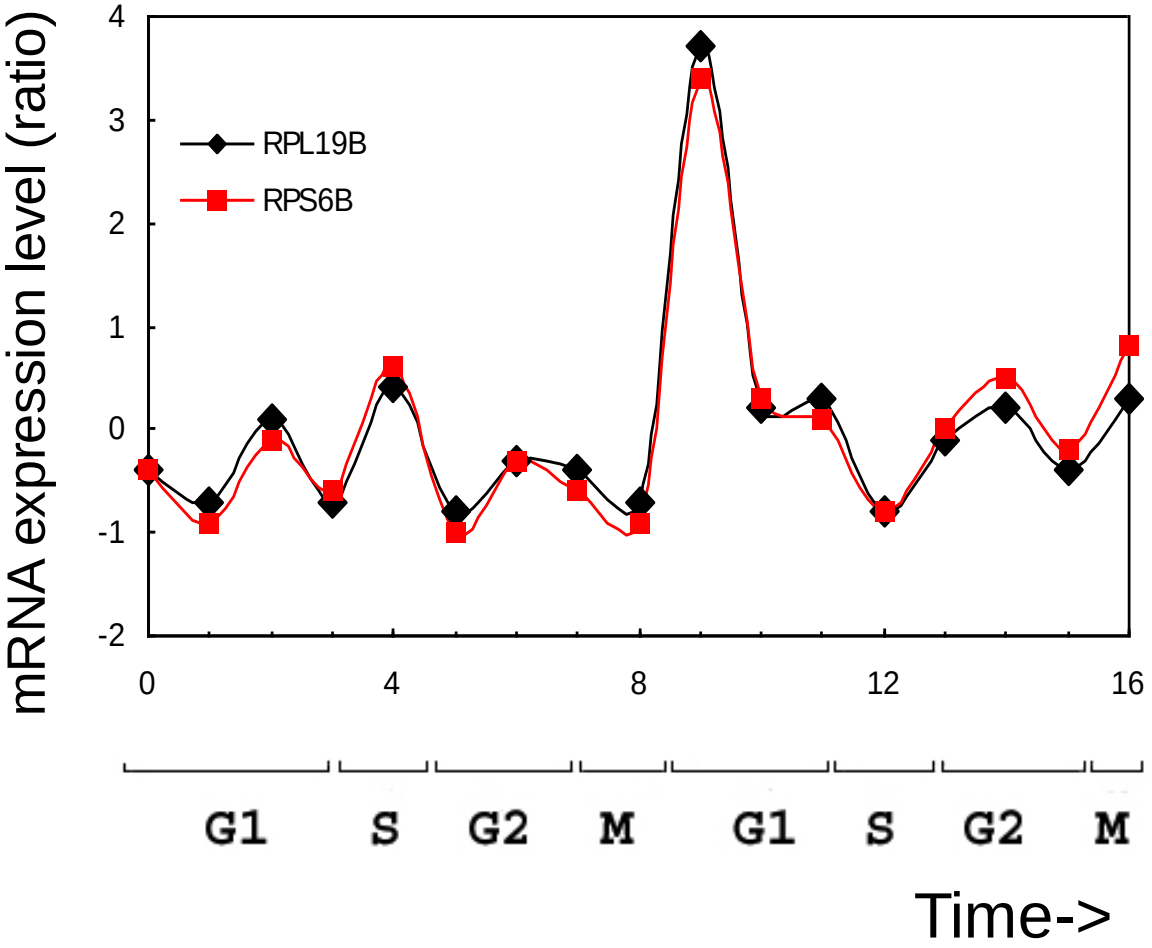




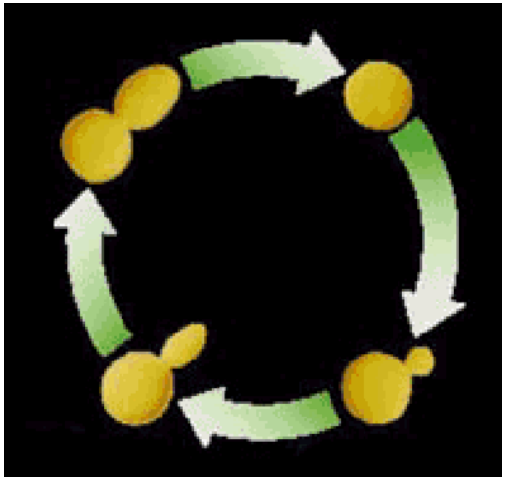
Clustering the yeast cell cycle to uncover interacting proteins

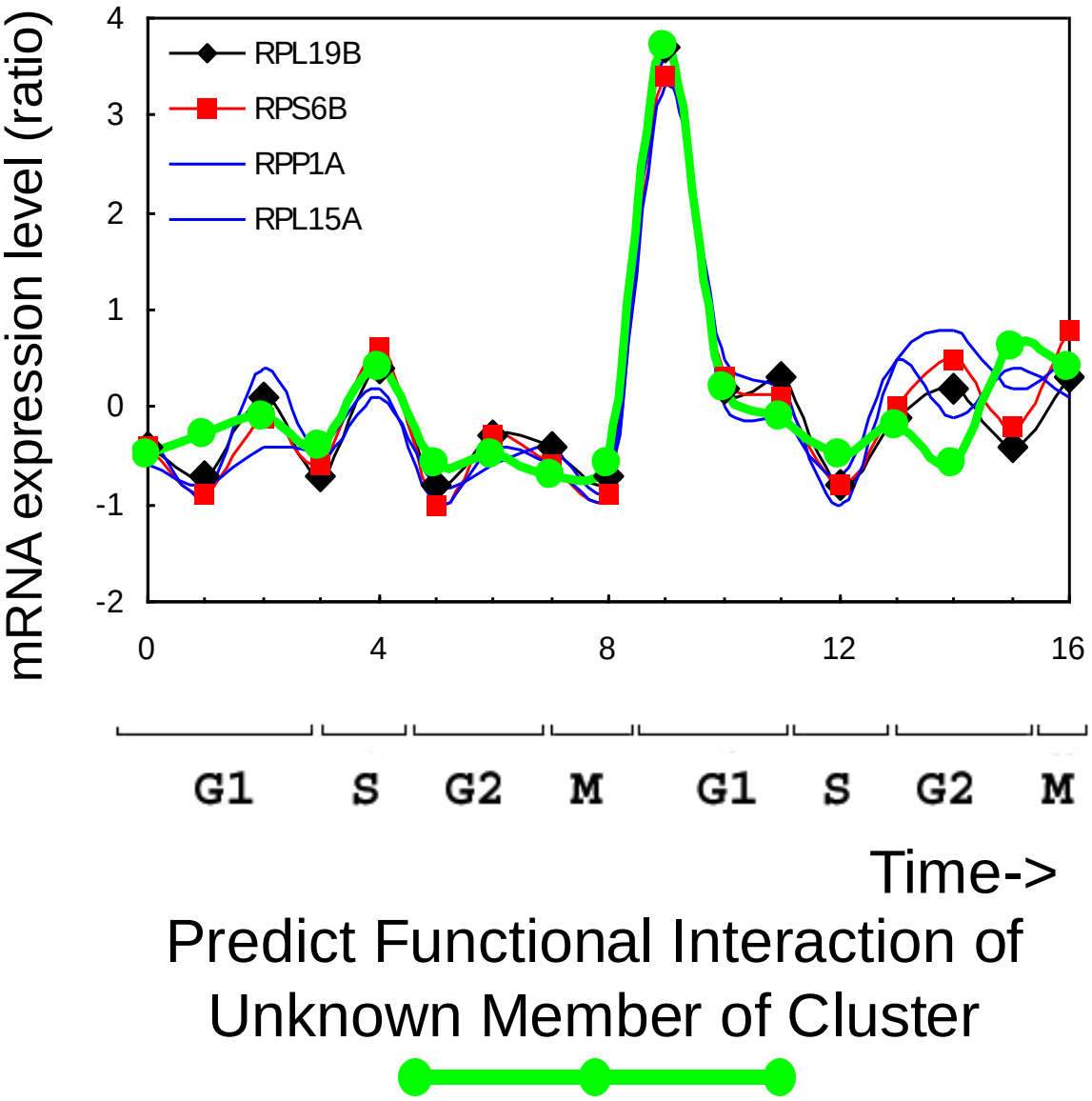
[Botstein; Church, Vidal]



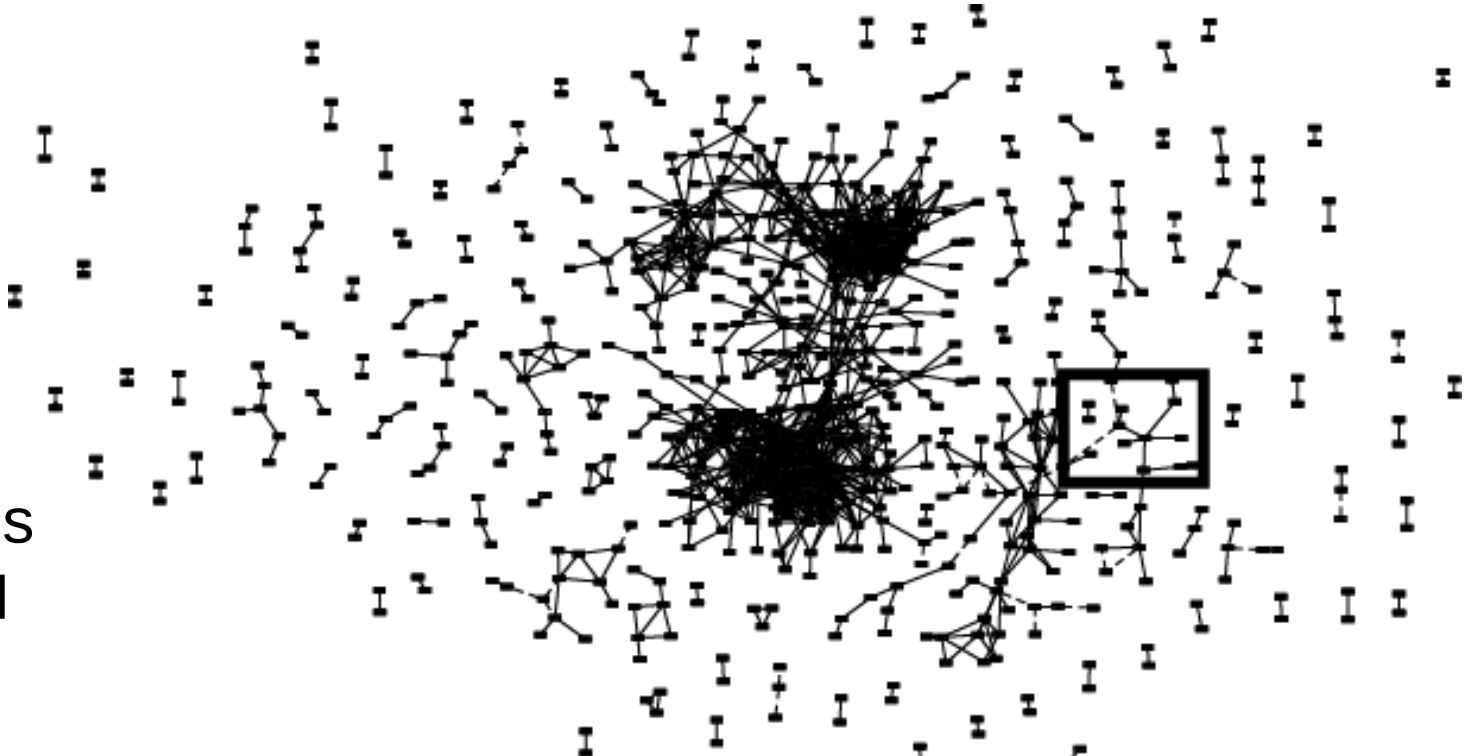


Clustering the yeast cell cycle to uncover interacting proteins





Global Network of Relationships



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

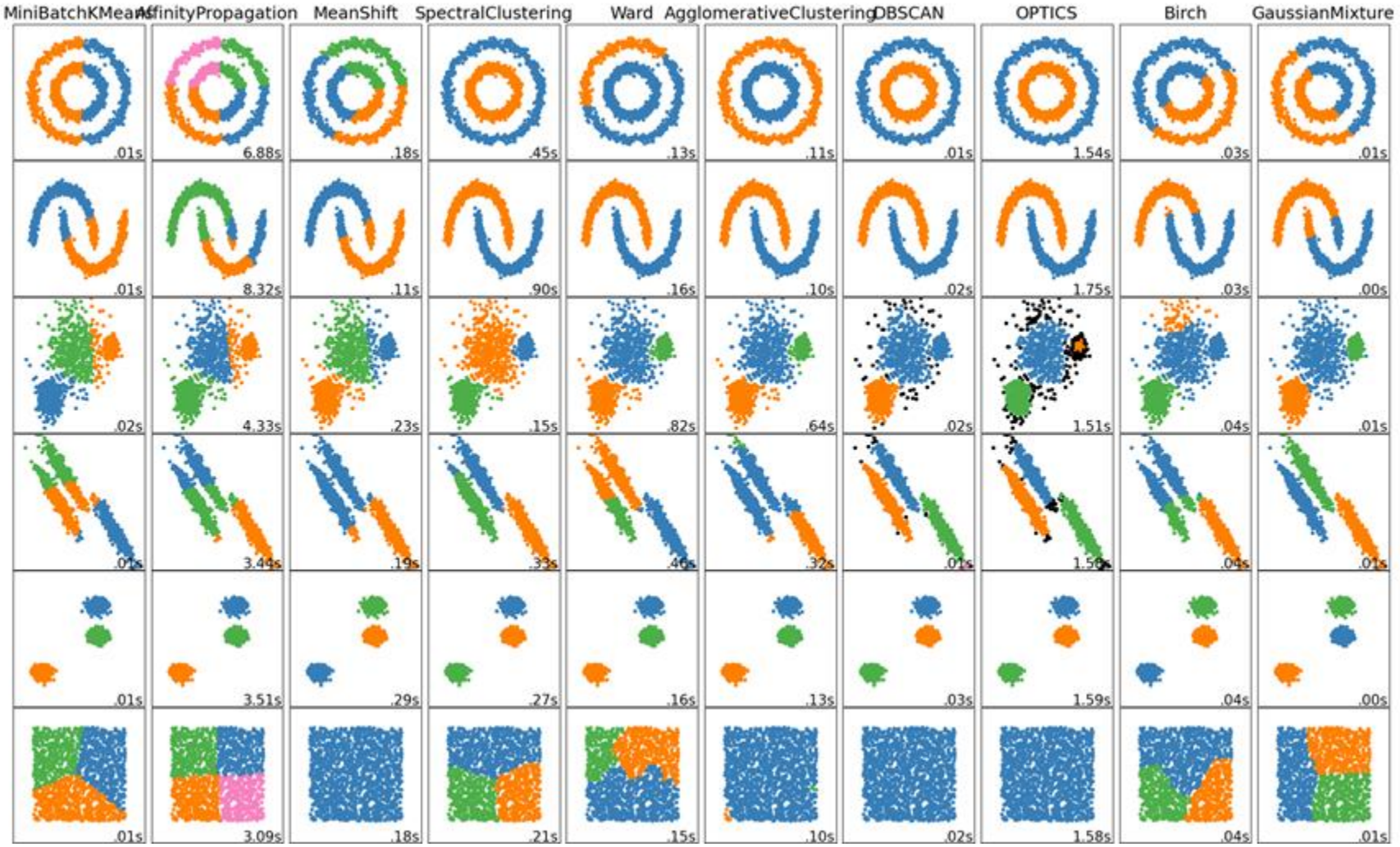
Unsupervised Mining

General Thoughts on Clustering

Overview of Clustering Methods (Very High Level)

- 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

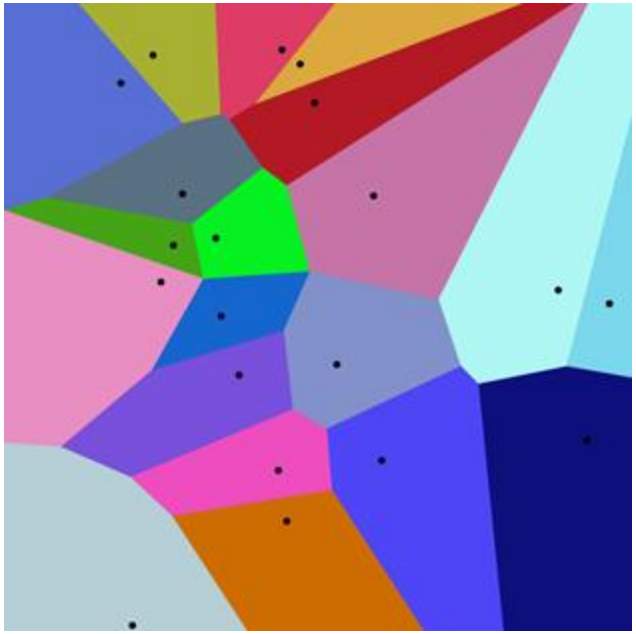


Image reference:
https://upload.wikimedia.org/wikipedia/commons/thumb/5/54/Euclidean_Voronoi_diagram.svg/1200px-Euclidean_Voronoi_diagram.svg.png

K-means

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

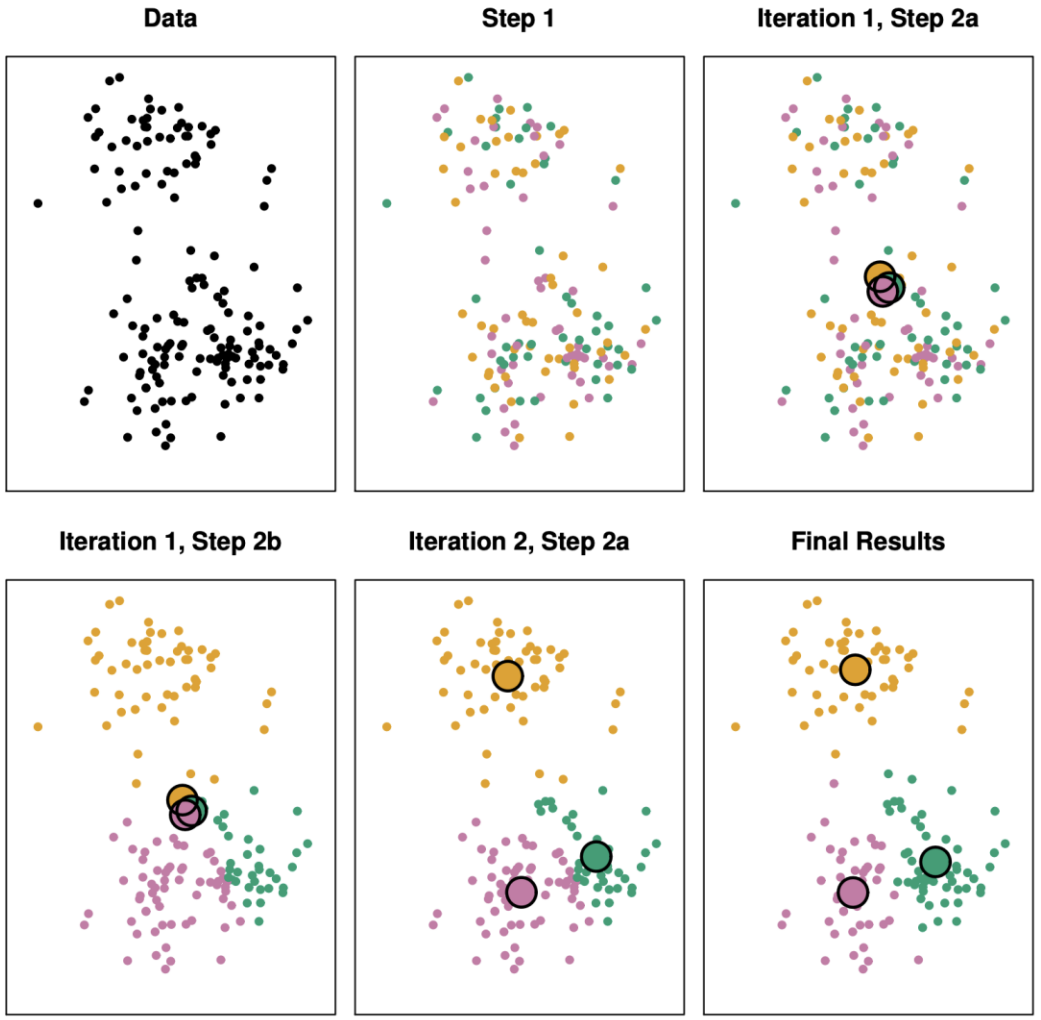


FIGURE 12.8. *The progress of the K-means algorithm on the example of Figure 12.7 with K=3.* Top left: *the observations are shown.* Top center: *in Step 1 of the algorithm, each observation is randomly assigned to a cluster.* Top right: *in Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.* Bottom left: *in Step 2(b), each observation is assigned to the nearest centroid.* Bottom center: *Step 2(a) is once again performed, leading to new cluster centroids.* Bottom right: *the results obtained after ten iterations.*

[ISLR, Chapter 12]

(More Later)

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

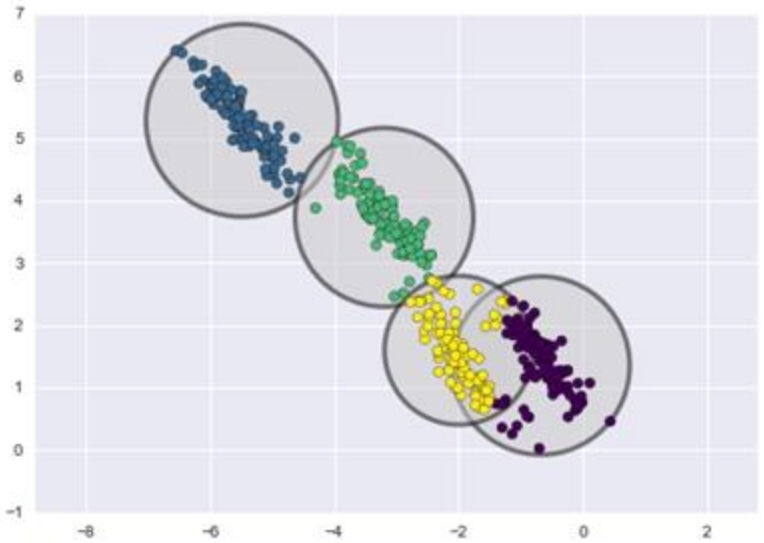
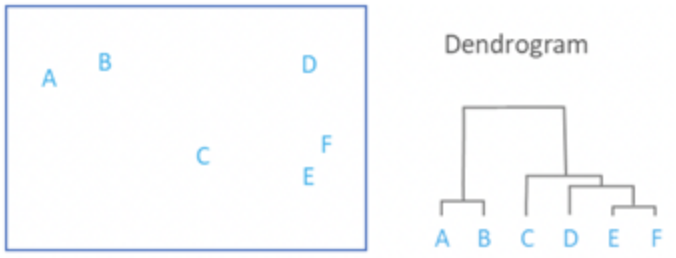


Image reference: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

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 & sequence trees (such as those for multiple alignment [mentioned earlier])



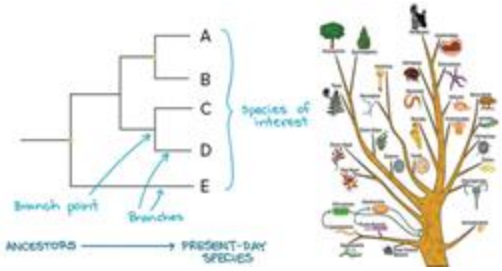


Image references:
https://46gyn61z4i0t1u1pnq2bbk2e-wpengine.netdna-ssl.com/wp-content/uploads/2018/03/Screen-Shot-2018-03-28-at-11.48.48-am.png
https://microbenotes.com/how-to-construct-a-phylogenetic-tree/

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

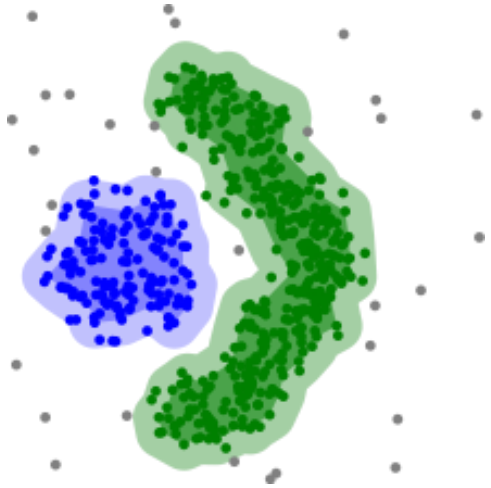
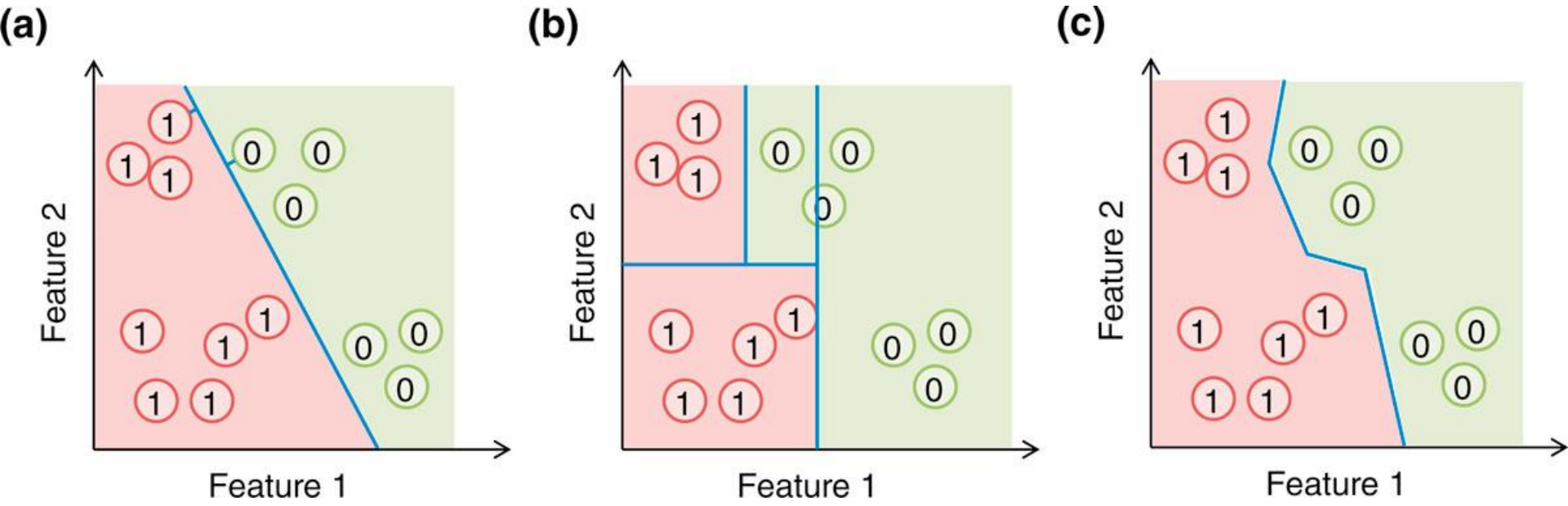


Image reference: https://upload.wikimedia.org/wikipedia/commons/thumb/0/05/DBSCAN-density-data.svg/440px-DBSCAN-density-data.svg.png

Semi-supervised Approaches

Decision boundaries: SVM v Tree v Nearest NBR



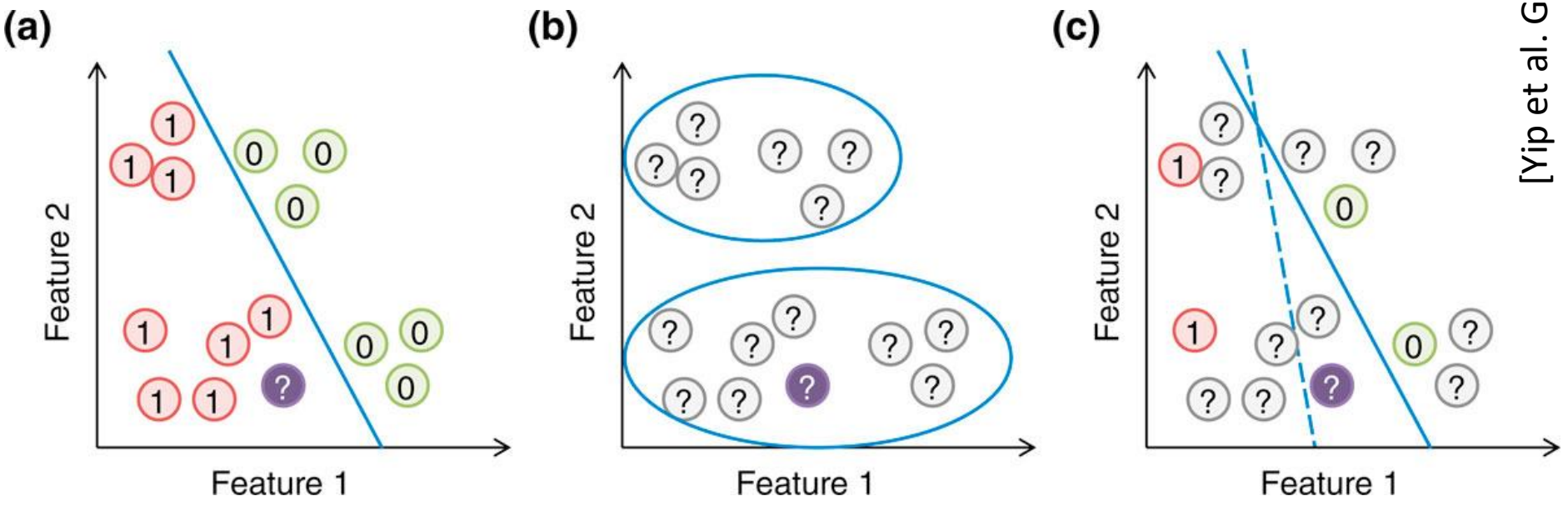
(a) A support vector machine (SVM) forms an affine decision surface (a straight line in the case of two dimensions) in the original feature space or a vector space defined by the similarity matrix (the kernel), to separate the positive and negative examples and maximize the distance of it from the closest training examples (the support vectors, those with a perpendicular line from the decision surface drawn). It predicts the label of a genomic region based on its direction from the decision surface. In the case a kernel is used, the decision surface in the original feature space could be highly non-linear. (b) A basic decision tree uses feature-parallel decision surfaces to repeatedly partition the feature space, and predicts the label of a genomic region based on the partition it falls within. (c) The one-nearest neighbor (1-NN) method predicts the label of a genomic region based on the label of its closest labeled example. In all three cases, the areas predicted to be positive and negative are indicated by the red and green background colors, respectively.

[Yip et al. Genome Biology 2013 14:205 doi:10.1186/gb-2013-14-5-205]

Semi-supervised Methods

- Supervised & Unsupervised:
  Can you combine them? YES
  - RHS (c) shows modifying the optimum decision boundary in (a) by "clustering" of unlabeled points

[Yip et al. Genome Biol. ('13)]



Supervised, unsupervised and semi-supervised learning. (a) In supervised learning, the model (blue line) is learned based on the positive and negative training examples, and the genomic region without a known class label (purple circle) is classified as positive according to the model. (b) In unsupervised learning, all examples are unlabeled, and they are grouped according to the data distribution. (c) In semi-supervised learning, information of both labeled and unlabeled examples is used to learn the parameters of the model. In this illustration, a purely supervised model (dashed blue line) classifies the purple object as negative, while a semi-supervised model that avoids cutting at regions with a high density of genomic regions (solid blue line) classifies it as positive.

References

- James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert
  An Introduction to Statistical Learning: with Applications in R
  [ISLR (2nd edition)]
  https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1071614177/ + https://www.statlearning.com
  (Chapter 12.1 and 12.4 gives background on Clustering.)

- Peixeiro, M. (2021, December 11). Towards Data Science. Medium.
  The complete Guide to Unsupervised Learning
  https://towardsdatascience.com/the-complete-guide-to-unsupervised-learning-ecf8b676f2af
  (Optional extra background on clustering)