Clustering Algorithms, Regulatory Networks

May 3, 2022

Acknowledgement: Materials in some slides are borrowed from Harvard STAT115 course taught by X. Shirley Liu.

Outline

- Clustering
  - Hierarchical clustering (e.g., WGCNA)
  - K-means clustering
  - Louvain method (e.g., scRNA-seq)
- Regulatory Networks
  - Gene Ontology
  - GSEA
  - BART

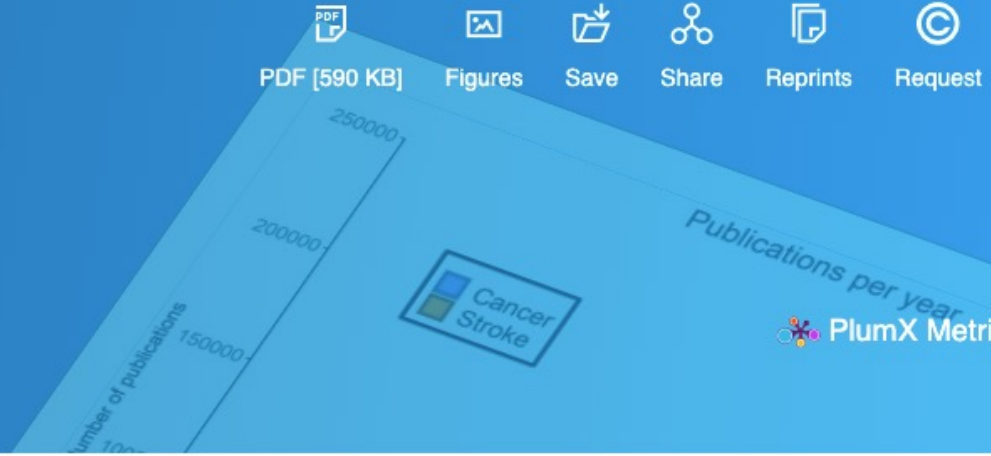
Why clustering?
- Genes do not work alone.

FORUM | VOLUME 38, ISSUE 3, P216-217, MARCH 01, 2022

Every gene can (and possibly will) be associated with cancer

João Pedro de Magalhães

Published: October 27, 2021 • DOI: https://doi.org/10.1016/j.tig.2021.09.005 • Check for updates



A PubMed analysis shows that the vast majority of human genes have been studied in the context of cancer. As such, the study of nearly any human gene can be justified based on existing literature by its potential relevance to cancer. Moreover, these results have implications for analyzing and interpreting large-scale analyses.

Keywords

genetics • network biology • oncology • research • science

Cancer as the most studied biomedical topic

Cancer is one of the most common diseases of modern times. In industrialized countries, cancer affects roughly one in two people at some point during their lives [1.] and cancer incidence and mortality is expected to continue increasing given the ageing populations worldwide [2.]. Not surprisingly, cancer attracts a huge amount of research funding from government, private, and philanthropic sources [3.]. At the time of writing, over 4 million of the over 30 million publications in PubMed mention cancer. For comparison, roughly 350 000 publications mention stroke. As of 2020, over 200 000 papers are

Clustering

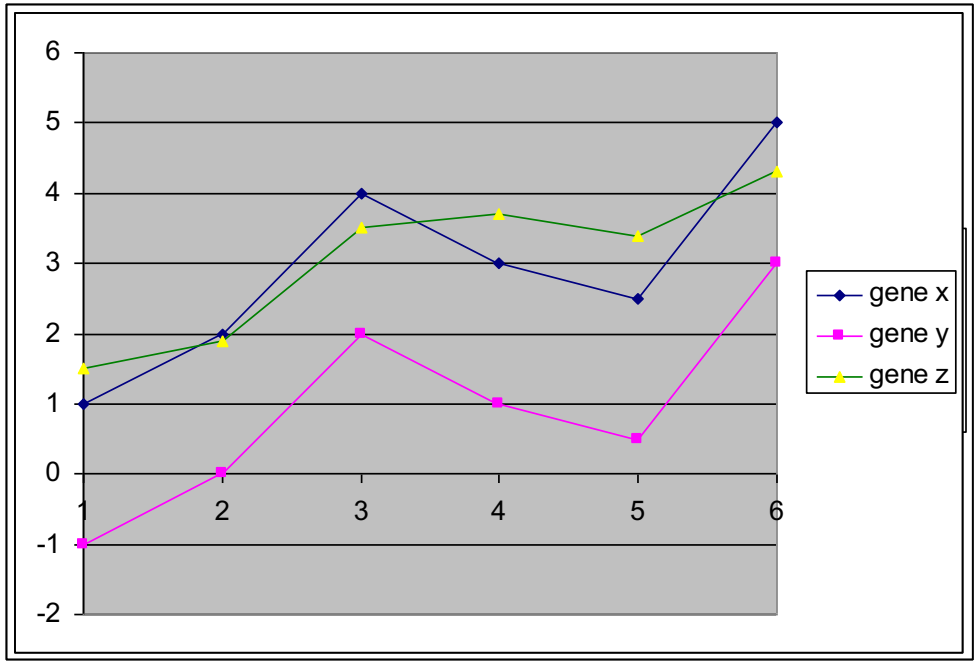
- We can cluster either genes or samples, or both
  - Genes: have similar expression profiles over different samples or conditions
  - Samples: have similar expression profiles over all genes

probe set	gene	Normal m4	Normal m4	Normal m4	Normal m	Normal m4	MM m282	MM m331a	MM m332a	MM m333a	MM m334a	MM m353a
31307_at	pre-T/NK c	28.53	32.61	29.56	36.59	33.19	25.1	32.79	34.3	35.44	28.48	29.55
31308_at	pre-T/NK c	69.14	53.69	52.78	62.07	58.74	67.88	85.82	83.54	85.91	60.93	62.82
31309_r_a	Human bre	16.9	67.7	27.61	46.16	51.46	45.62	35.57	32.62	35.14	96.18	45.94
31310_at	glycine rec	67.42	49.55	55.51	59.57	68.42	91.06	91.23	83.66	76.37	71.23	74.95
31311_at	Homo sapi	78.73	62.91	60.84	72.98	72.9	79.39	85.52	82.57	69.69	63.72	64.29
31312_at	potassium	66.68	59.46	55.47	61.79	69.92	75.28	85.53	97.91	69.92	74.77	71.83
31313_at	mannosyl	115.33	95.51	84.48	94.99	109.04	105.05	118.68	106.76	142.88	103.72	106.19
31314_at	bone morp	71.88	36.24	41.86	46.99	45.94	46.67	67.56	66.14	53.95	40.97	47.96
31315_at	immunoglo	103.99	86.27	83.81	81.81	254.63	87.12	99.11	109.56	86.37	75.03	74.97
31316_at	Human vac	16.79	10.08	9.53	16.48	11.98	12.8	16.7	18.76	11.25	12.09	18.89
31317_r_a	Human unp	316.79	269.61	254.92	352.61	342.4	327.12	366.39	346	308.43	279.81	312.4
31318_at	Stem cell f	32.68	19.79	27.45	29.56	28.34	26.55	38.04	41.05	31.91	22.76	23.58

Clustering

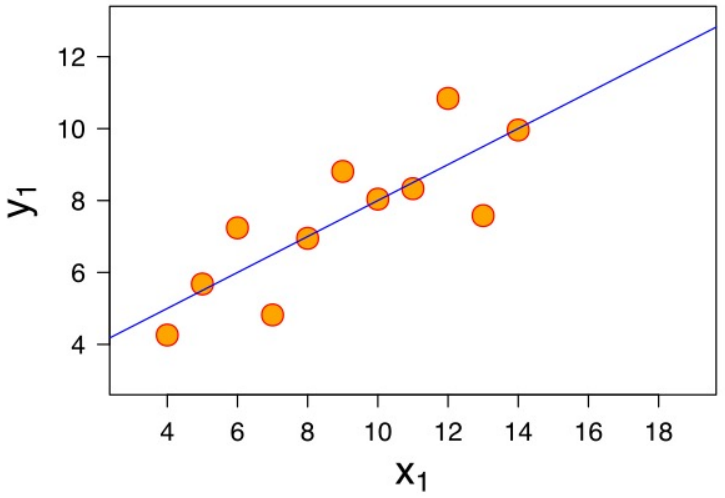
- Motivation for clustering:
  - Visualizing data, e.g. differential expression
  - Understand general characteristics of data
  - Make generalizations about gene behavior
  - Classify samples

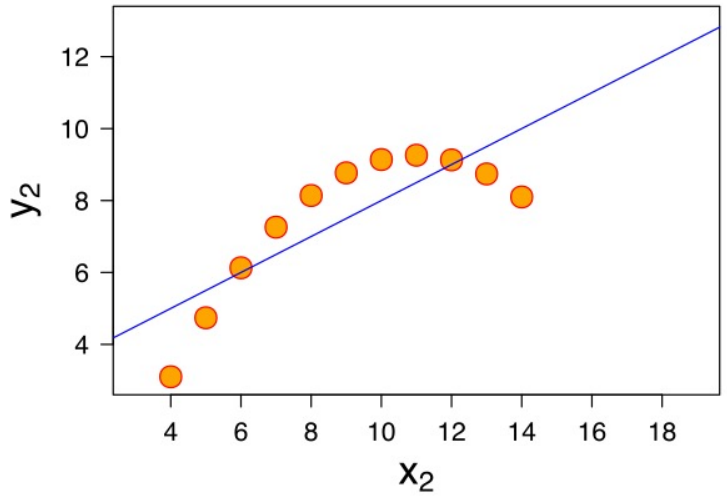
- Goal of clustering:
  - Maximize inter (between)-cluster distance
  - Minimize intra (within)-cluster distance
  - “Distance”: 1 – correlation

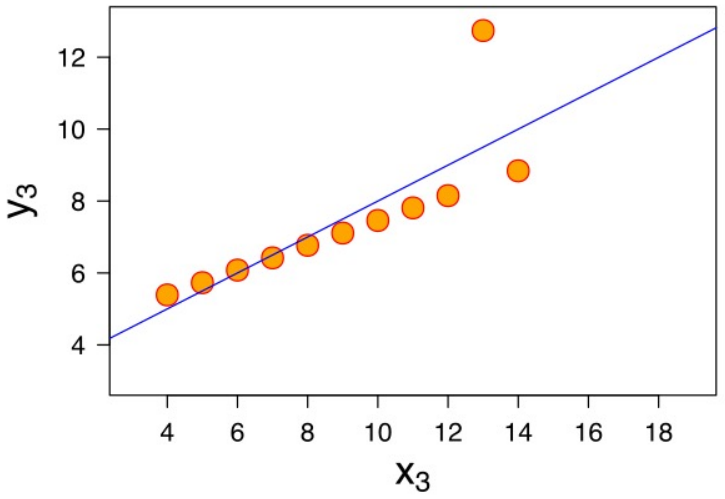


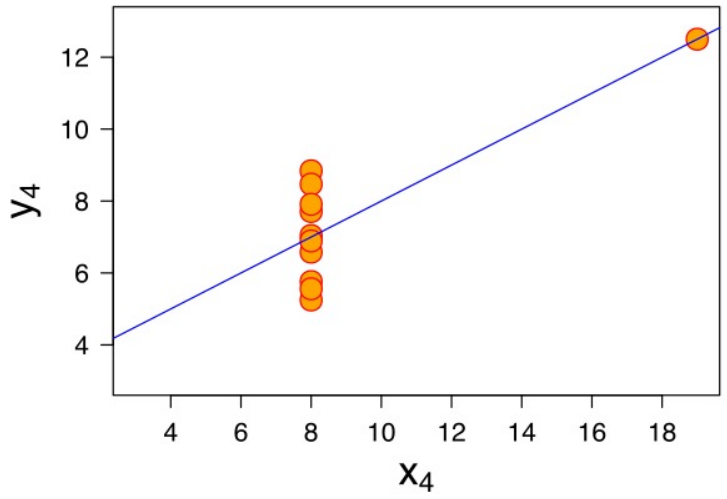
Correlation does not tell everything

Anscombe's Quartet 1973



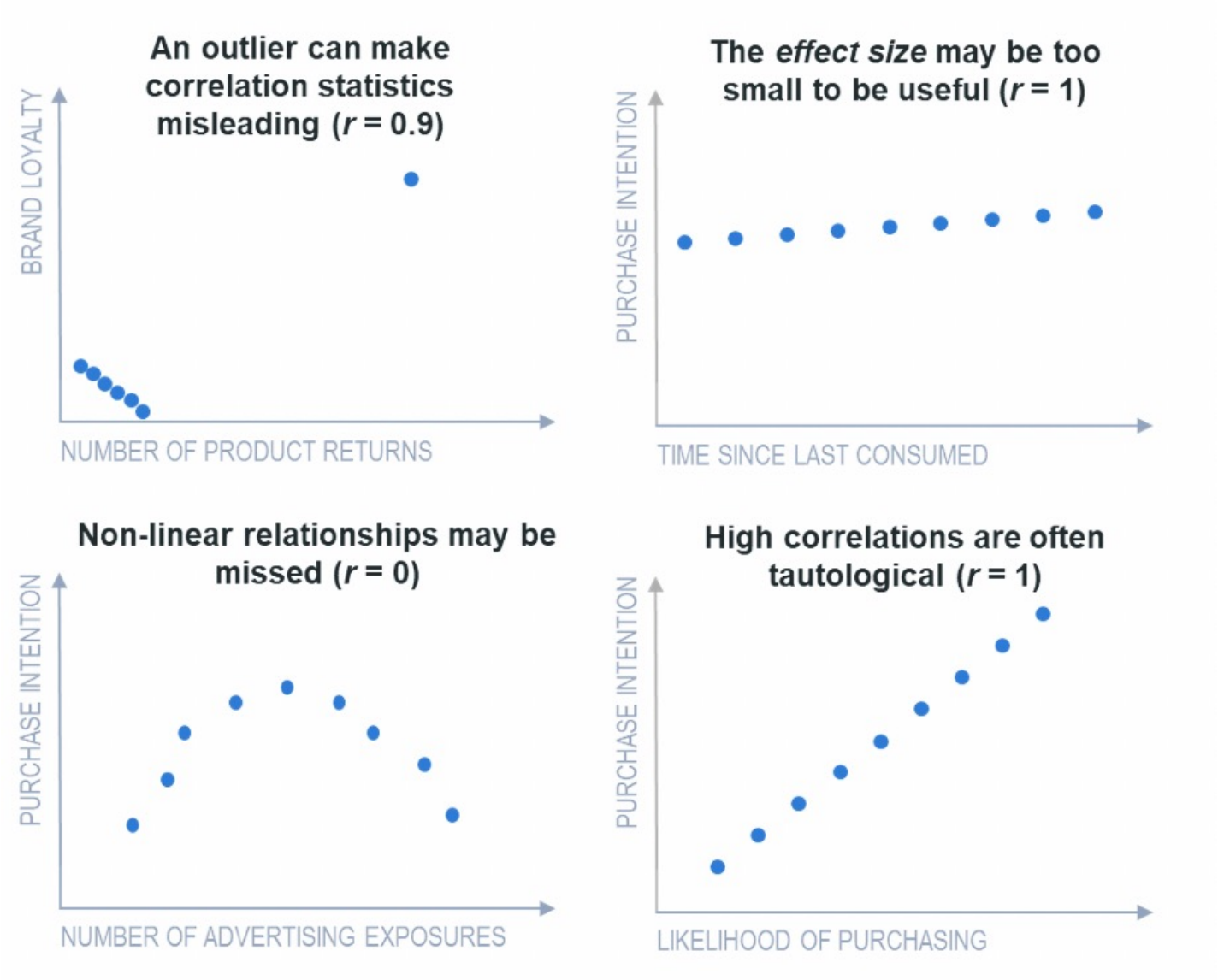




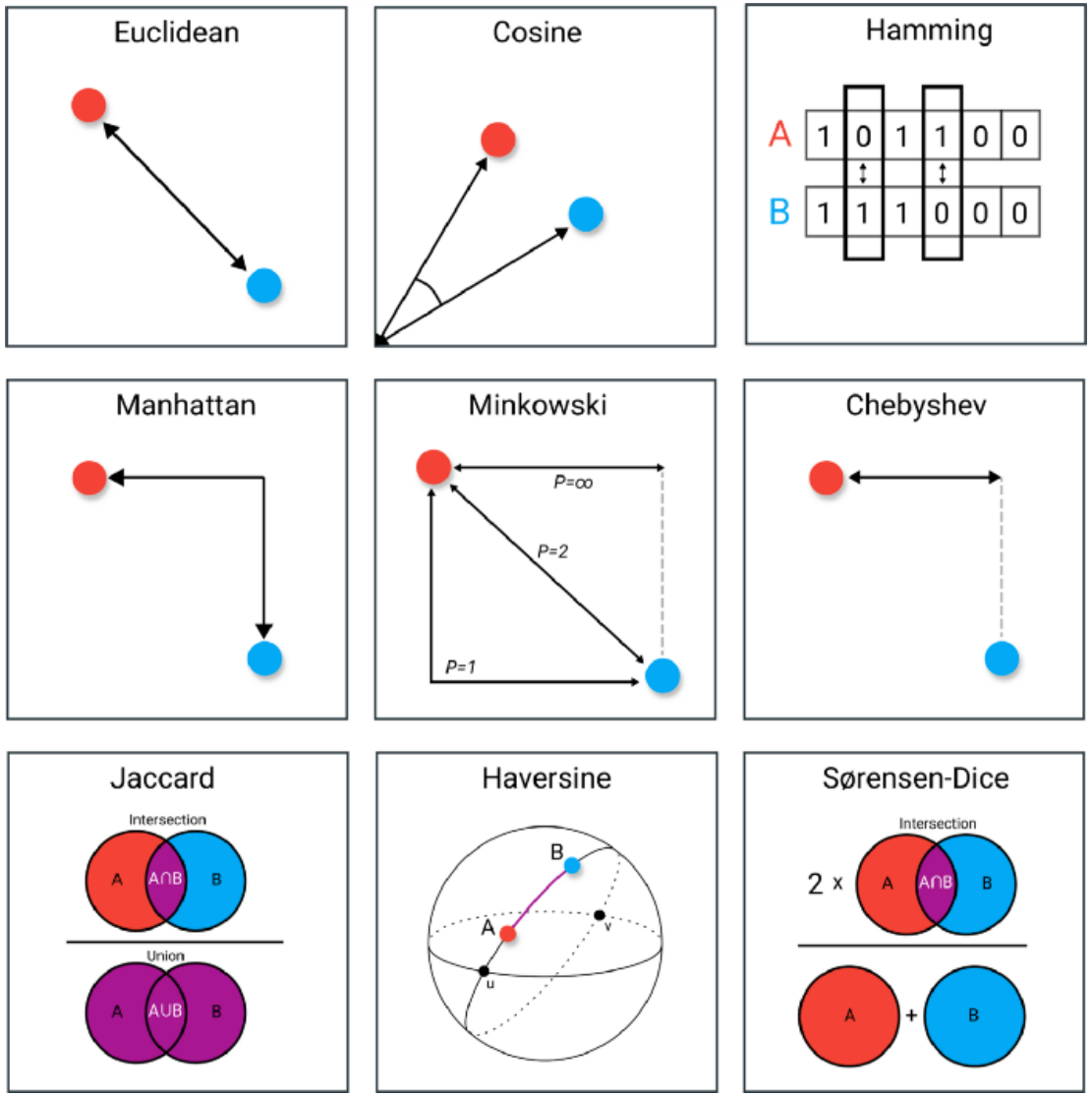


Property	Value
Mean of x	9
Sample variance of x : s_x^2	11
Mean of y	7.50
Sample variance of y : s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

Correlation does not tell everything



Distance Metrics



$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$D(x,y) = cos(\theta) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

$$D(x,y) = \sum_{i=1}^{k} |x_i - y_i|$$

$$D(x,y) = \max_i \left(|x_i - y_i|\right)$$

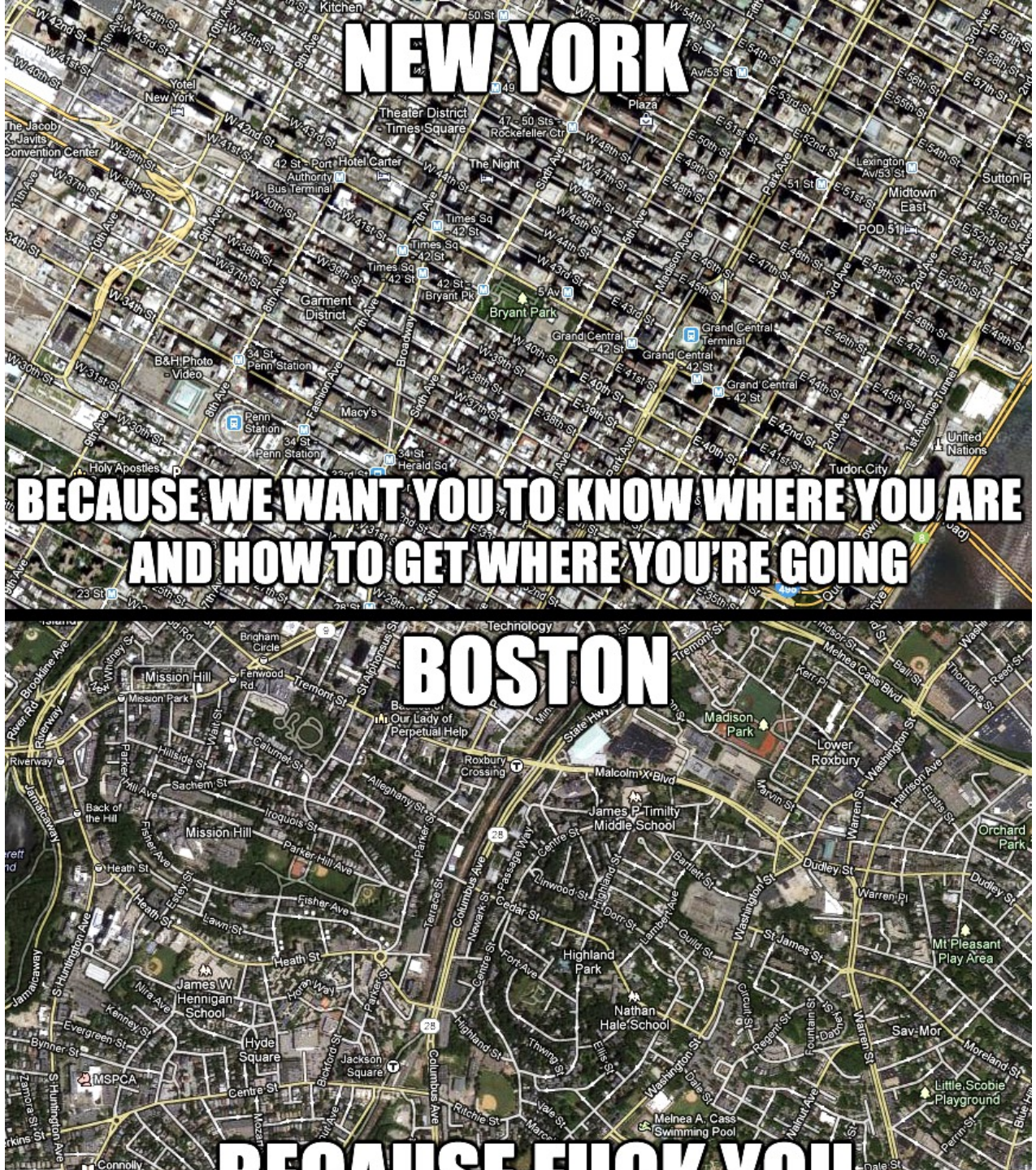
$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

$$D(x,y) = 1 - \frac{|x \cap y|}{|y \cup x|}$$

$$D(x,y) = \frac{2\,|x \cap y|}{|x| + |y|}$$

$$d = 2r\arcsin\left(\sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + cos(\varphi_1)cos(\varphi_2)sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

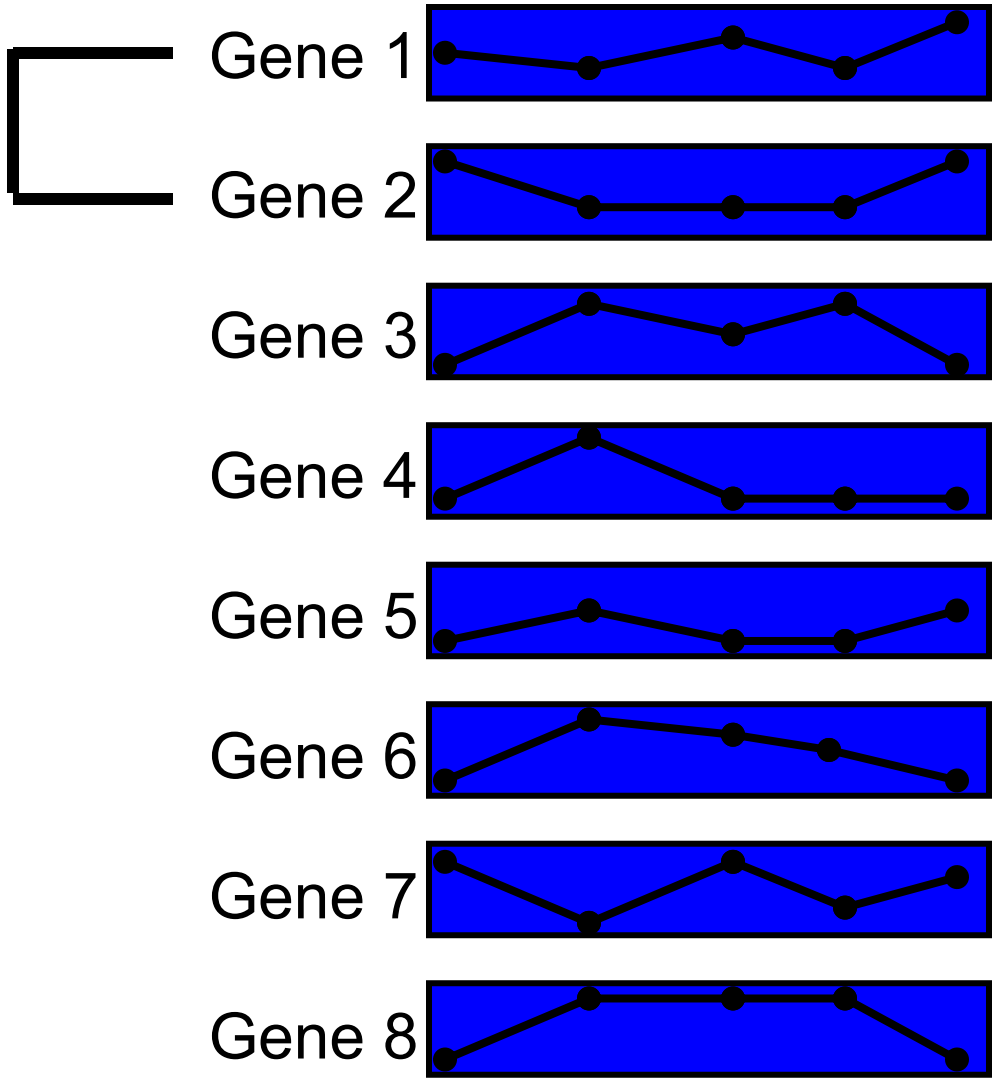

NEW YORK
BECAUSE WE WANT YOU TO KNOW WHERE YOU ARE
AND HOW TO GET WHERE YOU'RE GOING
BOSTON

Cluster Stability

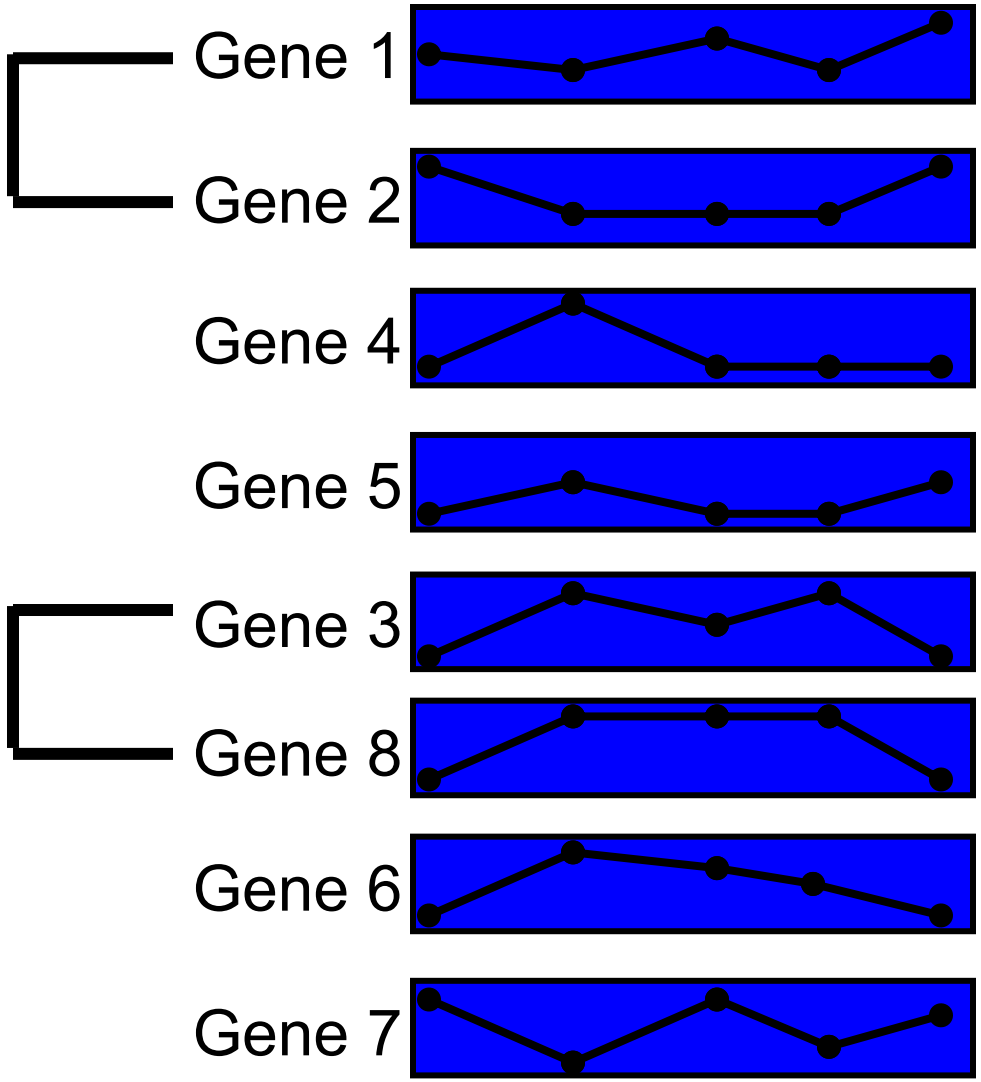
- See whether clustering gives the same result if:
  - Mask out some data (e.g. only sample a subset of genes or samples)
  - Introduce a little noise to the data
  - Change some parameters
- May select subsets of data points for clustering
  - Differentially expressed genes
  - Genes on certain pathways, etc.

Hierarchical Clustering

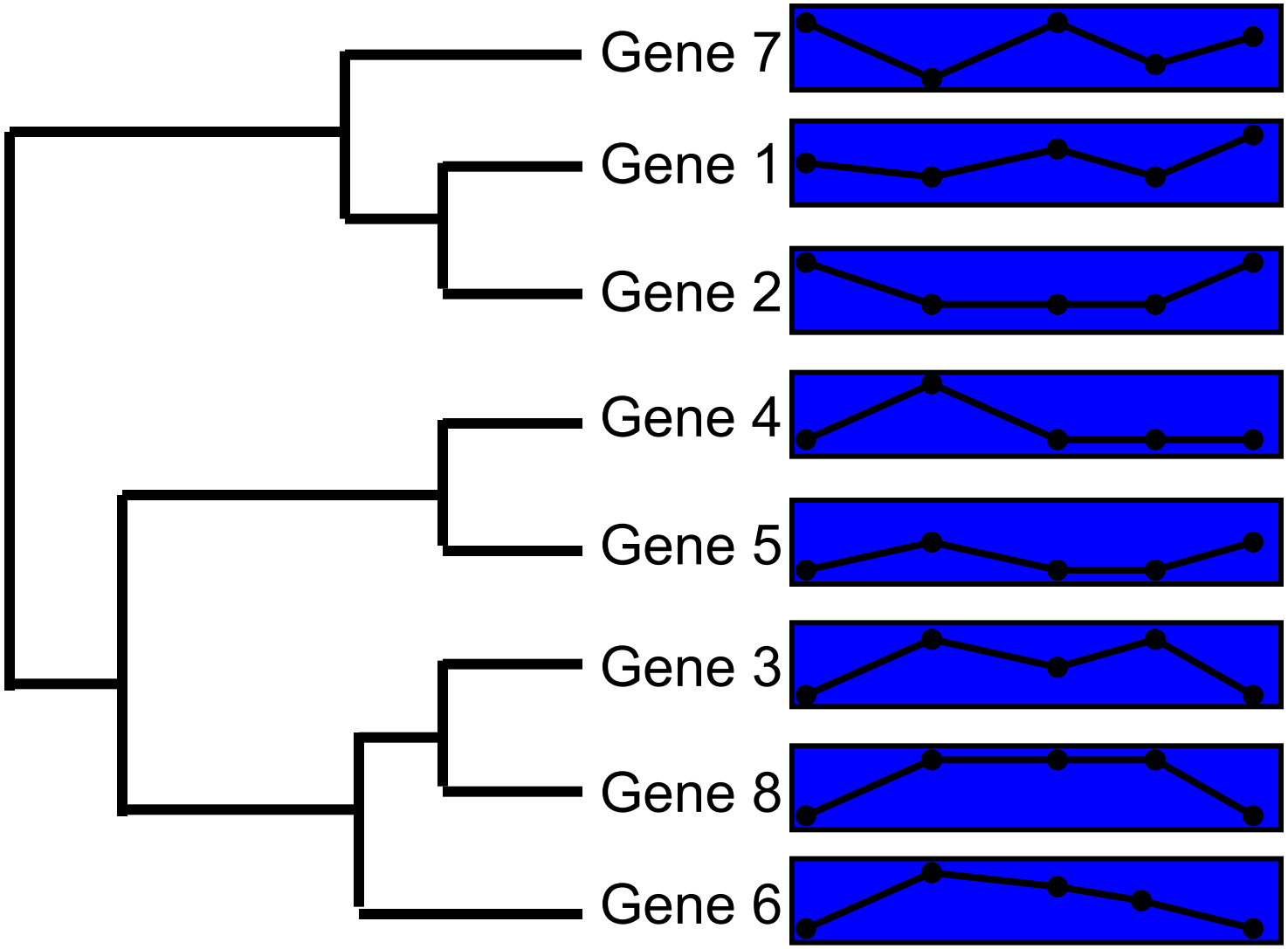
Hierarchical Clustering (Agglomerative)



Hierarchical Clustering (Agglomerative)



Hierarchical Clustering (Agglomerative)

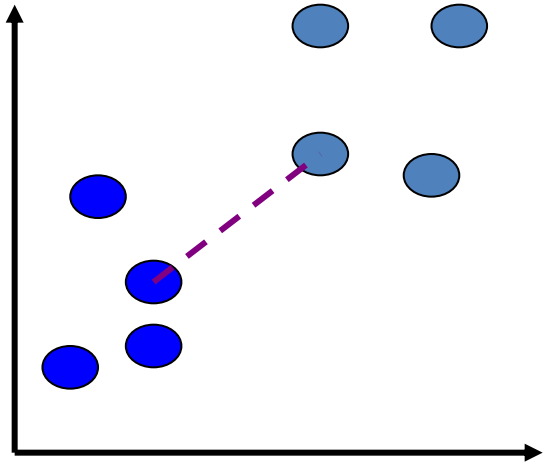


Hierarchical Clustering (Agglomerative)

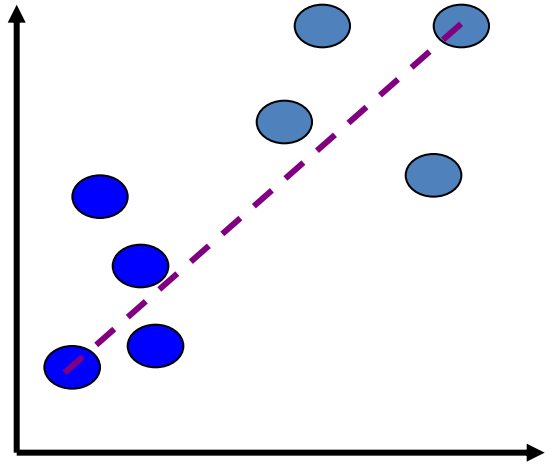
- Repeatedly
  - Merge two nodes (either a gene or a cluster) that are closest to each other
  - Re-calculate the distance from newly formed node to all other nodes
  - Branch length represents distance
- Linkage: distance from newly formed node to all other nodes

Hierarchical Clustering Linkage

Single

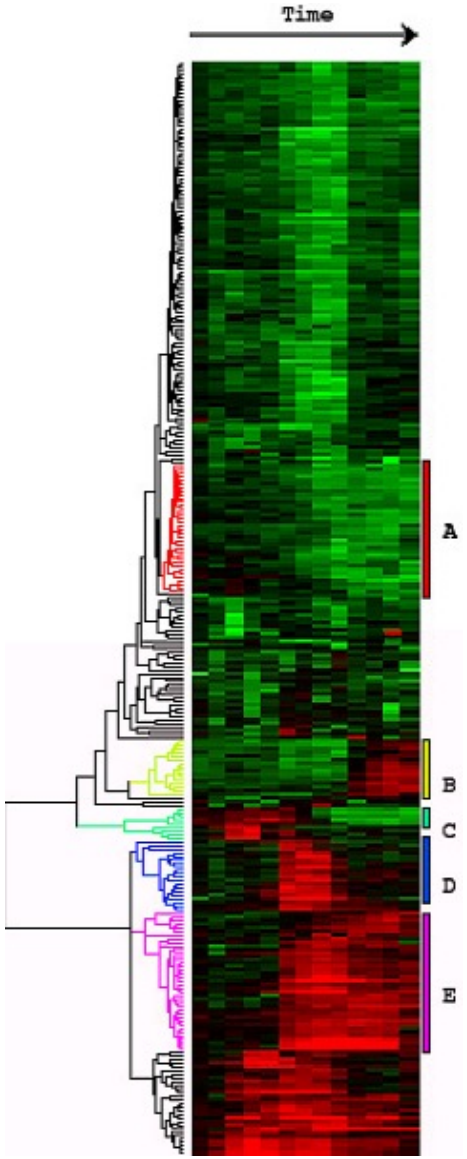


Complete



Average: pairwise distances

$$dist(c_j, c_k) = \frac{|c_u| \times dist(c_u, c_k) + |c_v| \times dist(c_v, c_k)}{|c_u| + |c_v|}$$

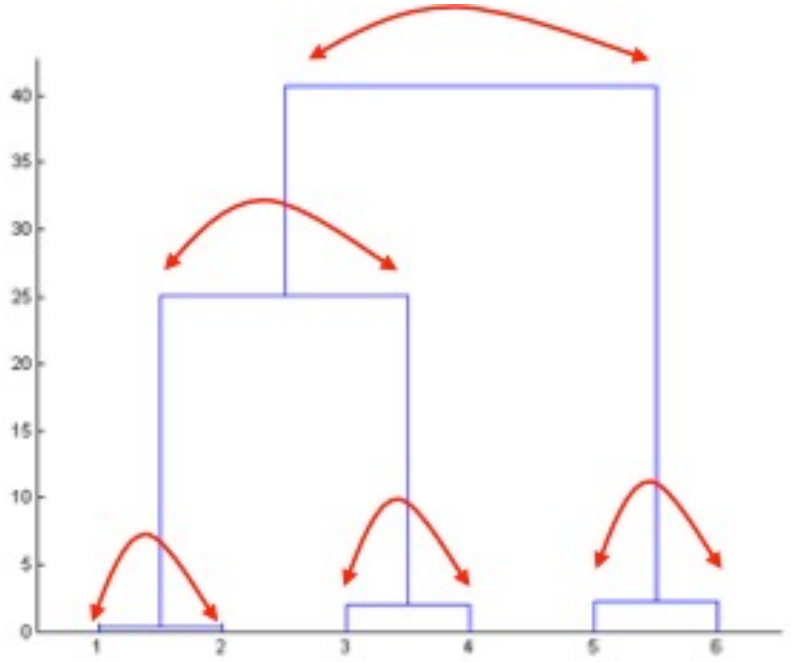


Brain Teasers

- If we have *N* data points

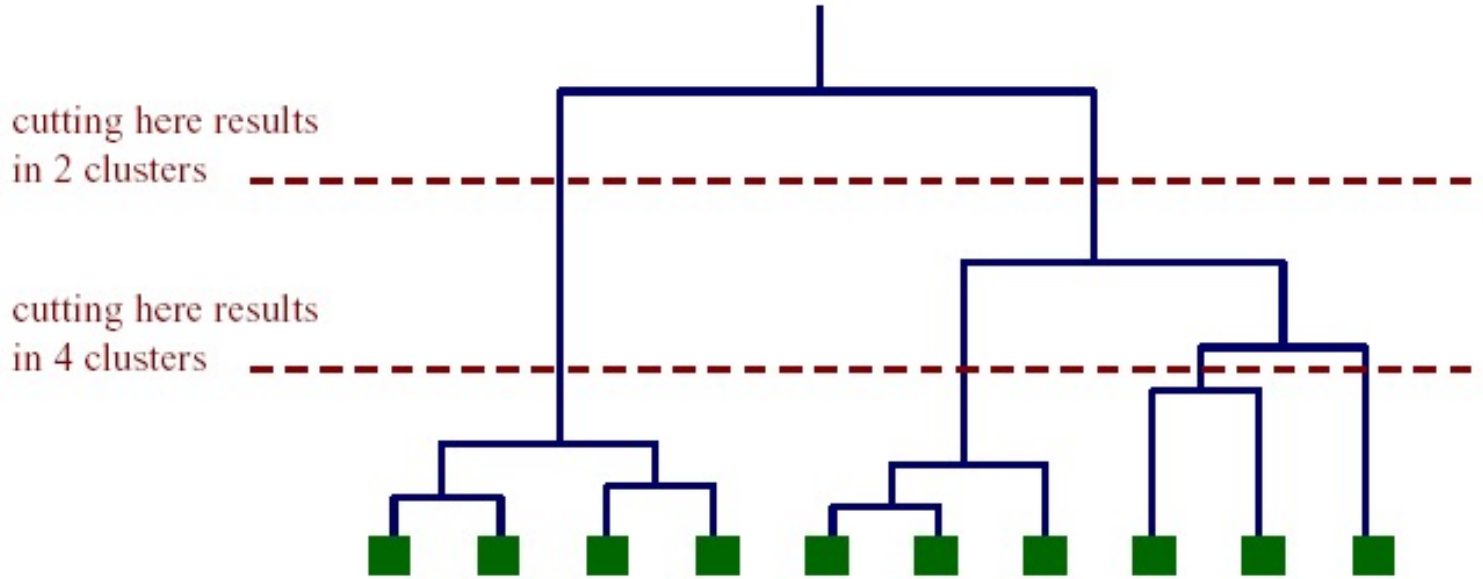
  How many internal nodes are in the H-cluster?

  How many possible ways to draw the H-cluster?



Partitional Clustering

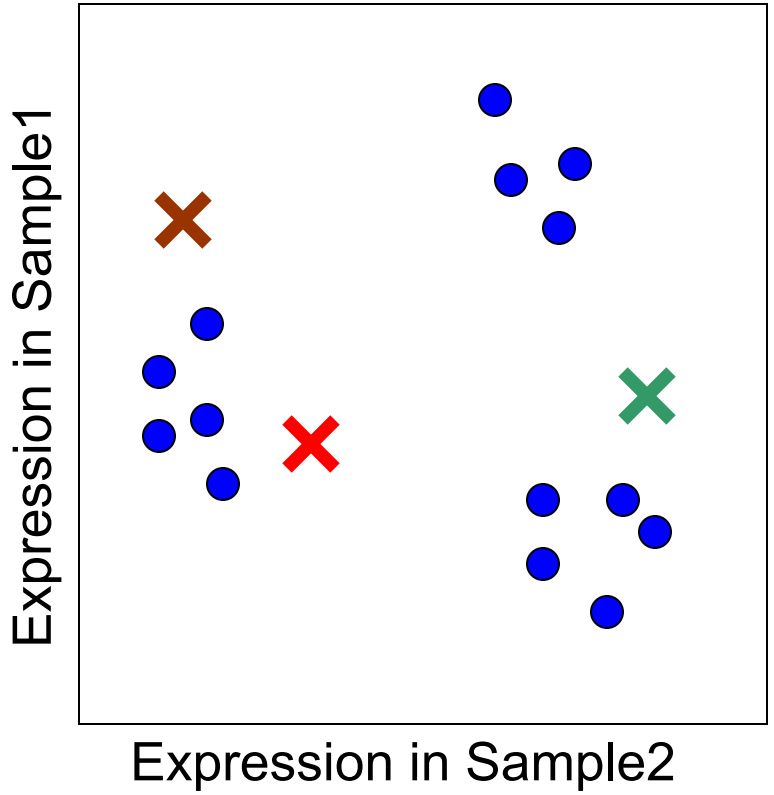
- Disjoint groups
- From hierarchical clustering:
  - Cut a line from hierarchical clustering
  - By varying the cut height, we could produce arbitrary number of clusters



K-means Clustering

K-means Clustering

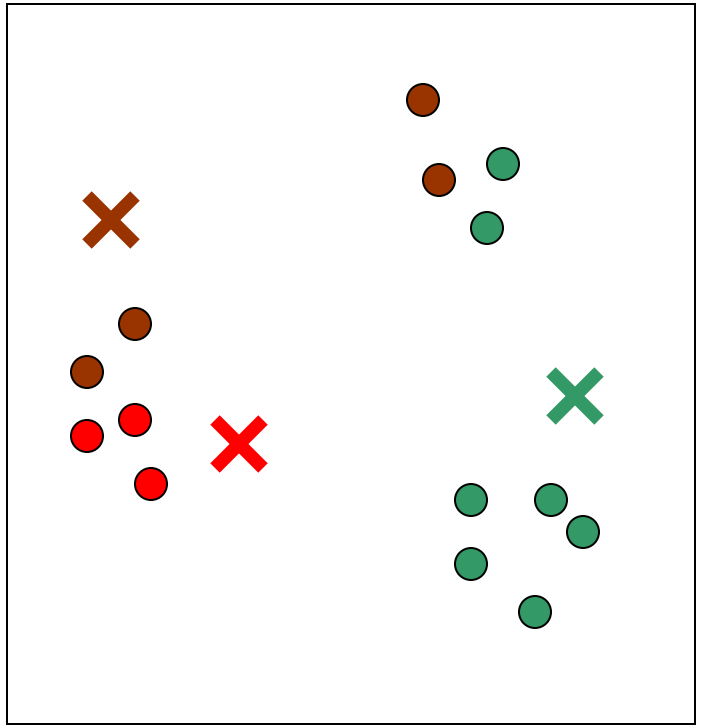
- Choose K centroids at random



Iteration = 0

K-means Clustering

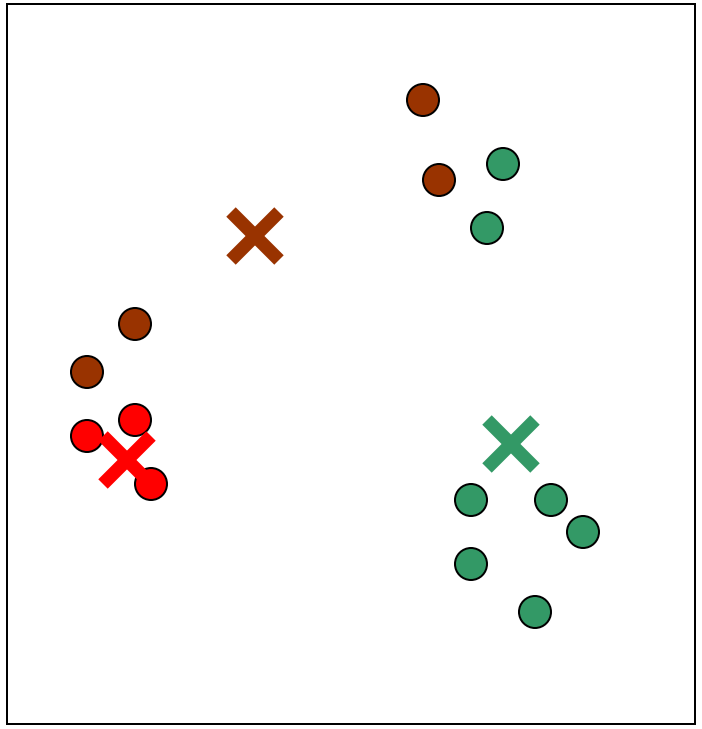
- Choose K centroids at random
- Assign object *i* to closest centroid



Iteration = 1

K-means Clustering

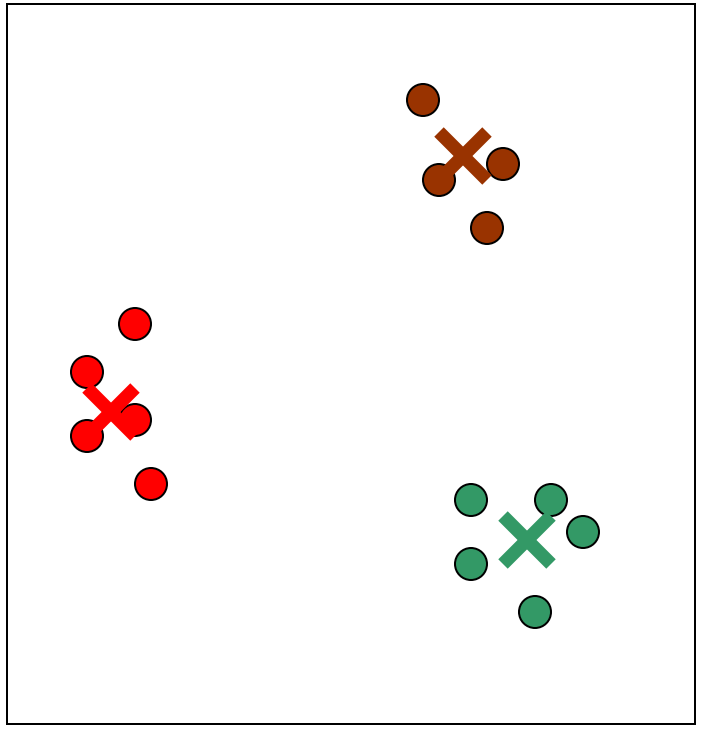
- Choose K centroids at random
- Assign object *i* to closest centroid
- Recalculate centroid based on current cluster assignment



Iteration = 2

K-means Clustering

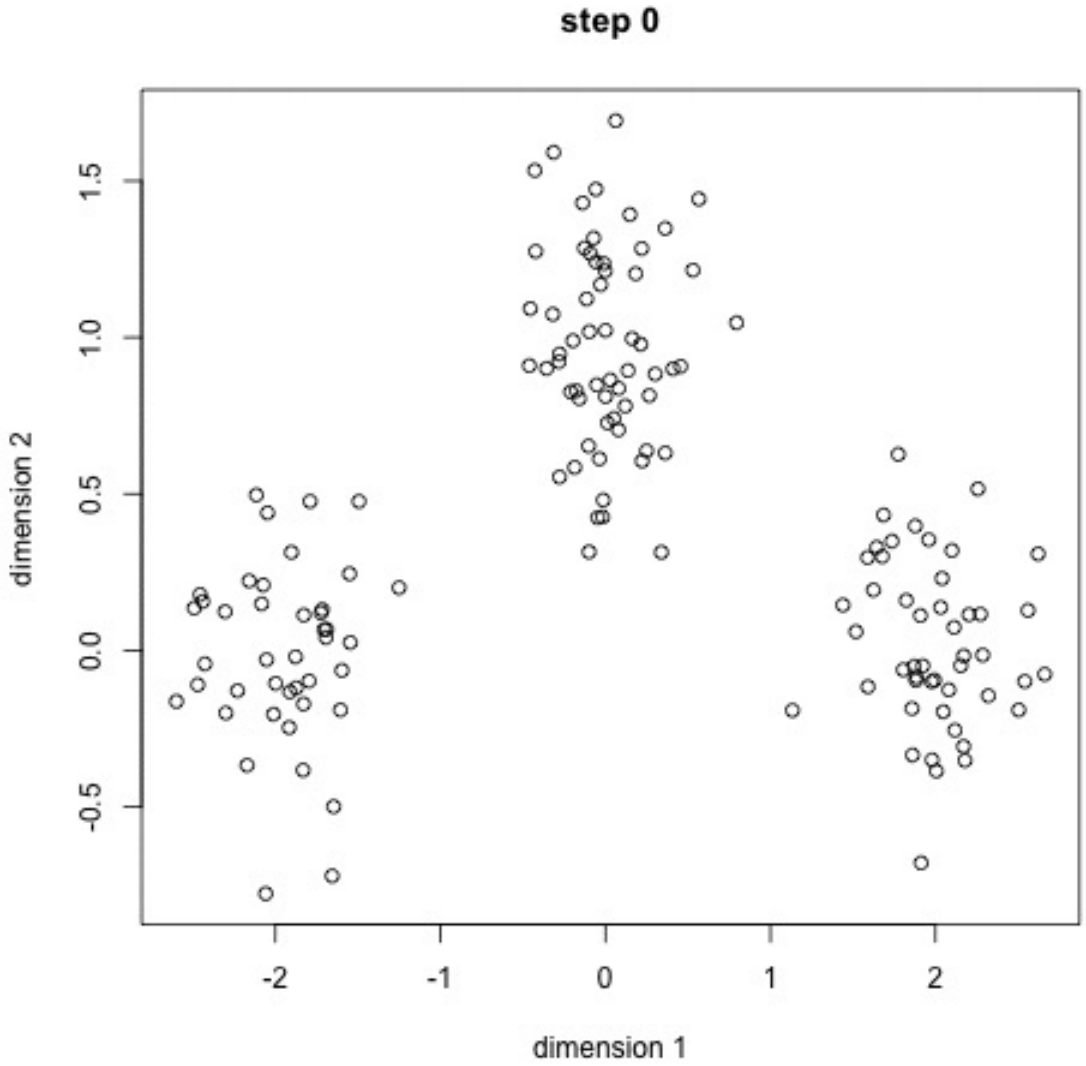
- Choose K centroids at random
- Assign object *i* to closest centroid
- Recalculate centroid based on current cluster assignment
- Repeat until assignment stabilize



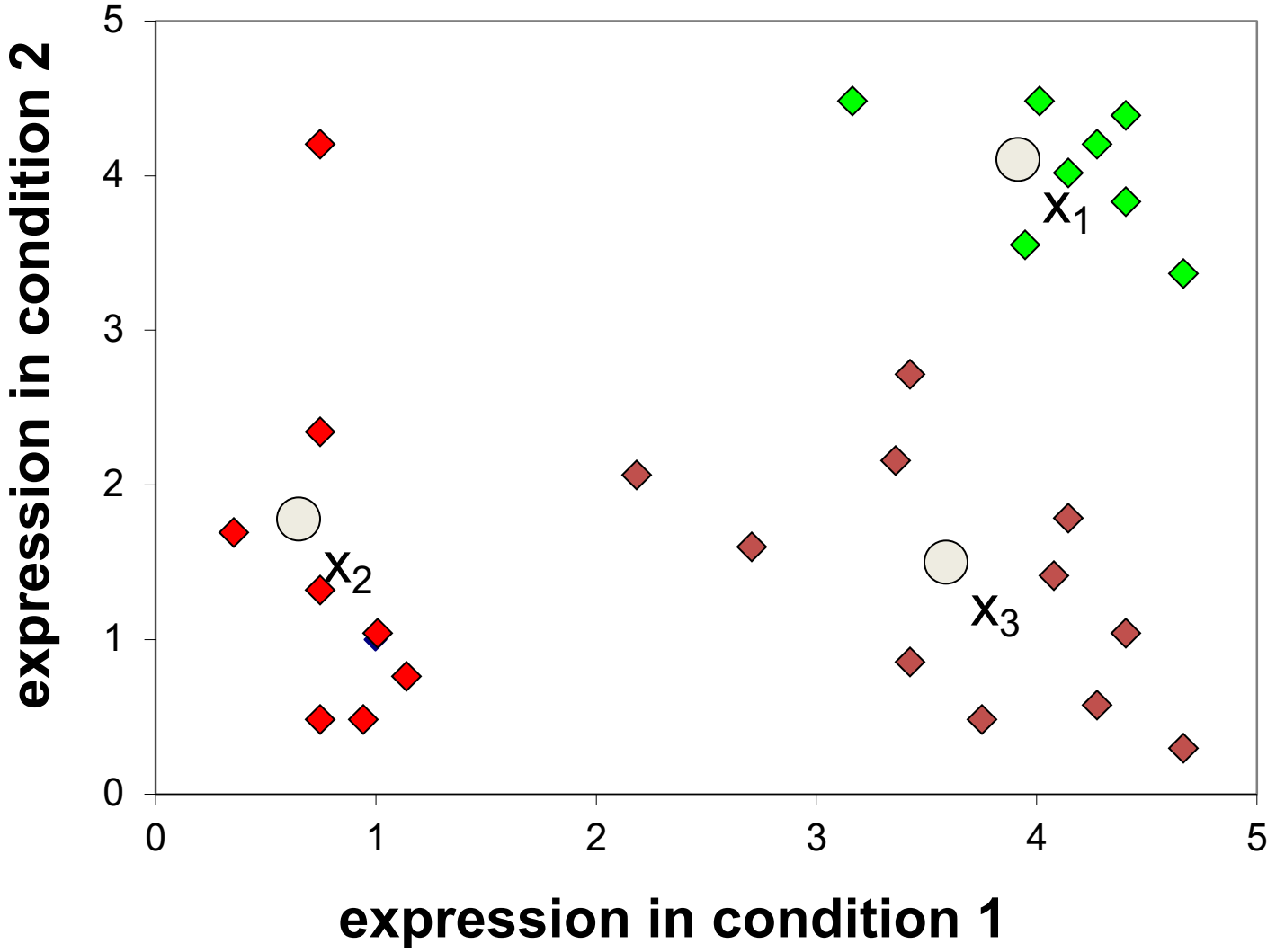
Iteration = 3

K-means Clustering

- Deterministic
  - Initial cluster centers are important
  - Can be trapped in local optimal
- How to pick initial cluster centers
  - Run hierarchical cluster, find cut line
  - Random start many times with different initial centers
- Might not be tolerant to outliers or noise



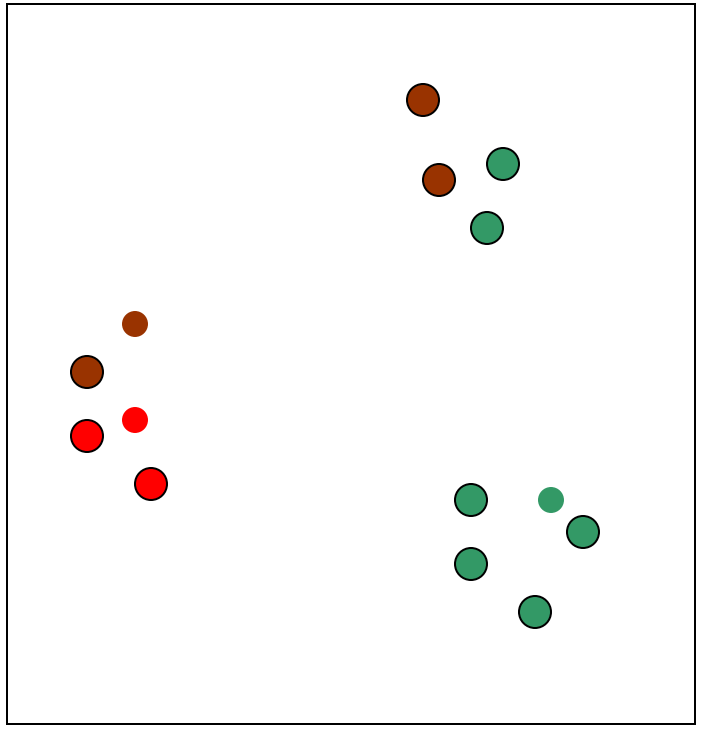
K-means Clustering: Problem With Outliers



Partition Around Medoids

- Pick one real data point (closest to all) instead of average as the centroid of the cluster

- More robust in the presence of noise and outliers



How to Pick K

- K = 2, gradually increase
- Improvement: reduce within-cluster distance and increase between-cluster distance
- Cost: cost with each increase in K
- Compare the cost with improvement, stop when not worth it

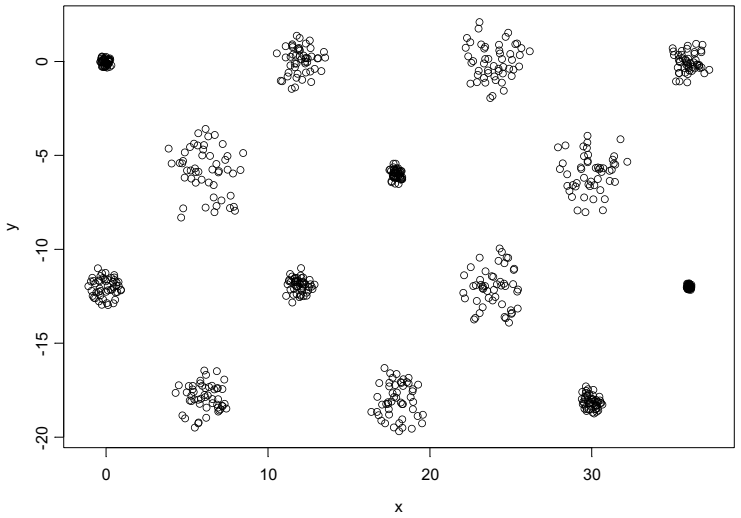
- *W(k)* = total sum of squares within clusters
- *B(k)* = sum of squares between cluster means
- *n* = total number of data points
- Calinski & Harabasz, 1974

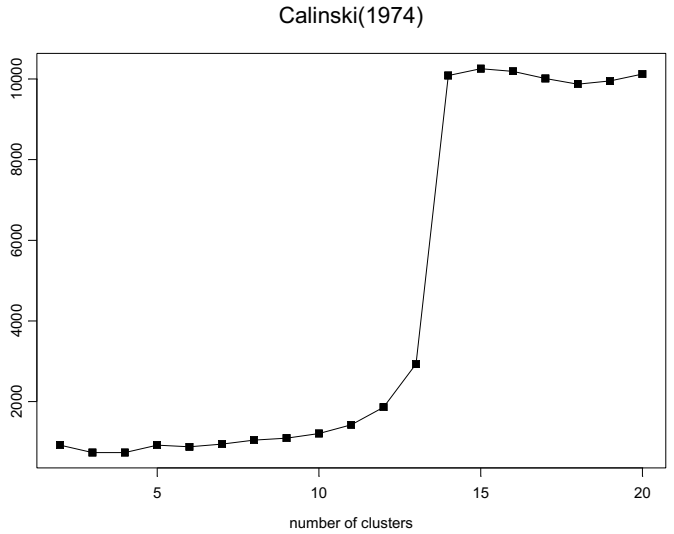
$$\max CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

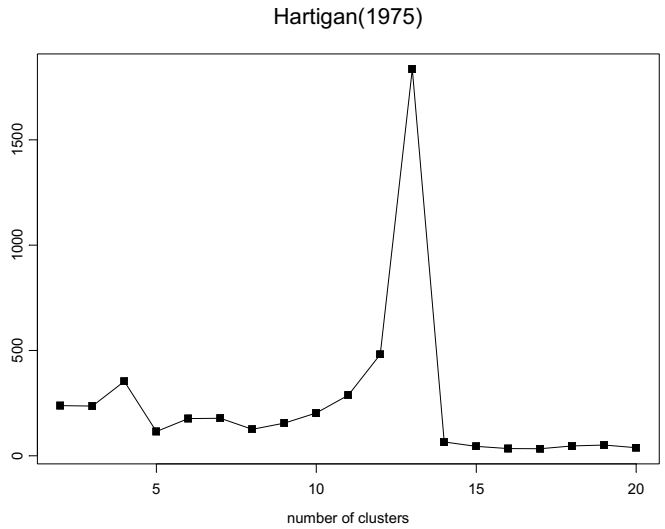
- Hartigan, 1975: stop when *H(k)* < 10

$$H(K) = (\frac{W(k)}{W(k+1)} - 1)(n-k-1)$$

How to Pick K







- In practice for genomics data:
  - Only cluster genes that are variable across samples
  - The magic number: **7**

Consensus Clustering

- Cluster ensembles

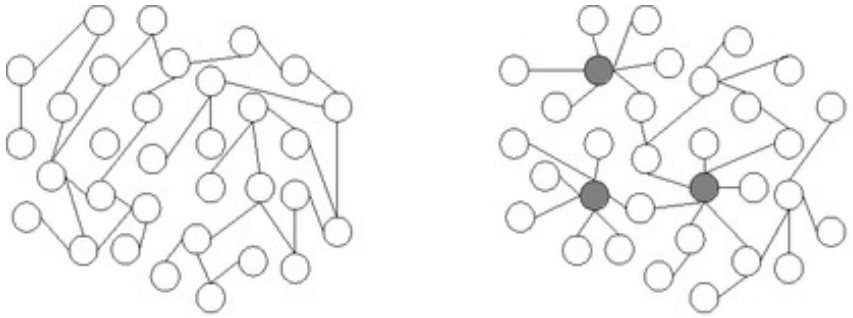
- Reconcile clustering information about the same dataset coming from different sources or from different runs of the same algorithm:
  - Tight Clustering (Tseng and Wong, Biometrics 2005)

- Reconcile clustering information about the same samples using different profiling techniques (data types):
  - iCluster (Shen et al. Bioinformatics 2009)

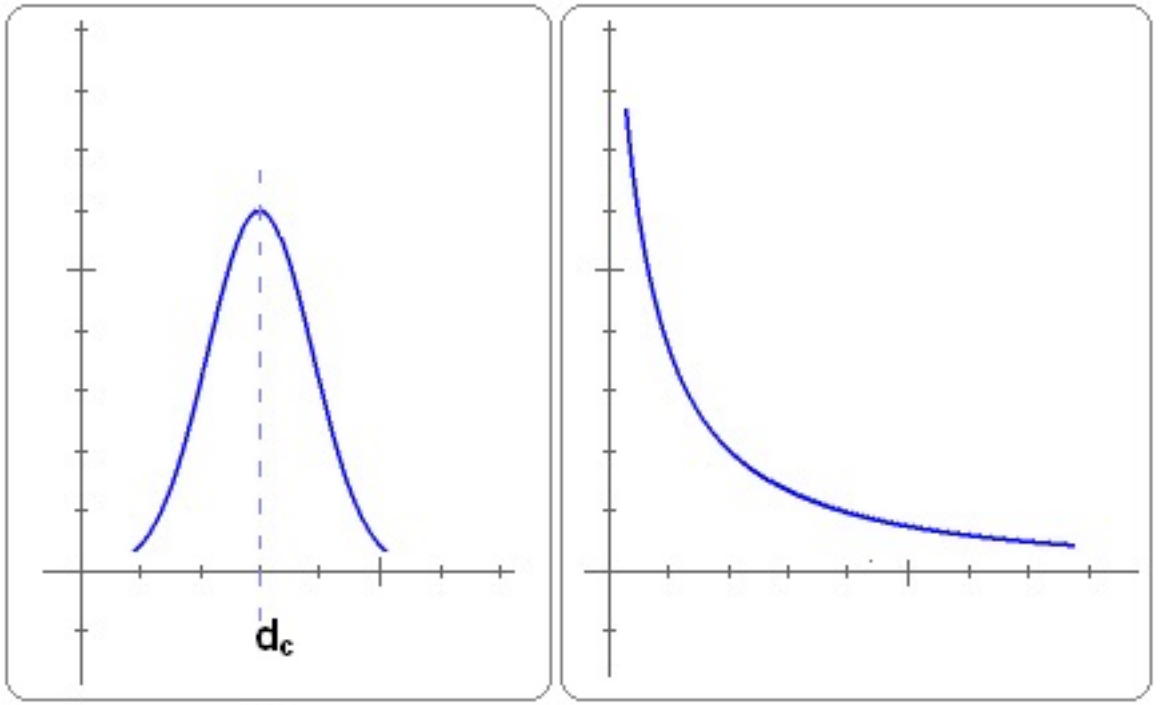
Louvain Method

Networks

- Network: nodes/vertices and edges
- Degree of a node
- Degree distribution of a network
  - Complete network
  - Random network
  - Scale free network: social network and most networks in nature



(a) Random network (b) Scale-free network



Louvain Method

- Network based
- Modularity:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

- A is the adjacency matrix
- m is the number of edges in the network
- k_i is the degree of vertex i
- c_i is the community of vertex i

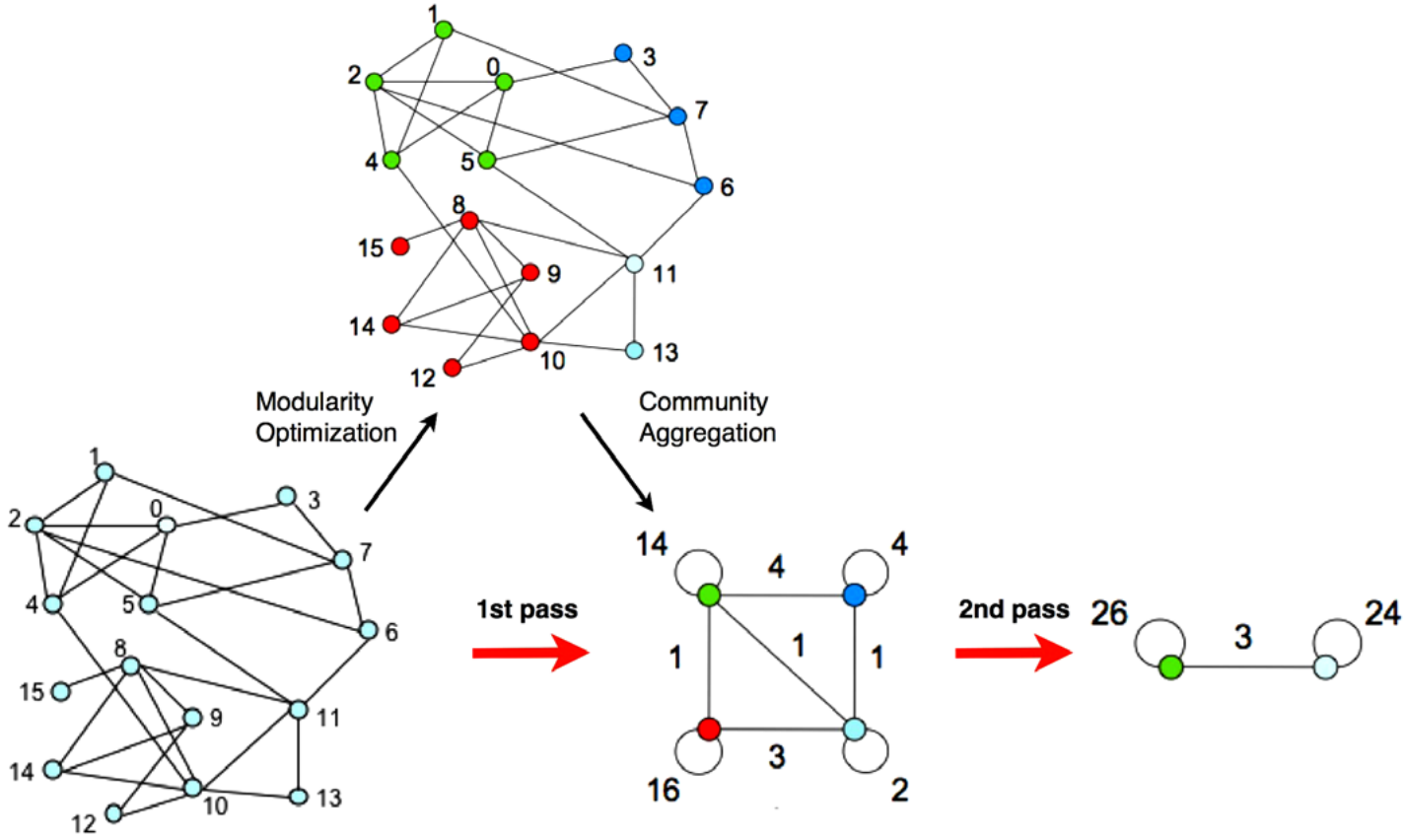
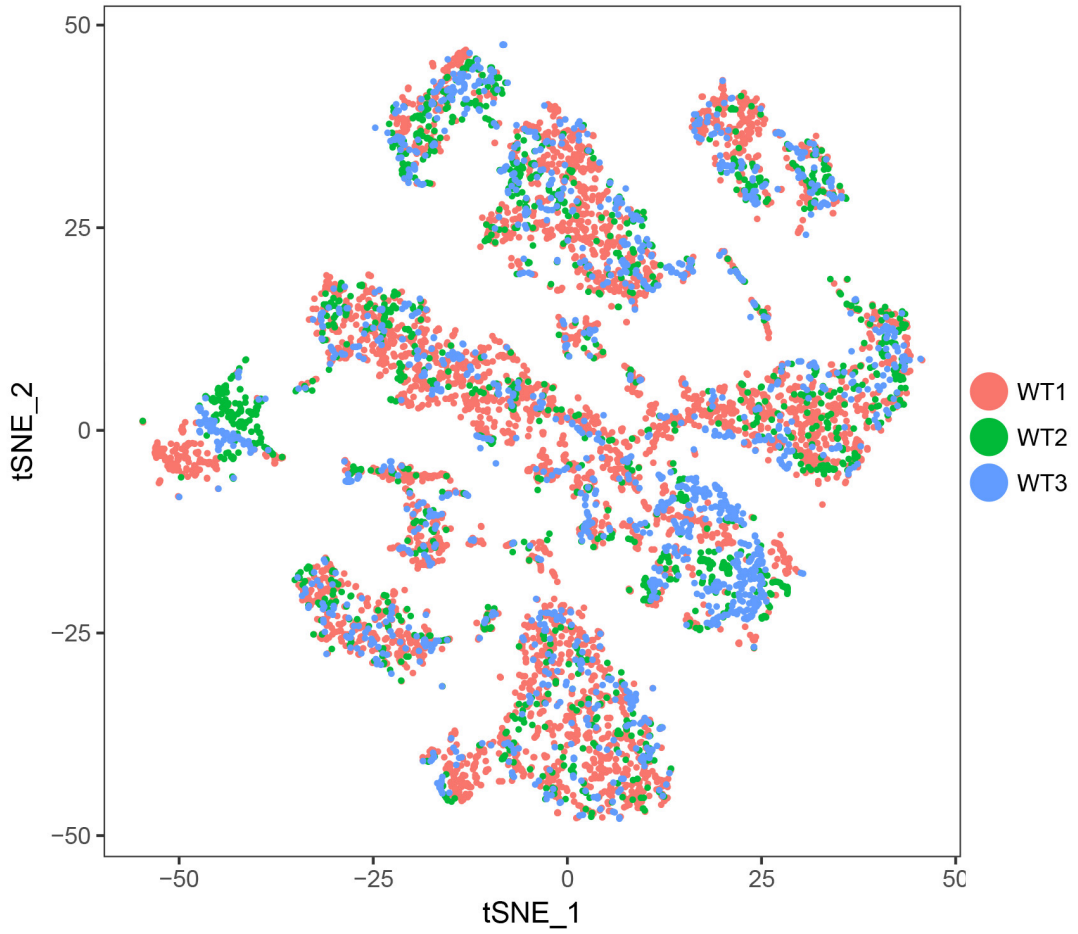
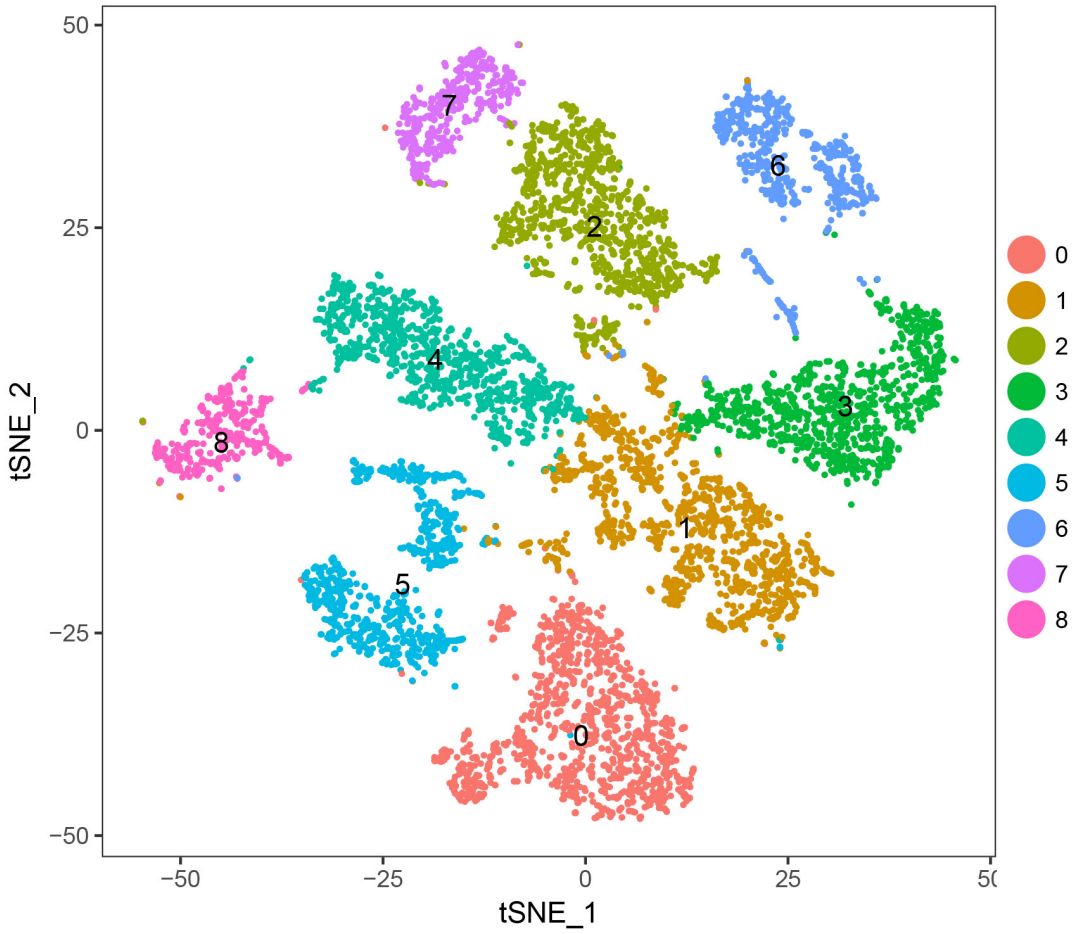


Figure 1. Visualization of the steps of our algorithm. Each pass is made of two phases: one where modularity is optimized by allowing only local changes of communities; one where the communities found are aggregated in order to build a new network of communities. The passes are repeated iteratively until no increase of modularity is possible.

Clustering vs. Visualization





Functional Analysis of Transcriptomics Profiling Data

Gene Annotation

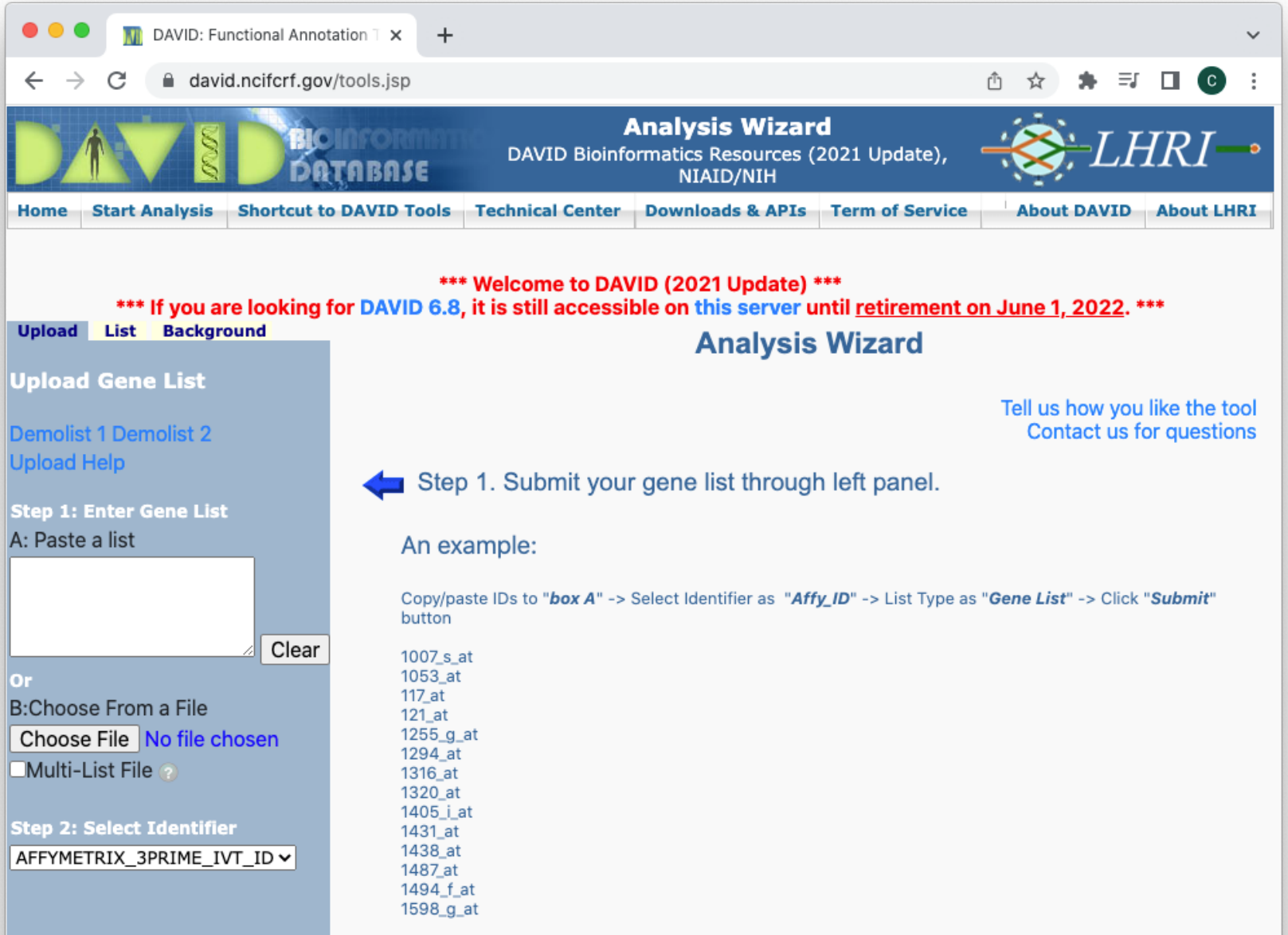
- How to report differentially expressed genes or gene clusters?
  - Enriched for certain pathways, certain functions, or proteins localized in the same complex, etc.?
- Gene Ontology Consortium
  - Ashburner et al. 1998
  - Annotate gene function in the human genome
  - Now extended to many model organisms
- Why do we care?
  - Effectively communicate biomedical knowledge
  - Organize and summarize annotations in a structured way
  - Allow effective and meaningful computation on gene annotations

GO Categories

- Molecular function
  - Describe a gene's jobs or abilities
  - e.g., transporters, transcription factor
- Biological process
  - Events or pathways
  - e.g., cell differentiation, maturation, development
- Cellular component
  - Describe locations (subcellular structures, macromolecular complexes)
  - e.g., nucleus, cell membrane, protein complexes

GO Tools

- DAVID



Gene Set Enrichment Analysis (GSEA)

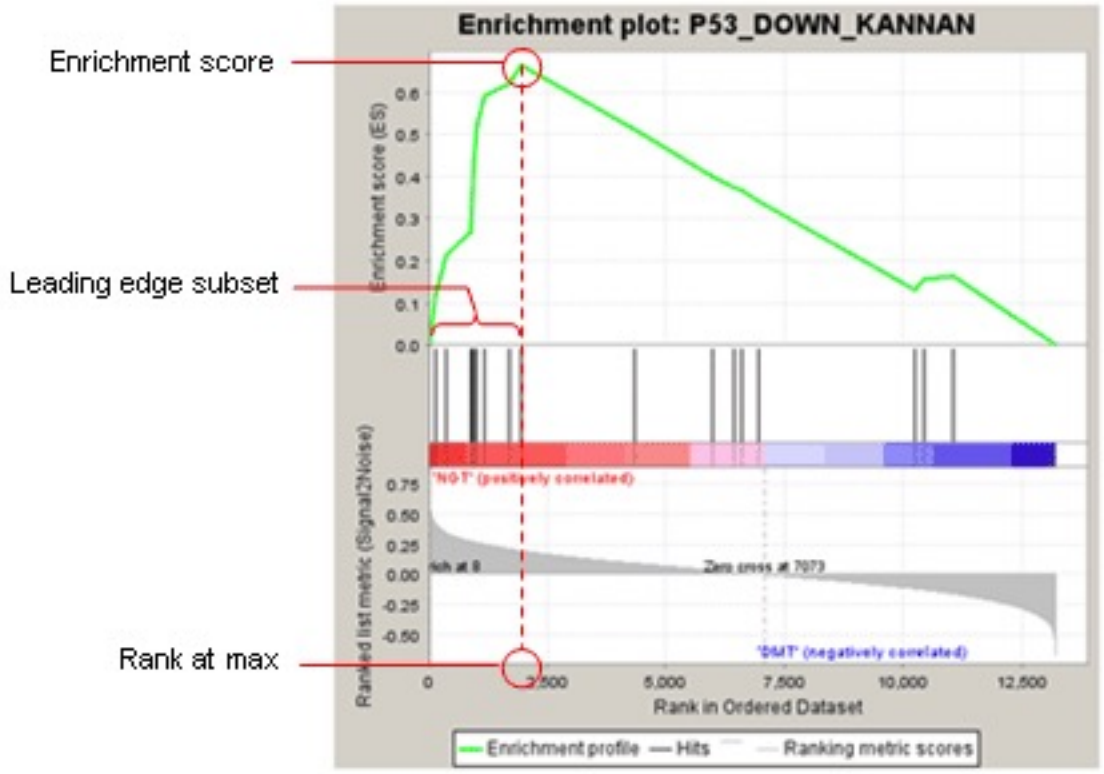
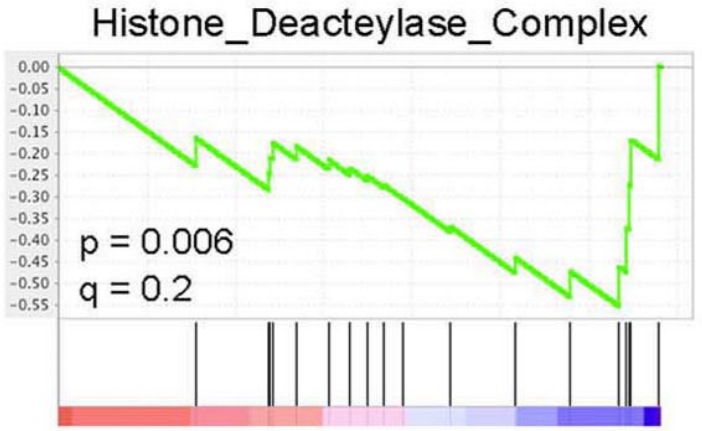
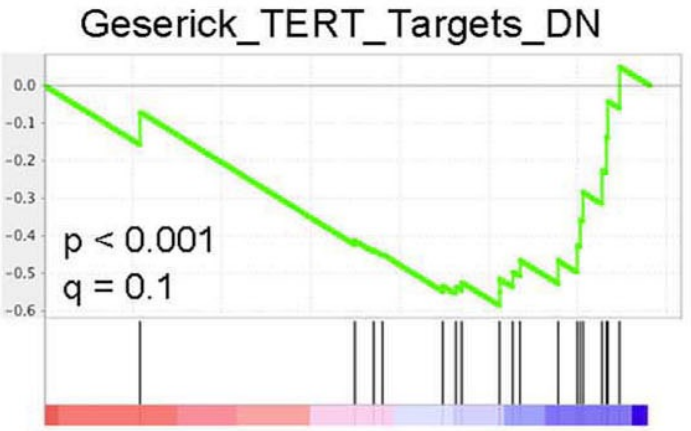
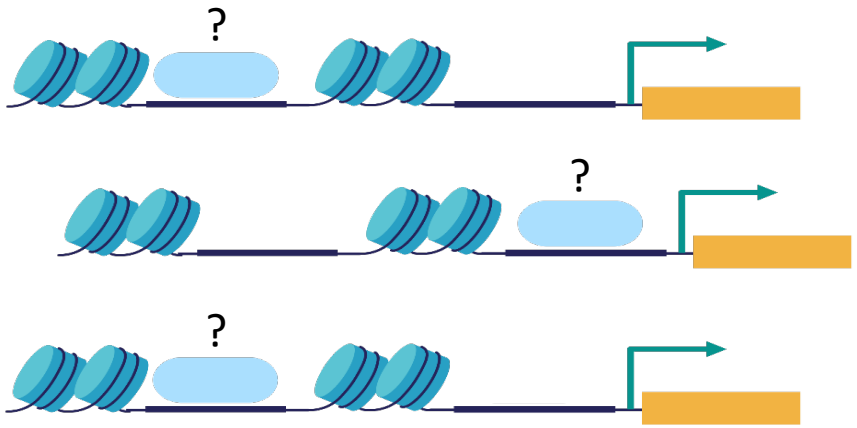


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

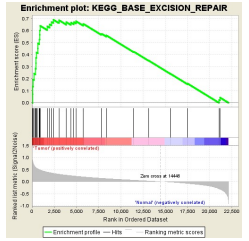




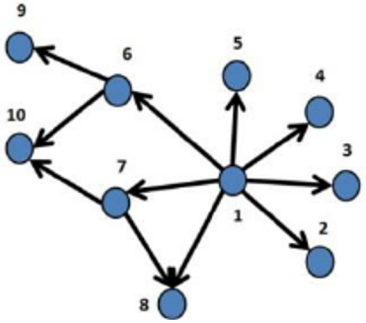
How to identify functional TFs?



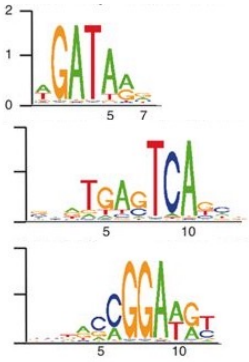
- Ontology based
  - Limited by existing database

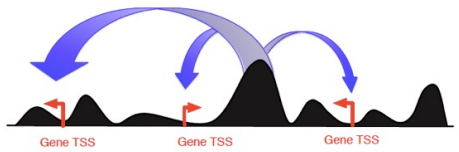


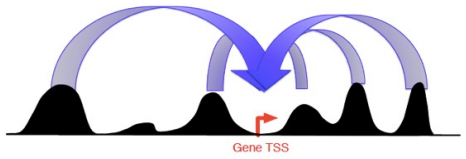
- Co-expression based
  - Expression of a TF ≠ Regulatory activity of the TF



- DNA sequence motif based
  - Motif occurrence ≠ TF binding
  - Difficult to tackle distal enhancers



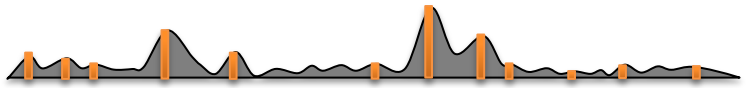




BART: Binding Analysis for Regulation of Transcription

gene set

Adaptive Lasso regression
+ semi-supervised learning



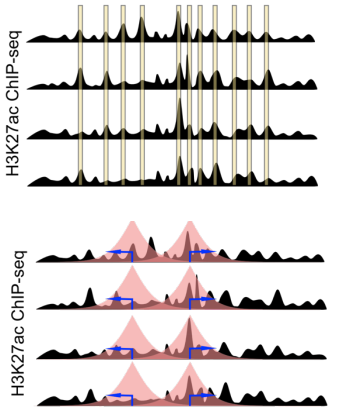
cis-regulatory (enhancer) profile

ROC association analysis,
background adjustment,
rank integration

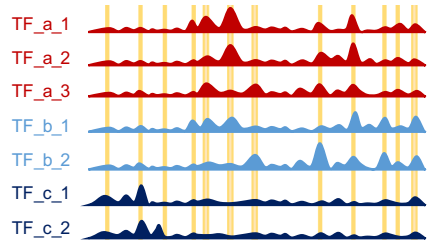
transcription factors

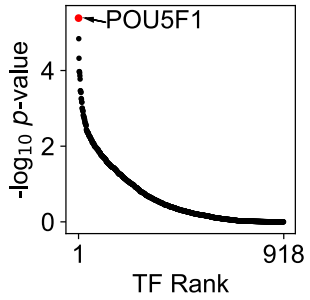
> 500 DNase-seq

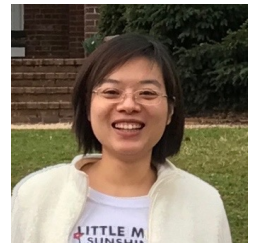
> 1000 H3K27ac ChIP-seq



> 13,000 TF ChIP-seq



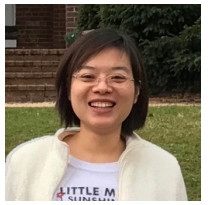




Zhenjia Wang

Wang *et al. Bioinformatics* 2018

BART web: infer transcriptional regulators from various inputs


Zhenjia Wang


Wenjing Ma

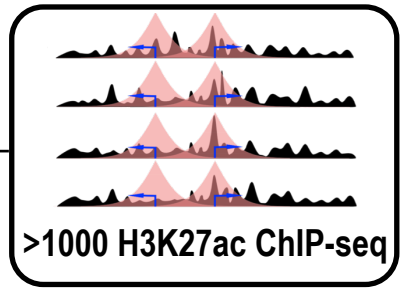
BART: Binding Analysis for Regulation of Transcription

User input

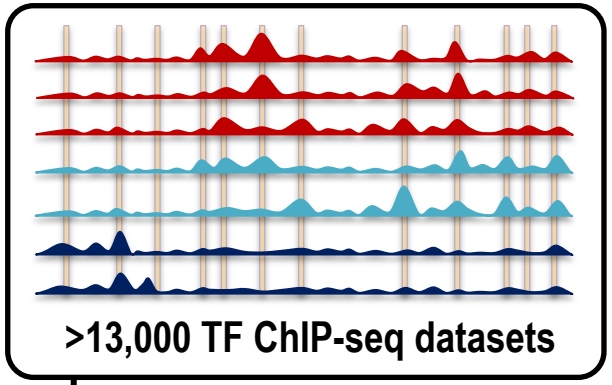
- **Gene list**
- ChIP-seq
- Region set
- Hi-C maps

Mapping

differential interaction



Adaptive Lasso regression

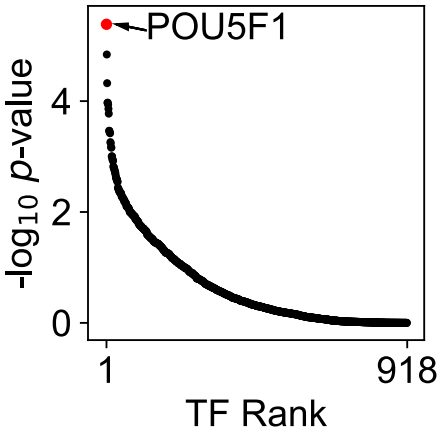


Cis-regulatory element repertoire
(2.7 million in the human genome,
1.5 million in the mouse genome)

ROC associations

Statistical tests,
Background adjustment,
Irwin-Hall rank integration

Output prediction



http://bartweb.org

Wang *et al., Bioinformatics* 2018
Wang *et al.*, *Bioinformatics* 2021
Ma, Wang *et al., NAR Genomics & Bioinformatics* 2021

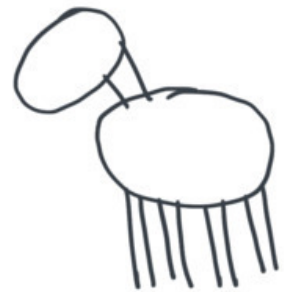
Summary

- Clustering
  - Hierarchical clustering (e.g., WGCNA)
  - K-means clustering
  - Louvain method (e.g., scRNA-seq)
- Regulatory Networks
  - Gene Ontology
  - GSEA
  - BART

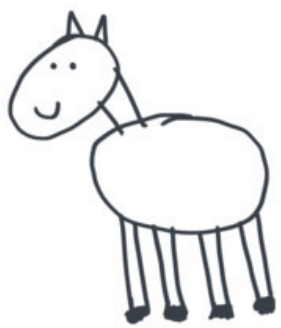
HOW TO:
DRAW A HORSE

BY VAN OKTOP

(1) DRAW 2 CIRCLES

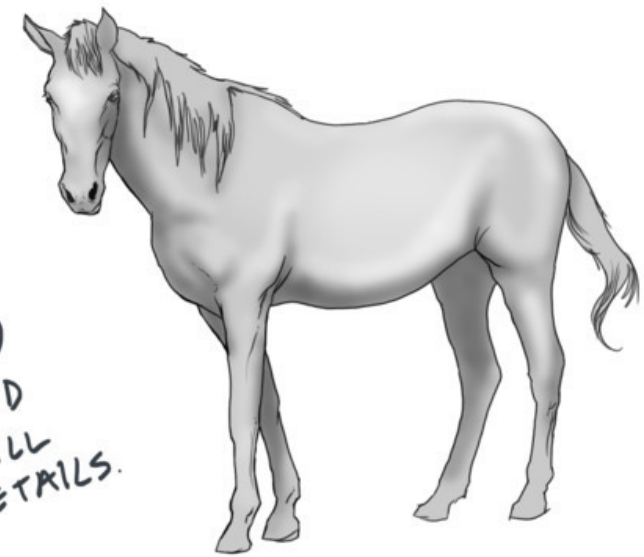


(2) DRAW THE LEGS



(3) DRAW THE FACE

(4) DRAW THE HAIR



(5) ADD SMALL DETAILS.