CMSC 423:
Data Clustering

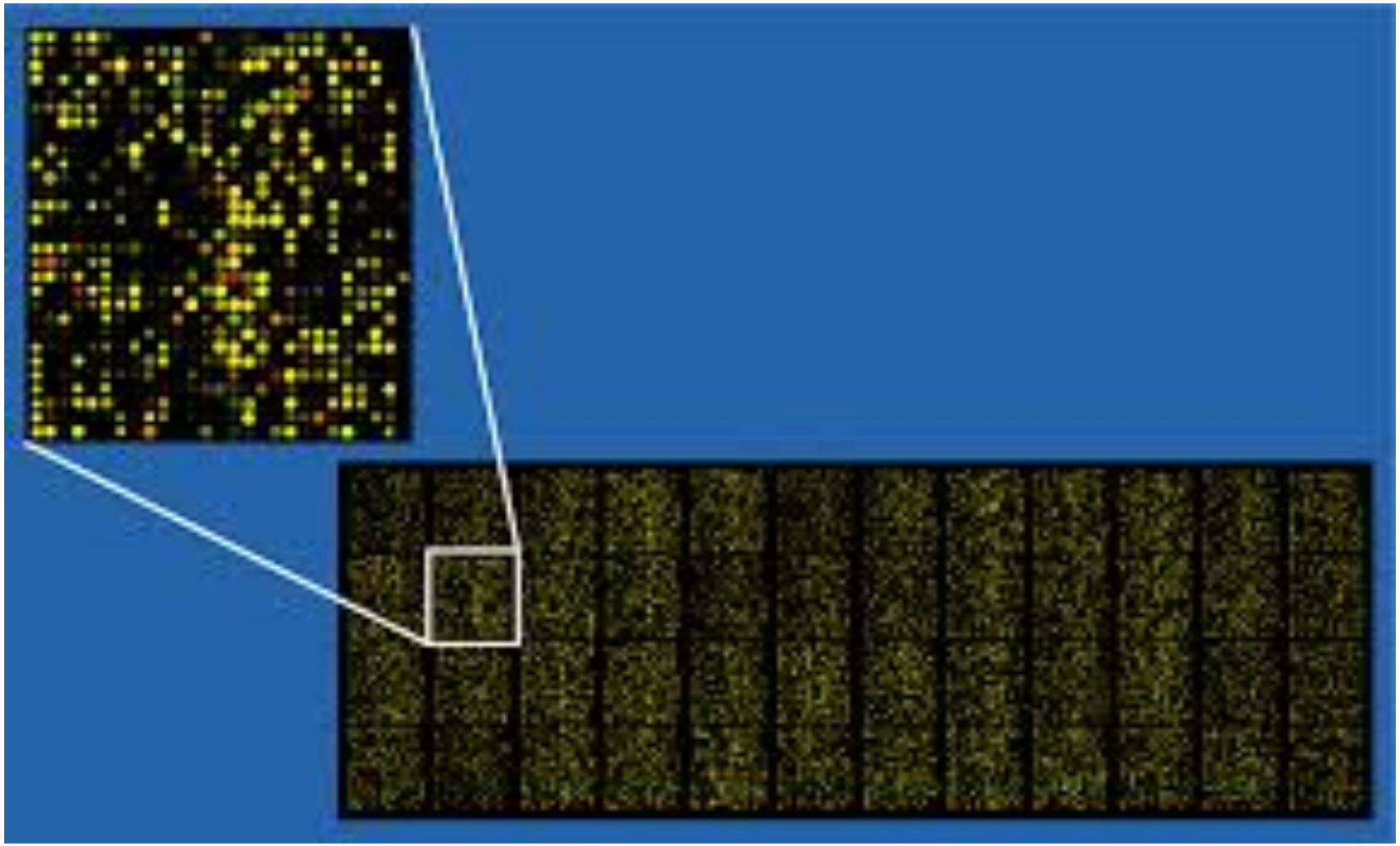
Part 1

Why data clustering?

This is a microarray

What does this mean?

How can we characterize gene expression data?



https://www.nature.com/scitable/definition/microarray-202/

Gene	Expression Vector						
YLR361C	0.14	0.03	-0.06	0.07	-0.01	-0.06	-0.01
YMR290C	0.12	-0.23	-0.24	-1.16	-1.40	-2.67	-3.00
YNR065C	-0.10	-0.14	-0.03	-0.06	-0.07	-0.14	-0.04
YGR043C	-0.43	-0.73	-0.06	-0.11	-0.16	3.47	2.64
YLR258W	0.11	0.43	0.45	1.89	2.00	3.32	2.56
YPL012W	0.09	-0.28	-0.15	-1.18	-1.59	-2.96	-3.08
YNL141W	-0.16	-0.04	-0.07	-1.26	-1.20	-2.82	-3.13
YJL028W	-0.28	-0.23	-0.19	-0.19	-0.32	-0.18	-0.18
YKL026C	-0.19	-0.15	0.03	0.27	0.54	3.64	2.74
YPR055W	0.15	0.15	0.17	0.09	0.07	0.09	0.07

Data clustering

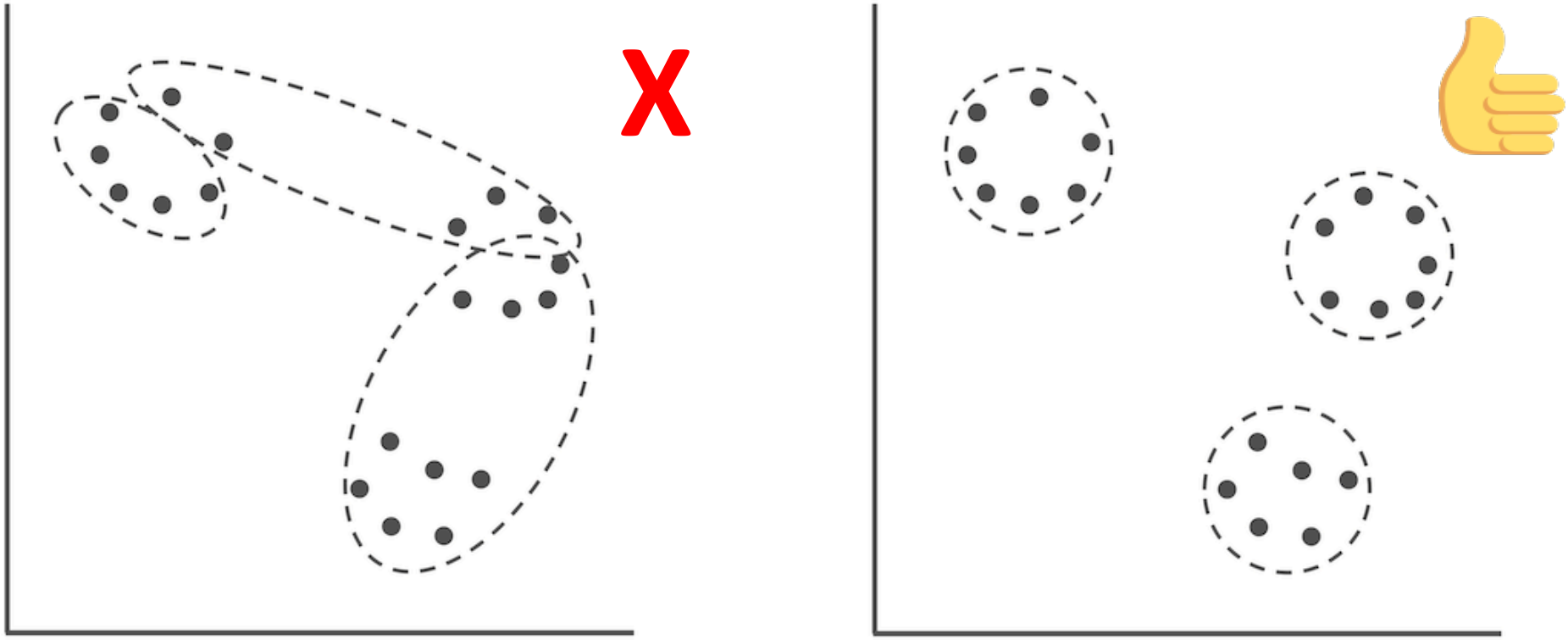
- Given a collection of data points, can we identify patterns?
- Data points:
  - DNA sequences
  - Gene expression levels
  - Abundances of organisms in an environment
  - Vitals
  - Much more…
- Patterns:
  - do certain points group together?

Types of clustering algorithms

- Agglomerative
  - Start with a single observation
  - Group similar observations into the same cluster
- Divisive
  - All data points start in the same cluster
  - Iteratively divide the cluster until you find good clustering
- Hierarchical
  - Build a tree
  - Leaves are data points and internal nodes represent clusters

The Good Clustering Principle

- Homogeneity: All points in the cluster must be similar
- Separation: Points in different clusters are dissimilar



Issues with clustering

- Good clustering may not be achievable
- Finding the optimal clustering is usually NP-hard
- In how many ways can you partition n points into 2 clusters?