Rice Data Science Bootcamp

Instructor: Natalie Berestovsky, PhD
Anadarko Petroleum

Supervised vs unsupervised

	Supervised	**Unsupervised**
Discrete	***Classification***	***Clustering***
	KNN **Datasets:** *Wine dataset*	K-Means Hierarchical clustering (Biclustering) **Datasets:** Synthetic data NCI60
Continuous	***Regression***	***Dimensionality reduction***
	Multilinear regression Ridge regression Lasso regression **Datasets:** *Boston housing dataset* *Synthetic data*	Principle Component Analysis (PCA) **Datasets:** Art data NCI60

Unsupervised learning best practices

- No “true” (response) data
- Exploratory analysis
  - Data driven discoveries
  - Hypothesis generating
- Show a data driven discovery is stable
  - Small changes in data, algorithm, parameter yield similar results
  - Multiple approaches yield the similar results
- Corroborate via existing knowledge / literature

1. Always visualize.
2. Use multiple techniques.
3. Validate discoveries when possible.
4. Communicate uncertainty.
5. Make your analysis reproducible.

Clustering

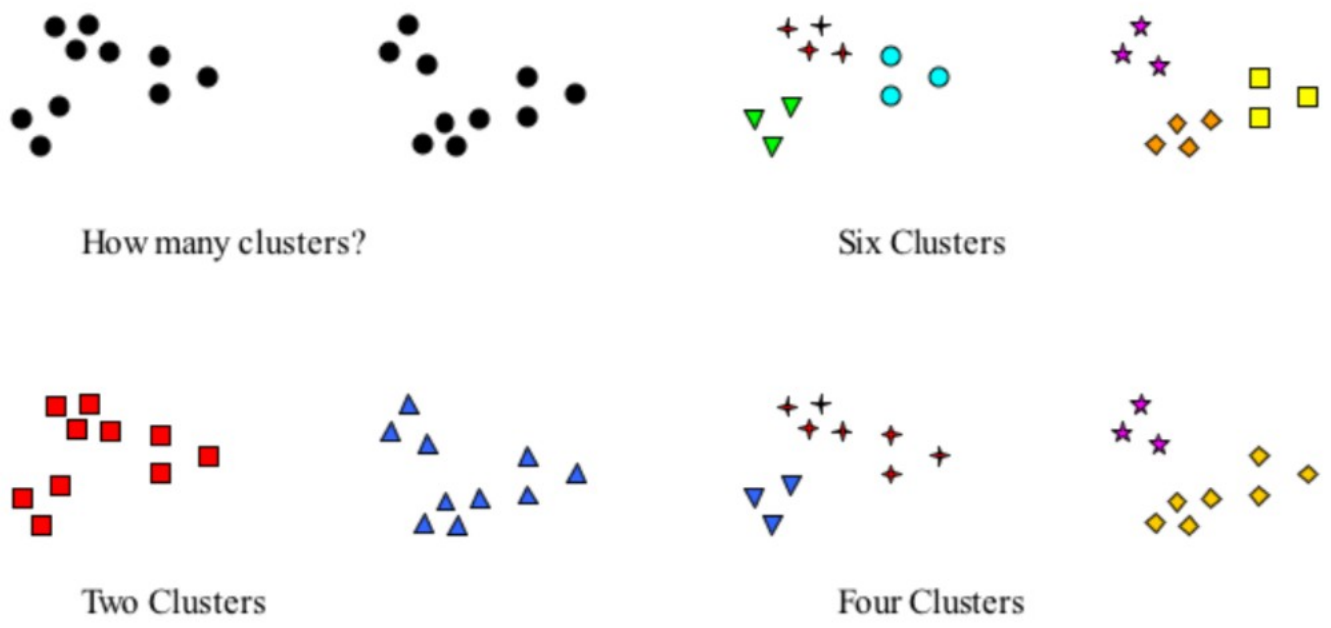
Objective:

- Definition: Group or segment the data set (a collection of objects) into subsets so that those within each subset are more closely related to others than those objects assigned to other subsets.
- Each group (subset) is called a cluster.

Challenging:

- What is a meaningful cluster?
- How do we validate clustering results?

What are meaningful clusters?



K Means

- Objective: minimize within cluster dissimilarity W(C)
  - *Using squared Euclidean distance*
  - *n* observations, *K* clusters
  - Initialize clusters OR centroids randomly

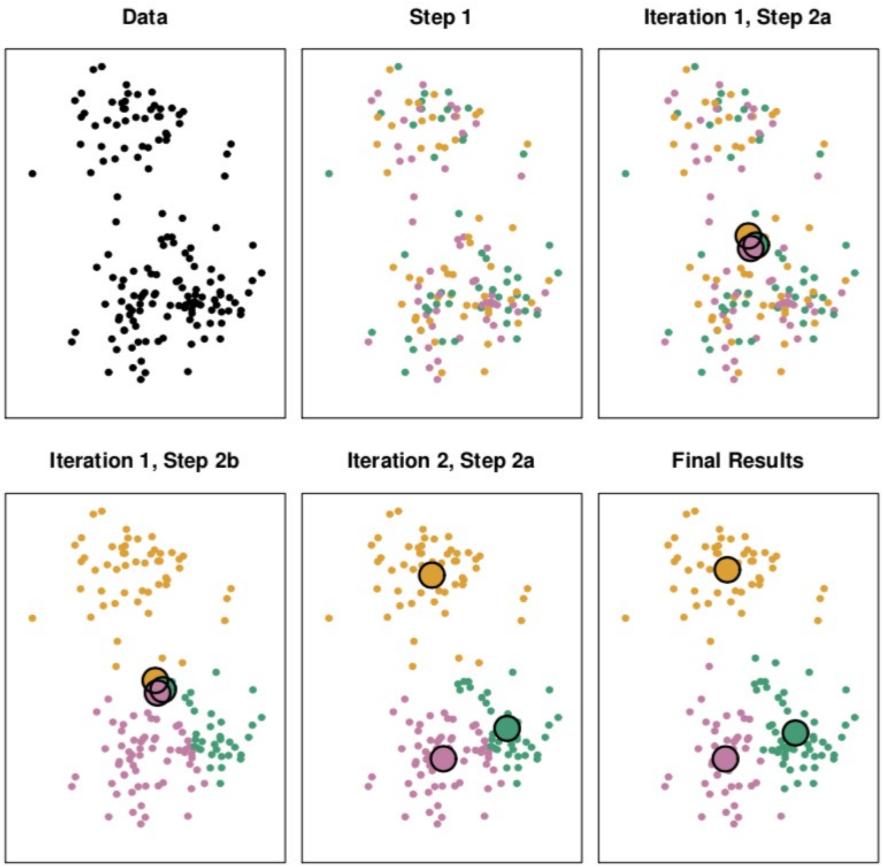
Idea:

- Augment $\mathrm{W(C)}$ with cluster means, $\mathbf{m}_{\mathrm{k}}$:

$$\mathrm{W}(\mathrm{C}, \mathbf{m}_{\mathrm{k}}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{n}_{\mathrm{k}} \sum_{\mathrm{C(i)=k}} ||\mathbf{x}_{\mathrm{i}} - \mathbf{m}_{\mathrm{k}}||_2^2$$

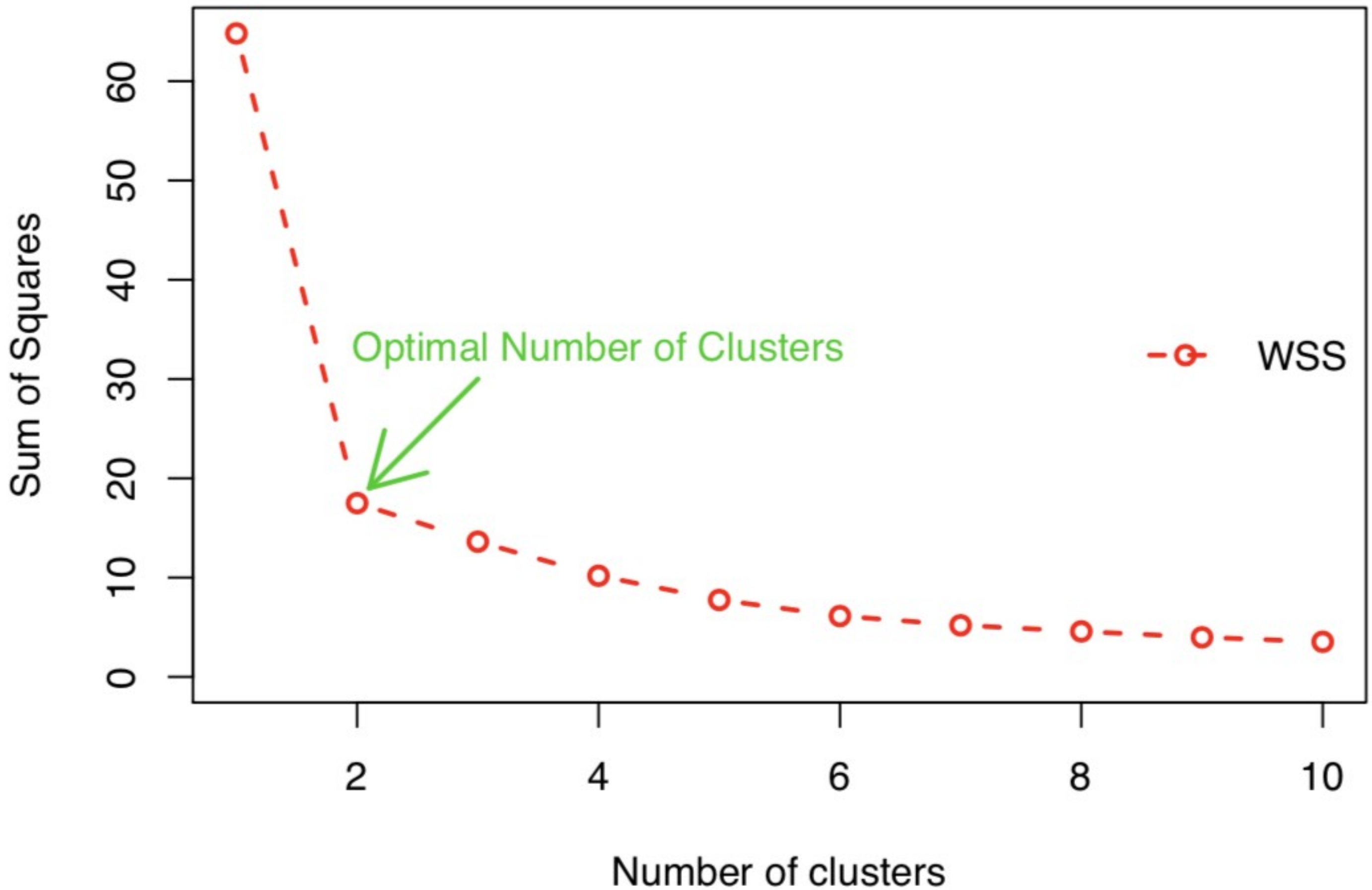
- Minimize $\mathrm{W}(\mathrm{C}, \mathbf{m}_{\mathrm{k}})$ by iteratively optimizing:
  1. Cluster means: $\mathbf{m}_{\mathrm{k}}$ (with $\mathrm{C(i)}$ fixed).
  2. Cluster assignments: $\mathrm{C(i)}$ (with $\mathbf{m}_{\mathrm{k}}$ fixed).

K Means



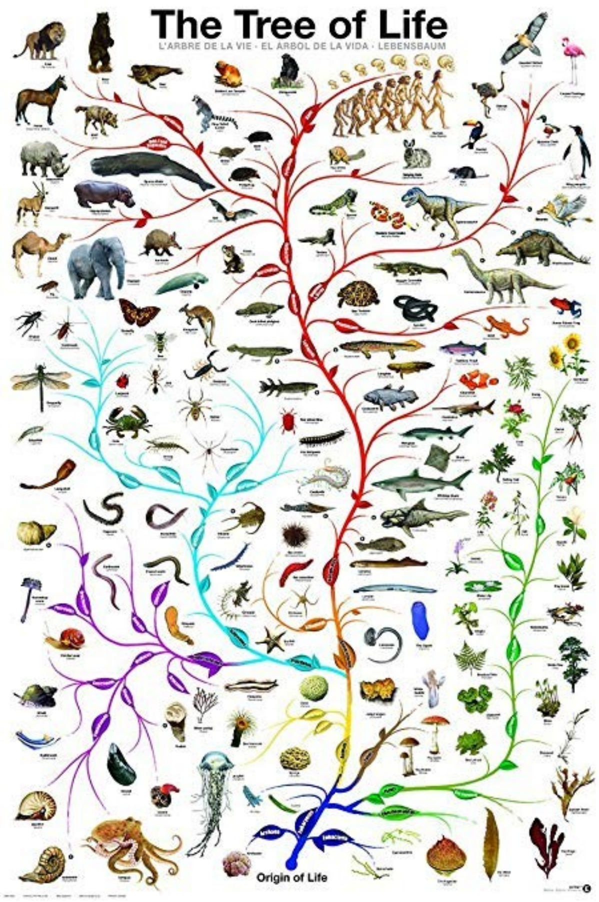
- Highly dependent on initialization
- Local solution
- Good for compact spherical clusters

- Elbow method



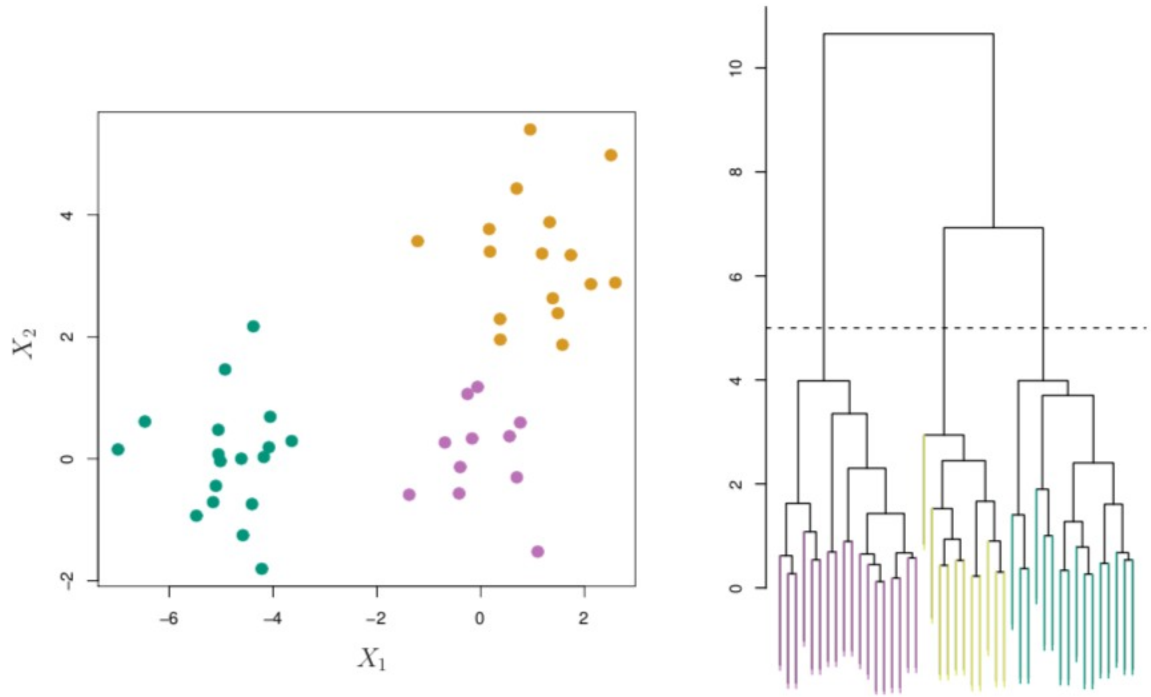
Hierarchical clustering

- Nested Clusters: Produce hierarchical representations in which the clusters at each level of the hierarchy are created by merging clusters at the next lower level.
- At the lowest level, each cluster contains a single observation.
- At the highest level there is only one cluster containing all observations.
- Two paradigms: agglomerative (bottom-up; most popular) and divisive (top-down; less popular).
- Use dendrogram to display the clustering result.



Hierarchical clustering

- Leaf for each observation.
- As we move up the tree, similar leaves begin to fuse into branches
- Observations that fuse near the top of the tree, can be quite different.
- The lower in the tree fusions occur, the more similar the groups of observations are to each other.



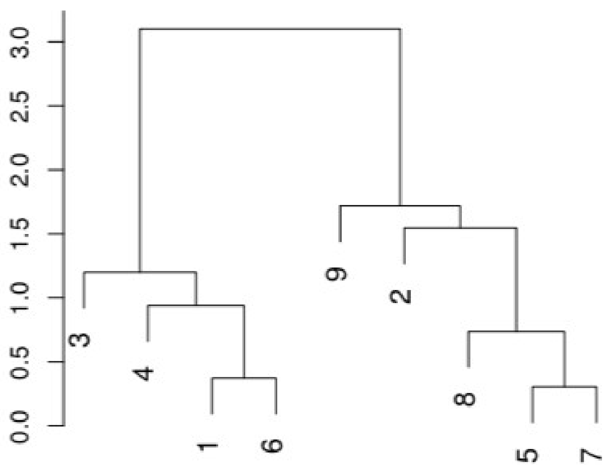
- **Height** of fusions indicate how **similar** objects are.
- Horizontal axis does not indicate anything

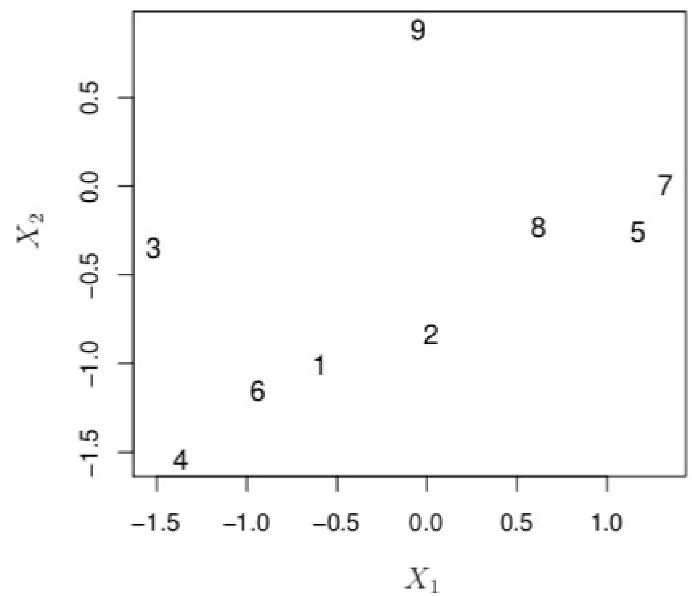
Agglomerative Clustering

- Begin with every observation representing a singleton cluster.

- At each step, merge two “closest” clusters into one cluster and reduce the number of clusters by one.

- Need a measure of dissimilarity between two clusters - called linkages.

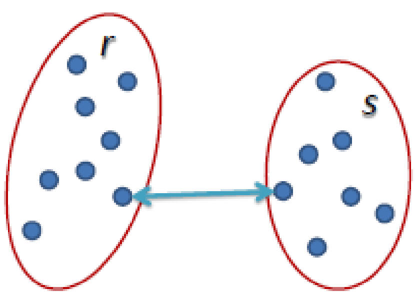




Linkage

Single linkage

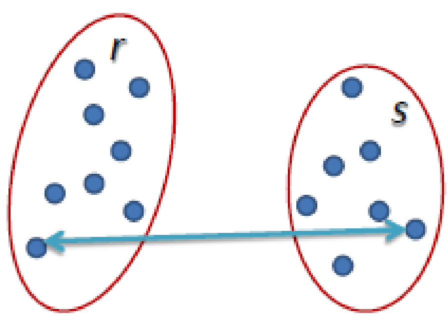
- the distance between two clusters is defined as the *shortest* distance between two points in each cluster



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete linkage

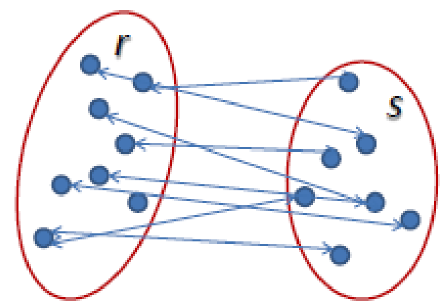
- the distance between two clusters is defined as the *longest* distance between two points in each cluster



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

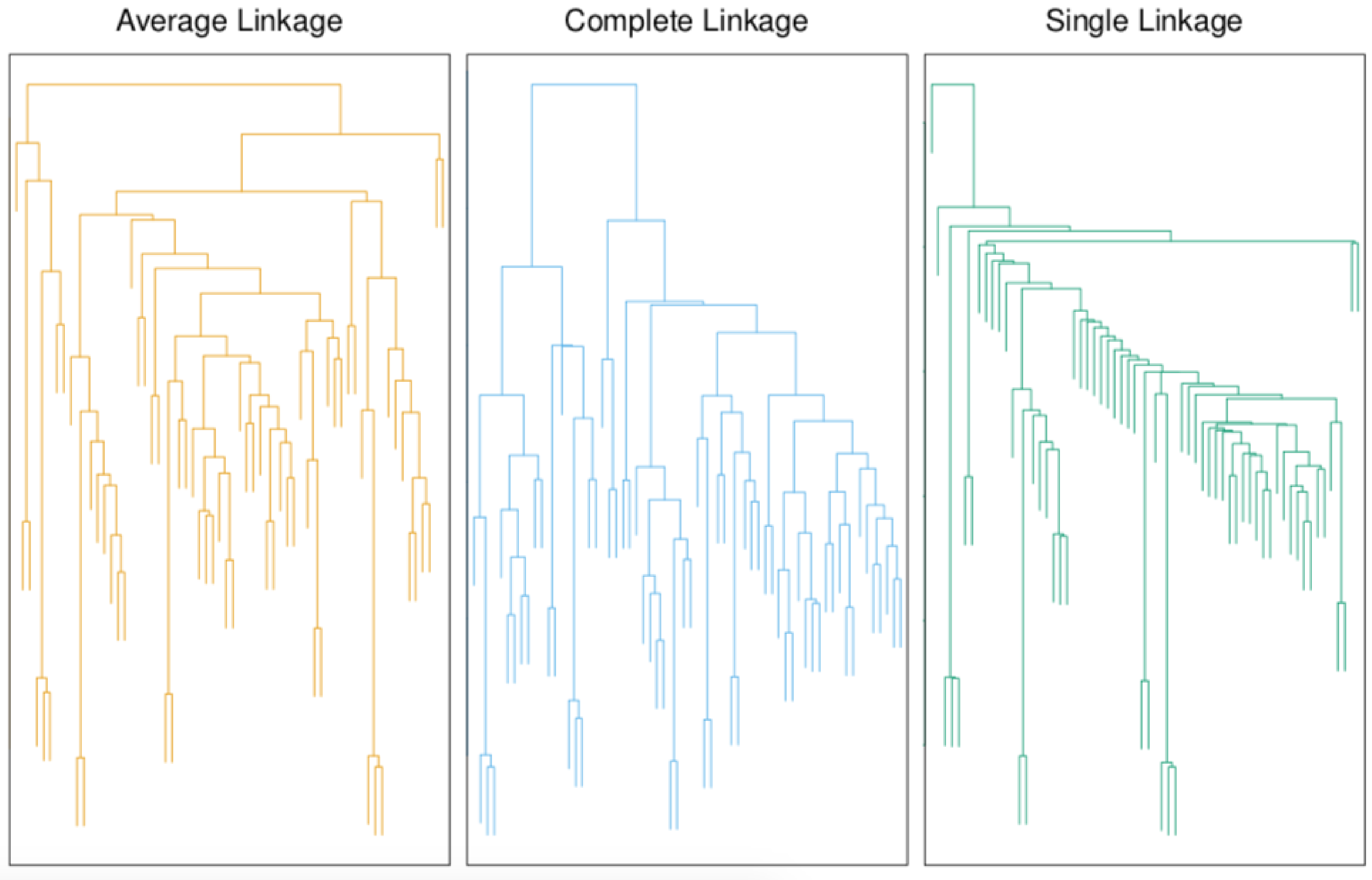
Average linkage

- the distance between two clusters is defined as the *average distance between each point* in one cluster to every point in the other cluster

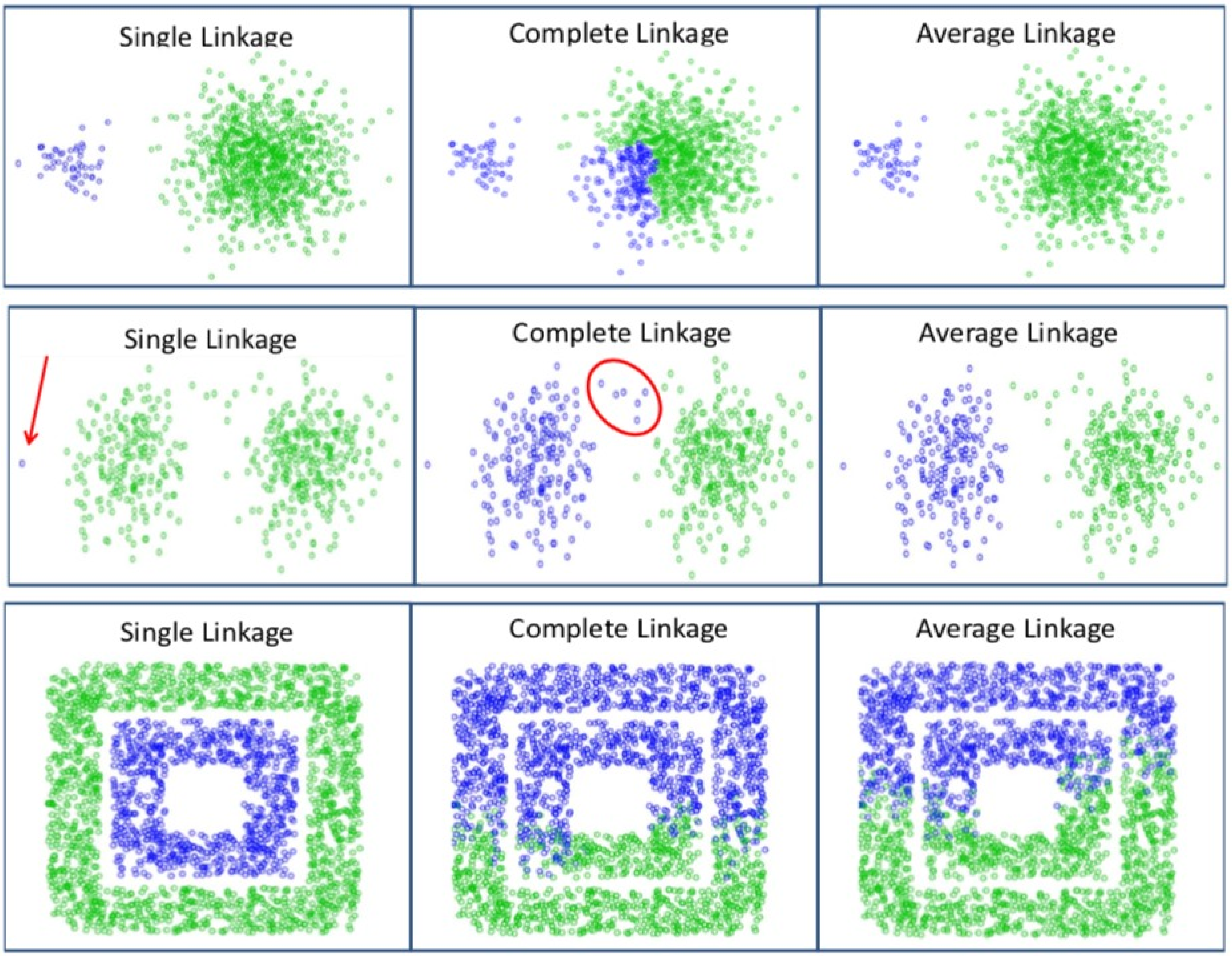


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Linkage



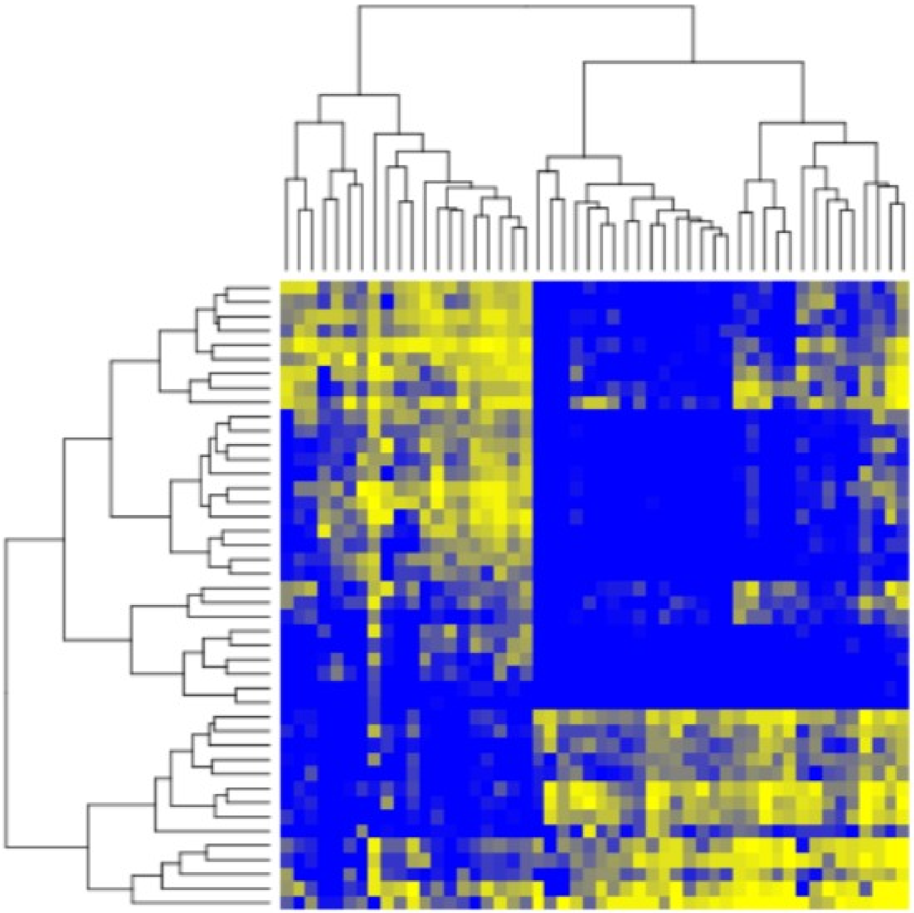
Linkage examples



- Complete linkage often gives comparable cluster sizes
- Single linkage is sensitive to outliers

Bi-clustering

- Find groups of BOTH observations & features.
- Clustering both rows and columns of data matrix.
- Applications in Omics data



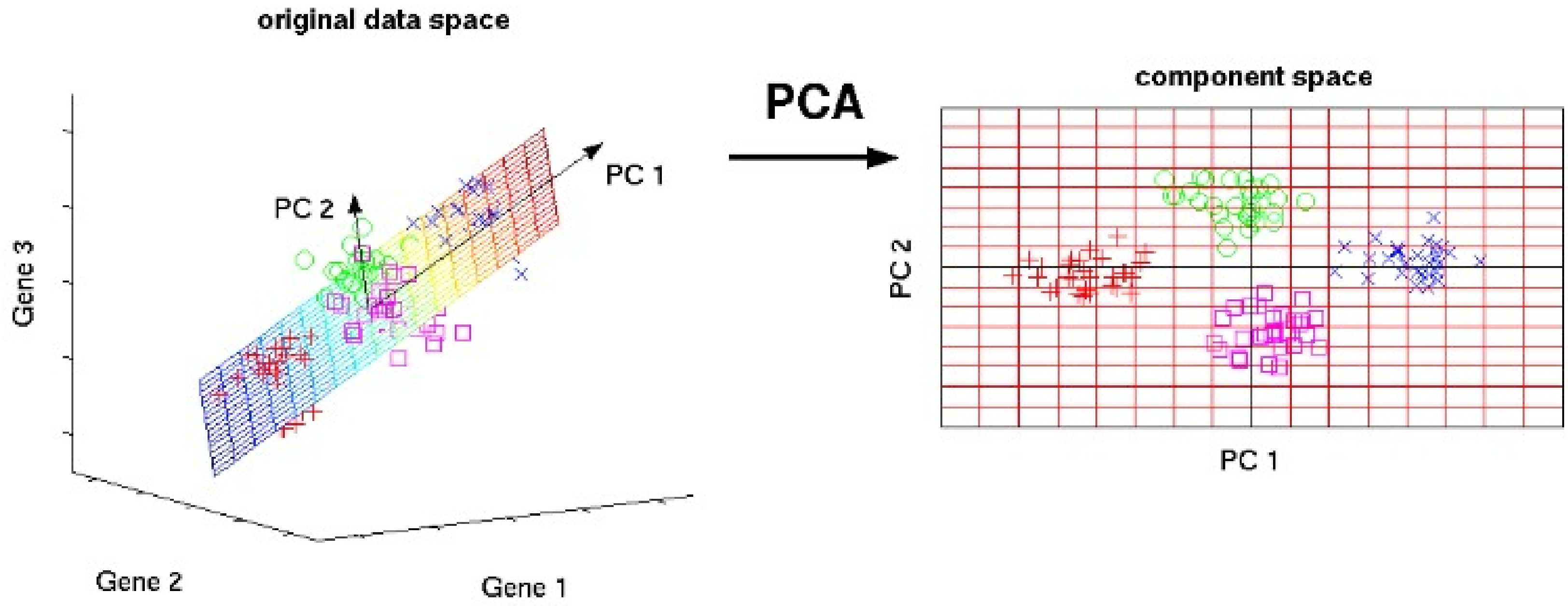
Supervised vs unsupervised

	Supervised	Unsupervised
Discrete	***Classification***	***Clustering***
	KNN **Datasets:** *Wine dataset*	K-Means Hierarchical clustering (Biclustering) **Datasets:** Synthetic data NCI60
Continuous	***Regression***	***Dimensionality reduction***
	Multilinear regression Ridge regression Lasso regression **Datasets:** *Boston housing dataset* *Synthetic data*	Principle Component Analysis (PCA) **Datasets:** Art data NCI60

Dimensionality reduction

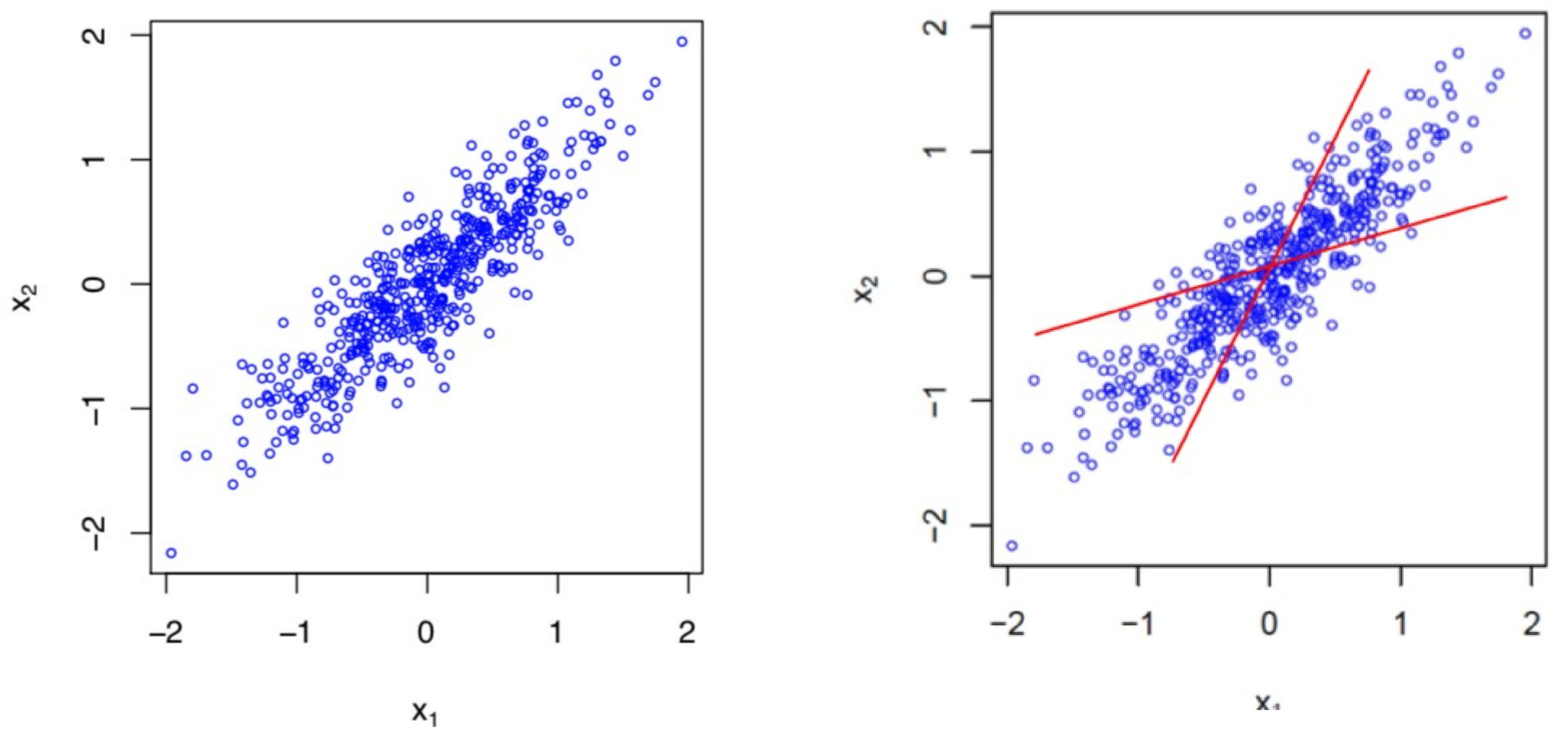
- For Big-Data:
  - Data visualization difficult!
  - High degrees of redundancy
  - Many features may be uninformative.
- Dimension Reduction Idea:
  - Map data into lower-dimensional space that retains important information.

Principle Component Analysis (PCA)



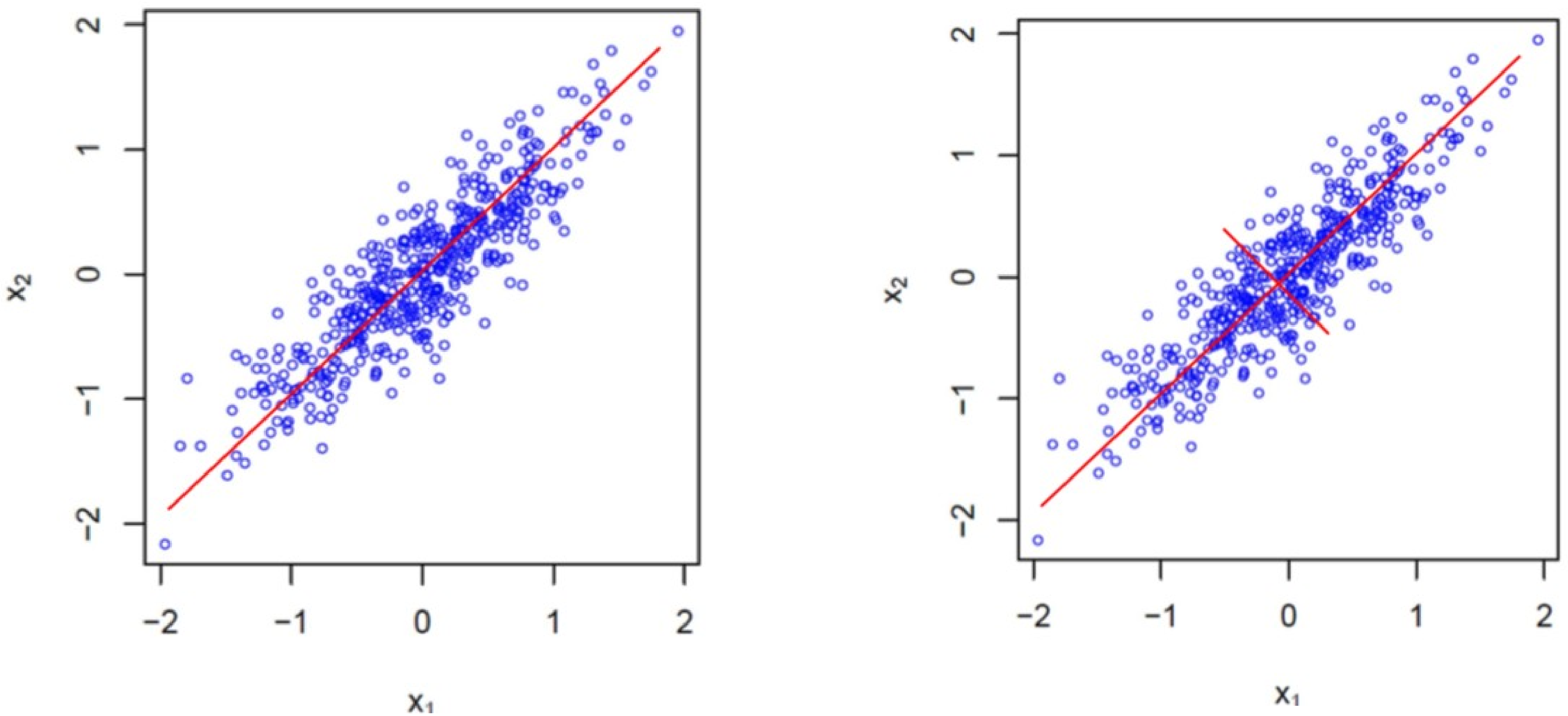
- Data matrix: $X_{n \times p}$, n observations and p features.
- Find low-dimensional representations that capture most of the variation in the data.

- What is a good 1D representation for this data?



PCA

- Find line that maximizes the variance of the data projected onto the line
- Subsequent components orthogonal (perpendicular).



PCA main idea

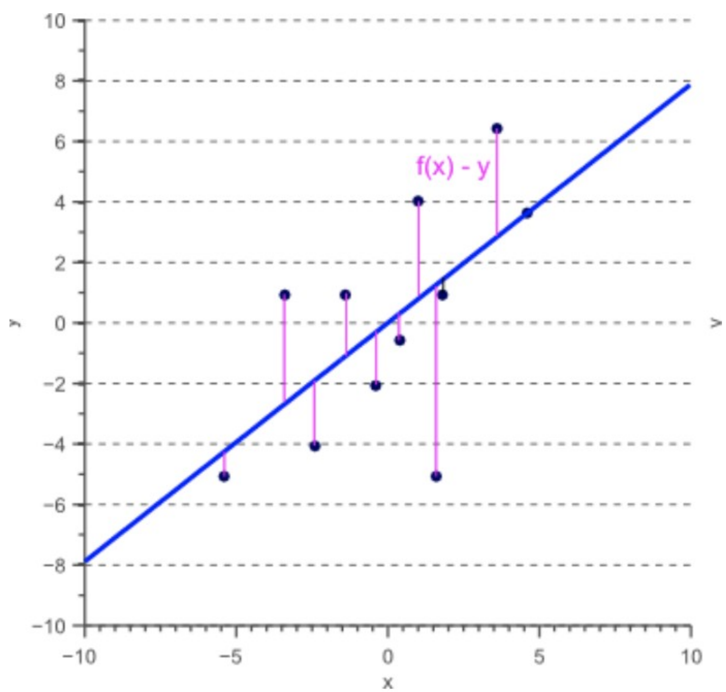


Figure 6. Linear regression where x is the independent variable and y is the dependent variable, corresponds to minimizing the vertical projection error.

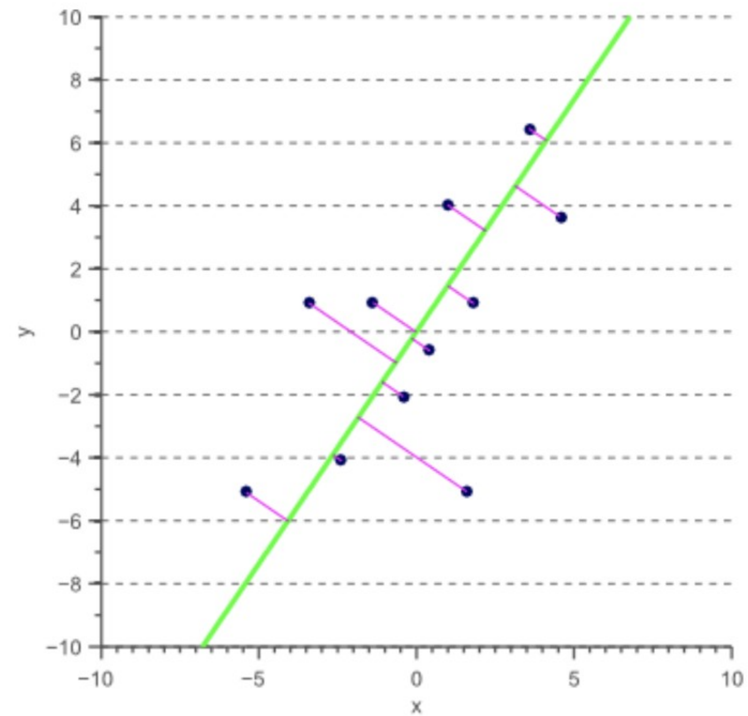
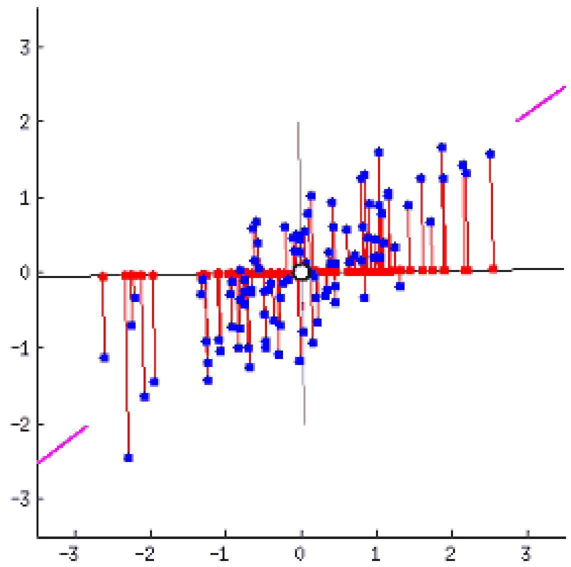


Figure 8. Linear regression where both variables are independent corresponds to minimizing the orthogonal projection error.

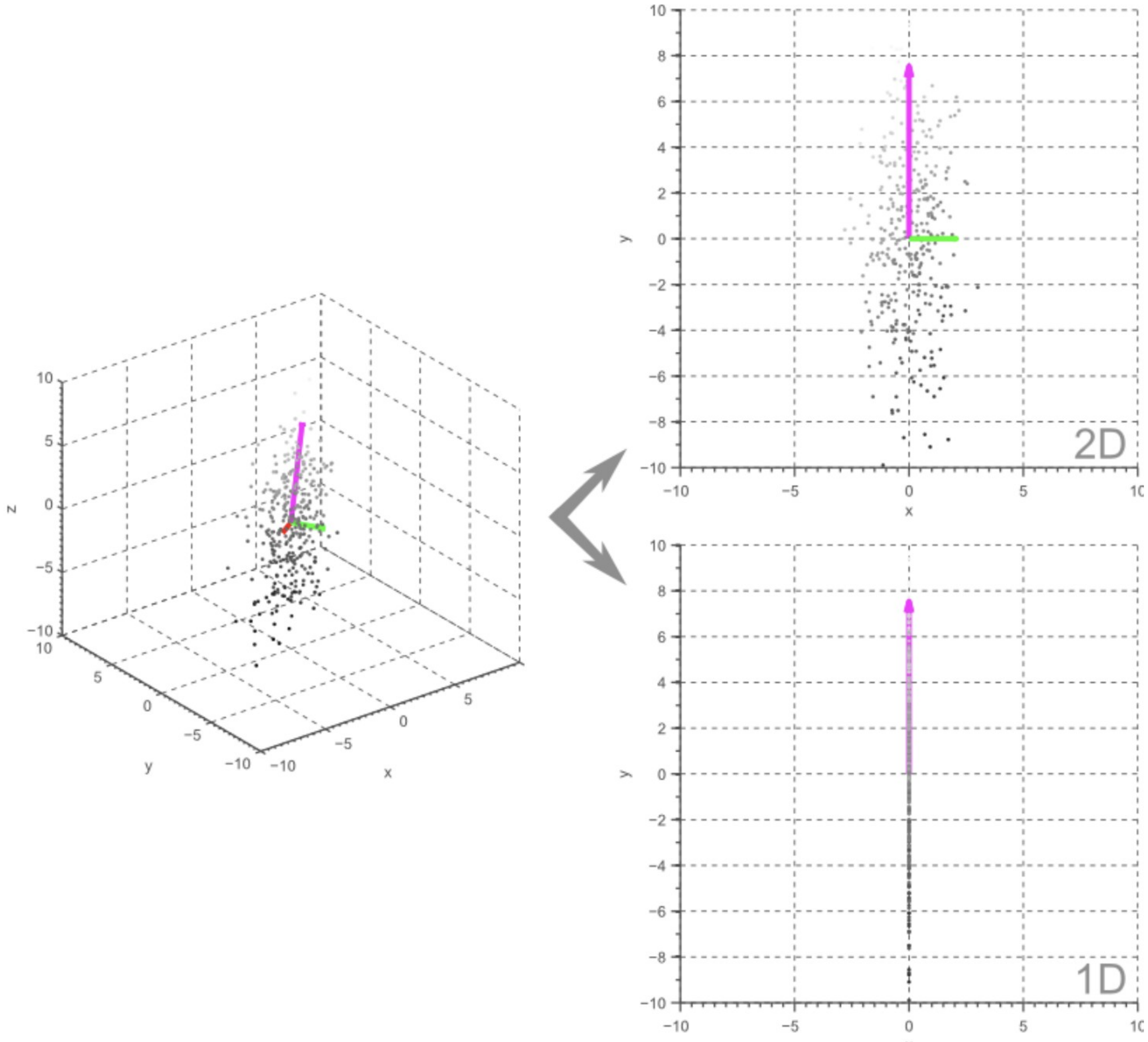
- PCA minimizes orthogonal projection onto line: $z = v_1 x_1 + v_2 x_2$.
- Note: Not same as OLS (ordinary least squares) which minimizes projection of y onto x!

PCA components



- The **first principal component** accounts for the **largest possible variance** in the data set
- The **second principal component** is calculated in the same way, with the condition that it is uncorrelated with (i.e., *perpendicular to*) the first principal component and that it accounts for the next highest variance.
- This continues until a total of **p principal components** have been calculated, *equal to the original number of variables*.

PCA visualization



PCA drawbacks

- Linear projections!

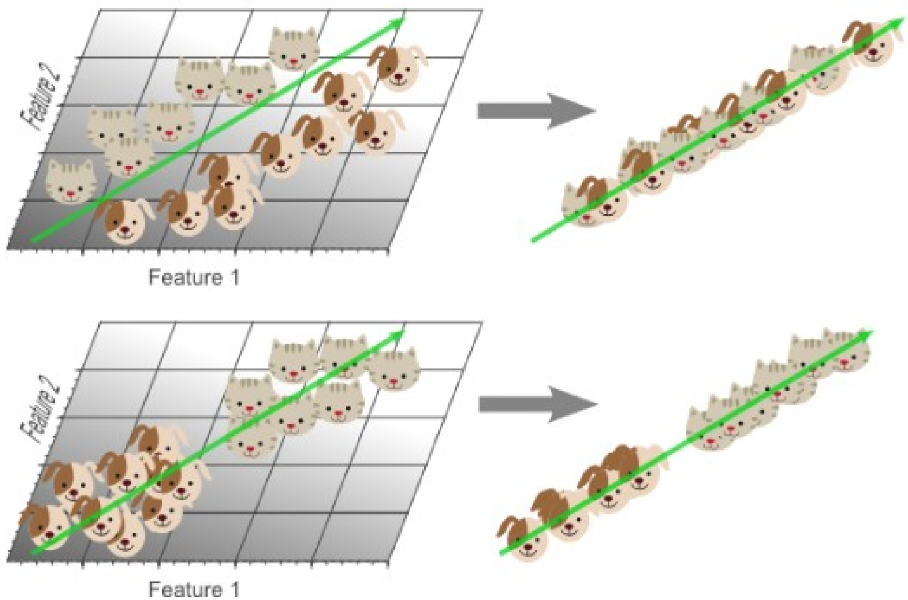


Figure 12. In the first case, PCA would hurt classification performance because the data becomes linearly unseparable. This happens when the most discriminative information resides in the smaller eigenvectors.

t-SNE

- t-Distributed Stochastic Neighbor Embedding
  - Laurens van der Maaten, 2008

- Step 1: In the high-dimensional space, create a **probability distribution** that dictates the relationships between various neighboring points

- Step 2: It then tries to **recreate** a low dimensional space that follows that probability distribution as best as possible.

- The “t” in t-SNE comes from the t-distribution, which is the distribution used in Step 2. The “S” and “N” (“stochastic” and “neighbor”) come from the fact that it uses a probability distribution across neighboring points.

Links

- https://idyll.pub/post/dimensionality-reduction-293e465c2a3443e8941b016d/