Week 4

Concepts in Machine Learning

fredhutch.io

Fred Hutchinson Cancer Research Center

Week 4 Learning Objectives

- Quick CRISP-DM Review
- Supervised Learning vs Unsupervised Learning
- Clustering
- Curse of Dimensionality
- PCA
- Transfer Learning

Definitions

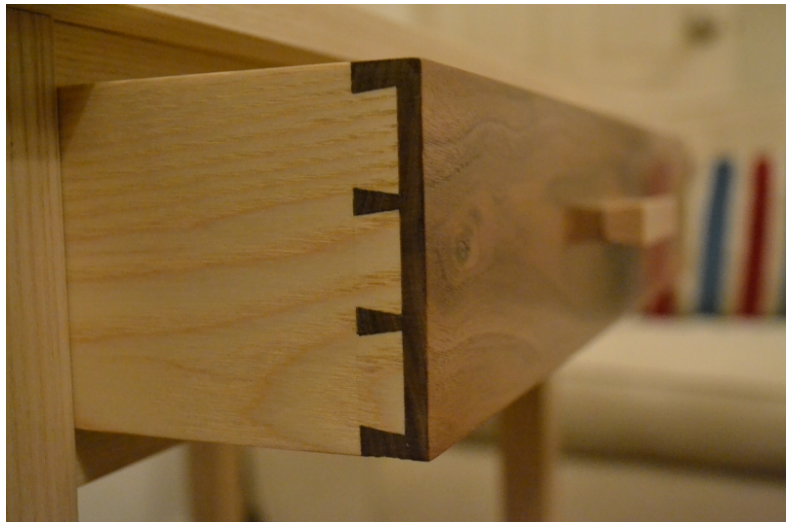
- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. (https://expertsystem.com/machine-learning-definition/)
- Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. (https://en.wikipedia.org/wiki/Machine_learning)
- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight

Experimental Design

- Difficult to master or even do well
- Close interplay between
  - Goals
  - Methods
  - Data
  - Execution
- Requires thoughtful approach and broad understanding

Capable Cabinet Maker

- Inspects and understands raw materials
- Uses the tools thoughtfully to shape and join materials
- Chooses approach and tools based on materials and goals
- Thoughtfulness born of experience



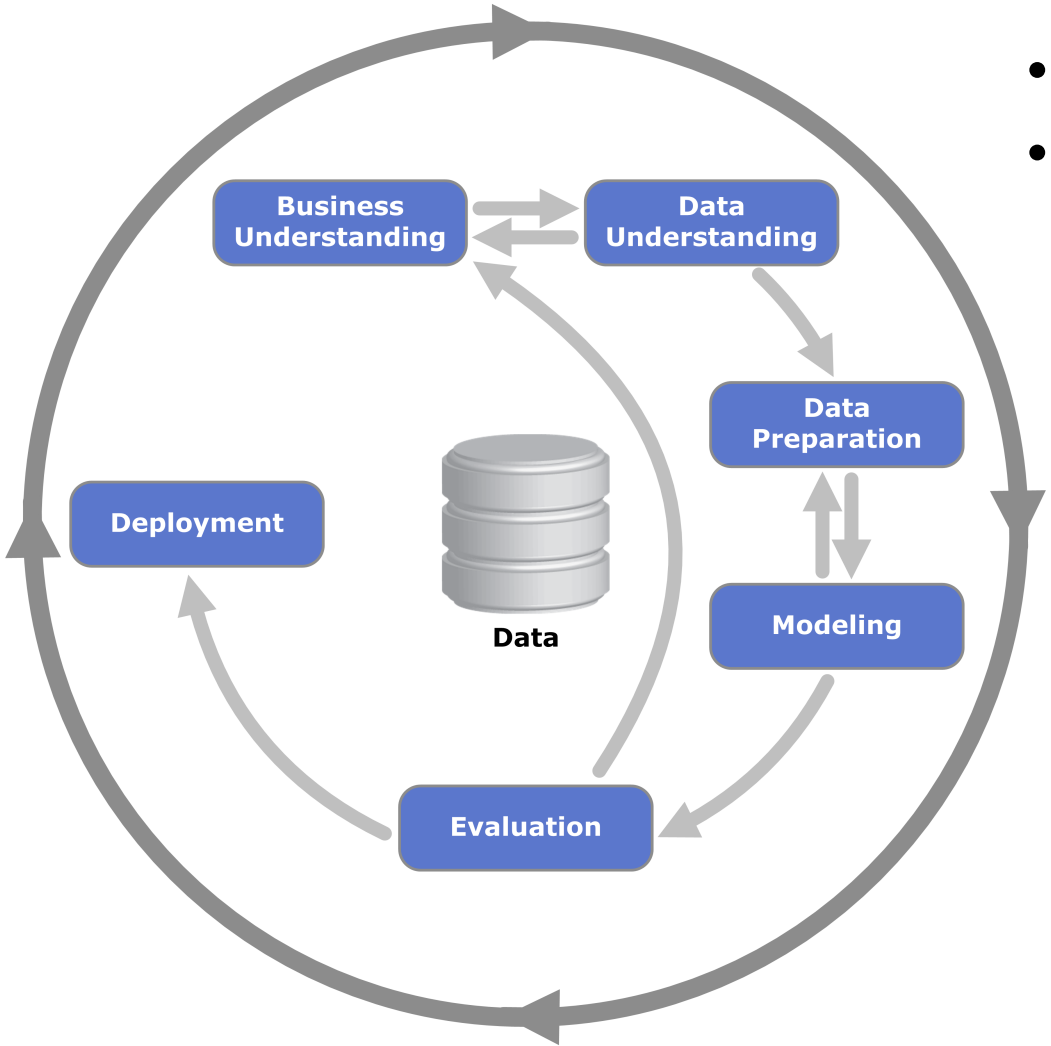
Cabinet Making?

- Storage Need + Raw Materials + Tools + Work ==> Cabinet

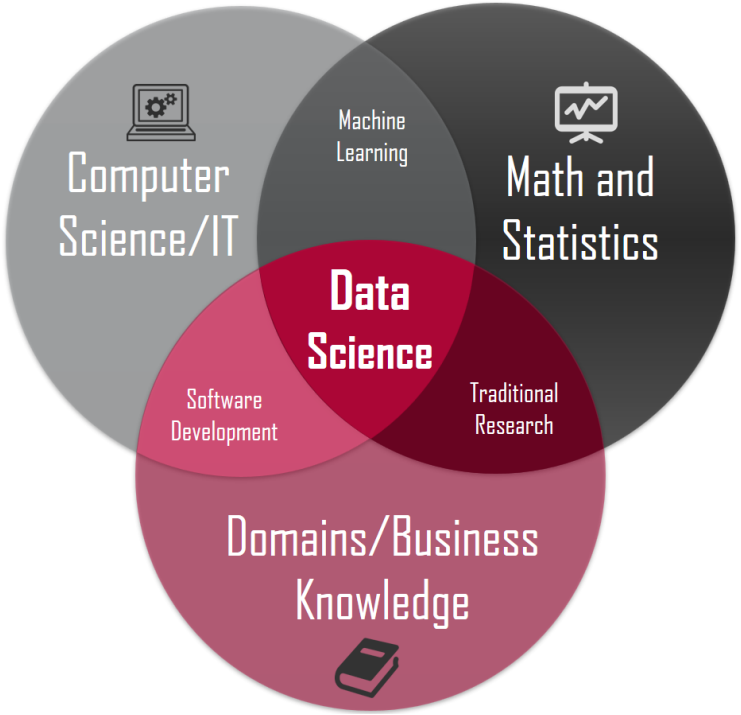
- SAT analogy format

  Tools : Cabinet Making as Algorithms : Machine Learning

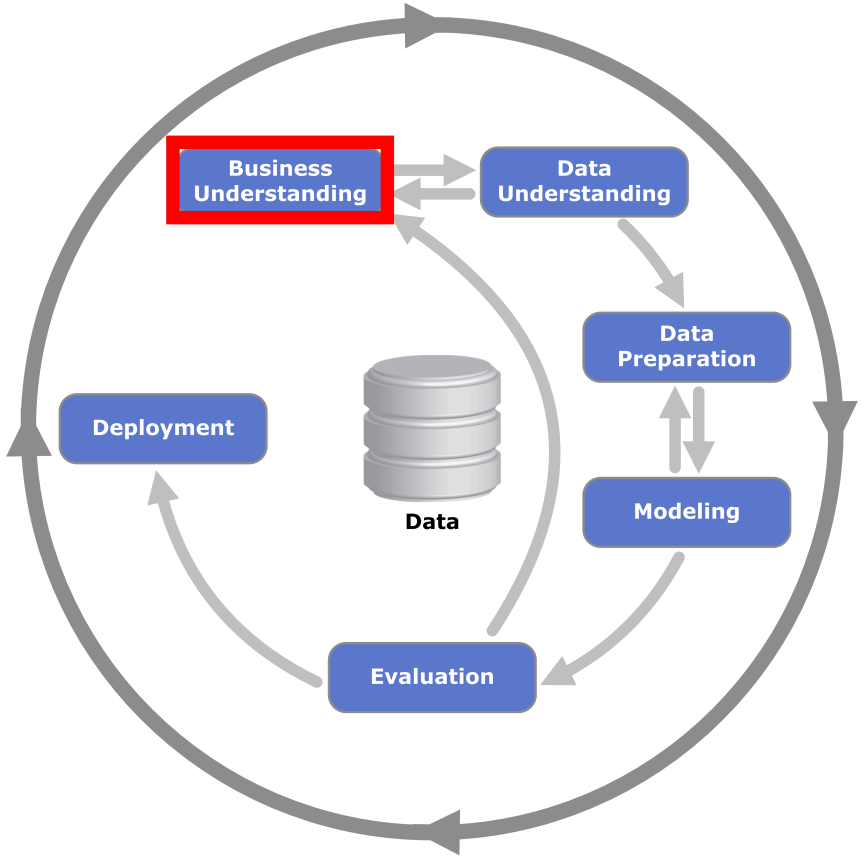
CRISP-DM
Cross-industry standard process for data mining



- Cyclical
- Iterative
- Connecting all 3 areas of the classic notion of “data science”

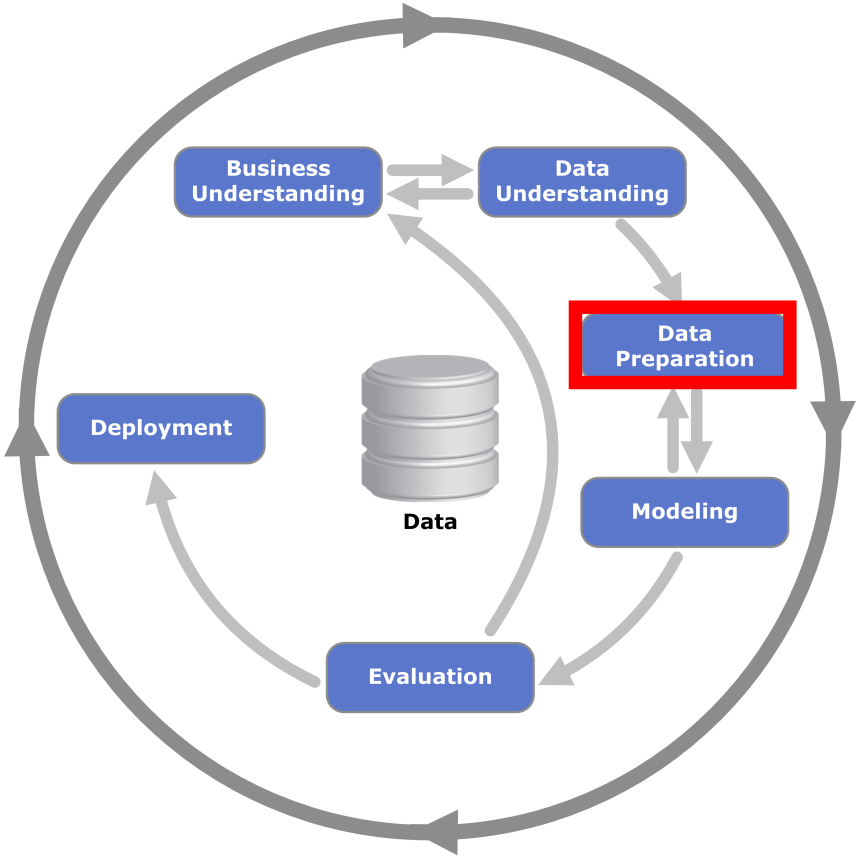


Business/Scientific Understanding



Data Understanding

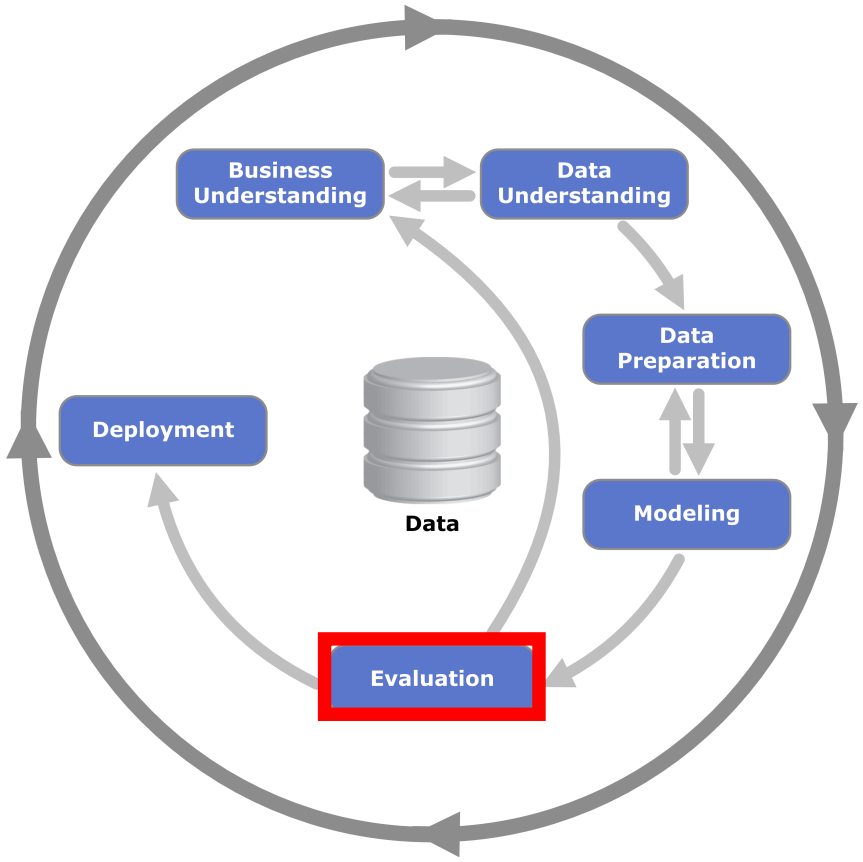
Data Preparation



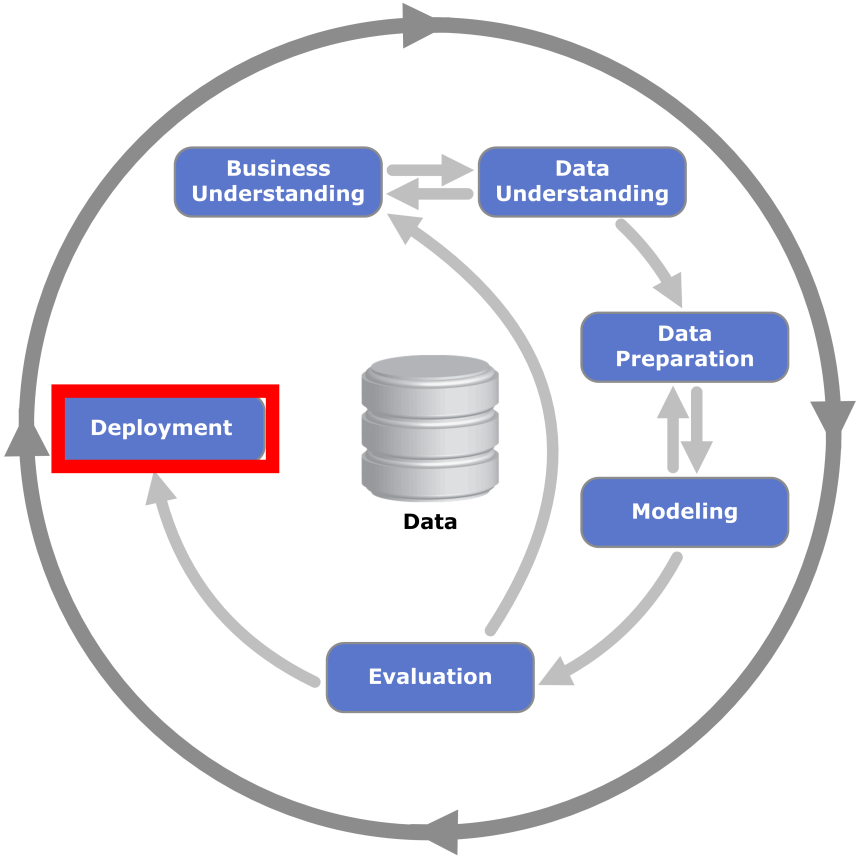
Modeling


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Evaluation

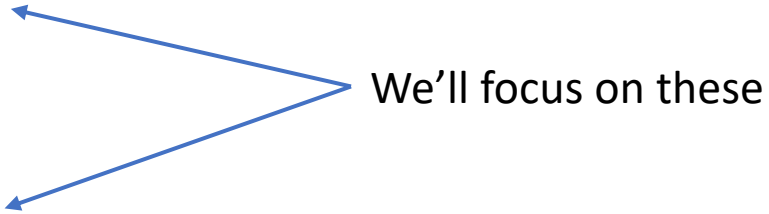

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning



- Reinforcement Learning
- Semi-Supervised Learning

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning

Unsupervised Problems

- What if you are doing more exploratory work?
- What if you want to make your supervised algorithm run more smoothly by reducing the dimensionality of your problem?
- What if you want to supplement your algorithm with additional variables?
- What if you have some instincts about sub-groups that may exist within your population?
- What if, for whatever reason, you don't have a label you can do supervised learning over?

Unsupervised Problems

- Can we find out anything interesting by comparing aspects of our data to each other?
- For our purposes, think of it as grouping our data, and assigning meaning to the groups after the fact

Unsupervised Problems

- Supervised problems can be characterized as having two phases: training and testing.
  - Show a kid many examples of planes, trains, and cars made from Legos, pointing out each one's type.
  - Show her a new Lego plane, train, or car and ask which it is
- Unsupervised problems have a single phase: fitting
  - Show a kid many examples of planes, trains, and cars made from Legos, without naming each one's type but asking her to put them in groups.

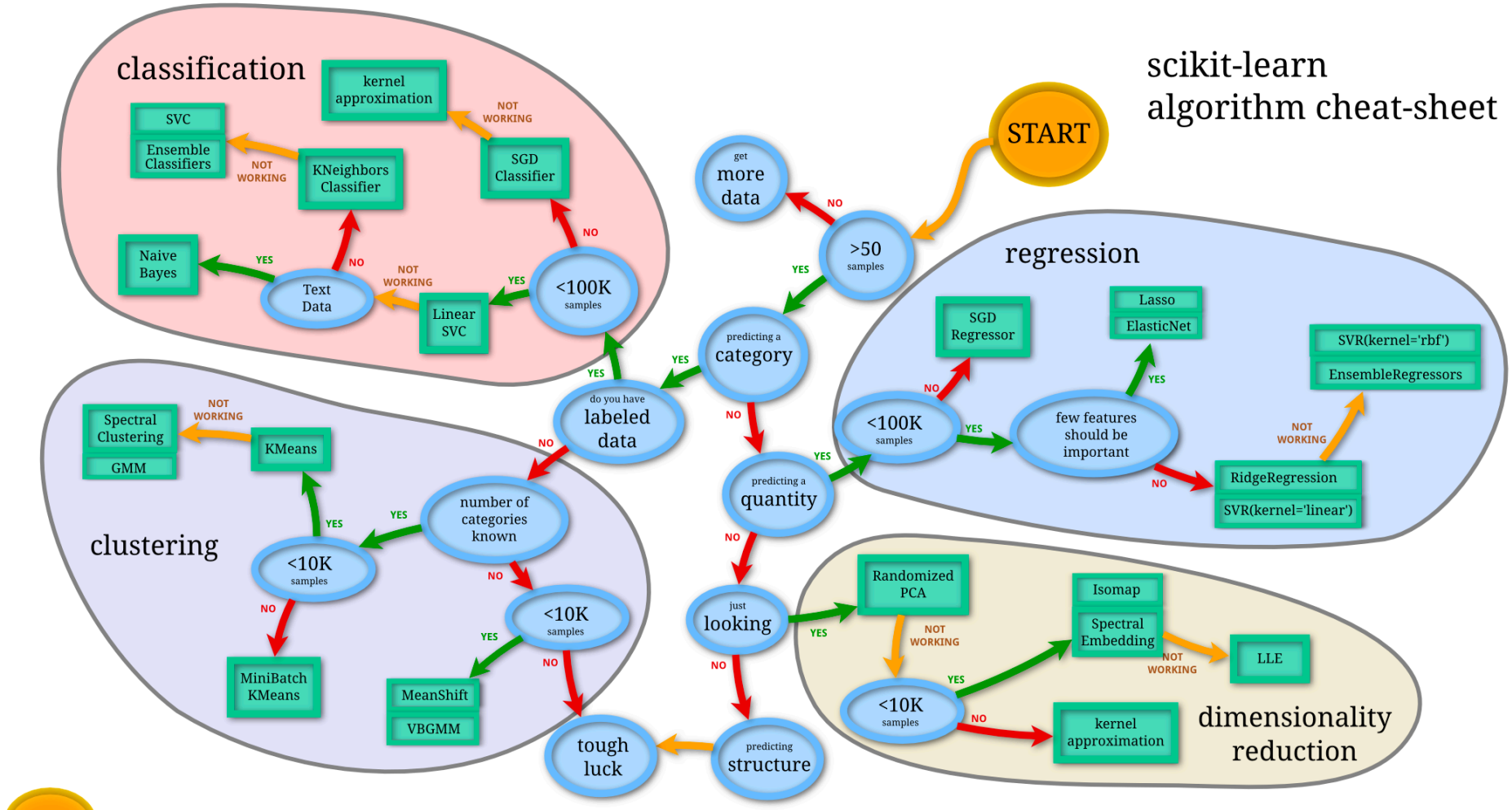
Question Time

- Why does a unsupervised problem only have a single phase?
  - Since there is no canonically “correct” target/outcome to predict, there is no “training” to give better predictions, or “testing” to evaluate how generalizable our model is.

Unsupervised Learning

- More subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
- But techniques for unsupervised learning are of growing importance in a number of fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements,
  - groups of shoppers characterized by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie viewers.

A Nifty Chart



A Nifty Chart

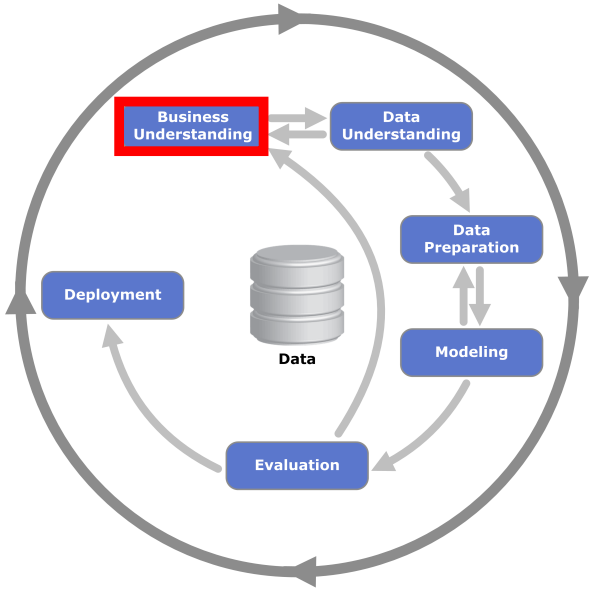
https://scikit-learn.org/stable/tutorial/machine_learning_map/

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other

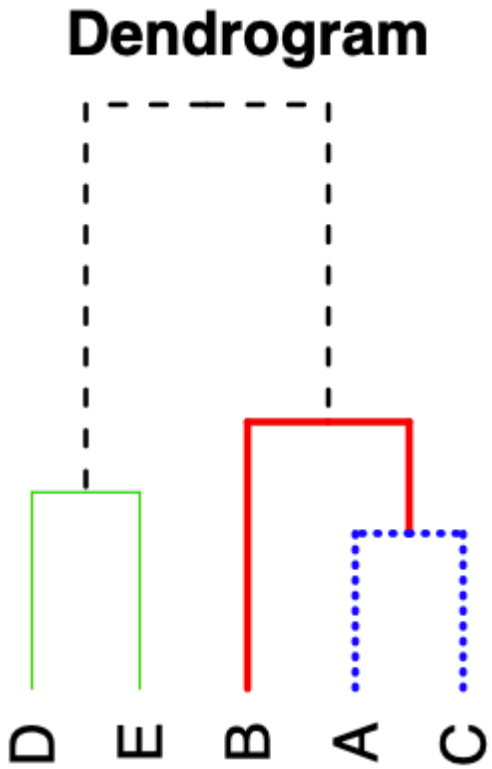
Clustering

- We must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.
  - CRISP-DM

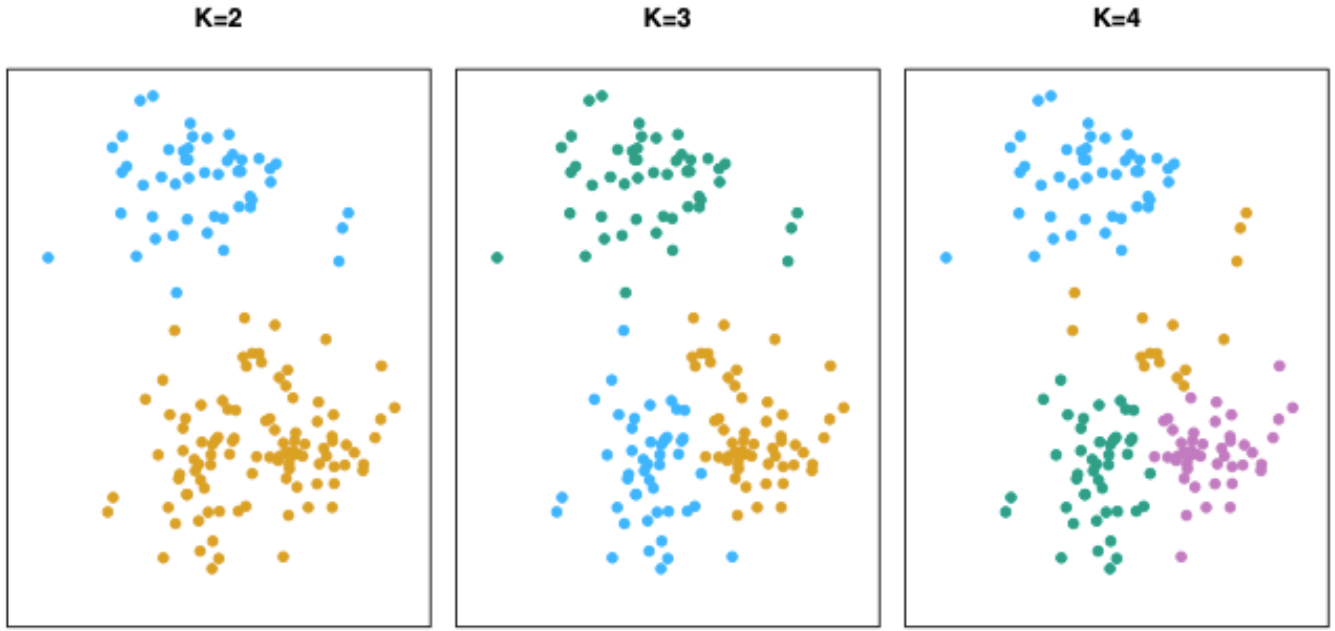


Two clustering methods

- K-means clustering
  - Seek to partition the observations into a pre-specified number of clusters.
- Hierarchical clustering
  - Do not know in advance how many clusters we want
  - We end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.



K-means clustering



- A simulated data set
  - 150 observations
  - 2-dimensional space
- Panels show the results of applying K-means clustering with different values of K, the number of clusters
- The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm
- There is no ordering of the clusters
  - The cluster coloring is arbitrary
  - These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

Details of K-means clustering

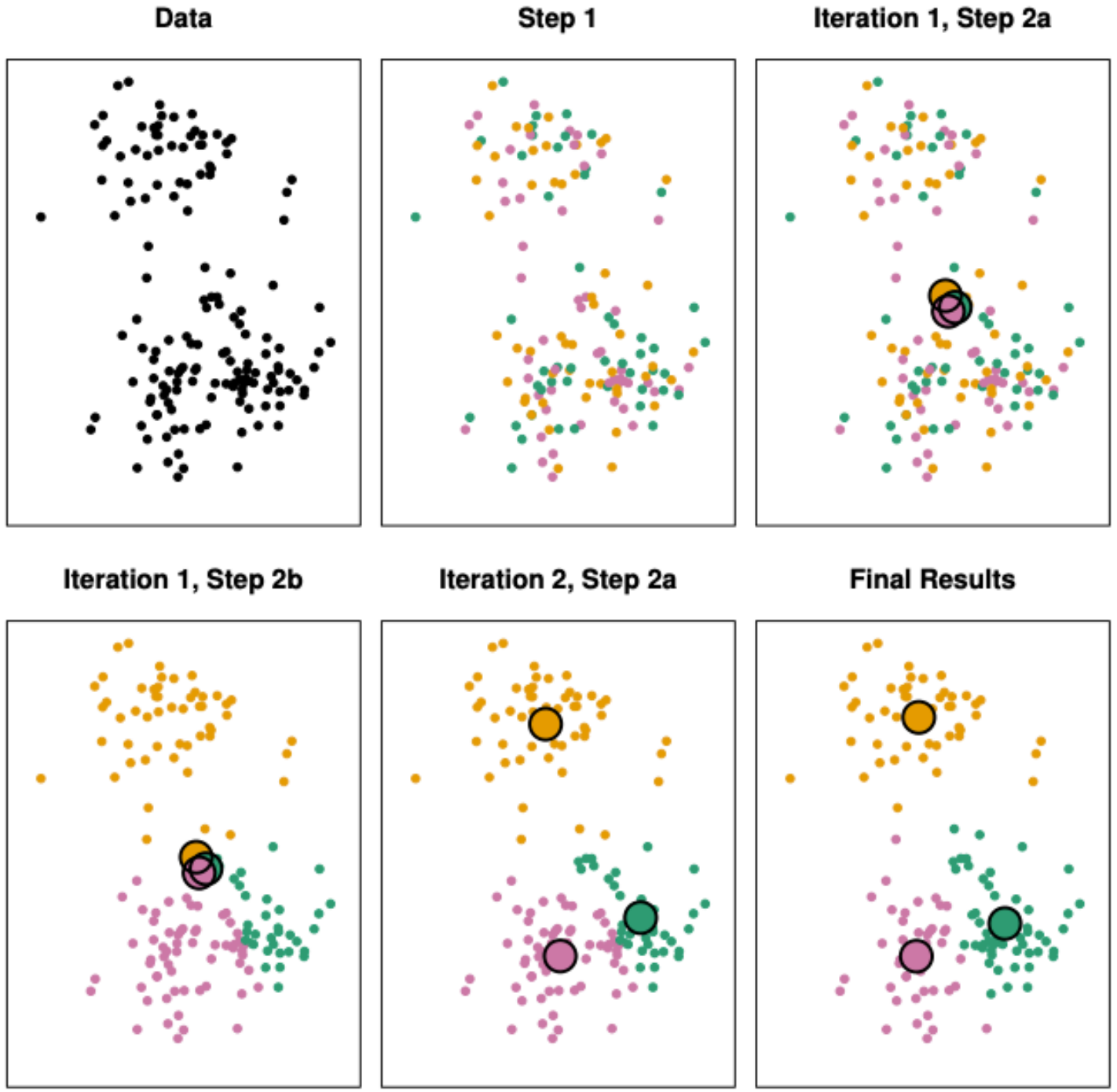
- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- What is within-cluster variation?
  - Typically, we define it using Euclidean distance

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

  - $|C_k|$ denotes the number of observations in the *kth* cluster

Example

- The progress of the K-means algorithm 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 10 iterations.

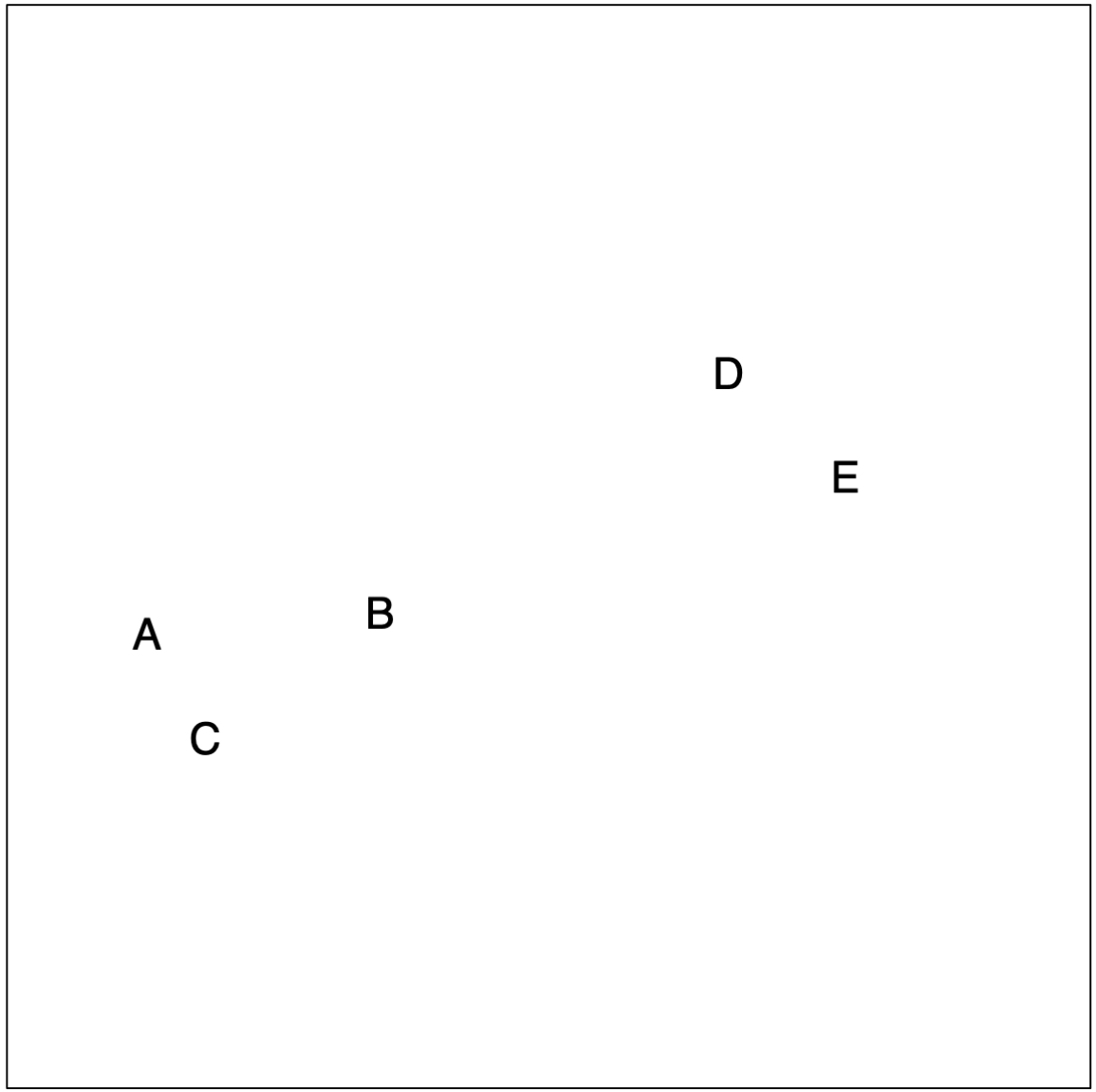


Hierarchical Clustering

- K-means clustering
  - Requires us to pre-specify the number of clusters K
- Hierarchical clustering
  - An alternative approach
  - Does not require that we commit to a particular choice of K.
  - A bottom-up or agglomerative clustering
  - This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

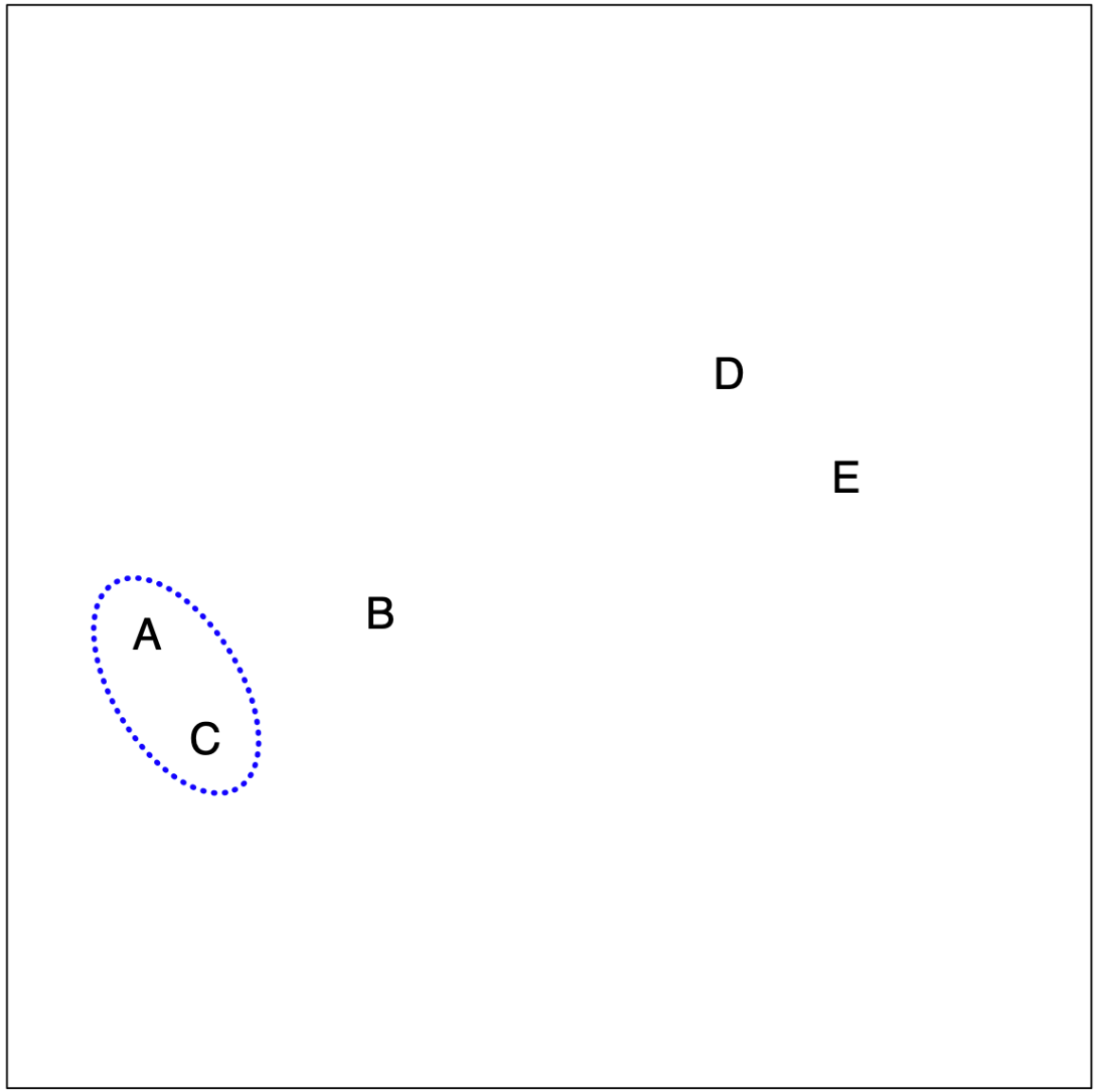
Hierarchical Clustering: the idea

Builds a hierarchy in a “bottom-up” fashion...



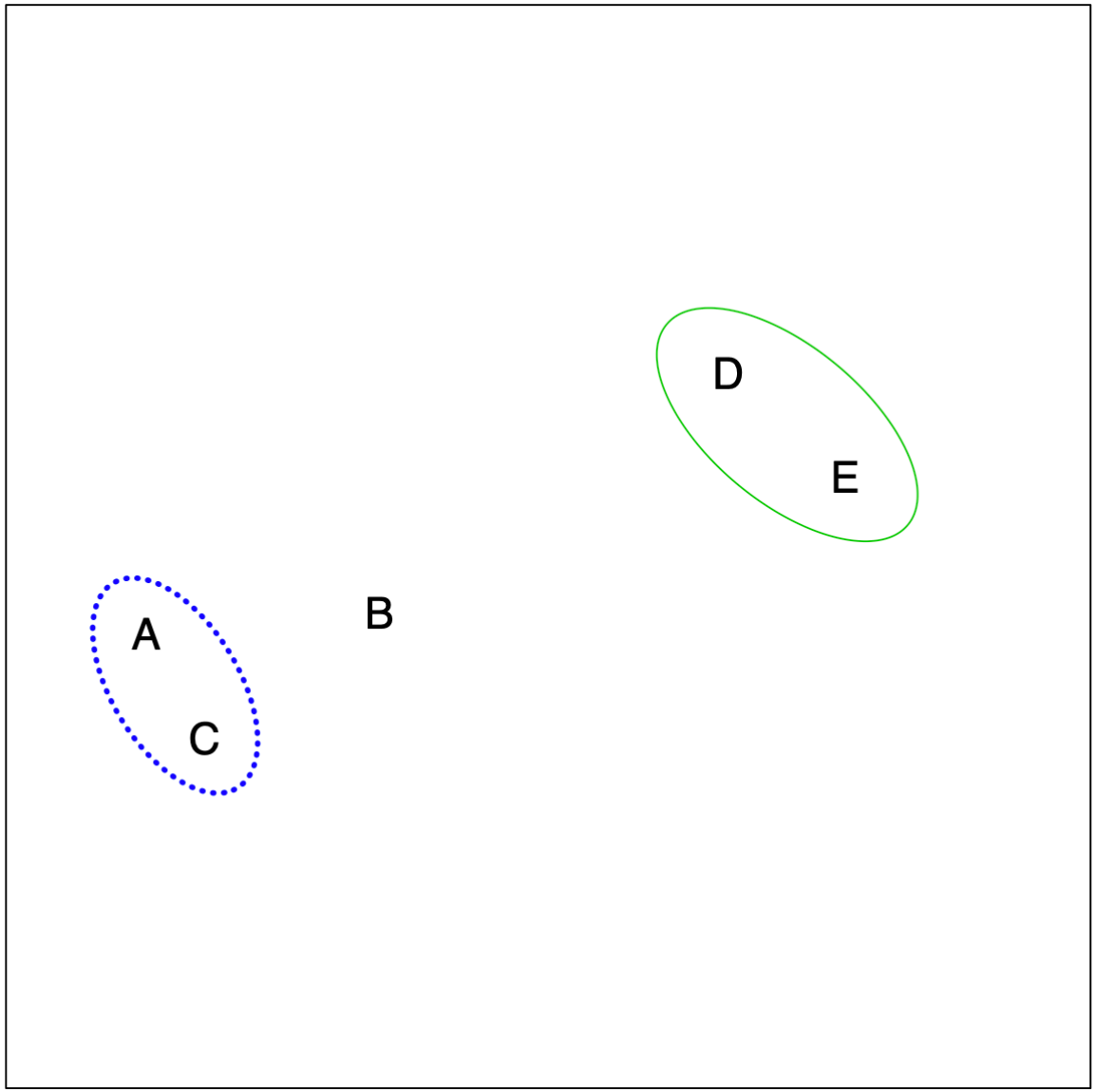
Hierarchical Clustering: the idea

Builds a hierarchy in a “bottom-up” fashion...



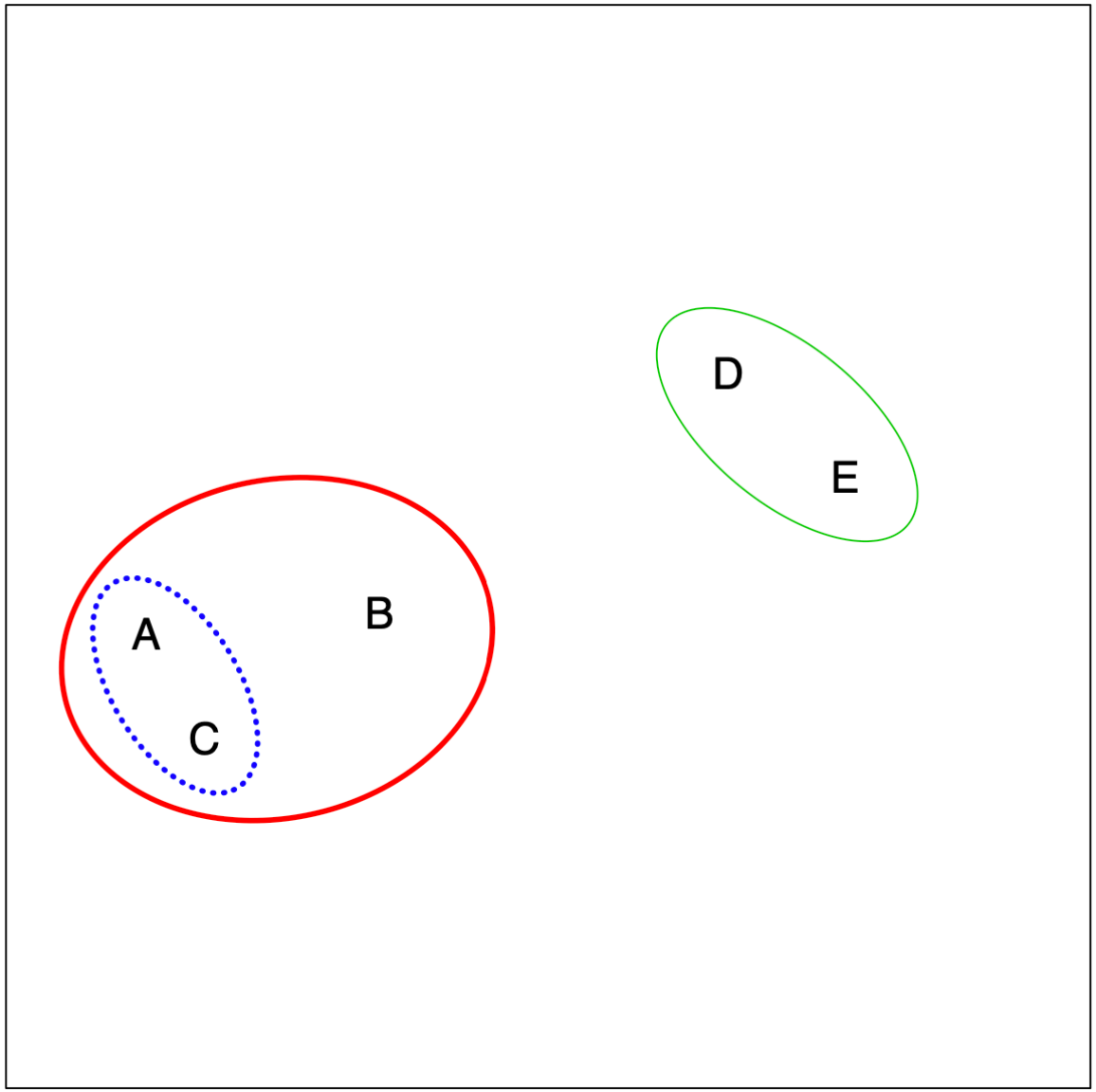
Hierarchical Clustering: the idea

Builds a hierarchy in a “bottom-up” fashion...



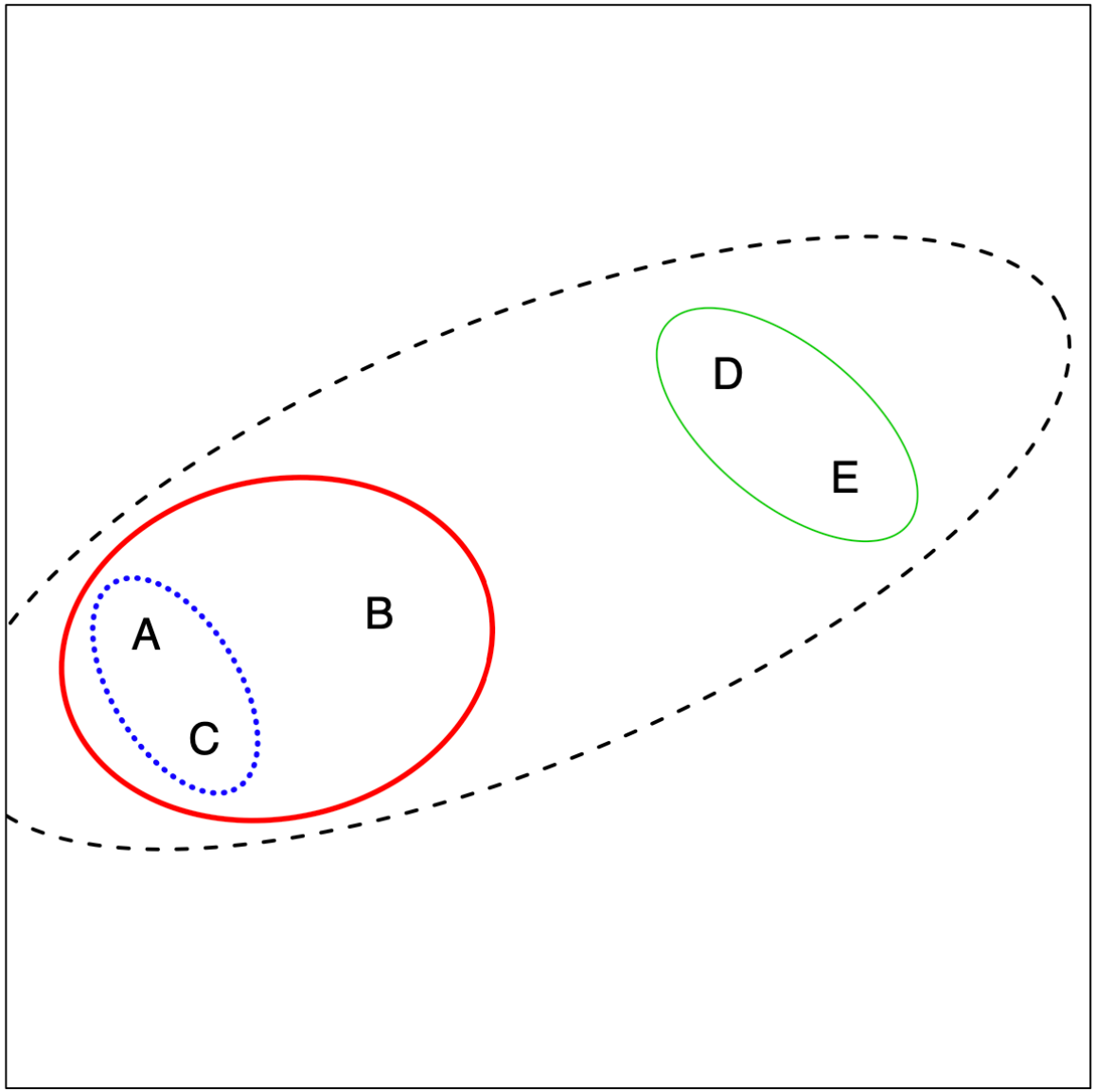
Hierarchical Clustering: the idea

Builds a hierarchy in a “bottom-up” fashion...



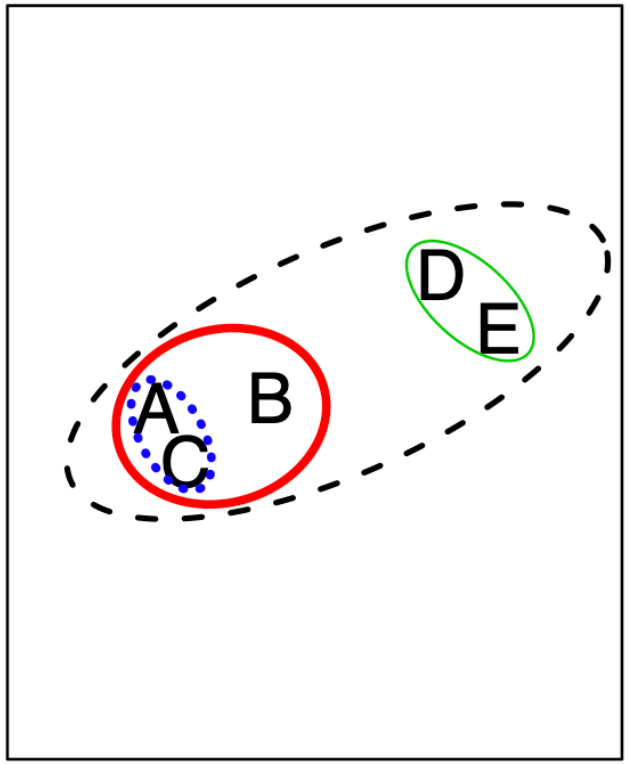
Hierarchical Clustering: the idea

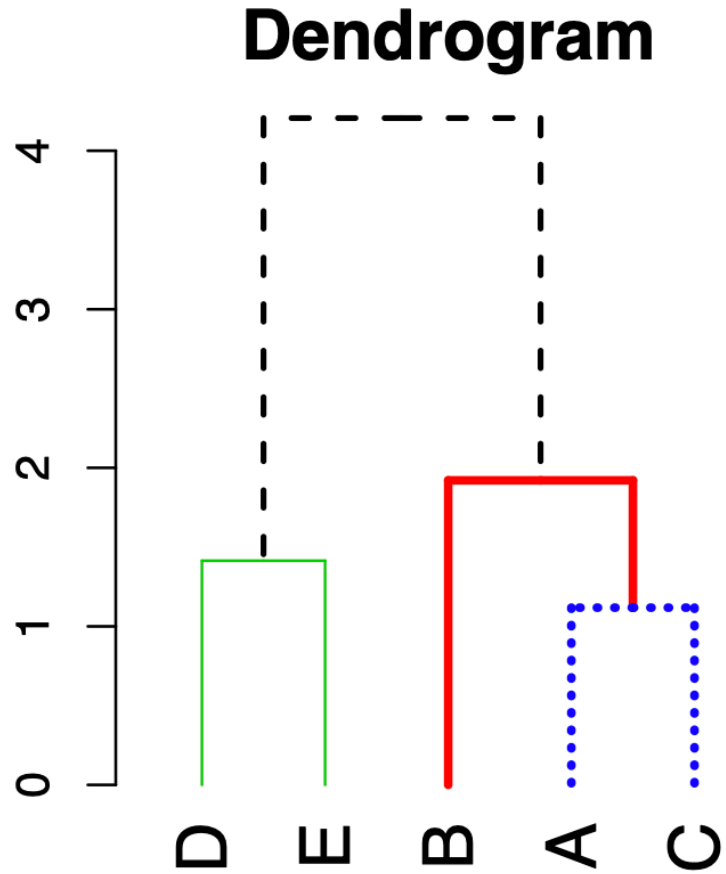
Builds a hierarchy in a “bottom-up” fashion...



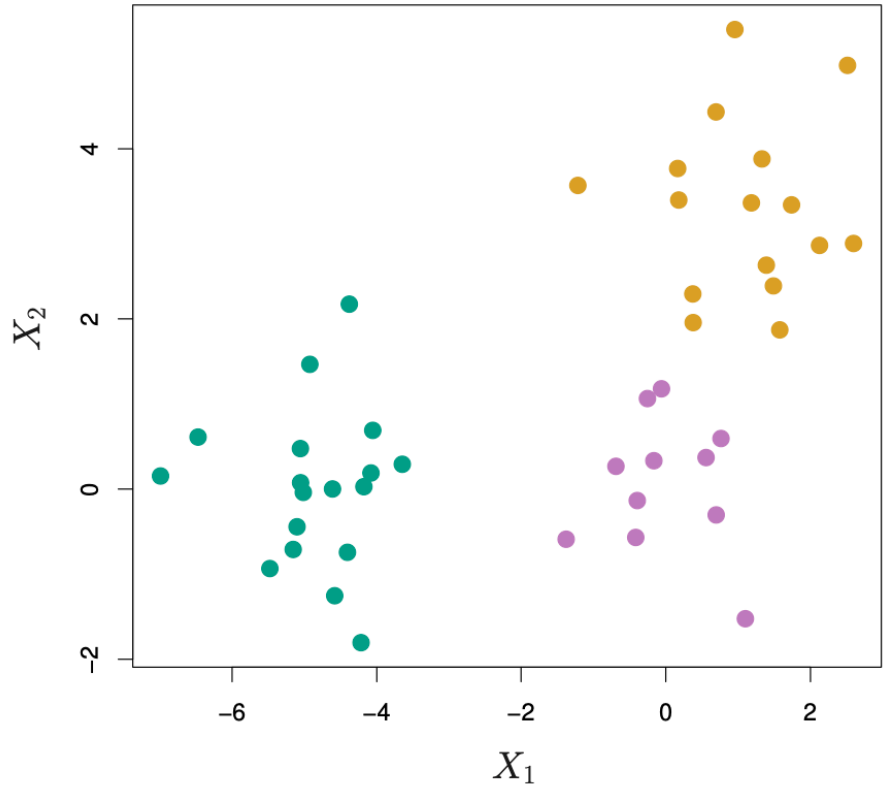
Hierarchical Clustering Algorithm

- The approach in words:
  - Start with each point in its own cluster.
  - Identify the closest two clusters and merge them
  - Repeat.
  - Ends when all points are in a single cluster.



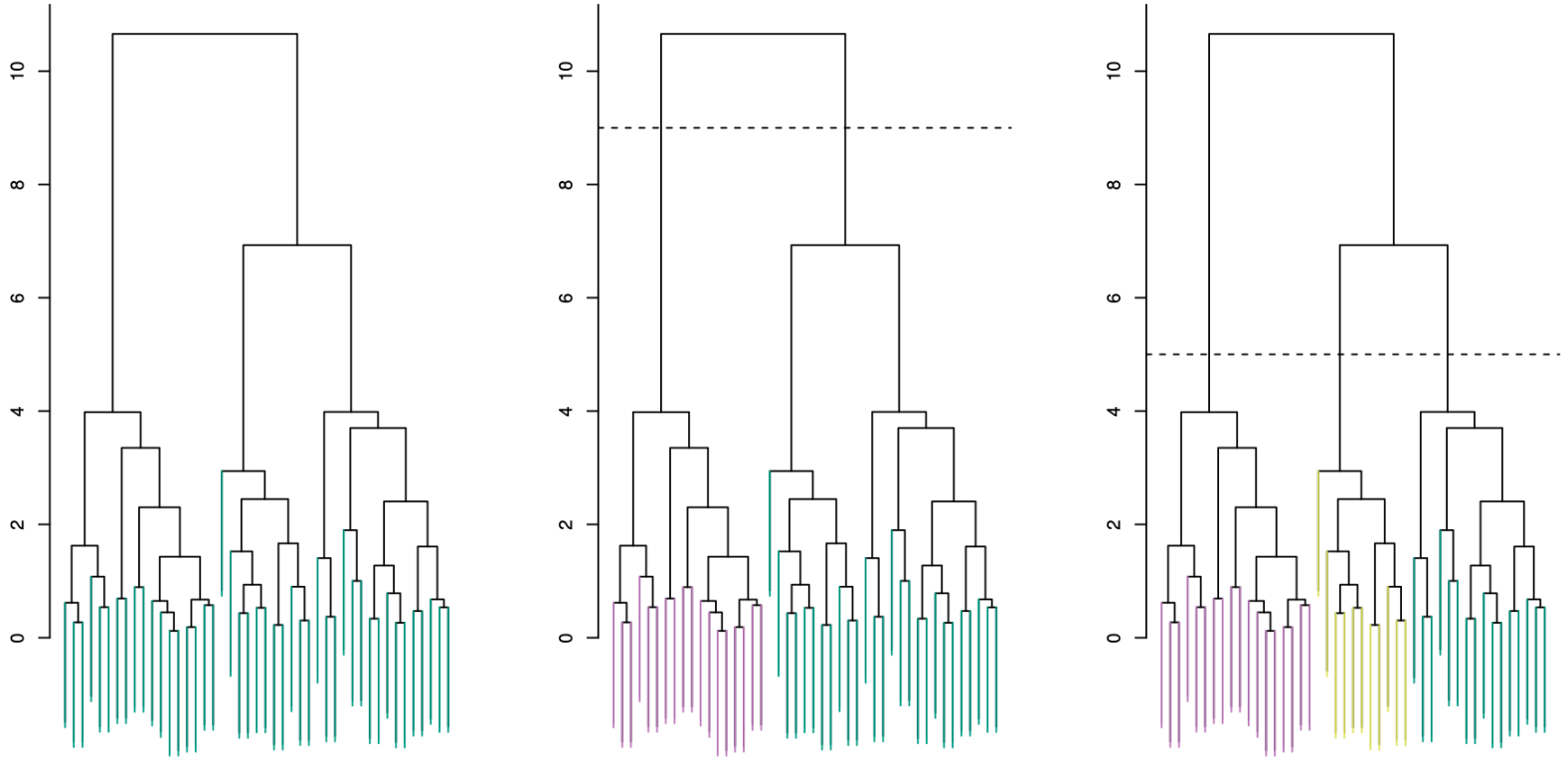


An Example



- 45 observations generated in 2-dimensional space
- In reality there are three distinct classes, shown in separate colors
- However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.

Application of hierarchical clustering



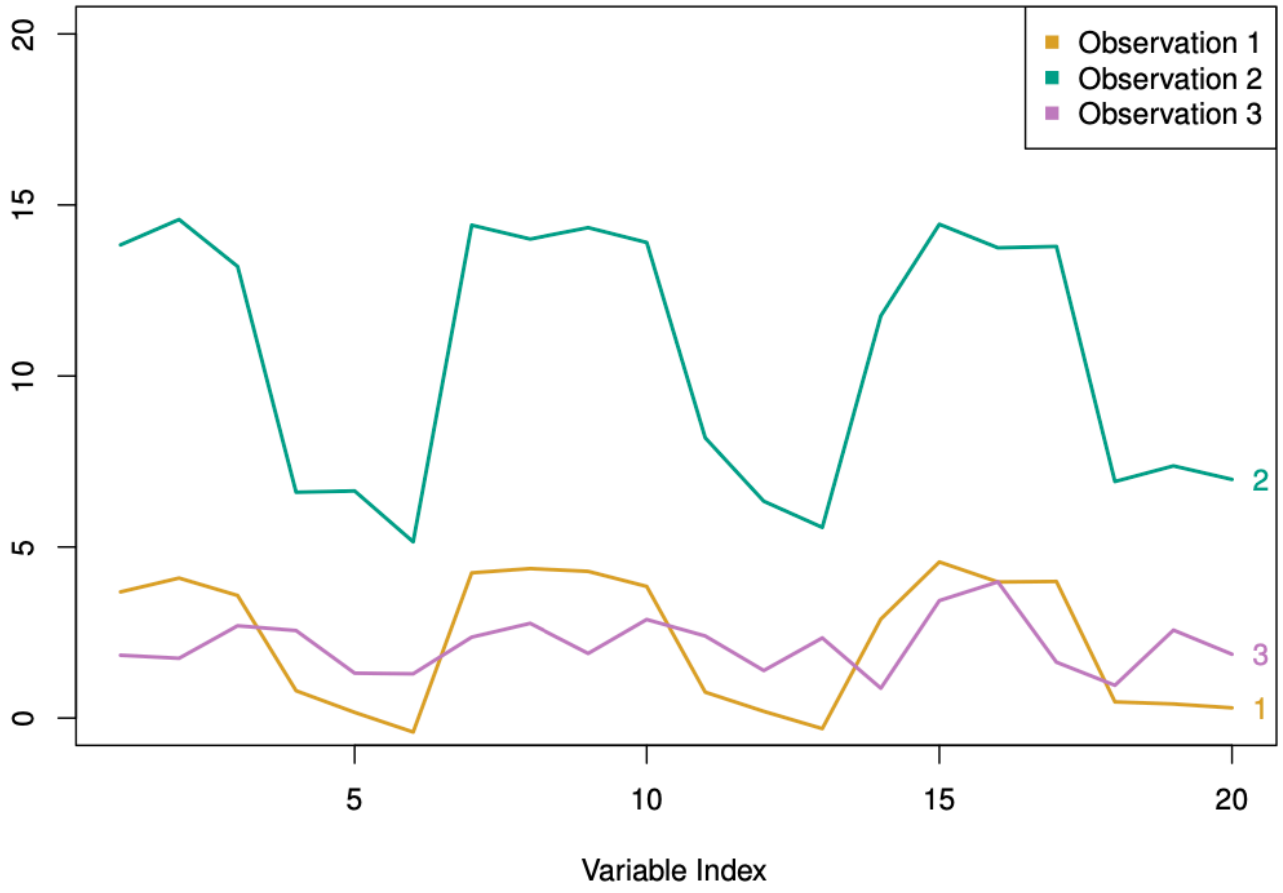
- Left: Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- Center: The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
- Right: The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure

Types of Linkage

Linkage	*Description*
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

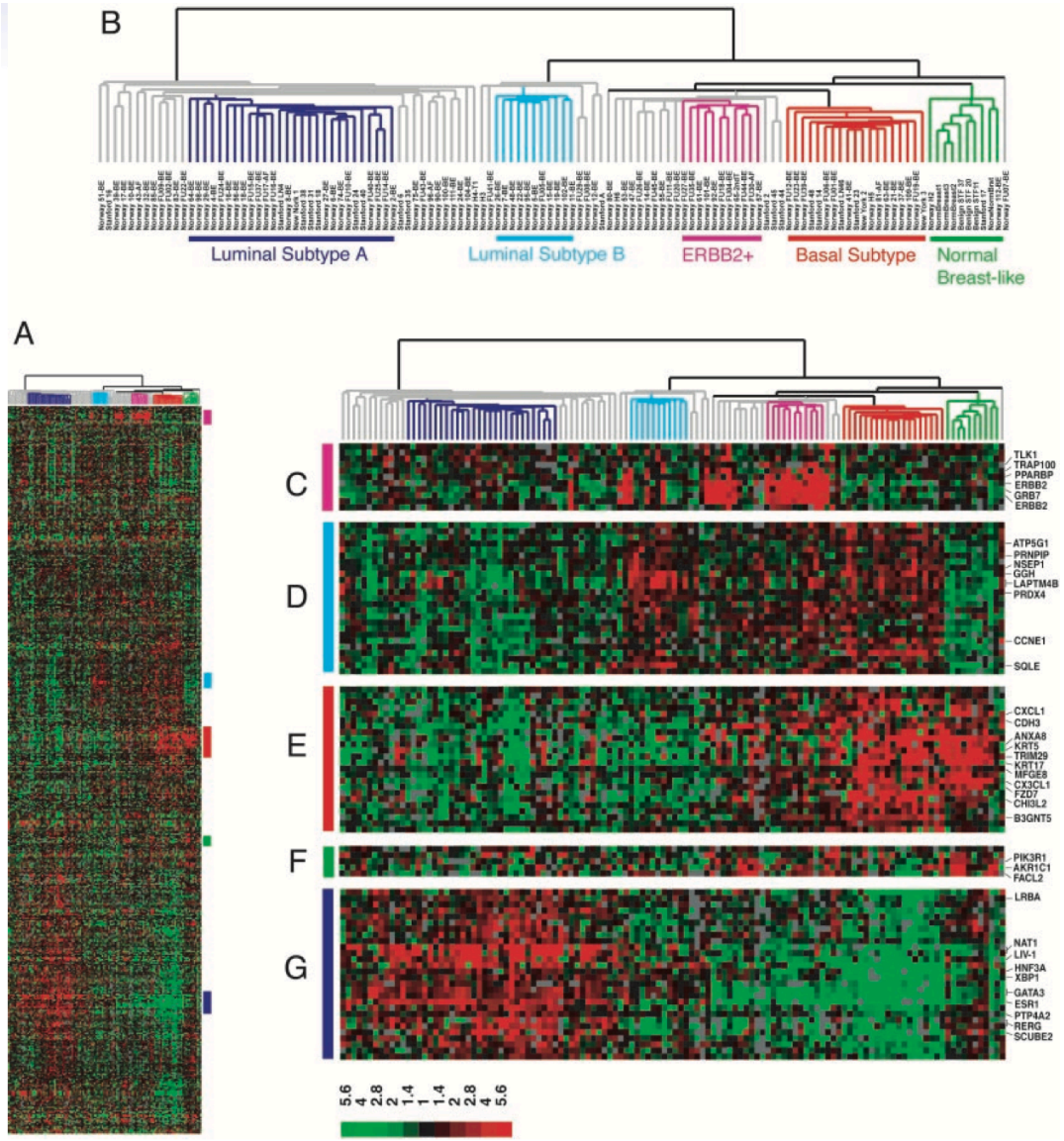
Choice of Dissimilarity Measure

- So far have used Euclidean distance.
- An alternative is correlation-based distance which considers two observations to be similar if their features are highly correlated.
- This is an unusual use of correlation, which is normally computed between variables; here it is computed between the observation profiles for each pair of observations.



Example: breast cancer microarray study

- "Repeated observation of breast tumor subtypes in independent gene expression data sets;" Sorlie at el, PNAS 2003
- Average linkage, correlation metric
- Clustered samples using 500 intrinsic genes: each woman was measured before and after chemotherapy. Intrinsic genes have smallest within/between variation.

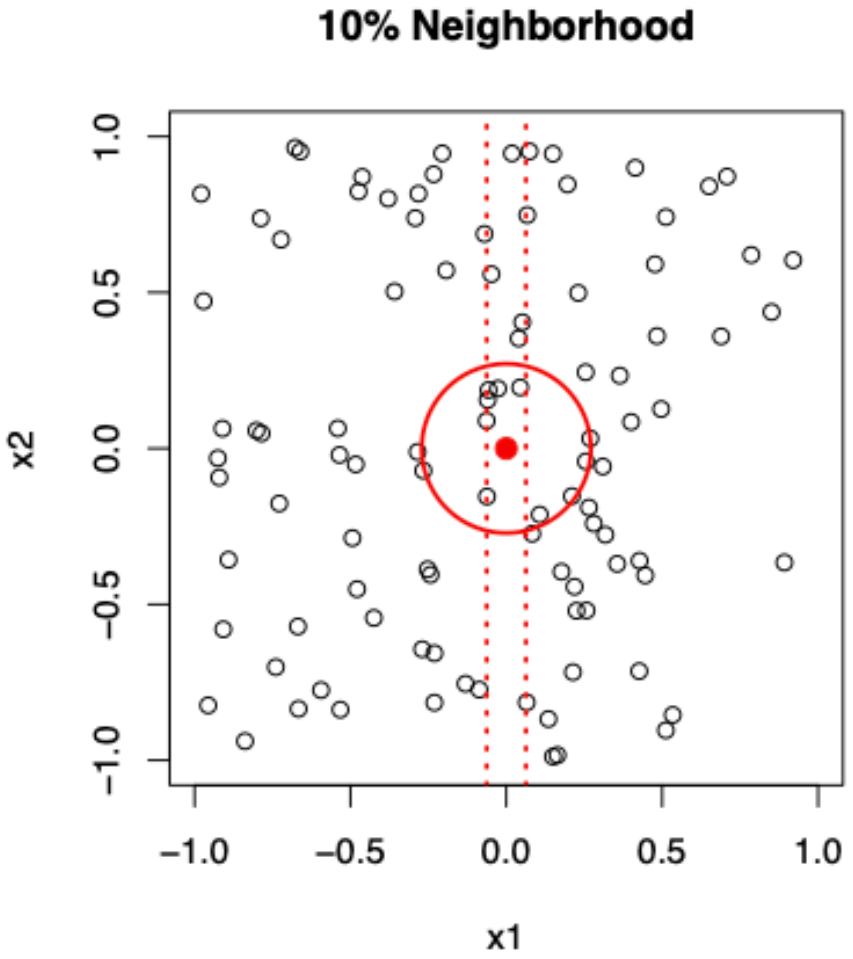


Practical issues

- Should the observations or features first be standardized in some way?
  - Variables centered to have mean zero?
  - Scaled to have standard deviation one?
- In the case of hierarchical clustering
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
- How many clusters to choose? (in both K-means or hierarchical clustering). Difficult problem. No agreed-upon method.

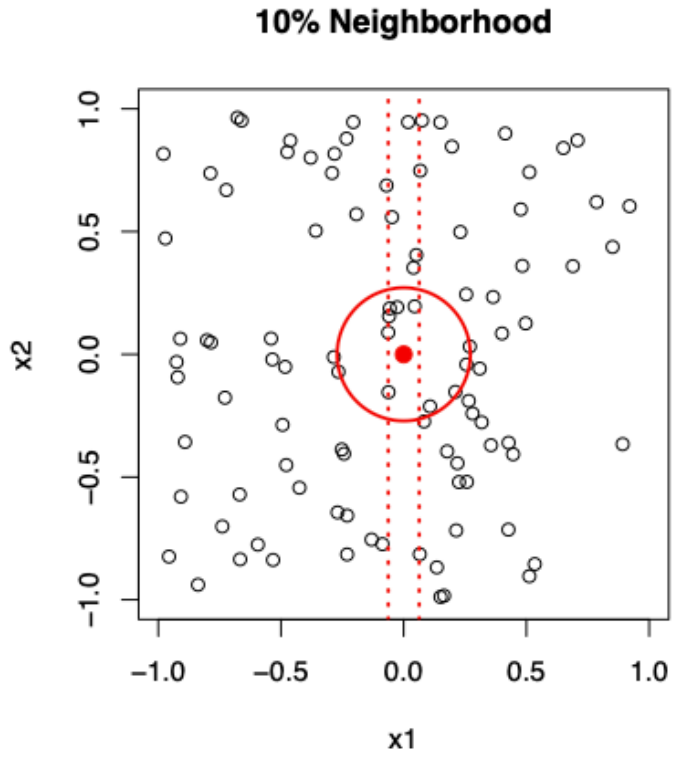
The Curse of Dimensionality

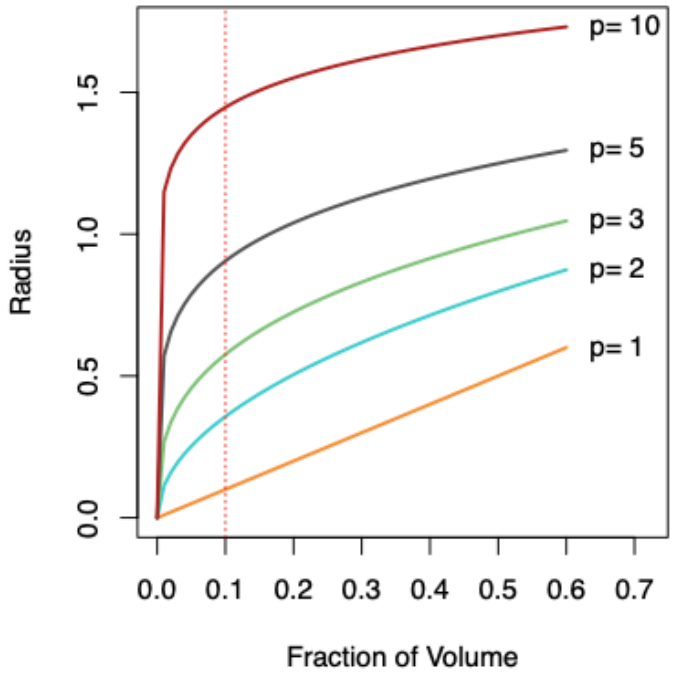
- We have points uniformly distributed throughout a series of regions with increasing dimensions.
- Given a n-dimensional sphere centered in the space
  - How large does the radius of the sphere need to be to capture 10% of the points in the region?



The Curse of Dimensionality

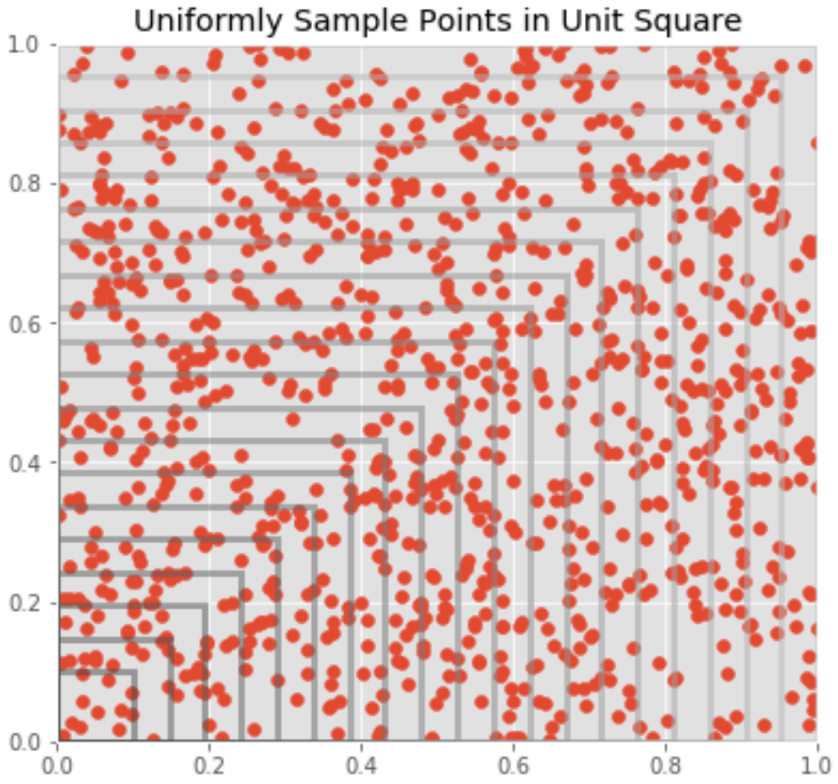
- What about when we move into 3-, 5-, and 10-dimensional space?
  - What is happening to the radius?

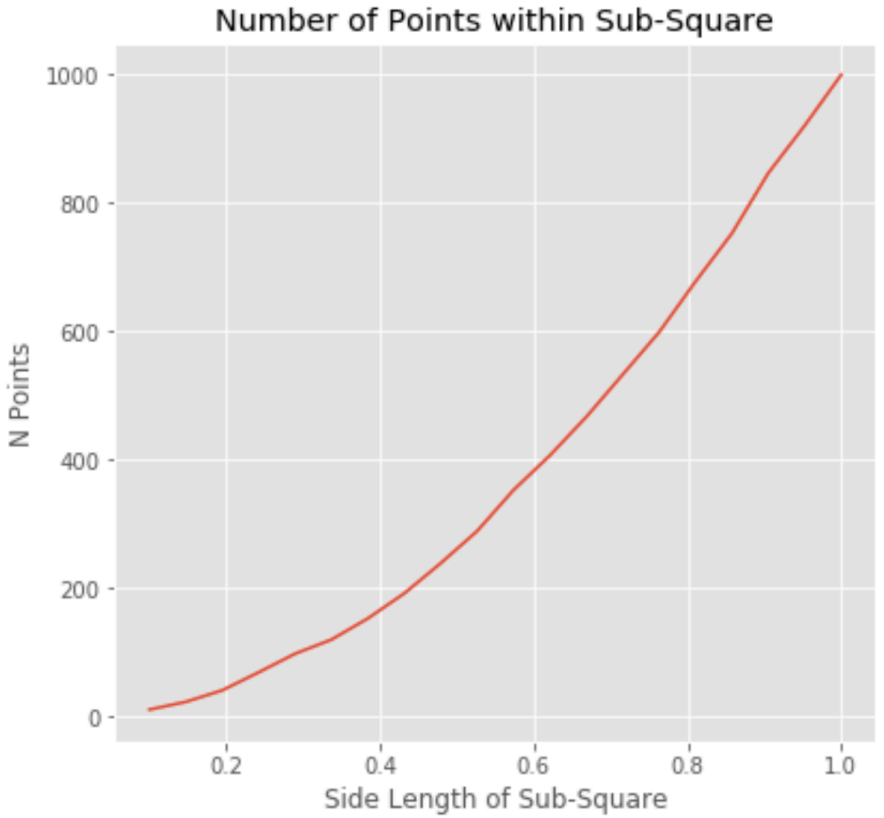




A Different Illustration

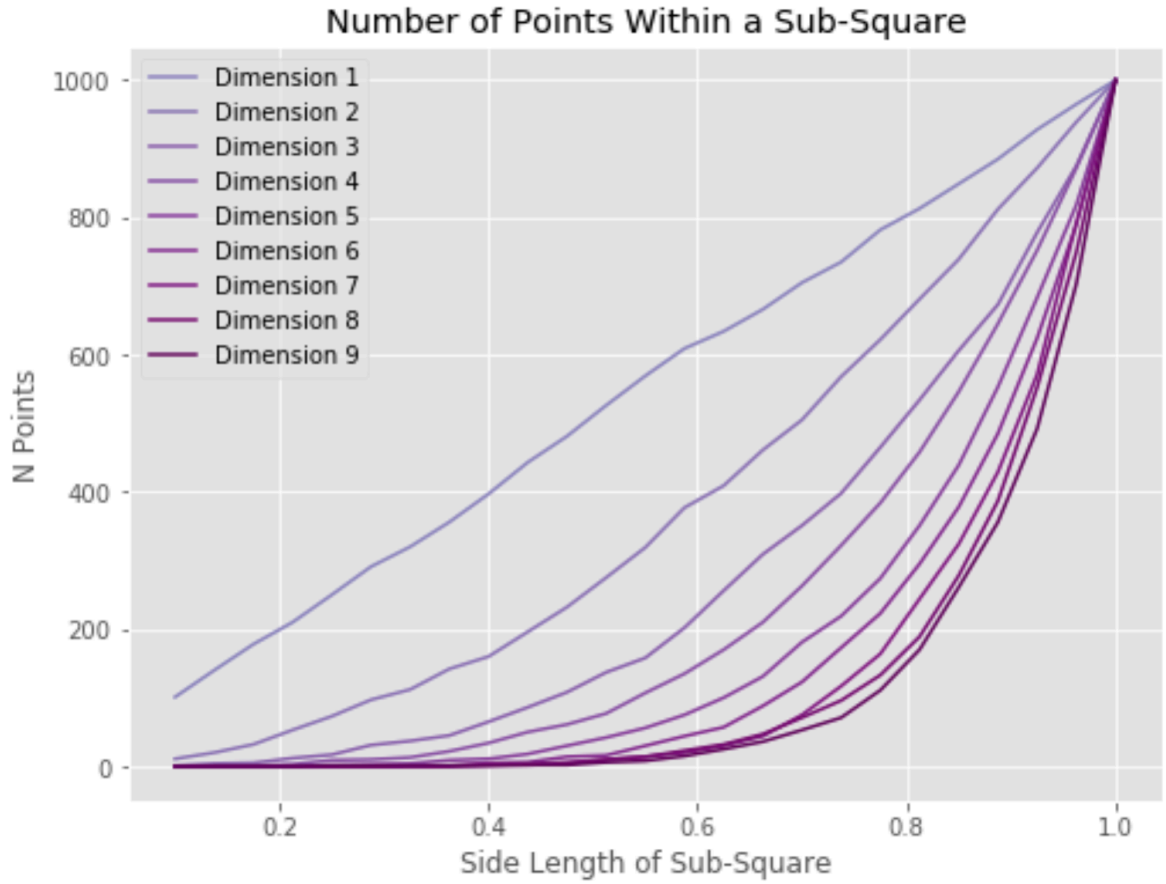
- Here, we're focusing on a sub-square in 2-dimensions





A Different Illustration

- What happens to the number of points within a sub-square at different sizes as we increase dimensions?



The Curse of Dimensionality

- End result is that nearby points tend to be far away in high dimensions!
- This makes clustering (and any other technique based on point closeness) much harder!
- In most cases, we will want some way to reduce the number of dimensions if we are going to use clustering
- How about…

Principal Component Analysis (PCA)

- Visualization
  - Natural to visualize 2 and 3 dimensions
  - 4 and 5 dimensions possible but much harder
  - How can we visualize some with hundreds or even hundreds of thousands of dimensions?
- Feature correlation
  - What if we have so many features that are so highly correlated that modeling becomes tremendously difficult?
- How can we know which features are most responsible for the variability in the data? Which ones are least uniform across our points?

PCA

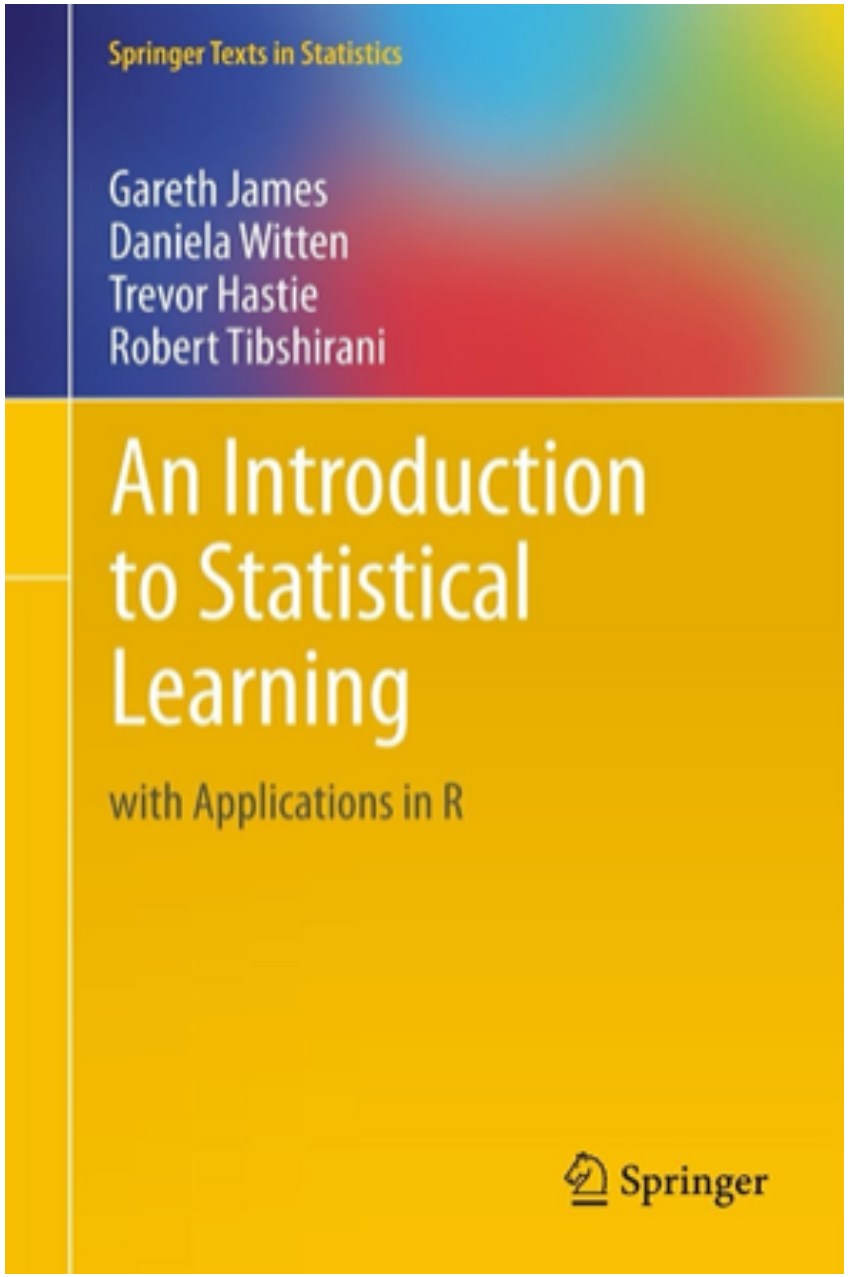
- A classic technique for dimensionality reduction
- Commonly used for
  - Data visualization
  - Data pre-processing before supervised learning techniques are applied

What is PCA?

- PCA is a technique that produces a lower-dimensional representation of a dataset
- It finds a sequence of linear combinations of the variables which
  - Have maximal variance
  - Are mutually uncorrelated.
- Can reduce a large set of variables to a small set that still contains most of the information in the large set.

PCA

- We won't go deeply into the math, but for an excellent exploration of PCA (and all other things Machine Learning), check out An Introduction to Statistical Learnining with Applications in R.

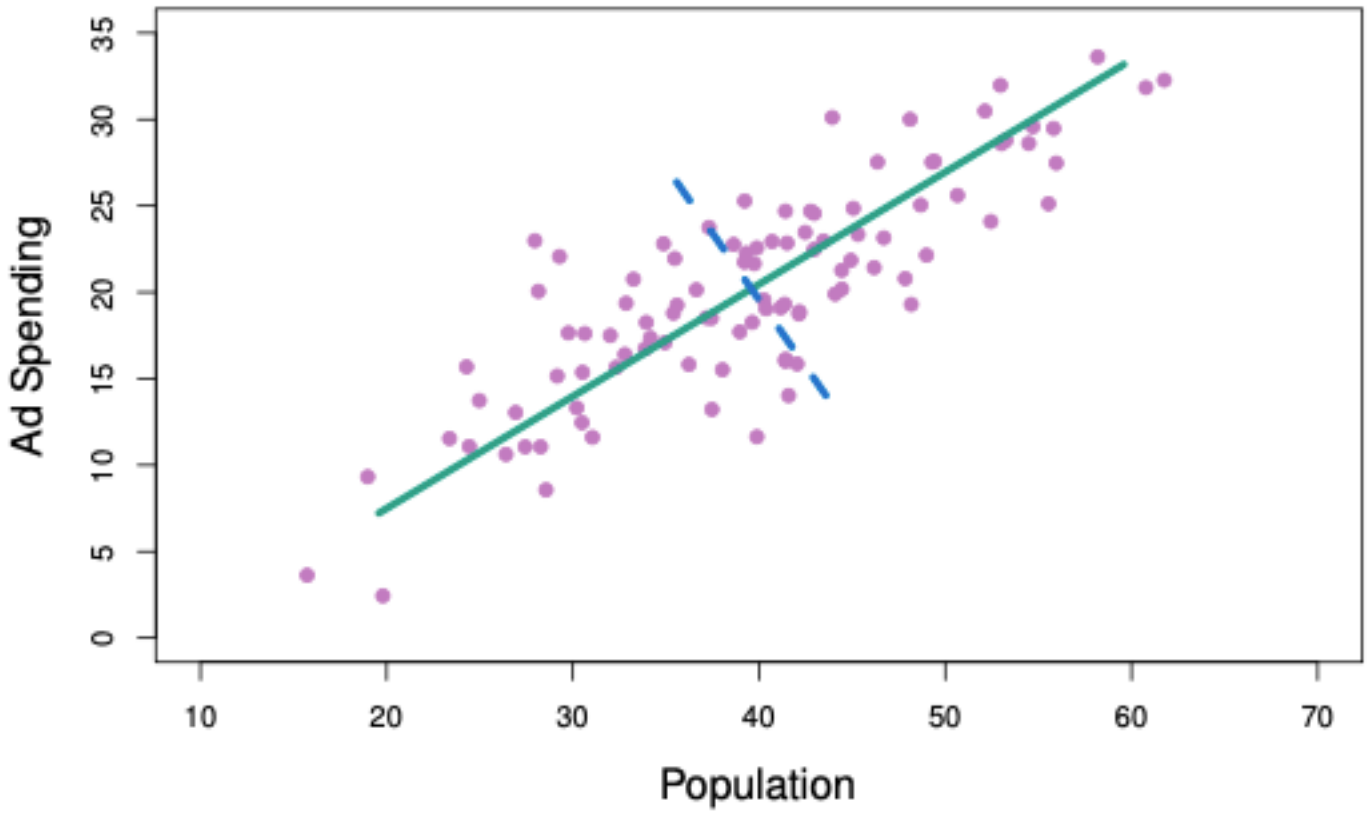


Significant portions of this lesson were adapted from this text

PCA example

- The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles

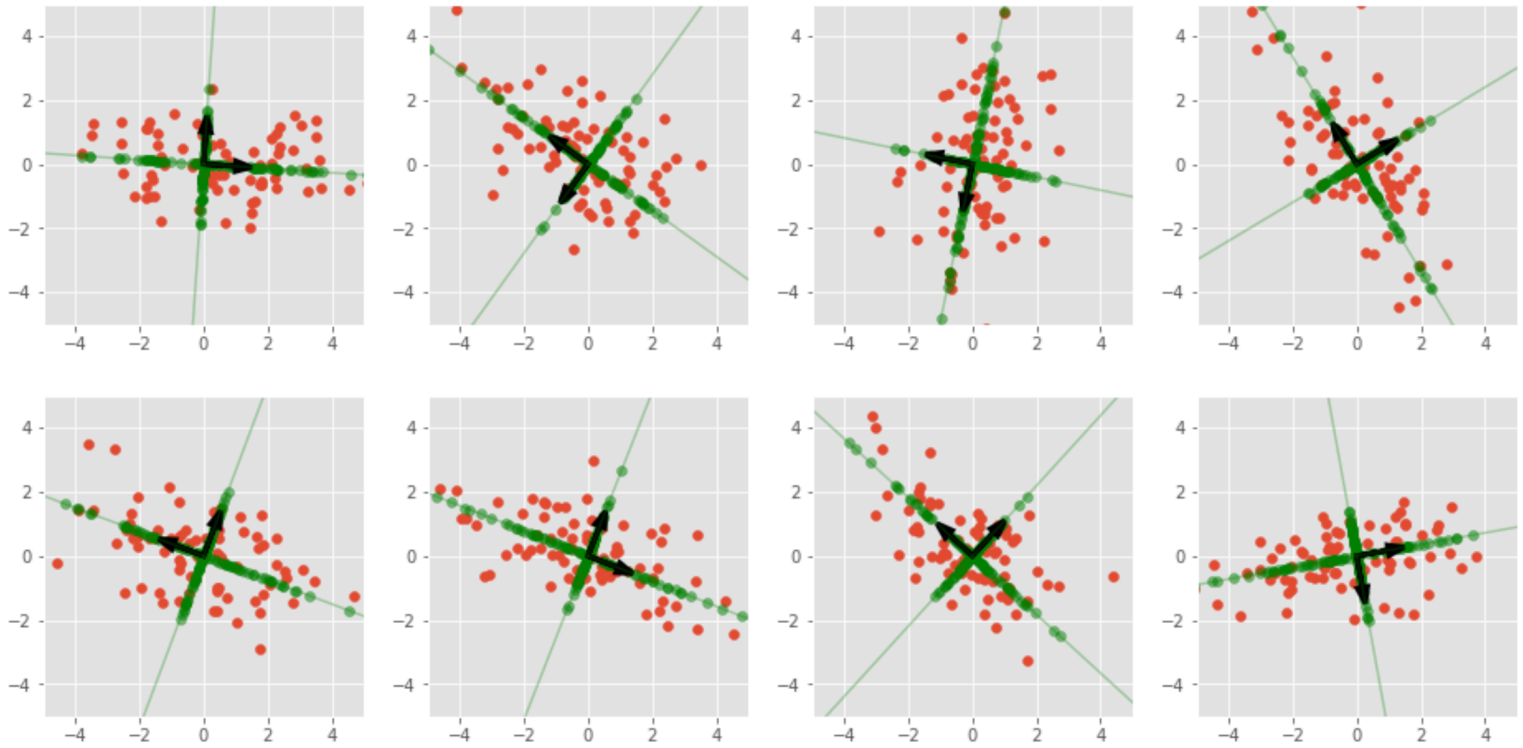
- The green solid line indicates the first principal component direction
- The blue dashed line indicates the second principal component direction.



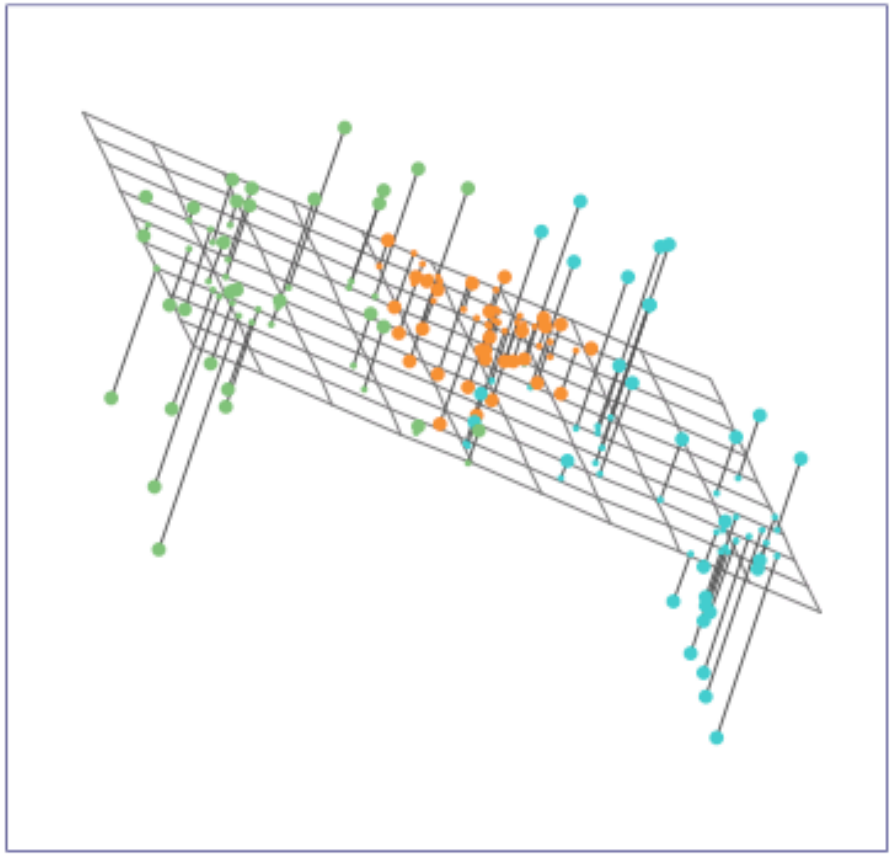
PCA example

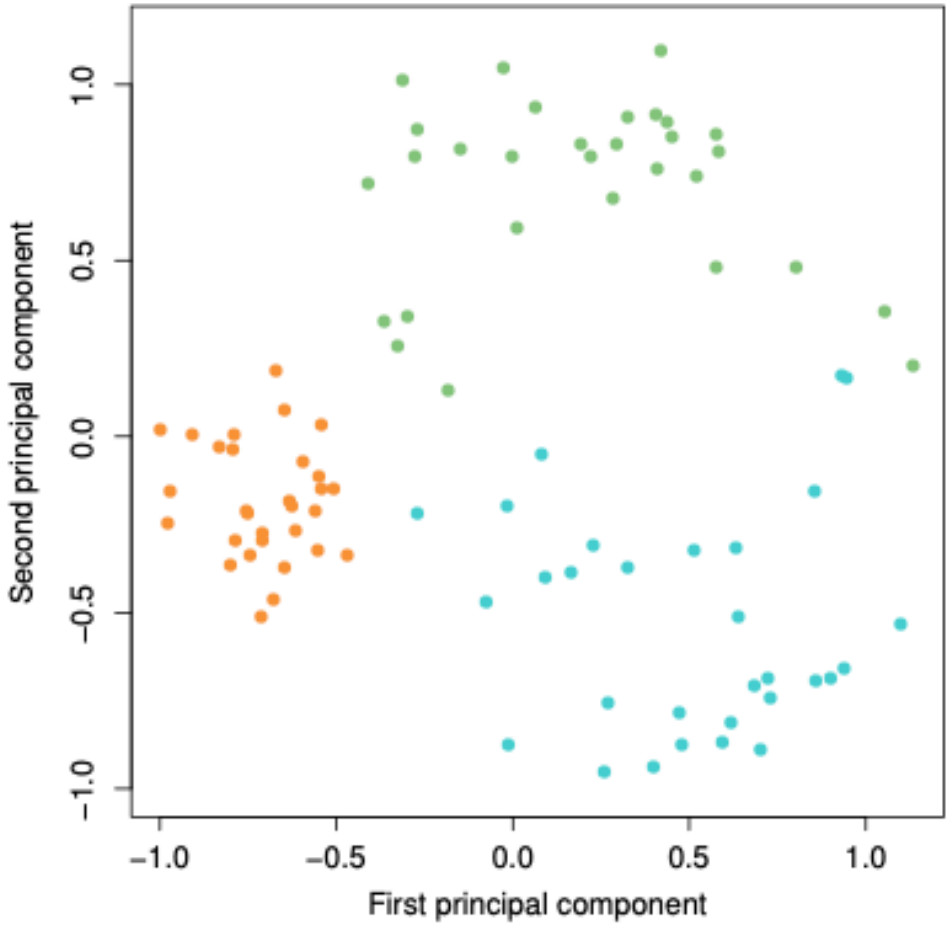
- In the below plots, we superimpose the first and second principal components onto a random scatter plot

- The **first principal component** determines a green line that maximizes the variance of the data's projection.
- The **second principal component** is orthogonal to the first, and maximizes the projection of the "leftover" data.



Another Interpretation of Principal Components

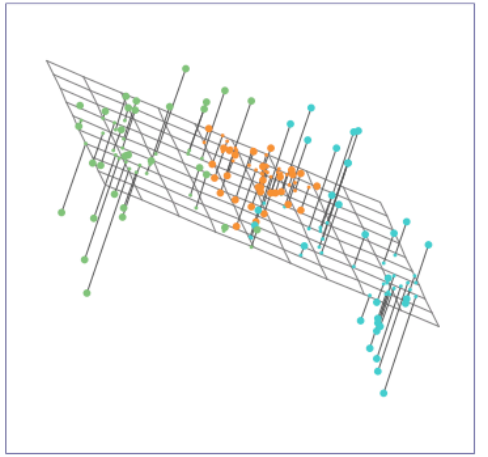


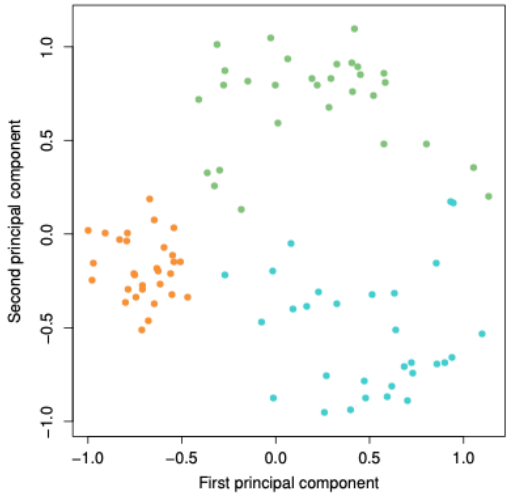


Another Interpretation of Principal Components

- The first principal component loading vector has a very special property: it defines the line in p-dimensional space that is closest to the n observations (using average squared Euclidean distance as a measure of closeness)
- The notion of principal components as the dimensions that are closest to the n observations extends beyond just the first principal component.

- For instance, the first two principal components of a data set span the plane that is closest to the n observations, in terms of average squared Euclidean distance.

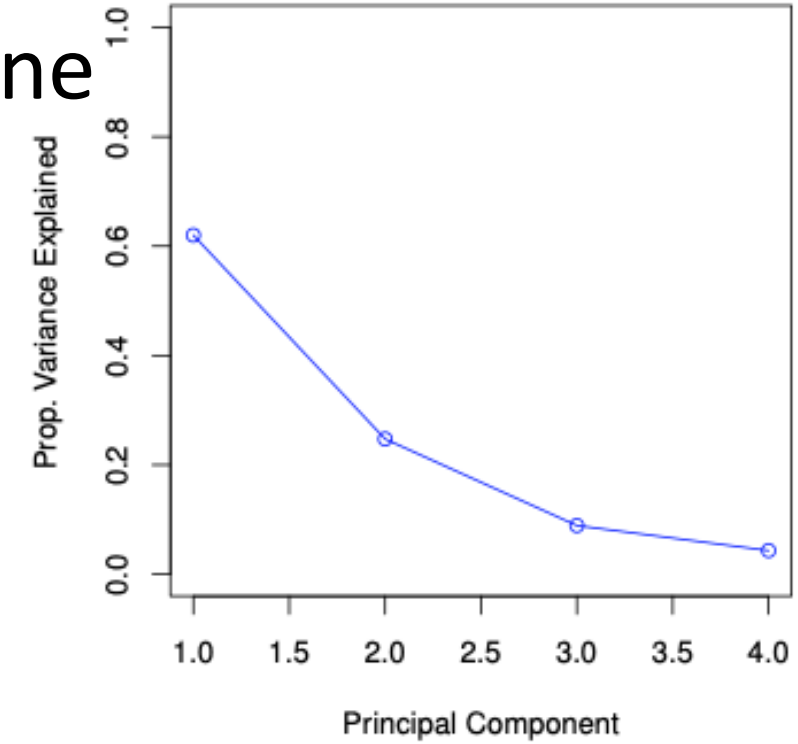


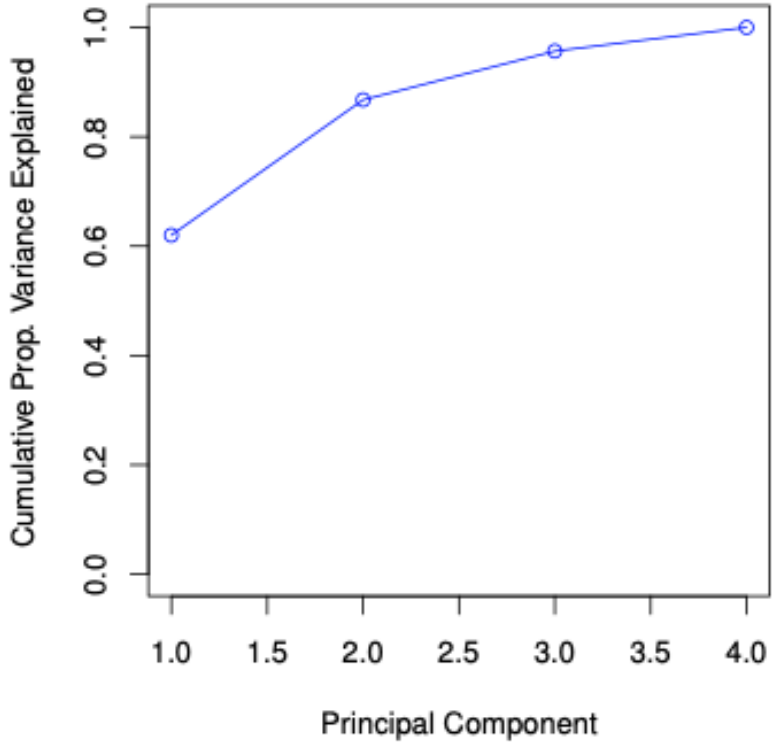


Proportion Variance Explained

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.

- The PVEs sum to one
- We sometimes display the cumulative PVEs.





How many principal components should we use?

- If we use principal components as a summary of our data, how many components are sufficient?
  - No simple answer to this question, as we don't have any simple way of directly evaluating…
  - Why not?
- The "scree plot" on the previous slide can be used as a guide:
  - We look for an "elbow"
  - Many other techniques

Summary: PCA vs Clustering

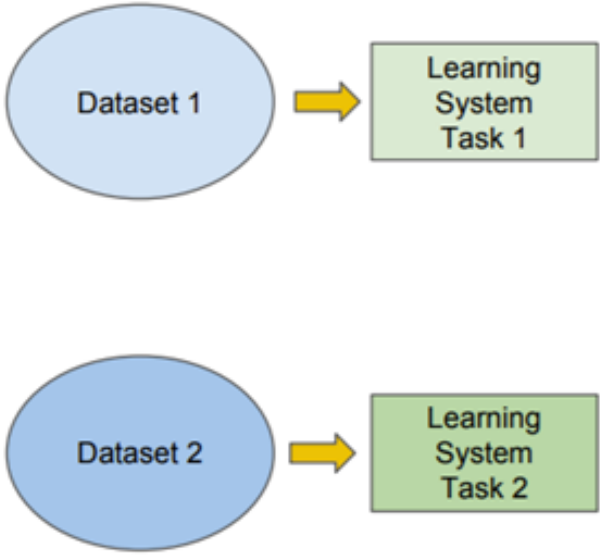
- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

Transfer Learning

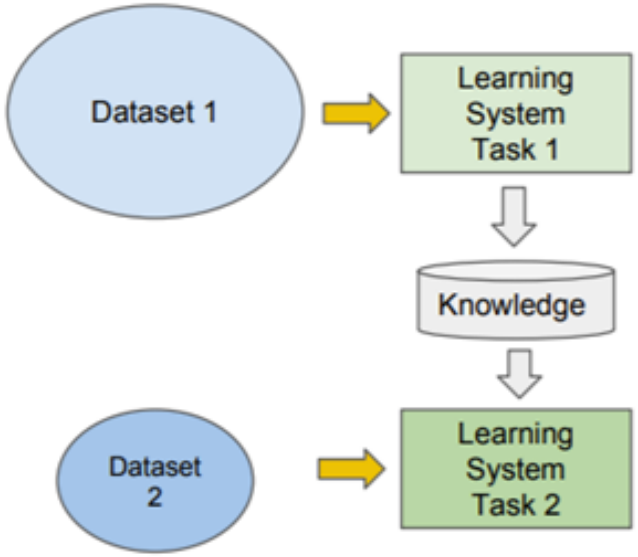
- Transfer knowledge gained while solving one problem to a different but related problem
  - Resembles the human ability to leverage past experience in evaluating new scenarios or topics
- Common pattern linking unsupervised and supervised learning models

Traditional ML vs Transfer Learning

- Traditional ML
  - Isolated, single task learning
  - Knowledge is not retained or accumulated
  - Learning performed without considering past learned knowledge in other tasks



- Transfer Learning
  - Learning of new tasks relies on the previous learned tasks
  - Learning process can be faster, more accurate, and/or need less training data



Transfer Learning with PCA and Clustering

- Specific use will depend on domain and goals
- Principal Component Analysis
  - Original features contributing heavily to first several PCs could lead to supervised learning methods with fewer features
  - Original features contributing heavily to the same PC could suggest potential engineered features for supervised learning
- Clustering
  - Clusters of unlabeled data could help leverage limited labeled data inhabiting the same clusters
  - Clusters of unlabeled data could suggest data subsetting to subsequently model specific regions of the data