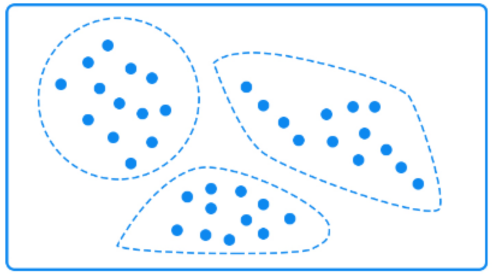
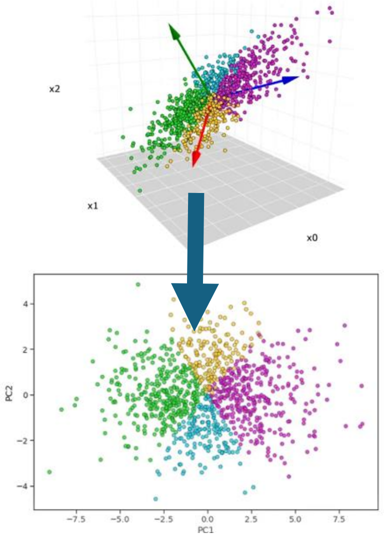
Module 5: Unsupervised Learning

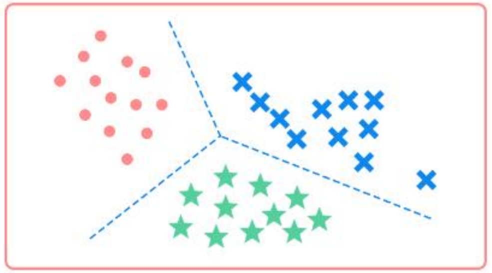
A smattering of options: PCA, permutations, bootstrap

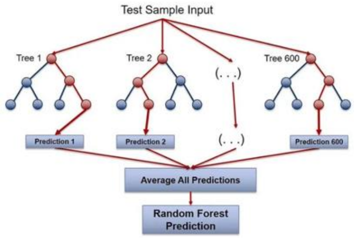
Machine Learning

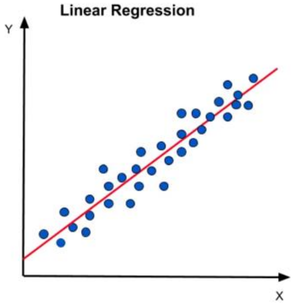
- Unsupervised Learning
  - Clustering
  - Dimension Reduction
  - Association
- Supervised Learning
  - Classification
  - Regression

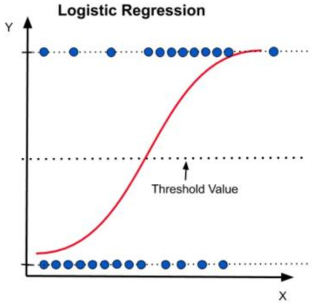








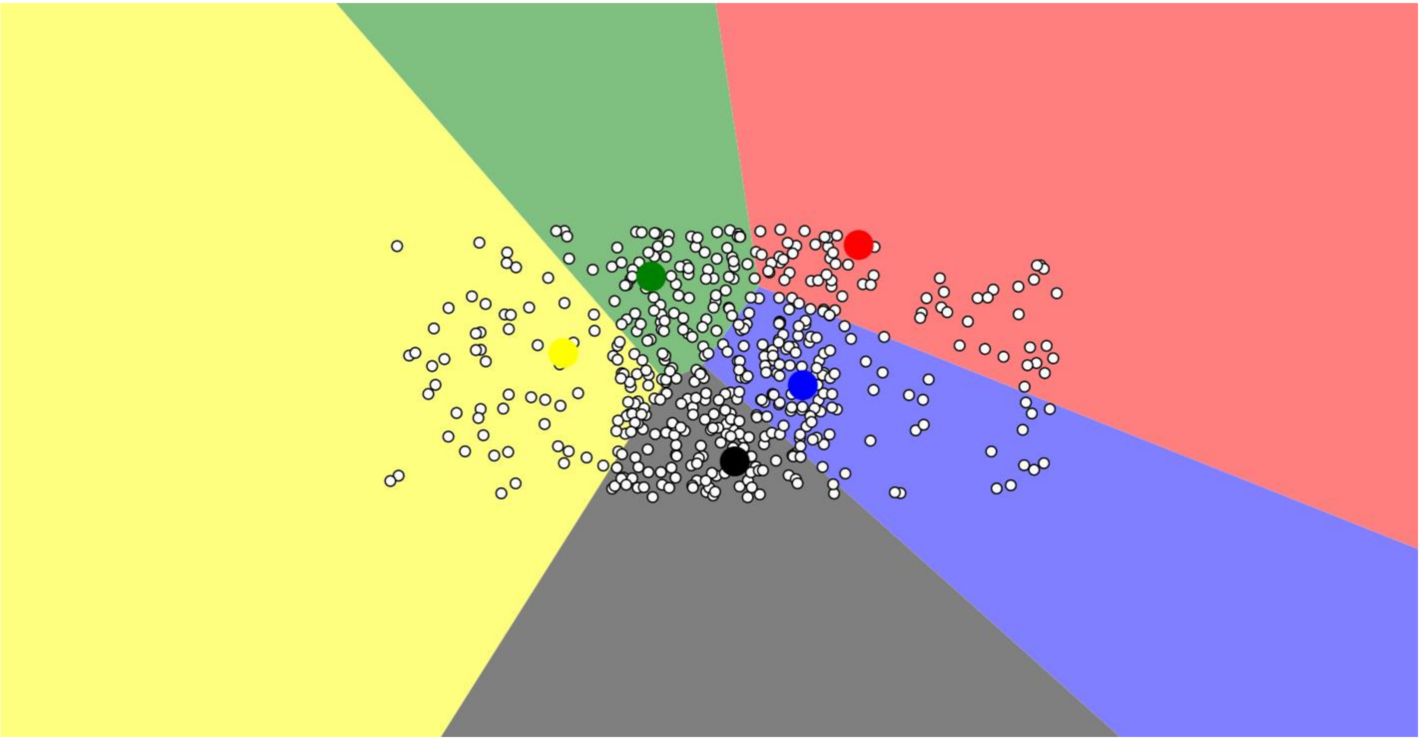




K-means clustering algorithm

Clusters group data points together that share similarities

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



Principal **C**omponent **A**nalysis (PCA)

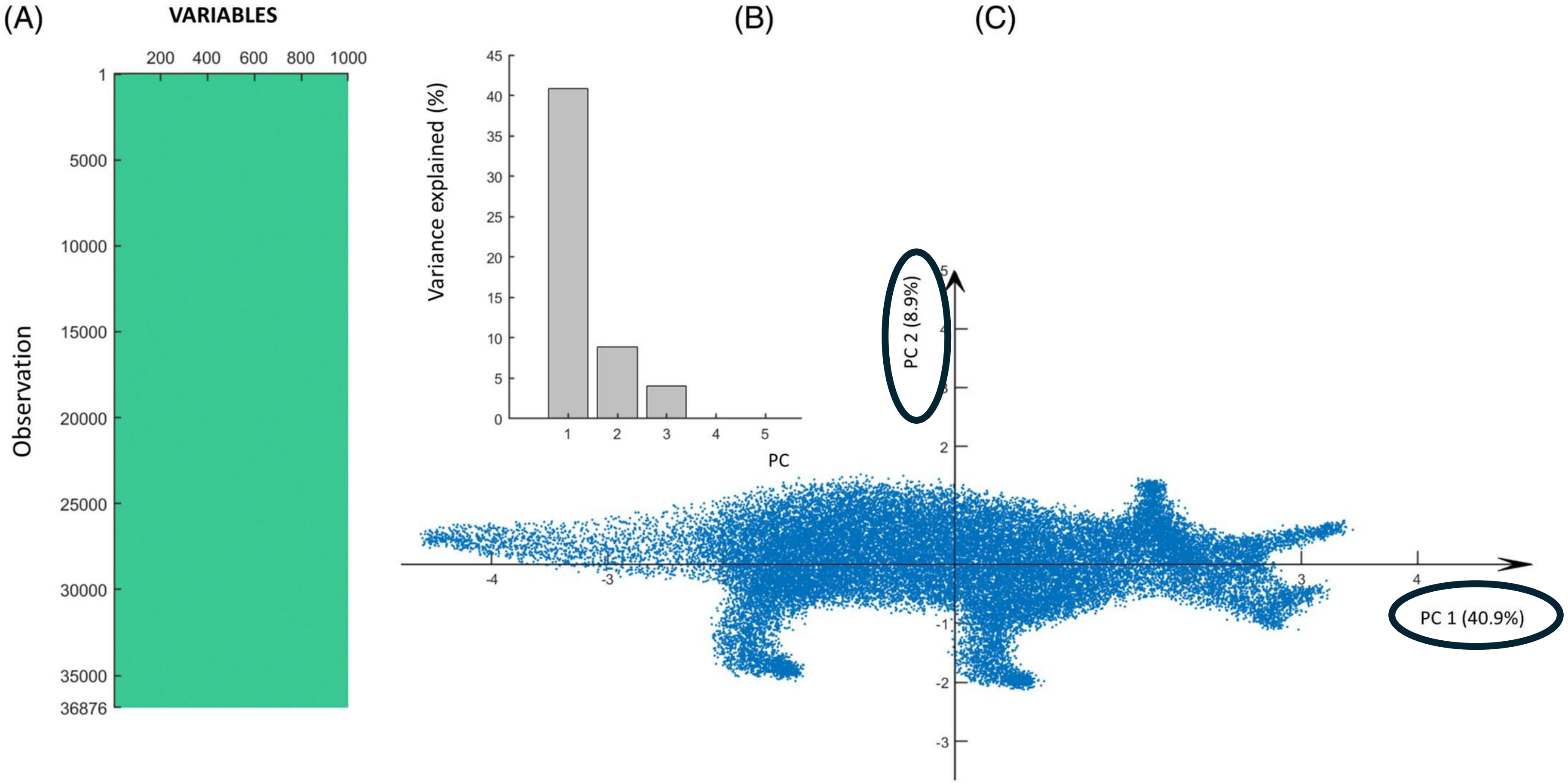
Identifies the major drivers of variation

Why PCA:

- Very few assumptions
- Non-parametric
- It **reduces** the dimensionality of your data
  - It may be surprising to you that you can reduce the dimensionality of your data without losing much information.
  - This occurs when the **variables are highly correlated**.
    - If you have included the following variables in your data set: arm length, leg length, height, you probably don't need them all – a linear combination of the three of them would capture the variation.
    - You can then use a smaller dataset of uncorrelated characteristics (or a smaller set of linear combinations of characteristics)
- Pearson, 1901 (yes, it is > 100 years old).

Principal Component Analysis (PCA)

Revisiting example, but with PCA:



Two major categories of computational methods

Null sampling distributions:

1. Simulation – hypothesis testing

2. Randomization/Permutation

Precision of estimates:

3. Bootstrapping – sampling distribution of estimate; the values for the parameter estimates that we might obtain and their probabilities.

Bootstrapping:

- ‘re-sampling’ the actual data
  - **Sampling with replacement**
  - Pick the original number of points for each group

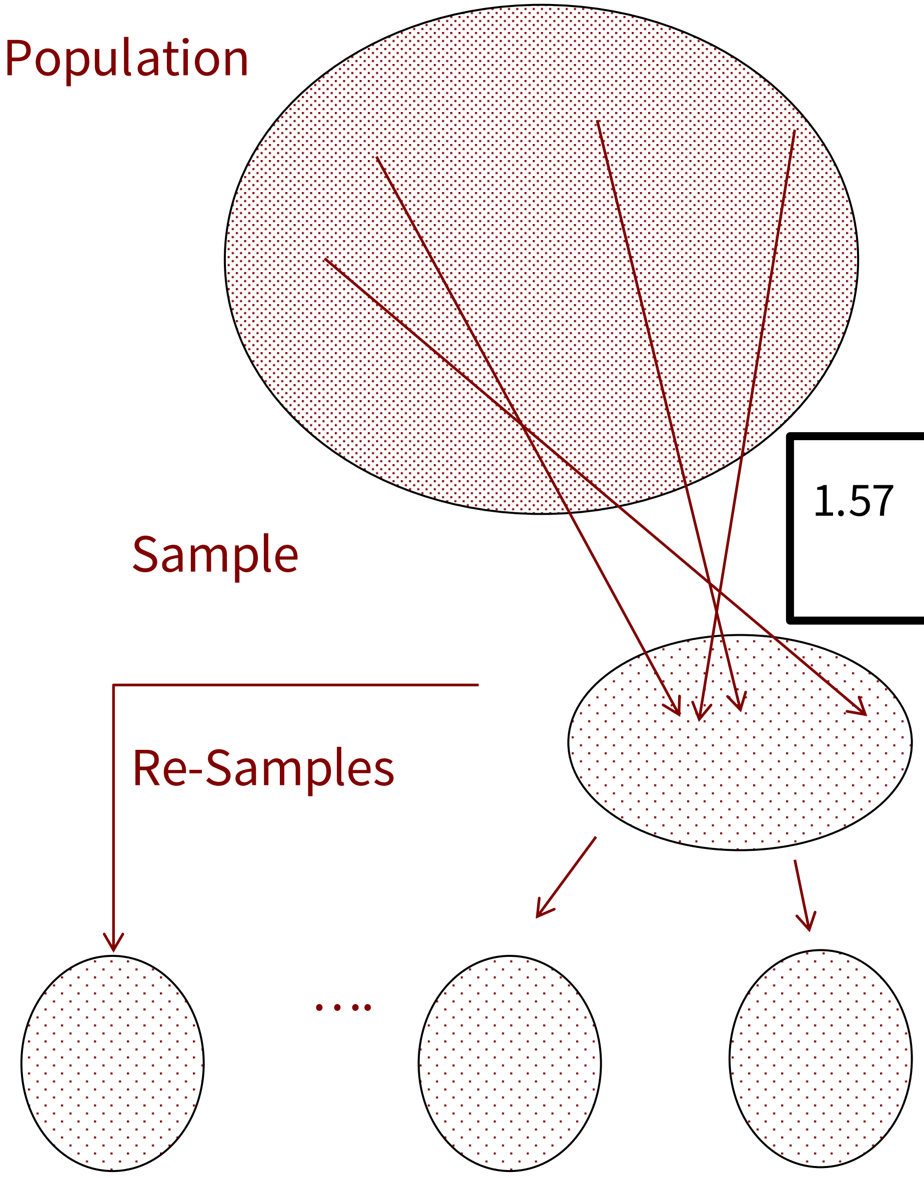
- Approximates the *sampling distribution* of an estimate
  - ***But NOT** the null (sampling) distribution as with simulation and randomization*

- Nonparametric and be applied to virtually any parameter – including means, proportions, correlations, linear model coefficients

- Used to find confidence interval and the bootstrap standard error
  - Precision method
  - Particularly useful when there is no ready formula for standard error (median, eigenvalue)

- Estimate uncertainty in phylogenies

Bootstrapping Method:



3.12 0.00 1.57 19.67 0.22 2.20
Mean = 4.46

1.57 0.22 19.67 0.00 0.22 3.12
Mean = 4.13

0.22 3.12 1.57 3.12 2.20 0.22
Mean = 1.74

0.00 2.20 2.20 2.20 19.67 1.57
Mean = 4.64

Two major categories of computational methods

Null sampling distributions:

1. Simulation – hypothesis testing

Determine the null distribution (from the parameters expected under the null hypothesis) by simulation of the sampling process

5 main steps

1. **Create and sample imaginary population**
   -parameters specified by null hypothesis
   -Same protocol that was used to collect real data
2. **Calculate test statistic on simulated sample**
3. **Repeat many times**
4. **Form the null distribution**
   - Gather simulated values for the test statistic
5. **Compare test statistic from the actual data to the null distribution**

This is a BROAD topic. Some of these simulations will be relevant: https://chi-feng.github.io/mcmc-demo/
Good blogpost with information that explains the above simulations: https://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/

Randomization/Permutation (a resampling method):

- Asks: **are two variables independent?**
- **Assumptions:** random sampling, distribution of variables have approximately same shape

- Versatile
  - Variables can be any combination of numerical or categorical
  - We don't need a null hypothesis *because we build it ourselves*. A randomization test generates a ***null distribution*** for the association between two variables.
  - **MWU test is a type of permutation tests** – but you lose power when you use ranks instead of the actual data

- Basis: **Permutation**
  - Sampling without replacement
  - Method:
    1. Create data set
       - Response variable of a test statistic measuring association **randomly assigned to Explanatory variable**
         - **You are effectively exchanging labels**
       - **All data points are used exactly once**
    2. Calculate measure of association for randomized sample
    3. Repeat randomization many times
       - A NULL distribution

Pretty much gives you a p-value and not much else!

Add to your methods flowchart!