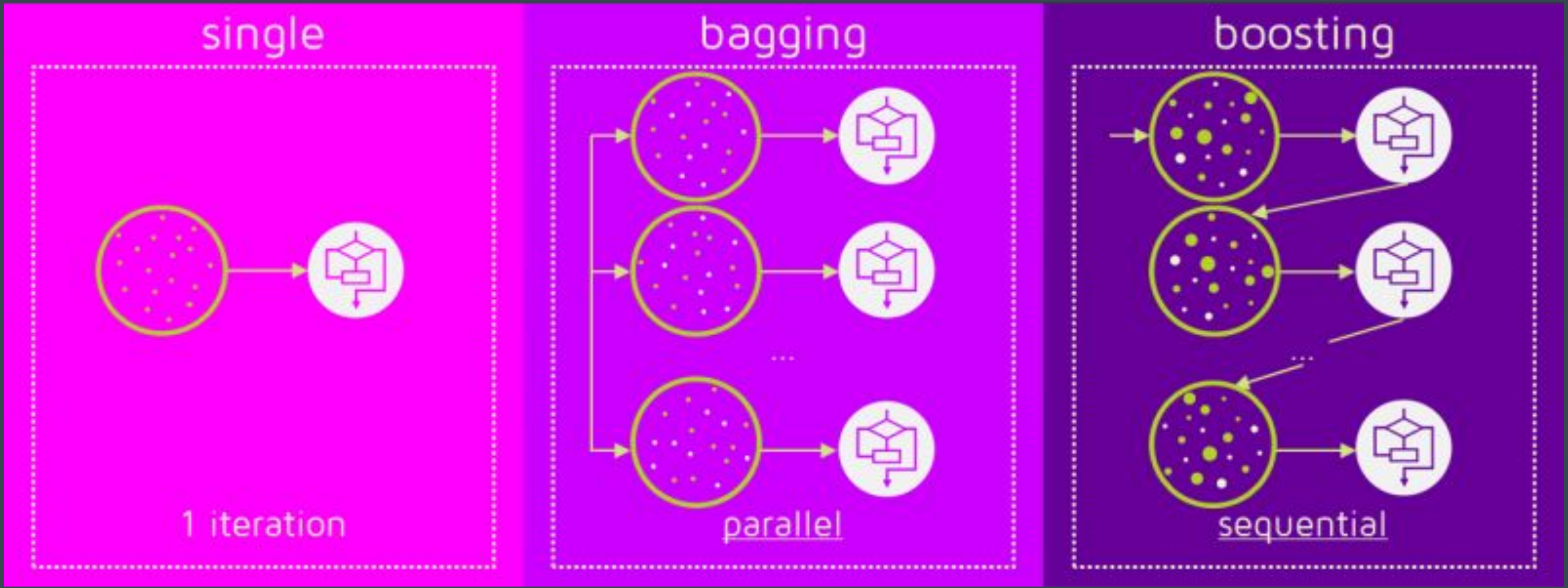
Recitation 4

●●●

Ensembles, Unsupervised Learning, & Fairness

Ensembles

- In statistics we like to average things
- How do we combine the predictions of multiple models together



Bagging

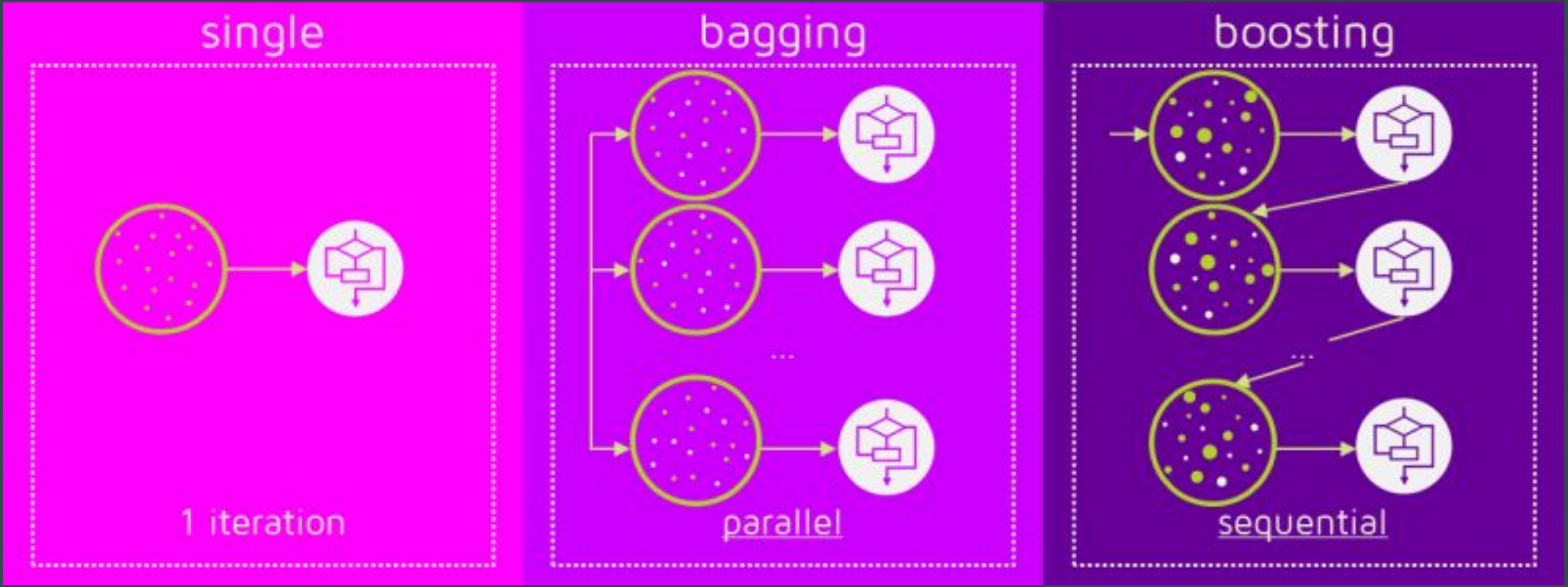
1) Create multiple datasets using sampling w/ replacement
2) Build a classifier on each of these smaller datasets (usually use the same classifier on each dataset)
3) Combine the predictions of the models to get an overall prediction
    a) Could use the Mode or Mean prediction

Boosting

1) Assign each point in our data equal weights
2) Create a subset of the data
3) Train a model on the data subset we have made
4) Reweight the dataset, giving higher weights to the points we are bad at predicting
5) Repeat steps (2) - (4) until you have enough models

- We weight the models at prediction time based on their overall training set accuracies

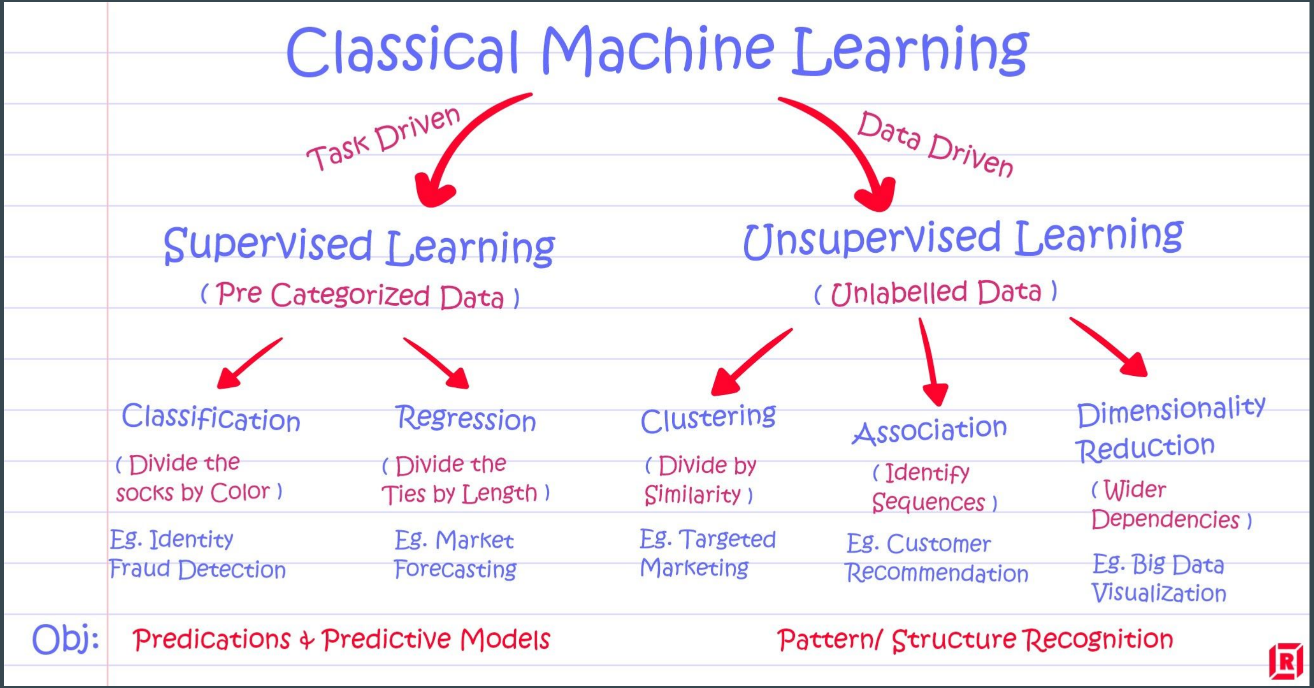
Boosting vs Bagging



Unsupervised Learning

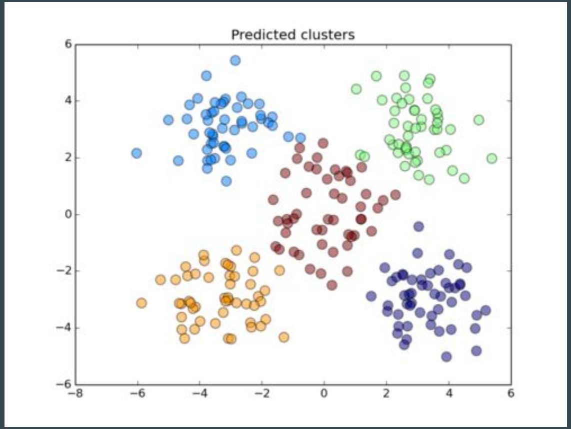
- We often have data without clear labels

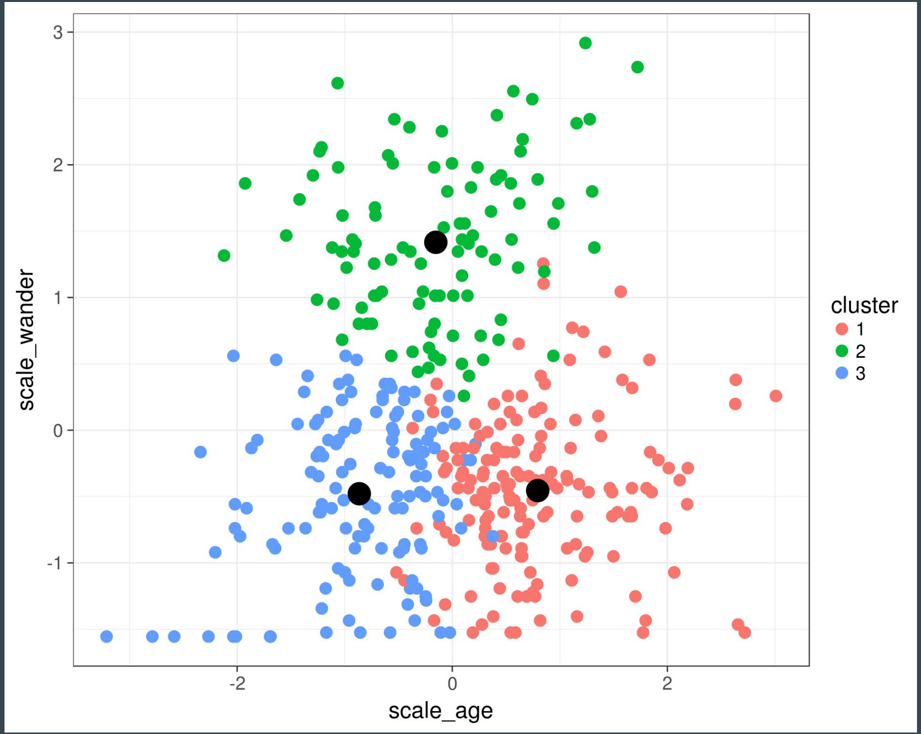
- How do we analyze this type of data?



Clustering (k-Means)

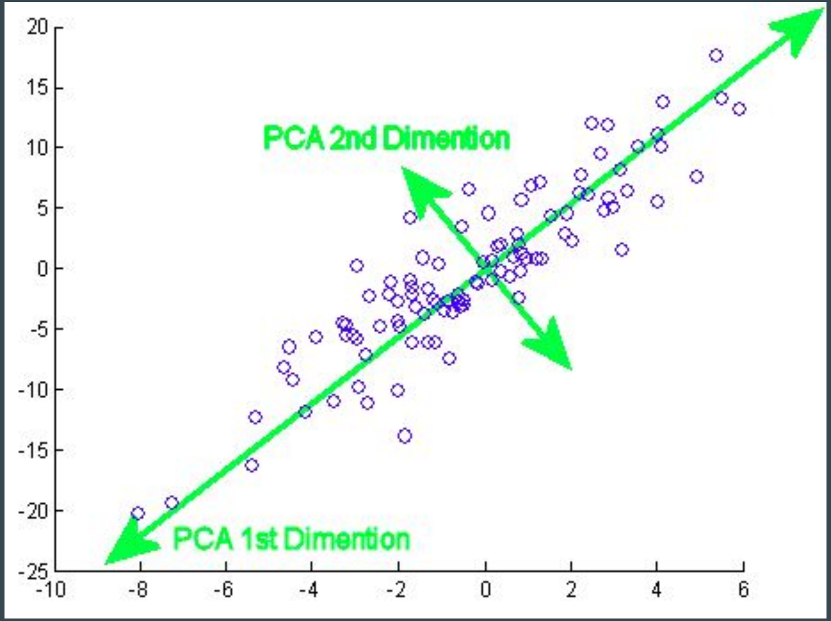
- Goal: Find k clusters in the data
- Strengths -- always converges to gives you k clusters, very explainable
- Weakness -- not good on unnormalized data, need to know k ahead of time





Dimensionality Reduction (PCA)

- Helps us create visualizations and to understand important features to our data
- Important to normalize data before doing PCA!



Fairness

- Go to fairness slides