Ensembles
Main Idea

- Fit many models to data
- Make prediction w/ each
- "Average" predictions

"Base Learner"
WHY?

\[ X_1, X_2, \ldots, X_n \sim F \]

\[ \mathbb{E}[X_i] = \mu \]

\[ \mathbb{V}[X_i] = \sigma^2 \]

\[ \bar{X} = \frac{1}{n} \sum X_i \]

\[ \mathbb{E}[\bar{X}] = \mu \]

\[ \mathbb{V}[\bar{X}] = \frac{\sigma^2}{n} \] + [Constant term]

\[ \therefore \bar{X} \sim N(\mu, \sigma^2/n) \]
Random Forest

- Fit many deep trees in parallel
- Bootstrap data for each tree
- Random splits
- "Big" file size
- Slow on fewer single trees to produce
- Feature importance
- Low tuning
Boosting

- Shallow trees in sequence
- Lots of tuning
- Win Kaggle!
  - XGBoost
  - LightGBM
  - CATBoost
Bootstrapping

Sample the rows with replacement.

"A Bootstrapping Example"