## Group 14 Team Members

Cameron Stover, Lyn Murph Talley, Megan Kuo, Herman Sun, Yunji Lee

## MLB Player and Batted Ball Data Proposal

- Goals
  - Use player statistics to predict the number of home runs a player will hit in a season
  - Use batted ball data to predict if a ball that is hit will be a home run
- Variables
  - Input for predicting number of home runs (2023 and 2024 data for qualified hitters)
    - Exit Velocity
    - Hard Hit Rate
    - Barrel Rate
    - Launch angle
    - Age
    - Bat Speed
    - Pull %
    - Fly Ball %
    - Others
  - Inputs for predicting if a ball that is hit will be a home run (2024 data)
    - Exit velocity
    - Bat speed
    - Launch angle
    - Swing length
    - Pitch speed
    - Pitch type
    - Others
  - Target variables
    - Number of home runs
    - Result of batted ball (single, double, triple, home run, out)
- Research questions
  - Can we predict the number of home runs a player will hit based on these variables?
  - Can we predict if a batted ball will be a homerun based on previous data from batted balls?
  - Can these results be used to advise players on what their approach should be when hitting (ex. swing hard, hit to pull side, should teams sign younger players)?
    - What variable is most correlated with success (more home runs)?
- Methods
  - Linear regression
  - Logistic regression
  - Decision tree
  - Random forest
  - SVM
- Data
  - <u>Baseball Savant</u>: Customizable dataset for player batting data. Can use a multitude of advanced metrics such as these batted ball statistics
  - Statcast: Shows data for every ball that is hit
  - Both datasets are downloadable as a csv file from the websites
  - Will likely need to use one hot encoding for categorical variables
  - df = pd.read\_csv("baseball\_data.csv")