Case Study I
Money Ball

Shan Wang

Lingnan College, Sun Yat-sen University

2020 Data Mining and Machine Learning LN3119

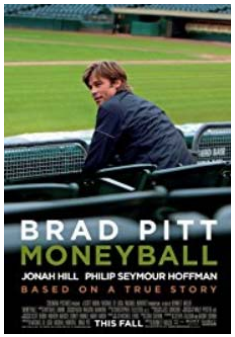
https://wangshan731.github.io/DM-ML/

The Moneyball Story

The Movie

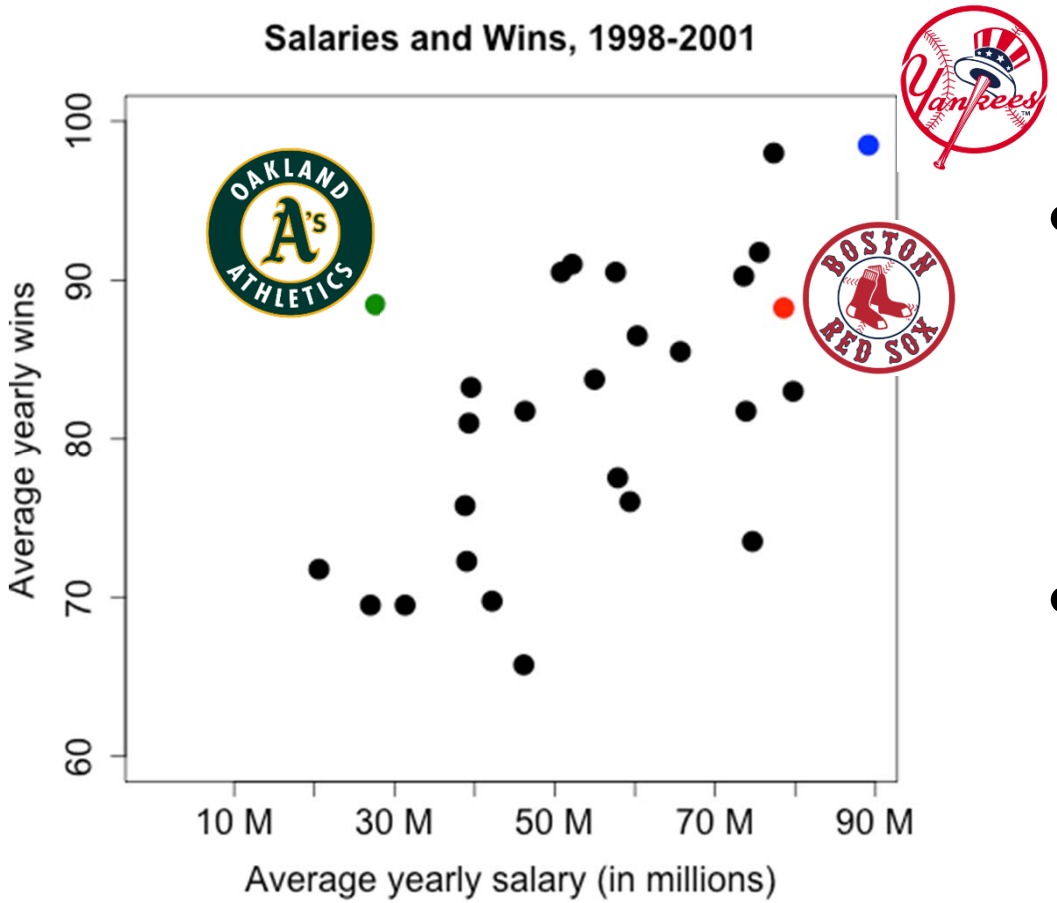
- *Moneyball* tells the story of the Oakland A's in 2002
  - One of the poorest teams in baseball
    - New ownership and budget cuts in 1995
  - But they were improving
  - How were they doing it?
  - Was it just luck?
  - In 2002, the A's lost three key players
  - Could they continue winning?

Year	Win %
1997	40%
1998	46%
1999	54%
2000	57%
2001	63%



https://www.bilibili.com/video/av58360594

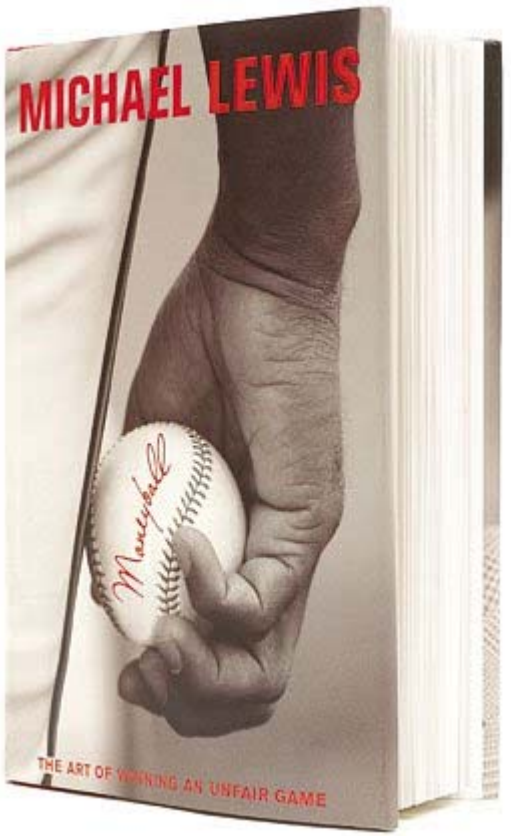
The Problem



- Rich teams can afford the all-star players
- How do the poor teams compete?

Competing as a Poor Team

- Competitive imbalances in the game
  - Rich teams have four times the salary of poor teams
- The Oakland A's can't afford the all-stars, but they are still making it to the playoffs. How?
- They take a quantitative approach and find undervalued players
  - The Art of Winning an Unfair Game



Challenging Optimization Problem

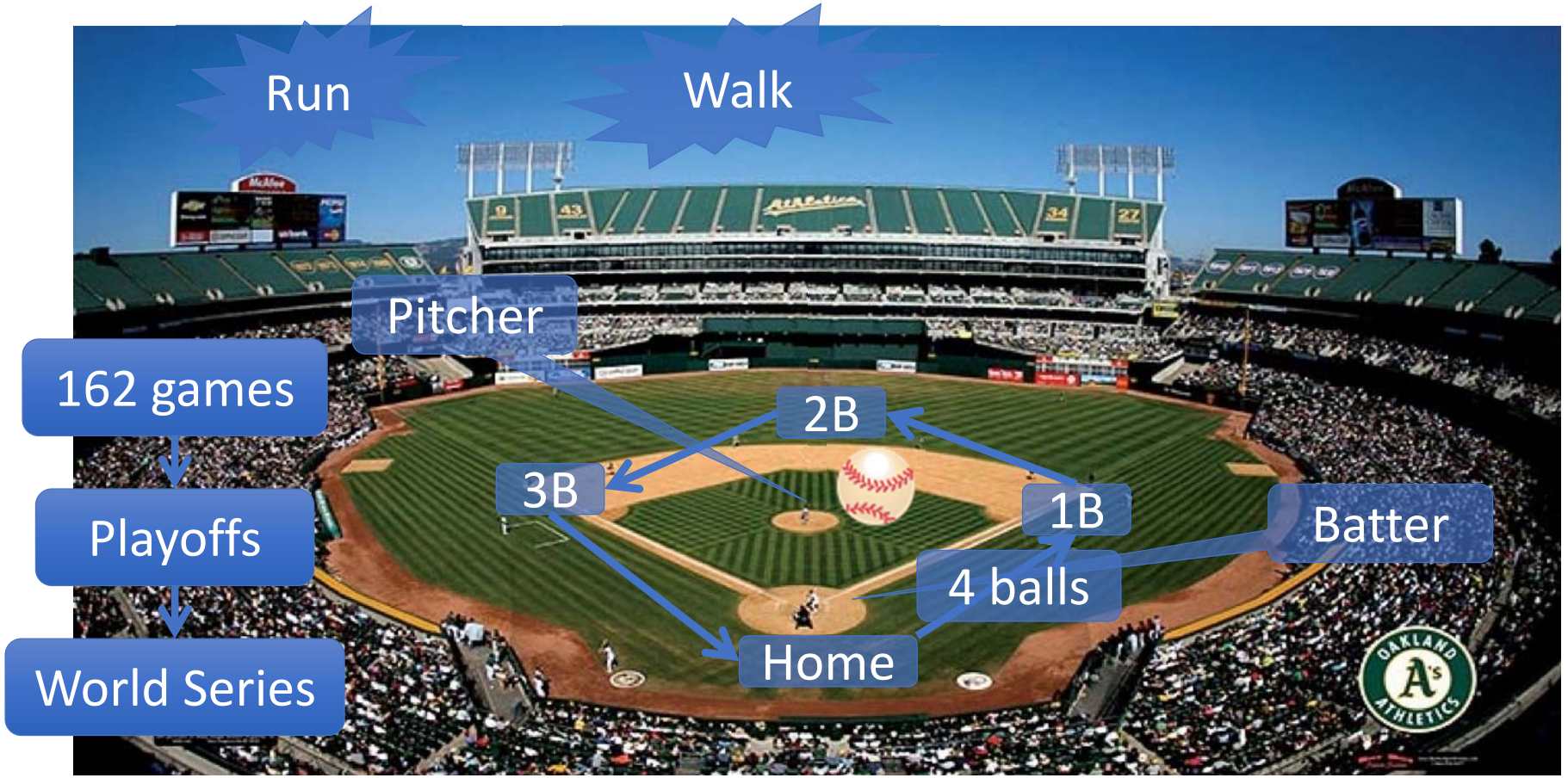
- Maximize the chance of getting into the playoffs
- Constraints
  - Budget
  - Roster
  - Etc.

Objective	Goal
$\text{Max } P(\sum_{p \in P}[[W_p \sim N(\mu_p, \sigma_p^2)]x_p] \geq T)$	Maximize probability of sum of wins being greater than or equal to amount needed to qualify for postseason with each player's wins contributed following a normal distribution
S.T.	**Rules**
$\sum_{p \in P} x_p = 25$	Roster must have exactly 25 players
$\sum_{p \in P} s_p x_p \leq B$	Sum of each player's salary must be less than or equal to the team budget
$\sum_{p \in P} SP_p x_p \geq 5$	Sum of starting pitchers must be greater than or equal to 5
$\sum_{p \in P} RP_p x_p \geq 7$	Sum of relief pitchers must be greater than of equal to 7
$\sum_{p \in P} P_p x_p = 12$	Sum of pitchers must equal 12
$\sum_{p \in P} C_p x_p \geq 2$	Sum of catchers must be greater than or equal to 2
$\sum_{p \in P} 1B_p x_p \geq 1$	Sum of first basemen must be greater than or equal to 1
$\sum_{p \in P} 2B_p x_p \geq 1$	Sum of second basemen must be greater than or equal to 1
$\sum_{p \in P} SS_p x_p \geq 1$	Sum of shortstops must be greater than or equal to 1
$\sum_{p \in P} 3B_p x_p \geq 1$	Sum of third basemen must be greater than or equal to 1
$\sum_{p \in P} LF_p x_p \geq 1$	Sum of left fielders must be greater than or equal to 1
$\sum_{p \in P} CF_p x_p \geq 1$	Sum of center fielders must be greater than or equal to 1
$\sum_{p \in P} RF_p x_p \geq 1$	Sum of right fielders must be greater than or equal to 1
$\sum_{p \in P} OF_p x_p \geq 5$	Sum of outfielders must be greater than or equal to 5
$\sum_{p \in P} IF_p x_p \geq 6$	Sum of infielders must be greater than or equal to 6
$x_p \in \{0,1\} \; \forall \; p \in P$	Binary variable constraints

Baseball Basics

https://www.youtube.com/watch?v=I8VGW0C_GO4

Baseball Basics



The Moneyball Approach

- The A's started using a different method to select players
- The traditional way was through scouting
  - Scouts would go watch high school and college players
  - Report back about their skills
  - A lot of talk about speed and athletic build
- The A's selected players based on their statistics, not on their looks
  - "The statistics enabled you to find your way past all sorts of sight-based scouting prejudices."
  - "We're not selling jeans here"

The Perfect Batter

The A's

A catcher who couldn't throw
Gets on base a lot

The Yankees

A consistent shortshop
Leader in hits and stolen bases

The Perfect Pitcher

The A's

Unconventional delivery
Slow speed

The Yankees

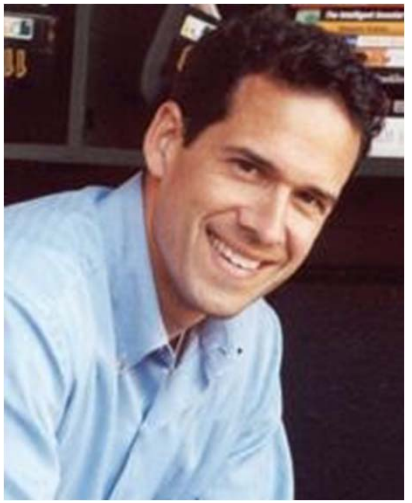
Conventional delivery
Fast speed

Billy Beane

- The general manager since 1997
- Played major league baseball, but never made it big
  - Sees himself as a typical scouting error
- Billy Beane succeeded in using analytics
  - Had a management position
    - In the 1980s and 1990s, analysts were hired by baseball teams but none of them had enough power to implement anything
  - Understood the importance of statistics
    - Hired Paul DePodesta (a Harvard graduate) as his assistant
  - Didn't care about being ostracized

Taking an Analytical View

- Paul DePodesta spent a lot of time looking at the data
- His opinion was that some skills were undervalued and some skills were overvalued
- If they could detect the undervalued skills, they could find players at a bargain



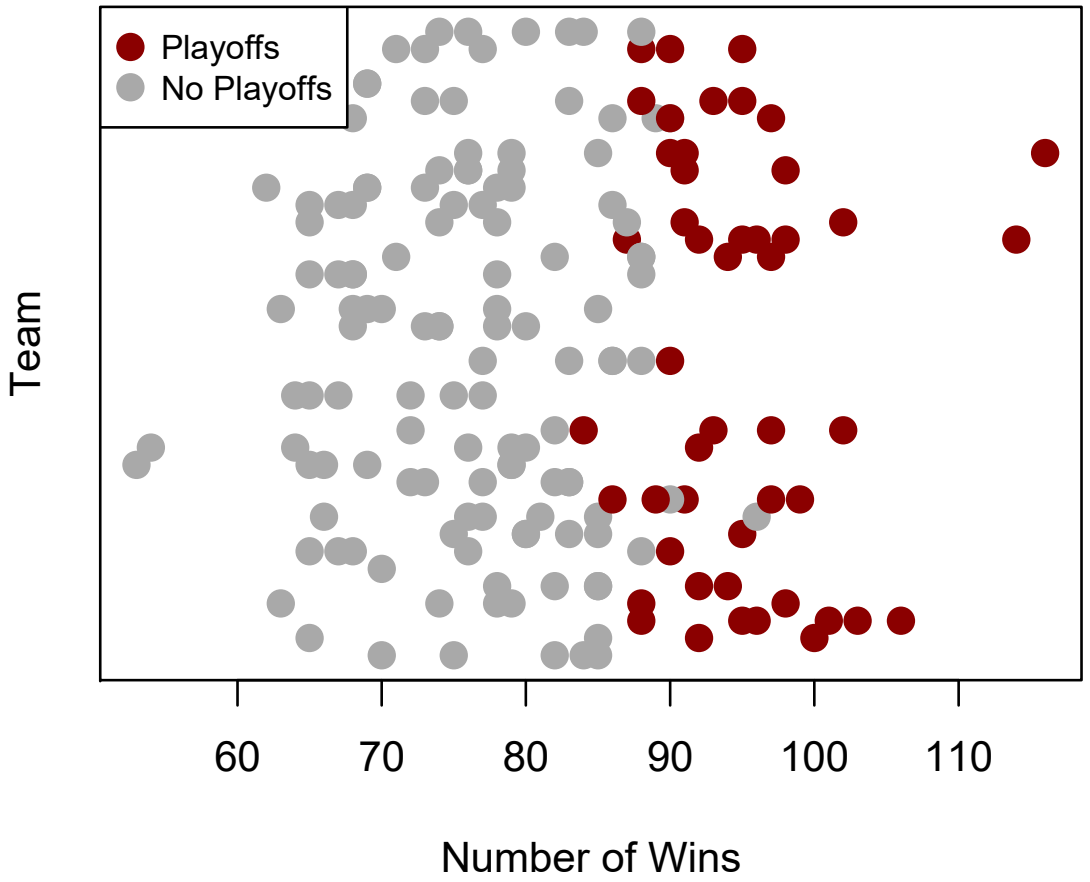
The Goal of a Baseball Team

Making it to the Playoffs

- How many games does a team need to win in the regular season to make it to the playoffs?

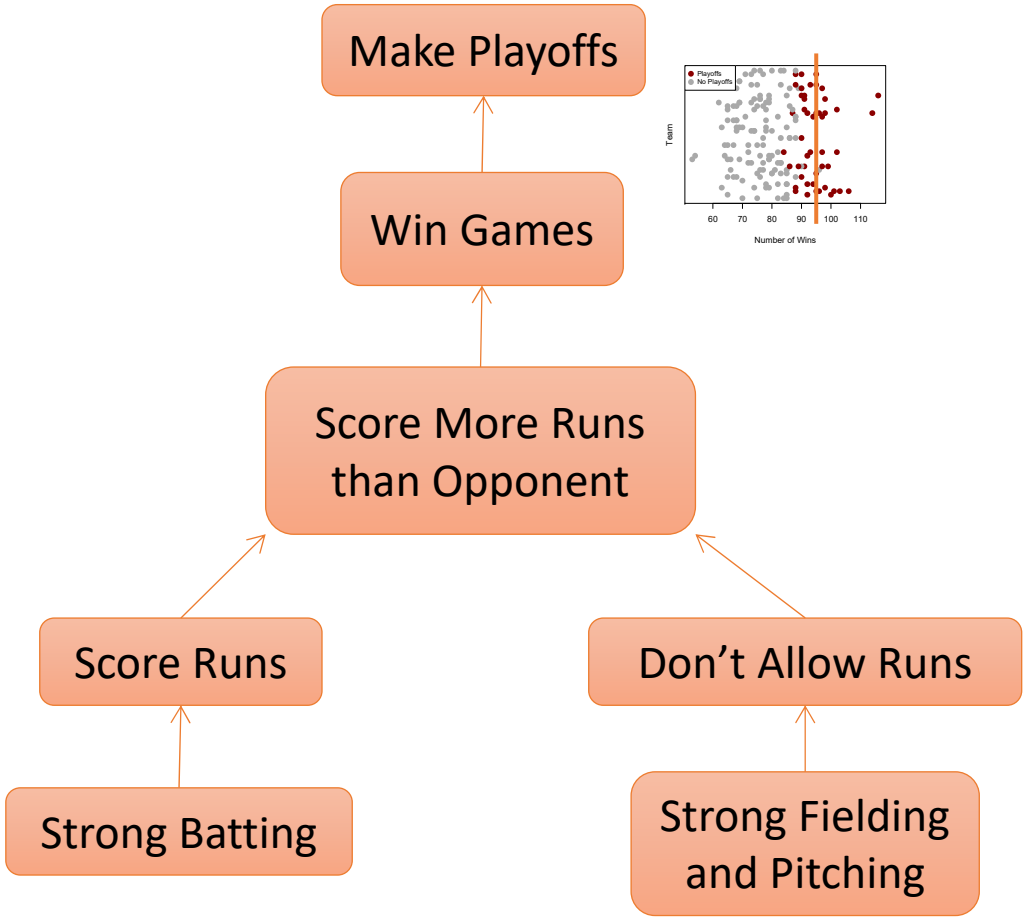
- “Paul DePodesta reduced the regular season to a math problem. He judged how many wins it would take to make it to the playoffs: 95.”

Making it to the Playoffs (cont.)



Data from all teams 1996–2001

The Goal of a Baseball Team



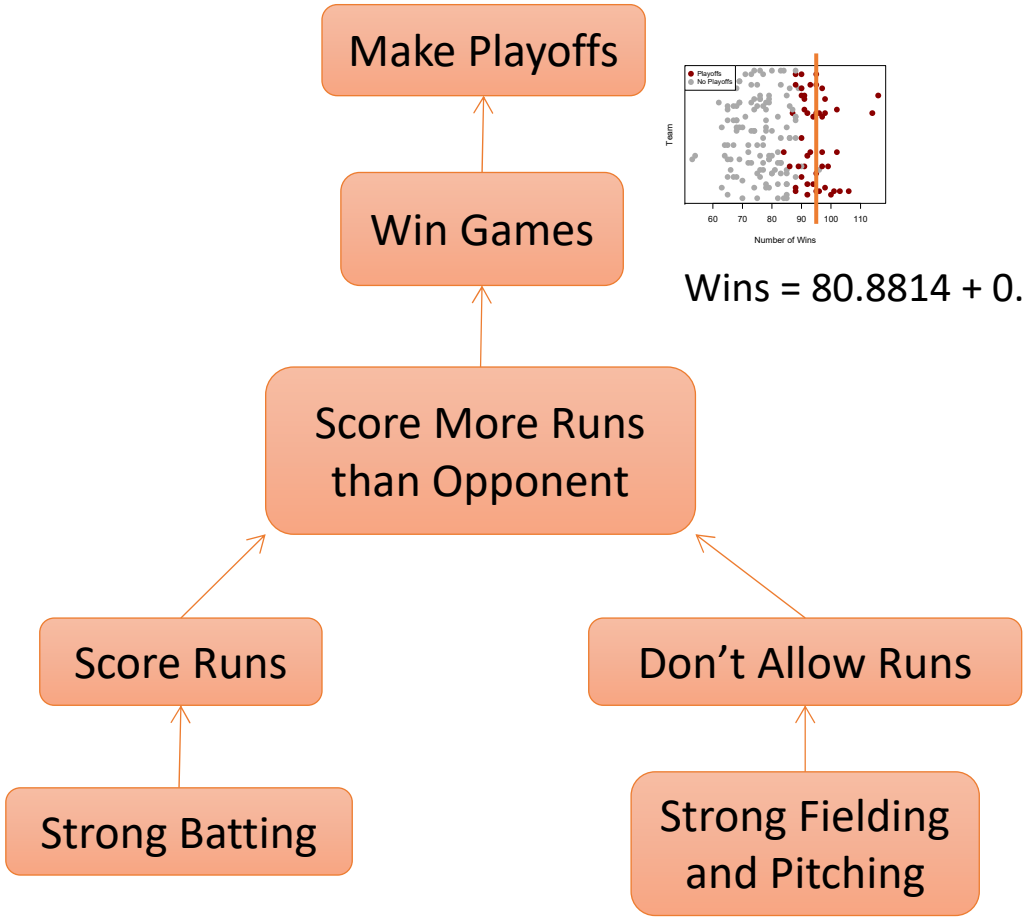
Winning 95 Games

- How does a team win games?
- They score more runs (points) than their opponent
- But how many more?
- The A's calculated that they needed to score 135 more runs than they allowed during the regular season to expect to win 95 games
  - How?

Using Linear Regression

- Data from 1962–2002
- Wins = 80.8814 + 0.1058(RS – RA)
  - RS: Runs scored (total points earned by the team in a regular season)
  - RA: Runs allowed (total points lost by the team in a regular season)
  - R^2 = 0.88
- We want
  - Wins ≥ 95
  - 80.8814 + 0.1058(RS – RA) ≥ 95
  - 0.1058(RS – RA) ≥ 95 – 80.8814
  - (RS – RA) ≥ (95 – 80.8814)/0.1058 = 133.4

The Goal of a Baseball Team



Wins = 80.8814 + 0.1058(RS – RA)

Runs Scored

- How does a team score more runs?
- The A's discovered that two statistics were significantly more important than anything else
  - On-Base Percentage (OBP)
    - Percentage of time a player gets on base (including walks)
  - Slugging Percentage (SLG)
    - How far a player gets around the bases on his turn
    - More weight to extra-base hits

Runs Scored

- Most teams focused on Batting Average (BA)
  - Getting on base by hitting the ball
- The A's claimed that:
  - On-Base Percentage (OBP) was the most important
  - Slugging Percentage (SLG) was important
  - Batting Average was overvalued
- Using linear regression:
  - RS = -804.63 + 2737.77(OBP) + 1584.91(SLG)
  - R^2 = 0.93

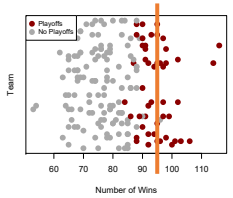
Variable selection and multicollinearity

Runs Allowed

- We can use pitching statistics to predict runs allowed
  - Opponents On-Base Percentage (OOBP)
  - Opponents Slugging Percentage (OSLG)
- Using linear regression again
  - RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)
  - R^2 = 0.91
  - Both variables are statistically significant

The Goal of a Baseball Team

Make Playoffs



Win Games

Wins = 80.8814 + 0.1058(RS – RA)

Score More Runs than Opponent

Score Runs

RS = -804.63 + 2737.77(OBP) + 1584.91(SLG)

Strong Batting

Don't Allow Runs

RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)

Strong Fielding and Pitching

Predicting Runs and Wins

- Can we predict how many games the 2002 Oakland A's will win using our models?
- The models for runs use team statistics
- Each year, a baseball team is different
- We need to estimate the new team statistics using past player performance
  - Assumes past performance correlates with future performance
  - Assumes few injuries
- We can estimate the team statistics for 2002 by using the 2001 player statistics

Predicting Runs Scored

- At the beginning of the 2002 season, the Oakland A's had 24 batters on their roster
- Using the 2001 regular season statistics for these players
  - Team OBP is 0.339
  - Team SLG is 0.430
- Our regression equation was

  RS = -804.63 + 2737.77(OBP) + 1584.91(SLG)
- Our 2002 prediction for the A's is

  RS = -804.63 + 2737.77(0.339) + 1584.91(0.430) = 805

Predicting Runs Allowed

- At the beginning of the 2002 season, the Oakland A's had 17 pitchers on their roster
- Using the 2001 regular season statistics for these players
  - Team OOBP is 0.307
  - Team OSLG is 0.373
- Our regression equation was

  RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)

- Our 2002 prediction for the A's is

  RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG) = 622

Predicting Wins

- Our regression equation to predict wins was

  $$\text{Wins} = 80.8814 + 0.1058(\text{RS} - \text{RA})$$

- We predicted
  - RS = 805
  - RA = 622
- So our prediction for wins is

  $$\text{Wins} = 80.8814 + 0.1058(805 - 622) \approx 100$$

The Oakland A's

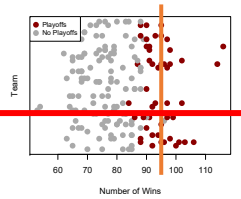
- We can use our models to make predictions for the 2002 season
- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800
Runs Allowed	622	650 – 670	653
Wins	100	93 – 97	103

- The A's set a record by winning 20 games in a row
- Won one more game than the previous year, and made it to the playoffs

The Goal of a Baseball Team

Make Playoffs



Win Games

Wins = 80.8814 + 0.1058(RS – RA)

Score More Runs than Opponent

Score Runs

RS = -804.63 + 2737.77(OBP) + 1584.91(SLG)

Strong Batting

Don't Allow Runs

RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)

Strong Fielding and Pitching

Let’s get our hands dirty!

Data Analysis

Let's do some basic data analysis using our WHO data.

Hide

```
WHO$Under15
```

```
  [1] 47.42 21.33 27.42 15.20 47.58 25.96 24.42 20.34 18.95 14.51 22.25 21.62 20.16 30.57 18.99 15.10 16.88 34.4
0 42.95 28.53
 [21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 30.17 40.07 48.52 21.38 17.95 28.03 42.1
7 42.37 30.61
 [41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 30.53 30.29 31.25 30.62 38.95 43.10 15.6
9 43.29 28.88
 [61] 16.42 18.26 38.49 45.90 17.62 13.17 38.59 14.60 26.96 40.80 42.46 41.55 36.77 35.35 35.72 14.62 20.71 29.4
3 29.27 23.68
 [81] 40.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 30.10 24.90 30.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.44 26.65 29.03 47.14 14.98 30.10 40.22 20.17 29.02 35.81 18.26 27.05 19.01 27.85 45.38 25.28 36.5
9 30.10 35.58
[121] 17.21 20.26 33.37 49.99 44.23 30.61 18.64 24.19 34.31 30.10 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.05 15.45 43.56 25.96 24.31 25.70 37.88 14.04 41.60 29.69 43.54 16.45 21.95 41.74 16.48 15.00 14.16 40.3
7 47.35 29.53
[161] 42.28 15.20 25.15 41.48 27.83 38.05 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 20.73 23.22 26.0
0 28.65 30.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.00 37.37 28.84 22.87 40.72 46.73 40.24
```

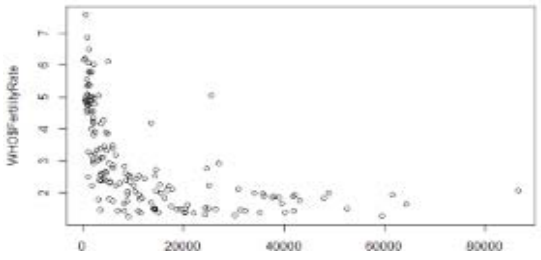
```
WHO$Country[which.min(WHO$Under15)]
```

```
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

Let's create some plots for exploratory data analysis (EDA). First, Let's create a basic scatterplot of GNI versus FertilityRate.

Hide

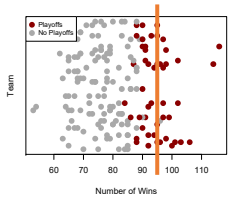
```
plot(WHO$GNI, WHO$FertilityRate)
```



Using Regression Tree

The Goal of a Baseball Team

Make Playoffs



Win Games

Wins = 80.8814 + 0.1058(RS – RA)

Score More Runs than Opponent

Score Runs

RS = -804.63 + 2737.77(OBP) + 1584.91(SLG)

Strong Batting

Don't Allow Runs

RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)

Strong Fielding and Pitching

Predicting Runs Scored

- Most teams focused on Batting Average (BA)
- The A's claimed that:
  - On-Base Percentage (OBP) was the most important
  - Slugging Percentage (SLG) was important
  - Batting Average was overvalued
- Using linear regression
  - Multicollinearity problem if all three variables are used
  - Conduct univariate analysis and try different combinations
- Let us try Regression Tree

Let’s get our hands dirty!

Data Analysis

Let's do some basic data analysis using our WHO data.

Hide

```
WHO$Under15
```

```
  [1] 47.42 21.33 27.42 15.20 47.58 25.96 24.42 20.34 18.95 14.51 22.25 21.62 20.16 30.57 18.99 15.10 16.88 34.4
0 42.95 28.53
 [21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 30.17 40.07 48.52 21.38 17.95 28.03 42.1
7 42.37 30.61
 [41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 30.53 30.29 31.25 30.62 38.95 43.10 15.6
9 43.29 28.88
 [61] 16.42 18.26 38.49 45.90 17.62 13.17 38.59 14.60 26.96 40.80 42.46 41.55 36.77 35.35 35.72 14.62 20.71 29.4
3 29.27 23.68
 [81] 40.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 30.10 24.90 30.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.44 26.65 29.03 47.14 14.98 30.10 40.22 20.17 29.02 35.81 18.26 27.05 19.01 27.85 45.38 25.28 36.5
9 30.10 35.58
[121] 17.21 20.26 33.37 49.99 44.23 30.61 18.64 24.19 34.31 30.10 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.05 15.45 43.56 25.96 24.31 25.70 37.88 14.04 41.60 29.69 43.54 16.45 21.95 41.74 16.48 15.00 14.16 40.3
7 47.35 29.53
[161] 42.28 15.20 25.15 41.48 27.83 38.05 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 20.73 23.22 26.0
0 28.65 30.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.00 37.37 28.84 22.87 40.72 46.73 40.24
```

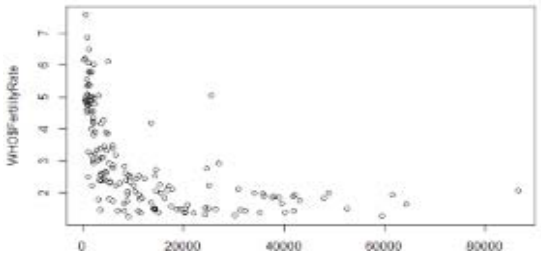
```
WHO$Country[which.min(WHO$Under15)]
```

```
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

Let's create some plots for exploratory data analysis (EDA). First, Let's create a basic scatterplot of GNI versus FertilityRate.

Hide

```
plot(WHO$GNI, WHO$FertilityRate)
```



The Goal of a Baseball Team

Why isn't the goal to win the World Series?

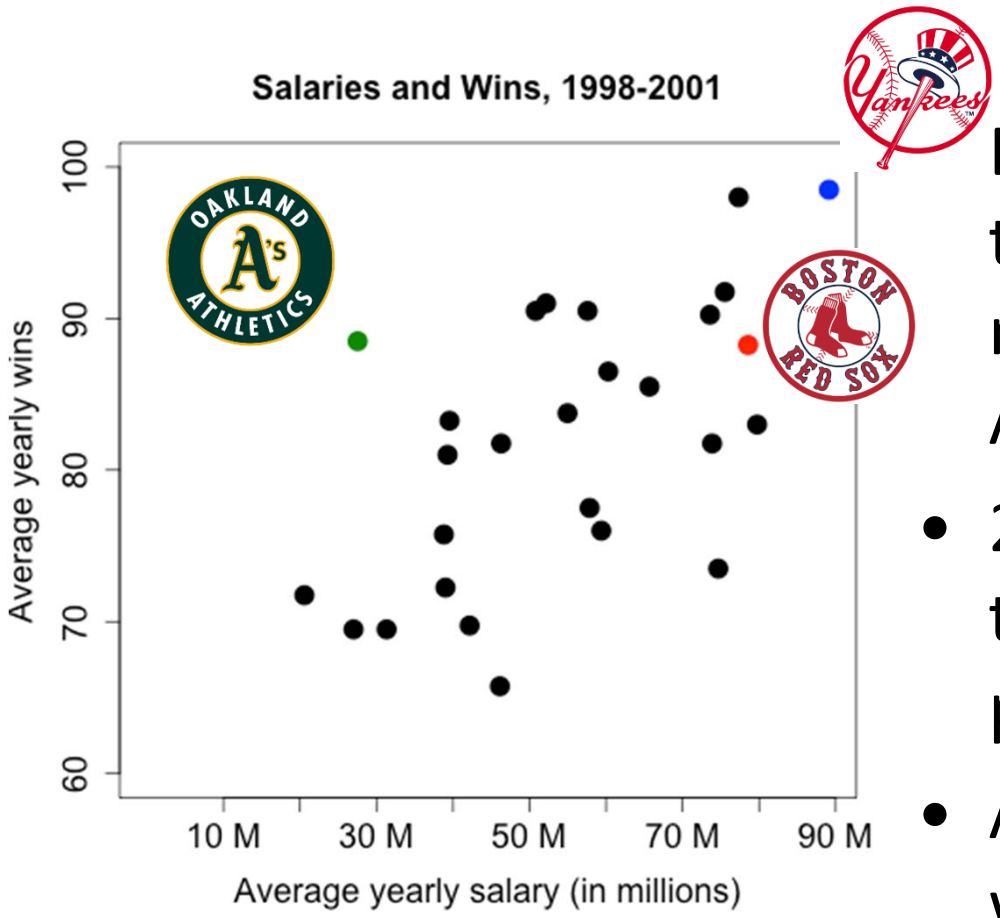
Luck in the Playoffs

- Billy and Paul see their job as making sure the team makes it to the playoffs – after that all bets are off
  - The A's made it to the playoffs in 2000, 2001, 2002, 2003
  - But they didn't win the World Series
- Why?
- "Over a long season the luck evens out, and the skill shines through. But in a series of three out of five, or even four out of seven, anything can happen."
  - *Law of large numbers*

Other Moneyball Strategies

- Moneyball also discusses:
  - How it is easier to predict professional success of college players than high school players
  - Stealing bases, sacrifice bunting, and sacrifice flies are overrated
  - Pitching statistics do not accurately measure pitcher ability
    - Pitchers only control strikeouts, home runs, and walks

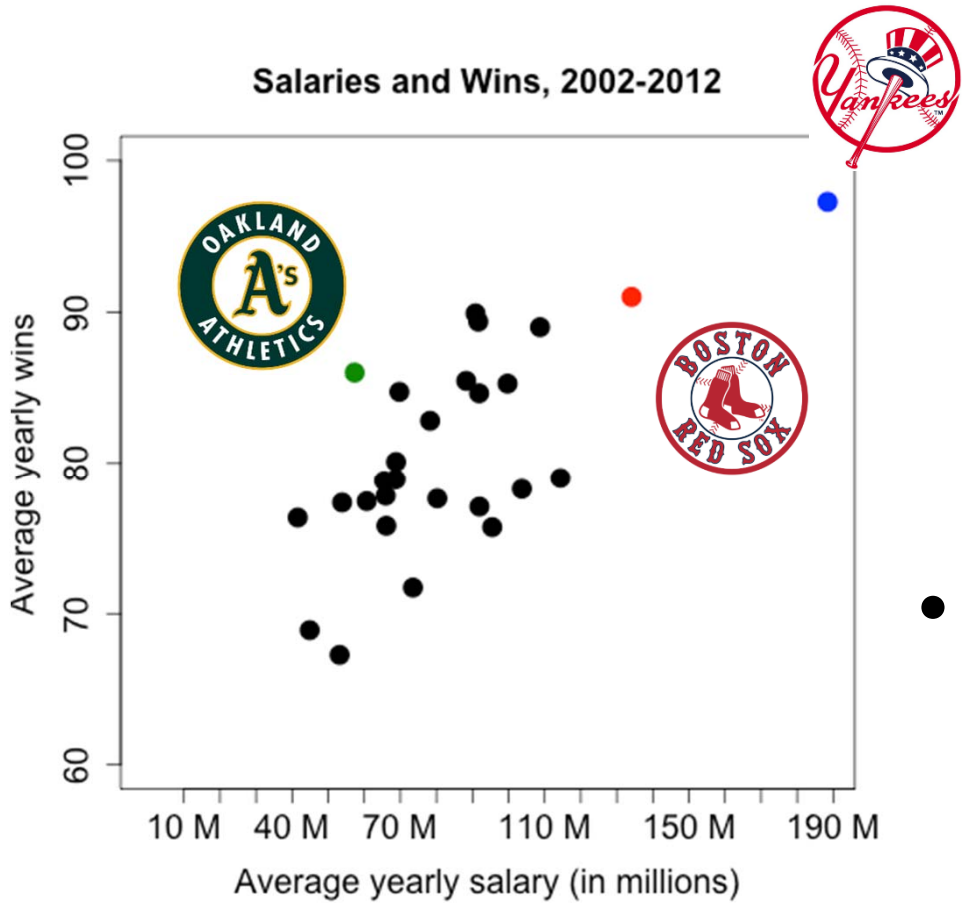
Where was Baseball in 2002?



Before Moneyball techniques became more well-known, the A's were an outlier

- 20 more wins than teams with equivalent payrolls
- As many wins as teams with double the payroll

Where is Baseball Now?



Now, the A's are still an efficient team, but they only have 10 more wins than teams with equivalent payrolls

- Fewer inefficiencies

Other Baseball Teams and Sports

- Every major league baseball team now has a statistics group
- The Red Sox implemented quantitative ideas and won the World Series for the first time in 86 years
- Baseball analytics grows into sabermetrics
- Analytics are also used in other sports

Billy Beane on Sports Analytics

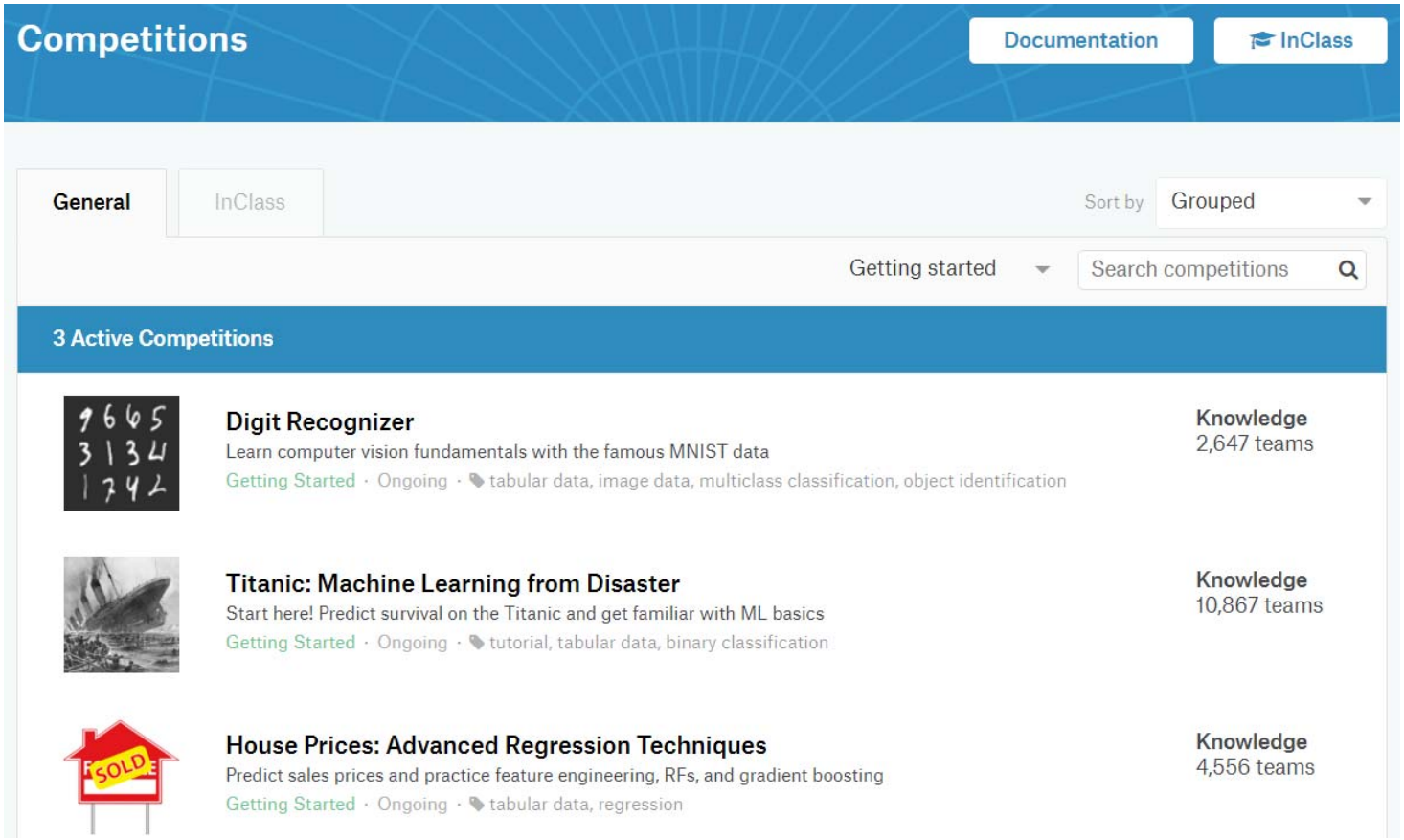
https://www.youtube.com/watch?v=42SxZIseImE

Kaggle Competition

Kaggle Competition

- **Dataset: Moneyball**
  - **Predict the value of "Playoff" from 2003-2010**
- **Techniques (prediction models)** are **restricted** to what you have learned so far in this course
  - You can use different packages for these techniques
  - You can search online for tips or tricks that make your coding more efficient
  - You can use any visualization techniques (no restriction) to help identify patterns and generate ideas to formulate your models
- https://www.kaggle.com/c/sysu-dmml-b/

“Getting Started” with Kaggle

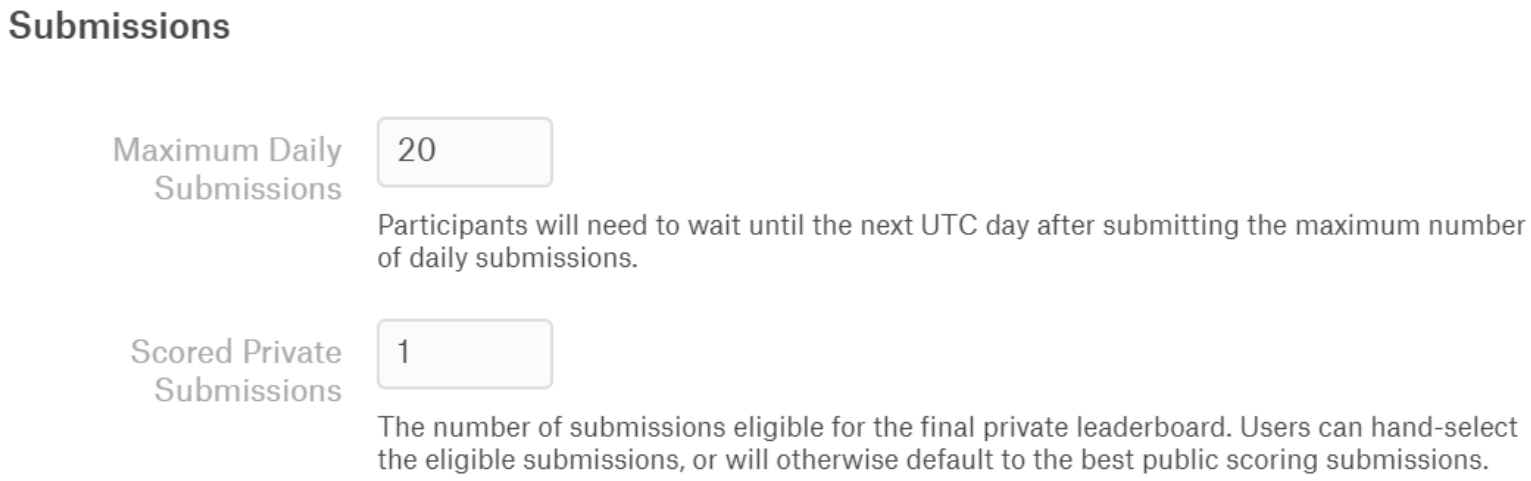


Grading

- **5%** on model (novelty/rigor) + **15%** on prediction performance by the deadline
  - Evaluated based on a private test set

Rules

- Each student is allowed to have only one account
  - Team name format: “Name_StudentID”, “Name_StudentID”
  - Submission under different names will not be considered
- Each student can submit maximum 20 attempts every day



Submission

- Predictions submitted to Kaggle
- R code in a single Notebook file with description on your approach
  - One file per student
  - Be able to compile with training and test data stored in the same location
  - Submit to TA
  - Due: 6/28/2020, 11:59pm

Next Lecture

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Unsupervised learning
  - Clustering
  - PCA (Dimension Reduction)
  - EM
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

2020 Data Mining and Machine Learning LN3119
https://wangshan731.github.io/DM-ML/

Questions?

Shan Wang (王杉)

https://wangshan731.github.io/