Welcome to SURE

Background and overview

June 6th, 2022

Meet the instructors

Meet the instructors

- Instructor: **Professor Ron Yurko** (@Stat_Ron)
  - @CMU_Stats PhD '22, BS '15
  - Incoming Assistant Teaching Professor
  - Industry experience: Pittsburgh Pirates '14, finance '16-'17, Zelus Analytics '21-'22
  - Research: statistical genetics, selective inference, clustering, and statistics in sports / sports analytics
  - Star Wars, Marvel, and Pittsburgh sports fan

Teaching Assistants!

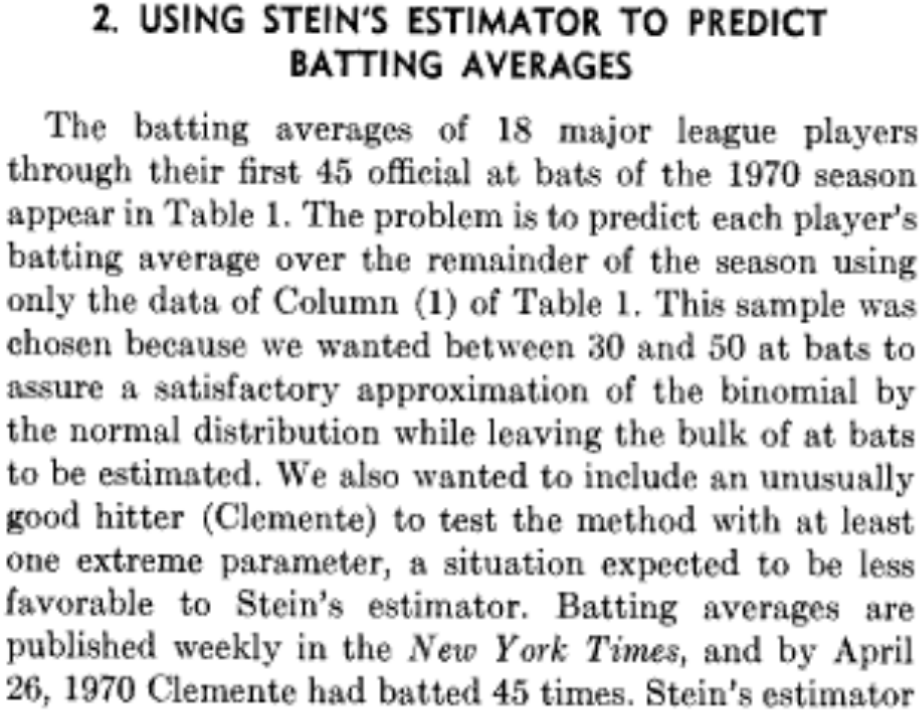
- Nick Kissel
- Wanshan Li
- Meg Ellingwood
- Kenta Takatsu
- YJ Choe

Statistics in sports research?

You might think statistics in sports or sports analytics research is relatively new...

Professors Brad Efron and Carl Morris disagree

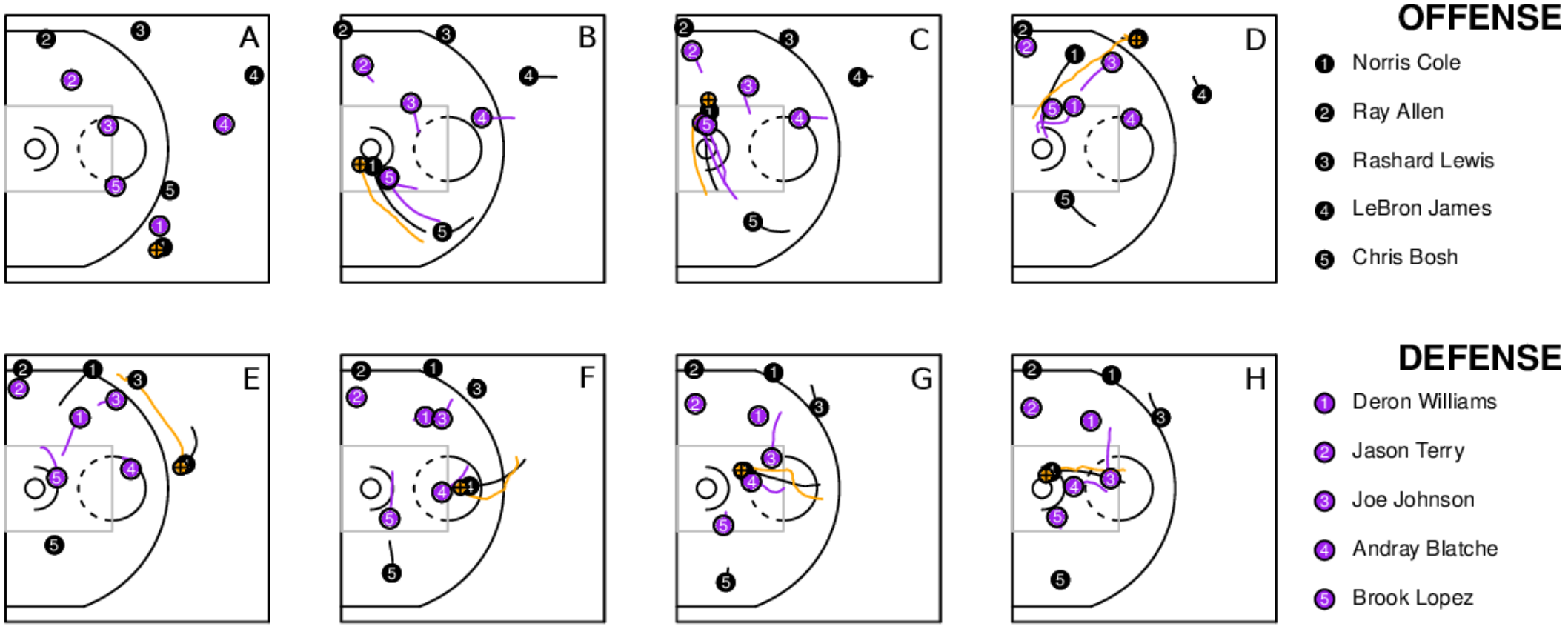
- "Data analysis using Stein's estimator and its generalizations"
- *Journal of the American Statistical Association* **(1975)**
- Introduction of **Empirical Bayes** to sports
    - Improve accuracy by pooling information from other players



2. USING STEIN'S ESTIMATOR TO PREDICT BATTING AVERAGES

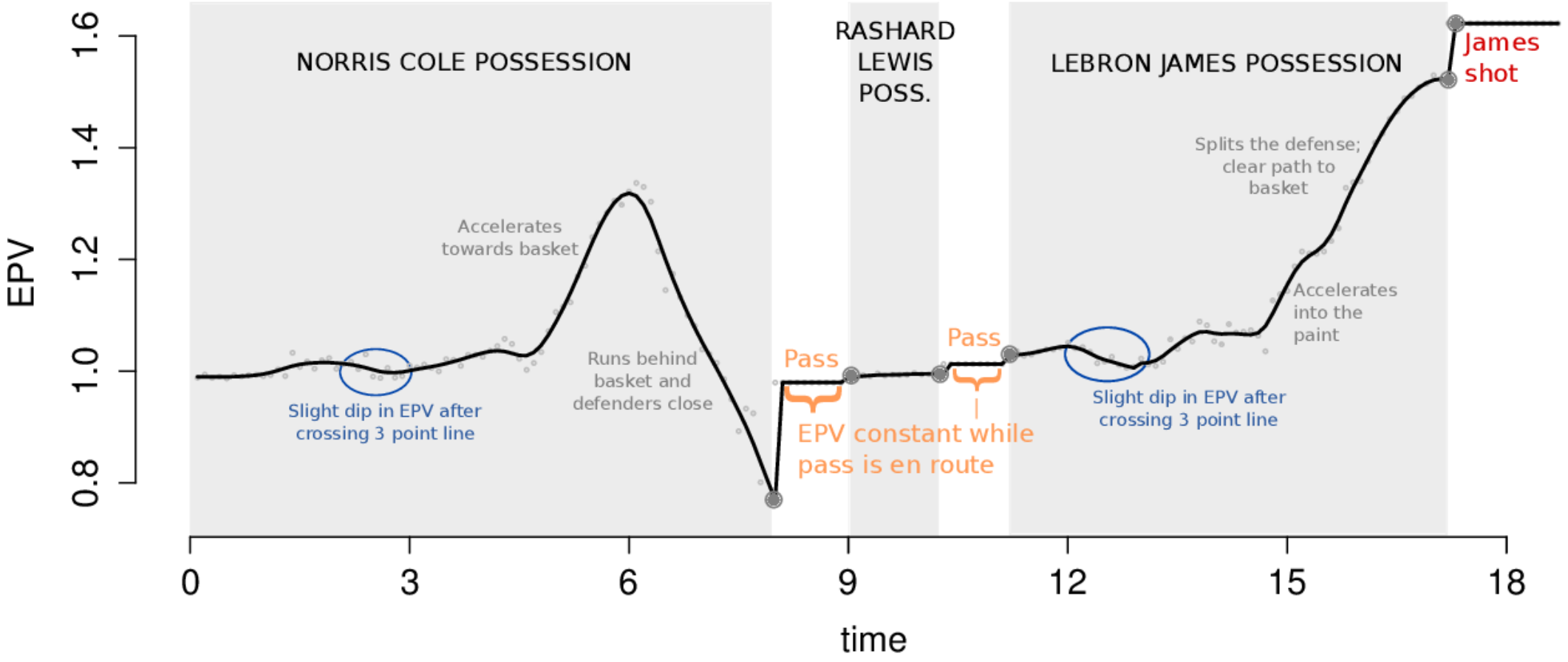
The batting averages of 18 major league players through their first 45 official at bats of the 1970 season appear in Table 1. The problem is to predict each player's batting average over the remainder of the season using only the data of Column (1) of Table 1. This sample was chosen because we wanted between 30 and 50 at bats to assure a satisfactory approximation of the binomial by the normal distribution while leaving the bulk of at bats to be estimated. We also wanted to include an unusually good hitter (Clemente) to test the method with at least one extreme parameter, a situation expected to be less favorable to Stein's estimator. Batting averages are published weekly in the *New York Times*, and by April 26, 1970 Clemente had batted 45 times. Stein's estimator

Everything **starts with the data**



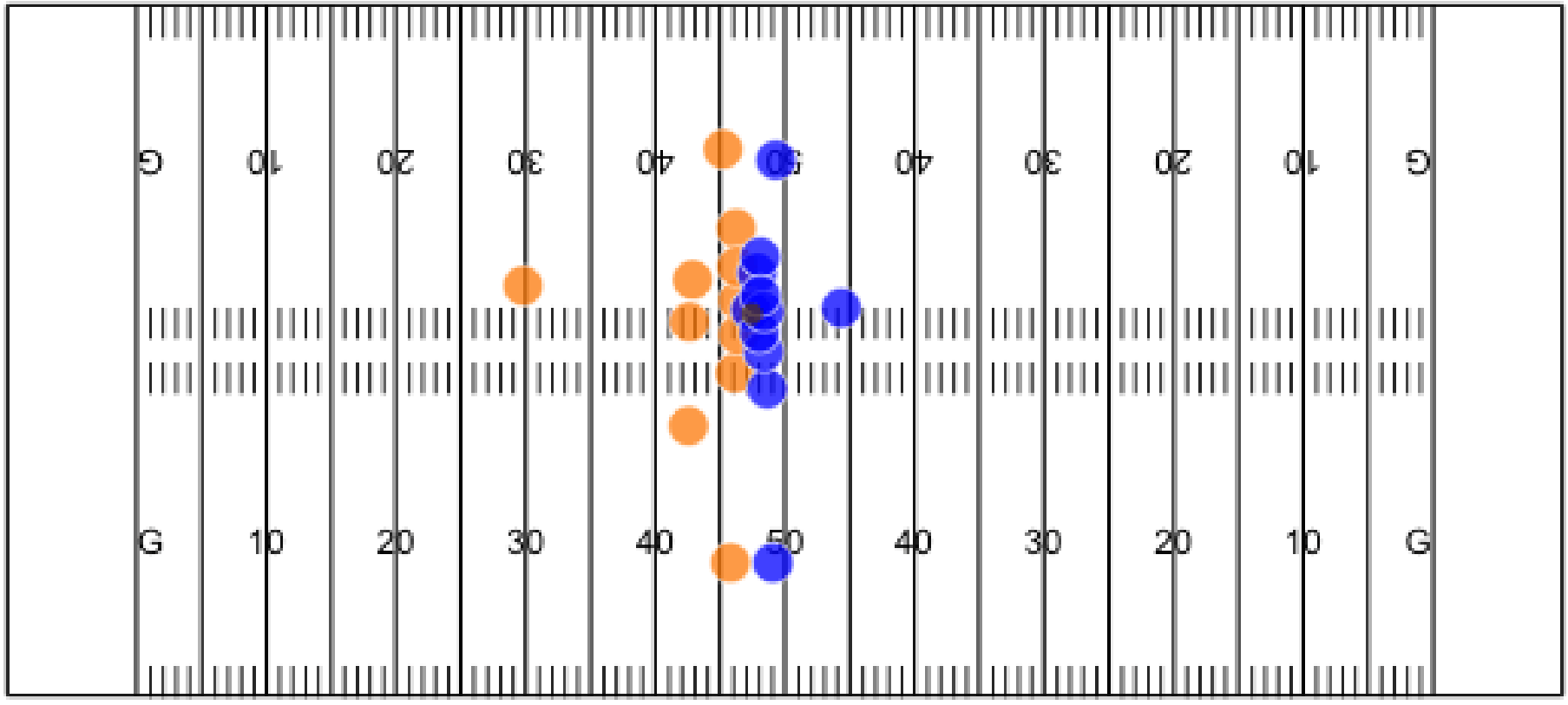
Cervone et al. "A multiresolution stochastic process model for predicting basketball possession outcomes." *Journal of the American Statistical Association* (2016)

Everything **starts with the data**



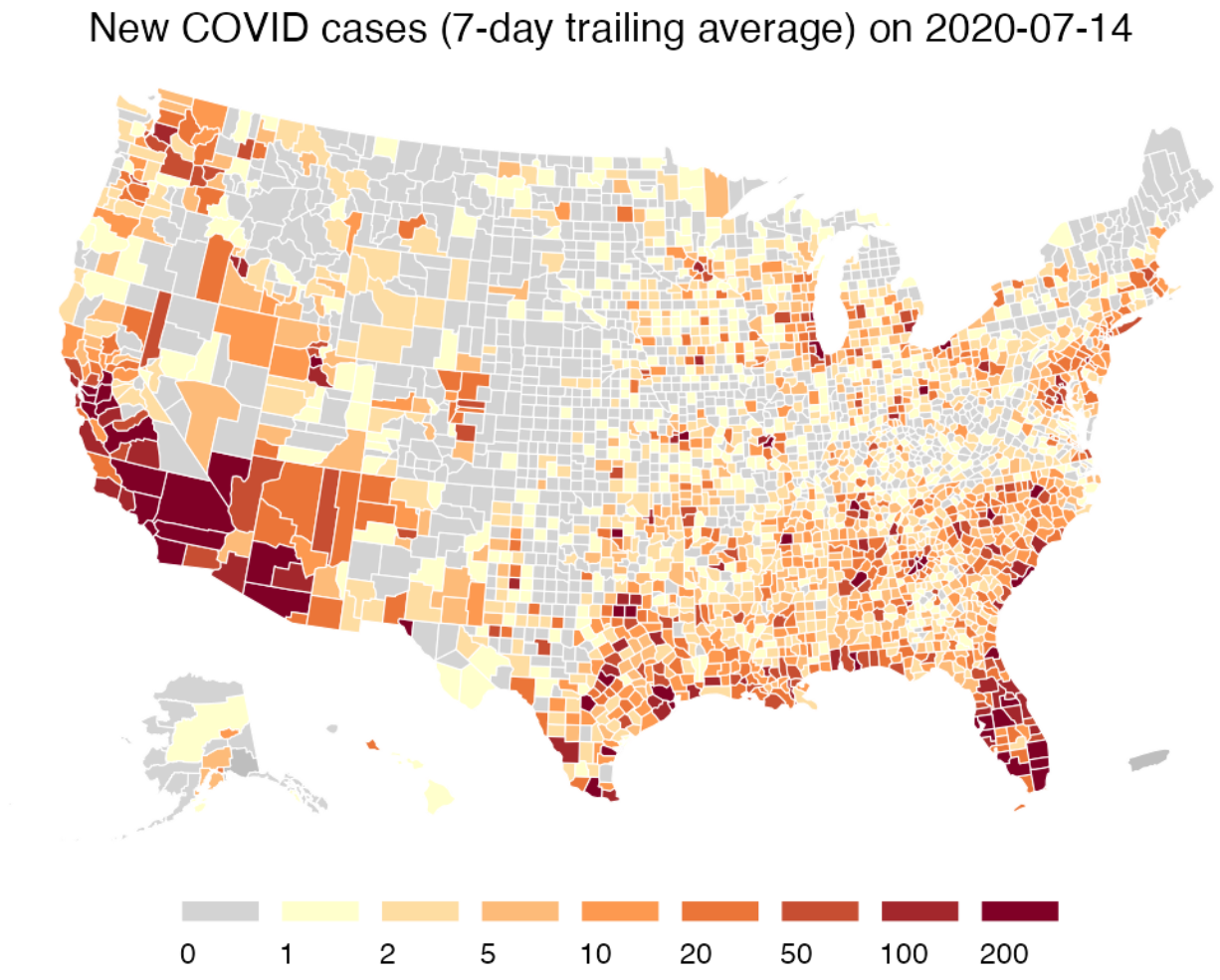
Cervone et al. "A multiresolution stochastic process model for predicting basketball possession outcomes." *Journal of the American Statistical Association* (2016)

NFL Big Data Bowl tracking data example



Yurko et al. "Going deep: models for continuous-time within-play valuation of game outcomes in American football with tracking data." *Journal of Quantitative Analysis in Sports* (2020)

CMU Delphi COVIDcast



General outline and key dates (subject to change, all times in EST)

Lectures: Monday to Friday, 10:30 AM to 12 PM

- Prof Yurko's office hours are Monday and Wednesdays 3:30 to 5:00 PM in 132D

Labs: Monday to Thursday, 1:30 to 3 PM (sports in 232M and health in 129A)

- Will begin with mini projects & practice presentations before shift to focus on main projects

- First two weeks, June 6-17:
  - EDA, data visualization,
  - Clustering
- June 21-30:
  - Linear models, model assessment, regularization
  - Splines, GAMs, and PCA
- July 6 - July 15:
  - GLMs
  - Tree-based models
  - Labs will shift focus to main projects
- July 18 - 29:
  - Special topics (e.g., survival analysis)
  - Focus on projects!

Plus many guest speakers (**check your email!**)

Goals for the summer

- Develop fundamentals research skills: data wrangling, visualization, modeling, communication
    - Become familiar with `R`, `tidyverse`, `ggplot2`, `markdown`, GitHub
- Complete statistical learning bootcamp
- Create a portfolio of projects with GitHub and practice reproducible research
    - **All presentations will be made using `R` Markdown with xaringan!**
- Network with academic researchers and industry professionals
    - Optum speaker series, Wednesdays at 12 PM

Ask questions, learn, and grow

Sports Academic Advisor: Glenn Clune (gclune@andrew.cmu.edu)

Health Academic Advisor: Amanda Mitchell (ajmitche@andrew.cmu.edu)

Resources to remember!

- Program website: http://www.stat.cmu.edu/cmsac/sure/2022/materials/
- Check out the References tab for links to online textbooks and other useful references
- Data Sources tab for links to various public datasets
- We will also use slack to communicate, share interesting articles and materials throughout the summer
    - See previous email with the workspace invitation link

And now it's your turn...

(but we're here to help!)