# DS-SF-27

Final Project 2

# Project Problem & Hypothesis

- Problem
    - This project is about being able to predict when users are likely to be retained after using a mobile app for the first time.  If we are able to reliably predict this, we can then design features to improve retention.
- Machine Learning Model
    - This is a classification problem, and the outcome of the machine learning model will be the probability that a user will be retained after their first session.
- Impact
    - This will have a substantial impact on the user growth of our mobile app.  If we are able to retain a higher percentage of new users, then our MAUs will increase substantially as we acquire more users.

# Project Problem & Hypothesis continued

- Prediction variables
    - # of teams added and time spent in app in first session
- Hypothesis
    - As users add more teams and spend more time in app in their first session, the higher the probability they will be retained.

# Datasets

The data will come from the main Bleacher Report user DB.

- Sample
  - All new users
  - First session between 8/1/2016 - 10/1/2016
  - Features
    - First session date
    - # streams added
    - Time spend in first session
    - Retained 30 days later (boolean)

# Domain Knowledge

- Current PM of Growth for Bleacher Report
- Intimate knowledge of app architecture/feature set, and qualitative feedback
- Deep understanding of mobile app growth framework
- My background/domain knowledge is essential for this project

- This will be the first project of this kind for Bleacher Report

# Project Concerns

- I’m not sure I fully understand how and when to employ feature analysis
- I need more practice with logistic regression
- The features I’ve chosen are a subset of a massive amount of potential features that could predict retention in our mobile app ➜ the outcome of this project very well could be that I can’t create a good model from the data.
    - I haven’t been able to look at raw data up until now, so I have no idea what state the data is in. There could be a substantial amount of cleaning required, or they may be none.
    - The main cost if I can’t build a good model is my time…
- There could be collinearity between # streams added and time spent in first session
- From my understanding our main s3 data warehouse contains unstructured data, but we are also working on structuring the data.  I’m hoping I can access the structured data…

# Outcomes

- I expect the output to be a probability that a user will be retained
- My audience expects the same
- The model should not be overly complicated
- The model will be a huge success if I can predict if a user will be retained with over 85% accuracy