* Only use teams, location, and date of games to make predictions
* Compare results with random numbers as "null" model
* Possible ML (LSTM)
* Mixed effects model
* Poisson model
* Effect of back to back games/by-weeks
* K fold cross validation

What to do:

* merge into one big giant **super crazy** california pizza kitchen data set
* Do the thing mario did in class
* General cleanup; delete stuff that is missing
* (betting data will be done over break by Henoc <3)

5 datasets & their vars:

* Teams
  + ~~MIN\_YEAR~~, ~~MAX\_YEAR~~, ABBREVIATION, ~~NICKNAME, YEARFOUNDED~~, CITY ARENA, ~~ARENACAPACITY~~, OWNER, ~~GENERALMANAGER~~, HEADCOACH, DLEAGUEAFFILIATION
  + All of these can be aggregated to games if necessary
* Ranking
  + TEAM\_ID, LEAGUE\_ID, SEASON\_ID, STANDINGSDATE, CONFERENCE TEAM, G, W, L, W\_PCT, HOME\_RECORD, ROAD\_RECORD, RETURNTOPLAY
  + All of these can be aggregated to games (they’ll just be repeating)
* Games
  + GAME\_DATE\_EST, GAME\_ID, GAME\_STATUS\_TEXT, HOME\_TEAM\_ID, VISITOR\_TEAM\_ID, SEASON, TEAM\_ID\_home, PTS\_home, FG\_PCT\_home, FT\_PCT\_home, FG3\_PCT\_home, AST\_home, REB\_home , TEAM\_ID\_away, PTS\_away, FG\_PCT\_away, FT\_PCT\_away, FG3\_PCT\_away, AST\_away, REB\_away, HOME\_TEAM\_WINS
* DO NOT USE IN PRELIMINARY MODELS: Games\_details
  + GAME\_ID, TEAM\_ID, TEAM\_ABBREVIATION, TEAM\_CITY, PLAYER\_ID, PLAYER\_NAME, NICKNAME, START\_POSITION, COMMENT, MIN, FGM, FGA, FG\_PCT, FG3M, FG3A, FG3\_PCT, FTM, FTA, FT\_PCT, OREB, DREB, REB, AST, STL, BLK, TO, PF, PTS, PLUS\_MINUS
  + This has a lot of missing vals that will need to be filled in

Variables we want to include in the dataset:

Datasets to put into our data: (put links of stuff we are going to use here)

* [Games](https://raw.githubusercontent.com/SuperMarioGiacomazzo/STOR538_WEBSITE/master/Playoffs/Round%202/Starting%20Data/games.csv) - is this spread consistent ?
* [player id](https://github.com/djblechn-su/nba-player-team-ids) - makes other datasets with players easier to merge
* <https://www.oddsshark.com/nba/computer-picks> / <https://www.oddsshark.com/nba/offensive-stats>

[Dataset that gives player and team IDs for different databases](https://github.com/djblechn-su/nba-player-team-ids) (ESPN, Basketball Reference etc.)

[Poisson process to predict outcomes of Soccer games](https://towardsdatascience.com/making-big-bucks-with-a-data-driven-sports-betting-strategy-6c21a6869171)

Use betting site data to make predictions

<https://projects.fivethirtyeight.com/2023-nba-predictions/games/>

[LSTM In R](https://www.r-bloggers.com/2021/04/lstm-network-in-r/)

[Generative model that predicts Spread, rebs, and total score all in one model](https://rpubs.com/HeatWave2019/780799)

3 Y vars

* Spread=Home Points−Away Points
* Total=Home Points+Away Points
* OREB=Home OREB+Away OREB

Model Methodology Line

1. Saturated linear regression
2. forward/backward selection
3. Interaction or polynomial
4. Random forest
5. GLM poisson
6. Mixed effects model
7. K fold cross validation
8. INLA or STAN model
9. Neural Nets (?): DCGAN or LSTMm

**lmer(spread~previous matchup spread + (1|away team)+(1|home team))**

[**Ridge Regression**](https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r)