**Titanic Kaggle Challenge**

Main mission: Use ML to make a model that predicts which passengers survived the titanic shipwreck.

* What sorts of people were more likely to survive?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Profile of survivor so far:

* Female
* 1st class
* Departed from???
* Married??
* Kids??
  + Age of kids?
* Position on boat in 1st class cabin?
  + Was there a ticket location that had highest chance of survival??? (Back of boat 2nd class for example?)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* Clean data:
  + ~~Drop unnecessary columns.~~
  + ~~Deal with NaN values with imputing (k-NN)~~
    - ~~Age - 177~~
    - ~~Embarked – 2.~~
    - ~~Cabin - 687~~
  + ~~Change age to whole numbers (ints)?~~ – Stick with floats, use ints if it increases accuracy
  + ~~Turn Embarking column into numbers.~~
  + ~~Turn Sex into binary~~
* Process data
  + Use one hot encoding on:
    - ~~Pclass~~
    - ~~Embarked~~
  + Impute NaNs?
* Titles and names from marriage, if your married are you more likely to survive?
* If you have kids, are you more likely to survive?
* What age should your kids be to survive?
* Ticket number indicates room??? (place on boat)
* Create:
  + Functions that we can apply to both the train and test.
  + If imputing data, then create copy. Then we can run model on both easily.
    - Have two data frames, original (with NaNs) and imputed data.
* EXTRA: Clean then group ticket data
  + Group tickets by 100s
  + Do they have letters?
  + Ticket fare?
  + Married
* Create these models:
  + KMeans
  + Decision tree
  + Gradient boost