

# Titanic Survivorship Prediction

August 8, 2013

# Backstory

- April 15, 1912 the Titanic collided w/ an iceberg and killed 1502 of 2224 passengers & crew
- Not enough lifeboats
- Most that made it on to lifeboats were women, children, etc. (misinterpreted orders)

# What can the data tell us?

- Explore data - what relationships are there?
- What outliers do you see?
- Any particular group that was “doomed” or “protected”?



# What can we infer?

- If we know (or think) we have an idea of correlations, what can we infer?
- Could we predict one variable from another? e.g. survivorship from sex

# The Problem: Predict Survivorship

- Given a training set of data, predict whether people survived or perished in the crash
- <https://github.com/dsindy/kaggle-titanic>
- <http://stetzer.github.io/2013/06/14/random-forests-of-titanic-survivors.html>

# Baseline: Random Forests

- Baseline R implementation of random forests using features Sex, Pclass, Embarked, Fare, SibSp, and Parch should receive accuracy of  $\sim 77.5\%$
- Random forests have the additional benefit of not having to separate training data into training + test to get idea of accuracy



# Where can we improve?

- Alternative algorithms? e.g. SVMs
- Additional feature extraction? e.g. Title
- Hybrid approach?
- Suggestions?

# Results

- TBD