**Capstone Project Proposal**

**Mentor**: Raj Bandyopadhyay

**Mentee**: Ripu Jain

**Date**: 04/27/2016

# **To Movie or Not To Movie**

A movie revenue predictor based on variables like budget, genre, director rating, cast rating, expected movie rating, release month, # of theatres released in, etc

## Client

Movie Studios and Investors will be able to make critical business decisions such as whether to proceed with funding and producing a movie or not, by being able to predict and visualize potential revenue a movie can earn, based on certain parameters such as budget, genre, release month, etc. This prediction model will be based on historical revenue data of movies and the factors involved generating that revenue for the movie.

## Data

[IMDB](ftp://ftp.fu-berlin.de/pub/misc/movies/database/), [Boxoffice](http://www.the-numbers.com/movie/budgets/all), [Ratings dataset](http://grouplens.org/datasets/movielens/latest/), [Tweets Rating dataset](https://github.com/sidooms/MovieTweetings)

Approach

* Collect and merge all the data - this will require significant wrangling.
* Scope the problem down to make it both more tractable, as well as relevant.
* Model it as a regression problem to predict revenue (in $). Train the model on an earlier subset of a data set and test it on later subsets.
* Evaluate effectiveness of the model.

## Deliverables

Code, Report, Slide Deck

# **Frequent Flying Tweeter**

A real time tweet sentiment classifier for airlines, based on past airline-related tweets categorized as positive, neutral, negative.

Client

A minor customer service error such as running out of soda on flight to a major problem such as flight getting cancelled can send the Twitterverse into a frenzy of complains about the airline. Airlines can use this app to analyze sentiments of airline related tweets and be able to classify them as Neutral, Positive or Negative, and take necessary action in timely fashion. The sentiment analysis of tweets will be based on historical airline related tweets by users, which are already classified as Neutral, Positive or Negative.

Data

[Airline Twitter sentimen](http://www.crowdflower.com/data-for-everyone/)t

Approach

* This data set may not need a lot of wrangling
* Expect to spend a lot of time in exploration and feature design
* Model as a 3-class classification problem (or a 1-vs rest).
* Evaluate effectiveness of various classification techniques

Deliverables

Code, Report, Slide Deck