DATA 643 Project 1 | **Global Baseline Predictors and RMSE**

In this first assignment, we'll attempt to predict ratings with very little information. We'll first look at just raw averages across all (training dataset) users. We'll then account for "bias" by normalizing across users and across items.

You'll be working with ratings in a user-item matrix, where each rating may be (1) assigned to a training dataset, (2) assigned to a test dataset, or (3) missing.

	Signal and the Noise	Weapons of Math Destruction	The Undoing Project
Hilary Mason	5	4	1
DJ Patil	4	4	?
Hadley Wickham	2	?	5

Please code as much of your work as possible in R or Python. You may use standard functions (e.g. from base R and the tidyverse or from a standard scientific Python tool stack). Your project should be delivered in an R Markdown or a Jupyter notebook, then the notebook should be saved into a GitHub repository. You should include a link to your GitHub repository in your assignment submission link.

Preparation. Start by watching Parts K through P from this playlist from the Coursera/Stanford Networks Illustrated course (total run time is about 22 minutes):
https://www.youtube.com/playlist?list=PLuKhJYywjDe96T2L0-zXFU5Up2jqXlWI9

- Briefly describe the recommender system that you're going to build out from a business perspective, e.g. "This system recommends data science books to readers."
- Find a dataset, or build out your own toy dataset. As a minimum requirement for complexity, please include numeric ratings for at least five users, across at least five items, with some missing data.
- Load your data into (for example) an R or pandas dataframe, a Python dictionary or list of lists, (or another data structure of your choosing). From there, create a user-item matrix.
- If you choose to work with a large dataset, you're encouraged to also create a small, relatively dense "user-item" matrix as a subset so that you can hand-verify your calculations.
- Break your ratings into separate training and test datasets.
- Using your training data, calculate the raw average (mean) rating for every user-item combination.
- Calculate the RMSE for raw average for both your training data and your test data.
- Using your training data, calculate the bias for each user and each item.
- From the raw average, and the appropriate user and item biases, calculate the baseline predictors for every user-item combination.
- Calculate the RMSE for the baseline predictors for both your training data and your test data.
- Summarize your results.

You may work in a small group (2 or 3 people) on this assignment.