Lab 4E - This model is big enough for all of us!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

## Building better models

* So far, in the labs, we've learned how to make predictions using the *line of best fit*
  + Which we also call *linear models* or *regression models*.
* We've also learned how to measure our model's prediction accuracy by cross-validation.
* In this lab, we'll investigate the following question:

*Will including more variables in our model improve its predictions?*

## Divide & Conquer

* Start by loading the movie data and split it into two sets (See Lab 4C for help). Remember to use set.seed.
  + A set named training that includes 75% of the data.
  + A set named testing that includes the remaining 25%.
* Create a linear model, using the training data, that predicts gross using runtime.
  + Compute the MSE of the model by making predictions for the testing data.
* **Do you think that a movie's runtime is the only factor that goes into how much a movie will make? What else might affect a movie's gross?**

## Including more info

* Data scientists often find that including more relevant information in their models leads to better predictions.
  + Fill in the blanks below to predict gross using runtime and reviews\_num.

lm(\_\_\_\_ ~ \_\_\_\_ + \_\_\_\_, data = training)

* **Does this new model make more or less accurate predictions? Describe the process you used to arrive at your conclusion.**
* **Write down the code you would use to include a 3rd variable, of your choosing, in your lm().**

## Own your own

* **Write down which other variables in the movie data you think would help you make better predictions.**
  + **Are there any variables that you think would not improve our predictions?**
* **Create a model for all of the variables you think are relevant.**
  + **Assess whether your model makes more accurate predictions for the testing data than the model that included only runtime and reviews\_num**
* **With your neighbors, determine which combination of variables leads to the best predictions for the testing data.**