**W261 HW1, Megan Jasek**

HW1.0.0. Define big data. Provide an example of a big data problem in your domain of expertise.

Big data is when a person or company has data that is too big to store and process on the system that is available. In terms of storage, if one has too much data to store on computers that are available now, then a machine with enough storage that will hold all of the data will need to be obtained or multiple machines will need to be obtained and the data will need to be divided among the new systems. In terms of processing, if the processing time is too slow to process the data that one has, then the data is considered big data. Add these terms and define them. Volume, velocity, variety and veracity.

Possible type example:

After incurring low win rates for new client acquisitions, a financial services firm turned to big data in order to better identify which new client opportunities warrant the most investment. The company supplemented its customer demographic data with third party data purchased from eBureau. The data service provider appended sales lead opportunities with consumer occupations, incomes, ages, retail histories and related factors. The enhanced data set is then applied to an algorithm which identifies which new client leads should receive additional investment and which should not. The result has been an 11 percent increase in new client win rates while at the same time the firm has lowered sales related expenses by 14.5%. - See more at: http://www.crmsearch.com/retail-big-data.php#sthash.4MGBIIA6.dpuf

Generally, one could say that data is big when it is too big to store on one computer: a laptop where one can store about 1 terabyte of data or a server where one could store about

Processor, storage, time

HW1.0.1.In 500 words (English or pseudo code or a combination) describe how to estimate the bias, the variance, the irreducible error for a test dataset T when using polynomial regression models of degree 1, 2, 3, 4, 5 are considered. How would you select a model?

Type out the formula for bias, variance and irreducible error with variables for the degree of the polynomials

Without any additional information I would select a model based on a balanced variance and bias ??and error. ??maybe there are applications where a higher/lower variance/bias might be a good thing.