# W4249 Intro to Data Science

## Homework #2

## 1. Sergei's problem

Dataset: http://bit.ly/cufeb10data

| | |
|---|---|
| impression_id | - unique impression id |
| user_id | - integer representing a unique user |
| day_of_week | - a string representing the day of the week |
| hour | - the hour of the day (0 to 23) |
| site_id | - an integer id of a particular site where the ad was displayed |
| ad_size | - size of the creative (image) in pixels |
| browser_id | - an integer representing the user's web browser |
| state | - 2-letter abbreviation of the state for this zip code |

The data consists of categorical variables with values from events. We would like to create a (sparse) binary matrix with dummy variables coding for levels of categorical variables.

**format**: row_id, column_id, value (see package Matrix)

For small number of categories (everything except site) use all categories, e.g., 24 columns for 24 hours. For large number of categories (site_id)

- use the top sites covering 99% of data points
- create an "other" column

Columns should be numbered sequentially (keep a map).

*Store this dataset in this format for easy retrieval, we will use it later.*

## 2. T & H, page 39, ex. 2.2

Assume the training data set similar to the one in the text, i.e. a mixture of normals with normal means but with 3 colors/ categories. Show how to compute and fit the Bayes decision boundary for the simulation example. Also fit the regression and nearest neighbor boundaries. Do the simulation and the plots. Comment.