# STATISTICS W4242 - INTRODUCTION TO DATA SCIENCE
**Grading Policy**

Basis for grades:
1) Homework assignments (40%)
2) Final project (40%)
3) Final (in-class) exam (15%)
4) Class community participation (5%)

**1) Homework assignments**
Some homework assignments will be completed individually, but for most assignments you will *have the choice* of working in teams. One of the major themes of the class is that most real-world data science requires teams, and we would like to encourage that in the assignments. On the other hand, we recognize that this is not quite the real world, and individual schedules, preferences, and other constraints might limit a person's ability to work in a team.

Whether working as a team or as an individual, the requirements and grading method described below will be applied in the same way, with the exception of F. Write-ups should include a brief description of your goal and the approach you take (*e.g.*, "we implemented Naive Bayes with the goal of creating an effective spam filter"); relevant background information (*e.g.*, "an important measure of effectiveness in the real estate market is time on market"); discussion of the data and how that affects your analysis (*e.g.*, "the data were generated by users navigating Company X's website, so we're interested in examining their behavior in the following ways..." or "the data are from DNA microarray experiments on samples from different patients..."); the results of the analysis; and conclusions. Feel free to structure the write-up in a way that you feel best presents the information. Assignments will be scored equally along multiple dimensions:

A. **Domain expertise**. Does your write-up include relevant background information? Have you applied that information in a way that informs your analysis?
B. **Technical implementation/analysis**. Does your code do what it should do? Does it produce useful results? Is your code clearly organized and easy to interpret?
C. **Data visualization**. Are the results of your analysis visualized effectively? Have you found a balance between too much and too little? Does each visualization tell a relevant story?
D. **Communication/presentation**. Is your report well-organized? Is it concise? Does it present the data/analysis/visualization in a cohesive, informative way?
E. **X-factor**. Not everything can be captured by four categories. This will attempt to measure the otherwise unmeasured.
F. **Personalized score based on peer feedback**. See below.

The score from each dimension will be combined into an overall score.

Write-ups should include the main write-up and two appendixes: one that includes a printout of any code written for the assignment (please comment and format in Courier font with a minimal amount of wrapped lines); the other is one of: a team skills inventory and a summary of how the team worked (see below); or a personal skills inventory and narrative of your approach to the assignment and any problems or limitations you had to overcome. In either case (individual or team), the second appendix should be concise; the intention is to get you to reflect on the process and to record your successes and shortcomings in a useful way. Groups need only turn

in one write-up for the entire team. Please include names and UNI's at the top of the write-up.

**Peer feedback / teamwork**. At the start of each assignment, teams should discuss and record a detailed set of responsibilities and what is expected of each team member. As a part of this, teams should create an inventory of team members' strengths and the areas they would like to develop as part of this class, and use this inventory to guide in deciding each member's role. A natural way to divide responsibilities is by each of the grading categories, but keep in mind that there is substantial interplay between each one. The team should not function as a set of independent workers who combine disparate products into a report. Rather, there should be frequent communication amongst team members. Domain expertise informs analysis; analysis and visualization inform each other, etc. With this in mind, team members should have clearly defined roles that may evolve throughout the assignment. There is plenty of flexibility here -- utilize strengths and give members the opportunity to develop skills they don't already have. (*Example*: someone skilled in programming but wanting to develop visualization skills could work closely with someone who has experience with visualization).

For each assignment, one person will be responsible for documenting each team member's responsibilities and, ultimately, their contributions to the final product and how well they fulfilled (or exceeded) expectations. This will involve recording the initial roles/expectations (and how those roles evolved, if at all), collecting and summarizing feedback from team members, and documenting any problems -- technical or otherwise -- that arose. There aren't any strict requirements on format, so long as the relevant information is communicated. Feel free to be creative! We (the instructors) are extremely interested in how the teams interact and accomplish their goals, so this is a valuable exercise. Teams can decide how to distribute this work -- it might rotate amongst team members, or it could be the same person from week to week.

**Submitting your assignment**. Assignments are due by 6:00 pm of the due date. Write-ups should be emailed to Ben (<reddy@stat.columbia.edu>) with the subject line "W4242 Homework #N", where N is whatever number assignment is being turned in. (Submissions without this subject line are liable to end up in Ben's spam filter.) Please include names and UNI's in the body of the email.

**2) Final project**.
The final project will be a Kaggle competition that incorporates a machine learning challenge with a visualization/interpretation component. More information will be provided later in the semester.

**3) Final exam**.
We are required to give a written final exam. It will primarily be a test of whether or not you were paying attention in class.

**4) Class community participation**. One of the primary goals of this class is to foster a data science community here at Columbia. In order for this to happen, your (enthusiastic) participation is necessary. The grading criteria for this category is intentionally vague; commenting on the class blog, talking in class, suggesting links to be posted on the class blog, providing solutions to technical issues -- all of these are useful ways to participate. But don't limit yourself to these things! If you have an idea for how you could participate, talk to the instructors about it.