Class Project Instructions -- Practical Data Science, Fall 2012

Your class project will be on a topic of your choice. It will be undertaken in groups, selected as described below. You will envision, design, execute, and report on a data-science-oriented study based on some interesting data available from the web, from your company, or from elsewhere. You will propose the project to us about ⅓ of the way through the class, based on a 2-page written proposal. We will give you feedback. Once the project is approved you will proceed in earnest.

The main requirement of the project is that it exercise concepts covered in the class. Generally, you should do some significant subset of the following:

- get some data, ideally in a non-trivial manner
- process the data
- build a non-trivial system for storing, querying or otherwise managing data.
- produce some graphs to support some analysis
- do other interesting visualizations
- discover interesting relationships or phenomena
- build some model or models
- evaluate the models, both quantitatively and visually
- operationalize your results in some way: make them accessible in some system
- solve a real problem, do something practically useful or desirable, etc

Don't worry too much about covering all those things. It is more important for you to work on a project that you are excited about (and can excite us about) or one that would be particularly useful. For example, in the latter case you might build a significant component of a project you are working on professionally (that you can share with your team, of course).

⇒ You will turn in a 2-page proposal to us no later than Class 4 (email to Foster & Josh is fine, with your group number included). The proposal should present the basic question/questions to be investigated, the planned data sources, any preliminary work on obtaining the data, and your ideas for planned analysis and expected results. Please feel free to interact with us before then to develop your ideas. Also, please feel free to look forward in the syllabus and book to see the sorts of concepts that we will cover. Also, search the press for articles about what companies are doing with data science. We will provide additional resources along these lines, linked to the class web page.

⇒You should plan on keeping us in the loop as your projects progress. Remember that we have a good deal of experience with data science projects, and may have useful suggestions. Also, we would like to see that you are actually working on your projects, and not just waiting until the last minute. Data science projects have a reputation for having unforeseen complexities--you don't want to discover that you're going to need to rethink a major component a week before the deadline.

⇒You will turn in a status report no later than Class 9 (email to Foster & Josh is fine, with group number included); the status report should present the question(s) you are addressing, describe the data activities to date, and provide some preliminary results.

⇒You will produce a final 15-20 page report documenting your findings, plus one or more appendices detailing (i) how you did each of the above-mentioned steps, (ii) any auxiliary analyses or visualizations that did not fit into the main body of the report, (iii) all code written,

and (iv) a detailed description of each group member's specific contributions to the project. The due date for the report is listed on the syllabus schedule.

Please feel free to ask us in class to give examples of what might be interesting project ideas.

Groups will be formed as follows. We will “seed” approximately 16 groups, each with one class member whom we have chosen based on the responses to the class survey. Then the rest of the class will self-select into these groups. Each group will have 3 members, the seed plus two others. The seed is not (necessarily) the group leader. The seed simply is a class member we have chosen, based on his or her stated skills.