MEMORANDUM on Projects **DIRE WARNINGS BELOW- READ CAREFULLY**

To: Students in Big Data Analytics BDA18

Subject: Managing a data mining project for speed.

From: Your de facto boss's boss for the course, Roger Bohn

Date: May 3, 2017; updated June 5, 2018

Bottom line message:

You are probably skirting the edge of a large, dangerous, hole! For various reasons, students tend to work first on creating a "perfect" data set. Only after it is finished do they start to think about the problem and analyze the data. This is exactly backwards, and it generally leads to serious problems. STOP working on perfect data. Take your existing, incomplete, data and start using it in a crude model!

Introduction

Managing projects is a key life skill, and it's something that you will never stop improving. These memos are intended to help you manage your projects insightfully, and to learn from your management experience. They contain insights that make Data Mining projects successful overall, even though they don't correspond to a particular formula or R function. By comparison, the weekly project assignments are intended more as step-by-step guides.

Early in this course it may seem difficult to know what projects will be feasible, and therefore to write a proposal. You don't yet know what techniques will be taught, you don't know how to manipulate data in R, and so forth. It will turn out that these are *not* big difficulties in successfully completing projects. Rather, the big issues are:

A. Can you find, or create from the web, a large data set with interesting variables in it? The best data sets have event-level data, not aggregated data. For example for crime, there is an entry for every reported crime. For e-commerce, there is an entry for every transaction, or every item in the catalog, or every customer. (All 3 would be ideal.) This is now easy - there are huge data sets, publicly available on numerous topics.

Is the data suitable in various ways? Not confidential, it must be clean or cleanable, etc. It does not have to already be in the right format, just some format that you can get your hooks on.

B. Do you have some interesting questions/issues to investigate that this data contains information about? This is limited mainly by your imagination and your search skills in Proquest and Google Scholar. Look for papers about analogous issues in other countries/industries/data sets. Search their references (backward linking), and papers that reference them (forward linking).

C. Specific mining techniques, like Random Forests versus Nearest Neighbor, are just tools, and they will *not* make or break your project. Not having interesting questions can break your project.

D. Once you are moving, the biggest issues for most teams are:

1. Project- specific data mining concepts and techniques. Each project relies on a few key methods that go beyond what the course covers. Examples are geographic analysis, time series, scraping data from the web, or text analysis. Find a narrowly targeted book or web tutorial that already has R code in it. Use those methods where appropriate.
2. Figuring out how to incorporate chunks of R code without having to write them yourselves.
3. Managing the research project: Assembling the data, doing the analysis, writing up your results.
4. Running out of RAM (memory) in your computer. This is actually a minor problem for almost everyone, but you will need to learn a few tricks to make it go away. Look for separate memoranda on this topic.

Skirting the Pit of Perfectionism

Here is a warning that I gave a team in week 5. This team has great data (potentially), covering multiple years with roughly 100,000 observations each year. They reported that "By our next weekly report, we hope to have merged all four years of data, and be able to produce some charts from the data." Here is what I wrote back to them:

The most important feedback is that **you are skirting the edge of a large pit**, and don't realize it! For some reason, students tend to work first on creating a "perfect" data set. Only after it is finished do they start to think about the problem and analyze the data. This is exactly backwards. You will *never have perfect data.* You can only have different gradations of incomplete data. Eventually you will be forced to give up the quest for perfection - but by then, there will only be a few weeks left.

Please immediately *stop* working on a multi-year dataset. That probably sounds counterintuitive, but there are sound theoretical as well as practical reasons for it. Fortunately, it is early enough that you won't pay a high price if you straighten out now.

Explanation: Don't put everything else on hold to finish one task, even if you currently think "I must do this." That is the "*waterfall model of software development*" (look it up), and it's a disaster for R&D projects like yours. Instead, adopt the "*incremental prototyping*" approach. That means building an early, and very weak, analysis early, and then gradually improving it. Incremental prototyping beats the waterfall every time because *as you progress, you discover that some of your early decisions were based on hidden assumptions* that turn out not to be true.

Right now, don't spend more than a few days importing and cleaning up data. It seems logical and "efficient" to assemble a full, multi-year data set first, and and only then to start analyzing. This is exactly backwards! Some teams get mentally stuck, and spend weeks "cleaning up" the data, merging different data sets together, etc.

Even if your data is still crap, and you only have 20,000 out of a planned 1 million observations, start analyzing it anyway. What you generally discover is that the elaborate plans you initially had turn out to be partially irrelevant to what is in the actual data. And, more interesting questions come up than those you originally started with.

One way to view this that you are managing to keep your boss (professor) happy at the deadline. If you got sick, and only had a few days to throw together what you have already done, would you have any *results* to report? Or would you just talk about all the great things you *intended* to do, but ran out of time for? That's the difference between a grade of B+ or a B- . And you won't get sick, so it will be the difference between an A- and a B+. (Getting an A generally takes thoroughness and diligence for several weeks. We can discuss that elsewhere.)

Why you will probably ignore this advice: Unless and until you have lived through a few project "overruns," it is easy to think that managing R&D projects (such as in Big Data) is

like managing a term paper, or a conference, or some other relatively predictable event. Research is not like normal production!

There are also natural feelings such as:

- I'm too smart to let that happen.
- The data we are working on now is more messed up than we expected. But the other data that we need for this project comes from a government agency, so it will be perfect.
- At least I understand what I'm doing with the data. Other parts of the project are unfamiliar and scary, so I instinctively postpone them.
- I only need another day or two to finish this piece of the project.
- My other courses and life obligations won't take much of my time.
- My roommate has a friend who can fix everything. (Certainly you should ask them; but don't make your survival conditional on their expertise!)

If you hear any of these in your vicinity, call immediately for some management counseling!

I hope that this year, for the first time, every team will prove me wrong!