Learning from the World Bank's "Big Data" Exploration Weekend

by
Dennis D. McDonald, Ph.D.[1]
March 21, 2013

Big Data Weekend

If you're serious about data analysis there's probably no substitute for getting "down and dirty" with real, live, messy data. Sometimes you just have to sift through the numbers with your "bare hands" if you really want to extract meaning from descriptive statistics, predictive models, and fancy visualizations.

I was reminded of all this over the past weekend when I was privileged to participate in a "data dive" at the World Bank in Washington DC designed to quickly analyze financial, social, and survey data relating to poverty and development project fraud. I had fun, met some very smart people, and I learned a lot in the process.

The Problems

The weekend's activities were sponsored by the World Bank and partners UNDP, UNDB, UN

[1] *Copyright (c) 2013 by Dennis D. McDonald, Ph.D. Dennis is a Washington DC area consultant specializing in collaborative project management and new technology adoption. His clients have included the US Department of Veterans Affairs, the US Environmental Protection Agency, Jive Software, the National Library of Medicine, the National Academy of Engineering, Social Media Today and Oracle, and the World Bank Group. His experience includes the management of projects involving the conversion or migration of financial and transaction data associated with large and small systems. Contact Dennis via email at ddmcd@yahoo.com or by phone at 703-402-7382. The current version of this document is located here.*

Global Pulse, and Qatar Computing Research Institute (QCRI). About 150 participants (analysts, economists, programmers, statisticians, database managers, finance managers, and many others) split into eight teams to attack the following problems defined in advance by the World Bank's Open Finances team and DataKind, a nonprofit organization specializing in data and development projects:

1. **Web scraping global food prices.**
    - Problem: Can an "early warning system" for measuring food price inflation be developed from food prices published on the web?
    - Data: Food products tracked included bananas and rice. Multiple data sources were examined including Twitter, an Indonesian food store chain's published price list, and cached webpages. The project team used GitHub to share its data among team members. Several first-timers got an a opportunity to learn how to "scrape" websites.
    - Results: Yes, it is possible to track food price trends that anticipate official government price report.
2. **Arabic Tweet Analysis.**
    - Problem: can details of factors such as sentiment and gender be derived from an analysis of tweets in Arabic?
    - Data: This team started with a database of 100 million tweets in Arabic and investigated, using a variety of tools, whether was possible to determine factors such as emotional state and economic status of the source.
    - Results: The two Arabic speaking members of the team worked very hard with the rest of team members to develop rules for scraping and classifying words and phrases used in tweets. One complication: men and women tweeting in Arabic often use the same terms differently.
3. **Correlating Ground Based Lighting Data with Poverty Rates.**
    - Problem: is it possible to predict changes in poverty rates based on an example of geo-coded lighting data as observed from space?
    - Data: This team made extensive use of geolocation based data on poverty in relation to satellite observation of night lights from space.
    - Results: Even controlling for availability of electricity, it is possible to improve predictions of poverty rates based on standard types of social and economic data by also including lighting data.
4. **Using mobile tech to collect development data.**
    - Problems: Can mobile phones be used to obtain reliable census-type data?
    - Data: Official Peruvian data sets were obtained that provide answers, based on in-home census interviews, about health, employment, and other household events. These data were compared with data collected via cell phones for the same questions administered via phone interviews and via automated IVR-based interviews.

- Results: When the data sets were compared it appears that the cell phone data and responses provide a more valid picture of responses to health and employment questions than the personal in home interviews.

5. **Risk factors and development project outcomes.**
    - Problem: World Bank projects are typically scored by whether or not they accomplished their objectives. Usually they do, but it appears that between 10 and $12 billion per year are being spent on projects that appear to be underperforming.
    - Data: This data analysis focused on network and cluster based comparisons of high and low performing projects based on sponsor supplied project descriptions.
    - Results: Different cluster analysis techniques were used to begin identification of factors that differentiate between high performing and low performing projects.
6. **Heuristic auditing tool.**
    - Problem: Can a set of key project descriptors, when compared against a database of the ranges these values can take on measurement scales ranging from legitimate to fraud prone, be used to help evaluate potential projects for fraud?
    - Data: This project is built on a web-based project data collection process that lets a user upload project data for comparison with a series of variables that provide correlation with potential fraud and corruption measures.
    - Results: By focusing on ease of use and transparency of the rating process, one goal is to transform victims of fraud into "agents of change."
7. **Detecting Fraud and Corruption.**
    - The problem: The World Bank is always on the lookout for organizations that must be disqualified from bidding on development projects. It has been found that, despite being disqualified from bidding, companiesand staff members might "collude" with other organizations to continue bidding.
    - The data: This team looked at the identification of possible project involvement by disqualified organizations. The effort involved "scraping" relevant project data from 5000 online pages of World Bank World Bank project data. Another examined linkages between staff on disqualified companies and other companies that might be getting involved in development projects
    - Results: public data from sources such as Linkedin can help identify linkages between banned organizations, their employees, and attempted proposals associated with those same staff or related to them.
8. **UNDP Capacity Problem.**
    - Problem: Every year the UNDP supplies funds for local development projects around the world. Not all projects spend the money or complete the projects. This is a problem. Are there staffing characteristics related to this?

- Data: The UNDP supplied five years of development project data along with a separate file containing anonymized staff data for the offices around the world that run the projects.
- Results: After spending some time on cleaning and merging the data the team developed a series of presentations that visualized the relationship between project expenditures and budget overhead consumption. Data visualization software allowed for display of complex ratio relationships

Lessons Learned

I've been working with some friends on a proposal with a DC-area nonprofit for development of an index that tracks performance against a set of defined social welfare priorities. I've also been researching government program transparency. This weekend project seemed a good opportunity to experience firsthand what's involved in so-called "big data" analysis, especially since I've expressed some concerns about that topic and what I viewed as over enthusiasm from those who might not have personally experienced the trials and tribulations of having to deal with masses of live -- and messy -- data.

Having spent a good chunk of my professional life on statistical and survey types of projects I was pleased and impressed at the tools and data sources brought together over the weekend by a diverse and dedicated crowd of young people. Our ability to scan, crunch, clean, merge, manipulate, and visualize large data files so rapidly was most impressive. Also impressive was the ease with which data files can be shared. Folks used a wide range of collaborative tools for sharing including GitHub, Google Drive, and DropBox. Language hopping is also common; having once designed a survey of scientists and institutions in Egypt that was administered in both English and Arabic I was amazed at how far our tools have advanced for manipulating all kinds of data.

Some members of my group (UNDP Capacity Problem) jumped immediately into analysis. Others -- like me -- were a bit more formal in first seeking out descriptive statistics to get a handle on general trends in the major variables as well as a firsthand view of missing values, outliers, and coding challenges. A range of tools were used by the group. Eventually a subgroup formed to clean and merge data.

Had I been wearing my "project manager" hat I might have had the group hold off a bit longer on correlation and modeling efforts, at least till we resolved the distribution shapes for major variables and decided -- as early as possible -- what to do with missing or bad data. In my experience a lack of standardization of how to deal with weird, messy, and missing values can lead to very different conclusions given that different analysis efforts might actually be focusing on different data sets.

Holding back would have been a mistake. The analysis was a group effort. Folks were eager to crunch some data using available tools. I recognize the value of that. The saving grace was that the group was small enough to be sitting around the table where a rapid sharing of findings or problems could be surfaced. In my opinion, this reduced the need for stricter or more formal controls.

Put another way, in an "agile" effort like this you have to be willing to try a variety of approaches quickly. You also have to be willing to tolerate some inefficiency which arises when, for example, dirty data gets swallowed up by the analysis or people start inconsistently recoding or handling missing data differently. Collaboration becomes paramount in being able to root out and fix such issues.

Still, a key issue with a weekend effort like this concerns sustainability. How will the effort move forward? Can it be sustained when most of the participants have "day jobs" they need to return to Monday morning? Have any real conclusions been reached, or have we just planted the seeds for further research?

I would be more concerned if the sponsoring organizations didn't include organizations like the World Bank and the UN. These organizations, along with DataKind, put serious planning and thought into preparations for this effort, and they aren't going away. The participants during the weekend also included active participation by senior people with a stake in the weekend's analysis. This bodes well for follow on in some fashion, I think, and I intend to stay in touch with those involved.

Conclusions

One of my main takeaways is that there is value from throwing a group of talented people at a series of challenging projects and giving them an environment and support for creative thinking. At the same time, there is also a need to provide decent data, clear goal statements, and an effective collaborative environment with an appropriate balancing of structure and creativity.

For the most part I think this weekend exercise was a success and I look forward to understanding how this learning can be integrated with other efforts with which I'm involved.