DATAFEST: CELEBRATING DATA IN THE DATA DELUGE

Robert Gould
Department of Statistics, University of California, Los Angeles, USA
rgould@stat.ucla.edu

DataFest is an undergraduate competition in which student teams have just 48 hours to find and communicate meaning in a rich, complex data set. Many of the skills and practices of data science—working collaboratively with a team, organizing raw data, dealing with non-traditional data types, sorting through datasets with hundreds of variables—are hard to teach in a classroom setting. Assigning projects is one approach, but in our experience, many student projects were far below the level we hoped that students would achieve. DataFest, which now includes participants from fifteen U.S. colleges and universities, provides an opportunity for students to challenge themselves with realistic, large data sets in an intense, fun, and encouraging environment.

INTRODUCTION

The practice of statistics is changing in reaction to what some have called the Data Deluge (see *The Economist*, 2010). Modern economies require data in order to grow and survive, and the analysis of data is not just of interest to statisticians working in remote agricultural stations or government bureaucracies, but arguably of great interest to all sectors of the economy as well as to our social and cultural lives (Gould, 2010). A hallmark of the Data Deluge is what some have called Big Data, a subjective and contested term which I choose to interpret as "complex and rich data". For the most part, richness enters through the number of variables represented explicitly or implicitly within the dataset. And complexity comes from the structure of the data, which may be a continuous stream fed from sensors, may be highly nested, may represent complex relationships between objects, may represent dates and times of day, may represent locations, photos, or texts. One might go so far as to state that any activity in which we humans engage is represented somewhere in the virtual sphere with data.

The complexity of Big Data poses a challenge to the educator that can't be solved merely by updating the curriculum. Some guidelines to undergraduate (and graduate) curricula recommend providing students with exposure to realistic problems with real data, and acquiring such data and providing them to students is a considerable undertaking. In this paper I describe DataFest, which is one approach to providing students with a meaningful encounter with Big Data in a supportive, highly charged educational environment.

DATAFEST STRUCTURE

DataFest is modeled after the hackathon. Hackathons have become an integral part of the computer science culture. Although hackathons vary greatly, in essence a hackathon consists of a large number of people gather for a short amount of time (typically 24-48 hours) who work around the clock to solve a particular problem or general class of problems. Their popularity at undergraduate institutions in the U.S. is spreading. One writer describes them as becoming "the intramural or club sport of the 21st century." (Mathews, 2014). Students must form a team to participate in DataFest. The teams gather on a Friday evening and are introduced to the dataset and given a general charge or set of open-ended questions or directives. They work somewhat furiously until 2pm on Sunday, when they must stop work and present their findings to a panel of judges. They are allowed only two "slides" and 5 minutes for their presentation.

Prizes are awarded in three categories. The Best Insight award is given to the team that shows the judges something interesting and surprising that was gleaned from the data. Best Visualization is awarded to the team with the most successful graphics, and this usually requires a set of graphics that tell a compelling story. Finally, Best Use of External Data is awarded to the team that meaningfully enhances the context of the original data by merging it with a dataset that they found on their own.

Critical to the success of DataFest are volunteers, who wander the room to answer (and ask) questions, steer teams along productive paths, introduce teams struggling with similar problems, and generally act as a resource. These volunteers are mostly graduate students and

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

faculty, but a large percentage of them are data professionals from the community, recruited through the local chapters of the American Statistical Association, the local R Meetup Group (an informal organization that meets monthly to discuss the statistical software R), and through word of mouth. These visitors play a crucial role in the success of DataFest, as we will discuss below.

WHAT MAKES A SUCCESSFUL DATASET

Complexity and richness are essential ingredients for a DataFest dataset. In addition, the context of the data needs to be accessible to undergraduate students. Ideally, the "owners" of the dataset can themselves present the data to the students and make it clear that they really will listen to the students' findings. It is essential that someone familiar with the data be available (either in person or through email) to answer questions as they arise. Finally, the problems posed cannot be close-ended. For instance, one year the dataset (from eHarmony, a match-making website) begged to be presented as a prediction problem. (For instance, could the students develop an algorithm to predict which matches would succeed?) However, we worked with an eHarmony data scientist to craft a more generally worded challenge so that (a) students could participate if they did not yet have the technical knowledge to build predictive models and (b) so that students could perhaps surprise us, and themselves, with a surprising and insightful approach to the data. Students become fully invested into DataFest when they are essentially working to solve their own problems, answering questions that they themselves have crafted. In this sense, they are acting as true data scientists.

The first DataFest was in 2011 on the University of California, Los Angeles (UCLA) campus and involved roughly 30 students. The data were presented by Lt. Thomas Zak of the Los Angeles Police Department, and consisted of police reports on every criminal arrest made in Los Angeles for the last five years. The winners of the Best Insight award provided evidence that homicides are lower, controlling for external factors, in neighborhoods with active community centers, and the Best Use of External Data award went to a team that found publicly available maps of gang territories and created graphics to show that homicides tend to congregate on the boundaries of these territories.

Subsequent DataFests included data from Kiva.com (data about who contributes towards microloans to entrepreneurs in developing countries), eHarmony (data on 10 million hopefuls looking for long-term relationships) and, this year, GridPoint (hourly data on energy consumption for 100 commercial buildings around the United States over a three-year period.) The Best Visualization Winner illustrated how air conditioning and heating were used based on time of day, outside temperature, and type of business (retail, full-serve dining, fast-food dining) (Figure 1).

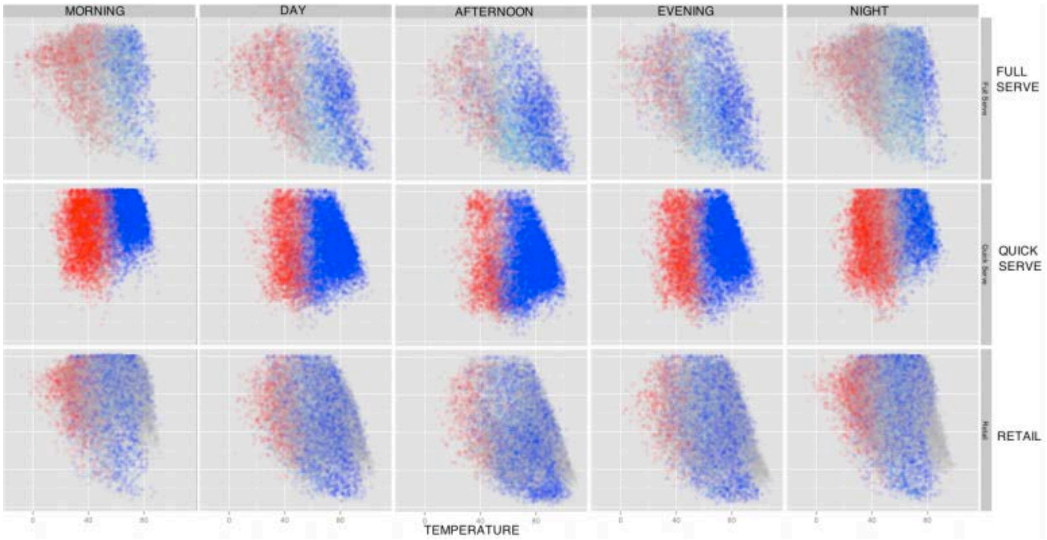

Figure 1. One panel from the graphic that won Best Visualization Award at UCLA DataFest 2014, by students Lev Golod and Jonhngyun Lee.

LEARNING COMMUNITY

DataFest provides an opportunity for students to learn from each other. This learning happens before, during, and after the event. Before the event, students seek teammates and share ideas and strategies. Successful teams sometimes include members from a non-statistics discipline such as mathematics, computer science, engineering, or even anthropology, and this teaches students about the usefulness of non-statistical skills. During the event, students often teach each other, and learn from our volunteers, new statistical methods, R packages, and programming languages. Li (2014) describes two sets of skills needed for working with Big Data: statistical skills and engineering skills. Students seem particularly adept at teaching each other the necessary engineering skills needed to help them solve problems. After the event, students often will exchange notes about what was and was not successful, and will make decisions about what they should learn in preparation for the following year.

The learning community extends to our own faculty, who have used DataFest as an informal formative assessment tool. After the first year, several of us recognized that the students were not as proficient at using statistical software as we had believed, and so we each added explicit software components to our upper division courses. This year, we discussed that students did not seem to be familiar with some basic strategies for confronting large data sets, and are now discussing how and where we can work this into our curriculum.

CAREER FAIR

DataFest also functions as an informal career fair. One reason we are able to get professionals to volunteer their time on the weekend is that it gives them a chance to see students working as part of a team while under deadline pressure. In other words, they get to watch students work in a situation that simulates the real world. Amy Deora, a Senior Manager at Summit Consulting, LLC put it like this:

> DataFest has been a very successful way for us to meet students who not only have the technical skills we need, but also have a true love of working with data in creative ways, and have a great work ethic. We have successfully hired full-time analysts as well as interns, and they have all made great contributions to Summit. (Personal Correspondence)

From the students' perspective, it gives them an opportunity to see how professionals from a variety of sectors think about problems, and to ask them questions in a very informal setting.

THE FUTURE OF DATAFEST

The 2012 report of the ASA Workgroup on Master's Degrees (Bailer, 2012) made several recommendations for educating future statisticians. These recommendations emphasized, in addition to mastery of core statistical understandings, the importance of skills in programming, communication, collaboration, teamwork and leadership. Recommendation 5 was that "Students should encounter non-routine, real problems....". While these recommendations are directed towards master's degrees, I feel they also hold true for a curriculum in undergraduate statistics. DataFest emphasizes to students the importance of these non-statistical skills for solving problems.

The first DataFest in 2011 had about 30 students participating. This most recent DataFest had roughly 400 students from 15 colleges and universities at five different locations in the U.S., all analyzing the same dataset. DataFests were hosted by Duke University (led by Mine Cetinkaya-Rundel), the Five Colleges of Massachusetts (Andrew Bray and Ben Baumer), Princeton University (Philipe Rigolett), and Emory College (Shannon McClintock). At all locations, these lead faculty were assisted by many others and by graduate and undergraduate students. This year, the American Statistical Association (ASA) agreed to be the "headquarters" of DataFest and I, along with the core team (those mentioned above in addition to several others) will work with the ASA to discuss how to help other institutions participate.

The growth of DataFest has been fueled mostly by the enthusiasm of the students who participate. I suspect that they feel a true hunger for being challenged, but within a supportive and fun environment.

PARTICIPATING INSTITUTIONS AND WEBSITES

UCLA hosted students from California Polytechnic University, San Luis Obispo; Pomona College; University of California, Riverside; and the University of Southern California. http://datafest.stat.ucla.edu

Duke University hosted North Carolina State University; Dartmouth College, and the University of North Carolina, Chapel Hill. http://stat.duke.edu/datafest

The Five Colleges: Smith College, University of Massachusetts Amherst, Mt. Holyoke College, Amherst College, and Hampshire College. http://www.science.smith.edu/departments/math/datafest/

Finally, Emory University and Princeton University each hosted a DataFest. http://www.quantitative.emory.edu/events/datafest.html, http://orfe.princeton.edu/datafest/

REFERENCES

Bailer, J., Hoerl, R., Madigan, D., Montaquila, J., & Wright, T. (2012). *Report of the ASA workgroup on master's degrees*. American Statistical Association. magazine.amstat.org/wp-content/uploads/2013an/masterworkgroup.pdf

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, *78*(2), 297-315. doi:10.1111/j.1751-5823.2010.00117.x

Economist, The (2010). *Data deluge: Special issue*. February 27, 2010. The Economist Newspapers Ltd.

Matthews, B. (2014). Are hackathons the classrooms of tomorrow? My journey to the frontier of education. *The Ubiquitous Librarian, Chronicle of Higher Education blog network.* April 28, 2014. http://chronicle.com/blognetwork/theubiquitouslibrarian/