GRADUATE STUDENT COUNCIL

Stephen Bailey
stephen.k.bailey@vanderbilt.edu

Interdisciplinary Data Science Fellowship

Problem

"Data Science" is currently one of the hottest career fields in the U.S. It was named by *Harvard Business Review* as the "Sexiest Job of the 21st Century" and is currently the #1 ranked job in satisfaction, starting salary and job availability. There is a huge interest in the field from Vanderbilt students: last October, the BRET office hosted an event entitled "Data Science in Industry" to a packed house – there was standing room only.

Data Science is fundamentally interdisciplinary, and the ideal practitioner has the hacking skills of a software engineer, the mathematical expertise of a statistician and the critical eye of a research scientist. Although data scientists perform a broad range of functions, their essential purpose is to leverage data to make informed business or research decisions. For example, an online shopping company that wants to revamp their website needs to be able to answer the question, "Does this change have any measurable impact on sales?" A wait-and-see approach could be extremely costly if, in fact, the change has a negative effect. So, a data scientist might set up an A-B test to see whether the change affects sales – or more subtle behavior, such as time on page or depth of clicking. They then clean the data, analyze the results and present it to their business audience. Similar questions in all domains -- education, social research, medicine, engineering and technology – could be asked, which is why the field is so in demand.

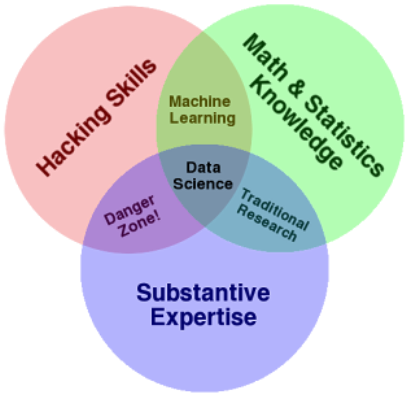

Figure 1: Data science skillsets

It is clear, then, that data scientists require skills from many disciplines; consequently, few comprehensive training programs exist. Various departments at Vanderbilt are beginning to address the demand, through the formation of a data science institute and

the offering of a “Big Biomedical Data Science” degree, for example. However, most graduate students who go on to be data scientists rely on online coursework (e.g. Coursera or DataCamp) or attend a professional “bridge” program upon graduation (e.g. Insight Data Science Fellows). These non-traditional routes are excellent ways for people to be quickly brought up to speed in multiple disciplines while building a project portfolio that can be leveraged for jobs.

One drawback to these programs is that they may rely solely on “canned” projects, whereas solving real problems with clients would provide a more valid and holistic experience. At Vanderbilt, there are hundreds of investigators across the graduate school, almost all of whom are collecting and analyzing data to some degree. Some disciplines, like those in the social sciences, are likely to have sophisticated quantitative expertise but less likely to have the expertise needed to implementing data software and analysis routines, which opens up an opportunity for collaboration between these aspiring data scientists to contribute to ongoing research goals at Vanderbilt.

Therefore, in the Vanderbilt community, we have:

1) Unmet demand for data science training and practice opportunities among graduate students in a range of disciplines
2) Ample opportunities for collaboration with ongoing research projects that have a lack of data science expertise.

Solution

A formal, Vanderbilt-based Fellowship program where selected students get basic data science training, build a foundation of skills and solve real-word problems would meet these needs.

Therefore, I am proposing a year-long Data Science fellowship, open to top students in all disciplines, for building a foundation for data science work. Different from a class, this fellowship would mirror “accelerator” programs such as Insight Data Science Fellows or Data Science Dojo, which are geared toward transitioning graduate students towards careers in industry. This fellowship would complement other BRET programs, such as the Management and Business Principles Module for Scientists, but would be more extensive in terms of the technical skills covered and body of work completed.

This fellowship’s capstone activity would include partnering with an organization at Vanderbilt to develop a fully fleshed-out data product, thereby producing value to the Vanderbilt community as well. If well-advertised and funded, I believe this Fellowship would attract some of the best students at Vanderbilt, and give them a forum to exercise their creativity, data-savvy and problem-solving skills in novel contexts. Labs could then “rent” this talent to solve problems, e.g. to develop a processing pipeline for a certain type of data being collected.

IMPACT

This fellowship would open up a new career path, especially for senior PhD students interested in data industry jobs -- the “sexiest” jobs of this century. Many of these students, especially from biological and social sciences, may have the critical knowledge base needed to be successful, but need additional formal training to be competitive in the job market. This fellowship would fill a huge gap for these students by introducing them to a community of data science researchers at Vanderbilt, and helping them build a portfolio of projects that can be used in interviews.

Individual research groups would also find in this Fellowship a valuable opportunity to “rent” data expertise, especially in departments where this is not largely present. This would be particularly useful in contexts where there is statistical, but not computing expertise. The situation might look like this: a sociology research group wants to perform a study that requires interviewing people in rural Tennessee. They work with Fellows to develop an app that can be used to record key information for each interviewee, record the interview, capture it in RedCap, and automatically summarize results in a dashboard.

GOALS

The specific deliverables of this Fellowship are:

1) Trainees that have a framework for how data scientists work through problems, and a foundation of mathematical, software and scientific expertise to build on.
2) Each trainee has a portfolio of 5 completed data science projects, some of which are “canned” and some of which are novel.
3) Partner labs/organizations have a finished data product that adds value to their mission.

IMPLEMENTATION

The Fellowship logistics are modeled off the Data Science Dojo bootcamp and the BRET module, Management and Business Principles Module for Scientists. The Fellowship will last, in total, for one school year, with the most intensive time investment coming at the beginning.

One-Week Bootcamp: The Fellowship starts at Vanderbilt with five consecutive days of training followed by one “hack” day where students are provided with real-world data science and big data problems to solve. A variety of topics such as data exploration, visualization, feature engineering, predictive analytics, predictive modeling, clustering, recommender systems, big data pipelines, event/message queues, real-time analytics, distributed databases, metrics, and A/B testing would be covered in this initial week. Course work would be completed using R and Python.

The bootcamp would occur just prior to the beginning of the Fall semester (early-mid August). Towards the end of the week, and especially during the hack day, students would compile a small portfolio (3 projects) which would demonstrate competence in key areas and which could be used with future employers.

VU Organization Partnership: At the end of the one-week bootcamp, students will be broken into groups, provided a faculty mentor (as 3rd party assistance and general guidance) and begin to work on their Vanderbilt partnership project. (Partner research labs/groups would have been previously identified.) An initial meeting between fellows and their partner would occur within two weeks of the bootcamp, and Fellows would be expected to provide an in-person, business-style report at least once a month to update partners on progress, troubleshoot problems and ensure expectations were aligned.

Ongoing Training: The final component would be bimonthly enrichment courses for fellows. These would cover more advanced topics that may not be usable for the fellow immediately (e.g. data ethics or true “big data” processing with Hadoop or Spark), but familiarity with which could prove important for future prospects. These two-hour classes could originate from a variety of departments and organizations across campus, such as ACCRE or the faculty in the Big Data specialty.

ESTIMATED COST

I estimate that, annually, the program would cost $110,000 to implement. The major costs are below:

Faculty Involvement: In its initial year, the program would require considerable time investment from one or more knowledgeable faculty members (probably on the level of a class). After a complete curriculum has been designed, faculty involvement would be restricted to teaching courses.

- Designing a curriculum ($5000)
- Teaching the bootcamp ($1000 * 2 instructors)
- Teaching ongoing courses ($250/session * 12 sessions)

Support and Administration: An existing BRET administrative program manager would facilitate logistics for applications, awards, meetings, events and publicity. All admissions and project decisions would be coordinated by a small advising faculty committee, with one member serving as the chair.

- Administrative officer (: $15,000

Fellowship Stipend: To entice the best students to join the fellowship, and because fellows will be providing a service to the Vanderbilt community, a stipend will be offered:

- Stipend: $5,000 * 15 fellows