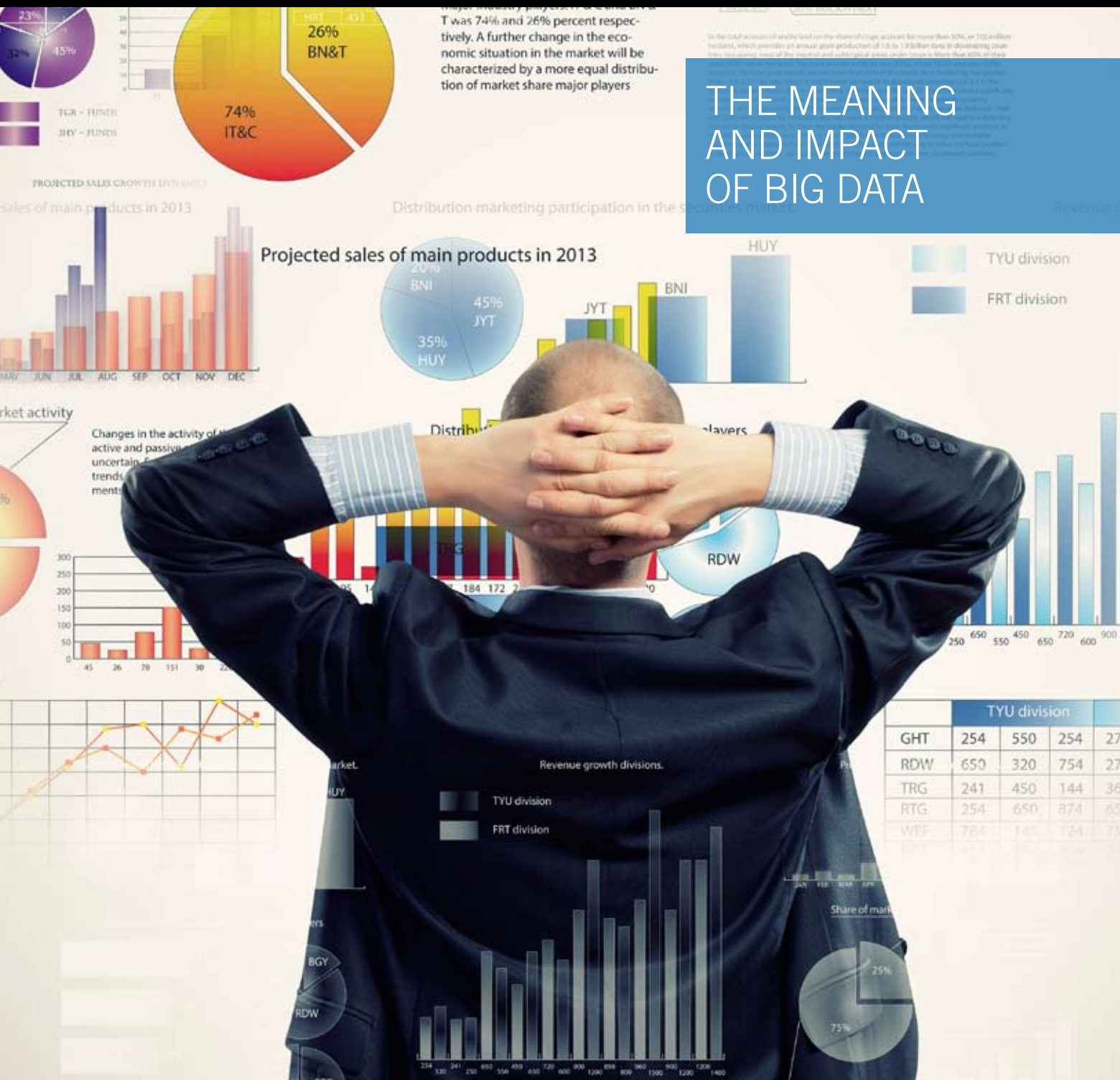
MAY
JUN
14
V 18 I N 03

information outlook

THE MAGAZINE OF THE SPECIAL LIBRARIES ASSOCIATION

THE MEANING AND IMPACT OF BIG DATA



Teaching Librarians to be Data Scientists

TRAINING LIBRARIANS AND INFORMATION PROFESSIONALS TO BE 'DATA SAVVY' CAN HELP THEM BECOME PARTNERS IN THE DATA-RELATED WORK OF THEIR ORGANIZATIONS.

BY CHRISTOPHER ERDMANN, MLIS

Since arriving at the Harvard-Smithsonian Center for Astrophysics in 2010, I have been interested in exploring new forms of digital librarianship. My interest stems in part from my involvement with the NASA Astrophysics Data System (ADS), a repository of astronomical literature that contains more than 12 million records and 4 million full-text documents. The ADS, located within the John G. Wolbach Library at the Harvard-Smithsonian Center, is an invaluable tool that I and other astronomy librarians use to perform digital curation, mining the literature and creating linked data.

Much of what we librarians do helps facilitate search and discovery in the ADS, but more importantly, we generate many of the data links that astronomers use on a daily basis. This curation activity also supports analyses of how telescopes and instrumentation are performing. I believe this type of work forms the backdrop of the data-centric library.

I soon decided that my staff needed formal training in creating and managing such a library. I made initial attempts at training my staff, but running in and out of the office, teaching in short bursts, is not the best way to go about it. Inspired by a staff member, Louise Rubin, I started a course titled Data Scientist Training for Librarians (DST4L)—but it did not come about without planning, hard work, and a bit of luck. In the spirit of the course, I would like to share with the library community how DST4L came to be, what we learned, and the next steps we plan to take. I hope this information will be of use to librarians and information professionals considering similar programs.

CHRIS ERDMANN is head librarian at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts. He can be reached at cerdmann@cfa.harvard.edu or @libcce.

Motivations for the Course

Several factors led me to develop the DST4L course. For one, I strongly believe that librarians should learn how to program. I think this skill has had a positive impact on my own career and my ability to improve services at the libraries in which I have worked. For this reason, I often encourage younger librarians to learn how to program, especially when they ask me what types of expertise and experience I am looking for in new hires.

Second, to inform our next steps in offering data services, we librarians need to dive into data, get our hands dirty in the research data life cycle, and experience the process of data science firsthand. By so doing, we can upgrade our skills and become true partners in the data-related work of our communities. By some estimates, data scientists spend up to 80 percent of their time translating data, leaving them less time to develop insights from it. Librarians, if

retooled, could play a vital role in this area.

Finally, David Dietrich of EMC, a company that provides IT storage hardware solutions, offered additional encouragement. At a Boston Data Science Meetup, I expressed my concerns to David that librarians are not experiencing data science firsthand but that, if trained, they could be valuable in particular aspects of data science, such as assisting with data clean-up and discovery. David agreed that this direction might be beneficial to libraries—that data-savvy librarians could play a valuable role in their communities.

Developing the Course

When I first set out to learn more about data science, I found the literature to be scarce. Still, the following resources helped greatly:

- *Data Jujitsu: The Art of Turning Data into Product*, by DJ Patil
- *Beautiful Data: The Stories Behind Elegant Data Solutions*, by Toby Segaran
- *The Data Journalism Handbook*, by Jonathan Gray
- *EMC Data Science and Big Data Analytics*, by David Dietrich
- *CS 194-16: Introduction to Data Science*, by Jeff Hammerbacher and Mike Franklin

The first three resources gave me a better understanding of the topic and prepared me for future conversations with experts such as David Dietrich. David's still-newly-minted EMC course on data science turned out to be an extremely helpful resource. He eventually came to speak at Data Scientist Training for Librarians, and his slides and a video of his talk can be found on the DST4L blog (see "Introductory Session to Data Scientist Training for Librarians, Round 2"). Course participants lauded his talk for providing a brief but complete summary of what data science entails. David's visit also highlighted another aspect of the course—the importance of bringing in experts from the local community to put the course into context.

The final resource in the list, which comprises material taken from the Introduction to Data Science course taught at the University of California at Berkeley, allowed me to think more deeply about what each section of DST4L might cover and find experts who could help teach these sections (see the DST4L syllabi). At the time, finding course material like Berkeley's was difficult; now, more and more data science programs are being taught, such as the Coursera Data Science Specialization course taught by Johns Hopkins University.

The two DST4L courses I have organized thus far would never have been possible without the contributions of the many guest instructors and speakers who have participated. I met many of these experts through local meet-ups, at conferences, in the Harvard University community, or just by chance. All of them were able to pick up the curriculum fairly quickly after I explained to them the relevance of librarianship to data science.

I hired a library student from Simmons College, Jennifer Prentice, to help with the planning, organization and reporting, and she did an amazing job helping me keep the course whole. I asked her to capture the essence of each class with a blog entry told from the perspective of a student, in keeping with my goal of making the course material and experience accessible to the outside world. (In later sessions, other students contributed to the blog as well.) Like the blog, our class notes, bookmarks, code, data, guest speaker videos, and everything else we used and produced were made available through the class WordPress site. An open approach to capturing and managing course material using Etherpad-like tools turned out to be the best solution, particularly when students needed to reference these resources later.

Several other resourceful Simmons Library School students joined with my staff in acting as teaching assistants. Though they had some experience with the technologies covered in the course, they, too, were learning along with the participants. It helped to have some experienced members of the group interspersed throughout the classroom to offer guidance to their closest peers and forestall some of the difficulties inherent in having one instructor address many problems.

The informal nature of the course helped minimize tension, though it could not be completely avoided. One such moment occurred during the setup of iPython, an interactive programming tool that allows you to step through code line by line. In the first DST4L course, we encountered a number of difficulties during this process; in fact, the experience almost completely scared participants away from iPython, especially when they witnessed the comparatively simple installation of RStudio. Since then, installation and setup have gotten a lot easier, with distributions like Anaconda and Web-based solutions such as Wakari.io.

Still, this experience made a lasting impression, and I was determined to make improvements in the next DST4L course. I turned to Software Carpentry, a group of volunteers who teach basic software skills to researchers. They conduct two-day "bootcamps" that teach core skills needed for productive research; for DST4L, they covered the shell, Python, version control, and regular expressions. We truly had an all-star cast providing the training, and it turned out to be a huge success. Two of the volunteers, Philip Guo and Matthew Ruttley, captured the event in blog posts.

When librarians ask me how they can conduct something similar to Data Scientist Training for Librarians, I often refer them to Software Carpentry as a first step (or the only step, if they cannot run a longer program). I also encourage them to invite other members of their community, especially graduate students, to the Software Carpentry bootcamps to network and foster connections around learning.

I felt strongly that the course needed to be open and diverse, a place where library school students could connect with librarians in the field.

Tools and Instructors

I owe a huge debt of gratitude to two instructors: Rahul Dave, a brilliant scientist who works for the NASA ADS, and Tom Morris, the talented lead developer for OpenRefine. Rahul taught data extraction (APIs and Web scraping), data manipulation, natural language processing and statistical analysis, which participants saw as the most daunting part of the course. He stressed the importance of toolkits like the NLTK, which can be used to automate the classification of documents. For data sources, we often used open, accessible repositories such as the Internet Archive and the NASA ADS or datasets in CSV format from data repositories such as Zenodo, figshare and Dataverse.

In both courses, Rahul used the programming language Python and the iPython Notebook, a Web-based interactive computational environment that allows you to combine code execution, text, mathematics, plots, and rich media into a single document. The academic community is using this tool increasingly to produce interactive textbooks that librarians should note (see the example books at the Notebook Viewer).

Tom Morris introduced participants to a GUI-based tool called OpenRefine, which is used for data extraction, cleaning, manipulation, and extension. OpenRefine is growing in popularity within the library community, especially for linked data projects, but it is still surprisingly unknown despite efforts by groups such as Free Your Metadata (whose members wrote a book titled *Using OpenRefine*) to increase its visibility. OpenRefine is a helpful stepping stone to the more advanced training in Python. The OpenRefine interface allows you to run simple functions and regular expressions while hiding some of the complexities of programming. It also allows you to perform some data analysis.

Many participants gravitated to OpenRefine as their tool of choice, but DST4L is designed to introduce technologies of varying complexity. In the most recent course, for example, the students used Excel, then OpenRefine, and, finally, Python. In the end, the students chose the tool with which they are most comfortable.

Two other instructors, Alex Storer and Lynn Cherny, did their best to keep their classes as simple as possible despite the ambitious schedule. Alex framed his sessions by trying to reproduce a library data-related visualization from IfWeAssume, a blog by James Davenport at the University of Washington. In the process of reproducing Davenport's work, Alex thoroughly explained each step at a manageable pace. He created an enjoyable experience, programming in R and RStudio. As a result, some participants adopted R as their preferred programming language moving forward.

Lynn started her training by showing the students how they could use Excel to perform data manipulation without knowing how to code and by demonstrating how pivot tables can be a powerful tool for exploring the shape of data distributions. She went on to discuss some of the principles behind data visualization and covered some of the tools that do not require much programming, pointing the students to sites such as Visualizing Information for Advocacy. Even her most advanced class on topic modeling, which culminated in the use of a tool called Gephi, was simple and easy to understand—I could see many of the students keeping pace with her and even exploring further during pauses.

Several guest speakers were invited to present their work and, in some cases, provide training. They included James Turk (Sunlight Foundation), Matt Carroll and Gabriel Florit (*The Boston Globe*), Erin Braswell (R Programming), Seth Woodward (Git), Raymond Randall (Tableau), and Jay Luker (MongoDB). Many of the instructors and speakers came from outside the Harvard community, and this was true of the students as well—they came from MIT, the University of Massachusetts, Simmons College, Brandeis University, Community Change, the Smithsonian Astrophysical Observatory, NASA, Boston University, the University of Connecticut, Bingham McCutchen, and the Federal Reserve Bank. I felt strongly that the course needed to be open and diverse, a place where library school students could connect with librarians in the field and librarians could connect with colleagues at other institutions. The only criterion for joining the program was to possess a genuine interest in the topic.

Sharing What We Learned

Both DST4L courses were offered free of charge. Many librarians took the course because they felt the content would be invaluable to their careers. Neither the informal, experimental nature of the training nor the steep learning curve limited attendance; nor, for that matter, did the lack of a certificate (though participants did receive a badge near the end of their training). Perhaps the challenging nature of the program kept librarians from leaving the program, or maybe it was the learning environment—the realization that it was acceptable to not know the answer, to experiment, and even to fail.

The students mentioned the importance of their group work in keeping them motivated and engaged. I had created the group projects to give students the opportunity to apply what they were learning in class to similar situations they might face in their own work. The projects were overly ambitious,

At the most basic level, everyone involved in the program came away with a better understanding of the research data life cycle.

but they were meant to challenge the students. The desire to complete these difficult projects and not let down their fellow group members helped keep many participants engaged.

At the end of each course, the participants delivered presentations on their experience and work in a "tell all" session or created blog posts called data stories. These stories have largely been responsible for motivating others to consider the program and encouraging the Harvard Library to continue sponsoring the program.

After each course, students participated in a "share all" event where we discussed the pros and cons of the course and how it could be improved. Jennifer Prentice and Marc McGee summarized the events in two blog posts. In both sessions, it seemed the participants had an endless number of ideas to improve the training, but I was particularly struck by how engaged they were, how important they thought the program was, and how committed they were to continuing it. I believe that the courses made an impact with each individual on different levels.

At the most basic level, everyone involved in the program came away with a better understanding of the research data life cycle; each participant then applied this new knowledge to his or her own situation. Many librarians have since reported improved data-related interactions with their communities as a result of their increased knowledge.

Another goal of DST4L was to upgrade the skills of librarians, and many of the participants are now using their new-found skills. For instance, Vernica Downey has automated library processes using Python, Alex Holachek is helping the NASA ADS improve its visualization tools, and Katie Frey is implementing semantic technologies in astronomy.

The course also sought to address the culture within libraries and encourage the students to change the "library mindset" through abstract thinking, continuous learning, hacking, and other approaches. I hoped the participants in the course would start seeing challenges as opportunities, exploring whether they could do better, and helping to improve library processes and services through creative solutions. The students have achieved these goals to varying extents, but these outcomes are bonuses—the first goal was the most important.

DST4L is slated for further support from the Harvard Library, but in its next iteration, the emphasis will be on creating "data savvy" librarians. The current title creates the impression that librarians are trained data scientists once they finish the course, but the students who have taken DST4L relate better to being called "data savvy."

Former DST4L students and external parties will help develop the new curriculum. As before, it will address aspects of the research data life cycle, though it will place more emphasis on applying the training in library settings. We also hope to continue the hands-on application approach from previous courses.

As is currently the case, DST4L applicants will be asked to demonstrate a genuine interest in the topic and pledge to apply what they learn. An announcement will be forwarded to the library community to apply to the next training program. The students' communications will be made available via a Google group, which will be open to a greater audience than before. With any luck, the program will continue to grow a "data savvy" library community, steer us toward change, and benefit future data services in libraries. **SLA**