Data Scientist Training for Librarians

Christopher Erdmann (Harvard-Smithsonian Center for Astrophysics)
Colin van Alstine, Christine Eslao, Michelle Durocher, Scott Wicks (Harvard Library)

<abstract>

Recent studies suggest that there will be a shortfall in the near future of skilled talent available to help take advantage of big data in organizations. Meanwhile, government initiatives have encouraged the research community to share their data more openly, raising new challenges for researchers. Librarians can assist in this new data driven environment. Data Scientist Training for Librarians is an experimental course being offered by the Harvard Library to train librarians to respond to the growing data needs of their communities. In the course, librarians familiarize themselves with the research data lifecycle, hands-on, using the latest tools for extracting, wrangling, storing, analyzing and visualizing data. By experiencing the research data lifecycle themselves, becoming data savvy and embracing the data science culture, librarians can begin to imagine how their services might be transformed.

<technologies>

Unix Shell, Git, GitHub, Python, iPython, Excel, OpenRefine, SQL, Data Repositories, R, RStudio, Tableau, D3, NoSQL, MongoDB, Gephi & More

<about the course>

Makeup: Librarians from beginner to intermediate, 60+ total students (2/3 Harvard, 1/3 local institutions), 12 instructors/speakers, 9 TAs

Local institutions w/ participants: MIT, University of Massachusetts, Simmons College, Brandeis University, Community Change, Smithsonian Astrophysical Observatory, NASA, Boston University, University of Connecticut, Bingham McCutchen, Federal Reserve Bank

Open course: All material, including the syllabus, lessons, instructor notebooks, scripts, class notes, student blog stories, guest speaker videos and projects is accessible online, open and searchable. The exception, classes are not streamed and recorded. Tools used include WordPress, Etherpad, Google Apps, Dropbox, NBViewer, RPubs.

Student projects: Precooked ideas, students choose, form groups, work hands on w/ data, use lessons, learn from each other, work in collaborative environment, instructor/ TA/supplemental online assistance, hack events, present story on experience, methods, findings, visualizations.

Student feedback: Following each course, students are asked for their input about the course. Was it useful? Overall, they say, “Absolutely!”, but students have a lot to say about their experience and how the course can be improved. See http://altbibl.io/dst4l/dst4l-feedback-session/ & http://altbibl.io/dst4l/dst4l-tells-all/.

<course outline>

Extract

Obtain data via CSV, API, web scraping using Excel, OpenRefine, Python & R.

Wrangle

Clean up, convert messy data, export in open format, prep for analysis, share & deposit.

Analyze

Explore distribution, shape of the data, run & test models, create plots, maps, graphics.

Visualize

Review types of viz, discover underlying story of data & tell a story w/ viz (tools).

D3

Visit: http://altbibl.io/dst4l
Contact: cerdmann@cfa.harvard.edu