Data Scientist Training for Librarians #DST4L

LISA VII June 17-21, Naples, Italy

C. Erdmann | @libcce
Harvard-Smithsonian Center for Astrophysics

E-Science @ LISA VI

E-Science and Astronomy Faculty: Past, Present, and Future

Lee A. Pedersen

Brown University

http://bit.ly/SE2YdU

E-Science @ CfA

- Policies & Copyright Advice
- DMPs & DMPTool
- DM Training Programs
- DM @ Harvard Site
- Research Data Collaborative
- Data Curation Profiles
- e-Science Portal for NE Librarians
- DataCite/EZID
- E-Science Institute
- WH OSTP Response
- Survey: Story of Your Data?
- Data Repos (Zenodo, Dataverse, Figshare)
- Data Citation Principles

To Plan Library Data Services...

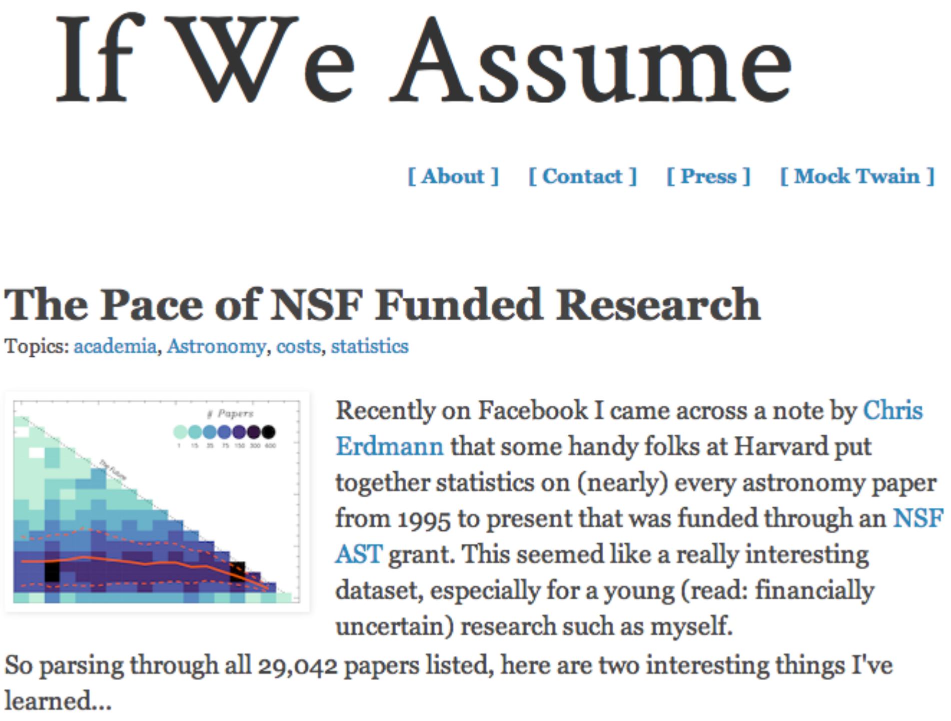
Librarians need to familiarize themselves with the research data lifecycle, get hands-on and use the latest tools for extracting, wrangling, storing, analyzing and visualizing data. By doing so, they can become data savvy, embrace the data science culture and begin to imagine how their services might be transformed.

Help with Data

“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any ‘analysis’ at all.”

Anonymous Data Scientist, Jeff Heer Study

What If We Could Be More?



If We Assume

[About] [Contact] [Press] [Mock Twain]

The Pace of NSF Funded Research

Topics: academia, Astronomy, costs, statistics

Recently on Facebook I came across a note by Chris Erdmann that some handy folks at Harvard put together statistics on (nearly) every astronomy paper from 1995 to present that was funded through an NSF AST grant. This seemed like a really interesting dataset, especially for a young (read: financially uncertain) research such as myself.

So parsing through all 29,042 papers listed, here are two interesting things I've learned...

Could we be of more use to scientists?

Sympathize with their data needs and offer assistance?

Become data savvy?

DATA SCIENTIST TRAINING
1
0
FOR LIBRARIANS
DIRK TIEDE 02.25.2013

Resources for #DST4L

26 Da 58 Sci 77 Gr

The Data Science Group

Preparation

David Dietrich, EMC

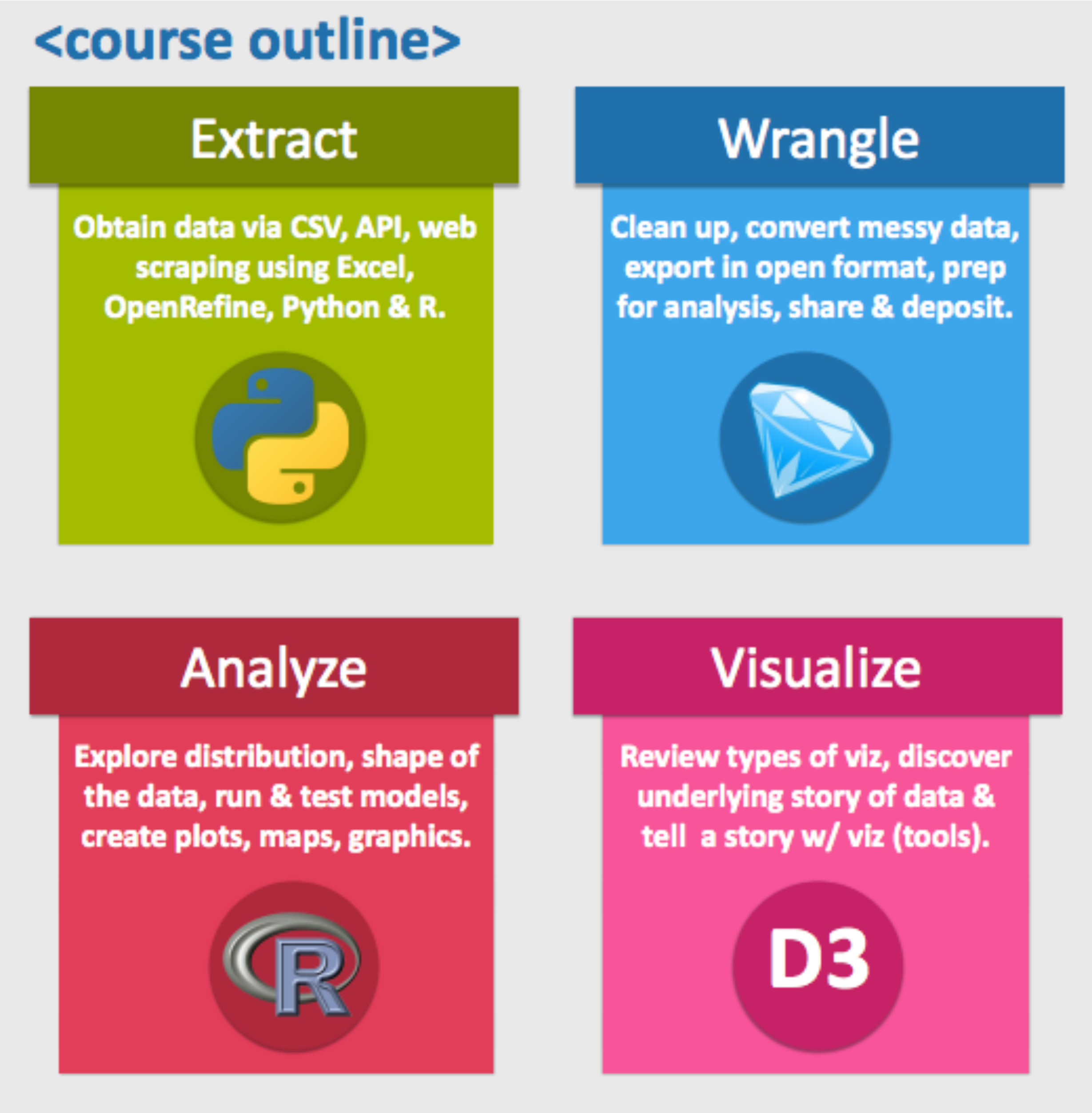
Overview, examples of Big Data Analytics & Data Science (Tools, Technology, & Skill Development) & EMC Data Science Courses

http://goo.gl/2Tp528

Greg Wilson, Software Carpentry Bootcamp

Hands-on training from experts on CLI, Git, Python

http://swcarpentry.github.io/2013-08-23-harvard/



<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).

http://www.dcc.ac.uk/sites/default/files/documents/IDCC14/215DST4L-idcc_14.pdf

Outline, Expanded

Class Date	Topic
2013-08-23	SW Carpentry
2013-08-24	SW Carpentry
2013-08-28	Starting with Excel: Data Manipulations & Graphing
2013-09-05	OpenRefine: Analyze, Clean & Reformat Data
2013-09-10	Linked Data Ecosystem OpenRefine: Reconcile, Extend, & Publish
2013-09-19	Wrapping up with Refine / Python Basics
2013-09-25	Python: Working with APIs & Web Scraping
2013-10-01	PyMARC, PDF Extraction: Text & Tables
2013-10-08	Math and Stats for Journalists
2013-10-15	Pandas: Munging, Stats & Visualization
2013-10-22	Stats1: Basics with NumPy, Matplotlib, scikit-learn
2013-10-29	Stats2: Text processing, Naives Bayes with NumPy, Matplotlib, scikit-learn
2013-11-5	Principles of Visualisation Design: Infographics, Interactives, Exploratory, Storytelling...
2013-11-12	Non-Interactive Visualization: Tools, Use Cases, Examples
2013-11-19	Interactive Visualization: Tools, Use Cases, Examples
2013-11-26	Feedback Session
2014-01-11	Hackathon

Syllabus 1: http://goo.gl/OpbIVa
Notes: http://goo.gl/hI2gEH

Technologies

Command Line

Git (GitHub)

Excel

OpenRefine

Python (pynb)

R (RStudio)

Tableau

SQL (SQLShare)

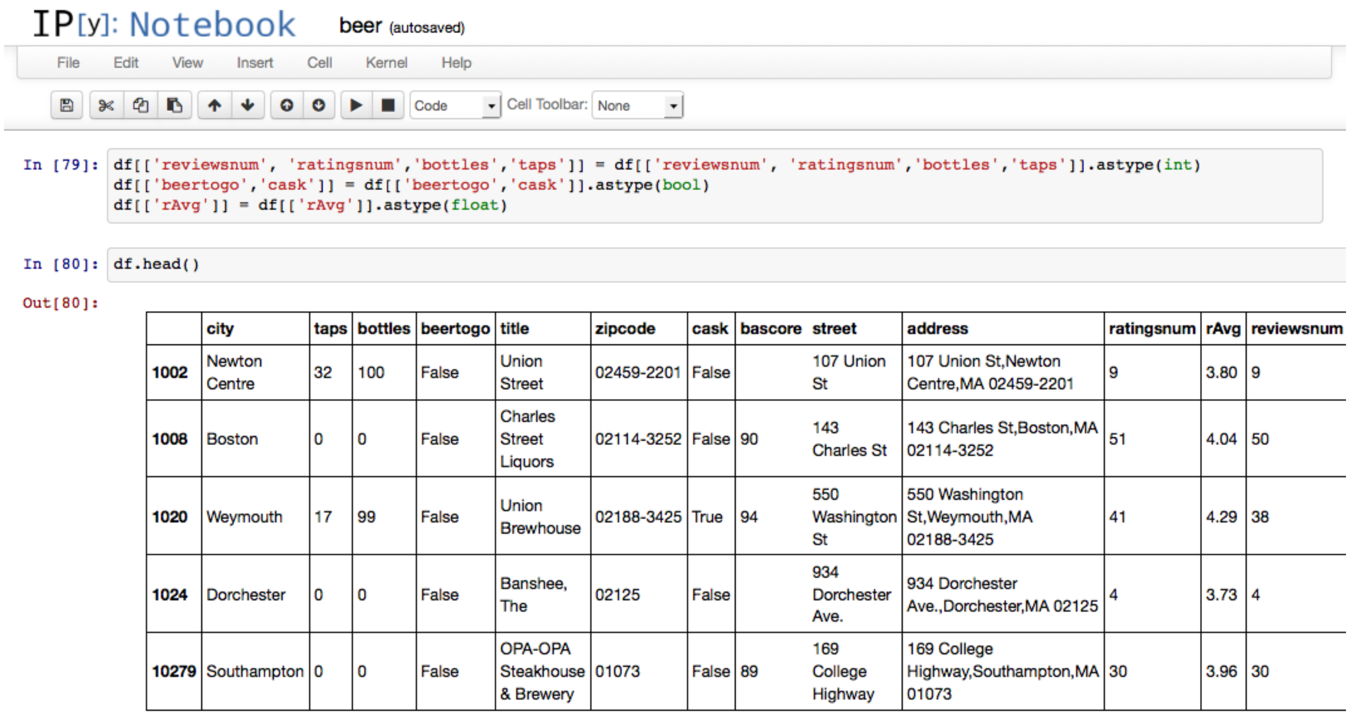
MongoDB

Data Repositories

D3

Gephi

Highlight: iPython Notebook



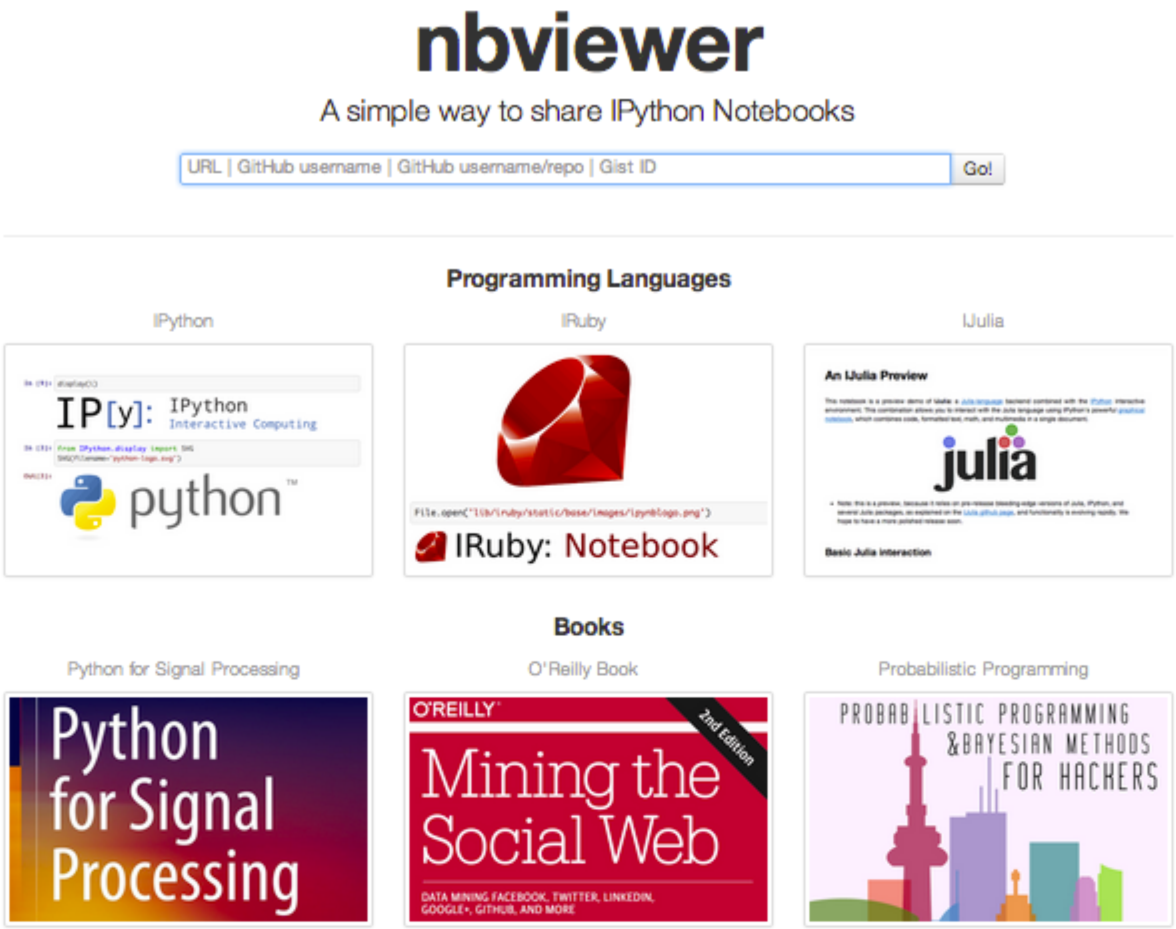
```
In [79]: df[['reviewsnum', 'ratingsnum','bottles','taps']] = df[['reviewsnum', 'ratingsnum','bottles','taps']].astype(int)
         df[['beertogo','cask']] = df[['beertogo','cask']].astype(bool)
         df[['rAvg']] = df[['rAvg']].astype(float)
```

```
In [80]: df.head()
```

Out[80]:

	city	taps	bottles	beertogo	title	zipcode	cask	bascore	street	address	ratingsnum	rAvg	reviewsnum
1002	Newton Centre	32	100	False	Union Street	02459-2201	False		107 Union St	107 Union St,Newton Centre,MA 02459-2201	9	3.80	9
1008	Boston	0	0	False	Charles Street Liquors	02114-3252	False	90	143 Charles St	143 Charles St,Boston,MA 02114-3252	51	4.04	50
1020	Weymouth	17	99	False	Union Brewhouse	02188-3425	True	94	550 Washington St	550 Washington St,Weymouth,MA 02188-3425	41	4.29	38
1024	Dorchester	0	0	False	Banshee, The	02125	False		934 Dorchester Ave.	934 Dorchester Ave.,Dorchester,MA 02125	4	3.73	4
10279	Southampton	0	0	False	OPA-OPA Steakhouse & Brewery	01073	False	89	169 College Highway	169 College Highway,Southampton,MA 01073	30	3.96	30

Highlight: NBViewer



nbviewer

A simple way to share IPython Notebooks

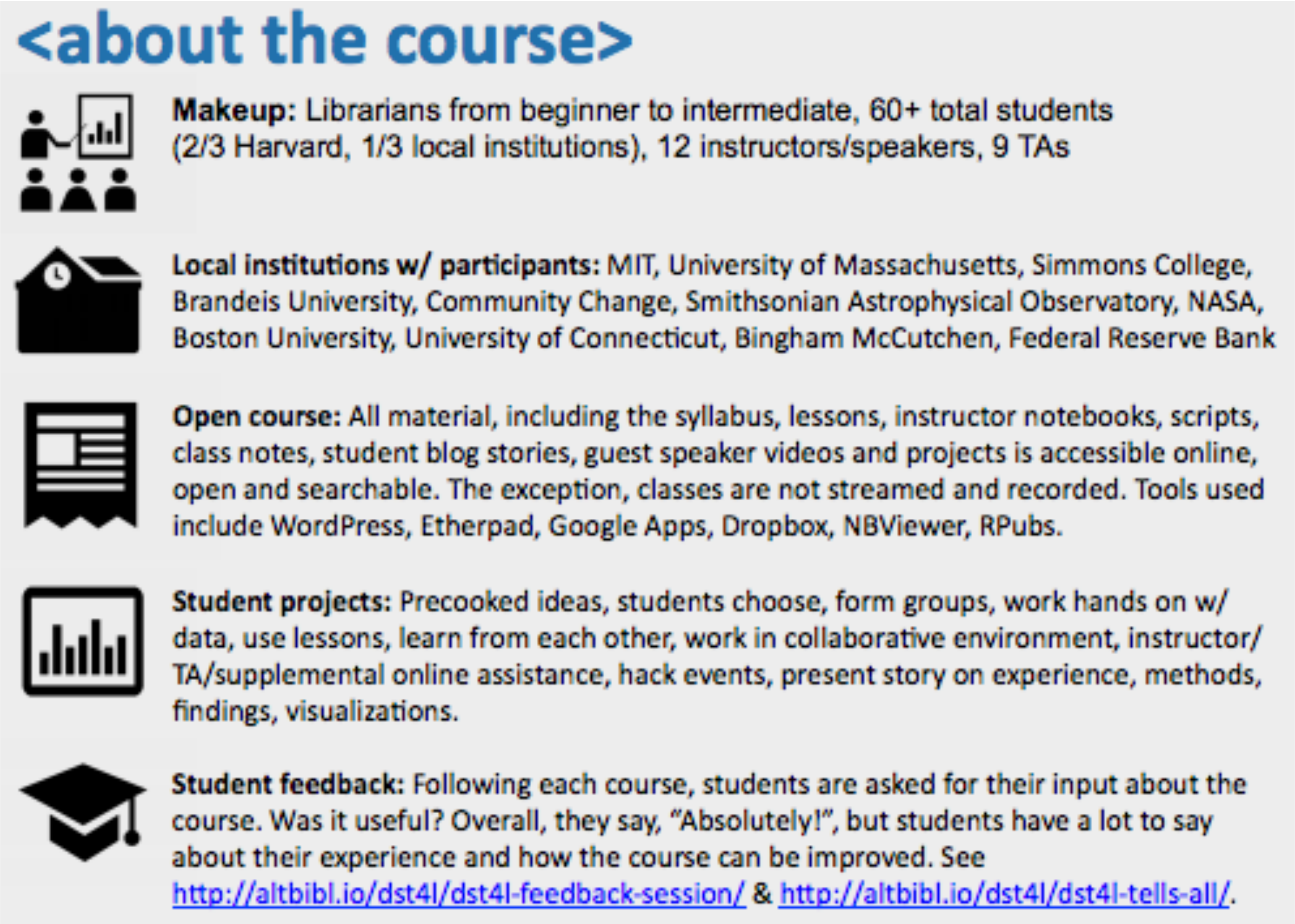
URL | GitHub username | GitHub username/repo | Gist ID
Go!

Programming Languages

IPython
IP[y]: IPython Interactive Computing
python
IRuby
IRuby: Notebook
IJulia
An IJulia Preview
julia
Basic Julia interaction

Books

Python for Signal Processing
Python for Signal Processing
O'Reilly Book
O'REILLY
2nd Edition
Mining the Social Web
DATA MINING FACEBOOK, TWITTER, LINKEDIN, GOOGLE+, GITHUB, AND MORE
Probabilistic Programming
PROBABILISTIC PROGRAMMING & BAYESIAN METHODS FOR HACKERS



<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.

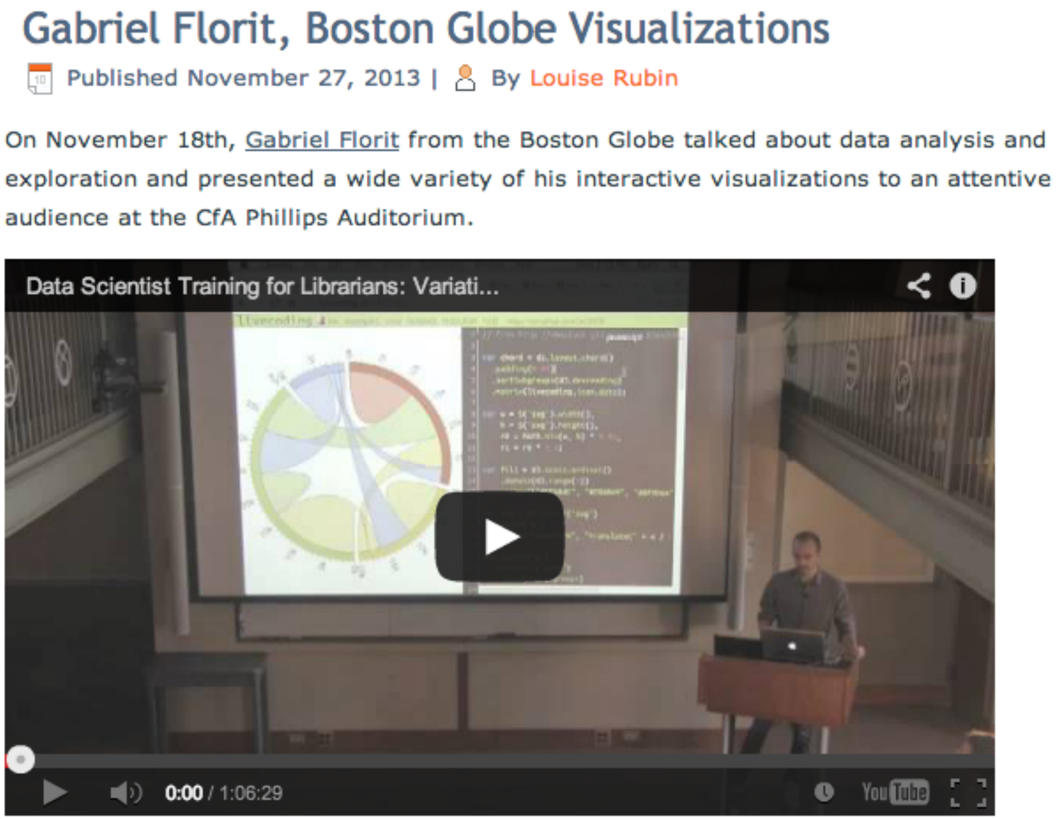
Student projects: Precooked ideas, students choose, form groups, work hands on w/ data, use lessons, learn from each other, work in collaborative environment, instructor/TA/supplemental online assistance, hack events, present story on experience, methods, findings, visualizations.

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/.

More About The Course

- 1 class/week for 3-4 months
  & 1 hands on session (projects)
- Consistent class location/hours
- Lead organizer (context), expert instructors, assistants
- In-person, not virtual or recorded
- Some homework, final presentations/projects
- 1st class free, informal vs 2nd class funded, structured
- Supplementary material listed throughout notes

Real World Examples



Gabriel Florit, Boston Globe Visualizations

Published November 27, 2013 | By Louise Rubin

On November 18th, Gabriel Florit from the Boston Globe talked about data analysis and exploration and presented a wide variety of his interactive visualizations to an attentive audience at the CfA Phillips Auditorium.

http://altbibl.io/dst4l/gabriel-florit-boston-globe-visualizations/

Blog/Projects

Image Courtesy of Trimble Navigation Limited © 1996 -

Blog/Student Perspective
http://altbibl.io/dst4l/category/blog-posts/

Course 1 Projects & Presentations
http://altbibl.io/dst4l/category/data-stories/

Course 2 Hackathon (Beer Example)
http://goo.gl/Ly6dxU

Response

Highlighted student comment:

http://www.youtube.com/watch?v=U5ZYM085bNo&t=1m21s

...very helpful, preparing for long career, going to see it more and more, will keep using skills set, like doing it, fun problem solving...

How Do I Start?

Software Carpentry Workshop

Online Learning (MOOCs)

Invite Local Experts

Tap Your Own Community

Questions?