



Teaching statistics for the future

The MOOC revolution and beyond

Brian Caffo

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Teaching statistics for the future

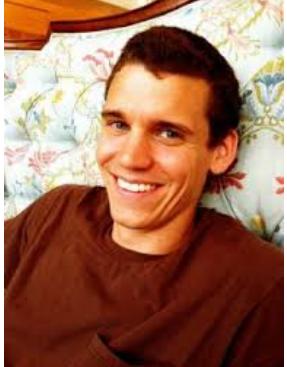
Outline of the talk

1. Who the heck am I?
2. A brief taxonomy and history of online educational models
3. Massive Open Online Courses (MOOCs)
4. JHU Biostat involvement in Coursera
5. Novel moving target directions of the field statistics
6. **Data Science series**
7. SWIRL

About these slides

- HTML5 using (customized) Google io2012 style (<https://code.google.com/p/io-2012-slides/>)
- Created using slidify (<http://slidify.org>)
- Appear on github at (<https://github.com/bcaffo/MOOCtalk>) (<https://github.com/bcaffo/MOOCtalk>)
fork if you'd like
- Jointly written with my collaborators Jeff Leek and Roger Peng
- CC licensed by-nc-sa

Core team



Plus generous contributions from the

- Department of Biostatistics (<http://www.biostat.jhsph.edu>)
- Center for Teaching and Learning (<http://www.jhsph.edu/offices-and-services/center-for-teaching-and-learning/>)
- Bloomberg School of Public Health (<http://www.jhsph.edu>)
- Johns Hopkins University (<http://www.jhu.edu>)
- Coursera (<http://coursera.org>)
- Team SWIRL
- Lauren and Ethan (Brian's 2013 interns)
- Contributions from github pull requests
- Tolerant families!
- A half of a million intrepid self learners

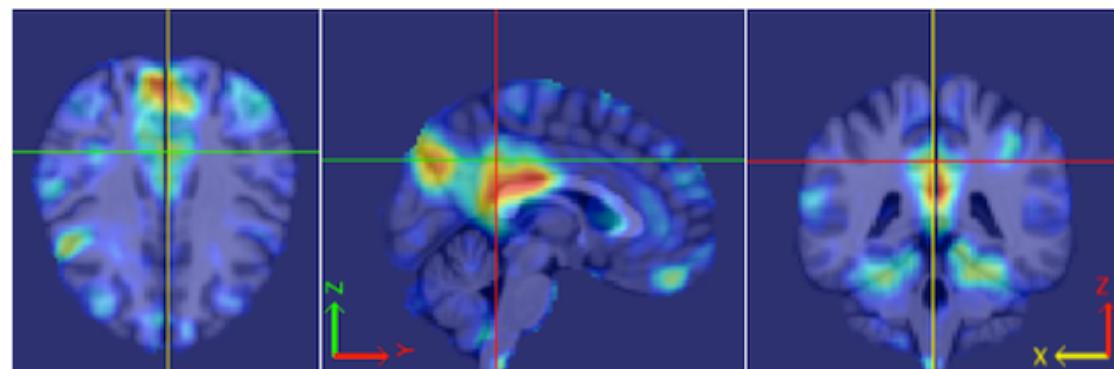
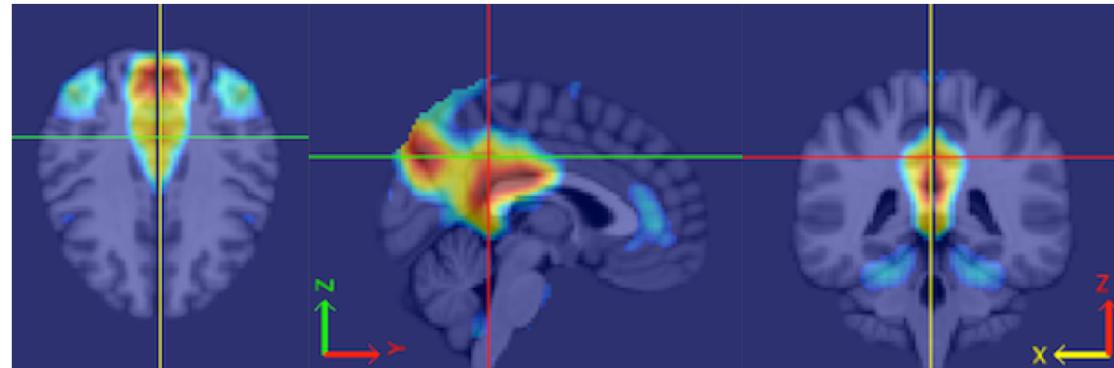
My day job(s)

SMART (www.smart-stats.org (<http://www.smart-stats.org>))

The screenshot shows the SMART website homepage. At the top, there is a dark header with the SMART logo (green 'M' and 'A'), 'Johns Hopkins University', and a subtitle 'STATISTICAL METHODS AND APPLICATIONS FOR RESEARCH IN TECHNOLOGY'. A search bar and a 'Log In' button are also in the header. Below the header, a navigation menu includes 'HOME', 'CALENDAR', 'PEOPLE', 'GRANTS', 'PAPERS', 'BLOGS', and 'WIKI'. A main content area features a banner for the 'ADHD-200' competition, showing a logo with 'ADHD' and '200' and the text: 'SMART research group led by Brian and Ciprian wins the ADHD-200 competition.' Below the banner, a paragraph discusses the competition's purpose and outcomes. At the bottom of the page, there are links for 'Previous', 'Pause', and 'Next'. The main content area contains a 'Home' section with a brief description of the group's mission and a 'Navigation' sidebar with links to 'Statistical methods', 'Scientific areas of interest', 'Software & Tutorials', and 'Social media'. Navigation links at the bottom include 'About Us', 'Contact Us', and 'Logout'.

Connectomics

resting state fMRI



JHU Biostat onsite degree programs



1. PhD program

- Around 50 students with around 10 matriculating per year.
- Around 200 applications per year.

2. ScM program

- Around 25 students with around 10 matriculating per year.
- Around 50 applications per year.

3. Concurrent MHS program

- Typically around 10 students with around 2 matriculating per year.
- 2 - 4 applications per year

4. Standalone MHS program

My person teaching

- Biostat 751 and 2
 - 16 weeks (8×2) of classes
 - Two 80 minute lectures per week
 - Two tests per term
 - Four homeworks per term
 - One TA
- About 10 students
- Covers intro methods and (mostly) linear models at the doctoral level
- Recently I've introduced the flipped classroom model

(Incomplete) characteristics of educational systems

- Online / in person / blended
- Active/participatory/interactive learning
- Scalable / non-scalable
- Low cost / high cost / freemium
- Student paced / teacher paced
- Open / restricted access
- Flipped / lecture style / blended
- Open / closed source content
- Instructor interaction
- Credentialing
- Funding model

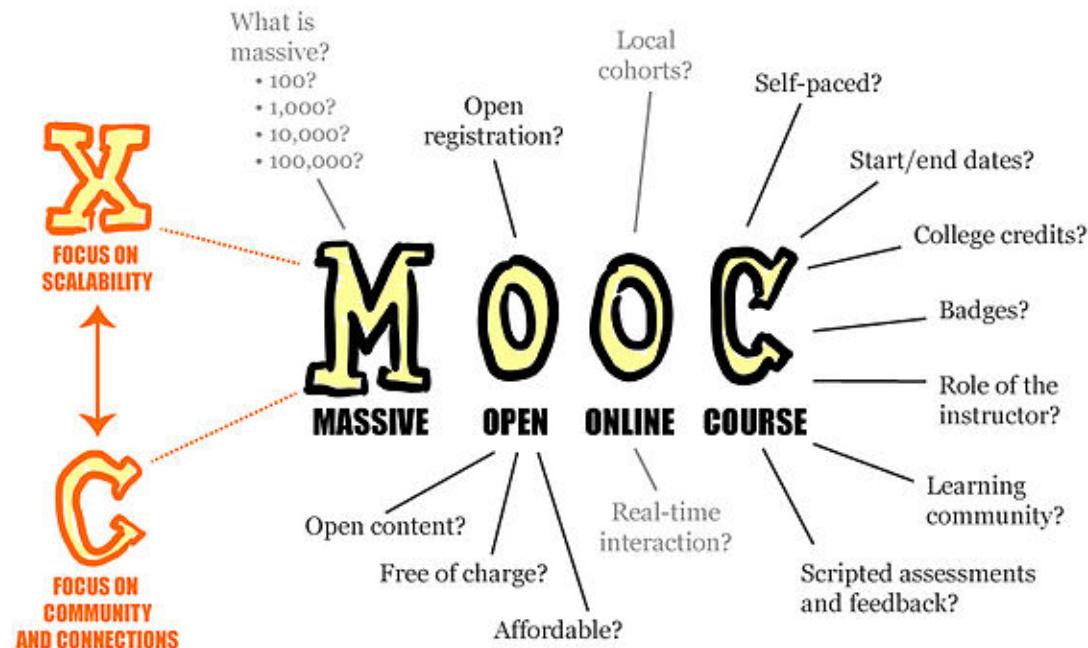
Examples

- Traditional in person teaching generally has characteristics of: in person, lecture style, non-scalable, high cost, restricted access with a large amount of instructor interaction valued credentialing
- "Traditional" online courses are online or blended online and in person and otherwise try to approximate traditional in person classes.
- Online interactive learning (OIL Code School, Code academy) primary characteristics are student-paced interactive learning
- Intelligent tutoring systems (swirl), like OILs just not necessarily online
- Khan Academy is online, interactive, scalable and low cost
- Other modalities : iTunes U, OpenCourseware, Udemy

MOOCs

Primary characteristics are open access, low cost, scalable, online

(every letter is negotiable, from Wikipedia citing Mathieu Plorde)



(<http://www.flickr.com/photos/23311795@N04/8620174342>)

Most visible MOOC instruction sites

edX Take great online courses from the world's best universities



(<https://www.edx.org/>)



(<https://www.udacity.com/>)

udemy

(<https://www.udemy.com/>)

coursera

Courses Partners About ▾ | Sign In Sign Up

The Coursera homepage features a large banner with a graduation cap and the text "Take the world's best courses, online, for free." Below the banner is a search bar with the placeholder "What would you like to learn about?" and a magnifying glass icon. Text at the bottom includes "Join 4,820,952 Courserians.", "Learn from 447 courses, from our 87 partners.", and a yellow link "How it works »". A small "Tecnológico" logo is in the bottom right corner.

Take the world's best courses,
online, for free.

What would you like to learn about?

Join 4,820,952 Courserians.
Learn from 447 courses, from our 87 partners.
[How it works »](#)

Tecnológico

Also

Several university/organization-specific sites, platforms and content delivery systems

- Stanford, CMU, Duke, Harvard, MIT, google ...
- Varying degrees of content/delivery
- EdX platform has been open sourced
- Google course builder (now contributing to EdX)
- Massive amount of development going into platforms and instruction sites/portals

Coursera platform, videos

The screenshot shows a web browser window displaying a Coursera course page. The title bar reads "Lecture 1A: Biostatistics and Experiments". The URL in the address bar is <https://class.coursera.org/biostats-002/lecture/3>. The main content area is titled "Table of contents" and lists four sections: 1. Biostatistics, 2. Experiments, 3. Set notation, and 4. Probability. To the right of the table of contents is a decorative graphic of a graduation cap with mathematical symbols like $X - \mu$, σ , and Z . Below the table of contents is a video player interface. The video title is "Lecture 1A: Biostatistics and Experiments (12:36)". The video progress bar shows it is at 00:14 of 12:36. Below the video player are links to other lectures: "Lecture 3A: Expected Values (12:00)", "Lecture 3B: Rules About Expected Values (10:10)", "Lecture 3C: Variances and Chebyshev's Inequality (18:49)", "Lecture 4A: Random Vectors and Independence (17:33)", "Lecture 4B: Correlation (10:41)", and "Lecture 4C: Variance Properties and Sample Variance (22:15)". On the left side of the page is a sidebar with course navigation links: Home, Syllabus, Grading Policy, Faculty, Github Repository, Video Lectures, Homework, Quizzes, Discussion Forums, Course Wiki, and Join a Meetup. At the top right of the page are links for Courses, About, and the user profile "Jeff Leek".

Example videos (on YouTube)

- Example from data science inference (<https://www.youtube.com/watch?v=ZD7kR4QLFnE#t=269>)
- Ad hoc phone recording (<https://www.youtube.com/watch?v=ZeS-ELmY7Fk>)

Equipment

- Cintiq 22inch display (<http://www.wacom.com/en/us/creative/cintiq-22-hd>)
- Yeti usb microphone (<http://bluemic.com/yeti/>)
- Camtasia (<http://www.techsmith.com/camtasia.html>)
- Note 2 (<http://www.samsung.com/global/microsite/galaxynote/note2/index.html?type=find>)
- Lecture notes (<https://play.google.com/store/apps/details?id=com.acadoid.lecturenotes>)
- ffmpeg (<http://www.ffmpeg.org/>)



Coursera platform, quizzes

The screenshot shows the Coursera quiz editor interface. At the top, there's a navigation bar with links for Courses, Admin, About, and Jeff Leek. Below the navigation, the course title is "JOHNS HOPKINS BLOOMBERG SCHOOL OF PUBLIC HEALTH" and the subject is "Data Analysis" by Jeff Leek.

The main area is titled "Editing Quiz: Weekly Quiz 8". It includes buttons for Save, Preview, Save and Exit, and Exit Without Saving. Below these are links for Edit Quiz Settings, Edit Quiz Preamble, and Edit Raw XML.

On the left, there are three question cards:

- Question 1**: Variation 1. Text: "Suppose this is the result of 85 hypothesis tests: Click and drag to re-order questions". Buttons: Add Variation.
- Question 2**: Variation 1. Text: "Generate P-values according to the following code: Click and drag to re-order questions". Buttons: Add Variation.
- Question 3**: Variation 1. Text: "Suppose I want to generate data from the following model with a simulation: Click and drag to re-order questions". Buttons: Add Variation.

In the center, there's a "Variation 1" section for Question 1. It shows "Question Type: Radio" and a checked checkbox for "Randomize Display Order". Below this is a "Text" field containing: "Suppose this is the result of 85 hypothesis tests:" followed by a table.

	$\beta = 0$	$\beta \neq 0$	CLAIMS TOTALS
Claim $\beta = 0$	50	10	60
Claim $\beta \neq 0$	5	20	25
Hypothesis Totals	55	30	85

To the right of the table, the text "What is the (observed) rate of false discoveries? What is the (observed) rate of false positives?" is visible.

Below the table, there's a "Question-level Explanation" text area and a "Points for this question:" input field.

Coursera platform, peer grading

The screenshot shows the Coursera Admin interface for managing assessments. The URL in the browser is https://class.coursera.org/dataanalysis-001/admin/human_grading/show?assessment_id=4. On the left, a sidebar lists various administrative tools: Log Viewer, User Administration, IMPORT TOOLS (selected), Assignment Submissions, EXPORT TOOLS, Quiz Summary, Detailed Quiz Responses, Assignment Submissions, Peer Assessment Grades, Class Gradebook, STATUS MONITORING (selected), Activity Tracking, Course Overview Statistics, Export Statistics, and Video Status.

The main content area is titled "Questions". A specific question is being configured:

Question

Please either enter the body of your data analysis in the text box or upload a pdf file with your analysis. This file should both contain the main text of your analysis and the numbered list of references. It may be no more than 2000 words.

Max Words: 2000

Enable students to upload files (supports text, image, PDF, and whitelisted files)

Evaluation Criterion 5 / 85

Evaluation Type:

Quantitative
 Qualitative

Does the analysis have an introduction, methods, results, and conclusions section?

POINTS	DESCRIPTION
0	No serious attempt to answer complete the assignment
1	None of these elements are present
2	Only one of these elements is present
3	Only two of these elements are present
4	All three of these elements are present
5	All four elements are present.

Coursera platform, forums

Main source for student interaction
(Forums can be brutal)

Anonymous · 3 months ago % 

Hmm, this could be one of the reasons why i found these lectures quite annoying to listen to, i actually get a headache. I even searched the forum to see if anyone else experiences this and came across this thread. No disrespect intended, i find the timbre of his voice somewhat unpleasant, and find the videos difficult to listen to for an extended period of time.

 1  · flag

[REDACTED] 

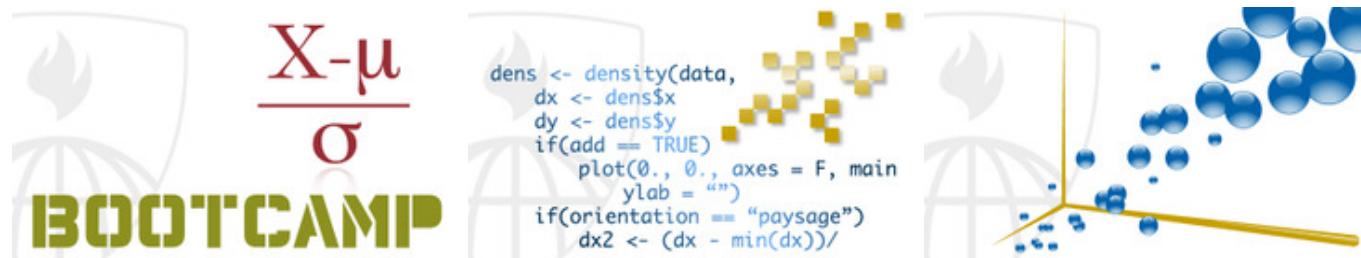
Poor bloke - he's getting a lot of flak today. FWIW, I found the narration a little hard to listen to as well, but I put that down to tiredness. He does sound pretty exhausted in some lectures (there are a couple of stifled yawns in one), which reflects how much work he's put into all this, much of it at night from the sound of it. Slowing down a bit and recording when bright & breezy would make it sound fresher, but maybe this is all done on personal time.

 1  · flag

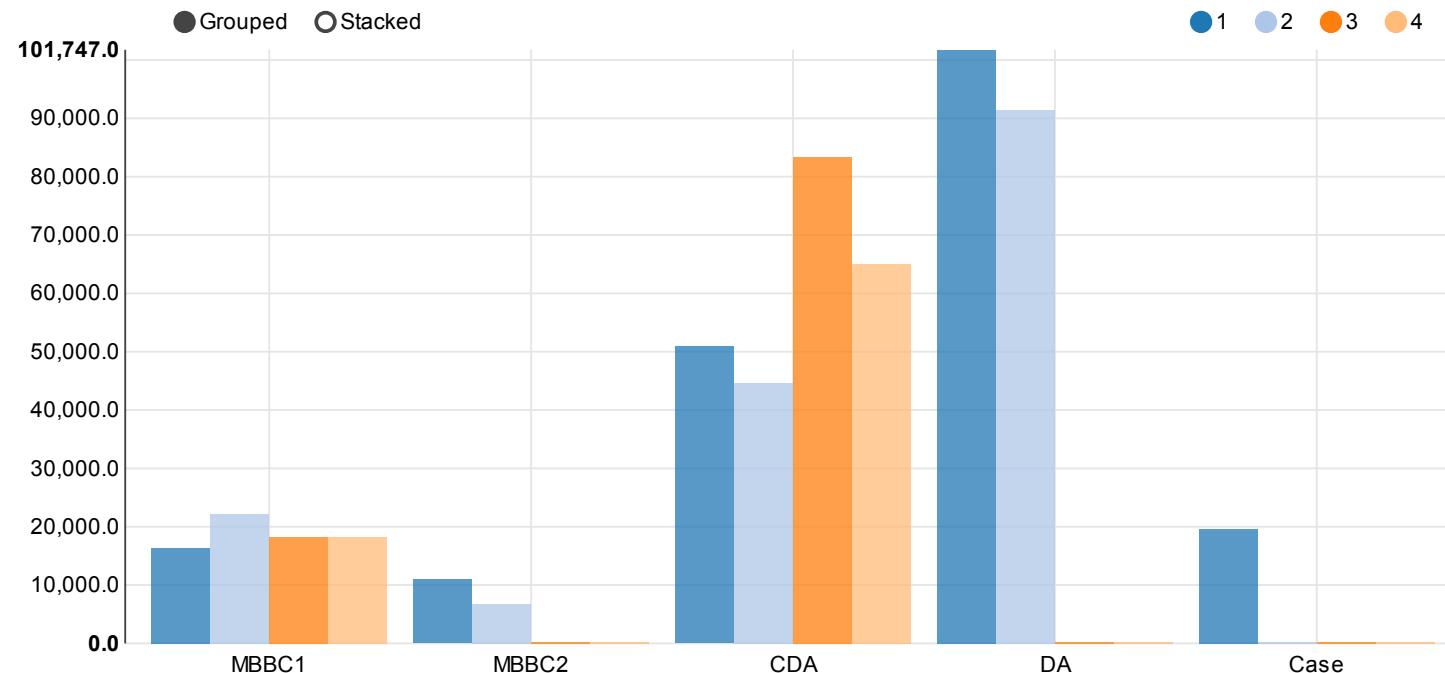
Johns Hopkins Biostat Coursera classes

Original three

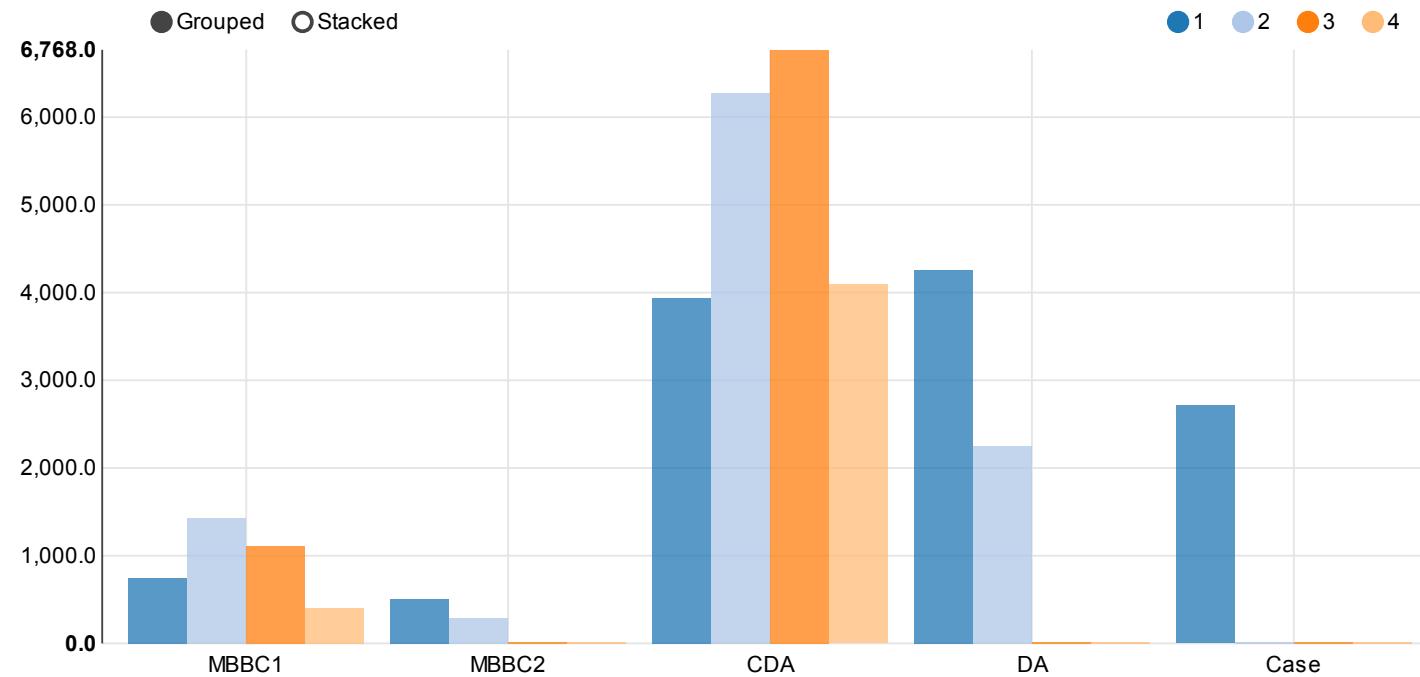
- Brian Caffo, Roger Peng, Jeff Leek
- Run 09/2012, 09/2012, 01/2013



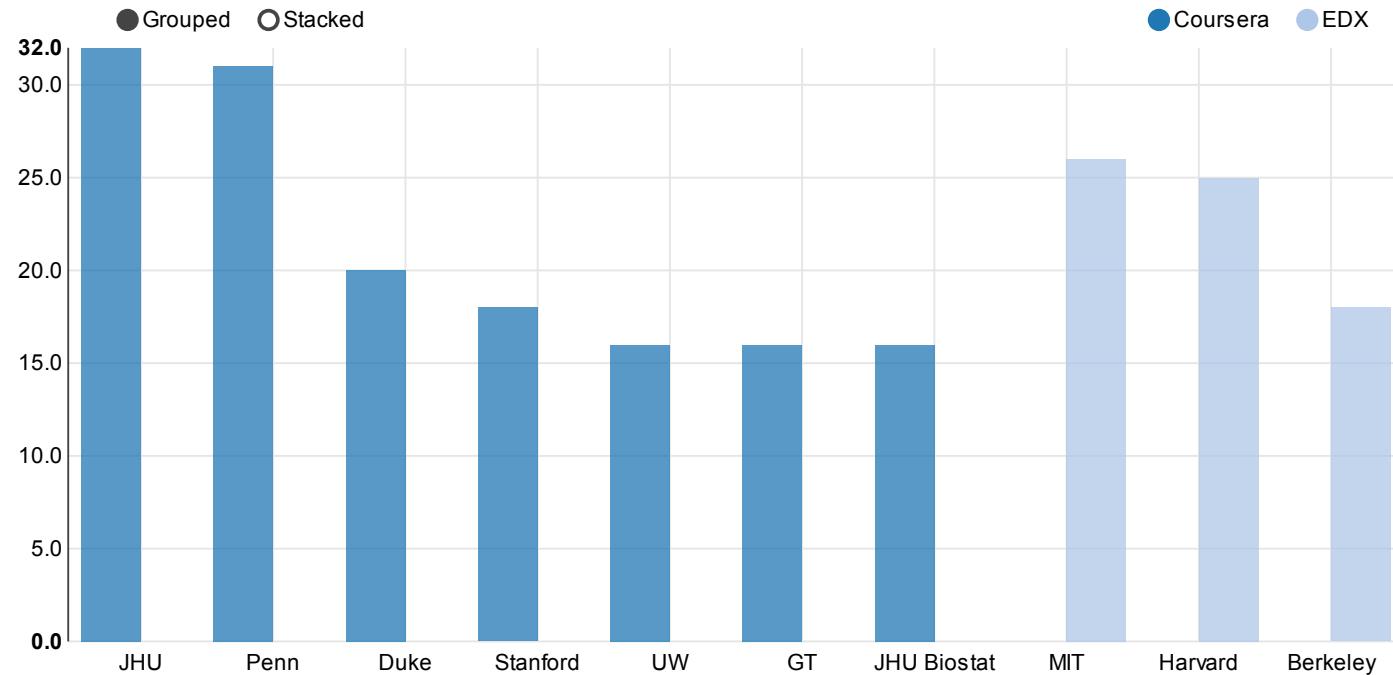
Enrollments by class and offering



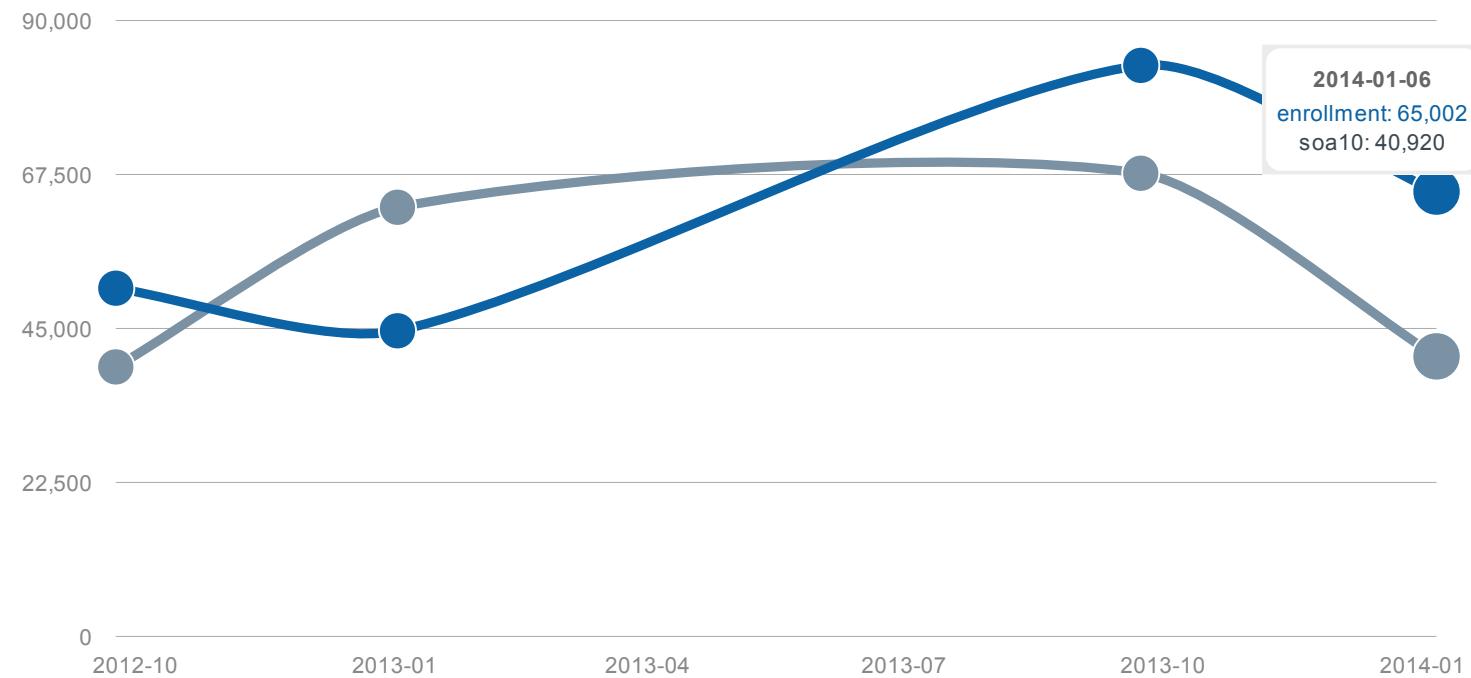
Statements of accomplishment by offering



Over time, MBBC 1



Over time CDA



Important consideration about completion rates

- Students participate in MOOCs for a variety of reasons
- Numerous students sign up for a course, but do not actively participate
- Current (ongoing) MBBC 1
 - 1% of enrolled students have taken any quiz whatsoever
 - 25% of enrolled students have watched any video content

Some summary statistics

- Classes considered are MBBC1, MBBC2, CDA, DA, Case
- A total of 549,542 students enrolled
- 13 class offerings
 - Average of 42,272 students per class.
- Minimum class size of 6,742 for class MBBC2 offering 2
- Maximum class size of 101,747 for class DA offering 1.

doing this

- Scott Zeger introduced class *Cased Based Introduction to Statistics*
- Brian introduced *MBBC2*
- Martin Lindquist introduced *Statistical Analysis of fMRI Data*
- John McGready introduced *Statistical Reasoning for Public Health*

Case studies

Hi Roger.

I did your computing for data analysis course on Coursera in January of this year. I wanted to thank you personally for providing it. I think it has taken my life in a strange and fantastic whole other direction.

Following your course I also did Jeff Leaks Data Analysis course and started to play around with R and data in general more. I was doing a Business Degree at the time which I wasn't that excited about. So I ended up using R a lot for various personal projects including some college stuff which impressed my lecturers tremendously.

I then got an interview for a Google internship partially due to the R stuff I had done, of this there is no doubt. Then when I was interviewed it turned out my interviewer had also completed your course so we hit it off pretty well. So I ended up in London in Google for the summer as a financial analyst intern but I spent most of my time coding, learned JavaScript and built a big batch regression forecasting system using R amongst other things.

Case studies

"I was reading the coursera course description you are teaching on Mathematical Biostatistics Boot Camp. I have a question for you. I have a son in prison in Georgia, and I would like him to take this course because he has the background for this class. The problem I have is that he is not allowed to use the internet. Would it be possible for me to download the necessary information for him? He would be able to read and write the homework assignments, and I will write his answers for you."

Statistics, big data, data science

06/28
2011

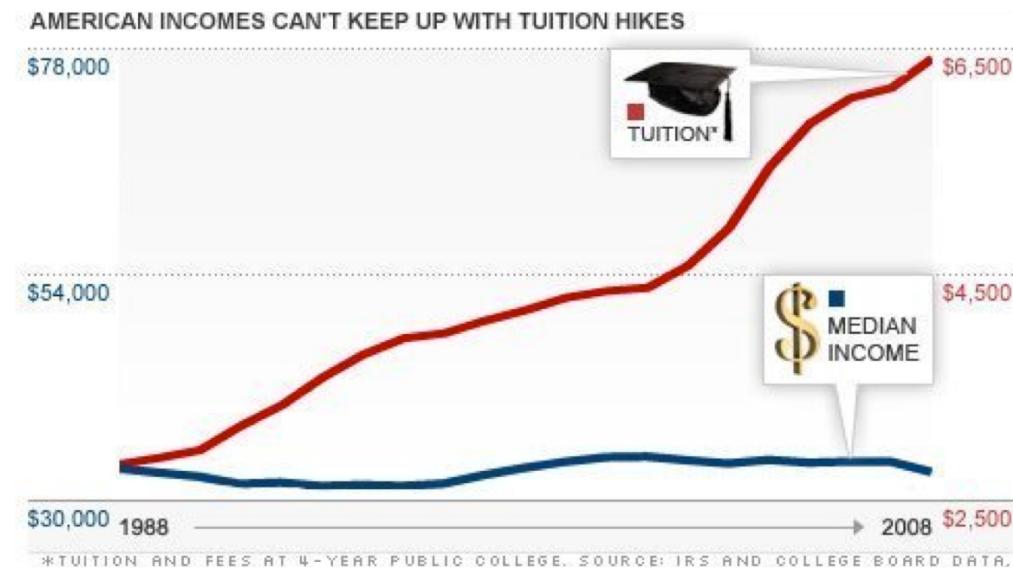
Critical Shortage Of “Data Geek” Talent Predicted By
2018

McKinsey&Company

New research by the McKinsey Global Institute (MGI) forecasts a 50 to 60 percent gap between the supply and demand of people with deep analytical talent. These “data geeks” have advanced training in statistics machine learning as well as the ability to analyze data sets. The study projects there will be approximately 140,000 to 190,000 unfilled positions in data analytics experts in the U.S. by 2018 and a shortage of 1.5 million managers and analysts who have the ability to understand and make decisions using big data.



Complimentary problems



Johhs Hopkins Data Science Specialization

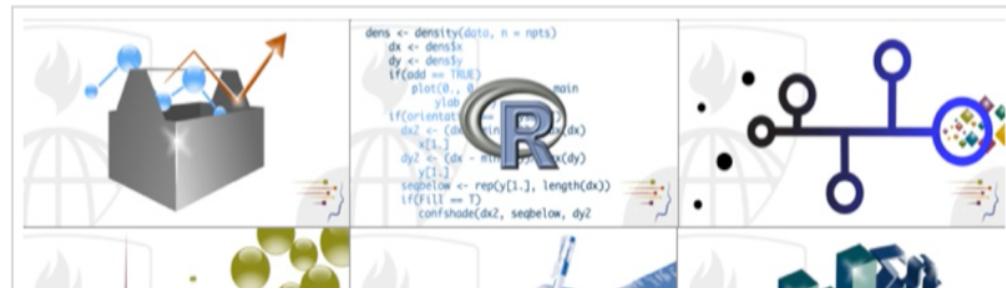
Codirected and taught by Roger Peng, Jeff Leek and Brian Caffo



The Johns

Courses

- 1 The Data Scientist's Toolbox
- 2 R Programming
- 3 Getting and Cleaning Data
- 4 Exploratory Data Analysis
- 5 Reproducible Research
- 6 Statistical Inference
- 7 Regression Models
- 8 Practical Machine Learning
- 9 Developing Data Products
- Capstone Project



34/51

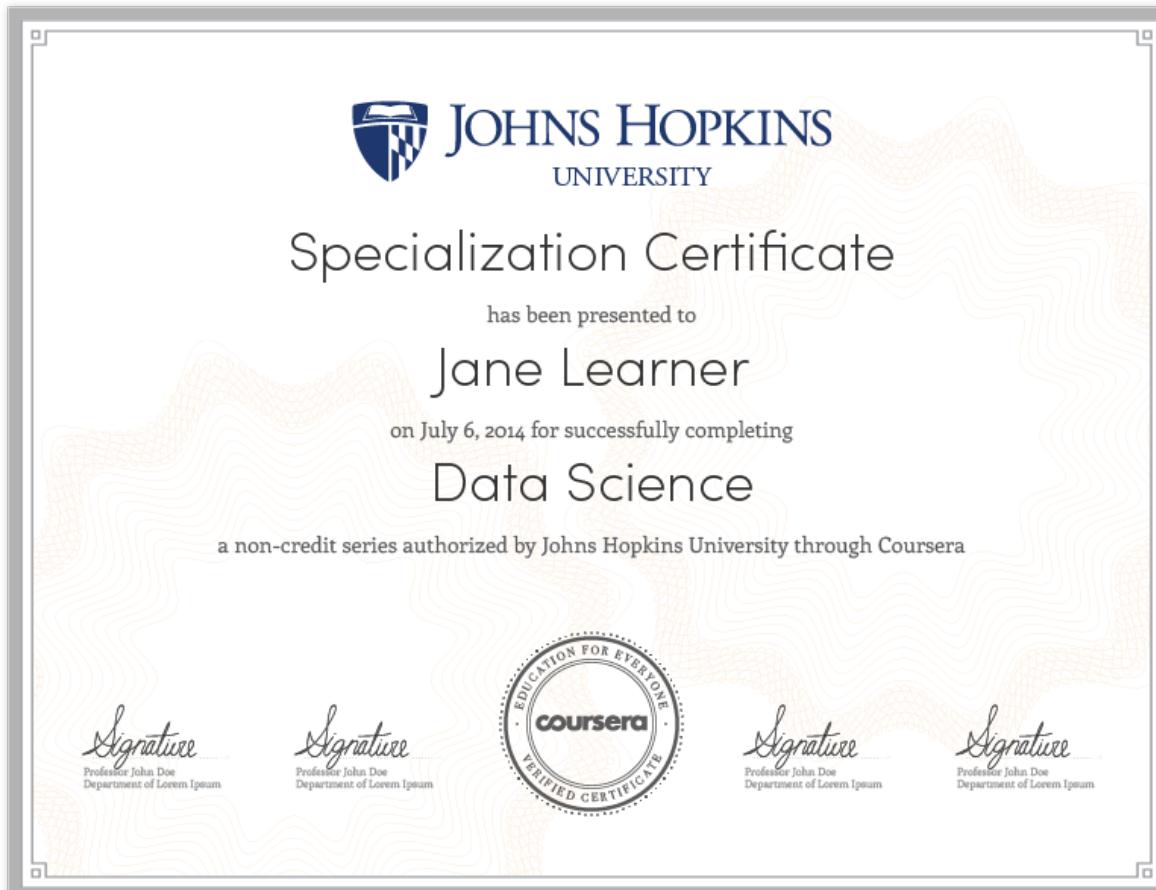
Course 4 Exploratory Data Analysis

Upcoming Session: May 5th 2014

Duration: 4 weeks

After successfully completing this course you will be able to make representations of data using the base, lattice, and ggplot2 packages; apply basic principles of data graphics to create rich analytic graphics; work with various types of datasets, construct exploratory summaries of data in question, and create visualizations of multidimensional data using multivariate statistical techniques.

Specialization certificate



Unique aspects of the program

- Completely redesigned stat curriculum
- 9 signature track courses
- 1 capstone project course
- Total cost (modular) \$490
 - \$49 per sig track for 10 classes
- Each class is four weeks
- Quizzes, in video quizzes and peer assessment projects
- Run monthly after initial rollout
- All content open source

Platform choices

- Everything done on Coursera
- All programming in R
- All lecture notes done in Slidify (common theme)
- All content open source
- Version control through git and github
- (Students will learn and use git)
- RStudio as an IDE
- knitr for reproducible documents and report writing

Standard and non-standard stat content

- Basic probability and math stat
- Statistical inference
 - Hypothesis tests, confidence intervals, likelihood
 - Brief intro to Bayesian analysis
- Regression and generalized linear models
- Statistical machine learning
- EDA
- Data analysis
- Reproducible research, report generation
- Presentations
- Interactive graphics (rgl, rCharts, shiny, manipulate)
- Data munging, obtaining data
- Programming
- Plotting (ggplot2, rCharts, R base graphics)
- Capstone project

Statistics With Interactive R Learning

<http://swirlstats.com> (<http://swirlstats.com>)



In the R console

```
> swirl()  
  
| Welcome to swirl! Please sign in. If you've been here before, use the same  
| name as you did then. If you are new, call yourself something unique.  
  
What shall I call you? Brian  
  
| Thanks, Brian. Let's cover a couple of quick housekeeping items before we  
| begin our first lesson. First off, you should know that when you see '...',  
| that means you should press Enter when you are done reading and ready to  
| continue.  
  
... <-- That's your cue to press Enter to continue
```

Class selection

```
| Please choose a course, or type 0 to exit swirl. We recommend Intro to R,  
| which is the only course we are actively developing.
```

- 1: Data Analysis
- 2: Intro to R
- 3: Mathematical Biostatistics Boot Camp
- 4: Open Intro
- 5: Test Modules

Selection: 2

Getting started

| In this module, you'll learn how to create sequences of numbers in R.

...

| The simplest way to create a sequence of numbers in R is by using the `:` operator. Type `1:20` to see how it works.

```
> 1:20
```

```
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
```

Feedback

| You're the best!

| That gave us every integer between (and including) 1 and 20. We could also use
| it to create a sequence of real numbers. For example, try `pi:10`.

```
> pi:1
```

```
[1] 3.141593 2.141593 1.141593
```

Feedback

| Nice try, but that's not exactly what I was hoping for. Try again. Or, type
| `info()` for more options.

| Enter ``pi:10`` and see what happens. ``pi`` is a predefined constant in R that
| takes on the value 3.1415....

```
> pi:10
```

```
[1] 3.141593 4.141593 5.141593 6.141593 7.141593 8.141593 9.141593
```

Feedback

- | You nailed it! Good job!
- | The result is a vector of real numbers starting with pi (3.142...) and
- | increasing in increments of 1. The upper limit of 10 is never reached, since
- | the next number in our sequence would be greater than 10.
- ...
- | What happens if we do this: `15:1`? Give it a try to find out.

Getting help

```
> info()
```

- | When you are at the R prompt (>):
- | -- Typing skip() allows you to skip the current question.
- | -- Typing play() lets you experiment with R on your own; swirl will ignore what you do...
- | -- UNTIL you type nxt() which will regain swirl's attention.
- | -- Typing bye() causes swirl to exit. Your progress will be saved.
- | -- Typing info() displays these options again.

Play mode

```
> play()  
  
| Entering play mode. Experiment as you please, then type nxt() when you are  
| ready to resume the lesson.  
  
> 1:50  
[1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25  
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50  
> 15:-15  
[1] 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1  0 -1 -2 -3  
[20] -4 -5 -6 -7 -8 -9 -10 -11 -12 -13 -14 -15
```

Starting up again

```
> nxt()  
| Resuming lesson...
```

Skipping stuff

- | What happens if we do this: `15:1`? Give it a try to find out.
- > `skip()`
- | I've entered the correct answer for you.
- | Keep up the great work!
- | It counted backwards in increments of 1! It's unlikely we'd want this behavior, but nonetheless it's good to know how it could happen.

Thanks!

