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

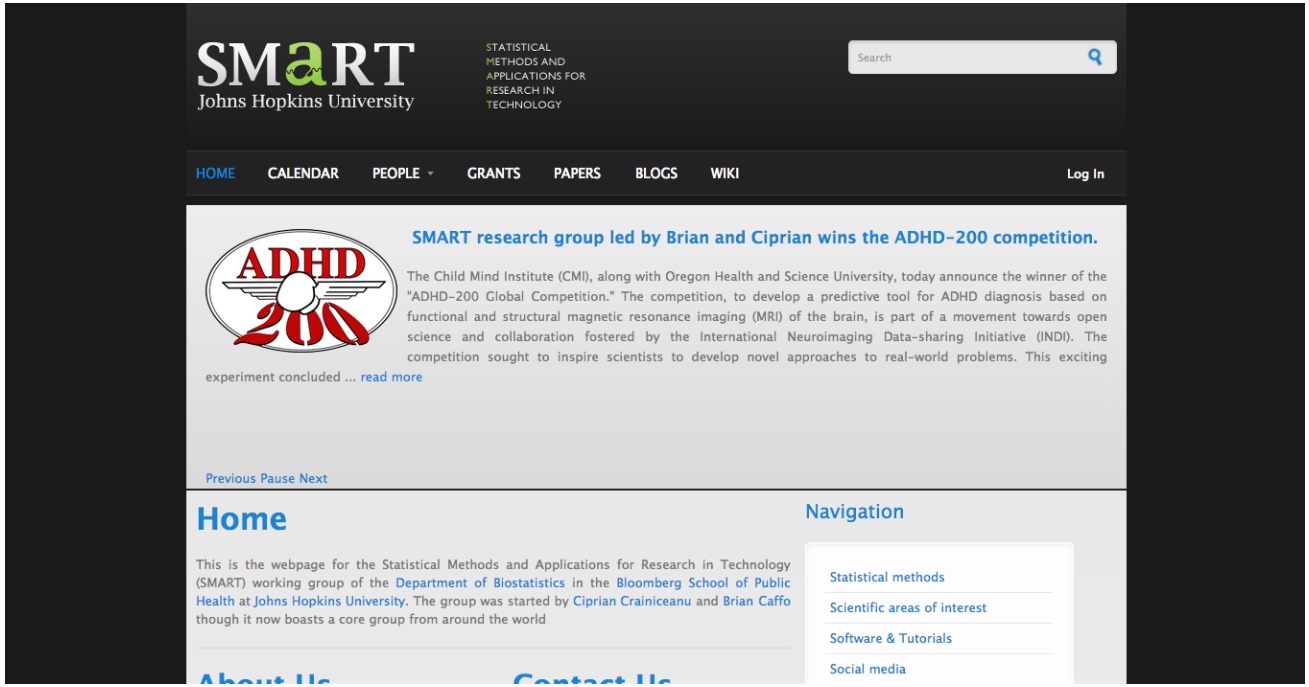
Core team

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

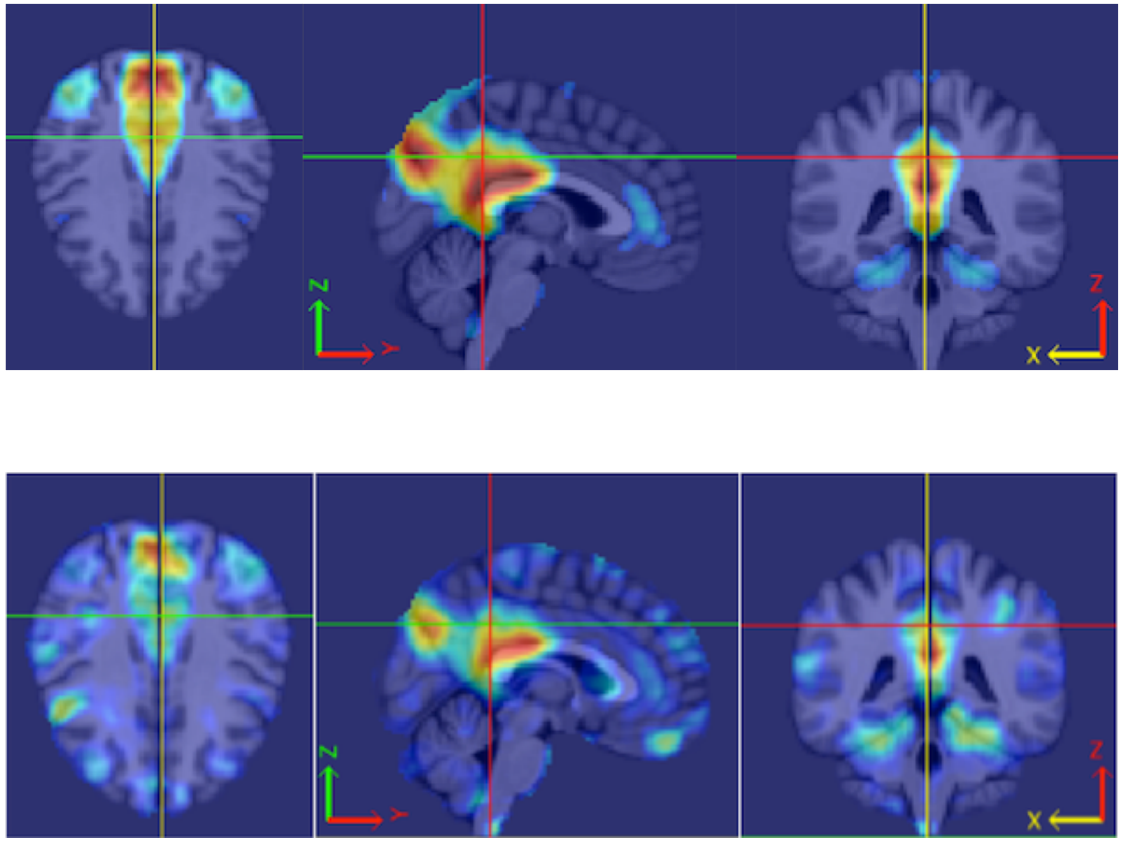
My day job(s)

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

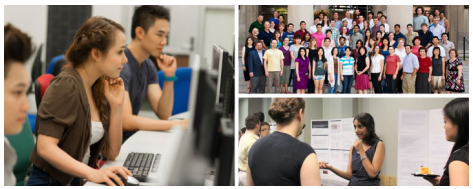


Connectomics

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 x 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

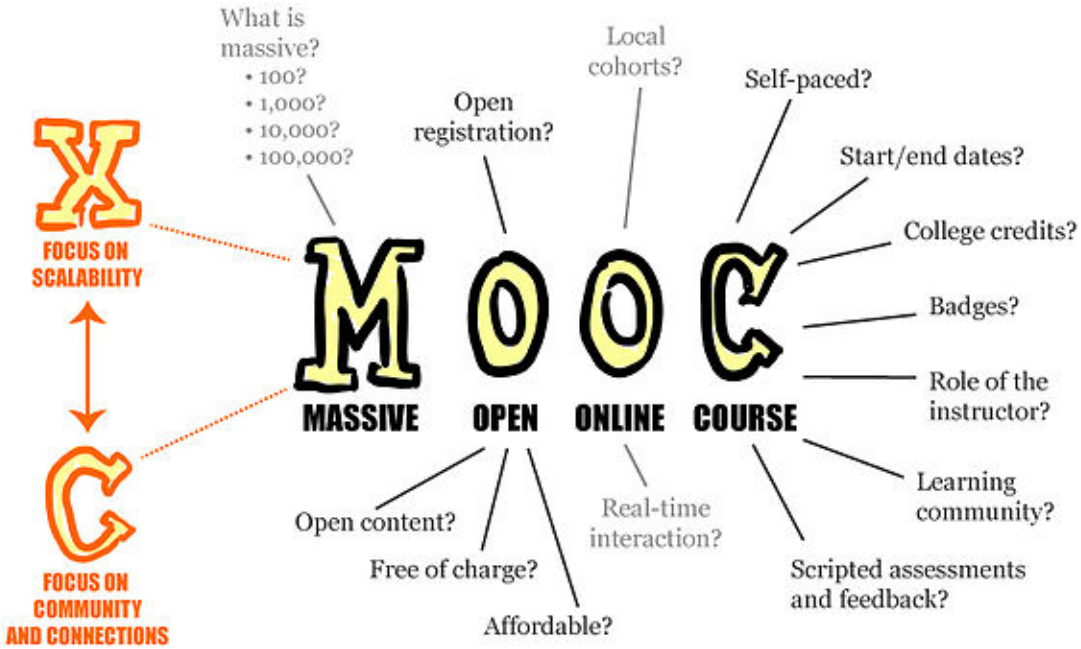
Examples

- Traditional in person teaching generally has characterists of: in person, lecture style, non-scalable, high cost, restricted access with a large amount of instructor interaction valued credentialing
- "Tranditional" 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)

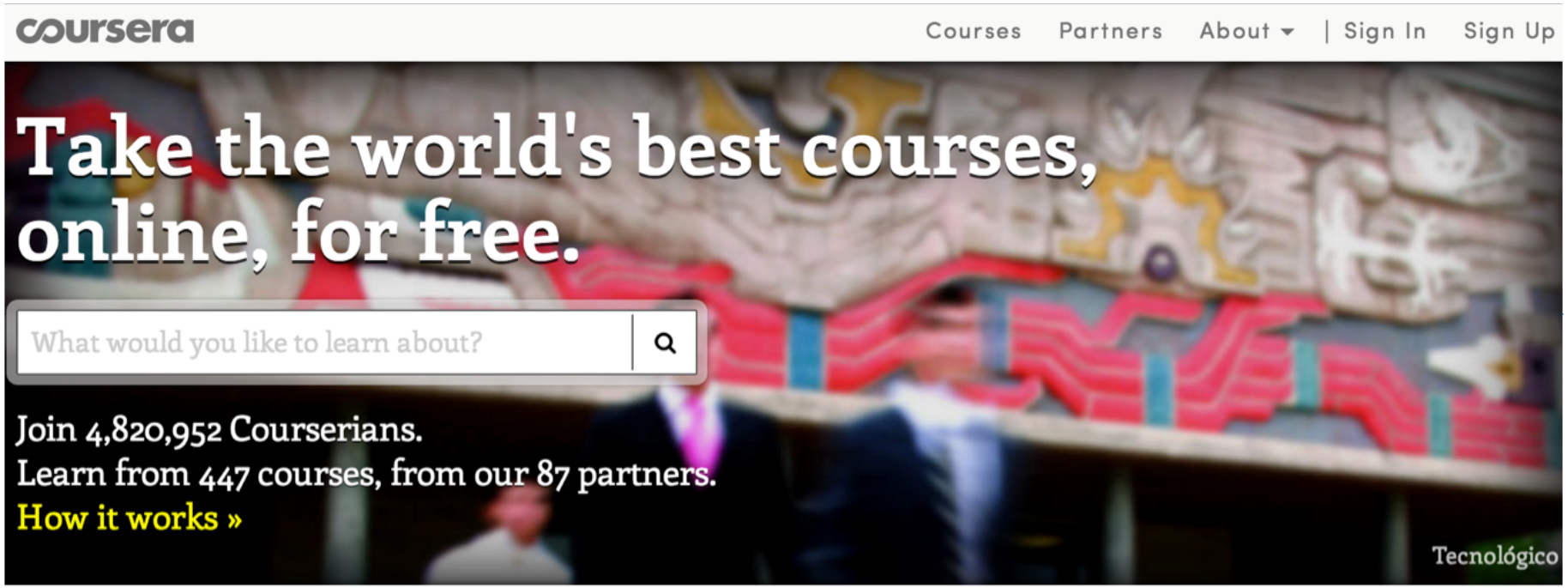
Most visible MOOC instruction sites

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

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

udemy

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

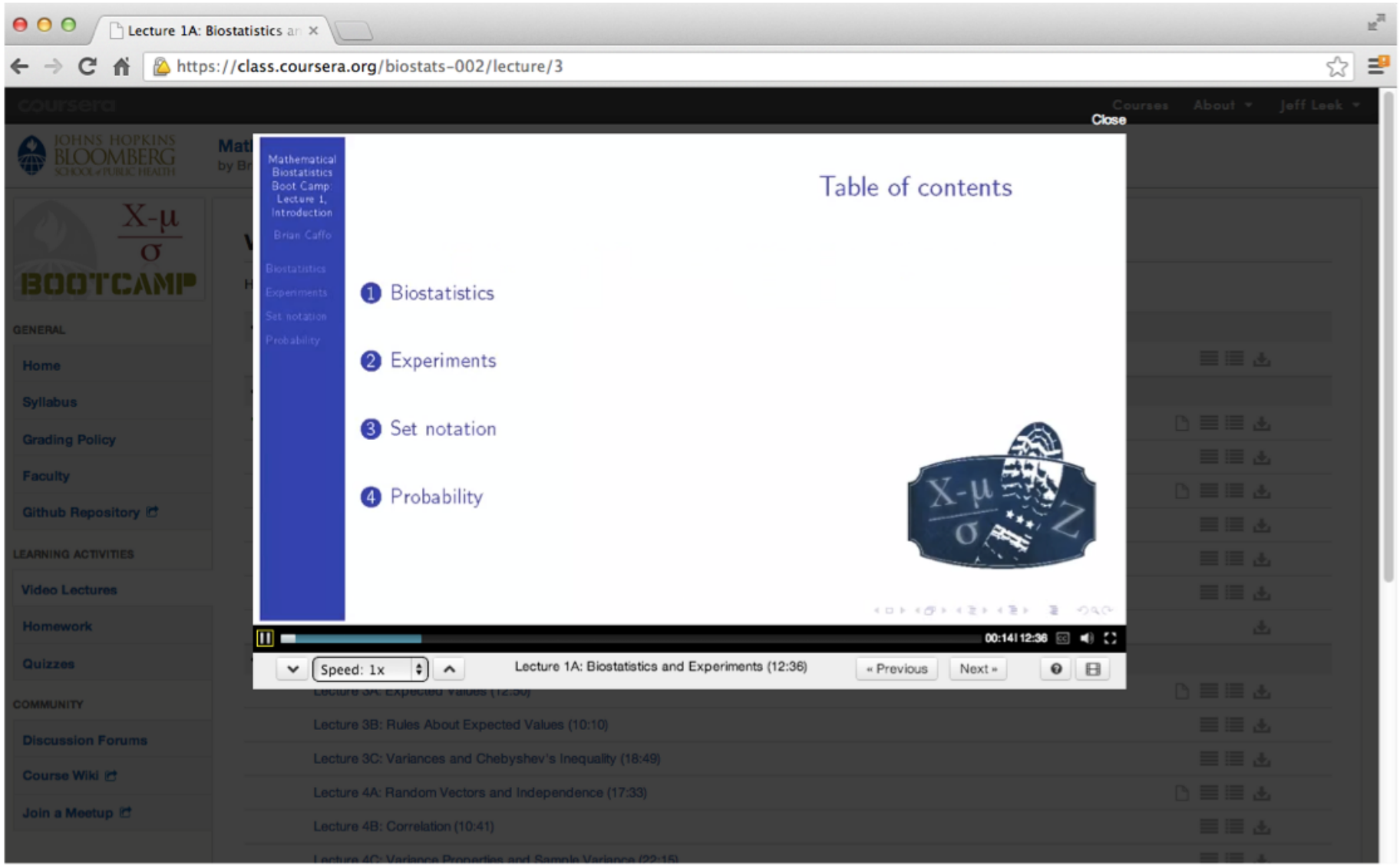


Also

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



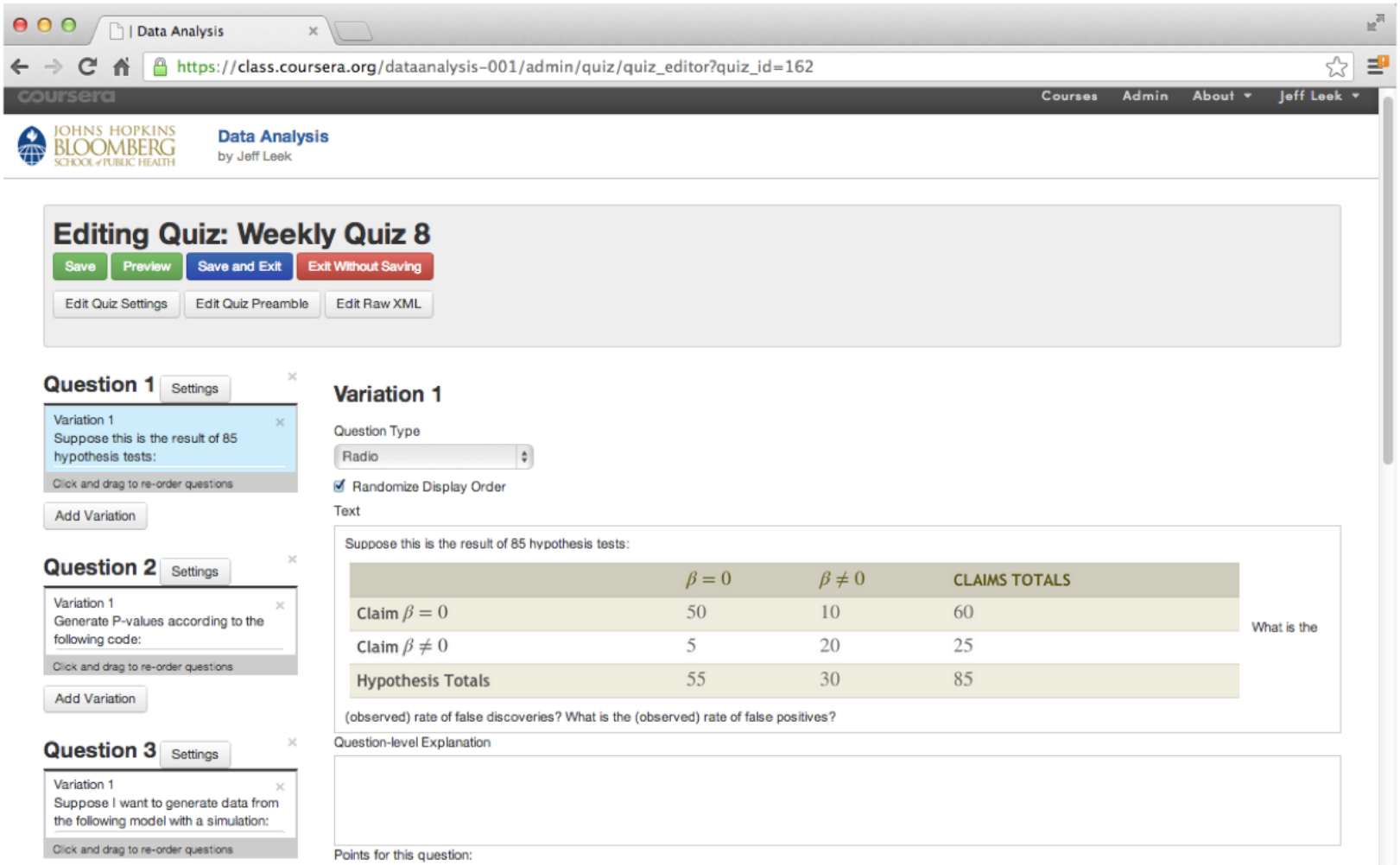
Example videos (on YouTube)

- 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)

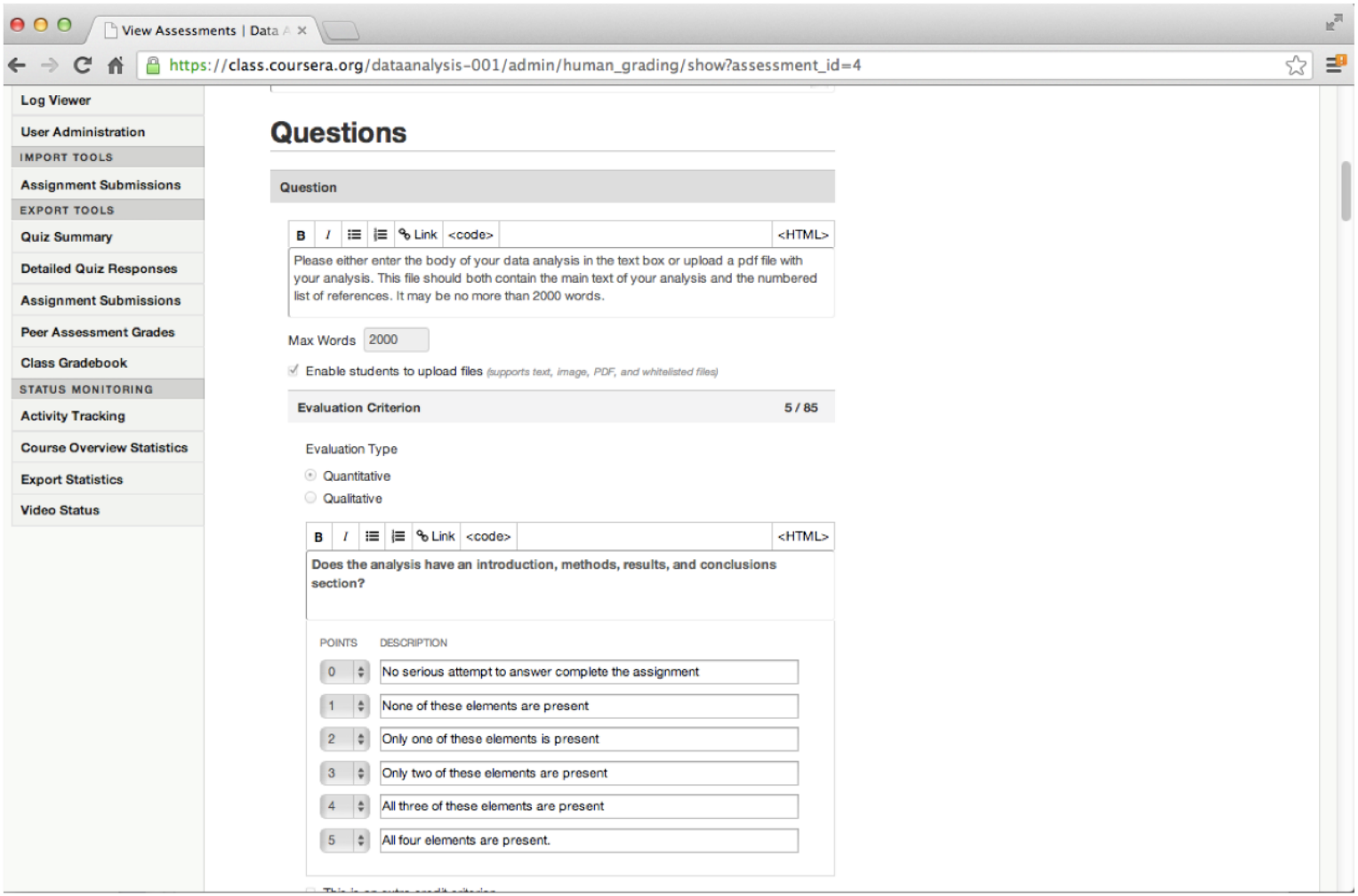
Equipment

- 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



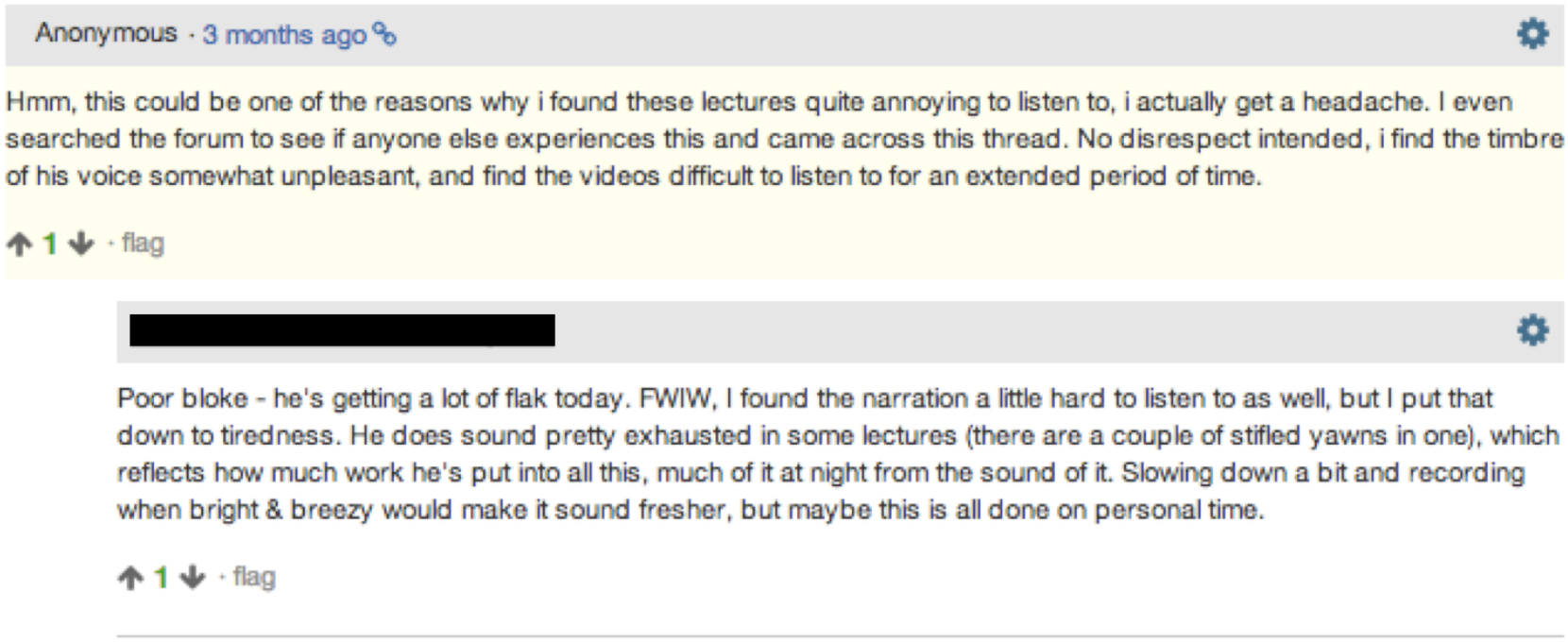
Coursera platform, peer grading



Coursera platform, forums

Main source for student interaction

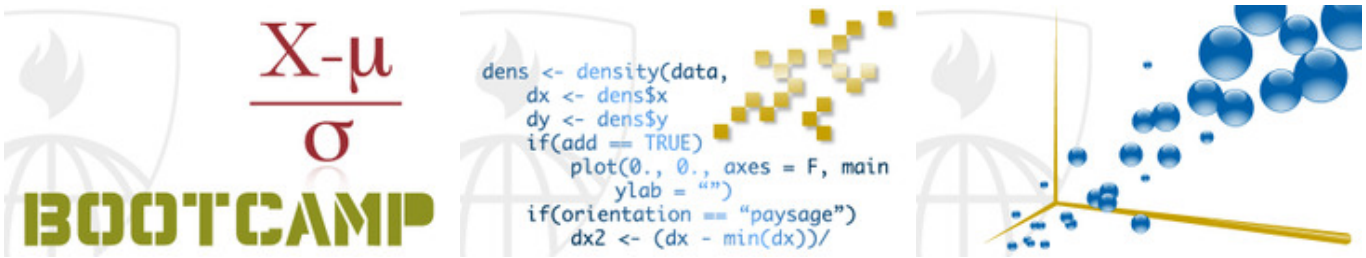
(Forums can be brutal)



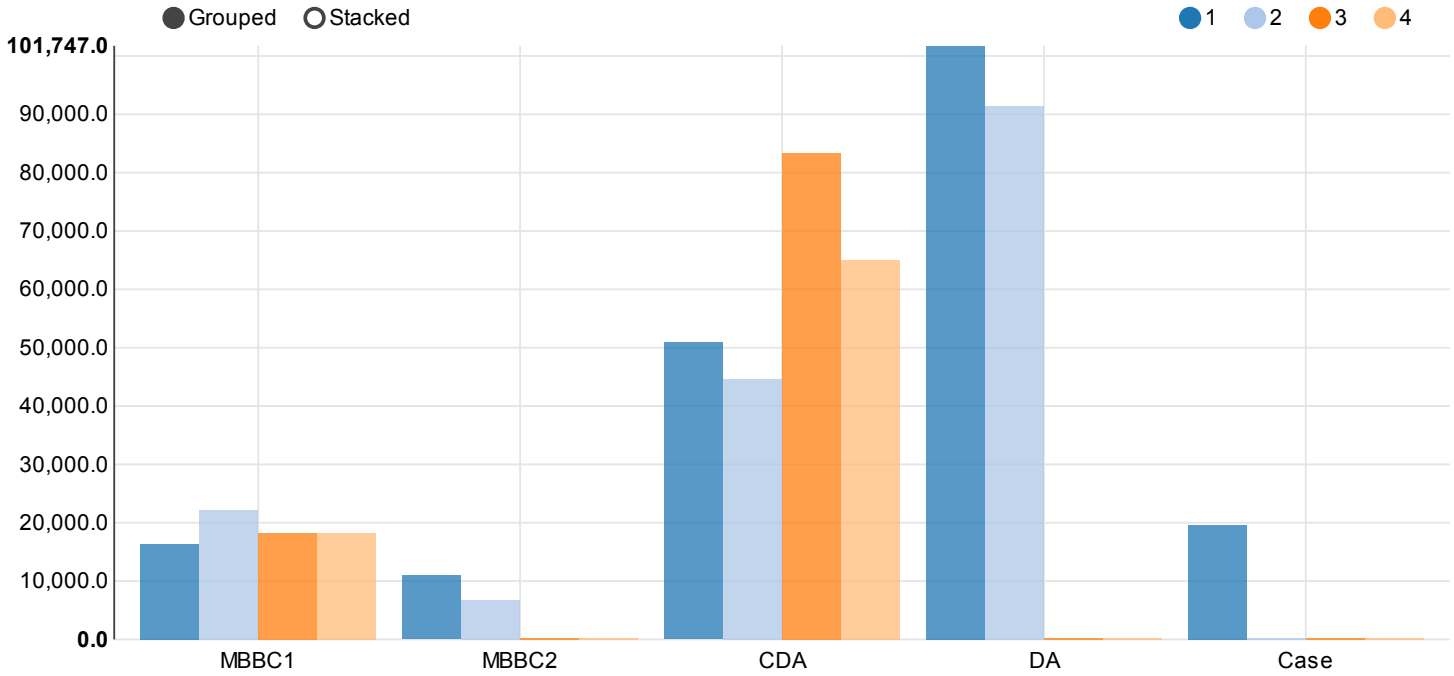
Johns Hopkins Biostat Coursera classes

Original three

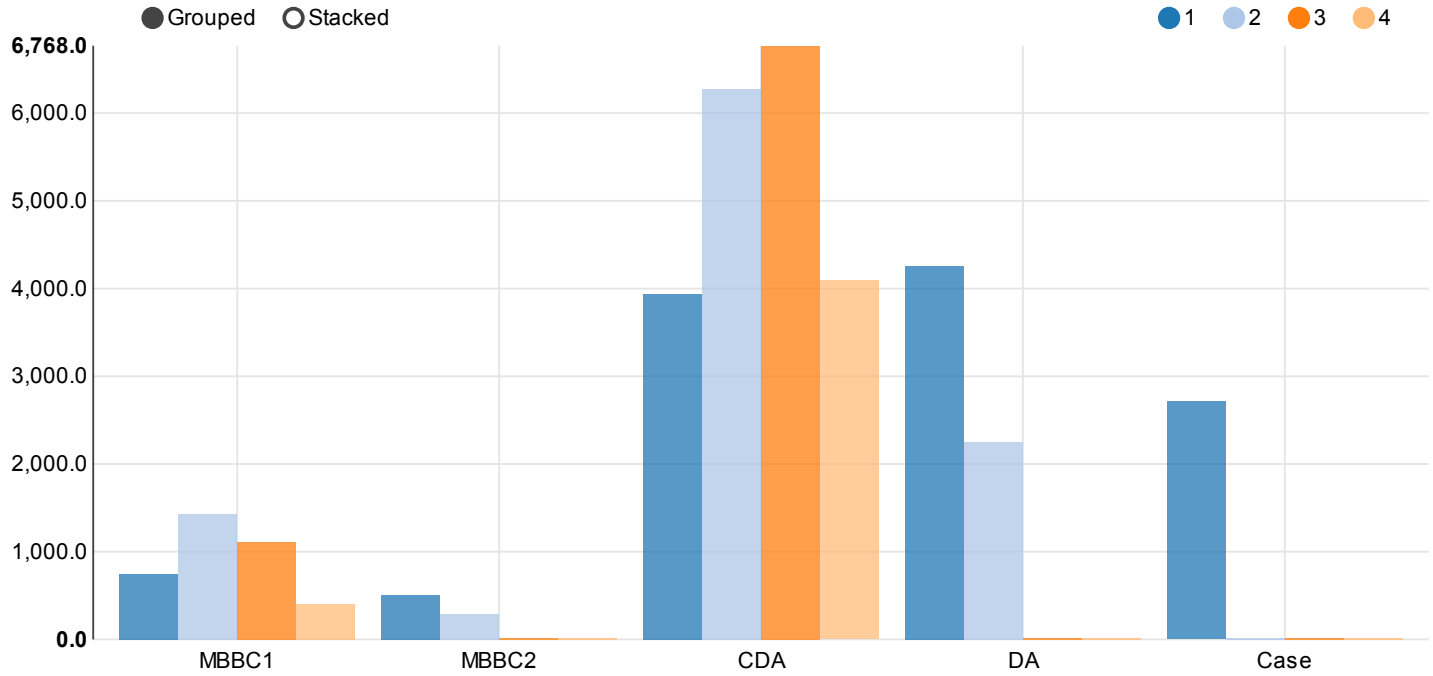
- Brian Caffo, Roger Peng, Jeff Leek
- Run 09/2012, 09/2012, 01/2013



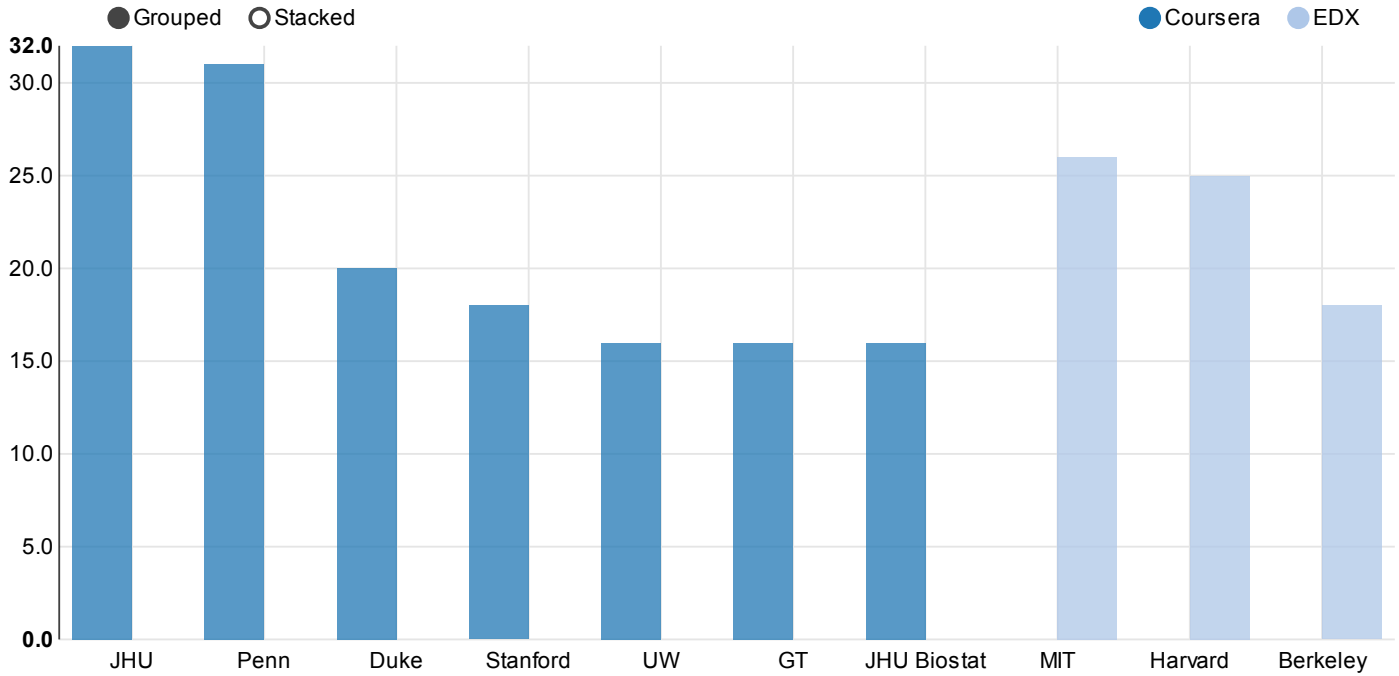
Enrollments by class and offering



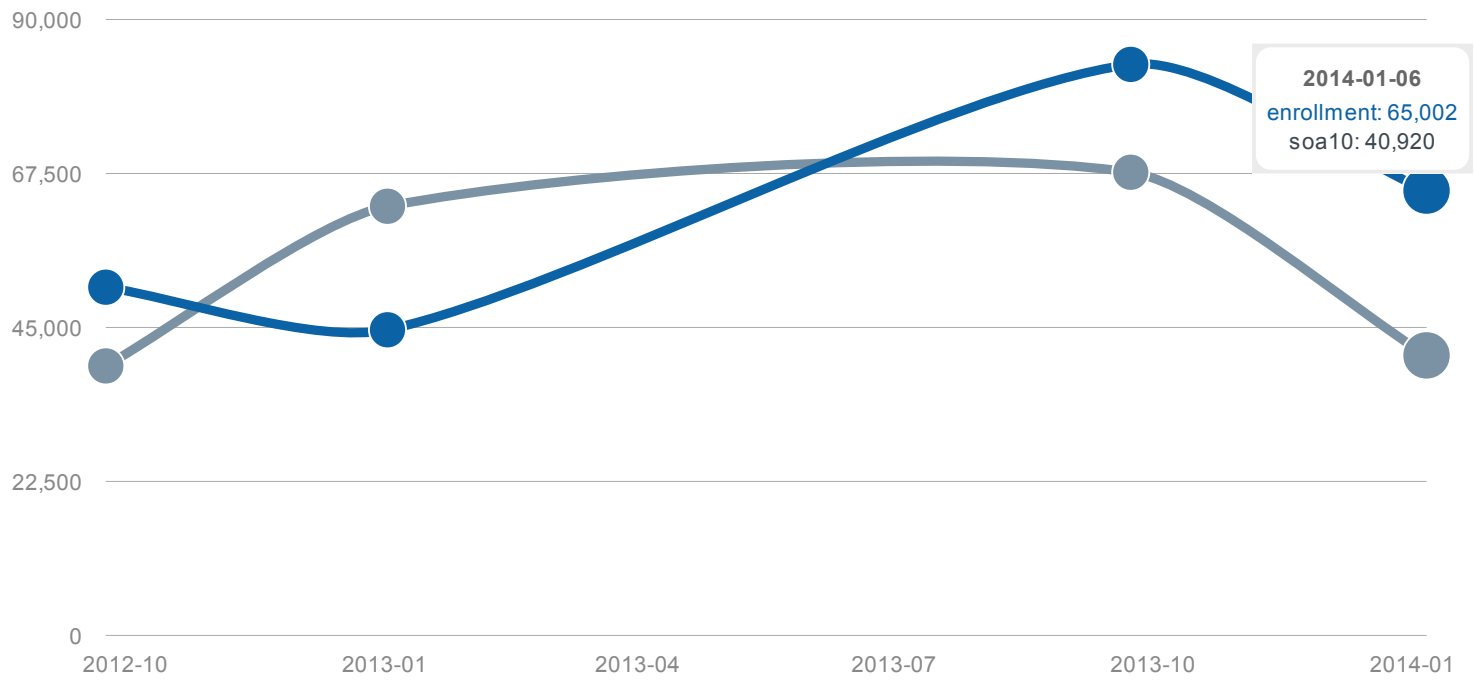
Statements of accomplishment by offering



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 introducted *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 down log 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

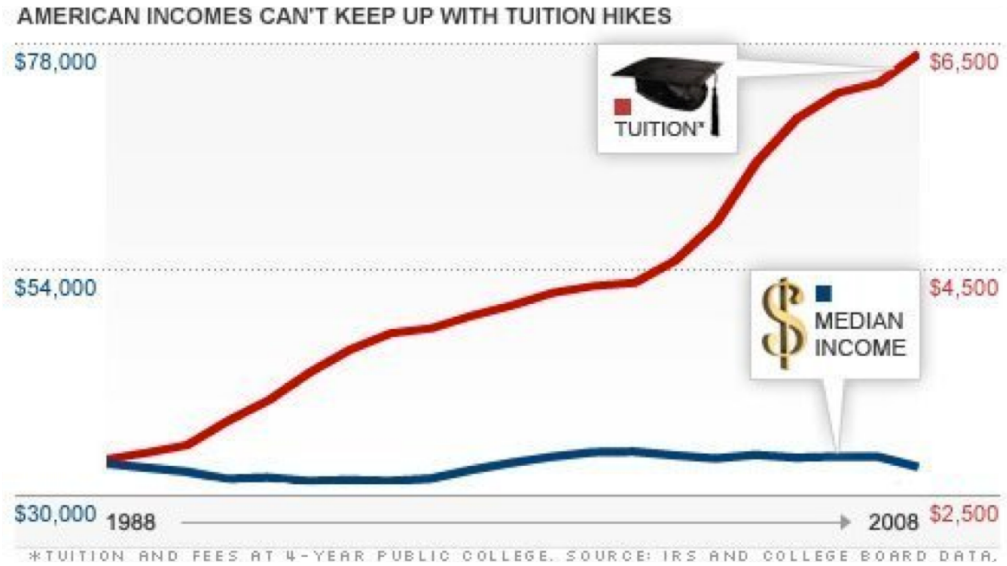
06/28
2011

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 analyz data sets. The study projects there will be approximately 140,000 to 190,000 unfilled positi 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
10. Capstone Project

Course 4

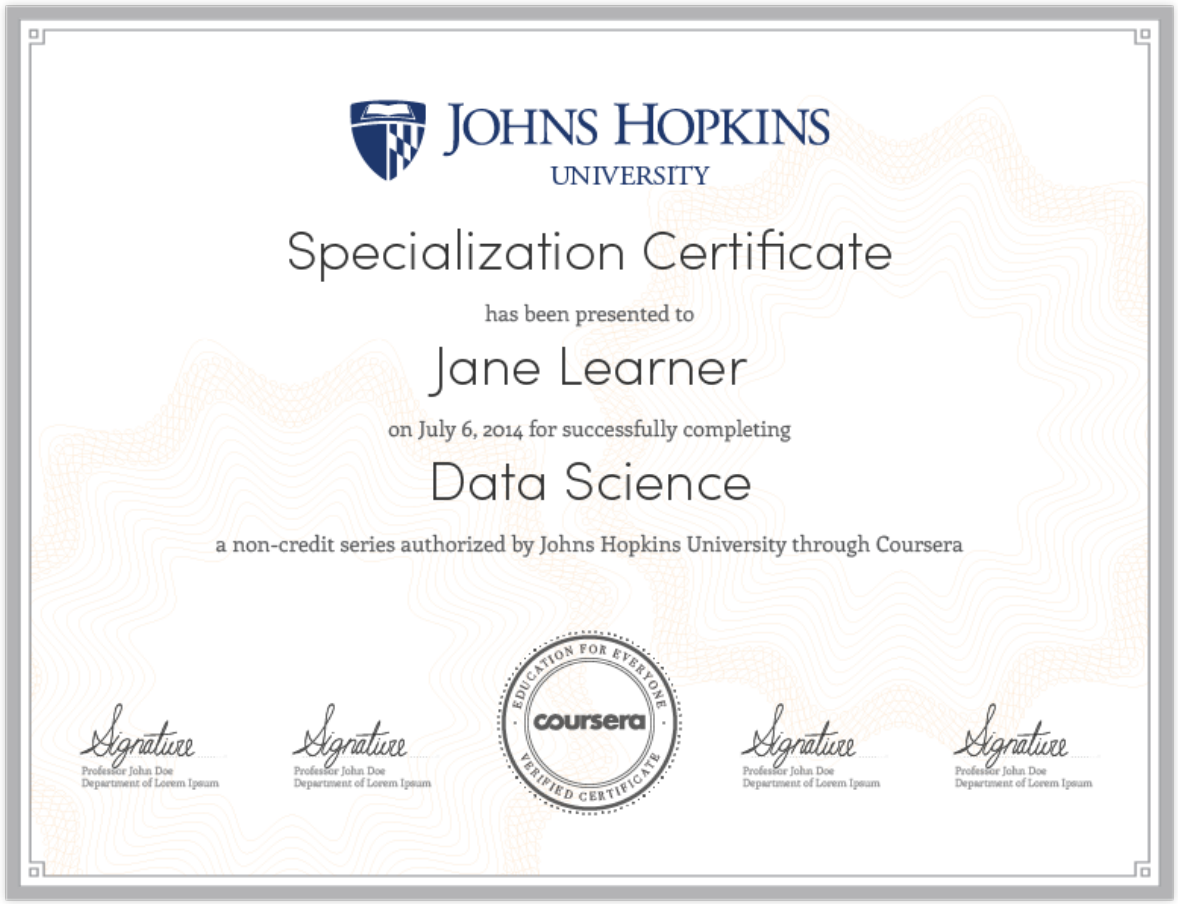
Exploratory Data Analysis

Upcoming Session: May 5th 2014

Duration: 4 weeks

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

Specialization certificate



JOHNS HOPKINS
UNIVERSITY

Specialization Certificate

has been presented to

Jane Learner

on July 6, 2014 for successfully completing

Data Science

a non-credit series authorized by Johns Hopkins University through Coursera

Signature
Professor John Doe
Department of Lorem Ipsum

Signature
Professor John Doe
Department of Lorem Ipsum

EDUCATION FOR EVERYONE
coursera
VERIFIED CERTIFICATE

Signature
Professor John Doe
Department of Lorem Ipsum

Signature
Professor John Doe
Department of Lorem Ipsum

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!