The democratization of data science education

Sean Kross
Department of Cognitive Science, The University of California San Diego
and
Roger D. Peng
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
and
Brian S. Caffo
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
and
Ira Gooding
Center for Teaching and Learning, Johns Hopkins Bloomberg School of Public Health
and
Jeffrey T. Leek
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

September 18, 2019

Abstract

Over the last three decades data has become ubiquitous and cheap. This transition has accelerated over the last five years and training in statistics, machine learning, and data analysis have struggled to keep up. In April 2014 we launched a program of nine courses, the Johns Hopkins Data Science Specialization, which has now had more than 4 million enrollments over the past three years. Here the program is described and compared to both standard and more recently developed data science curricula. We show that novel pedagogical and administrative decisions introduced in our program are now standard in online data science programs. The impact of the Data Science Specialization on data science education in the US is also discussed. Finally we conclude with some thoughts about the future of data science education in a data democratized world.

Keywords: Education, Statistical Computing, Applications and Case Studies

1 What is the Johns Hopkins Data Science Specialization?

Data has become dramatically cheaper to collect and store over the last several decades. This phenomenon has touched fields ranging from the government, to social sciences, to molecular biology (O'Connor 2014). As data have become commoditized and democratized, there has been a dramatically increased demand to analyze and extract information from these data. The demand for data analytic training was made clear when three Stanford computer science professors – Daphne Koller, Andrew Ng and Sebastian Thrun – opened their Stanford courses focusing on data analysis for free to the world and enrolled hundreds of thousands of students (Lewin 2012).

Recognizing the demand for these classes – later labeled massive online open courses (MOOCs) - these three professors launched companies, Coursera (*Coursera* 2016*a*) and Udacity (*Udacity* 2016), to offer similar courses in partnership with professors at other universities. These companies, and the nonprofit EdX (*edX* 2016) built on earlier efforts, such as OpenCourseware (OpenCourseWare 2016), for delivering top level content to a broad audience.

As biostatistics professors focusing on the development and in particular the application of data analytic methods to solve problems in biomedicine, we were also enthusiastic and equipped to teach this type of class. The effort that culminated in the Johns Hopkins Data Science Specialization occurred over a period of approximately 1.5 years. It was the work of three faculty at the Johns Hopkins University Department of Biostatistics with the assistance of key staff members who will be mentioned during the description of the program.

Our online educational efforts were spurred by a partnership that was formed between Johns Hopkins and Coursera (Gooding et al. 2013). Brian Caffo, Roger Peng, and Jeff Leek elected to teach three classes on Coursera: *Mathematical Biostatistics Boot Camp*, *R programming*, and *Data Analysis*. These three classes reflected course materials that were developed for students at Johns Hopkins in mathematical statistics, R programming, and the art of the analysis of data. In development, the new MOOCs were direct translations

of the associated courses to this new format. *Mathematical Biostatistics Boot Camp* and *Computing for Data Analysis* launched in September 2012 and *Data Analysis* launched in January 2013.

The lectures were recorded and distributed as videos. All three courses included multiple choice assignments. Two of the courses included unique assessment features. The R programming class included programming exercises where code output was automatically evaluated. The Data Analysis class included a final project that was peer assessed. Each submitted project was randomly assigned to a small number of other students, who graded the project using a fixed rubric (Leek 2017). Students then received the median score from these peer assessments on their project. These assessment tools, and related ones based on automation and crowdsourced grading, are common in MOOC platforms to address scalability of project based learning. They played a significant role in the development of the Johns Hopkins Data Science Specialization.

Propelled by the popularity of early MOOCs (Jordan 2014), the enrollment in these first three classes was extraordinary (Table 1) and was an early indication of the demand for data science focused courses. This demand has since been validated by the popularity of not only online courses in data science, but a host of boot camps (Eggleston 2017), internships, and a dramatic increase in enrollment across statistics, machine learning, and data science undergraduate programs (Guzdial 2017).

Table 1: **Number of Students in Initial Online Courses**

The size of the student enrollment and completion numbers from the initial courses we created for Coursera.org strongly encouraged our development of the Data Science Specialization. These courses launched between September 2012 and January 2013, with the final sessions occurring in July 2014.

Course	Enrollment	Completions
Mathematical Biostatistics Boot Camp 1	109,789	4,150

Course	Enrollment	Completions
Mathematical Biostatistics Boot Camp 2	23,842	944
Computing for Data Analysis	243,987	21,069
Data Analysis	193,126	6,500
Biostatistics Case Study	39,140	3,322

Seeing the demand for data science spiking, we were inspired to create an entire data science curriculum that would deviate from both the in person program at Johns Hopkins and other statistics and biostatistics programs nationally. This coincided with an effort by Coursera to create sequences of MOOCs or "Specializations". The launch date for the original set of specializations was April 2014. Using our lightweight production process we created a series of nine courses and launched one of the first Specializations on the Coursera platform. Each course is designed to be completed over four weeks, and the topic for each week is enumerated below. The following is an overview of the classes in our program:

1. The Data Scientist's Toolbox (*Coursera* 2016*b*)
    1. Installing software, overview of the specialization.
    2. Git, GitHub, and Markdown.
    3. Understanding data and experimental designs.
    4. Creating your first GitHub project.
2. R Programming (*Coursera* 2016*c*)
    1. Types and data structures in R.
    2. Control flow and functions.
    3. Mapping functions and debugging.
    4. Programming simulations.
3. Getting and Cleaning Data (*Coursera* 2016*d*)
    1. Loading raw data from files.
    2. Getting data from the internet.
    3. Manipulating data frames.

    4. Dates and regular expressions.

4. Exploratory Data Analysis (*Coursera* 2016*a*)

    1. Exploratory graphs.
    2. ggplot2 and the Grammar of Graphics.
    3. Clustering and dimension reduction.
    4. Case studies in scientific graphs.

5. Reproducible Research (*Coursera* 2016*e*)

    1. Structure of a data analysis.
    2. Literate programming with R Markdown.
    3. Principles of reproducible research.
    4. Case study in communicating results.

6. Statistical Inference (*Coursera* 2016*f*)

    1. Probability and expected values.
    2. Distributions and asymptotics.
    3. Confidence intervals and p-values.
    4. Power and the bootstrap.

7. Regression Models (*Coursera* 2016*g*)

    1. Least squares and linear regression.
    2. Multivariate regression.
    3. Residuals and diagnostics.
    4. Logistic and Poisson regression.

8. Practical Machine Learning (*Coursera* 2016*h*)

    1. Prediction, errors, and cross validation.
    2. The caret package in R.
    3. Trees and random forests.
    4. Regularized regression and combining predictors.

9. Developing Data Products (*Coursera* 2016*i*)

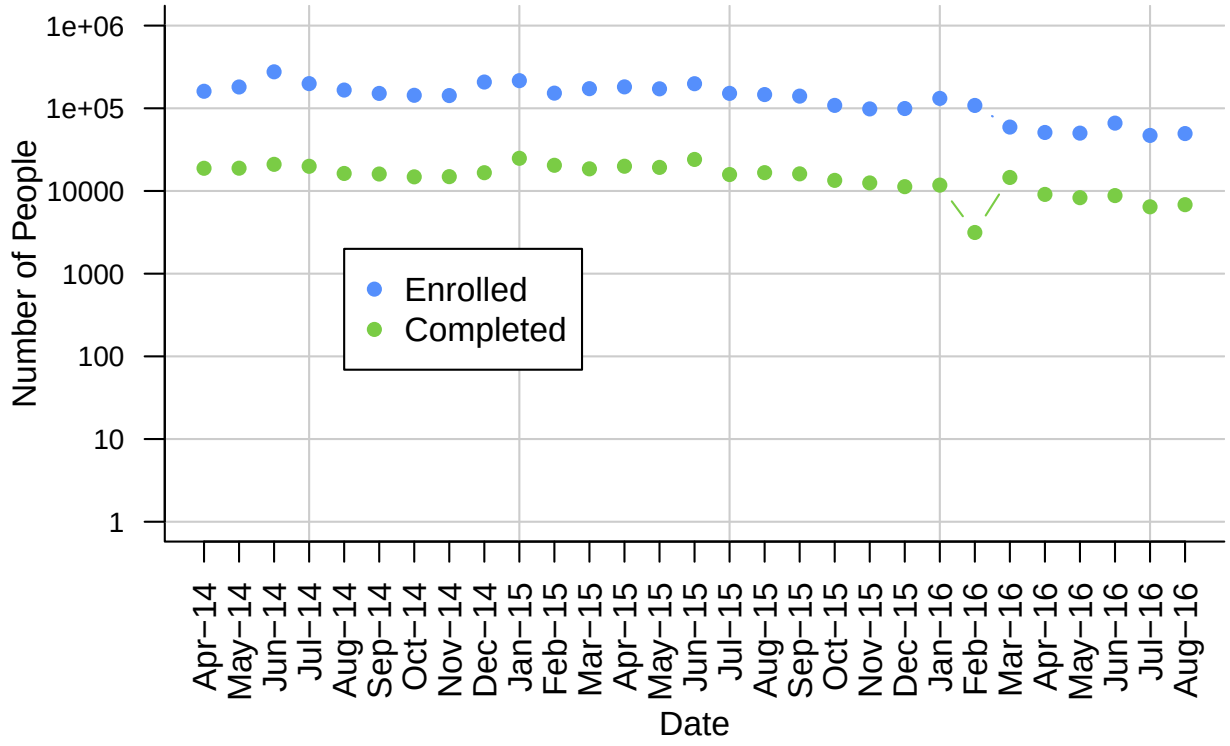

Figure 1: **Number of students in the Johns Hopkins Data Science Programs** The number of students enrolling (blue) and completing (green) courses in MOOCs created by the Johns Hopkins Biostatistics Department over time. Coursera launched their new platform in February 2016.

1. Web applications and interactive visualizations.
2. Dashboards and maps.
3. Writing R packages
4. Case study in web application development.

This program has now been running since April 2014 and has enrolled more than 4 million students and over 875,000 people have completed a course. It is the largest MOOC program in history.

Over time enrollments have appeared to reach a steady state (Figure 1). A large number of students continue to enroll in the program, now one of the most widely recognized sequences of MOOCs. Despite the well-known criticism that MOOCs have low completion rates (Clow 2013), the number of completers of the JHU Data Science Specialization - at the time of this writing (n = 6,235 completers) - is almost double the entire output of

masters students from the ten largest programs in statistics and the ten largest programs in biostatistics combined (n = 2,056 degrees) over the years 2011-2013 (Higher Logic 2016).

The courses can be audited on Coursera for free, where students have access to all course materials and all assessments that do not require peer feedback. All lecture materials and textbooks for the courses are provided for free via Leanpub and GitHub, two platforms that we will discuss in detail in later sections. A subscription to the specialization can be purchased for $50 per month, which grants students access to peer graded assignments and enables them to earn a certificate for each course they complete. Financial Aid is also available from Coursera which students must apply for and is approved by Coursera on a case-by-case basis.

The funds that we have raised through this program have been treated by our school and department much like a grant award, which has given us a great amount of flexibility. We have hired a full time videographer and a full time software and content developer, in addition to partially funding other staff. Our combined efforts have resulted in the creation of more online courses where we have reapplied and refined the innovations of the Data Science Specialization. We hope that reinvesting in online education will allow us to educate audiences we would not have reached otherwise.

Part of the reason for this success was simply the serendipity of launching the program at the right time. But some of the innovations produced by the program made it unique and have now led other groups to build on those ideas. In the following sections we summarize the differences we think gave us an advantage and propose some ideas for the future of data science education.

2 Philosophical principles behind our innovations

One of the key reasons for the success of the *Johns Hopkins Data Science Specialization* is our focus on a few key principles:

1. focusing on developing large quantities of quality open source content,
2. targeting broad audiences as opposed to niche experts,
3. thinking about the users of online education.

From the conception of the program, content was focused on our hypothesized motivations of online data science students. This lead to several strategic decisions - for example to reduce production time to increase the speed of generating content. While the resulting reduction in production value could be seen as a negative, it was turned positive by developing open source, modular, and public content that could be updated quickly. A second strategic decision was to orient programs for introductory and general audiences, rather than focusing on more technical material only accessible to experts with many prerequisites. This is part of a key general strategic decision to focus on the typical needs of online education users. For example, the modal MOOC students take courses in their spare time and arrive at the material with little background (Zheng et al. 2015, Dillahunt et al. (2014)). Such learners would not be interested in courses and programs that have typical universities requirements that a full time student would be able to meet.

The Data Science Specialization also differs from traditional university programs in the amount of individual attention that each student receives from the professors and the teaching staff. The ratio of instructors and staff to students fluctuates around one to one thousand, therefore it would not be possible to scale traditional lectures, office hours, and assignments without hiring as many staff as a small university. We attribute the success of the program and our relatively small team to the innovations we discuss in the following sections.

The pace of our courses is also much faster than a university course, which means that students may take more than the prescribed four weeks in order to get through all of the content in a particular course. Thankfully Coursera's platform has built in flexibility which allows learners to complete the course at their own pace. The relatively low ratio of course staff to the number learners means that our courses can be offered at a much lower cost compared to traditional programs. To compensate for the reduced individual attention available to learners, we leveraged several emerging platforms that allowed our team to scale data science learning to millions of learners.

Based on the previously discussed motivating principles we adapted our platform, content, structure and community to maximize our reach. We now detail several choices we made in each of these categories.

3 Innovations in Platform

3.1 *Leanpub*

The high cost of textbooks with frequent edition updates is one of the biggest justifiable complaints from students (Nicholls 2009). From the beginning we had released all course material under an open source license on GitHub. However, we needed a mechanism for releasing affordable books and course notes. We used the Leanpub (*Leanpub* 2016) platform, which offered a variable pricing model with no minimum payment required. Textbooks for several of our courses including Statistical Inference for Data Science, The Elements of Data Analytic Style (Leek 2016), and R Programming for Data Science (Peng 2016) have sold collectively hundreds of thousands of free and paid copies on the platform. The self-publishing model and markdown based writing format made it easy to translate notes written for courses into textbooks and rapidly release both at the same time. Since then, Brian Caffo, Roger Peng, and Jeff Leek have published a collective 12 books on Leanpub, all available for free on a variable pricing model.

3.2 *Capstones*

After developing 9 courses for the data science specialization, all that was left was to build a capstone project in which learners could apply all of the skills that they had learned through the series in a larger project. The first capstone project was a collaboration between Johns Hopkins University and SwiftKey, a company that made a predictive keyboard app for Android phones. Partnership with industry brought on several political and legal challenges but provided the opportunity for students to work on a project that was guided by experts in the field.

The basic idea behind the capstone project was that students would have to take a large corpus of text data and build a predictive model for text, essentially a streamlined version of a task SwiftKey would have to do at production scale. Students had to build a web application using the Shiny (Chang et al. 2015) framework for R. Shiny allows students to rapidly prototype and share web applications which are written purely in R code. We instructed students to build a web application where a user could enter a string

of text into a textbox, and then the user would see the next predicted word. In order to build the app students had to parse and clean a corpus of text data. In order to make a algorithm students created n-gram models using any method of their choosing, though they were encouraged to consider the runtime and the amount of memory that their algorithm would require. Ultimately, they had to build a Shiny app that demonstrated the algorithm to others. Each Shiny app was hosted on RStudio's shinyapps.io platform (*shinyapps.io* 2016), and RStudio was generous to provide limited-time free hosting for all learners in the capstone. Students graded each others assignments via Coursera's peer grading system. The rubric that students used to perform the peer assessment includes criteria for their app responding quickly and being able to process input from new articles and social media. SwiftKey's engineers helped to develop the project and were gracious with their time by offering to a live Google Hangout "office hours" where they answered questions from the learners in the capstone course.

Since our original capstone project, we have developed a second capstone with Yelp. Learners use data from the Yelp Dataset Challenge (Yelp 2018) which includes information about the location of restaurants, user ratings, reviews, and more. In the capstone learners must produce a written report and a slide presentation including R code and data visualizations in order to explore some part of the dataset. Learners have more freedom in terms of what data they want to look at, and project topics have included how restaurant review sentiment is related to ratings, how the social networks of Yelp users effect their influence on reviews, and how business can use Yelp data to forecast sales.

Now, the capstone projects rotate between the two, giving learners an opportunity to choose what type of project they wish to complete. Throughout the development of capstone projects we focused on making the problems real, as ambiguous as possible so that students would have to struggle with the same problems as practicing data scientists, and interesting so students would be motivated to complete them.

3.3 *Swirl*

One of the challenges of providing personalized activities for large numbers of students is the lack of interaction with an instructor. One approach we took to solving this problem was

to develop the swirl platform (Carchedi et al. 2014). The *swirl* R package walks students through a linear interactive programming tutorial in the R console. Students progress through *swirl* lessons by typing R command into the console, answering multiple choice questions, or by writing their own R scripts. If a student's answer is correct according a set of answer tests provided by a *swirl* lesson author, then the student is allowed to move on to the next exercise. If the student's answer is incorrect then *swirl* offers them a hint. Lessons for *swirl* can be written by anyone by combining a yaml structure for each lesson that specifies the order in which questions are asked, the question type, and the correct answer for a question.

swirl was built so student could learn programming, R and statistics in R ("learn R in R"" is the catchphrase for *swirl*). It was developed by Nick Carchedi, Brian Caffo, Lauren Williams, Ethan Schwartz, Gina Grdina, Sean Kross and Bill Bauer.

A key step was to incorporate swirl into Coursera via open education APIs. *swirl* has since become a main component of the program and one of the most popular R packages with over 600,000 downloads and it has garnered a community that has translated the package into eight languages.

4 Innovations in Content

There are several elements in our data science program that are not formally addressed as a part of a traditional statistics curriculum. We highlight these differences in Figure 2 by contrasting topics covered in the Data Science Specialization with courses from comparable academic programs. We include the Master of Data Science program from the University of British Columbia in this comparison to provide a richer contrast between statistics and data science curricula.

The first course in our program, called The Data Scientist's Toolbox, aims to set up the computational environment for students who are new to scientific computing. In this course we address material that is usually left to students to figure out on their own or in office hours including how to configure R and RStudio, the command line, Git, and GitHub, all across different operating systems. Consider the rate at which new data science technologies are being released, we believe that teaching students to set up their own programming

environment is an essential skill to have as a data scientist.

Raw data is often not easily parsed or interpreted by computer programs, and preparing raw data for analysis, modeling, and visualization is often the job of a data scientist. We created a course called Getting and Cleaning Data in order to emphasize the importance of these data preparation skills. Data preparation involves two general steps: first the data must be imported into a programming environment where it can be manipulated, and then the data is transformed into a structure that eases its later use. The importance of this data janitorial work often neglected by traditional programs despite the fact that it often occupies a significant amount of the time spent working with data (King & Magoulas 2017).

The importance of ensuring that one's research is reproducible has entered the scientific mainstream in recent years (Stodden et al. 2014), however the topic is often presented as yet another infrastructure burden that is expected of a data scientist. We designed our Reproducible Research course through a different philosophy by emphasizing how tools like Markdown and knitr can aid a data scientist while developing an analysis. This topic has been increasing in popularity and is now fundamental to many data analytic training programs (Teal et al. 2015).

The results of a data analysis deserve to be communicated as widely and coherently as possible, and this belief motivated the creation of our Developing Data Products course. A large proportion of this course is dedicated to teaching Shiny, a web application framework for R. Shiny allows students to create interactive web applications simply by writing R code, which can there be shared online. This allows our students to share models and data visualizations that app users can use without any knowledge of R. We also teach students how to create their own stand-alone websites with R Markdown, featuring interactive graphics and maps powered by the plotly and leaflet libraries for R. We believe that the creation of data products is integral to the skill set of any data scientist.

While many of these skills - programming, getting and cleaning data, reproducible research, and data products - were covered very well in individual classes at different universities or in one off programs (Bryan 2015), they were not integrated into core curricula in most statistics, biostatistics, and machine learning programs. One advantage we had in distributing our content via Coursera was the ability to avoid vetting our program

Institution	Program	Inference	Modeling	Programming	Data Products	Data Cleaning	Reproducible Science	Exploratory Analysis
Stanford	MS Statistics	Introduction to Statistical Inference	Regression Models and Analysis of Variance	Programming Methodology	NA	NA	NA	NA
CMU	MS Statistical Practice	Advanced Methods for Data Analysis	Applied Linear Models	Statistical Computing	Statistical Practice	NA	NA	NA
NYU	MS Applied Statistics	Applied Statistical Modeling and Inference	Applied Statistical Modeling and Inference	Statistical Computing	NA	NA	NA	NA
Columbia	MA Statistics	Multivariate Statistical Inference	Regression and Multi-Level Models	Statistical Computing and Intro to Data Science	NA	NA	NA	Topics in Modern Statistics: Statistical Graphics
Harvard	AM Statistics	Statistical Inference	Linear and Generalized Linear Models	Statistical Computing	NA	NA	NA	NA
Illinois	MS Statistics	Statistical Analysis	Applied Regression and Design	Statistical Computing	NA	NA	NA	NA
Georgia Tech	MS Statistics	Math Statistics I	Regression Analysis	Computational Statistics	NA	NA	NA	NA
Indiana	MS Applied Statistics	Introduction to Statistical Theory	Applied Linear Models	Statistical Computing	NA	NA	Managing Statistical Research	Exploratory Data Analysis
Johns Hopkins	Data Science Specialization	Statistical Inference	Linear Models	R Programming	Developing Data Products	Getting and Cleaning Data	Reproducible Research	Exploratory Data Analysis
UBC	Master of Data Science	Statistical Inference and Computation I	Regression I	Programming for Data Science	Capstone Project	Data Wrangling	Data Science Workflows	Data Visualization I

Figure 2: Comparison of Traditional Statistics Programs to the Data Science Specialization. The role of statistical practices is integral to data science, however the scope of data science programs extends beyond many of the courses offerings by traditional statistics programs. We include the Data Science Specialization for comparison in addition to the Master of Data Science program from the University of British Columbia.

through a curriculum committee, which allowed us to quickly insert content we believed was important but was not considered "core" to any specific discipline.

5 Innovations in Structure

Applying the lessons learned from their first three courses proved prescient. Several choices in the structure of the program have contributed to the popularity of the Johns Hopkins Data Science Specialization.

- **Timing** - the courses were created on an expedited time scale so the JHU DSS was one of the only programs available at the time of its release.
- **Cost** - we insisted the course certificate be priced at $50 for a total program cost of $500 including the capstone course.
- **Frequency** - before Coursera's courses could be taken *on demand*, we decided to run every course in the program every month. This was in contrast to previous MOOCs that ran once or at irregular intervals modeled after university semesters.
- **Length** - we designed their courses to be one month long with a 2 week project involved in each course. This fit into the schedule of career changers and people doing the courses part time.
- **Integration** - the sequence was not cobbled together from existing courses, but designed to follow the data science pipeline from beginning to end.
- **Portfolios** - each class had a project that was pushed to GitHub and peer graded, resulting in a portfolio of work for each student completer.

Several of these innovations have now been adopted as standard practice at Coursera and other MOOC platforms (Coursera 2015). Originally Coursera courses were offered simultaneously with university courses on a semester schedule, or they were offered on an irregular schedule. The Data Science Specialization was the first program on Coursera where courses were specifically designed to be completed in four weeks, and every course was offered every month, unlike a traditional semester system. This system of offering every course every month has become known as offering courses *on demand*, meaning that students could enroll in a course at any time. If a student can not complete a course because

they enrolled in the middle of the month and do not have enough to finish, they are free to finish the course the next month without any penalty. Coursera's new platform and course format makes these high-frequency sessions the default across its offerings, based on the design principles of the Johns Hopkins Data Science Specialization.

At the time, the *on demand* format was strongly discouraged because of the fear of increased labor required to migrate content and set up nine new course sessions on a monthly basis. However after a trial period it became clear that running courses frequently was greatly appreciated by students and resulted in no drops in quality of student experience and enrollment. Another lesson was that making content more atomic across the board was preferable in this format. In other words, whether for readings, lectures, modules or classes, splitting was always better than lumping. Atomizing content created flexibility in terms of organizing the course, so that if we felt that the order of course modules should be changed partway through developing a course, we could easily rearrange the content.

6 Innovations in Community

One of the most inspiring parts of this work has been the opportunity to witness the development of the community that has grown up around the specialization. We began with free-for-all discussion forums that were specific to each course session where students could discuss topics in class and seek help and feedback. Quickly these conversations spilled over into other online venues like blogs and social media. Thanks to Coursera's efforts to harness this dynamic community's eagerness to both give back and deepen their own knowledge by helping others, we have had the pleasure of recruiting and relying on some truly incredible Community Mentors (originally called Community TAs) who volunteer their time to moderate the forums and guide new learners.

Like many instructors on the Coursera platform we still have significant in-person teaching and research responsibilities. Therefore the Community Mentors constitute the front lines of directly helping learners and answering their questions. Since learners are participating in the course around the world, it is not practical for us to hold office hours. Instead, Community Mentors and the Coursera staff closely monitor the forums in each course, where they reply to forum posts by answering questions for students and directing

them towards resources that have been helpful for students in previous iterations of the course. In this way, the Community Mentors provide a social memory of issues that have faced students in the past. In the event that a Community Mentor cannot provide an answer to a student's question, they can elevate that question to our attention so that we can address it ourselves as instructors. To reach more students and to provide them with more "face time," we have experimented with answering some of the more complex questions we have received in YouTube videos and in blog posts.

We also learned that many of these mentors and other learners were starting to produce help documents, lecture notes, study guides, and other materials to supplement the specialization and assist new learners. Some people worried that this somehow constituted cheating, but we saw it as a great way for the community to help guide one another and fill in some of the gaps that we missed when developing the courses. In order to encourage these activities, we created the Data Science Specialization Community Site on GitHub (*DataScienceSpecialization* 2016) and asked learners to contribute their work so that others could easily find and use the materials.

7 The future of data science education

Since the development of our program there has been an explosion of data science programs around the country (Tate 2017). The ubiquity of data and the training gap in students and employees who can analyze this data means that data science education is likely to grow over the short term. But given the similarity between these programs we believe that data science education is becoming commoditized. The real value is no longer in generation of content - which is now abundant and available for free on the internet. In the future value will be driven by interactive education platforms such as swirl. It is also likely to be driven by the associated value of being around experts in the field who can help to improve the data literacy of students when they run into problems produced by rapidly changing open source software.

Most importantly the large majority of traditional data science programs target one major audience - technically minded undergraduate or early graduate students. But there is a major opportunity to develop data literacy beyond these groups - both extending

earlier in the curriculum and reaching across different disciplines. We imagine that much of the growth in data science education will focus on these populations in the coming years. Regardless of the platforms, content, or outcomes, the democratization of data demands a democratization of data science education that started with the Johns Hopkins Data Science Specialization.

References

Bryan, J. (2015), 'Stat 545', `http://stat545.com/`. Accessed: 2017-8-22.

Carchedi, N., Bauer, B., Grdina, G. & Kross, S. (2014), 'Swirl: Learn r, in R'.

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J. & Others (2015), 'shiny: Web application framework for r, 2015', *URL http://CRAN. R-project. org/package= shiny. R package version 0. 11* .

Clow, D. (2013), MOOCs and the funnel of participation, *in* 'Proceedings of the Third International Conference on Learning Analytics and Knowledge', LAK '13, ACM, New York, NY, USA, pp. 185–189.

Coursera (2015), 'Coursera update: Striking a balance with start dates and deadlines', `https://blog.coursera.org/coursera-update-striking-a-balance-with-start/`. Accessed: 2017-8-22.

Coursera (2016*a*), `https://www.coursera.org/learn/exploratory-data-analysis`. Accessed: 2016-9-2.

Coursera (2016*b*), `https://www.coursera.org/learn/data-scientists-tools`. Accessed: 2016-9-2.

Coursera (2016*c*), `https://www.coursera.org/learn/r-programming`. Accessed: 2016-9-2.

Coursera (2016*d*), `https://www.coursera.org/learn/data-cleaning`. Accessed: 2016-9-2.

Coursera (2016*e*), https://www.coursera.org/learn/reproducible-research. Accessed: 2016-9-2.

Coursera (2016*f*), https://www.coursera.org/learn/statistical-inference. Accessed: 2016-9-2.

Coursera (2016*g*), https://www.coursera.org/learn/regression-models. Accessed: 2016-9-2.

Coursera (2016*h*), https://www.coursera.org/learn/practical-machine-learning. Accessed: 2016-9-2.

Coursera (2016*i*), https://www.coursera.org/learn/data-products. Accessed: 2016-9-2.

DataScienceSpecialization (2016), https://github.com/datasciencespecialization. Accessed: 2016-9-2.

Dillahunt, T., Chen, B. & Teasley, S. (2014), Model thinking: Demographics and performance of mooc students unable to afford a formal education, *in* 'Proceedings of the First ACM Conference on Learning @ Scale Conference', L@S '14, ACM, New York, NY, USA, pp. 145–146.

edX (2016), https://www.edx.org/. Accessed: 2016-9-2.

Eggleston, L. (2017), '2017 coding bootcamp market size study', https://www.coursereport.com/reports/2017-coding-bootcamp-market-size-research. Accessed: 2017-8-21.

Gooding, I., Klaas, B., Yager, J. & Kanchanaraksa, S. (2013), 'Massive open online courses in public health', *Frontiers in Public Health* **1**, 59.
URL: *https://www.frontiersin.org/article/10.3389/fpubh.2013.00059*

Guzdial, M. (2017), ''generation CS' drives growth in enrollments', *Commun. ACM* **60**(7), 10–11.

Higher Logic, L. (2016), 'Largest U.S. master's programs in statistics and biostatistics - amstat', http://community.amstat.org/blogs/steve-pierson/2014/02/09/largest-graduate-programs-in-statistics. Accessed: 2016-9-2.

Jordan, K. (2014), 'Initial trends in enrolment and completion of massive open online courses', *The International Review of Research in Open and Distributed Learning* **15**(1).

King, J. & Magoulas, R. (2017), '2016 data science salary survey', http://www.oreilly.com/data/free/files/2016-data-science-salary-survey.pdf. Accessed: 2017-8-22.

Leanpub (2016), https://leanpub.com/. Accessed: 2016-9-2.

Leek, J. (2016), 'The elements of data analytic style', https://leanpub.com/datastyle. Accessed: 2016-9-2.

Leek, J. (2017), 'jtleek/dataanalysis', https://github.com/jtleek/dataanalysis. Accessed: 2017-8-21.

Lewin, T. (2012), 'MOOCs, large courses open to all, topple campus walls', *The New York Times* .

Nicholls, N. (2009), 'The investigation into the rising cost of textbooks'.

O'Connor, G. (2014), 'Moore's law gives way to bezos's law', *GigaOm, April* **19**.

OpenCourseWare, M. (2016), 'MIT OpenCourseWare — free online course materials', http://ocw.mit.edu/index.htm. Accessed: 2016-9-2.

Peng, R. (2016), 'R programming for data science', https://leanpub.com/rprogramming. Accessed: 2016-9-2.

shinyapps.io (2016), http://www.shinyapps.io/. Accessed: 2016-9-2.

Stodden, V., Leisch, F. & Peng, R. D. (2014), *Implementing Reproducible Research*, CRC Press.

Tate, E. (2017), 'Data analytics programs take off', https://www.insidehighered.com/digital-learning/article/2017/03/15/data-analytics-programs-taking-colleges. Accessed: 2017-8-22.

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K. & Pawlik, A. (2015), 'Data carpentry: Workshops to increase data literacy for researchers', *International Journal of Digital Curation* **10**(1), 135–143.

Udacity (2016), https://www.udacity.com/. Accessed: 2016-9-2.

Yelp (2018), 'Yelp dataset challenge', https://www.yelp.com/dataset/challenge. Accessed: 2018-11-01.

Zheng, S., Rosson, M. B., Shih, P. C. & Carroll, J. M. (2015), Understanding student motivation, behaviors and perceptions in MOOCs, *in* 'Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing', CSCW '15, ACM, New York, NY, USA, pp. 1882–1895.