Article from

Predictive Analytics and Futurism

December 2015
Issue 12

Johns Hopkins Data Science Specialization courses: A review

By Shea Parkes

Data science is a hot buzzword in many industries today, but its definition can be nebulous. Some definitions of a data scientist include:

- A person who is better at statistics than any software engineer and better at software engineering than any statistician;
- A person with an equal blend of computer science, statistics, and domain knowledge; and
- An applied statistician who is rebranding.

Even if nobody agrees on the specifics, the concept of data science can still facilitate a thought exercise in what blend of skills is most useful for data analytics. Actuaries are solidly grounded in statistics and domain knowledge as part of the examination and continuing education process. However, actuaries are traditionally weaker in regards to computer science skills than might be optimal to grow our presence in modern data analytics. This includes some blended skills, such as machine learning and predictive modeling, which require both applied statistics and computer science skills.

Computer science skills can bring a lot of value to a classically trained actuary. These skills can help:

- Make answers more transparent, reproducible, and reusable;
- Answer bigger questions than before;
- Answer smaller questions faster and more efficiently; and
- Present answers more visually and interactively.

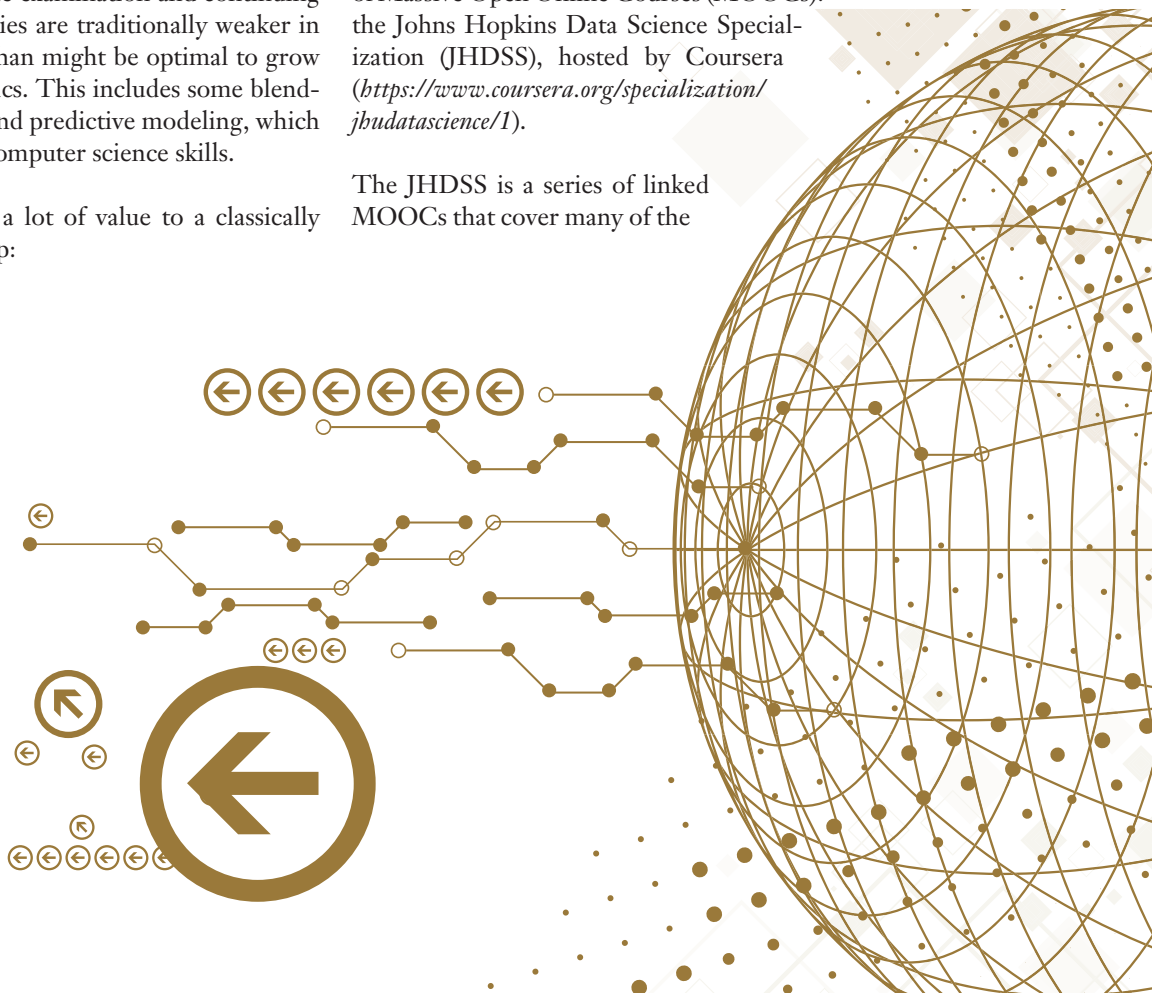
For actuaries interested in rebalancing their skill portfolios toward the data science blend, there are many resources available. I personally was drawn to the data science balance and explored it along a rough path that included:

- Repeatedly attempting (and failing) at cutthroat online predictive modeling competitions (such as those hosted on *http://www.Kaggle.com*) with my coworkers. Every failure was an excellent learning opportunity and after a couple of years we could consistently place in the top 10 as long as we exerted enough effort for a given contest. (Consistently winning was an echelon we never reached.); and

- Forced self-learning while helping carve a new product group out of a large consulting staff. We consumed countless books and other resources on best practices for development of prioritization techniques, software lifecycle management, and gritty details of source control tools and strategies. By the end of the year we reached workable solutions based on ideas such as Scrum, Kanban, Git, and Continuous Integration.

I think this rocky road was actually an excellent way to learn more about machine learning, computer science, and software engineering, but I don't believe it's available or appropriate for everyone. Just about the same time we felt we had found a paved road, a new opportunity was presented in the form of a series of Massive Open Online Courses (MOOCs): the Johns Hopkins Data Science Specialization (JHDSS), hosted by Coursera (*https://www.coursera.org/specialization/jhudatascience/1*).

The JHDSS is a series of linked MOOCs that cover many of the



traditional data science topics in which actuaries might be weakest. The JHDSS courses are not the only MOOCs of their kind, nor are they necessarily the best, but they appeared polished enough to make me interested in trying them. The JHDSS creators are prolific and respected contributors to the data science community in their own blogs and journals.

By the time I had signed up I was already proficient in most of the topics, but I still completed the courses as an external validation of my new skills and also to evaluate them as a continuing education resource for other employees at our company. I completed all of the JHDSS courses with a coworker in a little less than a year. We ultimately deemed it useful enough to make available to all of our staff alongside the actuarial exams and other credentialing opportunities.

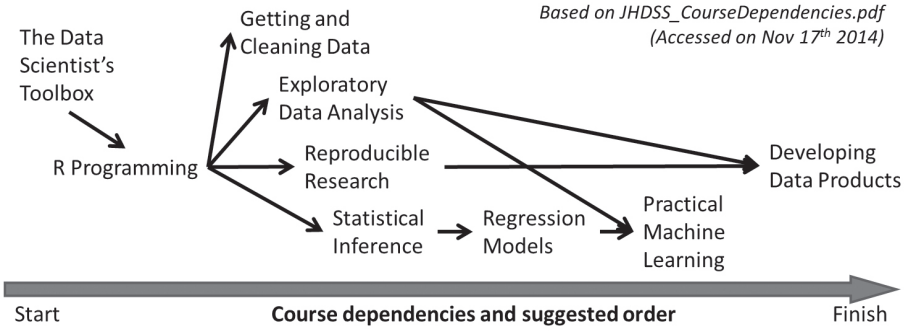
The majority of the JHDSS courses each take a month to complete and require roughly 10 hours of work per week. They include many aspects of standard MOOCs such as:

- Prerecorded lectures with accompanying notes and slides;
- Active discussion forums (an invaluable resource for any student);
- Weekly quizzes; and
- Peer-graded projects on real data.

The peer-graded projects were some of the richest learning opportunities, especially when it was your turn to grade your peers' submissions.

The modest time commitment (10 hours per week) allows for working professionals to complete the specialization somewhat comfortably. The amount of content provided is not enough to make anyone an expert, but it does equip the student with rough tools and ideas that can be sharpened and honed via application and experience.

FIGURE 1: DATA SCIENCE COMPONENT COURSES



Source: *https://www.gitbook.com/book/gdhorne/data-science-boot-camp-survival-manual/details.*

"The modest time commitment (10 hours per week) allows for working professionals to complete the specialization somewhat comfortably."

The chart in Figure 1 visualizes the names of the component courses and their suggested dependency order.

There is an optional fee to take each of the courses (well under $100 at the time of writing this). The courses can be taken free, but in that case verified "certificates" for resumes are not issued. I personally paid the fee to leverage the sunk cost fallacy and trick myself into committing more to the courses. I also thought the fees were a bargain for a working professional and I wanted to support the content creators who put a lot of effort into a good idea. The opportunity costs of your time will likely be the largest fee.

The remainder of this article presents each course's official tagline and my own brief review of the usefulness and quality of the content.

COURSE 1: THE DATA SCIENTIST'S TOOLBOX

Official tagline: "Get an overview of the data, questions, and tools that data analysts and data scientists work with."

MY REVIEW:

This is a very gentle introduction to some tools that can revolutionize the way you approach solutions. For example, I feel like I'm driving without my seatbelt now if I ever complete work without source control. Source control is a tool that tracks code changes at a very detailed level and greatly facilitates collaboration and quality. The source control tool covered in this course was the very popular GitHub (*http://www.github.com*). Learning a tool like GitHub can be very intimidating, though, and this might serve as a needed boost to get over the initial hump. Some simple text markup language, such as Markdown, introduced in this course, is a necessary companion because classic document solutions like Microsoft Word do not play nicely with most source control tools.

Still, with no prior background or appreciation, this overly simple introduction could lose students' interest because no practical examples are explored. Some of the next few courses do force you to use GitHub and Markdown so you

can better internalize what you are exposed to in this course. The later courses just assume you will use source control on your own (and you definitely should).

COURSE 2: R PROGRAMMING

Official tagline: "Learn how to program in R and how to use R for effective data analysis."

MY REVIEW:

The greater difficulty of this course is in sharp contrast to the prior course. Many students might get disheartened if they don't have much prior programming experience. Learning programming is hard, and learning R is harder. However, I agree that R is an excellent domain-specific language (DSL) for data analytics and learning it is worth the effort. I considered myself proficient in R prior to this course, but I learned a few additional aspects of R as a programming language (such as the full nuances of closures). All of the remaining courses depend greatly on this course; you need to be at least somewhat committed to learning R if you are going to complete the JHDSS (and I consider that a good thing).

Because my coworker and I already knew R prior to this course, it is very hard for me to judge how useful this course would be as a beginner introduction to R. It seemed to strike an appropriate balance of explanation, difficulty, and application, but I had a biased view from my place higher up on the R learning curve.

COURSE 3: GETTING AND CLEANING DATA

Official tagline: "Learn how to gather, clean, and manage data from a variety of sources."

MY REVIEW:

This course sustains the high difficulty level of Course 2: R Programming and it continues to teach invaluable data science skills: how to acquire and deal with real data. Coursework intentionally forces you into reading documentation for specific R packages (third-party extensions to R that each add specific functionality) and consulting with Google and Stack Overflow (very good skills to practice).

This course was refreshing compared with the classic style of academic courses that just provide students with already scrubbed data and ask them to perform rote statistical analyses. However, some of the hardest parts of this course were working with data source types that actuaries would be unlikely to dig through. R is great for integrating with traditional data sources such as databases, but this course pushed into some more unusual areas like web services.

COURSE 4: EXPLORATORY DATA ANALYSIS

Official tagline: "Learn the essential exploratory techniques for summarizing data."

MY REVIEW:

I believe exploratory data analysis (EDA) is a chronically underemphasized topic in all forms of education. I have read the classical texts on the subject by John Tukey and William Cleveland and consider them required reading for any aspiring data scientist. We hand out copies of *Show Me the Numbers* by Stephen Few to all new employees at my office and periodically read through it in book clubs. Basically, I loved this course as soon as I read the title. I breezed through the coursework, and I believe it was easier (or at least more innately enjoyable) than the prior courses. They give the subject a respectable treatment and I think any student would benefit from it. I wish they had spent more time with the more advanced tools such as the ggplot2 package for R, but I respect focusing on the theory over the fanciest of tools.

COURSE 5: REPRODUCIBLE RESEARCH

Official tagline: "Learn the concepts and tools behind reporting modern data analyses in a reproducible manner."

MY REVIEW:
I have mixed feelings about this course. I think the concepts of reproducible research are very important and deserve a course of their own. I think the foundational tool-chain they chose (*http://yihui.name/knitr/*) was a solid choice. But I think they went too far when they tried to integrate automatic Web publishing with an unreliable cloud service (*https://rpubs.com/*); stability might improve in the future, but during my course the forums were full of students who had difficulties with the cloud service. I understand why they wanted to go there (theoretical ease of accessibility and "wow" factor), but I believe they should have spent more time covering the advanced capabilities of the fundamental tools instead of trying to layer them into web services.

COURSE 6: STATISTICAL INFERENCE

Official tagline: "Learn how to draw conclusions about populations or scientific truths from data."

MY REVIEW:
I often identify myself as an applied statistician these days (more often than I call myself a data scientist; less often than I call myself an actuary). I find statistics a fascinating topic, but I also find the average teaching of statistics to be rote and formulaic, and this course did not elevate itself above that. I think frequentist statistics has its place, but this course, like many, put it front and center and barely left room to discuss Bayesian viewpoints. I think data scientists should have a firm understanding of statistics, but I believe this course was inadequate to provide that on its own. However, I don't think I could have provided a better grounding in the same amount of time. Statistics is just too big and broad of a subject to dig into as deeply as a data scientist would need to in a single month.

> "The capstone project class will allow students to create a usable/public data product that can be used to show your skills to potential employers."

CHAPTER 7: REGRESSION MODELS

Official tagline: "Learn how to use regression models, the most important statistical analysis tool in the data scientist's toolkit."

MY REVIEW:
Ordinary least squares regression is so far from ordinary. George Box once said "in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, [a scientist] can often derive results which match, to a useful approximation, those found in the real world." Regression theory and models are a great jumping point from applied statistics to predictive modeling and machine learning. I believe this course did a pretty good job of balancing depth of theory while also covering important extensions such as generalized linear modeling. I think aspiring data scientists should spend even more time on this subject to keep a balanced knowledge portfolio, but the next topic (machine learning/predictive modeling) can be quite alluring. I think they could have focused a bit more in this course on relating classical statistical terminology to the corresponding machine learning terminology used later. Making those deep connections really helps understand both topics better.

COURSE 8: PRACTICAL MACHINE LEARNING

Official tagline: "Learn the basic components of building and applying prediction functions with an emphasis on practical applications."

MY REVIEW:
This is a very exciting topic to a large portion of the students that participated, and I think most of them left satisfied. Covering all of the top-tier algorithms is not attempted, nor should be. An appropriately large amount of time is spent focusing on the bias-variance trade-off and model tuning tools such as cross-validation. The exercises force students to build models, but I do think a bit more room could have been allowed for creativity. I introduced some flair into my solutions, but it was not required. I do think the course dependency chart in Figure 1 above is very important, though. This class is a culmination of all that came before and it would be much less without the journey. The courses that come after this are still good ideas, but they take things a subtly different direction (productization).

COURSE 9: DEVELOPING DATA PRODUCTS

Official tagline: "Learn the basics of creating data products using Shiny, R packages, and interactive graphics."

MY REVIEW:
I believe productization is a natural stepping stone in the data science curriculum, but it is a very complicated subject. This class covers a bit of the theory and then jumps into a specific tool (Shiny) used to make responsive web-based applications. Shiny (and its corresponding cloud hosting services) is a promising but young tool that is not without its rough edges. Still, it has the right level of accessibility and "wow" factor; you can learn it and feel proud of your results within the duration of this course. I personally think there should have been more focus on "hardening" advanced modeling code to work stably in a production environment, but that's a much less exciting subject.

CAPSTONE PROJECT: DATA SCIENCE CAPSTONE

Official tagline: "The capstone project class will allow students to create a usable/public data product that can be used to show your skills to potential employers. Projects will be drawn from real-world problems and will be conducted with industry, government, and academic partners."

MY REVIEW:

I believe this was a strong finish to the JHDSS; it was a full two-month project focusing on a single problem. They intentionally introduced an important subject not covered in prior courses (text mining in my sitting) to force you to practice learning something fundamentally new as part of a larger engagement (a common occurrence in the real world). The problem was interesting and the amount of guidance was just right. The ancillary tasks (e.g., quizzes) were surprisingly weak, but that didn't distract from the overall strength of the capstone project. It felt very much like my day job (the fun parts of it), and I think that's the best endorsement I can give it. The difficulty level was quite high, but most participants rose to the challenge.

CONCLUSION

If you, or someone you know, wants to learn more about the data science viewpoint, the JHDSS is a useful means to do so. The largest hurdle might be that participants would need to be committed to learning R, but I consider that a positive aspect of the specialization. Trying to cover these topics without diving deep into an appropriate computer language would have failed to give them the treatment they deserve. The JHDSS is not perfect, but I believe the general content is a really good mix, especially to complement classical actuarial training. ■

Shea Parkes, FSA, MAAA, is an actuary at Milliman in Indianapolis. He can be reached at *shea.parkes@milliman.com*