Educational Data Scientists: A Scarce Breed

Simon Buckingham Shum
Knowledge Media Institute
The Open University
Walton Hall
Milton Keynes, MK7 6AA, UK
s.buckingham.shum@gmail.com

Ryan S.J.d. Baker
Teachers College
Columbia University
525 W. 120th St., Box 118
New York, NY 10027, USA
baker2@exchange.tc.columbia.edu

John T. Behrens
Center for Digital Data, Analytics, & Adaptive Learning
Pearson
400 Center Ridge Drive
Austin, TX 78753
John.Behrens@pearson.com

Martin Hawksey
Jisc CETIS
c/o University of Strathclyde
16 Richmond Street, Glasgow
G1 1XQ, UK
martin.hawksey@strath.ac.uk

Naomi Jeffery
The Open University in Scotland
10 Drumsheugh Gardens,
Edinburgh
EH3 7QJ, UK
naomi.jeffery@open.ac.uk

Roy Pea
H-STAR Institute
Stanford University
450 Serra Mall, Building 160,
Stanford CA 94305-2055, USA
roypea@stanford.edu

ABSTRACT

The *Educational Data Scientist* is currently a poorly understood, rarely sighted breed. Reports vary: some are known to be largely nocturnal, solitary creatures, while others have been reported to display highly social behaviour in broad daylight. What are their primary habits? How do they see the world? What ecological niches do they occupy now, and will predicted seismic shifts transform the landscape in their favour? What survival skills do they need when running into other breeds? Will their numbers grow, and how might they evolve? In this panel, the conference will hear and debate not only broad perspectives on the terrain, but will have been exposed to some real life specimens, and caught glimpses of the future ecosystem.

Categories and Subject Descriptors

J.1 [**Administrative Data Processing**] *Education*; K.3.1 [**Computer Uses in Education**]

General Terms

Measurement, Documentation, Human Factors, Theory

Keywords

Learning Analytics; Educational Data Mining; Data Science; Educational Data Scientist

1. INTRODUCTION

While the learning analytics and educational data mining research communities are tackling the question of what data can tell us about learners, relatively little attention has been paid, to date, to the specific mindset, skillset and career trajectory of the people who wield these tools. Within business and government, Data Scientists are heralded as the new, scarce breed. A widely cited report from McKinsey Global Institute identified a widening talent gap in the workforce [1], while Harvard Business Review [2] declared Data Scientist to be "sexiest job in the 21st century"!

The U.S. Department of Education contextualized this talent vacuum within the educational ecosystem as follows:

> *Interdisciplinary teams of experts in educational data mining, learning analytics, and visual analytics should collaborate to design and implement research and evidence projects. Higher education institutions should create new interdisciplinary graduate programs to develop data scientists who embody these same areas of expertise.* [3]

Clearly, the LAK community has every interest in shaping this mindset and skillset: as MOOCs, conventional courses, and the Learning Analytics Summer Institutes[1] begin to bear fruit, graduates need careers, and research teams will be recruiting back from this pool of people as they build industrial track records.

So, over to our panellists, who bring a wealth of academic and hands-on experience to help move this debate forward.

2. WE NEED MORE EDUCATION DATA SCIENTISTS

I should probably track the number of phone calls and emails I get each month from companies that want to hire an education data scientist (sometimes called an educational data miner, or learning analytics engineer). I get a lot of emails. I try to figure out what the company needs (sometimes they don't know), and recommend a colleague, or post-doc, or student. I don't have enough recommendations to go around. I've had companies ask me if they

[1] Learning Analytics Summer Institutes: http://www.solaresearch.org/events/lasi

can hire someone without experience, and then have me train that person.

We need more graduate programs for education data scientists. A couple of graduate programs are already out there that focus on this: Carnegie Mellon University, with their new M.S. in Learning Science and Engineering; Worcester Polytechnic Institute, with their Ph.D. and M.S. in Learning Sciences and Technologies; and there are a lot of academic labs that produce graduates trained in this area even without an official program.

One of the tricky things for training a cadre of education data scientists – in my opinion – is that "data is not just data". Different kinds of data have different characteristics and affordances; different grain-sizes of analysis; different features that are key to engineer; different algorithms that just "tend to work well". Less-specialized data science and data mining graduate programs can be very good, and if they have good education data scientists teaching in them, they'll naturally support the development of this cadre. But where there's not scale, we need resources to support teaching. No one researcher is going to know about all the types of data an education data scientist needs to know about (I'm very much not a text miner, myself – just to give one example). It's way too early for standardized curricula, but sharing resources, creating textbooks, and talking to each other will definitely help. Summer institutes like the PSLC Summer School and the Learning Analytics Summer Institute will help. MOOCs (and hopefully multiple competing MOOCs) like the LAK13 MOOC on Learning Analytics will help.

To the future: an education data scientist (or several) in every company, well-trained, knowing what has already been tried and is awesome, and ready to try new and more awesome things.

Bio: Ryan Shaun Joazeiro de Baker is the Julius and Rosa Sachs Distinguished Lecturer at Teachers College, Columbia University. He earned his Ph.D. in Human-Computer Interaction from Carnegie Mellon University's School of Computer Science. Dr. Baker was previously Assistant Professor of Psychology and the Learning Sciences at Worcester Polytechnic Institute. He previously served as the first Technical Director of the Pittsburgh Science of Learning Center DataShop, the largest public repository for data on the interaction between learners and educational software. He researches student engagement, affect, meta-cognitive behavior, and robust learning in educational software, including in intelligent tutors, simulations, microworlds, and serious games, and his lab developed the first automated detectors of gaming the system, off-task behavior, and preparation for future learning in these contexts. He is currently serving as the founding President of the International Educational Data Mining Society, and as Associate Editor of the Journal of Educational Data Mining. http://www.columbia.edu/~rsb2162

3. HOW DO YOU BECOME AN EDUCATIONAL DATA SCIENTIST?

In 1999 I was finishing my undergraduate in Structural Engineering. Whilst there were moments of interest such as using finite element analysis to produce strength visualizations from structural simulations, the subject largely left me cold. By 2001 I returned to academia to do postgrad in multimedia and interactive systems. A notable incident in this period was when the programme leader distributed the transcript for the previous semester's grades in a Microsoft Excel spreadsheet. Data were anonymized by using student matriculation numbers. Few students appeared to pay attention to his data, I did not. Not only was I analyzing where I ranked in the class, I was predicting my final grade based on previous performance and all by using formula a formatting available in Excel. My curiosity also allowed me to decode the student matriculation numbers. By formatting the data and pasting it into our student webmail client the 'check names' feature conveniently reconciled matriculation number to student name. Now not only could I see my own performance, I could also now look my competition in the eye. I graduated with distinction and with a university medal.

This example underlines a number of the qualities I believe are important in becoming an educational data scientist. As noted by Stephen Brobst in a talk at Teradata Universe Conference in 2012:

> *A data scientist … is someone who: wants to know what the question should be; embodies a combination of curiosity, data gathering skills, statistical and modelling expertise and strong communication skills. … The working environment for a data scientist should allow them to self-provision data, rather than having to rely on what is formally supported in the organisation, to enable them to be inquisitive and creative.* [4]

We all have varying degrees of curiosity, inquisition and creativity. The key in my evolution towards educational data science has been the professional and personal opportunity (aka pushing pixels at 1am) to explore ideas and concepts, doing this allowing me to refine my skills in data collection and processing, evolve recipes to allow others to taste the sweetness of data science, and develop an understanding of my craft.

Bio: Martin Hawksey is an advisor at the Centre for Educational Technology and Interoperability Standards (CETIS), a national advisory and innovation centre funded by JISC, supporting the UK Higher and Post-16 Education sectors on educational technology and standards. He is a master of Google Sheets and a number of his templates are used internationally within education and business for collecting and analyzing data. His most notable example is the Twitter Archiving Google Spreadsheet (TAGS) which archives Twitter search terms for analysis and visualization. His current research is primarily around aggregating and analyzing secondary data sources from Massive Open Online Courses (MOOCs). He is a supporter of open educational practices and his work can be followed on his blog, MASHe, at http://mashe.hawksey.info

4. REMEMBER OUR PURPOSE

How do you define an educational data scientist? The answer varies considerably by post and by practitioner. The role is also rapidly evolving as new software developments chase the demands of the research community. For me, data science is a combination of statistics, computer science and information design, but as my role connects with so many communities and professions, good communication and collaboration skills are essential.

Many definitions specify working with 'big data' but I for one don't work exclusively with big data – only when it's necessary and appropriate. One important basis for working with big data is pattern recognition in noisy data, which is a very visual – and, for now, primarily human – skill. As a statistician I am determined that in manipulating and visualising data we remain honest and realistic, hence shifting focus to smaller, more definite, data sources can be both refreshing and an important re-grounding.

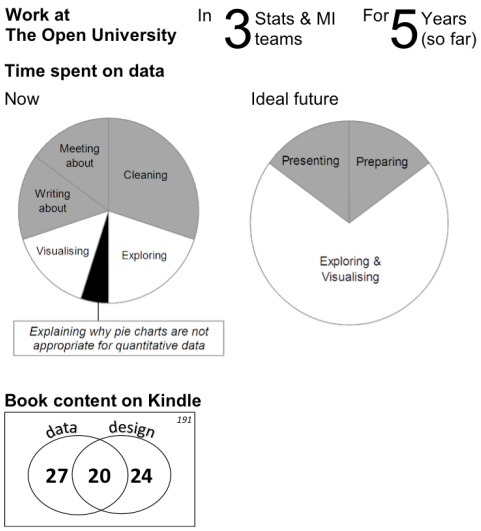
The educational data scientist must dive from Learning Analytics (LA) into Educational Data Mining (EDM) and resurface: exploring the real world, proposing meaningful measures, modelling the data, visualising the output, sharing the technique

and automating the process. Sharing new ideas and techniques takes enormous confidence and as many education data scientists are – like myself – not academics, promoting our work as individuals does not always come naturally or easily. Working with academic researchers who are more comfortable on the front line can be a relief… as long as credit is shared.

Keeping track of new LA publications (both formal and informal) is extremely important because an understanding of how to measure learning must come before any exploratory analysis of educational data. New pedagogies and technologies can render traditional analysis methods and conventional wisdom on learner behaviour patterns obsolete, hence the educational data scientist should always be scanning for new developments and considering how they may impact on (and ideally improve) the learner experience.

In the near future I think one focus of educational data science will be on developing animated, multi-dimensional techniques for mining and visualising data. Continuing issues will be data privacy and ethics, the danger of viewing learners as merely the incidental constructs of their own data identity, and the perfidious and ultimately ruinous double-standard that 'your' data should be shared but 'my' data must be protected. An exciting new frontier will be developing progressively more meaningful measures of learning as we engage with new data and are able to expand our quantitative conception of success beyond the course, and into individual learner aims for their career and life as a whole.

Bio: Naomi Jeffery's academic background is in statistics and computer science. She's a designer at heart and is working towards another degree in design and innovation. She has more ideas than time and more crafts than space. http://statisticiana.wordpress.com



5. BE A DATA ANALYST AND A PHILOSOPHER/ SCIENTIST

The science and art of data analysis concerns the process of learning from quantitative records of experience. By its very nature it exists in relation to people. Tukey & Wilk [4].

Though a famous and successful statistician, Tukey wanted to create a field that dealt with all data, even when it came in such poor shape that it was not amenable to statistical analysis. He called it "data analysis" and created the field called "Exploratory Data Analysis".

My undergraduate degree was in Psychology and Philosophy. I thought if I knew the logic of how we know things (epistemology) and understood the human lens through which all perception and thought occurs (psychology) I would have the fundamental layers of knowing from which to acquire more knowledge. After serving as a social worker and studying special education, I sought my Ph.D in Educational Psychology with a cognate called "Measurement, Statistics & Methodological Studies". I would approach it as applied epistemology: *How do we learn from data?* When I discovered Tukey's writings I knew I had found the right place. I conducted psychological studies on perception of statistical graphics and wrote about the logical foundations of data analysis. When I wrote such a chapter called "Data and Data Analysis" [6] people told me it was a silly title – data wasn't a subject, it's only a piece of the background to other sciences.

Philosophy is concerned with understanding meaning and the application of logic. The philosopher asks *What do we mean by 'data'? What do we mean by 'analysis'? If data are symbols that point to elements in the world, what kind of logic do we need to understand that linkage?* Like very good scientists, philosophers question the obvious. Such questioning may not be essential for what you do today, but it may open the door to do new ways of thinking you never imagined.

The successful learning analyst will avoid two common errors: Failure to understand the context and failure to become intimately familiar with the data. The first error is caused by lack of contextual knowledge. Studying the learning sciences, education, and related disciplines will help. The second is error is caused by a substitution of complex statistical or computational models for detailed mental models. We only build computational models or display to help our mental models. Question the assumptions of your work deeply. It is important that analysts understand their work is about "revelation" or "unveiling" the reality of the world. It is a special (at times prophetic) role in society and should be taken very seriously.

Do not think of data science as a set of techniques but as a collection of viewpoints (epistemic positions) and habits of mind. To undertake good visualization we need to know the techniques of data display, but also the psychology of perception, the anthropology of semiotics, the mathematics of fluctuation and the philosophy and art of aesthetic engagement. We will always need good technical analysts, but we need them to be (or at least understand) scientists, philosophers & artists as well.

Bio: John Behrens is Vice President and leader of the Center for Digital Data, Analytics, and Adaptive Learning at Pearson. He brings cognitive, statistical, computational and philosophical lenses to designing, deploying and analyzing data-intensive learning systems including instructional, assessment, and game environments. He continues to write and speak about data and data analysis and the implications of the digital revolution for methodology. He is in love with data and the worlds they reflect. Previously, John led product research and development in the Cisco Networking Academies. Serving 10,000 schools in 160 countries, John oversaw the first large scale use of Evidence Centered Design to drive the integrated use of curriculum, simulations, and gaming for on-line instruction and assessment. His work in Cisco certifications made simulation based assessment standard practice in the IT certification exam industry. Prior to Cisco, John was a tenured associate professor of Psychology in Education at Arizona State University. http://researchnetwork.pearson.com/digital-data-analytics-and-adaptive-learning

6. WHAT ARE KEY COMPETENCY DOMAINS FOR THE EDUCATION DATA SCIENTIST?

When I was responsible for designing and launching the first Learning Sciences doctoral program in the world in 1992 at Northwestern University with my colleagues in education, psychology, and computer science, we put considerable thought into how to frame the areas of scholarship and inquiry that our students needed to master to make the advances we felt the opportunities warranted.

What we developed at the time was a tripartite conceptualization of the Learning Sciences: *'Cognition'* – constructing scientific models of the structures and processes of learning and teaching by which organized knowledge, skills and understanding are acquired; *'Environments'* [now *'Context'*] – examining the social, organizational and cultural dynamics of learning and teaching situations, including schools and out-of-school settings, such as homes, museums and corporations; and *'Architectures'* [now *'Design'*] – building environments for learning and teaching, incorporating multimedia, artificial intelligence, computer networks and innovative curriculum and classroom activity structures. Each of our students tended to focus on going deep in one of these to make their contributions, but we insisted that they learn enough about the others to team productively. We also recognized that in principle and in practice, these distinctions between Cognition, Context and Design were as much about figure and ground than about hard and fast: embodied minds interact and learn in contexts involving designed socio-technical tools. And twenty years on, these distinctions are becoming even more indissociable in an increasingly hyperconnected world.

These three domains of competencies were prescient framings. Northwestern University revisited this scheme twenty years later and decided after critical reflection to keep to it, since it had served its graduates, faculty, and building the field of Learning Sciences well. At Stanford, we launched our Learning Sciences and Technology Design program in 2001. Today there are 40 or so postgraduate degree programs around the world in the Learning Sciences, and the 2nd edition of the *Cambridge Handbook of the Learning Sciences*, first published in 2006, is in the works. There is a growing International Society for the Learning Sciences (ISLS.org), and two archival journals that are among the highest impact journals in the education field: *The Journal of the Learning Sciences*, and the *International Journal of Computer-Supported Collaborative Learning*.

What is the relevancy of this history for what an education data scientist needs to know and be able to do? I argue that the education data scientist, too, needs to have an understanding of the cognitive, contextual and design aspects of the transactions that generate educational data and the interdisciplinary sciences that will contribute to an understanding of it – if he or she is to ask generative research questions that will advance the sciences and practices of education and learning.

What's new with education data science? A great deal: The foundational roles of developments in statistical computing for data analytics; the centrality of state-of-the-art interactive data visualization for exploratory data analysis; the vital roles of machine learning; more powerful and distributed computing architectures that enable sense-making and prediction with big data; deeper understanding of the links between types of research questions, education data types, and effective LA and EDM methodologies to suite the questions and data; greater need than ever for critical questioning of assumptions at every level, and for balancing learning data openness and transparency with ethical considerations. The technical and mathematical wizardry underlying these advances still needs theory and science from the Cognition, Context and Design domains. For example, the social and environmental contexts of technology-mediated learning need to be "sensed" and become part of our multimodal learning analytics agenda, as they are central to the environments which learners are experiencing when educational data is collected. In short, an education data scientist will have to be an active listener and open collaborator because no one will know and be able to do everything that will produce the highest quality work.

We should unite in a clarion call for universities, governments, philanthropists, and industry to all contribute to accelerating the field of education data science, and work collectively to attract the best minds of our generation to tackling problems and providing compelling solutions that would enable global personalized learning for all.

Bio. At Stanford University **Roy Pea** serves as David Jacks Professor of Education and the Learning Sciences (and, by courtesy, Computer Science), co-founder and Director of the H-STAR Institute (Human Sciences and Technologies Advanced Research), and founder and Director of the PhD program in Learning Sciences and Technology Design. He was on the founding board of ISLS and served as President, 2004-2005. His work in the learning sciences focuses on advancing theories, findings, tools, practices and interdisciplinary field-building for technology-enhanced learning of complex domains. He co-leads the "LIFE Center" (http://life-slc.org) whose studies seek to inform better bridging of the sciences of learning for informal and formal environments. He has been long been inspired by the Engelbartian quest for augmenting human intellect and performance by co-evolving human-computer systems, and learning analytics seems poised as a breakthrough area toward that long-term vision. Recently, he joined with Stanford's Vice Provost for Online Learning, John Mitchell, to serve as faculty co-director of an interdisciplinary student-initiated Lytics Lab, devoted to advancing learning analytics science, theory and tools, with special attention to improving MOOCs. http://www.stanford.edu/~roypea

6. REFERENCES

1. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute. http://bit.ly/McKinseyBigDataReport
2. *Data Scientist: The Sexiest Job of the 21st Century*. Harvard Business Review Magazine: http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1
3. *Expanding Evidence Approaches for Learning in a Digital World.* Office of Educ. Technology, U.S. Dept. Education: http://www.ed.gov/edblogs/technology/evidence-framework
4. Cooper, C.: *Analytics and Big Data - Reflections from the Teradata Universe Conference 2012.* Blog post (Apr. 27, 2012): http://bit.ly/CooperTeradataBlog
5. Tukey, J. W. and Wilk, M. B. (1966). Data analysis and statistics: An expository overview. *AFIPS Conf. Proc. 1966 Fall Joint Comp. Conf.* 29, pp. 695–709
6. Behrens, J. T., & Smith, M. L. (1996). Data and data analysis. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of Educational Psychology*, pp. 945-989. New York: MacMillan.