Who owns educational theory? Big data, algorithms and the expert power of education data science

Ben Williamson, Faculty of Social Science, University of Stirling, UK

ben.williamson@stir.ac.uk

[Post-print of article published in E-Learning and Digital Media at:
http://journals.sagepub.com/doi/full/10.1177/2042753017731238]

Abstract

'Education data science' is an emerging methodological field which possesses the algorithm-driven technologies required to generate insights and knowledge from educational big data. This article consists of an analysis of the Lytics Lab, Stanford University's laboratory for research and development in learning analytics, and the Center for Digital Data, Analytics and Adaptive Learning, a big data research centre of the commercial education company Pearson. These institutions are becoming methodological gatekeepers with the capacity to conduct new forms of educational research using big data and algorithmic data science methods. The central argument is that as educational data science has migrated from the academic lab to the commercial sector, ownership of the means to produce educational data analyses has become concentrated in the activities of for-profit companies. As a consequence, new theories of learning are being built-in to the tools they provide, in the shape of algorithm-driven technologies of personalization, which can be sold to schools and universities. The paper address two themes of this special issue: (1) how education is to be theorized in relation to algorithmic methods and data scientific epistemologies; and (2) how the political economy of education is shifting as knowledge production becomes concentrated in data-driven commercial organizations.

Keywords

algorithms; big data; data science; educational research; education data science; methodology

Introduction

The ownership of the technologies and algorithmic techniques required to analyse big data has emerged as a significant issue for research in e-learning and digital media. Recently Ruppert (2015) asked 'who owns big data?', noting that big data are the product of different actors and technologies involved in its generation and analysis. The actors involved in these practices and processes can in many ways be seen to 'own' big data, particularly as findings and insights are generated from big data using proprietary algorithms and analytics processes that are protected by intellectual property rights. These observations point to important issues about the ownership of the insights that come from big data as they are extracted from the everyday traces people leave as they interact with one another and transact with services digitally. For educators, this makes it essential to consider the more specific question of 'who owns educational big data?'

In this article, I analyse 'education data science' as an emerging field of methodological expertise in which ownership of educational big data technologies is being claimed by a handful of dominant actors. In particular, I examine the Center for Digital Data, Analytics and Adaptive Learning, a research centre launched by the commercial education company Pearson, and the Lytics Lab, Stanford University's research and development laboratory for learning analytics. These centres and labs have begun to act as methodological gatekeepers with the social, economic, and technical capacity to carry out new forms of research in e-learning by using large-scale datasets, big data analytics and algorithmic data science methods. These organizations have been selected as case studies because they have both generated significant visibility for educational data science albeit from very different positions, with Lytics Lab being an academic R&D initiative and Pearson a commercial edu-business. Though educational data science originates in the academic sector, it is being transferred to mainstream education through commercial vendors, such as Pearson, with the capacity to build and sell applications and products to schools and colleges. Lytics Lab and Pearson's centre also exemplify how education data science has become a trans-sector enterprise, applying its expertise and technologies across the spectrum of schools and Higher Education.

The focus for the analysis is how these organizations advocate educational data science, the extent to which they exemplify an emerging but increasingly shared vision of the future of data-driven educational research, and, in particular, what their activities indicate about the potential of big data-driven theory generation within the field of educational research itself. Pearson and Lytics Lab do not just own the big data technologies and the information they collect, but own the algorithms and analytics required to make sense of those data, and thereby potentially to generate novel theories about diverse processes and learning and education. According to these well-resourced institutions, big data and algorithmic forms of analysis are revealing a mismatch between the patterns of learning located in masses of data and existing conceptual frameworks to explain them. In addressing the theory gap, they are applying methodological approaches and epistemological assumptions from data science. This shift is beginning to reveal a political economy dimension to educational knowledge production and theory generation, with well-resourced commercial businesses like Pearson and prestigious institutions like Stanford University gaining legitimacy and authority through their expertise and technical capacity to generate insights algorithmically from big data.

The article addresses two themes of this special issue, (1) by querying how education is being theorized in the context of contemporary forms of algorithmic data scientific methods and epistemologies, and (2) by interrogating the political economy of education data science as authority, legitimacy and power in educational knowledge production and theory generation has begun to flow to a concentrated network of well-resourced big data R&D centres. The core argument is that as educational data science has begun migrating from the academic lab to the commercial sector, it is accumulating significant social, economic and cultural capital as a new field of expertise, and also that it is assembling a new kind of methodological capital that gives it the potential to gain competitive advantage over other methods and approaches to the study of e-learning and digital media.

Imaginaries, fields and methods

The article consists of two case studies of dominant sites in the emergence of the field of educational data science. One intention of the article is to construct a partial genealogy of this field, its emergence and its growth, interrogating the kind of 'imaginary' of the future of educational research it projects and the political economy that underpins it, mobilizing the concept of 'sociotechnical imaginaries' (Jasanoff, 2015). Sociotechnical imaginaries are socially shared visions of technologically mediated progress that have moved from single inspired individuals to much wider communities and fields of action. Educational data science is being animated by a particular sociotechnical imaginary of a desirable future of educational research, one being made seemingly attainable through the technologies and social practices of data science. Taking a genealogical view of the formation of this field—its convergences, alliances and juxtapositions—can help reveal the 'history of the present' of educational data science, as well as a 'history of the future' of the desirable imaginary it projects for itself as a field.

Considered as a 'field,' educational data science can be approached analytically through Bourdieu's (1993) notion of 'fields of power,' particularly as it has been applied in relation to education and methodological innovation. A field is a relatively autonomous domain of action in which agents and groups vie over socially valued resources, and it is through such processes that a field can develop its own 'distinction' or institutional niche from others (Simons, Olssen and Peters, 2009). Each field is a site for the creation of a particular kind of capital, or an institutionalized resource. These include economic capital in the form of money and other financial assets; cultural capital, or socially valued knowledge and credentials; and access to social networks and webs of relationships, or social capital. Taking this framework, educational data science can be understood in terms of its access to economic capital in the shape of funding and resourcing, its cultural capital in terms of the production of new specialist knowledge such as research findings derived from new expert methodologies, and the social capital it gains through its wider networks of affiliations, partnerships and connections. Educational data science is, in other words, a nascent methodological field of power creating its own distinction through a combination of economic, cultural and social capitals, all of it animated by a particular sociotechnical imaginary of a data scientific future for educational research and knowledge production.

The Bourdieusian approach to fields can also be applied to the formation of methodological fields of expertise. Sociologists have recently begun interrogating the 'social life of methods' (Ruppert, Savage and Law, 2013). This approach foregrounds methods themselves as objects of social scientific inquiry, and involves the genealogical examination of specific methods in order to identify the theoretical trajectories and assumptions that underlie their development—such as assumptions about what to measure and how to measure it—and then to identify how such methods contribute to the production of knowledge that might influence subsequent thinking and decision-making. In what follows, I offer a partial genealogy of educational data science that therefore pays attention to its financial resources, its knowledge production, and its social networks as an emerging methodological field of power. From the perspective of 'field theory, the aim is to analyse both the structures and relations within

fields and the dispositions and schemata of perception of the agents who inhabit the fields' (Simons, Olssen and Peters, 2009: 64). My own emphasis, then, is partly on locating the structures and relations within educational data science by conducting a genealogical survey of its formation as a distinctive field, but also on tracing the sociotechnical imaginary that animates the actors who inhabit it, treating such an imaginary as a kind of 'schema of perception' that galvanizes their social, discursive and material practices as technical and methodological experts.

Methodologically, constructing the genealogical case studies of the Stanford Lytics Lab and the Pearson Center for Digital Data, Analytics and Adaptive Learning has involved close documentary analysis of the various materials, reports, documents, websites, presentations and other outputs produced by actors from these sites. Doing so has allowed me to make sense of the kind of sociotechnical imaginary of the future of educational data science that animates them, and that they are seeking to materialize in practice, and to trace some of the ways in which these organizations have sought to create a kind of 'force field' of expert power around educational data science through their access to and deployment of economic, social and cultural capitals.

The social life of educational data science

Data science has emerged as a rapid growth discipline of the early twenty-first century, particularly with the emergence of 'big data.' In technical terms, big data refers to data sets that are huge in volume (at the scale of petabytes, exabytes and zettabytes), highly diverse in type and nature, generated continuously at great velocity in or near real time, exhaustive in scope, fine-grained in resolution, combinable with other networks of datasets, and flexible and scalable enough for new fields to be added and to expand in size rapidly (Kitchin and McArdle 2015). Whatever the size or speed that might define big data, making any sense of it requires sophisticated analysis. Hilbert (2015: 139) suggests that the 'full name' for big data is 'big data analytics,' since, 'independent from the specific peta-, exa- or zettabytes scale, the key feature of the paradigmatic change is that analytic treatment of data is systematically placed at the forefront of intelligent decision-making.' Building on established statistical and analytics methods developed in scientific settings and industry research over the last half-century, new data analytics and data mining technologies have been developed to detect, classify and extract associations and patterns from large datasets utilizing advances in information management and storage, data handling, algorithm design, and machine learning (Kitchin 2014; Mackenzie 2015).

Given the technical complexity of conducting big data analyses, and the far-reaching implications of data mining for a variety of social realms, the figure of the data scientist has been awarded particular expert status and power. Gehl (2015: 414) has characterized 'the rare subject capable of mining these messes':

> the Data Scientist, armed with … a large pile of data, algorithms and not a little genius. As with past generations of knowledge workers, the data scientist is called forth to tame the excesses of our constant sharing and mine it for new knowledge and produce valuable new techniques of social management.

As the ideal-type knowledge worker of the big data era, these new data scientific experts of the social world have been termed 'algorithmists'—multidisciplinary specialists in computer science, mathematics, and statistics, as well as policy, law, economics and social research—who can undertake big data analyses across commercial, political and scholarly sites (Mayer-Schönberger & Cukier 2013). Algorithmists are a new kind of super-class of scientific expertise, the knowledge workers who can work with algorithms to extract meaning from masses of data, visualize it for the consumption of others, and produce the insights, facts, and evidence that might lead to decision-making and other actions across diverse scientific and social domains.

However, data science is not simply a neutral and unbiased disciplinary and methodological field of expertise. Big data analyses and data science have often been accompanied by grand claims about a 'paradigm shift' in various kinds of research, one that emphasizes the inherent truthfulness and unbiased, impartial agnosticism of numbers (examples of such claims can be found in Kitchin 2014). The data-intensive statistical exploration and data mining of phenomena that are characteristic of data science reflects a 'data-ist' belief that there is no need for prior theory, models or hypotheses and that 'data can speak for themselves' unencumbered by human interpretation, bias and meaning-making (e.g. Anderson 2008). But data are always created through systems that are designed with very specific purposes according to particular scientific theories, methodological preferences and ways of working, while making sense of data is always framed by particular interpretive lenses:

> Even if the process is automated in some way, the algorithms used to process the data are imbued with particular values and contextualized within a particular scientific approach. … [D]ifferent analysts will draw different conclusion from the same analytics. Interpretation then is always in the eye of the beholder regardless of how neutral or value-free they claim to be. … As such, data never simply speak for themselves. (Kitchin 2014: 136)

These points highlight that data scientists, big data analysts and algorithmists work in specific disciplinary, professional and scientific contexts and a thought community with its own distinctive style of thinking, schema of perception, mode of expertise, specialist language, shared concepts, theories and practices that can be organized into explanations and arguments. It is through its empirical, data-driven schema of perception that data science is able to project itself as a distinctive field of power, with access to the resources, networks and expertise to produce new knowledge and understandings in diverse domains including healthcare, employment, crime and education.

The figure of the expert algorithmist is coming to occupy the field of educational research in the shape of the 'educational data scientist.' Educational data science itself is an emerging, transdisciplinary field, building on data scientific practices as well as existing knowledges from the learning sciences (itself a combination of psychological, cognitive and neurological sciences). Piety, Hickey and Bishop (2014) date the emergence of educational data science from around 2004-2007 as various forms of educational and learning analytics and data mining practices and communities combined. They particularly highlight how a community of researchers began to converge around educational data mining from about 2005, and more recently to team up with the learning analytics community to form a field that 'has begun to

receive combined attention from both federal policymakers and foundation funders and is often seen as the community dealing with "Big Data" in education' (Piety, Hickey & Bishop 2014: 3).

They term educational data science a 'sociotechnical movement' with shared interests that cut across the boundaries of its original communities. By sociotechnical movement what they mean is that 'the enabling conditions and key technologies emerge across sectors giving rise to multiple sets of innovations that may at times seem disconnected, but are often related and interdependent' (Piety, Hickey & Bishop 2014: 4-5). They also point out that a sociotechnical movement can gain traction when society's 'expectations are such that the innovations come at a time when there is other general interest in the kinds of changes that the innovations make possible' (Piety, Hickey & Bishop 2014: 5). Thus there has, in recent years, been both increasing capability to produce data and a greater public appetite for the use of data across many areas of education. They also highlight how new forms of evidence—log files, conversational records, peer assessments, online search and navigation behavior, and others—are raising big questions and disrupting traditional ways of working in educational research, 'acting in a way similar to *disruptive innovations* that alter cultural, historical practices and activity systems' (Piety, Hickey & Bishop 2014: 5).

These developments clearly have major implications in terms of technical and methodological expertise. In a collaborative presentation defining the field Piety, Behrens and Pea (2014) have traced its disciplinary origins to computer science techniques of computational statistics, data mining, machine learning, natural language processing and human-computer interaction. Commenting on this emerging field, Cope and Kalantzis (2016: 13) identify how 'big data and education data sciences may in time offer learners, teachers, and researchers new windows into the dynamics and outcomes of learning, finely grained in their detail, varied in their sources and forms, and massive in their scope,' though they caution that 'much work still needs to be done in the nascent field of education data sciences before the affordances of computer-mediated learning can be fully realized in educational practice.'

Anticipating the requirement for more expert educational data scientists, the Stanford University Lytics Lab founding director Roy Pea (2014) has called for much more support from governments for this sector, and details the need for new undergraduate and graduate courses to support its development. Pearson, for its part, established its own Center for Digital Data, Analytics and Adaptive Learning to practise educational data science in-house. The Connected Intelligence Centre at the University of Technology Sydney, the Institute of Technology at the Open University, and the LINK (Learning Innovation and Networked Knowledge) Research Lab at the University of Texas at Arlington, are other prominent sites of learning analytics R&D, and there is a substantial and fast-growing body of learning analytics literature (e.g. Siemens 2013; Clow 2013). The founder of the LINK lab, Siemens (2016) has also documented the emergence of the Society of Learning Analytics Research (SoLAR), a global members association of learning analytics researchers and developers with its own journals and an annual conference. These developments have been supported with

significant commercial sponsorship as well as academic partnership. As an association, SoLAR has had explicit goals both in terms of technical R&D and pedagogic innovation:

> Advances in knowledge modeling and representation, the semantic web, data mining, analytics, and open data form a foundation for new models of knowledge development and analysis. The technical complexity of this nascent field is paralleled by a transition within the full spectrum of learning (education, work place learning, informal learning) to social, networked learning. These technical, pedagogical, and social domains must be brought into dialogue with each other to ensure that interventions and organizational systems serve the needs of all stakeholders. (Siemens 2016)

These aims express the highly normative trajectory of educational data science, which aspires not just to produce insights from educational data but projects a strong and collectively shared future vision of educational innovation and reform.

Together, these institutions and alliances constitute an emerging technical, organizational and professional infrastructure for educational data science. In the remainder of this article I provide case studies of both the Stanford Lytics Lab and the Pearson Center for Digital Data, Analytics and Adaptive Learning as key sites in the emergence of educational data science, focusing on the implications for educational research, knowledge production and theory generation.

Stanford Lytics Lab

The Lytics Lab (short for Learning Analytics Laboratory) was established at Stanford University in 2012 to 'advance the science of learning through the use of data and digital technology' (https://lytics.stanford.edu/about-lytics). Founded by Stanford doctoral students under the direction of Professor Roy Pea—an academic with a long professional history in educational technology—the lab has subsequently become the unofficial research group of the university's Vice Provost for Teaching and Learning. Originally established to investigate the masses of data becoming available from the suite of massively open online courses (MOOCs) offered by the university, the lab now conducts highly diverse educational data science R&D based around online learning and learning analytics. Its projects focus on understanding online learners, including dropout prediction tools and analytics of attainment gaps, and studies that evaluate 'digital instruction,' as well as the development of new 'learning tools' such as social learning platforms and systems that 'provide feedback at scale.' Its staff include academics and doctoral students from a cross-disciplinary selection of 'Computer Science, Learning Science, Communication, Psychology, Statistics, Design, and Sociology.'

The Lytics Lab has been both a key driver of educational data science and the recipient of support from across academia, business and government. In 2014, the Stanford University Learning Analytics Workgroup published a report on 'building the field of learning analytics at scale' (Pea 2014). It was authored by Roy Pea, then director of the Lytics Lab as well as being the David Jacks Professor of Education and Learning Sciences and a courtesy professor of Computer Science. The working group was co-funded by the Bill and Melinda Gates Foundation and the MacArthur Foundation, two of the major grant-giving philanthropic foundations for educational technology in the US. The origins of the report actually lay in a 2011 event organized by the Gates Foundation at the University of Chicago Computation

Institute—from which Roy Pea coauthored a draft white paper (Pea, Childress & Yowell 2012)—and in a subsequent agreement with the National Academy of Education which led to a series of workshops and summits. These working group events were attended by Stanford University researchers who would subsequently establish the Lytics Lab including Roy Pea; John Behrens of Pearson (later the director of Pearson's Center for Digital Data, Analytics and Adaptive Learning); staff from many commercial MOOC platform providers (Khan Academy, Coursera); individuals from commercial computing businesses including Intel; philanthropic donors from the Gates Foundation; representatives of testing agencies such as SRI and ETS; academics working on learning analytics development in technology-focused research centres from other universities; and governmental officials from the White House Office of Science and Technology Policy and the US Department of Education's Institute of Education Sciences. In Bourdieusian terms, a great deal of economic and social capital—in the shape of funding and cross-sectoral social networks—became available to the Lytics Lab through its centrality to this working group and the various task forces it generated.

In the report of the working group, Pea (2014) proposed a new 'specialized' field combining the sciences of digital data and learning, and the construction of a 'big data infrastructure' for learning consisting of data science and computer science techniques that could be harnessed to the challenge of analysing large volumes of educational and learning data. The report established the need for a new kind of 'professional infrastructure in the field of learning analytics and education data mining, made up of data scientists (straddling statistics and computer science) who are also learning scientists and education researchers' (Pea 2014: 17). Specifically, it identified 'several competencies for education data science' that would contribute to this professional infrastructure, including:

- Computational and statistical tools and inquiry methods, including traditional statistics skills … as well as newer techniques like machine learning, network analysis, natural language processing, and agent-based modeling
- General educational, cognitive science, and sociocultural principles in the sciences of learning…
- Principles of human–computer interaction, user experience design, and design-based research
- An appreciation for the ethical and social concerns and questions around big data, for both formal educational settings and non-school learning environments (Pea 2014)

Expertise in psychometrics and educational measurement, cognitive neuroscience, bioinformatics, computational statistics, and other computational methods were also promoted in the report. These disciplinary practices and competencies offer a clear sense of the style of thinking underpinning educational data science, particularly its computer science and data science origins twinned with primarily psychological, cognitive and neuroscientific theories of learning—or 'learning science.' The report is a material instantiation of the sociotechnical imaginary of a big data-driven approach to educational research that animates the educational data science field.

The 'social life' of educational data science methods reveals its disciplinary as well as social and economic origins. However, methods are not just products but are also productive in the sense that they turn existing theories into instruments designed to measure the reality they

purport to explain. Pea (2014: 24) specifically highlights 'a pre-eminent objective' in educational data science of:

> *creating a model of the learner*. What characteristics are important as predictors for what is appropriate to support the learner's personalized progress? What are the classes of variables and data sources for building a learner model of the knowledge, difficulties, and misconceptions of an individual?

This field depends on the generation of models of learners, assembled from their digital data, which can be coded into pedagogic software tools and have the subsequent potential to shape the 'personalized progress' of learners.

The models and theories that galvanize educational data science are not all focused on cognitive aspects of learning. Pea (2014: 28) proposes using data scientific methods to engage with '"non-cognitive factors" in learning, such as academic persistence/perseverance (aka "grit"), self-regulation, and engagement or motivation,' that are 'improvable by appropriate practices.' Various techniques of measuring the 'emotional state' of learners include collecting 'proximal indicators that relate to learning' through such techniques as 'facial expressions detected by a computer webcam while learning' (32), plus other data sources like 'video, eye tracking, and skin temperature and conductivity' (46). Piety, Hickey and Bishop (2014: 3) also promote data science methods to measure 'student characteristics' including:

> cognitive traits like aptitudes, cognitive styles, prior learning, and the like, as well as the learners' non-cognitive characteristics such as differences in levels of academic motivation, attitudes toward content, attention and engagement styles, expectancy and incentive styles … persistence through adversity … [and] tenacity or grit.

The discourse in these texts of non-cognitive student characteristics of motivation, engagement, 'grit,' self-regulation, emotional state, and so on, is highly indicative of the strongly psychological genealogy of the education data science field. It particularly points toward the possibility of using digital devices to collect and calculate data about students' emotions during educational experiences, and then offering psychologically-defined prescriptions towards emotional maximization. These emphases on the non-cognitive aspects of learning also crucially align educational data science with emerging policy discourses of non-academic social and emotional learning, and emerging attempts to quantify the 'personal qualities' and 'character skills' of students as a measure of the effectiveness of school systems (Schechtman et al 2013; OECD 2015).

This brief genealogical survey of the 'social life' of the Stanford Lytics Lab indicates how educational data science has its origins as a field in a set of cross-sector concerns shared among academic, commercial, philanthropic and governmental actors and organizations. These organizations, between them, have begun to construct a vast big data infrastructure for the production of educational data scientific knowledge and theory. As a key catalyst and propellant for an increasingly shared and collective imaginary of the data-driven future of educational research, the Lytics Lab is itself a significant social actor with the economic, cultural and social capitals required to establish the field of educational data science.

Pearson Center for Digital Data, Analytics and Adaptive Learning

As already noted, one member of the Learning Analytics Working group that contributed to the report by Roy Pea was John Behrens of Pearson. Pearson plc is the world's largest educational publisher, and recently significantly extended its operations and ambitions to include R&D in digital learning and big data analysis in education. Its Center for Digital Data, Analytics and Adaptive Learning (CDDAAL), a research and development centre dedicated to the analysis and use of digital data for educational improvement, was established in 2012. Though the centre appears to have been closed during a restructure of Pearson in 2016—amid falling revenues and reputational decline—Pearson has retained a strong commitment to digital learning and educational data analysis. Crucially, Pearson is a key actor in the field of education data science because it represents how this field is being scaled up into mainstream practice in part by commercial actors with business plans to achieve, proprietorial products to sell and profits to secure.

Pearson established CDDAAL under the leadership of John Behrens to investigate how the billions of bits of digital data generated by students' interactions with online lessons and everyday digital activities could be combined to personalize learning. Its staff were described as 'research scientists' with expertise in data mining, computer science, algorithm design, intelligent systems, human-computer interaction, data analytics tools and methods, and interactive data visualization. In a methodological report for CDDAAL, Behrens (2013) claimed that educational research was increasingly under pressure to adopt new computational and data science methods. These methods would enable data manipulation and data visualization, including the mobilization of 'big data' to enable continuous tracking and monitoring of streaming data, rather than the collection of data through discrete temporal assessment events. They would include 'population analytics' techniques that can handle enormous, scalable samples of many millions of records of research data, make use of 'educational data mining' to extract patterns from it, and utilize statistical models for combining results from different datasets and to integrate new and existing data and information.

In another CDDAAL publication, data science was a positioned as a 'transformative' methodology:

> Once much of teaching and learning becomes digital, data will be available not just from once-a-year tests, but also from the wide-ranging daily activities of individual students … in real time. … [W]e need further research that brings together learning science and data science to create the new knowledge, processes, and systems this vision requires. (DiCerbo & Behrens 2014)

CDDAAL's researchers aimed to mobilize techniques of social network analysis to mine students' data for patterns, based on the understanding that, 'faced with a very large number of potential variables, computers are able to perform pattern identification tasks that are beyond the scope of human abilities … not only to collect information but also detect patterns within it' (DiCerbo & Behrens 2014). To do this, the report detailed how pattern recognition analysis could be used to trace and match patterns in learners' activities:

> Learner interactions with activities generate data that can be analysed for patterns. … Performance in individual activities can often provide immediate feedback … based on local pattern recognition, while

> performance over several activities can lead to profile updates, which can facilitate inferences about general performance. (DiCerbo & Behrens 2014)

As learners interact with systems and with other people, 'software records' every aspect of their activity so that as learners interact in digital environments, in formal and informal contexts, 'actionable data can be drawn from both':

> These developments have the potential to inform us about patterns and trajectories for individual learners, groups of learners, and schools. They may also tell us more about the processes and progressions of development in ways that can be generalised outside of school. (DiCerbo & Behrens 2014)

CDDAAL researchers explicitly mobilized pattern recognition methods including cluster analysis, natural language processing, Bayesian networks, neural networks and statistical analysis to reveal the hidden patterns of learning and build generalizable models of cognitive development. Behrens (2013: 18) even argued that insights extracted from the generation of huge quantities of educational data would challenge existing theoretical frameworks in the educational research field, as 'new forms of data and experience will create a theory gap between the dramatic increase in data-based results and the theory base to integrate them.' Through its big data analytics methods, Pearson proposed that it could generate new insights into and understandings of learning itself, using the results of data analysis to build new theories.

A significant aspect of the social life of Pearson's work in educational data science is its partnership with the learning analytics and adaptive learning platform provider Knewton. Since 2011, Knewton has provided the back-end analytics to many of Pearson's online learning and e-learning products. Through their partnership, Knewton combines the power and potential of its adaptive learning platform with Pearson's content and distribution, promising to 'usher in a new era of personalized and customizable education products':

> The Knewton Adaptive Learning Platform™ uses proprietary algorithms to deliver a personalized learning path for each student…. 'Knewton adaptive learning platform, as powerful as it is, would just be lines of code without Pearson,' said Jose Ferreira, founder and CEO of Knewton. 'You'll soon see Pearson products that diagnose each student's proficiency at every concept, and precisely deliver the needed content in the optimal learning style for each. These products will use the combined data power of millions of students to provide uniquely personalized learning.' (http://www.knewton.com/press-releases/pearson-partnership/).

Knewton's 'proprietary algorithms' have the capacity to predict students' probable future progress through predictive analytics processes, and then to 'personalize' their access to knowledge through modularized connections that has been deemed appropriate by the algorithm. For example, all content in the platform is linked by the 'Knewton knowledge graph, a cross-disciplinary graph of academic concepts' (Knewton 2013: 6). The 'knowledge graph' treats knowledge in terms of discrete modules of content that can be linked together to produce differently connected personalized pathways, enabling the Knewton platform to refine its 'recommendations through network effects that harness the power of all the data collected for all students to optimize learning for each individual student' (Knewton 2013: 8).

One question to be raised here is about how 'learning' can be counted in a database. In order for anything to be entered into a database, it first needs to be sorted into a classification system (Bowker 2008). Notably, the machine learning algorithms that underpin most analytics packages need to be trained to 'learn' from 'a data sample that has already been classified or labelled by someone. … The classification becomes what the data mining techniques seek to learn or model so that future instances can be classified in a similar way' (Mackenzie 2015: 433). Data analytics platforms, including adaptive learning analytics such as Knewton, produce algorithmically-learned knowledge that can be deployed to shape future activities, though this is dependent on the prior classificatory labour of the algorithm designers. This means that for Pearson and Knewton to make algorithmic calculations about learning processes, there needs to be a precise classification scheme available in advance into which various indicators of learning can be entered. Knewton's knowledge graph depends on techniques of content classification and taxonomization for its functioning. Without such processes of categorization taking place, content cannot be fitted into the knowledge graph. Once the classification of content has taken place, it can then be:

> organized in a graph-like structure, which means that the student flow from concept-to-concept can be optimized over time, as Knewton learns more and more about the relationships between them through data. Every student action and response around each content item ripples out and affects the system's understanding of all the content in the system and all the students in the network. (Knewton 2013: 14)

The human act of training the algorithm to identify and learn from things that have been classified or labelled for inclusion in the knowledge graph indicates how machine learning is both a form of automated knowledge production, but also one shaped by people working in specific labour conditions, within institutional frameworks, according to professional commitments, worldviews and disciplinary theories about the ways in which the world works. These contextual factors are consequential to the ways in which machine learning is trained, re-trained, and checked to ensure the accuracy and generalizability of its models (Mackenzie 2015). With Knewton specifically, the classifications of learning it uses are drawn from learning science, a field itself largely defined in terms of concepts and methods from the psychological and cognitive sciences. Therefore any inferences or insights drawn from its data analyses need to be understood as pre-defined by the theoretical, conceptual and classificatory systems of this particular field of educational research and its idiosyncratic epistemological commitments and schema of perception.

More recently, Pearson also partnered with IBM, one of the world's largest computing companies, to embed its 'cognitive computing' technologies in its courseware content. The key technology underpinning their ambitions is Watson, IBM's highly-publicized cognitive supercomputing system. IBM (2016a) describes Watson as 'a cognitive technology that can think like a human': it is able to analyse and interpret data, including unstructured text, images, audio and video; it can 'reason' and 'provide personalized recommendations by understanding a user's personality, tone, and emotion'; and can also 'learn,' utilizing machine learning to 'grow subject matter expertise,' and 'interact' through 'chat bots that can engage in dialog'. The partnership with Pearson will allow Watson to penetrate into educational institutions at huge scale, thanks to the massive reach of Pearson's courseware products. Pearson (2016) stated it would 'make Watson's cognitive capabilities available to millions':

> Pearson and IBM are innovating with Watson APIs, education-specific diagnostics and remediation capabilities. Watson will be able to search through an expanded set of education resources to retrieve relevant information to answer student questions, show how the new knowledge they gain relates to their own existing knowledge and, finally, ask them questions to check their understanding.

Strikingly, the partnership proposes that Watson will act as a 'flexible virtual tutor' which would be 'embedded in the Pearson courseware,' able to read the content and then spot patterns and generate insights in order to assess students' responses and guide them with automatically generated feedback and explanations in real time (IBM 2016b). Like Knewton, IBM's Watson is protected by intellectual property rights and patent laws. Both of these technical platforms and the organizations that developed them have become central to the enactment of educational data science, not just as a field of knowledge production, but as a field for the production of 'actionable insights' and automated intervention.

Although Pearson's CDDAAL disappeared with its 2016 organizational restructure, the company's commitment to data-driven digital education persists through its strategic partnerships with Knewton and IBM. Ultimately, Pearson is positioning itself as a big data gatekeeper in relation to the production of new knowledge about learning. It has a vast organizational, technical and expert infrastructure—in the shape of in-house analysts and developers, plus its partnerships with Knewton and IBM Watson—for conducting big data analyses in education. It is seeking to use the insights it generates from such analyses to construct new conceptual models and theories of learning that it can encode into new e-learning products. But these models and theories themselves reflect and reinforce existing styles of thinking that can be traced back to psychological learning sciences as well as to the methodological practices, schema of perception and presumptions of the data sciences.

This case study has shown how Pearson has successfully commercialized educational data science by embedding new data analytics applications in its courseware and e-learning products. Its partnerships with Knewton and IBM, which rely on highly proprietorial systems, are already enabling educational data science techniques and applications to penetrate public education through schools and colleges alike. Pearson's commercial role in educational data science therefore raises a very pressing concern for research in digital learning. It now appears that insight into learning itself is increasingly likely to emanate from private companies with their own proprietorial systems, intellectual property claims, and market needs. These companies are staking their claim to expertise in the conceptualization of learning through their ownership of the systems required to calculate big data.

Theory, patents and intellectual property

The algorithmic techniques of educational data science represent a set of distinctive challenges for educational researchers and their participation in knowledge production and theory development. Cope and Kalantzis (2016: 11), for example, note that:

> Statistical patterns in machine learning data are to a significant extent creatures of patterns already built into supervised training models. In the case of unsupervised machine learning, the statistical patterns make sense only when they are given explanatory labels. For these reasons indeed, theory is needed more than ever to frame data models, to create ontologies that structure fields for data collection, and for model tracing.

This point about increasing the importance of theory in big data analysis raises significant questions about the sites of expertise where such theory generation might occur. As the case studies of the Stanford Lytics Lab and the Pearson Center for Digital Data, Analytics and Adaptive Learning indicate, educational data science is being concentrated in well-resourced and highly-financed research centres, labs and partnerships.

This brings us back to the idea of educational data science as a 'field of power', whereby it can be understood as a distinctive set of relations and structures between a range of actors seeking to establish economic, cultural and social capital. As a field of expert power animated by a particular sociotechnical imaginary of a desirable future for educational research, educational data science has begun to accumulate significant economic capital in the shape of research funding and institutional resources. It has developed significant social capital through its own professional associations, conferences and journals, as well as its links to commercial industry, prestigious academic institutions, and governmental supporters. It has also accumulated cultural capital through its innovative methodological production of new knowledge, and has significant ambitions for future data-based theory generation too.

In this sense we can see educational data science as a field that is struggling for power and distinction through its pursuit of an imaginary of the future of educational research, with new theory generation as its target. The fact that this field is being concentrated in significantly well-resourced projects and partnerships such as those associated with Stanford and Pearson raises real questions about its capacity to generate new theories that might themselves reshape the ways in which processes of learning, aspects of cognition, and also the noncognitive aspects of learning such as social and emotional learning are known, understood and accepted more widely in the educational research field. Pearson and Stanford have the big data infrastructure to conduct the kinds of advanced data scientific studies that few education departments in universities can perform. The consequence is that as big data gains credibility as the source for educational knowledge production and theorizing, it is likely that legitimacy will flow toward those centres able to conduct such analyses. In the context of big data analytics more generally, Nielsen (2015) notes that it 'has become profitable to build a database containing the entire world's knowledge. The few for-profit companies that own the data and the tools to mine it – the data infrastructure – possess great power to understand and predict the world.' In other words, there is a political economy dimension to educational theorizing as it seems to be migrating toward commercial companies like Pearson, Knewton and IBM, or academic institutions with close industry and governmental connections such as Stanford.

By building systems that include proprietorial technologies and algorithms, these companies are not only constructing a technical and professional infrastructure for educational data science, but concentrating the means for the production of knowledge in their own hands. Ownership of these systems gives them extraordinary power to generate new insights from the data they have collected and analysed, and potential to generate new theoretical explanations. To the extent that data-driven education research is increasingly attractive in the current climate of 'evidence-based' policy, which seeks to legitimize specific forms of political action by referring to 'hard' statistical scientific evidence (Rieder & Simon 2016),

ownership of proprietorial systems could become a prerequisite for the production of the kind of evidence-based explanations and conceptualizations sought by government agencies and departments.

How learning is to be understood and theorized therefore looks increasingly to be led by actors with the economic, social and cultural capital to generate evidence and insights from big data. Some of them, like Pearson, Knewton and IBM, might then stand to gain commercially by designing and patenting e-learning software resources on the basis of the theories they have generated — essentially a case of locking-in a theory to a specific technical innovation. Watters (2016) provocatively suggests that the technological future of education is one where software patents *become* educational theory, which 'does not guarantee that these companies have developed technologies that will help students learn. But it might mean that there will be proprietary assets to litigate over, to negotiate with, and to sell.' It is in this specific sense that education data science as a field may be able to lay claim to ownership of educational theory. Ultimately, its explanations of learning are being built-in to the tools they provide, in the shape of algorithm-driven technologies of personalization which can be sold to schools and universities. Knewton's knowledge graph is one example of how a theory of learning based on a data-driven epistemology that data can reveal insights that previous theories have failed to explain has been coded into a platform that is now in use worldwide.

As platforms like Knewton then generate further data, these new data can be analysed and presented to prove their efficacy and effectiveness, while organizations like Pearson can then seek to insert those data into evidence-based policy decision-making in relation to the use of big data software in education. These technology-mediated theories then have the potential to flow back into the pedagogic spaces of schools and colleges, reshaping how teaching and learning are conducted. Piety, Hickey and Bishop (2014: 9) recognize that educational data science technologies 'encode various theories of learning that manifest themselves in the data the tools provide.' The underlying theories of learning contained in the information architectures of educational data science, and the models of the learner constructed by experts in the field, are therefore consequential to ways of both theorizing and acting practically upon individuals.

As big data practices increasingly infuse educational research, and educational analyses are performed by profit-making companies with ownership of the relevant big data facilities and proprietary algorithms, the question of who owns educational theory is becoming one of serious concern. The ownership of educational big data, the ownership of educational theory, and the application of such theories within proprietorial systems and software patents may then be leading to a near-future scenario where private companies with market imperatives become government-approved sites of expertise into learning and teaching processes. In this context, how learning is conceptualized and understood looks likely to become a kind of intellectual property for privately-resourced research centres.

Conclusion

Originating in academic research in the mid-2000s, educational data science has recently migrated from academic labs to commercial organizations such as Pearson to become a cross-sector enterprise driven by shared visions and interests. As an emerging field of power in educational research, knowledge production and theory development, educational data science represents an attempt to replicate the figure of the data scientist, or the entrepreneurial algorithmist, in the educational field. This is a figure armed with new kinds of expertise, new epistemologies, and new methodologies associated with big data. It would not be accurate to characterize data science solely in terms of positivist biases, but it is clear that it is oriented around methods and assumptions that are challenging to mainstream educational research. Data scientists have particular epistemological outlooks, schemata of perception, professional ways of working, and methodological approaches that are often technicist, functionalist and instrumentalist, and that tend to neglect the social, cultural, political and economic factors that contribute to the phenomena they analyse (Kitchin 2014).

Data scientific knowledge production and theory generation in education is likely to reflect the idiosyncratic disciplinary and professional style of thinking of data science. These are by no means neutral, value-free or atheoretical approaches. They are the product of academic and commercial actors animated by a specific imaginary of a desirable future of educational research and development, working within a field of power defined by its economic, social and cultural capital, whose practices reflect a particular data scientific style of thinking that views learning in scientific terms as quantifiable, measurable, actionable, and therefore optimizable.

Perhaps the most successful sites of educational data science are commercial organizations such as Pearson, Knewton and IBM. These commercial education data science organizations now possess concentrated ownership of the technologies and proprietorial algorithms to generate and analyse educational big data, and are claiming to be applying their expertise to fill a ‘theory gap’ in contemporary educational thinking. Their theories of learning—that it can be optimized by being personalized, and that students’ encounters with content can be best managed through knowledge graphs—are coded-in to the technologies that some commercial participants in educational data science are then able to sell to schools and colleges. This means their underlying theories of learning can flow back into schools and colleges, reshaping the pedagogic practices of teachers and the learning processes of students. With the knowledge and insights gained from their systems, they are then positioning themselves to provide the hard scientific data-driven explanations required by a current emphasis on evidence-based policy, and securing themselves further capitals and power over the educational field by so doing.

As educational institutions generate increasing quantities of digital data, education data science might therefore be understood as assembling a new form of ‘methodological capital.’ The methodological capital of educational data science consists of competence in big data analyses, the ability to secure funding and strategic partnerships, and the capacity to produce knowledge and theory that may be effective in the competition for control over contemporary understandings of e-learning, digital media and education.

References

Anderson, C (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired*, 23 June 2008: https://www.wired.com/2008/06/pb-theory/

Behrens, J (2013) Harnessing the Currents of the Digital Ocean. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA., April, 2013.

Bourdieu, P (1993) *The Field of Cultural Production: Essays on art and literature*. Cambridge, Polity.

Bowker, G.C. (2008) *Memory Practices in the Sciences*. London: MIT Press.

Clow, D (2013) An overview of learning analytics. *Teaching in Higher Education* 18, no. 6: 683–695.

Cope, B & Kalantzis, M (2016) Big data comes to school: implications for learning, assessment and research. *AERA Open* 2, no. 2: 1-19.

DiCerbo. KE. & Behrens, JT (2014) *Impacts of the Digital Ocean*. Austin, TX: Pearson.

Gehl, R (2015) Sharing, knowledge management and big data: A partial genealogy of the data scientist. *European Journal of Cultural Studies* 18, no. 4-5: 413-428.

Hilbert, M (2016) Big data for development: A review of promises and challenges. *Development Policy Review* 34, no. 1: 135-174.

IBM (2016a) Go beyond artificial intelligence with Watson. http://www.ibm.com/watson/

IBM (2016b) IBM Watson Education and Pearson to Drive Cognitive Learning Experiences for College Students. http://www-03.ibm.com/press/us/en/pressrelease/50842.wss

Kitchin, R (2014) *The Data Revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Luckin, R & Holmes, W (2016) Intelligence Unleashed: An argument for AI in education. London: Pearson.

Mackenzie, A (2015) The production of prediction: What does machine learning want? *European Journal of Cultural Studies* 18, no. 4-5: 429-445.

Mayer-Schonberger, V & Cukier, K (2013) *Big Data: A revolution that will change how we live, work and think*. London: John Murray.

Nielsen, M (2014) Who owns big data? *Change: 19 Key Essays on How the Internet Is Changing Our Lives*. Open Mind: https://www.bbvaopenmind.com/en/book/19-key-essays-on-how-internet-is-changing-our-lives/

OECD (Organization for Economic Cooperation and Development) (2015) Skills for Social Progress: The Power of Social and Emotional Skills. OECD Skills Studies, OECD Publishing.

Pea, R (2014) A report on building the field of learning analytics for personalized learning at scale. Stanford: Stanford University.

Pea, R, Childress, S & Yowell, C (2012) The Demand for Education Data Scientists, A working paper by H-Star Institute. Stanford, CA. Stanford University.

Pearson (2016) IBM Watson Education and Pearson to Drive Cognitive Learning Experiences for College Students. https://www.pearson.com/news/media/news-announcements/2016/10/ibm-watson-education-and-pearson-to-drive-cognitive-learning-exp.html

Piety, PJ, Behrens, J & Pea, R (2013) Educational data sciences and the need for interpretive skills. American Educational Research Association, 27 April-1 May 2013, San Francisco.

Piety, PJ, Hickey, DT & Bishop, MJ (2014) Educational data sciences—framing emergent practices for analytics of learning, organizations and systems. *LAK '14*, March 24 - 28 2014, Indianapolis.

Rieder, G & Simon, J (2016) Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data and Society* 3, no. 1: http://dx.doi.org/10.1177/2053951716649398

Ruppert, E (2015) Who owns big data? *Discover Society*, 30 July 2015: http://discoversociety.org/2015/07/30/who-owns-big-data/

Ruppert, E, Law, J & Savage, M (2013) Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture & Society* 30, no. 4: 22–46.

Shechtman, N, DeBarger, AH, Dornsife, C, Rosier, S & Yarnall, L (2013) *Promoting Grit, Tenacity and Perseverance: Critical factors for success in the 21st century.* U.S. Department of Education, Office of Educational Technology.

Siemens, G (2013) Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist* 57, no.10: 1380-1400.

Siemens, G (2016) Reflecting on Learning Analytics and SoLAR. Elearnspace, 28 April 2016: http://www.elearnspace.org/blog/2016/04/28/reflecting-on-learning-analytics-and-solar/

Simons, M, Olssen, M & Peters, MA (2009) Re-reading education policies. Part 2: Challenges, horizons, approaches, tools, styles. In M. Simons, M. Olssen & M.A. Peters (eds). *Re-Reading Education Policies: A handbook studying the policy agenda of the 21st century*: 36-95. Rotterdam, Sense Publishers.

Watters, A (2016) Ed-tech patents: prior art and learning theories. Hack Education, 12 January 2016: http://hackeducation.com/2016/01/12/patents