*AERA Open*
*January-December 2021, Vol. 7, No. 1, pp. 1–12*
*DOI: 10.1177/23328584211052055*
*Article reuse guidelines: sagepub.com/journals-permissions*


# Education Data Science: Past, Present, Future

**Daniel A. McFarland**
**Saurabh Khanna**
**Benjamin W. Domingue**
*Stanford University*
**Zachary A. Pardos**
*University of California, Berkeley*

*This AERA Open special topic concerns the large emerging research area of education data science (EDS). In a narrow sense, EDS applies statistics and computational techniques to educational phenomena and questions. In a broader sense, it is an umbrella for a fleet of new computational techniques being used to identify new forms of data, measures, descriptives, predictions, and experiments in education. Not only are old research questions being analyzed in new ways but also new questions are emerging based on novel data and discoveries from EDS techniques. This overview defines the emerging field of education data science and discusses 12 articles that illustrate an AERA-angle on EDS. Our overview relates a variety of promises EDS poses for the field of education as well as the areas where EDS scholars could successfully focus going forward.*

Keywords: *data science, network analysis, natural language processing, machine learning, learning analytics, data mining*

## Past: How Did We Get Here? Why Data Science Now?

### *History of Data Science*

As early as the 1960s, precursors of data science began to emerge. Tukey's "The Future of Data Analysis" can be considered a first step in this direction. His references to exploratory and confirmatory data analyses set in motion a chain of events that saw Naur (1974) and Wu (1986) arguing for "data science" as an alias for computer science and statistics respectively (Donoho, 2017). Naur (1974) published the "Concise Survey of Computer Methods" that surveyed data processing methods across a wide variety of applications. He considered data science to be "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to *other fields* [emphasis added] and sciences" (p. 30). Despite this allusion to "other" fields, data science originated as a byproduct of work in STEM fields and especially computer science and statistics. The International Association for Statistical Computing (IASC) in 1977 was established with a "mission to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge." The first Knowledge Discovery in Databases (KDD) workshop was organized in 1989, which later became the annual ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). The 1992 statistics symposium at the University of Montpellier was one of the first acknowledgments of data science as an emerging discipline harnessing data in different structural forms (Escoufier et al., 1995). Furthermore, data science was first featured as a standalone topic at the 1996 International Federation of Classification Societies conference (L. Cao, 2017).

Approaching the present, data science has become an essential idea not limited by traditional disciplinary boundaries. This need for boundary-crossing is exemplified by an argument to expand statistics beyond mere theoretical arguments (Cleveland, 2001). As the popularity of data science grew with the dawn of a new century, both the *Data Science Journal* and the *Journal of Data Science* were launched by the Committee on Data for Science and Technology and Columbia University, respectively. The establishment of these journals has been one part of a process leading to the consideration of data science as a domain distinct from both computer science and statistics. There were several justifications for this emergent decoupling of data science from these well-established fields. Data science had risen to occupy a unique disciplinary position on account of (a) being more application oriented as it targets solutions to real-world challenges (Donoho, 2017), (b) coupling quantitative and qualitative research across disciplines (Dhar, 2013), and (c) being largely focused on digital structured and unstructured data (Silver, 2020).

### *The Emergence of Education Data Science*

What implications did this rise of data science as a transdisciplinary methodological toolkit have for the field

of *education*? One means of illustrating the salience of data science in education research is to study its emergence in the Education Resources Information Center's (ERIC) publication corpus.[1] In the corpus, the growth of data science in education can be identified by the adoption rate of (at least) five prominent keywords: Learning Analytics, Machine Learning, Artificial Intelligence, Data Science, and Natural Language Processing. The adoption rate of these five terms can be compared with the overall growth in education research, see Figure 1. Four trends can be observed. First, while articles containing the five keywords are present in the Education Resources Information Center corpus as early as 2000, they are used quite sparingly until 2010. Second, after 2010, there is a sharp rise in published articles using these keywords. This boost could potentially be attributed to the meteoric rise in the popularity of e-learning and MOOCs (Massive Open Online Courses) between 2008 and 2012 (Clow, 2013; Yuan & Powell, 2013). Third, the increasing slopes for each curve indicates that EDS growth rates are accelerating with time. The speed of growth of data science related publications in education can be judged from the fact that the number of articles referring to "Learning Analytics" increased from a mere 0.01% in 2010 to around 0.35% in 2020—a 35-fold increase. Fourth, the relatively small absolute counts of articles with these keywords reflects that EDS is still a nascent subdomain within the larger domain of education.

Figure 1 illustrates a qualitative shift in the degree to which education research is hosting data-intensive studies inspired by methodological innovations from computer science and statistics. We refer to the research leading this shift as "educational data science" (EDS). This topic captures several interrelated areas of growth. Consider first the rise of education data mining (EDM) and learning analytics. EDM arose as a community around 2005 from work around cognitive tutors and predictive modeling that could furnish fine-grained data on student activity (Baker & Yacef, 2009; Piety et al., 2014). EDM notions of prediction methods, structure discovery methods, relationship mining, model distillation, and distillation of data for human judgment were absorbed, and in some cases extended from contemporary computer science and cognitive science research (Baker, 2013). Learning analytics emerged 6 years later with its own conference (Siemens, 2013), focusing more broadly on the implications of a digital world on learning, often studying how nascent data from digital systems (Siemens et al., 2011) could be used to describe or facilitate aspects of the learning process or shed light on digital environments. The rise of MOOCs and use of Learning Management System data generated from large student samples across the globe gave another boost to the Learning Analytics community.

We conceptualize EDS as capturing a broader array of data and methods than related in prior review articles. While EDM and learning analytics have been closely associated with developmental and psychological studies in education, recent trends in EDS have seen broader application. EDS has started to play a role in the study of higher order topics like policies and organizations while also seeing continued application in the study of high-stakes test scores and fine-grained reading measures. Modern conceptions in this domain are emerging around demographics and examples of student work (Reardon & Stuart, 2019). Likewise, curricular studies and teacher education are gaining from AI (artificial intelligence)–enabled dashboards where a teacher can apply dynamic feedback to instructional practice or decisions (Rosenberg et al., 2020). And researchers in the United States have studied large scale student friendship networks and their implication for health risk behaviors (Harris, 2013) and racial segregation (Currarini et al., 2010). These extensions of data science into educational topics, though nascent, are quickly gaining breadth in terms of data types used, and increasingly offer more opportunities and avenues to understand educational structures and processes.

In a narrow sense, one could conceptualize EDS as the application of tools and perspectives from statistics and computer science to educational phenomena and problems. But we argue for a more expansive definition where EDS is an umbrella for a range of new and often nontraditional quantitative methods (such as machine learning, network analysis, and natural language processing) applied to educational problems often using novel data. We explore and emphasize this combination of novel data and/or methods by further discussing the kind of EDS research that are enabled by these emerging possibilities below.

*Novel Data.* Before the rise of EDS, quantitative research in education focused on administrative data regarding, for example, course enrollment and outcome summaries and longitudinal data from summative student assessments. Recent years have seen a rise in novel data, and one key feature of this novel data is that it is frequently "unstructured" or nontrivial to structure in a relational form. Large volumes of text data, clickstreams, interactions, videos, and audio recordings, for example, can now be incorporated into computational models to spotlight trends at an unprecedented speed and scale. The release of this unstructured data is crucial as it often forms the vast majority of data available from any organization, with proportions running as high as 80% (Shilakes & Tylman, 1998).

Fischer et al.'s typology of micro, meso, and macrolevel data is one framework for understanding unstructured data (Fischer et al., 2020). Microlevel data have a temporal component associated with it. It is available from MOOCs, simulations, games, and intelligent tutors wherein fine-grained interaction data with closely tied learner interactions can capture individual data from large samples of learners. Clickstream data logs are a common example. Mesolevel data have a limited temporal component but add more depth toward assessing learners' cognitive abilities, social traits,

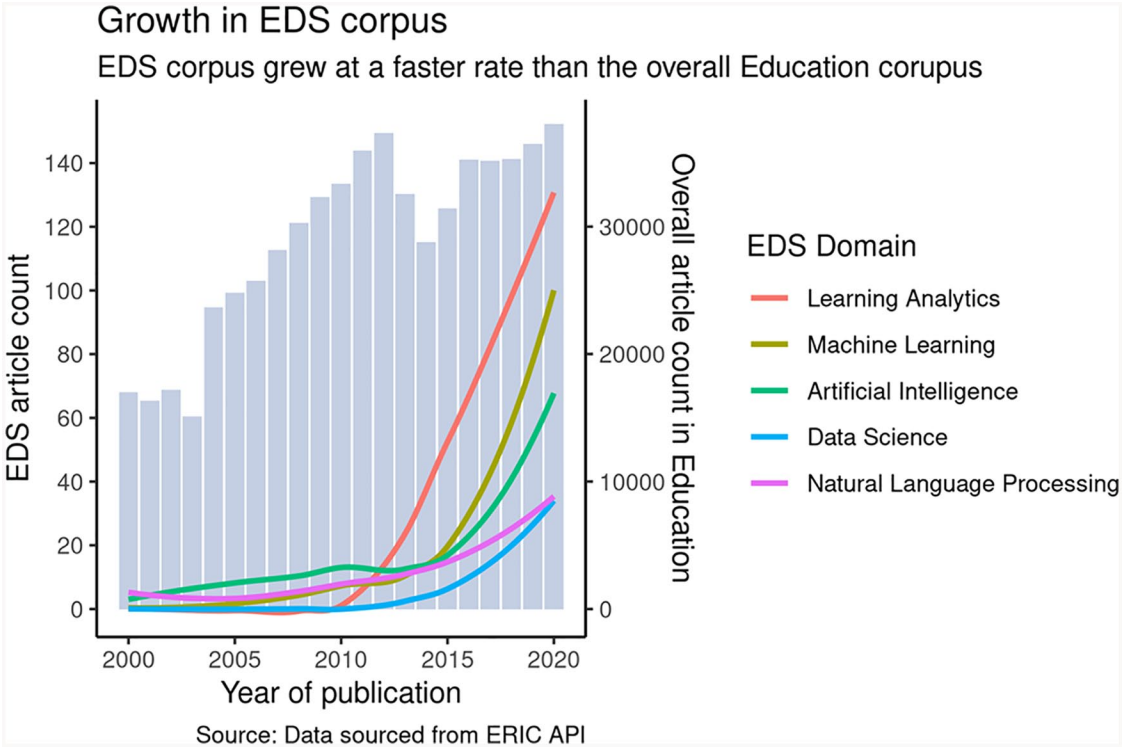


FIGURE 1. *Growth in education data science (EDS) corpus relative to education corpus. The EDS corpus has grown more than 30-fold in the past decade. The concave curves show that the rate of growth is increasing as well.*

and sustained relations. Multiple manifestations of text data, such as those generated from social media, MOOC forums, or digitized transcripts (all largely fed into natural language processing methods), fall into this domain, as do reports of friendship, affiliation, and other social network features. Macrolevel data sources work at the institution level and could be seen as a watershed between the structured data sets of the late 1990s and the unstructured, finer grained data generated from dynamic digital platforms. It can not only include static information, such as student demographics but also dynamic components, such as weekly attendance, engagement, and achievement scores (Fischer et al., 2020).

We are optimistic about the growth of novel data sources as EDS grows. Audio and video data can also be analyzed (beyond just the transcribed text) for embedded tone and body language, which could then be compared with findings from the transcribed text alone. We are also witnessing efforts within EDS to link multiple data sets together so as to inform research questions across domains (McFarland et al., 2015) or to identify domain similarities across digital learning platforms (Li et al., 2021). Reardon has linked educational achievement data with income tax data from the IRS to spotlight trends around achievement and income inequality in the United States (Reardon, 2016). Figlio and Lucas (2004) connect data from school report cards and housing markets to understand whether school grades affect families' residential locations and house prices. Research in higher education has also begun to use AI to link course transfer pathways across institutions (Pardos et al., 2019), link large-scale information on faculty research outputs to grants and patents (H. Cao et al., 2020; Manjunath et al., 2021), and the support of graduate training via grants to ensuing labor market returns in the wider economy (Weinberg et al., 2014). There are now also possibilities for researchers to combine data collected from firsthand surveys with other data that are open-sourced or otherwise available for public use. There is ample diversity in these public data sources—such as the College Scorecard (2013) data from the U.S. Department of Education, cross national data from the Programme for International Student Assessment (PISA; OECD, 2019), and student data from FreeCodeCamp (2014). These data linkage trends are especially promising and differ from the usual focused study of one dataset at a time (which has been a long-standing tradition in computer science). There is more focus on finding parallel findings across several corpora or linkages across them (and inferences therefrom). In the future, and in line with the train-test split analogy from machine learning, it might be interesting to see how well a big data model trained on one corpus performs in terms of outcomes predicted in a different corpus.

*Novel Methods.* The rapid growth of EDS is correlated with the growth in applications of "Machine Learning." Among machine learning algorithms, supervised learning algorithms

have found extensive application in the field of education (Dhanalakshmi et al., 2016; Mohammadi et al., 2019; Olivé et al., 2020). A supervised learning algorithm consists of a dependent variable that is being predicted from a given set of independent variables. Machine learning generates a function that maps inputs to desired outputs using both variable sets. Example algorithms include regressions, applications of deep neural networks, k-nearest neighbors, decision trees, and random forest. Unsupervised learning algorithms have also been applied (Liu & d'Aquin, 2017; Sathya & Abraham, 2013; Zhang et al., 2017). Since we do not have a dependent variable here, the approach often involves "clustering" a given population into similarity-based groups. Reinforcement learning has also been applied to education in multiple works (Bassen et al., 2020; Iglesias et al., 2009; Park et al., 2019; Doroudi et al., 2019). The approach involves exposure to an environment where the algorithm trains itself through multiple trial and error iterations, akin to what we see in a Markov Decision Process.

We see two clear factors contributing to the outsized growth of machine learning in education. First is the aforementioned explosion in textual and sequential data necessitating appropriate methodological approaches to accommodate their analysis (Pardos, 2017). With the rising popularity of MOOCs and e-learning, large volumes of data are produced through teaching and learning activities in online courses (Kizilcec & Brooks, 2017). These large data volumes are essential for performance gains through machine learning algorithms. These large data sets also establish much-needed test beds for interventions needing substantial training data and assessing predictive accuracy for machine learning algorithms. Second, the blurring of disciplinary boundaries has enabled talented computer scientists and statisticians to apply their data science skills toward addressing pressing societal challenges. Universities have further nurtured these trends with faculty positions and establishing Fellowship programs, such as CS + Social Good and the Data Science for Social Good Fellowships at several American universities at both the undergraduate and graduate levels.

Natural language processing (NLP), as a subdomain of machine learning, warrants particular attention. It has been extensively applied to text data in education settings—either collected firsthand or transcribed from media recordings. Broadly, NLP techniques can be applied to large text corpora in education to understand traits like sentiment embedded in the text, the novelty in information presented, and topics identified by topic modeling and topic classification approaches (Islam et al., 2012; Lucy et al., 2020). In a classroom setting, NLP algorithms can dynamically assess reading proficiency for a student and generate real-time feedback for improvement (Li et al., 2017). Modern NLP algorithms are being used to provide actionable feedback around prose, grammar, and general writing mechanics (Alhawiti, 2014; Shum et al., 2016) and to allow for examinations of, for example, the potential signature of class in written elements of educational materials (Alvero et al., 2021). In addition to a student-facing component, NLP platforms can provide a teacher-facing component as well. This helps enable teachers conduct robust formative assessments that might otherwise be difficult in classrooms with large student-teacher ratios (Burstein et al., 2014; Chapelle & Chung, 2010).

Another novel class of methods enabled by EDS centers around social network analysis. The medium of education is communication, and such communication forms social relationships, which are influenced in turn by established relationships, driving the interpersonal behaviors and attitudes of school participants. Network analytic methods are a means of representing these relations, interactions, and the interpersonal influences arising within educational settings and online platforms. Network analysis has long focused on the direct relations among education stakeholders—such as students in schools (McPherson et al., 2001), classrooms (McFarland et al., 2014) or lunchrooms (Moody, 2001), or teachers (Hawe & Ghali, 2008; Shaffer et al., 2009) but has recently been scaled up via new computational methods like node2vec, which reduce the complexity of interpersonal association to *n*-dimensional spaces (Grover & Leskovec, 2016). In addition, a fleet of inferential statistical methods have been developed to predict and model networks both as direct ties and as affiliational structures. Many of these new methods were developed with school and classroom data as their test cases, and they can be found extensively related in issues of the journal *Social Networks* (Cranmer et al., 2020). Some of the methods (stochastic actor-oriented models (SOAMs); Snijders, 1996) are even able to disentangle selection mechanisms from influence mechanisms, and identify, for example, whether improved grades arise from association with high-achieving friends, or if good students find high-achieving friends (Snijders, 2002; Stadtfeld et al., 2019). Such an approach has the potential to answer whether learning arises from social "pushes" on students or their own decisions to "jump." In general, social network methods appeal to ecological views of learning (Barron, 2003), and seem well adapted to relational databases and information on affiliations and interactions commonly represented in organizational records and web platforms.

In sum, novel data and novel methods have emerged as mutually reinforcing forces fueling the rapid growth of EDS. As more and more sources of unstructured data become accessible to humans, we expect to see methods in EDS evolve rapidly to match those challenges.

## Present

We now discuss 12 articles from the special topic in education data science that exemplify work in EDS. The set of works present theoretical, descriptive, predictive, and causal arguments, and they span levels from micro-interactions to

macrotrends. The *AERA Open* special topic offers a distinct perspective on education data science in comparison to prior summaries. Prior reviews emphasize the relevance of learning analytics and education data mining from either computer science, data science, or learning science angles (Fischer et al., 2020; Piety et al., 2014). Related works like that of Rosenberg et al. (2020) extend the view to include teacher education's concern with data science education, and Reardon and Stuart (2019) extend it to include the perspective of education policy in particular.

Many of the articles published in the *AERA Open* special topic focus on the mining of text data (via natural language processing) so as to better understand the variable success of experiments and policy implementation efforts. Other work uses digital technologies like web platforms and smartphone logs to acquire new forms of information. Such efforts serve to reveal previously hidden processes like considerations and interpretations that are integral to educational processes. Education scholars in this issue tend to focus more on macrolevel qualities of educational systems, like public opinion about reforms, and less on microlevel aspects of clicks and utterances focused on in learning analytics research. Last, all the efforts to predict individual outcomes via machine learning or other means are generally circumspect, cautious, and critical, and reflect a more mature and nuanced version of data science that is well aware of how complex educational phenomena can be. In general, the *AERA Open* special topic on education data science presents a line of heterogeneous research that veers more toward social and policy issues than learning science and individual learning seen in other journals; leans more on description and explanation than prediction; and offers more critique and the accomplishment of ethical goals than the creation of new algorithms and tools that can predict behavior or pressure it in certain directions.

Notably, any missing topics are likely the result of there being publication outlets already available in other subfields. Social network studies of education can often be found in *Social Networks* or mainstream sociology journals or say the *Journal of Adolescence*. Likewise, articles on learning analytics and education data mining can be found in conference proceedings of *International Conference on Learning Analytics and Knowledge (LAK)*, the *International Conference on Educational Data Mining (EDM)*, the *International Conference on Artificial Intelligence in Education*, the *ACM (Association for Computing Machinery) Conference on Learning at Scale*, the *International Journal of Artificial Intelligence in Education*, the *Journal of Educational Data Mining, IEEE Transactions on Learning Technologies*, the *Journal of Learning Analytics, Workshop on Innovative Use of NLP for Building Educational Applications* (BEA), and so on. *AERA Open*'s special topic is in many ways an outlet for education researchers and draws submissions from those in education departments and schools or from scholars eager to engage in conversation with education researchers more directly.

In what follows, we lay out five themes that summarize the main contributions of the 12 articles of the EDS special topic. However, please note that our brief summaries do not fully capture the contributions of the full articles. Careful reading of the individual articles will yield additional insights and inspirations the limited space afforded here did not permit.

### *Theory*

Shayan Doroudi (2020) in his article, "The Bias–Variance Tradeoff: How Data Science Can Inform Educational Debates", contributes a theoretical argument on how data science thinking can inform some of education's central debates and dualisms. In particular, a key concern of machine learning—the bias–variance tradeoff—is offered as an analogy for many vexing problems in educational research. On the one hand, much of education research emphasizes situated, or highly contextualized and rich qualitative cases, and on the other, it offers coarser, more generalized accounts using quantitative methods and large samples. The former approach tends to develop accounts that more closely fit the specific case and are less "biased" but have higher "variance" in that the specific examples may be poorly situated for understanding and prediction of behavior in other contexts. Conversely, the latter approach often adopted by quantitative scholars describe more cases relatively well (low variance), but they do not describe every specific case very well (higher bias). Doroudi's article argues that it helps reinterpret these persistent dualisms (and paradigm wars) via the bias-variance tradeoff. By seeing it as a "tradeoff" between competing inferential efforts, we can perhaps diminish divisions in education and adopt less of an essentialist and exclusive approach to research.

### *Data Mining Education Corpora*

A second theme is the application of natural language processing to large education corpora to reveal "hidden" patterns of language use. In the case of Lucy Li, Dora Demszky, Patricia Bromley, and Dan Jurafsky (Li et al., 2020; "Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks"), the authors study the language used within 15 U.S. history textbooks of the state of Texas (2015–2017). The work uses several methods (topic models, lexicons, and word embeddings) to identify the prevalent topics of textbooks, the actors discussed (gender, ethnicity) and their characterization (as passive or active via lexicons), and what sorts of contexts they are related to (embeddings). In so doing, the work reveals that history texts may be implicitly biased against and insensitive to today's students and their backgrounds. Through their choice of language, these texts may be inadvertently naturalizing historical inequities minority groups experience. As

such, the authors present a variety of means by which future education researchers can reveal hidden patterns of language usage and meaning so that more awareness can be had of the implicit narratives present in historical accounts.

In the article by Ha Nguyen and Jade Jenkins (2020; "In or Out of Sync: Federal Funding and Research in Early Childhood", NLP is applied to a large sample of nearly 16,000 articles and grants available in digital archives that reflect research on early childhood. Through the use of topic models, they identify the key topics of this subfield and those that emerge in grants before publications. In so doing, they potentially identify how federal funding motivates and catalyzes scholarly production and likely has strong effects on scholars' careers. This work is consistent with recent work using topic models to identify themes in math education (Inglis & Foster, 2018) and education research more generally (Munoz-Najar Galvez et al., 2020). Prior articles like these sought to develop greater field consciousness and critique by making collective intellectual pursuits visible via topic models and topic trends. Nguyen and Jenkins extend this perspective by revealing how research funding drives publishing. This line of work promises to make the sociology of knowledge a more immediate, reflexive, and critical research activity for entire knowledge domains in the years to come.

Social media also has information of relevance to schooling. Recent work finds that online school report cards can create disparities in housing values and reproduce social segregation (Hasan & Kumar, 2019). Likewise, the article on this special topic by Nabeel Gillani, Eric Chu, Doug Beeferman, Rebecca Eynon, and Deb Roy (Gillani et al., 2021; "Parents' Online School Reviews Reflect Several Racial and Socioeconomic Disparities in K-12 Education") focuses on the public written reviews that parents make on school rating websites as a means to understand how parents are making subjective assessments of quality. The authors use NLP methods (bidirectional encoder representations from transformers, BERT) to study textual snippets in half a million parent reviews concerning 50,000 K–12 public schools and identify those most associated with school characteristics (race, class, and test score) and school effectiveness (test score gains). They find that urban and affluent schools get more reviews, that review language correlates with test scores (and race and income) rather than improvement in test scores (i.e., effectiveness), and that reviews reflect racial and income disparities in education. As such, subjective online reviews by parents may be reproducing and reaffirming biased perspectives and achievement gaps. In short, the reproduction of educational inequality is performed not just institutionally by powerful leaders and elites but also by parents online.

Similarly, the article by Joshua Rosenberg, Conrad Borchers, Elizabeth Dyer, Daniel Anderson, and Christian Fischer (Rosenberg et al., 2021; "Understanding Public Sentiment About Educational Reforms: the Next Generation Science Standards on Twitter") explores the effects of public sentiment on education reforms. Whereas the Gillani et al.'s (2021) article uses half a million parental posts on greatschools.org, Rosenberg et al. (2021) focus on the sentiments expressed in 656,000 Twitter posts in relation to the Next Generation Science Standards (NGSS). As many reform scholars note, public buy-in to reforms (and teacher buy-in) is essential for reforms to be effective. To study such buy-in, Rosenberg et al. draw on social media posts and use sentiment analysis (and Machine learning methods for automated classification of manual coding) to ascertain how NGSS is being publicly perceived. They find that the NGSS is receiving increasingly positive support and contrast their findings with opinion polling on the Common Core State Standards. Such an approach to studying social media posts may prove a fruitful means to gauging public buy-in to educational reforms, and in an economical, scalable fashion.

### *Insight Into Interventions*

Another theme focuses on experiments and treatments employed in education research. This theme focuses on the use of data science to reveal why a treatment had varied success and reception, or how well a reform was implemented. The study by Nia Dowell, Timothy McKay, and George Perrett (Dowell et al., 2021; "It's Not That You Said It, It's How You Said It: Exploring the Linguistic Mechanisms Underlying Values Affirmation Interventions at Scale") uses NLP to identify features of essay interventions that make them successful in certain conditions and for certain disadvantaged groups (Jiang & Pardos, 2021). In particular, they look at value affirmation (VA) writings (a stereotype threat intervention) as a means of ameliorating gender disparities, and they identify various features of VA writings (using Coh-metrix to identify ideational and referential cohesion in texts, and LIWC [linguistic inquiry word count] to identify dictionary-based sets of terms reflective of affect, cognition, etc.) that distinguish successful from unsuccessful writing interventions. In addition, they ask what language and discourse features differentiate between successful male and female VA. With their array of NLP-derived features, they find certain combinations of language features—or principal components—best characterize their usage and best predict treatment/control conditions. In so doing, the work helps researchers learn why their psychological interventions have greater or lesser returns and especially for different groups (e.g., due to qualities of essay cohesion, affect, cognitive state, and social orientation).

The article by Lovenoor Aulck, Joshua Malters, Casey Lee, Gianni Mancinelli, Min Sun, and Jevin West (Aulck et al., 2021; "Helping FIG-ure It Out: A Large-Scale Study of Freshmen Interest Groups and Student Success") studies the impact of freshmen seminars on 76,000 University of

Washington students over 22 years. They use propensity scores to find matched samples of freshmen who were in the freshmen seminar (or freshmen interest groups, FIG) versus those who were not, and to see if the seminars have a positive effect. They find the seminars do have a positive effect on retention, especially for underrepresented minorities who are most likely to drop out. They then look at 12,500 open-ended survey responses from these students, use latent Dirichlet allocation (LDA) to identify topics or thematic resources mentioned, and use those to guide qualitative coding of specific resources FIGs' afford. Similar to the Dowell et al. (2021) article, they use data science tools to discern why an experimental condition has its observed effects (e.g., the seminars offer integration, belonging and information).

Whereas the Dowell et al. (2021) and Aulck et al. (2021) articles use NLP methods to understand how and why a treatment is effective or not, the article by Kylie Anglin, Vivian Wong, and Arielle Boguslav (Anglin et al., 2021; "A Natural Language Processing Approach to Measuring Treatment Adherence and Consistency Using Semantic Similarity") uses NLP to determine whether an educational reform is being delivered with fidelity. The work argues that the success of most reforms and interventions depends on whether it is actually implemented as proposed. The authors used data from five different randomized control trials to study mixed reality simulated classroom environments that entail coaching sessions (of ~100 persons each). They study the transcripts of video-taped coaching sessions and use simple NLP metrics (cosine similarity of word vectors across the protocol and the coaching session dialogue) to generate measures of intervention adherence and replication. They then ascertain the validity of their adherence and replication measures by seeing how well they correspond with survey responses on adherence and replication. They argue such NLP-based metrics could be used to determine how well reforms are being adhered to and replicated over time, and that such measurement may help education reformers better determine if implementation is to blame for the success or failure of education reforms.

A final article uses NLP to formatively assess an educational intervention. The article was written by Joshua Littenberg-Tobias, Justin Reich, and Elizabeth Borneman (Littenberg-Tobias et al., 2021; "Measuring Equity-Promoting Behaviors in Digital Teaching Simulations: A Topic Modeling Approach"), and concerns a large online course of 965 students that uses simulations to educate participants in diversity, equity, and inclusion. The study uses structural topic models (STM) to ascertain how attitudes and behaviors of the participants shift over the course of different equity simulation scenarios. In particular, the STM allows them to hypothesize and test whether persons with different equity attitudes converge over time. They find through this approach that desired attitudes and behaviors on diversity, equity, and inclusion do converge for all groups and toward the defined goals. So again, much like the prior work, data science is being used to formatively assess the effects or returns of various treatments. In this manner, traditional experimental research is augmented and advanced in useful and exciting ways.

### *New Data*

A fourth theme concerns the collection of new forms of data via new technologies like Web platforms and smartphones. From the recorded information on these platforms and the logs of these phones, insights regarding behaviors may become available that had heretofore been hidden due to the lack of appropriate data. For example, the article by Sorathan Chaturapruek, Tobias Dalberg, Marissa E. Thompson, Sonia Giebel, Monique H. Harrison, Ramesh Johari, Mitchell L. Stevens, and Rene F. Kizilcec (Chaturapruek et al., 2021; "Studying Undergraduate Course Consideration at Scale") develops a Web-based platform for college course exploration at Stanford. The clickstream data collected on this platform allows them to not only identify the courses that 3,336 Stanford freshmen took in 2016–2017 but also the courses they viewed or considered. The viewed set consists of only an average of nine courses (<2% available), but the authors interestingly find the narrow set of course considerations predict student majors 2 years later and net of their course taking. As such, data science can help us acquire new data and identify new mechanisms defining educational careers.

A second article, by René Kizilcec, Maximillian Chen, Kaja Jasińka, Michael Madalo, and Amy Ogan (Kizikcec et al., 2021; "Mobile Learning During School Disruptions in Sub-Saharan Africa"), also uses technology to acquire new data and new treatments. This work investigates how mobile learning technologies helped offset social disruptions to schooling for over 1.3 million students in sub-Saharan Africa. When schools were disrupted by violence during election cycles, the communities heavily relied on curricula and quizzes sent via smartphones to students, so as to sustain educational delivery in uncertain times. The reliance on big data and the log data of smart phones is what establishes this work as innovative education data science. Its application to more strained regions and populations (at scale) seems especially promising for education data science research going forward.

### *Critique of Machine Learning Predictions*

Were scholars to look at typical computer science outlets for data science articles, they would find a prevalence of machine learning models and efforts to predict various attitudinal or behavioral outcomes. Such a concern is less prevalent in the *AERA Open* special topic on education data science, but when and where it did arise, authors tend to be

more circumspect and critical, noting such approaches are fallible. Such caution, for example, can be found in the Fragile Families Data Challenge (Salganik et al., 2020) and prior efforts at predicting student achievement (Davidson 2019). Both efforts resulted in modest predictive power and were critical of efforts to predict complex social outcomes like poverty and student achievement. Often a simpler model does as well as a complex one, and usually, obvious factors play outsized roles (e.g., path dependence). The article exemplifying this in our special topic is written by Kelli Bird, Ben Castelman, Zachary Mabel, and Yifeng Song (Bird et al., 2021; "Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education"). Their article attempts to predict college dropouts using a variety of metrics and machine learning approaches and calls for greater transparency and critique of them. The authors argue there are good reasons to predict college dropouts using data science approaches: Many institutions are already using them and basing institutional decisions on their findings, but the methods and models used are all too often proprietary and lack transparency. The public cannot tell what metrics were used, what methods were employed, and what issues and problems arose. By performing such analysis on students in 23 community colleges in the commonwealth of Virginia (Virginia Community College System, VCCS), Bird et al. bring to light these potential concerns. They find that different predictive methods highlight different features salient to dropping out of college, but they tend to have relatively similar results differing on 600 cases out of the 300K modeled, and often the simplest predictive modeling approach (e.g., logit) works nearly as well as the most complex and computationally taxing (e.g., random forest, XGboost, and neural network). In sum, the work makes the strong case that educators should perform and investigate machine learning models used to inform and guide institutional policies so they are able to critique and assist institutions in their ethical use.

## Future (or Where We Go From Here)

We end with prospective guidance as to what opportunities the emerging field of EDS could seize upon so as to have a larger beneficial impact on education. The first such opportunity is that the field should draw on the rich set of traditions that inform education research; in particular, humanistic and social science traditions (McFarland et al., 2015). These fields are engaging in a similar watershed where data science and big data are entering and revolutionizing their fields. Confluences there reveal tensions and successes that perhaps EDS can learn from. One potential approach for integration of data science and education is through "adversarial collaboration" (see previous discussion in a different context in Martschenko et al., 2019). In a similar vein, projects in educational data science should aim to incorporate the best of the methodological traditions inherent in other disciplines alongside their theoretical and conceptual traditions.

Epistemological methods of economists have at times found tension with those from the machine learning community in the arena of data mining. In higher education, however, EDS methods are beginning to provide new complementary perspectives on institutional and student-level data (Chaturapruek et al., 2021). As these data become more readily available and joinable, machine learning can be used to synthesize and make comprehensible the rich contexts students traverse throughout their postsecondary paths (Pardos et al., 2019) and open up new avenues of intervention using predictive models (Bird et al., 2021). The introduction of data science approaches need not mean fields jettison well-earned advances and established tenets, that is, this is not a case of interfield colonization. Case in point, experimental designs from the discipline of economics can sometimes be applied as the gold standard for evaluating the effects of EDS-informed interventions in higher education.

As another example, a century's worth of psychometric research offers numerous modeling approaches that might be worth utilizing alongside modern machine learning approaches. Psychometric approaches may be both informative for subsequent methodological advancement (in terms of the features that may merit attention) and also useful as benchmarks for new methods from data science. Such a perspective reveals that the gains from machine learning approaches are often relatively minor (if they exist at all). Another example is for data scientists to listen and learn from a century's worth of careful work on sampling and to consider whether their all too often "found data" reflects known populations and to what extent (McFarland & McFarland, 2015). Much of the "shock and awe" from data science can be tempered and rendered more useful when we remember the insights our own fields impart.

One crucial question is how to best train students to enter the field of EDS. From our vantage, there should be a clear focus on problems of relevance to education that are potentially tractable given the data we have on hand. Even though there has been a relative explosion of data in education, we still have far less data (i.e., data rich with many fields) than is available in other domains, and this might limit the applicability of the most sophisticated algorithmic approaches (Bird et al., 2021). We should also work to build an ethic of responsibility in students. While many in technology have adopted a "move fast and break things" ethos, we think such an attitude would be highly inappropriate given the nature of education (i.e., the diversity of stakeholders and the care needed when dealing with issues affecting young people). Rather, we need to be more like physicians with their Hippocratic mandate (first, do no harm). Issues of equity, for

example, cannot be considered at the end but rather need to be central from the outset.

We close by offering two notes of caution. The first pertains to the limits of EDS. The computational approaches emphasized in EDS are exciting in that they may offer new insights into old problems or allow for novel kinds of data and perspectives to enter education research. However, these data and approaches will not be a panacea. The limitations of computational techniques can be seen in the recent shortcomings of the "Fragile Families Challenge" (Salganik et al., 2020). In that project, the addition of rich data and sophisticated modeling techniques did not substantially increase the predictability of several life course outcomes of relevance in the study of young people. We think these results are useful in terms of setting expectations: Behavioral science in general and educational science in particular are challenging. Most innovations on the data or computational side should be anticipated to bring only marginal improvements in our understanding.

The second pertains to the insidious problem of bias in computational approaches and the need to work diligently toward fairness. Notwithstanding early references to possibilities of bias in computation (Friedman & Nissenbaum, 1996), detrimental effects have only been thoroughly studied in the past 5 years—a phenomenon highly correlated with the rise of automation and machine intelligence. We see four potential problems around bias in EDS approaches that researchers need to be mindful about (Zou & Schiebinger, 2021). First, a growing majority of digital algorithms are automated based on data generated from a training set of past users. As this population is often skewed in favor of socioeconomically advantaged populations, this ends up further marginalizing the knowledge generated by traditionally disadvantaged social groups. Second, there is a problem with the nature of prediction itself embedded in most data science algorithms. Predictions look to the past to make guesses about future events. In an unjustly stratified world, methods of prediction could project the inequalities of the past into the future. Third, many of today's algorithms function on an unprecedented speed and scale. Google Translate serves over 200 million users a day (Prates et al., 2020). Previously expensive, slow, one-to-one functions can now be automated to become cheaper, faster and serve much larger audiences. This certainly means more people can benefit from automated algorithms. But a biased translation system could serve well over 200 million biased queries a day (Prates et al., 2020). Fourth, a lack of human control over what goes inside machine algorithms is sometimes considered indicative of impartiality. This assumption is problematic as fairness is not inherent in any algorithm. It is rather a quality that has to be carefully designed for and maintained.

A lack of attention to fairness concerns in EDS can potentially cause harm to both representation (when algorithms reinforce the subordination of information along the lines of identity) and allocation (when algorithms allocates or withholds certain groups an opportunity or a resource; Crawford, 2017). As Susskind (2018) cautions in *Future Politics*, "if you control the flow of information in a society, you can influence its shared sense of right and wrong, fair and unfair, clean and unclean, seemly and unseemly, real and fake, true and false, known and unknown" (p. 143).

In sum, EDS offers opportunities for innovating and advancing education research. The *AERA Open* special topic on education data science reveals the perspectives and concerns held by a subset of scholars genuinely interested in direct engagement with education research, and it offers an overview of the wider suite of data and methods beyond *AERA Open* that constitute EDS. While we have some degree of optimism about the research questions that will be tractable given the affordances of EDS, we also acknowledge a need for a healthy dose of realism and caution. As new scholars enter this emerging research area and become EDS practitioners, they should do so with their eyes wide open so they can make the best use and application of these innovative approaches. The particular article presented in this *AERA Open* special topic are seen as exemplifying these opportunities and the cautious, realistic engagement with them.

## ORCID iDs

Saurabh Khanna https://orcid.org/0000-0002-9346-4896

Zachary A. Pardos https://orcid.org/0000-0002-6016-7051

## Note

1. ERIC is a comprehensive, internet-based bibliographic and full-text database of education research and information.

## References

Alhawiti, K. M. (2014). Natural language processing and its use in education. *International Journal of Advanced Computer Science and Applications*, *5*(12). https://doi.org/10.14569/IJACSA.2014.051210

Alvero, A. J., Giebel, S., Gebre-Medhin, B., antonio, a. l., Stevens, M. L., & Domingue, B. (2021). *Essay content is strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications* (CEPA Working Papers). https://cepa.stanford.edu/sites/default/files/wp21-03-v042021.pdf

Anglin, K. L., Wong, V. C., & Boguslav, A. (2021). A natural language processing approach to measuring treatment adherence and consistency using semantic similarity. *AERA Open*, *7*(1). https://doi.org/10.1177/23328584211028615

Aulck, L., Malters, J., Lee, C., Mancinelli, G., Sun, M., & West, J. (2021). Helping students FIG-ure it out: A large-scale study of freshmen interest groups and student success. *AERA Open*, *7*(1). https://doi.org/10.1177/23328584211021857

Baker, R. S. (2013). *Educational data mining: Potentials and possibilities* [Paper presentation]. American Educational Research Association Annual Meeting, San Francisco, CA, United States.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Barron, B. (2003). Interest and self-sustained learning as catalysts of development: A learning ecology perspective. *Human Development*, *49*(4), 193–224. https://doi.org/10.1159/000094368

Bassen, J., Balaji, B., Schaarschmidt, M., Thille, C., Painter, J., Zimmaro, D., & Mitchell, J. C. (2020, April). *Reinforcement learning for the adaptive scheduling of educational activities* [Conference session]. 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, United States.

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open*, *7*(1). Advance online presentation. https://doi.org/10.1177/23328584211037630

Burstein, J., Shore, J., Sabatini, J., Moulder, B., Lentini, J., Biggers, K., & Holtzman, S. (2014). From teacher professional development to the classroom: How NLP technology can enhance teachers' linguistic awareness to support curriculum development for English language learners. *Journal of Educational Computing Research*, *51*(1), 119–144. https://doi.org/10.2190/EC.51.1.f

Cao, H., Cheng, M., Cen, Z., McFarland, D., & Ren, X. (2020). *Will this idea spread beyond academia? Understanding knowledge transfer of scientific concepts across text corpora*. Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.158.pdf

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, *50*(3), 1–42. https://doi.org/10.1145/3076253

Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301–315. https://doi.org/10.1177/0265532210364405

Chaturapruek, S., Dalberg, T., Thompson, M. E., Giebel, S., Harrison, M. H., Johari, R., Stevens, M. L., & Kizilcec, R. F. (2021). Studying undergraduate course consideration at scale. *AERA Open*, *7*(1). https://doi.org/10.1177/2332858421991148

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

Clow, D. (2013, April). MOOCs and the funnel of participation [Conference session]. Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium. https://doi.org/10.1145/2460296.2460332

College Scorecard. (2013). https://collegescorecard.ed.gov/

Cranmer, S. J., Desmarais, B. A., & Morgan, J. W. (2020). *Inferential network analysis*. Cambridge University Press.

Crawford, K. (2017, December 4–9). *The trouble with bias* [Conference session]. Conference on Neural Information Processing Systems, Long Beach, CA, United States.

Currarini, S., Jackson, M. O., & Pin, P. (2010). Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences*, *107*(11), 4857–4861. https://doi.org/10.1073/pnas.0911793107

Davidson, T. (2019). Black-box models and sociological explanations: Predicting high school grade point average using neural networks. *Socius*, *5*. Advance online publication. https://doi.org/10.1177/2378023118817702

Dhanalakshmi, V., Bino, D., & Saravanan, A. M. (2016). *Opinion mining from student feedback data using supervised learning algorithms*. In *2016 3rd MEC international conference on big data and smart city (ICBDSC)* (pp. 1–5). IEEE.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73. https://doi.org/10.1145/2500499

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Doroudi, S. (2020). The bias-variance tradeoff: How data science can inform educational debates. *AERA Open*, *6*(4). https://doi.org/10.1177/2332858420977208

Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward? *International Journal of Artificial Intelligence in Education*, *29*(4), 568–620. https://doi.org/10.1007/s40593-019-00187-x

Dowell, N. M. M., McKay, T. A., & Perrett, G. (2021). It's not that you said it, it's how you said it: Exploring the linguistic mechanisms underlying values affirmation interventions at scale. *AERA Open*, *7*(1). https://doi.org/10.1177/23328584211011611

Escoufier, Y., Hayashi, C., & Fichet, B. (Eds.). (1995). *Data science and its applications*. Academic Press/Harcourt Brace.

Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review*, *94*(3), 591–604. https://doi.org/10.1257/0002828041464489

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160. https://doi.org/10.3102/0091732X20903304

FreeCodeCamp. (2014). *Learn to code at home*. from https://www.freecodecamp.org/learn

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gillani, N., Chu, E., Beeferman, D., Eynon, R., & Roy, D. (2021). Parents' online school reviews reflect several racial and socioeconomic disparities in K–12 education. *AERA Open*, *7*(1), 1–16. https://doi.org/10.1177/2332858421992344

Grover, A., & Leskovec, J. (2016). *node2vec: Scalable feature learning for networks* [Conference session]. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, United States.

Harris, K. M. (2013). *The add health study: Design and accomplishments*. University of North Carolina at Chapel Hill. https://doi.org/10.17615/C6TW87

Hasan, S., & Kumar, A. (2019). *Digitization and divergence: Online school ratings and segregation in America*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265316

Hawe, P., & Ghali, L. (2008). Use of social network analysis to map the social relationships of staff and teachers at school. *Health Education Research*, *23*(1), 62–69. https://doi.org/10.1093/her/cyl162

Iglesias, A., Martínez, P., Aler, R., & Fernández, F. (2009). Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Applied Intelligence*, *31*(1), 89–106. https://doi.org/10.1007/s10489-008-0115-1

Inglis, M., & Foster, C. (2018). Five decades of mathematics education research. *Journal for Research in Mathematics Education*, *49*(4), 462–500. https://doi.org/10.5951/jresematheduc.49.4.0462

Islam, Z., Mehler, A., & Rahman, R. (2012, November). *Text readability classification of textbooks of a low-resource language* [Paper presentation]. 26th Pacific Asia Conference on Language, Information, and Computation, Bali, Indonesia. https://aclanthology.org/Y12-1059.pdf

Jiang, W., & Pardos, Z. A. (2021). *Towards equity and algorithmic fairness in student grade prediction*. In B. Kuipers, S. Lazar, D. Mulligan, & M. Fourcade (Eds.), *Proceedings of the Fourth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (pp. 608–617). ACM.

Kizilcec, R. F., & Brooks, C. (2017). Diverse big data and randomized field experiments in MOOCs. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 211–222). Society for Leaning Analytics Research.

Kizilcec, R. F., Chen, M., Jasińska, K. K., Madaio, M., & Ogan, A. (2021). Mobile learning during school disruptions in sub-Saharan Africa. *AERA Open*, *7*(1). https://doi.org/10.1177/23328584211014860

Li, X., Dunn, J., Salins, D., Zhou, G., Zhou, W., Schüssler-Fiorenza Rose, S. M., & Snyder, M. P. (2017). Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLOS BIOLOGY*, *15*(1), Article e2001402. https://doi.org/10.1371/journal.pbio.2001402

Li, Z., Ren, C., Li, X., & Pardos, Z. A. (2021). *Learning skill transfer models across systems*. In N. Dowell, S. Joksimovic, M. Scheffel, & G. Siemens (Eds.), *Proceedings of the 11th International Conference on Learning Analytics and Knowledge* (pp. 354–363). ACM.

Liu, S., & d'Aquin, M. (2017, April 25–28). *Unsupervised learning for understanding student achievement in a distance learning setting* [Conference session]. 2017 IEEE Global Engineering Education Conference, Athens, Greece.

Littenberg-Tobias, J., Borneman, E., & Reich, J. (2021). Measuring equity-promoting behaviors in digital teaching simulations: A topic modeling approach. *AERA Open*, *7*(1). https://doi.org/10.23328584211045685

Lucy, L., Demszky, D., Bromley, P., & Jurafsky, D. (2020). Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open*, *6*(3). https://doi.org/10.1177/2332858420940312

Manjunath, A., Li, H., Song, S., Zhang, Z., Liu, S., Kahrobai, N., Gowda, A., Seffens, A., Zou, J., & Kumar, I. (2021). Comprehensive analysis of 2.4 million patent-to-research citations maps the biomedical innovation and translation landscape. *Nature Biotechnology*, *39*(6), 678–684. https://doi.org/10.1038/s41587-021-00940-5

Martschenko, D., Trejo, S., & Domingue, B. W. (2019). Genetics and education: Recent developments in the context of an ugly history and an uncertain future. *AERA Open*, *5*(1). https://doi.org/10.1038/s41587-021-00940-5

McFarland, D. A., Lewis, K., & Goldberg, A. (2015). Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist*, *47*(1), 12–35. https://doi.org/10.1007/s12108-015-9291-8

McFarland, D. A., & McFarland, H. R. (2015). Big data and the danger of being precisely inaccurate. *Big Data & Society*, *2*(2). https://doi.org/10.1177/2053951715602495

McFarland, D. A., Moody, J., Diehl, D., Smith, J. A., & Thomas, R. J. (2014). Network ecology and adolescent social structure. *American Sociological Review*, *79*(6), 1088–1121. https://doi.org/10.1177/0003122414554001

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. (2019, February 11–13). Comparative study of supervised learning algorithms for student performance prediction [Conference session]. 2019 International Conference on Artificial Intelligence in Information and Communication, Okinawa, Japan.

Moody, J. (2001). Race, school integration, and friendship segregation in America. *American Journal of Sociology*, *107*(3), 679–716. https://doi.org/10.1086/338954

Munoz-Najar Galvez, S., Heiberger, R., & McFarland, D. (2020). Paradigm wars revisited: A cartography of graduate research in the field of education (1980–2010). *American Educational Research Journal*, *57*(2), 612–652. https://doi.org/10.3102/0002831219860511

Naur, P. (1974). *Concise survey of computer methods*. Petrocelli Books.

Nguyen, H., & Jenkins, J. (2020). In or out of sync: Federal funding and research in early childhood. *AERA Open*, *6*(4). https://doi.org/10.1177/2332858420979568

OECD. (2019). *PISA 2018 assessment and analytical Framework*. https://doi.org/10.1787/b25efab8-en

Olivé, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2020). A supervised learning framework: Using assessment to identify students at risk of dropping out of a MOOC. *Journal of Computing in Higher Education*, *32*(1), 9–26. https://doi.org/10.1007/s12528-019-09230-1

Pardos, Z. A. (2017). Big data in education and the models that love them. *Current Opinion in Behavioral Sciences*, *18*, 107–113. https://doi.org/10.1016/j.cobeha.2017.11.006

Pardos, Z. A., Chau, H., & Zhao, H. (2019). *Data- assistive course-to-course articulation using machine translation*. In J. C. Mitchell, & K. Porayska-Pomsta (Eds.), *Proceedings of the 6th ACM Conference on Learning @ Scale* (L@S) (pp. 1–10). ACM.

Pardos, Z. A., Fan, Z., & Jiang, W. (2019). Connectionist recommendation in the wild: On the utility and scrutability of neural networks for personalized course guidance. *User modeling and user-adapted interaction*, *29*(2), 487–525. https://doi.org/10.1007/s11257-019-09218-7

Park, H. W., Grover, I., Spaulding, S., Gomez, L., & Breazeal, C. (2019, July). A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 687–694. https://doi.org/10.1609/aaai.v33i01.3301687

Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014, March 24–28). *Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems* [Conference

session]. Fourth International Conference on Learning Analytics and Knowledge, Indianapolis, IN, United States.

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with google translate. *Neural Computing and Applications*, *32*(10), 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

Reardon, S. F. (2016). *School district socioeconomic status, race, and academic achievement*. Stanford Center for Educational Policy Analysis.

Reardon, S. F., & Stuart, E. A. (2019). Education research in a new data environment: Special issue introduction. *Journal of Research on Educational Effectiveness*, *12*(4), 567–569. https://doi.org/10.1080/19345747.2019.1685339

Rosenberg, J. M., Borchers, C., Dyer, E. B., Anderson, D., & Fischer, C. (2021). Understanding public sentiment about educational reforms: The next generation science standards on Twitter. *AERA Open*. https://doi.org/10.1177/23328584211024261

Rosenberg, J. M., Lawson, M., Anderson, D. J., Jones, R. S., & Rutherford, T. (2020). Making data science count in and for education. In E. Romero-Hall (Ed.), *Research Methods in Learning Design and Technology* (pp. 94–110). Routledge.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., . . . McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences of the U S A*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., Mislevy, R., & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*, *1*(2), 33–53. https://doi.org/10.1162/ijlm.2009.0013

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38. https://doi.org/10.14569/IJARAI.2013.020206

Shilakes, C. C., & Tylman, J. (1998). *Enterprise information portals*. Merrill Lynch.

Shum, S. B., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016, April 25–29). *Critical perspectives on writing analytics* [Conference session]. Sixth International Conference on Learning Analytics and Knowledge, Edinburgh, Scotland, United Kingdom.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., Duval, E., Verbert, K., & Baker, R. S. j. d. (2011). *Open learning analytics: An integrated and modularized platform*. Open University Press.

Silver, N. (2020, August 23). What I need from statisticians. *Stats and Data Science Views*. https://www.statisticsviews.com/article/nate-silver-what-i-need-from-statisticians/

Snijders, T. A. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, *21*(1–2), 149–172. https://doi.org/10.1080/0022250X.1996.9990178

Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, *3*(2), 1–40.

Stadtfeld, C., Vörös, A., Elmer, T., Boda, Z., & Raabe, I. J. (2019). Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences of the U S A*, *116*(3), 792–797. https://doi.org/10.1073/pnas.1811388115

Susskind, J. (2018). *Future politics: Living together in a world transformed by tech*. Oxford University Press.

Weinberg, B. A., Owen-Smith, J., Rosen, R. F., Schwarz, L., Allen, B. M., Weiss, R. E., & Lane, J. (2014). Science funding and short-term economic activity. *Science*, *344*(6179), 41–43. https://doi.org/10.1126/science.1250055

Wu, C. F. J. (1986). Future directions of statistical research in China: A historical perspective. *Application of Statistics and Management*, *1*, 1–7.

Yuan, L., & Powell, S. J. (2013). *MOOCs and open education: Implications for higher education*. JISC CETIS. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.354.108&rep=rep1&type=pdf

Zhang, N., Biswas, G., & Dong, Y. (2017). *Characterizing students' learning behaviors using unsupervised learning methods*. In E. André, R. Baker, X. Hu, M. M. T. Rodrigo, & B. du Boulay (Eds.). *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 430–441). Springer.

Zou, J., & Schiebinger, L. (2021). Ensuring that biomedical AI benefits diverse populations. *EBioMedicine*, *67*. Advance online publication. https://doi.org/10.1016/j.ebiom.2021.103358

## Authors

DANIEL A. MCFARLAND is professor of education, and by courtesy, sociology and organizational behavior at Stanford University. His current research is focused on the sociology of science and knowledge innovation.

SAURABH KHANNA is a PhD candidate in education policy at the Stanford Graduate School of Education. His research spans algorithmic fairness and social networks in the context of education in developing nations.

BENJAMIN W. DOMINGUE is an assistant professor at the Graduate School of Education at Stanford University. He is interested in psychometrics and quantitative methods.

ZACHARY A. PARDOS is an associate professor of education at University of California, Berkeley studying adaptive learning and artificial intelligence. His current research focuses on knowledge representation and recommender systems approaches to increasing upward mobility in postsecondary education using behavioral and semantic data.