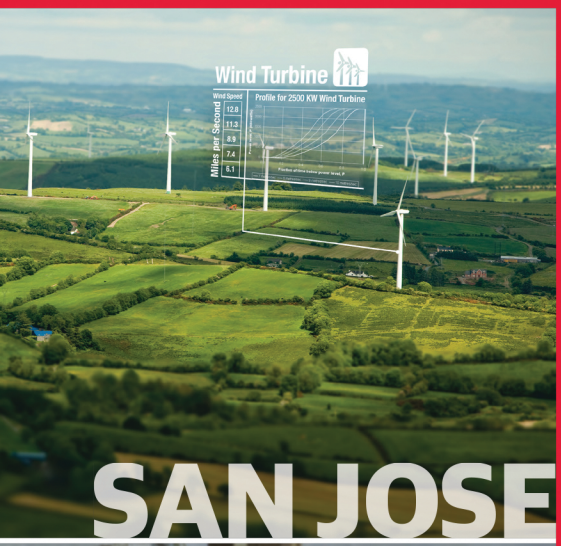
O'REILLY®

Educating Data

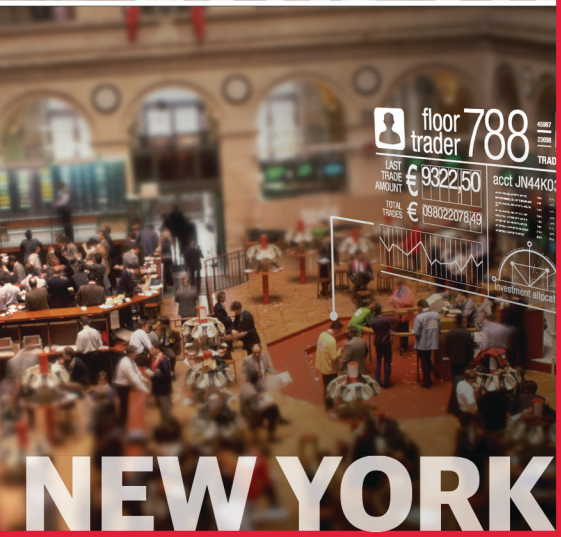
Using Data Science to Improve Learning, Motivation, and Persistence

Taylor Martin



SAN JOSE

LONDON



NEW YORK

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Educating Data

Using Data Science to Improve Learning, Motivation, and Persistence

Taylor Martin

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Educating Data
by Taylor Martin

Copyright © 2015 O'Reilly Media, Inc. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Tim McGovern
Production Editor: Dan Fauxsmith
Interior Designer: David Futato
Cover Designer: Randy Comer

September 2015: First Edition

Revision History for the First Edition
2015-09-01: First Release
2015-12-07: Second Release

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Educating Data*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-91893-7

[LSI]

Table of Contents

Educating Data

The use of large-scale and new, emerging sources of data to make better decisions has taken hold in industry after industry over the past several years. Corporations have been the first to act on this potential in search, advertising, finance, surveillance, retail, manufacturing, and more. Data is beginning to make inroads in the nonprofit sector as well—and will soon transform education. For example, GiveDirectly, an organization focused on managing unconditional cash transfer programs, and DataKind, an organization supporting data scientists who volunteer their time to social good projects, recently paired up to use data science to address poverty in the poorest rural areas in the world. They reduced the number of families that required face-to-face interviews by using satellite imagery, crowdsourced coding, and machine learning to develop a model that indicated villages most likely to be at the highest risk—based on the simple criterion of predominant type of roof in a village (villages with more metal than thatched roofs are at less risk). Education as an area of research and development is also moving in this direction. As Mark Milliron, Co-Founder and Chief Learning Officer of Civitas Learning, explains, "We've been able to get people from healthcare analytics or from the social media space. We have people who come from the advertising world and from others. What's been great is they've been so drawn to this mission. Use your powers for good, right?"

In this report, we explore some of the current trends in how the field of education, including researchers, practitioners, and industry players, is using data. We talked to several groups that are tackling a variety of issues in this space, and we present and discuss some of their thinking. We did not attempt to be exhaustive in our inclusion

of particular groups, but to explore how important trends are emerging.

The Promise

The promise of data science in education is to improve learning, motivation, persistence, and engagement for learners of all ages in a variety of settings in ways unimaginable without data of the quality and quantity available today.

Personalized Learning

Recommender and adaptive systems have been around for quite a while, both in and outside of education. Krishna Madhavan is an Associate Professor at Purdue and was a Visiting Research Scientist in Microsoft Research; he works on generating new visual analytic approaches to dealing with a variety of data—in particular educational data. He says, "The question is, there is a lot of work that has happened on intelligent tutors, recommender systems, automatic grading systems, and so on. So what's the big deal now?"

One answer is that, now, industry and research are developing personalized adaptive recommender systems around more open-ended, complex environments and information. Nigel Green is Chief Data Scientist at Dreambox Learning. They provide an adaptive learning platform for mathematics, primarily aimed at elementary school students. He describes it this way, "So many companies are looking at how many questions you got right or wrong. We actually care far less about whether the student gets the question right or wrong. We care about how they got their answer. That's the part that we're adapting on." Independent research studies comparing learning the same content with Dreambox's approach to other adaptive approaches have confirmed Green's idea.

Green's description of how they achieve this goal is important to understanding how personalized learning works.

As Green describes, every lesson in Dreambox targets small pieces of information or techniques, i.e., the knowledge students need to succeed in that area of mathematics. Dreambox calls these *micro-objectives*. For example, it is important that young students develop a basic understanding of numbers and what they mean. Green describes a task for younger grades this way, "Can you make the

number 6?" Dreambox assesses this with multiple smaller tasks that target the micro-objectives, for example dragging 6 balls into a shape and choosing the number six on a number line. Green says, "Those are two separate processes. And two separate questions that we're asking. One, can you actually move six balls in the right boxes, in the right order, and in the right locations. And then can you recognize the number 6 in the line below." In this way, Dreambox can assess each micro-objective separately. Following that, they can compare each student's state of knowledge to the average of all students their age who have completed the task, or the average of students who performed similarly to that student when they started using Dreambox, or the average of all students who are in remedial math. This allows them to direct students to the correct next task to maximize their learning. As Green says, "There are different ways of slicing and dicing those numbers. We want to make sure that we've got that student trending in their specific area. And then we can say, 'You know what, this student is taking nearly 2 standard deviations longer than the average student.'" If the student has done that repeatedly, Dreambox knows they haven't mastered the target content. At that point, they could increase the level of assistance provided to the student. In another case, a student might be taking so much longer than the average for their group that they will be unlikely to finish the lesson. At that point, Green says, "We might gracefully exit them out and take them to another lesson that practices content prior to or provides additional scaffolding for this lesson. In some cases, we move them sideways; we may have a lesson that's teaching or assessing exactly the same content in a different context. It may be that the student is not familiar with one context, or is more comfortable with one than another."

This is important for other businesses because so much of what we base the development of recommendation systems on is simplistic information. There are plenty of places where that is the right choice. In some cases, however (for example, in personalized health care), it may be better to follow Madhavan's and Dreambox's methods when developing algorithms and techniques, using more in-depth information to achieve an accurate picture. Piotr Mitros is Chief Data Scientist at edX, a provider of massive open online courses (MOOCs) in a wide range of disciplines to a worldwide audience. As he says, "The first course that we taught, a course was typically 20 hours a week, for about 14 or 16 weeks. That's a couple hundred hours of interaction. That is similar to video gaming compa-

nies, or to companies like Google perhaps. But it's not similar to most traditional industries where a person comes onto your website, interacts a little bit, and then leaves."

Another important way that what's happening now is different is that products build-in recommendation rather than just modeling what is likely to become a problem. Dave Kil, Chief Scientist at Civitas Learning, calls the latter the Forensic approach to education: "It's like we look at the patient and explain why he died, rather than analyzing and modeling the data to return it to the users in actionable form."

Another approach to personalized learning is tackling the entire college experience. Civitas Learning integrates many of the data sources that colleges and universities have available—e.g., Learning Management System (LMS) data, administrative data, and data on grades and attendance—creates predictive models based on the outcomes an institution has identified as most important to them—e.g., students graduating faster or more students passing introductory math courses—and then provides real-time feedback to students, instructors, and administrators to help the institution discover which interventions work best to reach those goals. They point to several important lessons they've learned along the way. One is the value of iterating until you get it right. Mark Milliron says, "some of our most exciting projects are projects that involve people testing trying, testing trying, testing trying until you really get that they've learned how to iterate." Civitas has had some of its best results with clients who pursue this sort of iteration. Another important lesson is not believing in the one-size-fits-all solution. As Milliron says, "Our second big challenge is really trying to solve the problem (for the college or university). A lot of people are trying to sell solutions they have developed—instead of solving a problem, they're trying to sell a solution."

Overall, current results are showing that personalized adaptive approaches are improving student learning and helping them navigate the complex world of college to graduate sooner and have a better probability of graduating. This area is likely to grow quickly as schools explore blended learning models and new companies pop up every year. These personalized adaptive approaches rely on being able to detect what students are learning on the fly, in real time, as they engage in learning activity. This leads to our next theme addressing automated assessment of learning.

No More Tests

There's no question that this goal is far off. However, it is exciting to think about the possibilities. "Wouldn't it be great if you could actually watch people do things and have some records of how they're actually doing them and relate what they're doing to the kinds of things they do or do not know?" as Matthew Berland puts it. Berland is a professor at the University of Wisconsin-Madison, researching learning from games and other engaging and complex environments. This is particularly exciting if it can be done in many of the evolving transformative learning environments such as games or even makerspaces.

The movement to reach this goal has been underway for some time, and there has been a lot of progress in the more structured environments Madhavan discusses. The struggle that presents, as Piotr Mitros, Chief Data Scientist at edX, points out, is that, "We're not yet really doing a good job of translating data into measurements of the types of skills we try to teach. We have some proxies for complex skills—such as answers to conceptual questions and simple problem solving ability—but they're limited. Right now, we have data on everything the student has done." With these data, Mitros and others hope to be able to find out more about complex problem solving, mathematical reasoning, persistence, and many skills employers mention as important, such as collaboration and clear communication while working on a team.

More open-ended environments present challenges in understanding the relationship between what people do and what they know. One challenge is capturing and integrating data. Clickstream data from environments is a common first step, but as Justin Reich from HarvardX, one of the partner institutions for edX, says, "You can have terabytes of information about what people have clicked and still not know a lot about what's going on inside their heads." In addition, Berland points out that, "There are missing aspects there (i.e., in clickstream data alone). Not least of which, what were their hands doing? What's going on with their face? What else is going on in the room?" It can be important to understand the context around the learning activity as well at what happens on the backend.

New efforts, such as Berland's ADAGE environment, aim to make these challenges easier. Berland says, "ADAGE is our backend system. It's something we agreed on as a way to format play data, live

play data. We also have an implementation of a server and a client system across formats like Unity, JavaScript, and a few others. The basic idea is 'Let's come to some common representations of how play data look. Then we have a set of tools that work with our ADAGE server, which is on our open-source software side of this."

It is exciting to think that more and more we will have the opportunity to directly assess what people know from what they do, rather than having to assess it by proxy based on their performance on tests. This will open up more possibilities for online learning and the wide deployment of complex engaging learning environments. We address this increasing access to learning opportunities next.

Access to Learning Opportunities

One of the greatest examples of the promise of big data for education is unprecedented access to learning opportunities. MOOCs are an example of a type of these opportunities. Organizations such as Udacity, Coursera, and edX offer courses ranging from Data Science to Epidemiology to the Letters of Paul, a Divinity School course.

As Justin Reich of HarvardX explains, "edX is a nonprofit organization that was created by Harvard and MIT. They provide a learning management system and then they create a storefront for courses on that learning management system and market those courses. So it's the individual university institutional partners who actually create the open online courses. HarvardX is one of those partners." The use of these courses has been significant. Reich says that, "In the past 2 years between Harvard and MIT, we've run 68 courses. They've had about 3 million people who've registered." Frequently, the image of these people has been either college students or people who already have a college degree. While this may be a largely true, Reich explains a more complex picture, "We now have an increasingly clear sense that in many of our courses many people already have a bachelor's degree; our median age is about 28. But we have people who are 13 years old. We have people who are octogenarians. And sometimes, even when groups are small percentages they can still be large numbers. So the about 30,000 of those users come from the UN's list of least-developed countries."

Mitros and Reich both described how many MOOCs are now attempting to incorporate features of the personalized and open-ended, complex learning environments discussed earlier. Building-

in tools to recognize what type of student a participant is critical here as so many different groups are participating in MOOCs. As Reich describes, currently, "Virtually every MOOC is sort of a one course for everyone kind of thing." He points out that this is problematic because it does not take into account multiple factors that we know affect learning, such as expertise in the course material, learner preferences, or learner goals in taking the course. To address the problem, Reich has been developing plans for a "recommender engine which would basically try to understand, for a particular course, what are not only the pathways of people who are successful but if we are to look at people across really important different dimensions; people who come in with high or low familiarity or high or low English language fluency. Then we ask, for different values of those characteristics, what are the pathways that successful people have taken? What do the most persistent learners do when they encounter difficulty or make errors? From that historical data and from what we know about how human learning works, can we recommend to folks what would be the effective strategies for when they get stuck."

At the same time, maintaining access while improving these courses is a key concern for MOOC providers. Reich: "There's a real tension between designing learning experiences that take full advantage of fast broadband access, and there's some cool things that you can build if you take advantage of that. But if you build those kinds of things, you may actually be cutting out many people in the world for whom watching a YouTube video or using a really complex simulation is a huge broadband cost. We need to think about that if we really want to serve people all around the world."

The promise of increasing access to a huge variety of learning experiences for lots of people and having many of those experiences adapt to learners' needs has great potential to transform education and improve outcomes for many students. Reaching this promise will involve addressing some key challenges however, and next we turn to those.

The Challenges

Privacy and Security of Student Data

Concerns over student privacy protection have increased in recent years. As an indicator of the times, just in the past month, Congress has introduced three new bills that address student data privacy. Some of this concern seems to be based on a lack of general knowledge about what is already clear from federal policies such as the Family Educational Rights and Privacy Act (FERPA). For example, all of those interviewed for this study have clear policies in place that follow FERPA standards. However, another problem is that FERPA was developed in the pre-digital age, and many feel that the guidelines do not address some critical issues for new types of data and new storage and analysis methods.

Part of the challenge involved in updating standards and practices around student data is technical. But a much larger problem is that what seems to be missing is a wide understanding of the value proposition of the promise of benefits of using data to improve outcomes for students, parents, schools, and teachers. It is the perspective of balancing usefulness of the data for helping people in huge ways and the real risks around data use and sharing. Governor Bob Wise is at the Alliance for Excellent Education, an advocacy and policy organization. For twenty years, their mission has been "that every child graduates from high school ready for college and career." They have been involved with high school reform and standards. Increasingly, as more educational products used in school are offered digitally, they have become involved in policy around student data. Governor Wise says that, "the best way to engage, at least for us, is to paint a basic picture of how a day in the life in a school room looks different when a teacher is using data effectively to benefit children—and the best messenger for that is the teacher saying 'This is how I use data.'" Overall, we are facing a cultural change in how we use data and how others use our data. This issue cuts across industries and sectors of the economy and is rapidly unfolding.

As folks on the more technical side of the issue point out, however, there is hope. There are technologies in place that have worked for other industries and groups. Ari Gesher is a Software Engineer and Privacy Wonk at Palantir and coauthor of the recent book, *The Architecture of Privacy* (O'Reilly). Palantir provides platforms for

integrating, managing, securing, and analyzing data at a large scale. He says, "the trade off between effectiveness of a system and the ability to preserve privacy is not an all-or-nothing proposition." He further asks, "What would it take to create an atmosphere where the anxiety around privacy risks is reduced to the point where data can be shared amongst institutions and researchers for the betterment of education while, at the same time, increasing the overall safety and privacy of the students about whom the data is recorded?" He points out that most data sharing in education currently is done by actually providing a de-identified copy of the dataset, but that we know that anonymization is fraught with issues. A better route is through providing access to data that the second party never actually has in their possession. Gesher: "Modern cloud-hosting environments are a cost-effective way to create such environments. An environment like this could not only include places to hold datasets, but also the analysis tools, instrumented for auditing, for people to work with the data." You need a combination of two things: you need access control to make sure that authorized people can see what they need to see, and that there are different levels of access rather than having an all-or-nothing model of access to data. On top of that, because any kind of access, even legitimate access, does represent a privacy risk, you need to have good auditing and oversight capabilities."

Data-Driven Decision Making

Providing a dashboard or other representation of what students are doing and learning is great, but what does a teacher or parent or student do with that information? That information is only helpful insofar as it can be used to affect critical outcomes. Building capacity for a variety of stakeholders to engage in data-driven decision making in education is a critical challenge to be met to fully realize the potential of big data for education.

Teachers

The perennial question is why educational technology has not taken off or taken hold at the level that it was trumpeted to do in the early days of wider Internet access and more computers in classrooms—and very often the answer is incredibly pedestrian: it takes too much time to get students going on any given computer-based instruction. Clever provides automated and secure log-in capability for students and teachers for Clever-enabled applications. Clever CEO Tyler Bos-

meny has an interesting approach to the question. Their answer was based in experience.

> When we actually went into the classroom, we saw what the day-to-day experience was like for teachers. And it blew our minds. Things that we take for granted or don't think about are actually huge impediments to using technology in the classroom. So, say a teacher wanted to use five different digital learning applications for their class, whether it's math or reading or another online assessment that meant that that teacher wouldn't have to enter all the data by hand. That teacher would have to register accounts for each of their thirty students in each of those five different applications. So they're spending hours and hours and hours just setting up applications so they can be used for the first time.

Bosmeny further points out that students change classes, schools, districts, and states and that, particularly for districts with high mobility, this causes a huge headache when the teacher has to be registering and then unregistering students frequently. Bosmeny: "Teachers told us, 'I feel like I'm a part-time data shuffler; just moving spreadsheets and uploading files; keeping the stuff up to date.'" This can be a huge drain on learning time in the classroom. Bosmeny: "Imagine a room of thirty second graders all in the computer lab trying to use software. They've got 30 different user names and passwords to manage and the teacher, instead of getting to spend time teaching, is spending a quarter of that class period just running around and helping their students get logged in."

Clever's solution to this problem is to do the integration for them. Bosmeny: "When an application is part of Clever we can integrate directly with the school student information systems which is where a lot of the information about students and classes lives inside a school district. And because of that we can automatically set up accounts for students in all of the different programs that their teacher wants them to use. So all of a sudden that process that teachers used to have to go through of downloading spreadsheets by hand and uploading them into different third-party applications, we've been able to completely automate that for them."

Beyond data integration for registration and logon, teachers are also required to perform as "human AIs," conducting data integration for inference. As Governor Wise points out, "Its one thing if you've got one dashboard, its something else if you've got four or five." This is an area that is just beginning to be explored. New data technologies are making it possible to bring educational resources to where they

are needed. MarkLogic, an Enterprise NoSQL platform provider, helps build systems that surface content in response to instructional needs for customers ranging from the textbook publishers to adaptive learning platforms. As Frank Rubino, Director of Solutions at MarkLogic, explained, "We integrate data across a variety of formats and from a variety of products, aggregate those data, and provide analytics to make meaning that teachers can act upon in the classroom."

Policy makers

Governor Wise points out, "there is a need for another group to understand the use of data; and that's the policy maker. Because at the end of the day, it's going to be that local school board member, that state legislator, even a member of congress reviewing FERPA or COPPA (the Children's Online Privacy Protection Act) that will make critical decisions that will affect the practitioner." There are some great examples around the country going on that show the power of data for helping policy makers make decisions. The Utah STEM Action Center, an organization housed in the Utah Governor's Office of Economic Development, aims to (1) produce a STEM (Science, Technology, Engineering, and Mathematics) competitive workforce to ensure Utah's continued economic success in the global marketplace; and (2) catalyze student experience, community engagement, and industry alignment by identifying and implementing the public- and higher-education best practices that will transform workforce development." As Jeff Nelson, Board Chairman of the STEM Action Center, says, "We have a legislative mandate to improve outcomes for Utah Students. The first legislation we have been working under has been focused on doing that by specifically using software interventions. What's been important about it is we could just be trying things, we could just throw some software at this and wait for end of year testing. But that isn't as strong as what we're doing." So far, they have contracted independent researchers to run several pilot studies comparing outcomes of students and teachers who receive the interventions the Center is funding to those who are not receiving these programs. As Nelson says, these results have been useful in policy-making situations. Nelson: "It's really been great, in fact we've been able to go to different interim committee meetings and show the data. Now it's interesting because in some cases the outcomes have been positive, in other cases there has been no difference versus the control group, but in all cases it's really

good information. We're in effect eliminating those things that don't work using the data and proving the things that do work." Nelson claims that this information was critical in the renewal of the legislation, as that process was largely based on data. Nelson: " I spent almost a whole day on the hill talking to legislators, telling them the story. We used a lot of that data that we had gathered for this purpose and so it was great to be able to say 'Hey look, here's what we're seeing, here are the students that are actually seeing some progress, and they're doing better than the control group.' And you know what, that's really good; that's the right conversation to be having. We all want the right outcomes for students and when the data can help you tell that story it's very meaningful." As Milliron from Civitas explains the phenomenon, "You have the leadership and cultural challenge of people being willing to leverage data."

Industry and academic researchers and developers

Educational data are often messy, not integrated, and come from many different environments. This presents challenges for making inferences about learning or engagement, providing recommendations based on these inferences, and helping teachers, policy makers, and others use the data to improve practice. It also presents challenges for training people to be data scientists working in education. There is no agreement on what would go into a program to train a data scientist in general. There is debate about whether we need separate academic departments or we just need to provide a summer program after someone finishes a PhD in say, Physics. The answer will probably be all of the above and somewhere in between. Currently, many people are learning on the job, whether it is in research or industry. In addition, people are leveraging partnerships to bring together content expertise and data science expertise. Berland: "So we have dozens of groups who want to take part in this. There are many people out there who don't have the background or don't know who to talk to. Part of what we're trying to do is just put people together who should be talking. Because there are parts that are really difficult and to some extent you do need people who know data analysis or map-reduce in some cases."

Conclusion

On the one hand, there is a sense of urgency about achieving results that will show the value of big data for education. As Dave Kil and

Krishna Madhavan mention, we need results that people understand and that show substantial value for educational outcomes soon, or we could be heading into the proverbial "Trough of Disillusionment" for any innovation. Madhavan: "Within the next 7 year window (because it takes a cohort about 4 years to graduate college), if we cannot establish that these methods can somehow bend the cost-curve towards spending less money on college and keeping people on time for graduation, we have lost the game."

On the other hand, as Piotr Mitros points out, "We're at the very early stages of using data to improve educational experiences and outcomes. Right now, most of what we're doing is at the level of what you would see in traditional web application and business analytics development where we can start to see data about what learners are actually using, where are students having problems, where students spend time. And we are just starting to do experiments, randomized controlled trials, in order to see what works better and what works worse."

To capitalize on the potential of big data, we need to go through the process of changing the field like other fields have. This involves (1) educating a new generation of data scientists in education, whether they are working in industry, teaching in or developing curricula for schools and universities, or in research; (2) building the infrastructure; (3) integrating the data sources; and (4) addressing the particular challenges education faces in the privacy area. Like all other fields that have made the jump, these changes take time.

Moving more slowly may just be okay though. Currently, education data really are not big data. Madhavan: "I would like to sort of also move away from this notion of big data. Educational data is not big. It's actually quite small and most of the time it's sitting in Excel spreadsheets. Think about just ten road sensors that are sitting on I-80 near Chicago. How much data do they collect versus how much data can you possibly collect in terms of all these tools working in a classroom or an educational setting? Just in a single day, in a few hours, they will dwarf the entire scale that we deal with. The complexity of the data comes from the format, the scale, or the dimensions. What I would try to move towards is this notion of smart data or multidimensional data."

About the Author

Taylor Martin is a professor of Instructional Technology and Learning Sciences at Utah State University. She researches how people learn from doing, or active participation, both physical and social. Particularly, Dr. Martin examines how mobile and social learning environments provided online and in person influence content learning in mathematics, engineering, and computational thinking, primarily using data science methods. Currently on rotation at the National Science Foundation, she works as a Program Officer for a variety of programs, including BIGDATA, Building Community and Capacity (BCC), DRK-12, STEM+C, Cyberlearning, and EHR Core Research. Dr. Martin focuses on a variety of efforts across the foundation to understand how big data is impacting research in education and across the STEM disciplines.