LEARNING WITH BIG DATA

THE FUTURE OF EDUCATION

VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER
authors of BIG DATA

LEARNING WITH

BIG DATA

Learning with BIG DATA

The Future of Education

VIKTOR MAYER-SCHÖNBERGER
and KENNETH CUKIER

AN EAMON DOLAN BOOK | HOUGHTON MIFFLIN HARCOURT
Boston *New York* 2014

www.hmhco.com

Book design by Melissa Lotfy

eISBN 978-0-544-35550-7

v1.0314

To our teachers and to our students
— V.M.-S. & K.N.C.

CONTENTS

1

DUSK

Dawa concentrates. he adds a bit of pigment to the tip of his brush. Then, with a careful stroke, he paints a thin black line. He does this again. And again. Slowly, as the hours pass, the thangka—a silk scroll-painting of the Buddha, with mesmerizing geometric detail—begins to take form.

Outside, the snow-covered summits of the Himalaya that surround Thimphu, the capital of the Kingdom of Bhutan, glisten in the late-afternoon sun. But inside, Dawa and his fellow students, all in their early 20s, in matching blue robes, have been focusing on their work under the watchful eye of their middle-aged instructor.

The training of thangka artists adheres to custom. Dawa and his fellow students are not there to have their minds broadened through education, but disciplined through apprenticeship. Learning is not about inquiry, but mimicry. Innumerable rules laid down centuries ago govern exactly what must be painted where and how.

Dawa's teacher makes sure the young artists follow his instructions precisely, to repeat what generations of thangka illustrators before them have done. Any deviation, any break from the rules, is not just frowned upon but prohibited. The best artist is the one who copies his master perfectly. The teacher constantly points out imperfections. But despite this immediate feedback, it is a form of learning that is largely devoid of data.

And it is a form of instruction that is fundamentally different to

how Andrew Ng, a computer scientist at Stanford University, teaches his class over the Internet on the topic of machine learning, a branch of computer science. Professor Ng (pronounced roughly as "Nnn") is a cofounder of Coursera, a startup company offering online classes. His approach is a harbinger of how big data is set to revolutionize education.

Professor Ng collects information on everything his students do. This lets him learn what works best and design systems that automatically parlay it back into his class: improving his teaching, his students' comprehension and performance, and tailoring education to everyone's individual needs.

For instance, he tracks students' interactions with his video lectures: when they watch them, if they press pause or fast-forward, or abandon the video before it's over — the digital equivalent of slipping out of class early. Professor Ng can see if they watch the same lesson multiple times, or return to a previous video to review material. He interlaces the video classes with pop quizzes. It's not to see if his charges are paying attention; such archaic forms of classroom discipline don't concern him. Instead, he wants to see if they're comprehending the material — and if they're getting stuck, exactly where, for each person individually.

By tracking homework and tests done on a computer or tablet, he can identify specific areas where a student needs extra help. He can parse the data across the entire class to see how the whole cohort is learning, and adjust his lessons accordingly. He can even compare that information with other classes from other years, to determine what is most effective.

It certainly helps that Professor Ng's classes teem with tens of thousands of students — so large that the findings he uncovers are statistically robust, not based on just a small number of observations, as most educational studies are. But the class size in itself is not the point. It's the data.

Already, he's tapped the data to extraordinary effect. For example, in tracking the sequence of video lessons that students see, a puzzling anomaly surfaced. A large fraction of students would progress in or-

der, but after a few weeks of class, around lesson 7, they'd return to lesson 3. Why?

He investigated a bit further and saw that lesson 7 asked students to write a formula in linear algebra. Lesson 3 was a refresher class on math. Clearly a lot of students weren't confident in their math skills. So Professor Ng knew to modify his class so it could offer more math review at precisely those points when students tend to get discouraged — points that the data alerted him to.

Another time, he saw that many students were repeating lessons on a certain topic. He literally saw this: he produced a data visualization in which the color intensity changed from dark blue to hot red when the statistical probability that a user progressed in the normal class sequence went out of kilter. Around lessons 75 and 80 something about the pattern was disrupted. Students were rewatching videos in a variety of orders. His takeaway: they were struggling to grasp the concepts. He realized that teachers armed with this insight could redo the lessons — and check the resulting data to make sure the situation improved.

A wealth of other data is tapped too. Online forum posts typically track how many people read them, and people are invited to rate them, to judge their usefulness. But Professor Ng ran a complex statistical study of his class forum posts to *really* judge their effectiveness. He looked at the percentage of students who, after getting a wrong answer related to a particular topic on a homework assignment or a test, upon reading a given forum post, produced a correct answer the next time they encountered the same question.

Thus, in a machine-learning class in 2011, thousands of students got an answer incorrect involving a "compute cost" in a linear regression. But those that read forum post number 830 had a 64 percent likelihood of correctly answering the question the next time they were presented with it.

From now on, the system can show that particular forum post to those students who get an answer on the topic wrong. It is a data-driven way to identify which forum posts actually work best for learning, not just which posts students judge to be the best.

And this big-data approach is not just restricted to Professor Ng's class at Stanford — this class is simply a front-runner of what is to come. Big data is invading all of education, with profound implications for how the world learns.

This e-book is about how big data changes education. Big data gives us unprecedented insight into what works and what doesn't. It is a way to improve student performance by showing aspects of learning that were previously impossible to observe. Lessons can be personally tailored to students' needs, boosting their comprehension and grades.

It helps teachers identify what is most effective: it doesn't take away their jobs but makes their work more productive, and probably more fun too. It helps school administrators and policymakers provide more educational opportunities at lower cost, important factors for reducing income gaps and social disparities in society. For the first time, we have a robust empirical tool with which to understand both how to teach, and how to learn.

This story is *not* about MOOCs, the "massive open online courses" like Professor Ng's at Stanford that have generated headlines in the past few years. The world has been captivated by the possibilities of these classes, which have democratized access to education. It is a wonderful development, to be sure. But in some respects, it is the same old education — "the sage on a stage" — only easier to access.

But there is one aspect of MOOCs that *is* new and powerful: the data they generate. The data can teach us what is most effective; it can tell us things we couldn't know before, since there was no way to unlock its secrets. But with big data we now can.

It helps that the marriage of education and technology is capturing the imagination of entrepreneurs and the wallets of investors. More than $1 billion in venture capital was poured into education in 2012 alone, a doubling from just five years earlier. In a sign that education technology has come of age, the industry is replete with its own arcane abbreviations, like LMS (learning management systems) and ITS (intelligent tutoring systems). Companies with cute names like Noodle, Knewton, and Knowillage Systems dot the landscape.

Old stalwarts like McGraw-Hill, News Corp., Pearson, and Kaplan have set up outposts in that territory too, having spent billions on research and development, as well as acquisitions. The e-learning market is estimated to be worth over $100 billion and growing by around 25 percent a year, according to GSV Advisors, a respected edtech market-research group. In the United States, spending on education overall is a hefty $1.3 trillion, or 9 percent of GDP, making it the second-largest area after health care.

Ultimately, this e-book is about more than education. At its core, it is about how one significant part of society and sector of the economy is adopting big data, as a case study for how big data is going to change all facets of life and of business. While here we will focus on the developments as they apply to education, the lessons are relevant to all industries, businesses, and organizations—be it a hospital, an oil company, a technology startup, a charity, or the military.

It also points at broader consequences for human knowledge—not just how we learn, but what we learn. Society must develop a deep understanding of the probabilistic nature of the world, not just the notion of cause and effect, which has permeated human inquiry throughout the ages.

So this book is intended as a guide for professionals of all stripes who are struggling to manage the epochal transition to big data that is now upon us. And it is for anyone who is interested in how people acquire knowledge in the big-data age.

In the next chapter, we consider three principal features of how big data will reshape learning: feedback, individualization, and probabilistic predictions. It looks at concepts like the "flipped classroom" popularized by the Khan Academy—where students watch lectures at home and do problem solving in class, the inverse of what's customary in traditional classrooms.

Chapter 3 considers the different platforms that are changing how we teach and learn, from online courses to e-textbooks. It delves into the idea of adaptive learning (in which the pace and materials are tailored to each student's individual needs) and learning analytics (which allows us to spot the most effective way to teach subjects). In

Chapter 4, we look at the potential dangers of big data in education, from worries over the persistence of data to its use in new forms of tracking, in which students fall victim to quantification, penalized for their propensities as much as their actual performance.

The e-book concludes by considering how the very content of education may change when we recast it with big data — as something that is more probabilistic than certain.

Bolting big data onto learning forces us to question a lot of assumptions about education. The school day and calendar were devised when most people worked on farms; new data may show that this is no longer appropriate. Students advanced in age-based cohorts, but a system of self-paced lessons makes such a lockstep approach less necessary — and the data may show it to be less effective than other approaches. So as we enter the big-data world, a burning question will be whether we are prepared to accept, and act upon, what we uncover.

Dawa looks at the black lines of the thangka he's traced as his master admonishes him. He tries again, to be as precise as the version he is being trained to copy. The process seems too mechanistic to even be called education. Yet the heritage of learning in the West was once rather like the training of Bhutanese thangka artists.

According to legend, French education ministers of yesteryear could look at their pocket watches and know exactly what every child across the country was learning at that very moment. In America, the U.S. commissioner of education in 1899, William Harris, boasted that schools had the "appearance of a machine"; that they instructed a young fellow "to behave in an orderly manner, to stay in his own place" — and other passive virtues.

Indeed, if a person from two or three centuries ago — say, Florence Nightingale in Britain, Talleyrand in France, or Benjamin Franklin in America — were to walk into a classroom today, it would feel perfectly familiar to them. Not much has changed, they'd probably say — even though everything outside the schoolyard has been transformed in almost unrecognizable ways.

At the same time, people have always seen in new technologies the chance to reform education, whether through CDs, television, radio, telephone, or computers. "Books will soon be obsolete in the public schools," Thomas Edison stated confidently in 1913. "It is possible to teach every branch of human knowledge with the motion picture. Our school system will be completely changed inside of ten years." Will big data really go where other innovations have barely made a dent?

For Professor Ng, the changes are happening faster than he could have imagined. On campus, his machine-learning class attracts several hundred students a semester. When he offered it online in 2011, more than 100,000 signed up. Around 46,000 started it and turned in the first assignments. By the end of the four-month course — some 113 ten-minute videos later — 23,000 had completed most of the work and 13,000 students received a high-enough grade to receive a statement of accomplishment.

A completion rate of around 10 percent may seem very low. Other online courses are more like 5 percent. Indeed, Sebastian Thrun, one of Professor Ng's Stanford colleagues, who cofounded a rival company to Coursera called Udacity, publically proclaimed MOOCs a failure in autumn 2013 because of the meager completion rates among those most in need of low-cost education. Yet such concerns miss a larger truth. Professor Ng's modest completion rate from a single course nevertheless comprises as many students as he could instruct in an entire lifetime of traditional teaching.

Big data is ripe to give education the transformative jolt it needs. Here's how it will happen.

2

CHANGE

Luis von Ahn looks like your typical American college student, and acts like one too. He likes to play video games. He speeds around in a blue sports car. And like a modern-day Tom Sawyer, he likes to get others to do his work for him. But looks are deceiving. In fact, von Ahn is one of the world's most distinguished computer science professors. And he's put about a billion people to work.

A decade ago, as a 22-year-old grad student, von Ahn helped create something called CAPTCHAs — squiggly text that people have to type into websites in order to sign up for things like free email. Doing so proves that they are humans and not spambots. An upgraded version (called reCAPTCHA) that von Ahn sold to Google had people type distorted text that wasn't just invented for the purpose, but came from Google's book-scanning project, which a computer couldn't decipher. It was a beautiful way to serve two goals with a single piece of data: register for things online, and decrypt words at the same time.

Since then, von Ahn, a professor at Carnegie Mellon University, has looked for other "twofers" — ways to get people to supply bits of data that can serve two purposes. He devised it in a startup that he launched in 2012 called Duolingo. The site and smartphone app help people learn foreign languages — something he can empathize with, having learned English as a young child in Guatemala. But the instruction happens in a very clever way.

The company has people translate texts in small phrases at a time,

or evaluate and fix other people's translations. Instead of presenting invented phrases, as is typical for translation software, Duolingo presents real sentences from documents that need translation, for which the company gets paid. After enough students have independently translated or verified a particular phrase, the system accepts it — and compiles all the discrete sentences into a complete document.

Among its customers are media companies such as CNN and BuzzFeed, which use it to translate their content in foreign markets. Like reCAPTCHA, Duolingo is a delightful "twin-win": students get free foreign language instruction while producing something of economic value in return.

But there is a third benefit: all the "data exhaust" that Duolingo collects as a byproduct of people interacting with the site — information like how long it takes someone to become proficient in a certain aspect of a language, how much practice is optimal, the consequences of missing a few days, and so on. All this data, von Ahn realized, could be processed in a way that let him see how people learn best. It's something we aren't very easily able to do in a nondigital setting. But considering that in 2013 Duolingo had around one million visitors a day, who spent more than 30 minutes each on the site, he had a huge population to study.

The most important insight von Ahn has uncovered is that the very question "how people learn best" is wrong. It's not about how "people" learn best — but *which* people, specifically. There has been little empirical work on what is the best way to teach a foreign language, he explains. There are lots of theories, positing that, say, one should teach adjectives before adverbs. But there is little hard data. And even when data exists, von Ahn notes, it's usually at such a small scale — a study of a few hundred students, for example — that using it to reach a generalizable finding is shaky at best. Why not base a conclusion on tens of millions of students over many years? With Duolingo, this is now becoming possible.

Crunching Duolingo's data, von Ahn spotted a significant finding. The best way to teach a language differs, depending on the students' native tongue and the language they're trying to acquire. In the case

of Spanish speakers learning English, it's common to teach pronouns early on: words like "he," "she," and "it." But he found that the term "it" tends to confuse and create anxiety for Spanish speakers, since the word doesn't easily translate into their language. So von Ahn ran a few tests. Teaching "he" and "she" but delaying the introduction of "it" until a few weeks later dramatically improves the number of people who stick with learning English rather than drop out.

Some of his findings are counterintuitive: women do better at sports terms; men lead them in cooking- and food-related words. In Italy, women as a group learn English better than men. And more such insights are popping up all the time.

The story of Duolingo underscores one of the most promising ways that big data is reshaping education. It is a lens into three core qualities that will improve learning: feedback, individualization, and probabilistic predictions.

Feedback

Formal education, from kindergarten to university, is steeped in feedback. We receive grades for homework, class participation, papers, and exams. Sometimes we get a grade just for mere attendance. Over the course of one's schooling, hundreds of such data points are amassed—"small data" signals that point to how well we performed in the eyes of our teachers. We have come to rely on this feedback as indicators of how well one is doing in school. And yet, almost every aspect of this system of educational feedback is deeply flawed.

We're not always collecting the right bits of information. Even when we are, we don't collect enough of it. And we don't use the data we've collected effectively.

This is ludicrous. Our iPhones are vastly more powerful than the NASA mainframe that flew astronauts safely to the moon and back. Spreadsheet software and graphing tools are amazingly versatile. But giving pupils, parents, and teachers an easy-to-use, comprehensive overview of student activity and performance remains the stuff of science fiction.

What's most curious about our current use of feedback in education is what we measure. We grade the performance of pupils, and hold them responsible for the results. We rarely measure — and certainly not comprehensively or at scale — how well we teach our kids. We do not grade the degree to which our techniques are conducive to learning, from textbooks and quizzes to class lectures.

In the small-data age, gathering data on these sorts of things was far too costly and difficult. So we measured the easy stuff, like test performance. The result was that the feedback went almost exclusively in one direction: from the teachers and schools to kids and their parents.

In any other sector, this would be very strange. No manufacturer or retailer evaluates just its customers. When they get feedback, it is largely about themselves — their own products and service, with an eye to how to improve them. In the context of learning, feedback is primarily about how well a person has understood her lesson as perceived by her teacher (culminating with an infrequent, standardized test), not how good the teacher or the teaching tools have been for a particular student. The feedback is about the result of learning, rather than the process of learning. And this is because of the perceived difficulty of capturing and analyzing the data.

Big data is changing this. We can collect data on aspects of learning that we couldn't gather before — we're datafying the learning process. And we can now combine the data in new ways, and parlay it back to students to improve comprehension and performance, as well as share it with teachers and administrators to improve the educational system.

Consider reading. Whether people reread a particular passage because it was especially elegant or obtuse was impossible to know. Did students make notes in the margins at specific paragraphs, and why? Did some readers give up before completing the text, and if so, where? All of this is highly revealing information, but was hard to know — until the invention of e-books.

When the textbook is on a tablet or computer, these sorts of signals can be collected, processed, and used to provide feedback to students,

teachers, and publishers. Little wonder, then, that the major educational textbook companies are piling into e-textbooks. Companies like Pearson, Kaplan, and McGraw-Hill want data on how their materials are used in order to improve them — as well as to tailor additional materials to students' specific needs. Not only will this improve student performance, but the firms will be better suited to compete with rivals on the basis of being more relevant and effective.

For example, one thing publishers hope to learn is the "decay curve" that tracks the degree to which students forget what they've previously read and perhaps had once been able to recall. This way, the system will know exactly when to review information with a student so she has a better chance of retaining that information. A student may receive a message that he is 85 percent more likely to remember a refresher module and answer correctly on a test if he watches the review video in the evening two days before an exam — not the night before, and never on the morning of the exam.

Developments like this change the educational book market. There, badly written materials do more damage than a boring novel that we put aside halfway through. Generations of frustrated students may struggle to reach their potential because they've been exposed to flawed teaching materials. One need only pick up an elementary school primer from the 1940s or so, with their small typefaces, arcane language, and oddball examples divorced from reality, to see the tragicomedy of what we taught children at the time.

Of course, school review boards today extensively vet educational materials. But these boards are often constrained in their evaluation. They can examine content for accuracy and bias, and compare it with accepted standards of pedagogy. But they have no easy empirical way to know whether such teaching materials work well for the students using them, or to see how students respond to specific parts of the textbook, so that any shortcomings can be fixed.

In contrast, textbook publishers hope to receive the analysis of aggregate data from e-book platforms about how students engage with their material, what they enjoy, and what annoys them. It is not that the authors would be forced to incorporate feedback, but just receiv-

ing it might give them a better sense of what worked and what did not. Writing is both an art and a craft, and thus is open to improvement based on a big-data analysis of feedback data gleaned from readers.

There is still a ways to go to make this a reality. In the United States, states as diverse as Indiana, Louisiana, Florida, Utah, and West Virginia allow districts to use digital textbooks in their classrooms. Yet although sales of e-books are approaching parity with paper-based ones, only 5 percent of school textbooks in the United States are digital.

Yet the potential gains are huge. Just as Professor Ng of Coursera can tap the clickstream data of tens of thousands of students taking his class at Stanford to know how to improve his lectures, so too can textbooks "learn" from how they are used. In the past, information traveled one way—from publisher to student. Now, it's becoming a two-way street. Our e-textbooks will "talk back" to the teacher.

However, not only will this information be used to redesign what already exists, but it can be analyzed in real time, to automatically present materials that are the best fit for the student's specific need at a particular moment. This is a technique called adaptive learning, and it is leading to a new era of highly personalized instruction.

Individualization

Learning has always been personal. We take what we see and hear and translate it into something to which we add to our own unique understanding of the world. But what we hear and see, what we are taught in schools or professional training courses, is packaged and standardized, as if one size fits all. This is the price we pay for making education more accessible, for transforming it from something that was once available mainly to the nobility, clergy, and wealthy, to something that is today within reach for most people.

As recently as two centuries ago, the idea of formal schooling was rare. Until university, the children of elites were individually tutored or sent to small, expensive academies. Education was in effect

custom-made to the student's exact needs at any moment. This obviously doesn't scale; only a handful of people could be taught in this way. When education became democratized in the nineteenth and twentieth centuries, it had to be mass-produced. Again, that was the price we had to pay.

Today, we enjoy tremendous variety for almost any category of consumer product. They may be mass-produced, but by choosing what best fits our personal preferences from a large selection of available goods, we can escape the one-size-fits-all mentality that led Henry Ford to quip, "Any customer can have a car painted any color that he wants so long as it is black." Yet the same sort of variety and customization that we've seen in other industries has not yet hit education at scale.

The reforms that have happened to date have been largely cosmetic. Students sometimes sit in circles; teaching is no longer strictly frontal. Students engage in group work, and are encouraged to learn from one another. Classrooms are welcoming and friendly. In developed countries, laptop and tablet computers are creeping into schools.

However, in one crucial dimension, learning has barely evolved. Modern education still resembles the factory era that accompanied its rise. Pupils are treated alike, given identical materials, and asked to solve the same problem sets. This is not individualized learning. Formal education still works essentially like an assembly line. The materials are interchangeable parts, and teaching is a process that — despite the best efforts of innovative and caring instructors — at its core treats all pupils similarly. Learning and teaching is benchmarked against a standard, based on an average, irrespective of individual preferences, qualities, or challenges. It reflects the mass-production paradigm of the industrial age.

Maintaining a consistent pace and presenting the exact same content at the same time, traditional education is geared to the interests of the instructor and the system, not the student. Indeed, most formal schooling is designed with the average student in mind — some fictional creature who learns slower than the whiz kid in the front

row but faster than the dullard in the back of the room. It's a category to which no one person actually belongs. But "average is over," as the title of a book by the American economist Tyler Cowen proclaims. That is, we now have technologies that let us tailor things to individual preferences and needs, not defer to the abstract homogeneity of yesteryear.

In fact, doing so is especially important, since in designing our education system for the average, we harm students on both sides of the bell curve. Optimizing for a mythical average student means that the quicker ones are bored out of their minds (or worse, become disciplinary problems), while the slower ones struggle to catch up. In reality, it is actually "one size fits few," in the words of Khan Academy's founder, Sal Khan, whose company is a leader in online instruction and individualization.

Instead, what we need is "one size fits one." And we can have it. We can individualize how knowledge is communicated, so that it better fits the specific learning context, preferences, and capabilities of individual pupils. It won't make rocket scientists out of everyone, and learning will continue to require concentration, dedication, and energy. But by breaking the homogeneity of one size fits all, we can optimize how people learn.

Tailoring education to each student has long been the aim of adaptive-learning software. The idea has been around for decades. In the past, however, the systems were of limited value. They harnessed computer technology to be faster and more personal. But they didn't learn from the data, to work in a bespoke way and individualize learning. This shift is similar to the change that happened in how computer scientists approached machine translation, from trying to code the proper word translations into software, to relying on data to get the computer to infer the most probable translation.

By tapping the data, adaptive-learning systems are now taking off. A report in 2013 commissioned by the Bill and Melinda Gates Foundation identified around 40 companies offering adaptive-learning software. Among them is Carnegie Learning. Its system for high school mathematics, Cognitive Tutor, decides what math questions

to ask based on how students answered previous questions. This way, it can identify problem areas and drill them, rather than try to cover everything but miss holes in their knowledge, as happens in the traditional method. In a highly scrutinized trial in Oklahoma with 400 high school freshmen, the system helped students achieve the same level of math proficiency in 12 percent less time than students learning math the traditional way.

The easiest wins are not in regular classrooms, where instructors are sometimes reluctant to adopt new approaches. (Teachers and their unions fear that the data may be used to rank performance or embolden school administrators to employ fewer teachers.) Instead, remedial classes are the perfect place to roll out these systems. These students are already behind the curve, so more drastic measures to improve learning are welcome since the traditional approach has clearly failed.

There, adaptive learning has shown substantial gains. "Students in these new-style remedial-ed courses outperformed students in conventional courses," observed Bill Gates to a conference of university trustees in 2013. "And colleges saw a 28 percent reduction in the cost per student," he added, for good measure.

The most impressive feature of individualized learning is that it is dynamic. The learning materials change and adapt as more data is collected, analyzed, and transformed into feedback. If one student has difficulties with fractions, future problem sets may incorporate them, so as to ensure she has enough opportunity to practice. This commonsensical notion is called "mastery learning," in which students move on to advanced material only once they have demonstrated they have a solid foundation.

For example, consider the classes at New York City's aptly named School of One, a math program operating at a handful of middle schools since 2009. Students get their own personalized "playlist," determined by an algorithm, each day—what math problems they will work on, tailored to their individual needs. "If I don't understand something, I can try and learn it in a new way and take my time. I don't have to learn it the same way everyone else does," says a School

of One student, Isabel Gonzales. Independent studies by the state and by a private educational service showed that students who went through the program did substantially better in math than students who did not.

If we can rip, mix, and burn our favorite music onto iPods, shouldn't we be able to do something similar with respect to our learning, where the stakes are higher? Clearly in the future, there will not just be one order and pace of study for a given textbook, subject, or course, but perhaps thousands of different combinations. In this, it is similar to online video games. There is not one version of Zynga's game FarmVille but hundreds, catering to the spectrum of customer interests and traits of play.

No longer will teachers select textbooks based on subjective beliefs about what works best pedagogically. Big-data analysis will guide them to select the materials that work best, which can then be further refined and customized for each individual student. To be sure, students in a cohort will still be exposed to the same material — after all, they'll need to pass the same test. But the material can be personalized.

This mass customization — the production of bespoke goods not much more expensive than mass-produced ones — has reshaped industries as diverse as car making and computers. It requires that detailed information flows from customers to producers, so that producers can create and offer customization options that are meaningful. Customers need to be able to express their preferences and choices easily and accurately. In the context of learning, individualization at scale demands even richer feedback data to flow to teachers and administrators. Individualization builds upon big-data feedback, and puts it right into practice.

Because we'll be collecting so much feedback data from so many students, we can continue to use the data to individualize in ways we did not think of when the data was collected. With small data, we collected only as much as necessary to answer a question we'd already posed (like test scores), because data collection and analysis was so costly. With big data, we have so much information, we can "let the

data speak"; that is, discover insights that were almost impossible to know before (like which forum posts improve exam results).

As a result, we will understand what in learning works and what doesn't — not only in general, but broken down by contexts, cohorts, and even down to the level of individuals. These systems will take the feedback results and dynamically adjust the materials and environment so that they are optimal for all students.

Probabilistic Predictions

With big data we gain unique insights into how people in aggregate learn, but much more importantly, into how each of us individually acquires knowledge. Yet these insights into education are not perfect. Our "learning about learning," so to speak, is probabilistic. We can predict with a high degree of likelihood what each individual needs to do to improve her educational performance: what kind of materials work best, what teaching style, and what feedback mechanism. Yet these are only probabilistic predictions.

For example, we may spot that teaching materials of a certain sort will improve a particular person's test scores in 95 percent of the cases, a very high degree of likelihood. Yet this still means that in one in twenty cases, we'll be wrong, and performance will not improve. That hardly means we shouldn't follow such predictions. They are clearly an improvement over classic, homogeneous education. They provide customization without the high cost that this normally implies. But in following these predictions, we must appreciate the limitations inherent in our insights. They are only probabilities; they do not offer certainty.

People are generally not very comfortable with probabilities. We prefer binary answers — yes or no; on or off; black or white. These answers offer direct and immediate guidance for decisions. What if big-data analysis tells us that switching to a particular textbook to teach our daughter Mandarin will improve her learning with 70 percent likelihood? Is that enough for us to have her switch? Are we willing to accept the risk of being wrong in three out of ten cases?

And what if the likelihood of the improvement is 70 percent, but the degree of improvement itself is relatively modest, say, a gain of 5 to 10 percent? Would we still have her switch if the effect for the people that it does not help is strongly negative, say, a full grade drop in test scores? Are we willing to take the chance of a high probability of some improvement over the small risk of a very negative effect? In a probabilistic universe, we will have to weigh such gains and risks and likelihoods often, and decide in the face of uncertainty.

This may be tolerable for recommendations from Amazon, or the results from Google Translate (both of which use probabilistic predictions based on big-data analyses). The consequences of being wrong are not debilitating. But it is potentially grave in decisions about people's education, which have a major effect on their future success.

Of course, we have always lived in a world of probabilities. We just failed to realize it. Whenever a teacher told concerned parents that their child needed to switch schools, or change subjects, redo a test, or use a particular textbook, these, too, were not absolutely certain truths, but probabilistic interventions. The big difference is that we can now measure these things, quantify them, and speak with greater precision. It shows not only how sure we are, but the limits of our certainty as well. In the age of big data, our chances become more visible. That may frighten people.

At the same time, as big-data predictions get more accurate and detailed, we should become more confident in the probabilities on which we base our decisions. Indeed, this may result in more specific and nuanced advice, leading to more tailored and perhaps less draconian interventions than in the past. So instead of mandating that a student spend the entire summer in remedial math, we can recommend with more gusto a focused, two-week refresher course on quadratic equations only.

However, the situation is exacerbated because of another necessary mental shift: from believing in our ability to uncover causalities to the realization that with big data, we'll often just see correlations. These correlations—seeming connections and associations among variables that we might not have otherwise known—do not tell us

why something is happening, only what is happening. But that is often good enough to help us make decisions.

For instance, Luis von Ahn's insight that Spanish speakers are better off learning different pronouns in English at different times — and when — is based on correlations. Likewise, Andrew Ng's method of ranking class-forum posts based on the degree to which students who have read them improve their test scores is wholly correlational. These things say nothing about the underlying reason at play, the causation. It's a matter of *what,* not *why.*

Relying on correlational insights is challenging. We are primed to see the world through the lens of cause and effect. Believing we have uncovered a cause is comforting for us; it gives us the sense that we comprehend the inner workings of the world. And yet, in reality, despite our efforts, we have discovered true causality in far fewer cases than we think. Often our quick intuitions of causal connections are just plain wrong when examined more closely.

That doesn't mean that the search for causality is wrong (or that we should give up looking for causes altogether). Far from it. But it suggests that we may need to be more humble in what we think we can understand of the world around us. Rather than hunting at great expense for an often elusive *why,* we may be better off with a more pragmatic approach, of aiming to first comprehend the *what* that noncausal analysis can reveal.

With big data, we can tap these predictions to improve how we teach and learn. The mythical one-room schoolhouse on the prairie is being replaced by electronic platforms. So it is there where we next cast our gaze.