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.

3

PLATFORMS

In 2004, a 28-year-old hedge fund analyst a year out of Harvard Business School was cornered by family members into tutoring his 12-year-old cousin Nadia in math. The only problem was that Nadia lived in New Orleans, and he lived in Boston. So he did it over the Internet—and forever changed the world of education.

The story of Salman Khan and the Khan Academy is fairly well known. A decade later, 50 million students from over 200 countries have used the site. It hosts over 5,000 video lessons on everything from math and science to art history. More than 4 million exercises are done on the site each day.

But what is less known is how the Khan Academy achieved its huge scale. Back in 2004, there weren't any videos yet—the tutoring took place live, at dedicated times. It was so effective for Nadia that family members petitioned Sal to tutor his other cousins, like Arman and Ali, and then still others. Soon, he was up to ten kids—and though he loved it, he couldn't keep up.

So he wrote a bit of software to manage the workload. It generated math questions and indicated whether the answers were correct. But it also collected data—troves of it. The software tracked things like how many questions each student got right and wrong, the length of time it took them to complete each problem, the time of day they did their work, and more.

"At first I thought of this as a mere convenience, an efficient way of

keeping tabs. Only gradually did the full potential usefulness of this feedback system occur to me," explains Khan. "By expanding and refining the feedback, I could begin to understand not only *what* my students were learning but *how* they were learning."

Thus, before he uploaded a single video for which the Khan Academy is famous, Khan himself designed a mechanism to harvest data from students' actions and to learn from them. If the charming 10-minute video lessons are the heart of the Khan Academy, the data analytics that constantly run in the background are its head.

The information, Khan said, allowed him to tackle questions that were difficult to even formulate in the past. Did students spend more time on questions they answered correctly or ones they got wrong? Did they work by perspiration or inspiration—that is, plod away until they uncovered the solution, or get it in a sudden burst of insight? Were mistakes made because they didn't understand the material—or because they were simply tired? These and many other fairly fundamental questions about how people learn could finally be asked. And perhaps answered.

Today, data is at the core of how the Khan Academy works. The nonprofit group ended 2013 with around 50 staff, nearly 10 of whom are focused on data analytics. The work is impressive. Teachers and "coaches" (perhaps a parent) get digital-dashboard reports on the students' progress. So do the students, allowing them to take an active role in their learning.

For instance, an overview pie chart identifies the subjects a student has watched lessons on, and another pie chart within it shows which she has taken tests on. For the teacher, a heat map displays how an entire class is performing, as well as each individual, broken down by the number of problems answered, the percentage correct, how many "hints" the student needed (and which ones worked best), among other things. A box beside the student's name turns red if the pupil is clearly stuck.

"Every interaction with our system is logged," says Khan, "and this data is used to give students, teachers, and parents real-time reports on student progress." Khan Academy runs a statistical model of every

student's accuracy on each exercise in order to determine if someone is "proficient" in an area. With over a billion completed exercises recorded on the site, this amounts to a lot of data indicating how students learn. The system also determines the most appropriate path of lessons one should take through a topic. So not only do students learn at their own pace, but in the sequence that works best for them as well.

The story of Khan Academy highlights how teaching and learning changes in the big-data era. It lets us see schools, classes, textbooks, and lessons in new ways — as platforms to collect data, analyze it, and parlay it into improved education.

This is not the way learning has happened before. The education sector is vertically integrated in terms of data and how it flows. The organizations that generate and gather data are mostly the ones that analyze it. Schools produce grades and other feedback data, and they are the entities that store and use the information. They incorporate it into their own decision making, and recommend decisions to others, from pupils and parents to potential employers and other schools. The institutions admit the kids, teach them, assess them, and ultimately bestow a credential on them as well.

Given the importance of the decisions at stake, the whole process is based on relatively little data. Indeed, it is data that has been gathered and analyzed not by an objective outsider, but by the quintessential insider. How can teachers and schools collect and analyze data objectively, when that data reflects their own abilities and failures in the classroom? And why do we still rely on such a system that, by the very way it is structured, will likely produce highly subjective and biased results?

Surprisingly, the institutions that collect the data and provide the analysis face only limited accountability for the data work they do. That is because their data gathering and modest analysis is just one of the services they provide. It is bundled together with what is seen as the primary service they offer: teaching.

Organizationally, this is odd. Companies have long known that

feedback and quality-assurance information should be collected by specialists who have no stake in the result. Otherwise, the process might be skewed. That's why quality control is often given to special units, whose task is to portray the situation as it is, not as management might like it to be. To achieve this, companies have disentangled responsibilities and data flows in their organizations. But even in a manufacturing plant, where quality control tends to measure the accuracy of machines more than of humans (and thus may be seen as less incriminating), such organizational separation of the data flows was fought over, and achieved only reluctantly.

In the education sector, we have seen nothing similar. In fact, the system has barely changed at all in centuries. Even the most basic tenets of the education system remain untested. The school day and annual calendar follow the cycle of agrarian life, though modern economies no longer do so. Classes are partitioned into regular segments separated by bells (as if in a factory), with no regard to whether this is the most effective way for material to be taught. Some classrooms impose a ban on digital devices, insisting that all work happen on paper, making the scholastic universe look rarified and antiquated compared to the social-media- and video-game-infused world the student interacts with everywhere else. And it also means that data can't be easily collated and analyzed.

To be sure, we have outside organizations to evaluate students using standardized tests, in the hope that this limits bias and subjectivity. Yet these test are taken only occasionally, capturing at best a fleeting moment in a student's learning. For example, in the United States, the No Child Left Behind Act of 2001 mandates that testing begin in the third grade (when children are around eight years old). It ushers in a host of penalties and benefits to schools based on their performance — all pinned on student tests that are just snapshots of a particular moment. In Europe and Asia, students are bludgeoned with national exams too.

The result of these sparse assessments is that the data is treated as a discrete, frozen image by which to rank a student — it's not used

as feedback in any real sense, be it to help students see how well they mastered the material, or to help teachers and administrators improve the learning environment, in terms of what materials to choose or how to structure the educational setting. It's like trying to monitor a patient not with an electrocardiogram that tracks 1,000 pulses a second, but with an old stethoscope, checking the heart rate once an hour. It's of meager use compared with the value that more frequent measurements could bring.

Yes, some schools are better at it than others. And yes, a handful of startups vie to serve as independent data-measurement firms or data platforms for schools. But they have yet to achieve any real scale. Educators use data to improve education the way our distant ancestors used cave paintings to communicate — it's still quite primitive.

This is poised to change, not because we are placing many educational activities into a digital setting, as with the Khan Academy and Andrew Ng's Coursera, though they certainly increase availability and lower cost. Rather, change is happening because when learning takes place digitally, we can collect data that was previously impossible to obtain. In so doing, big data has the potential to separate data generation from its processing and usage — to unbundle education informationally; to turn schools and textbooks into data platforms to improve learning.

Consider MOOCs and their scholastic siblings, SPOCs (small private online courses). They have given millions around the world access to high-quality instruction at low cost via video lectures from leading professors. This has democratized the distribution of and access to educational resources — a wonderful development fueled by the fact that many of the most prominent organizations, such as Coursera, Udacity, and edX (a joint venture of Harvard and MIT, with other universities), do so for free as part of their public mission.

These programs have fallen victim to the hype cycle. After being feted as one of the most important innovations since the discovery of fire, they have more recently been pummeled for not bringing about world peace quickly enough. In truth, the reality is somewhere be-

tween the extremes. Online courses are transformative, but probably as a supplement to more formal educational environments, not as a substitute for them.

Bill Gates makes the case well. All students everywhere can get lectures from the very best instructors in the world, he notes. And instead of delivering lectures that can't compete, other teachers may spend their time working directly with students. "The smart use of technology doesn't replace faculty—it redeploys them."

However, the democratizing element of these online courses is only one of their features. They are also platforms for the collection of data about learning—and they collect vast amounts of it. MOOCs are massive data-gathering platforms, bringing to individual learning a comprehensiveness and scale unprecedented in human history.

With massive online courses, the vertical integration of data flows will likely cease to exist. In its place a whole new ecosystem of data-gathering platforms may emerge. This opens up a huge space for innovation. Online courses may themselves mine the data for intriguing new insights, or they may give specialized third parties access to it. In fact, MOOCs may end up giving students access to their data, to let them decide through which third party they want it analyzed. (The Khan Academy already lets researchers tap some of its anonymized data on assessments, exercises, and videos via a "data sharing portal" after filling out a simple Data User Agreement.)

One could even see online courses becoming explicit data platforms, offering excellent lecturers a space to host their material. They would also accommodate a new cadre of instructors that would help students choose the right learning path—what courses to take, what materials to use, and how best to go through them—based on the feedback data they collect. Students could choose from a wide variety of lecturers as well as instructors, and select third-party service providers for data analysis. In fact, in such a porous ecosystem of open data flows, the differences between "inside" and "outside" the institution would likely diminish.

• • •

Where does this leave existing institutions of learning? On the one hand, schools and universities are already established units that could harness the power of big data. They have enough students to gather large amounts of data, and they have a vital interest in knowing how to improve learning for their pupils. This gives them an advantage in competing successfully against new entrants in the education arena.

On the other hand, schools and universities would have to change quite dramatically in order to benefit from big data. While they sit at the source of much data about learning, they have shown a limited ability to ingest that data, let alone analyze it effectively. In part, this may be due to regulatory restrictions concerning what data can be gathered and for what purposes. (Witness the outcry in many school districts over "data lockers" to store student information, and the opposition of teachers' unions to performance rankings—especially if they might be made public.) But in large part, it is because schools have never been challenged to use data effectively—though, to be fair, in the predigital age it would have been prohibitively expensive to do so.

The coming change will affect universities and other institutions of higher learning first. There, the audience is more mature, and thus knows better how to take material and transform it into something they can digest. After all, that's what learning to learn is all about. So the expectations in terms of pedagogy (compared with the content of the education) are lower, and thus even modest improvements in teaching will be welcomed.

Within universities, the changes will be first felt by the large undergraduate factories that churn out tens of thousands of graduates every year. They deliver education to the masses with limited resources—and are poised for disruption by innovators.

The story of Amazon versus Barnes & Noble is instructive. When Amazon first entered the book market, it attracted customers who enjoyed the convenience of shopping from home, and having access to a large inventory of books. Convenience and richness of content may be what draws students to MOOCs today. But this first incarna-

tion of Amazon did not endanger the Barnes & Noble Superstores.

Only after Amazon began using its data to make highly accurate and individualized recommendations did it create a unique buying experience that no Barnes & Noble Superstore could hope to replicate. It was primarily data, rather than convenience or richness of inventory, that led to the demise of the book superstores. Similarly, big data used by MOOCs (and other new entrants) will put tremendous pressure on mass universities, the brick-and-mortar establishments of the education sector.

Elite universities, with their outstanding faculty and valuable brands, might feel insulated (though a few have been early experimenters in online instruction). That may be true for the first wave of big-data applications in education. But they have to develop ways to capture and learn from data. As data starts to flow, the constituent parts of what makes a top university may fracture, perhaps to be reassembled in a different fashion, or perhaps to stay separated permanently.

Some top schools already get it. A high-profile MIT task force on the future of the institution released a preliminary report in November 2013 that identified edX as a critical part of its strategy to remain relevant. And mass universities may reinvent themselves as hybrid MOOCs, because administratively they are familiar with the scale that is necessary for mass-customized educational experiences.

In contrast, other top-brand universities and liberal arts colleges may have a harder time. They are unfamiliar with the challenges of scale, and their trusted but limited analog ways of individualizing the educational experience may not be good enough in the era of big data. Like small independent bookstores, they will find life hard. They will not all vanish, of course. Some independent bookstores, too, have survived Amazon's online advantage so far. But once big data hits them, they will need to do things differently. And these oceanic trends will eventually lap upon the shores of high schools and later elementary schools. No existing educational institution will be spared.

The response of some forward-thinking schools is to embrace e-learning, not because there are advantages with online instruction per

se, but because of the value of the data itself. The impact on student outcomes of harnessing data feedback, individualizing instruction, and relying on probabilistic predictions is just so great. For example, Khan Academy in California has partnered with public schools in places as diverse as Los Altos (extremely wealthy) and Oakland (relatively poor) in that state. The results have been amazing.

At Peninsula Bridge, a summer school program for middle-schoolers from poor communities in the Bay Area, Khan Academy's lessons are used to teach math. At the start of one session, a seventh-grade girl scored at the bottom of her class, and for most of the summer she was one of the slowest students. Clearly, she didn't "get" math. But then something clicked. She started making progress — fast. By the end of the session, she scored the second-highest in the class, far ahead of the smarty-pants who ranked way above her at the start.

Sal Khan was intrigued and dug up the records. He could see every question she answered, how long she took. The system created a graph that plotted her progress compared to her classmates. One sees a line hovering near the bottom for a painfully long period, until it bolts upright and surpasses almost every other line, representing all the other students.

For Khan, it marked a turning point in his thinking. The data clearly showed that what we consider a D student and an A student — based on how they perform on a single test at a single moment in time — says very little about actual ability. When students can work at their own pace, in an instructional sequence best suited to them, even those who seem the least capable may end up outperforming the very best. But in a conventional educational setting, based on small data, this one shy seventh-grader would have been relegated to remedial math and left to flounder — with consequences that could well have stretched across her lifetime.

In the future, there may be any number of companies that would compete to tailor instruction specifically to her, and companies would be rated on how well their data analysis predicted her performance and helped her succeed. That surely would be novel in education, though when one thinks about it, it shouldn't be.

So what will learning look like tomorrow? A little like the trends we are seeing in miniature today, only more of it. We can foresee data being generated independently of teaching. The basic idea is that data about how we learn (rather than just how we perform on occasional, formal tests) will be continuously collected and analyzed. It will be accessible not just to teachers, but to students, parents, and administrators. Educational materials will be algorithmically customized, to fit the needs of the student in terms of instructional sequence and the pace at which the individual most effectively learns. Moreover, the materials themselves will be constantly improved.

Schools will become, in effect, a cornerstone of a big-data ecosystem. One can even imagine institutions competing over their ability to harness data to improve student performance. And schools will be able to prove their worth not with flaky college rankings, but with solid data.

Applying data to education opens up the possibility for new, innovative organizations and business models for analyzing the information and applying its lessons. There is a creative space that is set to widen, since existing organizations that have all the data—or that could seemingly get it—currently lack the mindset to mine it effectively. So new entrants are poised to shake up the sector—which is ripe for shaking.

However, reaching this future requires surmounting a hurdle that's even tougher than school boards or teachers' unions. It is the dark side of big data applied to education. The consequences are profound for privacy and human freedom, in terms of the ever-presence of past data, and probabilistic predictions that can unfairly decide our fate and rob us of our future.

4

CONSEQUENCES

Arizona state university, like many colleges across the United States, has a problem with students who enter their freshman year ill prepared in math. Though the school offers remedial classes, one-third of students earn less than a C, a key predictor that they will leave before getting a degree. To improve the dismal situation, ASU turned to adaptive-learning software by Knewton, a prominent edtech company. The result: pass rates zipped up from 64 percent to 75 percent between 2009 and 2011, and dropout rates were cut in half.

But imagine the underside to this seeming success story. What if the data collected by the software never disappeared, and the fact that one had needed to take remedial classes became part of a student's permanent record, accessible decades later? Consider if the technical system made predictions that tried to improve the school's success rate not by pushing students to excel, but by pushing them out, in order to inflate the overall grade average of students who remained.

These sorts of scenarios are extremely possible. Some educational reformers advocate for "digital backpacks" that would have students carry their electronic transcripts with them throughout their schooling. And adaptive-learning algorithms are a spooky art. Khan Academy's "dean of analytics," Jace Kohlmeier, raises a conundrum with "domain learning curves" to identify what students know. "We could raise the average accuracy for the more experienced end of a learning

curve just by frustrating weaker learners early on and causing them to quit," he explains, "but that hardly seems like the thing to do!"

So far in this e-book we have shown the extraordinary ways in which big data can improve learning and education. Now our thoughts take a dark turn, as we consider the risks. Parents and education experts have long worried about protecting the privacy of minors and the consequences of academically "tracking" students, which potentially narrows their opportunities in life. Big data doesn't simply magnify these problems, it changes their very nature. Here, as elsewhere, the change in scale leads to a change in state.

Permanence of the Past

Many parents are viscerally alarmed by the huge stockpile of personal data that is starting to accumulate over the course of their children's schooling. For example, the nonprofit organization inBloom — backed with $100 million by the prestigious Gates Foundation and Carnegie — struck agreements with nine states to be a repository of student data. But after huge parental outcry in 2013, six of those states put the initiatives on hold.

Who can blame parents? Here's how the tabloid *New York Daily News* characterized the program in 2013: "In an unprecedented move, education officials will hand over personal student data to a new private company to create a national database for businesses that contract with public schools."

One might as well have suggested that school officials require students to walk barefoot atop cactuses while drinking paint thinner. The reality is that inBloom actually gives complete control to the schools to decide what information to store and who may access it — that control is the very point of its service. But still, many people are uneasy, because compiling so much data in one place is so new, and they may not be mentally prepared for the consequences. Our institutions aren't yet fully prepared to deal with it, either. For example, parent groups have rightly pointed out that uploading students' disci-

plinary information without proper checks on how it may be used is a recipe for problems.

Yet behind the intuitive opposition lies not just the conventional concern over privacy and data protection, but a more unique worry. Where traditional data protection has mostly been focused on addressing the power imbalance that results from others having access to one's personal data, here the concern is more about the threat posed by an unshakable past.

This is a particular concern, since we humans continuously evolve. Over time, we alter our views, realign our opinions, and even change our values. These changes are partly a function of age: young people tend to take more risks than older folks. It is partly because we exist in a particular context, and over time the opinions around us shape our own. And it is partly the simple result of further reflection, and — if we can call it this — mental and spiritual growth.

While we as individuals grow, evolve, and change, comprehensive educational data collected through the years remains unchanged. Even though we might have grown into the most even-tempered of individuals, if the data reveals an aggressive period in our school days long past, will future assessors be able to put that old data in the appropriate perspective? If not, we may remain forever beholden to our past — even though it represents a person that no longer really exists, and whose values bear little resemblance to our own. Constantly recalling that anachronistic data-shell of a person would not only be unjust, it would also produce incorrect results.

Think about records of student activism being stored and made available to prospective employers when an individual applies for a job a quarter of a century later. Today past records are very hard to access, save for high-profile individuals. But in the future this information will be routinely accessible for everyone. And it may not be just "snapshot" data like standardized college admissions tests — it may be every scrap of data related to our progress as a student, from number of sick days and visits to the guidance counselor, to number of pages read and passages underlined in *Huckleberry Finn*.

It isn't that data about our past is useless, but that it has to be understood in a wider context of who we are and what we have done. Our evaluators need to treat past data especially carefully, always judging whether the information has any relevance to who we are today. This seems obvious, and yet in practice it is incredibly hard. Often people have difficulty understanding time as a dimension of change.

Humans never really had to develop straightforward cognitive methods to put events from the distant past in appropriate perspective, because we had one of the best ways to do so built right into our brains: forgetting. Our brain constantly forgets details from the past that it deems no longer relevant to the present or useful for the future. Forgetting is mental house-cleaning that we never needed to care about consciously, and that helped us stay firmly wedded to the present. People who have difficulties forgetting describe their condition to researchers not as a blessing but a curse. It forces them to see only trees, never the forest—because any generalization requires us to forget details.

Even with report cards and requirements to save files in education, in the analog days, most of our academic information was stored in paper archives. These were so hard to locate, access, copy, and analyze that the records were in effect kept safe from inappropriate reuse by dint of the technical constraints of interacting with them.

With digital tools, and especially cheap storage and fast retrieval, the educational data of today's students will persist for much longer, and be much easier to access. Not only are recruiters Googling applicants, some have begun to demand their Facebook log-in details as well. This enables them to view almost a decade's worth of personal opinions, predilections, and ill-advised selfies. Perhaps more concerning, they can see what others have said about the applicant as well.

The permanence of this old data is the biggest worry. Confronted with this information, and with no way to put it into perspective, chances are that we'll look through the prism of persistent memory—a remembrance that can never forget. So though we might remind ourselves ten times before a job interview to disregard the fact

that an applicant got caught cheating long ago in high school, we still might not be able to give him the benefit of the doubt when it comes time to make a decision about whether to hire him. Worse, the fellow may himself carry that stigma with him everywhere he goes, like a scarlet letter, since society can't seem to forget the incident either. After all, people are habituated to remember the uncommon thing, not the mundane or the most recent.

Hence, the first significant danger with comprehensive educational data is not that the information may be released improperly, but that it shackles us to our past, denying us due credit for our ability to evolve, grow, and change. And there is no reliable safeguard against this danger. We can't easily change how we evaluate others, and what we take into account. Most of our thought-processes happen without our ability to fully control them rationally. On the other hand, not collecting or keeping the data would stunt the benefits that big data brings to learning.

Fixed Futures

The second danger is equally severe. The comprehensive educational data collected on all of us will be used to make predictions about our future: that we should learn at this pace, at this sequence, that we will have a 90 percent likelihood of getting a B or above if we review the material between 8:00 p.m. and 9:00 p.m., but it drops down to 50 percent if we do so earlier in the evening, and so on. This is probabilistic prediction — and the danger is that it may restrict our "learning freedom," and ultimately, our opportunity in life.

The huge promise of big data is that it individualizes learning and improves educational materials and teaching, and ultimately student performance. The data is parlayed as feedback to improve the product, not simply evaluate those who consume it. Today, the limited data that is collected is almost entirely used to assess students, the "consumers" of education.

We assess likely fit and potential success, from acceptance into accelerated programs in high school, to college admissions, to who

gets into which grad school. But such small-data predictions based on limited data points are fraught with uncertainties. As a result, admissions boards use them extremely carefully. Recognizing the imperfections of what the number represents — the smug bore who aced his SATs not because he's genuinely intelligent but because he memorized the review guide — these boards actively strive to add a dose of subjectivity into the assessment, to override what the data dictates when their judgment suggests that data isn't the whole story.

In the age of big data, however, these predictions will be far more accurate than today. This puts more pressure on decision makers, from admissions boards to job recruiters, to put more stock in what they foretell. In the past we could argue our case that a group to which we belonged might not apply specifically to us as an individual. For example, we might be placed in the cohort "good students who nevertheless had trouble in stats class," and warded away from majoring in economics. But we could still convince others that a prediction based on this grouping was inaccurate; we were different, and thus we might succeed where others in that cohort tended to fail. Because it was just a "small data" prediction, decision makers were primed to give us the benefit of the doubt. We could exculpate ourselves from guilt by association.

The danger now is that because the predictions are so accurate and inherently individualized, we are no longer guilty by virtue of the group we nominally belong to, but because of who we actually are. And thus all the convincing in the world may not be enough to sway decision makers in our favor. In fact, human judgment may be removed entirely from the decision-making process, as robotic algorithms simply access a spreadsheet, calculate the odds, and make a binding decision in a few milliseconds.

For example, some universities are experimenting with "e-advisors" — big-data software systems that crunch the numbers to help students graduate. Since the University of Arizona implemented such a system in 2007, the proportion of students who move on from one year to the next has increased from 77 percent to 84 percent. At Austin Peay State University in Tennessee, when students take a class for which soft-

ware called Degree Compass indicates they will get at least a B, they have a 90 percent chance of doing so, compared to around 60 percent otherwise.

These systems can make a big difference in graduation rates, considering that in the United States only about half of students graduate within six years. But they can have pernicious consequences too. What if the system predicts we're not likely to do well in one field, like bioinformatics, so subtly directs us toward another, like nursing? We may think it has our best interests at heart — providing us with a comfortable educational trajectory. But that may actually be the problem. Perhaps we should be pushed to succeed against the odds rather than feel content to advance along a smoother track.

In essence, these probabilistic predictions will enable decision makers — from admissions boards to job recruiters — to choose a safe route and minimize the risk of future disappointment. That proposition is very tempting, especially when compared with the alternative, an academic misstep that hurts us, such as failing to graduate, or choosing a major the rigors of which we can't handle. And institutions may even face potential legal liability if they don't follow the predictions that big data suggests.

Where probabilistic predictions may become most deeply rooted, and do the most harm, is in the area of tracking. For decades, many countries have divided students early onto different tracks, usually three: vocational for the academically challenged; regular classes for the average; and "advanced placement" for the high achievers. This has also long been controversial. It seemed to deny a person's fair chance to go to university if, right before high school, he fell outside the bounds of college prep. It could perpetuate social and economic divides through reinforcing educational divides, particularly if more women or minorities were culled from the top tier.

One hope, and it's just a hope, is that big data will make tracking disappear. As students learn at their own pace, and the sequence of material is algorithmically optimized so they learn best, we may see less need to formally track students.

But the reality could well be in the inverse. Customized education

may actually lock in these streams more ruthlessly, making it harder for one to break out of a particular track if they wanted to or could. There are now a billion different tracks: one for every individual student. The upside is that education is custom-tailored to each individual. The downside is that it may actually be harder to leap out of the canyon-like groove we're locked into. We're still trapped in a track, even if it is a bespoke one.

The system may have analyzed data across a million other students to base its prediction about one particular student's likelihood for success, and tailored the education directly to him—tracked him, in a way. But is this much better than if he were in a general class and had more of an opportunity to find and show his true skill level? The prediction may be accurate and in some cases helpful, but it is also unforgiving. He becomes a victim not of his abilities but of probabilistic predictions.

These constant forecasts of our likelihoods in areas big and small will not only affect our behavior, but will forever change what the future holds—transform it from a wide-open landscape to a terrain as predefined and immutable as our past. Would this not push our society back into a new form of caste system, an odd marriage of meritocracy and high-tech feudalism?

In the twentieth century, education was the great equalizer. Now, with big data, there is a risk that our predictions of potential outcomes, probabilistic outcomes, may make education the setting that widens inequalities.

Addressing the Anxiety

How to overcome these instinctual and rational fears of the dangers big data poses when applied to education? In most countries, some form of privacy law currently protects against the comprehensive collection and long-term storage of personal information. Generally, these laws require data users to inform people whose data they collect what it might be used for and get their consent for that use. But that approach doesn't work so well in the era of big data, where we of-

ten don't know to what use the information might be put years hence.

Much of the appeal of big data is that its value lies in its reuse for purposes that were scarcely contemplated when the data was initially gathered. So, informed consent at the time of collection is often impossible. Hence, if the spirit of the legal protections is taken seriously, much of big data's benefits in education will remain unrealized. Or, in the future, individuals will be asked to consent to vague descriptions covering almost any uses, which makes absurd the very notion of informed consent.

Right now, regulations focus on controlling the collection of data, making sure that people know their information is being gathered, getting their consent, and so on. But we must shift that focus to controlling how data is used. Informed consent must be augmented by direct accountability and responsibility on the part of companies and organizations that use the data and reap its value.

Policymakers in Europe and the United States are already discussing how to reform privacy laws to make those who use big data more accountable for any misuse of it. In return for taking on more responsibility (and thus more legal liability), data processors would be able to reuse personal information for new purposes. Of course, many obstacles will have to be cleared before this new and potentially more effective mechanism is ready for implementation, including which uses of personal data are permitted and restricted.

In education, this could permit the use of personal data to improve learning materials and tools, while using the same data to predict students' future abilities may be allowed only under much more stringent safeguards (such as transparency and regulatory oversight). It may require the explicit consent of the students themselves. It will also need tough enforcement, so that firms that use the data know that they cannot afford to break the rules. Here, there is a role for "algorithmists"—independent professionals trained in the art and science of big data (like statistics, data gathering, computer science, etc.). They will be able to scrutinize whether firms have implemented the big-data systems wisely, or serve as the experts that regulators rely upon to examine what's being done with datasets.

An individual's educational information is particularly sensitive—it goes to the heart of how each of us matures. Society rightly tends to forgive youthful transgressions more than it does adult misdeeds, in part because learning requires a modicum of experimentation, of trial and error. Hence, in the educational context it might be necessary to take big-data constraints one step further—especially as one cannot necessarily assume that stringently enforced data-user accountability will remain in place far into the future.

This sort of added protection we have in mind could be achieved by mandating that personal data related to one's education can only be stored and reused for a limited time. Different countries could determine what that period is, but the point is that it is not unlimited. This would give data users enough time to extract value from the information through reuse, but it prevents a data shadow from hanging over us permanently. It's a balancing act. We would be deliberately forgoing some of the benefits of future reuse in return for ensuring that we retain our ability to evolve beyond what the past data said.

Certain data could be kept longer if it were aggregated rather than recorded as individual data points (the grade-point average of a class, for example, rather than of particular students). And some individual data points could be stored for longer if they were stripped of obvious personal identifiers, such as names or social security numbers.

Such pseudo-anonymization of data is no silver bullet, however, as seemingly anonymized data can often still be reidentified by comparing it with data from other sources. (Similar patterns in different datasets can often be correlated and used to uncover a person's identity.) Thus, stripping obvious identifiers will only act as a speed bump—not a barrier—to someone (or some algorithm) intent on IDing an individual, and the data may have to be deleted altogether, eventually, to protect individuals' privacy.

Ultimately, how much big-data analysis we would like to see in education, and how we best protect against the dystopian dangers we foresee, will remain a delicate tradeoff between our desire to optimize learning and our refusal to let the past dictate the future.

5

DAWN

Sebastián díaz doesn't get it. He can't understand why people embrace statistics in other areas of life, but resist applying them to learning. "We take computational data and create fantasy football teams, which is computer simulation data," he says. "And at the same time, we claim it's too hard to bring quantitative accountability to education."

He is well placed to know. After getting a degree in chemistry and working for a year testing the groundwater at Walt Disney World in Florida, he pivoted and earned a master's degree in education. It evolved into a PhD—with a focus on applied statistics and measurement. (And he picked up a law degree along the way). As a professor at West Virginia University teaching stats and education law, he began wondering about what influenced the dropout rate of students, particularly in online classes. These classes were often specifically designed to help busy people fit schooling into their lives. But too many weren't making it through and getting a degree.

In the United States, millions of students even at brick-and-mortar universities choose online classes for their convenience and relatively low cost. In 2011, Professor Díaz and a handful of colleagues at several other institutions joined forces to figure out what was causing those troubling withdrawal rates. The answer, they believed, was in the data. With data from a half-dozen schools, and $1 million from the Gates Foundation, the team began compiling digital dossiers.

The records eventually grew to 1 million (anonymized) students, and a hefty 33 variables for analysis. Some were fairly basic, like age, gender, grades, and the degree being sought. Others were less obvious, like the size of the classes the students took, and military status. Still other variables were more uncommon, like the total number of courses for which the student had ever registered. Then there were those signals that could only be detected once education shifted to a digital setting, such as the number of days since the student last logged in to any class. Altogether, it amounted to many millions of data points, down to the level of the individual courses.

They were seeking the factors which best predicted that a student would drop out. Something surprising cropped up. A strong predictor that someone will continue with classes was not their age, or gender, or grades. It was, simply, how many classes they were taking. "Probabilistically speaking," says Professor Díaz, "they were more likely to persist if they took fewer courses simultaneously in the beginning."

Professor Díaz is quick to qualify these findings, as good academics do: it's early days and more study is needed. But they have serious implications for public policy. U.S. financial grants require the recipient to carry a full-time course load as a condition for support. The data suggests that this requirement may be deeply flawed. (In fact, it was almost certainly determined in the absence of empirical data, since that was how most education policies were made before big data arrived on the scene.)

If Professor Díaz's research holds true, then current U.S. policy isn't just wasting money by mandating more coursework than students can handle. It's wasting people's lives as well.

Big data transforms learning by rigorously examining aspects of education that have evaded empirical scrutiny for centuries. This is different than simply adding computer technology to schools. We've been doing that since the 1980s, with the chief result being that a lot of old IT junk needed to be carted away.

The current change is not technical, though information technology enables it. It affects what sort of data we can collect and how we

can crunch it to unearth new insights into learning, teaching, and the process of acquiring knowledge. So Isabel Gonzales at New York City's School of One can learn at her own pace and improve her performance. Professor Ng at Coursera can identify what parts of his lessons are least effective and improve them. Education policymakers can finally recognize the potential of students like the seventh-grader at Peninsula Bridge summer school, who went from the very bottom to the very top of her class after Khan Academy changed the process of instruction.

Yet the most meaningful wins are not improving what we already do, but doing things anew — giving a data-driven overhaul to instruction itself. An enthusiastic Sal Khan says that Khan Academy's focus on data and analytics allows for the creation of "an automated personal tutor." Some may feel that the use of software that constantly looks over a student's shoulder is tantamount to spying. But the opposite argument can be made, that such software gives necessary support to students who are too ashamed or nervous to ask for help and admit that they don't understand something. For those students — and they're far more common than we may like to admit — passive monitoring lets teachers track their progress and problems in a less threatening way.

Learning with big data brings three main changes. We can collect feedback data that was impractical or impossible to amass before. We can individualize learning, tailoring it not to a cohort of similar students, but to the individual student's needs. And we can use probabilistic predictions to optimize what they learn, when they learn, and how they learn.

As these changes unfold, we'll find that many of the tools and institutions we rely on must themselves change. The e-textbook, the digital lecture, the very university becomes a platform or nexus for the acquisition and analysis of data. This will lead to an unbundling of the educational experience. The monopoly that schools hold today is starting to resemble the monopolies once held by monarchy and Church. It is poised to crumble, as did those other seemingly impregnable institutions when the currents of an earlier information revo-

lution—printing—washed over them. This unbundling may bring competition to multiple facets of education, as academia goes from being vertically integrated and new players crop up in new areas.

However, the marriage of big data and learning also begets dangers. One is the permanence of information about evanescent aspects of our lives, which can give them undue significance. There's also the risk that our predictions may, in the guise of tailoring education to individual learning, actually narrow a person's educational opportunities to those predetermined by some algorithm. And, just as personal media has atomized us, undermining the notion of a set of interests and values shared by all, so too may probabilistic learning reduce education from a shared experience to one that is custom-made—but so snug that we're divorced from our neighbors and the wider society.

To remedy these worries, we call for a shift from regulating how data is collected to rules regarding how it's used. This will allow us to glean from data ways to improve the tools and methods of learning. At the same time, it will place strong constraints on big-data analyses that risk tarnishing a student's future through probabilistic predictions. We also argue for tough enforcement and skilled specialists—algorithmists—to assess the effectiveness and navigate the intricacies of big-data systems. And we recommend creating regulatory and technical speed bumps that place limits on how long and in what form sensitive educational data can be stored.

There is no simple solution to the challenges big data presents in education or in any realm. But by establishing multiple safeguards, we can hope that we are neither shackled to our past nor robbed of our future.

Big data will fundamentally alter education. By gathering and analyzing more information about how each of us learns, we'll be able to improve the techniques and tailor the materials to the precise needs of an individual student, a particular teacher, and a specific classroom. No longer will ignorance be a valid excuse to avoid improving our educational processes and institutions. The nature of education funda-

mentally changes, because with big data, society can finally learn how to learn.

And as we do, we will need to change *what* we learn, too.

Through collecting and analyzing large amounts of data, we improve our understanding of reality, the world around us. But what does that entail? Humans have always examined the world by observing it. We employ theories — generalizable ideas about how the world works — and apply them to various contexts, resulting in a hypothesis that can then be tested with data.

That's the essence of the scientific method that underpins (at least in theory!) most human discovery. Until now, however, collecting and analyzing data was both time-consuming and costly. So we collected as little data as possible to answer the questions we had, perhaps not appreciating the extent to which the information available to us determined the questions we asked. But when we change the amount of data we can collect, the result isn't just better answers to the same questions, but the capacity to ask different and better questions.

Big data helps us escape our mental constraints. Just as online classes, e-textbooks, and computer-based tests make it easier to collect data, so are we amassing an unprecedented amount of information in other areas of life. Big data won't necessarily explain the exact causes underlying all things (being largely correlational, it tells *what,* not *why*). Yet it will give us a more comprehensive and detailed perspective on the complexity of the world and our place in it.

Big data portends more than a change in how we learn, but in how we think about the world. Because of big data — and with its help — we will learn to set aside treasured notions of cause and effect.

In the big-data future, we'll still need theories. But rather than testing possibilities one at a time based on our preconceived notions, we can use big-data analysis to test not just one but a whole universe of possible hypotheses that our computers generate algorithmically, less encumbered by our extant beliefs. It is the difference between thinking one knows the answer and finding out through trial and error that one doesn't, and having the computer test all possible answers to find the best one using all the data available.

As we learn this new way of making sense of the world, we'll come to understand it much better, and thus improve our decision making. But this new understanding will also demand that we embrace probabilities, uncertainty, and risk much more than we have in the past. It's not that the world will have become more risky; it is that we will have come to realize that there are far fewer certainties than we previously thought.

This will require us to prepare future generations quite differently for their careers, but more importantly, for their lives. We will see that the world is vastly more complex, detailed, and uncertain than we'd once reckoned—yet we'll find that it is also more open to exploration and investigation than we've previously imagined.

Humankind is shedding new light in dark corners by collecting and analyzing big data, with which we can peer at something greater than ourselves and so can know ourselves better. We have done this before, with tools, technology, and new ideas like math, science, and the Enlightenment. Big data is just another step on a shadowy path along which we carry a torch, hoping to light our way, but perhaps occasionally burning ourselves as well.

This light will allow us to see past the intellectual shortcuts we used in the era of small data. These shortcuts worked well enough most of the time, even though they lacked complexity and detail. Newton's law of gravity is sufficient to build bridges and manufacture engines. But it wasn't precise enough to help us design the GPS system with which we determine exact locations on Earth. For that, a much more complex law of gravity, using Einstein's theory of relativity, was needed.

Similarly, through big data we will begin to understand that many of the so-called laws that we use to explain reality aren't exact enough. They worked for their time, solving the needs to which they were put, like a manual lever as compared to a hydraulic pump, or a rope pulley to a mechanical crane. But in the future, as more data can be collected and analyzed, these shortcuts will be replaced by a more complex but also more accurate understanding of the world.

Take probabilities. In the past, when asked what was the chance

that a coin tossed in the air would land with its face up, we'd say 50:50. That's a good approximation. But the reality is more complex, as every coin is slightly different and every person will throw it differently, too. With big data, rather than sticking with the ideal shortcut of 50:50, we'll just understand that each throw is an opportunity to learn a little bit more about reality. And after each throw, we'll use the result to improve our prediction about it. Slowly this will get us closer to the truth.

The age of big data will be one of continuous learning, of constantly improving our sense of what the world is, rather than believing that with a simple shortcut we have uncovered all there is to understand. We are fast entering a period in which everything we observe can and will be used to accumulate more knowledge, like the steady growth of a stalagmite in a cave.

This affects *what* we teach, not just *how* we teach. So as we improve the process of education, we need to change its substance as well.

As we deepen our understanding of the world and its complex beauty, and realize the power of discovery that big data provides, we also have to be aware of its limitations. More so than in the past, we'll have to learn about the inherent shortcomings of the tools that we use to make sense of the world—limitations that even with great care cannot be overcome or avoided.

And in our learning, too, we must continue to appreciate what the data cannot tell: the ideas produced by human ingenuity, originality, and creativity that no big-data analyses could have predicted. "Imagination is more important than knowledge," said Albert Einstein. "Knowledge is limited. Imagination encircles the world."

To ensure we keep these qualities alive, we will need to preserve a special space for ourselves, our irrationality, our occasional rebellion against attempts to quantify and qualify. Not because big data is amiss, but because even in a new age of learning, not everything can be learned.

NOTES

1. DUSK

page

1 Dawa concentrates — Observed during a visit to Bhutan, 2009.

2 Professor Ng collects information on everything — Interview with Andrew Ng, May 2012.

4 Venture capital was poured into education — "Catching On at Last," *Economist,* June 29, 2013, http://www.economist.com/news/briefing/21580136-new-technology-poised-disrupt-americas-schools-and-then-worlds-catching-last.

5 The e-learning market — e-Learning Centre, "2012–17 Market Predictions from GSV Advisors" http://www.e-learningcentre.co.uk/resources/market_reports_/2012_17_market_predictions_from_gsv_advisors.
Spending on education overall — Talk by Jace Kohlmeier, Mixpanel Office Hours, San Francisco. Video uploaded March 25, 2013, http://www.youtube.com/watch?v=Nvoc6atMpAw#t=30.

6 "Appearance of a machine" — Quoted in Joshua Davis, "How a Radical New Teaching Method Could Unleash a Generation of Geniuses," *Wired,* October 15, 2013, http://www.wired.com/business/2013/10/free-thinkers/all/.
It would feel perfectly familiar to them — This observation was made by Walter Isaacson at a Microsoft CEO Summit: Innovation in Education Part 1. Posted on YouTube, August 31, 2012, http://www.youtube.com/watch?v=vAQj9g8Igck.

7 "Books will soon be obsolete" — From an interview published in the *New York Dramatic Mirror,* July 9, 1913, which was reprinted in other newspapers at the time. Cited in Paul Saettler, *The Evolution of American Educational Technology* (Charlotte, NC: Information Age Publishing, 2004), p. 98.

As many students as he could instruct — Max Chafkin, "Udacity's Sebastian Thrun, Godfather of Free Online Education, Changes Course," *Fast Company,* December 2013/January 2014, http://www.fastcompany.com/3021473/udacity-sebastian-thrun-uphill-climb.

2. CHANGE

9 Luis von Ahn — Interview with von Ahn, May 2013.

14 E-books are approaching parity with paper-based ones — Laura Hazard Owen, "PwC: The U.S. Consumer Ebook Market Will Be Bigger Than the Print Book Market by 2017," PaidContent, June 4, 2013, http://paidcontent.org/2013/06/04/pwc-the-u-s-consumer-ebook-market-will-be-bigger-than-the-print-book-market-by-2017.

Only 5 percent — Darrell M. West, *Digital Schools: How Technology Can Transform Education* (Washington, DC: Brookings Institution Press, 2013), p. 24.

15 "Any customer can have a car" — Henry Ford, with Samuel Crowther, *My Life and Work* (Garden City, NY: Doubleday, Doran & Co., 1930), Project Gutenberg, http://www.gutenberg.org/cache/epub/7213/pg7213.html.

16 But "average is over" — Tyler Cowen, *Average Is Over: Powering America Beyond the Age of the Great Stagnation* (New York: Dutton, 2013).

"One size fits few" — Salman Khan, *The One World Schoolhouse: Education Reimagined* (New York: Twelve, 2012), p. 57.

A report in 2013 — Education Growth Advisors, *Learning To Adapt: Understanding the Adaptive Learning Supplier Landscape.* Interview with EGA's Adam Newman, "2013: The Year of Adaptive Learning," *Impatient Optimists* (blog), Bill and Melinda Gates Foundation, April 10, 2013, http://www.impatientoptimists.org/Posts/2013/04/2013-The-Year-of-Adaptive-Learning.

17 A highly scrutinized trial — "Catching On at Last," *Economist,* June 29, 2013, http://www.economist.com/news/briefing/21580136-new-technology-poised-disrupt-americas-schools-and-then-worlds-catching-last.

"Students in these new-style remedial-ed courses" — Bill Gates, keynote speech, Association of Community College Trustees, 44th Annual Leadership Congress, Seattle, WA, October 2, 2013, http://www.gatesfoundation.org/Media-Center/Speeches/2013/10/Bill-Gates-Association-of-Community-College-Trustees.

School of One — Mary Ann Wolf, *Innovate to Educate: System [Re]Design for Personalized Learning — A Report from the 2010 Symposium* (Washington, DC: Software & Information Industry Association [SIIA], 2010), p. 19, http://siia.net/pli/presentations/PerLearnPaper.pdf.

3. PLATFORMS

23 The story of Salman Khan — History of Khan Academy from Khan's *The One World Schoolhouse: Education Reimagined* (New York: Twelve, 2012), and from interview with Khan, "The Accidental Innovator," *HBS Alumni Bulletin,* March 1, 2012, https://www.alumni.hbs.edu/stories/Pages/story-bulletin.aspx?num=834.

50 million students — Metrics on Khan Academy from Khan Academy fact-sheet, December 2013, http://khanacademy.desk.com/customer/portal/articles/441307-press-room, and University of New Orleans, "Online Education Pioneer Salman Khan Wows Crowds at the University of New Orleans," May 21, 2013, http://www.uno.edu/news/2013/OnlineEducationPioneerSalmanKhanWowsCrowdsattheUniversityofNewOrleans.aspx.

"A mere convenience" — Khan, *The One World Schoolhouse*, p. 135. Emphasis in original.

24 Around 50 staff — Anya Kamenetz, "A Q&A with Salman Khan, Founder of Khan Academy," *Fast Company,* November 21, 2013, http://live.fastcompany.com/Event/A_QA_With_Salman_Khan.

"Every interaction with our system is logged" — Quoted in "The Accidental Innovator."

A statistical model of every student's accuracy — Jace Kohlmeier, "Khan Academy: Machine Learning — Measurable Learning," *Derandomized* (blog), May 10, 2013, http://derandomized.com/post/51729670543/khan-academy-machine-learning-measurable-learning.

25 A billion completed exercises — Talk by Jace Kohlmeier, Mixpanel Office Hours, San Francisco. Video uploaded March 25, 2013, http://www.youtube.com/watch?v=Nvoc6atMpAw#t=30.

28 All students everywhere can get lectures — Bill Gates, keynote speech, Association of Community College Trustees, 44th Annual Leadership Congress, October 2, 2013, http://www.gatesfoundation.org/Media-Center/Speeches/2013/10/Bill-Gates-Association-of-Community-College-Trustees.

29 Poised for disruption — For more on disruptive innovation in education, see Clayton M. Christensen's classic text *The Innovator's Dilemma: The Revolutionary Book That Will Change the Way You Do Business,* repr. (New York: HarperBusiness Essentials, 2011). Also Clayton M. Christensen, Michael B. Horn, and Curtis W. Johnson, *Disrupting Class: How Disruptive Innovation Will Change the Way the World Learns* (New York: McGraw-Hill, 2008).

30 MIT taskforce — *Institute-Wide Task Force on the Future of MIT Education: Preliminary Report,* November 21, 2013, http://future.mit.edu/preliminary-report.

31 A turning point in his thinking — Khan, *The One World Schoolhouse.*

4. CONSEQUENCES

33 ASU turned to adaptive-learning software — *Knewton Technology Helped More Arizona State University Students Succeed,* Knewton case study, 2013, http://www.knewton.com/assets-v2/downloads/asu-case-study.pdf.
Pass rates zipped up — Importantly, the improvements achieved may not all be attributable to Knewton's software: ASU changed its policies by letting students take the class over two semesters and retake the class for free. See Seth Fletcher, "Machine Learning," *Scientific American,* August 2013, http://www.nature.com/scientificamerican/journal/v309/n2/full/scientificamerican0813-62.html.
"We could raise the average accuracy" — Jace Kohlmeier, "Khan Academy: Machine Learning — Measurable Learning," *Derandomized* (blog), May 10, 2013, http://derandomized.com/post/51729670543/khan-academy-machine-learning-measurable-learning.

34 Huge stockpile of personal data — Natasha Singer, "Deciding Who Sees Students' Data," *New York Times,* October 5, 2013, http://www.nytimes.com/2013/10/06/business/deciding-who-sees-students-data.html.
"In an unprecedented move" — Corinne Lestch and Ben Chapman, "New York Parents Furious at Program, inBloom, That Compiles Private Student Information for Companies That Contract with It to Create Teaching Tools," *New York Daily News,* March 13, 2013, http://www.nydailynews.com/new-york/student-data-compiling-system-outrages-article-1.1287990.

36 Our brain constantly forgets — Viktor Mayer-Schönberger, *Delete: The Virtue of Forgetting in the Digital Age* (Princeton, NJ: Princeton University Press, 2009).

38 Experimenting with "e-advisors" — "Minding the Gap: Education Technology Helps Minorities Do Better at University," *Economist,* November 16, 2013, http://www.economist.com/news/united-states/21589924-education-technology-helps-minorities-do-better-university-minding-gap.

5. DAWN

43 "We take computational data" — Interview with Sebastián Díaz and his colleague, Phil Ice, November 2013. Since the research project described here began, Dr. Díaz was named Associate Vice President for Marketing Analytics at American Public University System (APUS). Dr. Ice, the Vice President of Research and Development at APUS, is the principal investigator of the project, called Predictive Analytics Reporting (PAR). See Phil Ice et al., "The PAR Framework Proof of Concept: Initial Findings from a Multi-Institutional Analysis of Federated Postsecondary Data,"

Journal of Asynchronous Learning Networks 16, no. 3 (June 2012), http://sloanconsortium.org/jaln/v16n3/par-framework-proof-concept-initial-findings-multi-institutional-analysis-federated-posts.

44 33 variables for analysis — Beth Davis, Sandy Daston, Dave Becher, and Jonathan Sherrill, *3 Million Course Records Walk Into an IRB Meeting: PAR — How We Did It* (Boulder, CO: WCET, October 2011), http://wcet.wiche.edu/wcet/docs/par/3MillionCourseRecordsPAR.pdf.

Implications for public policy — Paul Fain, "Using Big Data to Predict Online Student Success," *Inside Higher Ed,* February 1, 2012, http://www.insidehighered.com/print/news/2012/02/01/using-big-data-predict-online-student-success?width=775&height=500&iframe=true.

45 "An automated personal tutor" — David Rowan, "Online Education Is Redefining Learning Itself, Says Khan Academy Founder," *Wired UK,* August 27, 2013, http://www.wired.co.uk/magazine/archive/2013/08/start/reboot-the-teacher.

49 The age of big data will be one of continuous learning — This notion is not too far off the idea behind a method popular in big-data circles, Bayesian statistics, in which one constantly learns from additional data.

"Imagination is more important" — "What Life Means to Einstein: An Interview by George Sylvester Viereck," *Saturday Evening Post,* October 26, 1929, http://www.saturdayeveningpost.com/wp-content/uploads/satevepost/what_life_means_to_einstein.pdf.