Conclusion

Five questions that statistics can help answer

Not that long ago, information was much harder to gather and far more expensive to analyze. Imagine studying the information from one million credit card transactions in the era—only a few decades back—when there were merely paper receipts and no personal computers for analyzing the accumulated data. During the Great Depression, there were no official statistics with which to gauge the depth of the economic problems. Government did not collect official information on either gross domestic product (GDP) or unemployment, meaning that politicians were attempting to do the economic equivalent of navigating through a forest without a compass. Herbert Hoover declared that the Great Depression was over in 1930, on the basis of the inaccurate and outdated data that were available. He told the country in his State of Union address that two and a half million Americans were out of work. In fact, five million Americans were jobless, and unemployment was climbing by one hundred thousand every week. As James Surowiecki recently observed in *The New Yorker*, "Washington was making policy in the dark."[1]

We are now awash in data. For the most part, that is a good thing. The statistical tools introduced in this book can be used to address some

of our most significant social challenges. In that vein, I thought it fitting to finish the book with questions, not answers. As we try to digest and analyze staggering quantities of information, here are five important (and admittedly random) questions whose socially significant answers will involve many of the tools introduced in this book.

WHAT IS THE FUTURE OF FOOTBALL?

In 2009, Malcolm Gladwell posed a question in a *New Yorker* article that first struck me as needlessly sensationalist and provocative: How different are dog fighting and football?[2] The connection between the two activities stemmed from the fact that quarterback Michael Vick, who had served time in prison for his involvement in a dog-fighting ring, had been reinstated in the National Football League just as information was beginning to emerge that football-related head trauma may be associated with depression, memory loss, dementia, and other neurological problems later in life. Gladwell's central premise was that both professional football and dog fighting are inherently devastating to the participants. By the end of the article, I was convinced that he had raised an intriguing point.

Here is what we know. There is mounting evidence that concussions and other brain injuries associated with playing football can cause serious and permanent neurological damage. (Similar phenomena have been observed in boxers and hockey players.) Many prominent former NFL players have shared publicly their post-football battles with depression, memory loss, and dementia. Perhaps the most poignant was Dave Duerson, a former safety and Super Bowl winner for the Chicago Bears, who committed suicide by shooting himself in the chest; he left explicit instructions for his family to have his brain studied after his death.

In a phone survey of a thousand randomly selected former NFL players who had played at least three years in the league, 6.1 percent of the former players over fifty reported that they had received a diagnosis of "dementia, Alzheimer's disease, or other memory-related disease." That's five times the national average for that age group. For younger players, the rate of diagnosis was nineteen times the national average. Hundreds of former NFL players have now sued both the league and the makers of football helmets for allegedly hiding information about the dangers of head trauma.[3]

One of the researchers studying the impacts of brain trauma is Ann McKee, who runs the neuropathology laboratory at the Veterans Hospital in Bedford, Massachusetts. (Coincidentally, McKee also does the neuropathology work for the Framingham Heart Study.) Dr. McKee has documented the buildup of abnormal proteins called tau in the brains of athletes who have suffered brain trauma, such as boxers and football players. This leads to a condition known as chronic traumatic encephalopathy, or CTE, which is a progressive neurological disorder that has many of the same manifestations as Alzheimer's.

Meanwhile, other researchers have been documenting the connection between football and brain trauma. Kevin Guskiewicz, who runs the Sports Concussion Research Program at the University of North Carolina, has installed sensors on the inside of the helmets of North Carolina football players to record the force and nature of blows to the head. According to his data, players routinely receive blows to the head with a force equivalent to hitting the windshield in a car crash at twenty-five miles per hour.

Here is what we don't know. Is the brain injury evidence uncovered so far representative of the long-term neurological risks that all professional football players face? Or might this just be a "cluster" of adverse outcomes that is a statistical aberration? Even if it turns out that football players do face significantly higher risks of neurological disorder later in life, we would still have to probe the causality. Might the kind of men who play football (and boxing and hockey) be prone to this kind of problem? Is it possible that some other factors, such as steroid use, are contributing to the neurological problems later in life?

If the accumulating evidence does suggest a clear, causal link between playing football and long-term brain injury, one overriding question will have to be addressed by players (and the parents of younger players), coaches, lawyers, NFL officials, and perhaps even government regulators: Is there a way to play the game of football that reduces most or all of the head trauma risk? If not, then what? This is the point behind Malcolm Gladwell's comparison of football and dog fighting. He explains that

dog fighting is abhorrent to the public because the dog owner willingly submits his dog to a contest that culminates in suffering and destruction. "And why?" he asks. "For the entertainment of an audience and the chance of a payday. In the nineteenth century, dog fighting was widely accepted by the American public. But we no longer find that kind of transaction morally acceptable in a sport."

Nearly every kind of statistical analysis described in this book is currently being used to figure out whether or not professional football as we know it now has a future.

WHAT (IF ANYTHING) IS CAUSING THE DRAMATIC RISE IN THE INCIDENCE OF AUTISM?

In 2012, the Centers for Disease Control reported that 1 in 88 American children has been diagnosed with an autism spectrum disorder (on the basis of data from 2008).[4] The rate of diagnosis had climbed from 1 in 110 in 2006, and 1 in 150 in 2002—or nearly a doubling in less than a decade. Autism spectrum disorders (ASDs) are a group of developmental disabilities characterized by atypical development in socialization, communication, and behavior. The "spectrum" indicates that autism encompasses a broad range of behaviorally defined conditions.[5] Boys are five times as likely to be diagnosed with an ASD as girls (meaning that the incidence for boys is even higher than 1 in 88).

The first intriguing statistical question is whether we are experiencing an epidemic of autism, an "epidemic of diagnosis," or some combination of the two?[6] In previous decades, children with an autism spectrum disorder had symptoms that might have gone undiagnosed, or their developmental challenges might have been described more generally as a "learning disability." Doctors, parents, and teachers are now much more aware of the symptoms of ASDs, which naturally leads to more diagnoses regardless of whether or not the incidence of autism is on the rise.

In any case, the shockingly high incidence of ASDs represents a serious challenge for families, for schools, and for the rest of society. The average lifetime cost of managing an autism spectrum disorder for a single individual is $3.5 million.[7] Despite what is clearly an epidemic, we know amazingly little about what causes the condition. Thomas Insel, director of the National Institute of Mental Health, has said, "Is it cell phones? Ultrasound? Diet sodas? Every parent has a theory. At this point, we just don't know."[8]

What is different or unique about the lives and backgrounds of children with ASDs? What are the most significant physiological differences between children with and without an ASD? Is the incidence of ASDs different acrqss countries? If so, why? Traditional statistical detective work is turning up clues.

One recent study by researchers at the University of California at Davis identified ten locations in California with autism rates that are double the rates of surrounding areas; each of the autism clusters is a neighborhood with a concentration of white, highly educated parents.[9] Is that a clue, or a coincidence? Or might it reflect that relatively privileged families are more likely to have an autism spectrum disorder diagnosed? The same researchers are also conducting a study in which they will collect dust samples from the homes of 1,300 families with an autistic child to test for chemicals or other environmental contaminants than may play a causal role.

Meanwhile, other researchers have identified what appears to be a genetic component to autism by studying ASDs among identical and fraternal twins.[10] The likelihood that two children in the same family have an ASD is higher among identical twins (who share the same genetic makeup) than among fraternal twins (whose genetic similarity is the same as for regular siblings). This finding does not rule out significant environmental factors, or perhaps the interaction between environmental and genetic factors. After all, heart disease has a significant genetic component, but clearly smoking, diet, exercise, and many other behavioral and environmental factors all matter, too.

One of the most important contributions of statistical analysis so far has been to debunk false causes, many of which have arisen because of a confusion between correlation and causation. An autism spectrum disorder often appears suddenly between a child's first and second birthdays. This has led to a widespread belief that childhood vaccinations, particularly the triple vaccine for measles, mumps, and rubella (MMR), are caus-

ing the rising incidence of autism. Dan Burton, a member of Congress from Indiana, told the *New York Times*, "My grandson received nine shots in one day, seven of which contained thimerosal, which is 50 percent mercury as you know, and he became autistic a short time later."[11]

Scientists have soundly refuted the false association between thimerosal and ASDs. Autism rates did not decline when thimerosal was removed from the MMR vaccine, nor are autism rates lower in countries that never used this vaccine. Nonetheless, the false connection persists, which has caused some parents to refuse to vaccinate their children. Ironically, this offers no protection against autism while putting children at risk of contracting other serious diseases (and contributing to the spread of those diseases in the population).

Autism poses one of the greatest medical and social challenge of our day. We understand so little about the disorder relative to its huge (and possibly growing) impact on our collective well-being. Researchers are using every tool in this book (and lots more) to change that.

HOW CAN WE IDENTIFY AND REWARD GOOD TEACHERS AND SCHOOLS?

We need good schools. And we need good teachers in order to have good schools. Thus, it follows logically that we ought to reward good teachers and good schools while firing bad teachers and closing bad schools.

How exactly do we do that?

Test scores give us an objective measure of student performance. Yet we know that some students will do much better on standardized tests than others for reasons that have nothing to do with what is going on inside a classroom or a school. The seemingly simple solution is to evaluate schools and teachers on the basis of the *progress* that their students make over some period of time. What did students know when they started in a certain classroom with a particular teacher? What did they know a year later? The difference is the "value added" in that classroom.

We can even use statistics to get a more refined sense of this value added by taking into account the demographic characteristics of the students in a given classroom, such as race, income, and performance on other tests (which can be a measure of aptitude). If a teacher makes significant gains with students who have typically struggled in the past, then he or she can be deemed as highly effective.

Voilà! We can now evaluate teacher quality with statistical precision. And the good schools, of course, are just the ones full of effective teachers.

How do these handy statistical evaluations work in practice? In 2012, New York City took the plunge and published ratings of all 18,000 public school teachers on the basis of a "value-added assessment" that measured gains in their students' test scores while taking into account various student characteristics.[12] The *Los Angeles Times* published a similar set of rankings for Los Angeles teachers in 2010.

In both New York and LA, the reaction has been loud and mixed. Arne Duncan, the U.S. secretary of education, has generally been supportive of these kinds of value-added assessments. They provide information where none previously existed. After the Los Angeles data were published, Secretary Duncan told the *New York Times*, "Silence is not an option." The Obama administration has provided financial incentives for states to develop value-added indicators for paying and promoting teachers. Proponents of these evaluation measures rightfully point out that they are a huge potential improvement over systems in which all teachers are paid according to a uniform salary schedule that gives zero weight to any measure of performance in the classroom.

On the other hand, many experts have warned that these kinds of teacher assessment data have large margins of error and can deliver misleading results. The union representing New York City teachers spent more than $100,000 on a newspaper advertising campaign built around the headline "This Is No Way to Rate a Teacher."[13] Opponents argue that the value-added assessments create false precision that will be abused by parents and public officials who do not understand the limitations of this kind of assessment.

This appears to be a case where everybody is right—up to a point. Doug Staiger, an economist at Dartmouth College who works extensively with value-added data for teachers, warns that these data are inherently "noisy." The results for a given teacher are often based on a single test taken on a single day by a single group of students. All kinds of factors can

lead to random fluctuations—anything from a particularly difficult group of students to a broken air-conditioning unit clanking away in the classroom on test day. The correlation in performance from year to year for a single teacher that uses these indicators is only about .35. (Interestingly, the correlation in year-to-year performance for Major League baseball players is also around .35, as measured by batting average for hitters and earned run average for pitchers.)[14]

The teacher effectiveness data are useful, says Staiger, but they are just one tool in the process for evaluating teacher performance. The data get "less noisy" when authorities have more years of data for a particular teacher with different classrooms of students (just as we can tell more about an athlete when we have data for more games and more seasons). In the case of the New York City teacher ratings, principals in the system had been prepped on the appropriate use of the value-added data and the inherent limitations. The public did not get that briefing. As a result, the teacher assessments are too often viewed as a definitive guide to the "good" and "bad" teachers. We like rankings—just think *U.S. News & World Report* college rankings—even when the data do not support such precision.

Staiger offers a final warning of a different sort: We had better be certain that the outcomes we are measuring, such as the results of a given standardized test, truly track with what we care about in the long run. Some unique data from the Air Force Academy suggest, not surprisingly, that the test scores that glimmer now may not be gold in the future. The Air Force Academy, like the other military academies, randomly assigns its cadets to different sections of standardized core courses, such as introductory calculus. This randomization eliminates any potential selection effect when comparing the effectiveness of professors; over time, we can assume that all professors get students with similar aptitudes (unlike most universities, where students of different abilities can select into or out of different courses). The Air Force Academy also uses the same syllabi and exams in every section of a particular course. Scott Carrell and James West, professors at the University of California at Davis and the Air Force Academy, exploited this elegant arrangement to answer one of the most important questions in higher education: Which professors are most effective?[15]

The answer: *The professors with less experience and fewer degrees from fancy universities.* These professors have students who typically do better on the standardized exams for the introductory courses. They also get better student evaluations for their courses. Clearly these young, motivated instructors are more committed to their teaching than the old, crusty professors with PhDs from places like Harvard. The old guys must be using the same yellowing teaching notes that they used in 1978; they probably think PowerPoint is an energy drink—except that they don't know what an energy drink is either. Obviously the data tell us that we should fire these old codgers, or at least let them retire gracefully.

But hold on. Let's not fire anybody yet. The Air Force Academy study had another relevant finding—about student performance *over a longer horizon.* Carrell and West found that in math and science the students who had more experienced (and more highly credentialed) instructors in the introductory courses *do better in their mandatory follow-on courses* than students who had less experienced professors in the introductory courses. One logical interpretation is that less experienced instructors are more likely to "teach to the test" in the introductory course. This produces impressive exam scores and happy students when it comes to filling out the instructor evaluation.

Meanwhile, the old, crusty professors (whom we nearly fired just one paragraph ago) focus less on the exam and more on the important concepts, which are what matter most in follow-on courses and in life after the Air Force Academy.

Clearly we need to evaluate teachers and professors. We just have to make sure that we do it right. The long-term policy challenge, rooted in statistics, is to develop a system that rewards a teacher's real value added in the classroom.

WHAT ARE THE BEST TOOLS FOR FIGHTING GLOBAL POVERTY?

We know strikingly little about how to make poor countries less poor. True, we understand the things that distinguish rich countries from poor countries, such as their education levels and the quality of their governments. And it is also true that we have watched countries like India and China transform themselves economically over the last several decades.

But even with this knowledge, it is not obvious what steps we can take to make places like Mali or Burkina Faso less poor. Where should we begin?

French economist Esther Duflo is transforming our knowledge of global poverty by retrofitting an old tool for new purposes: the randomized, controlled experiment. Duflo, who teaches at MIT, literally conducts experiments on different interventions to improve the lives of the poor in developing countries. For example, one of the longstanding problems with schools in India is absenteeism among teachers, particularly in small, rural schools with only a single teacher. Duflo and her coauthor Rema Hanna tested a clever, technology-driven solution on a random sample of 60 one-teacher schools in the Indian state of Rajasthan.[16] Teachers in these 60 experimental schools were offered a bonus for good attendance. Here is the creative part: The teachers were given cameras with tamper-proof date and time stamps. They proved that they had showed up each day by having their picture taken with their students.[17]

Absenteeism dropped by half among teachers in the experimental schools compared with teachers in a randomly selected control group of 60 schools. Student test scores went up, and more students graduated into the next level of education. (I bet the photos are adorable, too!)

One of Duflo's experiments in Kenya involved giving a randomly selected group of farmers a small subsidy to buy fertilizer right *after* the harvest. Prior evidence suggested that fertilizer raises crop yields appreciably. Farmers were aware of this benefit, but when it came time to put a new crop into the ground, they often did not have enough money left over from the last crop to buy fertilizer. This perpetuates what is known as a "poverty trap" since the subsistence farmers are too poor to make themselves less poor. Duflo and her coauthors found that a tiny subsidy—free fertilizer delivery—offered to farmers when they still had cash after the harvest increased fertilizer use by 10 to 20 percentage points compared with use in a control group.[18]

Esther Duflo has even waded into the gender war. Who is more responsible when it comes to handling the family's finances, men or women? In rich countries, this is the kind of thing that couples can squabble over in marriage counseling. In poor countries, it can literally determine whether the children get enough to eat. Anecdotal evidence

going back to the dawn of civilization suggests that women place a high priority on the health and welfare of their children, while men are more inclined to drink up their wages at the local pub (or whatever the caveman equivalent was). At worst, this anecdotal evidence merely reinforces age-old stereotypes. At best, it is a hard thing to prove, because a family's finances are comingled to some extent. How can we separate out how husbands and wives choose to spend communal resources?

Duflo did not shy away from this delicate question.[19] To the contrary, she found a fascinating natural experiment. In Côte d'Ivoire, women and men in a family typically share responsibility for some crops. For long-standing cultural reasons, men and women also cultivate different cash crops of their own. (Men grow cocoa, coffee, and some other things; women grow plantains, coconuts, and a few other crops.) The beauty of this arrangement from a research standpoint is that the men's crops and the women's crops respond to rainfall patterns in different ways. In years in which cocoa and coffee do well, men have more disposable income to spend. In years in which plantains and coconuts do well, the women have more extra cash.

Now we need merely broach a delicate question: Are the children in these families better-off in years in which the men's crops do well, or in the years when the women have a particularly bountiful harvest?

The answer: When the women do well, they spend some of their extra cash on more food for the family. The men don't. Sorry guys.

In 2010, Duflo was awarded the John Bates Clark Medal. This prize is presented by the American Economic Association to the best economist under the age of forty.* Among economist geeks, this prize is considered to be more prestigious than the Nobel Prize in Economics because it was historically awarded only every two years. (Beginning with Duflo's award in 2010, the medal is now presented annually.) In any case, the Clark Medal is the MVP award for people with thick glasses (metaphorically speaking).

Duflo is doing program evaluation. Her work, and the work of

* I was ineligible for the 2010 prize since I was over forty. Also, I'd done nothing to deserve it.

others now using her methods, is literally changing the lives of the poor. From a statistical standpoint, Duflo's work has encouraged us to think more broadly about how randomized, controlled experiments—long thought to be the province of the laboratory sciences—can be used more widely to tease out causal relationships in many other areas of life.

WHO GETS TO KNOW WHAT ABOUT YOU?

Last summer, we hired a new babysitter. When she arrived at the house, I began to explain our family background: "I am a professor, my wife is a teacher. . . ."

"Oh, I know," the sitter said with a wave of the hand. "I Googled you."

I was simultaneously relieved that I did not have to finish my spiel and mildly alarmed by how much of my life could be cobbled together from a short Internet search. Our capacity to gather and analyze huge quantities of data—the marriage of digital information with cheap computing power and the Internet—is unique in human history. We are going to need some new rules for this new era.

Let's put the power of data in perspective with just one example from the retailer Target. Like most companies, Target strives to increase profits by understanding its customers. To do that, the company hires statisticians to do the kind of "predictive analytics" described earlier in the book; they use sales data combined with other information on consumers to figure out who buys what and why. Nothing about this is inherently bad, for it means that the Target near you is likely to have exactly what you want.

But let's drill down for a moment on just one example of the kinds of things that the statisticians working in the windowless basement at corporate headquarters can figure out. Target has learned that pregnancy is a particularly important time in terms of developing shopping patterns. Pregnant women develop "retail relationships" that can last for decades. As a result, Target wants to identify pregnant women, particularly those in their second trimester, and get them into their stores more often. A

writer for the *New York Times Magazine* followed the predictive analytics team at Target as it sought to find and attract pregnant shoppers.[20]

The first part is easy. Target has a baby shower registry in which pregnant women register for baby gifts in advance of the birth of their children. These women are already Target shoppers, and they've effectively told the store that they are pregnant. But here is the statistical twist: *Target figured out that other women who demonstrate the same shopping patterns are probably pregnant, too.* For example, pregnant women often switch to unscented lotions. They begin to buy vitamin supplements. They start buying extrabig bags of cotton balls. The Target predictive analytics gurus identified twenty-five products that together made possible a "pregnancy prediction score." The whole point of this analysis was to send pregnant women pregnancy-related coupons in hopes of hooking them as long-term Target shoppers.

How good was the model? The *New York Times Magazine* reported a story about a man from Minneapolis who walked into a Target store and demanded to see a manager. The man was irate that his high school daughter was being bombarded with pregnancy-related coupons from Target. "She's still in high school and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?" the man asked.

The store manager apologized profusely. He even called the father several days later to apologize again. Only this time, the man was less irate; it was his turn to be apologetic. "It turns out there's been some activities in my house I haven't been completely aware of," the father said. "She's due in August."

The Target statisticians had figured out that his daughter was pregnant before he did.

That is their business . . . and also not their business. It can feel more than a little intrusive. For that reason, some companies now mask how much they know about you. For example, if you are a pregnant woman in your second trimester, you may get some coupons in the mail for cribs and diapers—along with a discount on a riding lawn mower and a coupon for free bowling socks with the purchase of any pair of bowling shoes. To you, it just seems fortuitous that the pregnancy-related coupons came in

the mail along with the other junk. In fact, the company knows that you don't bowl or cut your own lawn; it's merely covering its tracks so that what it knows about you doesn't seem so spooky.

Facebook, a company with virtually no physical assets, has become one of the most valuable companies in the world. To investors (as opposed to users), Facebook has one enormous asset: data. Investors don't love Facebook because it allows them to reconnect with their prom dates. They love Facebook because every click of the mouse yields data about where users live, where they shop, what they buy, who they know, and how they spend their time. To users, who *are* hoping to reconnect with their prom dates, the corporate data gathering can overstep the boundaries of privacy.

Chris Cox, Facebook's vice president of product, told the *New York Times*, "The challenge of the information age is what to do with it."[21]

Yep.

And in the public arena, the marriage of data and technology gets even trickier. Cities around the world have installed thousands of security cameras in public places, some of which will soon have facial recognition technology. Law enforcement authorities can follow any car anywhere it may go (and keep extensive records of where it has been) by attaching a global positioning device to the vehicle and then tracking it by satellite. Is this a cheap and efficient way to monitor potential criminal activity? Or is this the government using technology to trample on our personal liberty? In 2012, the U.S. Supreme Court decided unanimously that it was the latter, ruling that law enforcement officials can no longer attach tracking devices to private vehicles without a warrant.*

Meanwhile, governments around the world maintain huge DNA databases that are a powerful tool for solving crimes. Whose DNA should be in the database? That of all convicted criminals? That of every person who is arrested (whether or not eventually convicted)? Or a sample from every one of us?

We are just beginning to wrestle with the issues that lie at the intersection of technology and personal data—none of which were terribly relevant when government information was stored in dusty basement filing cabinets rather than in digital databases that are potentially searchable by anyone from anywhere. Statistics is more important than ever before because we have more meaningful opportunities to make use of data. Yet the formulas will not tell us which uses of data are appropriate and which are not. Math cannot supplant judgment.

* *The United States v. Jones.*

In that vein, let's finish the book with some word association: fire, knives, automobiles, hair removal cream. Each one of these things serves an important purpose. Each one makes our lives better. And each one can cause some serious problems when abused.

Now you can add statistics to that list. Go forth and use data wisely and well!