DATA AND THE NEW WORLD OF EMPIRICAL SCHOLARSHIP

BY ROBIN I. MORDFIN
The rise of data-driven research has provided a dramatic new avenue of exploration for the prodigious research appetites of the University of Chicago Law School faculty. With enormous storage capacity and unprecedented speed, computer technology has made projects using millions of pieces of data possible, which also means that undertakings that involve hundreds of constitutions or hundreds of thousands of cases can be pursued.

“It used to be that when we wanted to do research on cases in a particular area, we would actually have to go down to the courthouses and haul out these really large books that had all the information in them. Then we would have to copy down the information we wanted to use and then take it back to school to figure out how to analyze it,” noted Thomas Miles, Clifton R. Musser Professor of Law and Economics. “It was very time consuming and limiting.”

The Law School is a pioneer and continual leader in empirical research. For more than seven decades, the Law School has led the way in Law and Economics, and in recent years it has made a substantial commitment in training and hiring the very best empirical faculty and providing them the resources they need to do the most cutting-edge empirical work.

Of course, having the technology and other resources does not mean that such research is simple: it is still necessary to manipulate massively complex collections of information. Fortunately, the Coase-Sandor Institute for Law and Economics employs savvy economists and statisticians who know how to apply advanced data and exploration tools that make these enormous data sets useful.

This empirical work is in the finest Coasean tradition: Ronald Coase was always concerned about the real world and the impact economics scholarship had on real problems. He came from a long tradition of British empiricism, and much of his most influential work dealt not with high theory but with complex practical problems such as the allocation of the electromagnetic spectrum.

The experts from the Coase-Sandor Institute work with the data sets and with the Law School’s scholars to bring these complex practically oriented projects to fruition. “Their work is so careful and so smart, it is just a pleasure to work with them,” Miles added. “Without them, I doubt I could do the work I am doing.”

When approaching a data-driven research project, the first step is to put whatever information is available into a workable form. “It all starts with the condition the data is in to begin with. It would be great if everything was downloadable into an Excel spreadsheet, but there is a lot of data that is in text, especially older data, that cannot be downloaded.” explained Daniel Marcin, a research professional with the Coase-Sandor Institute with a PhD in economics. “So we manually enter all of the information.”

Once the data is in an electronic form, it can be coded by assigning numbers to words or phrases, then the researchers determine which elements they want for their specific questions. When the coding is complete, other researchers can use the data, even those who are using different elements, because the data set has been coded with information that reaches beyond the specific question the initiating scholar is asking.

It would be great if everything was downloadable into an Excel spreadsheet, but there is a lot of data that is in text that cannot be downloaded. ... So we manually enter all of the information.

After coding, research specialists use available programs that analyze language or mood or a variety of other topics. “But of course they are never turnkey, they never work right away,” Marcin added. “You still have to write programs, or amend programs, on your own to make things all work together.”

One of the projects the researchers have been working on is Miles’s research on the Secure Communities program, a federal program launched at the end of 2008 that permits federal authorities to check the immigration status of every person arrested by local police. “The program helps the Department of Homeland Security to locate individuals who have been arrested by local police for crimes and allows them to determine the arrestee’s immigration status,” Miles explained. Previously, if an immigrant had been arrested for a crime, they would actually have to send an immigration enforcement agent to the local jail to scour the records and interview people. The new program streamlined the whole process by sending the fingerprints over to Homeland Security immediately. Critics argued that people who were getting systematically pulled in under the program for deportation proceedings would previously have been unknown to Homeland Security and would not have been prioritized for deportation.
The government claimed that Secure Communities was not an immigration enforcement program but rather was created to help with crime control. So Miles and his co-researcher, Adam Cox of New York University Law School, set out to determine which claim was the truth. The program was rolled out, for technological reasons, across the nation by county, rather than all at once. This was useful for research because it gave the researchers the opportunity to compare counties that received the program and those that hadn’t. It also allowed them to make comparisons within a county both before and after implementation.

Miles and Cox acquired some of their millions of pieces of data through Freedom of Information Act requests and from the program itself, which supplied information such as how many arrestees were transferred from local to federal custody and how many were deported. They also used the FBI’s Uniform Crime Reports data set, which provides information on the number of crimes reported and the number of people arrested for those crimes by county and by month across the country.

“Once we had the data set assembled, the first thing we did was try to figure out, since Homeland Security had the discretion to determine where they were going to roll out the program first and where later, what their priorities were. Therefore, if this really were a crime control program, you would think they would have activated first in places with the highest crime levels. And if it’s really an immigration enforcement program, they should turn it on first in places that have the largest fraction of immigrants. And what we found was really the latter,” Miles noted.

The program began in places that were close to the southern border with Mexico, and once they controlled for basic demographic characteristics including race and age across the nation, there was practically no impact on crime rates. They therefore concluded that Secure Communities was in fact an immigration enforcement program and published their findings in the University of Chicago Law Review. They next looked at whether there were any crime differences in a particular county before or after the program was activated there. Homeland Security had created four categories for those who were arrested: one category was noncriminal, and the other three were for different types of criminals.

“Homeland Security was really supposed to move to incapacitate only the most dangerous offenders and not bother with minor offenders. Therefore, we were looking for a positive change in that crime in those communities that had a lot of crime before the program began. But what we found was that there was no evidence of a decline in either violent or property crime,” Miles explained. That research is published in the November 2014 edition of the Journal of Law and Economics.

The duo is currently working on two additional papers connected to the Secure Communities data. First, they are looking at FBI crime statistics in Secure Communities counties to see if there has been a shift in the nature of the offenses for which local police make arrests. If the critics are correct, police are engaging in more sweeps of low-level offenders in order to apprehend more Hispanics and get their fingerprints checked. This would mean that less serious offenses should comprise a larger share of arrests. They are still examining that data to see if such a change has taken place.

Miles and Cox are also looking at the claim that Secure Communities really makes immigrants less willing to cooperate with the police. Therefore, they are looking at clearance rates of crimes—how many reported crimes result in arrests—in areas with high numbers of immigrants. So far, no one else has considered whether the program would create a backlash that would reduce cooperation. If the program is detrimental to trust, there should be a change in the clearance rate in areas like El Paso but not in areas like New Hampshire.

“But we found an interesting result, which is that we don’t see any reaction from the program on clearance rates, which surprised us. But the reality is that in these communities with large foreign-born population, clearance rates are already low, and there is no drop because that population doesn’t cooperate with law enforcement very much to begin with.” Miles said. “However, in the president’s recent speech on immigration reform, he announced that he is canceling Secure Communities, but what that really means is that it will be reworked in some way and the name will be changed. We will have to see what it becomes.”

While Miles and Cox were looking at a program that was well underway, Anup Malani, Lee and Brena Freeman Professor of Law, is looking at a very large program as it rolls out for a new population. Malani, with the help of eleven other researchers from around the world, is looking at whether and how the Indian government should provide universal health insurance to its population. In India, nearly 63 million people are forced into poverty because of medical expenses each year. In 2008, the Indian government began offering health insurance to the poorest quartile of its citizens, 300 million people, through a program called Rashtriya Swasthya Bima Yojna (RSBY). By 2012, the program was already covering more than 150 million persons. But those just above the poverty level are still not covered.
“So we reached an agreement with RSBY and the state of Karnataka to use RSBY insurance to study the effects of public health insurance on health and poverty and how different types of coverage expansions affect insurance uptake and government expenditures,” Malani said. “We are conducting a large, randomized control trial of different insurance options. The study has enrolled roughly 60,000 people in 12,000 households, making it one of the largest social science experiments ever conducted. We are looking at two different districts, one in Central India and one in Southern India, one that is more impoverished and one that is less impoverished, so we can generalize our results to other regions of the country, and indeed to other lower-income countries. Clearly, this is a big job, and we are fortunate to have some of the best economists and statisticians in the world working on this, as well as the cooperation of the Indian government and researchers.”

The study is focusing on families who are above the poverty line to figure out whether having insurance will prevent households from falling into the bottom quartile when hit with a medical emergency. The hope is that these people will benefit from having insurance and will not have to sell income-producing assets to get through the crisis. Perhaps having this cushion will also enable some members of the population to make seemingly risky investments such as education or starting a new business that can even improve their financial standing.

The study also examines the impacts of insurance on health. “For example, are there people who were unable to finance needed hospital treatment and therefore avoided it? Will having insurance change their decision and improve their health?” Malani explained. They are also looking at some novel, secondary outcomes, such as cognitive capacity. The researchers conjecture that serious illness—even nonmental illness—may erode a person’s cognitive capacity, causing him to make other bad decisions, which can lead to economic impoverishment on top of illness. The team will examine whether the insurance will ameliorate these impacts because the person will have less to worry about, at least from a financial point of view.

The researchers are primarily using surveys, over the course of the three-year study, that break down all the participants by sex, age, location, income, and many other details. The study will take place in three stages—enrollment, midline, and at the endpoint—and will ask participants to answer questions about the assets, liabilities, and consumption practices of each family. They will also look at the willingness of participants to pay for insurance, the state of health and healthcare participation, and their cognitive states. Right now the team is currently finishing the enrollment phase of the study.

“Imagine how much we could improve health and welfare in India if we had reliable evidence on the impact of different policy reforms. Imagine how we could change policy and meet people’s needs,” Malani exclaimed. The research is certainly timely, as the new government of Prime Minister Modi announced a plan to provide universal coverage in India in the next few years, perhaps by opening the RSBY program to all citizens. The team

Imagine how much we could improve health and welfare in India if we had reliable evidence on the impact of different policy reforms.
has already begun discussions with the Indian government about what can be learned from their study.

Both Malani and Miles earned PhDs in economics along with their JDs, which puts them at the forefront of Law and Economics research. They thrive at the intersection of economic practice and legal scholarship. Another JD/PhD member of the Law School faculty is William Hubbard, who spent the early parts of his career examining education and the returns on education, specifically legal education. But in the past few years, his attention has turned toward civil procedure.

While we take notoriously messy data and help to create clean data sets, the selection of what to leave in and what to take out is critical.

“My first interest was in what kind of effects do Supreme Court decisions have on the actual practice of attorneys on the ground, in terms of how litigators change their behavior in response to what are perceived as major Supreme Court decisions,” Hubbard explained.

Hubbard’s first major project looked at pleading standards, the standards that courts can use to dismiss civil cases that have been filed without having a trial or any further proceedings. This particular issue came to a head in 2007 when the Supreme Court decided *Bell Atlantic v. Twombly*, which heightened the pleading requirement for federal civil cases and required that plaintiffs include enough facts in their complaints to make the case plausible so that they will be able to prove facts to support their claims.

“The decision was controversial because people were worried about plaintiffs who may have a real injury but who don’t have a lot of information about the issue and are hoping to get that information through the civil discovery process in the course of the lawsuit,” Hubbard said. “I was fairly skeptical that it would have a big impact because in my practice experience, plaintiff lawyers would write very detailed complaints, even though they weren’t required to, because they wanted to impress upon the judge and the defendant that it was a really good case and that you should take me seriously.”

Hubbard is using a data set from the Administrative Office of the US Courts that allows him to look at all the civil cases filed at the federal level since the *Twombly* decision. Each year, about 250,000 new cases are filed in the federal court system. Right now, Hubbard’s data is sitting on a special secure computer in the Coase-Sandor Institute. That computer has about 8 million observations on it.

“The data show that the case had a fairly minor effect on the filing of civil cases,” Hubbard added. “But that controversy continues to rage on to this day.”

As a result, Hubbard is working on a follow-up, by combining the Administrative Office data with docket numbers to generate a random sample of cases before and after the *Twombly* decision. He is then accessing the electronic court filings database called the PACER-ECF System to look up all the cases, which allows him to download all the complaints. He is now using text analysis software to see if the complaints have changed over time—if they have more adjectives, if they are longer, if they have more paragraphs.

“Of course, none of these systems are designed to be useful for academic research—they are basically just court records. So in order to turn them into something we would see as useful for quantitative analysis, we sometimes have to write code, sometimes using several different languages, to get the data from the Internet to our computers and turned into something that we can process.”

Even more data-driven projects are underway and on the horizon for the Law and Economics faculty at the Law School. But their work is not received without criticism.

“There is a concern that authors are reporting only the models that produce significant results, that there is a bias that we are only reporting the good stuff and not the bad stuff,” commented Tom Ginsburg, Leo Spitz Professor of International Law. “People will ask, how many regressions did you have to run before you got this one that turned out to have a statistically significant result.”

Others are concerned with which information is included in the studies and which ones are omitted. “The process of mining the data is the part that some economists don’t like,” explained Joseph Burton, Executive Director of the Coase-Sandor Institute. “They are looking for a sound theoretical basis for imposing the structure the researchers have chosen. It is necessary to have a behavioral reason to explain data, you can’t just draw a line between two elements and say there is a correlation. So while we take notoriously messy data and help to create clean data sets, the selection of what to leave in and what to take out is critical.”

As a result, researchers who are working data-driven projects are now working to be as transparent as possible, by even leaving a record of every keystroke that has been made in processing the data.
“Our data is online and anyone can download it for their own research. We have a code book that explains all of our decisions, and we want everyone looking at our result to understand the steps we took,” Ginsburg said of the Comparative Constitutions Project. Social scientists spend a lot of time thinking about issues of measurement. Unfortunately, according to Ginsburg, many of the new data projects are not as careful as they should be, either about the quality of the measurements they are taking or in the way in which they are interpreting their data.

“Think about the US News & World Report law school rankings, which purport to be a clear ordinal ranking of law school performance, but they obfuscate many relevant dimensions that people ought to know about when evaluating a law school. In some sense, it is the illusion of good data, but really bad indexing. There is a lot of virtue in being careful today,” he added.

Ginsburg and his research partners, Zachary Elkins of the University of Texas and James Melton of University College London, are using their carefully mined data to look at more than 900 constitutions that have been written since 1789. To conduct their research, they are considering three primary questions, the first of which is how do ideas from one constitution, from one area, spread to others? A second question was what makes some constitutions endure, while others do not? This was addressed in their award-winning 2010 book, The Endurance of National Constitutions. And a third is under what circumstances do constitutions actually work, do they actually change practice in government and behavior?

When beginning the project, the researchers did not know there were so many constitutions but forged ahead to find them and get them translated. “The main thing is that we got all those old texts and we digitized them and began to analyze them. The analysis involved actually creating data—you can think of our work as a giant spreadsheet with hundreds of questions in hundreds of rows for different countries in different years, and these intersect with thousand of columns that are the various provisions that might be present in a constitution,” Ginsburg said.

The team developed special software at the University of Illinois that included an online interface for coding that allows users to look up a constitution and then answer various questions about it in the system, which then stores the data so that it can be spit out in various forms for statistical analysis. They have already written 27 articles with this data, which they began to gather in 2007.

“We have had lots of different results. I recently finished a paper that shows that countries in which the legislature is involved in war-making decisions, such as by declaring war, get into fewer wars that countries without that. But I also found that when these same countries do get into a wars, they tend to lose them more often, which was slightly surprising,” Ginsburg commented. The team also discovered that the average predicted lifespan for a constitution is only 19 years for all countries, which makes the US constitution extraordinary. This means that American advisors should be careful in drawing on our experience in working with other countries that are trying to establish new constitutions, such as the countries of the Arab Spring.

Another issue under consideration is amendment difficulty. Interestingly, Ginsburg and his colleagues have found that the difficulty of amending a constitution doesn’t predict the frequency with which it is amended. So while the US constitution is very difficult to amend, it has been amended on 18 separate occasions. On the other hand, Japan’s constitution is relatively easy to amend and has never been changed.

Beyond constitutions, the trio is also looking at treaties, because they have found that elements that show up in international treaties tend to make an appearance in constitutions written after the treaties have been enacted. “So, for example, if a new human rights treaty is established, even countries that don’t sign the treaty will have a bill of rights that looks a lot like the treaty,” Ginsburg said.

In addition to the projects mentioned here, these academics have many other data-driven research studies underway. Miles is working on a paper that considers different ways of evaluating judicial performance by looking at how attorneys review judges. Ginsburg is working with the World Justice Project, which collects data on the perception of the rule of law in 99 different countries, while Hubbard is examining how using federal or state procedural rules affects civil cases under different circumstances.

“Both Law and Economics and cutting-edge scholarship are synonymous with the University of Chicago Law School, so it is unsurprising that the most innovative and fascinating work that data-driven research has made possible is taking place here,” remarked Dean Michael Schill. “Our faculty members both inside Law and Economics and in many other disciplines are taking on larger and larger data sets in order to find bigger, more substantial insights than were ever possible before. This work will not only change the law and our understanding of it but change the world as well.”