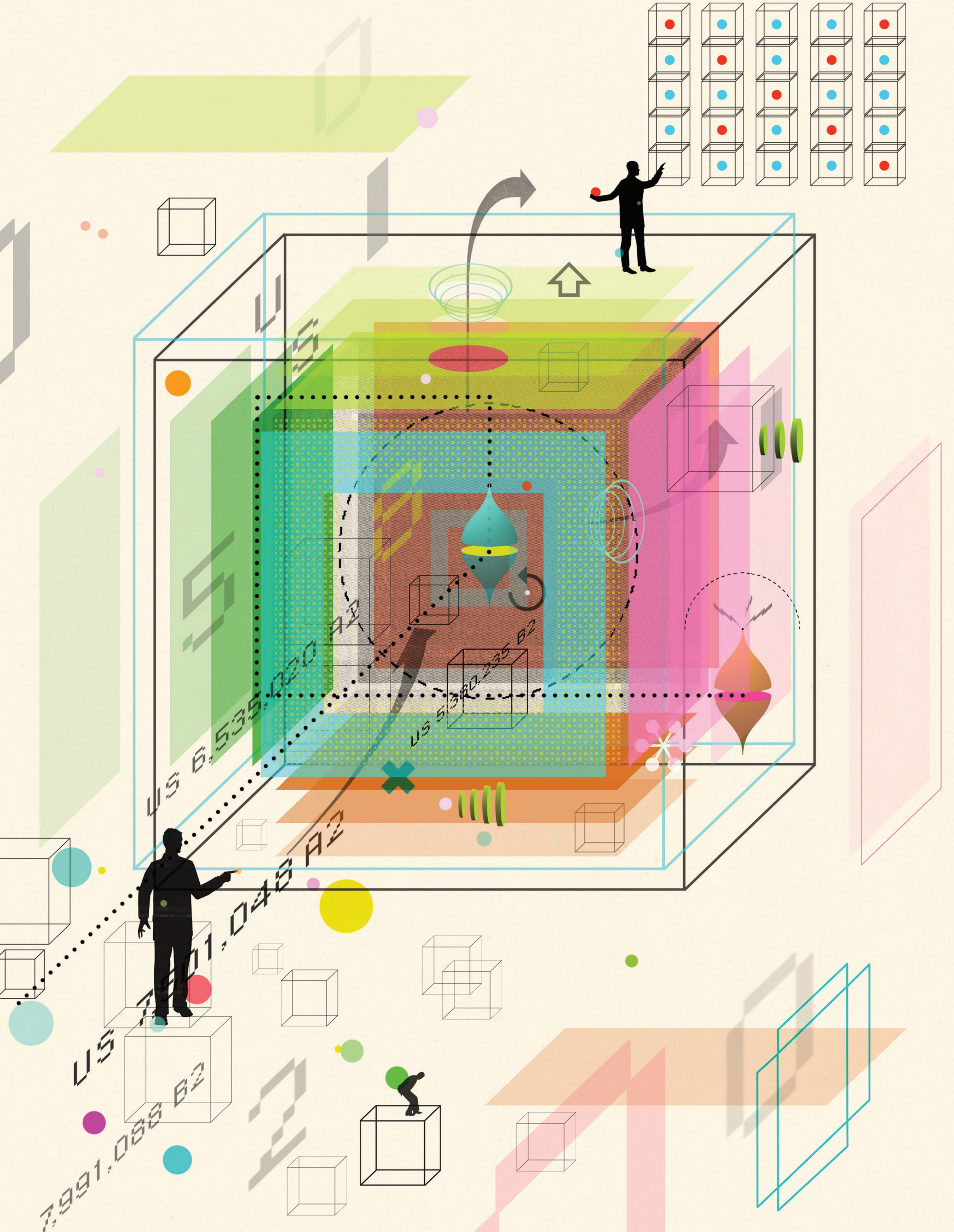

US 6,535,020 A2
US 5,380,235 B2
US 7,601,048 A2
US 7,991,088 B2

The LAW MACHINE

HOW A SILICON VALLEY START-UP AIMS TO OVERHAUL INTELLECTUAL PROPERTY LITIGATION

by Tam Harbert

ILLUSTRATIONS BY MARK ALLEN MILLER

IN A LOW-RISE BUILDING IN MENLO PARK, CALIF., JUST upstairs from a Mexican restaurant and a nail salon, a Stanford University spin-off is crunching data in ways that could shake the foundations of the legal profession.

Here, a small group of patent lawyers and computer scientists is applying the latest in machine learning and natural-language processing to reams of documents related to intellectual property lawsuits. The result is a massive statistical database on IP litigation like nothing the world has seen before. Which attorney has the best track record in defending against semiconductor-related infringement claims? Has a particular judge ruled on cases involving patent trolls, and if so, what was the outcome? Which companies tend to go to trial, and which settle out of court? By offering up such information, the database provides corporate lawyers, law firms, and government agencies with hard numbers that will reduce the guesswork, as well as the enormous expense, of patent litigation. In short, the company is building a "law machine," from which comes its name: Lex Machina.

"Law is horribly inefficient," says Mark Lemley, a professor at Stanford Law School, director of the Stanford Program in Law, Science & Technology, and cofounder of the company. "And in some ways, it is inefficient by design." After all, lawyers get paid by the hour, so inefficiency is rewarded, says Lemley. And some are rewarded richly: Top lawyers charge north of US $1000 per hour.

Lex Machina is in the vanguard of an emerging field known as legal analytics, according to Daniel Martin Katz, an associate professor of law at Michigan State University who writes the blog Computational Legal Studies and advocates overhauling the practice of law through technology. Practitioners of legal analytics statistically parse the practice of law in search of data that can be used to augment, or in some cases replace, the more qualitative judgment of human lawyers.

THE LAW MACHINISTS: Lex Machina CEO Josh Becker [left] and cofounder Mark Lemley aim to make intellectual property law more efficient.

"There's been a quiet transition going on in the legal world," Katz says. And that transition will shake up the legal profession. "Human reasoning, at least some part of it, is going to be replaced by machine-based prediction." If Lex Machina succeeds, there will eventually be fewer frivolous lawsuits—and maybe fewer lawyers too.

WE'RE THE MONEYBALL OF IP LITIGATION," SAYS Josh Becker, Lex Machina's CEO. Bespectacled and unassuming, he looks more like a professor than a savvy Silicon Valley player. With law and MBA degrees from Stanford, he served as press secretary for a Pennsylvania congresswoman, worked at the Internet start-up EarthWeb/DICE and at Netscape, and founded a venture capital firm before turning his attention to Lex Machina.

Becker is also a huge baseball fan who's made a careful study of Michael Lewis's 2004 best-selling book, *Moneyball*, which tells how Oakland Athletics general manager Billy Beane used nontraditional statistics, called sabermetrics, to make judgments about players and game strategy. Looking at the numbers, for instance, Beane determined that two popular baseball plays—bunting and stealing bases—don't contribute significantly to a team's chance of winning, so he banned them. Such decisions based on sabermetrics contributed to the Athletics' making it to the playoffs in 2002 and 2003.

That approach is basically what Lex Machina is doing for law. But while baseball is known for its reliance on statistics, Becker says, law has long been a profession that is more art than science. "Some people went to law school to avoid data," he quips.

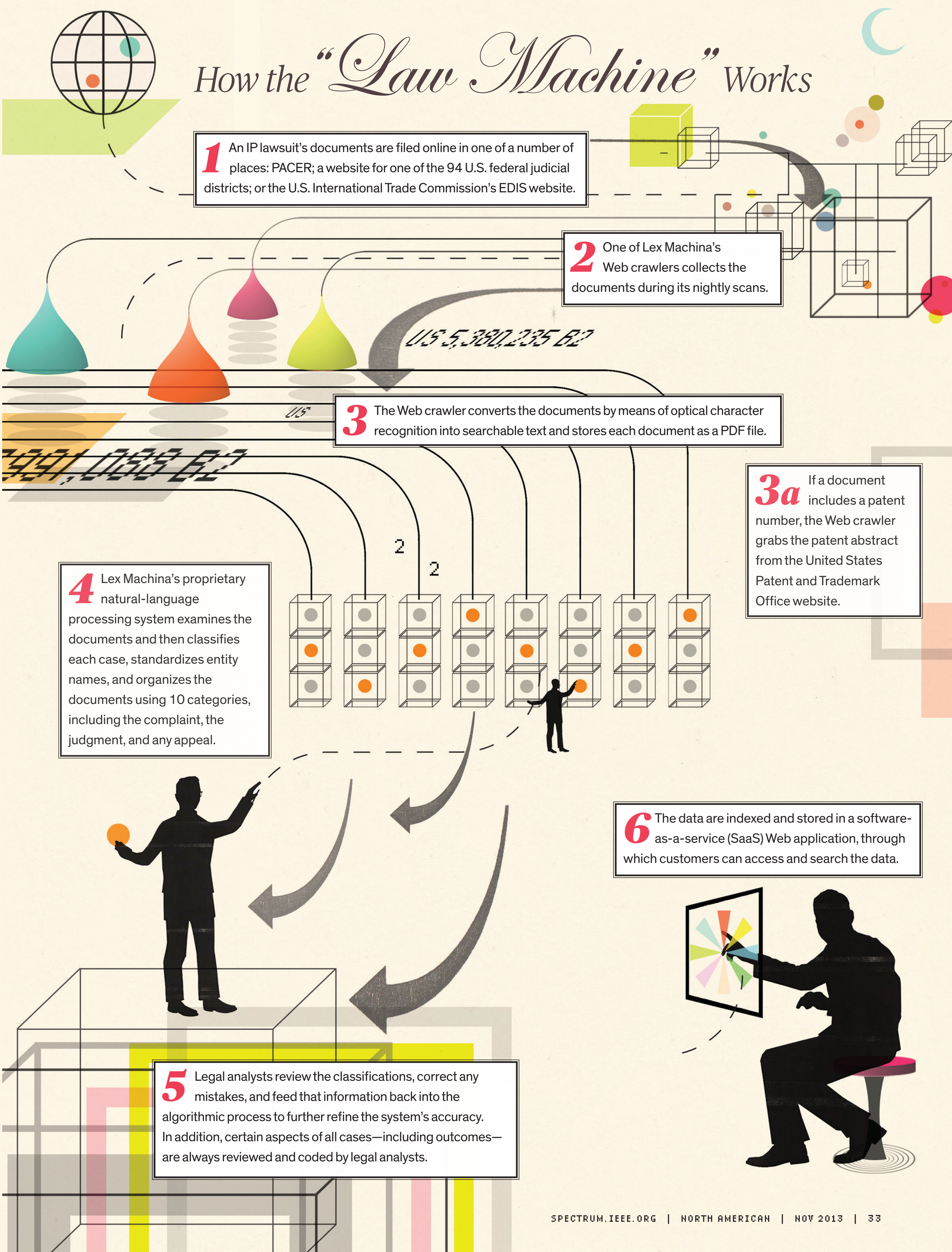
Lex Machina aims to change that. According to the company, its database covers more than 130 000 U.S. IP and antitrust cases dating back to the year 2000, including information on more than 1400 judges, 340 000 litigants, 100 000 attorneys, and 30 000 law firms. At present, it covers only the United States, but it may eventually include international patent cases as well.

With patent wars raging in every sector of the technology industry, IP litigation is big business and getting bigger all the time. The number of patent lawsuits in the United States skyrocketed between 2010 and 2012, from around 3200 filings to more than 5000, according to the Administrative Office of the United States Courts. One recent study, by James Bessen and Michael J. Meurer of the Boston University School of Law, found that defending against "nonpracticing entities"—sometimes called patent trolls—cost companies some $29 billion in 2011. Corporations are looking for a way to cut those costs.

Traditionally, a company that's been sued for patent infringement, or is thinking of suing because its own IP has been infringed, will hire top attorneys to pursue its case. Yet

PHOTOGRAPH BY Jonathan Sprague

How the "Law Machine" Works



1 An IP lawsuit's documents are filed online in one of a number of places: PACER; a website for one of the 94 U.S. federal judicial districts; or the U.S. International Trade Commission's EDIS website.

2 One of Lex Machina's Web crawlers collects the documents during its nightly scans.

3 The Web crawler converts the documents by means of optical character recognition into searchable text and stores each document as a PDF file.

3a If a document includes a patent number, the Web crawler grabs the patent abstract from the United States Patent and Trademark Office website.

4 Lex Machina's proprietary natural-language processing system examines the documents and then classifies each case, standardizes entity names, and organizes the documents using 10 categories, including the complaint, the judgment, and any appeal.

5 Legal analysts review the classifications, correct any mistakes, and feed that information back into the algorithmic process to further refine the system's accuracy. In addition, certain aspects of all cases—including outcomes—are always reviewed and coded by legal analysts.

6 The data are indexed and stored in a software-as-a-service (SaaS) Web application, through which customers can access and search the data.

the process of deciding whether, how, and even where to file such a suit is often driven by gut instinct rather than facts. Even the best patent attorney has seen maybe tens of cases that are similar to the client's. "Humans are limited. People haven't seen 10 000 cases or 100 000 cases—a human can't hold that kind of information," Katz says.

But Lex Machina can. For an annual subscription fee of around $50 000, its customers get access to 13 years of U.S. IP litigation. Just like the sabermetrics described in *Moneyball*, Lex Machina's database can aid in the formulation of broad strategy as well as the selection of players, says Becker. The company's stats reveal, among other things, which attorneys do the best against a particular patent troll, how much time and money it typically takes to fight a troll versus settling out of court, and even which judge you'd want to hear your case. The data might tell a company being sued that its peers have been settling similar lawsuits early, thereby saving money. Even if a company believes it's in the right, says Becker, a prolonged legal battle and "fighting to the death" may not make good business sense.

STANFORD LAW SPIN-OFFS

Despite Stanford University's legendary spin-off history—Cisco, Google, and Hewlett-Packard, to name a few, originated in its engineering and computer science departments—Lex Machina was the first to come out of Stanford's law school. Since that happened, in 2009, there have been several other law spin-offs. "I think Lex Machina broke the ice, showing the commercial potential of collaboration between the law, business, and engineering schools," says Clint Korver, a partner at Ulu Ventures, which has invested in Lex Machina and two other law start-ups.

Like Lex Machina, many of the newcomers rely on artificial intelligence and big-data technologies:

LawGives
was founded in 2011 and has developed a platform to match people needing legal help with an appropriate attorney. Users can get legal advice for free, then pay fixed fees for common legal services; attorneys pay a fee to be listed. The platform uses machine learning to automatically interpret the questions that clients enter into the system so it can match them with the right type of attorney.

SIPX
was founded in 2012 and offers access to copyrighted material. The company is initially targeting the higher education market, where there is a lot of confusion over tracking and managing the copyrights for teaching and training materials, a problem that has worsened with the rise of online education.

Ravel Law
was also founded in 2012. It is developing a legal search technology that uses sophisticated data visualization to speed up legal search and add context and clarity to the complex Web information.

SO HOW DOES LEX MACHINA do what it does? It started with documents—millions of pages of legal documents that, in theory at least, are available to anyone, free of charge. In practice, though, before Lex Machina came along, there was no easy way to collectively consider that vast body of information. Figuring out how to extract relevant data from countless files and then building a comprehensive database took years of dedicated effort on the part of Lex Machina's small and eclectic team. Among its 18 employees are 6 people with law degrees, 6 with computer science degrees, and 1 who has both.

The company began as an academic research project called the Intellectual Property Litigation Clearinghouse, launched by Lemley in 2006 as a collaboration between Stanford's law school and its computer science department. As Lemley explained during an interview on the sunny terrace of Stanford Law's William H. Neukom Building, "The industry was having all these debates about how to fix the patent system, and none of them were based on actual evidence."

Lemley hoped that a law database would foster decisions based on fact rather than assumption. The tech industry was enthusiastic about the project, as evidenced by more than $3.5 million in donations from companies like Apple, Cisco, and Microsoft, as well as several law firms, the Kauffman Foundation, and Stanford Law School. Lemley recruited Joshua Walker, a cofounder of CodeX: The Stanford Center for Legal Informatics. Walker in turn hired George Gregory, then a Stanford graduate student with expertise in natural-language processing and machine learning.

Several technology developments had come together that made collecting and interpreting the raw data possible. First, the documents were already available online. In the early 2000s, all 94 U.S. federal court districts adopted electronic case-filing systems, which let parties file documents pertaining to lawsuits online and make them available through the courts' websites. Other sources of data included PACER, short for Public Access to Court Electronic Records, which gives the public online access to case and docket information from federal appellate, district, and bankruptcy courts, and the Electronic Document Information System (EDIS) of the U.S. International Trade Commission. (This last source has become increasingly important in recent years, as many companies now file patent infringement claims at the USITC in addition to the courts because USITC administrative judges have the power to bar the importation of infringing products.)

Second, the growth in computer processing power and the drop in server prices had allowed data farms to crunch terabytes of data inexpensively. And third, processes and tools for machine learning and natural-language processing had advanced sufficiently to handle the complexities of legal information. Natural-language processing, also called computational linguistics, involves developing computer algorithms so that machines can understand language. Machine learning, a branch of artificial intelligence, is about constructing systems that can learn from data. The Lex Machina team uses machine-learning techniques to identify specific legal terms and phrases and then builds natural-language processing algorithms to encode the results.

Collecting and coding all that legal data was an overwhelming task. Fortunately, researchers at the Stanford AI Laboratory were eager to take on the challenge. Christopher Manning, a professor of linguistics and computer science at the AI lab, says the project offered an opportunity to extend machine learning beyond just understanding words to understanding phrases

| CONTINUED ON PAGE 53

CONTINUED FROM PAGE 34 | and contexts. For instance, locating all cases related to *patent infringement* was complicated by the fact that the exact term didn't always appear in a document's text. "It was a matter of translating upwards and understanding the concept of infringement regardless of the words they were using," Manning says.

Another difficulty the researchers encountered was that each court website uses its own variant of electronic filing. They therefore had to design a Web crawler for each one. Once collected, the data then had to be standardized to account for the variations in the way courts file data. And in many instances, the data were just plain wrong; in more than half of all cases, the final decision in the case had been incorrectly coded, according to Walker, who served as Lex Machina's first CEO and is now an attorney at the law firm Simpson Thacher & Bartlett. The way the cases were tagged–as patent, copyright, or trademark infringement, for example–was also often wrong. And there were no tags for certain types of cases, such as those involving trade secrets.

The researchers had to manually sort through, categorize, and correct the data. "There were hundreds of thousands of legal judgments that had to be made" as they sifted through the information, Walker says. In total, it took the team about 100 000 hours.

Once the team had cleaned up the data and understood its many complexities, the engineers designed algorithms to automatically review each document and sort the results. Similar algorithms are used in Web searches, but interpreting legal documents requires more sophistication. "From a science perspective, the baseline [of experience] was zero," says Lex Machina's former chief technology officer, Mihai Surdeanu, who had previously worked on natural-language processing at Yahoo Labs. "There was nobody doing this."

Existing machine-learning techniques don't work very well on litigation data, he says. Even a relatively simple process like normalizing names poses a challenge. The computer has to recognize, for example, that *IBM*, *International Business Machines*, and *IBM Corp*. all refer to the same company. More problematic were law firm names, because firms sometimes change their names when they merge with other firms or when partners join or leave. Even a firm's own attorneys can get the name wrong, Surdeanu says. "One of the firms in our database has 89 different legal spellings," he adds.

The system must also be able to handle complex legal constructs. Unlike baseball, which is a numbers game, the legal world is based on qualitative information, subtle distinctions, and most of all, words. "People argue about the meanings of words and make arguments with paragraphs of text," explains Manning. The machine needs to understand phrases and strings of commonly used legal language as well as context so that it can tell the difference between, for example, the summary judgment document (in which a judge determines which party wins the case or at least certain issues in the case) and a minor procedural filing that simply mentions the summary judgment.

To help parse the legalese, Lex Machina has developed a set of rules–a sort of legal grammar for the machine. The company does this through an iterative process: A legal analyst reviews the algorithms'

results and, if necessary, corrects them, and then an engineer tweaks the algorithms [see illustration, "How the 'Law Machine' Works"].

The result is "this ontology of terms that has been developed over the years" and continues to be refined every night, when the system crawls the Web to collect the latest data, says Owen Byrd, Lex Machina's chief evangelist and general counsel. So far, the company has coded more than 6 million docket entries.

LEX MACHINA'S DATABASE IS AVAILABLE TO ANYONE who can afford its annual fee. The pharmaceutical company Impax Laboratories uses it to guide its strategy for bringing generic drugs to market. Introducing new drugs is a highly structured and litigious process, with specific time limits for each step, so knowing the history of a judge–in particular, how fast cases move through his or her court–is critical, says Huong Nguyen, senior director of IP at Impax (and an adviser to Lex Machina).

Nguyen also uses the database to look up the litigation history of the maker of a brand-name drug to find out which attorneys it uses and how successful they've been in defending their patent positions. And she uses the database to evaluate outside counsel: She can see how many cases they're working on at any given time and which cases they have won or lost, not only for her company but for other clients as well. "We have a stable of outside counsel that we go to constantly," she says. "I want to know what kind of job performance they have across the board."

John Dragseth, a principal at Fish & Richardson (the most active IP litigation firm in the United States, according to *Corporate Counsel* magazine), credits Lex Machina's database with helping him spot meaningful but otherwise hidden trends in IP litigation–and he won't give details. "If you published it, then people on the other side would know," he says.

Typically, Dragseth says, when he reviews cases with clients, "they just nod their heads." But when he starts reeling off statistics like how a particular judge tends to rule in certain types of cases, "they lean forward, put their elbows on the table, and start asking questions," he says. "Clients go crazy about that stuff."

It's not just about the bottom line, though. Lex Machina gives its data, at no charge, to courts, government agencies, academic institutions, and media outlets. That's an important part of fulfilling the mission of Lemley's original research project: improving the legal system.

"In the short term, people will think more intelligently about whether to file suit or when they get sued, how to react: What lawyer should they hire? Should they settle the case early?" says Lemley. Ultimately, he says, people will be able to make informed decisions, not just in individual lawsuits but also in shaping policy and in bringing badly needed reform to the patent system. "My hope is that once everyone has access to the data, some number of lawsuits will go away." ■

POST YOUR COMMENTS at http://spectrum.ieee.org/lexmachina1113