WORD

PLAY

By mining a database of the world's books, Erez Lieberman Aiden is attempting to automate much of humanities research. But is the field ready to be digitized?

BY ERIC HAND

Erez Lieberman Aiden is standing on the sun deck of his town house, rocking back and forth on the balls of his bare feet as he belts out a blessing. The Hebrew words echo across the quiet courtyards of Harvard University in Cambridge, Massachusetts. The sky has turned indigo as the light and warmth leak away from this day in late April. *Shalom aleichem*, he sings. Peace be upon you.

Lieberman Aiden — molecular biologist, applied mathematician and, at 31 years old, the precocious doyen of the emerging field known as the digital humanities — could do with a little peace. The cries of his 10-month-old son have abated — for the moment — and he has had just enough time to throw on a pair of frayed black trousers and a shiny synthetic pullover before his guests arrive. A five o'clock shadow darkens the terrain between his thick goatee and unkempt hair. The night before, he caught a late train back from Princeton University in New Jersey, where he, the geeky scientist, had the delicate task of informing a room of erudite historians that his efforts at mining a database of 5 million books, about 4% of all those ever published, had made much of what they do trivially easy. The scrupulous tracking of ideas across history, for instance — work that has consumed entire careers — can be done in seconds with tools that Lieberman Aiden and his colleagues have invented.

SAM OGDEN

Yet his role as evangelist for change in the humanities — or doomsday prophet, depending on your point of view — is just one of the many parts played by Lieberman Aiden. He is also: the inventor of a groundbreaking protocol that reveals how DNA can be tightly wound and yet untangled enough to orchestrate life; the chief executive of iShoe, a company that is testing sensor-stuffed shoe inserts to help the elderly with their balance; and the co-founder, with his wife, of Bears Without Borders, which sends thousands of stuffed animals to children in the developing world. (Barely concealed in the couple's basement are mounds of donated animals awaiting delivery.) In pouring his energies into all the projects that excite him, Lieberman Aiden doesn't transcend disciplinary boundaries so much as ignore them. And although he is still technically a postdoctoral researcher at Harvard, Lieberman Aiden seems to publish the results of those projects almost exclusively on the covers of *Science* and *Nature*; hung in the stairwell below the sun deck, he has framed blow-ups of the magazine covers to prove it.

But that is work, and this is Shabbat dinner, the start of the Jewish Sabbath: a time for rest. The light switches will remain untouched, leaving the house illuminated through the night; the hot plate in the kitchen, on which the meal is being warmed, is on a timer. Three candles have been lit, one for each member of the household. Lieberman Aiden sings unabashedly in a hearty baritone that is not at all like his reedy, excitable speaking voice. He gazes at his wife, Aviva Presser Aiden, who grins back at him, holding her sweater tight to herself in the chilly night air. She too has reason to rest contentedly. The week before, she learned that she had won a US$100,000 grant from the Bill & Melinda Gates Foundation in Seattle, Washington, to build a microbial fuel cell that could charge mobile phones in Africa. The project means a year-long break from her studies at Harvard Medical School in Boston, where she is adding an MD to her PhD in genetics.

It is only by comparison with this academic power-couple that the other dinner guests — two young, self-assured Harvard physics graduates — look a bit lost, but

that probably has more to do with their unfamiliarity with the Shabbat rituals. They flip through the Hebrew prayer books and try to follow along. But Lieberman Aiden, who in his 20s toyed with becoming a rabbi, has no need for the book. These are the texts that he has studied for years. These are the words he knows best.

READING VERY NOT-CAREFULLY

As a reader with a finite amount of time, Lieberman Aiden likes to say, you pretty much have two choices. You can read a small number of books very carefully. Or you can read lots of books "very, very not-carefully". Most humanities scholars abide by the former approach. In a process known as close-reading, they seek out original sources in archives, where they underline, annotate and cross-reference the text in efforts to identify and interpret authors' intentions, historical trends and linguistic evolution. It's the approach Lieberman Aiden followed for a 2007 paper in *Nature*[1]. Sifting through old grammar books, he and his colleagues identified 177 verbs that were irregular in the era of Old English (around AD 800) and studied their conjugation in Middle English (around AD 1200), then in the English used today. They found that less-commonly used verbs regularized much more quickly than commonly used ones: 'wrought' became 'worked', but 'went' has not become 'goed'. The study gave Lieberman Aiden a first-hand lesson in how painstaking a traditional humanities approach could be.

But what if, Lieberman Aiden wondered, you could read every book ever written 'not-carefully'? You could then show how verbs are conjugated not just at isolated moments in history, but continuously through time, as the culture evolves. Studies could take in more data, faster. As he began thinking about this question, Lieberman Aiden realized that 'reading' books in this way was precisely the ambition of the Google Books project, a digitization of some 18 million books, most of them published since 1800. In 2007, he 'cold e-mailed' members of the Google Books team, and was surprised to get a face-to-face meeting with Peter Norvig, Google's director of research, just over a week later. "It went well," Lieberman Aiden says, in an understatement.

Working with Google and his chief collaborator, 29-year-old Harvard psychology postdoc Jean-Baptiste Michel, Lieberman Aiden built a software tool called the *n*-grams viewer to chart the frequency of phrases across a corpus of 500 billion words. A 'one-gram' plots the frequency of a single word such as 'feminism' over time; a 'two-gram' shows the frequency of a contiguous phrase, such as 'touch base' (see 'Think outside the box').

Google unveiled the tool on 16 December 2010, the same day that Lieberman Aiden and his colleagues published a paper in *Science*[2] describing how the tool could be used, for example, to identify the verb that has regularized the fastest: 'chid' and 'chode' to 'chided' in some 200 years (see 'The fastest verb on the planet'). "We found 'found' 200,000 times more often than we finded 'finded'," they wrote, with characteristic playfulness. "In contrast, 'dwelt' dwelt in our data only 60 times as often as 'dwelled' dwelled." Interspersed between the jokes were real discoveries — many of which had nothing to do with verbs. By comparing German and English texts from the first half of the twentieth century, the team showed that the Nazi regime suppressed mention of the Jewish artist Marc Chagall, and that the *n*-grams tool could be used to identify artists, writers or activists whose suppression had hitherto been unknown. Lieberman Aiden and Michel called their approach culturomics, a reference to the genomics-like scale of the book database, and a nod to the future, when they hope that more of the media that underpin culture — newspapers, blogs, art, music — will be folded in.

In the first 24 hours after its launch, the *n*-grams viewer (ngrams.googlelabs.com) received more than one million hits. Dan Cohen, director of the Roy Rosenzweig Center for History and New Media at George Mason University in Fairfax, Virginia, calls the tool a "gateway drug" for the digital humanities, a field that has been gaining pace and funding in the past few years (see 'A discipline goes digital'). The name is an umbrella term for approaches that include not just the assembly of large-scale databases of media and other cultural data, but also the willingness of humanities scholars to develop the algorithms to engage with them. "These tools are revolutionizing the way we work and the kinds of arguments we can make," says Dan Edelstein, a historian at Stanford University in California, who has used mapping software to show unexpected patterns in the way that Voltaire's letters spread through Europe during the Enlightenment.

"THESE TOOLS ARE REVOLUTIONIZING THE WAY WE WORK."

Yet some humanities researchers in the traditional camp complain that their field can never be encapsulated by the frequency charts of words and phrases produced by the *n*-grams tool. "I think saying all books equal the DNA of human experience — I think that's a very dangerous parallel," says Cohen. How do you factor in the cultural contributions of furniture, or dance, or ticket stubs at a movie hall, he asks. What about all the books that were never published? Or the culture as experienced by the world's vast illiterate populations?

Other scholars have deep reservations about the digital humanities movement as a whole — especially if it will come at the expense of traditional approaches. "You can't help but worry that this is going to sweep the deck of all money for humanities everywhere else," says Anthony Grafton, a historian at Princeton and president of the American Historical Association, who uses a giant, geared wooden reading wheel to help him manage his oversized, Renaissance texts. He wants researchers to hold onto the power that comes with intimately knowing their primary sources, right down to the scribbled notes in the margin that would elude the book scanners. "You don't want to give up what is your own core activity," he says.

FOLLOWING TRADITION

Back at the Aiden house, the Shabbat dinner guests have all laved their hands with a glass of water and returned to the sun deck for matzo-ball soup. Lieberman Aiden explains some of the trepidation he felt when he and Michel talked to the historians at Princeton about their work. "I was a little bit nervous going in," he says. "I really thought that we were going to get denounced at one point."

Although Lieberman Aiden and Michel are sensitive to the feelings of traditional humanities scholars, they are also too young, restless and deeply ambitious to slow their own pursuits. Lieberman Aiden says that the influence of technology on the humanities is already past a tipping point. The tools and methods that it provides, he says, will be impossible for researchers to ignore. And yet he doesn't think that the old approaches will ever disappear. "I think you should use the best methods available — and all of them," he says. "And I think that includes carefully reading texts and trying to get behind what authors think."

Daniel Koll, one of the dinner guests, shyly interrupts. "Erez? Do you think you're maybe partly influenced in that kind of thinking by your religious upbringing? From my limited outsider's perspective, Judaism has a very strong interpretive component. There is no single authority on a text, and so on." He wonders whether Lieberman Aiden, like any good humanities scholar, enjoys wrestling with the ambiguities of religious texts as much as he enjoys cool, hard data.

Clearly, the answer is yes — why else would his host have spent a year of his life at Yeshiva University in New York, studying the Talmud and Jewish case law? But Lieberman Aiden, who prefers to talk about other people and their

A DISCIPLINE GOES DIGITAL

The humanities mine cultural databases

The digital humanities — the use of algorithms to search for meaning in databases of text and other media — have been around for decades. Some trace the field's origins to Roberto Busa, an Italian priest who, in the late 1940s, teamed up with IBM to produce a searchable index of the works of thirteenth-century theologian Thomas Aquinas.

But the field has taken on new life in recent years. Journals have sprouted up and professional societies are blooming. Some universities are now requiring graduate students in the humanities to take statistics and computer-science courses. Funding — far harder to come by in the humanities than in the sciences — flows slightly more generously to those willing to adopt the new methods. This year, the US National Endowment for the Humanities, in collaboration with the National Science Foundation and research institutions in Canada and Britain, plans to hand out 20 grants in the digital humanities, worth a total of US$6 million.

Many researchers in the digital humanities use textual databases composed primarily of books — as Erez Lieberman Aiden does in his 'culturomics' project (see 'Heavy-duty data'). Franco Moretti, a literary scholar at Stanford University in California, has shown that genres of fiction — Gothic novels, for example, or romance — have a textual 'fingerprint' that is apparent even in simple frequency counts of nouns, verbs and prepositions. "These genres are different at every scale," he says, "not only in the huge scene of being held captive by a Count."

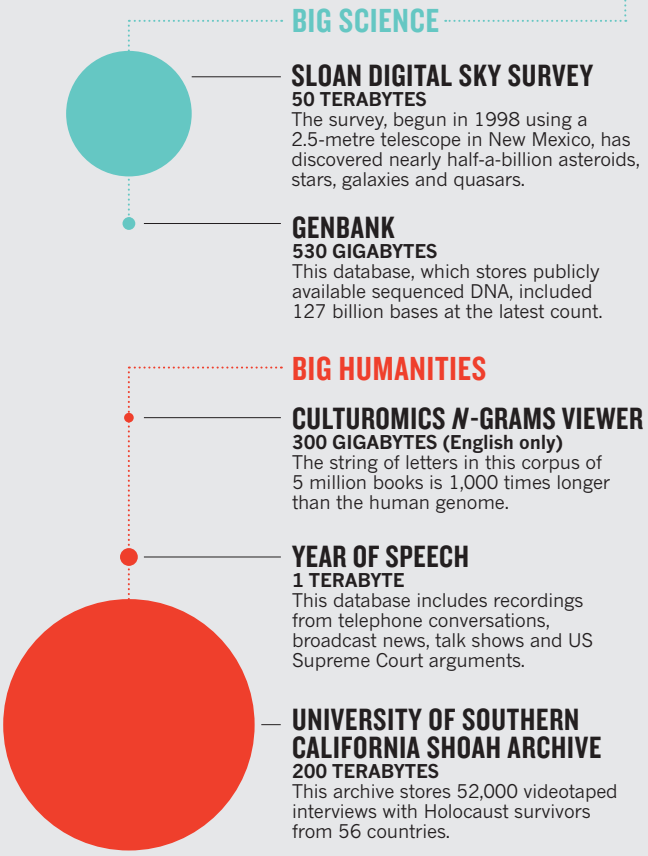
Some researchers are busy digitizing other forms of cultural data. John Coleman, a phonetician at the University of Oxford, UK, is putting 5 million spoken words — about 3 months of speech, end to end — into a database, down to the level of the individual phonemes. Collected largely as recordings made with Sony Walkmen in the 1990s, it contains all sorts of things typically ignored by linguists: neologisms, slurring and sub-verbal honks and snorts. Coleman is already learning how conversation partners take pacing cues from each other, and how pitch of voice reflects attitude. And, he says, he can prove that women and men talk at the same speed. The linguistics textbooks, he says, "are going to have to be rewritten".

Ichiro Fujinaga, a music technologist at McGill University in Montreal, Canada, is trying to do something similar for music. In a project known as SALAMI (Structural Analysis of Large Amounts of Music Information), Fujinaga is finding the common structural patterns (such as verse–chorus) in 350,000 pieces of music from all over the world. With more than 7,000 hours of Grateful Dead recordings in the database, he says, his team will be able to answer the all-important question: "Did the guitar solos get extended over the years or did they get shorter?" **E.H.**

LARGE HADRON COLLIDER
13 PETABYTES (2010)
The proton collider, near Geneva, Switzerland, generates about 15 petabytes of data per year — even after rejecting 99.9995% of collisions.

HEAVY-DUTY DATA

The computer-storage space required to support projects in the digital humanities is now starting to rival that of big-science projects.

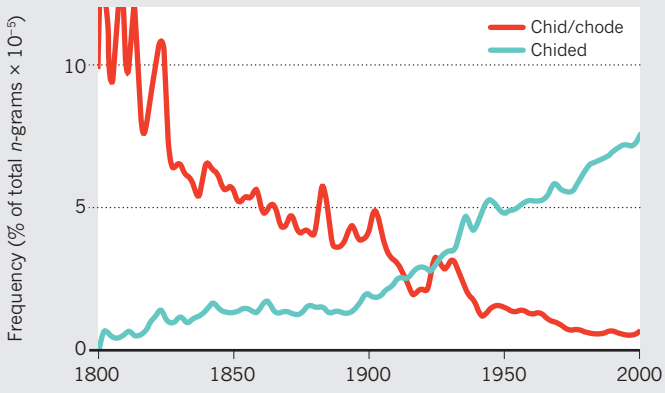


1 petabyte = 1,024 terabytes = 1,048,576 gigabytes

SOURCE: REF. 2

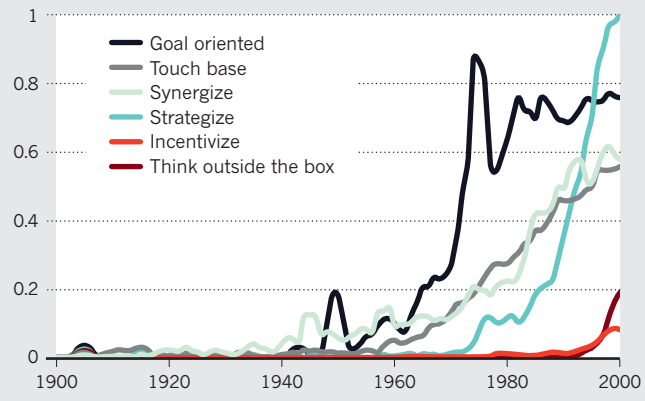
THE FASTEST VERB ON THE PLANET

Rarely used verbs regularize quickly; the *n*-grams viewer reveals that 'chide' has changed fastest of all.



THINK OUTSIDE THE BOX

Text analysis using the *n*-grams viewer shows the infiltration of corporate speak into the English language.



SOURCE: GOOGLE N-GRAMS VIEWER

ideas rather than himself, provides an indirect answer by way of history. He tells the story of Isaac Casaubon, a sixteenth-century Protestant scholar, who undermined the presumed Egyptian provenance of a set of religious texts by identifying a reference to a Greek play on words — something that could only have been written hundreds of years later. "That point is as objective an interpretive remark as any remark a scientist might make," says Lieberman Aiden. "So the methods of humanists are very, very formidable. And I think the degree of insecurity they have over whether these methods are here to stay is not really befitting."

TWO CULTURES

From the day he was born in a New York City hospital, Lieberman Aiden was steeped in the cultures of both language and technology. The son of a Hungarian mother and Romanian father, both émigrés by way of Israel, Lieberman Aiden grew up in a community of Satmar Jews, a branch of Hasidic Orthodox Judaism. English was his third language, after Hungarian and Hebrew, and by the age of nine, he was helping his father, a self-taught inventor, with the English contracts for the family saw-manufacturing business. Lieberman Aiden studied at a religious high school in Brooklyn, but soon found that video games held more allure. In his second year there, he found himself flunking classes, and his addiction to *X-COM: UFO Defense* was consuming so much time that he eventually had to quit, cold turkey. "It was an amazing game, actually," he says, ruefully.

> "HE'S A NON-CONFORMIST BY DESIGN AND HE REVELS IN IT."

Lieberman Aiden soon found more edifying outlets for his energies: he was allowed to skip school one day a week to study in a molecular-biology lab at Brooklyn College, and he began his own computer-repair business. The family was quite secular by Hasidic standards, going to synagogue only for the High Holidays of Rosh Hashana and Yom Kippur. One day in high school, he went to Burger King for his usual bacon cheeseburger, and decided to respect kosher rules by forgoing the bacon — not realizing that mixing dairy and beef in the cheeseburger itself was not kosher at all.

Lieberman Aiden went to Princeton as an undergraduate, where he wasn't content to study just maths and physics, but also fulfilled all the requirements for a philosophy degree. And even as he took five, six, seven classes a term, he managed to squeeze in creative-writing courses, specializing in haiku poetry. While at Yeshiva University after Princeton, he taught maths on the side to pay for a master's degree in history, and completed the first year of rabbinical studies. "He's a non-conformist by design and he revels in it," says his Talmudic study partner, Avi Bossewitch. And yet, he says, "he's the least arrogant person I've ever met".

The allure of science eventually proved too strong. Lieberman Aiden left Yeshiva to begin a PhD at the Broad Institute of the Massachusetts Institute of Technology (MIT) and Harvard in Cambridge, under the supervision of famed geneticist Eric Lander. But even while mastering molecular biology, he put his maths skills to use. He realized that a knot-free shape described in a 120-year-old maths paper — a fractal globule — could describe the way that the 2-metre-long human genome folds up into the cell nucleus, a space one million times smaller. He then developed a protocol to prove that this was true. The results, published in the first of his papers to make it onto the cover of *Science*[3], showed that the fractal globule allowed widely separated sections of DNA to unfold and interact. "There are no bounds to what he can be interested in," says Lander, who, like others, suspects that culturomics might eventually be merely a sideline for Lieberman Aiden in his ascent in mathematical biology.

It was during his time in the Lander lab that he met Aviva Presser, a shy young woman from Los Angeles who was also working towards her PhD. They married in 2005, with Lander giving one of the blessings. Rather than one of them taking the other's name, they decided to share a common new name: Aiden, which means Eden in Hebrew and, in Gaelic, little fire.

Those fires stoked, they were faced with another naming issue last June: what to call their son. The working title, *in utero*, was Snedley Balagan (Balagan means 'fiasco' in Hebrew). But they soon settled on Gabriel Galileo Aiden. Presser Aiden says that no one believes them when they say that they had no idea that their son's initials formed the DNA code for a sweet amino acid.

WORK AND PLAY

Towards the end of the Shabbat dinner, Gabriel decides to wake up, adding to Presser Aiden's obvious exhaustion. But her husband doesn't want to miss out on the highlight of the night, an Aiden-family staple: Dessert Face-off, in which each guest competitively designs a slice of brownie within a theme. Given that guests Koll and his girlfriend, Larissa Zhou, are interested in molecular gastronomy, Lieberman Aiden decides on the theme of food science. A box of edible Betty Crocker decorations is laid out.

Koll turns his brownie into the cross-section of a wok. Zhou transforms hers into a pig — definitely not kosher. But Lieberman Aiden's construct is perplexing: he has created a starry night-time skyscape, with multicoloured sprinkles as stars and the Milky Way. What does this have to do with food science? "Well, you know," he says, a little proud of himself, "gastronomy is just one letter away from astronomy."

Gabriel has returned to bed, and Presser Aiden looks ready to follow, but her husband isn't quite finished yet. As midnight approaches, Lieberman Aiden is holding forth on the mathematical beauty of ramen noodles — as depicted in the 2008 film *Kung Fu Panda* — and remarking on a piece of maths software, KnotPlot, that could help his wife to braid some really radical challah bread. The guests finally do Presser Aiden a favour and excuse themselves to go home.

But Lieberman Aiden just might keep revving and riffing through the night: he can work for 70–80 hours straight, fuelled by Diet Coke and junk food. He has big plans for culturomics, as he and Michel add more languages, books and other media to the *n*-grams database. He also has new projects to think over, such as one currently under way with Ed Boyden, a young, acclaimed neurobiologist at MIT, in which the pair are developing a way to detect the genes expressed in thousands of individual cells at a time. But tonight and tomorrow he'll keep his computer off, although it is not religious conviction that makes him abide by the Sabbath rules. Doing so forces him to detach, clear his mind and go for walks in the park with his wife and son.

And yet the boundary between work and play — just like that between the sciences and the humanities — is not one that Lieberman Aiden respects. That might just be what makes him successful, says Lander. For centuries, the best science has come from the most playful scientists, he says. Think of Watson and Crick shirking the lab in favour of tennis; think of Einstein and his wild-haired bike rides.

"What do children do?" says Lander. "They learn, they're curious, they're stimulated. The problem is, at some point, many people get in a rut. They're not really interested in learning more. They're not able to be fascinated and delighted by everything around them. Erez — he hasn't lost the playfulness." ■ SEE EDITORIAL P.420

Eric Hand *is a reporter for* Nature *in Washington DC.*

1. Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. A. *Nature* **449,** 713–716 (2007).
2. Michel, J.-B. *et al. Science* **331,** 176–182 (2010).
3. Lieberman-Aiden, E. *et al. Science* **326,** 289–293 (2009).