ARTICLE COLLECTION

Technology's Impact on the Research Enterprise

With Support From SMU Moody School of Graduate & Advanced Studies

THE CHRONICLE OF HIGHER EDUCATION

SMU

THE 21ST CENTURY IS DIGITAL. THIS IS A TRANSFORMATIONAL TIME FOR OUR FUTURE-FOCUSED ACADEMIC ENTERPRISE AS WE INVEST IN EMERGENT TECHNOLOGIES TO BUILD ON OUR TRADITION OF ACADEMIC EXCELLENCE THAT SHAPES HOW WE PERFORM RESEARCH, ADDRESS SOCIETAL PROBLEMS AND CREATE ECONOMIC OPPORTUNITIES THAT UPLIFT COMMUNITIES.

Now in our second century of achievement, SMU is an interdisciplinary research hub, prominently located in the heart of Dallas, that supports visionary faculty and Ph.D. students with the opportunities and the access to emergent technologies to achieve their goals. The new Moody School of Graduate and Advanced Studies, established in 2019 with a $100 million gift from the Moody Foundation, is a bold investment in and commitment to supporting SMU's research mission. By attracting and supporting outstanding graduate students – the workforce behind groundbreaking discoveries that bolster our doctoral and research ecosystem – the Moody School helps shape a University-wide community of scholars who leverage SMU's strengths in supercomputing to solve problems and drive impactful ideas on the Hilltop and beyond.

As a leader in the region in the research applications of 21st century digital technologies, SMU is investing even more in supercomputing as a primary and essential resource for research and education across the disciplines. Our collaboration with accelerated computing leader NVIDIA puts us in the fast lane for artificial intelligence (AI). Our $11.5 million investment to supercharge our AI infrastructure with an NVIDIA DGX SuperPOD™ will give our faculty, students and external research partners the ability to integrate sophisticated AI technology across a wide array of research disciplines, ranging from computational biology to human performance, from national defense to digital humanities. Increasing our computing capability also will provide real benefits for North Texas, as the region continues its growth as a technology hub.

> "SMU is investing even more in supercomputing as a primary and essential resource for research and education across disciplines."

We recognize that research universities like SMU have an obligation to actively engage in the development and application of AI for the good of humanity. We take that responsibility to heart, and I couldn't be prouder of our intrepid supercomputing community for its active engagement in tackling some of the world's most pressing problems to create a better future for all.

Sincerely,

ELIZABETH G. LOBOA, PH.D.
Provost and Vice President for Academic Affairs

THE CHRONICLE OF HIGHER EDUCATION®

Technology's Impact on the Research Enterprise

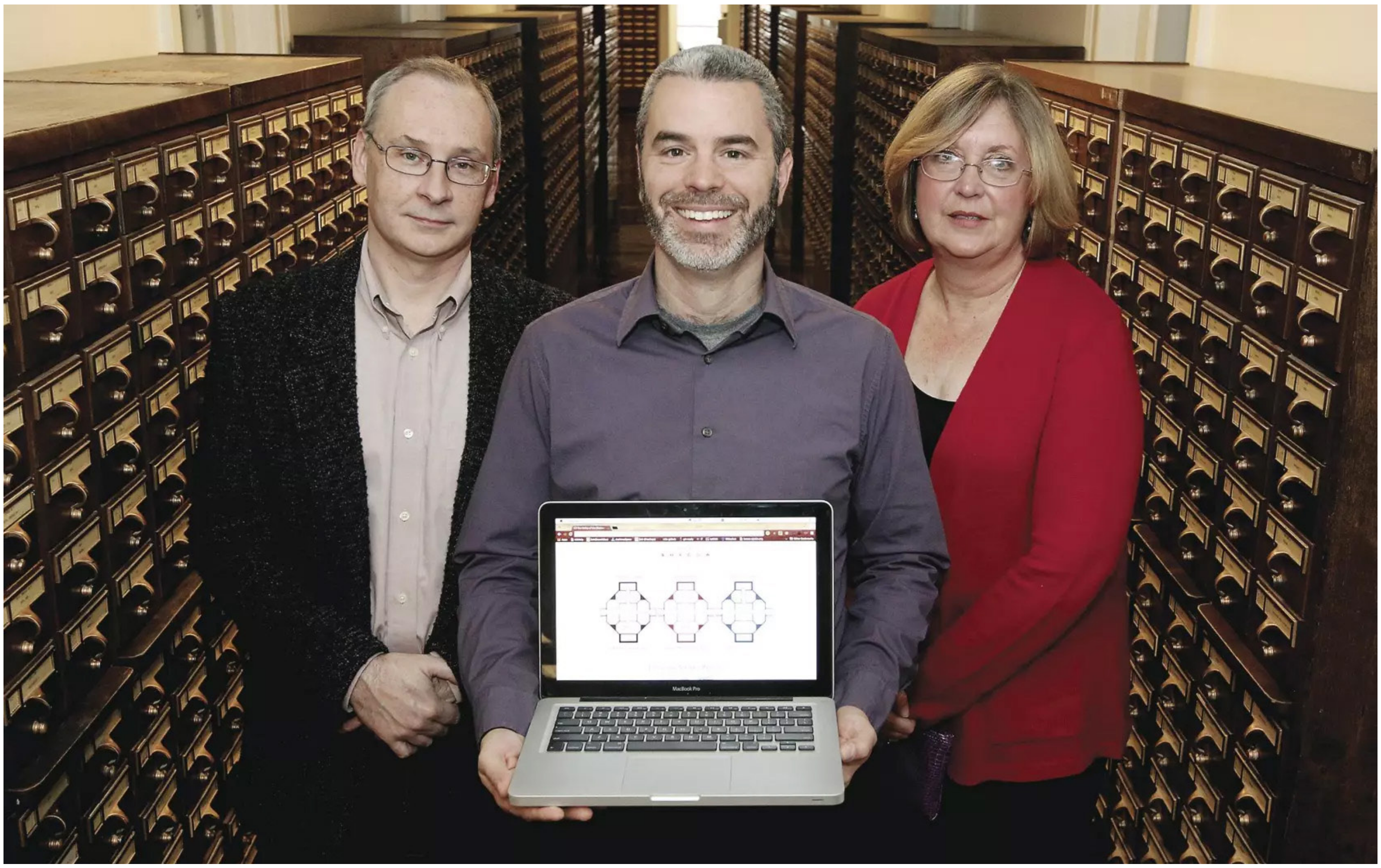
In the past decade, advancements in computing power and artificial intelligence have had a profound effect on how university faculty members pursue their research endeavors. The ability to mine large data sets, harness machine learning, and use other digital tools has created new opportunities — as well as controversies — in the sciences and humanities. This collection of *Chronicle* news and opinion articles explores the changes in how research is conducted and disseminated, offering a glimpse into how the roles of the researcher and the research institution have been impacted by tech during the last decade.

Cover photo by iStock

For questions or comments about the collection, email ci@chronicle.com

Born Digital, Projects Need Attention to Survive

BY JENNIFER HOWARD



ANDREW SHURTLEFF FOR THE CHRONICLE

Bradley Daigle, a digital curator at the U. of Virginia, and his colleagues Matthew Stephens and Lorrie Chisholm were in charge of preserving an early digital archive on the Civil War.

Publish a scholarly book and, absent a flood or other disaster, chances are it will last as long as a library has space for it — long enough to become part of the conversation in its field if it's notable enough. But create a pioneering work of digital scholarship, and how to preserve it becomes more of a challenge — in fact, one of several. While online scholarship often has dazzle — dynamic maps, data visualizations, or other features that invite interaction and exploration — it can have a harder time catching the eye of scholars who are used to arguments packaged in articles and monographs. Build it, and the experts won't necessarily come — at least not yet in great numbers.

The first challenge is making sure people can get to the work when they do want to come. Analog or digital, no work will have much influence if it doesn't stick around to be cited or argued with. The technological advances that make digital-humanities work possible also put it at risk of obsolescence, as software and hardware decay or become outmoded. Somebody — or a team of somebodies, often based in academic libraries or digital-scholarship centers — has to conduct regular inspections and make sure that today's digital scholarship doesn't become tomorrow's digital junk.

Bradley J. Daigle, director of digital curation services at the University of Virginia Library, calls this "digital stewardship." It's an essential but easily overlooked element in any digital-humanities project. Born-digital work can die. Digital stewardship "involves care and feeding" to make sure that doesn't happen, he says. "My unit essentially pays attention to the life cycle of the digital object."

"Most people conceive of preservation as backups," Daigle says. But tending a piece of digital scholarship involves much more than just dumping a copy in an archive.

In the past few years, Daigle's team has gotten a lot of experience in digital stewardship, as early digital-humanities work by Virginia faculty members and graduate students has begun to show its age. One notable example is a Civil War project called Valley of the Shadow, brainchild of Edward L. Ayers, a historian who is now president of the University of Richmond.

When Ayers began work on the Valley project, in the early 1990s, he was a professor of history at Virginia. He wanted to build an online library of primary-source documents that would shed light on two 19th-century American communities, one Northern and one Southern, from the time of John Brown's raid through the war and Reconstruction. (A lot of digital-humanities work in the 1990s involved the creation of online archives and editions.) Visitors can dip into life in Augusta County, Va., and Franklin County, Pa., before, during, and after the war via newspaper articles, letters and diaries, church and tax records, maps and images, and statistics.

The site invites users to do the searching and interpreting of the materials it includes. "The whole point is that you are supposed to come up with the interpretation," Ayers says.

It sounds like a simple idea, but "it took 14 years to build, and there's probably a million dollars in it," he says, much of that used to pay for the student labor that built it. By the time the last touches were put on the site, in 2007, technology had moved far beyond where it had been a decade and a half earlier. "When we began, there were no such things as PDF files," Ayers says. At one point, his team build a CD-ROM for the publisher W.W. Norton, "finishing just in time for nobody to use CD-ROMs" anymore.

"Think about the life cycle of preservation."

One of the graduate students who worked on the project, Andrew J. Torget, is now an assistant professor of history and director of the Digital History Lab at the University of North Texas. "That thing was a hairy beast because it was one of the earliest projects," he says. "It was built and rebuilt over time."

In 2009, Daigle and the digital-curation unit at Virginia's library were recruited to get the "hairy beast" back into shape technologically. Every element of the project had to be examined.

"What we essentially had to do was standardize it all," Daigle says. He compares the process to what auto mechanics used to do in the 1950s. "We basically swapped out all the parts and rebuilt the engine," he says. "We took the entire site and atomized it into several hundred thousand individual files," then analyzed them to see if they were damaged or in still-usable formats. Monitoring software now keeps tabs on the site to make sure it continues to function well. Users can email the library to report problems they encounter.

Most of that labor will be invisible to anyone who visits the Valley site, which looks a lot like it did when it was new. Some digital-humanities projects are designed to be open-ended, becoming platforms for subsequent additions, enhancements, and layers of work. (See, for example, the Perseus Digital Library, which took shape decades ago as a digital collection of Greek and Latin texts and has become an ever-expanding teaching-and-research hub for classics.) Others, like the Valley project, have natural limits.

Sometimes, to preserve its integrity as a scholarly resource, a completed piece of digital scholarship not only needs to be kept in good working order but also should look the way it did when its original builders finished it, like a specific edition of a published book that persists over time. The presentation as well as the content becomes an important part of the work's intellectual value.

The stewardship that the Valley project required was neither cheap nor easy. According to Daigle, the work took two years, a $100,000 grant from the university, and the contributions of three full-time employees and several graduate students — resources that many academic libraries cannot throw at an individual project. Even though the work was expensive, it has made subsequent projects easier to handle, he says.

"With Valley, we were able to create the manual that we can use and apply to other forms of scholarship," Daigle says. "We've become more sophisticated in how we approach these things. We're better mechanics now."

Scholars who undertake ambitious digital work have a hard reality to face: Not every project can or will get the kind of full-service care that the Valley of the Shadow received. "You can't save everything," Daigle says. "We had to make some core decisions about whether this project or this scholarship is worth preserving."

The UVa library gets "a steady stream" of old and new digital projects it could archive. If a faculty member wants to keep tinkering with a site, or if the project relies on a format, like Adobe Flash, that isn't intended for the long haul, the library might have to say no or give it a bare-bones treatment — taking a snapshot of it or making it available as a download, for instance, rather than doing an engine rebuild, as it did with Valley.

Daigle advises scholars who want to pursue digital-humanities work to consult with their librarians and put long-term archiving strategies in place early on. "Think about the life cycle of preservation," he says. "The more you do that, the longer it's going to be around, and that is time well spent."

Preserving something doesn't guarantee that anybody will use it, of course. That's just as true of digital scholarship as it is of print monographs. But digital work has the advantage of existing online, potentially within reach of anyone with an internet connection.

The Valley of the Shadow often gets assigned in history courses, according to Ayers and Daigle. In one 24-hour period, the site received almost 600 visits — not bad for a digital resource that began life 20 years ago, before eye-catching visualizations and the kinds of interactive approaches one sees now. "We know that we've reached millions of people with this, and we know that it's used around the world, even though we haven't changed it in six years," Ayers says.

He's less sure that Valley has had a direct influence on scholarship. "I'm not sure you're a pioneer if nobody follows you," he says. Aside from his own book *In the Presence of Mine Enemies: War in the Heart of America, 1859-1863*, which won the Bancroft Prize in 2004, Ayers doesn't know of scholarly books that use Valley of the Shadow as a base. More people may be citing primary sources they find on the site than citing the site itself.

The project has had an undeniable influence as a training ground for a younger generation of scholars moving up the academic ranks and taking the technical skills and approaches they learned with them. "It's fundamentally shaped my research," says Torget, of North Texas, who had no tech skills to speak of when he went to graduate school at UVa to study the South.

Working with Ayers and the Valley group gave him "the confidence to learn new technologies as I needed them for my own work," he says. He learned how to manage a project and build a team of collaborators, which is vital to much digital-humanities work but isn't part of traditional humanities training.

For his dissertation, Torget created the Texas Slavery Project, "animating and exploring" the history of emancipation in the Texas borderlands. The work was heavily informed by Valley of the Shadow. And like Valley, the borderlands site has appealed to people who aren't professional historians. Torget gets more email about it from genealogists than from any other group.

It makes sense, he says, because much of the scholarly give-and-take in history is "a very slow book-by-book kind of thing." Digital scholarship "invites a broad spectrum of people who wouldn't be reading our stuff in T*he Journal of American History* or *The American Historical Review*," two of the field's leading scholarly journals.

Still, Torget hopes that scholars as well as genealogists will pick up on what he's done. The Texas Slavery Project gets used "for classes and for research," he says, but as far as he knows it hasn't yet been cited by scholars. "I would love to have somebody take the materials I've put together and use them," he says. "It's something I've come to believe: that the most interesting thing someone's going to do with your data is something you've never thought of."

Although it comes in a multitude of forms, digital scholarship can be cited and referenced like more traditional work. (The day is coming when phrases like "more traditional work" won't be useful anymore, as digital approaches become more common and visualizations or other nonmonographic treatments of literary and historical data start to look as familiar as book-length arguments.) And while citations aren't yet plentiful enough to satisfy many digital scholars, there's immediate proof of their influence: Their work keeps inspiring new work. Look at the next round of digital scholarship under construction at places like the University of Richmond's Digital Scholarship Lab. It just released an interactive digital edition of Charles O. Paullin and John K. Wright's *1932 Atlas of the Historical Geography of the United States*. Robert K. Nelson, the lab's director, describes the atlas as a "prelude" to a much larger attempt to rethink what a historical atlas ought to be and do in the 21st century.

Nelson got his start as a digital humanist on another seminal first-generation digital project, the Walt Whitman Archive. The lab's associate director, Scott Nesbit, worked with Ayers on Valley of the Shadow.

Nelson and Nesbit describe how digital-humanities work has begun to evolve from its early emphasis on editorial projects — building online collections and editions of primary materials, for instance. "You do that and you realize you want to do things that are interpretive," Nelson says.

That spirit animates another of the lab's creations, Visualizing Emancipation. It maps specific "emancipation events" — people fleeing slavery, for instance — that touched individuals all over Civil War-era America, not just in the halls of politics or on the battlefield but also on farms and plantations, in cities and towns. The work keeps expanding; at its heart is a data set of 3,400 "documented places where we've found slavery changing in some way during the war years," Nesbit says.

Presenting those data as different, manipulable layers provides "a much richer picture of how emancipation works," he says. "It did not run in one direction. It looked very different at the level of the individual than it did to the nation as a whole. We wanted to be able to build a map that reflected this complexity."

Unlike Valley, which is more self-contained, the site makes other kinds of connections as well. It points users to relevant digital-humanities sites elsewhere, as long as those sites are "robust and mature" enough to be reliable, Nesbit says, invoking the need for good digital stewardship.

The Visualizing Emancipation site wouldn't exist without precedents like Valley of the Shadow, according to Nelson. But the field has only begun to explore the kinds of fresh analyses that digital approaches have put within humanists' reach.

"That's probably one of the modest disappointments of digital humanities 20 years into this enterprise — that it hasn't spawned more broad scholarship," Nelson says. "The challenge and the opportunity for the digital humanities is to start making arguments and producing interpretations that are going to be of interest to people who are not necessarily invested in the digital humanities as an enterprise."

Jennifer Howard was a senior reporter at The Chronicle.

Originally published on January 6, 2014

EDUARDO LUZZATTI FOR THE CHRONICLE

Peer Review in Flux

The internet era has changed the landscape.

BY PAUL BASKEN

Beaten down by technological change and economic pressures, the long-held notion of scientific peer review is losing its status as the "gold standard" measure of scholarly reliability.

The problem facing universities in 2018, however, isn't so much that peer review has inevitably evolved, but that scientists collectively have failed to respond with a better replacement.

Among the many troubles for peer-reviewed publications:

- Subscription-based journals are proving far too slow for the speed of scientific exchange in the internet era and have long generated resentment about costs, while more streamlined open-access models raise widespread questions about their sustainability and reliability.

- Publishing priorities and the financial rewards of research breakthroughs discourage scientists from reporting negative results and don't sufficiently guard against potential bias.
- Universities too often reward faculty members based on the quantity rather than the quality of publications.
- Citation-based measures of journal reputation — a proxy for peer-review quality — have long been recognized as flawed and susceptible to manipulation.
- Overworked researchers show a growing resistance to serving as reviewers or devoting adequate time to the task.

One result: The notion of what it means to have a highly respected "peer reviewed" work of science has become diminished, if not lost entirely. Another: Scientists caught up in uncertainties over the meaning and standards of "peer reviewed" research aren't doing all they can to share their work and collectively advance their fields.

The solution for scientists, say analysts studying the problem, lies in helping scholars — and their employers and funders — better understand how researchers can collaborate, share, and self-correct their work, and be credited for it.

"We have a too-narrow focus on peer review at the stage of publication," says Brian A. Nosek, co-founder and director of the Center for Open Science, "at the cost of appreciating how evidence becomes credible over time with all the other parts of continuous peer review in the community."

Traditionally, peer review has meant the formal evaluation process of a scientist's manuscript — by academic counterparts of the author — as a condition for journal publication.

Now, with advanced electronic methods of communication, the concept of peer review is evolving to mean any number of ways that a scientist receives useful feedback from colleagues, from the earliest stages of project design to post-publication critiques. The nonprofit Center for Open Science alone offers at least 15 preprint servers (online repositories for publicly sharing manuscripts with no pretense of peer review) in fields that include business, education, engineering, law, and the life sciences.

Major academic publishers, including Elsevier, are also joining in, offering a variety of online tools to help scientists record and immediately share their notes and data findings with colleagues around the world.

That's a good thing, many experts argue. "It allows us to keep going, to be current, to be at the vanguard, and to understand what's happening," says Harlan M. Krumholz, a professor of medicine at Yale University who studies accuracy in science.

At the same time, however, many journals and universities cling to the idea that a final published article that passes some measure of "peer review" remains a defining measure of academic accomplishment — even in the face of growing evidence that the standards of those reviews are slipping.

At last year's quadrennial Congress on Peer Review and Scientific Publication, Krumholz called on leading academic journals to tolerate the open sharing of findings among scientists and to stop making such activity a disqualification for the eventual

TAKEAWAY

Open Science Needs Further Review

- Scientific-journal peer review, long revered as the "gold standard" of scientific reliability, is increasingly seen as failing in the accelerating whirl of the internet era.
- Key factors include heavy demand for published work due to financial pressures for scientific breakthroughs and stubbornly formulaic university reward structures.
- Solutions may lie in greater "openness" and in making the scientific process itself more transparent and susceptible to sharing and feedback.
- Universities, which control researchers' salaries and other financial incentives, must be willing to make changes in their reward systems.

publication of a manuscript. "If we wait for peer-review publication," Krumholz said of his own research team, "we'll be years behind in the field."

Howard C. Bauchner, editor in chief of the *Journal of the American Medical Association*, pushed back, saying there had not yet been enough study of whether online sharing prior to peer-reviewed publication might produce more harm than benefit in fields like medicine. Nonscientists, for example, might see a preliminary finding and act upon it, with harmful results.

Many journals and universities cling to the idea that a final published article that passes some measure of "peer review" remains a defining measure of academic accomplishment.

"I know it always feels better if we're more transparent, if there's more science, if there's more information out there," Bauchner, a professor of pediatrics at Boston University, told Krumholz. "But I think we've seen, over the last 10 or 15 years, there is the real capacity to do harm."

Amid such fundamental disagreements, there appears to be little coordinated effort to determine what, exactly, "peer review" should look like in the future. Even among journals that make a good-faith effort at peer review, there's no common understanding of whether the process should mean a single reader giving a quick scan for obvious errors, a team of highly qualified reviewers offering multiple rounds of feedback to the author, or something in between.

That uncertainty has helped erode our collective trust in our science, says Bruce V. Lewenstein, a professor of science communication at Cornell University. The solution, in the eyes of many reformers, centers on greater openness. But in the world of academic publishing, debates over "openness" have mostly meant the push to eliminate subscription fees, rather than the openness of peer review and broader scientific processes.

Some journals are experimenting with notions of crowdsourced peer review. Preprint servers may be the most developed form of that idea. But other variants not yet widely adopted include the open publication of exchanges between authors and their reviewers.

Advocates of that idea include Erin K. O'Shea, president of the Howard Hughes Medical Institute, who outlined the approach at a conference the institute hosted last year. Along with publishing peer reviews — either anonymously or with attribution — O'Shea called for journals to establish systems that display "robust post-publication evaluations." She also suggested that authors, rather than editors, decide whether their manuscript is ultimately published, "removing the notion that publication itself is a quality-defining step."

But in the end, says Michail Kovanis, a French researcher who studies ways of improving peer review, universities themselves hold the power over the future of peer review, because they control promotions and salaries, and therefore can insist on practices that reflect quality rather than quantity.

If they fail to do that, says Kovanis, a data scientist at Inserm, the French Institute of Health and Medical Research, journals will continue to grow beyond the realistic capacity of their reviewers to meaningfully evaluate scientific work. "The ones who give money," he says, "are the ones who can enforce."

Paul Basken was a government policy and science reporter with The Chronicle.

Originally published on March 4, 2018

OPINION

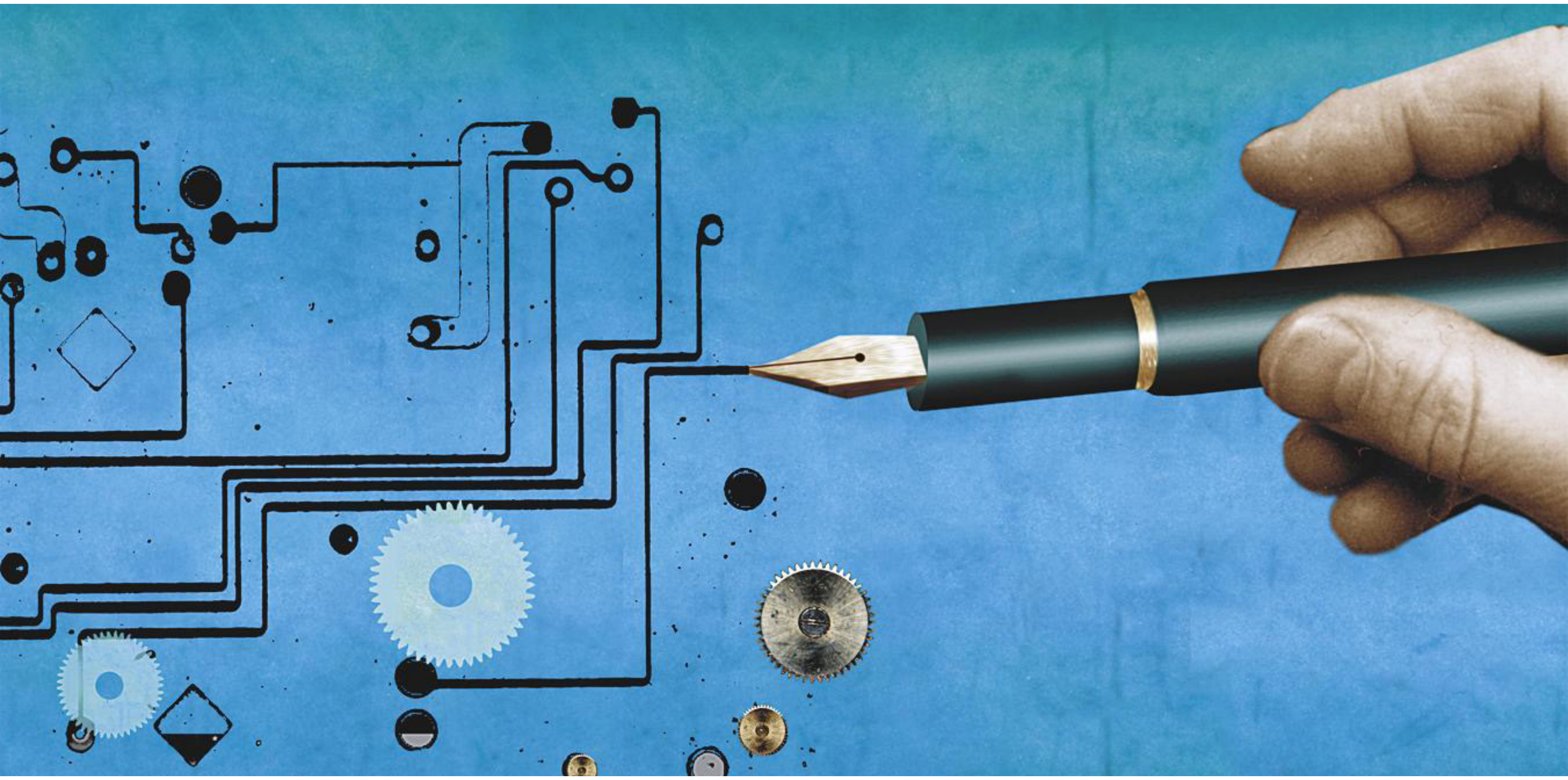
Artificial Intelligence Is a House Divided

A decades-old rivalry has riven the field. It's time to move on.

BY MICHAEL WOOLDRIDGE

The sun is shining on computer science right now, especially the subfield of artificial intelligence. Not a day goes by without the press breathlessly hailing some new miracle of intelligent machines. The leaders of the field are garlanded with honors, and seem to enjoy a status few academics have ever reached. Eye-watering amounts of money pour into AI, and new technology empires are being forged before our eyes. In 2014, DeepMind, a U.K. company with apparently no products, no customers, no obvious technology, and only about 50 employees, was acquired by Google for the reported sum of $600 million; today, DeepMind employs more than 1,000 people.

Given all this, along with the financial hard times that have hit so many other fields, AI, from the outside, must appear a happy ship. In fact, it's hard to imagine how things could be rosier. But look a little



MICHAEL MORGENSTERN FOR THE CHRONICLE

closer, and you'll see that all is not well in the field. AI is a broad church, and like many churches, it has schisms.

The fiercely controversial subject that has riven the field is perhaps the most basic question in AI: To achieve intelligent machines, should we model the mind or the brain? The former approach is known as *symbolic AI*, and it largely dominated the field for much of its 50-plus years of existence. The latter approach is called *neural networks*. For much of the field's existence, neural nets were regarded as a poor cousin to symbolic AI at best and a dead end at worst. But the current triumphs of AI are based on dramatic advances in neural-network technology, and now it is symbolic AI that is on its back foot. Some neural-net researchers vocally proclaim symbolic AI to be a dead field, and the symbolic AI community is desperately seeking to find a role for their ideas in the new AI.

Tribalism and mindless dogma are not the way forward.

The field of AI was given its name by John McCarthy in 1956. The founder of Stanford University's AI lab, McCarthy was the most influential and outspoken champion of the idea that the route to AI involved building machines that could reason. AI requires that we have computer programs that can compute the right thing to do at any given moment. In McCarthy's view, computing the right thing to do would reduce to logical reasoning: An AI system, according to him, should deduce the correct course of action. (If this makes you think of a certain Mr. Spock, well, you are in good company.)

McCarthy's version of AI is called symbolic AI because the reasoning involves manipulating expressions that are the mathematical equivalent of sentences. These expressions are made up of symbols that mean something in the real world. For example, a robot built according to the McCarthy model might use the symbol room451 to refer to your bedroom, and the symbol cleanUp to refer to the activity of cleaning. So when the robot decides to cleanUp(room451), we can immediately see what it is going to do: clean your bedroom.

There is lots to love about McCarthy's dream. It is simple, elegant, and mathematically clean, and it is transparent. If we want to know why one of McCarthy's robots cleaned your room, we can simply examine its reasoning. McCarthy's dream of AI was at the rather extreme end of the symbolic AI spectrum — it was not even widely accepted in the symbolic AI community, many of whose members believed in slightly "weaker" (and more practical) versions of the dream. But his basic ideas formed the AI orthodoxy for 30 years, from the founding of the field through the late 1980s. And while symbolic AI is no longer center stage for academic programs today, it remains an active area of research.

AI excels in developing ideas that are beautiful in principle, but which simply don't work in the real world, and symbolic AI is perhaps the canonical example of that phenomenon. There are many problems in making McCarthy's vision a reality, but perhaps the most important is that while some problems are well suited to this version of AI (proving mathematical theorems, for example), it just doesn't seem to work on many others. Symbolic AI has made only limited progress on problems that require perceiving and understanding the physical world. And it turns out that perceiving and understanding the physical world is a ubiquitous requirement for AI — you won't get far in building a useful robot if it can't understand what is around it. Knowing where you are and what is around you is by far the biggest obstacle standing in the way of the long-held dream of driverless cars.

By the late 1980s, the problems with the purest versions of symbolic AI caused it to drift out of favor. (McCarthy, a remarkable individual by any standards, never gave up on his dream: He remained committed to it right until his death in 2011 at the age of 84.)

A natural alternative to symbolic AI came to prominence: Instead of modeling high-level reasoning processes, why not instead model the brain? After all, brains are the only things that we know for certain can produce intelligent behavior. Why not start with them?

In AI, this approach is called neural networks. The name derives from neurons, the massively interconnected cell structures that appear in brains and nervous systems. Each neuron is an extremely simple information processing device. But when huge numbers of them are connected together in massive networks, they can produce the miracle that is human intelligence. Neural-net researchers build software versions of these networks, and while they aren't literally trying to simulate brains, the idea is that their networks will learn to produce intelligent behavior, just as in humans.

Neural networks are actually a very old idea — they date from the 1940s, and the work of Warren McCulloch and Walter Pitts, who realized that the natural neural networks that appear in human and animal brains resembled certain electrical circuits. However, McCulloch and Pitts had no means to actually build the structures they hypothesized, and it was not until the 1960s that the idea began to take off.

Frank Rosenblatt, a Cornell psychology professor, developed a model of neural networks that goes by the gloriously retro name of perceptrons — this was the first neural-network model to actually be built, and the model remains relevant today. But work in the nascent field was effectively snuffed out by the publication of a 1969 book, *Perceptrons*, by MIT professors Marvin Minsky and Seymour A. Papert, who were staunchly in favor of the symbolic approach. Their book drew attention to some theoretical limitations of Rosenblatt's model, and it was taken to imply that neural models were fundamentally limited in what they could achieve. Rosenblatt died in a boating accident just two years later, and neural networks lost their most prominent champion. Research into neural networks went into abeyance for nearly two decades.

There is still palpable bitterness about Minsky and Papert's book today. When the book came out, the church of AI was divided, and the two sides have never quite reconciled. When symbolic AI began its slow decline in the late 1980s, neural nets swung into favor for a decade, when new techniques for "training" neural nets were developed, and computers were at last powerful enough for neural nets big enough to do something useful. But the resurgence was short lived. By the end of the 1990s, neural nets were yet again in decline, having again hit the limits of what computers of the day could do. A decade later, however, the pendulum swung again, and this time the interest in neural networks was unprecedented.

It is easy to get over-excited by recent progress. Deep learning alone will not take us to the ultimate dream of AI.

Three ingredients came together to drive the new neural-network revolution. First were some scientific advances, called "deep learning" (basically, bigger and richer neural networks). Second, computer-processing power got cheap enough to make large neural networks affordable. Third, and just as important, was the availability of lots and lots of data: Neural networks are data hungry. And we are, of course, now in the era of "big data."

The last decade has seen an unprecedented wave of success stories in AI, and it is these successes that have led to the current AI frenzy. In 2016, DeepMind famously demonstrated a Go-playing program that could reliably beat world-champion players. This fall, a DeepMind project called AlphaFold made a giant step forward in biology by better predicting protein structures ("'It Will Change Everything,'" began a headline in *Nature*). Elsewhere, rapid progress has been made on driverless-car technology — last year, Waymo, Google's driverless-car company, launched a completely driver-free taxi service in Phoenix.

Recognition for the leaders of the new AI came in 2018, when Geoffrey Hinton, Yann LeCun, and Yoshua Bengio, three of the most prominent champions of neural networks, who had stuck with the technology throughout the lean years, were awarded the Turing Award — often described as the Nobel Prize for computing — which comes with $1 million in prize money. There could have been no clearer signal that, at last, neural networks had been accepted into the mainstream.

All these successes are predominantly the successes of deep learning. Symbolic AI has played a part in some of these — but strictly in a supporting role, never center stage.

While the media tend to generically apply the "AI" label to all recent advances, some members of the deep-learning community profoundly dislike it. They identify it with a long list of failed ideas that have characterized the history of AI, of which the symbolic AI project, they believe, is the most prominent, and most egregious.

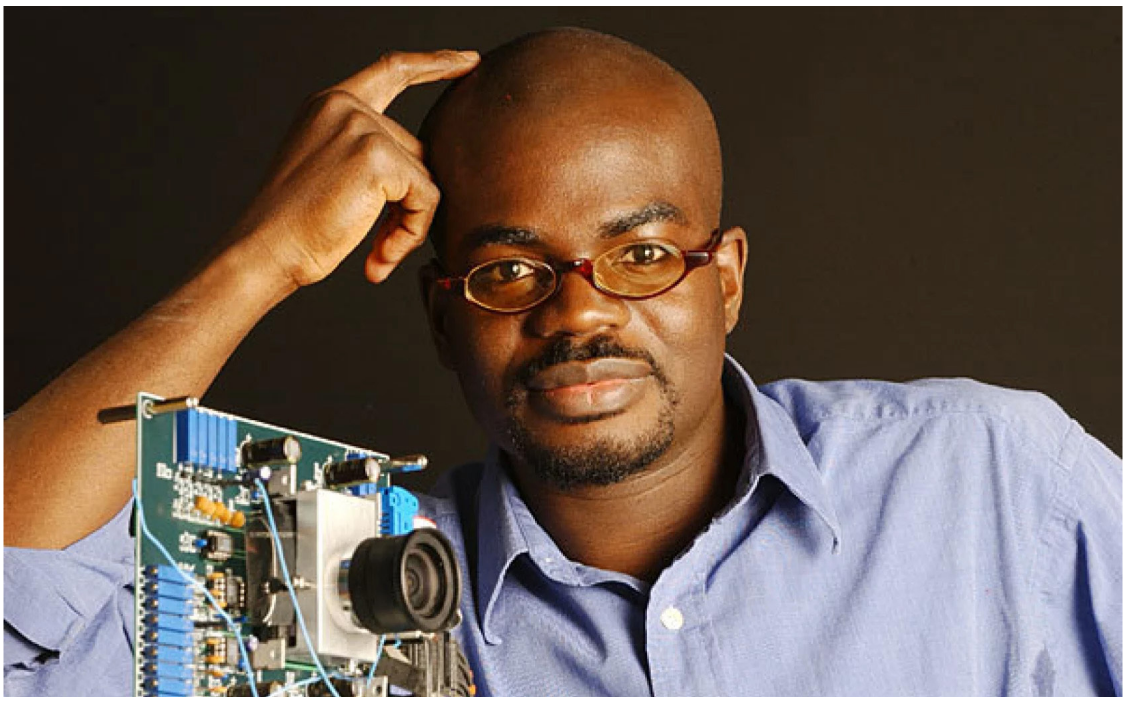
The successes of deep learning this century are real and exciting and deserve to be celebrated and applauded. And those researchers that stuck with neural networks through the lean years deserve our admiration for their vision and determination in the face, at times, of ridicule or scorn from their academic peers. But it is easy to get overexcited by recent progress. Deep learning alone will not take us to the ultimate dream of AI. It is surely one of the key ingredients, but there will be many others — some of which we probably cannot imagine right now. For all the progress we have made, we will not achieve the dream soon — if we achieve it at all. There is, I believe, no silver bullet for AI. Neural networks and symbolic AI each succeed with different aspects of intelligent behavior. Tribalism and mindless dogma are not the way forward: We must consider each other's ideas and learn from them. And to do this, we must first cast away the bitterness of ancient rivalries.

Michael Wooldridge is the head of the department of computer science at the University of Oxford. This essay was adapted from his book, A Brief History of Artificial Intelligence *(Flatiron Press).*

Originally published on January 20, 2021

An Immigrant Scholar Leads the Charge Against Computing's Biggest Roadblock

BY PAUL BASKEN



U. OF PENNSYLVANIA

After 30 years of studying the brain as a guide to building faster computers, Kwabena Boahen may have given his fellow researchers a much-needed template for finishing the job. His story, as a Stanford professor and Ghanaian emigrant, exemplifies what America has to gain — or lose — by inhibiting immigration.

You might not know it from the promising talk of conversational computers, self-driving cars and lifelike human prosthetics, but computing is confronting a crisis. After decades of rapid acceleration, the speed of transistors — computerization's fundamental building blocks — is hitting a wall.

But to the relief of researchers and industry leaders, a superstar scholar who immigrated to the United States from Ghana may have found a way forward. It concerns neuromorphic computing — mimicking the processes of the human brain — which has long been recognized as both a hugely daunting technological challenge and the likely key to solving the transistor problem.

Amid broad scientific uncertainties over how best to proceed, Kwabena Boahen, a professor of bioengineering and electrical engineering at Stanford University who is one of the field's pioneers, has now outlined a badly needed road map for finishing the job.

It appears to be a "transitional moment" for neuromorphic computing, said R. Stanley Williams, a senior fellow at Hewlett Packard Labs.

The central barrier to a future of ever-smarter products is well understood: Given the simple constraints of matter, transistors are now being made almost as small as possible. Researchers sounding the alarm bells "have been at this for a long time," said Randal E. Bryant, a professor of computer science at Carnegie Mellon University. As Bryant acknowledged at a recent scientific conference, "there's a certain point where atoms are atoms."

That's where Boahen and his fellow "neuromorphs" come in. Neuromorphic computing is an extremely complicated mix of advanced sciences, including human biology, physics, mathematics, and chemistry. The goal is to build automated systems far faster and more efficient than current computer-chip technologies allow, largely by using the human brain as their template.

A key insight is a back-to-the-future concept: Give up our heavy reliance on digital systems, and return to analog.

The Analog Brain

At its foundation, the transistor is a simple device: three wires sticking out of a tiny sandwich of materials arranged so that a sufficient electrical signal applied to one of those wires enables current to flow between the other two.

By taking the basic on-or-off switching ability of a single transistor — a signal that could be represented as either a zero or a one — and multiplying it thousands of times within a device, scientists and engineers have made generations of astounding products, including the computers that helped put humans on the moon.

The yes-or-no character of those signals provides some clear advantages over analog, which refers to a signal that varies within a broad range of possibilities. Old analog TVs, for instance, were notoriously susceptible to electrical interference from changes in weather, regularly giving viewers fuzzy and faded pictures as slight electrical disturbances from distant broadcast antennas were reflected in the televised images.

Digital signals, by contrast, put a priority on being flawless. For a TV, or a computer, one strategy to avoid transmission errors is to send signals at a relatively high rate of power to ensure each unit of data is clearly read as a zero or a one.

But that demand for power limits size and performance. And a reliance on digital is just one way in which modern computer systems remain, at least as compared to the brain, highly inefficient. While modern digital computers feed their processing tasks into thousands of "cores," the brain's army of processors — neurons — number in the billions.

"During his talk you could see people's eyes growing to the size of whale's eyes, and smoke coming out of their ears."

The brain is also, in computer terms, an ever-adjusting mix of software and hardware that grows and disposes of its processing equipment as needed. A digital representation does help computers process data very accurately. But a computer's inability to seamlessly trade precision for speed, and to widely share and adjust processing duties, impose far more critical limitations.

The brain knows these things. A product of billions of years of evolution, the essential organ of humanity is a marvel of efficiency. Boahen calculates that the human brain has the computational power of a refrigerator-sized supercomputer that is about 50 times heavier, takes up 100 times more space, and consumes 100 times more power than the brain.

The brain does use a type of binary signal to relay data, he said. But those signals are many thousands of tiny blips of electrical information that are processed by the brain in essentially an analog, or continuous, fashion. That combination of digital and analog, Boahen said, is "fundamental to the difference between the computer and the brain."

A Neuromorph's Manifesto

Boahen has been working for 30 years to make computers act more like brains. But it was in October, at the Institute of Electrical and Electronics Engineers' first International Conference on Rebooting Computing, that his efforts gained a significantly new level of appreciation.

The son of Albert Adu Boahen, the professor of history at the University of Ghana who helped lead his nation to democratic rule, Boahen did not arrive at the session as a revolutionary. But he may have left as one.

"During his talk," said Williams, who is also an adjunct professor of chemistry at UCLA, "you could see people's eyes growing to the size of whale's eyes, and smoke coming out of their ears."

In a 53-minute presentation, Boahen outlined five main challenges to producing a working computer based on neuromorphic principles. Each point was highly technical — covering such challenges as developing circuits that "gracefully" respond to signals and work without external timing cues — but the effect was of a salvo.

Many in the audience had spent years pursuing neuromorphic computing, Williams said, and Boahen had just given the research community a concise manifesto making clear which avenues of exploration deserved more attention and which were probably a waste of time.

"He effectively convinced everybody else in the room, who thought they were doing neuromorphic computing, that they didn't actually know what it was," Williams told last month's annual conference of the American Association for the Advancement of Science in Boston. "In his talk, he defined for the first time really what neuromorphic computing is."

The presentation also showed how far ahead of the pack Boahen appears to be. He already has built a small robot with a functioning mechanical arm using neuromorphic chips. Among the five challenges he listed, Boahen already has largely solved four, Williams said in an interview. "Others maybe have gotten to one or two," he said.

The last remaining obstacle, Boahen explained, involves figuring out ways to accurately convey the continuously changing signals in the series of "spikes," which are those blips of electrical signal that the brain uses to pass along data. Patterns of those spikes are like super-precise versions of the bar codes found on supermarket items, and reading their timing and placement is central to the brain's internal communication.

Boahen's success is a tribute to the value of thinking slowly and carefully, and not worrying about "the latest trend," Williams said. It's also a reminder, at this moment in American history, of the value of immigrants, he said.

A 'Traumatizing' Climate

After a recent speech at the University of Illinois at Urbana-Champaign, Boahen met with a graduate student from Lebanon who had a long list of questions about his work. The two also talked about U.S. politics, which had become "really scary for him," Boahen said of the student. Lebanon is not among the majority-Muslim countries affected by President Trump's travel-ban proposals, and the student is Christian.

But the student spoke of fear that his elderly parents might die if denied entry at a U.S. airport and forced to make back-to-back 20-hour flights. He described being warmly accepted on campus but encountering racism not too far out of town. And the student wondered if those conditions, combined with a lukewarm American commitment to funding research, mean that he'll eventually have to take his computing talents elsewhere.

Boahen said he had tried to reassure the student, who asked not to be identified by name, that circumstances change. "But it's really a shame," Boahen said, "that people who just came here don't have that context, and it's really traumatizing what we are putting them through."

At least Boahen is not worried for himself: He obtained U.S. citizenship last year. "I could see the writing on the wall," he said. "My father fought a dictatorship, and I don't want to try."

Paul Basken was a government policy and science reporter with The Chronicle of Higher Education.

Originally published on March 8, 2017

Twitter for Scientists: an Idea Whose Time Has Finally Come?

BY PAUL BASKEN

Tweeting has long posed a dilemma for scientists.

There's abundant evidence that widely sharing a research finding in just one or two simple sentences greatly increases its use and effectiveness.

But, ugh, that usually means Twitter — in the eyes of many, a low-attention-span cesspool of trolls, political partisans, and amateur comedians known more for braggadocio and snark than reason and facts.

Now, with federal backing, there's another option.

Known as Polyplexus, meaning "a network of many," it's a compilation of 300-character summaries of research findings, created with the idea of driving crossfield discoveries and spawning public and private funding for follow-up studies.

Unlike Twitter, it's meant to be "a professional environment for research-and-development professionals," said a Polyplexus developer, John A. Main, a program manager at the Pentagon's Defense Advanced Research Projects Agency, or Darpa.

Darpa and its $3-billion annual budget operate well beyond the military realm, and Polyplexus is an example. The newly established website is based on the sense that all fields of science and technology could move faster if researchers spent less time on the between-projects chores of synthesizing and disseminating findings.

Michael Goldblatt, an entrepreneur and former official at the Pentagon's Defense Advanced Research Projects Agency, came up with the Polyplexus concept, "the idea of trying to create research into digestible chunks to drive citations and drive understanding and learning."

"You're not going to accelerate a research project once it starts," said Main, a former associate professor of mechanical engineering at the University of Kentucky. "The time to accelerate it is before it starts — you want to get to the start date faster than you've ever done before, and that's what we're going after."

How eagerly scientists will actually use it is an open question.

The first researcher to submit a posting to Polyplexus — known as a "micropub," or micropublication — was Victor A. Benjamin, an assistant professor of business at Arizona State University. In his submission Benjamin summarized a 2016 study by researchers at the Stevens Institute of Technology who found that even a small number of false online reviews can significantly harm a hotel.

Benjamin said he had written that micropub, even though he wasn't part of the team that conducted the hotel-review study, largely because Darpa, in a bid to publicize Polyplexus, is promising several $100,000 grants based on ideas that are derived from micropubs. It's one of about a dozen micropubs that Benjamin submitted to Polyplexus within its first week of operation.

Once that introductory-grant incentive vanishes, Benjamin said, he's not sure how active he'll be on the site. A relative newcomer to the business school, he is an expert in machine learning and computational linguistics, studying how to mine useful intelligence from online conversations. His academic experience to date, he said, validates Darpa's understanding of the need to somehow accelerate the scientific process.

In his days as a graduate student in computer science, Benjamin said, social media was necessary to keep current on machine-learning technology. As a professor, though, engaging with platforms such as Twitter has struck Benjamin as a far lower priority than getting published in traditional academic journals.

The 'Guy in His Garage'

That calculation appears especially common in places such as the business school, Benjamin said. As in many fields, he said, business-school incentives tend to reward journal-publication rates and private-sector contract work that offers limited value to the outside world. "A lot of them are not interested in it," he said, describing his business-school colleagues and their attitudes toward social media.

That reflects a common misprioritization in academe, said Andrew M. Ibrahim, a University of Michigan medical doctor who studies ways to cut costs and improve quality in health care. Ibrahim advocates not only that scientists use social-media platforms such as Twitter, but also that they use them better. He co-authored a study, published last year in the *Annals of Surgery*, that showed tweets with a graphic data summary attract eight times as many views as do tweets that provide only the title of a journal article.

"You want to get to the start date faster than you've ever done before, and that's what we're going after."

"It's worth a couple of more hours of your time to put it in clear-enough form that makes your work more accessible," said Ibrahim, a staff surgeon at the university.

With or without the graphical element, the process of whittling down a research finding to a concise thought helps authors improve both their writing and their scientific thinking, Ibrahim said. Some scientists, after attending a workshop on producing more-effective tweets, have found they want to rewrite the journal articles they've been trying to describe, he said. "It's a hidden curriculum of teaching people to write more clearly," Ibrahim said of the workshop.

That's one insight that motivated Michael J. Goldblatt, an entrepreneur and former Darpa official who came up with the Polyplexus concept and now has a Darpa contract to build its web presence. "The idea of trying to create research into digestible chunks to drive citations and drive understanding and learning is a very useful thing," said Goldblatt, whose career has also included a stint teaching law at the University of California at Davis.

Goldblatt's and Darpa's more central goal, however, is to greatly expand the active scientific community, beyond the usual suspects in academic or corporate settings.

Although participation in Polyplexus is limited to academic scientists, Darpa plans to open it to anyone involved in legitimate scientific exploration, including the proverbial "guy in his garage," Goldblatt said.

From Evidence to Conjecture

How to judge legitimacy? Darpa plans to subject each proposed micropub to a three-question vote in which other participants must agree on the baseline quality of its evidence and sourcing.

It's a hidden curriculum of teaching people to write more clearly.

And beginning next month, participants will be asked to take two existing micropubs describing a research finding (known as "evidence micropubs") and create another (known as a "conjecture micropub") that briefly describes a key relationship or synthesis of the two evidence micropubs. The hope, Main said, is to make the writing of conjecture micropubs a curiosity-driven game — the kind of thing someone might do with idle time on a phone while riding a train.

For the third and final stage of Polyplexus, Main said, both public and private funders of science will be asked to read the theories outlined in the conjecture micropubs and see them as ideas for financing future research projects.

At the moment, however, the funders — even Darpa's fellow federal science agencies — seem a bit tentative. Only three companies have signed up to work with Darpa on Polyplexus. No government agencies have done so.

Initial interest among researchers appears a bit better. A first round of 260 invitations to college and university scientists has attracted 64 registered users of Polyplexus. Benjamin is not sure how many more of his fellow scientists will join, though he sees good reasons for them to consider it, especially if they are still early in their careers. Research funding tends to flow to scientists and fields with existing connections to decision makers, Benjamin said, and Polyplexus might start to redraw those lines. For scientists hunting opportunities, he said, "maybe this could help them break in."

Paul Basken was a government policy and science reporter with The Chronicle of Higher Education.

Originally published on March 19, 2018

As the Drive to Share Data Intensifies, Can Standards Keep Up?

BY PAUL BASKEN

Patrick J. Curran struggles with the problem when studying alcoholism in families. Quynh C. Nguyen sees it when analyzing housing-voucher programs. And the Nobel laureate Harold E. Varmus encounters it while developing genomic databases for cancer patients.

NIH

The work of integrating data from different sources can be expensive. For now that cost is often considered part of individual grant support, says William Riley, director of the National Institute of Health's Office of Behavioral and Social Sciences Research.

Their trouble isn't with sharing their data — all three professors are eager participants in the open-data revolution.

Instead, the problem is confidently sharing and interpreting data — huge amounts of it — with relevance and accuracy.

As they and other scientists embrace sharing, they're finding that computer systems are quite good at storing and easing access to the enormous quantities of information they generate. But comparing and synthesizing all that data, in differing formats and styles and methods, requires human skill and judgment. And even the best aren't sure how to do it, raising questions of whether the nationwide rush toward open data will really mean a momentous revolution in scientific progress or just a whole new level of gnarly reproducibility issues.

Curran is a professor of psychology at the University of North Carolina at Chapel Hill who studies the effects of alcoholic parents on their children. He combines findings from multiple studies and sees a challenge lurking in the varied scientific meanings and assessments that professional colleagues apply to terms such as "anxiety" and "depression."

"The thing that keeps me up at night," he said, "is, Am I making a substantive theoretical conclusion that is based on some artifact of how we scored the scale?"

Such questions represent a much-less-discussed aspect of the push for open science: Some researchers believe that the intricacies of translating and synchronizing data accurately are getting too little attention, even though reproducibility is already a major struggle. And those details will only grow more important as scientists begin writing the code for a future in which computers routinely extract answers from data piles far too big for any human to handle.

Varmus, a university professor of medicine at Weill Cornell Medical College and former director of the National Institutes of Health, is one of the world's best-known cancer experts. His concerns over data synchronization include the ways of describing for computerized processing the multitude of ways that patients develop tumors and respond to drugs or radiation or other therapies.

"It becomes a very complicated business that has to be settled, in my view, at a very early stage of having accepted terms for what we do to patients and how they respond," Varmus told an NIH conference on open data this month.

Accuracy and Precision

Over all, the benefits of computerizing data are expected to vastly outweigh the worries, in virtually all academic fields. In medicine, it could mean defeating devastating diseases by gleaning crucial insights into patient experiences that have already occurred, during thousands of costly clinical trials, rather than spending many years and millions of dollars conducting new trials.

Before those benefits can be realized, however, scientists face the daunting task of meshing the many differences in the questions, procedures, notational styles, and measurement units seen in vast data collections.

Nguyen, an assistant professor of health promotion and education at the University of Utah, uses social-media data to assess the health of urban neighborhoods and the success of government policies such as housing vouchers. The accuracy-related challenges, she said, include handling the differences in neighborhood boundaries used by various public agencies. Some researchers, Nguyen said, accept broad overlaps in residential identifications for the sake of convenience. But that comes as a cost, she said. "You definitely lose accuracy and precision that way," she said.

Such case-by-case assessments are one issue. Potentially bigger decisions await as notational styles and conversion formulas for each academic field develop and harden. Then, some scientists fear, they may be in the position of using statistical shortcuts that become automated and amplified during the process of combining databases.

For his work on alcoholism in families, Curran has been using three different data sets housed at Arizona State University, the University of Michigan, and the University of Missouri. All three are aimed at studying children with and without an alcoholic parent, but involve different developmental periods, different measures of core behaviors, and widely differing community types — suburban Phoenix, rural areas around Columbia, Mo., and urban Detroit and Lansing.

"The thing that keeps me up at night is, Am I making a substantive theoretical conclusion that is based on some artifact of how we scored the scale?"

It's a major job just to consider the immediate challenges involved in synchronizing such data, Curran said. "I hadn't really thought about the impact, especially the longer-term impact, of how laying down some of these common definitions and terminologies sets up expectations and directions for research" in the future, he said.

That same process is taking place across many academic fields, said Brian A. Nosek, a co-founder and director of the nonprofit Center for Open Science. A single set of field-specific data-conversion standards probably won't emerge for many academic disciplines until there's enough shared data in open formats to make such standards absolutely necessary, Nosek said.

Major funding agencies such as the NIH and the National Science Foundation have been financing work to create the standards and the conversion systems for databases that already exist. That process is fairly advanced in some fields, such as genomics, where the variables are relatively discrete. But a lot more progress is needed in the social sciences, where terms of reference tend to be highly subjective.

And in many cases, Curran said, the amount of grant support for data integration is unrealistically small. "It turns out this is vastly more complicated than you anticipate," he said.

Neither the NIH nor the NSF could provide figures on how much money they spend on such work. It's difficult to count because the job of synchronizing terminologies is often considered part of individual grant support, said William T. Riley, director of the NIH's Office of Behavioral and Social Sciences Research.

One approach being evaluated by the Defense Advanced Research Projects Agency, which specializes in novel solutions, involves the use of artificial intelligence strategies. As an experiment, the agency, which is known as Darpa, gave a team from the Rensselaer Polytechnic Institute some 300 terabytes of largely unlabeled data from tests of how a collection of composite metal samples made in various ways performed under tests related to flight worthiness.

Using data-analysis strategies that a law firm might use to extract information from a large collection of emails, the RPI scientists not only figured out what the data represented but used it to make predictions of future tests of such metals. The idea, said William C. Regli, deputy director of Darpa's Defense Sciences Office, is to let researchers share their data without the burden of also trying to ensure that some future user can interpret it.

"It's clear that to address this problem in kind of the existing way, we're going to drown," Regli said. He acknowledged, however, that standardization regimes may remain essential in fields such as the social sciences that rely on data that reflects largely subjective measures.

The Standardization Problem

Scientific standardization has long suffered from insufficient attention, said Kai R. Larsen, an associate professor of information management at the University of Colorado at Boulder. It's a big part of the reason why scientific journals are flooded with studies that often repeat the same basic findings, over and over again, just using different terms, he said.

As one recent example, Larsen was asked to review a paper on the "internet of things" — the growing network connectivity of everyday objects. Just looking through the paper, he said, he quickly recognized it as essentially a repetition of past analyses of patterns of new-technology acceptance. As such, the paper represented to him another argument for creating a coherent database of known behavioral patterns. "I asked to be relieved of the job of reviewing this paper," he said, "because I knew I would be very negative toward it."

It's "absolutely" the case that the availability of data is ahead of the availability of tools that can process that information accurately and completely.

Computers clearly can do better than humans in recognizing such patterns, said-Larsen, who works with NSF support to develop automated text-mining technologies for behavioral studies. But first humans need to create the systems for doing that, he said. "There's tens of thousands of behavioral and social-science researchers out there, producing papers as fast as they can, and there's literally a handful of projects out there subsisting on minimal grants trying to organize what they're doing," Larsen said. "We can't keep up with that."

Ultimately, such experts said, the move toward greater data sharing — with accurate standards for combining databases — will depend on whether researchers can begin to show progress in using such techniques and whether universities reward them for it.

But with relatively little support from funders such as NIH and NSF, the creation of data standards is tedious and expensive, said Mark A. Musen, a professor of medicine at Stanford University and director of the Stanford Center for Biomedical Informatics Research. "Right now we're in a situation where people do this out of the goodness of their heart," Musen said.

That's a recipe for reproducibility problems, said Douglas A. Mata, a clinical fellow at Harvard Medical School who uses meta-analyses to study depression in medical-school students. Meta-analyses are the more traditional method of summing up existing studies, which involves combining summary-level data from previously published studies. A future of robust data sharing could instead allow deeper findings based on analyses of individual patient-level data that form each of the published studies.

Mata said automated programs that routinely introduce inaccuracies into data-sharing protocols aren't likely to be a major problem, because future researchers will probably have a variety of choices and methods for handling their analyses. That said, a tendency to take shortcuts — such as not using statisticians in comparison studies — is a major cause of the current reproducibility crisis in science and could continue to cause problems in the future, he said. Too many research groups are "already making use of prepackaged tools where they can just kind of unthinkingly click a button and accept whatever the program puts out," Mata said.

Despite such warnings, the institutional push for open data in science still tends to focus more heavily on expanding access to data than on figuring out how to accurately handle huge amounts of data once it becomes available.

Last week, eight leading private funders of scientific research announced the creation of the Open Research Funders Group. The group, whose members include the Bill & Melinda Gates Foundation and the Alfred P. Sloan Foundation, said in its announcement that its members "are committed to using their positions to foster more open sharing of research articles and data."

But the group's project coordinator, Greg Tananbaum, acknowledged in an interview that ensuring accuracy needs even more attention. It's "absolutely" the case, Tananbaum said, that the availability of data is ahead of the availability of tools that can process that information accurately and completely. "There's no doubt about it," he said. "I don't think anyone could argue otherwise."

Paul Basken was a government policy and science reporter with The Chronicle of Higher Education.

Originally published on December 20, 2016

ROLAND SÁRKÁNY FOR THE CHRONICLE

Why Researchers Shouldn't Share All Their Data

BY NATHAN SCHNEIDER

In 1925, the year Gertrude Stein published her 1,000-page book, *The Making of Americans*, she felt the need to explain, in lectures at the Universities of Cambridge and Oxford, why it was so long. Her aim, she said, was to depict a tense she described as the "continuous present." Her method for doing so — and the culprit for her verbosity — was "using everything."

The thought of "using everything" should send a chill down the spine of any researcher. Producing publishable results from an investigation typically requires managing far more material than can fit into the publication format. A secret realm of dark data resides in the notebooks and hard drives of the data-gatherers; they judge that data to be excess, but who knows? What if it is not? In this excess, in this "everything," surely there are the ingredients of unrealized cures and upheavals. And in this excess, every researcher knows, are parts of our process we would rather not share. There, we are vulnerable.

During the years when I worked primarily as a reporter, this excess haunted me. I have hours-long interview transcripts from which only a few words, if that, appeared in an article — not because those words were the only ones of value, but because of the needs and constraints of that particular article.

Eventually I started to collect my reporting notes into public notebooks, including one that is the basis of my next book and of this article. Doing so has become an easy way to share what I gather with people who want more than what the published work can hold. It has also inclined me to take better notes, and to notice more threads of connection among disparate projects. But I have also found myself holding back. I hide my detailed reading notes behind a password. To protect my sources, interview recordings and transcripts remain offline altogether. Field notes stay in paper notebooks.

Predictably, the hard sciences have charged ahead on this curve. Far-flung research teams frequently collaborate in examining common data sets. Some government grants come with the requirement of publishing open data as well. The resulting demand has warranted open-notebook software like Jupyter, Observable, and Zenodo. Researchers frequently post their own code on platforms like GitHub or GitLab. These are based on Git, a tool designed for large groups collaborating on open-source software. Among other features, Git keeps a meticulous record of a given project's version history. It remembers every change and every bug. Likewise, open-source software communities tend to regard maximal transparency as an intrinsic good.

Some humanists have followed suit. The Rice University historian W. Caleb McDaniel, for instance, has developed a system that feeds his research notes into a public wiki, thanks to a mix of open-source tools and scripts he had to code for himself. Scholars across many fields share their bibliographies online using tools like Zotero, which was developed through an academic collaboration. Hypothesis, a nonprofit platform, enables users to make, collect, and share annotations on nearly any website. Requiring my students to use it, I've found, is a handy way of checking that they're doing their reading assignments and getting them to debate their interpretations.

Among journalists, there has been talk at times of "open journalism" as a new paradigm for reportage that extends beyond just the polished report. In 2011, as an executive in residence at the University of Southern California, the former *Sacramento Bee* editor Melanie Sill published a report called "The Case for Open Journalism Now: A New Framework for Informing Communities." Yet her call has not been widely answered, and it remains the definitive work on the subject. As editor in chief of the British daily *The Guardian*, Alan Rusbridger adopted open journalism as his strategy for the newspaper, but he left the job in 2015. Organizations such as BuzzFeed and ProPublica, at least, publish code and data sets on GitHub.

Every Bitcoin transaction is recorded in the open, and the same mechanisms could record acts of scholarly research, writing, and certification.

The emerging opportunities for self-exposure extend from research to the writing process. Kathleen Fitzpatrick, now a professor of English and director of digital humanities at Michigan State University, undertook a widely publicized "open review" process for her 2011 book, *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. She waited until she had finished a full draft, but one need not do so. I version-tracked the entire drafting of my latest book in Git, which means nearly my whole process of writing and revision could become immediately public if I simply pushed it to GitHub.

I don't think I will do that.

The "blockchain" technology underlying Bitcoin, which makes possible secure databases with no centralized authority, could open the doors of transparency still farther. Every Bitcoin transaction is recorded in the open, and the same mechanisms could record acts of scholarly research, writing, and certification. Natalie Smolenski, an anthropologist who works for the blockchain start-up Learning Machine, wants to use such tools to transform how we register academic achievements. Yet in her paper "Academic Decentralization in an Era of Digital Decentralization," Smolenski reserves some of her most arresting words for transparency.

"Transparency," she writes, "is socially pornographic and facilitates violence." It can mean revealing data about ourselves without the context we might otherwise provide. It can objectify the researcher and the process, inviting viewers to feel a false sense of intimacy, of inside knowledge.

This is a sentiment I've sometimes come across as a minority opinion in hacker communities I've studied. It's expressed most often by participants representing vulnerable identity groups, people for whom more self-exposure can mean more vulnerability. In the academy, I've heard it from those on "watch lists," whose every move is scrutinized for political reasons, in search of what might be construed as a misstep. Graduate students are often taught to be careful what they publish, for fear of being pigeonholed too early. Too much self-exposure might compromise a career. It might also muddle one's message.

"Meaning is not transparent," Smolenski told me in an email; rather, she stresses that meaningful communication happens through context and time. She contrasts the exposure of radical transparency to what the more careful, intentional cultivation of intimacy allows: "provisionality." Without the requirement of transparency, one can try on ideas, see how they look and work, then take them off.

Feminist techies, while sympathetic to calls for open-sourcing everything, have also recoiled at the most extreme demands to be transparent. As Ellen Marie Dash, a software developer, wrote in the magazine *Model View Culture*, for those accustomed to harassment online, the call for openness feels like a call to invite more harassment. "The only way to handle this sort of problem properly," Dash contends, "is by explicitly placing consent and safety over openness and transparency."

Dash also questions whether dumping vast amounts of information online counts as transparency in the first place: "What you wind up with is a company that produces so much unorganized, uninteresting and irrelevant data that you can't find meaningful information."

It's the old paradox of Jorge Luis Borges's Library of Babel, which contains such multitudes that little of use can be found. And this is the trouble with reading Gertrude Stein, as soon as you're ready to leave her bewildering "continuous present." The tools that afford us new opportunities for openness and collaboration also come at the risk of obfuscation and danger.

Nathan Schneider is an assistant professor of media studies at the University of Colorado at Boulder who wrote Everything for Everyone: The Radical Tradition That Is Shaping the Next Economy*, (Nation Books).*

Originally published on April 8, 2018

Powering a community of supercomputer users

AS A COMPREHENSIVE UNIVERSITY WITH BROAD EXPERIENCE IN THE APPLICATION OF DATA SCIENCE AND ARTIFICIAL INTELLIGENCE (AI) IN RESEARCH, SMU IS DEEPLY COMMITTED TO THE ACTIVE APPLICATION OF EMERGING TECHNOLOGIES FOR SOCIETAL GOOD WHILE PROVIDING WORLD-CLASS EDUCATION.

Since our founding more than a century ago, our academic enterprise has valued future-focused investments. Our strategic plan identifies high-performance computing as a primary and essential resource for research and education across the disciplines, including research-intensive doctoral programs through SMU's new Moody School of Graduate and Advanced Studies.

At the heart of investments that drive research with impact and our leadership in the innovative world of artificial intelligence research and applications is our supercomputer, which has served as an essential research tool for scholars across disciplines for over a decade.

Investments in supercomputer upgrades have resulted in establishing SMU as the home of one of the most powerful university computing clusters in Texas, while ensuring free and open access to all members of its academic community. Over the past several years, our highly active user base of faculty and students across campus has trended upward with a relentless increase of GPU hours to perform impactful research that utilizes AI and deep learning. We're collaborating with accelerated computing leader NVIDIA, whose DGX SuperPOD™ technology boosts SMU's AI and machine learning speeds 25 times faster than current levels. In combining accessibility and cutting-edge technological capabilities, SMU is fostering a vibrant community of supercomputer users, supporting world-changing work that shapes how we learn from the past, affect the present and prepare for the future.

The impact of this initiative brings advanced artificial intelligence power to the Dallas region, opening new doors for AI-reliant campus researchers such as Stephen Sekula, chair of the Department of Physics and professor of physics at SMU. Sekula's work focuses on the Higgs particle, studying its behavior through interactions with other building blocks of matter. His research is central to SMU's participation in ATLAS, a major Large Hadron Collider experiment performed at the CERN laboratory in Geneva, Switzerland. With the SMU supercomputer's enhanced AI infrastructure, Sekula is able

to reproduce exact simulations of CERN's experimental data. "The fact that we are able to produce simulation that is perfect fidelity with simulation from CERN is a huge asset for SMU," he says. "The supercomputer gives us seamless and flexible access to what we need, when we need it, and allows new ideas to go forward without dependence on external resources to provide what we need." Empowered by the unparalleled access and advanced enhancements of SMU's supercomputer, Sekula moves closer to unlocking key insights into the fundamental structure of the universe, shaping how we conceive space, time and life itself.

In the Roy M. Huffington Department of Earth Sciences, Professor Zhong Lu uses supercomputing in his study of geophysics. Each year, heavy rains pour into the Pacific Northwest, resulting in hundreds of landslides that endanger lives and damage infrastructure. In collaboration with the United States Geological Survey, Lu and his team monitored these events, archiving millions of high-resolution satellite images. Utilizing SMU's supercomputer, they were able to analyze this immense archive, exploring its data to detect and predict where landslides might occur next. As the world grapples with climate change, geohazard research will remain critical to the livelihoods of all people. In optimizing supercomputing to help mitigate these risks, Lu's work moves us toward a safer, more prepared tomorrow.

SMU's supercomputer also serves as an essential tool that connects scientific laboratory experiments with high-quality computer modeling. It enables the initiatives of the Computational and Theoretical Chemistry group (CATCO), led by Chemistry Department Chair Elfi Kraka, to help solve pending problems in chemistry and beyond via computer simulations. A flagship of CATCO is the Unified Reaction Valley Approach (URVA), developed to monitor complex subatomic reactions in remarkable detail. Running these simulations on SMU's supercomputer, CATCO has performed over 800 chemical reactions, fueling discoveries that spur the design of new materials with the potential to help develop promising drug candidates for a healthier world.

As SMU's academic community continues to flourish through supercomputing and AI, the University is establishing its place as a regional leader in 21st century digital technologies and supercharging pathways to data-driven discoveries with the potential to dramatically drive educational and socioeconomic changes across communities.

THE CHRONICLE
OF HIGHER EDUCATION®

1255 23rd Street, N.W., 7th Floor
Washington, D.C. 20037
202 466 1000 | Chronicle.com