Science & Technology

The Ghost Files

US historians have long complained about gaps in the National Archives. Can big-data analysis show what kinds of information the government is keeping classified?

By
David J. Craig
|
Winter 2013-14

Davide Bonazzi

Matthew Connelly had an idea for a book.

The Pentagon, he realized, was one of the first organizations ever to undertake a large, scientifically based effort to predict the future. During the Cold War, it had invested billions of dollars into the development of computer-based war games, statistical models, and elaborate role-playing exercises in hopes of anticipating Soviet military activity. How successful had the Pentagon's program been at

predicting the Soviets' next moves? And how had the Pentagon's predictions been skewed by the group dynamics of the generals, intelligence analysts, diplomats, and statisticians involved? Did they tend to push more cautious or alarmist conclusions? Did they favor predictions that were too forward-looking to be proved wrong while they were still on the job? These were questions that had never before been thoroughly investigated.

"I thought this would provide insights into how all sorts of predictions get made today, whether about climate change, disease outbreaks, or rogue states acquiring nuclear weapons," says Connelly '90CC, who is a professor of history at Columbia. "How seriously should we take these predictions? And what's the best way to gauge their relative validity? The US government has been in the business of forecasting the future for fifty years, so it seemed logical to evaluate its record."

He didn't get very far. In the spring of 2009, a few months after starting his research, Connelly decided it would be impossible to tell the story that he envisioned. Too little information was available. Connelly had spent long hours researching the Pentagon's forecasting efforts at the National Archives in College Park, Maryland, and at other government archives around the country. He had found a decent amount of material related to the program's beginnings in the 1960s, but few records from later decades.

"The Pentagon was certainly making forecasts throughout the course of the Cold War," says Connelly, the author of the 2008 book *Fatal Misconception: The Struggle to Control World Population*. "So it was pretty obvious that the records from the 1970s onward were incomplete."

What Connelly experienced was something that researchers had been complaining about for years: that the National Archives' contemporary holdings had more holes than a donut factory. The problem was that the US government was not releasing classified documents on schedule. Although federal policy requires that most documents labeled "Confidential," "Secret," or "Top Secret" be released within thirty years, by the time George W. Bush left office some four hundred million pages of classified material had been sitting in filing cabinets and on computer hard drives for longer than that. This was evident from the National Archives' own annual reports.

To many people who study the declassification process, this was a startling abrogation of the government's responsibility to act as its own archivist. The only

classified documents that were supposed to be kept hidden for more than three decades were those whose disclosure would pose a serious risk to national security, such as by revealing details of an ongoing military or intelligence operation. “Very few of those four hundred million pages could possibly have met the standard for remaining secret that long,” says Steven Aftergood, a transparency advocate who directs the Federation of American Scientists’ Project on Government Secrecy. “This was very troubling. The government’s prerogative to classify sensitive materials is supposed to be a temporary refuge from public oversight, not a permanent shield.”

Connelly, when confronted with the gaps he saw in the National Archives, did what he says most scholars do: he muddled through. After reading the documents that were available to him, he cobbled together the best history he could, soon publishing a paper about the power struggles among the CIA, the FBI, and the State Department over whose organization got to issue the authoritative interpretations of the military forecasts made early in the Cold War.

But afterward, Connelly couldn’t put the experience out of his mind. He wanted to know how long it would take the government to release those records. He also wondered: what other stories were hiding in those millions of backlogged documents? Other historians were asking similar questions, but Connelly grew angrier than most. The way he saw it, the government was not just standing in the way of new books being written; it was delaying a revolution in historical scholarship. Connelly was among a small but growing number of historians who believed that the future of his field was in using computers to analyze huge volumes of documents. For years, he had been going into archives with a digital camera and taking photographs of paper records. He would then turn those images into text files and feed them into software that in the aggregate could show him, for instance, where the paths of certain people, institutions, and companies had overlapped at different points in history. He was excited about the prospect of using similar techniques to analyze US government records from the digital era. A lot of sensitive electronic records should have already been declassified, since some federal agencies had embraced digital communications and record-keeping as early as the 1970s.

“There was all sorts of stuff that *should* have been released,” says Connelly, a slight man of forty-five with a boyish smile. “But the vast majority of it was still stuck in the pipeline somewhere. So on the one hand we have this amazing potential to study the inner workings of our government with a level of detail that is astonishing.

Yet we’re still waiting for the floodgates to open.”

In early 2012, Connelly put aside his research on the Cold War and began studying US secrecy policy. He learned everything he could about how federal records are created, maintained, and released to the public. He learned that since the 1970s, the government’s budget for reviewing and declassifying sensitive documents had failed to keep pace with the production of new ones. The backlog of secrets had grown significantly following the September 11, 2001, attacks, when federal employees were instructed to be more cautious in deciding whether to release old documents. After Barack Obama ’83CC became president, the glut shrank a bit, as government censors were told to relax their standards. By the end of Obama’s first term, though, progress plateaued and the size of the backlog stabilized at about 360 million pages.

Davide Bonazzi

Then Connelly had an idea: could he use data mining to infer what types of information were being left out of the public record? In theory, this seemed plausible, if he could compile enough materials to work with. He figured he could start by asking Columbia Libraries to give him special access to several commercial databases that the University licenses from academic publishers and which contain federal records. He could then download a wealth of material from government websites. Maybe he could even gather up documents that fellow scholars, journalists, and citizens had acquired directly from the government under the Freedom of Information Act (FOIA). No one had ever tried to analyze the entire

corpus of government records as one big database before. The promise of data mining now made it seem like a worthwhile endeavor to Connelly. He thought that if he were to recruit an interdisciplinary team of data analysts and fellow historians, he might create the first system for highlighting gaps in the National Archives. Perhaps this would even shame the government into releasing more classified materials.

“I thought if this were possible, it would be the most important thing I could do,” he says. “I’d go back to writing books later.”

Connelly would soon cast a new light on why the US government was slow in releasing its secrets. In doing so, he would thrust himself into a debate that had previously been taking place behind closed doors — a debate about whether the free flow of information and national security are on a collision course.

Toeholds and teamwork

In a small apartment in Harlem, a young mathematician named Daniel Krasner ’10GSAS sits at his kitchen table, staring into the soft blue light of his laptop. On the screen is a line graph depicting the number of teleconferences that Henry Kissinger participated in each day while serving as Richard Nixon’s secretary of state. “You see this spike here in late 1973?” says Krasner, pointing to a brief period when Kissinger was holding fifty to sixty teleconferences a day. “That has a pretty obvious explanation — it’s during the Yom Kippur war. But what about these spikes, here in 1975, or these in 1976? They could be worth looking into.”

Krasner, who earned a PhD in mathematics at Columbia, is among a half dozen computer scientists, mathematicians, and statisticians now working with Connelly on a multimedia research project they call the Declassification Engine. For the past year, this team has been gathering up large numbers of federal documents and creating analytic tools to detect anomalies in the collections. Several of the tools are on the project’s website and available for anyone to use. The one Krasner is developing is intended to find evidentiary traces of important historical episodes — a diplomatic crisis, say, or preparations for a military strike — that scholars until now have failed to notice. The Columbia researchers suspect that by spotting something as subtle as an uptick in a diplomat’s telephone activity they may be able to reveal the existence of historical episodes that the US government has largely suppressed

from the public record.

“If you can make out something happening in the shadows, then we can ask: does it seem curious that little information about this event is available in the public record?” says David Allen, a PhD candidate in history at Columbia who is working on the project.

Some of the material that Krasner is analyzing comes from a collection of 1.1 million telegrams, airgrams, telephone transcripts, and other communication records of American diplomats from the mid-1970s. The database, called the Central Foreign Policy Files, is available today on the National Archives’ website, where people can search its contents in rudimentary ways. Connelly, with the help of Columbia’s Digital Humanities Center, got his hands on the raw text files from the government. Now he and his colleagues are picking apart the documents using their own software.

“We can also analyze all of the language in these documents as what we call a ‘bag of words,’” says Krasner. “By seeing what terms tend to occur together in the same documents at certain times, we could spot interesting episodes.”

The Central Foreign Policy Files data set is an unusual collection in that it covers only material from 1973 — which is when the State Department implemented its first electronic records system — to 1976 — which is as far as the department’s employees have progressed in an ongoing effort to translate the files into a format that is Internet-friendly. But the collection has a couple of key advantages. The first is that it is comprehensive for its time period, containing all records of a particular type. Most collections of government documents are, by contrast, curated by archivists and editors to contain only materials thought to be of particular interest to scholars. The inclusiveness of the Central Foreign Policy Files would help the Columbia researchers spot conspicuous gaps.

The other reason Connelly sought out this collection was because of something he remembered seeing in the US State Department’s physical files at the National Archives. Often, when looking in a box of diplomatic records, he would find a single sheet of paper, slipped in between the others, that described the rough outlines of a document that appeared to be missing. This sheet usually contained only a date, a title, or subject, and sometimes the name of the sender and recipient. Connelly learned that this was the metadata of a classified document that had been rejected

for release — either upon turning thirty or when someone requested it through the FOIA.

“They’re not very interesting when viewed one at a time,” says Connelly. “You wouldn’t think much of them.”

But what if you had a quarter million of them? That’s how many were in the electronic version of the Central Foreign Policy Files. Every single diplomatic communication that had been transmitted between 1973 and 1976, marked as classified and later rejected for release, was represented by a metadata file.

It was when Connelly acquired this database, in the fall of 2012, that he began to recruit the help of professional number crunchers. First he called up Columbia statistics professor David Madigan, a versatile researcher who had previously developed algorithms that predict the side effects of medications. Then he brought in several members of Columbia’s computer-science department who specialize in finding patterns in large amounts of text. Within a few months, they would receive a $150,000 award from the Brown Institute for Media Innovation, a joint enterprise run by Columbia and Stanford that promotes interdisciplinary projects between journalists and data scientists.

“I’d worked with scientists before, but never like this,” says Connelly. “This would be as far as I’d ever strayed from the old model of history I grew up with, where Leopold von Ranke is standing alone atop a mountain, surveying the landscape of time with nothing but the facts in his head and a healthy dose of intuition.”

Ethnic profiling, ’70s style

Last spring, Connelly and his colleagues began inspecting those 250,000 metadata records to see what terms appeared on them most frequently.

“Basically, we were fishing around,” says Connelly. “We were modeling our technology.”

Once they did the analysis, one word stuck out: *boulder*. It appeared on thousands of cards.

Davide Bonazzi

Connelly soon concluded that this was a reference to “Operation Boulder,” a Nixon-era program that involved spying on Arab- Americans and scrutinizing visa applicants with Arab-sounding names. Initiated after the killing of eleven Israeli athletes by Palestinians at the 1972 Munich Olympics, Operation Boulder was roundly denounced by national-security experts for being ineffectual at improving the nation’s security. It was disbanded by the State Department in 1975. Few details about the program had emerged since. But what little information had been released provided Connelly and his colleagues the clues they needed to recognize the documents’ subject. The cards that contained the word *boulder*, when looked at in the aggregate, were also rich with references to visa applications, for example.

“There’s no doubt that these missing files are about the Nixon program,” says Connelly. “We can tell by looking at documents that have been released about the program. They also tend to mention visas.”

Why would the government release some documents about Operation Boulder and keep others secret? The Columbia researchers can shed light on this, too. Their analysis shows that before 2002, documents about Operation Boulder often got released when they came up for review. And then, abruptly, in April of that year, hardly any such files were declassified. Is it possible that the Bush administration blocked these releases to avoid comparisons between the antiterrorism measures

that it was pursuing at the time, such as its no-fly list, and Nixon's failed policy?

"It's not a smoking gun," Connelly says, "but it's suggestive, isn't it?"

David Pozen, a Columbia law professor who is an expert on government secrecy, says that this floating of trial balloons, this dropping of hints, is a valuable contribution to scholarship in itself. He says that the Declassification Engine, by revealing what types of information the US government is keeping secret, is likely to encourage scholars, journalists, and citizens to file more public-record requests. Furthermore, he says that the project's discoveries could help people win these petitions.

"One of the challenges in getting information through FOIA is that you need to describe what you're looking for in considerable detail," he says. "If you can show that an agency is sitting on thousands of documents related to a particular topic, well, the government may find it much less politically feasible to reject you."

All of it, not some of it

The Declassification Engine will soon provide its visitors access to more declassified US government documents than have ever been available in one place.

Many of the materials on its site have so far come from commercial vendors. These include a set of 117,000 records produced by various US departments and agencies from the 1940s to the present; this database, known as the Declassified Documents Reference System, is considered by scholars the most important of its type, based on the historical significance of its individual items. It is on loan to the Columbia researchers from the publishing company Gale.

In terms of sheer volume, though, the project's most impressive acquisition is yet to come. The Internet Archive, a nonprofit digital library based in San Francisco that collects all manner of public-domain content, from books to music to court transcripts, has agreed to give the Declassification Engine access to tens of millions of federal documents that its employees have trawled from government websites. These files will be accessible on the project's website later this year.

To keep the site growing, Connelly is also trying to create a sort of electronic catch basin for collecting documents as soon as the government releases them. One way he aims to do this is by collaborating with nonprofit organizations that have sprung up in recent years to help people file public-record requests. An organization called FOIA Machine, for instance, provides easy-to-use electronic submission forms and then tracks people's requests for them; when the government meets a request, the materials come to an e-mail account hosted by FOIA Machine. Connelly is now working with the organization to get access to those files. He says that migrating the documents to the Declassification Engine will allow researchers to study them alongside other declassified records using sophisticated analytic tools for the first time. Only a tiny percentage of documents that are released under FOIA, he points out, ever wind up in databases on library or government websites.

"Often, the person who receives material from the government is the only one in the world who now has a digital version of those records," he says. "That's a waste. Why not bring them all together?"

Old bars and stripes

One of the tools now operating on the Declassification Engine is ideally suited to gleaning insights from this influx of fresh material. Powered by software created by Columbia PhD candidate Alexander Rush, it can detect when multiple versions of the same document reside in the Declassification Engine's databases. Connelly says it is common for slightly different versions of the same record to be floating around, because the government will often release a document with lots of text blacked out and then put out a cleaner version, say, in response to a FOIA request, years later. He says researchers can gain insights into the political sensitivities of past US presidents by seeing what language was blacked out under their watch and subsequently restored by their successors.

"Sometimes it's the older, more heavily redacted version you're hunting for," Connelly says. "I've met historians who've spent years trying to track down all the versions that may exist of a particular memo."

Analyzing thousands of pairs of documents in this manner might also reveal political schisms within a sitting president's administration, say the Columbia researchers,

because sometimes one federal agency, in response to a FOIA request, will release a more complete version of a document than will another agency in response to similar requests.

“A classic example of this occurred in the aftermath of the Abu Ghraib scandal, when the FBI was eager to show that it had had nothing to do with torture and so it released a lot of information showing that other agencies were responsible for it,” says Connelly. “We hope that by analyzing huge numbers of documents, we’ll be able to identify the kinds of information that tend to get withheld by one or another agency, and thereby correct for the inherent bias in the public record.”

Truth and consequences

Is Matthew Connelly the next Julian Assange?

That’s a question he gets a lot. His answer is an emphatic “No.” He and his colleagues are only gathering documents that have been publicly released. And they are careful not to reveal any information that would endanger US security. They say their goal is merely to highlight broad categories of information that the federal government is keeping classified.

“Everybody involved in this project appreciates that some information needs to remain secret,” Connelly says. “On the other hand, lots of information is kept secret to avoid embarrassments, for political reasons, or simply because the government isn’t investing properly in reviewing and declassifying old documents. We want to help the government to uphold its own secrecy laws.”

That said, the data-mining technology that Connelly and his colleagues are developing could conceivably be adapted to generate statistically based guesses about what terms lie beneath redactions. And this is where things get tricky. Connelly described this possibility for a few journalists last spring. Their reports, appearing in *Wired*, the *New Yorker*, *Columbia Journalism Review*, and half a dozen other publications, posed riveting questions: Could a computer’s guess about the content of blacked-out passage be considered a leak? Would it matter if it guesses right or not?

Connelly and his colleagues have so far refrained from doing this kind of research while they evaluate its legal and ethical implications. They have formed a steering committee of historians, computer scientists, and national-security experts that will convene in January to help them decide whether to go ahead with it. If they did, Connelly says, they might rig the technology so that when it produces guesses about what lies beneath a redaction, it would exclude names of people and other highly sensitive types of information.

“The last thing we want to do is out the name of a CIA agent,” Connelly says. “Our main goal, even with this kind of research, would be to discover what *types* of information are getting classified, and why.”

But who is to say what information is safe to disclose? And might historians, by taking it upon themselves to decide this, inadvertently provoke the US government into releasing even less information so that they have fewer clues to work with?

It is conceivable that the US government will tighten its grip on classified information in response to Connelly’s work, according to several Columbia professors. They worry that the Declassification Engine, by demonstrating a capacity for redaction cracking that US intelligence experts have long feared that foreign spies would develop, might strengthen the hand of federal officials who are inclined to keep the lid on information.

“Those who advance a conservative approach to declassification could say, ‘Look, now there’s this small band of academics who are able to break down our redactions; can you imagine what others are capable of?’” says law professor David Pozen. “My concern would be that government officials might now say, “OK, instead of releasing these documents with redactions, we just won’t release them at all.”

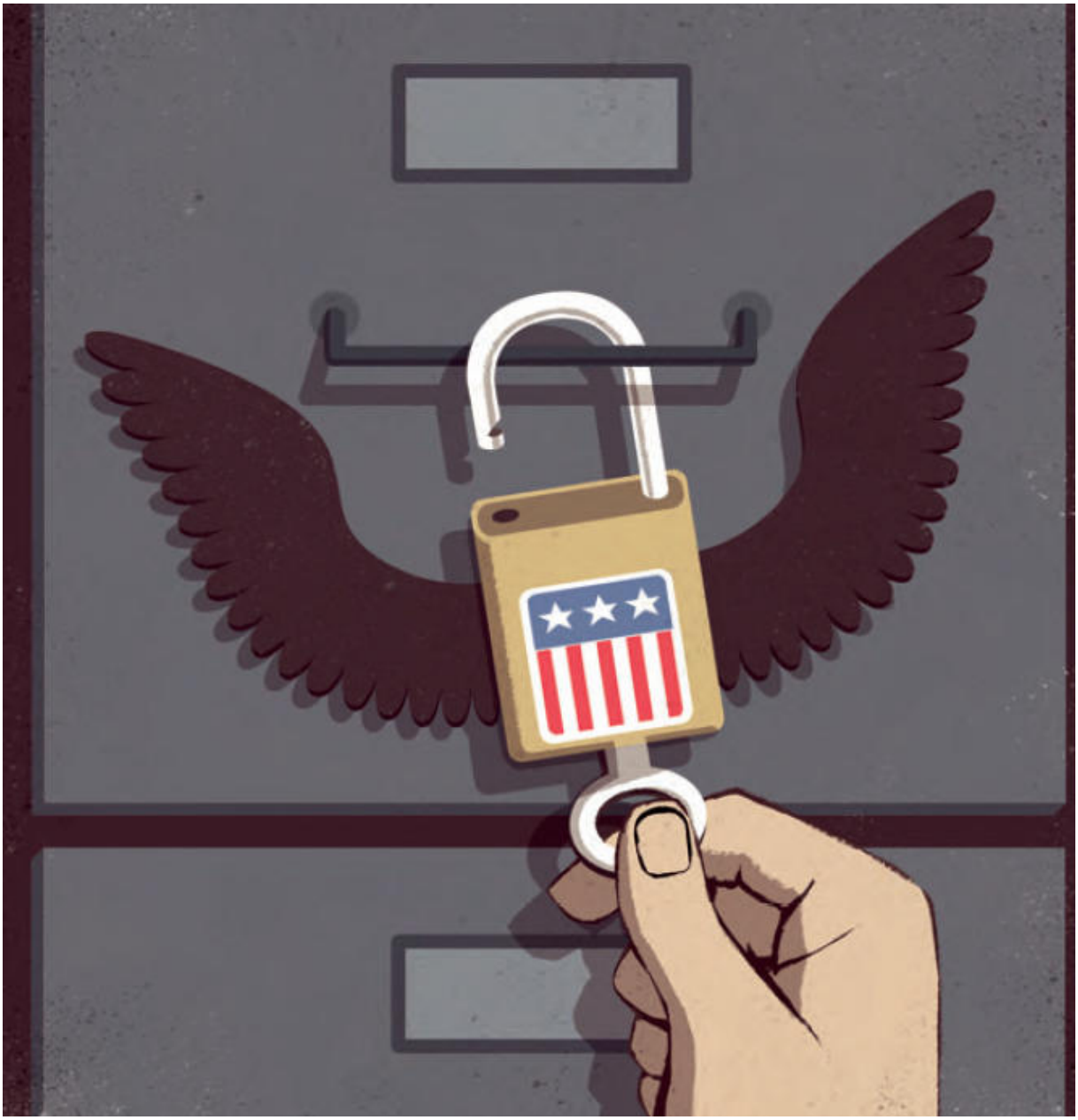
Yet these same Columbia experts say that the US government has for years been quietly taking steps to limit the information that it releases, specifically to frustrate any attempts to examine its records with data-mining techniques. One of the best things that could result from the Declassification Engine, they say, is that it will provoke debate about when it is justifiable to limit access to federal records as a way of offsetting this perceived risk. That this public conversation will take place soon seems inevitable. The analytic tools that Connelly and his colleagues are developing embody some of government censors’ worst fears of data mining — fears that, according to these Columbia experts, likely contributed to the enormous

backlog of declassified documents that inspired Connelly’s work in the first place.

Removing clues

Pozen’s own research has shown that the US government began fighting FOIA requests in court more aggressively in the early 2000s to avert the threat of computer-savvy spy craft. He has found that when FOIA cases go to court, Justice Department lawyers have often argued that documents that look innocuous in isolation ought to remain classified, because if they were to be analyzed in conjunction with a lot of other documents, vital secrets could be revealed. A hypothetical example goes like this: a document that references a café is released, and then is analyzed against another one that references a waiter, another a street, another a city, another an unnamed CIA informant, until, finally, a computer generates a list of people who could be that informant.

According to Pozen, this sort of hypothetical is plausible but is often treated by courts as a pretext for deferring to the government. “I don’t think judges carefully weigh the validity of this argument in each case, and they often don’t understand the technology that’s involved,” he says. “On top of this, they’re generally inclined to err on the side of caution whenever national-security concerns get raised. The result is they’ve tended to side with the government whenever they hear this argument.”



Davide Bonazzi

Pozen has argued in several papers that judges ought to take more time to consider these cases and push the government harder to justify why FOIA requests ought to be rejected on these grounds. But he says there has been little discussion of the issue among legal scholars or the judiciary so far. “It remains a pretty esoteric topic,” he says. “Anything that drums up some discussion about it will be a benefit to the legal community.”

Robert Jervis, a Columbia political-science professor who for the past ten years has chaired the Central Intelligence Agency’s Historical Review Panel, a role in which he advises the agency on which of its classified materials ought to be prioritized for

review and potential release, adds another twist to the story: he says that CIA officials worry that the Declassification Engine, by making available on its website huge numbers of federal documents that are drawn from disparate sources, could enable foreign spies or terrorist groups to conduct more powerful data-mining analyses of the nation's public record than they could otherwise. Jervis says it is partly to prevent enemies of the United States from data-mining old intelligence reports that the CIA's main digital repository for declassified documents, CREST, is not accessible on the Internet but only on computer terminals located at the National Archives in College Park — an inconvenience that has long irritated scholars.

The specter of data mining, Jervis says, could also cause some CIA officials to work more slowly while reviewing documents.

"These guys would love to have the budget that's necessary for reviewing all the documents that are before them carefully and getting them all out on time," Jervis says. "But they're not going to do anything that endangers an agent or his informants. So they're looking at this technology that's out there now, and they may say to themselves, 'We're going to have to work more scrupulously than ever.'"

Costs of complacency

On a recent Friday afternoon, Connelly sat behind his desk in Fayerweather Hall, quietly observing a group of graduate students who had gathered to work in a lounge outside of his office. Some were historians, others computer scientists. It was impossible to tell who was who, based on their conversations, which flowed with references to Nixon, Kissinger, Saigon, mean probabilities, gap-time distributions, and applets.

"Twenty years from now, when historians are writing the story of our time, their archive is going to include Google and Facebook," Connelly remarked. "They're going to need to understand data-mining techniques to do that work. I'm trying to develop those tools."

It had been a busy day. Connelly was preparing for talks with representatives of several federal agencies, including the State Department and the National Security Agency. He planned to address any concerns they had about his research. He would

also offer to demonstrate his team's analytic techniques in case the government had any interest in using them. Connelly had come away from previous conversations with federal officials convinced that the same tools his team is using to analyze the public record could help the government better manage its secrets. The government, too, is sifting through enormous numbers of documents and trying to make categorical assessments about their contents. In the government's case, this means determining which of the millions of classified documents that come up for review every year ought to be released, with or without redactions, and which ought to remain locked up in drawers. Federal employees do this work by reading documents one at a time, page by page, using black felt pens to ink over sensitive passages. Connelly said that many officials he has spoken to believe this needs to change soon; in order to process the tidal wave of electronic records that are coming due for review in the next few years, the government will need to implement its own data-mining system. One strategy that Connelly and many others have advocated to the government, he says, would involve screening large numbers of documents for language that is associated with sensitive topics. Human censors could then inspect these documents carefully, while funneling the others straight into the public domain.

"This would be a risk-management approach, and it would start from the position that it's impossible to catch everything, and that it's a mistake to try," Connelly says. "Time and time again government boards have proposed using technology in this way to make the declassification process more efficient."

That the US government would even consider releasing large numbers of sensitive documents, sight unseen, may sound surprising. Yet the current system may already be collapsing under its own weight. Connelly, echoing an argument that many experts on US secrecy have made, says that the rash of illegal leaks that the US government has experienced in recent years is partly a manifestation of a cynicism that has taken root about the government's perceived lack of transparency. When the government classifies too much information for too long, he says, the irony is that none of it is safe.

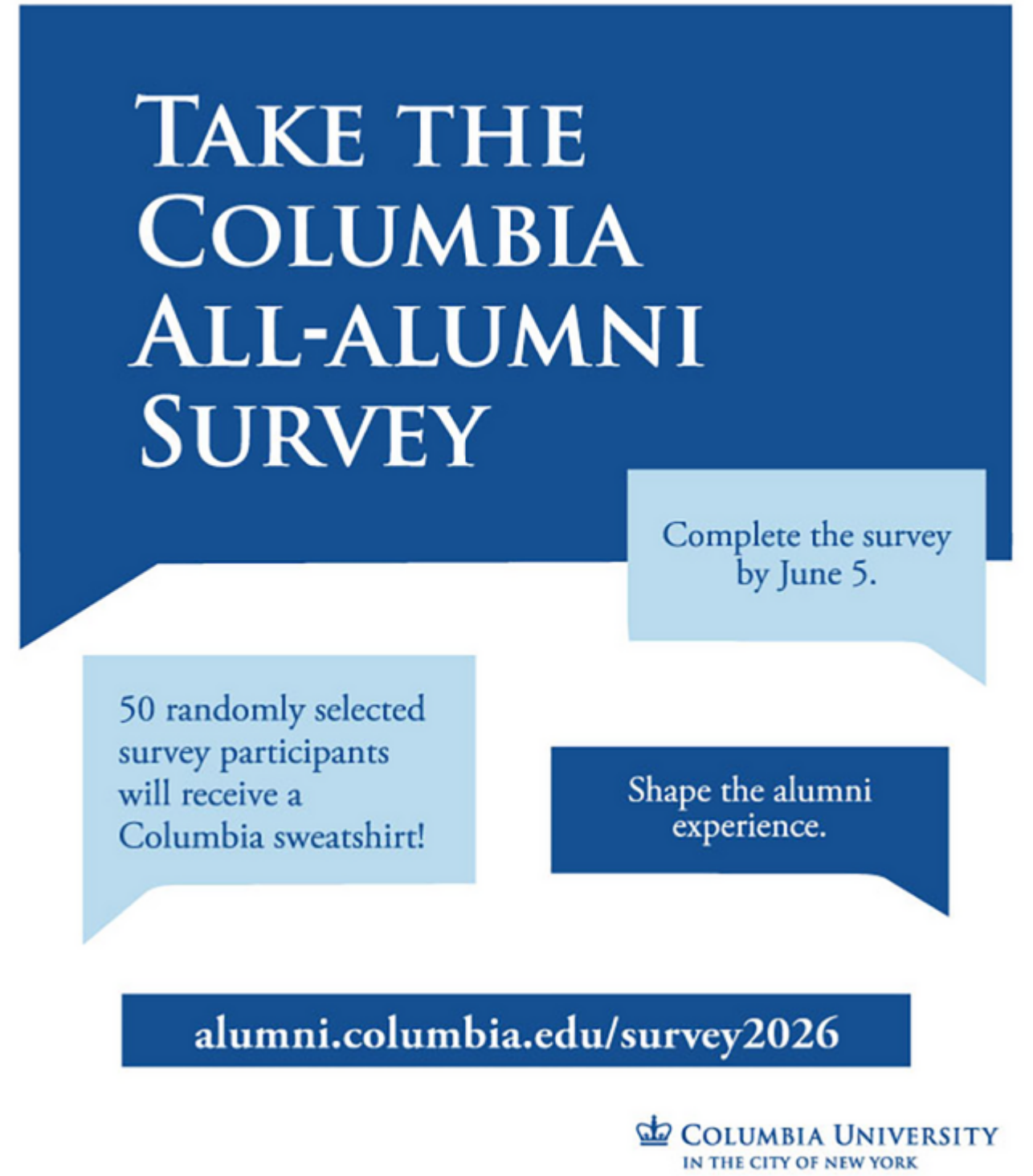
"What we need is a system that protects those secrets that are truly sensitive and releases the rest," he says. "Right now, neither of these goals is being accomplished. Technology has to be part of the solution."

Exactly how the Declassification Engine team could help the US government is unclear. Today, it is widely assumed by academics who study secrecy that the government must be pursuing its own data-mining research to speed the declassification process. It is also assumed that if this kind of research is taking place it is poorly funded, as most work related to declassification is perceived to be. It is hard to know for sure, though.

Why is that? Connelly pauses, and one can almost hear a drum roll. “The research is all classified.”

Read more from ***David J. Craig***

Guide to school abbreviations



All categories >

*Read more from **David J. Craig***