Can these government efforts crack the code for DNA storage?

The intelligence community is funding research into storing our ever-expanding troves of data in DNA. Their efforts could change how everything from our photos to Google searches get saved.

BY SARAH SCOLES | PUBLISHED SEP 29, 2022 9:00 AM

THE TOWN SLOGAN of Bluffdale, Utah, is "Life Connected." That's both innocuous and general—*so* innocuous and general that the two words are also, for instance, the slogan of a Colorado therapist and the title of a BBC tech column. However, in Bluffdale, under the shadow of the Wasatch mountain range, the words take on a slightly different tinge. Because this place is home to a facility code-named "Bumblehive."

The fact that a facility has an alias at all certainly gives away something about its nature. The Bumblehive is formally known as the Utah Data Center. It belongs to the Office of the Director of National Intelligence—the central organization of the official US spy set—and stores data for the infamous National Security Agency. "If you have nothing to hide," a sign in front of the Bumblehive ominously reads, "you have nothing to fear."

Those outside its fences don't know precisely how much data lives inside the 1 million-square-foot facility, but it's estimated to be in the exabytes. One exabyte, for context, is equivalent to about 330 billion Taylor Swift songs.

Obviously and controversially, the spy set isn't the only crew with a pretty significant interest in holding on to large chunks of information. Just think of all the info Google, Microsoft, Amazon, and Facebook (and their friends) have to store to make the internet run and to know which weird T-shirts to advertise to you. Facebook, for instance, is expanding a Texas facility that will take up 2.6 million square feet and 150 acres when it's finished this year, at a cost of around $1.5 billion.

Nobody *really* wants all that space wasted on humming servers and cooling systems and boring buildings that suck up lots of power and money. That's especially true because such storage methods could eventually become obsolete.

What if—instead of having to build a hive of any sort—all that information could fit in your hand, in a form that wouldn't degrade, go out of style, or break the bank?

The intelligence community would like to figure out how to make that almost laughable vision a reality. And they'd like to share their solutions with the private sector's data hogs. To make that happen, for spies and corporations, the Intelligence Advanced Research Projects Activity (IARPA [the IC's DARPA, if you want more acronyms]) is currently in the midst of a four-year project called Molecular Information Storage, or MIST. Contracts awarded to two teams in 2019 total around $48 million.

MIST's goal is to harness a biological form of storage: DNA. Genetic sequences can encode an entire human in a package too small to see, which is a much better job than a chip or CD can do. In the same way that computers use 0s and 1s to represent pictures, images, and documents, the nucleotide bases that make up DNA—adenine (A), cytosine (C), guanine (G), and thymine (T)—can also stand in for that same information. Each A, C, G, or T simply takes on new, coded meaning.

DNA storage is also very efficient: All human knowledge (such as it is, and as if it could be measured) could be stored inside a small room of DNA, whereas housing that information on magnetic tape would take millions of acres. Plus, as *Jurassic Park* will attest: DNA lasts a lot longer than magnetic tape or CDs

(RIP) or flash drives. And all of that is what IARPA is interested in.

R&D agencies like IARPA take on high-risk-high-reward challenges whose outcomes may be too iffy for other organizations. They call problems like fitting the entire Utah Data Center on a tabletop, appropriately and self-congratulatorily, "IARPA-hard."

Nevertheless, you don't breeze past tabletops of double-helixes containing the whole of Wikipedia (do you?). That's because it's still not practical *enough*, or cheap *enough*—two things MIST aims to alter.

If MIST succeeds, some of today's big-data warehouses could one day be just a bunch of double helixes. The program will ideally produce a prototype system that can encode 1 terabyte of data into DNA and extract 10 terabytes back from DNA in 24 hours, for less than $1,000, using less than 1 kilowatt of power.

IARPA's program is meant to put a shot in the arm of academic research and commercial industry—boosting them toward a goal that suits the intelligence community's interests, while appealing to their intellectual curiosity and their future checkbooks. If MIST succeeds, it can demonstrate to the private sector both interest from a potential large customer (spies) and a success that other companies can build on.

Two teams—one led by the Broad Institute and the other by the Georgia Tech Research Institute—have won contracts to try to make such DNA data storage more than a neat trick, so that everyone can live life connected.

THE AMOUNT of data is increasing faster than people can create cost-effective storage, meaning some info ends up in the trash. "Anybody with a massive data storage burden has this problem," says David Markowitz, the MIST program manager. Though nobody knows for certain, he estimates that the globe produces around 30 zettabytes of data a year. "There are 1,000 exabytes in a zettabyte," he says, "So that means we're only producing enough new tape to archive 0.3 percent of data produced annually, and more than 99 percent of new data couldn't be retained even if we wanted to."

Intelligence agencies naturally have an outsize interest in catching, and keeping, it all. "You don't always know in advance what data is going to be most useful for solving a mystery," Markowitz says. "Who is responsible for some of the events that happened in the future?" Untangling that time warp requires finding needles in haystacks and often, Markowitz says, "digging through a lot of historical data."

The idea of using DNA to hang on to it dates back decades as a what-if. In 1988, for one, an artist named Joe Davis (with helpers at Harvard) created a piece called "Microvenus." He embedded a 35-bit image into the nucleotides of *E. coli*, showing an old Germanic character meaning "female Earth."

Davis is now an affiliate at the Harvard lab of scientist George Church (the gene-editing pioneer who's currently trying to resurrect the woolly mammoth). In 2011, Church enmeshed 700 kilobytes' worth of a book (humbly, one he had cowritten) into DNA, and he worked on a similar project the next year, adding images and JavaScript code. In 2013, European Bioinformatics Institute researchers demonstrated that they could encode more than 625 kilobytes in DNA, with few errors.

One of the two teams working on that problem as part of MIST is led by Georgia Tech, and also includes Twist Bioscience, Roswell Biotechnologies, and the University of Washington in collaboration with Microsoft. The group calls its solution SMASH: Scalable Molecular Archival Software and Hardware.

Once SMASH is summoned into existence, it will work like this: Software will translate information into genetic sequences, spitting out strings of As, Gs, Cs, and Ts that represent the data. Then, a computer sends those strings of letters to a semiconductor chip—essentially instructions for which bases of DNA to build in which order.

That chip is filled with tiny wells, just a few hundred nanometers deep. Each well is a diminutive DNA synthesizer, able to grow genetic sequences, base by base, according to the instructions. The wells each build up their sequences in parallel, like chickens laying eggs next to each other. Once a given set of DNA strands is finished, it gets washed off into a droplet—like an information-dense Hershey's kiss. It can be put away wet, or dried out for longer-term keeping.

To extract information *from* that DNA, a sequencing chip then measures the electrical fingerprints of individual GATTACA molecules. Then scientists just have to reverse-decode the DNA, put the strings in the right order, and correct for errors.

Voilà! In the future, this could be how archival farmers' almanacs get shelved and then picked up again.

The other MIST team is led by the Broad Institute, in collaboration with Harvard University and the company DNA Script. "What we hope to do is have the systems that would be high-throughput enough that you could start to deploy them and archive data that might not be data you need to access every day but you definitely want to keep," says project lead Robert Nicol. He cites, as an innocuous example, sports. "Every baseball game is very high-throughput," he says. "There are very high-definition cameras all over the stadium." Catching every player, every spectator. Maybe, 20 years from now, people will want video of the audience's reactions to a retiring superstar's super hit.

It's unlikely, of course, that a grand slam is IARPA's jam.

NEITHER PROJECT is in full working order yet, but by the end of MIST—a couple years from now—officials hope they will be. To figure out how well it works, IARPA employs testing and evaluation partners: outside organizations with related expertise that create a kind of rubric for judging the new technology.

For MIST, one of those partners is Los Alamos National Laboratory. There, in a team led by Tracy Erkkila, scientists write a test for the teams and then create the answer key. At the lab, they themselves encode files into an electronic DNA archive. The teams must then write that archive back into DNA. "They'll essentially provide us with a pile of liquid DNA," Erkkila says. The evaluators will then read their liquid DNA piles by sequencing them. They then score the results, looking out for translation errors. The lab also tries to decode that DNA back into the original encoded information.

While Erkkila can't give too many details about the test (that would be like a senior passing an old copy of a 10th-grade quiz to a new sophomore), he will say that it includes video, audio, pictures from the Hubble Space Telescope (because, he says, "we're in love with some of those images"), and a 3D model of a rabbit figurine called the Stanford bunny.

Animals, as it happens, are a reason Erkkila cites for pursuing the possibilities of DNA data storage. Imagine you're a wildlife researcher, he says, and you want to plant a camera in remote Alaska. "I want to record for two years in a row," he says. "How are you going to store that information?" As to why spies might want the ability to hang onto massive data troves, he doesn't say.

Putting a DNA synthesizer on a glacier, though, would be a good punchline.

DNA data storage isn't a joke, even if it used to be. When Markowitz first started exploring whether DNA might be a good fit for the intelligence community's data in 2016, it wasn't a popular idea. "The few people who were working in the DNA space, they would get up at a conference and talk about it, and they would get laughed off the podium," he says. "Really. By people from the conventional storage industry."

Today, by contrast, the DNA Data Storage Alliance—an industry and academic collaborative group—has dozens of members, including IBM, Dell, and Microsoft. "Nobody," says Markowitz, "is laughing anymore."

If they were, Markowtiz could perhaps look back on it in 50 years, having kept a record of that conference session, and every other one, in As, Gs, Cs, and Ts, and smirk himself.

We hope you enjoyed "Overmatched," a new column exploring how government-funded research could transform everyday life. Check back on PopSci+ for future installments.

Sarah Scoles

Sarah Scoles is a freelance science journalist and regular Popular Science contributor, who's been writing for the publication since 2014. She covers the ways that science and technology interact with societal, corporate, and national security interests.

DARPA | ENGINEERING | MILITARY | MEMBERSHIP