

Deep Fried Data

Maciej Ceglowski

I run a small web archive for about twenty thousand people. Being invited to speak at the Library of Congress is like being a kid who glues paper fins to a cardboard tube and then gets asked to talk to NASA about rocket propulsion. As every speaker has correctly said, it is a signal honor to be here.

It also feels strange to be speaking in D.C., at the seat of government. In most of the talks I give, the U.S. government is an adversary.

But today I am at a government institution that champions not just freedom, but the fundamental right to privacy, and the dignity that that entails. During the panic that followed September 11, Carla Hayden, then head of the American Library Association, took a principled stand against [provisions in the Patriot Act](#) that required librarians to reveal what their patrons were reading. She did it in the face of ridicule from Attorney General Ashcroft and the administration. And of course just a few days ago, she became our new Librarian of Congress.

[AUDIENCE GIVES IT UP FOR THE INCOMPARABLE [CARLA HAYDEN](#)]

It saddens me that those provisions in the Patriot Act, which seemed so threatening and un-American at the time, look almost quaint today. And this time it's not the government, but the commercial Internet that has worked so hard to dismantle privacy.

Librarians are bound by their code of ethics to protect patron privacy. But this protection means little when Amazon knows every ebook you've read down to the page, Google has your complete correspondence and web history, and your phone company tracks your movements with a device you willingly carry in your pocket.

This information, once collected, becomes part of a permanent, indelible record maintained without accountability or restriction.

The commercial Internet is an amazing achievement. But its values are the opposite of a library. Where libraries exist to inform, those of us who run online businesses do our best to extract information from you. Where libraries try to be impartial, we practice constant manipulation. Every link has an ulterior motive. Click this, read that, view this ad, punch this monkey, and above all, share everything with us, no matter how private, forever.

However, these un-librarians have made genuine technical breakthroughs in dealing with large data collections. For people just bringing their collections online, this raises the question of how to engage with a world whose values are repugnant, but whose expertise is alluring.

MACHINE LEARNING

Today I'm here to talk to you about machine learning. I'd rather you hear about it from me than from your friends at school, or out on the street.

Machine learning is like a deep-fat fryer. If you've never deep-fried something before, you think to yourself: "This is amazing! I bet this would work on anything!"

And it kind of does.

When I was in college, friends who worked the snack bar conducted extensive research along these lines. They would deep-fry cheese, candy, pens, their name tag. And all of it came out tasting great.

In our case, the deep-fryer is a toolbox of statistical techniques. The names keep changing—it used to be unsupervised learning, now it's called big data or deep learning or AI. Next year it will be called something else. But the core ideas don't change. You train a computer on lots of data, and it learns to recognize structure.

These techniques are effective, but the fact that the same generic approach works across a wide range of domains should make you suspicious about how much insight it's adding.

And in any deep frying situation, a good question to ask is: what is this stuff being fried in?

In the seventies, we had a fellow who brought a high-pressure deep-fryer from Italy and set up a chicken shack in a Polish resort town. He called this undertaking Frico Polo. Since it was under communism, you had to bring your own chicken.

Frico Polo could pressure-fry a chicken in three minutes. It was greasy and hot and the best thing you ever tasted. People stood in lines with their chickens for hours. Then one day the health department arrived and shut things down. It turned out the operator had never once changed the cooking

oil, which dripped from the machine like roofing tar. That's where all the unique flavor was coming from.

So what's your data being fried in? These algorithms train on large collections that you know nothing about. Sites like Google operate on a scale hundreds of times bigger than anything in the humanities. Any irregularities in that training data end up infused into in the classifier.

For this reason I've referred to machine learning as [money laundering for bias](#). It's easy to put anything you want in training data.

For example, if you go to Google translate and paste in an Arabic-language article about terrorism or the war in Syria, you get something that reads like it was written by a native speaker of English. If you type in a kid's letter from camp, or an extract from a novel, the English text reads like it was written by the Frankenstein monster.

This isn't because Google's algorithm is a gung-ho war machine, but reflects the corpus of data it was trained on. I'm sure other languages would show their own irregularities.

Prejudice isn't always a problem. Some uses of machine learning are inherently benign. In an earlier talk, we heard about [identifying poetry in newspapers based on formatting](#), an excellent use of image recognition. OCR is another area where there are no concerns.

Others, though, would be problematic. I'd be very wary of using "sentiment analysis" or anything to do with social networks without careful experimental design.

I find it helpful to think of algorithms as a dim-witted but extremely industrious graduate student, whom you don't fully trust. You want a concordance made? An index? You want them to go through ten million photos and find every picture of a horse? Perfect.

You want them to draw conclusions on gender based on word use patterns? Or infer social relationships from census data? Now you need some adult supervision in the room.

Besides these issues of bias, there's also an opportunity cost in committing to computational tools. What irks me about the love affair with algorithms is that they remove a lot of the potential for surprise and serendipity that you get by working with people.

If you go searching for patterns in the data, you'll find patterns in the data. Whoop-de-doo. But anything fresh and distinctive in your digital collections will not make it through the deep frier.

We've seen entire fields disappear down the numerical rabbit hole before. Economics came first, sociology and political science are still trying to get out, bioinformatics is down there somewhere and hasn't been heard from in a while.

Before you spend a lot of time peeling and julienning your data, consider—is this really the best way to go?

Computers eliminate drudgery. But the excitement is the human potential. Today, for the first time, we can make things available to anyone on the planet who has an internet connection. I don't think we've internalized the enormity of that step.

Just throwing data online is not sufficient. Some years ago I worked as a program officer at the Mellon Foundation, and one of our big projects was JSTOR. I remember learning that half of the collection had never come up in a search result. Most of the collection had never been viewed, but half of it had never even shown up on a search page. That half may as well not have existed.

Part of the issue was extremely restrictive agreements with publishers. But part of it was a failure of imagination. We had digitized every journal article under the sun, but were making no attempt to connect that data to people outside "the academy", narrowly defined.

We missed so many opportunities! For example, we completely flubbed Wikipedia. Nobody could imagine a pseudonymous, collaborative effort like that succeeding. It wasn't scholarly! My boss eventually considered printing and publish a hard copy of it; that's as far as we got with Wikipedia at Mellon.

Later on I saw librarians fail to engage with vibrant communities at Flickr and Delicious, services they would later grow to love, because of their unstructured approach to tagging. There was a lack of trust and openness to an experiment that would have produced a remarkable collaboration.

DATA GARDENING

A lot of the language around data is extractive. We talk about data processing, data mining, or crunching data. It's kind of a rocky ore that we smash with heavy machinery to get the good stuff out.

In cultivating communities, I prefer gardening metaphors. You need the right conditions, a propitious climate, fertile soil, and a sprinkling of bullshit. But you also need patience, weeding, and tending. And while you're free to plant seeds, what you wind up with might not be what you expected.

If we take seriously the idea that digitizing collections makes them far more accessible, then we have to accept that the kinds of people and activities those collections will attract may seem odd to us. We have to give up some control.

This should make perfect sense. Human cultures are diverse. It's normal that there should be different kinds of food, music, dance, and we enjoy these differences. Unless you're a Silicon Valley mathlete, you delight in the fact that there are hundreds of different kinds of cuisine, rather than a single beige beverage that gives you all your nutrition.

But online, our horizons narrow. We expect domain experts and programmers to be able to meet everyone's needs, sight unseen. We think it's normal to build a social network for seven billion people.

I think of this as the [Mormon bartender problem](#). To understand what people need takes at least a little visceral experience.

Let me give you an example from my own work of what it's like to surrender control. I run a very vanilla bookmarking site, where you can save URLs for later. It's a personal search engine for scholars, journalists. I even have a priest who uses it to prepare his weekly sermons.

But one of the biggest groups of users is writers of fan fiction. Half of you are librarians, but we can pretend that I need to explain to you what fanfic is. This is a vibrant subculture of people who write stories, often highly erotic, set in various fictional universes. If you always thought there were sparks between Holmes and Watson, have I got a hobby for you.

Fanfic authors adapted the tagging system on my site so that they could use it as a search engine and publication tool. They do a lot of additional work to make it suit their needs.

It was like watching bees arrive and set up their hive. All I could do was observe in wonder, and try not to get stung. In return, I got an industrious

and extremely positive group of users, and learned a lot about online privacy.

The Internet needs to get much weirder. People out there are tired of deep-fried data, too, and want substance. They'll do interesting things. But you have to trust them.

In an institutional setting, this can be frightening. It takes courage to ask for a grant to bring a collection online with no measurable outcome other than the hope that it attracts interesting use. It takes even more courage to award that grant. It takes courage for a young faculty member to devote time and energy into projects that someone might use to make cat videos. At most institutions, that is not the royal road to tenure.

And it takes courage to commit to maintaining these collections, and staying engaged with the people who use them, for years to come.

But the search for intelligent life on the Internet means putting away some preconceptions about who our communities of use are.

I thought the fanfic authors on my site were just pursuing a harmless, quirky hobby.

What I didn't realize is that online, the frivolous blends with the serious. Fanfic authors tend to be women. Britta Gustafson has called fandom a secret seminar on feminism. Young fans use stories to explore the issue of gender identity (in some cases, they're finding out for the first time that there is such a thing a gender identity). They learn to deconstruct plot elements in a way that would make Russian structuralists blush. They coach each other to improve writing. And they also coach each other in technology.

It's important to realize that serious people can have frivolous hobbies. And in settings where young people share a hobby with high-caliber scholars—maybe some of the very people in this room—ideas can percolate down.

My friend Sacha Judd, in [an upcoming talk](#), describes something similar with fans of the boy band One Direction. This band has an obsessive following of young women, and in chronicling the lives of their beloved band members, the reach heights of technical achievement that rival anyone working in professional media. They are de facto professional archivists, developers, video editors, and journalists. But since “real” technologists don't take their interest seriously, these women don't recognize their own achievements. They would never dream of applying to the kind of jobs that they already excel at in their role as fans.

There is this dark matter of talented, motivated, interested people online. I'm convinced our time is much better spent trying to reach them and engage with them, than to use the same tools.

SOCIAL MEDIA DATA

So far I've talked about bringing collections online, but there's also the question of how to deal with social data that's born on the Internet. There is an awful lot of it, and it's fascinating.

One approach is to go to the people who control the data—the big companies—and partner with them to study it.

It's awkward because the very thing the Librarian of Congress objected to in the Patriot Act—the intrusive surveillance—is the bread and butter of online services. Much of the valuable information is collected in ways that would never pass ethical standards in academia, and ways that even the NSA would be legally prohibited from collecting.

But the data is there, and you can hear it calling to you.

"Study me" it coos. "Preserve me," it pleads. since fly-by-night companies obsessed with growth assuredly won't.

"Analyze me."

You can try to cover your qualms with layers of ethical review, like a coat of paint. But you can't hide the ugliness underneath.

I worry about legitimizing a culture of universal surveillance. I am very uneasy to see social scientists working with Facebook, for example.

People are pragmatic. In the absence of meaningful protection, their approach to privacy becomes "click OK and pray". Every once in a while a spectacular hack shakes us up. But we have yet to see a coordinated, tragic abuse of personal information. That doesn't mean it won't happen. Remember that we live in a time when a spiritual successor to fascism is on the ascendant in a number of Western democracies. The stakes are high.

Large, unregulated collections of behavioral data are a public hazard.

People face social pressure to abandon their privacy. Being on LinkedIn has become an expected part of getting a job or an apartment. The border patrol wants to look at your social media.

So in working with the online behemoths, realize that the behavioral data they collect is not consensual. There can be no consent to mass surveillance. These business models, and the social norm of collecting everything, are still fragile. By lending your prestige to them, you legitimize them and make them more likely to endure.

WEB ARCHIVING

Since it's bad ethics to work with the people who hoard it, you can always try to collect data in the wild.

A friend who works for a movie effects studios explained to me once how digital effects in CGI movies are preserved. The short answer is, they're not. Modern effects are rendered with Rube Goldberg like toolchains that become obsolete with each film. Studios upgrade the hardware, rewrite their software, and that's it. The bigger places have an in-house archivist to keep the artwork and digital assets around, but there's no way they could make them into a movie again.

Something similar threatens to happen on the web. Forgive me for being technical, but the average web page right now is a giant pile of steaming garbage. Only superhuman effort by browser developers makes it possible for anything work at all. Pages get stitched together at load time through dozens of intermediaries, with live dependencies, and the bulk of rendering done in JavaScript.

So what does it mean to archive that? You can save the rendered image in the browser, but what about the dynamic behaviors? Does autocomplete fall within the purview of archiving? Is a dynamic ad an annoyance, or a valuable insight into 2016 life that we should save for posterity? And if so, which ad do we save, and how, when it's pulled in at view time through a dozen different ad auctions and hypertargeted for the viewer?

Do we need to ultimately build an entire simulator for what it's like to use computers in 2016, and if so, will future generations forgive us?

Game developers have wrestled with this problem already, and have things to teach us. The early video games were hardware creatures, and even emulating them decades later is challenging. The only reason we can do it is

because there's a community of people who loves playing the old games enough to put in the work.

One problem that an institution like the Library of Congress faces is that it can't just roll up its sleeves willy-nilly and try some approach, because its practice will become normative. But we can't just ignore the web, either.

So another reason I pin my hopes on communities, rather than tools, is that they can help with the inevitable tasks of interpretation. A lot of this stuff can't be strictly preserved, it has to move across formats, and to prevent the essential being lost requires that the people immersed in the world participate in saving it.

LiveJournal, for example, was a blog site where users could have an avatar image, a thumbnail of which would appear next to each comment. You could choose from a set of avatar images when you posted your comment. At some point, the site changed the way these worked, so you were restricted to some small number of images.

What LiveJournal didn't realize was that people were using these images as visual commentary. The image was a gloss on the comment, and modified its meaning. People got very good at this form of subtext.

By capping the number of images, and making the change retroactive, LiveJournal destroyed a huge amount of information it didn't know existed.

This kind of mistake is obvious enough if you look at one service, but the challenge is that online communities can span multiple services, over many years.

We have to learn how to send out ambassadors to online communities, like they do with uncontacted civilizations in the Andaman islands. You go out, they throw a few spears at your helicopter, but eventually you get to be on speaking terms and can learn from one another.

The task is pressing, because we've lost so much from the web already. Not only does something like 5% of URLs disappear every year, but things go up in big conflagrations when a company goes under, or makes a terrible decision.

I've saluted the efforts of Archive Team and the Internet Archive, but their activity is like having a museum curator that rides around in a fire truck, looking for burning buildings to pull antiques from. It's heroic, it's admirable, but it's no way to run a culture.

WHAT TO DO

Focusing on communities means relinquishing control. Like I said, it's scary. But there are some other steps, too.

Most important is to make materials available in open formats, without restrictions, and with a serious commitment at permanence. These all require institutional courage, too. What if somebody grabs all this data, and does something with it that's not scholarly?

Well, that's what you want! A sign of life!

Publish your texts as text. Let the images be images. Put them behind URLs and then commit to keeping them there. A URL should be a promise.

It's not enough to get a caste of coders work with this data, and make their finished work available. By all means, do that, but don't pretend you are the sole arbiter of what your data is for. You aren't!

I will say a nice thing about programmers. You don't have to force us at gunpoint, or with huge financial bribes, to work with interesting data.

Many of us work jobs that are intellectually stimulating, but ultimately leave nothing behind. There is a large population of technical people who would enjoy contributing to something lasting. And we have a strong culture of collaboration that we can put to good use around projects.

This is an area where the Copyright Office can do a great deal of good, too, by aggressively advocating for fair use, and defending the framers' intent in creating copyright.

WISTFUL CONCLUSION

Like a lot of nerdy kids of my generation, I spent half adolescence at the library, or at home reading library books. Along with schoolbooks and crappy 90's TV, the Glenview Public Library was my window on the world.

I never reflected on why this unremarkable suburban library existed, who funded it, where its values had come from, or how long it would be around. It was as immutable a part of the world to me as Lake Michigan.

But it taught me that like everyone else, I had a right to learn and was welcome. That I could ask questions, and learn how to find my way to the answers. It taught me the importance of being quiet in public places.

For the generation growing up now, the Internet is their window on the world. They take it for granted. It's only us, who have seen it take shape, and are aware of all the ways it could have been different, who understand that it's fragile, contingent. The coming years will decide to what extent the Internet be a medium for consumption, to what extent it will lift people up, and to what extent it will become a tool of social control.

The way things are right now, the Internet is a shopping mall. There are two big anchor stores, Facebook and Google, at either end. There's an Apple store in the middle, along with a Sharper Image where they are trying to sell us the Internet of Things. A couple of punk kids hang out in the food court, but they don't really make trouble. This mall is well-policed and has security cameras everywhere. And you guys are the bookmobile in the parking lot, put there to try to make it classy.

My dream for the web is for it to feel like big city. A place where you rub elbows with people who are not like you. Somewhere a little bit scary, a little chaotic, full of everything you can imagine and a lot of things that you can't. A place where there's room for chain stores, room for entertainment conglomerates, but also room for people to be themselves, to create their own spaces, and to learn from one another.

And of course, room for big, beautiful, huge, tremendous libraries.