ISOJ 2018: Day 1, Afternoon Session

TRUST: Tools to Improve the Flow of Accurate Journalism

Chair: Jennifer Preston, Vice President of Journalism, **Knight Foundation**

- **Liza Fazio,** Assistant Professor, Vanderbilt University, **CrossCheck**
- **Frédéric Filloux,** Creator, **Deepnews.ai**
- **Darryl Holliday,** Editorial Director and Co-Founder, **City Bureau**
- **Joan Donovan,** Media Manipulation/Platform Accountability Research Lead, **Data & Society**
- **Cameron Hickey,** Producer, **PBS NewsHour**

Joan Donovan: So, as part of the Data & Society, I lead the research on media manipulation, also known as disinformation, and platform accountability. And so today, I want to talk to you about, how does the media get manipulated? But also, how are people using media to manipulate individuals and groups? And so, I'm going to introduce you to some of the most cutting edge research that we have right now, where we haven't even published on it yet. Because what I'm really interested in hearing from journalists [is], how they react to thinking about your source material getting hacked and the way that…. So, I'm going to talk today about breaking news. And during crises, we have very low information, so we're all scrambling to find different forms of evidence. We all go online, not just journalists, but the police go online looking for information, as well as everybody else. So, we have very high public interest, and then we have low information, and then we have a lot of mixed algorithmic signals.

And that's where the people that I study also are online, which we might call — I just lump them all into the category of the trolls. And what they're really trying to do is leverage the moments of disinfor- — of low information, high public interest, knowing that journalists and experts are looking for information, as well as the police, at those times. And so, after a breaking news event happens, usually within four-to-eight hours, if these trolls have been successful, someone—maybe even one of you in this room—has reported the wrong information. And that, and that…. They're doing this intentionally.

So, I just want to give you a little bit of a preview of some of the reports that we already have out at Data & Society. We have a report on navigating content moderation after fake news. So, we looked at over the last year, what are platform companies responses to [fake news], regulators responses to fake news, civil society organizations responses to the fake news crisis, and the crisis of legitimacy and accuracy of information online. Then we also have a report on how, if we were to try to fix this using a media literacy frame, what are the challenges that we face?

And one of the major challenges is, we have a lack of evaluation protocols for doing media literacy.

Third, we have a report on the lexicon of lies. If we're not going to call it fake news, what are we going to call it? Where does it come from? Why do propaganda studies and the institutions that were supposed to protect us from propaganda, why did they fail? So, there's a lot of historical information in there, as well as our landmark report by Becca Lewis and Alice Marwick on media manipulation and disinformation online. And that one is rather long, but definitely worth starting and starting to get through.

So, when I think about manipulation, I think about socio-technical systems. And I want to show you something that's a very plain image of something that some of us might have done in our younger years. But we take different pieces of technology and we plug them together in ways that the original designers never intended. So here, you have a record player that's playing out of a speaker. That's a normal use case. Then, you have someone using the telephone to broadcast to a friend, "Hey, listen to this really cool new song." And when we try to understand what platform companies are doing, what we notice the most is that they don't think about how their platforms connect into other platforms. And as well, journalists have to start thinking about, how does your content connect into other platforms? Right? You might think about it as a format issue. But it's also challenging to think about, well, who are the other people that might be connecting these things together?

And in terms of disinformation, each one of these components, we see disinformation happen in certain ways. So, something that might count as disinformation on one platform, may not count as disinformation on another platform. For instance, if you put up a blog post, and it's on a WordPress, and nobody is really monitoring it, is it disinformation? But when you then post that link to Twitter or Facebook knowing that it contains either libelous material…, then it becomes a problem for that other platform. Right? So, just hold this picture in your mind anytime you're trying to think about, what are we talking about when we talk about media manipulation?

So, one of the concepts that I've been working on is this idea of source hacking, which is a tactic where groups coordinate to feed false information to journalists and experts, usually during times of crisis. And I'm going to go through two quick examples of what I mean by this. Here's a lead. It's still online. This is a fake document that is on Twitter currently that is a forgery saying that Maxine Waters is calling on this bank to give her a $1-million donation for her campaign, in the promise that when she is reelected, she will bring 41,000 refugees to the area and the bank will be rewarded handsomely with mortgages. Maxine Waters has tried to get this taken down from the internet. She's even, like, added Twitter help and Twitter support and the FBI. There's been stories in Think Progress, as well as the LA Times, debunking this document, but it still lives on Twitter. It still lives on Reddit.

And so in our lab, we went back and we traced the origins of this document. No big surprise. It comes from the troll hole known as 4chan. It's also on 8chan. But when we see these kinds of events happening, we start to think about, in terms of case studies, where have we even seen this before? And this technique was also used during the Macron leaks, where trolls from 4chan also put out information suggesting that Macron had a hidden bank account in the Caribbean using a forged banking document.

So, they tend to repeat the same tactics over and over. And if anybody can get this document offline, please email me and tell me how you did it, so that I can have another, sort of, I can have another way of understanding how we can do this kind of content moderation when the disinformation stakes are so high.

The second tactic I want to talk about that's been in the news is this idea of keyword squatting. This is where they lay in wait and they hack social media in such a way that they just hang out in the keywords. So, the first instance here is a Blacktivist, which is a group that was now famously outed as being controlled by the IRA, the Internet Research Agency, the Russians. The Russians. [laughs/laughter] Like, I know, we're all skeptical, right? But in terms of this, they also had a page called Black Matters, right? So, if you typed in 'black lives matter,' 'black activism' into Facebook, these are the kinds of pages you'd be served. As well, CNN covered this story on the largest Black Lives Matter page belonging to an Australian.

And so in this instance, I've gone back and talked to other Black Lives Matter activists that run Black Lives Matter pages on Facebook, and this was a known known. For at least the past six months, activists have been flagging and trying to get Facebook to pay attention to this page and say, "This page is not real." Because there's no good way to verify who's running these pages, it's a difficult thing if you're covering a movement to be able to say if this is true and correct information, right? So now, you have to try to dig a little bit deeper and get at the sources. And of course, part of the keyword squatting here was motivated by money. They were selling mugs and t-shirts.

And so, I just wanted to bring your attention to a few of these different techniques around keyword squatting as well as disguising a forgery as a leak [and] show you that source material is much more difficult to vet in this age than it was prior, maybe even ten years ago or five years ago or even two years ago, because trolls are repeating the same tactics, using the same tactics, and they treat it as a game. And so when they win, if you cover the story and you cover the misinformation instead of covering or waiting or not publishing at all, then they take those stories and they compile them and they call them trophies, right?

So, what we're dealing with here is something akin to a media movement. And that's different from thinking about them as motivated by any other intent. If you think about them as an adversarial media movement that is really trying to destabilize the entire institution of journalism, then you are better equipped to address them and create a context in which you and your fellow journalist friends can lean on each other [and] create a network in order to vet this information as it's happening.

I'll tell you another one that is also something that was related to this keyword squatting. When we were going through after the election, we started noticing an uptick in Antifa Twitter accounts. And [they], too, have been thoroughly debunked saying, "Oh, a lot of these Antifa accounts on Twitter are run by trolls." And only a few weeks ago, someone from The New York Times cited one of those Antifa accounts to say, "Oh, look at the crazy left," right? And so, with impersonation and fission campaigns, we're starting to see social media be leveraged in different ways.

And so, I caution as you enter into breaking news zones, know that it's not just you looking for information. That these trolls haven't just thought about, how can I leverage the moment? But they are also laying in wait around certain topics and keywords.

So, I'm interested to talk to you about this. And it's a new idea. It's a new thing that we're working on at Data & Society. We're working on these tactics. We're working to try to figure out, what's the best way to package this threat model and give it to journalists? And so, I welcome all your feedback. Thank you.

[Applause.]

Frédéric Filloux: Thank you very much. Thank you very much for having me here. I'm quite intimidated, because not only I'm not a PhD, but also I'm French, so you'll have to endure my ignorance in many things; plus, you will have to endure my accent.

So, my project is called Deepnews.ai. The first name for it was News Quality Scoring. It was much better in the sense that it was much, much more self-explanatory; unfortunately, no one was able to remember this name, so I came up with something which is more marketable to some extent, which also reflects what we are trained to do, which is surfacing great journalism from the web by using matching learning algorithm, because our approach is not entirely but mostly driven by algorithm.

So, we are trying to solve three problems. The first one is detecting quality signals from the noise. My project is about lifting quality stories from the web. What do I mean by quality? Of course, part of the work, and it has been the initial part of the work, during the weeks, I spend time interviewing a whole bunch of people to ask, what's your definition of a good story? So in this room, we might have hundreds of different definitions. I came up with one which is, for what it's worth, which is basically the differentiation you have been a value-added story and a commodity story. In the value-added story, you have a huge amount of resources that are deployed by a news organization to differentiate its coverage from one to another. So, this is the kind of signals we would like to find out.

The second one is a very important one, because this project is also aimed at being a positive contribution to the sustainability of the news ecosystem. It is aimed at increasing the economic value for great journalism. Great journalism is kind of weird stuff. Digital publishing is kind of weird stuff. Today, there is no correlation whatsoever between the quality of a story and its economic value expressed in CPM. It's a kind of absurd thing when you think in terms of the global economic system, which is the production cost are not passed along to the user, whether the user is an individual or the user is an advertising company.

If you take a webpage, regardless of the content of a story, whether it is a 500-word story put together by a couple of interns about some kind of Hollywood gossip or whether it is a large story which requires the work of many talented writers and editors, the cost of the advertising that will be next to the story will remain exactly the same. This is the kind of absurdity we'd like to correct.

The third thing is, the importance of doing that at scale and automatically. Today, the internet looks like this, roughly. [laughter] So, this is a heavily polluted river. I was considering at some point putting some kind of label on the trash, which is on the foreground, but you might be able to recall. I'm going to mention that later in the presentation. But I mean, we have a problem of scale which could be summed up by this number—there are 100-million links per day, which are injected on the internet. Roughly half of them are in English. And today, the best way to retrieve the good from the bad is doing that manually. This is what fact-checking is doing, especially in this country, in which there is a heavy culture of fact-checking. Unfortunately, I tend to regret that in France we don't have the same culture when it comes to fact-checking. But in this country, you are doing that greatly. And yeah, there are outlets such as PolitiFact which are…[unintelligible]…the industry in that regard.

Still, manual retrieving, manual checking is not the solution. It is roughly like purifying the water of Ganges River one glass at a time. So, we need to find something else. We need to find some way to process the stream at scale. So, this is what the project is about.

Very practically, how the Deepnews platform will work. Let's say you are a publisher or a distributor of news. You would present a whole bunch of stories automatically through API to the machine. I wish it looked like this marvelous rocket, but that's not the case. But, and then you would get a score. Let's say, 3.6 out of 5. What then you do with this score? You can do a whole bunch of things. The first one is, as I mentioned, when it comes to the economic value, you can finally match the price of advertising to the quality of the content. No one has been able to do that before. So, the whole idea is, when a story has a rate of 4.5, for instance, out of 5, which is a super-good rate, that's how we should be able to detect that and be able to service the advertising accordingly, and of course, at a much higher price.

Second thing, our recommendation engine. Recommendation engine, our low-hanging fruit in our business. This is the most—the easiest way to actually increase the number of page views, increase the engagement of reader buy-in, by having more people looking at your content. Unfortunately, right now, it sucks a little bit. Most of the recommendation engine are keyword based. They are simply based on keywords that are detected by some kind of a search engine. And they lift the story based on keywords. We can do better than that. We can parse all the archives with platforms such as Deepnews. We can rank the story, put a score on the story, and be able to sell the story to the reader in accordance to their quality.

The third thing is personalization. I think that personalization, a clever personalization, is the future of the industry. It's kind of super important. Also, smart curation and aggregation are also super important editorial product that should not be left to third parties.

So, if we look under the hood, we built two things. The first thing we built is what we called the Human Scoring Interface, which is a way to actually score stories. We asked the reader—we asked anybody, for that matter…. The URL is available. Unfortunately, it didn't show, due to the stuff. But I will give it to you. So, you can score stories asking various things, such as, what kind of story type is it? What is your evaluation, subjective evaluation, of course, of the thoroughness of a story? What is your evaluation of the balance and fairness? Are all the polity represented in the story? What is the life span of a story? And what is the relevance? And we ask people to give a global score. This is an important part of the process, because I, again, as I say, the human factor is kind of super important.

Second thing we built, we built a deep learning model. So, we came up with something like 10-million articles are ranging from The Financial Times, The Economist, and so on. And also, we took super bad stuff such as Taboola and Outbrain, which are the worst publishers on the internet, as we just saw in the previous picture, and we find a way to find a hidden pattern that will draw clusters of editorial polity. We built a model that has 20,000 parameters right now. This model renders some kind of score percentage, which is translated into a score. So far, we have 90% accuracy. This stuff is, again, constructed to the human evaluation in the experimental phase. We are going to reinject what has been gone through the machine to submit that to human evaluation in order just to override the score.

So, what's next, finally? Our next goal is to create additional larger and diversified dataset, because this is the main problem. Second thing is to boost the human testing interface. Again, you can right now reach it at the https://evaluate.deepnews.ai. You're free to register and to score stories. Again, this is an important way, because this is by many extent, the most reliable way to score stories.

And finally, we'd like to explore new types of models relevant to components of quality stories. For instance, what about modeling the angle of a story? Which is a very important element. What about modeling the depth or the lifespan of a story? This is the kind of research we are doing.

So, thank you very much, everybody. [applause] And I thank the Knight Foundation and the John S. Knight Fellowship for their support.

Lisa Fazio: So, I am a cognitive psychologist. I study memory. My research is focused on how people learn true and false information from the world around them. And then once they do learn false information, what can we do to correct and fix those misconceptions?

So, this project, we were looking at the CrossCheck Initiative. If you're not familiar with it, it's a collaborative journalism project that was run around the recent French election. A bunch of newsrooms in France got together, decided that kind of fact checking was a public good, and so they were going to collaborate and produce these fact checks together. This is some of the partners that were involved in the project. They created these misinformation debunks, put them up on the web, and readers could get to them both from the CrossCheck website, plus from members websites, and even some local newsrooms as well.

So, this is what the debunks look like. There were all sorts of rumors that were kind of floating around in the French election. One of the rumors was actually the one about Macron's offshore bank account that Joan mentioned earlier. And what we did was pick ten of those rumors, presented them to readers both in the U.S. and then later in France, and looked at both whether or not they still believed the rumor after reading the debunk, but then we were also really interested in, what did they actually remember from these debunks? We're presenting this information to people. Do they recall it? Is it having a lasting effect on them? How are they responding to them?

So, in each study, readers would first do pre-ratings, where they saw all ten of the rumors, rated how accurate they thought the statements were. They'd then read one of the debunks. Read the article on how Macron was not wearing an ear piece during the debate. Re-rate the story. And then, answer a bunch of memory questions relating to the story.

So, the ones I'm going to talk about here are questions about kind of the key details in the story. How did the rumor originate? How was it debunked? And kind of one other key piece of information from each story.

So, here's what we found. So, these are the participants accuracy ratings, where zero is saying that the rumor is very inaccurate, ten is saying it's very accurate. The French participants are in orange, and the U.S. participants are in blue. You can see everyone starts out kind of not really knowing if the rumors are true or false. They're in the middle. Post-rating—everyone declines. These debunks do work. People think the rumors are less accurate after reading the debunks. But you can see it's much more effective in the society that doesn't already have preconceived notions, motivated reasoning, and other kind of internal reasons not to believe what's in the debunk.

One thing that's really important, though, is that we see no evidence of the backfire effect. So, there's been a lot of talk about backfire effects and how presenting debunks might cause people to actually become more entrenched in their views. There's some evidence that that can occur in very specific situations, but on the whole, the research community now agrees that that's not really something we have to worry about. So for some participants, these aren't effective in that they stay at the same belief, but we don't see this kind of backfire.

Other thing that's interesting is that one week later, when we went back to the participants in the U.S., they still remembered that these rumors were false. So, we are having a lasting effect. One week later, there's still—there is still evidence of them having read those debunks.

It's also interesting to look at their memory for those specific details. So, this is the proportion correct for those three memory questions. The questions for four alternative multiple choice. They were designed to be rather easy. But you can see that participants aren't that great. So both in the U.S. and France, they're answering kind of half to two-thirds of the questions correctly. So even when people are reading these stories, they are not reading them carefully, deeply, in a way that they're really remembering all of it later on.

We also tested whether or not we had the headline as a question or a statement. So, "Did Macron have an earpiece during the debate?" Or, "Macron was not wearing an earpiece during the debate." We so no difference between those two types of headlines.

Finally, one of the things that CrossCheck did was have logos over to the side that indicated which newsrooms had signed off on that particular debunk. It was designed to kind of be a signal for credibility. The more logos that were there, maybe the more credible the debunk was. But what you can see is that that didn't actually have much an effect on these readers' credibility ratings. And this is within the French sample.

But we've got pretty good reasons to think why it wasn't effective, and that's that participants didn't pay any attention to those logos over on the side. So, whether there was one logo, four logos, or seven logos, they thought there were about three. So, I think it's an important point that we need to pay attention to. What do we want readers to get out of the story or out of how we're presenting the story? And we need to design the features in a way that they'll actually pay attention and remember them.

So in conclusion, we think that this general model of trying something and then going back in the lab to test it, refine it, and then throw it back out in the field again, is really valuable. That researchers know a lot, practitioners know a lot, and by having this back and forth, we can really improve our situation here.

The other takeaways are that culture matters. So, you got much different results in France where they cared about these issues versus in the U.S. And then, that readers ignore kind of all the peripheral information. So, we had some other questions about other peripheral information, like, what type of misinformation was presented, and readers also weren't very good at that.

So, thank you.

[Applause.]

Darryl Holliday: Hello. Yes, my name is Darryl Holliday. I am the co-founder and labs director at City Bureau. Just some quick background before I get started. City Bureau is civic journalism lab that began in 2015 to address four interrelated issues: inequitable misrepresented local reporting, a lack of diverse perspectives in newsrooms, a distrust of media largely within communities of color, and the media business models to do the kind of work of addressing these particular issues.

So, in the last two-and-a-half years, we've designed a public space and digital space where people can engage, converse, and do the good work that needs to be done. We focus our work on the south and west sides of Chicago where we are based.

Quickly, we do this work through three programs. Our first program is a reporting fellowship. We basically bring together emerging journalists, some who don't have any journalism experience, some that do. Pair them up with more experienced reporters, and they work collaboratively to produce work that is published around the country.

Secondly, our public newsroom is a lot like what it sounds like. We open up our physical space once a week, every Thursday, and bring in a variety of co-hosts and guest, and bring the public in with those folks. But it is the Documenters Program that I want to spend time talking about today.

In a nutshell, the Documenters Program, we are paying and training people to go out and document public meetings. But a bit more on that. We have to get to the question, who are the documenters? Who are these people? Currently, there are 330 enrolled documenters. They come from all over the city of Chicago. 61% identify as female, 38% male, 1% non-binary, 36% are white, 33% are black, 11% are Latino or Hispanic, which is really close to parody in Chicago. And the age range is pretty wide. It's 16 to 73. So, they are a diverse set of people across the city.

So, we've been experimenting with the Documenters Program for the last two years or so, doing a lot of projects that, one, help journalists do their work, but two, really disaggregate the skills of journalism and distribute them among the population. But about a year ago, we began thinking of this larger question as a way to focus our work, and that was, who is making the decisions for Chicago, and how do we know where and when the decisions are being made?

The answer quickly came to us that it's public meetings. Public meetings are important spaces for democracy where any resident can participate and hold public figures accountable. The Documenters Program, like I said, pays and trains people to go to these public meetings, produce various types of content, and freely distribute that content to their communities and to our journalists.

Meetings look a bit like this in a sense. In Chicago, there are dozens of commissions, departments that are doing this work, so we're looking at the public school system, the Board of Education, essentially, the City Council, the Transportation Board, city colleges, public libraries. But those meetings are everywhere. They are spread across dozens of websites in different file formats, different types. Some don't have RSS feeds, some do. So, you'd have to look at about 20 different websites if you wanted, as a citizen, to figure out where decisions were being made.

Sometimes they aren't published at all, and sometimes the commissions don't bother to show up. So we sent out a documenter to a commission about two weeks ago, and he found that nobody was there. He was the only person in the room. So, he began live tweeting from that meeting, [laughter], and that's the only record that essentially happened of that meeting. That was a paid commission. I'm convinced that if he had not been there, nobody would have known that they were not actually at the meeting.

So, this initial question of, how do we solve the problem of these meetings being located in different spots? That was a tech solution. So, we began pulling together volunteer civic coders, supplying them with pizza, a good mission, a lot of time for conversation. They began to build scrapers that are scraping all these various websites, the meeting information, and putting them onto one single calendar. So, that will be documenters.org, which is part of our upcoming platform.

The backend looks a bit like this. So, this is what I see when I am seeing all of the…. We have about 2,000 meetings scraped at any given time. When I want to assign out a documenter to a meeting, I go to the second column here, assignment, and choose, maybe we want someone to live tweet or maybe we want someone to audio record, maybe we want someone to submit meeting notes. We send two documenters to every single meeting. Our goal, long term, is to have a documenter at every single meeting in Chicago; hence, knowing where those meetings are located.

This is an early mockup of how we're thinking through what the frontend of this tool will look like. So, the platform itself will contain the aggregator tool, the tool that is bringing together all of these public meetings, but also will allow documenters to login, create accounts, use message boards, and claim assignments. So, this is us thinking through, kind of, what that filter system looks like, what the wizard looks like. But this, I'm sure, will change in the next three months as we develop the project.

So, our question, really, aside from Chicago, is, what is your City Council talking about? What neighborhood is your local police department doing work in? Who governs your water? These are the answers that we can find out through attending these public meetings. And as local media is being gutted in a lot of places, staff is being fired, these local meetings are often the first to go. So, we're looking at, how do we help supplement that work so that journalists can continue doing the work of investigation, of contextualizing information, with engagement from the public.

So, we're starting that with Chicago. This is a map of documenters in Chicago currently. The red is where they are most heavily, so you can see that they are increasingly coming from all over the city. But we want to add more cities. That is the goal of the documenters platform. Once we have that, we will be able to spread to Detroit, which is next. North Carolina is underway. We're looking at a lot of different places.

Thank you.

[Applause.]

Cameron Hickey: My name is Cameron Hickey. For the last eight years, I've been producing science and technology stories with Miles O'Brien for the PBS News Hour. After the 2016 election, we began investigating misinformation on social media. We realized right off the bat that we were going to need a data-driven approach to understanding this phenomenon, if we're ever going to report on it effectively. So, I started building News Tracker.

It was clear right from the beginning that this challenge was going to be the game of Whack-A-Mole. Whoever is creating the misinformation, if it's some guy in his mom's basement in California, to a teenager in Macedonia, or a troll working for the Internet Research Agency, in some sense, they are all trying to avoid detection.

And as we dug deep into looking at the content itself, we realized we also needed a better name for it. Fake news just isn't right. The term itself has been coopted. It lacks a consistent definition. But more importantly, the problem is way bigger than just the fake stuff. So, we call it junk news. What constitutes junk news? If this is everything that's junk, it certainly includes everything that's fake. It also includes stuff that's clickbait, stuff that's hyper-partisan, anything that's plagiarized, and certainly anything that's misleading, and occasionally when satire is made to be tricky, that as well.

I'm going to skip through some slides here just to move along because I know we are running out of time. But so, these were our basic goals: identify news sources, track the content, analyze the patterns, and then make all the data accessible. I will skip my hypothesis and say that when we began looking at hundreds of sites that were identified by other researchers and journalists, all these entities here, we found tons and tons of junk. From there, we looked at all their Facebook pages to actually see what they were publishing on social media. But to solve this Whack-A-Mole problem that we identified, we had to find a way to identify new sources of junk, wherever they were popping up.

It turns out the key to finding new sources is this woman. Her name is Betty Manlove. She 86-years-old. She's a Christian conservative Trump supporter. She lives in Indianapolis, Indiana. And before you guys…. She's not a fake troll or a bot created by the Internet Research Agency. She's my grandmother. [laughter] And I'm not joking. She—she posted this this weekend. [Slide: Like and share if you have an amazing grandson!] [laughter] So, I recognized that Betty Manlove was a critical piece of this puzzle as I searched through junk news pages on Facebook. Betty Manlove had liked almost every one of them. But it was really bigger than that. She's liked over 1,400 stories or 1,400 pages on Facebook. When I started looking through all of those other pages, I found hundreds of new junk sites that no one had identified before.

Of course, it's not just Betty. Most Facebook users who engage with junk news pages have liked lots of other junk news pages as well. So by systematically reviewing the other pages that this audience liked, that would lead me straight to new sources of junk.

So, what do we build? We set up a system to automatically collect every post from every suspect junk news source we identified. There's over half-a-million articles at the moment and counting. We pulled this content from Facebook pages, and until last week, Facebook groups as well.

So, this is what News Tracker looks like. We do a number of different things. For each post, we collect the entire content of the URL as well as their Facebook engagement data, domain registration details, tracking metadata, and we link this URL to every other page that's been sharing it.

We built a simple interface for processing new sources that we collect from the social web, so that we can quickly classify them and track these pages after we review a sample of the content that they are publishing. Created a simple algorithm to come up with a crude junk score to help us sort which content should be reviewed first. And finally, we made it so that we're able to explore the misinformation networks that are created by the associations between Facebook pages and the domains that they are sharing.

As we built the prototype, we learned quite a lot. New domains were appearing constantly. This is a screenshot from this past weekend showing how quickly News Tracker picks up new domains that are sharing junk news. These two at the top were identified and pulled in News Tracker three hours after their domains were registered. We're seeing an average of 80 new domains that have just been registered every month.

In addition, we recognized that meme images can be a potent weapon in the spread of disinformation from these same pages. So, to analyze their content, we are using OCR to extract all the text so we can make that content searchable as well. And finally, to extract patterns that are emerging from the data, we are also using entity extraction on the content to identify popular and salient keywords. It gives us kind of a window into the hot trends in the junk news web.

So, this prototype is still a work in progress, but it’s already helping to serve this larger goal, which is, we’re identifying and tracking new misinformation narratives as they emerge.

I’m just going to jump ahead real quick. This project is going to be moving to FirstDraftNews and the Shorenstein Center at Harvard in order to make the insights that we’re gathering and our ability to investigate this information accessible to a wide range of newsrooms.

I think I can actually cut right there.

[Applause.]