Audio file

2020-07-04_DS4DR_Examples_[Name redacted]_69min.m4a

Transcript

00:00:20

I made him a spaceman gold.

00:00:31 Speaker 1

And let [Name redacted] be she needs.

00:00:33 Speaker 1

A break.

00:00:49 Speaker 1

I cannot. [Name redacted], you can sneak through.

00:00:53 Speaker 1

Second, OK, can you move it?

00:01:00 Speaker 2

This thing is not.

00:01:17

Leave it with water.

00:01:22 Speaker 1

And go put in the now.

00:01:30 Speaker 1

5 minutes. Why?

00:01:33

Oh, it is.

00:01:59 Speaker 1

Let me just double check that.

00:02:22 Speaker 1

Now we are.

00:02:24 Speaker 4

Awesome. So this is a super exciting opportunity to learn a bunch today about data science for disinformation response from [Name redacted]. But before we dive into that, I wanted to see last week, no, no problem.

00:02:46 Speaker 4

On Wednesday, we had the opportunity to dig a little bit into hive and talk about like what will make that more functional and just wanted to circle back and see Roger. If there's anything from there that we were either kind of moving forward with.

00:03:00 Speaker 4

Just to kind of close the loop.

00:03:03 Speaker 2

Yeah, I have my notes from that. I I haven't reached out to [Name redacted] to see if we can get those features added, but I did.

00:03:12 Speaker 2

Speak to his.

00:03:16 Speaker 2

Acquaintance. Partner. I'm not exactly sure who does programming for him, and it's on his list. So he's he's aware of the request for sub cases.

00:03:31 Speaker 2

Whether or not he himself will do it, I'm not sure. So there's a second discussion to have, but he might. So there's maybe, maybe some progress there.

00:03:39

Very cool.

00:03:40 Speaker 2

But that's.

00:03:42 Speaker 2

That's it. That's all I got for you.

00:03:45 Speaker 4

Awesome. And this is where I just tossed to [Name redacted] and say, [Name redacted], do that thing.

00:03:51 Speaker 3

Our recording because [Name redacted] can't make it and it being 4th of July.

00:03:56 Speaker 1

Annual I did just turn on the recording, so we're good. Speaking of which, come over for burgers and you're done.

00:04:03 Speaker 3

If you share the screen, then maybe.

00:04:08 Speaker 3

I can't.

00:04:09 Speaker 1

So the man I like.

00:04:10 Speaker 1

Know and.

00:04:13 Speaker 1

Meeting settings, meeting topic. Allow participants to share screen everybody but [Name redacted] got it. Perfect. Done. Alright, we're good.

00:04:26 Speaker 3

All righty.

00:04:29 Speaker 3

Share screen.

00:04:32 Speaker 3

Carefully not hitting the leave meeting button.

00:04:37 Speaker 3

You can see the Big Blue screen.

00:04:42 Speaker 3

And let's begin.

00:04:44 Speaker 3

So I've been promising you a series on data science for disinformation response for way too long. So this is this is it. Let's get on.

00:04:52 Speaker 3

With it.

00:04:55 Speaker 4

Just a quick pause.

00:04:56 Speaker 4

S can anyone see the screen?

00:04:57 Speaker 4

I don't think I can.

00:04:59 Speaker 2

Yep, I can see it.

00:05:00 Speaker 4

OK, sweet. I must be screwing.

00:05:02 Speaker 4

Something up. OK. Oh, now I.

00:05:03 Speaker 3

See it? My bad. OK, no problem.

00:05:06 Speaker 3

Let let let's get us all. I'm what I'm going to do is just in case this goes hinky.

00:05:14 Speaker 3

I'll also share the.

00:05:18 Speaker 3

Sync in the team chat.

00:05:23 Speaker 3

Damn it.

00:05:27 Speaker 3

This is me messing up.

00:05:30 Speaker 3

But what the hell?

00:05:32 Speaker 3

I should do this every time.

00:05:35 Speaker 1

It's not a true presentation until you actually.

00:05:38 Speaker 1

Screw up the slides.

00:05:42 Speaker 3

There you go. Have shared the slides and now I have no idea where my presentation went.

00:05:48 Speaker 3

So we can all sing together on this one.

00:05:57 Speaker 3

Somewhere on this, not I.

00:06:04 Speaker 3

Is it in zoom?

00:06:10 Speaker 3

I'm going to stop sharing for a second and find my slides.

00:06:17 Speaker 3

This is the second time this has happened to me.

00:06:21 Speaker 3

All right, I have completely lost my window.

00:06:26 Speaker 3

I'll probably find it as soon as I hit share again.

00:06:35 Speaker 3

So the deal seems to be if you are presenting.

00:06:39 Speaker 3

On the Mac.

00:06:42 Speaker 3

And then stop to look something else up like ohh your team window.

00:06:49 Speaker 3

Your machine decides you don't want to present anymore and hides the thing that you were.

00:06:57 Speaker 3

Desktop 2.

00:07:02 Speaker 3

Is it sharing now?

00:07:06 Speaker 3

Yes, got it. Cool. Bye. Let's begin again.

00:07:10 Speaker 3

OK so.

00:07:13 Speaker 3

This is the promise training to start off a series on.

00:07:20 Speaker 3

Data science and disinformation response.

00:07:23 Speaker 3

It's loosely based around a blog post I've been writing about the types of data science you find in disinformation.

00:07:34 Speaker 3

Of data science working on this information.

00:07:39 Speaker 3

Go one to three.

00:07:42 Speaker 3

I did promise you a series, so here's the series. So next week we have the hackathon.

00:07:50 Speaker 3

And for the hackathon, we have an introduction to this disinformation team.

00:07:58 Speaker 3

So this is the order I think we need to be doing this. So if you've not done any of these sessions before, this is a good order. So start with the introduction to the team.

00:08:10 Speaker 3

If you already know the team, then great.

00:08:11 Speaker 3

Hi, we're good.

00:08:13 Speaker 3

Look at the digital harm training. I'm going to write your piece on getting set up for disinformation data science.

00:08:22 Speaker 3

Which will have things like how do you get and set up your Jupyter notebooks? What does pandas look like? Where do you find the Python training you need?

00:08:31 Speaker 3

What are some of the disinformation basics? Here are some pointers to the ethics you need.

00:08:35 Speaker 3

To know about.

00:08:37 Speaker 3

The then we get into the cycle, so the data source is training we did last week.

00:08:43 Speaker 3

Tells you process of getting data. The places you can get data from the places you can put data.

00:08:50 Speaker 3

Before you start doing the day's science itself this session.

00:08:54 Speaker 3

Talking about what tactical data science is looking at, some examples of it, then we get into the weeds on some very specific pieces of data science.

00:09:05 Speaker 3

So how do you do data science on text?

00:09:09 Speaker 3

Specifically on social media text.

00:09:13 Speaker 3

So how do you look at social media text? How do?

00:09:15 Speaker 3

You make sense of that.

00:09:17 Speaker 3

Across networks, how do you do data science on images? So image analysis, but where you look at the images as data points themselves.

00:09:29 Speaker 3

How do you look at the relationships? Are we doing awful lot of relationship based work as as a team? I I think one of you's got a bit of background going on.

00:09:42 Speaker 3

[Name redacted] your window is like glowing a little.

00:09:46 Speaker 3

And having looked at those things then we head off into machine learning. So we're going to touch on machine learning all the way through.

00:09:55 Speaker 3

But it's OK let's let's go through some of the machine learning techniques that you can use and pick up and go and extend those things so you can go build your own design your own.

00:10:10 Speaker 3

The last thing on that chain probably happened earlier.

00:10:14 Speaker 3

Is having done these things, how do you tell people about it?

00:10:19 Speaker 3

So data science is really about the human.

00:10:23 Speaker 3

It's about asking good questions. It's about understanding what people need.

00:10:31 Speaker 3

And it's about being able to answer those questions.

00:10:35 Speaker 3

And get that information to other human beings in a form that they can.

00:10:39 Speaker 3

Do something with it.

00:10:41 Speaker 3

So communication.

00:10:43 Speaker 3

So the styles of communication you can use the places you can put it, the type of narrative styles you can use to communicate with different types of people.

00:10:51 Speaker 3

And things like visualization, because visualizations are a style of narrative.

00:10:56 Speaker 3

So it's just a way of getting a message from you to them.

00:10:59 Speaker 3

And I I don't some of you will have been on the conversation earlier this week where.

00:11:05 Speaker 3

I I had a bit of a reaction to sympatric jobs.

00:11:11 Speaker 3

Because as a data scientist I I just have this visceral reaction to pie charts with lots of teed weenie weenie wedges in.

00:11:17 Speaker 3

Because we're just trained not to do that. So it's about what does this mean? What does this mean? How do you tell good stories?

00:11:26 Speaker 3

Which is the other thing about data science is very much a storytelling art.

00:11:33 Speaker 3

There's a really lovely book called Storytelling with data. This is part of what we do.

00:11:38 Speaker 3

We ask those questions and we tell good stories with it.

00:11:43 Speaker 3

OK.

00:11:44 Speaker 4

That's yeah, this is.

00:11:45 Speaker 3

So exciting. I'm glad you're excited. I'm. I'm just me. Entering through. I mean, it's just like, hey, it was time we did this.

00:11:54 Speaker 3

So why are we?

00:11:55 Speaker 3

Doing this.

00:11:57 Speaker 3

So the first thing is because it doesn't exist.

00:12:02 Speaker 3

There are lots of little pieces of disinformation data science out there. We've basically set some about a few years back.

00:12:10 Speaker 3

But there's not really that much in the way of real time as in in the moment time to respond disinformation data science out there. So it's like we.

00:12:19 Speaker 3

Set the art.

00:12:21 Speaker 3

So part of this is actually forcing us to think about what we're doing.

00:12:27 Speaker 3

We want to train up our team.

00:12:29 Speaker 3

So CTI cookset collab.

00:12:36 Speaker 3

In this wonderful position, have having a bunch of people who want to.

00:12:39 Speaker 3

Learn how to do this stuff.

00:12:41 Speaker 3

Is that you guys too, I hope.

00:12:43 Speaker 3

To see you guys too.

00:12:46 Speaker 3

And while we're doing that.

00:12:48 Speaker 3

And kind of making it up as we go along, it is getting feedback on what we're building, whether it's useful ideas and what we should be doing. So we're doing that all the time anyway and you just kind of look at the course corrections, we do pretty much every week.

00:13:06 Speaker 3

Did we lose [Name redacted]?

00:13:07 Speaker 3

[Name redacted], you still with us?

00:13:10 Speaker 4

He's here.

00:13:10 Speaker 3

You are. Yeah, sorry.

00:13:12 Speaker 6

I'm here. I'm just listening. I had my microphone turned off.

00:13:15 Speaker 3

No, it it's OK my, my little puman thing shows me only five people at a time, so it was just like.

00:13:22 Speaker 3

Oh my God, I'll.

00:13:23 Speaker 3

Lose 1.

00:13:24 Speaker 3

And also I'm doing a university course on this in the fall and it's just like, OK, I'll do some practice.

00:13:31 Speaker 3

So sorry guys, you get to be practice students.

00:13:35 Speaker 4

Awesome. Which university?

00:13:38 Speaker 3

Probably WW, just like the little local 1 here.

00:13:44 Speaker 3

Who knows?

00:13:46 Speaker 3

Wherever I end up.

00:13:48 Speaker 3

So next.

00:13:50 Speaker 3

So that's why we're doing it.

00:13:52 Speaker 3

Sure. Then it were 1010 different sections.

00:13:56 Speaker 3

And next thing to do is kind of make sense of all the pieces and it's like I went and looked up the definitions for.

00:14:04 Speaker 3

Or OK.

00:14:05 Speaker 3

We're talking about data science, but we've also got we're working inside a threat intelligence team. What we do is essentially intelligence analysis, but we're overlapping the OSINT and the dark web and all the other pieces. And there's these long ******* definitions which, you know, feel free to read. I've left links.

00:14:25 Speaker 3

But this is me so it's like TLDR.

00:14:30 Speaker 3

So threatened allegiance.

00:14:33 Speaker 3

Recent current future threats.

00:14:37 Speaker 3

What's hitting us? What's what's going to be hitting us?

00:14:40 Speaker 3

Intelligence analysis. What's going on? So situation awareness.

00:14:44 Speaker 3

Who's involved in that?

00:14:47 Speaker 3

What are our best guesses so so like we we never get full information. So we're always guessing.

00:14:53 Speaker 3

So we guess about what's happening now, we guess, what about what's likely to happen next?

00:14:58 Speaker 3

There's actually another guess on top of that. As a data scientist, you have the time called the data science slider, so you're not just looking backwards at what's happened in the past. That's that's traditional statistics or.

00:15:13 Speaker 3

Looking at what's happening now, situational awareness or predicting ahead predictive analytics.

00:15:19 Speaker 3

You're also.

00:15:21 Speaker 3

Trying to influence your environment.

00:15:23 Speaker 3

So when we do things like counters, we're trying to change the information environment so that this **** doesn't happen so much.

00:15:31 Speaker 3

And if we do it in real time, we shape that environment. So we end up in these big mod player games.

00:15:39 Speaker 3

So intelligence analysis part of that too.

00:15:42 Speaker 3

But it's generally the what the hell?

00:15:44 Speaker 3

Is going on.

00:15:47 Speaker 3

It's about what you can do with.

00:15:49 Speaker 3

Public data so it's.

00:15:51 Speaker 3

That support part of intelligence analysis.

00:15:55 Speaker 3

So again, TLDR and that definition is.

00:16:00 Speaker 3

Doing analysis on its own, his own annalistic.

00:16:06 Speaker 3

A abuse of volunteers. It's an abuse of analysis.

00:16:12 Speaker 3

You do analysis so that you can do something with it, so you can have some action, some change in the world.

00:16:19 Speaker 3

So a lot of our scent is geared towards.

00:16:23 Speaker 3

Getting information to people who can do something in time for them to do it.

00:16:29 Speaker 3

And data science.

00:16:32 Speaker 3

It's something that supports all those things.

00:16:34 Speaker 3

So you've got like a whole bunch of neat skills.

00:16:39 Speaker 3

Which is looking backwards. I mean you've got domains, domain skills, you've got programming skills, you've got maths stats.

00:16:50 Speaker 3

You've got plans and juju's.

00:16:56 Speaker 3

You've worked out the questions you want to ask. You've got the data you've found, and it's that situation awareness. Again, it's like what is happening.

00:17:04 Speaker 3

And the bonus question for all of that is can you?

00:17:06 Speaker 3

Do this quickly. Can you do this in time for people to do something about it? So the data science is really a support for this?

00:17:15 Speaker 1

Like a good a good, a good.

00:17:18 Speaker 7

TLDR on the the bonus question is timely, relevant and actionable.

00:17:25 Speaker 3

Can someone write that down because we'll throw that into the notes?

00:17:29 Speaker 3

Yeah. Thanks.

00:17:35 Speaker 3

It's kind of smartish.

00:17:38 Speaker 3

So yeah, till the we're trying, we're trying to workout what the hell is going on.

00:17:44 Speaker 3

Recap from last week where we did the data data training.

00:17:50 Speaker 3

Data science has a process.

00:17:52 Speaker 3

So you're going from.

00:17:55 Speaker 3

Asking questions, working out what what your problem actually is and what you're trying to.

00:17:59 Speaker 3

Trying to learn.

00:18:01 Speaker 3

Through to getting your hands on the data that's relevant to that or that can proxy. So when I say data proxy, I mean sometimes you don't get exactly the data you need to answer your questions.

00:18:13 Speaker 3

But sometimes there's data that's close enough you can do a good estimate.

00:18:21 Speaker 3

There were some classic ones.

00:18:24 Speaker 3

Like food prices?

00:18:28 Speaker 3

Tie fairly closely to social media chatter. In our case, there's all sorts of Canaries. There's all sorts of anomaly indicators for disinformation going on.

00:18:44 Speaker 3

Go look at the data.

00:18:46 Speaker 3

Look for patterns in it.

00:18:49 Speaker 3

This is where people are starting to use models, so this is where machine learning comes in. It's like work out what's actually going on and then communicate it back, tell somebody in a way that they can do something.

00:19:03 Speaker 3

So clear, actionable stories. Visualizations find insight.

00:19:08 Speaker 4

Yeah, that's yeah.

00:19:09 Speaker 7

I think one of the one of the big things that that I've noticed doing a lot of like threat analysis over the years is.

00:19:17 Speaker 7

A lot of the times as an analyst.

00:19:19 Speaker 7

We get stuck.

00:19:20 Speaker 7

Doing the data entry data normalization before we can even get to the analysis.

00:19:27 Speaker 7

So if you can, if you can normalize it, you can collect it. Then you could spend more time getting that effective analysis so that you can effectively communicate.

00:19:37 Speaker 7

The situation on the ground for action.

00:19:40 Speaker 3

Yeah, I mean, we're we're lucky we're in an environment where the platforms aren't changing too much. So if we can do things like these, miss two or sort of social media to miss bots and.

00:19:54 Speaker 3

Other ways to ingest fast?

00:19:57 Speaker 3

It helps frank spots help things like that normally.

00:20:02 Speaker 7

I think the API's are the the API is now being able to hook into almost any application.

00:20:08 Speaker 3

Yeah, I mean normally as a data scientist, you're coming in raw into most environments and 80%.

00:20:16 Speaker 3

About 80% of my time generally is spent getting and cleaning data.

00:20:22 Speaker 3

It is just, it's not glamorous.

00:20:27 Speaker 3

And I think a lot of the work we've been doing over the last few months.

00:20:32 Speaker 3

Has been trying to cut down the amount of time we spent.

00:20:35 Speaker 3

Just getting data into our systems, yeah, so.

00:20:41 Speaker 3

Yeah. I mean, a lot of that is us trying.

00:20:42 Speaker 3

To speed up this part so we can get on with.

00:20:45 Speaker 3

The interesting part here.

00:20:47 Speaker 3

So yeah, data science process.

00:20:51 Speaker 3

And it's also a team sport.

00:20:54 Speaker 3

So I mean, I call myself a data scientist.

00:20:59 Speaker 3

Because I'm almost a Unicorn.

00:21:04 Speaker 3

Unicorns are people who can do all of the jobs in in the data team.

00:21:08 Speaker 1

Also, because you're a data scientist.

00:21:13 Speaker 3

Actually, most data scientists can do.

00:21:19 Speaker 3

Part of the set.

00:21:20 Speaker 3

So if you just look at the the the original three, I mean doing being able to do business analysis and being able to do math and statistics and being able to.

00:21:31 Speaker 3

Code and do all three of those well.

00:21:34 Speaker 3

You don't get that many people who do that, who can do all three.

00:21:39 Speaker 3

Adding people who can do good visualization on top of that really understand algorithms.

00:21:46 Speaker 3

Really understand strategy and know how to plump together the.

00:21:53 Speaker 3

The back end.

00:21:55 Speaker 3

That they they are. They are very, very, very rare.

00:21:58 Speaker 3

I know.

00:22:01 Speaker 3

Three or four.

00:22:03 Speaker 3

And most of them pulling down a huge amount of money.

00:22:08 Speaker 3

But generally you.

00:22:09 Speaker 3

Get problem people back end people, front end people.

00:22:12 Speaker 3

So it's like there's something for everyone in here.

00:22:17 Speaker 3

The data strategy is at top end of what do we actually want to do with all this data science. So we've been doing a lot of strategy work.

00:22:25 Speaker 3

Business analysis we're doing a lot of that too, but this is a very specific thing we're doing, so it's not.

00:22:33 Speaker 7

So would you say like the?

00:22:34 Speaker 7

The the problem, the problem people fall more.

00:22:37 Speaker 7

Into the tactical.

00:22:39 Speaker 3

Yeah, we're heading that way. We're heading that way, but we still need to do things like data engineering. We still need to work out where we put the data, what shape it's going to be when we, when, when we put it there, how it get back out again, how.

00:22:50 Speaker 3

We get it cleaned up. Think things like, you know, extract transform, load.

00:22:56 Speaker 3

Things like building the tools, the software stuff, that's all data work.

00:23:00 Speaker 3

This stuff, again with the glamorous, you know, doing the actual analysis, doing any machine learning, algorithm development, doing data visualization.

00:23:08 Speaker 3

That that is that 20% cool stuff. 10% cool stuff. Once you've done all the other stuff. So this is this is you need all the people.

00:23:18 Speaker 3

So I'm just like pointing out that it's team sport.

00:23:22 Speaker 3

Going back to the work we did, the section we did on incidents and harms.

00:23:29 Speaker 3

So we went through all the different types of harm, so this idea of physical, psychological, economic, reputational, cultural. So that things that could be harmed.

00:23:38 Speaker 3

This idea of disinformation having a falsehood somewhere in the system, not necessarily in content, but often in context, and this intent to harm.

00:23:50 Speaker 3

Went through why people did this, this idea of money, power, geopolitics, combinations thereof.

00:23:56 Speaker 3

And the different types of harms to nation states the different types of harms to businesses and.

00:24:04 Speaker 3

We went through the narrative list that we had for medical for COVID-19 specifically.

00:24:10 Speaker 3

So there were medical harms we're looking at, and we've been through the stark criteria for incidents. So, you know, is there whole sort is a potential harm, but we the.

00:24:20 Speaker 3

Best team to handle this?

00:24:22 Speaker 3

And the basic ethics of first do no harm.

00:24:28 Speaker 3

Thinking about what we do, thinking about the effects of what we do and and then.

00:24:34 Speaker 3

What do we start with?

00:24:35 Speaker 3

So usually in most of our work, we're starting with something like an image.

00:24:41 Speaker 3

Or someone has given us an image. They've given us a a text message of some form. Maybe we've got a domain.

00:24:49 Speaker 3

Occasionally we've got something I can add for it so we can pull the add the add tanks. Very rarely we get something like a queue and on command signal it's like they they're saying go do this thing and we're finding out they're.

00:25:02 Speaker 3

The other thing we have is the Canaries, so looking at things like you and on activity. So we have people looking at things like fire, right activity.

00:25:11 Speaker 3

And we have world events, so we know when vaccines come in that's going to kick up a whole series of disinformation.

00:25:22 Speaker 3

So we know where we're starting, so recaps.

00:25:26 Speaker 3

And recaps on.

00:25:29 Speaker 3

How we sort out those questions?

00:25:32 Speaker 3

So you know what's important to us? What do we care about? What are we starting with and what are we trying to produce and for whom? So that's all in the old trainings. I'm going to head.

00:25:42 Speaker 3

Oh God, this moral training.

00:25:44 Speaker 3

****** it. We're just going to say, OK, there's a data process, there's a whole bunch of places we can find data, and there's a whole bunch of places we put it.

00:25:54 Speaker 3

And get on to talking about, OK, we've got some data now what?

00:25:57 Speaker 3

The hell do we do?

00:25:59 Speaker 3

So DS4D data science for disinformation response. So tactical disinformation, data science.

00:26:12 Speaker 3

I use the term tactical because there are few people who do this real time.

00:26:18 Speaker 3

Is basically us.

00:26:21 Speaker 3

And the data journalists, some of the data journalists.

00:26:25 Speaker 3

So what you're seeing is these different levels. I mean, I've I've divided them out and strategic operational tactical to get a sense of time scales.

00:26:35 Speaker 3

So a lot of the work you're seeing out of places like Stanford, Shorenstein, [name redacted] and [name redacted], the Fr lab, Graphica.

00:26:44 Speaker 3

Those guys are taking months, sometimes years, to produce results, so they're doing data science, they're doing long form journalism.

00:26:57 Speaker 3

At it's around issues. It's around specific actors. It's really digging in.

00:27:04 Speaker 3

So some operational stuff.

00:27:09 Speaker 3

That tends to be.

00:27:11 Speaker 3

Much more around projects, so you'll seeing development teams on some of the AI machine learning tool. Developers are using data science just as an adjunct.

00:27:24 Speaker 3

To disinformation.

00:27:27 Speaker 3

So that's just like HDCD, if they're doing any sort of lean.

00:27:35 Speaker 3

Enterprise stuff.

00:27:37 Speaker 3

So we're not seeing too much of that in our space and the tactical stuff tends to be around incidents, something happens, kicks off a whole bunch of disinformation or somebody starts an incident.

00:27:48 Speaker 3

So you're then tracking down that incident?

00:27:52 Speaker 3

So we're seeing things like the New York Times report, so this this tends to be in our case ours.

00:28:01 Speaker 3

In the terms of journalists, it's maybe a couple of days. It might be an overnight if.

00:28:05 Speaker 3

It's a snap one.

00:28:07 Speaker 3

New York Times does a lot.

00:28:10 Speaker 3

There's us, some of the machine learning for sick people do this.

00:28:15 Speaker 3

And occasionally you get some of the crisis methods. They're not using that much in the way.

00:28:21 Speaker 3

Data science yet?

00:28:23 Speaker 3

But they've got some tool leaks, so you're going to see some something like the QCR it.

00:28:31 Speaker 3

Qatar Research Institute.

00:28:34 Speaker 3

Will be over and tactical so.

00:28:38 Speaker 3

New stuff guys and this is the kind.

00:28:41 Speaker 3

Of stuff that.

00:28:44 Speaker 3

We're doing and other people are doing.

00:28:46 Speaker 3

So we've got people doing fact checking, source checking.

00:28:52 Speaker 3

So we don't do fact checking.

00:28:57 Speaker 3

And I think we we do what we can in finding.

00:29:04 Speaker 3

If if there are fabrications out there we we do our best but.

00:29:08 Speaker 3

Doing full scale credibility verification is is hard.

00:29:13 Speaker 3

So there are teams out there doing that in near real time, especially if there's like disaster disinfo.

00:29:22 Speaker 3

Source checking that tends to be longer longer term.

00:29:27 Speaker 3

So yeah.

00:29:27 Speaker 7

Hey, [name redacted].

00:29:28 Speaker 7

Do we have ties?

00:29:30 Speaker 7

Currently within the league, if we want something so like a narrative keeps popping up, keeps going back down, keeps popping up. Is there somebody we can hand that over to do a deeper dive for the fact checking?

00:29:43 Speaker 3

We know people in snokes we know people through the credibility coalition, so if we need to we we've left a channel in the credibility coalitions slack.

00:29:53 Speaker 3

As an electing system for us.

00:29:55

OK.

00:29:56 Speaker 3

So if we want it, we got.

00:29:58 Speaker 3

And we're also in in everybody else's channels too.

00:30:04 Speaker 3

Network connection. So Pink slime network. Finding these tend to be more operational again.

00:30:10 Speaker 3

So it's is there a website network setup, so we're looking for those, we're looking for botnet.

00:30:19 Speaker 3

So university Indiana tends to be hunting, hunting big bot Nets.

00:30:28 Speaker 3

Computational amplification. So this is the bots like rising up there. There's some of that going on today.

00:30:36 Speaker 3

And it's political computational amplification. But you'll notice two big hashtags, one on either side that that have gone pretty pretty.

00:30:48 Speaker 3

Big overnight.

00:30:50 Speaker 3

Finding fake accounts. It's kind of part of that. And there's some sort of sub stuff. I think I need to clean this up a bit. Patterns of account creation dates. So there's kind of like a subtask.

00:31:03 Speaker 3

Of looking at.

00:31:08 Speaker 3

Finding those fake accounts, finding those networks, and then under that track, analyze narratives.

00:31:15 Speaker 3

So you're trying to workout what's going on.

00:31:20 Speaker 3

Apologies, this is not as polished as.

00:31:22 Speaker 3

It could be.

00:31:23 Speaker 3

But it's just starting to think about, OK, what are the things we're going to be trying to do? What are things other people are going?

00:31:29 Speaker 3

To be trying to do quickly.

00:31:32 Speaker 3

So how do we start supporting that? How do we scale what's been done?

00:31:39 Speaker 3

Over longer time spans with short smaller numbers of people 2 short time spans large.

00:31:43 Speaker 3

Numbers of people.

00:31:45 Speaker 3

You just basically I think, but if you look at the tactical tasks.

00:31:51 Speaker 3

Pretty much all of them have a tile on the Ammet port.

00:31:56 Speaker 3

So one way you can go look for them and think about them is go to Amit, pick a technique.

00:32:03 Speaker 3

And think is this something we do? Memes do we chase memes? Yeah, we chase memes.

00:32:12 Speaker 3

Fake experts. Yeah, we chase down fake fake experts. How do we how do we? How do we chase them down? How do we do this?

00:32:20 Speaker 3

That's one way to think and and to actually use the tools we.

00:32:23 Speaker 3

Have to think about the data science we need.

00:32:28 Speaker 3

Having wandered through.

00:32:32 Speaker 3

Where data science fits.

00:32:34 Speaker 3

Roughly what it looks like.

00:32:39 Speaker 3

How tactical short time span data science looks slightly different from.

00:32:47 Speaker 3

The longer time span.

00:32:49 Speaker 3

Data science that you'll see a lot of in.

00:32:54 Speaker 3

Disinformation research.

00:32:57 Speaker 3

I I've been wandering around our.

00:33:04 Speaker 3

Interest domain. I'm pretty much wandering around a lot of the strategic stuff.

00:33:10 Speaker 3

Thinking about how we can go tactical with it.

00:33:13 Speaker 3

So a A just a quick look at some of the things that were going on.

00:33:20 Speaker 3

So outside disinformation, data science. So here's here's some of the topics.

00:33:26 Speaker 3

So no surprise, there's a lot of network detection analysis.

00:33:31 Speaker 3

It's like hunt down the bot Nets. Everybody hunts bot Nets. It's like the pothole detector of the data science, the disinformation world.

00:33:41 Speaker 3

Having found accounts or found bot Nets, it's looking at.

00:33:47 Speaker 3

The activity in those accounts, so when are they active? What are they doing? What are they talking about? Do they change activities? I mean, do peoples do things like switch languages? Can you spot things? I mean one of the beautiful things, one of the nice things I found I I saw was somebody who found a four higher botnet.

00:34:07 Speaker 3

By looking at when all the bots changed what they did.

00:34:13 Speaker 3

So suddenly they change focus all at the same time.

00:34:18 Speaker 3

The different topic different language.

00:34:21 Speaker 3

So it's.

00:34:25 Speaker 3

That sort of thinking about what's happening in the world, that situation awareness, it's narrative tracking. So we have a set of rumors, a set of meta narratives.

00:34:37 Speaker 3

That we're watching. So how are those evolving? Where are they turning up what's coming through that's new?

00:34:43 Speaker 3

And we had some quite exciting papers on on narrative formation.

00:34:48 Speaker 3

That would be good to start making something of machine learning.

00:34:54 Speaker 3

So machine learning is not data science.

00:34:59 Speaker 3

Machine learning and artificial intelligence.

00:35:02 Speaker 3

I think of them in terms of.

00:35:11 Speaker 3

What a human would do.

00:35:14 Speaker 3

But doing it.

00:35:20 Speaker 3

And at scale and at speed.

00:35:28 Speaker 3

You're basically training something to.

00:35:35 Speaker 3

Crawl through very large amounts of data.

00:35:39 Speaker 3

And spot patterns that a human given enough time would but would get bored stupid doing.

00:35:51 Speaker 3

Can't remember the phrase I mean in robotics we had. They're dirty and dangerous, but.

00:35:57 Speaker 3

You you basically I I use machine learning at the point at which the humans get overwhelmed.

00:36:06 Speaker 3

Most of what we need.

00:36:11 Speaker 3

Isn't so much to replace what the humans do.

00:36:15 Speaker 3

It's to augment them.

00:36:17 Speaker 3

It's to produce tools that.

00:36:23 Speaker 3

Help the human makes sense of the data.

00:36:26 Speaker 3

That sift through for the first pass.

00:36:30 Speaker 3

Weed out all of the obvious stuff.

00:36:33 Speaker 3

Find the networks, find the places to look so the human can then do a much deeper, much more detailed analysis, or can study the things that are less clear.

00:36:46 Speaker 3

So, for instance, if you're looking at things like classifiers where you have, let's think about the binary classifier binary classifier you're working out whether something goes into one of two buckets.

00:37:01 Speaker 3

Is this a?

00:37:06 Speaker 3

False statement true or false and that is is really crappy example, but actually that's the most the cable data set.

00:37:14 Speaker 3

And some of your data is going to be really obviously this is OK. Some of your data is going to be really obviously this is backlit, crazy.

00:37:25 Speaker 3

Both a large pile of data is going to be I don't actually know and the computer won't be able know. Machine learning again won't be able to tell easily enough.

00:37:34 Speaker 3

And that stuff you give to a human and say, OK, you work it out.

00:37:38 Speaker 3

So this is where it becomes really useful and a lot of good machine learning stuff.

00:37:46 Speaker 3

In this base is basically hand it to a year ago you work it out.

00:37:50 Speaker 3

But just deal with the boring stuff, so tweak classification.

00:37:57 Speaker 3

Whether it's containing hate speech or whether it's looking at the sentiment.

00:38:04 Speaker 3

So when when we say sentiment, it's like it is the thing a happy thing.

00:38:09 Speaker 3

Is it positive about what it's talking about? Is it negative about what it's talking about?

00:38:15 Speaker 3

Remembering, of course, that machine learning is God awful with our.

00:38:22 Speaker 3

Yes, people who have tried sarcasm, detectors too, and GPT language generation detection. So this is interesting. So this is working at point out in a minute.

00:38:37 Speaker 3

I think.

00:38:40 Speaker 3

There's the fire team.

00:38:45 Speaker 3

This is training up machine learning on a whole bunch of.

00:38:51 Speaker 3

This information.

00:38:53 Speaker 3

So that it learns to generate stuff that looks credible enough at first glance.

00:39:00 Speaker 3

And so you're generating it. It also helps with detecting more of it. So it's.

00:39:06 Speaker 3

Kind of a 22 prong. Think article text classification. So there there's a classic Cagle data set you see.

00:39:11 Speaker 3

In a second, which was put out with fake news versus not fake news.

00:39:19 Speaker 3

If you're looking.

00:39:20 Speaker 3

For disinformation data science, you will find a.

00:39:23 Speaker 3

Lot of this.

00:39:25 Speaker 3

Not so much on the not Kaggle, but lots and lots of stuff on the Kaggle. Just look for the Kaggle data set. OK would you remove it in your?

00:39:34 Speaker 3

Oh, deep fakes.

00:39:36 Speaker 7

That's your quick question. What does GPT stand for?

00:39:40 Speaker 3

Ohh God I can't remember.

00:39:43 Speaker 7

OK, I'm going to look at.

00:39:44 Speaker 3

Yeah, yeah.

00:39:49 Speaker 3

Generative God, that no bring God.

00:39:53 Speaker 3

When you find it, just just.

00:39:54 Speaker 3

Yell at the class.

00:39:56 Speaker 3

And defect detection. So there's a lot of work on defect generation. There's also equally a lot of work on detecting the deep fakes happening. I.

00:40:05 Speaker 3

Mean. This is it's nice.

00:40:08 Speaker 3

But deep fakes are really our biggest problem.

00:40:11 Speaker 3

Most of.

00:40:13 Speaker 3

The interesting stuff in detecting disinformation is in detecting the structure around disinformation campaigns.

00:40:23 Speaker 3

It's it's that network theory. It's not actually the content.

00:40:27 Speaker 3

And then machine learning other so things like temporal analysis. So looking at how things change over time.

00:40:34 Speaker 3

And again.

00:40:34 Speaker 7

Generative pre train transformer.

00:40:37 Speaker 3

Generative. Thank you.

00:40:41 Speaker 3

We could have a whole session on those.

00:40:45 Speaker 3

Actually, we are gonna hold have a whole.

00:40:46 Speaker 3

Session on text, so we'll we'll add those in.

00:40:49 Speaker 3

OK.

00:40:53 Speaker 3

Looking at what specifically people are doing, there's an awful lot of our work, so there's a lot of people doing this by hand.

00:41:01 Speaker 3

They're they're going out. They're tracking down the counts, they're going spidering out by hand. So we've got fire eye looking at this Iranian stuff. We got the Jack, the Graphica work, DFR lab, similar stuff.

00:41:16 Speaker 3

Vice I don't know who does this, but I thought I'd at least include one deep fake link so these all clickable links on the slides and go look and see what people are up to.

00:41:28 Speaker 3

And then just looking at the sort of things people are doing, so University of Diana have data scientists working on team. So they're the guys behind tools like hoaxy. So they're doing network science.

00:41:42 Speaker 3

So these these are very very pretty.

00:41:46 Speaker 3

The diagrams.

00:41:48 Speaker 3

But you can look at who's dominating these networks.

00:41:53 Speaker 3

And this is kind of simple. You can look at who's got lots and lots.

00:41:56 Speaker 3

Of people connected to them.

00:41:59 Speaker 3

Like this, but they may not actually be.

00:42:03 Speaker 3

The dominant notes in this net.

00:42:07 Speaker 3

Because you can dominate the traffic through the.

00:42:08 Speaker 3

Net. So someone over here who has control of who comes across it.

00:42:14 Speaker 3

Has a different type of dominance.

00:42:16 Speaker 3

So they do a lot of interesting network analysis.

00:42:22 Speaker 3

Kaggle. Remember I mentioned the Kaggle data set so.

00:42:26 Speaker 3

This is.

00:42:29 Speaker 3

That data set so you could go look at it. It was about 270.

00:42:37 Speaker 3

News articles that somebody hand coded as fake news and not fake news.

00:42:44 Speaker 3

And just like four or five years ago now, I think so. Last election, last US election.

00:42:52 Speaker 3

And this is just somebody has done, I think latent Dirichlet analysis because so latent Dirichlet analysis is so topic models. So what they've done is.

00:43:05 Speaker 3

They've clustered.

00:43:07 Speaker 3

The text in those articles.

00:43:10 Speaker 3

And cluster those articles so they they.

00:43:14 Speaker 3

Worked out.

00:43:16 Speaker 3

What the biggest groups?

00:43:19 Speaker 3

Are and they've they've listed the main words in those groups.

00:43:24 Speaker 3

So government power, America, American country.

00:43:29 Speaker 3

Versus food cancer, health study, drug versus brilliant. Good liar. Nan. Nan. Oh, so we need to do to cleaning versus flu coming. Yeah. So it's it's of its time.

00:43:41 Speaker 3

But you just get a sense of the sort of stuff in there NBC put out the IRA data set from about that time 538 Russian troll trees. You're going to see these data sets a lot, but especially this one.

00:43:58 Speaker 3

I stress again.

00:44:00 Speaker 3

We did. It helped. It's really useful for looking at.

00:44:05 Speaker 3

Methods for classification.

00:44:08 Speaker 3

It's really not so useful for looking at what you need to understand on disinformation data science.

00:44:15 Speaker 3

Although examples, examples of interesting things, so this this is [name redacted], I think they're at Indiana.

00:44:24 Speaker 3

But they were looking at the narratives in COVID-19, so just pulled up a couple of pretty pictures from different different things links in the.

00:44:40 Speaker 3

[name redacted] said it again.

00:44:45 Speaker 3

They looked at just a really simple thing like account age.

00:44:49 Speaker 3

So how old are the accounts?

00:44:53 Speaker 3

Versus how but like are they?

00:44:57 Speaker 3

So you can kind of just look at and think about and. This is just like that data explore before you start explaining. And this is kind of like, OK, here's your.

00:45:09 Speaker 1

So I said.

00:45:09 Speaker 7

Does this kind of go into what we were talking about the two weeks ago? So this is kind of finding your outliers to to dig in. So if everything's baseline and then all of a sudden there's?

00:45:20 Speaker 7

A spike here, that's kind.

00:45:22 Speaker 3

Well, these are they've got a whole bunch of new bots and they've also got somebody did a whole bunch of creation about then. And these are these guys are real more real.

00:45:22 Speaker 7

Of a point.

00:45:31 Speaker 7

But it just helps you. Kind of, yeah. You're you're you're kind of analysis plan.

00:45:35 Speaker 3

Yeah, it is that explanation to someone who can do a do something about it.

00:45:40 Speaker 3

It's like, OK, go look at these guys.

00:45:44 Speaker 3

Oh God, the cagles. Ohh God, I should. I should put this earlier too.

00:45:48 Speaker 3

Remember I said weed the stuff out?

00:45:51 Speaker 3

And this was just like they somebody wrote a paper on all the different algorithms people use on those things.

00:45:59 Speaker 3

GPT 2.

00:46:02 Speaker 3

So this is the Kyrie paper. So these are interesting because.

00:46:09 Speaker 3

[name redacted].

00:46:14 Speaker 3

Does some interesting austint work.

00:46:18 Speaker 3

But he also has one of the MLC people in the same same building, so [name redacted] is partner. I'm wearing my camless T-shirt today, so in machine learning in infosec.

00:46:30 Speaker 3

Person got together with a.

00:46:36 Speaker 3

Researcher and they have an intern between them. Not like that.

00:46:43 Speaker 3

So they were looking at language modelling generation.

00:46:49 Speaker 3

And I think I'm kind of starting to run out of stuff, so.

00:46:54 Speaker 3

Wrap this up and just go.

00:46:56 Speaker 1

Because we're approaching happy hour.

00:46:56 Speaker 3

On to no wrap.

00:46:58 Speaker 3

So yeah, so I tried to make it fairly fairly short. So what are we up to? So this is where we are.

00:47:08 Speaker 3

We we've got OSINT, we're basically running OSINT processes at the moment.

00:47:13 Speaker 3

We've got data scientists cause I was being quietly doing data science pieces. We've got data science help.

00:47:21 Speaker 3

With getting our data data collection together, we've got toolkits coming together.

00:47:27 Speaker 3

But we need to work on our data science practice and this is basically one of the first.

00:47:33 Speaker 3

Stakes in the ground.

00:47:35 Speaker 3

Of getting our data science practice moving coming together.

00:47:42 Speaker 3

I have talked for way too long. It's your turn.

00:47:47 Speaker 3

Talk to me. What do we need to do?

00:47:50 Speaker 3

What did they miss?

00:47:52 Speaker 3

What do we?

00:47:52 Speaker 3

Want to do?

00:47:55 Speaker 7

So I.

00:47:55 Speaker 7

Think it's really going to be kind of diving into the the Jupiter notebooks and identifying the the different collections that we have.

00:48:05 Speaker 7

And then once we have something that comes in or even the narrative tracking, have those be consistent polls and run the the graphics into the data and the OR the?

00:48:16 Speaker 7

In the through.

00:48:17 Speaker 7

The data science pieces so that we can.

00:48:20 Speaker 7

Say, hey, here's where we need to dig in a little deeper. Here's what we need more of an understanding.

00:48:25 Speaker 3

Do we? Is there anyone taking notes?

00:48:35 Speaker 3

Thank you.

00:48:37 Speaker 3

So because when you say the Jupiter notebooks, I mean we have copies of Jupiter notebooks in our GitHub repo. At the moment we don't have a different notebook server.

00:48:44 Speaker 7

So I'm thinking.

00:48:47

OK, so I was.

00:48:49 Speaker 7

Thinking around like the Amet framework, so if we find something and we can map it there, then we could kind of kick that off from from whatever, whatever phase it is. And then that kind of builds out your collection plan.

00:49:06 Speaker 7

If that makes sense.

00:49:09 Speaker 3

It's just when you say Jupiter notebook, I mean that covers a lot of sins and Jupiter notebook is basically some Python with some pretty outputs.

00:49:19 Speaker 3

It's just a way to put code and outputs where these outputs can be.

00:49:28 Speaker 3

Numbers pandas arrays. They can be visualizations.

00:49:32 Speaker 7

But then the outputs can go in.

00:49:33 Speaker 7

The inputs for the.

00:49:36 Speaker 7

The following script right? So if if we find something and it gives us somewhere to start, then we plug that input into into that notebook and then run it from there.

00:49:49 Speaker 7

Through so you'll have output. You'll have. OK. Give me all of the accounts that are and I'll go. I'll go. Basic. Give me all the accounts in the last 24 hours that have been pushing this, this hashtag. And then once that output then it dumps it into it, dumps it into another. OK. When were the? When were the account?

00:50:09 Speaker 7

Creations. What other associated tags has it been pushing out? What other? So that kind of gets into the social?

00:50:15 Speaker 7

Network analysis piece.

00:50:16 Speaker 7

Of it. But I think that that that is is something that that we're.

00:50:23 Speaker 7

That we're missing. We get a lot of the from my perspective and I've been kind of out of the loop for a little bit and I think we've been getting better with the with the.

00:50:32 Speaker 7

With the.

00:50:33 Speaker 7

Threads but diving in and and working forward and backwards and kind of closing those gaps is definitely something that I've been trying to think through.

00:50:43 Speaker 3

OK. I mean that that's still moving.

00:50:47 Speaker 3

Yeah, I can. I can. I can see that's moving a little bit further from just data collection, but I think that's a conversation that we need to start having and involve [name redacted].

00:50:58 Speaker 3

Because [name redacted]'s already been building similar things to that.

00:50:59

OK.

00:51:01 Speaker 7

OK, awesome.

00:51:03 Speaker 6

Hello. Yes, I did guilty.

00:51:06 Speaker 3

You did have a forgotten.

00:51:07 Speaker 6

You know, we're still collecting data with that too. I mean, the database is growing.

00:51:13 Speaker 3

I know I keep looking at it.

00:51:15 Speaker 6

The well, I mean nothing. You what? You're looking at the results that are showing up in Slack, right?

00:51:21 Speaker 7

Yeah, you're talking about the actual.

00:51:21 Speaker 6

You know, I'm talking about the underlying database that that draws from.

00:51:25 Speaker 3

Ah, because that's three, three Twitter disinformation stream, yeah.

00:51:28 Speaker 6

The the stuff you see in Slack is a top ten top 20 list.

00:51:34 Speaker 6

It's it's. It's like it's like the top cut of of what's coming in. The actual database has all the records in it, all the hashtags, all the everything.

00:51:45 Speaker 3

Because we really need to.

00:51:48 Speaker 3

Look at what we're collecting on that because we've still got protest stuff on there.

00:51:52 Speaker 6

Right. And I've I've got it rigged now so that you folks can come in, read only and look at the whole database. I had one successful test. Everyone else has been kind of busy and not able to do it. So if you want to test the ability to come in there using like SQL desktop or what have you, I am rigged.

00:52:14 Speaker 2

Hey, [name redacted], have you ever worked with Elastic Elasticsearch?

00:52:18 Speaker 6

Yes, it was sort of the equivalent of shoving your elbow in your ear.

00:52:26

I'll give you.

00:52:26 Speaker 2

OK, I'll give you that. I'll, I'll give.

00:52:28 Speaker 2

You that one.

00:52:28 Speaker 6

Yeah, I I I used elastic search a long time ago with another company I worked with and I think 80% of my time was spent trying to figure out where the dropped records were going because it kept dropping records.

00:52:40 Speaker 1

Yeah, that that's a fair assessment.

00:52:41 Speaker 6

So I I I didn't really put a lot of time and effort into it all. All I'm doing right now literally is a Myra DB instance on a Raspberry Pi.

00:52:53 Speaker 2

OK.

00:52:54 Speaker 6

If you want to convert it into that, by all means I'll make the data available to you, but I cannot make you an elastic instance.

00:53:03 Speaker 6

And have it work with any amount of sanity that that's unfortunately outside of my ability right now.

00:53:12 Speaker 5

So is the assessment I'm hearing that.

00:53:15 Speaker 5

Elastic search is not ready for prime time. Useful.

00:53:21 Speaker 6

No. What I'm saying is that I can't make it behave. It's outside of my ability. Now there are other people out there that are much better at it.

00:53:31 Speaker 6

And there are other people that have had more success with it and they are.

00:53:34 Speaker 6

Not me, yeah.

00:53:35 Speaker 1

I don't know if you've heard the old saying. Unix is user friendly, it's just very selective about who its friends are.

00:53:43 Speaker 1

That goes doubly so for.

00:53:44 Speaker 1

Elasticsearch and also Logstash and Kibana but.

00:53:47 Speaker 1

That's another conversation entirely.

00:53:47 Speaker 6

So I'm I'm willing to work with anybody who wants to convert it into that. That's not a problem. It's just that.

00:53:55 Speaker 6

These don't look to me for any expertise on elastic search.

00:53:59 Speaker 2

It's a full time job, like it's, it's finicky. And I mean it scales well, it can handle, you know, just massive amounts of data. But it needs care and feeding, and that's the issue. But the the reason I bring it up is because, you know, the common thread here in in all of this is like the data ingestion bottleneck and and that we need to be able to perform like computation over.

00:54:19 Speaker 2

Standardized data, right? And to do that, we can't be collecting like random jsons in, you know, a collection of notebooks or.

00:54:28 Speaker 2

Different SQL databases here and there. It's not going to work right, so we need to standardize in something and so the decisions need to be taken on on on what that what that is. You know what works well for the data scientists. What's your ideal solution and then thinking about, you know, building the parsers, building the, the mappings or whatever it is.

00:54:48 Speaker 2

So that you know, what do we want to look up a user name? You know it's the same query.

00:54:55 Speaker 2

You know, 40 different social media platforms or whatever it is.

00:54:58 Speaker 2

We've we've built.

00:54:59 Speaker 2

There. That's my take. I mean, I don't want to speak to the group, right, but I think.

00:55:03 Speaker 6

No, all I'm saying is that whoever builds it, I will. I will help you.

00:55:09 Speaker 6

You know, I'll, I'll help you export the data and then import the data and all that, but unfortunately I'm not a DBA.

00:55:16 Speaker 2

OK, OK.

00:55:17 Speaker 4

Is that a good hackathon thing or is that are we?

00:55:19 Speaker 4

Too early to ask.

00:55:21 Speaker 4

Clearly enough for what we need there.

00:55:23 Speaker 2

It's a great hackathon thing and I'm sure.

00:55:25 Speaker 1

Yeah, actually, yes.

00:55:25 Speaker 2

I'm sure we haven't.

00:55:27 Speaker 2

I'm sure. I'm sure we have a.

00:55:28 Speaker 2

Ton of elastic.

00:55:31 Speaker 2

Folks in here I have one individual I'm trying to convince or con or whatever into coming in.

00:55:41 Speaker 2

Joining our group specifically for that purpose, but so far I haven't been successful but anyways.

00:55:53 Speaker 2

Yeah. It's to me. It makes a lot of sense, but it's not. It's not just my decision.

00:55:57 Speaker 2

Right. So I think.

00:56:00 Speaker 5

I like what you're saying here.

00:56:02 Speaker 4

Yeah. Thank you for raising it. And yeah, any other thoughts on elastic search?

00:56:11 Speaker 1

It's evil and must be punished.

00:56:13 Speaker 1

No. Yeah, no, these are.

00:56:14 Speaker 1

All great points.

00:56:15 Speaker 6

It's the closest thing to Emacs we have in our generation.

00:56:19 Speaker 4

[name redacted], did you have a thought? Yeah, just.

00:56:23 Speaker 5

Do we have enough of a handle of all the schemas of all the different data sources that we could build something to to take apart whatever data we grab and and put it into elastic search or?

00:56:36 Speaker 5

Whatever we use.

00:56:38 Speaker 2

Yeah. So part part of the work has been done by push shifts. They have mappings for Twitter, Reddit, TikTok, some telegram channels, a couple of other things, but but fundamentally an elastic mapping is very simply a JSON file.

00:56:54 Speaker 2

That maps JSON keys from an input data to a data type. That's all it is, so it's quite easy to build. It's the administration of the elastic cluster itself.

00:57:06 Speaker 2

That's where the Dragons are, so yeah.

00:57:16 Speaker 5

I also have some other point broken.

00:57:17 Speaker 3

The bad data engineering is hard, yeah.

00:57:22 Speaker 5

Who runs an elastic search cluster?

00:57:27 Speaker 3

I tried once.

00:57:30 Speaker 1

Our friends at Amazon Web Services.

00:57:33 Speaker 1

That's that's pretty much the end if I if I had to set up an elk stack or an elk just.

00:57:37 Speaker 1

An elastic search.

00:57:38 Speaker 1

Cluster answer just.

00:57:39 Speaker 1

You go set it.

00:57:40 Speaker 1

Up on what's it is a red ship.

00:57:44 Speaker 2

I believe that's well.

00:57:49 Speaker 2

It's I think was redshift deprecated? I think they replaced it with something. It could be. Yeah. I mean there are so many.

00:57:54 Speaker 2

Cloud services, I'm not sure. Yeah, whatever the new.

00:57:55 Speaker 6

There are so many cloud services inside of a WS that you will it's not impasto. Maybe you will go absolutely bug nuts insane trying to remember all of them for exams.

00:58:08 Speaker 6

Yep. And it's not like, you know, Storm Cloud, cloud front front cloud Lambda. You know this ********** that ********** it's like, you know ah.

00:58:22 Speaker 6

It's great when you're taking the test too, because it's like is it storm cloud, clown Storm, Trail, Cloud, Cloud Trail, or astral projection? And you know you have one answer to that question, and it consists of two words, and they're not happy birthday.

00:58:42 Speaker 4

So any other I guess one thing I'm wondering about is in specifically response to [name redacted]'s presentation, were there any major kind of like questions or ideas?

00:58:53 Speaker 4

That popped up.

00:58:56 Speaker 4

About the kind of like high level discussion of tactical data science that people want to drill down on more in like our following sessions.

00:59:08 Speaker 2

I'd really like.

00:59:09 Speaker 2

To drill down into mapping strategies or or whatever to.

00:59:19 Speaker 2

Techniques in the emit matrix, you know, like whether that's like text classification or defect detection or.

00:59:28 Speaker 2

I expect the bulk of it is going to have to do with network analysis. It would be really interesting to enumerate those things and then overlay them on the matrix so that with with miter attack for example, you have a you have a.

00:59:43 Speaker 2

A spot for?

00:59:46 Speaker 2

I want to say countermeasures, but that's the ammit word, but like mitigations maybe?

00:59:52 Speaker 2

Right. And so when we're doing like research on each of these, having the data science research.

00:59:59 Speaker 2

Like topic or whatever whatever it is would be very useful. I think that could help guide us in in.

01:00:06 Speaker 2

What we build next?

01:00:09 Speaker 2

Or find the right individuals to, you know, focus on the things that are important.

01:00:18 Speaker 4

Awesome. Yep, we can.

01:00:19 Speaker 3

Do this.

01:00:21 Speaker 3

I mean, this was me starting to lay out what data science actually means in practice here.

01:00:29 Speaker 3

So things like what's different about it, what does IT support?

01:00:34 Speaker 3

What does it actually mean to do it tactically at speed?

01:00:40 Speaker 5

All business model for somebody.

01:00:43 Speaker 3

Well, yeah, I mean help. Like I say it's it's not been done. So we're breaking ground.

01:00:50 Speaker 3

Which was interesting. Realizing we were doing something new, just writing some slides.

01:00:56 Speaker 3

But what the hell? That's that's what.

01:00:57 Speaker 3

We do.

01:00:58 Speaker 2

Yeah, I have another question.

01:01:02 Speaker 2

As you mentioned countermeasures and there's GPT 2 up here, are we going to look at the use of bots and you know, kind?

01:01:13 Speaker 2

Of like.

01:01:15 Speaker 2

NLP elves or whatever as a countermeasure like is it within our wheelhouse here to?

01:01:21 Speaker 2

To actually build those things and like.

01:01:23 Speaker 2

You know.

01:01:24 Speaker 3

It has always been within our wheelhouse, damaging those things.

01:01:29 Speaker 3

And there are bots already being used by teams like the Commons project.

01:01:36 Speaker 3

So the Commons project bots are used to identify people who could be brought back together.

01:01:43 Speaker 3

So they actually look at boundaries where communities split apart.

01:01:48 Speaker 3

And then look for people who are similar enough either side of a broken boundary.

01:01:55 Speaker 3

That and they they identify those people, they connect first to using bots to see if those people are willing to talk.

01:02:03 Speaker 3

And the ones that come back then get.

01:02:04 Speaker 3

A human connection.

01:02:08 Speaker 3

It it is, it really is.

01:02:11 Speaker 3

But so there is already precedent.

01:02:16 Speaker 3

In the use of bots for repair.

01:02:20 Speaker 3

There is already precedent in the use of bots in the disinformation space counter bots.

01:02:27 Speaker 3

So it it's not.

01:02:31 Speaker 3

Outside our wheelhouse.

01:02:33 Speaker 3

But it has to be done well.

01:02:36 Speaker 3

We we can't just kind of let a bunch of screaming bots out.

01:02:41 Speaker 2

Yeah, that makes sense, sure.

01:02:45 Speaker 3

Although the stands have done really well.

01:02:48 Speaker 2

That's like kind of a different beast, though, too. You know? Like that's not necessarily like counter messaging so much as just pollution and and distraction. Right. So that's.

01:02:56 Speaker 3

Yeah, yeah. I mean it's it's making sure that what you do is mindful enough that it doesn't produce another form of pollution.

01:03:09 Speaker 3

So you do things like look for data voids.

01:03:13 Speaker 3

And you drop things into the data voids so there is information to be found.

01:03:18 Speaker 3

Rather than spamming those spaces.

01:03:26 Speaker 3

And I hope somebody's writing that down because that's actually a useful use.

01:03:28 Speaker 4

I just wrote it down.

01:03:32 Speaker 3

Of bots bots in this space.

01:03:36 Speaker 2

So one more question, what do what do you need? What do you need from the team as [name redacted] for next steps on this like you know you talked about a place to host Jupiter notebooks. I I have a server.

01:03:46 Speaker 2

That I run respond.

01:03:47 Speaker 2

There's host. You know, there's the the hardware is more than capable to to do that, if that's what we want. It's the maintenance that's the issue.

01:03:54 Speaker 2

Because I'm.

01:03:56 Speaker 2

10% so.

01:03:57 Speaker 3

Actually, we've been given space on one of the CCC TI link service.

01:04:06 Speaker 3

So we've got that we don't have to maintain it.

01:04:06 Speaker 2

OK.

01:04:08 Speaker 2

Ourselves. So. So what do you need next?

01:04:13 Speaker 3

We need. I need out of [name redacted].

01:04:19 Speaker 3

To pull out of his head. What the hell he's talking about.

01:04:23 Speaker 3

Because I'm. I'm not sure we're talking about the same thing yet beside the.

01:04:26 Speaker 1

Oh boy, we're in for a wild ride.

01:04:30 Speaker 3

Yeah, I know, I know. I mean, if we kind of go back through the things I was trying to work through, I I will need some rubber ducking some serious rubber ducking.

01:04:43 Speaker 3

UM over the tasks that we do. Yes, please.

01:04:43 Speaker 4

OK, I'm happy to.

01:04:47 Speaker 4

Rubber duck in that conversation with you and.

01:04:49 Speaker 4

[name redacted] if that's.

01:04:49 Speaker 4

Useful I will just.

01:04:50 Speaker 3

No, I don't mean for the mean for that conversation. We need to do things like the looking through existing work, looking at the Amit framework and talking through what we would need.

01:04:58 Speaker 4

OK.

01:05:04 Speaker 3

Looking at the needs of the people on the back end so the our customer base, so we need to.

01:05:14 Speaker 3

Pull [name redacted] in.

01:05:16 Speaker 3

And talk to her about.

01:05:23 Speaker 3

In law enforcement hospitals, the people who receive our product.

01:05:30 Speaker 3

Who there?

01:05:31 Speaker 3

What they need?

01:05:35 Speaker 3

We need to talk about not just the tactical work, but also the resilience work.

01:05:44 Speaker 3

So there is a second part to the work that we do. So this is the shifts going on. We have to deal with that ****.

01:05:53 Speaker 3

But there's also the how do we help medical organisations groups?

01:06:00 Speaker 3

Possibly individuals become less vulnerable to disinformation based attacks.

01:06:08 Speaker 3

And that's a different thing.

01:06:12 Speaker 3

That adds a part of the preparation work that the rest of the CTI does.

01:06:18 Speaker 3

So it's again, it's matching it.

01:06:21 Speaker 3

And it's also I end up coding a lot.

01:06:26 Speaker 3

So some of this would just having somebody sit with me and.

01:06:30 Speaker 3

Work through.

01:06:32 Speaker 3

Some some of the stuff. Sometimes I'm just doing some really boring jobs.

01:06:35 Speaker 2

Yeah, I'm happy to. Like I think I have a good guess of some of the stuff wraps interested in and I definitely want to sit in on that conversation as well and I'm happy to help build out Jupiter notebooks or web scrapers or whatever it is. I just need to know what it is that.

01:06:55 Speaker 2

What it is you need and what the output looks like, you know so.

01:06:58 Speaker 3

I mean I need to go through the reposts. It's the other thing is at the moment we're using the disinformation repo for both a store of old.

01:07:11 Speaker 3

Disinformation, incident data, and.

01:07:19 Speaker 3

The code base.

01:07:21 Speaker 3

So we need to get that old instant data out thrown up into the hive or somewhere.

01:07:27 Speaker 3

Somewhere out of the UM out of the GitHub so there's some cleaning up to do.

01:07:35 Speaker 3

So we end up with a code base that's actually.

01:07:38 Speaker 3

We can do something with.

01:07:46 Speaker 3

It's also that thought about what we actually need.

01:07:51 Speaker 3

Is that top level? What do we need?

01:07:54 Speaker 3

So some strategy sessions.

01:07:56 Speaker 3

Sorry, you did ask. What, what, what, what? What you needed? What I needed from you.

01:08:00 Speaker 3

Yes, it's a list.

01:08:06 Speaker 4

We got it taken down.

01:08:08 Speaker 1

All right. One last question, how do?

01:08:09 Speaker 1

You want your burger.

01:08:11 Speaker 3

I media.

01:08:14 Speaker 3

I'll be over in a minute. I I give in. I give in. I'm going to. I'm going to come and I have.

01:08:15 Speaker 7

No problem.

01:08:19 Speaker 3

4th of July.

01:08:19 Speaker 3

With you come eat free food.

01:08:23 Speaker 4

I want free food.

01:08:30 Speaker 4

Awesome, yeah.

01:08:31 Speaker 3

I guess I sang for my supper. OK, I'm. I'm going to go and get some free food, guys, see.

01:08:35 Speaker 3

You in a while.

01:08:37 Speaker 5

See you.

01:08:39 Speaker 4

Cool bye.

01:08:43 Speaker 4

So I think we've moved into happy hour.

01:08:45 Speaker 1

Yes, now is the time in which we drink.

01:08:48 Speaker 3

Was that OK, by the way?

01:08:49 Speaker 1

It was great.

01:08:50 Speaker 4

Yeah, that was great.

01:08:52 Speaker 2

We're still recording, by the way.

01:08:54 Speaker 1

Oh, hold on.

01:08:54 Speaker 3

OK, stop there, [name redacted].