

Using Big Data to Understand Global Conflict

Kalev Hannes Leetaru

Senior Fellow, Center for Cyber & Homeland Security, George Washington University

Chair: Caroline Baylon

Research Associate in Science, Technology and Cyber Security, International Security Department,
Chatham House

24 April 2015

The views expressed in this document are the sole responsibility of the speaker(s) and participants do not necessarily reflect the view of Chatham House, its staff, associates or Council. Chatham House is independent and owes no allegiance to any government or to any political body. It does not take institutional positions on policy issues. This document is issued on the understanding that if any extract is used, the author(s)/ speaker(s) and Chatham House should be credited, preferably with the date of the publication or details of the event. Where this document refers to or reports statements made by speakers at an event every effort has been made to provide a fair representation of their views and opinions. The published text of speeches and presentations may differ from delivery.

Caroline Baylon

Hello, and welcome to today's event on 'Using Big Data to Understand Global Conflict'. I'm Caroline Baylon, I'm a researcher on cyber security here at Chatham House. I will be your moderator for today. I've been asked to start with a couple of announcements, which many of you are probably used to. First, the event is being held on the record. Second, you can tweet using #CHEvents.

If I can now introduce our speaker, which is Kalev Leetaru. He is currently a senior fellow at the George Washington University, within the Center for Cyber & Homeland Security. He founded the Global Data on Events Location and Tone project, which uses big data to track conflict globally and in real time. So that's what he's going to be speaking to us about today. Prior to that, Kalev was at Georgetown University, where he was a Yahoo fellow in residence and also an adjunct assistant professor in the School of Foreign Service. He has also worked with organizations such as the United Nations, the US government and the World Bank, and is a contributor to *Foreign Policy* magazine. Over to you.

Kalev Hannes Leetaru

Thank you so much. It's fantastic to be here today. It's a true honour to be here today, presenting to all of you.

I want to open with this image. This really puts in perspective my dream, my vision. Essentially, can we use massive computing power and massive data to scoop up all the world's public information? Not the NSA-style stuff, I'm talking about public things like news media, academic literature – public information. Apply massive computing power to try and understand the world around us. Just to put this in a little bit of perspective, some of my background – most of my work really has focused on essentially – I started my first start-up company in 8th grade, 20 years ago, and basically have been doing data mining and web-related work for a very long time now. Most of my work really has focused around this intersection, what I call 'reimagining our world through data'.

So a project I did a couple years ago, taking all of Wikipedia and pulling out every textual reference of a location anywhere in Wikipedia, every date reference, and essentially making an animated map of 200 years of world history, as seen through the eyes of Wikipedia. Doing some of the original work on the geography of social media, looking at how do we go beyond the 3 per cent of tweets that have an actual GPS coordinate to the rest of those 97 per cent of tweets? How do we induce geographic information from that? It's not enough to say someone is tweeting and saying, wow, this hurricane is really bad – you want to know, where is that hurricane that they're talking about? Looking at the spread of ideas through space over millions of books.

Looking at the declassified State Department cables – this is not Wikileaks, this is the Kissinger-era State Department cables. This is a project with Columbia University and a couple of other partners. Basically what we're doing is looking at about 1.7 million declassified US government cables from the State Department and looking at the flow of information. How were embassies communicating with each other? You see things – for example, that the embassy in Cairo, all of a sudden all the information from Sudan is flowing through that embassy and out to Washington. You see these fascinating changes in diplomatic behaviour that you can visualize when you do it spatially. I've done a lot of work with the group that has fed now what has become Google Constitute. Did some work with Google and a few others on what became one of the earliest studies of how social media is being used in conflict.

Done work with the Internet Archive on visualizing American television news. So this map, the second from the right on the bottom, what we did there, the Internet Archive has about five years of American television news. What we wanted to know is when the typical American turns on the television at night, what are they seeing about the world? Is it all about Syria and nothing about the Congo? What are they actually seeing? So this was actually an interactive map, where you can actually play it forward day by day, switch between CNN and Fox News and all the major stations, and really see what people are seeing. If you think about it from a policy perspective, governments make policy but most governments, at least democratic governments, policy is prefaced on selling that to their citizens. So what do people know? How informed are people about what's happening around the world?

Some work actually with NBC, Universal and SyFy – this is second from the left here. This essentially, they approached me and said: what would the future of interactive television look like? If you think about television as giving the people what they want, what would that look like? If you take, for example, everything said on Twitter about a show and actually remote-controlled a show, as people say, I want this character to do this, or this is my favourite character and I don't like this character. Imagine a show that's evolving day by day from that, over the course of an entire show. Essentially what this project ended up looking like was a leader board of most to least popular characters in the show. As this spread around Washington, I started getting a lot of calls from people saying, could you do this for heads of state? Could you rank all the world's heads of state? So I did that and I'll talk a little bit about that later.

Then some other work with the Internet Archive reimagining the book. Basically taking 600 million pages of books, dating back 500 years, pulling every image off of every page and the text that surrounds that, whether that's Latin or modern French or English, and putting all this up on Flickr, so you can search.

So you can see very quickly that most of my work focuses on big data, so to speak, large amounts of data. But not just large amounts of data – it's really about rethinking how we use data to understand society. Of course, today what I'm primarily here to talk to you about is my project called the GDELT project. The idea of this project is essentially to take the world's information – again, you'll never have all the world's information, but starting off, say, with news media. Taking as much of the world's news media (print, broadcast, web) every day, taking all this information and then – English-language material, sitting over here; anything that's not English, feeding over here and a machine translating that, live in 65 languages, into English. Doing other processing, depending on the language. Then taking all of that and using computers to try and extract what's this news media telling us about the world.

So basically, doing two things. Taking all this news media and, one, saying: what's the news media telling me? What's physically happening? Because if you think about the news media, the news media is two things. One, it's a record of what's physically happening around the world. It's a very biased record but it's a record. It's telling us what's physically happening. So having the computer go through an article like this and saying, well, this article is telling me two things. One, Iraq criticized Turkey, a diplomatic criticism. Two, the reason for that was because Turkey bombed Kurdish militants in northern Iraq.

So imagine basically taking this enormous pile of the world's news media, every 15 minutes, and basically making a live spreadsheet that updates every 15 minutes, cataloguing what's being reported across the world's news media. Again, you always have to remember the news media is not reality. There's always differences there. But being able to look at scale and being able to look across that – this map, this was actually done in 2013, I believe, on Egypt. Basically a live map. Pink is protests, red is violence. But essentially a live map that was updating constantly. It was zoomable, you could actually zoom in, click and say, what's happening down here? And actually pull up the original Arabic-language article and be able to say, wow, there's a report apparently that people are massing down here and protesting this.

But again, a lot of what we care about too is not necessarily physical activity. A lot of what we're really interested in oftentimes is understanding the narratives that are underpinning that. So for example, in Afghanistan right now, what are people saying about women's rights? This may not be a protest of people gathering for women's rights, this may not be attacks on women. This may just be the general climate. When people talk about women's rights, is it pro? Is it saying this is good, this is bad? You think about, for example, the umbrella revolution in Hong Kong. From a physical standpoint, you had all the right pieces for the country to collapse. You've got the right demographics, everyone is gathering here. But what the narrative tells you – and this is an oversimplification, but the narrative shows you when people go from sort of supporting those people to essentially saying, look, enough is enough, you've made your point, you're causing me problems now, go home. Again, that's a gross oversimplification of what happened there, but the ability to see when populations are for protest or against protest, when they're for their government or against their government – you see a lot of this. Even repressive regimes, you see what their neighbours are saying, what the rest of the world. You see the stratification of media. So again, that ability to really reach in deeply to the themes, the emotions of the world.

So from physical events, you can do interesting things. You can, for example, map out Nigeria. You can map out, say, three months of Nigeria and basically put a pushpin on a map for everything that's being reported, whether that's a riot or a protest or a gathering of people or even a diplomatic statement. You can make animated maps showing what's happening across the country.

Ukraine up there – this is actually a very controversial map I did. That map, pink are protests, red are violence, major ones. This I actually did the day before the president fled. This generated a lot of controversy in certain circles because people said: you've got a lot of stuff happening in Crimea, a lot of Crimea discussion, both activity and discussion there, that's really anti-government and pro-Russia. You're showing a lot of that in eastern Ukraine as well. This can't possibly be. The president just signed a peace deal with the protestors, they're already leaving. Ukraine is at peace. Now, I would not chalk that up to an intelligence failure. I think many policymakers, at least on the US side – and I think this is universal – tend to view the world through their own lenses of how they're seeing the world. So to them, if Ukraine needs to be at peace right now, naturally they will just filter out all that.

That's where I think data becomes powerful. Data allows us to play devil's advocate. It allows us to say, actually there's a [indiscernible] right here, you're watching all these protestors. This isn't made up. So I think that becomes very powerful.

We can also answer questions. A lot of questions that I received were: globally, have protests increased post Arab Spring? You look at the papers today, it seems like the whole world is falling apart. So we can do something. This graph, I'll skip a lot of the details, but essentially this is measuring the intensity of global protest activity. So higher means more intense activity, lower means less. The black line is basically kind of a smooth line. What you see is down there at the end, post Arab Spring, you do see a big spike in global protest activity. Of course, this is based on news media, so it could just be that the news media missed the Arab Spring so they want to be hyper-sensitive. But we see this across all the world's countries. We see an increase in reported protests worldwide.

What's interesting is if you look at it, we've been in kind of a 20-year lull where there hasn't been as much protest activity worldwide, at least as reported in the news media. Then we see, in the 1980s actually, we saw quite a bit. The fall of the Soviet Union, all these lovely things that occurred. So part of the reason that to a lot of people it maybe seems like the world is falling apart is that really it's because we've had kind of a quiet period, and we're not quite where we were at the fall of the Soviet Union. This bottom graph is showing Ukraine, so you can look in the context of what's happening right now and how does that fit into

what's happened in the past in Ukraine. So you can see – even I wasn't aware of Ukraine's history with Crimea wanting to pull apart in the past, and disillusionment with the constitution and all those pieces. So it's an ability to look across global media, across decades, and say: what's the picture of things? Where are we today and what does that look like compared to where we've been in the past?

Of course, the other half of that then – oftentimes questions are not what's physically happening but the narrative around that. So a question I got a lot when the Ebola crisis occurred was, at what point did the US start caring about Ebola? It turns out actually – this is volume of American television broadcast news about the Ebola outbreak. This is the WHO announcement, WHO formally announcing this. Just a little blip. Who cares about Ebola? It's not in our backyard. All of a sudden the Americans get it and oh my goodness, this is a huge deal now. Oh my god, we're all going to die now. This becomes very powerful because again, intuitively, reading American news media, you realize: wow, at a certain point, everyone started talking about Ebola, and I don't remember that before. To many Americans, if you ask them when was the Ebola outbreak, they say the fall of 2014 is really when everything happened. Because again, from the perspective of what American news media was covering, this is when it kicks off, when Eric Duncan arrives and you see the real surge there.

What's really interesting is this bottom graph. This is the tone, how positive or negative. I'll come back to this, because emotion is a very complex topic. Essentially, higher numbers are more positive, lower numbers are more negative. You notice this inflection point where all of a sudden news starts becoming more and more positive. That's the point when the Americans got it, and that miracle vaccine came out. All of a sudden the discussion is – because before, the very first scene in that article, was a very uplifting article about Ebola. Ninety per cent fatality rate, guaranteed death sentence. If it arrives on American shores, it will devastate the nation. A very upbeat article. What you saw was most of the early coverage was very negative. It was all about how Ebola is going to kill us all. Then the first two Americans get it and they're flown back to the US, this miracle vaccine comes from nowhere, and then you see American news media really start sharing and saying: American medicine to the rescue. We're going to save the world. We can do anything. That's where you see the 'America to the rescue' narrative.

That becomes very powerful, to be able to actually look at data and see that inflection point. The data will tell you: hey, there's an inflection point here. This is when things start shifting. As a human, you then drill into that and say: why is this? What is this 'America to the rescue' narrative? How does that fit into our psyche? So computers will never replace humans, at least for this type of work, at least for the time being. But where I see machines having enormous possibilities is being able to quantify these things, to pinpoint these inflection points for us.

This is an interesting graph. This was all the media coverage for maybe a six-month period about Wikileaks and extracting out every person mentioned in that news coverage, and then connecting the people that occur in articles together with each other. This isn't looking at who is actually related to whom, it's just saying anytime this person is mentioned in an article, here are the other people that are mentioned. You see big clusters here. The colour codes are names that are mentioned more often with each other than with other names. What's interesting is the blue is Russia, which is here. This is kind of diffuse. It turns out, if you actually start digging through the news coverage – at least, the English-language coverage of Wikileaks – it will say Snowden yadda yadda yadda, and it will say: and he's currently in Russia. So I don't know if that's Putin stage-managing things or if that's just how the media has been portraying it, but certainly the English-language media has been – Russia is really a backdrop. They're not positioning it as Russia versus the US, Cold War again. They're really portraying it as: look at all the stuff the US did, and he happens to be in Russia. So that's a very interesting thing to be able to see, the Russia perspective, how it's being angled and portrayed in the press.

This is on carbon capture and sequestration, a clean air technology that my father actually does a lot of work in. This is five years of English-language coverage of this particular energy technology. Again, what you're seeing is the cluster of names. What are the names that appear in different groupings? What's interesting though is in the periphery of each of these clusters are the journalists, the Wall Street analysts, the others that are more tightly connected with that. So for example, this cluster at the top, you probably recognize a lot of those names – British politicians that are mentioned in the context of this particular energy technology. Two of the names connected to that – Karolin Schaps, who is one of the key Reuters reporters who focuses a lot on the British politician perspective on that, and Alex Morales from Bloomberg News London. So again, a very rapid way of triaging, starting with a big pile – here's five years of news coverage of this type of energy, and then instantly getting – this is completely computer-generated – being able to instantly get this and then understand, wow, there's about six major clusters that this news is breaking into.

So ability to simplify and extract out those narratives and understand what they're telling us. And then we can scale up. This is probably more artwork than analytic. This is a piece I did for *Foreign Policy*, which is basically the first six months of 2013, every name extracted from all that news coverage and how they're all interconnected. It turns out actually, if you drill down, you can even get things like the military hierarchy of different nations. Again, you're not getting the average private, but anyone who's more senior who actually is mentioned in news coverage, especially local coverage. What you'll see, for example, is the people in the military who are mentioned in the same article as the president are probably pretty senior, and then on downward from that hierarchy.

Then of course, this was a map that looks at one day of global coverage of the Mexican drug cartels. Make a heat map of all the locations that are mentioned in global coverage about the cartels. Turns out, this actually matched the DEA's map very closely. So it turns out that news actually gives us a lot of indicators that you might expect but actually go beyond some of that.

And of course, bringing in – one of the things we do now is actually tag all of the world's news media that we can monitor in 65 languages, under a taxonomy – the World Bank taxonomy is one of those. This allows us to really get high resolution in this. So you can say, for example, give me all the coverage right now in Syria. Give me all the coverage about the education of Syrian refugees. Of course, most of that is in Arabic. So this ability – in the West, at least in the US, most of the coverage has basically been: there's a bunch of kids that have been displaced, those poor kids, there are probably going to be some problems educating them. The Arabic-language press has been giving a lot of coverage about that. So being able to really drill in and see perspectives from different parts of the world.

Sentiment, of course. Oftentimes what you care about is not the people who are talking about the government, it's whether they think the government is helping them or hurting them. Sentiment mining is a very complex topic. A lot of stuff that goes behind that. The bottom line is that most of the work that's done today – you probably hear of tone mining, positive-negative – but you talk to psychologists, they don't talk about positive and negative. They talk about all kinds of different psychological signals.

So one of the things that my work does is basically trying to get all these tools that different psychologists and scholars have built, and bring all those in. So instead of saying, I'm going to try and build the world's most powerful tone dictionary ever built, trying to bring all these in. Because if you think about conflict – when Saudi Arabia is bombing the rebels, if a bunch of rebels are killed – there's no such thing as universal truth. There is no positive or negative. From the rebels' perspective, a bunch of them just got killed, that's really bad news. From the Saudis' perspective, that's fantastic, they're winning the war. So you always have to think about, emotion is really who's side you're on, what your perspective is on things.

So being able to drill into things like anxiety – most coverage about Ebola is not saying Ebola is a great disease. Most coverage is going to be very negative about it. What you care about is when that moves from negativity to anxiety – people saying, oh my god, we're all going to die. These are just some of the tools it uses. Of course, you've got to look across languages, you can't just look at English.

This is a simple graph of American television news, intensity of anxiety. You can see the US government shutdown there, everyone getting very anxious about what's going to happen. You can do things – this is pretty interesting. This is the news tone of media coverage about Assad in Syria. This is very interesting. You can see that the tone is becoming very negative about him right before the Ghouta chemical weapons attack. Higher numbers are more positive, lower numbers are more negative. What's interesting here is this big surge right here. This is not positivity. This dictionary takes into account military invulnerability. So a leader who's killing everyone with impunity, that will be considered positive here, because essentially he's all-powerful. You see basically when the chemical weapons attack occurred, you saw the whole world basically say, essentially, his policies are going to start vaporizing shortly, and then you see – well, nothing happened. He got off scot-free. You see, for example, the Israeli media, you saw them light up and say, well, the US drew a red line about Iran too. Once you start saying if people do things, you're not going to deal with that, you've basically set a precedent here. Then you can basically see the major ups and flows of what he's been doing captured in this. There are some interesting pieces on this.

Of course, mass translation – we need to be able to reach into the world's languages. A lot of the world doesn't write in English, news media, if you want local coverage in many areas. So basically this uses massive machine translation. A lot of technical details that go behind this. But essentially right now we live-translate about 65 languages in real time. As the article arrives, the machine translates it and processes it.

This is why this is so important. Blue is English-language coverage of Yemen over the past week, red is a heat map of Arabic, basic locations being discussed in Arabic-language press. Obviously, English-language press, just a few areas. Arabic-language press, like checkpoints are not showing up in English-language press much, Arabic media are covering that a lot.

Then of course, cultural context. It's not just the news media, so processing things like Amnesty's back file, being able to process all this. So you can essentially make a heat map, drill in and say: what's been reported, say, in the human rights literature in a certain area?

Then of course, going to the academic literature. This is a project with the US Army [indiscernible]. We took 21 billion words of the academic literature and looked at a large fraction of what's appeared in over 2,200 journals in the social sciences and humanities about Africa and the Middle East. Geocoded all this. So essentially you can go in and say: the Hutus and the Tutsis fighting in this particular area, over this particular issue, who are the top five people most heavily cited about that? Or food security: make me a map of all the locations discussed with respect to food security in the academic literature over a 60-year period. Be able to say, for example, a lot of people look at this now and say: there's a lot of stuff happening here that's not in the map. That's because that's an area that the academic literature hasn't really covered. Or in South Africa, you see the difference between food security and food insecurity, areas where people can't eat at all and areas where there's a lot of food there but people can't afford to eat it. So being able to look at those different perspectives that are captured in the media.

Then other things that are interesting. Of course, the news media is an incredible source of imagery. Right now we have a number of partnerships that we've been doing where we work with human reviewers to go through – this is a project actually with Micro Mappers. What we do is we take, for example, a natural

disaster like Vanuatu, take all the imagery that's coming out of that area, and send that for humans to tag that imagery. Flag that this is a picture of a downed building. We're feeding that now to [indiscernible] network tools to go through those images – they can actually recognize the species of tree that's down. So saying this is a downed palm tree versus this is a banana tree. That can be very important to local economies.

Of course, being able to then say what's actually happening around the world. Making live maps of protests and violence around the world. Being able to show, for example, what's trendy in the last 15 minutes. It's been amazing going through this amount of data because you think about, at any given moment, what's everything that's happening around the planet right now? Even the stuff that's just being reported in the world's news media. When you go beyond western media, there's a lot of stuff that happens there.

So this is actually the BBC versus the *New York Times*, for all of March 2015. It's a heat map. You can see the *New York Times* kind of goes in a burst once a day. You can see the BBC has a lot more coverage of the world. So you can actually visualize and compare two different news outlets. This is actually a strong argument why you should never just use one news outlet for your information. In fact, a lot of academia historically has been based on just coding the *New York Times* for all world news. This is a good perspective of why you have to look across multiple media.

This is a visual of everything that we're monitoring for a two-day period but in 15-minute increments. So you can see the pulse as Europe wakes up and there's tonnes of coverage, and then it kind of goes to sleep. It pulses. You see these fascinating pulses. A lot of you have seen maps that look like this with Twitter but being able to then see that actually from news media, being able to look at these similar types of visuals.

Then of course, coming up to the end here, forecasting is critically important. What's so interesting about maps like this – there's a lot of complexity and I'm almost out of time, so I'll touch on this a little bit. What this project looked at is to say: are there cycles in world history? You usually hear that history comes in cycles. Can we actually assess this computationally, using all this data?

So today, GDELT has over 300 million records of about 300 categories of events, down to the city level, worldwide from 1979 to present. So could we actually look across all that data and see, are there regular cycles? Imagine saying today – so basically this black line right here was January 27th. Egypt is kind of getting into full swing. Imagine saying on that day, rewind the clock and say: what's going to happen right now? Is Egypt going to completely collapse, is it going to restabilize? What's going to happen in Egypt? Imagine saying on this day, on January 27th, make a timeline of the last two months of Egypt, everything happening in Egypt. Basically, make a timeline where you're measuring the stability of the country. Higher numbers are more bad stuff is happening, lower numbers are it's more stable, there's less stuff happening.

So essentially you make a timeline and you say, let's look across all world history, for every period in the last 30 years from any country, that's most similar to this particular timeline. Then look at all the periods in world history, rank them – find the top couple periods from world history, from any country, that's most similar to where Egypt is right now. What happened after each of those periods in the past? Then average those together.

So what you're seeing here: red is what actually happened in Egypt, the actual instability line. Green is Sweden right around December 2010 or so. The black line essentially would be – what you'd be doing is this part of it is where it's matching. The computer is trying to find a match. You see these lines. Then if

you look at the right-hand side, how do you run this on that day? The red is what really happened, the green is the prediction of what the machine thinks would have happened. Essentially what you see is it's not a perfect match, but you do see, even these spikes, where you see the spikes happening. So you see there's a lot of regularity.

Here's another example. This is the Ukraine conflict. You can see red is what actually happened in Ukraine, green is the average in the past, basically Turkey in 1999 and Lebanon in 2007. Both had periods that were very similar in Ukraine prior to the fall of the president. So the day the president fled, the two months leading up to the president fleeing, it turns out that these countries had similar things. Again, this is not looking at the type of things. Obviously Ukraine, the president fleeing, other things happening, did not happen in these countries. What it's looking at is the general trend of whatever did happen in this country, sort of fitting what happened and then looking at what happened in the past. It turns out there's a lot of similarities here. So it tells us that there are some really interesting patterns to history.

Of course, this is my contact info. Anyway, I know I just ploughed through a tremendous amount of material. I'm famous for doing 200 slides in 30-minute talks, so today I gave you a lot less than that. But hopefully this gives you an idea and we'll have an interactive session now. But hopefully this gives you a feel of what's possible when we start thinking about – oftentimes, we think about policy. I work a lot with historians and other folks who say you can't really reduce a protest to an entry on a spreadsheet, because it's really about the narrative, it's putting it in a social and cultural context. I completely agree with that. But this is kind of a first attempt at how can we use large amounts of data to allow us to both access what's happening today – and something I've been exploring a lot right now is, how do we surface what's happening right now around the planet? I didn't realize how many flags get burned at any given moment around the world. You think about ferries sinking, we're all probably looking right now in Europe at the migrant sinkings, but look at all the other maritime accidents that are happening at any moment around the world. It's really eye-opening when you start moving beyond your local western media and you start really looking at the world's media, and you look across all this media throughout the world, especially in local languages. You look at small, local media, local community broadcaster in the corner of a rural area. Maybe a two-watt local broadcaster. When you start looking at that material and you're actually seeing that and translating that, it's fascinating the perspectives that you see.

Even especially when you go beyond the physical event and you start looking at how people are talking about the world, what are the narratives? Some of the NGOs I work with say women's rights are a huge barometer, incredibly powerful indicator of the stability of a nation. Not just physical activity like women being attacked but just discussion of it at the beginning. Is this important, is this overblown, is this something we should strive for? These are incredible indicators to be able to get at – again, news media is news media. It can be very biased. But even in countries like North Korea, which have not really mastered the dictatorship control over media, all the neighbouring countries – you look at South Korea, Russia, China, you look at what the neighbours say about it. You get a lot of detail from this.

Anyway, I guess now we'll do a Q&A session.