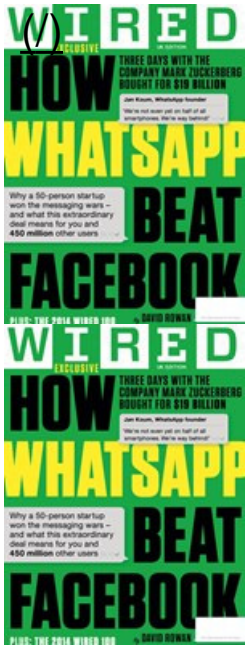
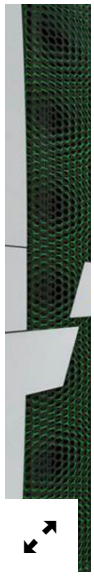


Big data and the death of the theorist

TECHNOLOGY (/BROAD-TOPICS/TECHNOLOGY) / 25 JANUARY 13 /
by IAN STEADMAN (/SEARCH/AUTHOR/IAN+STEADMAN)

Plenty of people have foreseen the death of the scientific theory at the hands of big data analysis, but when computers become good enough to understand literature, art and human history, will it spell the end for the humanities academic?



SGI

A lot has been written about the ways that big data has changed scientific enquiry, but as supercomputers increase in power and the tools to use them become less obtuse, whole new academic disciplines are beginning to feel the benefits of crunching data.

Believe it or not, some people even think we can forecast the future with big data. Predicting world-changing events is a possibility, some claim, if you treat society and history like a big data problem. It's how big data analyst Kalev Leetaru (http://www.kalevleetaru.com/) found where Osama bin Laden had been hiding, in a way.

Leetaru's work made the news (http://www.bbc.co.uk/news/technology-14841018) in 2011, as it claimed to pick up early clues to where the al-Qaeda leader had been living in Pakistan, just from publicly available sources of information. This was after the fact, of course, but the point was it could have found him, maybe, if someone had thought to look for him. That same method could, possibly, pick up where the next bout of social unrest will appear in the Middle East, or reveal a new history of the US Civil War -- or at least, that's the claim.

His research involves taking vast quantities of data -- usually on the scale of millions, if not billions, of individual data points -- and running algorithms that look for the connections between them on supercomputers. This is the essence of big data, a field with a name that both summarises the problem and offers nothing of what that actually means. One possible definition of it might be how humanity copes with all the information that it produces, and the web, and social media, means that there is a lot of information out there to look through. Exabytes upon exabytes.

"The thing about the reality about everything that happens on Earth, a hundred years ago we only understood the tiniest fraction of that," Leetaru explains, the day I met him. "Political scientists, over the last half century, have largely studied political unrest in other countries through the *New York Times*. Literally, paying teams of graduate students to open the *New York Times* each day (http://eventdata.psu.edu/papers.dir/AJPS94.pdf), read through it, and clip articles or check boxes on what they're seeing (http://polmeth.wustl.edu/methodologist/tpm_v14_n1.pdf). *The New York Times* is a wonderful paper, but it's not necessarily the best paper to understand the Liberian civil war through."

For scientists and mathematicians, working with supercomputers makes sense -- their information is numerical. It already exists in a language that machines can read. The interesting thing here for historians and sociologists and literature critics, and everyone else who works with language and the vagaries of the human condition, is that we've reached a

point where supercomputers are fast enough to crunch that data just as easily as anything else.

Leetaru does this with a host of scripts he knocks up on his home desktop, using a range of tone dictionaries that let his programs filter through tweets and judge them for sarcasm, irony, sincerity, rage, and so on. He does this by relying on the masses of information to find a general consensus, one that's statistically meaningful: "If someone is using the words 'wonderful', 'terrific', 'fantastic', more likely than not there's probably going to be a more positive meaning than 'horrible', 'horrific', 'awful'. Things can go wrong, but that's where the power of big data comes in. If you're looking at ten tweets and you're getting a few wrong, you've got problems. If you're looking at ten billion tweets, basically it washes out as noise. The real patterns are the ones that survive the noise."

He loads those same scripts onto a supercomputer and lets them loose on all the information he can find, linking topics that to a human researcher might seem disparate and unconnected. A tweet is just as much of a source for understanding humanity as a newspaper article, and the program recognises that.

He said: "The theoretical models that are out there predicted zero of the recent recognised civil wars (http://www.polsci.wvu.edu/faculty/hauser/Summer2011InternalConflict/WardEtAlPerilsPValueJPR2010.pdf). We've done very poorly at it, and a lot of that's because this theory is basically gut feeling. You have so little data, you look at ten data points and you have to use all ten of them to build your model, and all you're doing with testing your model is you're retesting on the same data you trained that model on, versus now -- we have enough data to say, OK I'll set aside this data, and I'm going to build my model on this data, and test it on new data it's never seen before, and see how accurate it is."

According to Leetaru, 50 percent of the CIA's intelligence (https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/volume-54-number-1/the-scope-of-fbis-and-bbc-open-source-media.html) is now culled from the web, which is how he found bin Laden -- or, rather, judge bin Laden to have been living within 200km of Abbottabad. It was simply mining the news that pours onto the

internet from agencies and local papers every day and finding the connections that human analysts would never have spotted.

Similarly, for the Arab Spring, news mining can track changes in the public discourse that might foreshadow social unrest. He said: "If you look at Egypt with the revolution, it would have held on if the Coptic church bombing (http://articles.latimes.com/2011/jan/01/world/la-fg-egypt-church-attack-20110102) hadn't happened. Those people died on New Year's Eve, and then what you see immediately is the social contract being discussed. You see this notion of 'we're not like those Americans, we don't have the freedoms and rights, but we live in a safe and secure society'. It's a social contract. All of a sudden it was 'we can't get security for our freedoms, why are we trading this?' You see that narrative explode and accelerate towards the downfall."

The big data approach to intelligence gathering allows an analyst to get the full resolution on worldwide affairs. Nothing is lost from looking too closely at one particular section of data; nothing is lost from trying to get too wide a perspective on a situation that the fine detail is lost. The algorithms find the patterns and the hypothesis follows from the data. The analyst doesn't even have to bother proposing a hypothesis any more. Her role switches from proactive to reactive, with the algorithms doing the contextual work.

"What I'm doing is saying, extract out every possible permutation of words and features out of this text," Leetaru explains. "Every combination of words out of this pile of text each day, and run it through everything. Try every possible permutation of every model. I've so far done half a million models and climbing, and this'll probably be up in the billions by the time I'm done, and what's happening is I'm finding results that are enormously significant. Way, way beyond the current state of the art. Basically take the news content, take today's coverage of Egypt, and forecast what's going to happen tomorrow."

This approach -- previously common in economic history -- has begun to appear in disciplines that you might not think are susceptible to it (and there's a fantastic piece on what some call "cliodynamics" in *Nature (http://www.nature.com/news/human-*

cycles-history-as-science-1.11078) by Laura Spinney about it that's well worth reading). There are unexpected crossover benefits, too. Only one percent of tweets are geotagged, but Leetaru can place them with contextual information. Someone's profile might mention being a student at the University of Illinois, once tweet about a coffee shop in Chicago, and then in another tweet mention a street name -- that information allows triangulation that geotags a tweet as effectively as anything else.

You can see this same process as work elsewhere, like with the University of Toronto team who are dating medieval manuscripts (http://www.technologyreview.com/view/509876 /the-algorithms-that-automatically-date-medieval-manuscripts/) by analysing their language and using peculiar turns of phrase as a kind of dating benchmark. Leetaru has worked with the digitised contents of the US Library of Congress to map changing attitudes to slavery and racism across the American south in the era of the American Civil War. Tweets and historical documents are all as accessible, as cross-referenceable, to the big data analyst, and all are as valid as sources because the vagaries of language and context aren't a problem, they're a boon.

As a joke, I suggest a dystopian future where terrorists set off bombs at the point when the signals from the public discourse show they'll cause the most damage. It's not a ridiculous suggestion, though. Leetaru even suggests, offhandedly, that in the future a government could monitor in realtime the economic health of its people, and swoop in with emergency economic measures before things get too bad.

It's the kind of realtime, analytical insight that feels more at home in science fiction. In essence, we're reaching a point where everyone can use big data, regardless of how comfortable they are with that situation. It's a future that instantly reminds me of Chris Anderson's " end of theory (http://www.wired.com/science /discoveries/magazine/16-07/pb_theory)" piece that questioned whether big data spells the end of the scientific method.

For science, it makes sense to see big data as a revolution. Algorithms will spot patterns and generates theories, so there's a decreasing need to worry about inventing a hypothesis first and then testing it with a sample of data. But the thing with Leetaru's work is that it isn't working with numerical data. It's working

with words, with jokes, with sarcasm and sincerity. Those are the kinds of things that we have humans for, because they are what makes us human. Right?

Karl Marx spent 12 years in the British Library developing both carbuncles and the intellectual framework for *Das Kapital*. While many of his ideas may not be fashionable in the economic mainstream, it's notable that he did predict that even the intellectuals would one day need to face up to being replaced with machines. It's doubtful, however, whether he would have foreseen an automaton one day being able to look through all of the sources that he used -- and millions more -- within a fraction of the time he spent, and being able to present its own models of history.

In the same way that the internal combustion engine spelled the end of the horse as a working animal, big data could be the tool to render host of academic disciplines redundant if it proves better at building better narratives of human society.

Speaking to academics about this, though, and there's scepticism that the big data scientists are coming for their tenure. Mark Graham (http://www.oii.ox.ac.uk/people/?id=165), from Oxford University's Internet Institute, points out one of the primary reasons to be wary -- even a sample of millions is still selective in some way. "There is no doubt that 'big data' approaches are allowing us to ask questions that have never been asked, see patterns that have never been seen, and get all sorts of new insights into the human experience, but there are many other types of speech and many other voices that are important to listen to," he stresses.

The move towards quantative (rather than qualitative) analysis is happening, but "when talking about 'big data' and the humanities, there will always be things that are left unsaid, things that haven't been measured or codified". He explains: "I

do get why people think that 'big data' will mean the end of theory, because you can now answer almost any conceivable question with large data sets and transactional data shadows, but irrespective of how big or complete our datasets are, they will always be selective and partial. We're talking about a classic 'if you have a hammer everything starts to look like a nail' issue here."

This makes sense -- not everyone tweets, and not everyone who tweets geotags their tweets. Even with the aforementioned contextual geotagging of tweets, that still leaves a sample of tweeters that isn't absolutely everyone. It's still a sample of "people with the capability and urge to tweet".

That's something that Melissa Terras (http://www.ucl.ac.uk/dis/people/melissaterras) from UCL's Centre for Digital Humanities agrees with -- and even big data patterns need someone to understand them. She said: "To understand the question to ask of the data requires insight into cultures and history. Just because you made a pretty map that looks pretty, it doesn't answer a question that improves our understanding of it. We're asking the big questions about society and culture."

The revolution that big data brings to the humanities -- and any subject that deals with humanity on a profound level -- is that it provides a new way to construct models and narratives. But we have to know if those narratives are equivalent to the truth, and the gut feeling there is surely that they're not.

Terras continues: "It's not the same as pointing a telescope to the sky and every second taking a reading from the telescope. The data is complex, the data is ugly, the data is problematic. Most digital humanities work isn't individual, and that's the big threat. The last 200 years of humanities has been the lone scholar in the ivory tower, writing the treatise on the fall of the Roman Empire. But now, when you're doing digital humanities stuff, you need your programmer, your interface expert, you need someone who can rig up a database, and so on."

It looks like the academics in the future might have to learn to code if they want to be able to do their research to a level that we'll accept as significant, but they're still going to be needed for their insight. "Important questions in the social sciences and

humanities about equality, power, voice, justice, fairness will always be around, will always require sustained and critical inquiry, and won't ever fully be answered by computers alone," Graham argues.

Listening to stories of forecasting the future (and being told that the important thing is to letting the data "tell us" what it means), it becomes incredibly tempting to foresee a future where the researcher is as redundant as the horse became after the invention of the internal combustion engine, but perhaps what we're seeing here is that the revolution that big data is bringing to science is different to the revolution it is bringing to the humanities -- it's not so much a change in method as a change in perspective.

"I always joke, to paraphrase the great Indiana Jones," Leetaru says, "that big data is about finding patterns, not truth -- if it's truth you're after, the philosophy department's down the hallway."

Image: Shutterstock (http://shutterstock.com)

Edited by OLIVIA SOLON

4 COMMENTS

LATEST ON WIRED.CO.UK