Making sense of big data

A galaxy of user-generated data points is providing a near-unimaginable quantity of data that can improve disaster preparedness and response. But there are some problems that must be overcome, warn **Ian Portelli**, **Ramin Bajoghli**, **Megan Mantaro** and **Amanda Horowitz**

The destruction of Hurricane Sandy in New York City in 2012 is well documented: flooded Subway stations, evacuated hospitals, thousands of homes and cars destroyed, widespread power failure and an estimated $18 (€16.19) billion in damages. Yet, despite the blackout, New Yorkers in lower Manhattan relied heavily on social media to disseminate information, co-ordinate relief with various government and non-governmental organisations, and to communicate with family and friends.

Twitter and Facebook quickly emerged as reliable outlets for Consolidated Edison, New York's primary electrical provider, to communicate with customers. In a 2013 report, the Federal Emergency Management Agency (FEMA) concluded that 20 million Sandy-related tweets were posted during the storm's peak. New Jersey's largest utility company, PSE&G, had its Twitter account suspended briefly because it far exceeded the number of posts allowed per day. The few businesses in lower Manhattan that owned gasoline-powered generators quickly set up charging stations, allowing people stranded by the storm to charge their phones. In a city with a massive electrical blackout, charging cell phones to stay connected with family, friends, and social media went from being a luxury to a necessity.

In times of crisis, organisations seek platforms to reach out to the largest number of people at once, and what better way than by targeting people on their mobile phones? During Hurricane Sandy, Twitter became an essential resource to many people by sending them updates on their mobiles. Twitter Alerts allowed people to receive updates on resource distribution locations and safe evacuation routes. These alerts were sent directly to cell phones as a text message, without the user even having signed up for a Twitter account, in order to make the information more readily available.

Unbeknown to some, the rising popularity of social media and smartphones, reconnaissance drones and satellites, and cloud-based networking, has created a galaxy of user-generated data points that has enhanced our capability to co-ordinate disaster response. The sheer volume of user-generated data from a natural disaster (20 million tweets alone during the peak of the hurricane) is an example of big data. Big data is an all-encompassing term that describes the availability of data that is generated every second by individuals, businesses, and governments. Data from structured (databases), semi-structured (metadata tagging), and unstructured (social media) sources contribute to the exabyte of data stored.

Traditionally, big data has been described as the three 'V's: the volume of data, the velocity of data processed and, as mentioned above, the variety types of data. In a 2011 report it was estimated that 1.8 trillion gigabytes would be collected in 2011 alone, and that the amount of big data would grow exponentially every year thereafter.

Data generation and collection have been occurring as far back as the 1970s, but the pace at which data is generated today is unfathomable compared to a decade ago.

In terms of natural disasters, government agencies have been employing big data to co-ordinate response and relief primarily by employing satellite imagery and other high-tech equipment. Previously, technology was mostly limited to military and government agencies and open-source coding was not largely embraced. As the technology sector expanded, hackers and other specialised computer programmers represented the second generation of network analyst. This

generation ushered in an era of open source coding and embraced real-time, user-generated content – such as crowdmapping – that expanded the pool of available data.

In a 2012 study *The American Journal of Tropical Medicine and Hygiene* highlighted the potential effectiveness of big data and disaster response. The study concluded that, with the help of social media, public health officials could have predicted the 2010 cholera outbreak in Haiti two weeks earlier than initially reported. Automated surveillance platforms were used to compare social and traditional media posts related to cholera with the officially-reported cases. The authors of the study found that the number of media posts correlated with the number of official cases being reported. If analysts had analysed the data properly in real time, health care officials may have responded to the outbreak sooner.

In the spirit of using user-generated data to monitor and predict events, Google unveiled Google Flu Trends (GFT) in 2008. GFT tracks Google search queries and analyses them to predict the prevalence of a regional influenza outbreak. The algorithm utilised by Google analyses trigger words related to influenza, such as 'cough', 'flu', 'fever' and 'flu-like symptoms', among others. For example, if GFT identifies an increase in search words related to the flu in Queensland, Australia, it will predict an increase of the influenza virus for that region. GFT was created to complement, not replace, the standard monitoring of the influenza efforts by The Centers for Disease Control and Prevention (CDC).

Initially, GFT was applauded for its early success in predicting influenza trends; however, it has been criticised for its declining success in predicting outbreaks. Regardless of GFT's shortcomings, it is a good illustration of the tools at the disposal of epidemiologists to curb the spread of diseases.

Health officials are not alone when it comes to analysing the data available to them via social media posts. The Central Intelligence Agency (CIA) of the United States pores over five million tweets a day, cross-referencing social media posts with news from various parts of the world. It has been reported that using this analytical approach, the CIA predicted the Egyptian uprising in 2011 and gauged the world's reaction to the killing of Osama bin Laden.

Skewed conclusions

Further, the New York Police Department (NYPD) has partnered with Microsoft to unveil the Domain Awareness System (DAS), a real-time database that merged numerous city official databases, 3,000 closed-circuit television cameras, hundreds of license plate readers, and radiation detectors that are simultaneously being used around New York City. DAS allows the NYPD to gather real-time information from various sources in one convenient location.

Although big data has tremendous potential in the prediction and monitoring of disasters, analysts of big data face a myriad of challenges. The amount of available data is overwhelming, and each year this grows exponentially. Before being analysed, big data must be organised and compartmentalised systematically, while taking note of signal error and confirmation bias. Signal error (overlooking large gaps of data) and confirmation bias (the use of data to support a pre-existing viewpoint) can cause analysts to reach skewed conclusions that lead to detrimental outcomes. Succumbing to either of these errors could lead to a misunderstanding of the severity or nature of a disaster, or to analysts sending supplies to an area in less need than another.

Further, big data analysts using Twitter face additional hurdles combing through relevant data. Twitter messages, or tweets, often lack proper grammar, use abbreviations, and contain slang words, which impede an analyst's efforts to locate relevant tweets. Moreover, these analysts can face signal error, owing to the demographics of Twitter users. Only 16 per cent of the population uses Twitter, and this demographic tends to be younger, wealthier and more urban than the ▶

▶ general population. When analysing Twitter data during a disaster, response agencies must ensure that this does not result in poorer, less urban areas being overlooked.

The data gathered from Twitter during Hurricane Sandy illustrates the difficulties of using Twitter for big data. Over 20 million tweets were posted about the hurricane, with a majority originating from Manhattan. Seaside Heights and Midland Beach were much more affected by the hurricane than Manhattan, but fewer tweets came from these areas because of power outages, uncharged phone batteries, and a lower overall concentration of Twitter users. If an analyst looking at the Twitter data had falsely concluded – based on the number of tweets – that Manhattan sustained greater damage from the storm, supplies could have been sent to the wrong areas and it is likely there would have been a higher death count in Seaside Heights and Midland Beach.

Many tools are being developed to address the overwhelming task of organising and analysing the multitude of available data. One such example is the free and open source platform called Artificial Intelligence for Disaster Response (AIDR). AIDR is a three-pronged platform that helps to identify Tweets related to natural disasters, tags them, and customises – or 'trains' – the system to identify relevant posts. AIDR was utilised during the 2013 earthquake in Pakistan with mixed results. The epicentre of the earthquake was located in a remote area in south-western Pakistan, which had little to no social media activity. Yet 1,000 tweets were tagged and AIDR was able to identify, with up to 80 per cent accuracy, whether the post was relevant to rescue efforts. If such accuracy can be maintained, AIDR will be a welcome addition to the arsenal of tools available for disaster response teams.

Violations

Another major concern with big data is the issue of privacy. A study published in *Nature* in March 2013 revealed that even data that has been stripped of personal identifiers can be used to identify individuals. The study examined 1.5 million phone records, determining that 95 per cent of individuals can be identified with a mere four data points of when and where a call is made. With only two data points, half of the individuals could be identified. To put this in perspective, identifying an individual with a fingerprint requires 12 data points.

Moreover, the issue of privacy within big data is not restricted to phone logs. Social media site providers like Facebook and Twitter employ the practice of 'notice and consent' to obtain their customers' permission to share their information with third party companies. However, as The United States President's Council of Advisors on Science and Technology notes, this practice is unfairly skewed in the provider's favour as they offer complex, non-negotiable terms that the user must evaluate, realistically, within a few seconds. Thus, the burden of privacy protection falls almost entirely onto users, who subsequently face increased violations of their privacy.

Big data sources do not stop at Facebook and Twitter, in fact they also include other social interfacing applications that many people carry around in their pockets, regularly. Smartphone applications like Instagram, Tumblr, and Snapchat allow individuals' locations and information to be sent and identified by one simple swipe of the geotag filter. This expanded pool of data fosters the question of how can we use this information to enhance disaster relief in times of crisis while maintaining the personal privacy and wellbeing of individuals?

Privacy concerns are not restricted to the use of social media. Simple, everyday tasks have been made more accessible thanks to the mobility of smartphones. Tasks like linking debit cards to urban transport passes to top them up automatically when the balance is low, create data points that provide a baseline for tracking and monitoring the movement of an individual. Once a connection to that person's bank account has been established, other data points can be extrapolated and cross-referenced, creating a complete portfolio of an individual that includes spending habits, travel history, and possible relationships with other individuals. A single data point created with a swipe of a Subway pass provides the gateway to breaching a person's privacy.

With the exponential growth of big data and the endless opportunities to do both good and harm, it is vital to regulate and ensure the integrity of the data generated. Disaster response and humanitarian teams often work in harsh conditions in countries with authoritarian governments that can use big data to monitor illegally and/or harass workers in the field. The safety of these workers needs to be fully guaranteed and addressed when developing applications of big data as tools to supplement their work. Neither are democratic governments immune from illegally using data to monitor their citizens. The leaking of classified information over the last couple of years has exposed the scope of illegal data gathering on individuals by the United States and other democratic countries. For better or worse, the ethical use of big data is a concern that needs to be addressed by all the players involved.

A single data point created with a swipe of a subway pass provides the gateway to breaching a person's privacy

Yet, despite the aforementioned problems, the potential for big data and disaster response is endless. To address some of these issues, the US National Science Foundation (NSF) and the Japan Science and Technology Agency (JST) are currently working to overcome the challenges of using big data for disaster management. In particular, the agencies aim to improve big data capture and processing and to ensure that data analysis is dependable, even through the power and system failures that are frequently brought upon by the disasters under scrutiny for the purposes of this undertaking.

To achieve this, the two agencies will fund various teams to research and find solutions for problems faced by analysts of big data. The NSF will fund approximately six to eight projects worth up to $300,000 (€269,524) each over three years. As these agencies work to overcome big data's challenges, analysts who seek to use big data for disaster management should be aware of the possible problems with the data and should not automatically take the results of big data analysis as fact. Despite its vast potential, big data currently remains one element of the wider disaster and humanitarian response picture, and is not the complete solution. CJR

■ *Portelli, Bajoghli and Mantaro are* CRJ*'s R&D bloggers on www.crisis-response.com*

Authors

Ian Portelli *PhD, MSc, is the Director of Clinical Research and Analytics for Health Quest Systems Inc, Director of Clinical Research at Vassar Brothers Medical Center, and an Assistant Professor of Emergency Medicine at the New York University School of Medicine;*

Ramin Bajoghli *is a second year medical student at The University of Queensland in Brisbane, Australia. He has also worked as a photo/video journalist for several international news organisations;*

Megan Mantaro *is an honours freshman studying International Business with a minor in German at Northeastern University, Boston, MA, USA;*

Amanda Horowitz *is a student at Vassar College studying the sciences behind society, economics, and biology*