Qadir *et al. Journal of International Humanitarian Action* (2016) 1:12
DOI 10.1186/s41018-016-0013-9

Journal of International Humanitarian Action

REPORT Open Access

Crisis analytics: big data-driven crisis response

Junaid Qadir[1], Anwaar Ali[1*], Raihan ur Rasool[2], Andrej Zwitter[3], Arjuna Sathiaseelan[4] and Jon Crowcroft[4]

Abstract

Disasters have long been a scourge for humanity. With the advances in technology (in terms of computing, communications, and the ability to process, and analyze big data), our ability to respond to disasters is at an inflection point. There is great optimism that big data tools can be leveraged to process large amounts of crisis-related data (in the form of user generated data in addition to traditional humanitarian data) to provide an insight into the fast-changing situation and help drive an effective disaster response. This article introduces the history and the future of big crisis data analytics, along with a discussion on its promise, enabling technologies, challenges, and pitfalls.

Keywords: Big data, Crisis response

Introduction

Motivation

According to Solnit 2009 disasters, besides being devastating, can bring people closer that consequently creates an altruistic community where people help each other and fight the ongoing calamity together. Here, we take the example of Haiti earthquake to provide motivation for big data driven crisis response and the formation of a community powered by digital technology that helped the affected people of the earthquake. In 2010, Haiti faced a devastating earthquake that left a huge number of people dead and many homeless. Humanitarian aid was predominantly delivered in a different manner that time, which was mostly driven by digital technology. The concept of *digital humanitarianism* (Meier 2014) became popular after this incident. Digital humanitarians (the people who participate or volunteer to deploy technology for the humanitarian aid) used a number of emerging technologies—e.g., read-write web, big data analytics, participatory mapping, crowdsourced translation, social media, and mobile technology (we will discuss these in detail in the upcoming sections of this paper)—to catalyze an effective response after the Haiti disaster.

Huge volume of data related to the crisis—including short message service (SMS) from onsite victims, social media data from citizens, journalists, and aid organizations—was subsequently collected. Sifting through the voluminous "*big crisis data*" (big data collected during a crisis situation (Meier 2014)) to find information about the affected population was a challenging task. This challenge was tackled by the digital humanitarians by employing techniques such as crowdsourcing to acquire data and producing *crisis maps*. The online-crowdsourcing platform used to collect data for the Haiti earthquake was Ushahidi 2016. Ushahidi provides a mobile-based platform for developing "*crowd maps*" through collecting, visualizing, and mapping citizen-supplied (or crowdsourced) data. After Haiti, we observe that data and technology driven disaster response has become a norm. This led to the emergence of a new kind of distributed intelligence: *Disaster Response 2.0* (Crowley and Chan 2011).

The humanitarian aid and emergency response is not limited only to situations involving natural disasters. Data-driven digital humanitarianism is changing the face of humanitarian aid and development in a wide variety of fields: be it (1) a migration/refugee crisis (e.g., the Syrian migrants crisis); (2) an epidemic crisis (such as the dengue crisis in Punjab, Pakistan); (3) a natural disaster (such as the Haiti/Kashmir Earthquakes); (4) crowd control problems (such as the Mina crush crisis in Hajj 2015); (5) terrorist attack (such as the Peshawar, Pakistan school attack); (6) civil wars (such as in Iraq and Syria); (7) public violence (such as post election violence in Kenya);

*Correspondence: anwaar.ali@itu.edu.pk
[1] Electrical Engineering Department, Information Technology University (ITU), Lahore, Pakistan
Full list of author information is available at the end of the article

and (8) other disaster situations (such as public disorder, industrial accidents, and infrastructural failures).

The importance of disaster response in the perspective of big data and technology can be realized by the fact that the 2013 World Disasters Report (Vinck 2013) dedicated itself to the topic of *humanitarian information and communication technologies (HCIT)*. It highlights the importance of various enabling technologies ranging from mobile phones and social media to big data. It discusses different case studies where these technologies can be used for the effective disaster response and management along with various challenges that these technologies entail. Next, we discuss different stages of a crisis and what role data plays in each of these.

Crisis life cycle

The overall crisis life cycle can be divided and analyzed into three stages (Fig. 1) namely before, during, and after a crisis. Let us call the first stage *Pre-Crisis Preparedness* that deals with the in-advance preparation of a potential crisis (in terms of logistics, planning, and warehousing of emergency relief goods regionally and internationally) along with efforts for disaster risk reduction (Twigg et al. 2004). Big data analytics can be useful in this stage for emergency prediction either before the initiating event (such as an earthquake/tsunami) or as part of the unfolding of an ongoing crisis (such as prediction of the refugees' flow in the aftermath of a natural disaster such as the Haiti earthquake). Any efforts focused on preventing impending crises will also be part of this stage.

We call the second stage *During-Crisis Response* and it deals with coordination (among aid agencies) in the aftermath of a crisis to address any gaps or overlaps that may be affecting the effectiveness of the response. This stage also involves adjusting and monitoring aid to ensure accountability and effective dispatch of resources so that it is available at the right time and at the right place.

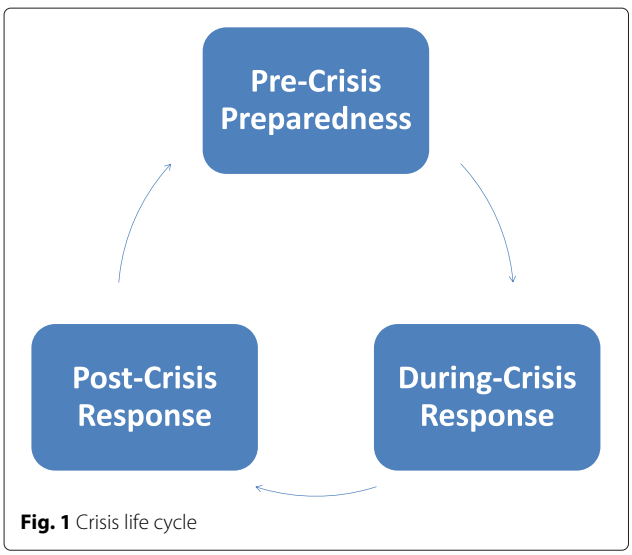

Fig. 1 Crisis life cycle

Finally, we term the third stage *Post-Crisis Response* that deals with reporting, auditing, and accounting of the overall humanitarian response. This stage looks at the crisis response in a more comprehensive, analytical and insightful way so that lessons are learnt and insights developed for any future crises. Big data can help ascertain the efficacy of relief work and make it more accountable by using actively acquired data (gathered through, e.g., unmanned aerial vehicles (UAVs) or social media). Big crisis data analytics has the potential to transform all the stages of a crisis life cycle (i.e., pre-, during-, and, post-crisis) and can be useful for emergency prevention, preparedness, response, and recovery. Next, we summarize how our discussion in this paper is structured.

Road map

The focus of this article is on reviewing how big data analytics and digital technologies can be used to respond in emergencies in ways that can mitigate or even avoid a crisis. In Section "Big data and crisis analytics", we first explain what big (crisis) data is and how we can acquire it. In Section "Enabling technologies", we discuss modern technologies along with machine learning techniques that enable data-driven crisis analytics. In Section "The big crisis data ecosystem and case studies", we present how the modern world is reshaping itself to create an ecosystem where crises can be fought based on big data along with two case studies that highlight the importance of such an ecosystem. In Section "Why big crisis data analytics is challenging?", we discuss different challenges in the way of bid data-based crisis response. After presenting future directions in Section "Future directions for big crisis data analytics", we finally conclude our paper in Section "Conclusions".

Big data and crisis analytics

In this section, we first discuss, in detail, what is meant by big data. Why it is gaining so much importance these days; how it is collected, stored and used in the perspective of crises and humanitarian emergencies. At the end of this section we explain what we mean by *crisis analytics*, which is basically deploying different analytical and machine learning techniques on the big data collected during a crisis situation.

What is big data?

In the modern world, we are inundated with data as the massive amounts of public data is being generated on a daily basis (James et al. 2011). The glut of data is further increasing exponentially due to the rapid proliferation of mobile phones and the increased digitization

of all aspects of modern life due to technologies such as the Internet of Things (IoT); which deploy sensors, for example in the shape of wearable devices, to provide data related to human activities and different behavioral patterns. The commoditization of data collection with the advancement in digital technology has led companies to collect all kinds of data, such as the digital trail of an online user, the transaction history of a customer, and the call details record (CDR) of a mobile-phone user. Companies such as Facebook, Google, Twitter, Yahoo, and Microsoft routinely deal with petabytes of data on a daily basis. It is estimated that we are generating an incredible 2.5 quintillion bytes per day (Siegel 2013).

In order to understand what is new about big data based crisis analytics, we will have to understand, first, what big data is. *"Big Data"* refers to our emerging ability to collect, process, and analyze massive sets of largely *unstructured data*—e.g., word documents, emails, blog posts, social, and multimedia data that can not be stored in an organized fashion in relational databases (Leavitt 2010)—from a multitude of sources to find previously inaccessible insights. With the transformative role, big data has played in diverse settings, there is a lot of interest in harnessing the power of big data for development and social good. We can define big data to be data that exceeds the processing capacity of conventional database and analytics technology (in that the data is unstructured, too large or too fast). Now, we briefly discuss modern database technology that is being used to store big data.

Database technology for big data

As mentioned before, big data is dominantly unstructured data. The traditional relational databases can not store such data, which require data to be structured (that resides in fixed fields, e.g., spreadsheets) and stored in a conventional relational sense. A novel approach is required to store and access unstructured data. NoSQL (or non-relational) databases have been developed for the said purpose (Leavitt 2010). As compared to relational databases, NoSQL databases are distributed and hence easily scalable, fast, and flexible. Big companies also use NoSQL databases, like Amazon's Dynamo (DeCandia et al. 2007) and Google's Bigtable (Chang et al. 2008), for data storage and access. One potential downside, though, in using NoSQL databases is that they usually do not inherently support the ACID (atomicity, consistency, integrity, and durability) set (as supported by the relational databases). One has to manually program these functionalities into one's NoSQL database. Now, we describe what we mean by the term *big crisis data*.

Big crisis data

Big crisis data refers, simply, to big data collected during crises or mass emergencies. Big crisis data, just like big data, can be of two types: *structured* and *unstructured* (with the latter being predominant). It has been suggested that the biggest gap in big crisis data informatics currently is in the extraction of structured information from the huge amount of unstructured information. Big data analytics is well-suited to deal with unstructured data: in fact, a major motivation of its development is the fact that traditional data tools are inflexible about structure and cannot process unstructured data well. In the following two subsections, we will first study the various sources of big crisis data followed by how big data analytics can be used to process this avalanche of information.

Sources of big crisis data

Here, we consider six different types of big data sources that, broadly speaking, principally drive the big crisis data revolution. These include data exhaust, online activity, sensing technologies, small data/MyData, public/governmental data, and crowdsourced data. This is illustrated in the taxonomy presented in Fig. 2. In this figure, apart from the data sources, the enabling technologies, future directions, pitfalls, and challenges of crisis analytics (introduced later in this paper) are also shown.

Data exhaust

It refers to information that is passively collected along with the use of digital technology. Our lives—with our ubiquitous use of digital devices—are increasingly online and leave behind a traceable digital trail. The most important example of data exhaust for big crisis data analytics is the mobile call details records (CDRs), which are generated by mobile telecom companies to capture various details related to any call made over their network. Data exhaust also includes transactional data (e.g., banking records and credit card history) and usage data (e.g., access logs). A promising use of data exhaust in crisis analytics is to use mobile money transactions[1] where such services are established (as in Africa, e.g., the M-PESA (Big data in action for development 2014) is one of the most profitable banking services in Kenya). Most of the data exhaust is privately owned by organizations (such as mobile service operators); such data is mostly used in-house for troubleshooting and planning and is seldom shared publicly due to legal and privacy concerns.

Online activity

This encompasses all types of social activity on the Internet such as user generated data on the Internet (e.g., emails, SMS, blogs, comments), social media (such as Facebook comments, Google+ posts, and Twitter tweets), search activity using a search engine (such as Google) etc. Not all online activities are used alike in crises; it has been shown in literature that Twitter and SMS are used differently in crisis situations (e.g., SMS is used mostly on the

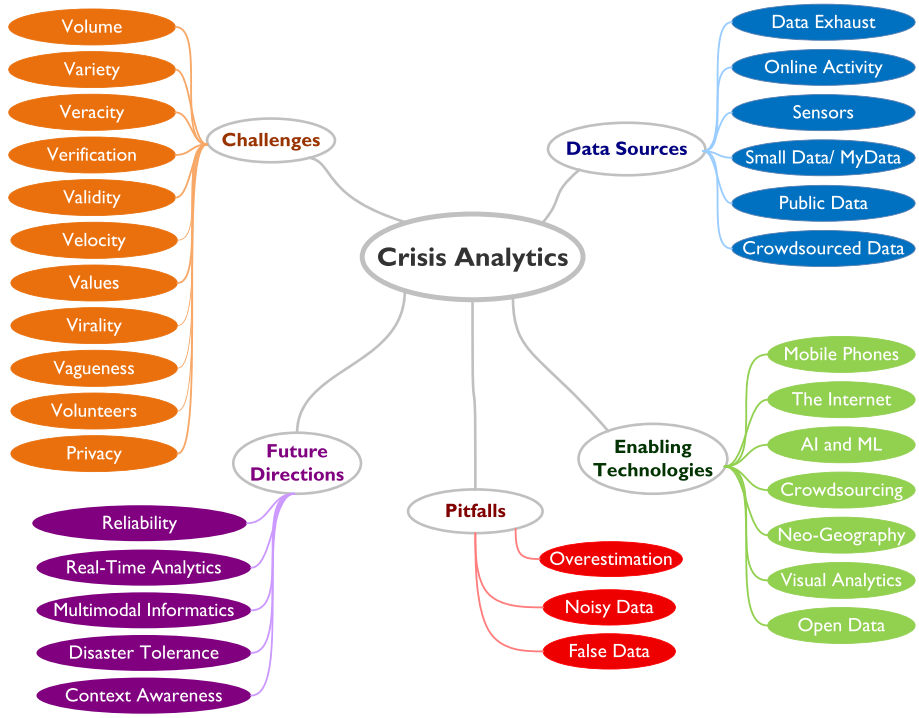

Fig. 2 Taxonomy of crisis analytics

ground by the affected community, while Twitter is used mostly by the international aid community) (Munro and Manning 2012). One advantage of online data is that it is often publicly available, and thus it is heavily used by the academics in big crisis data research.

Sensing technologies

Sensing technologies mostly gather information about social behavior and environmental conditions using a number of sensing technologies. With the emergence of architectures such as the IoT (explained in the next paragraph), it is anticipated that sensor data will match or even outgrow social data soon. There are a number of sensing technologies such as (1) remote sensing (in which a satellite or high-flying aircraft scans the earth in order to obtain information about it (Big data in action for development 2014)—e.g., for measuring the traffic, weather, or other environmental conditions); (2) networked sensing (in which sensors can perform sensing and can communicate with each other—as in wireless sensor networks); and (3) participatory sensing technologies (in which everyday devices—such as mobile phones, buses, etc.—are fit with sensors; in particular, modern phones are equipped with multiple sensors such as accelerometer, gyroscope, microphone, camera, GPS, and can effectively work as sensors). Various cyber-physical sensing systems—such as unmanned ground, aerial, and marine vehicles—can play an important role in disaster management. These devices can act as an active source of big data by providing on-the-ground information in the form of images or sensed data. Sensing data is mostly (but not always) public. Advances in open mapping technologies (such as the OpenStreetMap (Haklay and Weber 2008)) are helping in further democratizing the access to sensing data.

Now, we briefly discuss the Internet of things (IoT), which is a new field fueled by the hype in big data, proliferation of digital communication devices and ubiquitous Internet access to common population. In IoT, different sensors and actuators are connected via a network to various computing systems providing data for actionable knowledge (Manyika et al. 2015). *Interoperability*, which is the harmony of data (specifically its format) from one system with another, is a potential challenge in the way of IoT expansion. IoT finds its application in healthcare monitoring systems. Data from wearable body sensory devices and hospital healthcare databases, if made interoperable, could help doctors to make more efficient (and near real-time) decisions in diagnosing and monitoring chronic diseases. On a bigger level, the health status of a local region can be monitored in real-time. We believe that this can be used to evaluate the efficiency of a healthcare mission in a region.

Small data and MyData

Big data by its nature has to deal with torrents of data, and thus the field's tools of choice revolve around analyzing

filtered and aggregated data. With big data, the units of sampling and analyses are vastly dissimilar (e.g., the unit of sampling is at the individual level, while the unit of analysis can be at the country level). In "*small data*" (Estrin 2014), even if the datasets are large, the unit of analysis is similarly scoped as the unit of sampling. Personal data, or "*MyData*", can help optimize individual human lives (in terms of health, productivity, etc.). There is ongoing research on how to use digital trails of information that users leave as they go about the routines of their daily lives; there is emerging interest in using small data and MyData for personalized solutions such as those that focus on health (e.g., a startup named *Open mHealth* (Open mHealth) led by the Cornell Tech's professor Deborah Estrin (Deborah 2016)) and sustainable development (e.g., the Small Data lab at the Cornell Tech. and the United Nations University (Small Data Lab - UNU; The Small Data Lab at Cornell Tech)). Hospitals now have started providing individual patients with access to their individual medical records data. The availability of health MyData can revolutionize healthcare through personalized medicine.

Public-related data

A lot of public-related data—that can be very valuable in the case of a crisis—is already being collected by various public/governmental/or municipal offices. This includes census data, birth and death certificates, and other types of personal and socio-economic data. Typically, such data has been painstakingly collected using paper-based traditional survey methods. In recent times, advances in digital technology has led people to develop mobile phone-based data collection tools that can easily collect, aggregate, and analyze data. Various open source tools such as the Open Data Kit (ODK) (Hartung et al. 2010) make it trivial for such data to be collected. While public-related data is not always publicly accessible, increasingly governments are adopting the Open Data trend, with which public access to public-related data is provided to promote their innovative usage and general accountability and transparency.

The trend of "*open data*" (Manyika 2013), i.e., public sharing of data from various public and private organizations in searchable and machine-readable formats, is a positive step towards data-driven humanitarian development. Governments, e.g., in the USA (Hoffmann 2012; The home of the U.S. Government's open data) and the UK (The home of the U.K. Governmen's open data), are increasingly adopting open data projects to fuel innovation and transparency. The *United Nations Global Pulse* (United Nations Global Pulse) initiative, by the Secretary-General of the UN Ban Ki-moon, has the explicit goal of harnessing big data technology for human development (Pulse 2012). Kirkpatrick, the director of the UN Global Pulse innovation initiative, presents the concept of "data philanthropy" (Kirkpatrick 2013)—companies that gather massive amounts of data can volunteer with the UN so that their data, particularly mobile phone and social media data, can be harnessed to fight humanitarian crises such as hunger and poverty. Here, we note that data philanthropy, although aimed at the general public's wellbeing, does not always imply that a database is made public. Organizations can directly partner with the UN to provide their data for its missions.

Another related example of open data is International Aid Transparency Initiative (IATI) (International Aid Transparency Initiative). This initiative aims to make the flow of aid from the donor to the developing countries and the process of development transparent. Organizations such as NGOs share their data in a standard IATI format (an agreed electronic XML format) (IATI Standard) that can be accessed by other similar organizations to gain insights or to compare their own data with it.

Crowdsourced data

Finally, the method of *crowdsourcing*—in contrast with the aforementioned sources of big crisis data—is an active data collection method. In crowdsourcing, applications actively involve a wide user base to solicit their knowledge about particular topics or events. Crowdsourcing platforms are usually publicly available and are widely used by big crisis data practitioners. Crowdsourcing is different from the traditional *outsourcing*. In crowdsourcing, a task or a job *is* outsourced but not to a designated professional or organization but to general public in the from of an open call (Howe 2006). In this way, crowdsourcing usually combines (a) digital technology, (b) human skills, and (c) human generosity and it is a technique that can be deployed to gather data from various sources such as text messages, social media updates, blogs, etc. Social media, in particular, provide good opportunities for disaster relief efforts that can be planed based on data collected from these sites (Gao et al. 2011). This data can then be harmonized and analyzed in mapping disaster-struck regions and to further enable the commencement of search operations. As described before, this technique helped during the 2010 Haiti earthquake that consequently gave rise to *digital humanitarians* (Meier 2015; 2014).

Big crisis data analytics

First, we describe big data analytics that is a field that combines faculties from computer science, statistics, signal processing, data mining, and machine learning (in particular, techniques for computational linguistics (see Section "AI/ML domain areas relevant for crisis informatics"), geospatial technologies, and visual analytics), to mine insights from big data. The *big crisis data analytics*, broadly speaking, aims to leverage big data analytics

techniques, along with digital platforms (such as mobile phones/Internet), for efficient humanitarian response to different crises. There are many thematic applications of big crisis data including (1) data-driven digital epidemiology (in which public health research is conducted using CDRs and social media) (Salathe et al. 2012); (2) population surveillance and urban analytics (Boulos et al. 2011) (in which big crisis data is used for tracking the movement of crisis-affected population during a crisis); (3) crisis informatics and sociology (Palen et al. 2007) (in which data, along with participatory mapping and crowdsourcing technology, is used for analyzing the sociological behavior of the affected community through behavioral inference and "*reality mining*"—that is data mining to extract and study social patterns of an individual or a group of individuals).

We can also consider the various different *big crisis analytics tasks* that can be performed on big crisis data such as (a) various discovery tasks—e.g., clustering (to obtain natural groupings), outlier detection (to detect any anomalies), or affinity/correlation analysis (to detect co-occurrence patterns); and (b) various predictive tasks—such as classification (in which a category is predicted), regression (in which value of a continuous variable is predicted), and finally recommendation (in which some preference is predicted).

Enabling technologies

After explaining big crisis data, various sources that generate this data and big crisis data analytics, we are in a position to discuss various enabling technologies and techniques that either generate data or help in analyzing it. These technologies are shaping today's age in which data driven crisis response and humanitarian development are being considered the future of human progress and wellbeing. Table 1 can be consulted to gain an insight into the relation between different big crisis data sources (discussed in the previous section) and corresponding enabling technologies and example scopes where (big data driven) actionable knowledge can be applied.

Table 1 Big crisis data sources and enabling technologies

Enabling technologies	*Example(s)*	*Example scope(s)*
Data source: data exhaust		
Mobile phones, machine learning	Call detail records (CDRs)	Epidemiology, social network analysis
Data source: online activity		
The internet, machine learning	User-generated data (e.g., emails, comments, search engine and social media activity)	Sentiment analysis, opinion mining, search and rescue
Data source: sensing technologies		
IoT, data visualization, machine learning	Satellites, UAVs, sensor networks, wearable devices	Social behavior and enviornmental analysis
Data source: sensing technologies		
IoT, the internet, machine learning	Wearable devices, social media	Healthcare (e.g., personalized medicine)
Data source: public data		
Open source and open data, machine learning	Governmental data (e.g., census, public health and transportation data)	Effective policy making
Data source: crowdsourced data		
Crowdsourcing, Data visualization, neo-geography, machine learning	Crowdsourcing platforms (e.g, Ushahidi)	Crisis maps

Mobile phones

The rapid adoption of mobile technology has been unprecedented. With almost universal service of mobile phone technology[2], smartphones are rapidly becoming the central computer and communication devices in the lives of people around the world. Modern phones are not restricted to only making and receiving calls—current off-the-shelf smartphones can be used to detect, among other things, physical activity (via accelerometers); the speech and auditory context (via microphones); location (via GPS) and co-location with others (via Bluetooth and GPS) (Lane et al. 2010). This transforms the modern crisis response since modern smartphones can now act as general-purpose sensors and individuals can directly engage in the disaster response activities using some combination of cloud-, crowd-, and SMS-based technologies. This participatory trend in which the aid efforts are driven and centered on people—and the fact that aid workers have to work with large amounts of diverse data—makes modern disaster response completely different from traditional approaches.

Interestingly, mobile phones are as widely deployed in underdeveloped countries (which are often more prone to various kinds of crises) as in developed countries. The mass penetration of mobile phones, networked through communication technology, has led to vast amounts of data. CDRs are a type of mobile-based data collected by mobile telecom service providers for billing and troubleshooting purposes (e.g., by determining which cells are more congested). A CDR contains information such as

the time at which a call/message was received, as well as the identity of the cellular tower with whose antennas the phone was associated with—using the information contained in the CDR, the approximate locations of the users can be easily determined. The use of CDR for localization can help the study of mobility, epidemiology, and population dynamics at an unprecedented scale with very low cost. CDRs are attractive for mobile data analytics since (1) they involve negligible marginal cost (as the telecom companies are already collecting them); (2) they provide real-time information; (3) they are already working at a very large scale (as all the calls received and sent are being logged). The limitation of CDRs are that (1) the calls are recorded when they are made thus making the data sparse in time; (2) they can provide coarse-grained spatial granularity.

Let us explain the importance and applicability of CDRs with the help of an example scenario. As it usually happens that large disasters are often followed by large population movement. This complicates the task of humanitarian relief (in which there is the need to get the aid to the right people, at the right place and at the right time). CDRs have been used by digital humanitarians during various crises (such as the non-profit *FlowMinder*'s (Flowminder) work with anonymous mobile operator data during the Haiti earthquake to follow the massive population displacements) to not only point out the current locations of populations, but also predict their future trajectory (Bengtsson et al. 2011). FlowMinder used the same mobile analytics methodology during the cholera epidemic later that year to enable rapid monitoring of population movement from the outbreak area (Bengtsson et al. 2015).

Mobile phones are also excellent tools for gathering survey data (and can cheaply conduct socio-economic and health surveys). Mobile phone-based data collection can reverse the trend established in traditional humanitarian organizations on relying on paper documents. To facilitate such mobile-based surveys, various ;tools have been developed—such as ODK (Hartung et al. 2010), which is a free and open-source set of tools that provide an out-of-the-box solutions for building, collecting, and aggregating data. ODK can help organizations author, field, and manage mobile data collection solutions. Mobile phones can also be used for intelligence gathering using the crowdsourcing technique. In addition, mobile-based analytics can also act as surrogates for official statistics (e.g., on public health and food security) when not available or when more real-time information is needed.

With the all abovementioned benefits provided by the data collected from mobile phones, it is also of utmost importance to consider the potential harms and challenges that ensue this process. Taylor in (Taylor 2014) mentions that the analysis of data gathered from mobile phones has two dimensions: virtual and physical. The virtual dimension involves the interpretation of data whereas the physical dimension implies the monitoring, governing, and controlling of population. According to Taylor, a global standard to ensure sufficient privacy is lacking in mobile data analysis. Mostly individual entities are self-regulating themselves with custom made anonymization techniques which according to him is not enough. The author stresses on creating a balance between the *right to be visible* (i.e., in gaining benefits from the analysis performed on their mobile data) and *right to be hidden* (i.e., where there is minimum risk of the people's privacy and freedom being compromised). Another important argument raised by the author in (Taylor 2014) is that there should be an (incentivised) environment where social scientists could work in unison with the data scientists so that a standard framework of privacy for data analysis can be ultimately realized. As an example, McDonald in (McDonald 2016) also emphasizes the need for a standard framework to ensure the privacy of already vulnerable population of a crisis struck region for whom data is gathered and analyzed such as during epidemic breakouts like Ebola. McDonald describes that a lack of legal framework renders the data analysis process for a crisis incongruent as different aid organizations could react to a problem in non-organized and uncoordinated manner. There is also a risk of legal lawsuits against governments and NGOs that expropriate data of a region, specially CDRs that are rich with personal information despite the anonymization processes, whereas the local constitutional orders forbid doing so. This was what happened in Liberia during the Ebola outbreak in which organizations could not gain access in a fast enough and legal way which resulted in more spreading of the disease.

The internet, open source, and open data

These days the ubiquitous proliferation of the Internet connectivity is enabling the new Internet-based economy that is spawning a paradigm shift in terms of how institutions operate. The "*open source*" culture (the paradigm underlying projects such as Linux and Wikipedia) that has been popular on the Internet has ushered in a new era that relies more on collaboration and volunteerism (than on formal organizations). Instead of a rigid structure constrained by scarcity of human resources, the new paradigm is driven by abundance and cognitive surplus (due to technology-driven pooling of volunteer human resources). The open source trend is also visible in humanitarian development in various manifestations such as digital humanitarianism, user-generated information, participatory mapping, volunteered geographic information, open source software (such as *OpenStreetMap* (Haklay and Weber 2008)), and open data. Many volunteer and technical communities (also called collectively as V&TCs) have exploited these open standards to link

data from disparate sources and create a *mashup* (which is defined as a web page/application that uses or combines data or functionality from multiple existing sources to create new services) (Initiative et al. 2010).

Another important trend emerging in the modern era is the unprecedented commoditization and opening up of data. A number of countries from all over the world (more than 40 countries in 2015) have established "open data" initiatives to open all kinds of datasets to public to promote innovative use of such data by volunteer data scientists and thereby lead to more transparency, less corruption, and economic development (Sanou 2013). Open data can also lead to improved governance through the involvement of public and better collaboration between public and private organizations. As an example, in the aftermath of the Haiti crisis, volunteers across the world cobbled together data from various sources—including data from satellite maps and mobile companies along with information about health facilities from the maps of the World Health Organization, and police facilities from the Pacific Disaster Center—and plotted them on open source platforms such as the OpenStreetMap. The importance of OpenStreetMap can be gauged from the fact that data from it eventually became the most reliable information source that was widely used by both governments and private aid workers for supplies to hospitals, triage centers, and refugee camps.

The real power of the open data will start to emerge when the data is exchanged in standardized formats through which we can link, merge, and repurpose data according to our needs. The term *Linked Data* is used to refer to the recommended best practices for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using standard data sharing formats such as uniform resource identifier (URI) and resource description framework (RDF) (Bizer et al. 2009; Heath and Bizer 2011). Linked Data has been envisioned as an essential component of the vision of Semantic Web (pioneered by Berners-Lee, the inventor of the World Wide Web). With streaming sensing devices becoming an important source of big crisis data, Linked Stream Data can be created by integrating the sensed data with data from other sources by utilizing their semantic descriptions. Despite its enormous potential, there is little exploration Linked Stream Data, which can be consider as a potential open issue for the future work.

Visual analytics and neo-geography

Visual analytics is an interesting branch of big data exploration that borrows techniques from computer vision and geospatial technologies which can be used to deal with crisis images and videos. The aim is to support analytical reasoning through visual interfaces. With geographic data being critical for disaster response and sustainable development, the ability to visualize and map geographic data becomes very important. Neo-geography ("new geography") prominently uses the Web 2.0's democratic approach of participatory mapping along with crowdsourced user-generated volunteered geographic content (Turner 2006). Unlike conventional maps, with neo-cartography (i.e., "new cartography"), the data is not controlled by an elite few. A concrete neogeography/neocartography example is the participatory mapping approach taken by the free and open source software OpenStreetMap, which is based on geographical data collected from volunteers. The importance of OpenStreetMap can be ascertained from the fact that soon after the earthquake, OpenStreetMap had become the de facto source of data to map the crisis afflicted parts of Haiti for most of the UN agencies (Initiative et al. 2010). A related project, which aims to support the operation of OpenStreetMap, is Missing Maps (Missing Maps). This project focuses on the developing world and aims to map the most vulnerable places of such regions. It provides a good resource for the NGOs and other aid agencies to get the relevant map-related information to steer their aid activities in the right direction. The novel trends of newgeography and neocatrography have encouraged the development of the field of *"crisis mapping"* which we discuss next.

Crisis mapping

Crisis mapping (Ziemke 2012) is the newly developed interdisciplinary research field. The crisis maps contain up-to-date satellite imagery that are often collected and annotated using crowdcomputing and crowdsourcing techniques (that are discussed later). The important feature of these maps are that they are real-time source of information and provide effective situational awareness with respect to spatial and temporal dimensions by giving a dynamic bird's-eye view to guide the response efforts. The process of crisis mapping is overseen by dedicated organizations such as the International Network of Crisis Mappers (CrisisMappers 2016b) (an organization, co-founded by Patrick Meier and Jen Ziemke in 2009, which has more than 6000 members worldwide) and managed through Crisis Mappers Humanitarian Technology Network; crisis mapping even has its own international conference (the International Conference of Crisis Mappers, ICCM).

We have already seen example of how Ushahidi was used in the Haiti crisis. To demonstrate the versatility of crowd mapping, we will briefly describe how Ushahidi—only one tool of a growing big crisis data mapping ecosystem—has been used for a number of diverse tasks such as (1) by Al Jazeera for gathering mobile reports from citizen journalists in Uganda to uncover human rights violations and amplify local voices; (2) by the United Nations

to improve peacekeeping operations through collection of mapped reports from field staff and local community; (3) by PakReport.org (Chohan et al. 2012; Pakreport 2016) to support Pakistanis in collecting, mapping, and sharing crisis information during the floods in 2010.

Another example of crisis mapping is *HealthMap* (HealthMap) that combines data from different sources to create a unified and comprehensive picture of the current global state of infectious diseases. Healthmap, which is accessible on the web and on mobile devices, utilizes a number of informal online sources for disease outbreak monitoring and real-time surveillance of emerging public health threats. The HealthMap is widely used, with the World Health Organization (WHO) (Heymann et al. 2001; Wilson and Brownstein 2009) and regional groups like the European Center for Disease Prevention and Control being prominent users who utilize HealthMap for digital disease detection.

Satellites are also used to capture high-resolution imagery of ground locations to provide near-real-time situational awareness to the concerned authorities about a potential crisis. As an example, the Harvard Humanitarian Initiative (HHI) in 2010 monitored and analyzed the visual imagery captured by the Satellite Sentinel Project (SSP). The HHI analysts monitored the precarious situation related to mass atrocities along the border of Sudan and South Sudan and provided a real-time public reports of the situation to be used by the government and aid agencies. The HHI team in (Raymond et al. 2013) summarizes its experience with the SSP project. The HHI team concludes in this report that although the SSP project provided novel ways of collecting information related to conflict zones, the overall success of this endeavor was not a significant thing. The overall analysis of the HHI reports were used in a biased fashion and moreover by making all the reports public inflicted the risk this information being exploited by the armed groups involved in the conflict. The authors of (Raymond et al. 2013) emphasize the need for a standard framework that can be deployed on global basis in which sensing technologies like SSP project can be used in a secure and responsible manner (Think Brief: Building data responsibility into humanitarian action 2016).

Leveraging the wisdom and the generosity of the crowd

Broadly speaking, there are only a few ways we can go about problem solving or predicting something: (1) experts, (2) crowds, and (3) machines (working on algorithms; or learning from data). While experts possess valuable experiences and insights, they may also suffer from biases. The benefit of crowds accrues from its diversity: it is typically the case that due to a phenomenon known as "the wisdom of the crowds" (Surowiecki 2005), the collective opinion of a group of diverse individuals is better than, or at least as good as, the opinion of experts. Crowds can be useful in disaster response in at least two different ways: firstly, crowdsourcing, in which disaster data is gathered from a broad set of users and locations (Boulos et al. 2011); and secondly, crowdcomputing, in which crowds help process and analyze the data so that it may be effectively displayed and thereafter utilized (Meier 2014).

Crowdsourcing

Crowdsourcing (as discussed in detail in Section "Crowdsourced data") is the outsourcing of a job traditionally performed by a designated agent (usually an employee) to an undefined—generally large group of people—in the form of an open call. In essence, crowdsourcing is the use of the open source principles to fields outside of software. Crowdsourcing has been used in the context of disaster response in multiple ways (Munro 2013): including crowd searching, microtasking (explained next), citizen science, rapid translation, data cleaning and verification, developing ML classifiers, and election monitoring (Boulos et al. 2011).

Crowdcomputing (crowdsearching and microtasking)

Crowdcomputing is a technique that utilizes crowds for solving complex problems. A notable early use of crowdcomputing was the use of crowdsearching (a subproblem of crowdcomputing where crowds participate in a search problem) by MIT's team at the DARPA's "I Spy a Red Balloon" network challenge (Tang et al. 2011). The task was related to the geo-location problem in which the participants were to find 10 red weather balloons that were at secret locations across the continental US The MIT's team solved a time-critical problem in the least time using an integration of social networking, the Internet, and some clever incentives to foment crowd collaboration.

We note that microtasking works by dividing a bigger task into a series of smaller tasks that can be completed in parallel (usually by many people over the Internet). In contemporary times, a number of "*microtasking*" platforms have emerged as smarter ways of crowdsearching. A familiar example is the commercial microtasking platform "Amazon Mechanical Turk" run by Amazon that allows users to submit tasks for a larger job (that is too large for a single person or small team to perform) to be distributed to a crowd of a large number of volunteer global workers who are remunerated in return for performing these microtasks. Apart from commercial solutions such as the Amazon Mechanical Turk (Amazon Mechanical Turk 2016), a number of free and open-source microtasking platforms such as Crowdcrafting (Crowdcrafting 2016c)—which was used by Digital Humanitarian Network (DHN) (Digital Humanitarian Network) volunteers in response

to Typhoon Pablo in the Philippines—have been developed. However, since Crowdcrafting was not optimized for humanitarian aid, the big crisis data community has developed humanitarian-aid-focused microtasking tools, which we will discuss next.

MicroMappers 2016, developed at the Qatar Computing Research Institute (QCRI), was conceived as a fully customized microtasking platform for humanitarian response—a platform that would be on standby and available within minutes of the DHN being activated. MicroMappers is a platform for tagging the mass of online user-generated multimedia content (in various formats such as text, images, videos) related to a disaster to establish its importance. Micromappers performs this tagging operation through volunteer digital humanitarians who leverage a series of microtasking apps called Clickers to focus on different tasks. As an example, the *TranslateClicker*, *VideoClicker*, *ImageClicker*, and *TweetClicker* focus on the microtasks of translating, classifications of videos, images, and Twitter tweets.

Artificial intelligence and machine learning

With humans and experts, the bottleneck is human nature (which is fickle, humans can get bored and make mistakes). Machines, on the other hand, do not suffer from boredom and can work on menial algorithmic tasks with high throughput and no error. Machines can utilize well-crafted algorithms for various problems, whose solutions can be exactly defined step by step. However, for many (mostly non-trivial) problems, exact algorithms are not known; for such problems, the machine learning approach—which is based on learning from data—is appropriate. In humanitarian technologies, there is a need to combine the wisdom of the crowd with the efficiency and automation of artificial intelligence (AI) and machine learning (ML). While crowdsourced technologies, such as crowdcomputing, will continue to play a big role in processing big crisis data, but crowdcomputing alone will not be enough—the efficiency of the crisis response can be significantly enhanced if the crowds are also complemented by the machines. AI/ML techniques are especially important since they can lead to automated processing and extracting useful information from big crisis data (and thereby compensate for the lack of human experts). Broadly speaking, we can categorize ML algorithms into two types: *supervised* and *unsupervised*.

Supervised learning

It is a subset of ML algorithms, that uses techniques to infer a function (or generalize a relationship between the input and the output) from the given training data that contains labeled examples of the data and the corresponding correct category. A number of supervised learning techniques exist including the popular and easy to use techniques of support vector machines (SVM) and random forests. Supervised learning works very well when labeled data is available. However, it fails to scale to different settings (in terms of the type of crisis and socio-geographic setting)—since the classifier generated by the ML algorithm is tightly coupled with the data being used and cannot be reused when the setting is changed. The need of a supervisor (or a human who will produce the labeled data) limits the use of supervised ML.

Unsupervised learning

Unsupervised learning, on the other hand, refers to the set of ML techniques that do not depend on the availability of training data; such techniques can automatically find hidden structure in unlabeled data. Clustering is an example of unsupervised learning. Clustering differs from classification in that classification aims to assign data instances to predefined classes, while clustering aims to group-related data instances together without labeling them. Common examples of unsupervised learning techniques are the K-means and expectation-maximization (EM) clustering algorithms. Unsupervised learning can also be used for automatically learning features for the data. Unsupervised learning is becoming more important in the big data era to alleviate the inconvenience of handcrafting features and providing labeled examples.

In the perspective of unsupervised learning, the new and emerging field of deep learning holds a lot of promise. Deep learning deals with deep and complex architectures (Bengio 2009; Schmidhuber 2015). There are multiple processing layers in these architectures. Each layer generates non-linear response against data that is fed into it as an input. The data given as input is processed by several small processors that run in parallel. These processors are termed as *neurons*. Deep learning has shown promise to be efficient in pattern recognition, image, and natural language processing (Collobert and Weston 2008). Deep learning applications range from healthcare to the fashion industry (Knight 2015). Many technological giants such as Google, IBM, and Facebook are deploying deep learning techniques to create intelligent products.

AI/ML domain areas relevant for crisis informatics

The two most important AI/ML application domains that are relevant for crisis informatics are computational linguistics and computer vision, which we discuss below:

Computational linguistics Computational linguistics deals with the application of intelligent computational techniques and algorithms to gain insights from the (written or spoken) human speech. The most important application of AI-based computational linguistics for big crisis data analytics is automated analysis of social media

using sentiment analysis and *opinion mining* (Pang and Lee 2008). While there have been many advances in the field of natural language processing (NLP), applying NLP for crisis informatics is still non-trivial. Computational linguistics based AI techniques have been utilized extensively in humanitarian responses to many previous crises. As an example, the *CrowdFlower* platform (CrowdFlower 2016d), which provides services for crowdsourced translation, has been used in the response to the Haiti crisis (by the Mission 4636 project (Mission 4636)) and to the Pakistani flood crisis (by the PakReport project (Chohan et al. 2012; Pakreport 2016)). There is a lot of work needed and research scope for applied big crisis data computational linguistics. One problem is that computational linguistics systems optimized for one language or mode of communication do not generalize to others. For example, it has been reported that systems optimized for one type of short message service (Twitter or SMS) may not work well for the other. In addition, systems optimal for one language are unlikely to be optimal for other languages (Munro and Manning 2012).

Computer vision There is a lot of interest in using AI-based techniques from the field of computer vision for automated analysis using aerial images (via UAVs or satellites). This interest has received a further fillip through the "Imagery to the Crowd" initiative by the Humanitarian Information Unit at the US State Department (Map Give: imagery to the crowd)—this initiative aims to democratize access to high-resolution commercial satellite images purchased by the US government in a format that public volunteers can easily map into an open platform such as the OpenStreetMap. There are various projects focused on automated analysis of global high-resolution satellite imagery (e.g., projects on this topic at the European Commission's Joint Research Center (JRC)).

Tomnod (Tomnod) (owned by the satellite company DigitalGlobe 2016) and MicroMappers 2016 are microtasking platforms that can be used for computer vision applications. Unlike MicroMappers, which handles multiple media (such as text, images, video), Tomnod is exclusively focused on microtasking satellite imagery. Tomnod provides an ability to slice and dice the satellite imagery into (potentially millions of) very small micro-images. To demonstrate the value of a satellite image microtasker, consider the use of Tomnod for tracking the missing Malaysian Airlines flight 370 incident in 2014. Within days of the launch of the project on Tomnod, the crowdsearching participants had swelled to 8 million volunteers, who had tagged more than 15 million features of interest in 1 billion satellite images in just four days. Tomnod then used an indigenous CrowdRank algorithm for deciding which of the tagged features to escalate to on-the-ground rescue times through DigitalGlobe (by looking at the consensus of the taggers that have the highest levels of crowd consensus.)

Interfacing human and artificial intelligence

Previously, we have seen that analytics can be performed through exploiting three sources of intelligence: humans, crowds of humans, and machines (that are either programmed by humans or learn from data). The most potent approach to big crisis data analytics is to leverage both human and machine intelligence. Crowds are better at finding patterns while machines can process faster than humans (and even crowds). This hybrid approach can exploit the complementary competencies of humans or crowds and machines: in particular, with such a hybrid approach both the pace and accuracy of crisis analytics can be improved.

This implies that in humanitarian aid, AI/ML algorithms will not obviate the need of human involvement. AI/ML can be used to ensure that valuable human resources are optimally utilized: this is done by outsourcing the onerous grunt work to the machines and taking the onus off the human beings for routine things (that the machines can do more efficiently). Human beings can then focus their time on assisting algorithms in codifying unstructured information and recognizing unusual patterns. Also, people from diverse set of fields can get involved in the analysis of visual reports—which are, let us say, produced after an ML algorithmic operation finishes running on a dataset. In this manner, the technological savvy personnel will only be a part of a larger team that may comprise people from as diverse fields as healthcare, public policy making, and government agencies etc.

A promising approach to bridging crowdsourcing and ML, which has been recently used by humanitarian organizations, is to outsource the labeling task and thereby the classifier generation to crowds (using microtasking-based crowdcomputing approaches). This is the underlying principle behind the AI for disaster response (AIDR) platform (Imran et al. 2014), which seeks to crowdsource ML classifiers by providing a platform that makes it trivial to generate new ML classifiers. AIDR can be used to automatically classify tweets (after it has been taught initially through the building of the classifier). Since with automated machine learning, the possibility of erroneous classification is present, AIDR also provides a confidence level with its auto-tags (which nicely allows integration of human and machine intelligence; e.g., by bringing the human into the loop when the machine's confidence level is below, let us assume, 70 %). We note here that AI-based platform is certainly not limited to automatic analysis of only tweets: AI/ML can be used for all kinds of media including, most notably, short messages (such as SMS), and images (such as satellite/UAV images, as well as web-based images).

The need for such a hybrid technology (combining human and machine learning) is stronger in areas such as computational linguistics and NLP in which the data has vague semantics (Munro et al. 2012). While NLP is better done through ML/AI, hybrid ML systems in which human feedback is incorporated currently perform much better than pure ML-based NLP systems. A grand challenge for the AI-based NLP research is that the developed tools do not generalize to different settings (in terms of languages, modality of data, and the type of crisis). The current way of building specialized tools is both expensive (in time, money, and resources) as well as non-scalable (since there is no reuse of existing software). With the lack of total automation through current AI tools, the hybridization of ML and humans is especially needed when humanitarian workers are working in less resourced languages.

The big crisis data ecosystem and case studies

The humanitarian response ecosystem principally includes two kinds of actors. Firstly, traditional humanitarian groups, which include groups such as the International Committee of the Red Cross (ICRC), United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA), United Nations High Commissioner for Refugees (UNHCR), as well as non-government organizations (NGOs) like Doctors Without Borders (also known as MSF), World Vision, and Care International etc. Secondly, digital humanitarian groups, exemplified by the DHN —which was activated by the UNOCHA—whose aim is to empower a technology-enhanced 21st century humanitarian response; this is achieved by forming a consortium of volunteer and technical communities (called V&TCs) (Initiative et al. 2010), which then interface with formal humanitarian organizations to aid their response.

More specifically, the digital humanitarians can assist in coping with the big crisis data overload by tracking media (e.g., by monitoring social networks); software development for producing programs to organize and streamline crisis response; performing statistical analysis on data; mapping data; and translating local crowd media so that its accessible more broadly. A number of V&TCs exists such as Translators Without Borders (Translators without borders), Geeks Without Bounds, Datakind and ESRI etc. In addition, a number of technology-focused organizations (such as the SBTF 2016, OpenStreetMap 2008, GISCorps, CrowdFlower 2016d, Crisis Mappers 2016b, Crisis Commons 2016a, MapAction) have sprouted up that utilize big crisis data techniques for remotely accelerating humanitarian assistance through a hybrid of technology, collaboration, and volunteerism.

Also relevant to big crisis data analytics are the many *big data for development* initiatives that focus on mining, understanding, and harnessing data for improving the world in both crisis settings (e.g., by providing data-driven assistance) and non-crisis settings (e.g., for battling hunger and poverty). Such initiatives overlap with the digital humanitarian initiatives and include (1) social computing groups (such as the UN Global Pulse, and QCRI); (2) groups initiated by for-profit companies (such as Google.org—through it's Google Crisis Response, Google People finder, and Google Crisis Map projects, and Microsoft Research); and (3) academic humanitarian initiatives (the MIT Humanitarian Response Lab, and, Harvard Humanitarian Initiative and Peace Informatics Lab in The Hague, Netherlands); and (4) non-profits such as DataKind (formerly, Data without Borders).

We now discuss two case studies: the first case study is about the ongoing migrant crisis that is caused by civil war in Syria. Secondly, we discuss the application of big data in the filed of healthcare. These case studies highlight the important positive role that big data plays in mass emergencies and humanitarian wellbeing and shed light on the importance of the development of an effective big data-driven crisis response ecosystem.

Case study: migrant crisis

Here, consider a case study of the recent migrant crisis—brought squarely in the world's limelight by the case of Aylan Kurdi. Aylan, a three-year old Syrian toddler, was tragically found lying face down on a Turkish beach in 2015. He was washed ashore lifeless after he, led by his mother and brother, had attempted in desperation to reach the Greek island of Kos in a dinghy. The heart-wrenching image of Aylan, which rocketed around the world through social media, galvanized considerable public attention to the refugee crisis, highlighting the extraordinary risks migrants take to escape various crises. Aylan was able to attach a human face to the problem of migrant crisis, helping vitalize the international response to the problem. Around the world, there are millions of refugees (including many kids like Aylan) taking comparable risk to flee from war/violence to safer pastures. The various stakeholders around the world are looking towards big data technology to play an important role in thwarting human tragedies.

Aleppo, the industrial capital of Syria, is one of the most affected cities in the conflict. Neighborhoods all over the city are divided among, and under the control of, Syrian government, militia and rebel parties actively engaged in fighting with one another. This act has caused many civilian casualties and displacement. Before any response, it is imperative for the aid agencies to get the sense of the dynamics of such a region. If the conflict is *mapped* properly, only then the right aid can flow into the right regions in time. Caerus Associates, a research consultancy entity, with the technological help from First Mile Geo, which provides online data collection, analysis and

sharing services, mapped the dynamics of the conflict in Aleppo in 2014 (Caerus 2014; Mapping the conflict in Aleppo 2014). Caerus deployed a team of locals, equipped with tools connected to cloud storage services, to survey local residents of the city of Aleppo over the period of 4 months (September 2013 to January 2014). This survey included the opinion of people about the government and rebel forces, position of checkpoints (established by government and rebels at different roads of the city), status of bakeries and the status of a neighborhood (in terms of which entity controls it). All the data was analyzed through First Mile Geo and a *hyper-local time-series* map of the region came out as a result. This map showed the people's opinion about who is right in their claims in this conflict (to which a vast majority said "no one"), strongholds of rebel forces (which spread in the eastern side of the city) and government (largely in the western side of the city). More details on this project, and the related interactive maps can be seen at the link provided in reference (Caerus 2014), the data collected for this project is also available to download for research purposes.

This information provides the status of the political situation of the city. The information about the checkpoints tells about the public movement, which is strictly controlled and restricted through these checkpoints. During the conflict, the government forces particularly attacked the bakeries in the rebel-held areas. This resulted in food shortages. People in the rebel-held area had to resort to indirect means and routes to obtain bread to survive. The information related to bakeries combined with that related to checkpoints let the aid agencies know about possible areas where there is, or will be a, food shortage.

May people are displaced, from all over Syria, who are forced to leave their country in search of immediate shelter and assistance. The most likely target for the refugees to go to are the countries in the immediate neighborhood of Syria. Jordan (Syria's neighbor in its South) hosts more than half a million Syrian refugees. It is critical that the right and real-time information about the location, type, and capacity of aid agencies are communicated to the refugees. Such type of information is important for the humanitarian organizations as well. Through this information each agency could know what other agency is working on so that double response to the same problem can be avoided. About 60 such agencies have been in Jordan to assist the incoming Syrian refugees.

Besides Jordan, a large number of people are traveling to the European countries often through dangerous land and marine routes. Many go missing, and unfortunately a few also die during this process. It becomes very difficult to reconnect the separated people of a family. Reconnecting is especially rendered difficult, as the displaced people mostly remain mobile, moving from one place/camp to another. There are also delays and language barriers in registration processes. To overcome this difficultly, at least to some extent, an initiative taken by the Red Cross society by the name of *Trace the Face* (Trace the Face). This is a website that collects data in the form of photographs and publishes photos of the people who are looking for their families and the missing people whose families are looking for them. Besides being published online, these pictures are printed on posters that are hung all over Europe especially in places where the migrants gather. In this way, the broken family links are restored through *crowdsourcing*.

Case study: healthcare

We now highlight the importance of big data and the establishment of a data driven healthcare ecosystem. As it usually happens that a patient might be seeing different specialists for, seemingly, different medical reasons. These specialists, then further, can prescribe different types of clinical tests resulting in different kinds of results. If, however, some protocol or a system is developed to integrate these fragments together and run analysis on them collectively then a clear and big picture of a patient's current health can be extracted. If the issue of fragmentation (Elhauge 2010) (i.e., data having heterogenous formats and residing at potentially disconnected silos) is resolved then this can not only speed up the diagnostic process but also has the ability to provide personalized treatment (realization and the application of small data and MyData—discussed in Section "Small data and MyData") that is most suitable for the patient under consideration. It is also important to address the issue of fragmentation for different NGOs and field health workers responding to a crisis. So that data collected by different organizations or teams can be corroborated to avoid double response in which two entities redundantly respond to the same problem.

An *Individualized Health Initiative* (Duffy 2015) is taken at Johns Hopkins University aimed at collecting, integrating, analyzing and sharing patients' data among healthcare providers for better and more informed decision making. Collecting data from a large population having a same disease helps in analyzing and building computational models through which a new patient with similar early symptoms can be categorized more accurately and put on the right track of medication at early stages of an ailment. An example of this process can be the treatment of prostate cancer (Duffy 2015). In the usual course of medical treatment, most of the patients are put on a similar regimen without properly analyzing risk factor of the disease. As it can happen that a person has a tumor that might not grow big enough during the course of his lifetime that would further endanger his life. Still such a patient is passed through the rigors of the cancer treatment, suffering many unnecessary side affects and mental agitation. This is what doctors

call an *overdiagnosis*. Imagine a situation where there is a lot of related data of many similar cases in a database and a computational model is present that can categorize a patient and predict a potential trajectory that his current state of disease will follow. In this way, a lot of hassle can be avoided and a personalized course of treatment can be prescribed. This approach eliminates the guess work, that is solely performed based on a medical practitioner's memory and replaces it with computational reports crafted after integrating huge amounts of data from various related sources that range from pathological clinical tests to diagnoses from different medical experts.

There is a healthy aspect of *democratization* of health data (Lohr 2015), which implies that by combining medical and the data science, the study of disease has the potential to flourish. This can lead to a coverage that involves a whole country to construct a *Learning Health System (LHS)* (Friedman et al. 2014) in which stakeholders from government, healthcare, engineering and technology are brought together to analyze the prevailing health situation/emergency, in a country, more accurately and fight diseases/epidemics fast. The creation of such ecosystem will have the capability to establish an environment in which research and clinical practice are not performed in isolation; instead new research and analysis are directly applied to patients in a near real-time manner. If we expand this notion, then the whole world can be monitored and the present state of any of the world's countries health can be analyzed and the early warning signs of imminent viral epidemic outbreaks can be predicted and/or detected.

IBM's Watson has been developed by the IBM as the realization of a fully advanced cognitive computing system. Watson, as opposed to the current AI-based computing mechanisms, does not look for key-words in a line of text to make decisions. Watson combines machine learning, natural language processing, and statistical analysis (IBM 2015) to glean *context-based* information from all the data that is fed into it. One of the most important applications of Watson is expertise scaling and democratization, especially benefitting the healthcare field. It manages information and bridges it among the experts of a field helping them to make more informed and relevant decisions quickly.

Why big crisis data analytics is challenging?

The "Vexing Vs"

The technical challenges associated with processing big data have traditionally been summarized using the four Vs: (1) *volume* (large amounts of data that cannot be processed on a single machine or with traditional database tools); (2) *variety* (data in many formats: structured, semi-structured, and unstructured—with mostly the data being unstructured); (3) *velocity* (streaming data, with milliseconds to seconds to respond); and (4) *veracity* (uncertainty in the data being "true" or consistent with reality). With big crisis data (such as social media), the list of "vexing Vs" is even longer: we also have to deal with (5) *vagueness* (dealing with natural language); (6) *virality* (understanding cascading crisis information); (7) *volunteers* (motivating and coordinating digital volunteers); (8) *validity* (mitigating the biases and pitfalls of social media); (9) *values* (ensuring privacy and ethical use of crisis data); and finally, (10) *visualization* (how to best visualize big crisis data such as crisis maps). We will cover some of these vexing Vs next; for the sake of brevity, we focus only on a subset of these challenges.

Volume and variety

The sheer volume of big crisis data makes its analysis very challenging. Just like the absence of information can debilitate rescue efforts, the overflow of information (as seen in voluminous big crisis data) can derail crisis response efforts. Humanitarians working with this tsunami of data face a difficult situation: their efficiency can benefit from the crisis data, but the high volume makes big data difficult to tame—humanitarians working with big data are akin to a thirsty person drinking from a fire hose. The volume can be overwhelming since the data is not being provided only by the hordes of ordinary citizens, but also from other sources (such as government and journalists). As an example of voluminous big crisis data, there were more than 20 million tweets with the hashtag of Sandy within less than a week about the deadly Hurricane Sandy that swept through the Caribbean and the East Coast of the USA (Twitter).

Big crisis data is highly diverse and varied. Big crisis data comprises social media, data exhaust (such as CDRs), satellite and aerial imagery, crowdsourced maps, and more generally any information shared by on-the-ground crisis-hit community as well as news/reports shared by journalists and aid organizations. Apart from these sources, big crisis data also needs to integrate surveys of various kinds (such as land, environmental, population), and models (such as climate, environmental, geomorphological, disease, and all kinds of hazard models). Apart from the well-known fact that data can be unstructured or structured, another diversity related problem unique to emergency-response is the diversity of languages—which can complicate NLP-based crisis informatics tasks since English-based NLP solutions do not always generalize to NLP problems in other less resourced languages. For example, researchers have noted that the tropics regions that contribute to 90 % of the pathogens that lead to epidemics also have immense linguistic diversity (the languages in these areas comprise 90 % of the world's languages).

Veracity, verification, and validity

The problem of data veracity addresses the need to verify humanitarian/crisis data, so that the credibility of data sources is established (e.g., by verifying metadata—which is data about data, that describes the content and context—to ensure that data is reliable) and humanitarian efforts are not derailed through the spreading of incorrect or stale information. Verification and validation are important since it has been noted that citizens have a tendency to exaggerate under extreme stress. For verification and validation, standard verification techniques that have been developed for information verification for humanitarian purposes should be followed (Silverman 2013). As an example, information coming from several independent sources (e.g., through social media and traditional media) may be considered more reliable than information from dependent sources (e.g., retweets of the same original tweet). In addition, we note that the collection of metadata and data provenance is a challenge when data is being collected in stressful situations by a large number of actors (but such data is necessary for appropriately dealing with heterogeneous big data). Existing verification projects have mostly combined crowdsourcing with ML to build verification projects. The *Verily* project (Verily) crowdsources the task of verification by seeking rapid crowdsource evidence (in terms of an affirmative or negative answer) to answer verification questions. The *TweetCred* is a web- and ML-based plug-in that can be used to automatically detect non-credible tweets and fake images being shared on Twitter in real time (Gupta et al. 2014).

The data collected can also be false in terms of poor quality, malice, or bias. The presence of false data dilutes the signal to noise ratio, making the task of finding the right information at the right time even more challenging. As an example of the inherent bias in big data, we note that the Google Flu Tracker overestimated the size of the influenza pandemic by 50 %, miscalculating the severity of the 2013 flu and predicting double the amount of flu-related doctor visits (Butler 2013). As an example of quality problems, we refer to (Meier 2014) in which a study focusing on social media content relating to the Boston bombing in 2013 is cited; this study (which was based on an analysis of 8 million unique tweets sent in 2013) showed that only 20 % of the tweets that went viral relayed accurate information; the remaining tweets either propagated fake content (30 % of the viral tweets) or offered only general comments/opinions (50 % of the overall tweets).

This problem noisy data is particularly problematic for crowdsourced data in which the noise may be injected intentionally or unintentionally. Intentional sources of noise may come from pranksters or more sinisterly through cyber-attacks (this is particularly a risk during man-inflicted disasters, such as coordinated terrorist attacks). Unintentional sources of noise also creep into disaster data (e.g., through the spreading of false rumors on social networks; or through the circulation of stale information about some time-critical matter).

It is therefore in order to rely on big data-based analysis with caution. A promising technique for addressing the signal to noise problem in big data is to use human microtasking (for increasing the signal to noise ratio) so that the amateur volunteers can reduce the noise while the experts can focus on following the potential signals more closely. It is also a good idea to verify the output of crisis analytics with plain-old verification techniques that journalists have honed over many years (Silverman 2013).

The problem of validity also addresses "representativity": i.e., it aims to reassure that the digital traces appropriately capture the overall population (which is especially important in developing countries, where the data may not capture everyone's voice due to socio-economic differences/preferences). It is possible that population sub-groups with limited access to technology may be systemically underrepresented in data—thus the need of sound statistical analysis is not obviated due to the large size of data. Much like the "dog that didn't bark" that tipped off Sherlock Holmes in one of his stories, sometimes the data that is not captured is more important than what was captured. This sampling bias is always present in social media and must be accounted for (i.e., when the dog does not bark, one must dig deeper to investigate more closely).

Values

It is important that the big crisis data community and the digital humanitarian community emphasize value-based and ethical humanitarian service. In this regard, these communities can leverage the collective knowledge of the existing humanitarian organizations available in the form of the *"humanitarian principles"* (Humanitarian Principles) that define a set of universal principles for humanitarian action based on international humanitarian law. These principles are widely accepted by humanitarian actors and are even binding for UN agencies. The big crisis data analytics community also needs to adopt the same or similar principles to guide their own work. The humanitarian principles are described below:

- *Humanity* means that the humanitarian imperative comes first–aid has to be given in accordance to need. Currently, the trend is to collect and retain data in the hope that future data analytics will help us understand conflicts and disasters better. The specific vulnerabilities of people in need might require, however, a very restrictive data policy for their own protection that in many instances should comply

with medical standards of restricted data use and deletion. For humanitarian big data this could mean that the collection of data might strictly be governed by a "need to know" principle.

- *Neutrality* means that humanitarian actors must not take sides in the conflict. The big data collection and analysis posits the problem that data needs to be collected in such a way that it represents the actual need and not, for example, the statistic about the access to the Internet or the availability of cell phones which might be different between parties of a conflict (see also above Veracity, Verification, and Validity).
- *Impartiality* means that aid should be delivered without discrimination as to the nationality, race, religious beliefs, class or political opinions. More specifically, this means that data collection needs to be sensitive to: (a) reducing biases in specific types of data, (b) the specific data sensitivities with regards to vulnerable groups, (c) processing and analytics in order to remain truthful, and (d) being cautious not to perpetuate or increase the vulnerabilities of groups by emphasizing differences.
- *Independence* means that "humanitarian action must be autonomous from the political, economic, military or other objectives that any actor may hold with regard to areas where humanitarian action is being implemented" (Humanitarian Principles). Big data is collected in the form of CDR data and mobile phone cash transfer data by telecommunication service providers and financial institutions with commercial interest. The same agencies provide data as "charity" to humanitarian analysts in order to improve humanitarian response. However, one consequence of the humanitarian data analytics is that it might expose vulnerable groups as targets for (specifically) tailor-made services, which the service providers might want to profit from. This would be a particularly exploitative attempt to increase sales (e.g. offering international roaming discounts for cell phone or the Internet usage to refugees). Such exploitative practices that profit from the vulnerabilities of people in need are incompatible with the principle of independence.

With the rise of digital humanitarianism, and various applications of big crisis data analytics emerging, it becomes important to ensure that these humanitarian principles are followed, and that the advice of ethical experts is sought where necessary.

Volunteers

Crowdsourcing information and solving complex big data tasks using microtasking sounds wonderful; however, it is not without problems. Sustaining a volunteer task force for crowdsourcing and croudcomputing is difficult beyond a certain time frame (as the volunteers have other needs—such as jobs, families—to attend). Two complex problems related to volunteered crowdsourcing are: (1) how to maintain the quality of crowdsourced data? (2) how to motivate micotaskers to work on the crowdsourced problem? To be truly successful, big crisis data analytics has to leverage not only crowdsourcing, but also the knowledge of human experts and the automation of ML. In existing literature, various incentives have been studied. One interesting approach is to use gaming approaches to "gamify" crowdsourcing and thereby incentivize volunteer participation. Such an approach has also been used for citizen science (e.g., Massively Multiplayer Online Gaming (MMOG) techniques have been used to incentivize volunteer participation; such an approach helped gamers on a crowdsourced gaming science site http://fold.it decoded an AIDS protein in 3 weeks, a problem that had stumped researchers for 15 years!).

The process of efficient crowdsourcing, particularly for disaster relief, is marred with several other challenges as well. One of the biggest challenges is coordination among different organizations, or agencies, that collect and analyze data from social media for disaster relief. As there is no infrastructure provided by the social media sites for such organizations, there can be two organizations responding to a same problem determined from the data gathered from a particular site, at the same time. Another issue is the spatial problem. Many users update their status with the information related to a crisis sitting, all together, at a different geographical site. This is a challenge in pinpointing an actual place of crisis for which the information was provided at the first place. Although, scientists are working on trust management systems for the verification of the information gathered for an appropriate action: Fraudulent information and entities can still infiltrate the information network. This information can then be treated like normal data and has the potential to diffuse and infect other connected entities of the information network. This vulnerability is primarily caused by the connected nature of information producing and consuming entities, this *vulnerability of connectivity* and cacasding errors/failures are discussed by Barabasi in his book (Barabási 2015).

Policy challenges: ensuring privacy and preventing abuse

The field of big data promises great opportunities but also entails some great risks of abuse and misuse. With big crisis data, there is always the danger of the wrong people getting hold of sensitive data—something that can easily lead to disastrous consequences. For example, PakReport.org had to restrict access to its crowd sourced crisis maps after the Taliban had threatened to use that data to target humanitarian organizations (Chamales 2012). The

importance of ensuring data privacy and protection can be gauged from the fact that UN Global Pulse has a Data Privacy Advisory Group (Data privacy advisory group). This group involves experts from academia, private sector and civil society so that the protection and privacy of data that is being used for UN Global Pulse's missions is ensured. The development of appropriate policy can help manage this dilemma between the opportunities and risks of big data. Some of the big questions that big data policies should address are: (1) what data to open up? (2) who should be able to access which data? (3) which data should be publicly accessible? (4) how can the data be used, reused, repurposed, and linked? The devised policies must also include prescriptive steps that ensure that data is used ethically (and not misused by malevolent actors and crisis profiteers). In particular, we should take steps to ensure that crisis victims do not expose themselves or others to further harm unwittingly (e.g., in countries beset with a civil war, or sectarian violence, a request for help with personal information may also be used maliciously by malevolent actors for violent purposes).

The above discussion stresses the need for a formal framework to be developed in which data can be used responsibly in humanitarian actions. The notion of establishing a standard framework that could be applicable across the board goes further than individual assurances of data privacy and protection by NGOs and similar entities. A recent report by UN's Office for the Coordination for Humanitarian Affairs (UN-OCHA) (Think Brief: Building data responsibility into humanitarian action 2016) stresses the need for standard set of principles and processes that will ensure the responsible use of data by all for humanitarian purposes. Specifically, the report outlines a four-step process that can be followed to achieve data responsibility:

1. ***Context evaluation:*** It should be emphasized that data must be used given a need and a context not just because it is available to use.
2. ***Data storage:*** Questions related to the location, security and access of the data should be answered properly.
3. ***Risk assessment:*** The vulnerabilities in data should be pre-assessed.
4. ***Risk mitigation plan:*** From the assessment performed in the above step effective and fool-proof measures should be taken to minimize any future risks.

Uneven data and digital divide

Among other issues, *digital divide* (Hilbert 2011) is related to the uneven proliferation of technology through out the world. The result of this divide could harm nations that lack the infrastructure, economic affordability, and *data-savvy* faculty. This unevenness, among other issues, poses a challenge for privacy as well. The concept of *information asymmerty* is important to ensure privacy in today's big data culture (O'Leary 2015). Information asymmetry implies that all the entities that gather user data to analyze, should ideally be limited in gaining insights about their users. The author in (O'Leary 2015) first discusses the uneven nature of big data in the perspective of well known big data Vs (volume, velocity, veracity, and variety). This uneven nature of big data inherently introduces a bias that could further lead to inefficient data analysis. As an example, if we consider the volume aspect of data then it could happen that some individual leaves sparse digital footprint (may be because of her personal habits of technology usage) as compared to a few others, resulting in a *non-all-inclusive* data analytics that could lead to biased policies. Secondly, there are various approaches that act in reducing the information asymmetries and hence these are potential privacy caveats, such as corroborating different information items from various sources can potentially reveal insights about an individual or an organization (which it never meant to make public) that can harm its privacy (O'Leary 2015).

Future directions for big crisis data analytics

Real-time crisis analytics

During disaster response, time is of paramount importance (for example, in the case of an earthquake, first responders can typically rescue someone trapped under the rubble in the first few days after which factors such as dehydration, untreated injuries will likely take toll). While it is true that these massive amounts of data can play a big part in providing clues to the best way to respond to a situation, accessing the right data entails the challenging task of sifting through the date deluge and filtering out the required information in real time. To complicate the big data analytics problem (which is akin to the problem of finding a needle in the haystack), the data quickly becomes out of date due to a rapidly and dynamically changing environment (and thus our problem becomes more difficult: the needle in the haystack has to be found in real-time). Futhermore, as time passes, the data may not only become useless, but can also be harmful (e.g., the limited resources of the disaster response may be incorrectly routed due to wrong information).

Real-time big data analytics can have many applications for enhancing disaster response. In many cases, various crises have many cascading stages in which one problem may trigger other problems (e.g., a tsunami approaching a nuclear power plant). In such scenarios, real-time analytics can help prioritize the most urgent problems before any further damage is caused. Critical information can also be broadcasted to the public (e.g., using

radio technology or through SMS) so that any preventable follow-up hazards are thwarted. In addition, real-time analytics can assist aid organizations and first responders to coordinate with other stakeholders.

Secure, reliable, disaster-tolerant crisis analytics

There is a need to develop highly secure and reliable, real-time, crisis response infrastructures. Given the critical nature of crisis situations, it is important to ensure that big crisis data computing and analytics systems are highly disaster-tolerant (in terms of reliability and availability). This requires the system to have foolproof mechanisms to withstand extremely adverse conditions due to natural disasters such as flooding, earthquakes, physical damage to infrastructure, power outages, etc. It is also important to ensure proper authentication, authorization, and accounting. This will be important since we would like only authorized emergency personnel to have remote access to resources (in the case of tele-operation such as tele-health) and sensitive data (such as the location and identification of affected people). We also need mechanisms that will ensure integrity of the data so that we are confident that the correct data is being used to drive the emergency response.

A combination of mobile technology and cloud computing (in the form of *mobile cloud computing*) can be useful here as cloud computing naturally complements big data technologies and is well-suited for reliable storage and analysis of big data. However, since the weakest link is the bottleneck for security, reliability, and availability, a number of technical issues need to be worked out for the desired solution: in particular, we have to pay equal attention to the security and reliability of the cloud infrastructure as well as the edge devices that may be acting as sensors. This requires the usage of redundancy as well as diversity at all the architecture components (such as the computing, storage, communications, sensing, and power components). Special attention should be paid to the power issue: many onsite communication/computation devices (such as smartphones) continuously need battery charging or will otherwise cease to function. System architects of crisis informatics systems should incorporate these factors into the design of their applications.

AI-based predictive/context-aware crisis analytics

Big crisis data analytics can be performed to answer various kinds of questions. Most of the work on crisis analytics has been hindsight-focused, including works on *Descriptive Analytics* (to answer "what happened or is happening?") or *Diagnostic Analytics* (to answer "why did it happen?"). Relatively little work has focused on forward-looking analytics such as *Predictive Analytics* (to answer "what will happen?") and *Prescriptive Analytics* (in which we answer "how can we make it happen?"). With predictive and prescriptive crisis analytics, we can obtain accurate advance information on developing disasters in an easy-to-understand and reliable way, such that the affected community can accept and act on the data. Some data science-based works have been proposed in literature for predicting impending conflicts or crises. As an example, Perry et al. have proposed predictive policing mechanisms for crime forecasting as a tool for riot/violence mitigation (Perry et al. 2013). Predictive policing involves statistical analyses in order to provide indicators to law enforcement agencies that they can intervene a potential criminal situation or solve a past crime. Applied mathematics, statistical physics and network science (Barabási 2013) can also be used to better understand and model the dynamism of criminal activities and thereby using predictive policing to prevent crimes from happening (D'Orsogna and Perc 2015). Mohler et al. provide an example of it in which they relate crime to earthquakes in nature, which are recurrent, and model it with Epidemic Type Aftershock Sequence (ETAS) point process. It has been shown to perform better than the traditional methods used by criminal analysts by 1.4–2.2 times (Mohler et al. 2015). Similarly, concepts specially from statistical physics and network science (Barabási 2013) can be used to model the behavior of epidemics and methods of predictive policing can be deployed to fight the spread of such diseases. However, extreme care should be taken while deploying the predictive policing methods particularly to prevent crimes. False positives in such processes can aggravate the situation even more. There are risks of specific ethnic/racial groups being falsely detected as the targets by a particular algorithm. Some security experts are of the view that infrequent criminal activities (such as terrorist attacks) cannot be efficiently predicted using data mining (Schneier 2015). Good quality prior data is a must, i.e., there needs to be a significant instances of events related to a criminal activity. In the case of terror attacks, as an example, this is not always the case. Hence the error rate in detection could be large. Another challenge is the presence of unique events in data, which do not correlate strongly to the past events, and since machine learning and data mining fields mostly make decisions based on past occurrences of an event that could potentially lead to false positives.

With the advances in AI, and the widespread availability of Internet and GPS with smartphones, it is possible (assuming that legal and policy issues can be suitably addressed) to push out *context-aware personalized crisis information* to the crisis prone or crisis affectees (e.g., using AI for automated processing and SMS for dissemination). A possible scenario is an earthquake: a simple application could be to send out notifications according to the information about the earthquake's epicenter

and the location of the user; in more advanced applications, other sources of data (such as information on a social network) can be integrated to create applications that leverage knowledge about location of the user and the user's friends to push messages to the user and his/her friends. Facebook already offers a similar social networking based application called "*Safety Check*" that users can use to inform and know about the safety of their friends and family after a disaster. For phones with active users, appropriate escape routes could also be pushed out to users using facilities such as the *Google Crisis Response*.

Multimodal big crisis data informatics

Big crisis data incorporates both very large data and also a large number of sources of data (that may be providing diverse kinds of data). Each of these big crisis data sources provides a unique (but necessarily incomplete) window to onsite developments. As popularized by John Godfrey Saxe's poem on "the blind men and the elephant", it is true that different experiences of the same reality can lead to different perceptions. This poem tells a story of the six blind men who went to "see" an elephant and ended up in complete disagreement; the poem goes on to say that the six blind men "disputed loud and long, each in his own opinion; exceeding stiff and strong, though each was partly in the right, and all were in the wrong!". This highlights the dangers of siloed big crisis data analytics. To obtain more complete understanding of the reality of the crisis, it becomes important that we tackle the challenging task of reconciling and combining the various distinct modalities of crisis information (such as text, images, speech, video, maps, crowdsourced and formal reports). This field of multimodal big crisis data analytics defines the real frontier of research in big crisis data analytics and promises to be a fertile area of future research in the overall field of crisis informatics.

As an example the Humanitarian Data Exchange (HDX) (Humanitarian Data Exchange) is an open platform where humanitarian data can be shared, accessed and analyzed to gain insights into a crisis. Recently the HDX team has provided the facility to analyze multiple datasets related to a crisis together to get a better and bigger picture of the dynamics of a crisis. They have developed a 'map explorer' prototype which allowed to analyze the Lake Chad Basin crisis by visually exploring the relationships between diverse datasets such as related to displacement, population fatalities, conflict locations, humanitarian funding and food security etc. Another similar effort is by World Food Programme with the name of "mobile Vulnerability Analysis and Mapping (mVAM)" (mVAM). This project collects data remotely using brief surveys conducted using SMS, telephone calls and Interactive Voice Response (IVR) system and then corroborates these data streams to provide visual reports that are important to understand the status of food security in different regions of the developing world.

Conclusions

This article has reviewed the use of big data analytics for processing and analyzing big crisis data. We discussed which technologies can enable a efficient data driven crisis response. We also outlined various sources of big data that can be utilized to gather useful information related to a crisis. We have highlighted the current state of the art along with some future directions with a discussion on the accompanying challenges and pitfalls.

As we observed in this article that even though big data can provide the answers to the most important questions facing modern humanitarian response, reaping the full benefit of big crisis data requires work on many different fronts. It must be understood that big crisis data analytics is but one cog in the entire crisis informatics ecosystem. With many stakeholders in the crisis response ecosystem, there is also a need of harmonic smooth handover of responsibilities. A friction-less interworking of various stakeholders should be aimed at by incorporating big crisis data analytics for the Linking Relief, Rehabilitation and Development (LRRD) process (which is focused on ensuring the interworking of humanitarian aid and medium- and long-term development organizations).

Endnotes

[1] A solution based on the digital currency Bitcoin named BitNation has been developed to provide banking for the refugees and the crisis affectees (https://refugees.bitnation.co/).

[2] According to International Telecommunication Union (ITU), by the end of 2013, there were almost as many mobile-cellular subscriptions as there are people ($\tilde{7}$ billion). The estimated number of active mobile-broadband subscriptions was 2.1 billion in 2013 (approximately, 1 in every 3).

Authors' contributions

JQ carried out most of the research work. AA assisted with the whole writing process along with providing ideas related to the content and the overall structure of the paper. RR provided many useful resources and information related to big data processing and analytics along with pointing out a few of the challenges in the big data ecosystems. AS and JC provided many useful insights and ideas related to the ethical use of data. AZ provided his detailed feedback on the paper's manuscript along with information related to the ethical use of data. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Electrical Engineering Department, Information Technology University (ITU), Lahore, Pakistan. [2] King Faisal University, Hofuf, Kingdom of Saudi Arabia. [3] University of Groningen, Groningen, Netherlands. [4] Computer Laboratory, University of Cambridge, Cambridge, UK.

Received: 13 April 2016 Accepted: 26 July 2016
Published online: 17 August 2016

References

Amazon Mechanical Turk (2016). https://www.mturk.com/mturk/welcome. Accessed 24 Feb 2016
Barabási AL (2013) Network science. Philos Trans R Soc London A 371(1987):20120375
Barabási, AL (2015) Network Science. Cambridge University Press. [The text is currently in press at Cambridge University Press. Online text; Accessed 10 Sept 2015]. http://barabasi.com/networksciencebook/
Bengio Y (2009) Learning deep architectures for ai. Found trends® Mach Learn 2(1):1–127
Bengtsson L, Lu X, Thorson A, Garfield R, Von Schreeb J (2011) Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS Med 8(8):1128
Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, Rebaudet S, Piarroux R (2015) Using mobile phone data to predict the spatial spread of cholera. Sci Rep 5. doi:10.1038/srep08923
Big data in action for development (2014) Technical report. The World Bank
Bizer C, Heath T, Berners-Lee T (2009) Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts. Information Science Reference (an imprint of IGI Global), USA, pp 205–227
Boulos MNK, Resch B, Crowley DN, Breslin JG, Sohn G, Burtner R, Pike WA, Jezierski E, Chuang K-YS (2011) Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. Int J Health Geograph 10(1):67
Butler D (2013) When google got flu wrong. Nature 494(7436):155
Caerus First Mile Geo (2014) Aleppo Overview. http://aleppo.firstmilegeo.com/. Accessed 19 Feb 2016
Chamales G (2012) Lives on the line - securing crisis maps in Libya, Sudan, and Pakistan. https://media.blackhat.com/bh-us-11/Chamales/BH_US_11_Chamales_Lives_on_the_Line_WP.pdf. Accessed 18 Mar 2016
Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: a distributed storage system for structured data. ACM Trans Comput Syst (TOCS) 26(2):4
Chohan F, Hester V, Munro R (2012) Pakreport: Crowdsourcing for multipurpose and multicategory climaterelated disaster reporting. CTs, Climate Change and Disaster Management Case Study. Climate Change, Innovation & ICTs Project. Center for Development Informatics (CDI), University of Manchester, UK
Collobert R, Weston J (2008) A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. ACM, Helsinki, Finland. pp 160–167
CrisisCommons (2016a). http://crisiscommons.org/. Accessed 25 Feb 2016
CrisisMappers (2016b). http://crisismappers.net/. Accessed 23 Feb 2016
Crowdcrafting (2016c). https://crowdcrafting.org/. Accessed 24 Feb 2016
CrowdFlower (2016d). http://www.crowdflower.com/. Accessed 24 Feb 2016
Crowley J, Chan J (2011) Disaster relief 2.0: the future of information sharing in humanitarian emergencies. Harvard Humanitarian Initiative and UN Foundation-Vodafone Foundation-UNOCHA. Harvard Humanitarian Initiative; United Nations Foundation; OCHA; The Vodafone Foundation
DataKind. http://www.datakind.org/. Accessed 25 Feb 2016
Data privacy advisory group. http://www.unglobalpulse.org/data-privacy-advisory-group Accessed 27 June 2016
Deborah E (2016). https://tech.cornell.edu/people/deborah-estrin. Accessed 26 Feb 2016
DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W (2007) Dynamo: amazon's highly available key-value store. In: ACM SIGOPS Operating Systems Review. ACM, Stevenson, Washington, USA Vol. 41. pp 205–220
Digital Humanitarian Network. http://digitalhumanitarians.com/. Accessed 24 Feb 2016
DigitalGlobe (2016). http://www.digitalglobe.com/. Accessed 24 Feb 2016
D'Orsogna MR, Perc M (2015) Statistical physics of crime: a review. Phys Life Rev 12:1–21
Duffy J (2015) Personalizing health care through big data. http://hub.jhu.edu/magazine/2015/spring/individualized-health-through-big-data. Accessed 19 Feb 2016
Elhauge E (2010) The fragmentation of U.S. Health Care: causes and solutions. Oxford University Press, New York. Oxford Scholarship Online, 2010. doi: 10.1093/acprof:oso/9780195390131.001.0001
esri. http://www.esri.com/. Accessed 25 Feb 2016
Estrin D (2014) Small data, where n= me. Commun ACM 57(4):32–34
Flowminder. http://www.flowminder.org/. Accessed 15 Mar 2016
Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, Gunter C, Musen M, Platt R, Stead W, et al. (2014) Toward a science of learning systems: a research agenda for the high-functioning learning health system. Journal of the American Medical Informatics Association. The Oxford University Press. doi:10.1136/amiajnl-2014-002977
Gao H, Barbier G, Goolsby R (2011) Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell Syst 26(3):10–14. IEEE Computer Society, Los Alamitos, CA, USA. http://doi.ieeecomputersociety.org/10.1109/MIS.2011.52
Geeks Without Bounds. http://gwob.org/. Accessed 25 Feb 2016
GISCorps. http://giscorps.org/. Accessed 25 Feb 2016
Gupta A, Kumaraguru P, Castillo C, Meier P, McFarland D (2014) Tweetcred: real-time credibility assessment of content on twitter. In: Aiello LM (ed). Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings. Springer International Publishing, Cham. pp 228–243. doi:10.1007/978-3-319-13734-6_16. http://dx.doi.org/10.1007/978-3-319-13734-6_16
Haklay M, Weber P (2008) Openstreetmap: user-generated street maps. Pervasive Comput IEEE 7(4):12–18
Hartung C, Lerer A, Anokwa Y, Tseng C, Brunette W, Borriello G (2010) Open data kit: tools to build information services for developing regions. In: Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development. ACM, London, United Kingdom. p 18
HealthMap. http://www.healthmap.org/. Accessed 24 Feb 2016
Heath T, Bizer C (2011) Linked data: Evolving the web into a global data space. Synth Lect Seman Web: Theory Technol 1(1):1–136
Heymann DL, Rodier GR, et al. (2001) Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases. Lancet Infect Dis 1(5):345–353
Hilbert M (2011) The end justifies the definition: the manifold outlooks on the digital divide and their practical usefulness for policy-making. Telecommun Policy 35(8):715–736
Hoffmann L (2012) Data mining meets city hall. Commun ACM 55(6):19–21
Howe J (2006) CROWDSOURCING. http://www.crowdsourcing.com/. Accessed 19 Feb 2016
Humanitarian Data Exchange. http://docs.humdata.org Accessed 28 June 2016
Humanitarian Principles. https://docs.unocha.org/sites/dms/Documents/OOM_HumPrinciple_English.pdf. Accessed 17 Mar 2016
IATI Standard. http://iatistandard.org/ Accessed 28 June 2016
IBM (2015) IBM Research. http://www.research.ibm.com/cognitive-computing/watson/index.shtml#fbid=O1rcZI4aQWK. Accessed 17 Sept 2015
Imran M, Castillo C, Lucas J, Meier P, Vieweg S (2014) AIDR: Artificial intelligence for disaster response. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, Seoul, Korea. pp 159–162
Initiative HH, et al. (2010) Disaster relief 2.0: The future of information sharing in humanitarian emergencies. In: Disaster Relief 2.0: The Future of Information Sharing in Humanitarian Emergencies. HHI; United Nations Foundation; OCHA; The Vodafone Foundation
International Aid Transparency Initiative. http://www.aidtransparency.net/about. Accessed 28 June 2016
James M, Michael C, Brad B, Jacques B (2011) Big data: the next frontier for innovation, competition, and productivity. The McKinsey Global Institute. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation
Kirkpatrick R (2013) Big data for development. Big Data 1(1):3–4
Knight W (2015) Deep learning catches on in new industries, from fashion to finance. http://www.technologyreview.com/news/537806/deep-learning-catches-on-in-new-industries-from-fashion-to-finance/. Accessed 26 Feb 2016
Lane ND, Miluzzo E, Lu H, Peebles D, Choudhury T, Campbell AT (2010) A survey of mobile phone sensing. Commun Mag IEEE 48(9):140–150

Leavitt N (2010) Will nosql databases live up to their promise? Computer 43(2):12–14
Lohr S (2015) Using patient data to democratize medical discovery. http://bits.blogs.nytimes.com/2015/04/02/using-patient-data-to-democratize-medical-discovery/?_r=1. Accessed 25 Feb 2016
Manyika J (2013) Open data: unlocking innovation and performance with liquid information. McKinsey. http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information. Accessed 04 Aug 2016
Manyika J, Chui M, Bisson P, Woetzel J, Dobbs R, Bughin J, Aharon D (2015) The Internet of things: mapping the value beyond the hype. Technical report. McKinsey Global Institute. http://www.mckinsey.com/business-functions/business-technology/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world. Accessed 04 Aug 2016
MapAction. http://www.mapaction.org/. Accessed 25 Feb 2016
Map Give: imagery to the crowd. http://mapgive.state.gov/ittc/. Accessed 24 Feb 2016
Mapping the conflict in Aleppo Syria (2014) Technical report, Caerus Associates
McDonald SM (2016) Ebola: a big data disaster. Technical report
Meier P (2014) Digital humanitarians: how big data is changing the face of humanitarian response. CRC Press
Meier, P (2015) Digital humanitarians. How big data is changing the face of humanitarian response. CRC Press
MicroMapper (2016). http://micromappers.org/. Accessed 24 Feb 2016
Missing Maps. http://www.missingmaps.org/about/ Accessed 27 June 2016
Mission 4636. http://www.mission4636.org/. Accessed 24 Feb 2016
Mohler GO, Short MB, Malinowski S, Johnson M, Tita GE, Bertozzi AL, Brantingham PJ (2015) Randomized controlled field trials of predictive policing. J Am Stat Assoc 110(512):1399–1411
Munro R (2013) Crowdsourcing and the crisis-affected community. Inform Retriev 16(2):210–266
Munro R, Manning CD (2012) Short message communications: users, topics, and in-language processing. In: Proceedings of the 2nd ACM Symposium on Computing for Development. ACM, Georgia, Atlanta. p 4
Munro R, Gunasekara L, Nevins S, Polepeddi L, Rosen E (2012) Tracking epidemics with natural language processing and crowdsourcing. AAAI Spring Symposium Series. http://www.aaai.org/ocs/index.php/SSS/SSS12/paper/view/4337. Accessed 04 Aug 2016
mVAM. http://mvam.org/info/ Accessed 28 June 2016
O'Leary DE (2015) Big data and privacy: Emerging issues. Intell Syst IEEE 30(6):92–96. doi:10.1109/MIS.2015.110
Open mHealth. http://www.openmhealth.org/. Accessed 26 Feb 2016
Pakreport (2016). http://www.pakreport.org/. Accessed 24 Feb 2016
Palen L, Vieweg S, Sutton J, Liu SB, Hughes AL (2007) Crisis informatics: Studying crisis in a networked world? In: Proceedings of the Third International Conference on E-Social Science
Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inform Retriev 2(1–2):1–135
Perry WL, McInnis B, Price CC, Smith S, Hollywood JS (2013) Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Rand Corporation, Santa Monica, CA. http://www.rand.org/pubs/research_reports/RR233.html
Pulse UG (2012) Big data for development: Challenges & opportunities. mayo, Naciones Unidas, Nueva York
Raymond NA, BLCZAAILB, Davies BI (2013) While we watched: Assessing the impact of the satellite sentinel project. Georgetown J Int Affairs 14(2):185–191
Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL, et al. (2012) Digital epidemiology. PLoS Comput Biol 8(7):1002616
Sanou B (2013) The world in 2013: Ict facts and figures. International Telecommunications Union
Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Netw 61:85–117
Schneier B (2015) Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World. WW Norton & Company
Siegel E (2013) Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, Or Die. John Wiley & Sons
Silverman C (2013) Verification handbook. The European Journalism Centre (EJC). http://verificationhandbook.com/. Accessed 04 Aug 2016
Small Data Lab - UNU. http://cs.unu.edu/about/small-data/. Accessed 26 Feb 2016
Solnit R (2009) A paradise built in hell. Viking, New York
Standby task force (2016). http://blog.standbytaskforce.com/. Accessed 25 Feb 2016
Surowiecki J (2005) The Wisdom of Crowds. Anchor
Tang JC, Cebrian M, Giacobe NA, Kim HW, Kim T, Wickert DB (2011) Reflecting on the darpa red balloon challenge. Commun ACM 54(4):78–85
Taylor L (2014) No place to hide? the ethics and analytics of tracking mobility using mobile phone data. http://bordercriminologies.law.ox.ac.uk/no-place-to-hide/. Accessed 04 Aug 2016
The home of the U.K. Governmen's open data. https://www.data.gov.uk/. Accessed 24 Feb 2016
The home of the U.S. Government's open data. http://www.data.gov/. Accessed 23 Feb 2016
The Small Data Lab at Cornell Tech. http://smalldata.io/. Accessed 26 Feb 2016
Think Brief: Building data responsibility into humanitarian action (2016) Technical report, UN-OCHA
Tomnod. http://www.tomnod.com/. Accessed 24 Feb 2016
Trace the Face. http://familylinks.icrc.org/europe/en/Pages/Home.aspx. Accessed 19 Feb 2016
Translators without borders. http://translatorswithoutborders.org/. Accessed 25 Feb 2016
Turner A (2006) Introduction to Neogeography. O'Reilly Media, Inc.
Twigg J, et al. (2004) Disaster Risk Reduction: Mitigation and Preparedness in Development and Emergency Programming. Humanitarian Practice Network, Overseas Development Institute
Twitter SuperstromSandy. https://2012.twitter.com/en/global-town-square.html. Accessed 17 Mar 2016
United Nations Global Pulse. http://www.unglobalpulse.org/. Accessed 04 Oct 2015
Ushahidi (2016). http://www.ushahidi.com/. Accessed 19 Feb 2016
Vinck P (2013) World Disasters Report 2013: Focus on Technology and the Future of Humanitarian Intervention. International Federation of Red Cross and Red Crescent Societies. http://www.ifrc.org/PageFiles/134658/WDR%202013%20complete.pdf. Accessed 04 Aug 2016
Verily. https://veri.ly/. Accessed 25 Feb 2016
Wilson K, Brownstein JS (2009) Early detection of disease outbreaks using the internet. Can Med Assoc J 180(8):829–831
Ziemke J (2012) Crisis mapping: The construction of a new interdisciplinary field? J Map Geograph Libr 8(2):101–117