Big Data, Communities and Ethical Resilience: A Framework for Action

By 2013 Bellagio/PopTech Fellows Kate Crawford, Gustavo Faleiros, Amy Luers, Patrick Meier, Claudia Perlich and Jer Thorp

Draft Date: Oct. 24, 2013

In August 2013, a multidisciplinary group gathered at the Rockefeller Foundation's Bellagio Center to address the theme of "Community Resilience Through big data and Technology." Creative and critical thinkers were selected from the technology sector, academia, the arts, humanitarian and ecological spheres. Over ten days, we explored how data could be used to help build community resilience in the face of a range of stresses — environmental, political, social and economic. Large data collection and analysis may support communities by providing them with timely feedback loops on their immediate environment. However, the collection and use of data can also create new vulnerabilities and risks, by enabling discriminating against individuals, skewing evidence, and creating dependencies on centralized infrastructure that may increase a system's vulnerability. After analyzing these risks and opportunities, we developed a framework to help guide the effective use of data for building community-driven resilience. In this framework, we propose six domains: ethics, governance, science, technology, place and sociocultural context. We believe that by considering all six domains together, organizations can safeguard against predictable failures by exposing project weaknesses from the outset rather than in hindsight.

Introduction

What is resilience?

The term “resilience” is both complex and shifting, and can be used to mean many different things. First, we define resilience as the ability to absorb and cope with unexpected shocks. To this end, resilient communities are those with a particular capacity to respond to disturbance or hazards. In the popular discourse resilience often focuses on resources and infrastructure and overlooks issues of power, ethics and accountability — which recent research suggests are critical.[1] Resilience is not a normative term: systems characterized as resilient may be either desirable or undesirable. As a result, if ethics is not consciously considered at the inception of data projects, steps taken to increase community resilience could in fact create more vulnerability and thus do harm. This explains why our focus is on *ethical resilience* – and how we might apply this idea to data-driven community projects.

What is ‘big data’?

“Big data” is a hype-laden concept, but it points to a multi-faceted, emerging phenomenon.[2] Data collection and processing have changed dramatically in the last five years. Where once data was collected only for a specific purpose, now massive amounts of data are opportunistically and passively collected and reused in multiple contexts, over an indefinite time frame, often without informed consent. Beyond a shift in scale and persistence, data have become more granular and intrusive, and profoundly interlinked. Using data for predictive modeling can allow us to “fill in the blanks,” anticipating information about others without even needing to collect it. big data is thus challenging our assumptions about what can and can’t be private. And as recent studies have shown, even anonymized data sets can be used to re-identify people.[3] In such a context, even the idea of “privacy” is not enough — we need to think more broadly to assess how people might be affected by big data projects. How will these new attributes of data analysis and predictive

[1] O'Brien, K., B. Hayward, and F. Berkes. 2009. Rethinking social contracts: building resilience in a changing climate. Ecology and Society 14(2): 12. [online] URL: http://www.ecologyandsociety.org/vol14/iss2/art12/Synthesis

[2] boyd, d. and K. Crawford. 2011. Six Provocations for big data. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. [online] URL: http://blogs.gartner.com/svetlana-sicular/big-data-is-falling-into-the-trough-of-disillusionment/

[3] Montjoye, Y., C. Hidalgo, M. Verleysen, and V. Blondel. 2012. Unique in the Crowd: The privacy bound of human mobility. Scientific Reports 3. [online] URL: http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

modeling affect community resilience? How can we enhance the capacities of communities without doing harm?

Framework

To address these questions and others that arise when conducting data-oriented community resilience projects, we propose a framework emphasizing six domains: Ethics, Governance, Science, Technology, Place, and Sociocultural context. This framework is depicted in *Figure 1* below.

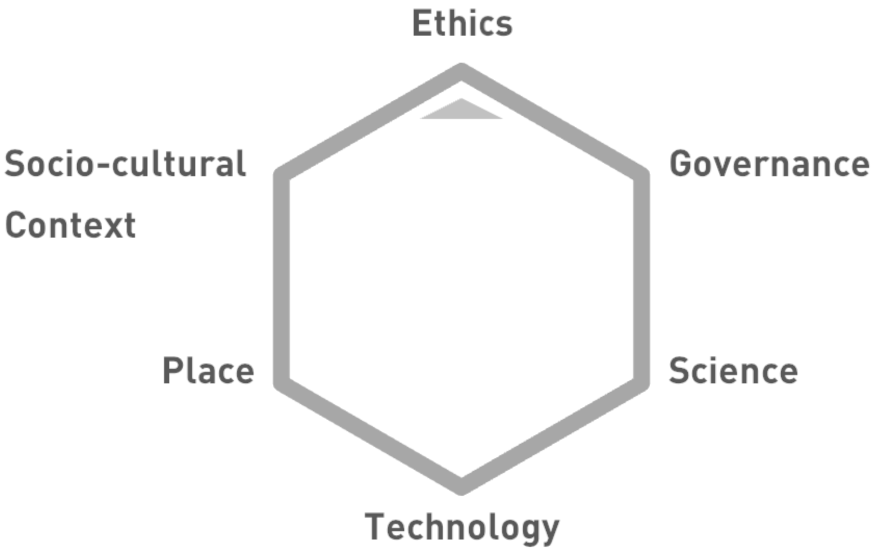

Figure 1: The arrangement of the six domains above reflects both similarities between domains (through adjacency) and differences (through oppositions).

Ethics is placed at the top of the hexagon, to emphasize its importance when designing community-based data projects. It is the primary guiding principle, which should be considered first and then in connection to the other domains. By doing so, we argue that the most substantial risks can be fully assessed and considered before data projects commence. What follows is a detailed explanation of each domain, as well as a discussion of how multiple domains can usefully interact.

1. Ethics

What is it?
Fairness, Agency, Consent, Social Justice, Participation

Why does it matter?
Ethical practices allow people to participate in how, where and for what purposes their data is used, which strengthens agency and leads to stronger communities.

Ethics is currently one of the most overlooked elements of big data science. Developing a strong set of ethical principles is essential to ensuring that community-based data projects do no harm. While privacy is important, it is too narrow a concept to adequately address the risks of what big data projects can do. By focusing on ethics, we can more broadly consider the human impact — both short and long term — of how data is being gathered and used. Unfortunately, ethical considerations are often sidelined because gathering more data is seen as necessarily better, and concerns about how data might be abused are only considered after the fact. The result is data sets that are overly intrusive, that collect personally identifying information without informed consent, and may have serious unintended consequences, particularly when brought together with other kinds of personally identifying data sets.

At its core, an ethical framework for data aims to enhance *agency*: the ability for individuals and communities to be able to make decisions about how, where, why and for how long their data is used. By emphasizing agency for community resilience projects, we are underscoring the responsibility for data projects to allow both granular control of data collection and retention, and the ability for users to opt out. However, the burden of responsibility should not be placed solely on individual users to understand and respond to the privacy guidelines or data retention practices of a project. Agency means giving people the capacity to engage with their data and participate, but the onus of responsibility still lies with the project developers. Ensuring this combination of data control, ethical engagement and responsibility builds trust and confidence in community data projects. We also think it makes for better data science.

The negative impacts of data collection and analysis are not evenly distributed. Some communities may be more exposed than others, and vulnerable individuals and groups can face serious risks. For example, the use of crowdsourcing for crime reporting in Mexico resulted in drug cartels tracking and lynching participants.[4] Predictive policing models, which use big data to create neighborhood ‘hot spots’ where crime is expected to occur, risk reinforcing existing racial and class prejudices and institutionalizing differential policing.[5] Finally, community activists in many countries are subject to email and cellphone monitoring.[6] The inability to escape data surveillance affects people’s freedom to express themselves, to engage in political activism or to trust their governments.

Ethical principles need to be incorporated from the outset when designing or assessing projects that focus on supporting community resilience. These principles should include informed consent, data ownership, accountability, transparency, data protection and data access.[7] Informed consent goes beyond merely making people aware of the terms of service or that data is being gathered about them, but a clear articulation of how data might be used, whether third parties will get access to that data, and how people can opt out or limit how much of their data is gathered. Data ownership, for example, means that a community owns the data it generates, as well as the learning algorithms and other derivatives from data projects. Best practices of accountability and transparency would enforce greater responsibility about how data is used, and allow participants to see where their data goes and alter or remove their data where possible. Ideally, data protection and data access should give people more certainty that their data is being stored as securely as possible, while also giving individuals access to their own data.

While not all projects will be able to achieve these goals, we suggest that they can serve as a set of best practices. We also suggest the need to consider big data due process: ensuring fairness when big data is used to make decisions that affect individuals and communities.[8]

[4] Monroy-Hernandez, A., E. Kiciman, D. Boyd, and S. Counts. 2012. Tweeting the Drug War: Empowerment, Intimidation, and Regulation in Social Media. HCIC. [online] URL: http://research.microsoft.com/apps/pubs/default.aspx?id=168809

[5] Robertson, J. 2013. How big data Could Help Identify the Next Felon - Or Blame the Wrong Guy. [online] URL: http://www.bloomberg.com/news/2013-08-14/how-big-data-could-help-identify-the-next-felon-or-blame-the-wrong-guy.html

[6] For example, see Solove, D. 2011 Nothing to Hide: The False Tradeoff Between Privacy and Security. New Haven: Yale University Press.

[7] For example, see the Fair Information Practice Principles: http://www.ftc.gov/reports/privacy3/fairinfo.shtm

[8] Crawford, K., and J. Schultz. 2013. big data and Due Process: Towards a Framework to Redress Predictive Privacy Harms. Boston College Law Review. 55(1). [online] URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2325784

2. Governance

What is it?
Public Policy, Corporate Policy, Social Networks

Why does it matter?
Good governance ensures that communities are placed at the center of data-driven projects

Governance, in an ethical resilience context, focuses on how public policy and corporate data policies can better prioritize and serve local communities by supporting community resilience projects that are genuinely people-centered, community driven and locally owned.

Data-driven approaches to community resilience run the risk of promoting a technical discourse that crowds out local participation: emphasizing technology rather than the multiplicity of approaches and perspectives.[9] We are concerned by definitions of resilience that emphasize external expertise and technological solutions as these are often misrepresented as politically neutral. More informed policies at a local and national level can prioritize community-driven participation and indigenous knowledge rather than solely emphasizing technology and infrastructure investments. Policies for ethical resilience can foster a balance between external assistance and expertise and the pre-existing capacity of local communities to help themselves. In sum, the responsibility of public policy for ethical resilience is to foster an environment conducive to respectful reciprocity, self-governance, improvisation and mutual aid.

The governance domain also encompasses corporate governance, which typically dictates data management policies such as access, ownership and privacy. Corporate data collection policies that are formulated without direct input from local communities may violate community norms and practices. As noted above, digital intrusion and surveillance erodes trust and threatens community resilience. Corporate governance should therefore promote respectful data management policies that enable individuals to determine if, how and for what purposes and for how long their data can be used.

[9] Meier, P. 2013. How to Create Resilience Through Big Data. iRevolution. [online] URL: http://iRevolution.net/2013/01/11/disaster-resilience-2-0

At best, governance models can facilitate and protect environments conducive to local and ethical self-organization. This means acting as a mediator between local and external interests. Good governance can support the transparency and accountability structures necessary for a just and ethical approach to community resilience.

3. Science

What is it?
Hypothesis Testing, Purposeful Data Collection, Projections, Theory

Why does it matter? Science allows people to assess causality, as opposed to simple correlation from big data, and to identify vulnerabilities and opportunities for resilience in complex systems.

Scientific knowledge is critical for building resilience as it enables society to understand trends, identify thresholds and anticipate vulnerabilities. However, the rise of data science, which emphasizes finding meaning in patterns, has begun to threaten elements of traditional scientific method, which purposely collects data to test hypotheses about how a system works, and conducts experiments that are reproducible. Without these characteristics of science we may create vulnerabilities in our knowledge base.

Big data science is used by government, corporations and international development organizations to identify patterns to better “serve the people.” Indeed, correlations made with big data do provide an opportunity for improved decision making by increasing the speed of analysis, enabling more targeted response and facilitating greater opportunities for citizen participation in the production of knowledge. However, taken in isolation outside of the traditional modes of scientific inquiry, data science can increase inherent biases and lead to erroneous conclusions when correlation is mistaken for causation.[10] There have been notable advances in data science around reliable causal analysis from observational data that has to be considered.[11]

Perhaps the biggest challenge regarding the focus on big data for building community resilience is that much of this is heavily centered on analysis data

[10] Crawford, K. 2013. Hidden Biases in big data. Harvard Business Review [online] http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data

[11] Van Der Laan, M., and S. Rose. 2011. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics

sensed in human systems — search queries, social media, and mobile activities. Such data sources are inherently linked to the here and now. Without a longer term understanding of the social and biophysical systems in which these data are captured, they provide no ability to understand future trends.[12] For example, consider the challenge of managing the risks of increasing shocks from climate change. While the collection and processing of huge amounts of data has been the foundation of climate change science, enabling a picture of possible futures, the collection and analysis of social data over the last decade has provided little guidance on how social and natural systems are likely to respond to these changes.

When using data science to inform and build community resilience in the face of uncertainty of future shocks, it will be critical to not take this in isolation from traditional scientific method.

4. Technology

What is it?
Modes of Data Collection, Storage, Analytics, Coordination

Why does it matter?
Technology provides new and more opportunistic modes to collect potentially valuable data, provides tools for coordination but can create dependencies and introduce new vulnerabilities

The main drivers of the recent big data trend are advances in technology around data sharing, data collection, predictive modeling and storage. While data and data analysis has been used for centuries, big data approaches have changed the nature of how information is stored and analyzed. The size of data is less relevant than the vastly increased level of granularity and the ability to capture not just information about entities and people but on the very specific connections between them. With this new richness and complexity of big data has come a growing expectation that technology should play a large part in emergency response and resilience.

In order to meet this expectation, we need to understand that technological systems are often very fragile. For example, many communications and power systems come under enormous stress during crisis events, leading to instabilities in technologies that rely on them. Another consideration is the correlation of the

[12] Luers, A. 2013. (Appropriate) big data for Climate Resilience?. Stanford Social Innovation Review. [online] URL: http://www.ssireview.org/blog/entry/appropriate_big_data_for_climate_resilience

availability of networked technology (resilient technology in particular) with economic means. In such cases, sensing and response solutions need to consider and actively pursue approaches that include community members on the lower economic scale, who may not use smartphones or social media.

One important consideration in developing technology for resilience is the acknowledgement that the goal has to be to empower resilient characteristics in communities rather than developing top-down technological solutions. This must include an integration of local leadership and structure as much as the need to have local skills and knowledge to sense, interact with, and evaluate data. Technological approaches should also consider the broader environment. Many large data projects require data centers, which draw considerable resources like water and energy. Making these processes more sustainable will be a priority for community resilience in the coming years.

To facilitate a tight integration of the community and to allow local knowledge sharing, data solutions should give preference to broadly available and open source tools (R, hadoop) as well as open standards (XML, JSON) where possible. Not only are these tools low cost and therefore more accessible to communities, they are often in the long run more stable as they do not depend on the financial well-being of a company. Their code is also open to inspection and auditing. Human skill remains the main success factor in data-driven projects. Being able to attract, train, and maintain that human capital locally is key to developing resilient data-driven solutions; relying on non-local expertise that may not have strong ties to a region or community creates new kinds of vulnerability and does not build data skills at a local level. A more detailed discussion of important technology consideration and an outline of best practices in non-commercial data is provided here.[13] Above all, the design of data projects for communities should emphasis broad participation, shared awareness, and strong boundaries for privacy and ethics.

[13] 'Seven Principles for Big Data and Resilience Projects', http://irevolution.net/2013/09/23/principles-for-big-data-and-resilience/

5. Place

What is it?
Natural Environment, Built Environment, Infrastructure, Climate, Cities, Neighborhoods

Why does it matter?
Community resilience is determined by place based peculiarities. Data generated on the local level strengthens decision making

A consideration of place is critical for building resilience. Both the natural and built environment define vulnerabilities and determine opportunities for resilience. The application of data science to increase ethical community resilience may consider the influence of historical changes made in the landscape and the accumulated knowledge of locals about their own environment. The relationship of communities with place can be a source of resilience as well as risks.

Weather patterns, landscape transformation and availability of natural resources shape local knowledge and define the capacity to absorb shocks. Nevertheless, climate change and the depletion of forests, water stocks, among other environmental stresses, are redefining established conditions. Constant assessments are needed. The access to more location specific data, like high resolution satellite imagery or geo-tagged media (photos, reports) can help to build greater situational awareness and can help to improve predictions and response. Geolocated data can serve as a tool for local decision making.

The ágora, the city, the borough, the prefecture or the council, among other instances of local governance, should be forces working towards stronger and more resilient communities. But there can be considerable tensions between local and global impacts: a hydroelectric dam can be considered a secure source of energy for a region, but also can act as a disruptive event for local communities that will be displaced. Meanwhile, an investment in extracting oil may improve the medium-term economic conditions of one specific location, but undermines the need to reduce emissions of greenhouse gases at a global level.

Because of these tensions between local and global impacts, hyperlocal data generated by sensors in many cities may not be effective for strengthening resilience unless an ethical approach is taken to give individuals and communities a choice as to whether they will participate or not in sensor data collection and use.

6. Sociocultural Context

What is it?
Customs, Language, Ideologies, Art, Media, Journalism

Why does it matter?
As data becomes culture, and vice versa, a deep understanding of sociocultural conditions becomes critical.

Sociocultural context encompasses concerns related to a specific population, and the behaviors that define them. This domain includes a variety of concepts, including social customs, language, and art. It also includes a number of broadly important considerations for communities and individuals, such as gender, race and class, as well as values, ideologies, and norms. The sociocultural domain also addresses media and journalism, religious practice, and cultural output such as literature, performance or film.

In the frame of data and community resilience, there are many ways in which sociocultural context is important. Norms around usage of personal data can vary drastically from one culture to another. Technologies or methods that do not infringe on an individual depending on their gender, race and class may nonetheless be dangerous to a different individual within the same community (for example, New York City's controversial "stop and frisk" policy).[14] Thus, an understanding of sociocultural diversity must be built into ethical community resilience thinking.

Sociocultural context should not only be seen as a set of conditions which should be considered in thinking around data and resilient communities. For example, the arts can play an active role in promoting and facilitating data literacy. By moving data discussions from restricted pockets of academia, corporations and governments to the public sphere, data practices can become part of the broader cultural discourse. As a result, communities as a whole can be more informed when making decisions based on their own data and that of their fellow citizens. This in turn builds a system of trust, leading to more resilient communities.

An example of such an approach is *Nuage Vert*, a public art project by Finnish artists Helen Evans and Heiko Hansen.[15] In this project, real-time data about electricity consumption in Helsinki is projected onto vapor emissions from a local

[14] Editorial Board. 2013. Racial Discrimination in Stop-and-Frisk. New York Times. [online] URL: http://www.nytimes.com/2013/08/13/opinion/racial-discrimination-in-stop-and-frisk.html

[15] Andrews, K. 2009. Nuage Vert "Green Cloud" Illuminates Emissions. [online] URL: http://inhabitat.com/green-cloud-hehe-helsinki-environmental-art/

smokestack. Visible for miles, the project makes consumption data a part of the built environment. In doing so, it generates discussion amongst the community, and ultimately seeks to persuade people to change patterns of consumption.

An important additional factor to consider in the context of culture is narrative. The way that resilience and data are framed within a society can greatly influence the trust and cooperation generated within the community. The fostering of community-specific narratives may strongly influence the overall success of resilience initiatives. Projects that may originate from outside communities need to consider that local narratives may differ significantly from those of the originating group.

Conclusion

This document offers a framework for thinking and action for big data community resilience projects. In our conversations, we focused on how future projects might avoid the failures of previous data-oriented community undertakings. Of primary importance is to focus on people-centered, community-driven approaches. The discourse of big data and community resilience often excludes local participation by less powerful or technically literate populations. As a result, external experts may reduce complex social problems like community resilience to terms that are suited to technological solutions. This crowds out local knowledge, participation and agency, which undermines trust, social connectedness and resilience. Clear public policy and corporate governance frameworks are needed to foster a generative and inclusive environment that is conducive to local communities participating in their own data projects.

Human skill is a key element of effective long-term data-driven projects. This makes local human collaboration an essential component of data projects for resilience. To enable greater local skills development and engagement, we suggest a combination of using open standards for data, common tools for the analytical components and educational workshops that develop data skills in the local community. To ensure the effectiveness of more community-oriented data projects, a set of ethical guidelines should be drafted at the beginning - rather than considering ethics and privacy merely as a “compliance process” at the end of a project. We hope this framework will serve as a provocation to consider the many possible human impacts of data projects much earlier in their design.

Lastly, while community-based projects are typically memorialized by their successes, there must be openness around and communication about failures as well. By embracing the possibility of failure, risks and challenges that comes with the process of embracing new technological approaches, we can all better learn from each other.

We are at a critical time in the emergence of big data science and its application to enhance the resilience of communities. We need to ask searching questions about how best to do this in a way that enhances the agency and capacity of communities without exposing them to new risks. By considering ethics as a primary guiding principle, and then the connections with governance, science, technology, place and the wider sociocultural context, we can produce groundbreaking data projects that can truly enhance the ability of communities to thrive.