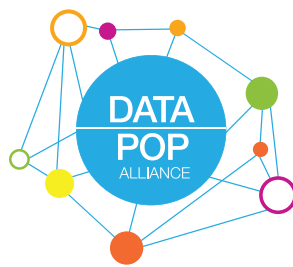


DATA-POP
ALLIANCE



TOOLKIT



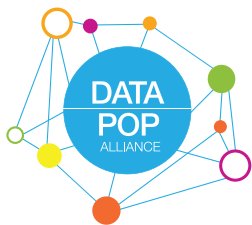
Leveraging Big Data for Sustainable Development

Financial support from



ABOUT US

Promoting a people-centered Big Data revolution



About Data-Pop Alliance

Data-Pop Alliance is a global coalition on Big Data and development created in 2014 by the Harvard Humanitarian Initiative (HHI), MIT Media Lab and the Overseas Development Institute (ODI), joined in February 2016 by the Flowminder Foundation as a fourth core member, to “promote a people-centered Big Data revolution.” Data-Pop Alliance started with seed funding from the Rockefeller Foundation and has since then been supported by a range of bilateral and multilateral donors, philanthropic foundations, and private companies.

Data-Pop Alliance brings together researchers, experts, practitioners, and activists working around the globe across different disciplines and industries around a common vision: making Big Data a force of positive social change in the 21st Century. To that end, Data-Pop Alliance focuses on the applications and implications of Big Data for development and seeks to ensure that Big Data contributes to improving decision-making processes and fostering citizen empowerment in ways that avoid elite capture, widening inequities, and the dehumanization of public policies.

Data-Pop Alliance works through 3 main modalities, deployed on the ground in various regions of world: (1) collaborative research, with both empirical and ‘White Papers’, etc.; (2) capacity building, through trainings, workshops, Data Expeditions, and more; and (3) policy and community engagement, including through our involvement in the Global Partnership for Sustainable Development

Data, the UN World Data Forum and in-country partnerships and activities. Our core domains of expertise and focus include official statistics and governance; urban dynamics; demographic and economic processes; peacebuilding and social cohesion; climate change and resilience; and data literacy and ethics.

Data-Pop Alliance functions as a distributed network with its headquarters and core team hosted at ThoughtWorks in New York City, a Bogotá-based team, and directors in New York City, Cambridge, London, and Geneva. In addition, Data-Pop Alliance relies on the intellectual resources of its inner circle of close to 30 Research Affiliates, experts and scholars located in more than 15 countries, and a network of more than 30 technical partners.

TABLE OF CONTENTS

- 1 Overview and Learning Objectives
- 3 Contexts & Concepts**
 - 4 The 3 C's of Big Data
 - 6 Big Data Sources for Development
 - 7 Big Data and Official Statistics
 - 9 The Global Partnership for Sustainable Development Data (GPSDD)
- 11 Methods & Tools**
 - 12 Stages of Data Use
 - 14 Unpacking the "DNA" Behind Big Data Projects: Sequencing the Stages of Data Use
 - 19 Stages of Data Use in Practice
- 21 Design & Strategy**
 - 22 Project Archetypes for Big Data and Development
 - 34 Navigating Your Big Data Innovation Strategy Roadmap
- 37 Ethics & Engagement**
 - 38 Menlo Report: Ethical Principles Guiding Information and Communication Technology Research
 - 40 Risks and challenges in big data projects
 - 47 Operationalizing "Do No Harm" Innovation in a Privacy-Centered World

This toolkit is not yet for publication or widespread circulation without written permission. Please be in touch at contact@datapopalliance.org with any inquiries or requests.

Trust
me



i'm
a data scientist

OVERVIEW AND LEARNING OBJECTIVES

Corresponding to what we consider to be the four building blocks of data literacy, this toolkit is structured around four key dimensions, or “building blocks” for developing practical data literacy skills for practitioners and policy-makers to build inclusive Big Data-driven innovation projects, policies and partnerships.

The following summary highlights the key learning points related to each building block:

C² Context & Concepts

- 1** Decode key terms and buzzwords in the Big Data and development landscape
- 2** Discuss Big Data within the broader political context of the post-2015 sustainable development framework and data for social good
- 3** Translate development problems into specific data objectives

M+T Methods & Tools

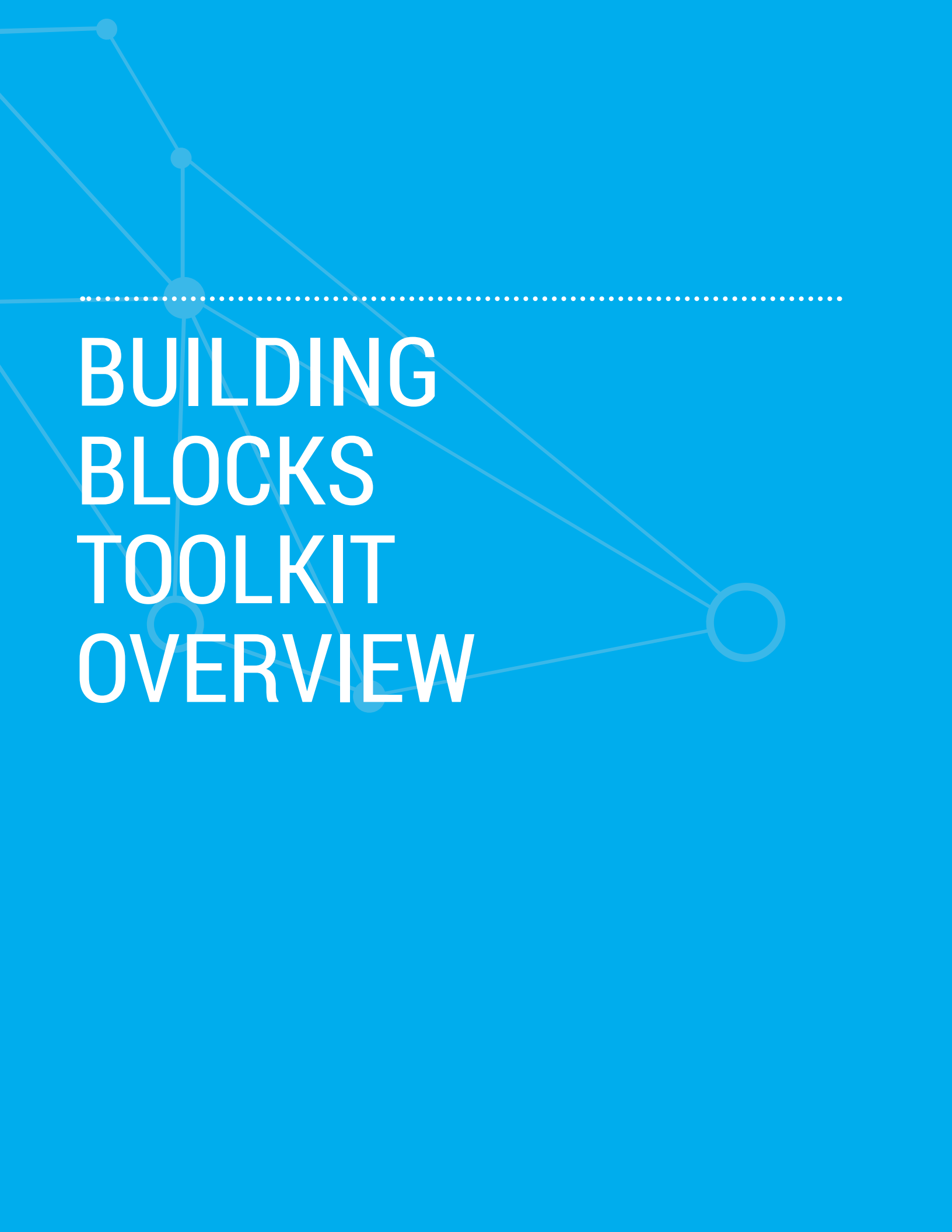
- 4** Understand existing methods and tools used to leverage Big Data
- 5** Assess data representativeness, biases and insights within Big Data-driven approaches and methods
- 6** Identify applicable tools by assessing the value add of Big Data for specific development problems

D*S Design & Strategy

- 7** Identify individual and organizational objectives towards a Big Data strategy
- 8** Understand how to operationalize Big Data as projects, partnerships and policies
- 9** Recognize individual and organizational next steps towards Big Data applications

Ε Ethics & Engagement

- 10** Identify models for prioritizing inclusivity, transparency and accountability in data public-private-people partnerships
- 11** Articulate and assess ethical, privacy and legal implications of Big Data applications
- 12** Understand key principles for effective data communication and story-telling



BUILDING
BLOCKS
TOOLKIT
OVERVIEW



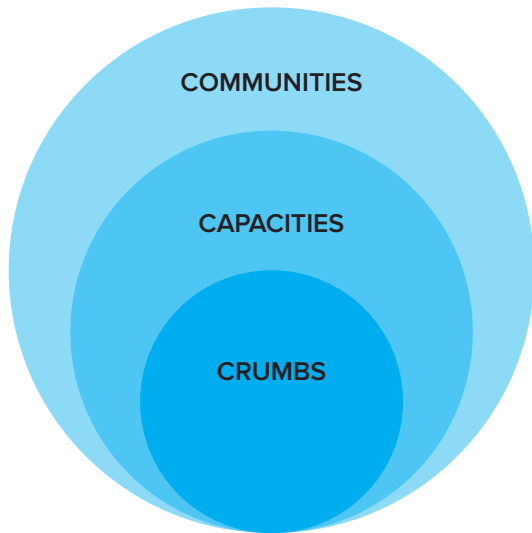
CONTEXTS & CONCEPTS

Understanding key Big Data ideas in order to translate development problems into specific data objectives

Key learning points

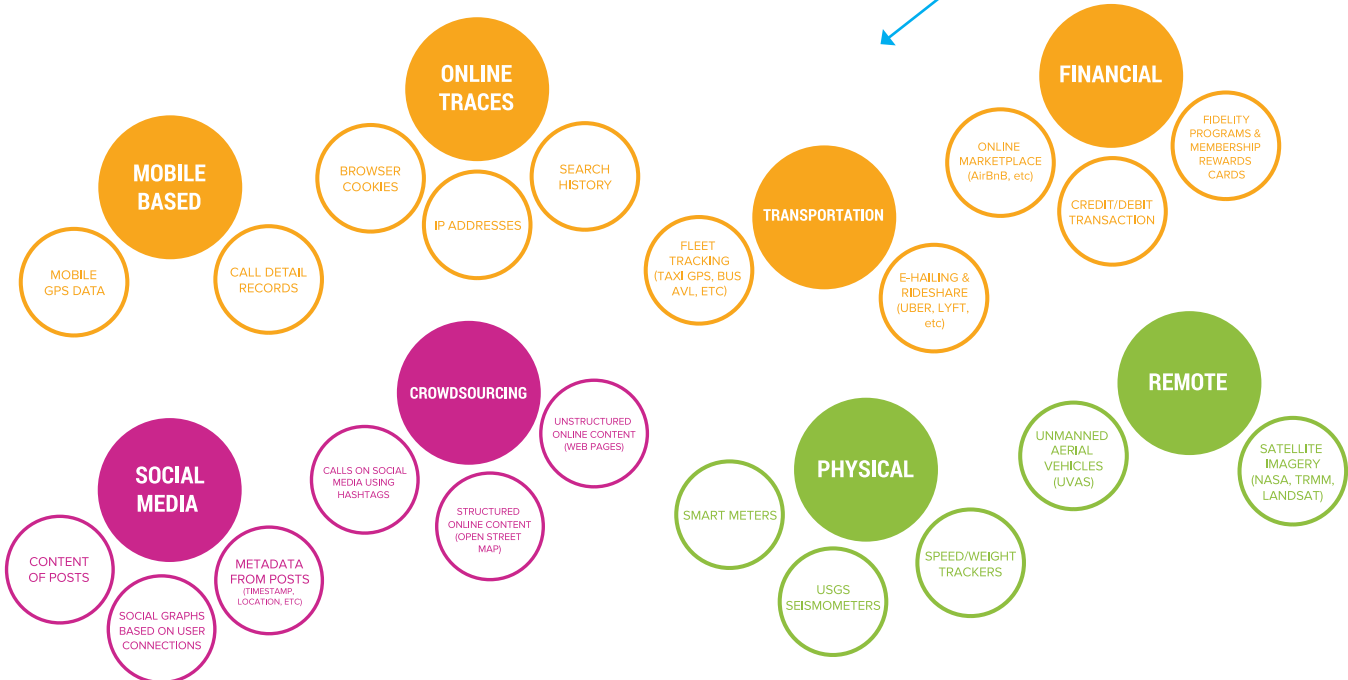
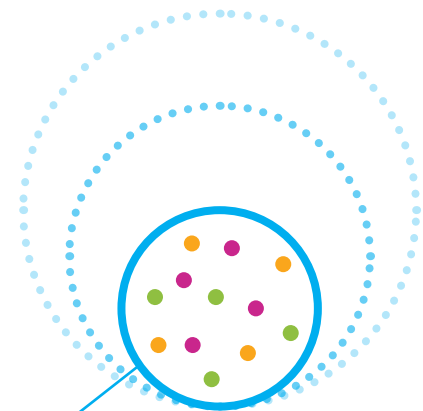
- 1 Decode key terms and buzzwords in the Big Data and development landscape
- 2 Discuss Big Data within the broader political context of the post-2015 sustainable development framework and data for social good
- 3 Translate development problems into specific data objectives

THE 3 C's OF BIG DATA

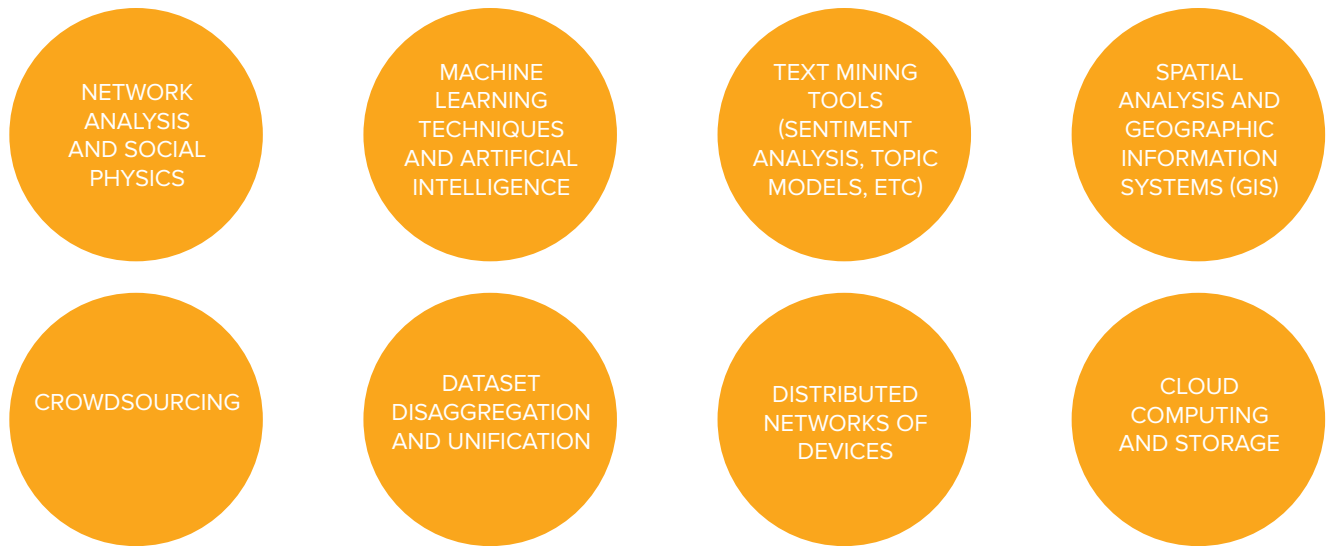
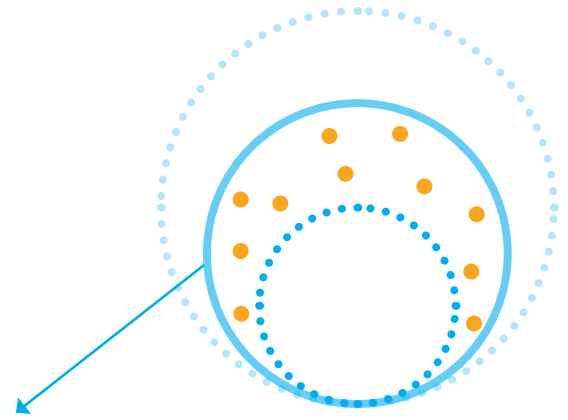


Originally framed as the “3 V’s” (volume, velocity and variety) in the early 2000s, we consider Big Data as an ecosystem of “3 C’s”: digital “crumbs” i.e. digital translations of human actions and interactions captured by digital devices; powerful capacities to collect, aggregate and analyze data; and communities involved in generating, governing and using data, including data generators, end users, policy-makers, experts, privacy advocates and civic hacker communities.

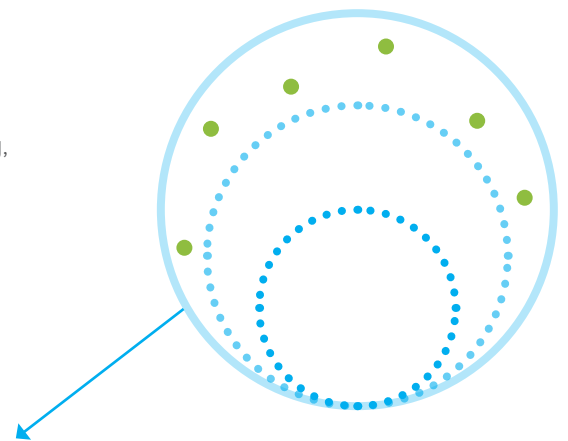
The **C of crumbs**—i.e. “digital bread crumbs” (a term coined by Alex “Sandy” Pentland), translations of human actions and interactions both passively emitted and captured by digital devices and other devices. At the center of our information societies is the production of massive amounts of data through connected platforms, social networks, and machines. This feature is important as it presides over a fundamental qualitative shift as much as a quantitative one and gives Big Data its deeply political nature.



The **C of capacities**—i.e. tools and methods to collect, aggregate, analyze and visualize. ‘Capacities’ act as a bridge between the crumbs and communities, helping to analyze both innovative, alternative uses of data as well as affecting how Big Data communities engage with and crowd around new knowledge and methodologies.



The **C of communities**—i.e. all stakeholders involved in generating, governing and using data, including data producers, end users, policymakers, civil society, experts, privacy advocates and civic hacker communities, as well as anyone represented in a dataset.



BIG DATA SOURCES FOR DEVELOPMENT

| | TYPES | EXAMPLES | OPPORTUNITIES |
|-----------------|------------------------------|--|--|
| Exhaust | Mobile-based | Call Details Records (CDRs) GPS (Fleet tracking, Bus AVL) | Estimate population distribution and socioeconomic status in places as diverse as the U.K. and Rwanda |
| | Financial transactions | Electronic ID E-licenses (e.g. insurance) Transportation cards (including airplane fidelity cards) Credit/debit cards | Provide critical information on population movements and behavioural response after a disaster |
| | Transportation | GPS (Fleet tracking, Bus AVL) EZ passes | Provide early assessment of damage caused by hurricanes and earthquakes |
| | Online traces | Cookies IP addresses | Mitigate impacts of infectious diseases through more timely monitoring using access logs from the online encyclopedia Wikipedia |
| Digital Content | Social media | Tweets (Twitter API) Check-ins (Foursquare) Facebook content YouTube videos | Provide early warning on threats ranging from disease outbreaks to food insecurity |
| | Crowd-sourced/online content | Mapping (Open Street Map, Google Maps, Yelp) Monitoring/ Reporting (uReport) | Empower volunteers to add ground-level data that are useful notably for verification purpose |
| Sensing | Physical | Smart meters Speed/weight/mail trackers USGS seismometers | Analysis of shipment trackers to infer trade patterns and trends |
| | Remote | Satellite imagery (NASA TRMM, LandSat) Unmanned Aerial Vehicles (UAVs) | Satellite images revealing changes in, for example, soil quality or water availability have been used to inform agricultural interventions in developing countries |

BIG DATA AND OFFICIAL STATISTICS

National Statistical Offices (NSOs) remain a pivotal actor in the ongoing evolution of official statistics and the achievement of the SDGs. New data sources—such as social media data, cell-phone data, satellite data, etc.—have created new opportunities and challenges for the production of statistics, their dissemination and engagement with beneficiaries; and leading to discussions about a new set of responsibilities that goes beyond pure measurement towards informing or even creating knowledge within societies¹. Simultaneously, NSOs are getting prepared for a new task: the “Data Revolution”; this global development puts them at the center of the Post-2015 agenda, and their contribution in measuring the Sustainable Development Goals (SDGs) will inevitably be important.

There is certain excitement that Big Data could be one element to help NSOs fulfill their responsibility. The advent of Big Data will influence the business of organizations whose core business lies in the production of statistical data. Not surprisingly, the discussions on “Big Data and Official Statistics” originated within NSOs statistical systems are well-established. However, in developing countries, many NSOs are still struggling with basic operating challenges, such as access to administrative data, poor collaboration between different governmental agencies, poor financial resources and capacities, and the absence of legislative frameworks. These challenges question the extent to which NSOs might be capable to actively engage with the Big Data.

NSOs are governed by legal democratic frameworks and have the general tools and know-how to work

with data in the most sensitive manner, under the premise of contributing to the wellbeing of societies in accordance with the first of the Fundamental Principles of Official Statistics. That is why they need to be key players in shaping the Big Data ecosystems of their respective countries and regions. In countries where they are recognized as trusted third parties, they will be crucial in the context of sharing data and forming a counterbalance to the interests of the private sector and governmental actors, including safeguarding privacy and the quality of the data.

Even from an opportunistic perspective, it would only be reasonable for NSOs to engage with Big Data as it becomes more important and as governments across the globe exercise influence in this field. If NSOs show leadership and become authorities on Big Data, they might receive the recognition and prioritization from governments they so urgently need (and with that more resources). Big Data can be strategically important to NSOs in several other respects. Given their likely higher level experience in developing techniques and standards related to data collection, curation and release (for example, metadata and data anonymization), NSOs will have a clear role to play in issuing guidelines in these areas for their own statistical products and for other agencies in the national statistics system.

To fulfill this role, NSOs need to actively engage the Big Data ecosystem to ensure that the yet-to-be-defined path of Big Data leads towards societal progress. The measurement of the SDGs will be an important task for the next fifteen years, and there is certainly evidence that Big Data could help NSOs fulfill this responsibility.

Major challenges and barriers persist for NSOs to leverage Big Data:

- 1 Institutional barriers to innovation and change management**, including the lack of an internal digital culture, skeptical institutional outlook on new data sources, and lack of coordination among stakeholders;
- 2 Constraints to data access and completeness**, particularly in access and continued use of private sector data, lack of public private partnerships and limited ownership rights involving people and their relationships with data;
- 3 Technical challenges**, including infrastructure for capturing, cleaning, processing, analyzing and visualizing both structured and unstructured data as well as adoption of specific IT tools and techniques;
- 4 Human capacity gaps**, including talent discovery, data literacy, limited data science training programs, and limited involvement of universities and other academic institutions;
- 5 Methodological challenges**, including challenges in data representativeness, biases, and the lack of standards and guidelines;
- 6 Ethical and political risks**, including risks to privacy and weak legal frameworks.

1. Giovannini, Enrico. "Statistics 2.0 - The next level." In: 10th National conference of statistics. 2010. URL: http://en.istat.it/istat/eventi/2010/10_conferenza_statistica/.

.....

THE GLOBAL PARTNERSHIP FOR SUSTAINABLE DEVELOPMENT DATA (GPSDD)

Reliable and up-to-date data has improved significantly over time due to new technologies, which have increased the volume, level of detail and speed of available data on societies, the economy, and the environment. Yet despite considerable progress in recent years, there are still entire groups of people, organizations, and governments who are excluded due to the lack of resources, knowledge, capacity and opportunity, which has caused growing inequalities related to the access and use of data.

In response to this changing technological landscape, UN Secretary-General Ban Ki-moon convened 24 international experts in August 2014 to propose ways to improve data for achieving and monitoring sustainable development. In the report “A World That Counts,” the Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG) highlight two global challenges related to the current state of data. The first is the challenge of invisibility (i.e. gaps in what we know from the data, and when we find out), and second the challenge of inequality (i.e. gaps between those with and without information, and what they need to know to make their own decisions).

In response to these challenges, the IEAG report calls for a UN-led effort to mobilize a “data revolution” for all people and the whole planet in order to monitor progress, hold governments accountable, and foster sustainable development.

The IEAG also offers several key recommendations on how to mobilize a data revolution for sustainable development and address these challenges:

- 1** Fostering and promoting innovation to fill data gaps.
- 2** Mobilizing resources to overcome inequalities between developed and developing countries and between data-poor and data-rich people.
- 3** Leadership and coordination to enable the data revolution to play its full role in the realisation of sustainable development.

Regarding the third point on leadership, the IEAG proposed the creation of the Global Partnership for Sustainable Development Data (GPSDD) to “mobilise and coordinate the actions and institutions required to make the data resolution serve sustainable development.” The Global Partnership was launched in September 2015, and adopted by world leaders during the 70th session of the United Nations General Assembly. The GPSDD is an open, independent, and multi-stakeholder partnership comprised of a network of more than 150 data champions, representing both data producers and users working around the world to harness the data revolution for sustainable development. Members include governments, companies, civil society groups, international organizations, academic institutions, foundations, statistics agencies and data communities.

The Global Partnership serves as a forum to to convene, connect and catalyze key stakeholders to exchange best practices, learn from one another, and drive progress to fill data gaps, and improve the accessibility and usefulness of data in order to achieve the Sustainable Development Goals and other national development priorities. The primary goals of the Global Partnership are to:

- **INCREASE** the demand and supply of credible, accurate, dynamic and disaggregated data.
- **ADDRESS** data gaps in terms of quantity, timeliness, credibility and quality.
- **INCREASE** the capacity, accessibility and effective use of data for decision-making, empowerment and accountability.
- **PROMOTE** broad dissemination and scaling of new data innovations and technologies.
- **CONNECT** stakeholders across sectors as well as across data communities to harness the data revolution for sustainable development and leave no one behind.

The Global Partnership convenes several working groups dedicated to filling data gaps and supporting national data revolution roadmaps, action plans and collaboratives, crafting data principles and protocols, strengthening data architectures and platforms, as well as mobilizing and aligning resources. The Global Partnership is also working with Data Champions to develop a toolbox that supports country-led efforts to develop and implement roadmaps for sustainable development.

Until recently, an Interim Steering Group of Anchor Partners governed the Global Partnership while a small Interim Secretariat managed day-to-day operations. Claire Melamed was named the permanent Executive Director of the Global Partnership in September 2016. The United Nations Foundation is the institutional host for the Global Partnership. Major funding for the Global Partnership is provided by the William and Flora Hewlett Foundation, Ford Foundation, International Development Research Centre and World Bank, as well as the Children's Investment Fund Foundation and US government, through the President's Emergency Plan for AIDS Relief.

-
1. Independent Expert Advisory Group on a Data Revolution for Sustainable Development. (2014). A World That Counts: Mobilizing The Data Revolution for Sustainable Development (Data Revolution Group). United Nations Independent Expert Advisory Group on a Data Revolution for Sustainable Development. Available online: <http://www.undatarevolution.org/report/>
 2. UN Data Revolution Group. Data Revolution Report: A World That Counts. <http://www.undatarevolution.org/report/>
 3. For the list of Data Champions, see: <https://static1.squarespace.com/static/55f7418ce4b0c5233375af19/t/579928ce2e69cfde233ba9be9/1469655246973/GPSDD+Data+Champion+List%28WEB%29.pdf>
 4. For more on the Data4SDGs Toolbox, see: <http://www.data4sdgs.org/toolbox>
 5. Global Partnership for Sustainable Development Data. (2015). FAQs. Available online: <http://www.data4sdgs.org/faq/>



METHODS & TOOLS

Applying Big Data methods and tools to yield insights for specific development problems

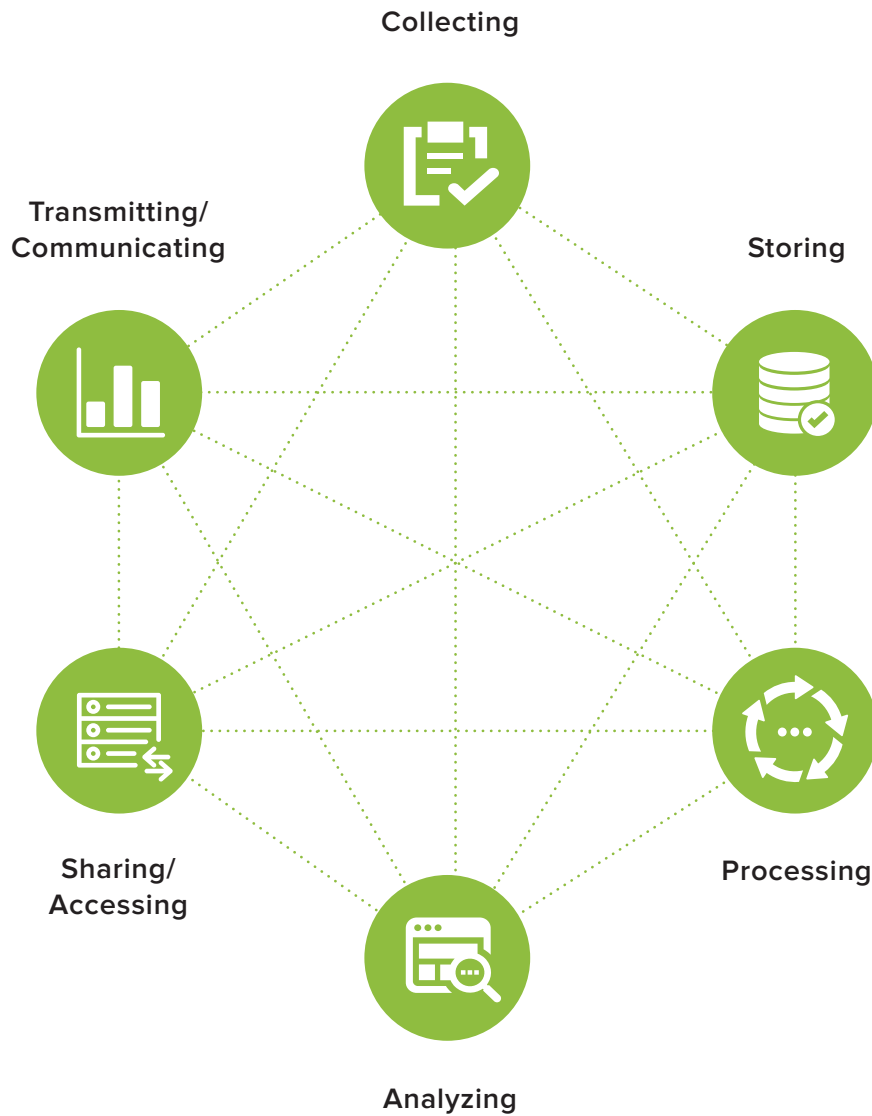
Key learning points

- 1 Understand existing methods and tools used to leverage Big Data
- 2 Assess data representativeness, biases and insights within Big Data-driven approaches and methods
- 3 Identify applicable tools by assessing the value add of Big Data for specific development problems

STAGES OF DATA USE

When using big data for a development project, it typically goes through a number of stages. The following diagram represents these stages of data use and provides key examples of tools that relate to each stage.

Stages need not occur in a specific order, and can be repeated, omitted, or occur simultaneously — in fact, as we will see in the following section, much of the art of project design involves choosing the appropriate combination sequence of stages to carry out a project that is feasible, ethical, impactful, effective and scalable, while taking into account the constraints, needs, and capabilities of the various stakeholders and beneficiaries.



Collecting

Gathering data from people or devices

Key Examples:

Tools for collecting data from people:

- Crowdsourcing tools: uReport, OpenStreetMap, Ushahidi.
- Survey tools: Kobo Toolbox.

Tools for collecting data from devices:

- Sensors on smartphones: Bandicoot, Funf library.
- Remote sensing: Google Earth

Storing

Keeping data for future use

Key Examples:

Tools for storing large volumes of data/information:

- Database tools: Hadoop, MySQL.
- Cloud storage solutions: Amazon S3, Dropbox, Google Drive, OneDrive.
- Version control systems: Git, SVN.

Tools for storing sensitive data securely:

- Secure data stores: OpenPDS.
- Trustless/decentralized storage: Bitcoin blockchain.

Processing

Turning raw data into intermediate data

Key Examples:

Tools for machines to execute instructions:

- Cloud computing platforms: Amazon Web Services, Google Earth Engine.
- Decentralized computing: MIT Enigma.
- Batch automation: IFTTT.

Tools for people to make sense of raw data:

- Data wrangling tools: OpenRefine, DataBasic.io.

Analyzing

Going from data to insights

Key Examples:

Tools for programming and statistical analysis:

- Programming languages: MatLab, SPSS, STATA, Python, R, Julia.
- Open source libraries: Scipy, Numpy, Pandas, Matplotlib, Plyr.
- Software and utilities: Rstudio, Jupyter, Anaconda scientific stack.

Tools for applying models and methodologies:

- Machine learning tools: Scikit-learn, Spark, Accord.NET.
- Sentiment analysis tools: Natural Language Toolkit.
- Geographic information systems: ArcGIS, QGIS, Google Earth Engine.

Sharing/Accessing

Making stored data/information available to others

Key Examples:

Tools for sharing/accessing data & code:

- Code hosting platforms: GitHub, BitBucket, CDNJS.
- Data pooling platforms: UN HDX.
- Application programming interfaces: Twitter Streaming API, Bit.ly Social Data API.

Tools for sharing/accessing sensitive data:

- Privacy architectures: OpenPDS.
- Governance structures: OPAL Project.

Transmitting/Communicating

Making stored data/information available to others or getting data shared by others

Key Examples:

Tools for transferring data:

- Standard formats: csv, json, geojson, xml, shp, xls.
- Encryption tools: AES, RSA, OpenPGP.

Tools for communicating information:

- Information and communications technology infrastructure: World Wide Web.
- Visualization tools: D3.js, Rshiny, Tableau.



UNPACKING THE "DNA" BEHIND BIG DATA PROJECTS: SEQUENCING THE STAGES OF DATA USE

Big Data Project DNA

To guarantee the success of a data project in development, the stages in which data is used must be carefully thought out in advance. Although there is some flexibility in how the stages are arranged, there are a number of specific sequences that help address particular complexities or requirements of a project. These sequences of stages -- which can be thought of as the DNA of a successful project -- provide templates addressing these challenges and make that a project is feasible, ethical, impactful, scalable, etc. Much of the innovation in the area of Big Data and Development, particularly as new data sources and uses cases emerge, relates to discovering new ways -- new DNA -- to structure a project while taking into account the constraints, needs, and capabilities of the various stakeholders [and beneficiaries].

A - Processing data in real-time



Overview:

In certain projects, some initial data processing/analyzing has to happen simultaneously with the collection process. This may be the case, for example, when more raw data is collected than can be stored/

transmitted safely and/or affordably. In such cases, it is often preferable to opt for an approach where the incoming data is processed on-the-fly: as the data is collected, it is processed automatically to extract key metrics or search for specific patterns that can help determine whether the data is worth retaining. The system then stores some combination of raw data and computed data (rather than retaining all the raw data) for future analysis.

Motivations:

This approach is especially interesting when the volume and/or velocity of incoming raw data make it impractical to store or transmit; or when there are security implications that require immediate deletion of the raw data.

Sensors such as cameras and accelerometers, for example, often collect more data than is actually needed in the long run; it can therefore be advantageous to assess, upon collection, and process the data enough to determine what can/should be retained. The limiting factor in what can be retained depends on the context: storage (ex: we have limited local storage space); bandwidth (ex: we have limited capacity to transmit data to remote storage); security (ex: we have restrictions on what is too sensitive to keep).

Challenges:

If raw data is not retained, it becomes difficult to go back and correct calculations, make changes to the process/algorithms, or allow others to verify/validate the process. Typically, these steps require access to the

raw data, which is not necessarily possible if only the computed results have been retained.

Real-time processing without data retention can therefore inhibit validation and replicability of the analysis. Such approaches therefore require careful initial design and testing to ensure that the processed data is reliable and useful.

As noted below, the removal of raw data is sometimes an advantage: if the data is sensitive, its deletion is often intentional.

B - Analyzing data behind a firewall



Overview:

In many contexts, the data an organization is storing on its servers can be shared with other groups to analyze it and come up with valuable insights for development projects - but sharing the data often has important ethical, legal, and strategic implications, which prevent the group from sharing raw data. It may, however, be possible to carry out some analysis in-house, and share the results in aggregated and anonymized form.

Motivations:

Approaches that involve conducting analysis on stored data before it can be shared are especially common when dealing with private sector data, such as call detail records, transaction data, etc. In such cases, the raw data is not allowed to leave the secure servers on which it resides, so if anyone wants to explore the data or run it through a model, this must be done behind the firewall. In order for outputs to be shared outside of the secure computing environment, they must meet the organization's guidelines, which typically require the extracted information to be highly

processed and aggregated, so as to avoid exposing any private/sensitive/secret information contained in the protected data.

This approach of analyzing data behind a firewall is therefore more complex than other methods (where data is shared directly for others to analyze it), but may be more suitable for projects where sharing is limited by ethical, legal, and strategic considerations.

Challenges:

The group hosting the data must have sufficient computing capabilities and expertise to carry out the analysis sufficiently. Alternatively, analysis from one of the partner groups may be brought in to assist in the process (through on-site or remote access to the data), which introduces a number of overhead, personnel, and compliance costs and constraints.

Additionally, such approaches may add complexity to the analysis process without necessarily eliminating all risk. For instance, simple anonymization (ex: removing the name, phone number, etc.) isn't always sufficient to ensure privacy, as it has been shown that many anonymisation approaches currently in use still make it possible to re-identify an individual with only a small number of data points.

C - Sharing analysis through a visualization dashboard



Overview:

Raw data can often be hard to explore and interpret, especially for people who don't have expertise in the particular formats and collection methods employed for a given dataset; yet many organizations have rich datasets that they wish to share in order to disseminate

the information and ideas that can be supported by the data. One approach that can help an organization share a dataset while making them more accessible to users is sharing the data through a visualization dashboard.

Motivations:

Building a dashboard allows the organization hosting the dataset to decide which analysis tools would be appropriate, and provides a level of control about how the data is communicated to users. For example: How is it aggregated? Which data points are available? What are the best ways to summarize/represent the information the data conveys? What models/tools can be used to make sense of the data?

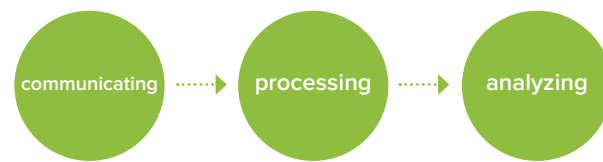
Dashboards can be interactive, allowing users to control a predefined set of parameters: variables, time periods, aggregation methods, smoothing, clustering, etc. If necessary, computations are performed on the host's server and the results of this analysis are communicated to the users (typically through a web browser or in an app).

The data analysis/communication tools are connected to the dataset, so changes to the input data can be reflected in the output in realtime.

Challenges:

This approach is not applicable to all situations. To start, the data in question must be appropriate for sharing, which is not always the case. This also requires computational power on the host's end, especially when the underlying data is being updated frequently, and the user is limited by whatever analysis functions have been set up and made available by the host. In some cases, communicating a message through data may be done more effectively through in-house analysis (i.e. analyze the data and figure out what to communicate, and then share).

D - Cloud computing



Overview:

Although personal computers and handheld devices can now perform fast and powerful computations, larger specialized computers are typically even faster and more powerful. In some cases, when processing and analytical power are very important to the project, it can be advantageous to transmit information to a specialized computer, often referred to as being “in the cloud” because it does not need to be located onsite. Cloud computing refers to the use of high performance computing environments that perform processing and analysis for the devices they are connected to, often with the aim of increasing performance and/or expanding the possibilities of what analysis can be performed on the data.

Motivations:

Even for real-time applications (for example, when data is collected on a handheld device, analyzed, and displayed to the user on the device), it is sometimes advantageous to have the analysis performed “in the cloud”. This may be the case, for example, if the device is not powerful enough to accomplish the analysis or if doing so would take too long or put strain on its resources (battery, memory, etc).

Using cloud computing can also add a degree of flexibility to the project, because there is currently a market for cloud-based resources, meaning that organizations can usually rent cloud-based resources as their needs increase/decrease. This helps organizations avoid steep upfront costs (resources and expertise) of setting up and maintaining their own high performance computing environment.

Challenges:

One major challenge of cloud-based approaches is that they require connectivity. Any process relying on resources in the cloud will therefore be unavailable or limited if either the device or the server itself goes offline, or if bandwidth is limited. There is thus a major tradeoff: centralized systems can potentially be more powerful, but also require connectivity and bandwidth (i.e. functionality is highly dependent on the central server) meaning that cloud computing may not be the best approach for all applications.

As mentioned above, organizations often outsource the task of setting up and maintaining their high performance tools by renting cloud-based services from companies that specialize in that market. Using a third-party service of this type can have security and quality implications, and it is up to the organization to exercise judgement on whether or not that is appropriate for their context and type of data. In cases where security precludes the use of third-party services, organizations can look into hosting their computing environment to perform analysis on local servers.

E - Using crowdsourcing tools to gather data



Overview:

Crowdsourcing (a portmanteau of ‘crowd’ and ‘outsourcing’)¹ refers to “a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.”² The tasks performed by the user typically involve collecting/creating quantitative or qualitative information, and reporting it to the organization through an app, web interface, or SMS platform.

Motivations:

Crowdsourcing, or soliciting data from the public, is a key opportunity for gathering a broad range of data. The data users are asked to report can be factual information about their surroundings (ex: features of business they’ve visited, locations and names of streets and parks in their community, etc) as well as more qualitative and subjective information (such as their perception of food insecurity in their community or the extent of property damage after a disaster). For example, citizen science literature has demonstrated that certain types of scientific data can be reliably gathered by distributed networks of non-specialists. In particular, even a small number of dedicated users can bring large increases in understanding in data-scarce areas.³

Crowdsourcing approaches can also be used to complement other data sources: when using traditional data as a baseline, crowdsourced data can help provide an extra layer by contributing information that may be less robust but more fine-grained and dynamic. In some contexts, an approach that combines both types of data can produce richer results.

Challenges:

Not all information can be crowdsourced: there are specific instances in which it makes sense to leverage non-experts to gather information and insight, but the challenges of dealing with data from disparate sources (ex: not everyone reports things the same way; not everyone has the same information or knowledge; not all the data can be verified or questioned; you always know much about those who participated, and you know nothing about those who didn’t) may sometimes outweigh the benefits.

Studies on prevalence of social media in developing countries have shown that analyses that work well in upper and middle income countries may falter in poorer countries with much thinner and more skewed user bases⁴. Since social media is often used for

crowdsourcing data, it is important to stress that crowdsourced data is not always representative, and this fact should be taken into account when opting to crowdsource data, when designing and implementing the collection process, and when analyzing and acting upon the results.

F - Data pooling



Overview:

When organizations have data to share, they can choose to host the datasets on a server and make them available for download. However, there are also existing platforms they can upload their data to. Such platforms are used as data aggregators, to gather and store datasets (usually focused on a common topic or source), and make it easier for potential users to find and access the datasets. Data pooling platforms can host raw data, processed data, or some combination of both, depending on the intent of the platform.

Motivations:

Data pooling platforms typically put the data through basic processing before it is stored and made available for access. Processing may involve, for example, standardizing the variable names and data formats (ex: dates, decimal numbers, measurements with units, etc) and computing descriptive statistics (min/max/mean/median/mode of numerical columns, number of missing values in each column, list of possible values for each column, etc). This makes it easier for users of

the platform to explore datasets and decide which ones to download; the standardization aspect also makes it easier to merge multiple datasets from the data pool.

Pooling multiple datasets in one platform, especially if there is a thematic link between them, makes it easier for users to find data they need, and also gives data providers more visibility since they can benefit from the name recognition of the platforms.

Although many data pooling platforms have open access for both uploading and downloading data, platforms can also allow editors to curate the content by approving/rejecting datasets, and host discussion forums about the dataset so that users can discuss the datasets, ask for clarifications, point out errors, and showcase projects they've done with the data.

Challenges:

The main limitation of data pooling platforms is that not all can be safely shared in this way. They also put a lot of responsibility on the platform host to maintain the storage and monitor the activity in the data pool (if curation and/or moderation are part of the platform's offering). One particular consideration is whether access to the platform will be provided free of charge or whether users/contributors will be asked to share the setup/maintenance costs.

Processing the data also requires making decisions about how the data will be used. Indeed, this decision reflects a tradeoff: processing data before it is distributed offers the possibility of standardizing it so that it is easier to explore and work with, but also gives users less liberty with what analysis they can run on the data since they are not starting from raw data.

-
1. <https://en.wikipedia.org/wiki/Crowdsourcing>
 2. Howe, Jeff (June 2, 2006). "Crowdsourcing: A Definition". Crowdsourcing Blog. Retrieved January 2, 2013.
 3. Data-Pop Alliance. "Big Data for Resilience: Realising the Benefits for Developing Countries". Synthesis report. September 2015.cite: DFID report
 4. Roth & Luczak-Roesch 2015, in Data-Pop Alliance. "Big Data for Resilience: Realising the Benefits for Developing Countries". Synthesis report. September 2015.

STAGES OF DATA USE IN PRACTICE

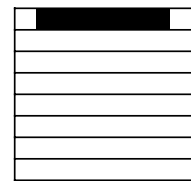
Layering Geographic Data

Using multiple datasets can produce nuanced, real-time insight into populations in specific geographic areas. The example of using geographic data for assessing vulnerability of populations is shown below to reflect the stages of data use for a specific application. The process of extracting indicators from census and topology/weather data and explore real-world applications of spatial analysis are:



Collecting

Gather data from census, topographic/ weather data.



Is everyone represented (accurately) in the data?

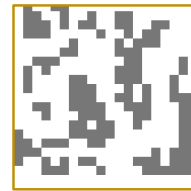


Processing

Clean and process data to build layers of vulnerability with a common geographic structure.



Bio-physical vulnerability



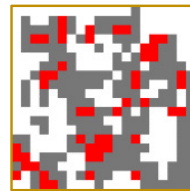
Social vulnerability

How is vulnerability defined, and is that definition relevant in context?



Analyzing

Merge layers and choose vulnerability thresholds to identify high-risk areas.



Does intersecting these two layers accurately capture the phenomenon we were looking for?



Communicating

Build and package map as a shapefile to communicate results about highly vulnerable areas.



Bio-physical vulnerability

Can and should these results be translated into policy action or further research?

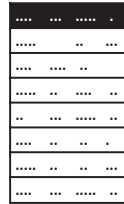
Call Detail Record Analysis

Other data sources can be layered or added to provide further insights. For example, mobile phone data can be used to understand how patterns observed across time and space can serve to observe the effects of a large-scale disaster.



Collecting

Gather call detail records.

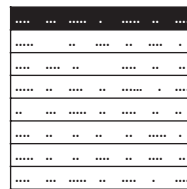


How do we ensure privacy and obtain user consent for this type of study?



Processing

Clean and process data to build indicators of call volume in each area.

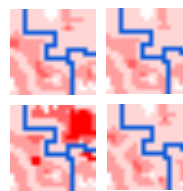


Does our cleaning/processing take into account how people actual use their phones? Is our process for inferring a user's home location effective and relevant in context?



Analyzing

Analyze indicators across time and space to identify characteristic signs of people responding to a flood.

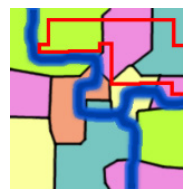


How do we develop robust models for differentiating the high activity right before a disaster from, say, national holidays where high activity is expcted?



Communicating

Build and package map as a shapefile to communicate results about where the flood occurred.



Can the results of this type of study be deployed effectively and responsibly on a large scale (for example, to assess awareness about the oncoming flood)?



DESIGN & STRATEGY

Operationalizing

Big Data as inclusive projects,
partnerships and policies

Key learning points

- 1 Identify individual and organizational objectives towards a Big Data strategy
- 2 Understand how to operationalize Big Data as projects, partnerships and policies
- 3 Recognize individual and organizational next steps towards Big Data applications

PROJECT ARCHETYPES FOR BIG DATA AND DEVELOPMENT

Citizen Science Tools

Examples: Crowdsourcing platforms (Ushahidi, U-Report, Open Street Map, etc), Hackathons (Kaggle)

| Purpose | Resources | Design |
|--|--|---|
| Gather data and derive knowledge from dispersed network of nonprofessional experts | <p>Technical: platform/dashboard, user-centered design and interface; technical maintenance</p> <p>People: context specialists, survey designers, community organizers, developers, data scientists, data ethics/governance specialists, communications specialists</p> <p>Data: crowdsourced data, open data</p> | <p>Key considerations based on real-world experience:</p> <ul style="list-style-type: none"> • Quality of data collected • Privacy protection and informed consent • Individual and group privacy risks • Representation and diverse participation • Citizen engagement campaigns |

Case Study: Using U-Report to Respond to Ebola in Liberia

In March 2014, Liberia began facing an unprecedented EVD (Ebola Virus Disease) outbreak that lasted for more than one year, affecting 10,048 people, claiming the lives of 4,421. In November 2014, UNICEF launched a text-message based interactive social media platform, U-Report, to give a voice to youth to access and provide information about Ebola. The initiative equipped mobile phone users with the tools to establish new standards of transparency and accountability in development programming and services, using simple text messages. U-report is one of the applications using Rapid-Pro, an open-source software platform. UNICEF partnered with a local organization present in all 15 counties in Liberia through youth networks, FLY (Federation of Liberian

Youth), to help recruit U-reporters. They brought in the UNICEF C4D (Communication for Development) team to expand recruitment to all levels of the community with the objective of: designing better interventions, monitoring activities, tracking rumors and empowering audiences. U-report allows citizens to provide feedback through a forum to amplify their voices, send alerts to key stakeholders about issues their constituents are facing, so they are empowered to work for change and improvements in their localities themselves. A few months after the launch, U-Report caught on quickly, and by June 2015, almost 51,000 U-reporters were feeding live information from their communities.

Source: <https://www.rapidsms.org/projects/ureport/>

Data Challenges

Examples: D4D Challenges, BBVA Challenge, Telecom Italia Big Data Challenge, Yelp Dataset Challenge

Purpose

Provide specific datasets to derive knowledge and interpretation from dispersed network of academics, professionals and nonprofessional experts

Resources

Technical: API/batch datasets for data access;

People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement

Data: Open government data, proprietary data

Design

Key considerations based on real-world experience:

- Public-private sector partnership strategy and implementation
- Informed consent
- Individual and group privacy risks
- Responsible data governance

Case Study: Orange Telecom's Data for Development Challenge

The First Orange D4D (Data for Development) Challenge was organized in Cote d'Ivoire over the span of a few months in 2012 and culminated at the Third NetMob Conference with a 1-day dedicated event at MIT in May 2013. It generated a surprising response, which revealed and spurred huge interest in the field of CDR (Call Detail Record) analytics. A total of 250 teams submitted proposals, out of which 83 submitted papers. The generated papers addressed questions about migration, poverty, public health, urban development, crisis response, demographics and economic statistics and more. There were four winners and 30 teams were granted permission to keep the data for further collaborative work.

Based on this success, Orange organized a second challenge in 2014-2015, focusing on Senegal. This challenge was designed keeping in mind feedback that came from the results of the challenge in

Cote d'Ivoire: including greater involvement and engagement of Senegalese authorities and a prize for the best ethical project.

While running the D4D challenges, Orange learned important lessons, such as: having the local community be a part of the challenge, seeking results that lead to concrete implementation in the area desired to benefit. Lastly, keeping issues of justice and ethics in mind, including aspects of consent and privacy is crucial throughout the process. Because there is a lack of legal and institutional infrastructure surrounding access to CDRs, cases are mostly solved on a case-by-case basis, based more on personal relationships than on existing norms that strictly regulate the way in which data is used. This exemplifies the shortcomings of the current state of the field and challenges moving ahead.

Intelligent Products

Examples: Business-to-business (B2B) solutions (Flux Vision, SmartSteps), data-driven decisionmaking tools (e.g. Google Flu Trends)

Purpose

Provide data-driven knowledge products based on closed datasets for analysis and decision-making by external experts

Resources

Technical: platform/dashboard, user-centered design and interface; technical maintenance

People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement

Data: Proprietary/closed data, data analysis and visualization

Design

Key considerations based on real-world experience:

- Translation from data to insights
- Transparency and accountability mechanisms (e.g. feedback loops)

Case Study: Flux Vision - the leading Big Data solution among tourism professionals

Designed by Orange Business Services, Flux Vision analyzes in real time more than 4 million pieces of new information per minute, and translates them into statistical indicators to provide metrics on geographic area visits and population movement. Today, more than 70 businesses and communities in fields such as tourism, transportation, trade and commerce rely on Flux Vision. In the field of transportation for example, Flux Vision provides experts additional insight into travel patterns and schedules, mapping out congestion points and delays or providing segmented data based on age and gender.

The algorithm, which ensures irreversible anonymity, was developed by engineers at Orange Labs to process technical data of the mobile network – including real time location. These technical data are displayed as indicators without any possibility to trace back to any individual. The development of Flux Vision has been regularly submitted to the “Commission Nationale Informatique et Libertés” (French regulatory authority in charge of protecting the confidentiality of personal information) at different design stages.

Source:

<http://www.orange-business.com/en/press/flux-vision-the-leading-big-data-solution-among-tourism-professionals>

Research Initiatives

Examples: World Bank, Universities (SafariCom-Harvard, Karolinska Institute, MIT-Twitter, Data-Pop Alliance), Flowminder/WorldPop Project

Purpose

Conduct data-driven research to derive knowledge, through the creation of new data methodologies and knowledge products

Resources

Technical: Batch datasets, secure storage & processing,
People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement
Data: Open government data, restricted proprietary data access, remote sensing, traditional data (including systems/records data, survey data, etc)

Design

Key considerations based on real-world experience:

- Ground truth verification
- Validity, representativeness and bias
- Informed consent
- Public benefit rationale

Case Study: Flowminder/World Pop - Nepal

The Flowminder/World Pop team has continuously distributed reports on population displacement in Nepal since the April 2015 earthquake. Reports have been produced together with UN Office for the Coordination of Humanitarian Affairs (UN OCHA), and been a key information source on the large-scale displacement taking place after the disaster.

Flowminder and Ncell (the largest mobile operator in Nepal) formed a partnership in December 2014, which would enable them to respond to future earthquakes and support long-term development objectives in Nepal. With support from the Rockefeller Foundation, the project was initiated, and rapid response capacity was set up by a Flowminder team in Kathmandu one week before the earthquake happened. After the earthquake, people fled affected areas, and a large number of people slept in open areas after losing their homes and over fear of aftershocks.

Flowminder researchers pioneered the use of anonymous mobile operator data to assess population displacement after the earthquake to understand where affected people were located - an essential factor for effective humanitarian response operations. The results were released as a report to the United Nations and a range of relief agencies in less than two weeks after the earthquake. Flowminder continues to analyze post-earthquake population movements in Nepal, and they are working with key organizations to ensure that capacity to respond with similar analyses in future earthquakes and disasters can be maintained.

Source: <http://www.flowminder.org/case-studies/nepal-earthquake-2015> CA.

Labs

Examples: Consortia (Urban Lab, UN Global Pulse Labs, HDX Labs), Innovation nodes and government taskforces (Etalab, White House OSTP, 18F, etc)

Purpose

Create and manage data-driven projects and initiatives relevant to the public sector through research and knowledge products, knowledge-sharing platforms and events

Resources

Technical: platform/dashboard, user-centered design and interface; batch datasets; technical maintenance

People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement

Data: Open government data, remote sensing, traditional data (including systems/records data, survey data, etc)

Design

Key considerations based on real-world experience:

- Transparency and accountability mechanisms
- Public benefit rationale

Case Study: Urban Lab

The Urban Living Lab, a partnership between the Bogota Mayor's Office and the Economic Development Secretary, the Bogota Chamber of Commerce and Data-Pop Alliance, was launched on May 27, 2016. The main objective of the partnership is to establish a joint basis and guidelines to create and shape a living lab that will be an open space of interaction and collaborative work, in which the interdisciplinary analysis of Bogotá's urban challenges results in innovative, applicable, and replicable solutions. This space seeks to be at the forefront of people-centered public innovation, by creating solutions in an inclusive manner, engaging the public and private sectors, academia, and above all, citizens.

The project will build on three main pillars:

Fostering citizen-driven innovation; Increasing citizen participation and empowering people to report, analyze and monitor their own community

Use of ITC's to understand people's needs and problems through data;

Creating collaboration networks among existing and emerging initiatives, in order to offer solutions to the city's relevant issues, resulting from the cooperation of the public and private sectors, academia and civil society.

These pillars will contribute to achieving the objective by increasing data literacy among individuals; becoming a resource that provides real-time data to support mass media content; and creating a platform that will allow citizens to feed in real-time information. While this partnership is fairly new, an example of a desired outcome for this joint effort is to work with the creation of policies to support the homeless population in Bogotá in collaboration with the Secretaría de Integración Social (Department for Social Inclusion).

Source: <http://datapopalliance.org/urban-living-lab-in-bogota-project-launch-a-collaboration-between-bogotas-economic-development-secretary-bogota-chamber-of-commerce-data-pop-alliance/>

Capacity and Community Building Activities

Examples: Workshops and trainings (School of Data, DataKind), Data access grants (Twitter Academic Data Grants, Yelp Academic Datasets, etc)

Purpose

Promote knowledge sharing on data and data literacy skills development through knowledge products, platforms, and events

Resources

Technical: platform/dashboard
People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement
Data: Sample data sets

Design

Key considerations based on real-world experience:

- Inclusive participation
- User-centered design
- Content adaptation
- Contextualization
- Impact assessment

Case Study: School of Data

School of Data is a network of data literacy practitioners - organizations and individuals - implementing training and other data literacy activities in their countries and regions. Members of School of Data work to empower civil society organizations (CSOs), journalists, civil servants and citizens with the skills they need to use data effectively in their efforts to create better, more equitable and more sustainable societies. They have produced or are in the process of producing dozens of articles, lessons and hands-on tutorials on how to

work with data. This, along with trainings events and mentoring have helped individuals and organizations become more data driven. Over the past four years, School of Data has succeeded in developing and sustaining a thriving and active network of data literacy practitioners in partnership with our implementing partners across Europe, Latin America, Asia and Africa.

Source: <http://schoolofdata.org/team/>

Events

Examples: Conferences, convenings and meetups (UN World Data Forum, Cartagena Data Festival),

Purpose

Promote collaboration, knowledge sharing and idea generation involving data through event activities and presentations.

Resources

Technical: digital communication and coordination tools

People: government, NGOs, civil society, private sector, academia, advocacy

Data: sample datasets

Design

Key considerations based on real-world experience:

- Inclusive participation of leaders in the field
- Stakeholder analysis and engagement

Case Study: Cartagena Data Festival

The Cartagena Data Festival, a three day festival in Colombia, was organized by ODI, Africa Gathering, CEPEI, Data-Pop Alliance, PARIS21, UNDP and UNFPA. The Festival brought together 300 participants from across the world – including government representatives, civil society organizations, technical innovators, academics and data activists – to join the global conversation and ensure the data revolution is informed by perspectives at every level. The event focused on solving critical gaps in coverage, access and analysis of data, thereby contributing to the global effort to

drive progress in the post-2015 agenda. Conference objectives included:

Driving the changes that are needed to advance a data revolution by bringing together the people and organisations whose innovations, resources, expertise and influence can make them happen;

Developing concrete solutions and practical tools to produce long-term and sustainable progress through a data revolution;

Building the ideas, innovations and partnerships needed to monitor the sustainable development goals

Source: <http://www.cartagenadatafest2015.org>

Coalitions and Collaboratives

Examples: GPSDD, PARIS21, Data-Pop Alliance, Health Data Collaborative, Data2x, Responsible Data Forum, Making All Voices Count

| Purpose | Resources | Design |
|--|--|--|
| Form multi-stakeholder collaboration around data applications and policies | <p>Technical: communications platform/dashboard, technical maintenance</p> <p>People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement</p> <p>Data: Sector-specific</p> | <p>Key considerations based on real-world experience:</p> <ul style="list-style-type: none"> • Partnership and incentives assessment • Short-term and long-term strategy • Inclusive participation of leaders in the field |

Case Study: Global Partnership on Sustainable Development Data (GPSDD)

The Global Partnership for Sustainable Development Data is an open, multi-stakeholder network working to harness the data revolution for sustainable development. With over 150+ partners, they seek to galvanize political commitment, align strategic priorities, foster collaboration, spur innovation and build trust in the booming data ecosystems of the 21st century.

The Global Partnership is working to build an enabling environment for harnessing the data revolution for sustainable development by:

- Strengthening data ecosystems by supporting countries to develop and implement roadmaps for harnessing the data revolution for development
- Mobilizing collective action through global data collaboratives and mechanisms; and helping catalyze new collaboratives where needed

- Mobilizing stakeholders to develop global data principles and protocols for sharing and leveraging privately held data

- Engaging in global meetings on statistics, data, the SDGs, and sustainable development to spur innovation and collaboration among various data communities

- Defining a clear action plan for data revolution resource alignment and mobilization

- Harmonizing data specifications and architectures, and helping ensure the interoperability of technology platforms for assembling, accessing, and using data.

Source: <http://www.data4sdgs.org>

Data Governance Frameworks and Policies

Examples: Policy directives (national data policies, strategies and briefs), standards associations (Creative Commons), oversight mechanisms, consortia (International Data Responsibility Group, Responsible Data Forum, Data & Society)

Purpose

Promote the implementation of responsible data governance principles, policies and frameworks

Resources

Technical: communications platform/dashboard, technical maintenance

People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement

Data: N/A

Design

Key considerations based on real-world experience:

- Inclusive participation
- Existing legal, ethical and policy frameworks and gaps
- Data ownership
- Governance and implementation structures

Case Study: International Data Responsibility Group

New data streams offer enormous potential to improve the livelihoods of affected populations in the developing world. However, as a growing number of actors attempt to harness this potential, the risk of irresponsible, unethical and inappropriate data use grows too. The potential negative implications of such misuse are significant and include loss of property, liberty, and life. To ultimately develop a credible framework and minimum standards for responsible data use, building a shared narrative is a crucial first step. No organization can single-handedly conduct the research required as well

as formulate and support such a narrative. The members to the International Data Responsibility Group (IDRG) are taking on this challenge collectively. The group is a global network of experts and organizations working on guidelines and frameworks towards responsible data use and sharing in the context of humanitarian action, sustainable development, and peace and justice. Its members seek to build a knowledge platform that enables responsible release, processing and use of data while minimizing risk for affected communities.

Source: <http://www.responsible-data.org/>

Platforms

Examples: Dashboards, Platforms, Mappings; UN Humanitarian Data Exchange, KoBo Toolbox

Purpose

Provide ease-of-use digital tools to collect, use and analyze data derived from multiple sources

Resources

Technical: platform/dashboard; technical maintenance

People: Data scientists, social science researchers, data ethics/governance specialists, partnership development, community engagement

Data: Survey data, open data, crowdsourced data

Design

Key considerations based on real-world experience:

- User-centered design and interface
- Accessibility and ease-of-use features

Case Study: Kobo Toolbox

Quickly collecting reliable information in a humanitarian crisis – especially following a natural disaster such as a large earthquake or a typhoon taking place in a poor country – is the critical link to saving the lives of the most vulnerable. Understanding the population’s needs is often neglected because of the lack of quick means to gather and analyze this crucial information. KoBoToolbox, developed by the Harvard Humanitarian Initiative, is a free, open source suite of tools for data collection and analysis in humanitarian emergencies and other challenging environments that was built to address this gap.

In September 2014 Kobo Toolbox launched a significant new phase of their software suite in coordination with the United Nations and the International Rescue Committee (IRC) to make electronic data collection more standardized, more reliable, and easier to use in humanitarian crises. Many additional features are still being added to the software every week to make the products more useful and relevant to a growing number of people.

Source: <http://www.kobotoolbox.org>

Mixed Media

Examples: Visualization, Data Journalism, Art installations, etc

Purpose

Provide ease-of-use digital tools to visualize data derived from multiple sources for communication purposes

Resources

Technical: context-specific

People: Data scientists, developers, graphic designers, artists, communications specialists, context specialists, journalists/media

Data: Open data, crowdsourced data, survey data

Design

Key considerations based on real-world experience:

- Communication and visualization strategy
- Representativeness and bias
- Advocacy considerations

Case Study: Time Machine

UNICEF's Data, Research and Policy division and Domestic Data Streamers have been working together to transform the lack of data on children into a tangible experience in order to shed light on their current situation and on the vital information we are missing on them today. The Time Machine was an experience that aims to make the world leaders and other participants at the 71st General Assembly look back to their own childhood and understand how they can change the childhood of millions

of children today. To do so, they have created a commitment contract and a commitment album. The commitment contract is a non-legally binding pledge that each participant signed between themselves and themselves as a child, committing to improve children's related data that will be publicly exhibited. The commitment album is a software that transforms each participant's data into a personalized song created from their childhood memories.

Source: <http://makingvisible.org>

Open Source Tools

Examples: Open source libraries (D3.js, Funf, Bandicoot); GitHub, CRAN

Purpose

Create incremental knowledge and leverage open collaboration models towards data-driven products and further data use

Resources

Technical: context-specific

People: Developers, testers, academia, data scientists

Data: context-specific

Design

Key considerations based on real-world experience:

[applicability in broad contexts; documentation and user engagement; R&D → integrating feedback and/or accepting contributions from the community]

Case Study: Bandicoot

Bandicoot, developed by researchers at the MIT Media Lab, is a python toolbox that provides a complete, easy-to-use environment for data-scientists to analyze mobile phone metadata. Bandicoot indicators fall into three categories: individual (e.g. number of calls, text response rate),

spatial (e.g. radius of gyration, entropy of places), and social network (e.g. clustering coefficient). With only a few lines of code, people are able to load their datasets, visualize the data, perform analyses, and export the results.

Source: <http://bandicoot.mit.edu/>

NAVIGATING YOUR BIG DATA INNOVATION STRATEGY ROADMAP

Big Data and Innovation

Big Data has increasingly been viewed as a critical lever of innovation for society. Much of the societal usefulness of big data comes from discoveries in secondary or alternative uses of passively collected data from companies, such as the use of call details records (CDR) or location data for humanitarian response and disease tracking. Enthusiasts of the “big data revolution” underline how big data can help society spot socially valuable insights, unlock new forms of economic value in data and uncover human and social dynamics.

However, the rules governing the Big Data ecosystem have been a source of constant debate in light of widespread corporate and government use of data that counter an individual’s right to privacy. In today’s increasingly connected, “big data” world, the emergence of big data problematizes several of our governing principles around data collection, sharing and consent; individual users—largely accustomed to ubiquitous use of data-emitting digital devices—are largely unaware of how and through which channels their data is used and processed, and the mechanisms used by companies to provide notice and consent (e.g. terms and condition agreements, privacy policies, etc.) have often failed to provide meaningful choice.

Private Sector Data Sharing for Public Good

In the last seven years, several case studies have emerged involving companies sharing “private sector data” through various modalities with the public sector towards social benefit (e.g. the Orange Data for Development Challenge, BBVA Innova Challenge, Twitter-MIT Laboratory on Social Machines, etc.); these

experiments have been presented as exciting new opportunities to understand evolving societal behavior and improve policy and practice at large.

These projects highlight specific dimensions of big data-driven innovation:

- **The potential of big data-driven innovation:** largely centered on issues of **anonymization** and **data protection; objectivity** and **validity** of big data’s claims; and the **diversity of expertise** needed to explore existing potential solutions and develop new tools and safeguards towards optimal data use.
- **The practice of big data-driven innovation:** existing **infrastructure and frameworks** in place to address issues of **access, governance, and ethics**.
- **The promotion of big data-driven innovation:** lessons learned from years of **public-private-people data partnerships and projects**—their formation, structure, maintenance and impact—and insights for companies and governments initiating these projects. Overall, these discussions focus largely on issues of **transparency, accountability, consent and literacy**.

Public-Private-People Data Partnerships: Knowledge Gaps and Opportunities

These case studies have sparked several global discussions on big data driven innovation, and several knowledge gaps have been identified:

- 1 **Unpacking new discoveries and possibilities of big data.** The need to help both companies and citizens understand the possibilities of the data revolution, through capacity-building and case studies of companies using data for social impact;

- 2 **Protecting consumer data.** The need to recognize the relationship between consumers and their own data, and help promote civic understanding of the data revolution through legal safeguards as well as the promotion of transparency-enhancing, privacy-preserving tools and platforms;
- 3 **Incentivizing the private sector.** The recognition of the private sector as part of larger societal aspect of the big data revolution and consider criteria towards responsible data governance for social purposes;
- 4 **Encouraging greater collaboration towards understanding the complexity of big data ethics and operationalizing ethical frameworks.** The need for multi-stakeholder collaboration and infrastructure to deepen conversation and understanding around “do no harm”;
- 5 **Understanding “lessons learned” from existing data partnerships.** The need for further exploration and analysis of the emergence of public-private-people data partnerships, and their role in facilitating the evolution of principles and modalities around responsible data use.

Though public-private partnerships are in no way a new phenomenon in governance innovation, the nature through which data (and its value) is shared, as well as the norms and practices around data use, will require setting new rules for collaboration and public-private business relations that emphasize societal benefits and lead towards scalable innovation in the public interest.

Navigating the Big Data Innovation Strategy Roadmap

The following roadmap reflects the numerous insights from existing initiatives and participant discussions, and highlights the requirements and milestones for initiating big data-driven innovation projects.

Invest in Big Data internal knowledge and support

- 1 **Focus projects and resources on solving public problems with clearest ethical imperative and problem definition.** Identify issue areas and public problems in which your data can play an important role towards decision-making and/or resource optimization. Additionally, look for “low hanging fruit” – data from your organization or company could even play a small but critical role in a specific issue area rather than solve a much larger problem.
- 2 **Analyse current risks and tradeoffs in various forms of data sharing to make the case for internal support.** Evaluate existing case studies in which companies have previously shared similar data. What modalities did they use to share data? How have others handled privacy issues in using data towards this particular issue? What kinds of risks and tradeoffs would exist in the creation of this specific project? How can this project both make social impact as well as support the organization’s overall objectives? This could take the form of sponsoring or hosting a series of expert workshops, participating in larger forums on big data and social impact or soliciting help from external consultants or researchers who would be embedded into your organization and teams. Ultimately, answering these questions internally is critically both for gaining support as well as building internal momentum and enthusiasm.
- 3 **Recognize organizational challenges and constraints in initiating innovation.** What resource constraints exist for your organization or company? How much funding or support is available and what kind of projects will be possible?
- 4 **Invest in knowledge management and implementation of privacy engineering solutions.** Evaluate the existing legal framework around privacy and learn more through existing solutions how to incorporate privacy-preserving elements into your project.

Invest in developing mutually beneficial Big Data partnerships

- 5 Catalyze mutually beneficial partnerships with the right stakeholders with diversity of expertise and capacity for thought leadership.** Several practitioners stress that many of the existing big data projects would not exist outside of years of building trust and taking steps (and risks) with the right stakeholders. Identifying the right stakeholders involved both evaluating expertise and thought leadership, as well as willingness to take on risks and process alignment.
- 6 Focus partnerships on determining best use of private sector data to enhance existing traditional data.** Using data sources such as call detail records, transaction data and other new data sources will require some form of ground truth data from existing traditional sources. Identify partners that can also provide data or capacity to analyse data from public or open sources.
- 7 Adapt features of existing ethical frameworks and guiding principles toward initial “do no harm” governance model.** While organizations and companies may aim to “do no harm” in the development of their projects, what guiding principles or ethical frameworks govern their use, and what happens when problems arise? While researchers continue to develop and assess ethical frameworks for data use (e.g. the Menlo Report), several companies and organizations have experimented with various models to incorporate ethical standards into their projects. This has involved the development of codes of conduct (in the case of BBVA, for example), organizing ethical roundtables (in the case of the second Data for Development Challenge organized by Orange Telecom), and incorporating an ombudsman to oversee data sharing projects.

- 8 Create and encourage mechanisms for public feedback and consultation.** Consult stakeholder groups to help inform the project design and provide opportunities for the public to give feedback.

Invest in meaningful civic engagement through multi-stakeholder Big Data literacy promotion

- 9 Communicate project results through data visualizations.** Evaluate how to best communicate project results for both intended beneficiaries and the public, using infographics and data visualizations for example.
- 10 Enable meaningful mechanisms for opt out.** Assess applicable notice and consent solutions and provide option for users to remove their data from use in the project.
- 11 Promote opportunities for shared language and principles among stakeholders** through participation in cross-sector, cross-disciplinary collaboration, platforms and events led by intermediaries
- 12 Invest in long-term civic engagement and public understanding** of how data is being used through data literacy efforts through media, government and civic outreach.



ETHICS & ENGAGEMENT

Engaging key stakeholders and communities through ethical practices and effective story-telling

Key learning points

- 1 Identify models for prioritizing inclusivity, transparency and accountability in data public-private-people partnerships
- 2 Articulate and assess ethical, privacy and legal implications of Big Data applications
- 3 Understand key principles for effective data communication and story-telling

MENLO REPORT: ETHICAL PRINCIPLES GUIDING INFORMATION AND COMMUNICATION TECHNOLOGY RESEARCH

The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research was first published in December 2011 and amended in 2012. It was a result of an initiative by the US Department of Homeland Security, Science and Technology, Cyber Security Division to adapt established ethical principles in the context of the ICT (Information and Communications Technology) and data revolutions. It identifies four key ethical principles for computer and information security research:

1. Beneficence;
2. Respect for Persons;
3. Justice;
4. Respect for Law and Public Interest.

| Beneficence | Respect for Persons | Justice | Respect for Law and Public Interest |
|---|---|---|---|
| <ul style="list-style-type: none"> · Do not harm; · Maximize probable benefits and minimize probable harms; · Systematically assess both risk of harm and benefit. | <ul style="list-style-type: none"> · Participation as a research subject is voluntary, and follows from informed consent; · Treat individuals as autonomous agents and respect their right to determine their own best interests; · Respect individuals who are not targets of research yet are impacted; · Individuals with diminished autonomy, who are incapable of deciding for themselves, are entitled to protection. | <ul style="list-style-type: none"> · Each person deserves equal consideration in how to be treated, and the benefits of research should be fairly distributed according to individual need, effort, societal contribution, and merit; · Selection of subjects should be fair, and burdens should be allocated equitably across impacted subjects. | <ul style="list-style-type: none"> · Engage in legal due diligence; · Be transparent in methods and results; · Be accountable for actions. |

1. Beneficence: Understanding risks and benefits

The principle of beneficence refers to “a moral obligation to act for the others' benefit, helping them to further their important and legitimate interests, often by preventing or removing possible harms.” Under this principle, researchers and project leads must maximize the probability and magnitude of benefits to individual research subjects as well as to society.

However, what constitutes a benefit or a risk is not always straightforward or consensual and depends to a large degree on the actors considered. For example, CDRs are largely stored and handled by private companies, which are often the ones investing in transmission and storage infrastructures. Commercial considerations must therefore be taken into account in framing the risks and benefits of using and sharing CDRs.

2. Respect for Persons

The issue of consent is gaining attention and is central to the privacy concerns relating to the use and sharing of CDRs. Specifically, users of mobile phone handsets rarely grant formal permission for their personal data to be used and shared. If they do so, it is often with little to no choice, since not consenting would limit their access to the technology. Furthermore, the choice given to consumers is typically to either dissent or fully consent, regardless of what use may be made of the data several years later, or by a third party should it be accessed by them. There is little to no way for consumers to exclude specific usage of their data that they do not want, raising major questions around the secondary use of data.

For secondary use, it is generally agreed that uses that are consistent with the original context can carry the permission granted in that context, but that new uses should require new consent. Broad (unlimited) consent remains widely used despite strong opposition on moral, ethical, and legal grounds. An even more advanced model proposes that individuals would permanently “carry” a set of permissions that they grant to algorithms seeking to use their data—no matter what data, enabling them to modify access and permissions at any time.

3. Justice: Bias and inequalities

The principle of justice highlights issues of fairness and equal distribution of risks and benefits. Arguably one of its key aspects is that everyone must have an opportunity to contribute and benefit (e.g. from CDR analytics) even when unequal access to technology exists. Yet, whose data is considered in CDR analytics is inherently affected by unequal access to and use of mobile phones, creating inherent biases and violating the principle of justice.

This creates yet another tension in CDR analytics: it is especially relevant in otherwise data-poor environments, but it is precisely in these environments that access to technology is most unequal, which implies that CDRs are non-representative data. The underlying challenge is that CDRs will typically reflect structural inequalities in any given country: owning a cell phone is strongly correlated with socio-economic status, and even in countries with high mobile phone penetration, CDRs may be analyzed along criteria that would single out more affluent individuals or areas. These biases hinder the external validity of findings based on CDRs and may potentially reinforce structural inequalities (if, for instance, programs are based on data from areas with high cell phone usage).

4. Respect for Law and Public Interest

The fourth and last principle framing our discussion highlights the need to engage in legal due diligence; be transparent in methods and results; and be accountable for actions. However, inconsistency of the legal or regulatory environment guiding the opening and use of CDRs across countries is problematic where legal protections are insufficient to protect the individual, and where cross-border accountability is difficult to enforce (e.g. if an individual is put at risk because of a foreign organization’s use of his or her CDRs, what recourse is available to that individual?).

Telecommunication companies (Telcos) are especially concerned about their legal exposure if CDRs were to be used to identify, target and/or discriminate against specific individuals or groups. In repressive environments, for example, Telcos should consider as their first priority to protect the sources of information (their customers) and place-sensitive data beyond the reach of authorities, even where this may be against their financial and commercial interests. At the same time, Telcos which have access to potentially life-saving information may be morally, if not legally, required to make that information available.

RISKS AND CHALLENGES IN BIG DATA PROJECTS

Risks to Privacy and Legal Considerations

- Protecting privacy of individuals and groups in all stages of the data lifecycle through technical and governance mechanisms
- Finding, selecting, and following applicable guidelines and frameworks related to data use when data is from multiple sources or partners, or reconciling regulations when multiple countries/organizations are involved and implicated.
- Addressing complexities of privacy protection as a result of a wide range of data sources and unforeseen risks that may not exist or be apparent in the individual datasets

Risks in Responsible Data Governance

- Defining fair, appropriate governance and oversight mechanisms
- Following responsible, ethical and legal standards of data use across project lifecycle (e.g. uphold principles of do no harm, beneficence, etc.)

Challenges in Representativeness and Participation

- Identifying and accounting for biases -- both in collection, analysis and communication of results -- as well as weighing the balance of the utility of certain tools (e.g. latent or ignored bias).
- Ensuring inclusiveness across the project lifecycle and encouraging engagement and access for all relevant stakeholders

Challenges in Data Protection and Consent

- Defining and ensuring appropriate protections around data collection, storage and use and consent throughout the project lifecycle as well as when datasets are selected for use after collection, when data is repurposed, etc. at both an individual and group level.

Challenges in Transparency

- Balancing complementary or competing motivations of data generators, custodians, users, etc; and effectively communicating tradeoffs.
- Establishing and enforcing transparent governance of the project with open processes, components, and information to the furthest extent possible



Risks to Privacy / Legal

Data ownership/ stewardship: Who owns the data, or is the custodian of the data? What legal responsibilities and protections exist in that regard?

Privacy Protection in Legal Context: Do the project's data collection methods meet privacy protection guidelines? What frameworks govern the use of

personal and group data relevant to this project?

Regional Privacy Protection Laws:

If multiple organizations are providing data, what is the best way to balance the laws/policies that apply to each data source?

Risks in Data Governance

Latent group discovery: What risks are associated with the revelation of latent groups in using and aggregating user data with available open data resources?

Do people know how their data is being used?
What do people know about how their data is being associated in a group or aggregated?

If the product is deployed across multiple countries, what is the process for balancing data collection/

storage with the laws of each location?

What is the balance of crowdsourced, collected (ex: sensor, survey), and official data used in the project?

If multiple organizations are involved, what is the process to ensure that each data provider follows relevant guidelines, and that use complies with collective guidelines?

Challenges in Representativeness and Participation

Selection bias: Who is using this? Does everyone have the same access to the platform to contribute?

Sample bias assessment/correction: Does the project

account for bias in where the data comes from (ex: market share for mobile operators) or attempt to assess/ communicate it?

Challenges to Data Protection and Consent

What are the data collection policies? How are people informed of them?

Informed consent: Are users made aware of when and why their data is being collected, and how it will be used?

Is consent affected by the usefulness of the product? What are the consequences for the user of denying consent? Do they have reasonable alternatives if they opt-out?

If using datasets that were already collected, is the current use in line with the uses the user originally consented to?

If the project involves data about individuals and groups, what safeguards can be put in place to protect them? And how can consent be ensured?

Challenges in Transparency

Is it clear to users what product/use they are contributing data to? (Do they get a service/reward in return for their contribution?)

Process transparency: Is information collected about users for analytics purposes (now or in the future)? Are users informed in an intelligible way of current and future uses (and implications) of any collected data?



Data Stage Storing

Risks to Privacy / Legal

Would the information you collected expose users (or others) to risks if it were seen or found by unauthorized people?

Risks in Data Governance

What are the relevant laws and policies that should inform how the data will be kept/stored?

Challenges in Representativeness and Participation

Invisible populations: In working with existing datasets or from databanks/records, is there good representation about who exists in the data? How is the coverage of the data discerned?

Data breaches and leaks: How do you protect stored data that isn't intended to be accessed at all?

Challenges to Data Protection and Consent

Data retention: Are users aware of how / how long their data will be stored

How long can data be retained? Can users opt out?

Challenges in Transparency

Collected Metadata: Is metadata about methodologies, user consent, and acceptable uses recorded alongside the dataset itself?



Data Stage

Processing/Analyzing

Risks to Privacy / Legal

Does the analysis require using/exposing sensitive data, and does this create risks to individual and group privacy?

Are there individual and group privacy risks that emerge from studying and using the data (in particular when bringing together a broad range of datasets)?

Are there individual and group privacy risks that emerge when bringing together a broad range of datasets?

Are there individual and group privacy risks that emerge when bringing together a broad range of datasets?

Risks in Data Governance

Is the data being merged with other datasets? How do we assess new risks created by merging multiple rich datasets?

Challenges in Representativeness and Participation

Consistency/quality of the data: Do all users report things the same way?

Is there a risk of the analysis producing outputs (results, recommendations, pricing, etc) that disproportionately affect certain individuals or groups (ex: algorithmic discrimination)?

Community engagement: Are the intended beneficiaries involved in the insights/decision process?

Challenges to Data Protection and Consent

How do you ensure protection and consent for data from the public space (ex: traffic cameras, transit data, in urban labs)?

Challenges in Transparency

Replicability: Are the methodologies replicable / open-source? Or is there a mechanism for community review/approval/validation?



Data Stage

Sharing/Accessing

Risks to Privacy / Legal

What are the safeguards to prevent unauthorized access during sharing of sensitive/ proprietary data, in order to protect individual and group privacy?

Risks in Data Governance

What are the safeguards to prevent unauthorized access during sharing of sensitive/ proprietary data, in order to protect individual and group privacy?

Challenges in Representativeness and Participation

Access bias: If the models/ insights are proprietary, how is access determined? Do any people/groups face disproportionate barriers to access?

Challenges to Data Protection and Consent

User expectations and intended use: Does the distribution of the data for external R&D conform with user's expectations? (Is there informed consent?)

Challenges in Transparency

Are the data inputs / products made available externally (for replication / validation)?



Risks to Privacy / Legal

How to ensure that the guidelines/laws of all entities (orgs, countries, etc) involved are taken into account and balanced/reflected in the resulting framework?

What legal and policy frameworks exist to guide how data & info can/should be shared (ex: balancing confidentiality, proprietary, sensitive; with free expression, journalistic integrity & responsibility, etc.)?

Risks in Data Governance

How does the framework act as a tool for making decisions about the applications and implications of data collection, storage, analysis, access?

Challenges in Representativeness and Participation

Public benefit rationale: Does the project help raise awareness of issues, build capacities, etc? Does the project have an engagement component?

What mechanisms could be used to disseminate insights to potential beneficiaries? (ex: should you make an API?)

Representativeness: Does the data provide an unbiased picture of what is going on? Are there ways in which the product could be misused/ misinterpreted?

How does one ensure that the framework's metrics/ processes/checklists can be generalized or adapted to the specific context, in order to avoid oversimplification, edge cases, invisibility, false appearance of objectivity, etc.?

How can individuals/ communities use the products for engagement/ advocacy? Are there foreseeable risks of making these tools available (ex: unintended uses)?

Challenges to Data Protection and Consent

Are the terms of consent accessible or digestible by producers of information? What are the best modalities for consent? Are alternative or future uses of data effectively communicated?

Challenges in Transparency

Freedom of information laws: Do any of the data or the results become subject to freedom of information laws? What risks might that create?

What are the trust mechanisms for validating and reviewing the decisions that underpin the framework?

If the tool is intended to convey a specific message, is that communicated transparently, or does it have the guise of objectivity/ neutrality?

Selected References

Adarve, L. H., Acosta, & Cala, Lina. (June 2016). Data protection in Colombia: overview. *Practical Law*. Retrieved November 22, 2016, from <http://us.practicallaw.com/2-619-4326>

Royal Statistical Society. (2016). The Opportunities and Ethics of Big Data (Workshop Report). Retrieved from <http://www.rss.org.uk/Images/PDF/influencing-change/2016/rss-report-oppo-and-ethics-of-big-data-feb-2016.pdf>

Steen, M. (30 November 2015). Ethical Uses of Collected Data. *Markkula Center for Applied Ethics, Santa Clara University*. Retrieved from <https://www.scu.edu/ethics/focus-areas/business-ethics/resources/ethical-uses-of-collected-data/>

the engine room. (2016). The Hand-Book of the Modern Development Specialist: Being a Complete Illustrated Guide to Responsible Data Usage, Manners & General Department. *Responsible Data Forum*. Retrieved from <https://responsibledata.io/resources/handbook/assets/pdf/responsible-data-handbook.pdf>

OPERATIONALIZING “DO NO HARM” INNOVATION IN A PRIVACY-CENTERED WORLD

Background: Privacy Principles in the Age of Big Data

In 1980, the Organization for Economic Co-operation and Development (OECD) published its Privacy Principles, the first set of global guidelines on personal data collection and privacy that have largely shaped existing national, regional and international policies and discussions even in the age of Big Data.¹ The governing principles of these guidelines—largely centered around notice and consent, purpose specification, and use limitation—emerged out of an era of client-server and mainframe computing in which personal data was more simply stored on computers and actively shared from one party to another.

As personal interactions with computers and capabilities in data storage have evolved in the last four decades in scale and use, the term “big data” has been used to characterize the rapid global changes in personal data use. Originally framed as the “3 V’s” (volume, velocity and variety) in the early 2000s, big data has emerged as an ecosystem of “3 C’s”: digital “crumbs” (digital translations of human actions and interactions captured by digital devices); powerful capacities to collect, aggregate and analyze data; and communities involved in generating, governing and using data, including data generators, end users, policy-makers, experts, privacy advocates and civic hacker communities.²

In today’s increasingly connected, “big data” world, the emergence of big data problematizes the governing

principles established by the OECD. Individual users—largely accustomed to ubiquitous use of data-emitting digital devices—are largely unaware of how and through what channels their data is used and processed, and the mechanisms used by companies to provide notice and consent (e.g. terms and condition agreements, privacy policies, etc.) have failed to provide meaningful choice. Businesses can passively collect personal data in ways that go beyond the original principles of the OECD guidelines. Government and public sector innovators who are exploring new, innovative ways to harness data towards greater economic growth and public good inherently are not able to identify at the outset of their efforts all the uses of the data being used.

Although the free-flow of data does promise new business innovation and tailored and efficient services for customers, global competition has largely hindered a genuine discussion on data responsibility and transparency among big data and innovation enthusiasts. How can we make full use of data analytics in a responsible and human-centered manner? What new infrastructure and policies should be put in place for both the private and public sector entities that collect, store and use personal data?

While the economic and societal benefits of the big data revolution have been underlined by the EU Horizon 2020 program and forthcoming initiatives under the EU Digital Single Market, privacy protection has often been illustrated as the opponent of innovation within regional policy conversations. While

many point to the potential benefits, others point to unknown costs—to individual privacy via personally-identifiable information (PII) in addition to group privacy via demographically-identifiable information (DII). These concerns specifically lie at the center of the EU’s recently adopted General Data Protection Regulation (GDPR)—in which EU policymakers and officials have debated how to find the balance between the opportunities of big data use and the risks to individual privacy protection. Although anonymization and privacy by design techniques have been described as potential privacy-preserving solutions, academic research has revealed that assumingly anonymized data could be “de-anonymized,” calling into question the reliability of current policy measures.³

Social physics and big data

Enthusiasts of the big data revolution underline how big data helps us spot socially valuable insights, unlock new forms of economic value in data and uncover human and social dynamics. Much of the societal usefulness of the big data being collected comes from discoveries in alternative uses of passively collected data from companies, such as the use of call details records (CDR) for humanitarian response and disease tracking.

In her work, Professor Linnet Taylor underlines the emergence of the private sector as a new actor within a longer historical context of “social physics,” and the need to consider how to involve these new actors in

participating with other stakeholders in measuring and understanding data on populations and society. The term social physics has been recently popularized by Professor Alex ‘Sandy’ Pentland in his pioneering work on big data, collective intelligence and social influence.⁴ However, social physics dates back to the late 1700s, coming back into the prominent view both in the 1830s and during the 1950s—“each time as part of...politicking around the right to know; the right to visualize; the right to count, sort and categorize...[and further on] the way we see populations and how we conceptualize the right to intervene and the process of intervening.”

Today, social physics arises again in the context of big data where the focus is no longer only on government and survey data emerging from public and academic sector actors, but rather proprietary, commercial data from the private sector. Big data invites us to exciting new opportunities to understand behavior and society as well as forces us to consider old questions of governance—who categorizes, who sorts, who intervenes, and who ultimately should be held accountable or responsible in these matters?

Though several case examples have emerged where companies have experimented in sharing data with the public sector for social efforts, little has been proposed on what kind of governance is needed to allow such data sharing and use. So far, private companies have been left alone with the decision of how and when data could be used to tackle societal issues.

Box 1. Leveraging Mobile Phone Data to Infer Post-Disaster Movements

Source: Data-Pop Alliance, 2015. "Big Data for Climate Change and Disaster Resilience: Realising the Benefits for Developing Countries," Synthesis Report.

Data from mobile phones constitute a precious source of information to infer population mobility after disasters.

- Tracking the position of mobile phones in Haiti before and after the 2010 earthquake allowed experts to estimate population displacement and helped build an effective real-time monitoring system to track the outbreak of infectious diseases.⁵
- After a 2011 earthquake, Statistics New Zealand mapped population movements by tracking text messages and voice calls.⁶ The experiment showed 'which geographical areas attract high percentages of people, patterns of return movements over time, and flows of non-residents into the emergency zone'. However, it did not assist in 'verifying residential areas people leave from, which areas people relocate to following an event, or the actual number of people who relocate, temporarily or permanently'.
- More recently, Flowminder used a similar approach to study population movements in Nepal in the aftermath of the April 2015 earthquake

Big Data and 'do no harm'?

In ongoing global conversations on responsible data governance and sharing, researchers, policymakers and data sharing entities often bring up a common central question in regards to conceiving governing principles and minimum standards for responsible data collection and use: how can we, similar to doctors, "do no harm?"

'Do no harm' proves to be just as difficult for data scientists as it is for medical doctors. Just as Big Data inherently problematizes the use limitation principle of the former OECD guidelines, understanding the effects of all such future uses of data and determining harm in the present has proved to be unrealistic and difficult. Researchers are attempting to create frameworks to understand the boundaries of what they can understand, while policymakers feel the administrative

pressure to respond quickly and create protective regulations.

In order to even begin to gain ground on tackling this issue of "do no harm" even within specific domains and contexts, companies and other big data stakeholders—including privacy experts, policymakers, data scientists, data users, private sector and ethicists—need to co-create infrastructure and spaces for discussion, safe experimentation and transparent findings. This could involve the further adaptation of ethical modalities used in other fields of research (e.g. the use of institutional review boards (IRBs), as in the case with Flowminder and the Karolinska Institute) or the creation of international ethical committees (e.g. the creation of ethical committees during the 2nd Orange Data for Development Challenge involving external actors to review violations in use and potential negative impact of projects on host country).

“Do no harm” by understanding current notions of privacy

One dimension of a “do no harms” approach for companies would be to gain a better understanding of current notions of privacy among citizens. A recent study from the Vodafone Institute published in January 2016 highlighted European user uncertainty and skepticism of the benefits derived from the use of their information as a part of Big Data initiatives. Surveying more than 8,000 European digital users, a key insight from the Vodafone Institute study suggests that, though overall users were unsure of the impact of the data revolution, when asked about sharing their data and impacting privacy, users were willing to share personal data if the personal or societal benefits were made clear; for instance⁷:

- “53% said that they wouldn’t mind their data being analysed if it would help them or other people to improve their health;”
- “68% stated that they were in favour of smart meters to record data on building residents’ usage behaviour so that more eco-friendly heating practices could be introduced; and
- “55% said they were happy about data from their cars being transferred in order to receive personalised traffic reports.”

Taylor describes much different sentiment from focus groups in her work in Amsterdam on smart lamp posts that capture data on who is in the area, their behavior and other habits in the area. Though none of the data is personally identifiable, focus group participants flagged this as doing harm, in the form of violating their autonomy and privacy.

Group privacy

Much of the ongoing policy conversations in Brussels are focused on the collection of personal data and the protection of individual privacy. However, ongoing research have revealed the need to consider the impact of demographically identifiable information (DII) and group privacy as well.

Described in Nathaniel Raymond’s recent work, DII describes information that is “either individual and/or aggregated data points that allow inferences to be drawn, enabling the classification, identification, and/or tracking of both named and/or unnamed individuals, groups of individuals, and/or multiple groups of individuals according to ethnicity, economic class, religion, gender, age, health condition, location, occupation, and/or other demographically defining factors.”⁸

The release of DII can lead to a kind of group association that in some countries may largely result in unwanted targeted ads and other inconveniences in customer experience. However, for vulnerable populations in poor, conflict-affected environments, “identification and association with groups facing demographic-based discrimination can result in unchecked aggression against both actual and perceived group members.”

The Harvard Humanitarian Initiative’s Signal Program is often cited as a clear example of DII risk in using satellite data to identify human rights abuses. In order to trace the conflict in Sudan and classify the ongoing destructive behavior during the genocide, the program developed machine learning algorithms to distinguish local huts from other structures and specifically to observe whether these structures were destroyed by

human activity (via burned or shelled by artillery) or burned by natural wildfires. While the moral imperative was clear in the design of the program and none of the information was personally identifiable, the program unknowingly released a form of DII that, following numerous remote hacking from an unknown source in Africa, most likely fell in the hands of the Sudanese government who could target what villages had not been attacked and changing their strategy towards the next attacks.

-
1. Organisation for Economic Co-operation and Development. OECD Guidelines on the Protection of Privacy and Trans-border Flows of Personal Data, 1980. Available online: <https://www.oecd.org/sti/ieconomy/oecdguidelinesontheProtectionofPrivacyandTransborderFlowsOfPersonalData.htm>
 2. Letouzé, Emmanuel. "Big Data and Development Overview Primer". Data-Pop Alliance, SciDev.Net and the World Bank (2015)
 3. de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. "Unique In The Crowd: The Privacy Bounds Of Human Mobility".Sci. Rep. 3. Nature Publishing Group. doi:10.1038/srep01376.
 4. Pentland, Alex. Social physics: how good ideas spread-the lessons from a new science. Penguin, 2014.
 5. Bengtsson et al. 2011
 6. Statistics New Zealand (2012)
 7. <http://www.vodafone-institut.de/2015/09/vodafone-institute-survey/>
 8. Raymond, Nathaniel. "Beyond 'Do No Harm' and Individual Consent: Reckoning with the Emerging Ethical Challenges of Civil Society's Use of Data" in L. Taylor, L. Floridi, & B. van der sloot Eds., Safety in numbers? Group privacy and big data analytics in the developing world. Springer. (forthcoming)

