AN INTRODUCTION TO DATA COLLABORATIVES

CREATING PUBLIC VALUE BY EXCHANGING DATA

Stefaan Verhulst, Andrew Young and Prianka Srinivasan

GOVLAB unicef OMIDYAR NETWORK

Stefaan G. Verhulst

Data Collaboratives: Exchanging Data to Improve People's Lives

By Stefaan Verhulst and David Sangokoya, The GovLab

In late July 2014, a sick passenger from Liberia traveled to Nigeria and brought the Ebola virus to Lagos, Africa's largest city, with a population of 21 million. In response, government agencies, universities and hospitals collaborated with private telecommunications companies and healthcare organizations to collect and share data on infected patients and trace those who had come into contact with them. State government health officials also initiated emergency steps to share information on a daily basis among actors involved in stemming the crisis. After two months, the virus was contained in Nigeria and the country declared Ebola-free.

Several of society's greatest challenges—from addressing climate change to public health to job creation—require greater access to data, more collaboration between public- and private-sector entities, and an increased ability to analyze datasets. This relationship between data and public benefits was vividly demonstrated in case study after case study at the recently concluded Cartagena Data Festival.

Yet for all the potential, a limiting factor is that much of the data valuable for solving public problems actually resides within the private sector—for example, in the form of click histories, online purchases, sensor data, and, as in the case of the above example, call data records. Amid the proliferation of apps, platforms and sensors, data on how people and societies behave is increasingly privately owned. We believe that if we truly want to leverage the potential of data to improve people's lives, then we need to accelerate the creation and use of "data collaboratives."

The term data collaborative refers to a new form of collaboration, beyond the public-private partnership model, in which participants from different sectors—including private companies, research institutions, and government agencies—can exchange data to help solve public problems. In the coming months and years, data collaboratives will be essential vehicles for harnessing the vast stores of privately held data toward the public good.

"The term data collaborative refers to a new form of collaboration, beyond the public-private partnership model, in which participants from different sectors—including private companies, research institutions, and government agencies—can exchange data to help solve public problems.

In the coming months and years, data collaboratives will be essential vehicles for harnessing the vast stores of privately held data toward the public good."

Stefaan G. Verhulst - April 22, 2015 - Medium

medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a#.flib5frfh

DATA COLLABORATIVES IN ACTION

CHALLENGE

IMPROVING DISASTER RELIEF BY EXCHANGING MOBILE PHONE DATA

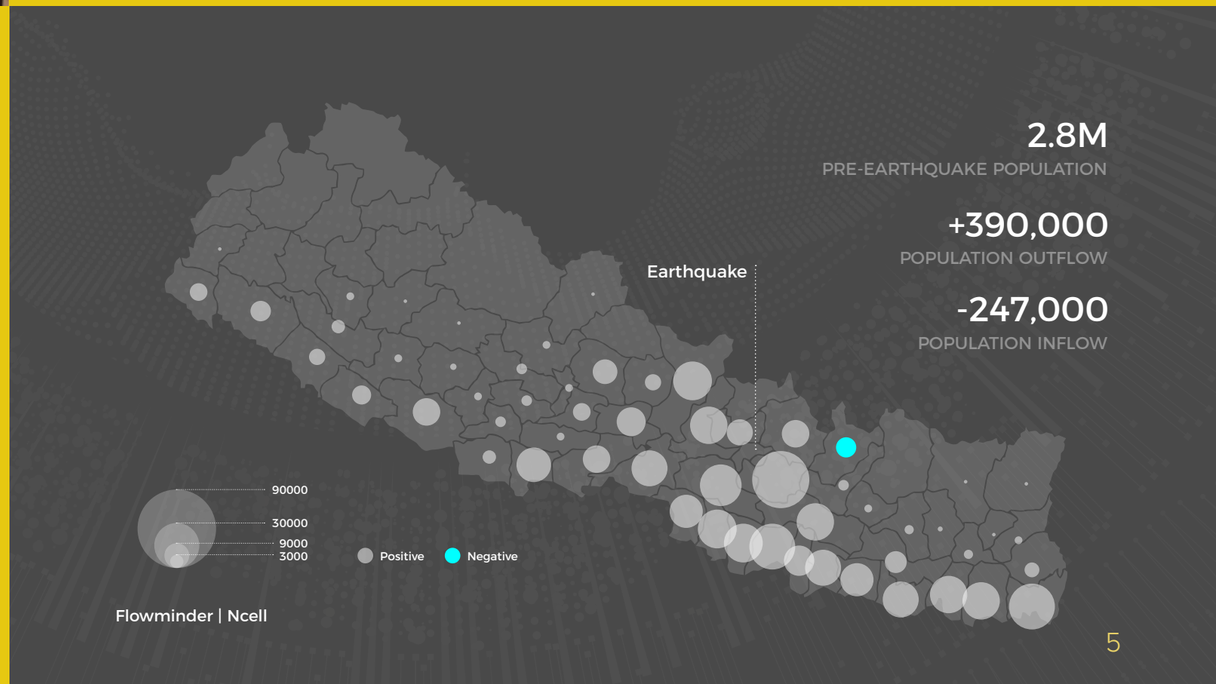
On April 25th, 2015, a violent earthquake hit Nepal—the worst of its kind since 1934. The damage left hundreds of thousands of people homeless and flattened entire villages. Ultimately, the Gorkha earthquake killed nearly 9,000 people and injured nearly 22,000.

Yet, the death toll could have been much worse.

DATA COLLABORATIVES

NCEL, Nepal's largest mobile operator, **shared anonymized mobile phone data** with the non-profit Swedish organization Flowminder. With this data, Flowminder mapped where and how people moved in the wake of the disaster, and shared this information with government and UN agencies to assist their relief efforts.

The Data Collaborative between NCEL and Flowminder allowed humanitarian organizations to better target aid to affected communities - saving hundreds if not thousands of lives.



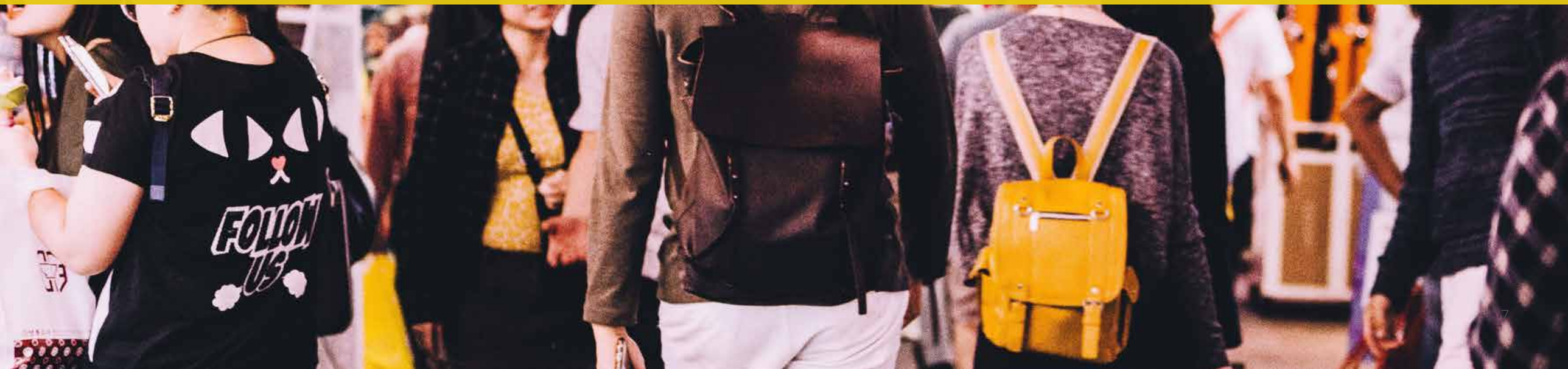
DATA COLLABORATIVES HAVE EMERGED AS A NEW FORM OF PUBLIC-PRIVATE PARTNERSHIPS TO ADDRESS SOCIETY'S MOST PRESSING PROBLEMS.

Global Fishing Watch, a partnership between Google, Oceania and Sky Truth, aims to stop illegal fishing by tracking the movement of over 35,000 sea vessels.

Simpa Networks, a company that provides pay-as-you-go solar energy to residents in India, shared its data with DataKind to **ensure more people have access to electricity.**

NetHope, by acquiring data from the private, public and humanitarian sectors, mapped the trajectory of new Ebola outbreaks in West Africa, **preventing further spread of the virus.**

CREATING PUBLIC VALUE



CREATING PUBLIC VALUE BY EXCHANGING DATA

DEMAND

What if we could find out how people move around in cities or where they go whenever a disaster strikes?

What if we could better understand how and where poverty changes over time?

What if governments had information to respond directly to people's needs?

What if research on disease treatment was shared globally?

SUPPLY

It's estimated that nearly 3 billion people will have smartphones by 2020

Over 1,000 active satellites orbit the earth

6,000 tweets are posted every second

The size of the global digital health data market is expected to increase to $233.3 billion in 2020

Access to different data sources allows us to understand public problems from **many angles**.

By creating data collaboratives, and opening corporate data, we can find **new, innovative and data-driven solutions** to combat public problems.

The potential of Data Collaboratives is immense.

FIVE WAYS DATA COLLABORATIVES CREATE PUBLIC VALUE

SITUATIONAL AWARENESS AND RESPONSE

A greater understanding and ability to track conditions on the ground can improve interventions, including especially during emerging situations like crisis relief.

Orbital Insights and the World Bank are using satellite imagery to measure and track poverty. Initial results showed, that in some instances, this technology "can be more accurate than U.S. census data."

PUBLIC SERVICE DESIGN AND DELIVERY

Access to previously inaccessible datasets can enable more accurate modelling of public service design and guide service delivery in a targeted, evidence-based manner.

Global mapping company, Esri, and Waze's Connected Citizen's program uses crowdsourced traffic information to help governments design better transportation.

KNOWLEDGE CREATION AND TRANSFER

Bringing more and more diverse datasets to bear can fill knowledge gaps and ensure that those responsible for solving problems have the most useful information at hand.

The National Institutes of Health (NIH), the U.S. Food and Drug Administration (FDA), 10 biopharmaceutical companies and a number of non-profit organizations are sharing data to create new, more effective diagnostics and therapies for medical patients.

PREDICTION AND FORECASTING

New predictive capabilities enabled by access to datasets can help institutions be more proactive, putting in place mechanisms based on sound evidence that mitigate problems or avert crises before they occur.

Intel and the Earth Research Institute at the University of California Santa Barbara (UCSB) are using satellite imagery to predict drought conditions and develop targeted interventions for farmers and governments

IMPACT ASSESSMENT AND EVALUATION

Access to additional datasets can help institutions monitor and evaluate the real-world impacts of policies and inform iteration.

Nielsen and the World Food Program have been using data collected via mobile phone surveys to better monitor food insecurity in order to advise WFP's resource allocations.

EXCHANGING DATA

EXCHANGING DATA: TYPES OF DATA

Registration records, data included in government transactions, and crowdsourced data. For example, patient health systems records shared by 10 biopharmaceutical companies in the Accelerating Medicines Partnership.

Personal information actively and intentionally shared by an individual, entity or group for a specific reason.

Internet usage data, commercial transactions like credit card data, and records of energy usage. For example, anonymized energy usage data shared by Dutch energy company, Enexis .

Information with potentially personally identifiable data that is passively collected by an entity prior to any use.

Information free from personally identifiable elements that is actively shared by an individual, entity or group for a specific reason.

Citizen science data, computer system logs, and data on the domain name system. For example, crop data shared through computer systems logs in Intel's Big Data for Precision Farming Initiative.

Information with no personally identifiable elements that is passively collected by an entity prior to any use.

Satellite and aerial imagery. For example, geolocational data on the movement of fishing vessels shared by Global Fishing Watch.

SIX TYPES OF DATA COLLABORATIVES

Data Cooperatives or Pooling

Prizes & Challenges

Research Partnerships

Intelligence Products

Application Programming Interfaces (APIs)

Trusted Intermediary

SIX TYPES OF DATA COLLABORATIVES

Problem: Lack of data diversity, leading to unrepresentative interventions.

Data Cooperatives or Pooling

Corporations and other important dataholders group together to create "data pools" with shared data resources.

Problem: Lack of external actors to apply data analysis skills within public sector.

Prizes & Challenges

Corporations make data available to qualified applicants who compete to develop new apps or discover innovative uses for the data.

Problem: Limited information and data for academic researchers, stymying their progress.

Research Partnerships

Corporations share data with universities and other academic organizations giving researchers access to consumer datasets and other sources of data to analyze social trends.

Problem: Inability or lack of resources to create data-driven products to solve a public problem.

Intelligence Products

Shared (often aggregated) corporate data is used to build a tool, dashboard, report, app or another technical device to support a public or humanitarian objective.

Problem: Inability to access useful data continuously from particular companies, like social networks.

Application Programming Interfaces (APIs)

APIs allow developers and others to access data for testing, product development, and data analytics.

Problem: Lack of expertise to analyze or use private sector data, even when given access.

Trusted Intermediary

Corporations share data with a limited number of known partners. Companies generally share data with these entities for data analysis and modelling, as well as other value chain activities.

SIX TYPES OF DATA COLLABORATIVES

Data Cooperatives or Pooling

MDEEP Project: In partnership with the United Nations and nonprofit organizations, Grameenphone shares anonymized telecom data as a part of the Mobile Data, Environmental Extremes and Population (MDEEP) Project with the United Nations and nonprofit organizations.

Prizes & Challenges

BBVA Innova Challenge: Spanish regional bank BBVA shares anonymized credit card transactions data with participating developers in its Innova Challenge.

Research Partnerships

T1D Data Exchange: T1D Exchange acts as a convener of clinicians, patients and caregivers working to improve the health outcomes of patients with Type 1 Diabetes by sharing data to accelerate research in the field.

Intelligence Products

Ethiopian Commodity Exchange: The Ethiopian Commodity Exchange provides real time price data, with dissemination mechanisms tailored to the needs of small farmers

Application Programming Interfaces (APIs)

Clever: Clever provides a single, universal API allowing developers and education companies to access siloed data within a school's student information systems.

Trusted Intermediary

Nethope: Assists in the development of public & private partnerships to deliver information technology solutions to the developing world. Data is shared with NetHope, who can then work with humanitarian organizations in the field to deliver data-centric innovations to their work.

MAKING IT WORK

MOTIVATIONS TO SHARE: THE SIX RS BEHIND CORPORATE DATA SHARING

RECIPROCITY

REVENUE

RESEARCH & INSIGHTS

REGULATORY COMPLIANCE

REPUTATION

RESPONSIBILITY

THE SIX RS BEHIND CORPORATE DATA SHARING

RECIPROCITY

Corporations may share their data with others entities across sectors for **mutual benefit**, especially when gaining access to other data sources that may be important to their own business decisions.

For instance, data pools created under the Accelerating Medicines Partnership (AMP) aim to overcome fragmentation in the pharmaceutical industry and improve innovation in drug therapy, ultimately allowing pharmaceutical companies to find new drug targets, and reduce wasteful repetition of testing found when companies work in-silo.

RESEARCH & INSIGHTS

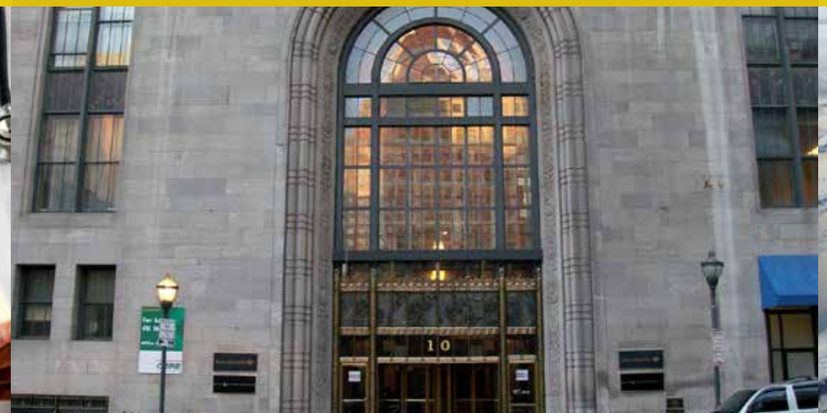
Opening up corporate data may generate new answers to particular questions providing companies insights that may not have been extracted otherwise. Just as with open source, sharing data (and in some cases algorithms) can enable corporations to tap into data analytical skills (often free labor) distributed beyond the boundaries of their own company. External users may interrogate the data in new ways and use the skills and methodologies not readily available in the company. It may also create the potential to identify and hire valuable talent that can emerge from the data sharing and use arrangement. In addition, these insights may enable companies to identify new niches for activity and to develop new business models.

For instance, Spanish bank BBVA's Innova Challenge allows participating developers to access BBVA's 'Big Data' to create apps and compete for awards for innovation both within and outside the company. By opening up their corporate data through this challenge, BBVA have supported research and innovation for the public good and for their private commercial interests.

REPUTATION AND PUBLIC RELATIONS

Sharing data for public good may enhance a firm's corporate image and reputation, potentially attracting new users, customers, employees and investors who value socially conscious corporate actors. It may also offer an opportunity to gain (free) media attention and increase visibility among certain decision makers and other audiences.

For instance, data pools created under the Accelerating Medicines Partnership (AMP) aim to overcome fragmentation in the pharmaceutical industry and improve innovation in drug therapy, ultimately allowing pharmaceutical companies to find new drug targets, and reduce wasteful repetition of testing found when companies work in-silo.



THE SIX RS BEHIND CORPORATE DATA SHARING

REVENUE GENERATION

Opening up corporate data does not always have to be for free. Under some conditions, corporate data may be offered for sale, generating extra revenue for firms in B2B and B2G arrangements.

For instance, as part of Telefonica's Insights arm, Telefonica Smart Steps releases anonymous aggregated mobile data for-profit to both the public and business sector. By sharing its data, Telefonica has expanded its role beyond simply a telecommunications provider to capitalize on the value of their data to aid public and commercial problem-solving.

REGULATORY COMPLIANCE

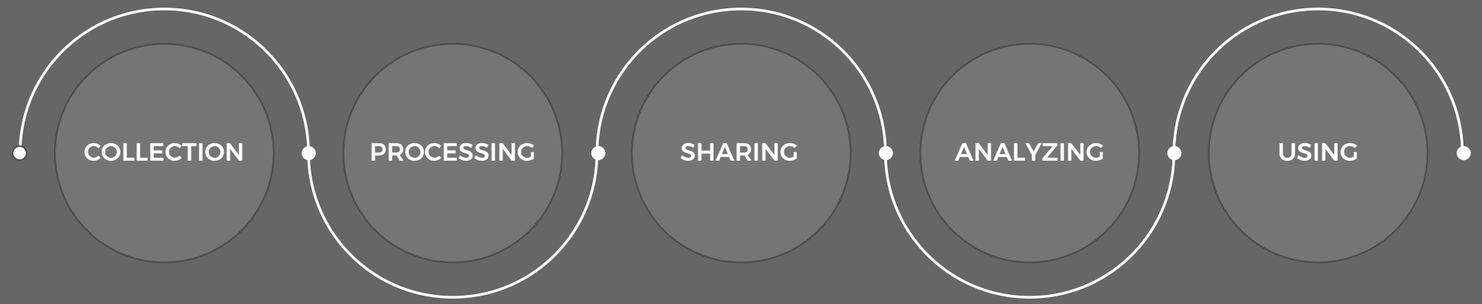
Sharing data can also help corporations comply with sectoral regulations and become more transparent and trusted. In addition, many corporations generate specific datasets for the sole purpose of regulatory compliance. Sharing and using that data in a responsible manner for public and private benefit may leverage more broadly the investment made to collect the data for a narrow purpose.

For instance, All US companies are required to release their employment data on race/ethnicity, gender and job categories to the Equal Employment Opportunity. Some companies, like Apple, Cisco, Dell and Google (among others), choose to release this information to the public, ensuring that the diversity in their workforce is interrogated, promoted and maintained.

RESPONSIBILITY AND CORPORATE PHILANTHROPY

Sharing corporate data achieves many of the goals sought by traditional corporate social responsibility or philanthropy. Companies can derive value from socially responsible behavior not just because of the positive image such an activity produces, but because opening up data can also improve the competitive business environment within which the business operates.

For instance, Nielsen releases food pricing information to Feeding America to assist them with their advocacy and food monitoring efforts. Nielson have entered this partnership as part of their Nielson Cares initiative to use their data for social good, adding to their corporate social responsibility efforts. In such a way, Nielson is able to use their existing data and expertise to benefit the wider society and simultaneously enhance their corporate image.



RISKS AND HARMS

DATA LIFECYCLE

Data risks exist at every stage of the data value cycle, from collection to use. They are often the result of

- technological weaknesses (e.g. security flaws);
- individual and institutional norms and standards of quality (e.g. weak scientific rigor in analysis);
- legal confusion or gaps, or misaligned business and other incentives (e.g. seeking to push the boundaries of what is societally appropriate).

While there are common elements across these risks – including both the root causes of such risks and the potential negative impacts – is useful to examine them according to the stage of the data lifecycle.

When not addressed (for instance, when dirty data doesn't get cleaned at the collection stage) risks accumulate and may lead to additional risks downstream (for instance, making flawed inferences from data analysis due to inaccurate data).

Collecting inaccurate, old or “dirty” data affecting data quality and ability draw meaningful insights

Unauthorized or intrusive data collection potentially leading to privacy harms;

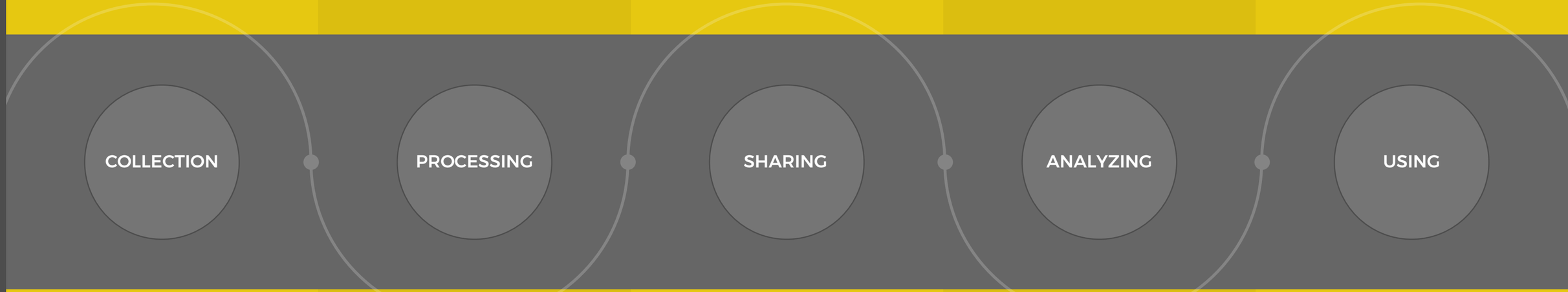
Insufficient, outdated or inflexible security provisions creating the potential for data vulnerabilities or breaches;

Lack of interoperable cultural and institutional norms and expectations, creating a difficult environment to collaborate toward mutual benefit;

Lack of data stewardship at both ends to ensure the responsible use of personally identifiable information as it travels across use cases and sectors;

Poor problem definition or research design, potentially leading to data being analyzed in a way that does not add value toward the ultimate objective

When data is ultimately put to use, risks emerge especially from collaborative organizations using shared data controversially or incongruously in relation to the original objective for its collection and/ or the original consent provided by the data subject (if any). Such risks can have negative results like the misinterpretation of data, the re-identification of individual data subjects, and decisional interference



Incomplete or non-representative sampling of the universe – e.g., ignoring “data invisibles,” or population segments with a limited data footprint – potentially leading to non-inclusive or unrepresentative approaches or interventions.

Aggregation and correlation of incompatible datasets can create ‘apples and oranges’ scenarios where the eventual sharing and analysis of commingled datasets are doomed for failure.

Improper or unauthorized access to shared data as it passes between entities, whether by unsanctioned actors inside or outside of collaborating organizations;

Conflicting legal jurisdictions and different levels of security within collaborating entities, making the eventual congruous data use difficult.

Inaccurate data modeling or use the of biased algorithms, which, like dirty data at the Collection stage, can lead to confidence in fundamentally flawed insights.

Additionally, at the Using stage, many of the risks from previous stages could yield true, identifiable harms for the first time – e.g., a negatively impactful policy decision being made based on faulty data from the Collection stage.

A DATA RESPONSIBILITY FRAMEWORK

IDENTIFY THE NEED

Data should never be used simply because they can be; the problem and potential benefits should be clear and defined.

MANAGE RISK TO (VULNERABLE) POPULATIONS

Participants (ideally, Data Stewards) should identify risks and harms to individuals and communities before the collaboration commences and adopt a plan to manage and mitigate those risks.

ASSESS CORE COMPETENCIES

Actors should identify what core competencies are needed to deploy a specific data-driven approach, and only proceed if those competencies are available to them.

ADHERENCE TO LEGAL AND ETHICAL STANDARDS

Data user are responsible for determining what legal and ethical standards apply to proposed applications of data in specific contexts, and for adhering to these to prevent potential violations of laws and rights.

See also “Building data responsibility into humanitarian action” (UNOCHA at http://www.unocha.org/node/227781)

DESIGNING A DATA COLLABORATIVE

DEMAND

Define the problem to be solved
Define the value proposition of the data collaborative

SUPPLY

Expertise: Data Science and Organizational Competency
Data Supply

COLLABORATION

Select the most promising potential supply-side data providers and identify specific incentives for them to participate

DESIGN

Define ideal type of data collaborative based on supply/ demand

Assess major risks, ethical concerns and potential challenges

Develop a multi-faceted risk mitigation strategy
Agree upon terms and conditions for arrangement

Establish a Governance Structure

IMPLEMENTATION

Agree upon expectations, roles, responsibilities, timeline and operational specifics of data-sharing process
Determine Resources: Cost and Funding Models

COMMUNICATION

Develop communications strategy
Determine audience and information sharing approach

LEARNING

Define a common baseline against which to measure progress
Measure progress against agreed-upon defined metrics of success
Impact assessment

ITERATION

Iterate as needed

- SITUATIONAL AWARENESS AND RESPONSE
- PUBLIC SERVICE DESIGN AND DELIVERY
- KNOWLEDGE CREATION AND TRANSFER
- PREDICTION AND FORECASTING
- IMPACT ASSESSMENT AND EVALUATION

TYPES OF DATA COLLABORATIVES

- DATA COOPERATIVES AND POOLING
- PRIZES & CHALLENGES
- RESEARCH PARTNERSHIPS
- INTELLIGENCE PRODUCTS
- APPLICATION PROGRAMMING INTERFACES (APIS)
- TRUSTED INTERMEDIARIES

TYPES OF DATA

- DISCLOSED PERSONAL DATA
- OBSERVED PERSONAL DATA
- DISCLOSED NON-PERSONAL DATA
- OBSERVED NON-PERSONAL DATA

MOTIVATING THE PRIVATE SECTOR

- RECIPROCITY
- RESEARCH & INSIGHTS
- REPUTATION & PUBLIC RELATIONS
- REVENUE GENERATION
- REGULATORY COMPLIANCE
- RESPONSIBILITY AND CORPORATE PHILANTHROPY

RISKS AND RESPONSIBILITY

- RISKS ACROSS THE DATA LIFE CYCLE
- COLLECTION
- PROCESSING
- SHARING
- ANALYZING
- USING
- DATA RESPONSIBILITY

LEARN MORE AT

datacollaboratives.org

#DATACOLLABORATIVES

info@thegovlab.org

GOVLAB unicef OMIDYAR NETWORK™