DATA COLLABORATIVES

CREATING PUBLIC VALUE BY EXCHANGING DATA

GOVLAB

USING MOBILE PHONE DATA FOR DISASTER RECOVERY

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.

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.

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

What if we better understood how individuals experienced poverty?

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

What if research on disease treatment was shared globally?

Ncel's collaboration following the earthquake in Nepal is just one instance of how private data can be leveraged for greater public good. By opening corporate data, and **engage** with a wide variety of analysts , **we can find new, innovative and data-driven solutions to combat society's problems**.

These **Data Collaboratives** have significant potential to create public value by exchangin data and expertise available in the private sector.

The potential of Data Collaboratives is immense.

DATA COLLABORATIVES

Data Collaboratives are a new form of public private partnership focused on the exchange of (corporate) data and expertise to help solve public problems.
Data Collaboratives is Corporate Social Responsibility for a Data Age

Data Collaboratives: Matching Demand and Supply

DEMAND: As public problems grow in complexity and urgency, the need for accurate and timely data has also intensified. Without this data, large holes remain in problem-solvers' toolkit, compromising the ability to make meaningful change for the public good.

SUPPLY: On the other hand private companies collect, analyze and store vast amounts of data about individuals, societies and their environments. From number of website visits, to customer habits, to the movement of peoples across the globe, the private sector is virtually awash with data about the public. Sometimes, this data is used to better coordinate business strategies and yield higher profits. Often, the value of this data remains wasted and lost in the sea of 'Big Data', with little opportunity for the wider society to benefit from the Data Age.

US companies with more than 1,000 employees have at least 200 terabytes of stored data

McKinsey estimate in 2009

Today the private sector is the prevalent collector of personal information. Walmart's data warehouse on consumer buying habits exceeds 500 terabytes of data.

Mehdi Khosrowpour, 2007

"For a long time governments had an effective monopoly on big data... This is no longer the case."

World Bank, 2015

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 it data with DataKind to improve its customer service, and to ensure more people have access to electricity.

NetHope, in partnership with the private, public and humanitarian sectors, mapped the trajectory of new Ebola outbreaks in West Africa, preventing further spread of the virus to other individuals.

MONITORING

Situational and response awareness

GOVERNANCE

Governance - policy making and public service delivery

PROBLEM SOLVING

Problem solving - based upon evidence

ANTICIPATION

Anticipation - prediction and forecasting

Data Collaboratives can improve public problem-solving by providing more targeted solutions for policy-makers at all levels of society.

These societal benefits can include improvements in efficiency, effectiveness, transparency and accountability of public services. This means that data collaboratives can both improve the performance of public services (i.e., effectiveness, efficiency) while also taking steps to improve the legitimacy of governance (i.e., transparency, accountability).

These public benefits can arise from increased accessibility of information, improved public service design, better input from key stakeholders, and bringing more human capital and expertise to bear in addressing a range of public challenges. These benefits are already manifest in sectors including **humanitarian issues, urban planning, natural resource stewardship and disaster management.**

PUBLIC VALUE THROUGH MONITORING

SITUATIONAL AND RESPONSE AWARENESS

Meaningfully tracking poverty remains a major challenge for governments, particularly in developing countries. As it stands, door-to-door surveys are the primary method to measure poverty levels. **Orbital Insights** and the **World Bank** are using satellite imagery to measure and track poverty. Initial results are promising--the company showed, for instance, that artificial intelligence which analyzed satellite imagery to count cars in a retailer's parking lot "can be more accurate than U.S. census data at predicting the retailer's quarterly earnings."

There is great value in sourcing **timely information**, particularly in crisis situations (such as disease proliferation or weather disasters) that benefit from quickly accessed, refreshed, disaggregated data. Timely information can lead to a more **equitable and targeted allocation of financial resources, expertise, and strategy** for addressing the issue.

The innovative use of data collaboratives can also close the loop between service providers and beneficiaries, where people can more easily **provide feedback** to institutions regarding their lived experiences and struggles. Real-time monitoring includes the use of dynamic, tech-based data collection methods, so often used by the private sector, which can revolutionize how public services and humanitarian interventions are monitored and evaluated.

PUBLIC VALUE THROUGH BETTER GOVERNANCE

POLICY MAKING AND PUBLIC SERVICE DELIVERY

Data collaboratives enable **more accurate modelling** of public service design, delivery and evaluations through data analytics. Improved tracking of policy outcomes to demonstrate government effectiveness can be enhanced by, in particular, observed personal data (e.g. commercial transactions, internet usage and other tracking data) typically held by the private sector.

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

PUBLIC VALUE THROUGH PROBLEM SOLVING THAT IS BASED UPON EVIDENCE

Data-driven decision making may be a common aim for institutions and policy makers, but often the most useful data is locked away in corporate databases. Creating spaces for the private sector to share useful, sector-specific data relevant to problem solving efforts can arm decision-makers with insights impossible to generate based on public sector data alone. The ability to **blend proprietary data with existing public datasets** can provide a more holistic picture of the problem and solution space, improving the likelihood of effective interventions.

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.

PUBLIC VALUE THROUGH BETTER ANTICIPATION – PREDICTION AND FORECASTING

Intel is working 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.

Private organizations often use data to make predictions about their future business, consumer interest or societal trends. Governments and international organizations can similarly take advantage of such predictive capabilities that big data affords us through data collaboratives.

By **anticipating crises** or problems as projected by data, organizations can better prevent or prepare for these events. Projects using such data are no longer reactionary, responding to disasters or societal problems as they emerge, but rather can become **more proactive**, putting in place mechanisms based on sound data that avert crises before they occur.

EXCHANGING DATA

Data Collaboratives exist both as formal, long-term collaborations, where agreements are made between organizations to govern long-term data sharing practices. For example, Clever, an educational program and app, has an agreement with US' School Districts to securely collect and share in a way that is compliant with FERPA (the Family and Educational Rights and Privacy Act).

...or as more opportunistic, informal partnerships where data is shared on more event-based, incidental conditions. For example, following the Haitian earthquake, a global team of data scientists used mobile phone data to map population displacement. Such ad-hoc arrangement, however, often require risk mitigation strategies, describe more below

6 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 “collaborative databases” 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.

EXCHANGING DATA: TYPES OF DATA

Corporations collect, hold and could potentially share many different types of data. To help inform the development of a targeted data collaborative and, especially, to help craft a meaningful approach for handling data responsibly, it can be useful to examine the supply of data according to **whether the data was actively disclosed or passively observed; and whether or not the data contains personally identifiable information.**

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.

Personally identifiable 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.

Making It Work

MOTIVATING THE PRIVATE SECTOR: THE 6 R'S OF CORPORATE DATA SHARING

Though the societal benefits of Data Collaboratives are vast and valuable in themselves, **without the voluntary commitment of private companies, such partnerships are doomed to fail.**

When and why corporations contribute their data differs according to the context in which the data is being requested or shared, the question access to their data may answer and the corporate and legal culture of the firm. Different corporations also have different views regarding the expected benefits and risks from sharing their data. As such, **when firms extend themselves and share their data they seek to satisfy a variety of motivations.**

Understanding the unique persuaders of businesses and corporations, and using this information to elicit a data sharing mechanism with the private sector, is therefore fundamental to ensuring Data Collaboratives are both responsible and successful.

We have identified 6 central motivators that encourage for-profit corporations to enter into cross-sector Data Collaboratives.

RECIPROCITY

RESEARCH & INSIGHTS

REPUTATION

REVENUE

REGULATORY COMPLIANCE

RESPONSIBILITY

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.

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, Orbital Insights has collaborated with the World Bank in its Measuring Poverty from Space project. By addressing needs from both the private and public sector, Orbital Insights has generated significant investor interest, recently raising $20 million from venture capital groups and In-Q-Tel, an American non-profit firm specializing in investments for intelligence products.

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.

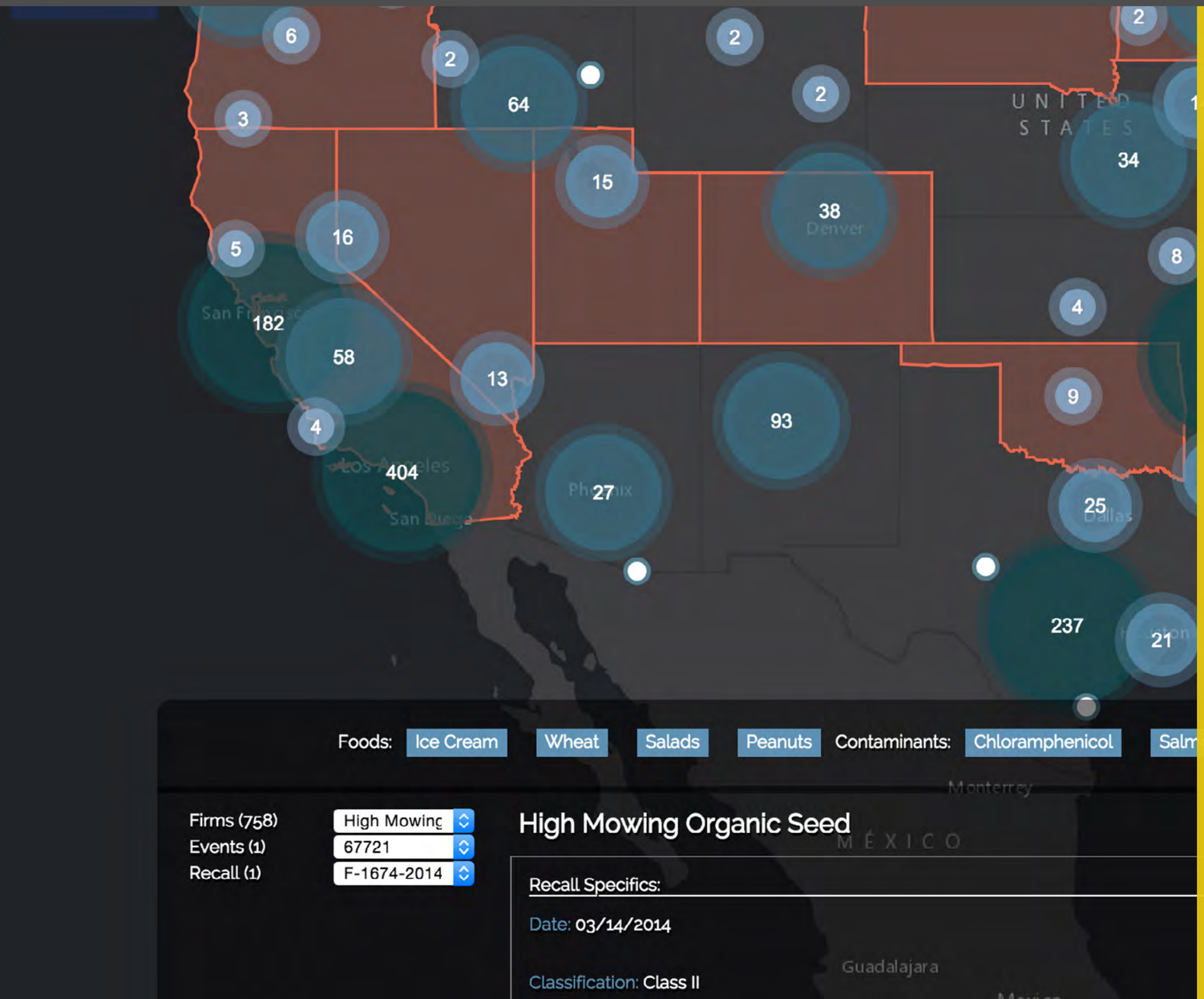
For instance, as part of Telefonica's Insights arm, Telefonica Smart Steps releases anonymous aggregated mobile data for-profit to both the public and business sector. By sharing its data, Telefonica has expanded its role beyond simply a telecommunications provider to capitalize on the value of their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

CONCLUSION: WIN-WIN VALUE PROPOSITION BEHIND DATA COLLABORATIVES

PUBLIC VALUE OF SHARED PRIVATE DATA

- Monitoring – situational and response awareness
- Governance – policy making and public service delivery
- Problem solving – based upon evidence
- Anticipation – prediction and forecasting

PRIVATE-SECTOR MOTIVATIONS TO SHARE DATA

- Reciprocity
- Research & Insights
- Reputation & Public Relations
- Revenue Generation
- Regulatory Compliance
- Responsibility and Corporate Philanthropy

MAKING IT WORK – RISKS

THE RISKS AND POTENTIAL HARMS OF SHARING CORPORATE DATA

The sharing of corporate data poses a number of risks that **could lead to harms for the organizations** involved in the exchange, the **individual subjects of shared datasets** (if any), and **those intended to benefit** from the use of the data. To tackle these risks and potential hams **risk mitigation strategies** are needed when sharing – including for instance limiting data access to specific, pre-approved uses, or “bringing the algorithm to the data” arrangements wherein private-sector datasets never leave corporate databases and instead are processed and analyzed using external algorithms.

MAKING IT WORK – RISKS

THE RISKS AND POTENTIAL HARMS OF SHARING CORPORATE DATA

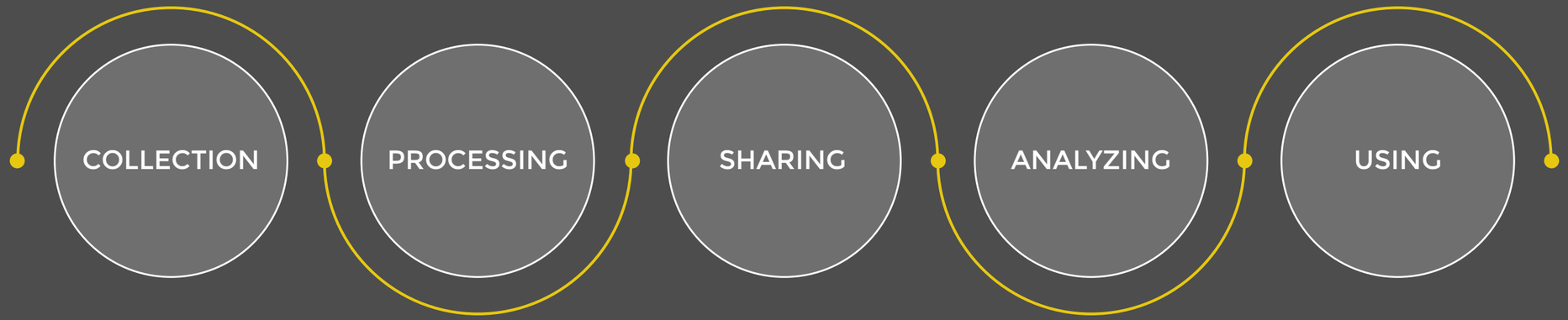
The sharing of corporate data poses a number of risks that **could lead to harms for the organizations** involved in the exchange, the **individual subjects of shared datasets** (if any), and **those intended to benefit** from the use of the data. To tackle these risks and potential hams **risk mitigation strategies** are needed when sharing – including for instance limiting data access to specific, pre-approved uses, or “moving the algorithm rather than the data” arrangements wherein private-sector datasets never leave corporate databases and instead are processed and analyzed using external algorithms.

EXAMPLE: The Open Algorithms (OPAL) project offers an open platform and ready-made algorithms allowing private companies to transform their own data in their own secure environments using pre-defined code. This method gives businesses full control of the data analysis process, reducing their workloads through access to external software, while simultaneously preventing external parties from accessing original data sources.

RISKS ACROSS THE 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 value cycle.

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).

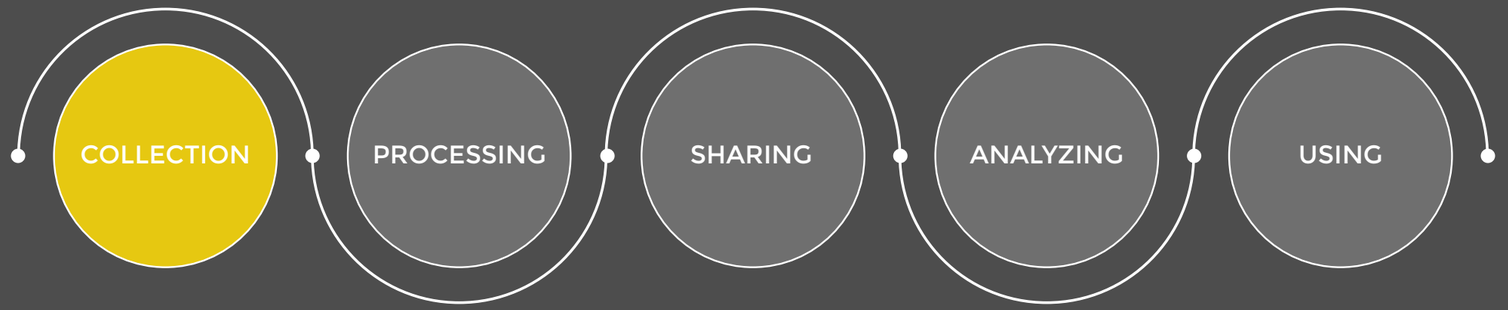


RISKS – AT THE COLLECTION STAGE

Collecting inaccurate, old or “dirty” data affecting data quality and thus affecting the ability to leverage and draw meaningful insights from the data;

Unauthorized or intrusive data collection from individuals and organizations – including the use of flawed or misleading consent mechanisms – potentially leading to privacy harms;

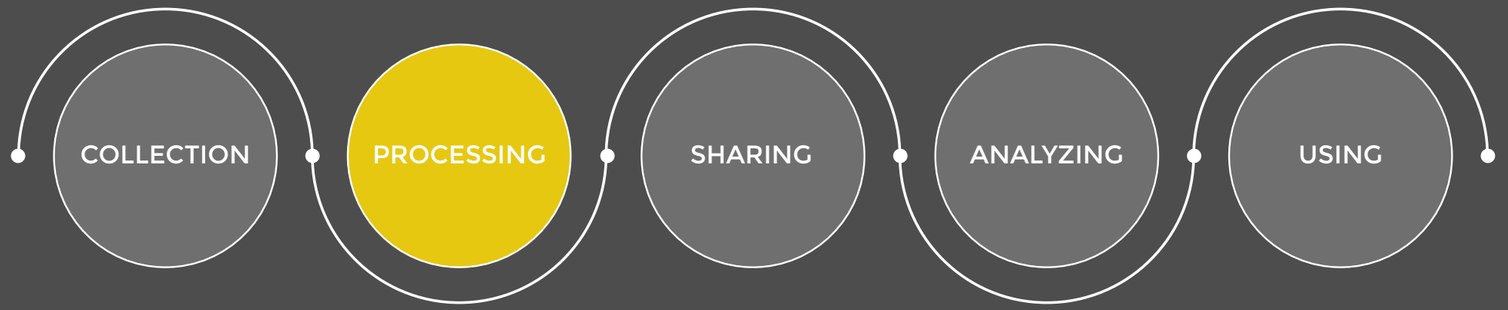
Incomplete or non-representative sampling of the universe – e.g., ignoring “data invisibles,” or population segments with a limited data footprint – potentially leading to non-inclusive or unrepresentative approaches or interventions.



RISKS – AT THE PROCESSING STAGE

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

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.



RISKS – AT THE SHARING STAGE

Of particular importance to Data Collaboratives, as opposed to data-driven decision-making more generally, risks at the Sharing stage include, for instance:

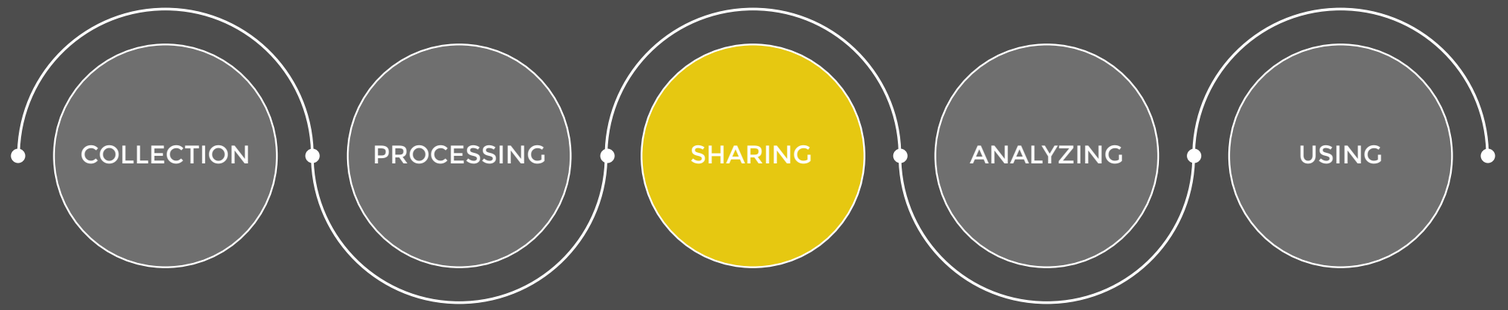
Lack of interoperable cultural and institutional norms and expectations, creating a difficult environment to collaborate toward mutual benefit;

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

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.

Public concerns surrounding the collection and use of personal data, particularly data about children, led to the dismantling of nonprofit inBloom. Initiated in 2011 through $100 million in seed money from the Bill and Melinda Gates Foundation and the Carnegie Corporation of New York, inBloom aimed to store, clean and aggregate student data for states and districts, making data available and standardized for approved third party applications and software designed for educators. Backlash came from increased privacy concerns from parents and educators over the collection, use and storage of personally identifiable information on students, causing many partners to back out of initial agreements. inBloom ended its services in April 2014.

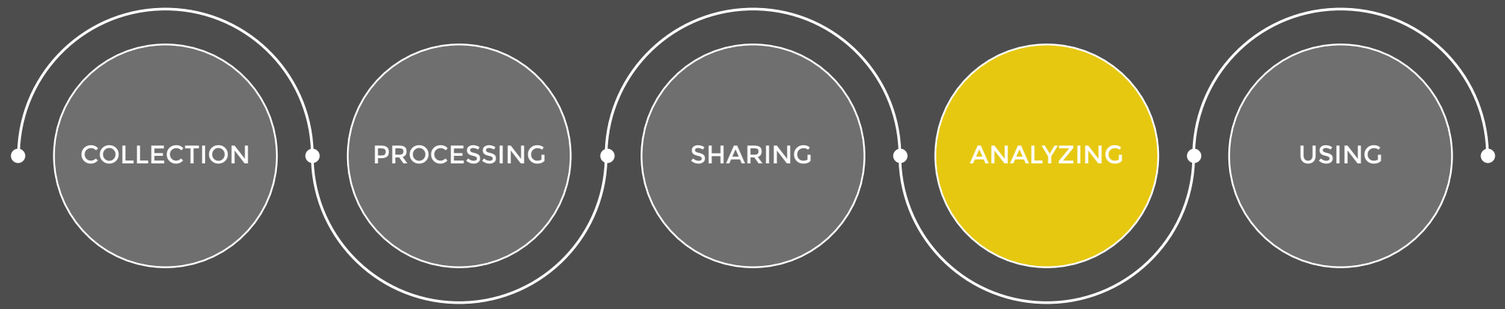


RISKS – AT THE ANALYZING STAGE

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

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.

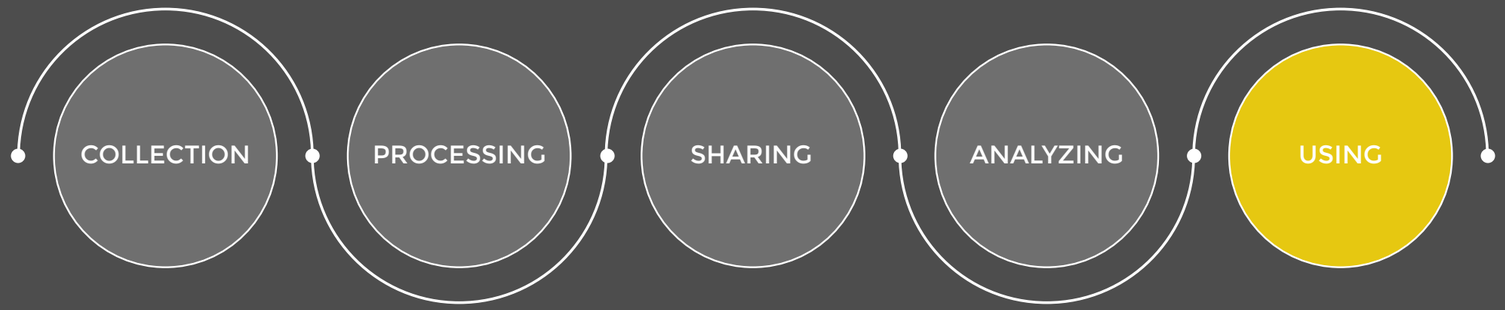
Example: Google Flu Trends provides a good example of a data collaborative that was weakened by flawed data modelling. The 2013 initiative attempted to provide real-time predictions of flu prevalence through analyzing Google's search terms. However, in its inaccurate algorithm which was too broad and mistakenly identified seasonal search terms, like "high school basketball," as flu predictions. As a result, Google Flu Trends missed the peak flu season by 140 percent.



RISKS – AT THE USING STAGE

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** (i.e., certain datasets influencing decisions, like insurance claims, that they should not affect)

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.



MAKING IT WORK: DATA RESPONSIBILITY, AN ESSENTIAL PROCESS

In order to mitigate risks and avoid potential harms of data collaboratives, participating organizations should take steps to establish Data Responsibility:

Evaluate the context and purpose within which data is being generated and shared;

Take inventory of the data and how it is stored;

Pre-identify risks and harms associated with a proposed use of data before data is collected;

Develop strategies to mitigate those risks.

AS PART OF THIS DATA RESPONSIBILITY PROCESS, COLLABORATORS SHOULD:

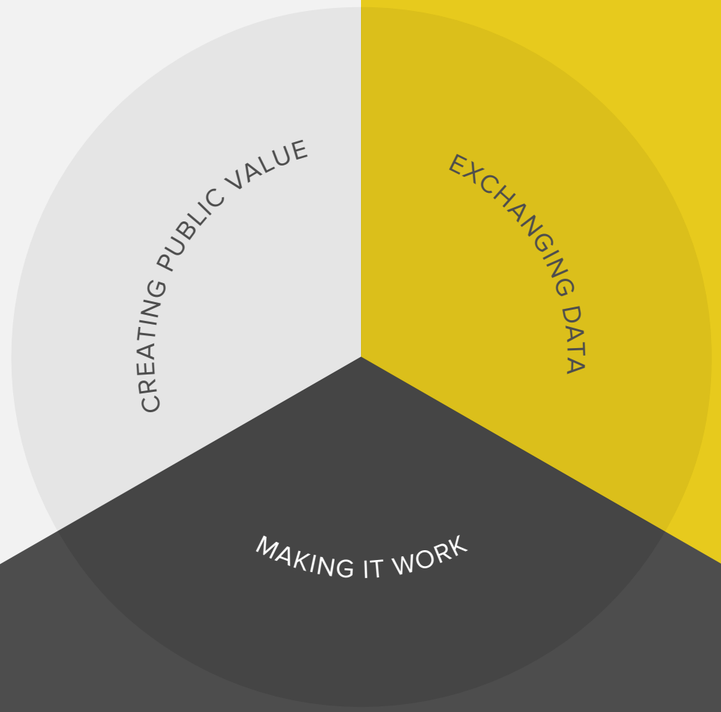
IDENTIFY THE NEED: Data should never be used simply because they can be; the problem and potential benefits should be clear and defined.

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.

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.

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.

- Monitoring – situational and response awareness
- Governance – policy making and public service delivery
- Problem solving – based upon evidence
- Anticipation – prediction and forecasting



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