GOVLAB

DATA COLLABORATIVES

LEVERAGING PRIVATE DATA FOR PUBLIC GOOD

A Descriptive Analysis and Typology of Existing Practices

Stefaan G. Verhulst
Andrew Young
Michelle Winowatan
Andrew J. Zahuranec

OCTOBER 2019

GOVLAB

ABOUT THE GOVLAB

The GovLab's mission is to improve people's lives by changing the way we govern. Our goal is to strengthen the ability of institutions – including but not limited to governments — and people to work more openly, collaboratively, effectively and legitimately to make better decisions and solve public problems. We believe that increased availability and use of data, new ways to leverage the capacity, intelligence, and expertise of people in the problem-solving process, combined with new advances in technology and science can transform governance. Housed at New York University (NYU) Tandon School of Engineering, The GovLab is funded by various donors and partner organizations.

ABOUT DATACOLLABORATIVES.ORG

DataCollaboratives.org launched in January 2017 to share the fruits of The GovLab's research into these emerging partnerships. Resources collected on the site include: (1) a collection of 150+ Data Collaboratives launched around the world, (2) curated resources defining data collaboration's current field of research and practice, (3) materials guiding stakeholders toward more effective and responsible data stewardship, and (4) a guide to designing and implementing a data collaborative.

ABOUT THE AUTHORS

Stefaan G. Verhulst is Co-Founder and Chief Research and Development Officer of The GovLab at NYU where he is building a research foundation on how to transform governance using advances in science and technology–including data and collective intelligence. He is Editor-in-Chief of the Data & Policy (Cambridge University) and Curator-in-Chief of thelivinglib.org. Before joining NYU full time, Verhulst spent more than a decade as Chief of Research at the Markle Foundation, where he continues to serve as a Senior Advisor. Previously, he co-founded PCMLP at Oxford University where he was also the UNESCO Chairholder in Communications Law and Policy.

Andrew Young is the Knowledge Director at The GovLab, where he leads research efforts focusing on the impact of technology on public institutions. Among the grant-funded projects he has directed are a global assessment of the impact of open government data; comparative benchmarking of government innovation efforts against those of other countries; a methodology for leveraging corporate data to benefit the public good; and crafting the experimental design for testing the adoption of technology innovations in federal agencies.

Michelle Winowatan is a Research Assistant at The GovLab, where she focuses on exploring ways in which data and collaboration can improve policymaking. Previously, she worked at several international nonprofit organizations, such as Search for Common Ground and Human Rights Watch, implementing projects that focused on conflict and human rights issues. Michelle holds a Master's degree in Public Policy from NYU Wagner, which she pursued while on a Fulbright scholarship. She also holds a Bachelor's degree in International Relations from Universitas Pelita Harapan.

Andrew J. Zahuranec is Research Fellow at The GovLab studying how advances in science and technology can improve governance. In previous positions at the NATO Parliamentary Assembly and National Governors Association, he worked on issues as far-ranging as election security, the commercial space industry, and the opioid epidemic. He has a Master of Arts in Security Policy Studies from the George Washington University and a bachelor's degree in Political Science and Intelligence Studies from Mercyhurst University.

ACKNOWLEDGEMENTS

The authors of this report would like to express their deep appreciation for the many people who reviewed and contributed to the many iterations of this report. The GovLab's Eve Marenghi, Carolyn Yang, Stephen Tubman, and Michael Zachor conducted invaluable research. We'd also like to thank Akash Kapur, Senior Fellow at The GovLab, for his important editorial support. We appreciate Claudio Mendonca's contributions to the report's design and graphical elements. Finally, thanks to the several members of the Data Stewards Network who provided useful insight that guided this work, including Jacqueline Lu in particular.

GOVLAB

TABLE OF CONTENTS

Cover image by Chris Liverani on Unsplash

DATA COLLABORATIVES

LEVERAGING PRIVATE DATA FOR PUBLIC GOOD

A Descriptive Analysis and Typology of Existing Practices

Stefaan G. Verhulst
Andrew Young
Michelle Winowatan
Andrew J. Zahuranec

OCTOBER 2019

GOVLAB

INTRODUCTION

Data collaboratives are an emerging form of collaboration in which data held by an entity in the private-sector is leveraged in partnership with another entity (from the public sector, civil society and/or academia) for public good.

Photo by Yanalya on Freepik

LEVERAGING PRIVATE DATA FOR PUBLIC GOOD

To address the challenges of our times, we need both new solutions and new ways to develop those solutions.[1] Data will play a central role in this process.[2] Yet, much of the most useful, timely and comprehensive data that could help transform the way we make decisions or solve public problems resides with the private sector in the form of call detail records, online purchases, sensor data, social media data, and other assets. If we truly want to harness the potential of data to improve people's lives, we need to understand and find ways to unlock and re-use this private data for public good.

1 Stefaan G. Verhulst and Andrew Young. (2018, January 23). How the Data That Internet Companies Collect Can Be Used for the Public Good. Harvard Business Review. Retrieved from https://hbr.org/2018/01/how-the-data-that-internet-companies-collect-can-be-used-for-the-public-good

2 Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. 2016. "The Age of Analytics: Competing in a Data-Driven World." McKinsey. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world; Eric T. Meyer, Jon Crowcroft, Zeynep Engin, Anne Alexander, "Data for Public Policy", Policy & Internet, 2017. https://doi.org/10.1002/poi3.147.

In what follows, we analyze the current practice of "data collaboratives," an emerging form of collaboration in which a private-sector entity's data is leveraged in partnership with other entities from the public sector, civil society or academia for public good.[3] The GovLab coined the term "data collaborative" in 2015.[4]

The potential and realized contributions of data collaboratives stem from how the supply of and demand for data are widely dispersed—spread across government, the private sector, and civil society—and often poorly matched. While most commentary on the data era's shortcomings focuses on the potential misuse of data, one of the key challenges of our data age actually lies in a persistent failure to re-use data responsibly for public good. This failure results in tremendous inefficiencies and lost potential.

Data collaboratives, when designed responsibly, are key to addressing this shortcoming. They draw together otherwise siloed data and a dispersed range of expertise, matching supply and demand and ensuring that relevant institutions and individuals are using and analyzing data in ways that maximize the possibility of new, innovative social solutions.

While we have seen an uptake in normative discussions on how data should be shared, little analysis exists of the actual practice. Over the last few years, we have identified, curated and documented more than 150 data collaboratives deployed around the world to address societal challenges as varied as urban mobility,[5] public health,[6] and corruption.[7] These cases are stored on our Data Collaboratives Explorer, the largest such repository on the topic.[8]

This paper seeks to answer the central question: What institutional arrangements and operational dynamics enable private-sector data holders to collaborate with external parties to create new public value?

3 Iryna Susha, Marijn Janssen, and Stefaan Verhulst. 2017. "Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development." https://doi.org/10.24251/HICSS.2017.325.

4 Stefaan Verhulst and Andrew Young, "The Potential of Social Media Intelligence to Improve People's Lives," accessed February 11, 2019, https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a.

5 Populus. 2018. Measuring Equitable Access to New Mobility: A Case Study of Shared Bikes and Electric Scooters. Retrieved from https://research.populus.ai/reports/Populus_MeasuringAccess_2018-Nov.pdf

6 Flowminder — Guiding Malaria Elimination Strategies in Namibia. (n.d.). Retrieved May 14, 2019, from http://www.flowminder.org/case-studies/guiding-malaria-elimination-strategies-in-namibia

7 Angelico, F. (2017, May 12). Brazil: Open data just made investigating corruption easier. Retrieved May 14, 2019, from Transparency International website: https://www.transparency.org/news/feature/brazil_open_data_just_made_investigating_corruption_easier

8 "Data Collaboratives Explorer." Data Collaboratives, 2019. http://datacollaboratives.org/explorer.html.

In what follows, we analyze the emerging universe of data collaboratives and develop a typology of practice areas to provide insight into current applications and to inform processes to establish future data collaboratives. With this foundation, we will focus next on key industry sectors, common business and governance models, technical features and especially impactful use cases. Importantly, responsible data collaboration and re-use of data requires responsible data collection. The GovLab's work on data collaboratives begins from the assumption that actors' initial data collection processes are legitimate and ethical.

Taken together, the typology and set of variables introduced here illustrate there is no "one size fits all" approach for designing a data collaborative. This paper focuses on the more operational aspects of these practice areas and intentionally does not provide a detailed typology of data collaborative governance models. We do not, for example, reflect on concepts like "data trusts," which involve both operational and governance concerns.[9] Data trusts take the concept of a legal trust and applies it to data, providing independent stewardship of some data for the benefit of a group of organizations or people.[10] The technological platforms and foundations of data collaboration are similarly out of scope for this report, though we do provide a brief overview of some key technical concepts in Box 2. The governance and technological bases for data collaboration are complex and important considerations that warrant future dedicated analyses.

9 "Defining a 'Data Trust' – The ODI." n.d. Accessed October 7, 2019. https://theodi.org/article/defining-a-data-trust/.; "Digital Civil Society Lab–A Framework for Data Trusts–Stanford PACS." n.d. Accessed October 7, 2019. https://pacscenter.stanford.edu/research/digital-civil-society-lab/a-framework-for-data-trusts/.; Wylie, Bianca, and Sean McDonald. "What Is a Data Trust?" Centre for International Governance Innovation, October 9, 2018. https://www.cigionline.org/articles/what-data-trust.

10 Hardinges, J. and Wells, P.. 2018. "Defining a 'data trust'". From Open Data Institute website, https://theodi.org/article/defining-a-data-trust/

VARIABLES OF ENGAGEMENT AND ACCESSIBILITY

	Open Access	Restricted Access
Independent Use	**Public Interfaces**	**Trusted Intermediary**
Cooperative Use	**Data Pooling**	**Research and Analysis Partnership**
Directed Use	**Prizes & Challenges**	**Intelligence Generation**

Figure 1: Data Collaboratives Matrix of Engagement and Accessibility

There exists a wide array of variables involved in the operational and institutional dynamics of a data collaborative (see Section IV). To support the analysis of the broad and fluid field of current practice, we focus initially on two defining variables: Engagement and Accessibility. By determining the level of engagement between parties and the accessibility of private-sector data, we can define the contours of a descriptive typology of data collaboratives.

Engagement: the degree to which the data supply and demand actors co-design the use of corporate data assets.

- **Independent Use:** The analysis and use can be entirely independent, meaning the private-sector data holder has little direct involvement in the data's re-use.
- **Cooperative Use:** Second, it can be cooperative, meaning data suppliers and users decide the focus of the data use and analysis in partnership.
- **Directed Use:** Last, it can be directed, meaning the data holder seeks out partners to derive specific, prioritized types of public value from the data.

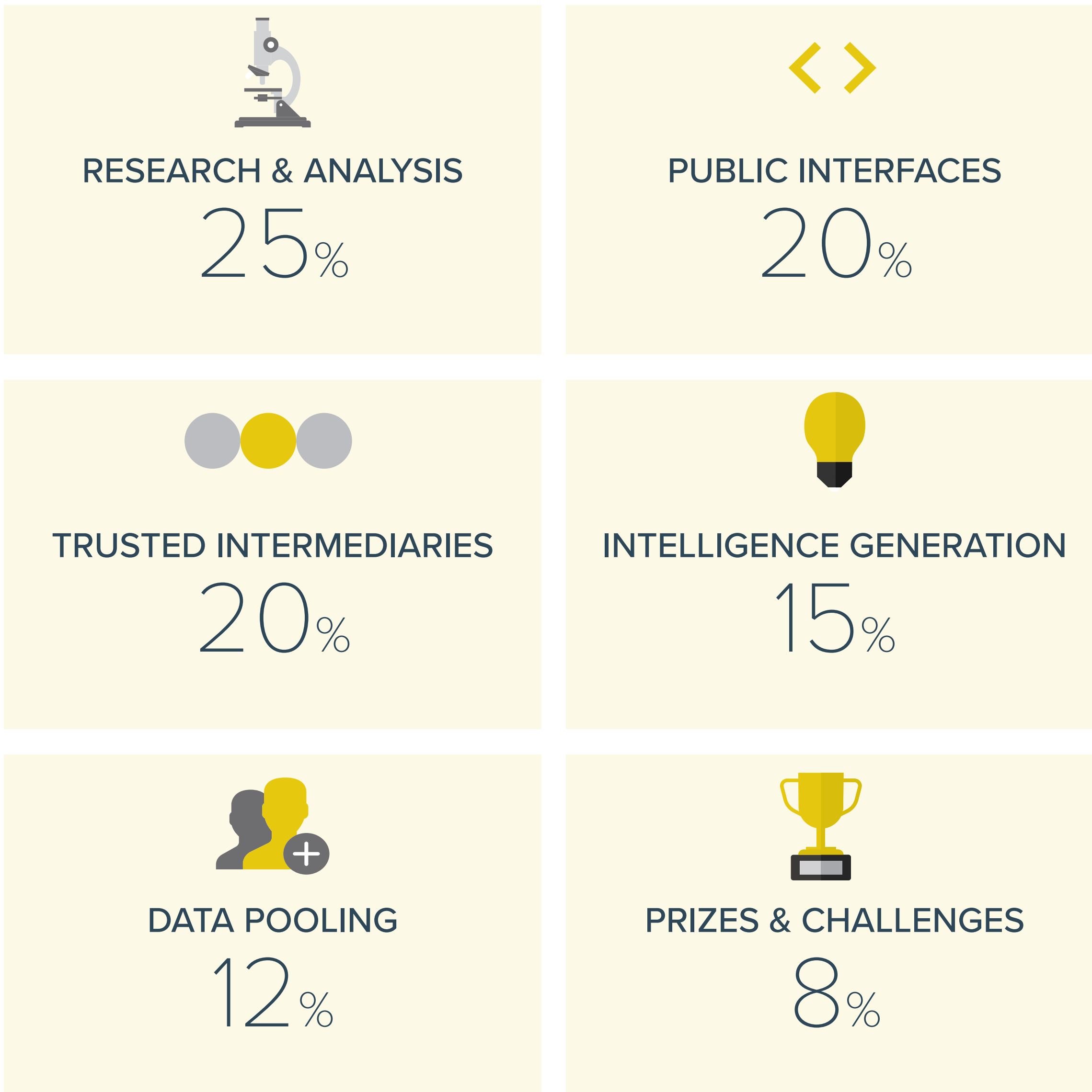
Accessibility: the conditionality of accessing private data by external parties.

- **Open Access:** These data collaboratives place very few restrictions on leveraging private-sector data, in some cases allowing for the general public to view or download certain data assets.
- **Restricted Access:** In more restricted access data collaboratives, only pre-selected partners receive access to corporate data assets.

Importantly, both of these variables exist on a spectrum. Differing levels of engagement and shades of accessibility manifest in different approaches or projects. Yet, when considered according to a matrix (as done in Figure 1), several models emerge:

- **Public Interfaces:** Companies provide open access to certain data assets, enabling independent uses of the data by external parties. Current approaches include: APIs and Data Platforms.
- **Trusted Intermediary:** Third-party actors support collaboration between private-sector data providers and data users from the public sector, civil society, or academia. Current approaches include: Data Brokerage and Third Party Analytics Projects.
- **Data Pooling:** Companies and other data holders agree to create a unified presentation of datasets as a collection accessible by multiple parties. Current approaches include: Public Data Pools and Private Data Pools.
- **Research and Analysis Partnerships:** Companies engage directly with public-sector partners and share certain proprietary data assets to generate new knowledge with public value. Current approaches include: Data Transfers and Data Fellowships.
- **Prizes and Challenges:** Companies make data available to participants who compete to develop apps; answer problem statements; test hypotheses and premises; or pioneer innovative uses of data for the public interest and to provide business value. Current approaches include: Open Innovation Challenges and Selective Innovation Challenges.
- **Intelligence Generation:** Companies internally develop data-driven analyses, tools, and other resources, and release those insights to the broader public.

Figure 2: Distribution of Practice Areas in the Data Collaboratives Explorer



Over the last three years, The GovLab has identified and documented data collaborative efforts from around the world in its Data Collaboratives Explorer. The repository is the most comprehensive mapping of data collaboratives worldwide including over 150 projects as of September 2019. The projects are distributed across six practice areas as shown in Figure 2, with Research and Analysis Partnerships as the most prevalent practice area documented followed by Public Interfaces.

TYPOLOGY OF DATA COLLABORATIVE PRACTICE AREAS

Below, we seek to expand our understanding of the six existing practice areas of data collaboratives and provide an overview of specific approaches implemented across those categories. For each practice area and approach, we include several representative examples derived from our Data Collaboratives Explorer. These examples are included in the interest of capturing current efforts, not necessarily to rank individual practices.

< > PUBLIC INTERFACES

Companies provide open access to certain data assets, enabling independent uses of the data by external parties.

	Open Access	Restricted Access
Independent Use	**Public Interfaces**	**Trusted Intermediary**
Cooperative Use	**Data Pooling**	**Research and Analysis Partnership**
Directed Use	**Prizes & Challenges**	**Intelligence Generation**

Data collaboratives in this practice area involve a single company providing public access to certain types of pre-processed data and/or data-driven tools, such as maps or dashboards. In many cases, these collaboratives enable certain communities or actors (e.g. city government officials or independent software developers) to make informed decisions or create new innovations via increased access to data. Although data assets are publicly accessible, these tools are often created with a particular audience or type of use case prioritized.

VARIOUS APPROACHES

APPLICATION PROGRAMMING INTERFACES (APIs)

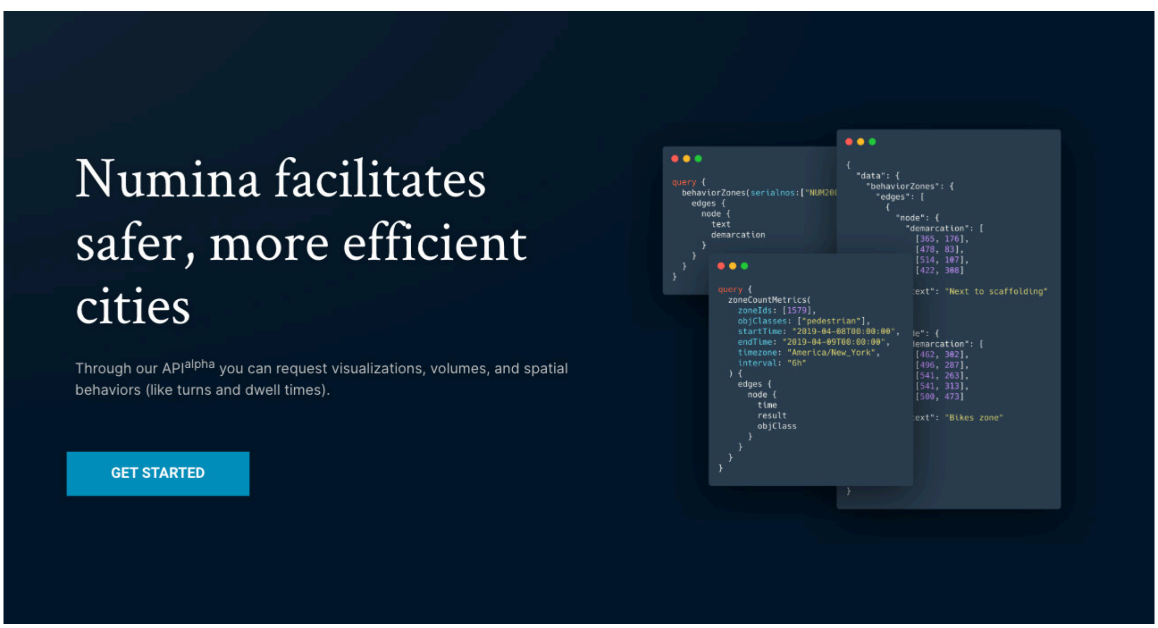
Description: Application Programming Interfaces (APIs) are digital protocols that publish data in an automated fashion on a near real-time data basis. While data providers can create certain conditions on acceptable uses of their APIs, the technology can allow for open, ongoing access to certain machine-readable information and enable a wide array of use cases that are largely independent of direct involvement of the data provider.

Examples:

Climate Corporation's Field View Products: Field View provides an API through which farmers can access agricultural data such as those related to weather, yield analysis, and field health data. It seeks to enable precision farming by using data. Field View provides basic datasets for free and charges a fee for full API access.[11]

Google Earth Outreach: Google Earth Outreach's API provides nonprofits and the public sector with geospatial data and analytic tools to analyze that data to help developers generate new tools and insights. The Ashoka Trust for Research in Ecology and the Environment used this platform to provide forest information to policymakers and researchers to protect tigers and elephants in forest reservations in India.[12]

Numina Street Intelligence API: Numina is a tech startup that offers insights about people, vehicles, and the street environment using sensors placed on outdoor fixtures such as lamp posts. The sensors use edge processing so that only anonymous, aggregated is transmitted from the sensors to Numina. Insights prevented trash bins from overfilling, identified where to locate mid-block crossing, and measured traffic during a car-free day. Numina's public API gives developers access to the data in order to enable them to build apps that could improve city street planning.[13]



numina.co

11 Climate. 2019. "Digital Farming Decisions and Insights to Maximize Every Acre." https://www.climate.com/.

12 "Google Earth Outreach (GEO) Initiative." 2019. http://datacollaboratives.org/cases/google-earth-outreach-geo-initiative.html.

13 "Numina | API." 2019. Numina (blog). Accessed October 7, 2019. https://numina.co/api/.

Reddit's API for Public Health Research: The social website Reddit provides a public API to enable researchers and developers to extract information from its website.[14] Several researchers use this tool to study public health through online discussions. Researchers at the University of Utah used the API to publish a paper tracking discussions about Ebola, electronic cigarettes, influenza, and marijuana on Reddit's forums.[15]

We Feel: We Feel is a research initiative of the Digital Economy Programme of the Commonwealth Scientific and Industrial Research Organisation's (CSIRO) Digital Productivity Flagship. This project uses several Twitter APIs and analyzes public tweets for emotional context to obtain a signal of the world's emotional state. We Feel visualizes the result and makes it accessible through its website. The project also shares its raw data to researchers using its REST-full API.[16]

DATA PLATFORMS

Description: Data platforms make private-sector data assets and tools accessible to the general public through web or mobile applications. Often, companies develop these platforms with certain users in mind (e.g. humanitarian actors or city planners), but data assets are generally made accessible to any user. Data platforms enable similarly flexible uses but often require less data or software-development expertise relative to APIs.

Examples:

Disease Surveillance And Risk Monitoring (DiSARM) Application: DiSARM is an open-source platform that provides data on the spread of diseases to inform disease-control programs to target interventions and improve health service coverage. The project has an online application, through which health workers can access the aforementioned data to inform their malaria-eradication program.[17]

Open Diversity Data: Open Diversity Data is a project from Double Union, a hacker and maker collective, that collects diversity reports, which provide demographic data on employees at private US companies. It publishes this information on its platform to encourage greater di-

14 "Reddit: The Front Page of the Internet." 2019. https://www.reddit.com/r/TheoryOfReddit/wiki/collecting_data.

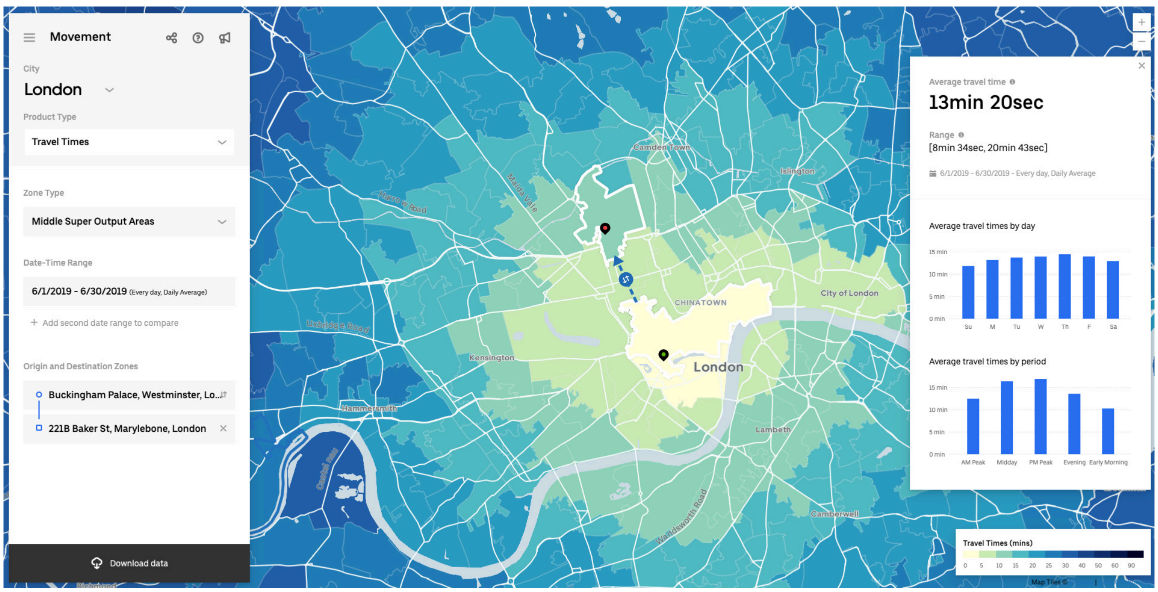
15 Park, Albert, and Mike Conway. 2018. "Tracking Health Related Discussions on Reddit for Public Health Applications." AMIA Annual Symposium Proceedings 2017 (April): 1362–71.

16 Larsen, Mark E., Tjeerd W. Boonstra, Philip J. Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. "We Feel: Mapping Emotion on Twitter." IEEE Journal of Biomedical and Health Informatics 19, no. 4 (July 2015): 1246–52. https://doi.org/10.1109/JBHI.2015.2403839.

17 This project also provides an algorithm that is accessible via an API.See: "Home Page." Disease Surveillance And Risk Monitoring, 2019. https://www.disarm.io/.

versity in the workforce. Platform users can find diversity reports from individual companies. They can also send a tweet to thank companies that have released their diversity report or call for releases from companies yet to do so.[18]

Uber Movement: The Uber Movement platform provides traffic patterns, average speeds, and other insights using data collected from the more than two billion trips made around the world on the Uber app. The service intends to inform urban planning and research into traffic impact, expediting decisions on issues like bridge or road closures.[19]



movement.uber.com

Waze Connected Citizens: Waze Connected Citizens provides mobility data and insights derived from the Waze navigation app's users to city governments and nonprofits through an online, account-protected platform. City governments and non-profit organizations, such as New York City, Washington, DC, Rio de Janeiro, Brazil and Ghent, Belgium use Waze's data to study problems related to traffic and mobility.[20]

18 Union, Double. 2019. "Open Diversity Data." http://opendiversitydata.org

19 "Uber Movement: Let's Find Smarter Ways Forward, Together." 2019. https://movement.uber.com/?lang=en-US.

20 "Driving Directions, Traffic Reports & Carpool Rideshares by Waze." 2019. https://www.waze.com/ccp.

BOX 1

WEB SCRAPING

Description: Web scraping is a process where users manually download data available on websites or platforms (such as social media networks) for local analysis. Web scraping is a fully independent use of private-sector data and is done without the data holder taking any action to make the data available for analysis. As web scraping approaches are not truly data collaboratives but fully independent uses of private-sector data, we do not include them as part of this typology. The legality of web scraping has also been challenged in certain cases. US courts have rejected these challenges.[21] Web scraping can also raise questions regarding the fully independent secondary uses of data generated for other purposes and not intentionally shared for further analysis. The field is still wrestling with these types of ethical questions.[22]

Nevertheless, there are many instructive uses of web scraping in the public interest. The following examples illustrate how the analysis of private-sector data can create public value, even in cases where the data supplier did not actively participate at any stage.

Examples:

Predicting Flood with Flickr Metatags: A research article released in 2017 sought to improve flood prediction by scraping and analyzing tags from Flickr, the photo-sharing website. The paper demonstrates how "volunteered geographic data can be used to provide early warning of an event before its outbreak."[23] This finding provides a blueprint for data-driven prediction tools to improve situational awareness and deliver more accurate forecast of natural disasters.[24]

Tracking Anti-Vaccination Sentiment with Social Media Data: In collaboration with the Indonesian ministries of development planning and health, UNICEF, and the World Health Organisation, Pulse Lab Jakarta filtered and analyzed publicly available tweets about vaccines and immunization to better understand public perceptions of immunization among the Bahasa-speaking population of Indonesia. The research suggests ways to address rumors and misinformation online.[25]

21 Adi Robertson, "Scraping public data from a website probably isn't hacking, says court," The Verge, September 10, 2019, https://www.theverge.com/2019/9/10/20859399/linkedin-hiq-data-scraping-cfaa-lawsuit-ninth-circuit-ruling

22 Densmore, James. 2019. "Ethics in Web Scraping." Medium. July 23, 2019. https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01.

23 Tkachenko N, Jarvis S, Procter R (2017) Predicting floods with Flickr tags. PLoS ONE 12(2): e0172870. https://doi.org/10.1371/journal.pone.0172870

24 Ibid.

25 "Understanding Public Perceptions of Immunisation Using Social Media. 2014. United Nations Global Pulse." https://www.unglobalpulse.org/projects/immunisation-parent-perceptions.

TRUSTED INTERMEDIARY

Third-party actors support collaboration between private-sector data providers and data users from the public sector, civil society, or academia.

	Open Access	Restricted Access
Independent Use	**Public Interfaces**	**Trusted Intermediary**
Cooperative Use	**Data Pooling**	**Research and Analysis Partnership**
Directed Use	**Prizes & Challenges**	**Intelligence Generation**

Trusted intermediaries provide data users with opportunities to unlock the public value of data while maintaining strict access controls. Some intermediaries facilitate data collaboration by matching supply and demand actors, ensuring that both public and private objectives can be achieved in a responsible manner. Other intermediaries can provide additional technical expertise to a collaboration by analyzing data from the supply side and passing on actionable insights to users representing the demand. This approach can advance data collaboration without requiring resources to be expended by data providers or data users beyond their capacity.

VARIOUS APPROACHES

DATA BROKERAGE

Description: Data Brokerages involve third parties who facilitate connections and match the supply of data (i.e. private-sector data holders) with the demand for it (i.e. by public or nonprofit institutions tasked with addressing social problems). These connections are generally purpose-bound and time-bound. While brokers tend to collaborate closely with the supply and demand sides of these arrangements, the users of data generally do not coordinate directly with the supply side.

Examples:

Consumer Data Research Center (CDRC): The United Kingdom Economic and Social Research Council established the CDRC to act as an intermediary making data held by consumer-related businesses (such as retail, transportation, and banking) available to researchers. CDRC connects researchers from the University of Leeds, University College London, University of Liverpool, and the University of Oxford to this data to inform their social and economic research about the United Kingdom.[26]

Social Science One: Social Science One is an organization established by the Social Science Research Council that tries to enable research partnerships between academic researchers and the private sector, including a partnership with Facebook. It is run by a commission of senior academics. This commission acts as a trusted intermediary and determines which datasets would be useful for researchers. They also manage partnerships with data holders, issue requests for proposals on specific topics, and support projects deemed to have academic and social merit.[27]

Stats NZ's Data Ventures: Data Ventures is the commercial arm of New Zealand's statistical agency, which functions as a trusted intermediary that pulls datasets from various sectors for later re-distribution to the platform's customers. The platform collects statistical data, government data, and private sector data, such as that from telecommunications companies. For its first project, it aggregated datasets from multiple businesses to better estimate population density in different regions.[28]

Yale University Open Data Access (YODA): Clinical data holders can share their data through YODA to be used for various health research objectives. YODA, then, facilitates the sharing of this data with other researchers inside and outside of the Yale community. YODA has the rights to grant or deny data access request, based on the agreement that it made with the data holders.[29]

26 "CDRC Home." 2019. CDRC. Accessed October 7, 2019. https://www.cdrc.ac.uk/.

27 "Social Science One." 2019. Accessed October 7, 2019. https://socialscience.one/home.

28 "Stats NZ's Data Ventures." 2019. Accessed October 7, 2019. http://datacollaboratives.org/cases/stats-nzs-data-ventures.html.

29 FAQS. 2019. "The YODA Project." Accessed October 7, 2019. https://yoda.yale.edu/welcome-yoda-project.

THIRD-PARTY ANALYTICS PROJECTS

Description: Third-party analytics projects see trusted intermediaries—research organizations, data analytics businesses, and nonprofits—access private-sector data, conduct targeted analysis, and share insights, but not the underlying data, with public or civil sector partners. This approach enables public interest uses of private-sector data while retaining strict access control. It brings to bear external data expertise that would likely not be available through more direct bilateral collaboration between data holders and users.

Some of these intermediaries offer algorithmic querying services. These services provide high data security by allowing only remote third party analysis of data that remains on a company's servers. Data never flows out of the company representing the supply. Rather, an algorithm is brought to the data, and the third-party providing this analytical capability only shares the insights generated with eventual data users. This model is prevalent in our Data Collaboratives Explorer among collaboratives using highly sensitive personal data such as Call Detail Records (CDRs).

Examples:

Dalberg's Food Security Manager: Dalberg Data Insights helps development and humanitarian organizations make data-driven impact by facilitating analysis of big data held by private companies using Dalberg's analytics tools. Dalberg developed the Food Security Manager with the Ugandan NGO Hunger Fighters to identify food insecurity risk by analyzing mobile phone money spending in Uganda. Hunger Fighters used this insight to inform their hunger eradication program.[30]

Deloitte's ConvergeHEALTH, Allergan, and Intermountain Healthcare: Through its ConvergeHEALTH program, the professional services company Deloitte formed a partnership with the pharmaceutical company Allergan and the Utah-based hospital network Intermountain Healthcare centered on improving the health of female patients. ConvergeHEALTH applied Allergan's data about intrauterine devices to its OutcomesMiner platform and provided subsequent insights to the hospital network. Participants hoped the relationship would reveal unmet needs of patients to improve clinical care and patient health outcomes.[31,32]

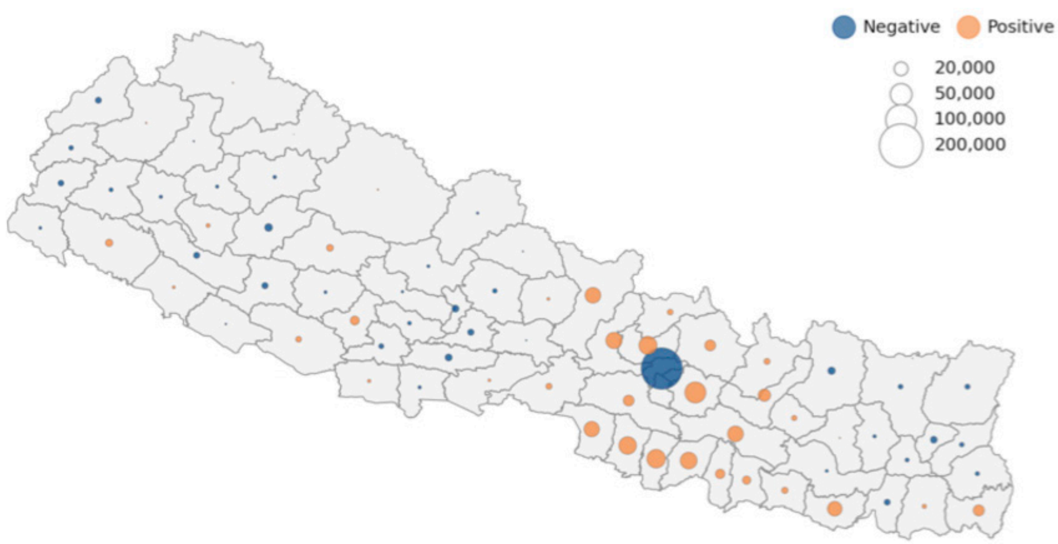
---

30 Dalberg Data Insights "Dalberg Data Insights Identifies Areas At-Risk for Food Insecurity Using Mobile Phone Data | Dalberg." 2018. Accessed October 7, 2019. https://www.dalberg.com/our-ideas/dalberg-data-insights-identifies-areas-risk-food-insecurity-using-mobile-phone-data.

31 "Life Sciences and Health Care Software Solutions — ConvergeHEALTH." 2019. Deloitte United States. Accessed October 7, 2019. https://www2.deloitte.com/us/en/pages/consulting/topics/convergehealth.html.

32 "ConvergeHEALTH by Deloitte and Intermountain Healthcare Expand Real World Evidence Collaboration with Allergan to Focus on Women's Health." 2015. MarketWatch. Accessed October 7, 2019. https://on.mktw.net/31QbEkV

Flowminder's Data-Driven Humanitarian Actions: Flowminder is a nonprofit organization based in Sweden that provides data analytics capability to help development and humanitarian organizations achieve their mission. It acts as a third party, analyzing data held by private companies and providing the result to actors who can provide on-the-ground humanitarian assistance. Flowminder analyzed anonymized cell phone records provided by Namibia's largest network provider, Mobile Telecommunications Limited, to help health workers' efforts to eradicate malaria.[33] It has also analyzed SIM card data to map population movement in Nepal after the 2015 earthquake to coordinate aid delivery,[34] collaborated with UNFPA and WorldPop to obtain more robust population data in Afghanistan,[35] and supported relief efforts in Haiti during the 2010 cholera outbreak by tracking its spatial spread.[36]



flowminder.org

33 "Tracking Malaria in Namibia with Cell Phone Data." 2019. Accessed October 7, 2019. http://datacollaboratives.org/cases/tracking-malaria-in-namibia-with-cell-phone-data.html.

34 "Nepal's Telecom Data & Post-Earthquake Mobility." n.d. Accessed October 7, 2019. http://datacollaboratives.org/cases/nepals-telecom-data—post-earthquake-mobility.html.

35 "WorldPop :: Case Studies." 2019. Accessed October 7, 2019. https://www.worldpop.org/case_studies/mapping_afg_pop.

36 "Flowminder — Haiti Cholera Outbreak 2010." 2015. Accessed October 7, 2019. https://web.flowminder.org/case-studies/haiti-cholera-outbreak-2010.

Open Algorithms (OPAL) Pilot Projects in Colombia and Senegal: OPAL is a not-for-profit project founded by Data-Pop Alliance, Imperial College London, MIT Media Lab, Orange, and the World Economic Forum with a stated goal of securely and ethically collecting private-sector data, analyzing it, and providing insights on various development issues for government or humanitarian organizations. OPAL's platform pseudo-anonymizes selected datasets from company servers. Researchers can then uses OPAL's algorithms to analyze the data in OPAL's secure local database. OPAL had two pilot projects with the governments of Senegal and Colombia, using data from Orange Sonatel and Telefonica Colombia to address development problems.[37]

37 "About OPAL." 2019. OPAL Project. Accessed October 7, 2019. https://www.opalproject.org/about-opal.; DNicolas de Cordes, Orange Data for Development & OPAL. 2016. Accessed October 7, 2019. http://cepei.org/wp-content/uploads/2017/12/D4D-and-OPAL-for-GPSDD.ppt.pdf

DATA POOLING

Companies and other data holders agree to create a unified presentation of datasets as a collection accessible by multiple parties.

	Open Access	Restricted Access
Independent Use	**Public Interfaces**	**Trusted Intermediary**
Cooperative Use	**Data Pooling**	**Research and Analysis Partnership**
Directed Use	**Prizes & Challenges**	**Intelligence Generation**

Collaboratives within this practice area tend to allow for open access to data, either between partners contributing to the pool or among the broader public. Data uses can either be highly cooperative or more independent depending on the level of access and the specific objectives of the pool.

VARIOUS APPROACHES

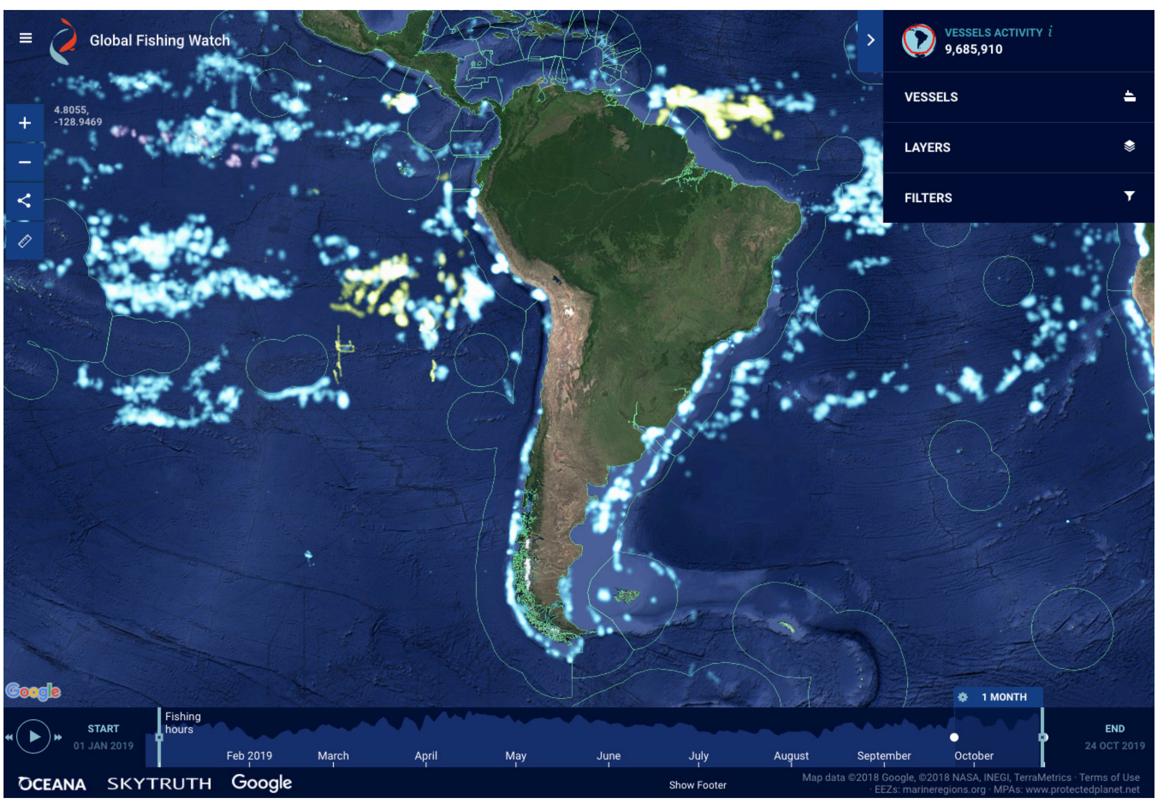
PUBLIC DATA POOLS

Description: Public data pools co-mingle data assets from multiple data holders—including companies—and make those shared assets available on the web. Pools often limit contributions to approved partners (as public data pools are not crowdsourcing efforts), but access to the shared assets is open, enabling independent uses. Nonetheless, the pools are usually developed primarily to provide utility to contributing partners or other user groups such as medical researchers or humanitarian actors.

Examples:

Accelerating Medicines Partnership (AMP): AMP partners—including government, industry, and non-profit health organizations—make genomic and molecular data and analyses publicly accessible to the broad biomedical community through an online portal. The AMP initiative combines public–private expertise and pooled data to reduce the time and cost of developing biomarkers for disease treatment. The project attempts to overcome fragmentation of data assets in the pharmaceutical industry and improve innovation in drug therapy. By combining data, the AMP portal can find new drug targets and reduce wasteful repetition of testing found when companies work in silos.[38]

Global Fishing Watch: Global Fishing Watch is a collaboration among SkyTruth, Oceana, and Google to map and measure fishing activity worldwide using data from the Automatic Identification System (AIS), a ship-tracking system used by large fishing vessels. A map of this data is available to anyone with an internet connection and allows users to monitor when and where commercial fishing is occurring around the world. Governments can use this data to ensure fishing regulations are upheld, allowing them to respond to illegal fishing rapidly and efficiently.[39]



globalfishingwatch.org

38 "Accelerating Medicines Partnership (AMP)." 2019. Accessed October 7, 2019. http://datacollaboratives.org/cases/accelerating-medicines-partnership-amp.html.

39 "Global Fishing Watch." 2019. Accessed October 7, 2019. http://datacollaboratives.org/cases/global-fishing-watch.html.

Global Forest Watch: Global Forest Watch collects crowdsourced, geographic data to track environmental degradation around the world. It includes an ArcGIS online location data portal as well as thousands of authoritative datasets, references, and thematic maps about hundreds of topics. Governments, NGOs, academia, and industry provide data to track deforestation and campaign for more sustainable environmental practices.[40]

Humanitarian Data Exchange (HDX): Launched in 2014, HDX aims to make humanitarian data easy to find and use for analysis undertaken by humanitarian workers across sectors and regions. Actors share the data through an open platform. A team within the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) manages it. The site houses more than 9,800 datasets from 253 locations and 1,200 sources, from companies such as Facebook and the mapping and analytics company Esri. Datasets containing personal information or demographically identifiable information about populations or aid workers are only made available by request for approved users.[41]

PRIVATE DATA POOLS

Description: Partners from different sectors pool data assets in a controlled and restricted access environment. Unlike public data pools, this approach limits data contribution ***and*** data access to only approved partners. Private data pools tend to be highly topic-specific with development and maintenance aimed at serving a particular user group.

Examples:

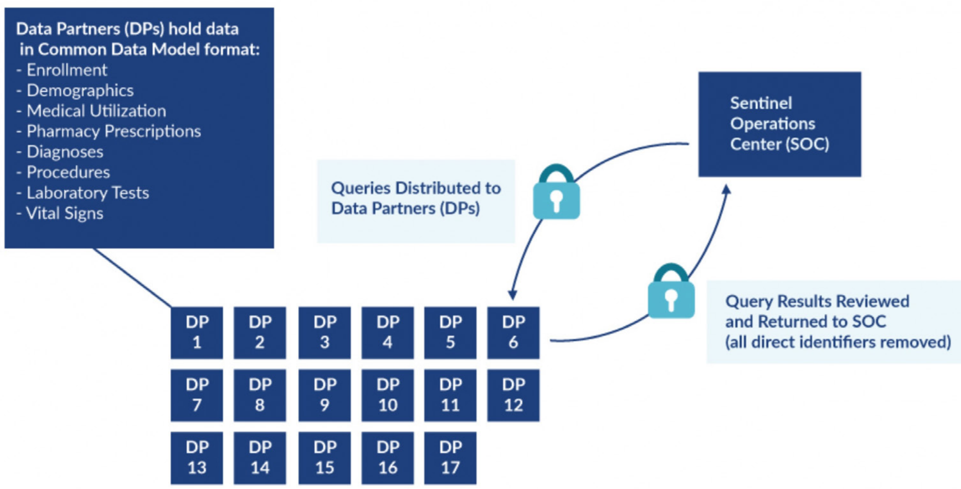
Mobile Data, Environmental Extremes and Population (MDEEP): Grameenphone, a Bangladeshi Telecom company, pooled anonymized mobile CDRs and other data to analyze population movements before and after a climate-driven extreme event, Cyclone Mahasen, which affected 1.3 million people when it struck the southern districts of Bangladesh in May 2013. The project brought together relevant Bangladeshi ministries, telecommunications industry leaders, and the research and disaster response communities to improve the response to future extreme weather events. It is a first-stage project in a longer-term initiative.[42]

40 Ibid.

41 "The Humanitarian Data Exchange." 2019. Accessed October 7, 2019. http://datacollaboratives.org/cases/the-humanitarian-data-exchange.html.

42 "Mobile Data, Environmental Extremes and Population (MDEEP) Project." n.d. Accessed October 7, 2019. http://datacollaboratives.org/cases/mobile-data-environmental-extremes-and-population-mdeep-project.html.

Sentinel Common Data Model (SCDM): The US Food and Drug Administration established the Sentinel Initiative with operations overseen by the Harvard Pilgrim Health Care Institute. It uses a distributed database through which the FDA can run analytical programs on local databases of health providers, such as Humana, Inc. and Blue Cross Blue Shield. The model securely shares and analyzes five data types: administrative data; clinical data; registry data; inpatient data; and mother-infant linkage data. Sentinel intends to actively monitor adverse reactions of medical products after they are on the market.[43]



sentinelinitiative.org

OpenTraffic: Founded by the World Bank, transportation consulting and software company Conveyal, and open-source mapping company Mapzen, OpenTraffic facilitates the exchange and use of data from transportation operators and makes it accessible to select partners.[44] Data contributors can apply to participate in a larger community engaged on sharing anonymized traffic statistics while making use of open source software tools.[45]

UNICEF Magic Box of Data Collaboration: Magic Box is a data collaboration hub created by UNICEF. Magic Box pools data from multiple private companies such as Telefonica, Google, IBM, Amadeus, and Red Hat to allow users at UNICEF to analyze the data and gain insights into various development and humanitarian issues.[46]

43 "Sentinel's Distributed Database | Sentinel Initiative." 2019. Accessed October 7, 2019. https://www.sentinelinitiative.org/sentinel/sentinels-distributed-database

44 Mapzen Open Source Data and Software for Real-Time Mapping Applications to Become A Linux Foundation Project" https://www.linuxfoundation.org/press-release/2019/01/mapzen-open-source-data-and-software-for-real-time-mapping-applications-to-become-a-linux-foundation-project/

45 "OpenTraffic." 2019. Accessed October 7, 2019. http://opentraffic.io.

46 "Data Science and Artificial Intelligence | UNICEF Office of Innovation." 2019. Accessed October 7, 2019. https://www.unicef.org/innovation/Magicbox.

RESEARCH AND ANALYSIS PARTNERSHIPS

Companies engage directly with public-sector partners and share certain proprietary data assets to generate new knowledge with public value.

	Open Access	Restricted Access
Independent Use	**Public Interfaces**	**Trusted Intermediary**
Cooperative Use	**Data Pooling**	**Research and Analysis Partnership**
Directed Use	**Prizes & Challenges**	**Intelligence Generation**

This highly cooperative practice area involves companies transferring data to partners at universities or government statistical offices or external researchers embedding in companies to conduct targeted analysis of certain proprietary data. From a corporate perspective, this practice area can support research that could be used to augment the reach of existing business capabilities, incubate new product ideas, or analyze questions that are out of the scope of internal business operations.

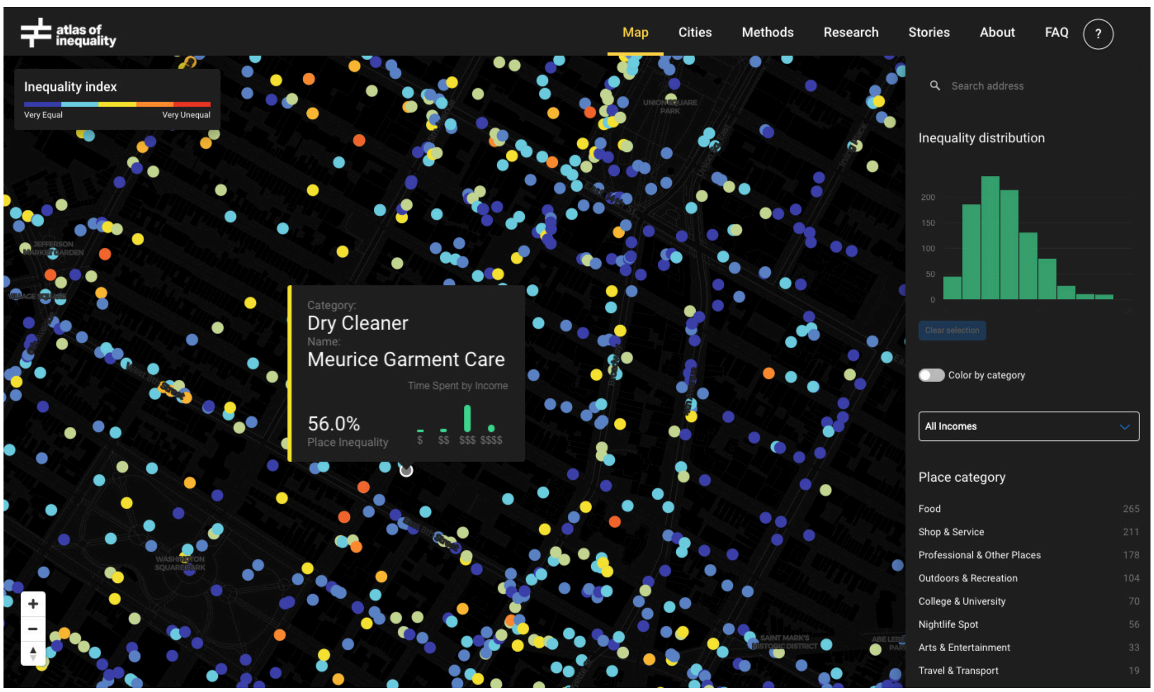
VARIOUS APPROACHES

DATA TRANSFERS

Description: Companies provide data to partners for analysis, sometimes under the banner of "data philanthropy." Access to data is highly restrictive, with only specific partners able to analyze the assets provided. Approved uses are also determined in a cooperative manner, often with some agreement outlining how and why parties requesting access to data will put it to use.

Examples:

Cuebiq Data for Good for MIT Atlas of Inequality: Through its Data for Good Initiative, Cuebiq, a location intelligence company, shared its location data to researchers at MIT Media Lab. These researchers analyzed the data and visualized it with a map to show how economic inequality segregates movement across public spaces.[47]



inequality.media.mit.edu

Dutch Consumer Price Index: Supermarket chains partnered with Statistics Netherlands to create the Consumer Price Index (CPI). The index measures price variation and market changes of goods and services within a certain time period. This index is then used to measure the level of inflation.[48]

Statistics Canada Electricity Consumption Analysis: Statistics Canada obtained a data-sharing agreement with two anonymous smart meter distribution companies to help them better understand electricity consumption patterns. The companies transported

47 "Data Science and Artificial Intelligence | UNICEF Office of Innovation." 2019. Accessed October 7, 2019. https://www.unicef.org/innovation/Magicbox.

48 "The Use of Supermarket Scanner Data in the Dutch CPI–Netherlands .2010. (Consumer Price Index, CPI, Index Number Theory, Netherlands, Scanner Data).". Accessed October 7, 2019. https://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10379.aspx.

the data using an encrypted hard drive to Statistics Canada data storage which was then processed by the agency's analysts.[49]

Valassis's Junk Mail Contribution to Post-Katrina Repopulation Effort: Valassis, a marketing company, shared its massive mailing address database with nonprofit Greater New Orleans Community Data Center, which then used Google Maps and Street View to visualize and track the block-by-block repopulation rate after Hurricane Katrina.[50]

DATA FELLOWSHIPS

Description: Companies establish opportunities for selected individuals or parties to access data assets for a fixed period of time. These fellowship opportunities often involve independent researchers embedded within corporations to analyze corporate assets or corporate staff embedded within public or nonprofit entities and providing their data science expertise to the partner institution.

Examples:

AWS and Azavea Open Source Fellowship Program: Azavea is a social enterprise that uses geospatial technology to tackle various civic, social, and economic issues. In collaboration with Amazon Web Services (AWS), Azavea recruits fellows to work in its office and provides them access to AWS's earth data such as Landsat, Sentinel-2, Sentinel-1, and NEXRAD, among others.[51]

Google.org Fellowship: Google.org, the 501(c)(3) philanthropy operated by Google, offers grant funding as well as the company's technology and expertise to charitable organizations.[52] One of Google.org's offerings is its fellowship program, which embeds Google data scientists, software engineers, and product managers in charitable organizations on a pro bono basis to help them develop their operations. Google Fellows previously worked with the job-placement organization Goodwill Industries International and Thorn, an organization that seeks to prevent child sexual abuse.[53]

---

49 Ma, Lily. "A Big Data Pilot Project with Smart Meter Data (abridged version)." https://www.statcan.gc.ca/eng/conferences/symposium2014/program/14274-eng.pdf

50 Neff, J. "'Junk Mail' to the Rescue in New Orleans." 2008. August 19, 2008. https://adage.com/article/news/junk-mail-rescue-orleans/130414.

51 "Azavea Open Source Fellowship–About." 2019. Accessed October 7, 2019. https://fellowship.azavea.com/about/.

52 "Home." 2019. Google.Org. Accessed October 7, 2019. https://www.google.org/.

53 "Introducing the Google.Org Fellowship." 2019. Google. January 15, 2019. https://migration-dot-gweb-uniblog-publish-prod.appspot.com/outreach-initiatives/google-org/googleorg-fellowship/.

IBM Science for Social Good Fellowship: Started in 2016, this fellowship program provides graduate students and postdoctoral researchers working on social issues such as health, education, and justice mentoring in data science and access to IBM tools. The program intends to help the fellows build and implement data-driven projects. As of September 2019, IBM has awarded 36 social good fellowships.[54]

MasterCard Data Fellows: The Mastercard Center for Inclusive Growth is an independent subsidiary of Mastercard pursuing cross-sector collaboration to further equitable economic development. Its Data Fellows are well-known and well-respected professionals from a variety of disciplines. They include Vint Cerf, Vice President and Chief Internet Evangelist at Google and Melanie Walker, former Director and Senior Advisor at the World Bank Group. Fellows work with MasterCard's data scientists to identify patterns, develop research papers, and glean insights to drive economic growth for underserved segments of society. The insights resulting from this program will be made broadly available following a privacy and data protection review.[55]

54 "Science for Social Good." 2019. Accessed October 8, 2019. http://www.research.ibm.com/science-for-social-good/

55 "Mastercard Center for Inclusive Growth Announces Inaugural Class of Data Fellows."2019. MasterCard Social Newsroom. Accessed October 7, 2019. https://newsroom.mastercard.com/press-releases/mastercard-center-for-inclusive-growth-announces-inaugural-class-of-data-fellows/.

PRIZES AND CHALLENGES

Companies make data available to participants who compete to develop apps; answer problem statements; test hypotheses and premises; or pioneer innovative uses of data for the public interest and to provide business value.

	Open Access	Restricted Access
Independent Use	**Public Interfaces**	**Trusted Intermediary**
Cooperative Use	**Data Pooling**	**Research and Analysis Partnership**
Directed Use	**Prizes & Challenges**	**Intelligence Generation**

Some competitions provide open access to data assets and encourage any user to apply for recognition while others require an upfront application prior to the corporation making that information accessible. Like prizes and challenges in the field of governance innovation more generally, the intention is to use and target expertise to address challenges or opportunities defined by the project's organizers.[56]

VARIOUS APPROACHES:

OPEN INNOVATION CHALLENGES

Description: Companies provide open access to datasets to attract self-selecting participants to develop data-driven solutions, premises, or insights to public challenges. This approach provides limited barriers to entry for data access and analysis. The challenge and data release seek to direct use cases toward addressing particular issues or opportunities as defined by challenge organizers. However, given the openness of data assets, more independent uses of the data could occur, though they would fall out of the challenge's intended scope.

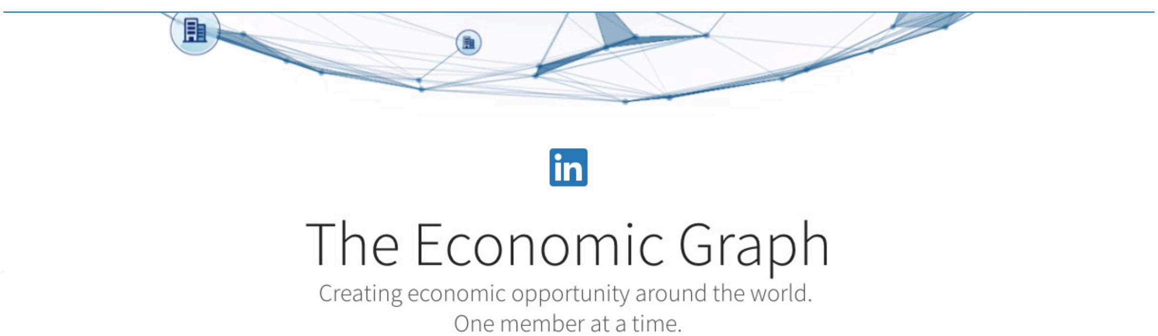
56 Verhulst, Stefaan G. 2015. "Governing through Prizes and Challenges." Medium. January 27, 2015. https://medium.com/@sverhulst/governing-through-prizes-and-challenges-677f3ef861d1.

Examples:

DrivenData's Competition Platform: DrivenData, in collaboration with civil society or international organizations, hosts data competitions on developing tools that create public value, such as an earthquake damage model, heart disease predictor, and electricity use anomaly detector. Participants receive data from both public and private institutions, including the French energy corporation Schneider Electric, the US National Oceanic and Atmospheric Administration, and Cleveland Heart Disease Database.[57]

Kaggle's Data Competitions: The platform Kaggle hosts data-driven competitions by organizations such as Google Research and the mobility company Lyft. These competitions are open to public participation with data that can be downloaded from the website.[58] Example projects look for innovative ways to help autonomous vehicles detect objects,[59] improve the detection of actions in videos,[60] and classify cloud structures from satellite images.[61]

LinkedIn Economic Graph Challenge: In 2017, the online professional network LinkedIn hosted a LinkedIn Economic Graph Challenge. Any U.S. resident over the age of eighteen could submit a proposal to LinkedIn requesting use of its data. After assessing proposals for novelty, impact, and feasibility, LinkedIn awarded winning proposals a $25,000 research prize. Participants had six months to complete their research.[62]



economicgraph.linkedin.com

57 DrivenData. 2019. "Competitions." DrivenData. Accessed October 7, 2019. https://www.drivendata.org/competitions/.

58 "Competitions." 2019. Kaggle. Accessed October 7, 2019. https://www.kaggle.com/competitions.

59 "Lyft 3D Object Detection for Autonomous Vehicles." n.d. Accessed October 7, 2019. https://kaggle.com/c/3d-object-detection-for-autonomous-vehicles.

60 "Understanding Clouds from Satellite Images."2019. Accessed October 7, 2019. https://kaggle.com/c/understanding_cloud_organization.

61 Ibid.

62 "Economic Graph Challenge. 2014. LinkedIn. Accessed October 7, 2019. https://specialedition.linkedin.com/details.

Challenge for participating researchers to build tools and provide research on urban trends and behavior. The dataset includes reviews, locations, restaurant names, and photographs. In keeping with the challenge aspect of the project, Yelp provides a cash prize to students who use the data in a way that demonstrates technical depth, rigor, and relevance. In addition, Yelp keeps updated links to past winners, highlights notable past uses of the data and suggests possible future research avenues to keep individuals engaged. [63]

SELECTIVE INNOVATION CHALLENGES

Description: Selective innovation challenges provide restricted data access to external parties that proposed approved uses of the data to address a public problem. This approach allows data holders and organizers to more closely collaborate with the parties putting their data to use. It can also enable data holders to provide select participants with access to more sensitive data, given increased access controls. Applications often require an articulation of the data protection or responsibility approach to be used by applicants.

Examples:

Data for Climate Action: Data for Climate Action called upon researchers to harness data provided by the multinational banking company BBVA, the global telecommunications provider Orange, Waze, and others for climate-related activities. After submitting a research proposal, participants were called to use the datasets in some way, with the most relevant, impactful, and methodologically sound submissions receiving prizes.[64]

GBDX for Sustainability Challenge: Inspired by the United Nations Sustainable Development Goals, the GBDX for Sustainability Challenge was created to provide innovative problem-solvers with access to geospatial and satellite imagery data as they tried to address problems across topics like disaster response, food security, and poverty elimination. Applicants proposed their ideas on how they would leverage the satellite imagery company DigitalGlobe's geospatial big data platform (GBDX) to solve important sustainability problems. Participants were provided access to GBDX for two months to test their project concepts.[65]

63 "Yelp Dataset." 2019. Accessed October 7, 2019. https://www.yelp.com/dataset/challenge.

64 "Challenges & Hackathons." United Nations Global Pulse. n.d. Accessed October 7, 2019. https://www.unglobalpulse.org/data-for-climate-action.

65 "GBDX for Sustainability Challenge." 2019. Accessed October 7, 2019. http://datacollaboratives.org/cases/gbdx-for-sustainability-challenge.html.

Türk Telekom Data for Refugees (D4R) Challenges: Turkey's state-owned telecommunications company Türk Telekom made anonymized call detail record data available to research groups from non-profit organizations and universities to develop ideas for improving living conditions for the more than 3.5 million Syrian refugees in Turkey. Research groups first submitted research proposals, which a panel of experts evaluated for research merit and addressed one of five prioritized subject areas: safety and security; integration; education; unemployment; and health. Approved organizations then received data access. The competition further aimed to improve applied knowledge of ethical guidelines to big data problems and help capacity-building at a local and international level for creating technological applications to help governments, NGOs, and other stakeholders in dealing with refugee crises.[66]

Orange Telecom Data for Development Challenge (D4D): The D4D Challenge was an international competition in which Orange Telecom offered its data to researchers seeking to address development problems in the Ivory Coast and Senegal. Research teams could apply by sending in a research proposal. Orange then relied on a panel of outside experts representing academic, business, public, and civil society interests to review project proposals and flag those that posed substantial ethical risks under the previously agreed framework. These bodies allowed Orange to see questionable research proposals before they became liable for them.[67]

66 "Turk Telekom Data for Refugees (D4R) Challenge." n.d. Accessed October 7, 2019. http://datacollaboratives.org/cases/turk-telekom-data-for-refugees-d4r-challenge.html.

67 "Orange Telecom Data for Development Challenge (D4D)." 2019. Accessed October 7, 2019. http://datacollaboratives.org/cases/orange-telecom-data-for-development-challenge-d4d.html.

INTELLIGENCE GENERATION

Companies internally develop data-driven analyses, tools, and other resources, and release those insights to the broader public.

	Open Access	Restricted Access
Independent Use	**Public Interfaces**	**Trusted Intermediary**
Cooperative Use	**Data Pooling**	**Research and Analysis Partnership**
Directed Use	**Prizes & Challenges**	**Intelligence Generation**

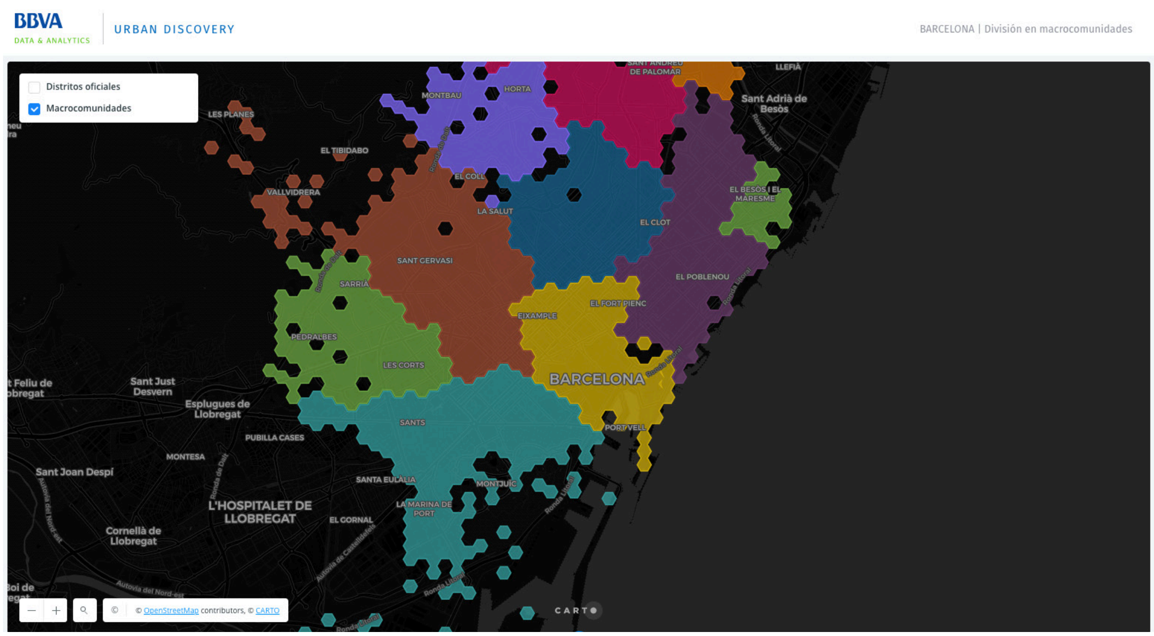
Intelligence generation is a fully restricted access model, with no data being shared with external parties. The analysis generating new intelligence is also highly directed, as it takes place entirely within the data holder's organization. Companies, often using purpose-built research divisions or subsidiaries, conduct analysis on data they hold. This work often aims to aggregate and disseminate insights and analysis to inform policy making and service delivery.

This model does not involve any direct cross-sector sharing or use of proprietary data assets, but it seeks to enable knowledge transfer and inform more evidence-based decision-making in different sectors while restricting access to underlying data. Intelligence generation outputs generally do not allow for peer review, which can complicate attempts to validate or disprove findings.

The intelligence generation practice area is distinct from the other five discussed in this report, but still demonstrates the potential public value that can be created through the analysis of private-sector data assets. Moreover, collaboration with parties from other sectors could occur around defining the focus area of internal data analysis or developing actions based on the outcomes of such analyses.

Examples:

BBVA Urban Discovery: Spanish banking group BBVA collaborated with CARTO, a company that provides spatial intelligence data and tools, to generate insights about city dynamics by analyzing credit card transaction data. The result of the analysis is visualized as a map of population activity, which reveals patterns of economic behaviors of city residents. These insights could inform urban planners or city officials about the ways through which a neighborhood organizes itself, regardless of the administrative boundaries set by the local government. The Urban Discovery initiative has provided visualization and insights of three cities: Madrid; Bar-



bbvadata.com/urbandiscovery

Mastercard Center for Inclusive Growth Donation Insights Report: The Center for Inclusive Growth, a subsidiary of MasterCard, conducts research and data philanthropy with the stated mission of "advanc[ing] sustainable and equitable economic growth and financial inclusion."[69] The Center's researchers make use of anonymized and aggregated MasterCard transaction data, which they analyze internally, and produce reports that summarize their findings. One of these documents is its annual "Donation Insights" report, an analysis of charitable giving patterns.[70]

68 BBVA Data and Analytics. 2019. "Discover New Functional Areas in Madrid Barcelona and CDMX with Anonymized Credit Card Data BBVAData CARTO." Accessed October 7, 2019. http://40.121.52.161/

69 "About the Center for Inclusive Growth." 2019. The Center for Inclusive Growth. Accessed October 7, 2019. https://www.mastercardcenter.org/about-us#our-mission.

70 "Mastercard Center for Inclusive Growth Donations Insights." 2019. Accessed October 7, 2019. https://partners.mastercard.com/donation-insights/.

JP Morgan Chase Institute's Insights Reports: The Institute uses its expertise and proprietary data from the multinational financial corporation to produce research reports on economic issues. These reports are intended for policymakers, businesses, and nonprofit leaders to address issues such as how families manage tax refunds and how local commerce adapts in the digital age.[71]

Using Big Data to Combat Air Pollution: Telefonica's data unit developed LUCA, an AI tool powered by big data, and used it to monitor and predict air pollution in Brazil. Using data from weather stations, air quality sensors, and traffic sensors combined with mobility data from Telefonica, the tool can predict pollution 24-48 hours in advance. Through this initiative, Telefonica sought to improve the city's traffic management and environmental planning.[72]

71 "Expert Insights for the Public Good." JPMorgan Chase Institute. 2019. Accessed October 7, 2019. https://www.institute.jpmorganchase.com/institute.

72 Juárez Carretero, Raquel. 2018. "Using Big Data to Combat Air Pollution in Brazil." Telefonica. 2018. https://www.telefonica.com/en/web/public-policy/blog/article/-/blogs/using-big-data-to-combat-air-pollution-in-brazil.

GOVLAB

TOWARD AN ANALYTICAL & SCOPING FRAMEWORK

The six categories described above illustrate the diversity of models in use across the data collaboratives landscape. They are categorized based on the two main variables: "Engagement" and "Accessibility." These variables, while important, have some limitations in portraying the different nuances in the operations of a particular data collaborative.

In this section, we introduce four categories of variables that can enhance our understanding of the practice areas; and can be used to scope the type of data collaboration needed. Our discussion of these variables—Accessibility, Data Attributes, Collaboration Dynamics, and Scope—is brief by necessity. It is intended to open a conversation that may lead to a more fleshed out analytical framework in subsequent research projects. Nonetheless, these additional variables can enrich the analysis of data collaboratives and their operations and enhance an organization's efforts to assess the practicality and viability of any particular practice area.

The variables can be utilized as a checklist of questions to identify needs and subsequently shape practices. They can guide decisions made in designing a data collaborative, pertaining to who to give access to, how much access to give, where to share the data, which datasets to share, how involved a company can be, or what the scope of the collaboration is.

DATA ACCESSIBILITY

The accessibility of a data collaborative can be gauged along two axes.

Access: Is the data made ***Open***, broadly accessible to potential partners, or ***Restricted***, wherein only pre-selected collaborators gain access?

Availability: Will data be made available to collaborators ***On-Site***, wherein parties access and analyze data without it leaving the company's servers and computer devices, and/or ***Online***, wherein data is made available through a portal, sandbox environment, or other sharing mechanism?

DATA ATTRIBUTES

The attributes of data included within a data collaborative may depend on four characteristics.

Data Temporality: Is the data made accessible to collaborators ***On-Demand***, wherein historic data is held by the provider, or on an ***Ongoing*** basis, in which the company provides data immediately after generation on a real- or near-real-time basis?

Type of Data: Are the accessible data assets ***Pre-Processed*** or in the form of ***Insights***, wherein the underlying data contributing to insights is not provided?

Data Providers: Does the supply side of the data collaborative involve a ***Single Data Provider*** or ***Multiple Data Providers***?

Data Variety: Is the collaborative built around making accessible a ***Single Dataset*** or ***Multiple Datasets***?

COLLABORATION DYNAMICS

The nature of the collaboration itself depends on four key factors.

Engagement: Is the use of data more ***Independent***, in which there is little alignment between the supply side of the collaborative and the eventual use of the data; ***Cooperative***, where terms of analysis and use are determined by the data-demand actor and data supplier together; or ***Directed***, in which the data holder seeks partners for specific, pre-defined uses of its data.

Flow: Will the transfer of data assets be ***Uni-Directional***, wherein data flows from one data provider to one or more demand side actors, or ***Multidirectional***, in which there are data assets provided by multiple entities that flow in more than one direction?

Relationship: Will data collaboration take place on a ***Bilateral***, ***Multilateral***, or ***Distributed basis***?

Sectoral Stakeholders: Do parties in the collaborative come from the ***Private Sector***, ***Public Sector*** and/or ***Civil Sector***?

SCOPE

The scope of a data collaborative can be assessed by considering two main attributes

Focus: Is the data collaborative ***Purpose-Bound***, designed to address a particular challenge, or ***Flexible***, having no upfront determination of how the data is intended to be used?

Timeframe: Is the data collaborative ***Time-Bound***, made accessible for a specific period of time, or ***Open-Ended***, with indefinite data access for collaborators?

The structure of a particular Data Collaborative approach is ultimately determined by the choices made regarding these variables. The visualization of these operational variables can act as a profile, and means of analysis of different approaches to data collaboration (See Figure 3).

FIGURE 3: TEMPLATE PROFILE OF A DATA COLLABORATIVE

DATA ACCESSIBILITY

Open Access	Restricted
On-Site	Online

DATA ATTRIBUTES

Open Access	Restricted
Pre-Processed Data	Insights
Single Data Provider	Multiple Data Providers
Single Dataset	Multiple Datasets

COLLABORATION DYNAMICS

Uni-Directional Data Flow		Multidirectional
Directed	Cooperative	Independent

SCOPE

Purpose-bound	Flexible
Time-Bound	Open-Ended

BOX 2

TECHNICAL CONSIDERATIONS FOR DATA COLLABORATIVES

Although this paper focuses specifically on the more operational components of data collaborative practice areas, technical considerations play a clear role in the viability, effectiveness, and security of data collaboratives. This table provides a non-comprehensive though instructive overview of some key technical considerations for data collaboration, as well as specific techniques currently being used in the field.

Access Controls: Process by which users are granted access and certain privileges to systems, resources or data assets.

Technical Component	Purpose
Homomorphic Encryption	Allows computation to be performed on encrypted data with the result of computation remaining encrypted; can allow collaborative engagement in computations.[73]
Multi-Factor Authentication (MFA)	Requires authorized parties to provide at least two identifying credentials before accessing data in order to protect data from unauthorized access.[74]
Multilevel Secure Database Management System (DBMS)	Seeks to provide users with different access levels to a shared database in order to protect data with varying levels of sensitivity.[75]

73 Homomorphic Encryption Standardization. 2019. Retrieved from https://homomorphicencryption.org/introduction/.

74 Back to basics: Multi-factor authentication (MFA). 2016. Retrieved from https://www.nist.gov/itl/tig/back-basics-multi-factor-authentication.

75 Rjaibi, W. 2004. An introduction to multilevel secure relational database management systems. Retrieved from https://www.researchgate.net/publication/221500703_An_introduction_to_multilevel_secure_relational_database_management_systems.

Data Preparation and Security: Processing data to ensure anonymization, confidentiality, and/or protection of individual privacy.

Technical Component	Purpose
De-identification	Removes identifying information from data to mitigate privacy risks to individuals.
Hashing	One-way encryption of sensitive data that is stored as a hashed value—a unique, randomly generated.[76]
Data Protection Audit	Assesses whether an organization follows best practice when processing personal data, e.g., by complying with GDPR and Freedom of Information rights.[77]
Differential Privacy	Allows the collection and sharing of aggregate information about groups of data subjects without revealing information about individuals in the dataset.[78]
K-Anonymization	Protects confidentiality by ensuring that every combination of identity-revealing characteristics occurs in at least *k* different rows of the data set.[79]
Synthetic Data	Data that is artificially created rather than generated by actual events and can be used to conduct analysis without the risk of a data leak or data breach.[80]
Pre-Computed Indicators	Indicators computed from raw or pre-processed data are aggregated and disconnected from the raw data to minimize privacy risks.[18]

76 Reeves, M. and McMillian, R. Secure Shouldn't Mean Secret: A Call for Public Policy Schools to Share, Support, and Teach Data Stewardship. 2019. Retrieved from https://www.naspaa.org/sites/default/files/docs/2019-09/NASPAA%20Data%20Science%20Curriculum_Reeves.pdf.

77 A guide to ICO audits. 2018. Retrieved from https://ico.org.uk/media/for-organisations/documents/2787/guide-to-data-protection-audits.pdf.

78 Tu, S. Advanced Topics in Data Processing: Introduction to Differential Privacy. 2013. Retrieved from https://people.eecs.berkeley.edu/~stephentu/writeups/6885-lec20-b.pdf.

79 Templ, M., Meindl, B., Kowarik, A., and Chen, S. Introduction to Statistical Disclosure Control (SDC). 2014. Retrieved from http://www.ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf.

80 Montjoye, Y.A., et al. On the privacy-conscientious use of mobile phone data. 2018. Retrieved from https://www.nature.com/articles/sdata2018286?proof=trueJul&draft=journal.

Data Storage and Transfer: Holding data within a single database; copying or moving data between multiple databases	
Technical Component	**Purpose**
Data Enclaves/Safe Sandboxes	A data storage and experimentation technique that provides access to data on a remote server while restricting users from downloading the data or otherwise moving it outside of the secure environment.[81]
Distributed Ledger/ Blockchain	Introduces immutable properties to databases, proving data has not been tampered with.[82]
Edge Processing	Allows data to be processed close to where it is produced (e.g. on a sensor) instead of transmitting it long distances to data centers or the cloud for processing.[83]
Secure Cloud	Stores data securely in the cloud by encoding user data with a specific encryption key and requiring password authentication to access the key.[84]

81 What is the Virtual Data Enclave (VDE)? Data Sharing for Demographic Research. Retrieved from https://www.icpsr.umich.edu/icpsrweb/content/DSDR/faqs/what-is-the-vde.html.

82 Zyskind, G. Using Blockchain to Protect Personal Data. 2015. Retrieved from https://enigma.co/ZNP15.pdf.

83 What is edge computing and how it's changing the network. 2017. Retrieved from https://www.networkworld.com/article/3224893/what-is-edge-computing-and-how-it-s-changing-the-network.html.

84 Keogh, B. 2018. How secure is your data when it's stored in the cloud?. Retrieved from https://theconversation.com/how-secure-is-your-data-when-its-stored-in-the-cloud-90000

GOVLAB

RECOMMENDATIONS AND CONCLUSION

Photo by Brendan Church on Unsplash

Choosing a data collaborative approach is a complex and context-sensitive decision. As indicated by the above, there are various factors to consider; and there is no single approach that is best. Every organization has unique needs, capacities, and desired outcomes. Our analysis does not intend to evaluate or otherwise rank the practice areas on a hierarchical scale. There is no "better" or "worse" type of data collaborative. Each of these practice areas has strengths and limitations, and each has been implemented successfully and unsuccessfully.

Informed by this analysis and an increasingly focused view of the data collaboratives landscape, we offer three recommendations to further advance impactful and responsible data collaboration.

Developing a new scoping methodology for assessing the variables at play in a data collaborative to ensure fit-for-purpose implementations. As this paper makes evident, every data collaborative is the result of a series of choices regarding different variables—like access, engagement, and flow—and the interplay between those choices. There is currently no approach for assessing the benefits, challenges, and tradeoffs associated with the myriad "equations" available to practitioners designing and deploying a data collaborative in different contexts. Future work should leverage the above framework of variables to provide insight regarding what type of data collaborative may be fit for purpose while building an evidence base on the resource requirements, organizational commitments and timeframes, and real-world impacts of different approaches. Such insights will be instrumental in informing the development of a grounded methodology for the field.

Establish and empower data stewards. Given the complexity of potential avenues for collaboration outlined above, there is a need for dedicated individuals or teams within the private sector that can review and implement opportunities for unlocking the public value of a company's data: data stewards. Data stewards need to be established and empowered across the private sector to seek new ways to create public value through cross-sector collaboration. These individuals and teams could drive decision-making related to collaboration around their business's data assets. In partnership with the demand side of data collaboratives, and often with intermediaries and other ecosystem enablers, data stewards can play an essential role in determining the operational approach for collaboration. Indeed, data stewards' work will be key for informing the scoping methodology described in the previous recommendation, and data stewards will represent key users of that methodology.

Explore the creation of new intermediaries positioned to enable data collaboration and lower transaction costs for data suppliers and data users. Data stewards in the private sector and their counterparts in the public sector, civil society, and academia, cannot achieve the potential of data collaboration alone. As outlined above, some intermediaries are already active in the space, including data brokers and providers of third party analytics. New intermediaries are needed that can help define new business models and institutional structures; and to conduct applied research in responsible AI toward making data collaboratives more systematic, sustainable and responsible. In addition, we need the creation of communities of practice bringing together actors in the space, such as the Data Stewards Network.[85]

The analysis presented in this paper is a start, an informational basis describing current practice and providing clarity on the variables that determine a data collaborative. But moving forward, we need more "data about the use of data,"[86] and we need more analysis to achieve actionable intelligence on how to structure data collaboratives given certain requirements and needs.

At The GovLab, we intend to continue expanding and enhancing our Data Collaboratives Explorer, and also to deepen and experiment with our methodology for establishing data collaboratives,[87] which focus on establishing the operational, governance, and technical bases for collaboration. We will complement these products with detailed case studies from around the world to enrich our understanding of these practice areas to help address the most vexing public problems.

85 Young, Andrew. 2019. "About the Data Stewards Network." Medium. March 26, 2019. https://medium.com/data-stewards-network/about-the-data-stewards-network-1cb9db0c0792.

86 Data for Impact "To Turn the Open Data Revolution from Idea to Reality, We Need More Evidence | Apolitical." 2019. Accessed October 7, 2019. https://apolitical.co/solution_article/to-turn-the-open-data-revolution-from-idea-to-reality-we-need-more-evidence/.

87 "Designing a Data Collaborative." DataCollaboratives.org. Accessed October 7, 2019. https://datacollaboratives.org/canvas.html.

GOVLAB

APPENDIX

FULL DATASET OF EXAMPLES

	Public Interfaces	Trusted Intermediary	Data Pooling	Research Partnerships	Prizes and Challenges	Intelligence Generation
23andMe Patient-Centric Research Portal				✓		
Accelerating Medicines Partnership (AMP)			✓			
Ag Data Commons Beta			✓			
Air France-KLM Open Data Portal	✓					
Amsterdam Data Exchange (AMDEX)		✓				
ArcGIS Living Atlas of the World			✓			
AWS and Azavea Open Source Fellowship Program				✓		
Barcode for Life Data (BOLD) System	✓					
BBVA - Measuring People's Economic Resilience to Natural Disaster						✓
BBVA Innova Challenge					✓	
BBVA Urban Discovery						✓
Beeline Crowdsourced Bus Service		✓				
California Data Collaborative (CaDC) Coalition of Water Utilities			✓			
Canada: A Big Data Pilot Project with Smart Meter Data				✓		
CapitalOne Hackathons					✓	
Center for Big Data Statistics: Identifying Population Movements				✓		
China Green Horizon						✓
CIESIN & Facebook: Open, Improved Settlement Data				✓		
Citiesense-Long Island City Partnership Engagement						✓
City Verve Data Portal			✓			
Climate Modeling in Colombia						✓
Clinical Study Data Request Program		✓				

	Public Interfaces	Trusted Intermediary	Data Pooling	Research Partnerships	Prizes and Challenges	Intelligence Generation
Collaborative Study to Diminish Abuse on Twitter				✓		
Consumer Data Research Centre		✓				
Copenhagen-Hitachi City Data Exchange		✓				
Copper Smith and Digital Impact Alliance (DIAL) Partner to Improve Health in Malawi		✓				
Counter-Trafficking Data Collaborative (CTDC)			✓			
Cuebiq Data for Good Initiative				✓		
Dalberg Data Insights		✓				
Data and Analytics Facility for National Infrastructure (DAFNI)		✓				
Data Does Good			✓			
Data for Climate Action (D4CA) Challenge					✓	
data2x Financial Inclusion for Women			✓			
DBpedia			✓			
Deloitte's ConvergeHEALTH, Allergan, and Intermountain			✓			
Digicel Telecom, Karolinska, and Columbia University Partnership		✓				
Digital Dog	✓					
Digital Ecologies Research Partnership				✓		
Digital Matatus				✓		
DigitalGlobe 8-Band Research Challenge					✓	
DiSARM: Disease Surveillance and Risk Monitoring	✓					
DrivenData					✓	
Dutch Consumer Price Index				✓		
Electronic Rothamsted Archive (e-RA)	✓					

	Public Interfaces	Trusted Intermediary	Data Pooling	Research Partnerships	Prizes and Challenges	Intelligence Generation
Enexis Open Data Agenda	✓					
Esoko						✓
Esri and Waze Open Data-Sharing for governments						✓
Estonia: Mobile Positioning Data for Tourism Statistics				✓		
Ethiopian Commodity Exchange	✓					
Facebook, OECD and World Bank Measuring Business Sentiment						✓
Feeding America and Map the Meal Gap						✓
GBDX for Sustainability Challenge					✓	
GlaxoSmithKline and Community Care of North Carolina Medicine Management	✓					
Global Fishing Watch	✓					
Global Forest Watch	✓					
Google DeepMind & NHS		✓				
Google Earth Enterprise	✓					
Google Earth Outreach (GEO) Initiative	✓					
Google Flu Trends						✓
Google Waze (Connected Citizens)	✓					
Google.org Fellowship				✓		
Grampian Data Safe Haven (DaSH)		✓				
GRI's Sustainability Disclosure Database	✓					
Harvard's Malaria Tracking Research		✓				
Health Data Collaborative	✓					
Humanitarian Data Exchange	✓					

	Public Interfaces	Trusted Intermediary	Data Pooling	Research Partnerships	Prizes and Challenges	Intelligence Generation
IBM Science for Good Fellowship				✓		
ImageCat Physical Exposure Database						✓
InBloom	✓					
Industrial Data Space			✓			
Intel's Big Data for Precision Farming		✓				
Johnson & Johnson–Yale Open Data Access Project		✓				
JP Morgan Chase Institute Insights Report						✓
Kaggle's Data Competition Platform					✓	
Kaiser Permanente Research Bank				✓		
Knuper Data Upcycling in Senegal				✓		
LA Tech Talent Pipeline						✓
Lilly, T1D Exchange Diabetes Research Collaboration				✓		
LinkedIn Economic Graph Challenge						✓
LIRNEasia Big Data for Development				✓		
Mapping Snow Melting in the Sierra Nevada				✓		
MasterCard Center for Inclusive Growth Insights Report						✓
MasterCard Center for Inclusive Growth Data Fellows				✓		
Measuring Poverty from Space				✓		
Microsoft Intelligent Network for Eyecare (MINE)			✓			
Microsoft Research Open Data			✓			
Missing Maps	✓					
MIT Laboratory for Social Machines (LSM)				✓		
MIT Senseable City Labs | Urban Exposures				✓		

	Public Interfaces	Trusted Intermediary	Data Pooling	Research Partnerships	Prizes and Challenges	Intelligence Generation
Mobile Data, Environmental Extremse and Population (MDEEP) Project				✓		
Mobile Operator Data for Haiti Earthquake 2010		✓				
Mobile Phone Data for Optimizing Bus Routes in Korea	✓					
Nepal's telecom Data and Post- Earthquake Mobility		✓				
Numina Street Intelligence API	✓					
NYU Langone fastMRI Datasets				✓		
oneTRANSPORT		✓				
Open Algorithm (OPAL) Project		✓				
Open Diversity Data	✓					
Open Mobility Foundation		✓				
Open Targets Consortium			✓			
Open Up Challenge					✓	
Opening up satellite data to track power plant pollution: WattTime						✓
OpenTraffic	✓					
Orange Telecom Data for Development Challenge (D4D)					✓	
Our Brain Bank		✓				
Predicting Consumer Confidence Using Social Media Data				✓		
Project 8: A Data Collaboration Platform For Sustainable Development			✓			
Project Data Sphere			✓			
Properati Data	✓					
Proximus, Statistical Agency of Belgium, Eurostat				✓		

	Public Interfaces	Trusted Intermediary	Data Pooling	Research Partnerships	Prizes and Challenges	Intelligence Generation
Reddit's API for Public Health Research	✓					
Salus Coop		✓				
Sanford Data Collaboratives				✓		
Sentinel Initiative			✓			
SharedStreets		✓				
Sidewalk Labs Urban Data Project						✓
Singapore Data Discovery Challenge					✓	
SoBigData Research Infrastructure				✓		
Social Science One		✓				
Statistics Canada Eletricity Consumption Analysis				✓		
Statistics Finland: Big Data for Official Statistics				✓		
Stats NZ's Data Ventures		✓				
Strava Metro						✓
Telecom Italia's Big Data Challenge					✓	
Telenor Big Data for Social Good				✓		✓
Telkomsel's Big Data for Indonesia's Tourism Statistics						✓
The Climate Corporation's Field View Products	✓					
The Netherland's Center for Big Data Statistics				✓		
TM Forum Open API	✓					
Trackin Anti Vaccination Sentiment with Social Media Data	✓					
Tracking Malaria in Namibia with Cell Phone Data		✓				
Turk Telekom Data for Refugees (D4R) Challenges					✓	
Twitter's Academic Data Grant				✓		

	Public Interfaces	Trusted Intermediary	Data Pooling	Research Partnerships	Prizes and Challenges	Intelligence Generation
Uber and City of Boston Partnership				✓		
Uber Movement	✓					
UK-Brazil collaboration on leveraging crowdsourced and sensor data to support urban resilience				✓		
UN Global Pulse & DataSift				✓		
UN Global Pulse & Twitter				✓		
UNFPA, WorldPop, and Flowminder Helps Map Afghanistan Population		✓				
UNICEF Magic Box of Data Collaboration			✓			
UNICEF and Facebook Zika Campaigns						✓
Using Big Data to Combat Air Pollution in Brazil						✓
Using Mobile Phone Data to Track Puerto Rico's Hurricane Migration		✓				
Utilizing mobile analytics to inform emergency disaster response in Turkey						✓
Valassis Data-Driven Disaster Recovery in New Orleans				✓		
Washington D.C. works with Populus to Use Bike-Share Data to Improve Mobility		✓				
We Feel	✓					
World Bank and LinkedIn Data Sharing Initiative	✓					
Yale Open Data Access (YODA)		✓				
Yelp Dataset Challenge					✓	
Yelp LIVES score		✓				
Zillow Research	✓					

DATA COLLABORATIVES

LEVERAGING PRIVATE DATA FOR PUBLIC GOOD

A Descriptive Analysis and Typology of Existing Practices

Stefaan G. Verhulst
Andrew Young
Michelle Winowatan
Andrew J. Zahuranec