THE DATA ASSEMBLY

Responsible Data Re-Use Framework

By Andrew Young, Stefaan G. Verhulst, Nadiya Safonova, and Andrew J. Zahuranec

November 2020

THEGOVLAB

HENRY LUCE FOUNDATION

Authors

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.

Stefaan G. Verhulst is Co-Founder and Chief Research and Development Officer of The GovLab, where he is building an action-research foundation on how to transform governance using advances in science, data, and technology. Verhulst's latest scholarship centers on how technology can improve people's lives and the creation of more effective and collaborative forms of governance. Specifically, he is interested in the perils and promise of collaborative technologies and how to harness the unprecedented volume of information to advance the public good.

Nadiya Safonova is a Research Assistant at The GovLab. Nadiya is a member of the Data Program research team and focusing especially on the Data Assembly project, which seeks to increase our understanding of the varied needs and perceptions regarding the responsible re-use of data. Previously, she worked for the Open Government team at the Treasury Board of Canada Secretariat, helping to support the implementation of Canada's National Action Plans on open government. She is a recent masters graduate from the Faculty of Public Affairs at Carleton University. She also holds a BA in International Relations from Mount Allison University, in Canada.

Andrew J. Zahuranec is a Research Fellow at The GovLab, where he is responsible for 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 from Mercyhurst University.

Acknowledgements

The authors would like to thank Mariko Silver and Toby Volkman at the Henry Luce Foundation whose support and guidance were instrumental. This work would not be possible without the participation of the members of our three mini-publics; we thank them for their thoughtful reflections and contributions. Thanks also to our partners and advisors in the New York Public Library, Brooklyn Public Library, and New York City Government, including particularly Oliver Bjornsson, Zachary Feder, Diana Plunkett, Brent Reidy, Kathleen Riegelhaupt, Luke Swarthout, and Adrienne Schmoeker. We also appreciate the expert counsel we received from Larissa De Lima, Mahmud Farooque, Ana Kreacic, Panthea Lee, Yves Mathieu, and Lee Rainie. Finally, we are grateful to our GovLab colleagues Beth Simone Noveck, Michelle Winowatan, Anirudh Dinesh, and Mary Ann Badavi for their strategic, research and design contributions.

TABLE OF CONTENTS

Key Takeaways 6

Introduction 9

- Data Re-Use and COVID-19 9
- The Opportunities and Risks of Data Re-Use 10
- Why We Need a Data Assembly 11
- Insights and Outputs of the Data Assembly 12

Methodology 14

- Data Re-Use Exhibits 15
- Design Wheel of Data Re-use 17

The Responsible Data Re-Use Framework: Emerging Principles from the Data Assembly 19

- Why 19
- What 20
- Who 20
- How 21
- When 21
- Where 21

Lessons Learned from the Three Mini-Publics 22

- 1. New Yorkers Mini-Public 23
- 2. Rights Groups and Advocacy Organizations Mini-Public 31
- 3. Data Holders and Policymakers Mini-Public 35

Appendix 1: Remesh Questionnaire Used in New Yorkers Mini-Public 40

Appendix 2: Further Reading on Public Perceptions and the Responsible Re-Use of Data for COVID-19 50

KEY TAKEAWAYS

The Data Assembly is an initiative from The GovLab supported by the Henry Luce Foundation to solicit diverse, actionable public input on data re-use for crisis response in the United States. The initiative began in the summer of 2020 with an initial focus on the response to the COVID-19 pandemic in New York City. The GovLab, New York Public Library, and Brooklyn Public Library co-hosted remote deliberations with three "mini-publics" featuring data holders and policymakers, representatives of civic rights and advocacy organizations, and New Yorkers from across the five boroughs. These deliberations yielded five key takeaways regarding the responsible re-use of data for COVID-19:

1. Match Urgency with Accountability

Participants in all three mini-publics expressed a willingness to tolerate increased surveillance for public health purposes. However, this expanded support for data collection and re-use does not excuse organizations from abiding by responsible data practices and other basic duties of care. Organizations should provide mechanisms that guarantee public oversight of their actions and provide opportunities for public input and accountability.

2. Support and Expand Data Literacy

Though recent events have prompted more awareness of data re-use, many government leaders, community groups, and members of the public lack knowledge of certain data practices and terminology. As such, meaningful public participation (including informed consent) in a data re-use effort depends on all communications being clear, well-justified, and broadly understandable. Various actors, including public libraries, could play an important role in fostering data literacy.

3. Center Equity

Data re-use can yield substantial benefits for a community, but these benefits are not always distributed to those who need them most. In the mini-publics, participants noted the ability for data projects to miss subsets of the population or otherwise exacerbate existing inequalities. To address these problems, organizations should consider whether the data they intend to re-use misses or under-represents any groups or whether the methods have the potential to otherwise cause harm.

4. Engage Legitimate, Local Actors

Participants in the mini-publics highlighted the need for effective public engagement and leadership from local actors in government and civil society. The deliberations also pointed to the importance of involving trusted intermediary organizations working at the local level that can help to engage with and solicit input from target beneficiary communities.

5. Develop Positions for Responsible Data Re-Use

Data re-use projects are complex undertakings that require coordination with various actors inside and outside an organization. Dedicated positions devoted to coordinating data re-use can allow organizations to better respond to new circumstances as they arise. In The GovLab's work, we call the people in these positions "data stewards."

INTRODUCTION

DATA RE-USE AND COVID-19

The COVID-19 pandemic is an economic, political, cultural and, above all else, human tragedy without parallel in our recent history. The path out of this crisis remains unclear; the world is gripped by a profound sense of uncertainty about what comes next. Though vaccines, therapeutics, and non-pharmaceutical interventions each play a role, none alone are a silver bullet. We will need an arsenal.

Data is likely to be an important—perhaps even essential—component of this arsenal. At The GovLab, we have conducted extensive research in countries around the world to understand how data can be used responsibly and safely to help decision-makers navigate

Photo by Brendan Church on Unsplash

complex problems such as the one we are now facing. Our work has focused on the re-use of data, often through collaborative mechanisms and arrangements (both formal and ad hoc) between the public and private sectors and also between the private sector and civil society.

THE OPPORTUNITIES AND RISKS OF DATA RE-USE

Re-use can often take place through the framework of data collaboration,[1] which now occurs widely through a variety of institutional, contractual and technical structures and instruments. Borrowing in language and inspiration from the open data movement, the emerging data collaborative movement has demonstrated its capacity for creating public value, including specifically as it relates to the COVID-19 pandemic. Data re-use has the potential to improve disease treatment, identify better ways to source supplies, monitor adherence to non-pharmaceutical restrictions, and provide a range of other public benefits. From these applications, it is clear that data has tremendous potential to mitigate the worst effects of this pandemic.

As promising and attractive as re-using data might seem, data re-use also comes with challenges. For instance, privacy advocates, citizens' groups and other stakeholders are anxious that—absent a clear framework of checks and balances and strong oversight institutions—today's "new normal" of data sharing may turn into tomorrow's tools of surveillance and oppression. These concerns cannot be dismissed as alarmism or conspiracy theories.

In addition, while the potential of data sharing may be large, significant obstacles can impede its positive impact. Some obstacles are technical and stem from a lack of interoperable systems or problems with data quality. Others relate to a lack of capacity within organizations both supplying and consuming data to formulate questions that matter and make sense of data.

---

[1] Data collaborations are new forms of collaboration, beyond the public–private partnership model, in which participants from different sectors exchange their data to create public value. See: https://datacollaboratives.org/

Some of the hardest—and perhaps least understood—obstacles, though, stem from another area: governance. Broadly, these governance challenges manifest in difficulties that regulators, government leaders, and societies face when trying to determine an appropriate balance between the potential benefits and costs of re-using data and identifying a system of appropriate checks and balances. The questions raised by such difficulties are inherently normative and value-based; they involve a range of judgements, particularly about how to distribute benefits and risks evenly across different communities and stakeholders. Appropriate responses are highly context dependent. Solutions, thus, must be complex and come with great variability. Policymakers must find solutions that reconcile different (sometimes contradictory) needs of various stakeholders and groups.

WHY WE NEED A DATA ASSEMBLY

A central difficulty in achieving this reconciliation is that policymakers and data holders often have little understanding of how different communities of users feel about the underlying issues—especially the trade-offs between risk and benefit that are inherent to data re-use. As a result, regulators and government leaders often find themselves torn between competing impulses. On the one hand, they may adopt sharing and re-use policies that could endanger privacy and other rights of users, particularly those from traditionally marginalized communities. On the other hand, excessive caution may severely limit the options for data re-use out of fear of violating those rights, curtailing the wider societal benefits. This conflict between over-sharing and not sharing enough is the central conundrum faced by data governance today. It is one we hope to begin addressing with this project.

Signals about social attitudes and values toward data tend to come from op-ed pieces in newspapers and broad surveys of public opinion, such as those contained within the Pew Research Center's *Americans and Privacy* report. Such surveys can provide a useful snapshot of public opinion, but they tend to lack the nuance and contextual discussion enabled by more deliberative methods. Deliberative public engagement methodologies (e.g. citizens' juries, citizens' assemblies, and public dialogues) offer a more context-rich approach,

allowing us to understand how different constituencies make value judgements and how they perceive challenges and risks involved in data sharing.

To achieve broadly acceptable policy solutions that harmonize and address the needs of as many stakeholders as possible, we initiated deliberations with three cohorts involved in or impacted by data re-use:

1	A mini-public of New York residents that aims to be representative of various community groups and individuals;
2	An assembly of rights groups and advocacy organizations; and
3	A selection of data holders and policymakers operating in New York City.

INSIGHTS AND OUTPUTS OF THE DATA ASSEMBLY

The Data Assembly seeks to provide decision-makers in New York City and beyond with a clear understanding of the expectations of public stakeholders regarding the responsible re-use of data. This work intends to clarify the necessary conditions and institutional procedures and processes needed to mitigate the risks of data re-use for crisis management and policymaking. The insights and guidance generated by this effort also aim to instill trust among citizens that others are managing data in an accountable, transparent, and safe manner.

The core output of the Data Assembly is a *Responsible Data Re-Use Framework,* which seeks to inform decision-makers on how best to re-use data to solve public problems, such as COVID-19. This framework, presented below seeks to inform if, when and how the re-use

personal data can be aligned with people's expectations and societal values. While the initial focus of the Data Assembly falls on data re-use for COVID-19 measures in New York City, the framework is intended to be applicable to policymaking and data re-use project design in other contexts.

Going forward, the ***Responsible Data Re-Use Framework*** and insights generated through the Data Assembly will also help to inform data literacy programs in New York City with our partners at the New York Public Library and Brooklyn Public Library.

METHODOLOGY

The Data Assembly deliberations took place during July and August of 2020. The GovLab and its partners at the New York Public Library and Brooklyn Public Library facilitated 90-minute remote video conferences with the data holders and policymakers mini-public and the rights groups and advocacy organizations mini-public. Both of these consultations involved between 15–20 experts curated using the GovLab's Smarter Crowdsourcing methodology.

Photo by Alex Kondratiev on Unsplash

The New Yorkers Mini-Public deliberation occurred on Remesh, an online research and public engagement platform. This consultation featured 55 New York City residents, sourced through the Remesh sampling methodology, with a focus on diversity across age, gender, income, and borough of residence. The Remesh platform provided participants with the ability to respond to polling questions, free-form text prompts, and to indicate their support for the contributions of their fellow participants.

DATA RE-USE EXHIBITS

The Data Assembly presented participants in each of the three mini-publics with three generalized examples of data being re-used to support the response to COVID-19. These "Data Re-Use Exhibits" were grounded in reality, with each of them based on projects launched since the emergence of COVID-19 or data collaboratives initiated by real-world actors in similar contexts. Wherever possible, the exhibits did not include the names of particular companies or organizations to avoid biasing participants' reactions.

The three Data Re-Use Exhibits are summarized here.

EXHIBIT A Mobility Data Analysis	A telecommunications company collects location data from devices that downloaded certain apps and opted in to the app tracking where they are. Apps are maps, step counters, and weather apps unrelated to public health. The company processes data to remove identifying features such as names, phone numbers, and device identification numbers. It shares its aggregated data with health agencies, local government, and university researchers to help them assess adherence to lockdown policies. Data is kept for the crisis duration. Assessments of it may be used in a variety of ways, such as informing future policy changes or redistributing enforcement to different neighborhoods.

EXHIBIT B Consumer Data	Consumer spending during the pandemic falls drastically. A major credit card company analyzes its consumer spending habits throughout the pandemic and stages of reopening. The company follows an internally defined procedure to aggregate and anonymize the data. The company and a local economic recovery government agency announce a partnership that will allow the agency to conduct analyses of consumer data in a safe sandbox environment (an isolated environment on a secure network). Insights from the analysis are used for internal decision-making within the government agency.
EXHIBIT C 311 Data	A city government department is responsible for responding to and triaging non-emergency calls and complaints from city residents. During the pandemic, the number of calls related to social distancing and face covering-related complaints spike. The majority of these complaints are directed to the city's police department, which is one of the agencies supporting the enforcement of social distancing rules. Information about complaints and resolutions is recorded on the city's open data platform, including date, neighborhood, incident address, complaint type and description. This information is stored on the open data platform indefinitely.

DESIGN WHEEL OF DATA RE-USE

To further guide the deliberation, The GovLab and its partners presented an analytical tool outlining key considerations involved in a data re-use project. This Design Wheel of Data Re-Use (pictured below) provided a framework for participants to reflect on five key elements of such initiatives:

WHY
The purpose, scope, and limitations of a data re-use project;

WHAT
Data assets used in a project and their standards, formats, and technical requirements;

WHO
Actors, including data providers, data demand actors, data subjects, and intermediaries involved in these projects, as well as their custodial duties, data access criteria, rights and responsibilities;

HOW
The operational strategy and governance framework for data re-use;

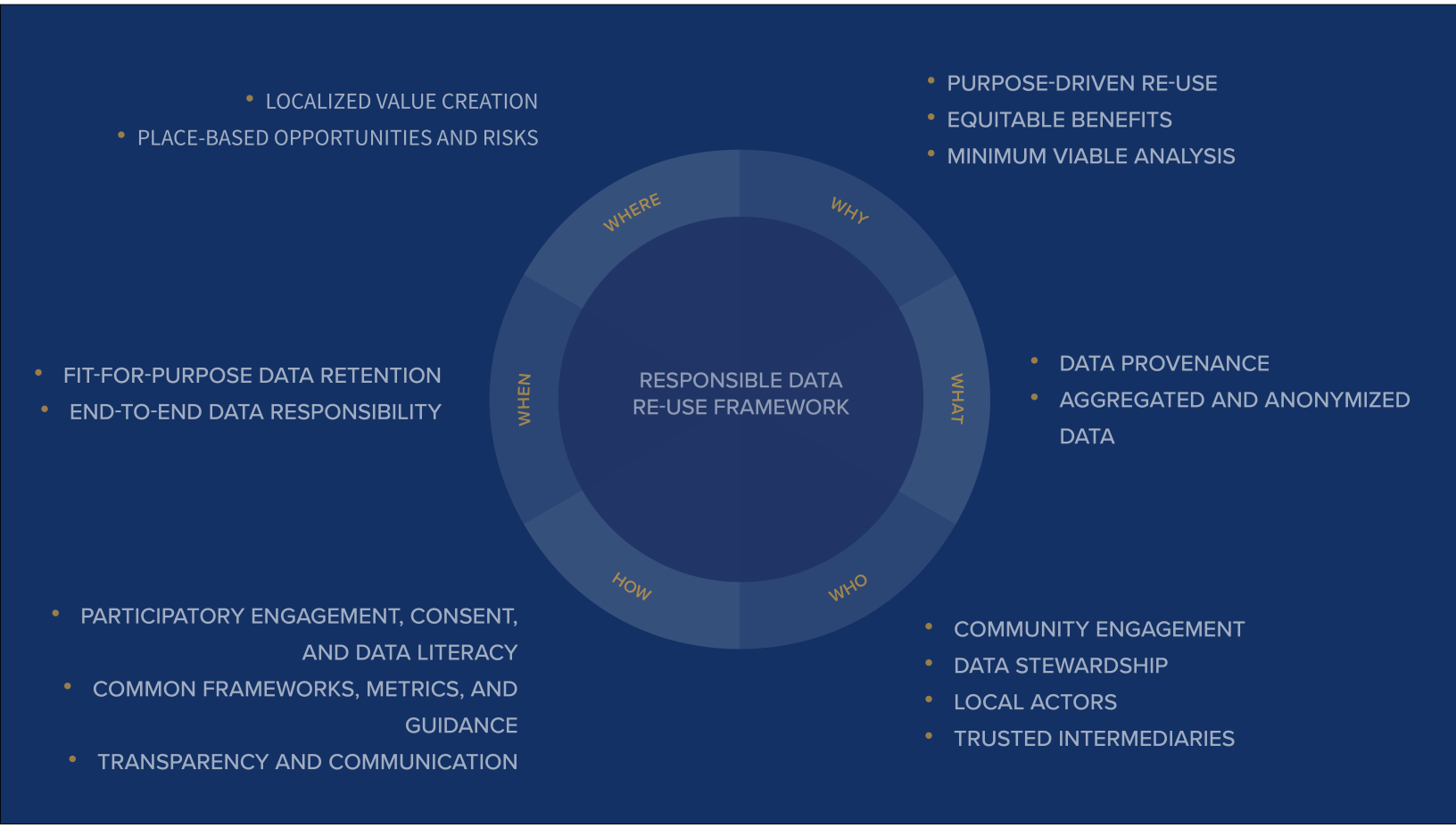
WHEN
The duration of the data re-use effort, including provisions for data retention, termination, and modification; and

WHERE
The regional focus, contextual and jurisdictional implications of the data re-use project.

The issues of *Why*, *What*, *Who*, *How*, *When*, and *Where* are subsequently used to organize the recommendations and findings outlined in this report.

In what follows we present the initial draft ***Responsible Data Re-Use Framework,*** which consolidates and synthesizes key insights from across the three mini-public deliberations. We then summarize the more detailed findings from each of the three mini-publics in turn.

THE RESPONSIBLE DATA RE-USE FRAMEWORK: EMERGING PRINCIPLES FROM THE DATA ASSEMBLY



WHY

- **Purpose-Driven Re-Use:** The re-use of data in the context of COVID-19 should be tied to a clear and well-defined purpose.

- **Equitable Benefits:** Practitioners should prioritize data re-use that benefits all people, including under-served populations and those who are "invisible" in many institutional datasets.
- **Minimum Viable Analysis:** Practitioners should only re-use data when it is the most direct, least invasive means to obtain the desired outcome.

WHAT

- **Data Provenance**: Practitioners should capture and communicate the origin, potential biases, limitations, and previous uses of datasets to ensure that those re-using the data are clear on what insights the data can and cannot provide.
- **Aggregated and Anonymized Data:** While recognizing that risks can never be fully erased, practitioners should ensure an adequate level of data aggregation to guard against group privacy harms and re-identification of individuals.

WHO

- **Community Engagement:** Community leaders and members of the general public should be involved in the planning stages of data re-use to help clarify what is "mission critical" and valuable to them.
- **Data Stewardship:** Data re-use should never be a fully automated process. Human actors need to be involved to ensure data quality and accuracy, and provide oversight throughout the data lifecycle.
- **Local Actors:** Where possible, the re-users of data should be actors in local governments, nonprofits, businesses, or academia in close proximity to the problems at hand and intended beneficiaries.
- **Trusted Intermediaries:** Beyond data suppliers and data re-users, trusted third parties should be empowered to help support responsible, ethical, and legally sound data re-use.

HOW

- **Participatory Engagement, Consent, and Data Literacy:** Practitioners should engage data subjects and community leaders at the planning stage of a project to steer responsible data re-use. They should also seek meaningful consent, with the ability to opt-out prior to the initiation of data re-use. Clear and accessible language and data literacy education can support these efforts.
- **Common Frameworks, Metrics, and Guidance:** No one-size-fits-all approach will suffice for responsible data re-use. Nonetheless, practitioners should seek out best practices and engage with stakeholders to create repurposable public resources to support peer-learning and collaboration.
- **Transparency and Communication:** Throughout the data re-use lifecycle, practitioners should communicate regularly with data subjects regarding how their data is being handled and how it is (or is not) contributing to the intended purpose of the work.

WHEN

- **Fit-for-Purpose Data Retention:** Data should only be held for as long as necessary to address the core issue or to answer the key question that is driving the re-use project. Future-oriented or exploratory analyses require new consent.
- **End-to-End Data Responsibility:** Opportunities, risks and challenges exist at all stages of the data re-use cycle. Policies, procedures, and oversight should be designed and deployed with a focus on navigating inevitable shifts in circumstance over time.

WHERE

- **Localized Value Creation:** Practitioners should prioritize re-using data to address local, community-based problems and opportunities first and foremost.
- **Place-Based Opportunities and Risks:** The re-use of geolocation data in particular can lead to emergent or unexpected risks and challenges. Data stewards should be tasked with assessing and mitigating place-based risks on a regular basis.

LESSONS LEARNED FROM THE THREE MINI-PUBLICS

As discussed, The GovLab hosted three mini-public deliberations for this effort: a mini-public of 55 randomly selected New York residents; a curated discussion of 17 rights groups representatives; and a curated discussion of 18 data holders and decision-makers. Summarized takeaways from these deliberations are provided below.

Photo by Colton Duke on Unsplash

1. NEW YORKERS MINI-PUBLIC

The first of these mini-publics involved 55 randomly selected residents of New York. As discussed above, this deliberation took place through the Remesh platform, allowing for additional quantitative analysis of participants' responses.

Level of Support for Data Re-Use Exhibits

Table 1: Support of Data Re-Use Cases Across Segments of the Mini-Public

Data Re-Use Exhibits	Participants Supportive of the Data Re-Use Case
EXHIBIT A Mobility Analysis	88% either supportive or very supportive
EXHIBIT B Consumer Data	29% either supportive or very supportive
EXHIBIT C 311 Data	65% either supportive or very supportive

Table 2: Support of Data Re-Use Cases Among Younger vs. Older Participants

Data Re-Use Exhibits	Participants between the ages of 18–34 (n=26)	Participants between the ages of 35–64 (n=28)
EXHIBIT A Mobility Analysis	81% either supportive or very supportive	92% either supportive or very supportive
EXHIBIT B Consumer Data	19% either supportive or very supportive	37% either supportive or very supportive
EXHIBIT C 311 Data	52% either supportive or very supportive	77% either supportive or very supportive

Across each of the three data re-use cases, older participants were more likely to be either supportive or very supportive, with younger participants less likely to be so.

Table 3: Support of Data Re-Use Cases Among Participants with an Immediate Family Member Who has Been Diagnosed with COVID-19	
Data Re-Use Exhibits	**Participants with a family member who has been diagnosed with COVID-19 (n=12)**
EXHIBIT A Mobility Analysis	84% either supportive or very supportive
EXHIBIT B Consumer Data	17% either supportive or very supportive
EXHIBIT C 311 Data	83% either supportive or very supportive

Participants with family members who have been diagnosed with COVID-19 — just under 22% of the sample — were slightly less likely to be supportive or very supportive of both the mobility analysis and consumer data re-use cases. They were, however, significantly more likely to be supportive of the 311 data re-use case.

Table 4: Support of Data Re-Use Cases Among Participants with an Immediate Family Member Who has Lost Their Source of Income Due to COVID-19	
Data Re-Use Exhibits	**Participants with a family member who has lost their source of income (n=28)**
EXHIBIT A Mobility Analysis	89% either supportive or very supportive
EXHIBIT B Consumer Data	26% either supportive or very supportive
EXHIBIT C 311 Data	71% either supportive or very supportive

Notably, people who have an immediate family member who lost their source of income were less likely to be supportive or very supportive of the consumer data re-use case compared to the full sample.

Participants' Familiarity with Key Concepts

Facilitators polled participants regarding their level of familiarity with key concepts involved in each data re-use exhibit. We then defined each concept to ensure that participants understood important components of each example. This process helped to contextualize participants' responses and to identify areas of focus for future data literacy and education efforts.

Table 5: Familiarity with key data re-use concepts

	I know what it is	I have heard of it but don't know what it is	I have not heard of it
Aggregated Data	29% of participants	47% of participants	24% of participants
Safe Sandboxes and Secure Data Processing Environments	17% of participants	30% of participants	53% of participants
New York City's 311 System	82% of participants	16% of participants	2% of participants
Open Data	34% of participants	33% of participants	33% of participants

Findings Across the Design Wheel of Data Re-Use

In what follows we delve into the detailed findings from across the Design Wheel of Data Re-Use.

WHY

- Overall participants were most supportive of mobility data analysis and saw the value in tracking aggregated movement patterns. They were least supportive of re-using consumer data from credit card companies
- Across the three use cases, participants made clear that the re-use of data should always be targeted at benefiting the community and informing decision-making that can slow the spread of COVID-19 and enable re-opening of the New York City economy.
- Participants generally did not support the re-use of credit card data because of the perceived invasiveness or riskiness of analyzing financial activity. Notably, many participants indicated their belief that the insights this type of analysis could provide could be attained through other, less invasiveness means. This demonstrates the importance of not just a clear and compelling purpose for data re-use, but also the need for identifying the most direct, least invasive pathway to the desired insight.
- While participants made clear their lack of support for individual-level data analysis and re-use, they indicated consistent support for data re-use toward enforcing social distancing provisions and the wearing of face masks. Participants indicated data re-use can be appropriate for supporting institutional decision-making and for keeping residents accountable to fellow New Yorkers.

> **"I think it is very important that researchers have specific knowledge of individuals where they go, what they do and how this epidemic spreads."**
>
> — Participant in New Yorker Mini-Public

WHAT

- Across the deliberation, participants reiterated that data needs to be aggregated so individuals cannot be identified.
- The appropriateness of re-using location data and 311 citizen reports were shown to be largely contextual, with varied support depending on the use case and partners involved in the effort.
- The re-use of credit card data, on the other hand, was deemed to be inappropriate in essentially all cases.

WHO

- Participants were particularly supportive of the data being re-used by researchers that could inform decision-making by actors in the public sector. This likely indicates the importance of trusted intermediaries being engaged in the data re-use process.
- Public health agencies and local government agencies were also supported by most participants across use cases.
- The majority of participants were supportive of the re-use of 311 data in particular by law enforcement.

> **"Academic researchers, nonprofits, and some government agencies should be allowed to re-use aggregated data."**
>
> — Participant in New Yorker Mini-Public

HOW

- Across the data re-use cases, three core issues repeatedly emerged as important to the perceived appropriateness of data re-use:
    - **Transparency** regarding the intentions, operations, parties involved, and outcomes of a data re-use effort;
    - The upfront ability to **opt out** of data re-use — including re-use of aggregated data; and
    - The presence of an independent **third party** empowered to influence data re-use and ensure that it aligns with the interests of the public.
- Participants were presented with a number of specific, operational strategies for ensuring responsible data re-use. Here we list their level of support for these efforts in descending order. Note that participants could select more than one response:
    - The ability of data subjects to opt out of any uses of their aggregated data — **63% supportive;**
    - Creating an independent council of experts to oversee and provide ethical guidance on a re-use project—**54% supportive;**
    - Making legal agreements between data suppliers (e.g. telecom companies) and data users (e.g. researchers) publicly accessible—**41% supportive;** and
    - Creating a representative panel of regular New Yorkers to help decide what types of data re-use would be appropriate—**41% supportive**.
- Participants were largely supportive of non-sensitive data being made open through the city's open data portal. Participants also indicated their interest in the outcomes of 311 data re-use in particular to be made open and transparent. This again highlights the need to center purpose and outcomes in the re-use of data.

> **"I agree that data should be made accessible to the public as open data, since we fund the service, therefore, we should have the access to the data that they collect. However, this can also come with a catch as well, given that data could be abused."**
>
> — Participant in New Yorker Mini-Public

WHEN

- Participants were supportive of the shortest period of data retention possible across the three data re-use exhibits. Like much of the discussion across the mini-publics, contextual considerations came into focus—especially the need for specific timeframes for different types of data re-use.
- In cases where data could provide analytical or research value beyond the period of the pandemic and response, participants indicated that actors should obtain new consent from data subjects in order to continue retaining and analyzing relevant datasets.
- As discussed above, participants made clear that individuals should be given the ability to opt out of data re-use prior to the initiation of data sharing or analysis.
- The deliberation raised some contradictory results related to the opening of data and the placement of 311 data on the city's open data portal. Most participants supported the opening of 311 data, especially since the system is taxpayer funded. A majority also indicated that the data should only be retained as long as the pandemic and response continues. These opposing messages signal the need for effective communication and engagement with citizens regarding how and how long data is made broadly accessible. Participants did not appear to recognize the issue of "the horse having already left the barn" as it relates to data retention periods and the opening of datasets to the public.

> **"I don't believe any assurance would be enough to provide 100% confidence. There's always human error."**
>
> — Participant in New Yorker Mini-Public

WHERE

- While equitable benefits were a common theme across the other two mini-publics, participants in the New Yorkers mini-public championed more localized and community-based value creation from data re-use cases.
- The majority of participants were not concerned about place-based data privacy and security issues. Participants were comfortable with the collection and analysis of aggregated location data drawn from potentially sensitive areas, such as protests or places of worship. This reaction differs significantly from insights drawn in both the dataholder and policymaker mini-public as well as the rights and advocacy group mini-public. This likely speaks to the composition of the mini-public, which was diverse across various criteria outlined above, but likely did not include a large number of individuals with specific sensitivities related to, for example, political activism or religious sensitivities.

2. RIGHTS GROUPS AND ADVOCACY ORGANIZATIONS MINI-PUBLIC

WHY

- Participants highlighted the importance of equitable benefits of data re-use, with a focus on including those usually missing or underrepresented in data sets.
- This cohort indicated that data re-users should closely consider the necessity of data re-use. Stakeholders should not introduce data risks, especially affecting already vulnerable communities in cases where the potential value to be created is tenuous or theoretical.
- This focus on necessity was seen as essential, in part, because the pandemic has predisposed a large portion of the population to allow additional surveillance and monitoring to support the perceived collective good. This relaxation of anti-surveillance viewpoints potentially opens the door to exploitation.
- Participants also encouraged stakeholders to consider carefully the ways data can be weaponized, including in cases where the intended re-use is meant to create social benefits.

> **"Do we need to know this? That is my first guiding principle."**
>
> — Participant in the Rights Groups and Advocacy Organizations Mini-Public

WHAT

- Participants indicated vulnerable populations tend to be underrepresented in many datasets — including smartphone-derived location data or credit card activity data. This data "invisibility" can limit individuals' ability to see the benefits of data in various contexts, not limited to COVID-related data re-use. Practitioners should take care to ensure that these "data invisibles" are not left behind when targeting service delivery.

- Importantly, being "data visible" can create risks and perceived risks for certain communities—such as undocumented immigrants and current or formerly incarcerated persons. Practitioners should similarly take care to assess the risks of increasing the data visibility of these populations, even in cases where data re-use is intended to benefit them. Additional protections for these datasets will likely be necessary.
- The support for data re-use, governance, and control at the local and hyper-local level also raised questions regarding effective anonymization. Practitioners should also take care to avoid the inadvertent re-identification of data subjects in relatively small samples.

WHO

- Participants expressed the importance of community engagement to define what types of data re-use would be most appropriate and what unique protections are necessary. This engagement should target both community leaders who can represent the interests of various populations and "regular" data subjects.
- The mini-public indicated that community-led data literacy training and education will likely be necessary for community engagement to be meaningful. Participants highlighted trusted intermediaries, such as public libraries, as key actors who could lead this work.
- Participants observed that in many cases, data subjects are more likely to trust community leaders and local government actors as re-users of data. This observation is, in part, because people tend to place more faith in actors who are more proximate to the problems at hand and familiar with the communities they serve in comparison to state or federal actors.
- There was a clear concern regarding the effective and responsible stewardship of data on both the supply and demand side of data re-use. Participants highlighted the need for actors to be tasked with ensuring data is clean, accurate, and handled carefully and ethically across the data lifecycle as well as when any new circumstances arise.

> **"Communities need to be involved in deciding what is mission critical and valuable to them."**
>
> — Participant in the Rights Groups and Advocacy Organizations Mini-Public

HOW

- Several participants reiterated that a one-size-fits-all approach to data re-use will not be workable. Three cross-cutting principles nonetheless came to the fore:
    - **Informed consent,** even for the use of aggregated data. Informed consent necessitates engagement and the ability to opt out prior to data re-use. Data subjects should be clear on which data activities are enabled by their consent.
    - **Transparency** of practices using approachable language. Transparency includes ensuring practitioners' use of language that is clear, accessible, and easily understandable for all. It is closely related to the question of data literacy.
    - **Contextualizing data** that will be re-used. Stakeholders analyzing and re-using data can gloss over important nuances if they do not take steps to understand the context in which the data was generated and came to be re-used. Practitioners should track data provenance to clarify what data actually represents—and what it does not represent—prior to its re-use. Similarly, practitioners should scope the decision points, or capture the "decision provenance" that impacts how authorities collected, processed, shared, analyzed, and re-used data.

WHEN

- Consistent with much of the deliberation, participants indicated that data retention should be determined on a case-by-case and user-by-user basis. Academia, for example, might be trusted to retain aggregated data for a longer period of time to enable their

analyses. Government actors, especially federal government agencies, should relinquish access or destroy data after they have generated the insight or made the decision for which the data was re-used.

- Participants were clear that, for all actors, data should only be held as long as is necessary to achieve the discrete, stated purpose of the data re-use. Future, exploratory analyses should be treated as new instances of data re-use necessitating new consent and subject to the ability to opt out.

> **"Instead of asking how long we need to retain this data, maybe we should ask what is the shortest time we need to do what we need to do?"**
>
> — Participant in the Rights Groups and Advocacy Organizations Mini-Public

WHERE

- As described above, the mini-public advocated for localized, community-oriented data re-use. They supported empowering local actors operating in close proximity to the problems at hand and intended beneficiaries of data re-use to steer these efforts, and respond to any place-based concerns as they may arise.

3. DATA HOLDERS AND POLICYMAKERS MINI-PUBLIC

WHY

- Participants argued equitable benefits should be the central goal when defining the purpose and intended outcome of data re-use. This clarity of purpose has implications throughout the data re-use lifecycle, including considerations of data relevance and quality and the operational dynamics of a data re-use project.

> **"We were able to show that there is a lot of value in making use of usually siloed data."**
>
> — Participant in the Data Holders and Policymakers Mini-Public

WHAT

- Data holders made clear that aggregated data re-use is necessary given the many privacy, data security, and organizational reputation considerations in play. More granular or identifiable forms of analysis and re-use simply raise too many risks.
- Regardless of the type of data, participants highlighted the importance of tracking data provenance to better understand and respond to any biases or limitations. Data provenance should also be communicated to collaborators on the demand side involved in the data re-use so that they can respond to those biases or limitations accordingly.
- Participants also indicated that a collaborative approach will often be necessary, as no single dataset or organization can provide all the answers to key questions. This type of multi-stakeholder approach can increase the validity of data re-use, and mitigate issues or bias and lack of representativeness in the data.

> **"Privacy is the biggest consideration for my team and, I hope, for the private sector. But it is hard to make everyone happy."**
>
> — Participant in the Data Holders and Policymakers Mini-Public

WHO

- Participants described how data stewards can help respond to emergent challenges and place "additional guardrails" on the data analysis. Even if upfront conditions and limitations are put in place to ensure responsibility, no one person or team can comprehensively predict risks or challenges that could surface over time. Participants indicated that human infrastructure, such as dedicated data stewardship roles, should be put in place to respond to new circumstances as they arise.
- Participants called for governments to take a leadership role in data re-use. This includes enabling the re-use of data the government holds where appropriate and potentially impactful; engaging with data holders in other sectors to create public value from that data; helping to increase the data literacy amongst the general public; and democratizing and governing the re-use of data in the public interest.
- For the public sector to take on this leadership role, participants believed that data literacy must be advanced within government, including especially outside of IT or data departments. As data can intersect with all aspects of an organization's operations, leaders and rank-and-file workers need to have some basic understanding of responsible data use to promote good practices.
- Participants called for the creation of new intermediaries explicitly and exclusively tasked with representing the public and their interest. These intermediaries are seen as the "missing player" in most data re-use projects. Such an intermediary could help ensure responsible re-use independent of the pressures public and private actors face—such as competitive balance or political pressures.

- The mini-public also highlighted the need for more legal talent in the data field. Data policies and laws, when they exist for a specific domain, are rarely black and white. Cross-sector data collaboration between private, public, and nonprofit entities can be subject to complex legal provisions that are not always familiar to actors on the supply or demand side. Intermediaries with legal knowledge could support more and more effective data re-use.

HOW

- Consistent with the other two mini-publics, participants called for proactive, upfront communication with data subjects to clarify how their data is being re-used. This includes participatory engagement and consultation at the design or planning stage, the pursuit of informed consent prior to project initiation, and transparency across the data lifecycle through eventual data re-use.
- Demonstrating their knowledge and experience in data re-use, this mini-public was the only one that raised issues of shared metrics, data glossaries, and operational guidelines. While acknowledging data re-use projects will not be "one size fits all," these tools could help raise the bar of data responsibility across contexts, and capture and operationalize lessons learned from other sectors.
- Participants also underlined the need for a stronger legal and regulatory framework for data re-use. They identified licensing agreements as one avenue for clarifying what types of data re-use are or are not appropriate and legal. These agreements are often, however, long and not read by users. New approaches to civil and contract law will be necessary to support responsible and transparent data re-use. Some participants also pushed for better regulation of data re-use so that companies are not forced to self-regulate or operate in an unclear regulatory environment.

WHEN

- Participants called for an end-to-end data responsibility approach, highlighting challenges, risks, and opportunities at the planning, collecting, processing, sharing, analyzing, and re-using stages of the data lifecycle. Such an approach, they argued, involves establishing data stewardship roles and points in time when responsible data efforts should be assessed and refined as necessary.

> **"The caveats and the issues with the original data need to be bundled with data throughout the process. By the time data can get to the re-use phase, that context can get lost."**
>
> — Participant in the Data Holders and Policymakers Mini-Public

WHERE

- Consistent with the other mini-publics, participants highlighted the need for case-by-case oversight and data stewardship, including as it relates to the re-use of geolocation data. New place-based sensitivities and challenges may arise, and fully automated processes that are not subject to human oversight and intervention might fail to effectively mitigate risks.
- Participants also reflected on technological means for automating place-based responsible data practices. Using the example of protests, participants indicated that sensitive areas can be flagged, and data can be destroyed at the moment of collection. The discussion clearly indicated the importance of both human- and technology-driven pathways to handling place-based information in a responsible manner.

“There needs to be some kind of oversight because there are always some sensitivities. There are broad principles you can apply but there are individual nuances”

— Participant in the Data Holders and Policymakers Mini-Public

APPENDIX 1: REMESH QUESTIONNAIRE USED IN NEW YORKERS MINI-PUBLIC

The following questionnaire was used to solicit input on the three data re-use exhibits during the New Yorkers Mini-Public deliberation of the Data Assembly. Please note that this version of the questionnaire does not include demographic questions used to inform segmentation analysis. GovLab facilitators also shared brief video presentations, graphics, and text to help guide participants through the deliberation. Finally, the questionnaire originally included five additional questions that were deemed redundant based on participant inputs and subsequently removed from the deliberation's discussion guide.

Exhibit A: Mobility Data Analysis

1. Are you familiar with the concept of "aggregated data" or "data aggregation"? [Poll]
    a. Yes, I know what aggregated data is
    b. I've heard the term, but not sure what it means
    c. No, I've never heard of aggregated data

2. Do you support or oppose an organization reusing your data in the project described in Exhibit A? [Poll]
    a. Very supportive
    b. Supportive
    c. Unsupportive
    d. Very unsupportive

3. Please explain your response and level of support for the project described in Exhibit A. [Open-ended]

4. **WHY**: The project described in Exhibit A seeks to give academic researchers new ways to study people's movement patterns in relation to government policies and

directives—such as curfews and shelter-in-place orders. Do you find this to be an important or unimportant purpose to re-use mobility data? [Poll]

a. Very important
b. Important
c. Unimportant
d. Very unimportant

5. **WHY**: Do you think that the re-use of aggregated location data by academic researchers is appropriate or inappropriate? Please explain why. [Open-ended]

6. **WHAT**: The project described in Exhibit A involves the reuse of aggregated location data drawn from smartphones. Are you supportive or unsupportive of this type of data being reused in general terms (not just for the project described in Exhibit A). [Poll]
    a. Very supportive
    b. Supportive
    c. Unsupportive
    d. Very unsupportive

7. **WHAT**: Please explain your answer. What makes you supportive or unsupportive of the reuse of aggregated smartphone location in general terms? [Open-ended]

8. **WHO**: The project described in Exhibit A involves a telecommunications company sharing aggregated location data with academic researchers. Would you be more, less, or equally supportive of that aggregated data being shared with a public health agency like the Center for Disease Control (CDC) or New York City Department of Public Health? [Poll]
    a. More supportive of data being reused by public health agencies
    b. Equally supportive of data being reused by academic researcher or public health agencies
    c. Less supportive of data being reused by public health agencies

9. **WHO**: Which actors should be allowed to reuse aggregated location data in the response to COVID-19 (e.g. businesses, academic researchers, nonprofits or charities, government agencies)? Who should make that decision? [Open-ended]

10. **HOW**: The project described in Exhibit A involves a private-sector telecommunications company working with various university researchers. Would any of the below actions increase your support for the project? Select as many options as relevant. [Poll]
    a. Creating an independent council of experts to oversee and provide ethical guidance on the project
    b. Making legal agreements between the company and participating researchers publicly accessible
    c. Creating a representative panel of regular New Yorkers to help decide what types of data analysis or research would be appropriate and valuable
    d. Designating internal authorities responsible for ensuring data reuse is effective and appropriate
    e. Oversight by a designated business association
    f. Management of your data by an trusted, independent third party
    g. Providing you with the ability to opt out of any uses of your aggregated data
    h. None of these

11. **HOW**: What other steps could be taken to increase your trust in such mobility data analysis? For instance, experts could regularly report to the public about the project or notify you when your data is aggregated for use in a research study. [Open-ended]

12. **WHEN**: The project described in Exhibit A involves short-term access to aggregated data. In other words, the researchers will lose access to the aggregated data once the analysis has completed. Are you more or less supportive of the effort knowing that the data will not be held indefinitely? [Poll]
    a. Yes, I am supportive of short-term data retention
    b. No, short-term data retention does not increase my support for the project

c. Short-term data retention has no impact on my views of the project

13. **WHEN**: Short-term data retention can limit the ability to conduct new analysis as time goes on. What are your thoughts on that trade off? [Open-ended]

14. **WHERE**: The project described in Exhibit A involves anonymous location data aggregated by city block, a common level of aggregation for location data analysis. Do you have a problem with data being categorized according to city blocks? [Poll]
    a. No, as long as the data is not personally identifiable
    b. Yes, aggregation at the city block level is too risky or invasive
    c. Not sure

15. **WHERE**: Are there certain places that you feel should be excluded from this type of analysis (e.g. church going, protests, sporting events)? Please describe any place-based considerations that you feel project organizers should take into account when doing this type of location analysis. [Open-ended]

Interstitial Questions:

1. Has anyone in your immediate family lost their source of income as a result of the pandemic? [Poll]
    a. Yes
    b. No
    c. Prefer not to say

2. Who should lead the response to COVID-19 moving forward? You are free to select multiple answers. [Poll]
    a. National Government
    b. State Government
    c. Local Government
    d. Public Health Experts
    e. Private Sector

f. Nonprofits and Charities
g. Community Organizations

Exhibit B: Consumer Data Analysis

1. Are you familiar with the concept of "safe sandboxes" and "secure data processing environments"? [Poll]
    a. Yes, I know what a safe sandbox or secure data processing environment is
    b. I've heard the term, but not sure what it means
    c. No, I've never heard of safe sandboxes or secure data processing environments

2. How supportive or unsupportive would you be of your credit card data being reused in the project described in Exhibit B? [Poll]
    a. Very supportive
    b. Supportive
    c. Unsupportive
    d. Very unsupportive

3. Please describe your level of support or opposition to an organization reusing your data in the project described in Exhibit B. Why did you answer the previous question as you did? [Open-ended]

4. **WHY**: The project described in Exhibit B is intended to give government economic development agencies the ability to understand consumer spending patterns to support the city's economic recovery from the pandemic and to target recovery investment to neighborhoods or industries in greatest need of support. Do you find this to be an important or unimportant objective? [Poll]
    a. Very important
    b. Important
    c. Unimportant
    d. Very unimportant

5. **WHY**: Do you think that providing insight on consumer spending patterns to economic development agencies is an appropriate or inappropriate purpose for reusing aggregated credit card data? Please share your thoughts on the value and importance of this objective. [Open-ended]

6. **WHO**: The project described in Exhibit B involves a credit card company enabling a government economic development agency to analyze consumer spending data. Would you be more or less supportive of that aggregated consumer data being shared with a non-governmental actor, like a nonprofit, charity, or academic research center? [Poll]
    a. More supportive of data being reused by a non-governmental organization
    b. Equally supportive of data being reused by a government agency or non-governmental organization
    c. Less supportive of data being reused by non-governmental organization

7. **WHO**: If this project did go forward, which actors should be allowed to reuse aggregated consumer spending data in the response to COVID-19 (e.g. businesses, academic researchers, nonprofits or charities, government agencies)? Why do you feel that way? [Open-ended]

8. **HOW**: The project described in Exhibit B involves an economic development agency analyzing credit card data in a "safe sandbox" where no aggregated or personally identifiable data can be extracted. Are you supportive or unsupportive of using technical approaches like safe sandboxes to enable government agencies to analyze potentially sensitive data in a controlled environment? [Poll]
    a. Yes, government agencies should be allowed to analyze sensitive and potentially valuable information in a controlled environment
    b. No, credit card data is too sensitive or risky, even in a controlled environment
    c. Not sure

9. **HOW**: Safe sandboxes and similar privacy-preserving strategies are highly technical, built on technologies and concepts that are not familiar to most people who do not work in the field of data science. What information or assurances would you need to feel confident in these types of privacy-preserving technologies? [Open-ended]

10. **WHEN**: The project described in Exhibit B involves indefinite data retention, with the potential for the government economic development agency to conduct multiple analyses over time. Would you be more supportive of this type of data reuse if you were alerted every time your data was used in this way? [Poll]
    a. Yes, I would be more supportive if I were alerted to new uses of my data
    b. No, I don't need to be alerted when my data is aggregated and analyzed in a secure environment
    c. Yes, but only if I had the option to opt out each time I was alerted my data was used
    d. Not sure

11. **CROSS-CUTTING:** What changes could be made to the project described in Exhibit B that would make you more convinced of its appropriateness and/or more confident in its value? [Open-ended]

Interstitial Questions

1. Have you or an immediate family member been diagnosed with COVID-19? [Poll]
    a. Yes
    b. No
    c. Prefer not to say

2. How long do you think it will take for life to return to "normal"? [Open-ended]

Exhibit C: 311 Data

1. Are you familiar with New York City's 311 system? [Poll]
    a. Yes, I know what 311 is

b. I've heard of 311, but not sure what it is
c. No, I've never heard of 311

2. Are you familiar with the concept of "open data" ? [Poll]
    a. Yes, I know what open data is
    b. I've heard of open data, but not sure what it is
    c. No, I've never heard of open data

3. How supportive or unsupportive would you be of your data being reused in the project described in Exhibit C? [Poll]
    a. Very supportive
    b. Supportive
    c. Unsupportive
    d. Very unsupportive

4. Please describe your level of support or opposition to an organization reusing your data in the project described in Exhibit C. Why did you answer the previous question as you did? [Open-ended]

5. **WHY**: The project described in Exhibit C involves a law enforcement agency's use of open data to determine where people are failing to follow social distancing and masking guidelines. Do you find this to be a valuable and important objective? [Poll]
    a. Very important
    b. Important
    c. Unimportant
    d. Very unimportant

6. **WHY**: Do you think that targeted law enforcement is an appropriate or inappropriate use of data collected through the 311 citizen reporting system? Please share your thoughts on the value and importance of this objective. [Open-ended]

7. **WHAT**: The project described in Exhibit C involves the reuse of citizen complaints to the city's 311 reporting system. In general terms, are you supportive or unsupportive of this type of data being reused in the public interest? [Poll]
    a. Very supportive
    b. Supportive
    c. Unsupportive
    d. Very unsupportive

8. **WHAT**: Please tell us more about your views on the reuse of citizen reporting data from the 311 system. What makes you supportive or unsupportive of reusing this type of data? [Open-ended]

9. **WHO**: The project described in Exhibit C involves citizen reporting data from the 311 system being shared with law enforcement. Do you consider law enforcement actors to be appropriate reusers of non-emergency citizen reporting data? [Poll]
    a. Yes, I think law enforcement's reuse of 311 data would be appropriate
    b. No, I do not think that law enforcement should reuse 311 citizen reporting data
    c. Not sure

10. **WHO**: Do you feel that certain restrictions should be placed on law enforcement's reuse of citizen reporting data from 311 related to COVID-19? Why or why not? [Open-ended]

11. **HOW**: Given that 311 is a service funded by taxpayers' money do you agree that the data collected should be made accessible to the public as open data? Why or why not? [Open-ended]

12. **WHEN**: As described in Exhibit A, shorter retention periods could limit future reuse of data. What are your feelings on the balance between holding COVID-related citizen reporting data for shorter periods of time versus retaining it indefinitely, potentially

enabling future analysis of, for example, how effective and equitable law enforcement's response to the pandemic was. [Open-ended]

13. **CROSS-CUTTING:** What changes could be made to the project described in Exhibit C that would make you more convinced of its appropriateness and/or more confident in its value? [Open-ended]

APPENDIX 2: FURTHER READING ON PUBLIC PERCEPTIONS AND THE RESPONSIBLE RE-USE OF DATA FOR COVID-19

Ada Lovelace Institute. No green lights, no red lines: Public perspectives on COVID-19 technologies. July 2020.

A report providing recommendations to policymakers informed by lessons learned from three public deliberations: The Citizens Biometrics Council, Community Voices Workshops, and online deliberation on technology use in the UK response to COVID-19.

Brooke Auxier. How Americans see digital privacy issues amongst the COVID-19 outbreak. Pew Research Center FactTank. May 4, 2020.

A collection of survey findings regarding Americans' views on privacy issues in the context of the pandemic, with a particular focus on location data collected through smartphones.

Douglas J. Elliott, Ana Kreacic, Lorenzo Miláns del Bosch, Lisa Quest. Data Sharing in the Time of Coronavirus. Oliver Wyman Forum. April 1, 2020.

Findings from a six-nation survey conducted by the Oliver Wyman Forum to gain insight into public perceptions on the sharing and re-use of personal health data in the context of the COVID-19 pandemic.

Frederic Gerdon, Helen Nissenbaum, Ruben L. Bach, Frauke Kreuter. Individual Acceptance of Using Health Data for Private and Public Benefit: Changes During the COVID-19 Pandemic. Harvard Data Science Review. September 3, 2020.

A paper presenting findings from a survey of over 3500 participants regarding their views on the use of data for COVID-19. The authors consider these findings in relation to a similar study conducted in 2019, prior to the COVID-19 pandemic, regarding public perceptions on health data sharing and privacy more generally.

Saira Ghafur, Jackie Van Dael, Melanie Leis, Ara Darzi, Aziz Sheikh. Public Perceptions on data sharing: key insights from the US and USA. The Lancet, Digital Health. July 24, 2020.

A paper sharing results from a survey conducted in the United States and United Kingdom to capture public perceptions regarding "data sharing, data access, and the use of AI in health care."

Helen Kennedy, Susan Oman, Mark Taylor, Jo Bates, Robin Steedman. Public understanding and perceptions of data practices: a review of existing research. Living with Data. May 2020.

A report synthesizing existing research (pre-dating the COVID-19 pandemic) into public understanding and perceptions of data handling practices. The findings will inform a future work supported by the Nuffield foundation to better reflect public perceptions in data initiatives, including those related to addressing COVID-19.

https://thedataassembly.org/