Collaborative Tools and Strategies for Data-driven Development Engineering

Jordan Freitas
Eric Brewer

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-18
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-18.html

May 1, 2021

Acknowledgement

I have had the honor and privilege of being advised by Professor Eric Brewer, who has been unfailingly wise, kind, practical, and helpful in guiding my work and career. I am grateful for the opportunity to have worked with and learned from him, and the Technology and Infrastructure for Emerging Regions (TIER) research group. This work is supported by the Development Impact Lab (US- AID Cooperative Agreement AID-OAA-A-13-00002), part of the USAID Higher Education Solutions Network; the Commission of Higher Education (CHED) of the Republic of the Philippines, through the Village Base Station (VBTS) project of the Philippine-California Advanced Research Institutes (PCARI); a SanDisk Fellowship; and the National Science Foundation through the CyberSEES program (Award 1539585).

Collaborative Tools and Strategies for Data-driven Development Engineering

by
Jordan Freitas

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division
of the
University of California, Berkeley.

Committee in charge:
Professor Eric Brewer, Chair
Professor Xiaodong Song
Assistant Professor Joshua Blumenstock

Spring 2019

Collaborative Tools and Strategies for Data-driven Development Engineering

Abstract

Collaborative Tools and Strategies for Data-driven Development Engineering

by

Jordan Freitas
Doctor of Philosophy in Computer Science
University of California, Berkeley

Professor Eric Brewer, Chair

Open data requirements and concern for privacy in data-driven international development projects are increasingly prevalent. Current practices typically attempt to balance the two by manually removing personally identifying information and publishing a view of the remaining data. Both practically and theoretically this approach fails to satisfy the open data objective of reusability, and fails to protect privacy of individuals in the data. This thesis explores how to improve both the utility of shared data and how well privacy is maintained with strategically designed tools and methods. We propose and evaluate these tools and strategies for collaborative data management to help navigate tensions between open data and data privacy in the context of international development engineering projects. We first share the results of interviews with individuals who work closely with data in one subfield of development engineering and analyze the results in terms of implications for building data management and data sharing tools. From there, we propose design requirements for workflow sharing tools based on four motivating use cases in different areas of development engineering and present our implementation of a tool to satisfy these requirements. We then provide an overview of privacy considerations and our improvement mechanisms. Both our workflow sharing tool and privacy strategies enable more fine-grained control over data and code sharing with an emphasis on usability. Finally, we situate this work politically and socially in the context of international development.

Contents

Chapter 1: Introduction **1**
1.1 Motivation . . . 1
1.2 Perspectives on Data Sharing . . . 2
1.3 Design Requirements for Data Sharing . . . 2
1.4 Pipelines and Usability Features . . . 2
1.5 Privacy Considerations and Strategies . . . 3
1.6 Participation and Development Context . . . 3
1.7 Contributions . . . 3

Chapter 2: Perceptions of the Value and Risk of Sharing Code and Data **5**
2.1 Introduction . . . 5
2.2 Background . . . 5
2.3 Interview Methods . . . 8
2.4 Interview Results . . . 11
2.5 Do Data Sharing Policies Address Trade-offs? . . . 17
2.6 Implications for Data and Code Management Tools . . . 20

Chapter 3: Design Requirements **23**
3.1 Introduction . . . 23
3.2 Motivation . . . 23
3.3 Examples . . . 24
3.4 Design Requirements . . . 27
3.5 Conclusion . . . 30

Chapter 4: Pipelines and Usability for Sharing Data Workflows **31**
4.1 Introduction . . . 31
4.2 Related Work . . . 32
4.3 Design . . . 35
4.4 Features . . . 40
4.5 Validation . . . 46
4.6 Conclusion . . . 49

Chapter 5: Privacy Considerations and Strategies **51**
5.1 Introduction . . . 51
5.2 Related Work . . . 52
5.3 Improvement Mechanisms . . . 56
5.4 Example Implementation and User Feedback . . . 59
5.5 Conclusions . . . 63

Chapter 6: Participation and Development Context **65**

6.1 Introduction 65
6.2 Motivation 65
6.3 Methods 66
6.4 Challenges 67
6.5 Discussion 74
6.6 Conclusion 77

Chapter 7: Summary and future work **79**
7.1 Summary of Findings 79
7.2 Future Work 80

References **80**

Appendix A: Interview Guide **87**

Acknowledgments

First and foremost, I have had the honor and privilege of being advised by Professor Eric Brewer, who has been unfailingly wise, kind, practical, and helpful in guiding my work and career. I am grateful for the opportunity to have worked with and learned from him, and the Technology and Infrastructure for Emerging Regions (TIER) research group.

I am also thankful for advising from Dr. Khalid Kadir, who helped me explore and think critically about important questions related to this work, and in doing so helped me grow as a scholar, teacher, and person.

Throughout my graduate career, Dr. William Bosl at University of San Francisco has provided an additional source of inspiration, guidance, and opportunities for which I am thankful. Dr. Andrew Nguyen at University of San Francisco has also been a generous and supportive colleague.

My research would not have been possible or enjoyable without my collaborators: Matthew Podolsky, Javier Rosa, Andreas Kipf, Isa Ferrall, Jonathan Lee, and the support and guidance of Professor Duncan Callaway. Orianna DeMasi has also provided extensive helpful feedback and encouragement.

I would like to acknowledge my Creighton University professors for the training that led me here and for convincing me I was capable, especially: Dr. Mark Wierman, Dr. David Reed, Dr. Carol Zuegner, Dr. Michael Cherney, and Dr. Jeffrey Hause.

Finally and especially, thank you to my husband Rodrigo Freitas for years of being supportive, reliable, and gracious, and to the family who raised me: Katherine Kellerstrass, Amy Kellerstrass, Albert Kellerstrass, Tracy Daily Moore, and Leslie Harder.

The work presented in this dissertation was made possible by the financial support provided by different sources including the Development Impact Lab (US- AID Cooperative Agreement AID-OAA-A-13-00002), part of the USAID Higher Education Solutions Network; the Commission of Higher Education (CHED) of the Republic of the Philippines, through the Village Base Station (VBTS) project of the Philippine-California Advanced Research Institutes (PCARI); a SanDisk Fellowship; and the National Science Foundation through the CyberSEES program (Award 1539585) for which I am also grateful.

Chapter 1:

Introduction

1.1 Motivation

The fundamental tension between data privacy and openness in research deserves careful consideration. Compulsory open data policies are increasingly popular among journals and funding agencies as a way of improving transparency and reproducibility as well as increasing potential impact by allowing other researchers to validate and expand on original work. At the same time, when the underlying research data is collected from human subjects, participants have a right to anonymity and privacy that cannot be guaranteed with de-identification alone. Data re-identification can be accomplished with automated analyses, so manually removing identifying attributes–a common practice–is both ineffective for privacy protection and forfeits key contextual information in the process, which is arguably a bad approach to science. Added to this is a complex landscape of stakeholders who have varying intentions with the data, and would ideally have access to varying levels of detail. Thoughtfully designed data management systems with dynamic privacy strategies will do a better job of both protecting privacy and sharing information.

The topic of this dissertation is collaborative data management tools as a means of more accurate exploration of social issues. The more freely different stakeholders of different kinds of data can question, validate, or expand on results, the more thorough our collective understanding will be. The more equitable voice different perspectives have in processes of collecting, analyzing, and presenting data, the more confidence we can have in results' validity and applicability. Collaborative data management research is a response to the widespread calls for open data and transparency along with corresponding uncertainty around privacy best practices and a growing acknowledgement of digital data rights such as the EU General Data Protection Regulation (GDPR). We focus on contexts of international development and social impact projects with minimal resources to allocate for proper data management and sharing. My research considers participation of and tensions among different stakeholders of the data, usability of existing infrastructure and interfaces, privacy strategies for protecting human subjects, the uses and impact of data analysis work, and aims to situate all of these factors in the social political contexts in which they live. There is a movement towards digital ownership of individuals' own data, and the right to opt in and out of clearly stated purposes for analyzing personal data. All of this depends on access to new tools for projects with limited resources.

1.2 Perspectives on Data Sharing

Chapter 2 documents relationships to data sharing among practitioners working in energy access along with perceived positive and negative consequences. The participants in the interviews are representative of the data professionals for whom we designed improved data sharing mechanisms discussed in the following chapters. We share what we learned from these stakeholders about their work and their understanding of issues that come up in their work around data collection, use, and sharing. A summary of interview results is preceded by an overview of similar studies on data sharing, background on the open data policies our participants were most likely to have encountered, how our participants fit into the landscape of data stakeholders for our contexts of interest, and the methods we used to conduct interviews and process transcript data.

Other data sharing studies have investigated many aspects of data sharing in a variety of fields, including best practices, compliance rates or effectiveness of official policies of journals and funding agencies, general motivations and disincentives, and even personality traits of researchers who are more or less likely to share data. Another perspective has been to focus on discoverability and usability of data once it is made public. This investigation is unique in its focus on interviewing experts in the area of energy access and planning, and in its purpose to inform the development of data sharing tools more so than practices.

Interview results are categorized into the following themes: reasons to share or not share data, how data sharing is accomplished or impeded, data collection and management issues, collaboration, and privacy. Each of these themes offer insight and implications for developing code and data management tools.

1.3 Design Requirements for Data Sharing

Chapter 3 articulates key design requirements for collaborative data sharing. We study four development engineering projects to consider how data is collected, stored, analyzed, and shared either as a dataset or as results. Design requirements are based on what these projects would have benefited from in the past or what they need to enable work going forward. The requirements include definitive specifications and existing semantics from the field of database systems, along with less precise values to be upheld. The motivating use case projects are summarized and then each design requirement is defined and justified based on aspects of the respective projects. The primary scenarios are data cleaning, monitoring and evaluation, and algorithm development. Each of these scenarios is present in more than one of the projects with varying centrality.

1.4 Pipelines and Usability Features

Chapter 4 elaborates on the features we built as one approach to meeting the design requirements specified in the previous chapter. We provide an overview of related projects in the space of block programming, other user interfaces for data analysis and collaboration, and other platforms for data sharing. From there we describe seven primary features of a tool we built for researchers to create data analysis pipelines that optimizes for usability

and collaboration requirements. The pipeline builder involves a drag and drop interface for assembling an analysis graph of operators from an operator library, where each operator has a designated input and output type and performs one task on the data. The library offers easy to use and reuse blocks of code, and also flexibility in the ability to add custom operators. A pipeline graph is used to generate Google Cloud Dataflow (now Apache Beam) code, which can run locally or in the cloud and offers managed parallelism and other benefits we explain further.

We revisit the use cases from which design requirements were derived in order to evaluate my implementation of this pipeline-based code generation tool for collaboration. The use cases reveal unique and partially overlapping sets of values and scenarios. This tool is particularly useful for data cleaning, monitoring and evaluation, and algorithm development.

1.5 Privacy Considerations and Strategies

This chapter begins by mapping out existing privacy strategies framed by trade-offs and then, given the contexts and applications of interest to the overall thesis, describes a combination of improvement mechanisms. Related work includes the alternative strategies along with implementations of privacy preserving systems. We then describe how we incorporate the improvement mechanisms for an application of interest that requires considering the privacy of multiple stakeholder groups. This work is validated by user feedback, which is laid out in a section on the privacy expectations of users. Through surveys and interviews, participants in the application provide some unanticipated considerations that invite reflection on common assumptions.

1.6 Participation and Development Context

The final chapter provides an overview of theories of development particularly relevant to technology and development and then applies this work to data management technology in particular. Data-driven decision making along with open data initiatives are popular among development agencies. We look at power dynamics in decision making structures as well as the role of consumers or intended beneficiaries in shaping technological deployments. Up to this point in the dissertation, when we refer to data sharing and collaboration, it is implied to be among peers and researchers. This chapter considers participation among different stakeholders, whose stakes in outcomes are different, whose values and freedoms are different, and whose decision-making powers are not equal. As such, this chapter explores the concept of equitable participation as it relates to data management tools in international development, what challenges and barriers are at play, and potential solutions offered as a discussion on existing literature applied to the scope of this dissertation work.

1.7 Contributions

The first chapter, *Perspectives on Data Sharing*, addresses a gap in the literature on energy access regarding what practitioners believe and experience by summarizing and ana-

lyzing interviews with 13 practitioners who have authority on data sharing practices within their work in the energy sector. The interviewees represent work in 4 countries, 3 of which are low and middle income countries (LMIC). Understanding the themes that emerge from the perspectives in this unique dataset enable and validate approaches to building tools to help facilitate benefits and mitigate risks of sharing data. Lessons from the perspective of these stakeholders provide insights that may be more broadly applicable to other fields and inform future work on incorporating the perspectives of other stakeholder groups in the design and implementation of data management tools. We also articulate the implications of the interview results on the work of developing features of data sharing and collaboration tools, such as those described in the chapters addressing usability and privacy issues.

The second chapter, *Design Requirements for Collaborative Data Management*, contributes novel design requirements for collaborative data management and analysis tools, especially for projects operating in contexts lacking technical and financial resources to invest in what's typically referred to as research and development (R&D) or monitoring and evaluation (M&E). We demonstrate how these design requirements are based on multiple use cases and how the achievement of such requirements would improve their work.

The contributions of chapter three, *Pipelines and Usability Features*, are the design and implementation of a tool we developed for sharing code and workflows, which is one possible realization of the design requirements described in the preceding chapter. We evaluate the tool based on the same case studies used to develop the design requirements.

Chapter four, *Privacy Considerations and Strategies*, explores the privacy considerations and strategies for data sharing. We include an overview of existing privacy strategies and their relevance to collaborative research and data sharing as a means of satisfying the open data policies of journals or funding agencies. We articulate who the consumers or stakeholders are of private data and develop a role-based access control approach to data sharing. The roles reflect those stakeholders and associated permissions are based on their intentions. The stakeholder roles we include are those who typically interact with or would like to interact with the data generated from development engineering research projects for several different reasons. We describe the permissions needed to satisfy those stakeholder intentions on a need-to-know basis. One of the permissions we describe is a novel combination of concepts from other strategies including uniqueness and privacy budgets, intended to achieve compromise between differential privacy and de-identification. Dynamic aggregation as a permission along with strong and enforceable data use agreements would be more fitting for collaboration among researchers than other existing strategies in a few important scenarios. We use this work to develop a privacy strategy for an interesting use case. These contributions will be the building blocks of an extensible framework for role-based access control with fine-grained control over permissions such as access to differentially private or dynamically aggregated query results.

Finally chapter five, *Participation and Development Contexts*, situates data management tools in some of the literature and history of technological international development interventions. We revisit the interview transcripts to specifically consider mentions of the power dynamics among stakeholders, and then bring this insight into conversation with theories of development. This work is an effort towards devising design principles for equitable, collaborative data management systems.

Chapter 2:

Perceptions of the Value and Risk of Sharing Code and Data

2.1 Introduction

We explore perceptions of value and risk related to sharing data and the implications of those perceptions for developing tools that enhance existing value and mitigate potential risk. A review of related data sharing studies reveals a variety of methods to better understanding several factors related to researcher behavior and motivations. Our goals are to understand the relationships among different stakeholders with different objectives, how data sharing can play out among them, and how to better manage the inherent tension between openness and privacy. Our primary contributions to this topic are the results of interviews we conducted with thirteen energy data practitioners and discussion of the implications for data sharing practices, tools, and policies. We also review several relevant data sharing policies and guidelines of funding agencies, and look to them for insight into funding agencies as a powerful stakeholder as well as to what extent these existing policies and some of their predecessors address data sharing trade-offs.

We interviewed 13 people, and processed 448 minutes of audio recordings and 395 response paragraphs. The next section provides a background on related studies on data sharing as well as an overview of who we consider to be stakeholders of the data in question in our interview conversations. We then describe in detail our interview methods, followed by a summary of results. An overview of open data policies is provided for insight into another stakeholder group: funding agencies. We conclude with our interpretations of the implications of our interview results on developing data management and sharing tools.

2.2 Background

2.2.1 Related studies on data sharing

Although reviewing privacy risks and strategies on a case by case basis is a widely accepted best practice, there is a push for open, public, raw data, and for balancing risks with benefits. This is in contrast with the privacy-first, rights-based approach preferred in the security and privacy community. Privacy-first implies systems are set up to ensure data privacy and protection first, and then data sharing and collaboration can be enhanced from that starting point. A lesson echoed several times throughout the related literature on data sharing practices and beliefs as well as our interview results is that privacy and data protection, and utility of shared data must be considered at the start of projects in order to

succeed.

Common methods of studying data sharing are outlined in a PLoS ONE article, *Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data* by Heather A. Piwowar[1] and include using surveys and interviews to analyze self-reported behaviors, attitudes, and opinions about data sharing requirements and incentives, which are most inline with the methods proposed for our own study. Other studies have focused on compliance rates by counting open data sets corresponding to a set of publications with the same data sharing expectations from a shared funding agency or journal. Piwowar collected studies that had generated biological gene expression microarray intensity data from Google Scholar along with 124 attributes of these studies to identify which attributes were correlated with whether or not the study had an associated open dataset[1].

Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide by Tenopir et al. is based on surveys conducted 3-4 years apart between 2009 and 2014 including participants from a variety of countries, age groups, cultures, and disciplines [2]. They discovered over this time period and increase in acceptance, willingness, and actual data sharing behaviors as well as an increased perception of risk. They note constraints and enablers to sharing data vary by discipline, which we find to be a good argument for having studies of both perceptions across research areas as well as deeper analyses within specific fields. They write, "Implications of these findings include the continued need to build infrastructure that promotes data sharing while recognizing the needs of different research communities" [2]. Their study, like ours, brings up questions of the value of shared data. A particularly interesting finding from this study was less willingness to share data when human subject research was involved, however what the actual perceived risks were did not vary much across disciplines whether human subjects were involved or not. Also interesting are the variation they illustrate across regions and cultures with more or less focus on data usage or asking permission, which brings up one of multiple issues around how to design tools for sharing data internationally.

Data sharing, small science and institutional repositories by Cragin et al. analyzes data collected through interviews and surveys of small-science researchers who also have an interest in data management or sharing [3]. The authors explore what counts as worthwhile and shareable data to their interview participants, private vs public data sharing practices, and avoiding data misuse. Data misuse is a notable deterrent from sharing data broadly and includes such incidences as publishing without permission, co-author and attribution issues, and unintentional misinterpretation or industry selecting among all the public data that which suits their commercial interests in a clearly unscientific fashion. They focus on small science for the sake of informing development on institutional repositories, as opposed to focusing on large science (e.g. physics or astronomy) fields where they note are already "served by disciplinary or nationally scoped infrastructure initiatives" [3]. Our study also serves the purpose of informing data management and sharing infrastructure, although we focus on only one field (electricity and energy access) which has the added complexity of human subjects.

Stakeholders' views on data sharing in multicenter studies by Mazor et al. focuses on stakeholders of health data sharing in multicenter studies, i.e. multiple healthcare providers agree to pool data in order to study broader populations [4]. The authors conducted 11

interviews with patients, researchers, IRB and regulatory staff, and multicenter research governance experts. Interestingly, the patient participants seemed more interested in their data being used to help develop medicine than concerned about privacy. We presume interesting and unexpected perceptions would also come up if we had conducted interviews with energy consumers and those who responded to surveys conducted by our own interview participants, who represented a more narrow selection of stakeholders.

An open mind on open data by Virginia Gewin as implied by the title is an optimistic exploration of data sharing among scientists, and pays special attention to how scientists may make themselves vulnerable to getting scooped and damaging their reputations in the process of making their data available [5]. Several of our own interviewees are academics and researchers, however perceptions about data sharing may not be broadly relevant to all studies involving human subjects. For example, data collected about some population's opinions in a psychology study has different likelihood of re-identification and associated risks than socioeconomic and health factors. Gewin includes a section on "Open-data pro tips" with both technical and organizational recommendations, however none of the recommendations relate to privacy or consent issues. This leads us to believe the article is aimed at physical sciences not involving human subjects despite a few examples from the field of psychology. An interesting finding from this article is the fact that whether requests for data were granted was influenced by the requester's seniority in the field.

Our study of data sharing perceptions and practices is limited to the area of research in energy access including electrification and grid management, and we focus on stakeholders who work with data on a regular basis although not necessarily concerned with data sharing. Not limiting our participants to those with an interest in data sharing shed light on many cases for which data sharing may not make sense. Our study and others [4, 6, 2], share a theme of investigating perceived benefits and risks of data sharing practices. We have an eye towards developing appropriate and helpful infrastructure while other studies take an interest in policy compliance or simply understanding what the attitudes are and why.

2.2.2 Stakeholders

Prior to conducting interviews, we established a set of user roles based on our own professional observations of who has interests in, and classifiable intentions towards, the kind of data typically collected in development projects. Namely this data comes from surveys about socioeconomic conditions, sensor monitoring [7], and project-level data. These user roles are described in Chapter 5: Privacy Considerations. It is worth mentioning here on whom they are based. See Fiture 2.1.

A **project owner** has complete access to their own raw data and responsibility to protect the privacy of participants. Without added incentives or requirements, a project owner has little to gain from taking the time to figure out how to appropriately share their data with a wider audience. The **colleagues** of a project owner are likely familiar with research goals, and have relevant experience and skills to interpret the data. They would be easily accessible and even included on the project's Institutional Review Board proposal, thus familiar with proper data management processes. Yet we think it's worth asking, *Do they need full access to raw data about participants? Is there any risk in sharing raw data with trusted colleagues?*

We explore technical alternatives to sharing entire and raw datasets among colleagues in Chapter 5.

Some form of the data will probably be made available to **funding agencies** who are keen to assess the impact of their sponsorship, and whether the project is fruitful. Funding agencies intend to calculate progress made towards institutional goals and maximize the impact of investment dollars.

We differentiate between **other researchers** and **potential collaborators** based on likelihood of direct involvement with project owner: while both are outside of the project owner's organization, other researchers will consume and want to compare results and potential collaborators may additionally be in a position to contribute insight, skills, or new context to the project.

Participants have an interest in their own data. They might gain personal insight by comparing themselves with overall results and they will likely be concerned about the consequences of their private information becoming visible to people with whom they did not intend to share it. Multiple interview participants of ours, along with [4], suggest people are also inclined to participate in research as a contribution to the advancement of that research.

An additional and important player in the consideration of data sharing issues is the **Institutional Review Board (IRB)**, tasked with making sure human subjects are protected in the research process. Researchers conducting a study involving human subjects must complete an ethics training and apply for IRB approval prior to conducting the study. IRB applications involve data protection protocols to minimize the chance of data leaks, as well as the potential for harm if data is leaked.

2.3 Interview Methods

2.3.1 Why Interview?

The Related studies on data sharing section describes related studies of data sharing along with their methods. It can be interesting and valuable to observe behaviors such as open data policy compliance rates, especially along with, for example, what characteristics of researchers are correlated with those behaviors [1]. In studying human phenomena for the sake of designing for humans, it is less relevant to know about behavior patterns than it is to understand underlying points of view. Observed behavior could be bonafide best practice, or a work-around. Behavior could be ingrained and inflexible or it could be easily influenced by changing circumstances, such as development of new technology. Surveys, observation of compliance, and literature review are valuable to learn about what people do. Interviewing is necessary to explore the meaning people make of their experiences [8], i.e. the significance of these behavioral trends. In turn, understanding the perceptions and motivations underlying specific behaviors is key to developing relevant and appropriate tools. Seidman points out, "[interviewing] is a powerful way to gain insight on social issues." Decisions about whether, to what extent, and with whom to share data are as much social as they are technical, if not more so. We dive into the social and political contexts of sharing data and enabling technologies in Chapter 6.

A survey could be made about how people perceive value and risk, however an interview

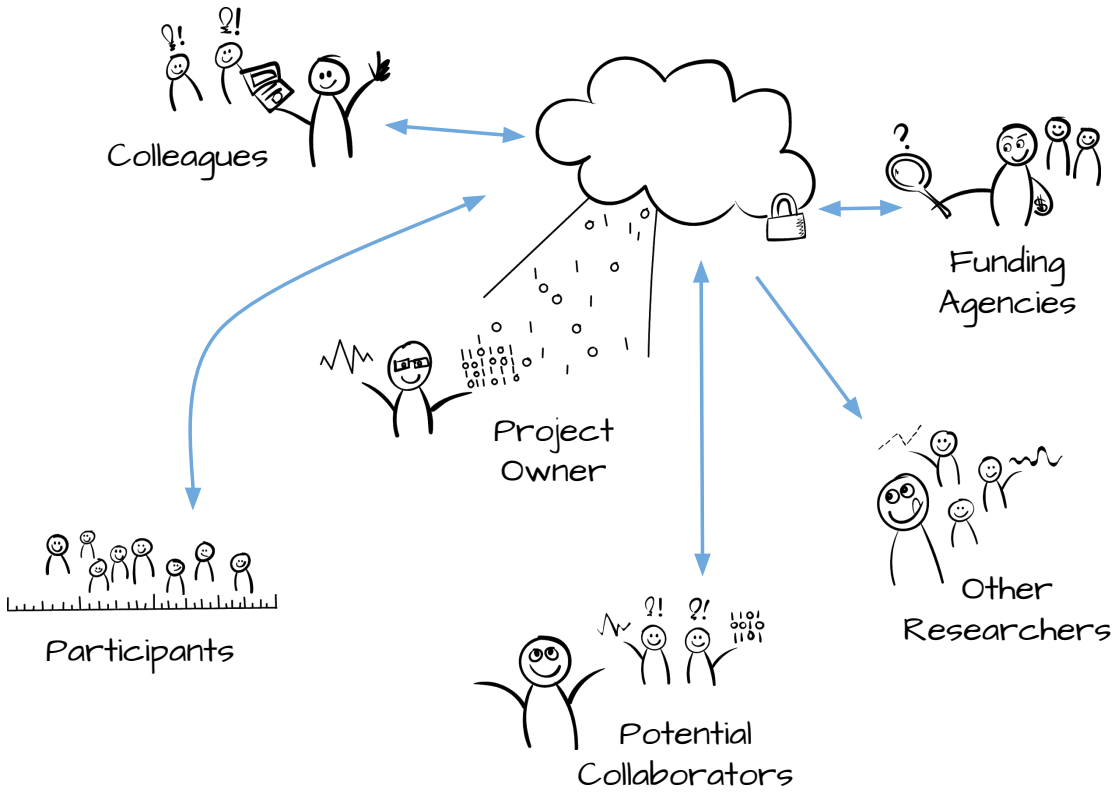

Figure 2.1: An illustration of the landscape of stakeholders.

allows the researcher to tune in to subtleties and probe for more when a participant merely hints at potentially key issues. A survey is also more likely to capture a public voice, the way a participant would communicate the information to the public. An interviewer can reframe questions in-the-moment to ask for more personal opinions, the thoughts most influencing their behaviors.

Based on existing data sharing studies, we notice a disconnect between affirmed value of sharing data, collaborating, comparing and expanding on results, and researchers' data sharing behaviors [9] although this has been moving in the direction of more sharing [2]. We conducted interviews in an effort to better understand this disconnect, as well as develop informed tools to reinforce the value and mitigate risks of open data.

2.3.2 Developing the Interview Guide

The prepared interview questions were established around three goals: 1) better understand the context of relevant data management processes, 2) learn about practitioners' perceptions of the value and consequences of data sharing practices, and 3) collect technical details, along with advantages and limitations, of how practitioners currently work with data. Given our interest in positive or negative consequences, we avoided leading questions. We did not follow the interview guide strictly if participants were more eager to discuss a subset of the questions. Our interview guide is attached as Appendix [...].

2.3.3 Recruiting Participants

Interviewees who work with energy data were identified through the professional networks of the Energy and Resources Group (ERG) and Technology and Infrastructure for Emerging Regions (TIER) research groups at UC Berkeley, as well as a one-day workshop in Nairobi, Kenya for energy practitioners in the region, sponsored by a grant led by PIs from both research groups. Participation was on a volunteer-basis and no compensation was provided.

2.3.4 Conducting the Interviews

Use of the interview guide questions was flexible, although the topics were consistent. Often a response to one question would answer another question, and guiding a participant to follow through with a train of thought while staying on topic provided more insight than a strict, brief question-answer format. Each participant was interviewed one time and interviews took between 30 and 45 minutes, depending on the availability of the participant. In some cases, the interviewer later followed up with participants to clarify or ask for additional information.

2.3.5 Limitations

The most notable limitation of our method is that we did not interview any potentially vulnerable stakeholders – those whose stake in the management of the data is personal as opposed to professional, and who have little influence over data processes other than opting in or out at the beginning of a project, e.g. customers of energy providers. Talking with

people from these perspectives could provide insight on issues such as informed consent: Do customers understand the nature of the information that can be inferred from energy usage? and power dynamics of who wins and loses based on the value of the data. We also know the burden of privacy risk is carried by consumers more so than the researchers or companies. We also have a limited view on how results and conclusions are verified by those outside of the project.

Interview-based social science research is often structured around multiple and longer interviews. One difference between our study and typical social science research is our narrow focus on a mostly non-personal topic: working with data at work. Longer and more interviews per participant could have provided more information, although given the nature of the topic, the shorter and single interviews do not seem to leave out too much context or insight.

Our participants are professionals working on energy access and sustainability issues. Another stakeholder group, energy consumers, has a different perspective based on inherently different priorities and motivating factors. There are plenty of potential ways energy consumers' perspectives could become skewed in how they are represented in the results of our interviews with the professionals who work with their data. Some of our interview participants described learning about customers perspectives from deployment managers, so in addition to the potential for misunderstandings or missing information, they are sharing second or third hand points of view. Our interview results reflect how professionals interpret consumers' perspectives, rather than the consumers' actual perspectives. The concluding sections are about implications for data and code management infrastructure. Further study of consumers' own perspectives is necessary for a more grounded analysis of these implications. That said, we chose to begin our exploration of perspectives with stakeholders who are presently making decisions about data management and analysis because they are the primary users of data management and analysis tools and will either adopt or reject new ones.

2.3.6 Interpreting Interview Data

To process the interview transcripts, we identified interesting and important excerpts, and then categorized those excerpts into themes. We then processed each theme by identifying and sorting them into subthemes often along the lines of challenges, strategies, or common issue within that theme. The results below are a summary of each of the themes identified in the collective interview data, followed by bulleted lists of important points that came up within each subtheme.

2.4 Interview Results

We coded interview transcript excerpts into four main themes of interest: data collection and management issues, motivations to share or not share data, how to share data, and perceptions related to privacy. Within each of these themes were multiple subthemes, which we summarizes in the following subsections.

2.4.1 Data Collection and Management

Interestingly, multiple interview participants pointed out infrastructure challenges when it came to collecting quality data, especially but not exclusively limited network bandwidth in rural areas. Some even expressed that the data collection processes are more challenging than figuring out how to share the data. Others emphasized more the cost of storing and sharing large amounts of data as the ability to collect it rapidly increases. Multiple participants also noted having more trust in the sensor data than survey responses when it comes to understanding electricity usage, however other participants brought up the ability to directly observe which appliances a household has when conducting the surveys in person. As such, the data quality may depend on what questions are being examined.

2.4.2 Motivations to Share or Not Share Data

Motivations aligned with whether the participants were primarily doing research or business, or if they were significantly interested in both their company and their research. Reasons for wanting to share data were primarily related to doing better science, i.e. enabling more transparency and reproducibility in research, or the extra business insight that may be learned from having access to more and different types of data. All of our interview participants affiliated with a business mentioned competitive advantages as the biggest factor in considering what to share and how. In line with other data sharing studies, our interview conversations revealed data was most commonly shared among individuals and whether or not to share data with others on an individual basis was influenced by trust and anticipated benefit such as contributions to the project.

2.4.3 How to Share Data

The most common form of sharing data involved granting access to individuals who sign data use or nondisclosure agreements, either in exchange for consulting services, as part of an academic research partnership, or both. When data is prepared for sharing more broadly, obvious identifiers are removed from the data. Multiple participants noted questionable utility of data made available as supplemental materials for publications or satisfying open data requirements at the end of a funded project, potentially due to purposefully minimal sharing or a lack of foresight, time, and effort in curating the data for reusability. One of our interview participants does spend significant time and effort to make data accessible and useful to other researchers, although the rest spoke about general reusability issues more so than their own efforts.

2.4.4 Perspectives on Privacy

Although misconceptions about privacy were not repeated enough across interview conversations to include in Table 1, some misconceptions did come up and most participants admitted not knowing the best ways to protect privacy so they simply try to do their best. The notable misconception that came up a few times was the impossibility of re-identification after names, addresses, and clearly identifying attributes were removed from the data, and along with this misconception was a frustration data was not made more freely available.

Most participants interestingly acknowledged the importance of privacy while pointing out it was less of a concern in the low and middle income countries (LMIC) where they work compared with the United States. Our interview participants also used similar language to talk about only sharing data on a "need-to-know" basis, and that the data they have could cause problems if it were "in the wrong hands."

2.4.5 Summary of Perceptions

The following lists summarize points made by interview participants. See methods section for how we identified themes and subthemes. Points repeated by more than half of interview participants are emphasized in bold. These points do not reflect the views of the authors, only those of our interview participants. All the points included here were either mentioned by multiple participants, or in a few cases particularly insightful even if only one participant brought up that point.

Data collection and management

- Data collection and data quality
  - Collect data on electricity usage patterns by asking people in a survey or home visit, using smart meters or other remote monitoring, or decomposing aggregate load data (varying degrees of involvement and accuracy)
  - Data collection is challenging due to infrastructure issues: network, meter error, time synchronization, energy system outages, or system artifacts that appear to be usage patterns but are not
  - Data quality issues are discovered at the point of analysis
  - Lack of existing nation-wide data in one low-middle income country (LMIC), e.g. how many solar companies are there
  - Our data quality is good
  - Sensor data is more trustworthy
- Scalability
  - As the volume of data increases, there will be more stakeholders and interested parties
  - Human resources are a scalability challenge, especially for building sensors
  - Network issues in remote areas bottleneck amount and frequency of data that can be collected
  - For projects that do not have a huge amount of data parallel processing systems are probably not relevant
- Tools

  - There is less readily available infrastructure in this LMIC country/ analysis would be better if we had access to better tools
  - We write our own analysis code because we have a capable team and existing tools are expensive and do not fit our needs
  - Synchronizing system updates with outsourced tools can be frustrating

Motivations to share or not share data

- Value or benefits of data sharing
  - Greater data sharing speeds up research and development (as in product development)
  - Desire to improve transparency, accountability, and scientific processes
  - There are always limitations, so having access to different types of data is helpful
  - Compare work with other researchers or companies
  - Research publication makes more sense when data is included
  - Advertising
- Reasons to potentially not share data or code
  - Data about energy consumption and how systems behave is interesting but also requires quite a bit of context
  - Suspicion of anyone who is not familiar
  - Competitive advantage
  - Fear of investor turn-off
  - Not much benefit
  - Privacy of customers
  - Intellectual property
  - Do not want data to be seen out of context (control the story by sharing summary of data instead)
  - Data is specific to the work of this company
- Obligation to open data policies
  - Not much resistance to requests for exceptions
  - Not historically enforced but seems to be changing
  - Good for replication
  - Funding usually comes with reporting and data sharing requirements
  - Obligation to release data in service contracts

- Tensions
  - Want open data but suspicious of data cleaning and filtering decisions made prior to final published version
  - Does not make sense for data to be made public unless it matters to everyone, e.g. air quality data
  - Different entities want different data
  - Happier to publicize data that makes company look good
  - Inclined to monetize/ sell the data, sometimes the data becomes more valuable than the product
  - Scholarly incentives to publish, not to honor scientific process
  - More available data could increase funding for the field at large, although competitors could benefit disproportionately

How to share data

- Making data available
  - Most common to have individual collaborators/ contributors sign data use or nondisclosure agreements and then give them access to the data
  - Protocols for sharing data are predefined in agreements at start of projects
  - Anonymized data included as supplemental materials to publications, possibly along with documentation and statistical analysis program files
  - Work-in-progress API access for third parties
  - An app on the app store facilitates access to some data
  - Might request exemption from data sharing policies in case of sensitive or proprietary data
  - Upload data to Google Drive Enterprise or sometimes dropbox
  - Curate data before making it public for data quality purpose
  - Files on encrypted cloud storage and collaborators have decryption keys
- Receiving data
  - Depends from contract to contract
  - Utility or energy company grants access via file sharing software or permission to login To their servers directly
  - US utility company created a process for researchers without monetary incentives to access data freely but under use agreements
  - Data use agreements can cause delays when parties disagree
- Challenges

  - Open data is not well defined so people will share the minimum amount of data allowable and not much energy is invested in making that data useful
  - Energy technologies have significant variability which complicates comparison across projects
  - Standardized metrics are not always applicable
  - Cost of time and money associated with sharing especially as more data can be collected
  - Technology changes faster than policies are written

- Navigating issues related to various stakeholders
  - Open data policies can be a source of uncertainty
  - Continually re-evaluate and adapt data protection strategies
  - Various levels of access depending on the stakeholder, the type of project, and the data use agreements
  - Disconnect between marketing and engineering teams in terms of communicating what the data means
  - Some say public utility data should be public, others entrust utility to make the best public use of the data
  - Lack of data protection laws beyond use agreement

- Enabling value & mitigating risks
  - Transparency in what data is collected and used for what purposes
  - Improved policies and legislation
  - Pooling data of several companies rather than sharing on individual bases
  - Use a disinterested third party to facilitate data sharing
  - Standardized metrics
  - Metrics that are better contextualized

Perspectives on privacy

- Relevance and attitudes
  - Privacy issues are critical/ data is sensitive and detailed
  - People are less concerned about privacy in these particular LMICs
  - Little control over which companies collect and use which data
  - Extra aware of and careful about data privacy when the plan is to make it publicly available
  - Should only share data that is necessary for research

- Deciding what to keep private
  - If the data does not concern everyone, it should not be shared broadly
  - In addition to minimizing potential re-identification of individuals or households, academic researchers sometimes also avoid disclosing which companies are in their studies
  - Deciding what to keep private depends on which stakeholders are gaining access
- Challenges
  - Risk to privacy depends on the environment, e.g. other available data and processing algorithms that are constantly changing, not just what data is published
  - Regulation struggles to keep up with industry practices
  - Households may not understand how their data could be used in harmful ways, or how much of their data is stored indefinitely
- Strategies
  - Store data and identifiers separately
  - Propose legislation that requires disclosure of data breaches
  - Update and evolve data protection protocols
  - Limit number of people who have access to personally identifiable information
  - More general purpose anonymization or aggregation techniques in the future
  - Do not share real-time data externally
  - Automate anonymization
  - IRB can enforce data protection in the event of disagreement between open data policy managers and researchers
  - Check to make sure there is not identifying information in free text responses

2.5 Do Data Sharing Policies Address Trade-offs?

As a means of understand the funding agencies as another stakeholder, we analyze their open data policies and guidelines at the time of writing. In particular, we are interested to see whether trade-offs such as the added costs of data curation and potential for confidentiality breaches are addressed by these policies.

USAID's open data policy – Automated Directives System 579 (ADS 579) – accounts for potentially sensitive data with access levels an implementing partner would propose at the time of completing a Development Data Library (DDL) Submission Form. If accepted, "Non-Public" limits access to one or more Federal Government program or agency, or "Restricted Public" which limits availability of data to researchers under certain conditions, such as data use and non-disclosure agreements, depending on the data and the potential for re-identification of individuals [10].

ADS 579 includes section 579.3.2.3 *Redacting Data and Exceptions to the Open Data Mandate*, which states

> "To the extent that the redaction process is likely to remove information that would prove useful during secondary data analysis, Operating Units must work within prevailing best practices to identify alternative redaction methods or consider ... assigning an access level of "Restricted Public" to the Dataset [10, p. 14]."

The same section of the policy references Executive Order 13642 including the following statement,

> "agencies shall incorporate a full analysis of privacy, confidentiality, and security risks into each stage of the information lifecycle to identify information that should not be released... It is vital that agencies not release information if doing so would violate any law or policy, or jeopardize privacy, confidentiality, or national security [10, p. 12]."

This policy language [1] points to the importance of secondary analysis alongside protecting privacy and confidentiality, leaving some ambiguity as to whom is responsible – the agency or the implementers – and without specific guidance on how to share enough information for secondary analysis without compromising privacy by releasing too much information.

In the *Final NIH statement on sharing research data*, the National Institutes of Health states,

> "[T]he rights and privacy of people who participate in NIH-sponsored research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects [11]."

The NIH data sharing policy and implementation guidance points out unusual variables will pose greater risk of being re-identified and suggests general privacy preservation techniques: withholding parts of the data, statistically altering the data in a way that will not compromise future analysis, restricting access to the data with a data enclave[2], or some combination of these. NIH provides a *Data Sharing Workbook* with examples of how other investigators have shared data. The data sharing statement, policy, and implementation guides acknowledge data sharing is complicated and point out a few high level privacy preservation techniques. Deciding among these strategies and ultimately taking responsibility for protecting privacy of participants is left to investigators, their IRB, and their institutions.

DFID has in place the *Open and Enhanced Access Policy* [13] focusing on removing monetary barriers to research outputs, especially publications and software but also including datasets, videos, audio, and images. The policy itself does not acknowledge privacy concerns other than presumably as potential and exceptional security or ethical constraints:

[1] See Glossary of Terms Used for USAID's Automated Directives System (ADS): http://www.developmentwork.net/component/glossary/USAID-ADS-Glossary-1/O/Operating-Unit-1279/

[2] NIH defines a data enclave as, "A controlled, secure environment in which eligible researchers can perform analyses using restricted data resources.[12]"

> "Exceptionally, exemptions may be granted to specific policy requirements. Generally, these will be granted only if doing so would lead to better development outcomes. Exemptions may also be granted on grounds of security, legal, ethical or commercial constraint . . . DFID will consider these requests and may grant an exemption [13]."

The corresponding *DFID Research Open and Enhanced Access Policy: Implementation Guide* [14] similarly focuses on published outcomes. In regard to datasets, the guide suggests how to choose or establish a repository and then briefly defers to a 2008 report by the Research Information Network, *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs* [15] simply for its discussion on raw versus derived data. This report is a summary of findings from interviews with over 100 researchers and data experts in the UK about their attitudes related to data-practices at that time. In 56 pages, the report has one paragraph, *Legal and ethical constraints* in the section about *Constraints on data publication and use*. The paragraph begins with these two sentences, with primarily licensing issues in mind, "It is not always clear to researchers whether or not they have the rights to make datasets publicly available. This rarely appears to prevent researchers sharing datasets on a one-to-one level, but gives pause for thought when it comes to publishing the data more widely [15, p. 27]." The same paragraph continues with the following about personal data and confidentiality concerns:

> "In areas of research where personal data are collected, issues of confidentiality and data protection come to the fore. There are anonymisation techniques available to mask the identity of survey participants – though the ESRC has identified a shortage of skills in this respect – but many researchers appear reluctant to obtain permission from interviewees to share the project's data, fearing that to do so might diminish the likelihood of interviewees continuing to participate in the study. Often consent is only sought [for] the purposes of the original project, precluding re-use of those data for other projects [15, p. 27]."

These findings on their own are not surprising, especially for a report on attitudes about data-practices ten years ago. Nonetheless for outdated and problematic attitudes and practices to be the only, granted indirect, mention of the handling of sensitive data referred to in DFID's policy or implementation guide might serve to reinforce them.

That said, researchers have made efforts to anonymize data shared on DFID's R4D platform [16] at their own discretion. More recently DFID also published a report, *DFID Digital Strategy 2018 to 2020: doing development in a digital world* [17]. The strategy references *The Principles for Digital Development* [18] – one of the principles being to, "Address Privacy and Security" – and names data security as one of many issues DFID will collaborate on with other government departments, without going into detail.

These data sharing policies and others like them are a driving force for data being shared more widely, the primary motivation being to increase impact and insights gained from investments by making data available for research findings to be validated and expanded. The policies range from brief mentions to explicit prioritization of privacy considerations over openness goals and vice versa. Some funding agencies reserve the rights to make final

judgments about what data gets shared and how, while others defer completely to any other applicable regulations and the researchers' best efforts. Many policies suggest expert involvement without specifying what qualifies as adequate expertise. We examine how intentionally balancing information with privacy considerations can improve both data sharing results and privacy protection over existing practices. Policies that address the information and privacy trade-offs will do more to help realize the visions of open data and open sciences.

For discussions on journal data sharing policies, attitudes about data sharing, and compliance trends we suggest the reader considers [9, 1, 2].

2.6 Implications for Data and Code Management Tools

Some researchers, especially economists, point to increasingly prominent pre-analysis plans as an effort to improve transparency, and many researchers we interviewed regardless of field pointed to issues with how useful open data can be. As such, it seems that researchers may be willing to take small steps at the beginning of studies to improve research quality. Open data policies could specify a requirement to create data sharing plans at an earlier stage in the project. A data sharing plan, not unlike data management plans some funding agencies already require, could identify key variables for reproducibility as well as privacy.

Our interview participants, inline with typical open data policy recommendations, perform de-identification manually by removing identifiable information–although they may not release the data publicly at all–and yet de-anonymyzation is usually accomplished with automated analysis scripts. As such, filtering data should be done in a rigorous, automated way. In the process, it may also be possible to remove less overall data that helps contextualize results. Automated anonymization is primarily based on how unique individuals in the dataset are, and could be tuned to additionally consider how sensitive an attribute may be. As in other fields, many of our interview conversations, though not all, reflected a real perception of manually de-identifying data as good enough, especially when the associated identifying data is stored separately and offline. The privacy literature says decisively that de-identification is not good enough, in mathematically sound terms. At the same time, we are curious about whether there may be a difference between practically and theoretically good enough anonymization strategies. Given the perceived risks and values of sharing their data our interviewees discussed, it may be interesting to explore what the threats and motivations to steal and re-identify individuals in the data. Assumptions made around what constitutes "good enough" data protection often lead to unanticipated data breaches. In developing data management and sharing technology, we thought about how data can be shared on a need-to-know basis such that data released to the public ought to be differentially private [19] and when highly accurate data is needed it can still be filtered in a rigorous way.

As we mentioned earlier, multiple participants also used the exact phrase, "in the wrong hands" to convey their data could cause problems if unauthorized parties gained access. Naturally this data cannot be made public, and so if some form of their data were subject to open data requirements it would be a heavily filtered version. A heavily filtered version of research data is likely not enough to reproduce results, so varied levels of access to data would help protect the data while also enabling reproducibility. Data use agreements are also

common and could be incorporated into levels of access in a systematized and helpful way. Erring in the direction of theoretically robust privacy strategies is also preferable because privacy depends not only on the data access but changes in the environment around the data in terms of availability of auxiliary data and inference potential of algorithms. For this reason, we advocate for sharing data in a monitored, interactive way as opposed to uploading and downloading sanitized views of the data.

Our interview results informed us of how data can become more interesting to a wider variety of stakeholders as it grows in size, which increases risks and exacerbates challenges. This supports the cases for automated anonymization as well as a cloud environment that can scale gracefully and provide cutting edge security at a physical level and beyond.

Chapter 3:

Design Requirements

3.1 Introduction

This chapter introduces the design requirements for sharing workflows and is based on work evolved from the authors' original contributions to a proposed end-to-end data management platform called Mezuri [20]. As such, the paper describing the Mezuri platform is cited heavily and sometimes referred to as the Mezuri paper. Some design requirements were adapted from Mezuri and others are based on new insight and slightly different goals. Mezuri is an end-to-end, fully featured data management and analysis platform. Tools for sharing workflows are an important component of such a platform. Some of the design requirements laid out in this chapter are inherited from Mezuri design requirements (Table 1 in [20]), and others are new or adapted (see 3.1) to more narrowly focus on the goal of controlling how code and data are shared among stakeholders in a useful way. We first motivate the need for well-designed workflow sharing tools in development engineering, and then describe four representative motivating use cases followed by the set of derived design requirements and their explanations.

3.2 Motivation

It is common for technology companies with abundant resources to spend large portions of revenue on research and development to continuously improve products[1]. In contrast, technology projects targeting emerging markets (often unfamiliar and vulnerable) tend to have limited resources to collect and analyze data beyond the reporting expectations of their funders. We are interested in collaborative data management infrastructure for the sake of enabling development practitioners such as these to better understand their own work and make use of available data in the interest of more appropriate and positively impactful work. As such, our example use cases come from the field of development engineering, also known as Tech for Dev or Information Communication Technology for Development (ICTD).

Our design requirements for data management infrastructure that works well for sharing workflows within this field are based on the specific workflows of four projects as well as broader values in this field such as usability, provenance, accuracy, consistency, and efficiency. Possible Health describes "efficiency as a moral must" in their culture code[2] which goes on to claim, "It's everyone's job to turn time into resources and possibility for our patients... We are obsessed with using simple tools to shrink the time we spend on 'work about work'.

[1] See https://www.theatlas.com/charts/N1Gs8E4v

[2] http://possiblehealth.org/wp-content/uploads/2014/02/Possible-Culture-Code1.pdf

There is a CRITICAL and constant push towards making our individual and team workflows as efficient as possible."[3]

A significant amount of technical expertise and hours goes into setting up data management infrastructure and moving data from collection tools, to processing tools, sometimes again to analysis tools. This work is often redundant, therefore inefficient, across similar projects and even within one project when team members have different preferences, for example different programming languages. Using an end-to-end system like Mezuri essentially bootstraps these initial set up efforts and enhances the ability to maintain these projects over time. Sharing workflows is an opportunity for projects to learn about success and failure from each other and reduce time and resources spent setting up data analysis tools.

3.3 Examples

The following are development engineering examples of data management and analysis workflows that would have benefited from data management infrastructure for sharing workflows. A few themes resurface for each example including heavily technical workloads and interdisciplinary teams. The case studies involve one or both of monitoring and evaluation (M&E) workflows and research-oriented algorithm development workflows. The teams behind the first two use cases (SweetSense and GridWatch) were also co-developers of Mezuri.

3.3.1 Water Pump Monitoring

SweetSense is an Internet of Things (IoT) product used to collect data on several types of applications[4] including the usage and function of water pumps and filters. A sensor detects when someone is pumping water as well as how much water is flowing, and thus infers whether the water pump is broken. The water-filter sensors are configured to only monitor how much water flows through the filters, i.e. how often they are used. Sensors report data points every so many hours or days, and in the case of the water pumps the data populates a map showing which water pumps seem functional or not, and which sensors have not been reporting. The company managing the sensors was collaborating with the organization responsible for maintaining the water pumps. An outside team of statisticians was incorporating the sensor data along with survey responses as part of a randomized control trial. At one point more than ten studies were ongoing to investigate impact of water filter and other interventions, along with studies of how to incorporate the sensors themselves in impact analysis work. This work was taking place in a country that mandated the data be shared with relevant government offices. The data about individuals was further protected under the Institutional Review Boards[5] of the home institutions in the US. Reporting to officials and protecting data privacy both require carefully organized and trackable workflows.

[3]The authors did not work with Possible Health. We reference their culture code because it captures a sense of urgency and efficiency as a form of responsibility towards intended beneficiaries, directly related to the data management tools they use. In other contexts, the "work about work" can be seen as more beneficial, fruitful, and worthwhile to spend time and resources.

[4]See http://www.sweetsensors.com/

[5]We describe Institutional Review Boards (IRB) as they relate to data-management in Section 2.2.2

With multiple ongoing studies, it is valuable to reuse key workflow components such as data cleaning scripts and statistical analyses for the sake of consistency and to avoid redundant work. Provenance of data analysis in this case and others involving sensors entails keeping track of sensor firmware versioning as well as data and processing code. Ideally a workflow could be published in such a way that readers and stakeholders of the research can understand and help validate conclusions without exposing proprietary code owned by the entities overseeing sensor deployments. This project involved programming and deploying sensors, collecting and managing data, and coordinating several staff. Any shared tool must be easy to use or it would disrupt rather than support the several ongoing processes among interdisciplinary and intercultural team members.

3.3.2 GridWatch

GridWatch monitors power outages using data reported from civilian smartphones running the application. Alternative sources of understanding power grid outages would be power companies who have little incentive for being transparent about less than ideal downtime, and smart meter deployments which can be cost prohibitive. GridWatch uses sensors built-in to standard smartphones to detect a local outage, and with enough participants in an area, the scale of an outage. A sister project, PlugWatch, has the same goals and reports similar data only from a dedicated, semi-permanent phone installation constantly plugged in, with the ability to be integrated with and collect data from commercial power meters, and send time-based reports instead of event-based[6].

Some of the data being collected by a phone running GridWatch includes whether the phone is plugged in, whether wifi is enabled and access points are available, acceleration, peak frequency from an audio sample that would correspond to the hum of electricity current, GPS, and timestamp. When noteworthy changes in these measures are detected, the data is reported as an event to a cloud application that collects and processes it along with other event reports, ground truth or other relevant data sources such as utility incidents, tweets–for example to the power company (e.g. KPC), and maybe weather. The first level of processing event reports is to do analysis and training to look for outage signals, and to tune the detection and classification algorithms. When the classification methods are determined, another cloud application can take the raw data, apply the classification methods, and output information about power outages over space and time. Power outage information can then be sent back to participants in real-time, and over a longer term used to study grid quality.

Some defining characteristics of GridWatch that influenced our design work include a highly technical academic team setting up a project that will generate data of serious interest to less technical interdisciplinary audiences, the sensitivity of energy usage data revealing personal behavior patterns, algorithm development workflows based on sensor and survey data, and a focus on international development.

[6]See https://github.com/lab11/PlugWatch

3.3.3 Cookstoves

A cookstove project led by our colleagues is a quintessential example of monitoring and evaluation in development. The project directly measured adoption of cookstoves by installing temperature sensors on a sample of 170 (out of 35,000) cookstove recipients at internally displaced persons (IDP) camps in Darfur, Sudan. Researchers compared this temperature time-series data with survey data from participants reporting how often they were using the cookstoves. Potential types of bias that show up in surveys on adoption rates include recall and appeasement. Comparing sensor data with survey results is a way to gauge the degree of influence these sources of bias can be expected to contribute and have a more accurate understanding of intervention impact.

The cookstove project first involved instrumenting the stoves with sensors called Stove Use Monitors (SUMs), built upon iButton data loggers. Time-series data from the sensors was collected in parallel with survey data. From there, "data processing operations included spot-checks, cleaning up data, normalizing data, generating the set of cookstove events using [a novel cooking event detection] algorithm, and finally creating summary statistics" [20]. Version-controlled R and MATLAB scripts were used to process both sensor and survey data. Excel was used for spot-checks in a mostly ad hoc and irreversible way.

Data cleaning, normalization, and statistical analysis involve common operators that could be found in the operator library. Building a processing pipeline from mostly existing operators would have enabled the cookstove researchers to focus more of their time on developing their cooking event detection algorithm within one custom operator, even if the algorithm involved several steps. They would also benefit from version control at the lower level of each processing step in the workflow of tuning the event detection algorithm because it would be more clear which updates had to do with which steps.

Development practice has had a broad interest in cookstoves as an intervention because of their potential to save fuel, save time collecting fuel, reduce indoor air pollution, and increase overall efficiency of cooking labor. Ideally disparate yet similar efforts to distribute and evaluate the impact of cookstoves could learn from each other. We believe missed opportunities to learn from other projects results have a significant effect on the realized impact of development work. The particular cookstove project that motivated our design work noted, "the techniques and environment used to process the data for this case study are not scalable or replicable in any meaningful way for other studies. Even if the processing scripts would be available via services such as GitHub, it is still hard for others to understand them" [20].

3.3.4 EEG labeling

This use case is not limited to international development or monitoring and evaluation for impact analysis. A platform is being developed for crowdsourcing electroencephalogram (EEG) labels [21]. EEG signals include features or patterns that are manually labeled and used in clinical and research workflows. The platform seeks to serve a global community including professionals in low- and middle-income countries.

A workflow for collecting and analyzing EEG labels involves an EEG recording being displayed to an EEG technician or technician in training who submits one or more labels

for each segment of the signal they are shown. This builds up a repository of EEG signal data and various labels which may or may not be correct. In fact, the most highly qualified EEG technicians will choose a label correctly around 80% of the time [21]. Metadata about a technician such as experience and past labeling accuracy within the platform is used to weigh all the labels and determine which is correct.

In order to study labeling accuracy, the initial EEG recordings shown to technician participants have already been labeled. The labeled data set is tens of gigabytes and cleaning the data involves formatting and filtering the data by types of labels. As such, complete data processing is not feasible on laptops. In order to study the categories of labels in a comparable way, it is important to reproduce the data cleaning steps exactly even though there may be several months between analysis of different types of labels. The data set is also publicly available and other researchers may also be interested in studying labels, it is important to document the data provenance in a readable way. Although data cleaning and filtering is an important step, ideally the majority of the project owners' time will be spent on the research, using clean and labeled data.

Besides studying label accuracy, a goal of the platform is to make newly labeled EEG data available to researchers. The labeling platform is also in a position to enable new collaboration practices in which researchers may contribute data or data processing components and workflows, in exchange for labels, more data, or new data processing components and workflows. This sharing of data and workflows requires varied levels of access such that the general public does not have access to sensitive information and collaborating researchers can execute meaningful data processing workflows.

3.4 Design Requirements

The table below is a list and description of design requirements along with which of the four use cases need each requirement. The "Derived from" column specifies whether the requirement is inherited or adapted from the Mezuri platform [20] or based on new understandings of sharing workflows. Sharing workflows includes sharing data and processing code, which is a focused sub-goal of the Mezuri platform. A workflow sharing tool that satisfies these design requirements could be a part of the Mezuri platform infrastructure without itself being a fully functional data management platform. Some requirement rows are italicized to designate they belong to the broader infrastructure and not necessarily a workflow sharing tool, e.g. data storage, however the workflow sharing tool needs to be compatible with components that satisfy these requirements.

Requirement	**Definition**	**Derived From**	**Use Cases**
Support survey + sensor data	Supporting survey and sensor data requires compatibility with a variety of data sources and formats	All of our example use cases incorporate both sensor and survey data	All four

Accuracy/ transparency	Changes to data and processing code are immutable and traceable such that the system could re-run a workflow and get the same results; workflows can be shared at varying levels of detail	Mezuri + Open Data goal of reproducibility	All four
Standardization, common schemas	Data types and schemas are compatible across workflows within or among projects	Mezuri + cases in which multiple parties contribute comparable data	EEG
Durability	*System failures do not result in loss of data or workflow components*	*Mezuri*	*All four*
Isolation	*Processes are isolated from each other to avoid side effects*	*Mezuri*	*All four*
Privacy	Sensitive data can be protected from unauthorized parties	The original Mezuri paper defined the privacy requirement as compatibility with most IRB requirements. Ideally a data management system could also proactively support projects handling sensitive data.	All four – SweetSense especially for code and EEG especially for data

Scalability	*Workflow infrastructure does not need to change as a project grows in amount of data or complexity of data processing*	*Mezuri and cases where data is collected from several sources as opposed to a few of particular interest*	*EEG, GridWatch maintains workflows on their own cluster and reported this being an adequate scale for now*
Sharing	Stakeholders working within and among other organizations can understand and sometimes reuse others' data and code	Mezuri + consideration of different stakeholders interacting with data along with open data requirements	All four
Flexibility	Workflow development must be highly customizable	Projects with access to technical expertise have a strong preference for being able to incorporate their own customizations	All four – even projects that lack full-time technical expertise can bring in temporary support
Efficiency	Avoid redundant work	The nature of projects with limited resources is to value the efficient use of time and money.	GridWatch, EEG, Cookstoves
Consistency	When workflows have overlapping steps, such as data cleaning, the steps should not cause discrepancies as a result of differing implementation details	Causes of inconsistent results within and across projects can be difficult to find and interpret	EEG, Cookstoves

Usability	Non-technical experts can contribute to workflow development	Stakeholders tend to be interdisciplinary	All four

Table 3.1: Design requirements and their explanations

3.5 Conclusion

This chapter focuses on design requirements for sharing data analysis workflows based on initial work on the Mezuri platform and four use cases of particular interest that involve social impact, technical, and international components. Sharing workflows is a central aspect of collaboration, and can help facilitate the scientific process approach to research. Sharing workflows in real-time with a user interface in the same way documents are shared in real-time with Google Docs has added benefits of easily keeping track of and incorporating feedback or minor contributions from several collaborators. Current practices often involve isolated workflows of individuals and selective sharing based on personal connections via email attachments, which hinders collaboration among interdisciplinary stakeholders and makes it hard to keep track of changes overtime and ultimately reduces the quality of the results.

Improved data and workflow sharing tools can be implemented in a variety of ways to satisfy these requirements. The next chapter describes the tools we designed and built.

Chapter 4:

Pipelines and Usability for Sharing Data Workflows

4.1 Introduction

Collaboration tools for teams, especially Google Docs and Slack, are increasingly ubiquitous and raise expectations for collaborative work. However it is still common to share data or processing code as downloadable files and email attachments, which makes it nearly impossible to keep track of who ends up getting access to what or to compare how different people come to different conclusions. Data collaboration can happen on other specialized shared data management systems such as DHIS2, however these tend to require the data be formatted in a certain way before being uploaded and target specific applications such as health or energy planning. We suggest more generalizable infrastructure in most cases for the sake of interdisciplinary collaboration starting at the point of data collection.

When two researchers come up with inconsistent findings as a result of data analysis it is difficult to track down the cause. Programming errors, using newer or older versions of the same datasets, fundamentally different understandings of the problem or opinions about methods to analyze it, among other issues can create confusion in trying to collaborate. Data analysts with different programming language preferences and levels of skill might also have trouble comparing workflows. Keeping track of all these potential points of contention is a challenge even when all parties organize their work carefully and sometimes impossible if not organized. In the best case scenario, revisiting details of one or more workflows to track down errors and inconsistencies costs time and redundant use of computing resources that are already a limited resource.

We propose pipeline-based workflows in which a data source is specified and data analysis steps–transformations, filters, visualization, etc.–makeup nodes in a directed graph that the data flows through until the output format is reached. A single step in the data analysis or node in the pipeline graph, what we call an operator, can be edited or updated without disrupting the rest of the pipeline so long as the input and output types remain the same. The operators are modules of code that can be reused in other pipelines and stored in a library with version control and descriptions of exactly what the operator code does. This option already exists as the programming model of Apache Beam [22], which evolved from Google Cloud Dataflow (CDF) [23]. One of the main advantages of using Apache Beam for data processing is managed parallelism, although here we focus on the collaborative benefits of using any pipeline-based approach.

Writing pipeline-based data analysis programs still requires some specific programming knowledge. Although the program structure is predictable and straightforward–the oper-

ators are written or imported and then the pipeline is described–unfamiliar programming paradigms are a barrier to realizing the benefits of cloud computing for projects with limited access to technical expertise. We leveraged the predictable program structure to build a code generation interface in which the user specifies where to get input data and draws the pipeline graph with drag-and-drop style operator blocks. The interface allows a user to save or download the Apache Beam program code or run it immediately. Options for running the pipeline such as local or cloud execution and where to store the output can be included in the generated program or specified at runtime. The pipeline code, or simply the graph and corresponding versions of input data and operators, can be saved as a snapshot. This makes reproducing and comparing workflows much easier. The pipeline model never updates the data it processes, only transforms it and generates new output data. This avoids loss of information and enables redoing of any of the steps. Pipelines make clear what the data provenance and each of the steps were, which are easy to lose track of otherwise.

This chapter names and describes the important features of workflow-sharing infrastructure we chose based on the needs of three main example use cases that encounter a variety of issues described in the previous chapter. We implemented these features as a tool for collaboratively building Google Cloud Dataflow[1] (CDF) pipelines. CDF is now also available as Apache Beam [22].

4.2 Related Work

Topics of related work includes block programming, other user interfaces for data analysis without block programming, and other programming tools for collaborative data management without user interfaces. The design of the interface and supporting features described in this chapter generates code for and can execute pipelines on Google Cloud Dataflow, which provides managed parallelism. Other block programming tools described below can be integrated with Apache Spark or Hadoop, or support cloud database sources, although this would require technical preparation. The CDF Pipeline Builder is meant to be fully cloud based. Internet connectivity is becoming more and more ubiquitous, and to use the Google Docs analogy again, an offline mode could help users progress between periods of connection. Schemas of the data and metadata for some subset of operators can be stored locally such that validity of processing pipeline arrangements can also be checked offline. Other user interfaces for data analysis are cloud-based although their support for complex data modeling can be limited, or their user interfaces for specifying the data processing pipelines can be complicated.

The “freemium” model is common for data management software, i.e. there is a limited free version and a costly, often subscription-based, fully-featured version. Licensing issues are outside the scope of this chapter, except to note 1) expensive software is generally unreasonable for resource-constrained projects, and 2) setting up data-management infrastructure to depend on tools with limits on the size of data or number of iterations to a processing pipeline is a generally unreasonable risk for real-world projects. In some cases the limitations are minor and unrelated to size of the data or number of iterations to analysis. Academic

[1] See https://cloud.google.com/dataflow/

licenses and reduced pricing for nonprofit organizations are also common.

There are endless tools for data management and analysis. Those described below are highlighted for their prominence and successful or similar implementation of features in our own pipeline building tool design.

4.2.1 Block Programming

Block programming is often thought of as an approach to teaching people, often young people, about programming concepts and functions, represented by blocks that can be assembled into programs. Block programming libraries can also used to make real applications (See https://snap.berkeley.edu/, https://scratch.mit.edu/, and https://developers.google.com/blockly/). The advantages of applying this paradigm to data analysis workflows revolve around having a visual representation of each step of data processing as opposed to less readable, less arrangeable analysis scripts. Indeed, block programming shows up in data management platforms.

RapidMiner Studio (https://rapidminer.com/products/studio/) is an enterprise-oriented platform for analytics teams with a sophisticated user interface for visually arranging data analysis processing graphs and includes a library of several algorithms and functions. The free version is limited to one processor and 10,000 rows, or an unlimited license is thousands of dollars per year.

EasyMorph (https://easymorph.com/) is a visual data transformation software with 70-80 built-in transformations that can be applied to a dataset. The free version has a limited number of transforms and iterations per project, and the unlimited version is also exorbitant.

Orange (https://orange.biolab.si/) is an open-source program for visually building data processing pipelines from a library of widgets, which are the equivalent of operators. Orange has a simple and elegant user interface and the ability to add a custom widget in the form of a python script. Interestingly, Orange is marketed as a teaching tool for data mining, situating it as a peer among other block programming tools that are not focused on data processing. Both Orange and RapidMiner AutoML offer predictive analytics that highlight most likely important or significant variables and correlations.

4.2.2 User Interfaces for Data Analysis

Popular user interfaces for data analysis are Tableau (https://www.tableau.com/) and DHIS2 (https://www.dhis2.org/) both of which facilitate the creation of online data dashboards and reporting. DHIS2 was originally designed for health data and has since been leveraged by projects in other sectors. Tableau has an extension in beta that allows python scripts to be included in workflows, enabling machine learning applications and more sophisticated data modeling.

Google Data Studio (https://marketingplatform.google.com/about/data-studio) offers several compelling features, especially for managing dashboards, data reporting, and collaboration, which is built on Google Drive technology. There is an interesting overlap of helpful data analysis features for the use cases of marketing and reporting to funding agencies. The CDF pipeline builder is designed to be very similar in user experience to Google

Data Studio, although instead of teams collaborating on a shared report, collaborators have a shared view of the data processing pipeline.

Ona (https://ona.io/) is a commercial extension of Open Data Kit that offers extra built-in features including role-based access control, and data filters and charts. Open Data Kit (https://opendatakit.org/) is a widely adopted open-source survey and data collection tool designed for resource-constrained projects.

Jupyter notebooks are essentially a user interface for python programming with the ability to execute code cells individually and view the output directly beneath the code cell. Code and output can also be documented with markdown cells. Jupyter notebooks are particularly helpful for data analysis as visualized output can be viewed and edited efficiently. A well-documented, version-controlled Jupyter notebook is useful for sharing code and keeping track of provenance. Anyone familiar with Python programming can quickly and easily learn to use Jupyter.

Excel, along with Google Sheets, is worth mentioning here because it is the dominant data-management tool for development projects that lack extensive technical support (which is most of them). When sensor data has few columns and thousands or tens of thousands of rows, or when a survey data has over 400 columns (questions) and relatively fewer rows (respondents), the usefulness of Excel becomes limited. For one, a data analyst can no longer get a sense of the data by skimming through raw data visually, and it becomes non-trivial to select and manipulate subsets of rows or columns.

4.2.3 Collaborative Data and Workflow Management Tools

In the scientific community, especially computational physics, chemistry, and materials science, increased processing capabilities and computational resources have lead to significantly more data generated from simulations, often consuming up to millions of compute hours per project. It is highly beneficial to the scientific community not to waste human and computer resources on redundantly generating this data in order to collaborate and reproduce or expand on analytical results. Government funding agencies such as DOE and NSF, as well as several journals, now also require making some form of raw data and processing scripts used to generate relevant plots available publicly or by request. This is a challenge in part because the output of large simulations tends to be gigabytes of unorganized data, which is then filtered down to a few text files with a few columns and hundreds to thousands of rows of data that ends up being relevant to publications and sharing. Another researcher wanting to study the same simulation data, filtered slightly differently, would depend on the original researcher having saved the original raw data and taken care to preserve provenance.

It is interesting to look briefly at how computational scientists perceive data sharing issues when privacy is not a concern. Although privacy dominates conversations about sharing data even indirectly linked to human subjects, data management and sharing trends in computational science remind us what the barriers are after we account for privacy. For example, there are methods being explored to generate fake datasets retaining several statistical properties of original raw data while individual rows are meaningless [24], at which point the barriers to data sharing and collaboration in social science might become more analogous to computational science for some cases.

We can learn about these barriers and proposed solutions from a framework for data and workflow management created from within and for the computational physics, chemistry, and materials science community called signac [25]. The signac application, part of the signac framework, essentially manages metadata for file-based data analysis workflows, which offers some properties comparable to a relational database, and ultimately facilitates a well-organized data space along with provenance as a researcher iterates through testing theories. The signac framework includes an additional tool for documentation and claims the combination of documentation and metadata management ensures interpretability, and as a result accessibility even for researchers not using signac. Although there are similar attempts to improve data and workflow management [26–28], signac focuses on collaboration as a primary goal.

4.3 Design

This section describes features we built into a tool we refer to as a "pipeline builder" to satisfy the design requirements laid out in the previous chapter. The primary insight we built upon is that cloud-based workflows, especially in Cloud Dataflow, offer several advantages that would be valuable to our target projects if cloud environments were more accessible to researchers other than computer scientists. This insight is based on a related observation: technical and non-technical researchers alike often spend a significant amount of time setting up data management and analysis infrastructure components that look similar to other projects, i.e. reusable data management infrastructure will save these projects time. We demonstrate how many of the design requirements are supported by building our tools on top of Cloud Dataflow and then describe features we built from scratch to address usability.

The pipeline builder does not address data collection or storage, however importing data from a custom source, such as Open Data Kit (ODK), only needs to be solved once and then that custom source can be reused among any projects collecting data with Open Data Kit. Custom sources and sinks, along with built-in support for common data sources and sinks, are a feature of Cloud Dataflow exploited by the pipeline builder. In lieu of a custom ODK source, data from ODK can be loaded into BigTable or Google Cloud Storage.

Requirements	Supporting features of pipeline builder
Support survey + sensor data	Google Cloud Dataflow (CDF) is compatible with varied data sources
Accuracy/ transparency	Version control for pipelines, and data processing creates new output rather than modifying original data
Standardization, common schemas	Out of scope, standardization is future work that requires either automated schema and data type conversions or large-scale coordination by data and project owners
Durability	*Out of scope because durability depends on underlying data storage, but this is solved in practice by using the cloud.*
Isolation	*CDF offers a secure processing environment*
Privacy	Of code: operators can be shared as black boxes; Of data: sharing pipeline can be independent of data sources, permissions can be used to require that a privacy-preserving data-filtering operator be used with input from a sensitive source
Provenance	Pipelines and data sources can be saved as snapshots, future git integration
Scalability	*CDF provides managed parallelism*
Sharing	UI for building pipelines can be shared like Google Docs
Flexibility	Custom operators can be added to the operator library
Efficiency	Pipelines and operators are reusable and easy to share
Consistency	Reusable library of operators
Usability	Readability of pipelines, shared user interface, code generation, and operator metadata that describes functions each contribute to overall usability for teams

Table 4.1: Design requirements from the previous chapter along with supporting features implemented as part of the pipeline builder tool. Italicized rows indicate requirements that would be applicable to other components of data management infrastructure and are not directly features of the pipeline tool.

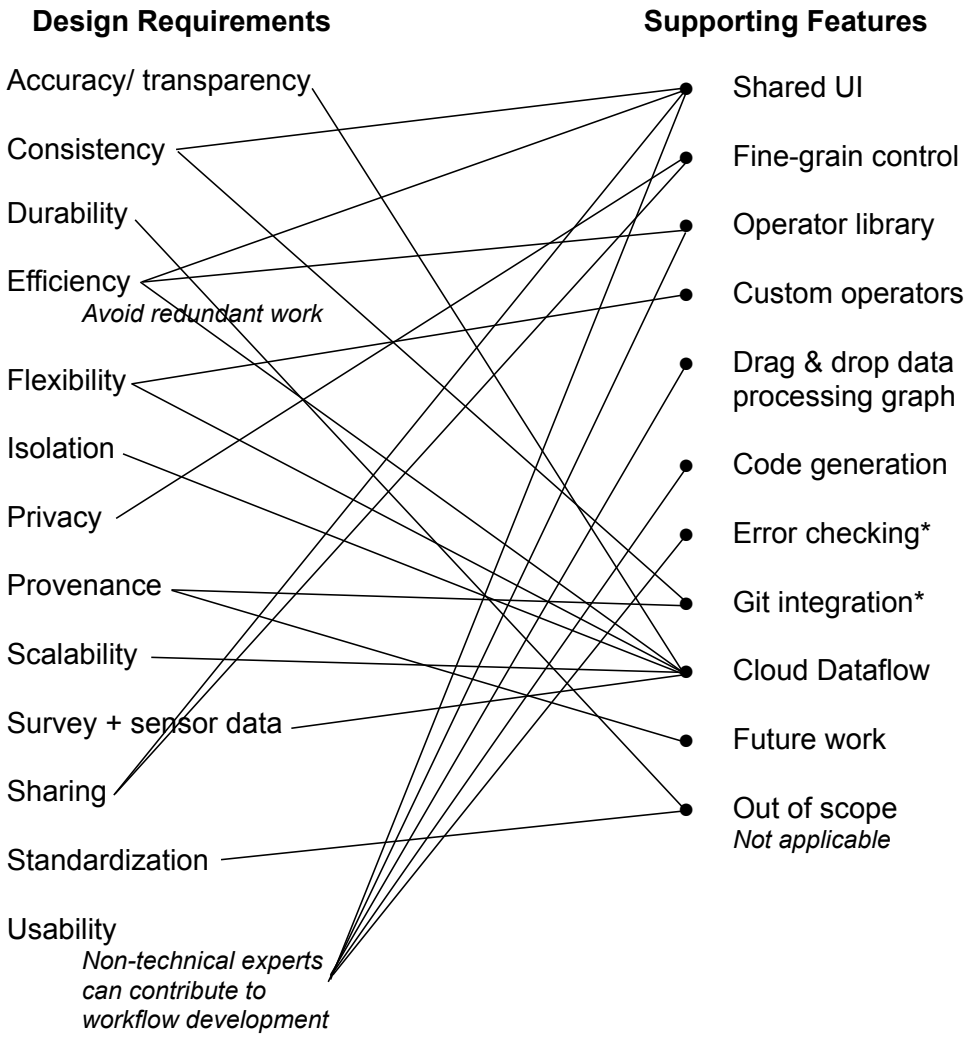

Figure 4.1: This diagram illustrates how each design requirement is satisfied by one or more feature.

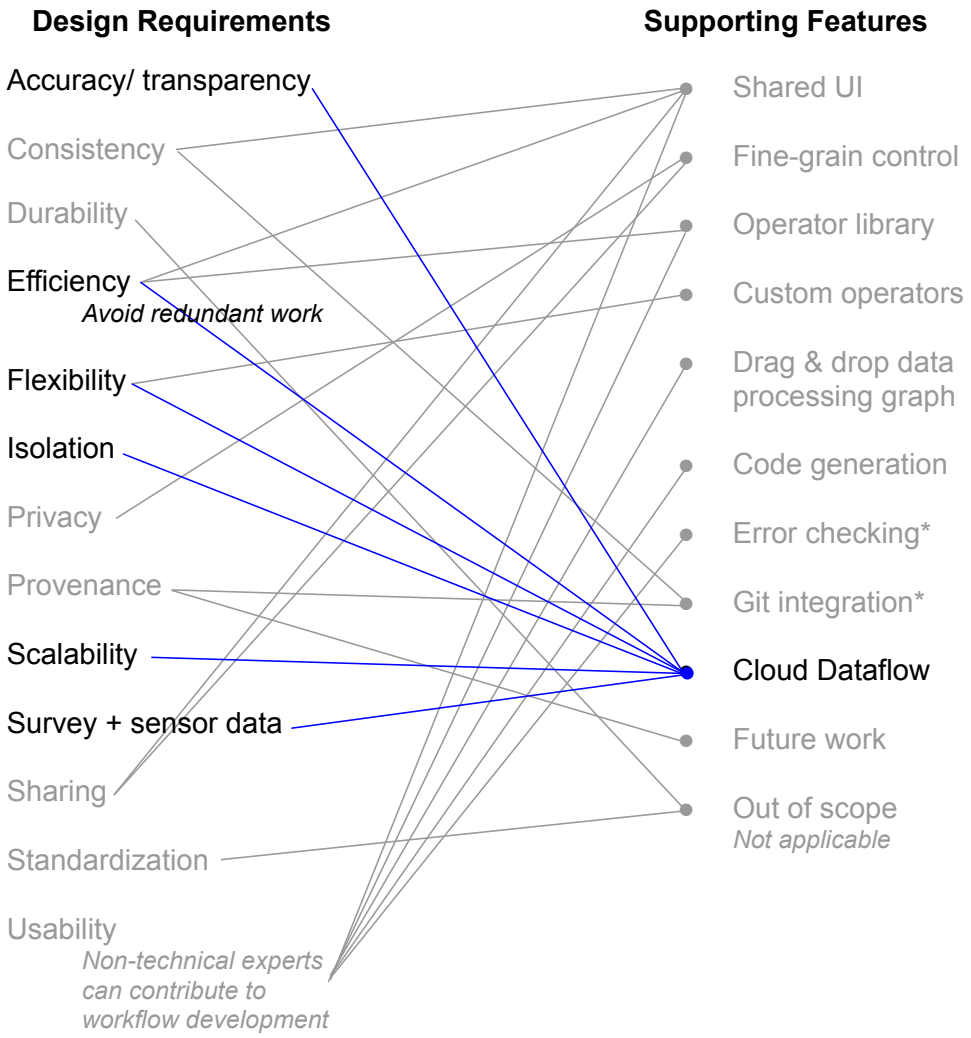

Figure 4.2: This version of the diagrams highlights how many design requirements are supported by incorporating Cloud Dataflow into our tool.

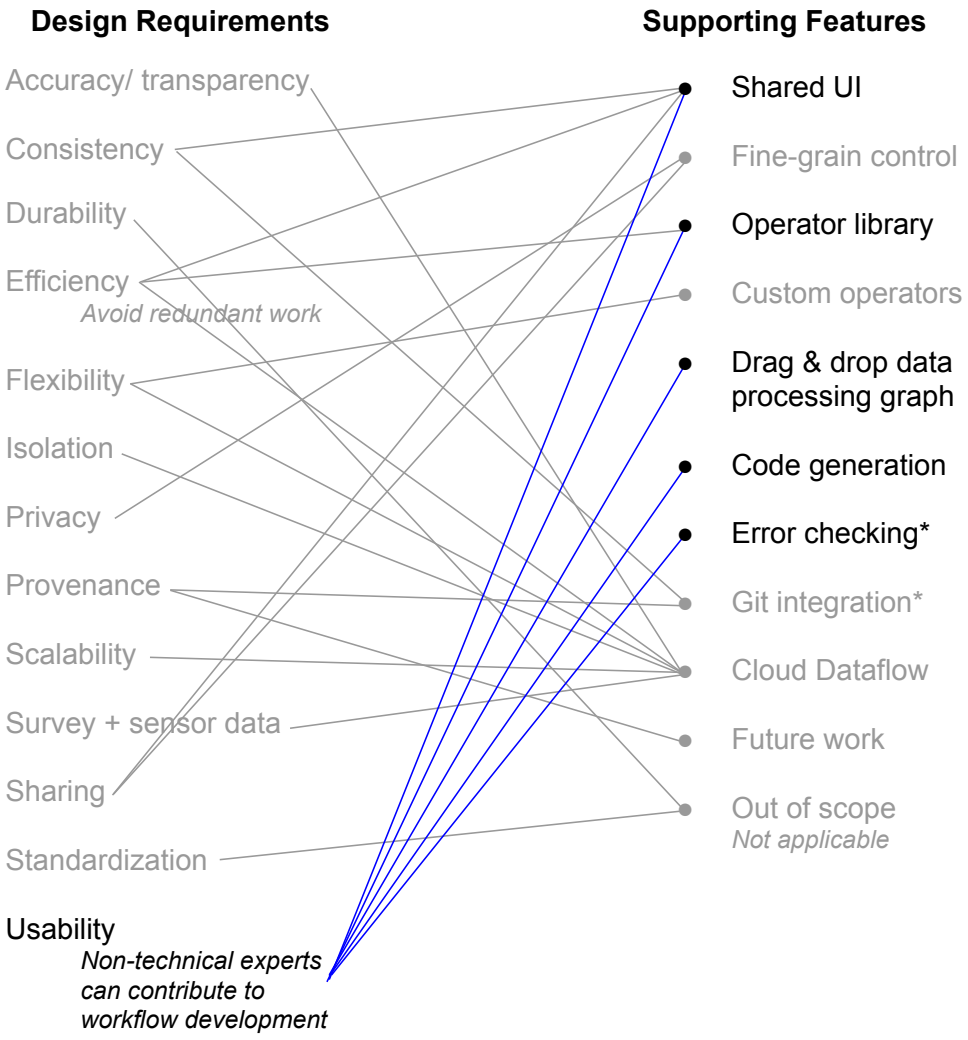

Figure 4.3: This version of the diagram highlights how many features we implemented for the sake of usability.

4.4 Features

4.4.1 Shared User Interface

We implemented the user interface with AngularJS and jsPlumb for the drag and drop graph-drawing functionality. The data source and location are specified at the top of the page. The library of operators is displayed in a scrolling panel at the left side of the workspace, from which they can be dragged and added to a pipeline graph. Operator blocks include input textboxes for parameter specification (see figure of steps). When a pipeline is assembled, it can be executed directly from the user interface or the generated processing code can be saved. Multiple users can edit or view pipeline documents, much the same way groups can share Google Docs documents or JupyterHub notebooks.

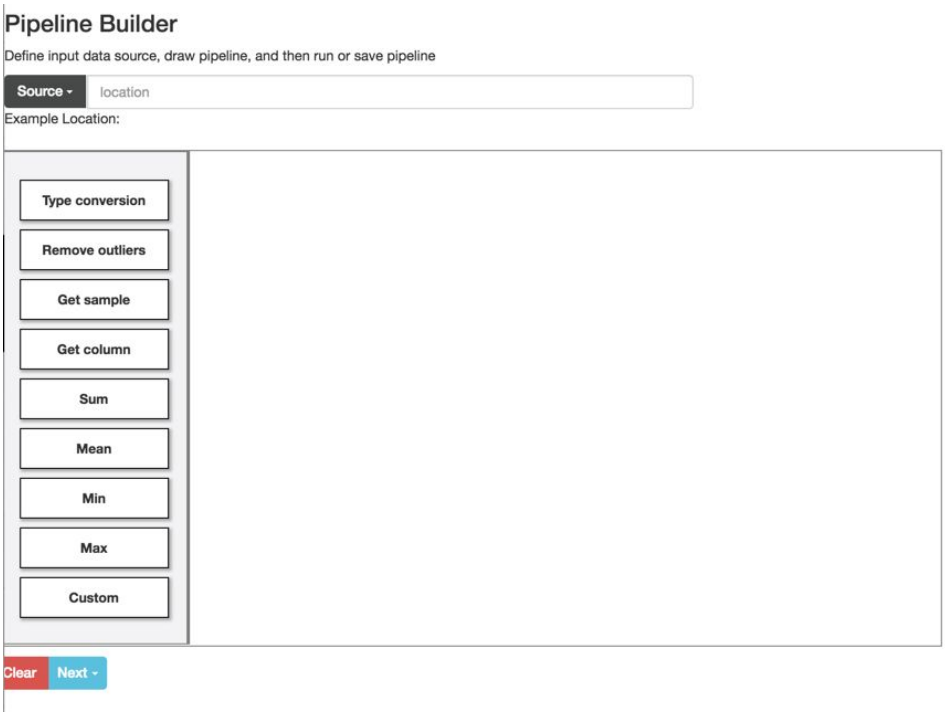

Figure 4.4: Blank pipeline document

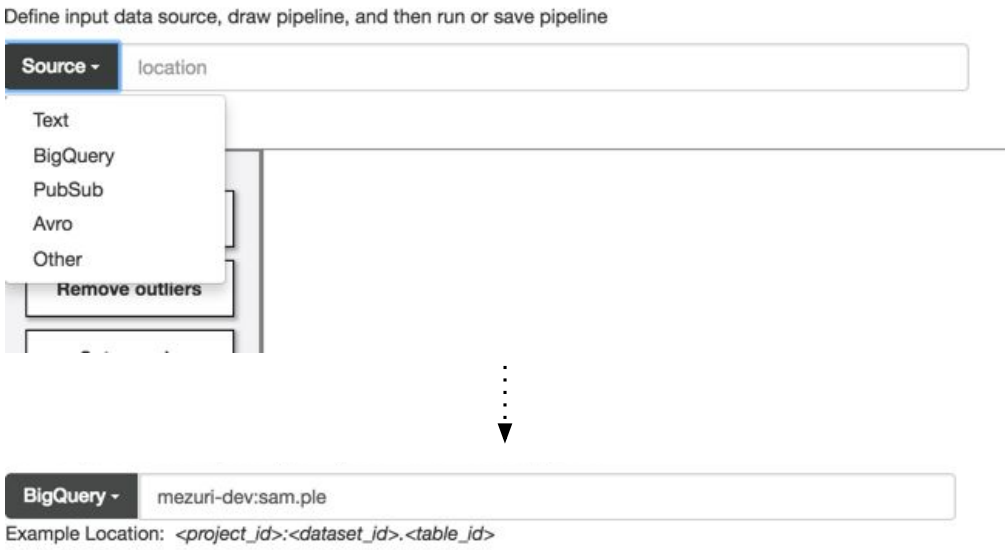

Figure 4.5: Step 1: Specify input data

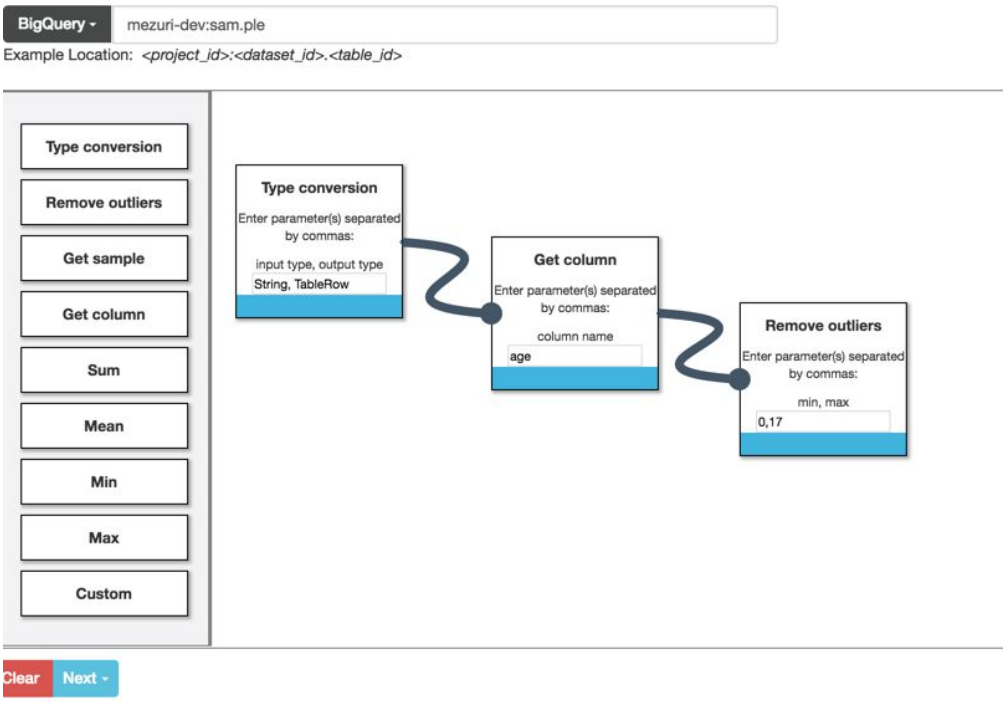

Figure 4.6: Step 2: Draw pipeline from operators in side menu (library of available operators)
Step 3: Specify output (write to file, table, or visualization)
Step 4: Run on in the cloud, locally, or simply save the Apache Beam code

4.4.2 Fine-grain Control

Rather than keeping track of disparate and entire data-processing programs, the pipeline model offers operator-level control. More fine-grained control has implications for privacy and useful data sharing in particular when the alternative is to expose all-or-nothing of workflows, data, and data processing code. Fine-grain control over sharing code and data as needed enables much of the value in verification and reproducibility while mitigating risks of data and proprietary code reaching the wrong hands. The pipeline builder interface provides an easy mechanism for researchers to publish detailed and readable information about their workflows and data schemas, which can be enough in many cases to accept or discuss validity of analysis. Generated pipeline code will not execute if the user does not have permission to access the designated input data. Multiple project contributors can build a pipeline with sample or fake data and then a project owner can simply change the input data and run the pipeline on project data. This effectively separates concerns of access to sensitive data and shared workflow development. In the same way, projects with similar data sets on different populations can share workflows by simply specifying their own input data sources. While many scenarios call for private repositories for analysis code, not sharing workflows makes conclusions difficult or impossible to validate.

4.4.3 Operator Library

The purpose of an operator library is to organize the blocks of code used to build each step of pipeline graphs. An operator library keeps track of built-in or imported operator code along with metadata on the name, description, input type, output type, version, and tags. Tags are searchable and help group operators by any characteristic users find helpful. A typical example is to tag operators by type of functionality such as filter, type conversion, statistical operation, or visualization. Libraries can be expanded by importing operators published to any code sharing repositories.

4.4.4 Custom Operators

Apache Beam includes an extensive library of common operators that can be incorporated into pipelines out of the box. More often than not, customized operators are also needed. Below is a simple example to show the format of a custom operator written in Java. The code generation tool finds the '//customcodehere' line in the template and inserts self-contained operator code such as this above that line. The generated pipeline code is one file. An advantage of Java for code generation is that inconsistencies between spaces and tabs in the template versus custom code will not cause runtime errors.

4.4.5 Drag and Drop Data Processing Graph

A feature of the user interface is adding and arranging operators into a data processing graph by dragging each operator from the library and dropping it into the workspace. Drag and drop functionality is a concept borrowed from block programming and implemented for usability. The visual representation of the resulting graph improves the ease of debugging logic errors and the readability of the program for collaborative workflows. We implemented

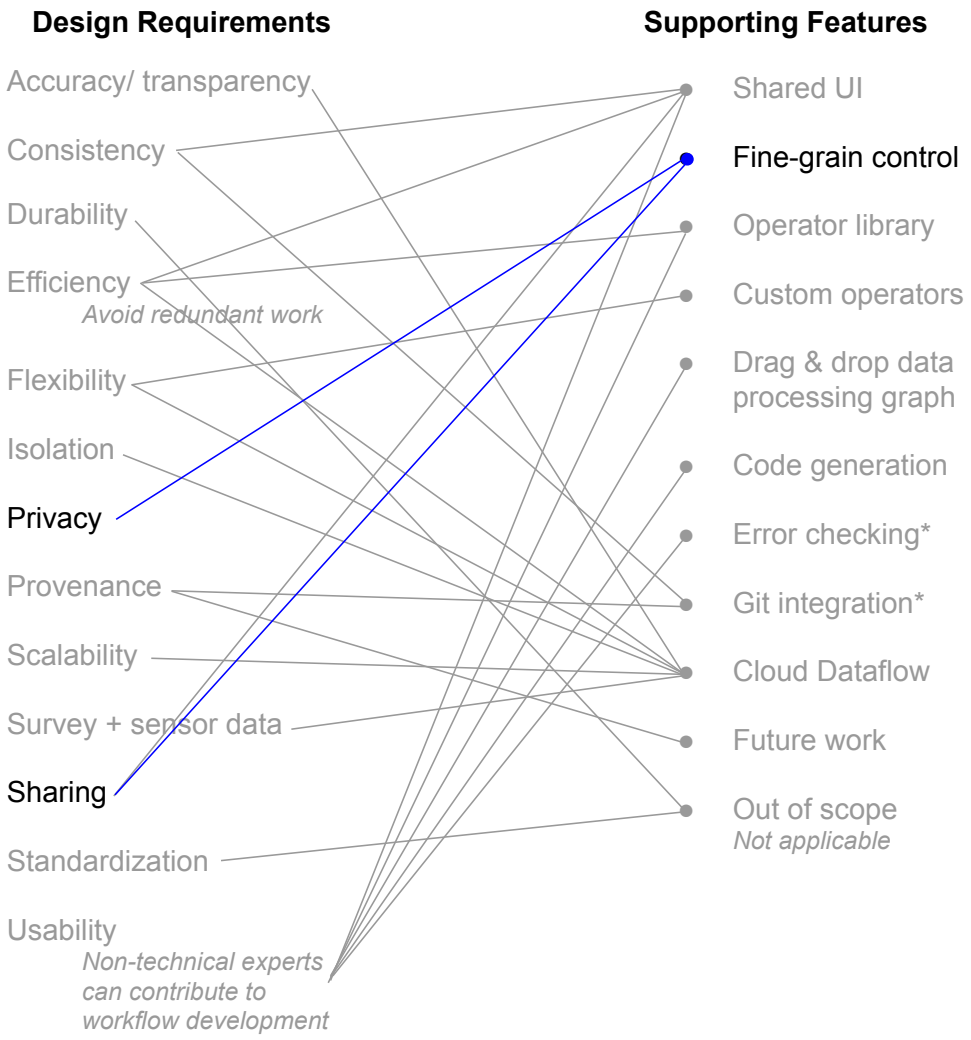

Figure 4.7: This version of the diagram highlights fine-grain control for both privacy and sharing, i.e. fine-grain control is the feature that enables balancing privacy best practices with sharing enough information.

FtoC.java

```
public static class FtoC
    extends PTransform<PCollection<Double>, PCollection<Double>> {
        @Override
        public PCollection<Double> apply(PCollection<Double> lines) {
            PCollection<Double> celsius = lines.apply(ParDo.of(new DoFn<Double, Double>() {
              @Override
              public void processElement(ProcessContext c) {
                c.output((c.element().doubleValue()-32.0)*5.0/9.0);
              }
            }));

            return celsius;
        }
}
```

Figure 4.8: Example of a custom operator. This one converts temperature data from Fahrenheit to Celsius.

this using the Community Edition of jsPlumb, which includes drag and drop as part of a collection of tools for visually connecting elements in a user interface.

4.4.6 Code Generation

Google Cloud Dataflow, now Apache Beam, code is highly structured. Rather than being executed directly, CDF or Beam code describes a pipeline of operators which is compiled into more a optimized program for parallel processing. The predictability of pipeline code makes it straightforward to generate based on a graph, even with user-added, custom operator code.

A user specifies the location of input data, drags and drops data processing operators from the library to draw a pipeline, and then specifies an output destination. We wrote a Python program to validate the output of each operator is compatible with the input of the next operator and then generate the Cloud Dataflow or Beam code to run the pipeline. Our code generation tool edits a CDF or Beam program template which includes import statements, the basic program structure, and an empty pipeline object, p, and then appends 'p.apply(...)' for each operator in the pipeline graph and inserts custom operator code that can then be appended to the pipeline with the same process for appending built-in operators. Custom operator code is added in the same program file, outside of the pipeline specification. The original program template includes flags to guide the correct organization of operators in the pipeline graph and insertion of the snippets of custom code. If a custom operator is uploaded as a file, an include statement in the pipeline program will run correctly as long as the file paths are not broken. Alternatively, and especially for simple operators, custom code can be pasted into the pipeline program. After a pipeline graph is drawn and validated, the generated CDF or Beam code can be executed in the cloud or saved.

4.4.7 Error Checking and Git Integration

On top of these baseline features already implemented, our design incorporates error checking and git integration to further facilitate usability and provenance. Error checking would be similar to spellcheck and indicate when operators have been added incorrectly. For incompatible output to input types the interface would offer to insert type conversions when possible. Other errors the user interface could flag are invalid parameters or anything that conflicts with operator metadata. Git integration would provide operator-level version control and enable rerunning pipelines months after original results are shared, for example in response to reviewer feedback. The code generation tool could be adapted to save custom operators to the operator library and then be included rather than pasted in the pipeline spec, to improve both usability and version control. Without version control on both data and operators, researchers can find it difficult to reproduce their own work.

4.5 Validation

To validate our implementation, we describe how our motivating use case workflows could have been created with and benefitted from these features.

4.5.1 Water Monitoring

As the water monitoring project involved a large-scale, sophisticated randomized control trial (RCT) based on both sensor and survey data, as well as several other studies, setting up the data sources and cleaning tasks to be reused several times would eliminate redundant work among researchers within the organization, and establish a consistent input. Cloud Dataflow supports both sensor and survey data, and the pipeline execution creates new output, rather than modifying original data such that multiple users of the same data sources will not interfere with each others' work.

The survey data is protected by Institutional Review Board agreements, so having fine-grain control over how the data is used is one way of keeping track and being careful with who is using what data for which purposes and how. Operator-level control complements our privacy strategies described in the next chapter. Importantly, the more readable data processing, the more easily a variety of stakeholders with different backgrounds can work with the data and compare results. For projects in which not all participants share the same first languages, readability of workflows enhances collaboration.

4.5.2 Grid Monitoring

Figure 4.9 is an example of a GridWatch pipeline that could have been achieved with the Cloud Dataflow Pipeline Builder.

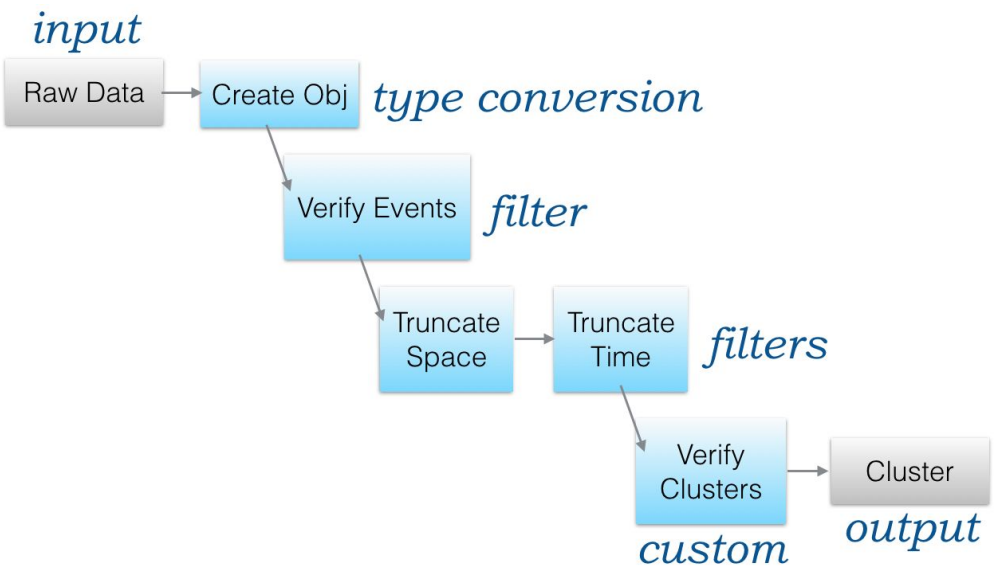

Figure 4.9: An example pipeline

The events in the GridWatch pipeline refer to measured activity possibly indicative of a

power outage. If multiple events are detected in a constrained area and time period, these events would be clustered and be indicative of a power outage. The clusters also provide insight into the duration and scale of the outages. In order to tune the power outage detection algorithms effectively, it is important that the underlying data remains consistent or that updates to sensors and data cleaning are trackable. This application also involves significant collaboration between engineers developing the technology and economists interested in studying impact of investments in energy infrastructure. As such, readable pipelines, similar to the example in Figure 7 can help everyone working on the project quickly understand this pipeline in its entirety and contribute to an individual component, for example tuning the clustering algorithm within the operator that verifies clusters. The pipeline model also ensures the preceding steps in the data analysis remain consistent, such that new results are clearly connected to specific operator updates. Otherwise, modified data cleaning or updated sensor firmware changes happening concurrently with tuning of the clustering algorithms can make it difficult to keep track of the causes for different results. In other words, the pipeline model enables readable provenance.

Because the data analysis involves several filtering steps, the library of reusable operators would also save time for the team. Researchers could easily review which filters have been used in which pipelines and quickly set up those steps of their pipeline, focusing most of their time and energy on new objectives. The sensors are live and publish new data regularly, which fits well with PubSub as the input data source. The pipeline builder enables setting up the process to connect to this data once and then reusing the first few steps, flattening the learning curve for researchers more familiar with batch processing workflows.

The GridWatch project aims to help policy makers understand the value of prioritizing investments in energy infrastructure. To do this, they will carefully make their research and data more broadly available. The pipeline builder makes it easier to not only share processing code within and outside of teams of original researchers, but to communicate the processes and findings.

This project also currently runs on private servers. While Cloud Dataflow pipelines can be run locally, building pipelines to run in the cloud has advantages. With the goal of being able to detect power outages, this application is sensitive to downtime. In a cloud environment, the researchers would defer responsibilities for server maintenance issues to the cloud provider.

4.5.3 Cookstoves

The first point of scalability is addressed in our design by leveraging Cloud Dataflow, and specifically the managed parallelism of data processing pipelines it offers without the user needing to think about parallel programming. Notably, any workflows involving Excel and scrolling through spreadsheets of data will not scale. The next points about replicability and the anticipated difficulty of other researchers understanding the original researchers' techniques are addressed by the greatly improved readability of pipelines over a collection of scripts as well as having a shared user interface. The original cookstove researchers could share a copy of their data processing pipeline without exposing any custom operator code they might not want to share. Fine-grain control allows operators to be shared as

black boxes or transparent blocks of code. Others could then interact with the replicated processing pipeline by using their own data as input, adding operators to format input data in a compatible way. Any disparities in data processing conclusions could be discussed over a shared view of the workflow.

Fine-grain control also refers to different ways of sharing the data, especially survey and personal or household data protected by an IRB. Chapter 5 describes mechanisms of sharing information while preserving privacy. In addition, the process of sharing pipelines described here enables other researchers to tweak operators and filter data based on varying alternative criteria and review the data in a dynamic way as opposed to static views of results or setting up their own data processing either from scratch or based on scripts which are often hard to read. The process of connecting data sources to processing code is built-in to the user interface and mitigates the need for other researchers to have the technical expertise needed to set up any other data processing infrastructure.

4.5.4 EEG Labeling

The EEG labeling platform does use Apache Beam for its initial data cleaning, in this case extract, transform, load (ETL), pipelines. Public EEG signal data along with corresponding labels and doctor notes were downloaded from the source and saved as files in Google Cloud Storage. The same few operators are reused to read the raw data files and a custom operator is used to filter data files according to types of labels and types of EEG signals. After filtering, relevant data is loaded into a database which will eventually populate the labeling platform. For these pipelines, collaborators reused operators by copy and pasting code sometimes and importing an operator from a shared folder other times. This could be problematic, especially as the code base grows, in that updates to operators are not disseminated to the copies being reused in other pipelines. In contrast, the operator library would keep track of updates and provide the option for researchers to use the latest versions or easily keep track of discrepancies in operator versions across otherwise similar pipelines.

Automated managed parallelism is a major asset to these ETL pipelines in that tens of gigabytes of data can be filtered in minutes or less, without the need to optimize the project in any way around the amount of data. The original data set is only one source of EEG data and the platform aims to serve a global community of researchers and clinicians. If the underlying infrastructure were setup to best support the current size of the project, these considerations would need to be revisited as the project grows. Instead, Apache Beam will continue to automatically calculate how many more processing nodes are reasonable depending on the amount of data and complexity of processing.

4.6 Conclusion

This chapter has described features of a collaboration tool for sharing workflows in the form of pipelines. These features represent one approach to satisfying the design requirements in the previous chapter. The pipeline builder user interface can be summarized as a usability layer built on top of Google Cloud Dataflow to leverage the advantages of its programming model, parallelism, and cloud environment and to make these advantages more accessible

to interdisciplinary collaborators. We made design choices based on the actual workflows as well as stated values of four example projects in development engineering. The lessons and ideas in this chapter can help existing and future data management infrastructure serve the projects similar to those we highlight.

Chapter 5:

Privacy Considerations and Strategies

5.1 Introduction

The spirit of open data is not at odds with protecting privacy, however in practice increasing how much data we share in a useful way without compromising the privacy of study participants is hard to do. As such, data sharing policies acknowledge the importance of mitigating risks to individuals and tend to either exempt or limit access to datasets with potentially sensitive information, accept filtered/de-identified versions of the data, or defer judgment to internal review boards and other local and federal policies.

Two problems with omitting and manually filtering data are 1) de-identification is not an effective method of protecting privacy (see impossibility result in [19]) and 2) the data that is left to be published is lacking context which would influence results and interpretations. Methods of filtering, de-identifying, and limiting access to data are decided on a case by case basis, without much shared understanding of the trade-offs and implications. This chapter explores those trade-offs, identifies a gap between the gold standard and alternative strategies in which several applications are lacking adequate approaches to data sharing, and describes a possible solution.

There is a spectrum of effective privacy preservation techniques and best practices, which we review in the related work section. The challenges of implementing privacy strategies while opening access to datasets are magnified in contexts where collecting and analyzing data in the first place can be a stretch of limited resources. Technical expertise is more likely to be focused on analyzing data and generating reports than preparing data to be shared. Among practitioners, data is often over-shared among trusted colleagues, resulting in privacy violations, or not shared at all, representing missed knowledge sharing and research validation opportunities.

Reliable methods of balancing privacy and information in data sharing will lead to improved participation, collaboration, and transparency in data-driven impact analysis or research, accompanied by better decision making or research outcomes. The more different stakeholders participate in collecting, analyzing, and interpreting data, the more results represent varied perspectives and reality. Collaboration and transparency also have the potential to reduce errors as well as the proliferation of "knowledge" based on incomplete or inaccurate information.

NPR recently highlighted a story in which a telemedicine project received multiple awards and $23 million from the Gates Foundation to scale efforts in India and over the course of three years showed zero evidence of impact [29, 30]. Meanwhile, there are several examples of impactful and sustainable telemedicine in India [31, 32]. The disconnect between failure and success was not only for lack of data management infrastructure, however we expect

better tools would help facilitate and enforce higher standards.

Precisely how to improve knowledge sharing with better tools and behavior changes is a generally underserved research area this chapter aims to explore and motivate. Section 5.2.1 provides an overview of the trade-offs for existing methods of protecting privacy, namely information versus privacy. We propose and describe three mechanisms of improvement over common practice along with evidence of their relevance in Section 5.3, followed by how these can be implemented for an interesting use case in Section 5.4. Although we describe mechanisms for navigating the tension and achieving balance between information and privacy, it is important to note: strategies opting to make privacy sacrifices for the sake of sharing more or more accurate information must be accompanied by data use agreements. Likewise, sharing less information does not guarantee privacy and reducing privacy protections does not imply shared data will yield more useful information.

5.2 Related Work

5.2.1 Trade-offs of Existing Privacy Strategies

The tension between openness and privacy is reflected in the spectrum of privacy techniques with the essential trade-off of different methods being how much information can be learned about groups of interest versus how well privacy of individuals in the data set is protected. On one end of the spectrum is not sharing data at all, the only perfectly private strategy. The opposite is publicly or broadly disseminating raw data, which maximizes the potential reuse and secondary analysis as well as the likelihood of exposing personal information about individuals in the data. Everything in between, (see Figure 5.1) is an attempt at balancing privacy and information. Determining whether a privacy strategy is appropriate depends on the sensitivity of the data, the data analysis needs of the application, and the audience receiving access. In Figure 5.1, we use dotted lines to represent thresholds for minimum viable information and privacy. If the data is sensitive and the audience is broad or untrusted, the privacy threshold is higher. If the data analysis depends on complex and precise answers from relatively smaller data sets, the information threshold is higher than it would be for simple statistics based on larger populations in the data.

De-identification usually refers to removing personally identifiable information (PII) column by column. Names, social security, credit card, and phone numbers, and other directly identifiable information should be omitted completely, although other PII might include birth dates and location information. Age representation in a population and proximity to resources or hazards could be invaluable context. Beyond directly identifiable values, individuals can be re-identified in a dataset when their data is combined with auxiliary information or when the individuals are relatively unique in the remaining attributes [33]. The differential privacy work includes an impossibility result demonstrating that once a dataset has been removed of columns to be sufficiently private, there would not be enough information left to be considered data, i.e. de-identified data is not a possibility [19].

Differential privacy [19] is a theoretical framework involving a measured addition of noise to query results to optimize accuracy under the restraint that the differentially private query results are skewed enough that they cannot be used to determine whether or not an individual

exists in the dataset. The amount of noise turns out to be very small when asking broad questions about a large dataset, while a significant amount of noise is added to specific queries about a small population. Differential privacy is considered the gold standard definition for privacy protection. There are multiple implementations continually improving the theory's usability in practice [34–37]. Because differential privacy involves adding noise, it may not be a good fit for tuning sensitive algorithms or in cases where the recipient of shared data has authority to require accuracy, such as a funding agency. Population statistics are good enough for many scenarios and fundamentally not enough information for others. The latter category includes combining data with other dependent datasets, and building models from data, including machine learning. These likely need to be run as workflows on full datasets.

Data aggregation techniques fall somewhere in between de-identification and differential privacy in terms of privacy protection and retaining interesting information (see Figure 5.1). Aggregation reduces granularity of data by summarizing the data in statistics or replacing precise values with ranges. To what extent aggregation preserves information and protects privacy depends on the strategy.

K-anonymity[38], l-diversity[39], and t-closeness[40] are aggregation techniques that reduce granularity of values in potentially sensitive columns enough so that no group of individual k, l, or t rows in the dataset are re-identifiable based on that column. These approaches are an improvement over de-identification for both privacy and retaining interesting information however still vulnerable to leaking information and have a significant limitation in that range boundaries must be predetermined and cannot be changed once published. For example, if an updated age column now includes the range 18-25 as a replacement value for individual ages between 18 and 25, then the k-anonymized data could not be used to study people aged 21 and older.

On the other hand, summary statistics are not much information on their own and still vulnerable to differencing, as described in this differential privacy reference [41], where statistics can be compared with each other to infer fine-grained information. When the summary statistic is based on a large enough population, such as election results by county, the risk is negligible. To publicize data about voter breakdown by several demographics or exit poll responses could easily become problematic in smaller communities. Latanya Sweeney demonstrated that 87% of adults in the United States were likely uniquely identifiable by their five-digit zip code, sex, and date of birth [33]. We propose and discuss dynamic and measured aggregation as one of our improvement mechanisms in the next section.

The blue dashed line in Figure 5.1 represents a threshold of how interesting the data needs to be after being sanitized per se and depends on the application. For example, differential privacy might not disclose enough information for algorithm development so the blue line would fall to the right of the point marking differential privacy for those applications. On the other hand, k-anonymity might provide more than enough information for describing broad population statistics, and for this application the blue line would fall to the left of the k-anonymity point. The red dotted line in Figure 5.1 represents a safe-to-share threshold, which depends on the role of the person or group requesting the data. For example, this red dotted line would shift up on the y-axis to represent what is safe to share with the public and down to represent whomever already owns the raw data. Appropriate methods of filtering data for given applications and audiences then would fall in the top right quadrant formed by

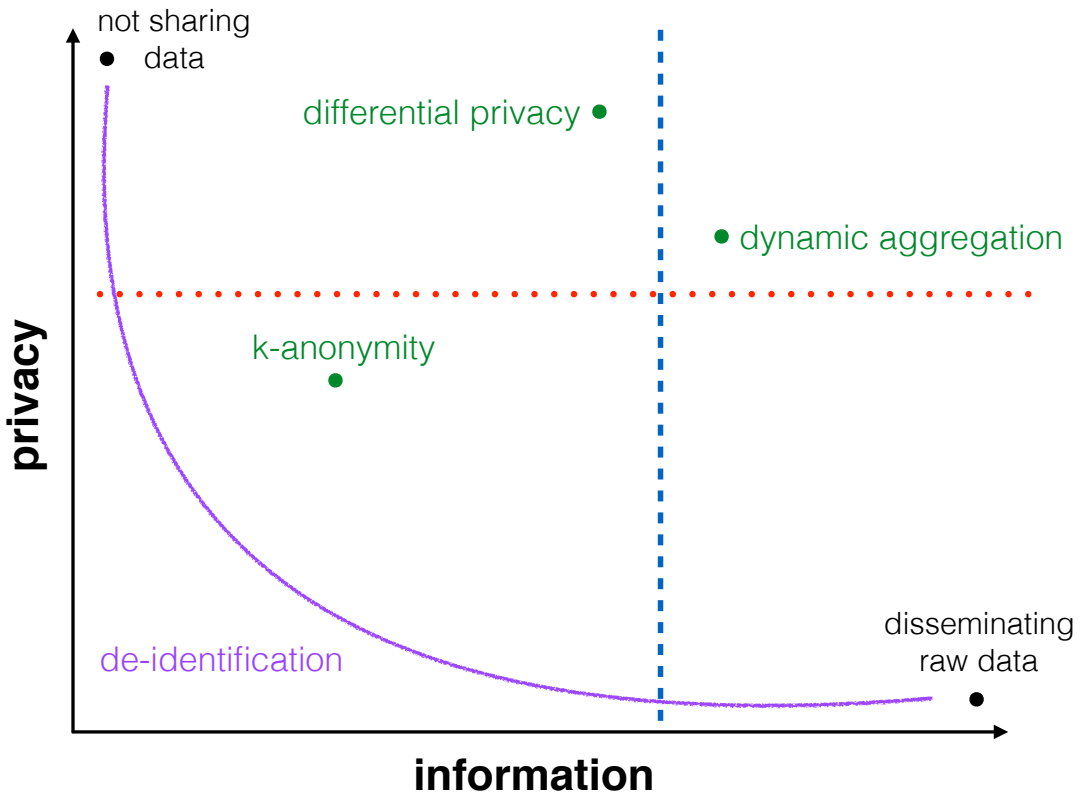

Figure 5.1: This plot illustrates the relationship between protecting privacy and retaining interesting information for a few methods of protecting privacy. The blue dashed and red dotted lines represent sliding thresholds for information and privacy, depending on the application and audience respectively. Simple de-identification methods are rarely adequate.

the blue and red lines representing those applications and audiences. Thresholds in Figure 5.1 define a space where some privacy guarantees of differential privacy would need to be sacrificed for precise analysis although the expectation of privacy is still high.

5.2.2 Related Work in Access Control Mechanisms

The Dataverse Project [42] proposed and made popular in, although not limited to, social sciences for "sharing, discovering, and preserving data" includes three levels of access control: public with terms of use, partially restricted, and restricted. For partially and fully restricted studies, data owners (also known as authors) can selectively grant access to individuals or institutions for example via groups of IP addresses or shared passwords. Such access control levels reject the oversimplified all-or-nothing attitudes towards open data. Funding organizations' open data policies currently tend to acknowledge the need to remove personally identifying information, leaving methods of doing so up to the discernment of project leaders, and then require the project leaders to justify if restricted access is necessary.

As the landscape of stakeholders described in Chapter 2 demonstrates, there is value in enabling several more levels and types of access. Considering requests to access restricted data on individual or institutional bases, can involve arbitrary discernment factors such as personal feelings and association biases. Data owners considering requests based on the intention behind a request and having a set of privileges prepared to match can make opening access more efficient and impartial. In addition, granted requests are more specific and informed about how much data is shared with whom, i.e. they improve balance of information and privacy protections.

Another approach to access control was implemented around the purpose of accessing certain data [43]. Purpose-based access control incorporates role-based access control to determine and enforce restrictions on what data is accessed for what purposes, and as such is similar to our approach of basing permissions in role-based access control on presumed intentions, or purposes. Purpose-based access control involves labeling data by appropriate and inappropriate purposes and then rewriting queries, whereas our approach focuses on data filtering mechanisms enforced by permissions. Purpose-based access control could be implemented within our version of role-based access control mechanisms as a subset of permissions applicable to certain user roles or groups of user roles.

5.2.3 Implementations of Privacy-conscious Data Management Systems

Along with privacy researchers at UC Berkeley, Uber has recently deployed and made open source its system for implementing differential privacy called Chorus [37]. The system accepts as input and rewrites a query such that it returns differentially private results. A tool within the system measures Elastic Sensitivity[34] of the original query based on database metrics. The Elastic Sensitivity measurement is then used to determine how much noise to add to results to satisfy differential privacy.

The Open Science Platform [44]vision includes six privacy levels based on risk and has associated security mechanisms such as user authentication, password and 2-factor authentication, and data use agreements. A related tool, DataTags, is a model for automating

the determination of these risks and policy suggestions that comply with legal and technical standards even in the absence of security and privacy expertise [45].

5.3 Improvement Mechanisms

In this section we elaborate on our two proposed mechanisms to improve information access as well as privacy over current data sharing practices, namely 1) a role-based access control list of user roles and corresponding privileges based on intentions and 2) a special permission we describe as dynamic aggregation. These two mechanisms necessitate a third and existing improvement of using a system to run analysis behalf of people seeking to use the data and return results, as opposed to sharing data by publishing "sanitized" datasets. Our goal in developing these improvement mechanisms is to enable more fine-grained control as well as accuracy in cases that preclude adding noise but should not settle for de-identification. One such case is labeling electroencephalogram (EEG) signal data by identifying subtle features. The labeled data is used for training medical assessment algorithms, further implying a low tolerance for error margins. Managing privacy considerations for both the patients and the technicians who contribute labels, for reasons we explain later, necessitates fine-grained control.

5.3.1 Role-based ACLs

The intention of sharing data is to learn as much as possible about issues and populations while respecting privacy rights of individuals. When multiple organizations independently release anonymized data about overlapping populations, all of the data is vulnerable to composition attacks [46]. As such, it is best to limit new access to data to only what is necessary to accomplish reasonable data analysis goals. Since reasonable intentions depend on the role of the data analyst and the application, we recommend fine-grained access control based on user roles and corresponding intentions and privileges as demonstrated in Table 1, which we originally described in this paper [20]. Role-based access control can be used to implement the concept of sharing data on a need-to-know basis even when the need is to learn as much as possible or to know a precise measurement of the population in the dataset.

Each collaborator can be assigned one or more user role based on intentions and minimum necessary access privileges to satisfy those intentions. Collaborators from the same organization can be assigned to a user group to avoid potential work around schemes of combining individuals' limited access to learn more than intended. For example, the concept of a privacy budget [19] can be shared by group members. The mechanism for measuring aggregation would compare a request with information that has been granted to the individual as well as the group.

Access based on the sensitivity of the data, i.e. potential risk to individuals, should be implemented in addition to permissions. The Open Science Platform and DataTags approaches described in Related Work are examples of this.

User Roles	Intentions	Privileges
Project owners	Full access, zero overhead, manage collaborators	Read original data + collaborators' aggregated data, write , add or revoke users
Participants	Learn about myself	Read data about myself + synthetic, noisy and/ or aggregated population data
Data collectors	Contribute data	Append
Funding agencies/ partnering NGOs	Learn about the population, monitor deployment progress	Read aggregated data
Colleagues/ researchers	Test code	Read synthetic, noisy and/ or aggregated data
Potential collaborators	Contribute data, test code, learn about the population	Append, Read aggregated data + own contributions
Other researchers/ the public	Learn about populations	Read synthetic, noisy and/ or aggregated data

Table 5.1: User roles, intentions, and privileges for access control. Adapted from [20]

5.3.2 Dynamic Aggregation as a Privilege

Rather than publishing a view of the entire dataset with some values substituted with value ranges or a dataset of high-level aggregated statistics, individual queries can be analyzed at the time of request, dynamically, and compared with the privileges of the user role requesting along with previously granted requests to measure the granularity of cumulative query results and determine if the information revealed exceeds a privacy budget for that user. We refer to this strategy as dynamic and measured aggregation, which can be used as an alternative to differential privacy in some cases. The workflow is very similar to systems enforcing differential privacy [34, 35] in that a system analyzes the query request and manipulates the results to satisfy privacy requirements.

The query can be answered, denied, or answered in part by reducing granularity. Comparing new requests with information the user already has been granted is a form of privacy budget, another concept from differential privacy[19]. The point when a user is no longer able to get queries answered represents when sharing any more information would increase the probability of violating privacy of individuals in the data set beyond a threshold connected to the privacy budget. The privacy budget or aggregation strategy can be tuned to be more

or less restrictive depending on the user and the data. In addition to the privacy budget, the aggregation strategy enforces rules, such as information concerning groups of fewer than n individuals will not be exposed by query results or combinations of query results.

5.3.3 Using a system to run analysis on behalf of collaborators

HIPAA Privacy Rule includes a De-identification Standard[1] allowing for two methods of de-identification. The first is, *Expert Determination*,

> "(1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
>
> (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
>
> (ii) Documents the methods and results of the analysis that justify such determination ... [47]"

The second method, *Safe Harbor*, is to remove 18 types of identifiers, 14 of which would be directly linked to specific individuals such as names and phone numbers, and the other 4 being related to too specific locations, too specific dates, device ids, and URLs. This method is still subject to the impossibility theorem from differential privacy referenced in Section 5.2.1.

Interfaces to datasets that help project owners navigate privacy risks while allowing just enough access to potential collaborators to validate or expand on original research would be a major improvement over personally selective sharing of static views of the data. A system can accept queries and analysis code, access the original database on behalf of the collaborator, and calculate how to return output based on aspects of what the query is requesting, who the collaborator is along with his or her relationship to the dataset, a log of previous requests, and metadata about uniqueness of certain values and sensitivity of certain features in the data.

Comprehensive data management platforms have also been proposed [44, 20] for shared access to data, tools, and workflows. Provenance that enables reproducibility is a key benefit of using systems that necessarily keep track of users, requests, and ideally versioning of data and processing code too. Privacy preserving data analysis systems can be validated for a variety of circumstances and become familiar to institutional review boards and policy writers, encouraging higher standards and more consistency.

[1] Where de-identified means, "Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information [47]"

5.4 Example Implementation and User Feedback

5.4.1 EEG Labeling Platform

In this section we describe how mechanisms of balancing information and privacy and improvements over current data sharing practices can be implemented for a crowdsourcing system for research of EEG annotation. This is a proposed system for professionals and professionals in-training to annotate EEG signal data, to learn about the accuracy of annotations, and simultaneously to build up an internationally accessible research database of labeled EEG data that can be used for development of clinical decision-support algorithms when analyzed along with patient data. We use the terms label and annotation interchangeably for this use case.

The EEG labeling approach and prototype of the system was presented for the first time at AMIA 2018 [21]. We formulate an approach to implementing access control and describe privacy considerations based on several conversations with William Bosl and Andrew Nguyen, the system's creators who are also well established within the target communities of EEG professionals, neurologists, and health informaticists. First, William Bosl motivates and contextualizes the system as follows[2],

> The motivation for crowd-sourcing and annotation research in our case comes from a traditional use: that of reviewing continuous EEG readings in an ICU setting (cEEG-ICU). Continuous EEG monitoring is a relatively new and growing practice used mostly in larger medical centers. The challenge is that busy neurologists and even neurodiagnostic technologists do not have time to adequately review the data as often as medically necessary. A neurologist typically reviews the EEGs every 12 hours, and a technologist may do a brief review every few hours. This is not often enough to catch emergency seizures that occur without any clinical signs ("electrographic" seizures).
>
> As such there is a rather urgent need for algorithms to screen the data continuously and, ideally, send an alert when something appears to require more serious attention. In this case, the algorithms need to continuously monitor the EEG data streams to detect the signal features that a neurologist would be searching for: spikes, slowing, rhythmic activity. These are known to be indicators of seizure activity. (An aside: for mental disorders, no visible EEG features are known, so annotation by humans is not relevant, or possible.)
>
> In order to begin to develop and train algorithms for this task, annotated data must be available to researchers. The same problem arises again. A neurologist colleague at Beth Israel Deaconess Hospital in Boston has cEEG data from several thousands of patients. Unfortunately, neither she nor her residents have time to go through such large amounts of data and annotate it. They would need a large grant and dedicated research staff just to annotate. Furthermore, if this data will be used for research, the annotations must be reliable. How reliable? How reliable is a resident doing the labeling? How reliable is a trained neurophysiologist with

[2]Quote is taken from personal correspondence and used here with permission

> 10 years experience? None of these questions have been adequately addressed. This is the primary driving force behind the need for annotation.
>
> In addition, the emergence of a new generation of EEG devices that are easier to use *and* lower in cost will absolutely bring EEG into community and primary clinics in low-income regions. In many places, epilepsy is not treated, not because low cost AEDs are not available (they are), but because a qualified neurologist is not available to review and diagnosis epilepsy. Antiepileptic drugs can have powerful side effects, thus should not be given unless warranted. The ability to use algorithms to screen for epilepsy in the LIC settings would be very beneficial. Having annotated EEG data to train screening algorithms will be necessary. The goal in this case is to fill in for the lack of highly trained professionals for epilepsy screening in order to plan appropriate therapy or medications.

Individuals interacting with the EEG labeling system fill a variety of roles: those who contribute EEG data to be labeled and potentially some form of corresponding clinical data, those who annotate EEG signals either for training or to contribute expert annotations, those who evaluate accuracy of collections of annotations and labelers, and those who leverage the repository of labeled EEG data and corresponding clinical information in research and algorithm development. Table 5.2 assigns these stakeholders roles from Table 5.1. These user roles enable sharing data based on the intention of usability in a way that is efficient for data owners. The alternative of making one or maybe two forms of the data openly accessible to satisfy open data policies often fails to be useful to those who discover the data set, as we learned from our interviews described in the Perspectives on Data Sharing chapter. Another common alternative is to grant access to all or most of the data on a case by case basis, which has the drawbacks of being time-intensive for the project owners and less discerning about who sees what and how much. Multiple roles are assigned to some individuals who interact with the EEG labeling system. For example, the project owners also evaluate accuracy of collections of annotations and labelers. They will interact with the system using whichever role is appropriate for the task. Project owners will typically engage with the system as a researcher with limited access to collaborators' raw data, unless an issue with the system itself requires higher order permissions to solve. Data use agreements will determine what under what circumstances project owners may invoke which roles. For this project, there are two types of researchers: those studying the annotation process and those using the annotations. The former does not involve clinical patient data; they are analyzing labels and personal data about labelers. Those using the annotations are analyzing the labels along with clinical patient data. As such, there would be two implementations of the "colleagues/ researchers" role in which the permissions would be similar but applied to different subsets of the data.

The EEG labeling system inherently requires keeping track of the identities of the labelers and evaluators, with labelers intending to practice or contribute annotations and evaluators analyzing the accuracy of labelers' annotations. Privacy considerations then extend beyond managing patient data within the system to how data and performance metrics about the labelers is handled. A labeler needs to view discrete windows of visualized EEG signal data and contribute new annotations, which can be satisfied with simple read and write

Stakeholder description	Role
Those who contribute EEG data to be labeled and potentially some form of corresponding clinical data	Potential collaborators
Those who annotate EEG signals either for training or to contribute expert annotations	Participants
Those who evaluate accuracy of collections of annotations and labelers	Colleagues/ researchers
Those who leverage the repository of labeled EEG data and corresponding clinical information in research and algorithm development	Colleagues/ researchers, or other researchers/ the public
Those who set up and maintain the EEG labeling system	Project owners
Those from whom the EEG data was collected	Participants
Those who have funded and will evaluate the EEG labeling system and surrounding studies	Funding agencies

Table 5.2: Stakeholders of the EEG labeling platform and their corresponding user roles.

permissions. Someone evaluating accuracy needs to analyze and compare sets of labels with what is referred to as gold standard labels, as well as perform experiments to learn about inferring correct labels from group consensus and how to weigh labels based on attributes of labelers.

In a system meant to be widely used by a large research community, it would be reasonable for labelers to expect only certain people to have access to their labeling history and other attributes meant to quantify their personal expertise such as certifications and average distance from correct labels. Evaluators may also be overseeing professionals using the system for training, so their permissions in the system should account for access to raw data about those relatively few labelers they have a relationship to and access to information about other labels and labelers in the system that is accurate without exposing the identity of other labelers. Not including labelers the evaluator does know, data about any given individual labeler must be indistinguishable from some other number of other labelers to be considered private.

The above requirements can be implemented with query analysis and measured aggregation as a permission for relevant user roles (as described in Section 5.3). Query analysis parameters include the query itself, precomputed statistics and metadata on uniqueness in the dataset, and the identity of the user making the request along with his or her role and a log of previously granted results. The measured aggregation step involves a threshold for minimum bin size, in this case how many labelers must share the same attribute values – i.e. be indistinguishable from each other – for the query results about those attributes not

to expose too much information about any of those labelers. Minimum bin size can be based on trust and sensitivity of the data, and becomes less of a burden to analysis as the amount of data increases.

Data use agreements are still important. Access to auxiliary data cannot be controlled, and therefore any accurate query results cannot be guaranteed to preserve privacy. An evaluator role would not have access to any clinical data in the system, however could be assigned an additional researcher role with permissions enabling clinical research if applicable. Someone with multiple roles can choose to interact with the system using any one role and its associated privileges at a time.

Developing algorithms to recognize atypical neurodevelopment in children involves analyzing their ages in months if not weeks. Age values more granular than years are enough to preclude a training dataset from being shared in certain circumstances without additional and expert privacy measures under HIPAA's Privacy Rule and De-identification Standard. This is where dynamic aggregation as a permission is an invaluable improvement mechanism. Access to the repository of labeled EEG data and even minimal corresponding clinical information can be limited to those assigned a user role based on clinical research that carries more stringent data use agreements and with the measured, dynamic aggregation permission tuned to account for more sensitive data.

Finally there is a user role for interacting with the EEG labeling system by contributing raw data in the form of EEG signal data and, optionally, limited corresponding clinical data. Once the data is uploaded, these contributors should have access to their original raw data along with correct annotations if the system has collected or calculated them. If the EEG labeling system or clinical research database system were to expand to include other built-in analysis tools, contributors would have access to use those within the system with their own data and compare findings with any published results based on this database and these analysis tools. As mentioned above, a contributor might additionally take on a researcher, evaluator, or labeler role.

This use case incorporates basing access on relationships to data, being intentional about how data is aggregated, and having the analysis run within the system on behalf of users rather than disseminating data itself. Privacy components implemented in a system like this one should be evaluated based on what risks are mitigated that might have been otherwise tolerated and the extent to which data sharing and collaboration are expanded without increased privacy risk. In this case, runtime efficiency, ease of use, and scalability are also critical.

5.4.2 Privacy Expectations of Users

To design and validate ideas in the case study implementation, we interviewed and surveyed practitioners consisting of EEG technicians, those who label EEG data in clinical settings, and medical researchers who use labeled EEG data in their research work. Practitioner feedback revealed additional contextual information and beliefs about privacy issues.

The platform for crowdsourcing EEG labels will be used as a training tool and has the potential supplement certification requirements. As such, it will be necessary to capture and selectively share metrics related to how accurately or how often an EEG technician or

technician-in-training labels the signals. On the other hand, use of the platform is highly motivated by an opportunity to practice and improve labeling skill. If performance metrics could inadvertently be used to penalize a practitioner, this could create a disincentive to participate thereby undermining the goal of crowdsourcing many labels. This dilemma can be mitigated by ensuring those contributing labels are able decide to what extent and with whom to share the data generated about their use of the platform. Users being in control of their own data is an increasingly prevalent requirement for data management systems, and effective realization will depend on better and more accessible tools for sharing data while protecting privacy.

An interesting concept was proposed and reinforced in the interviews with EEG practitioners around the country, who meet each other at professional development conferences and highly value the sense of community promoted at these events and in their profession more broadly. As such, technicians may find it worthwhile to be discoverable by other EEG practitioners in order to learn from each other. Interpreting EEG data is influenced by patients' age and situation, for example if they have certain injuries or illnesses. A technician who practices in a pediatric or emergency setting may be happy to assist others who are learning to develop their own skills in these areas. In essence, if the platform incorporated social networking features, this may provide value and incentive as well as promote sharing of certain performance metrics and other personal metadata related to a technician's training.

The overall attitude towards privacy considerations of both EEG signal data and labeling performance was casual and generally open, however multiple practitioners acknowledged "in the wrong hands" the data from this platform would cause problems. The strengths in our approach to managing data privacy of participants' performance and metadata come from three key aspects: 1) we are crafting and implementing privacy and data sharing mechanisms before collecting the data, 2) users will opt-in and opt-out of sharing performance and metadata, and 3) we incorporate robust identity management to reinforce role-based access control.

In terms of the EEG signals in particular, there already exist public datasets with correlating labels and doctors' notes. All respondents described these datasets as posing no risks to patient privacy, in contrast to theoretical definitions from the privacy literature. It is worth considering to what extent current practices may in fact be good enough in some cases. Practical concerns about theoretical privacy have typically revolved around implementation challenges, as opposed to value in practical settings. On the other hand as the amount, complexity, and availability of medical data for research increases, the bar will almost certainly be raised for which privacy strategies are adequate.

5.5 Conclusions

This chapter focuses on trade-offs between sharing data and information and protecting privacy, whether policies address these trade-offs, the motivation for balancing them well, and mechanisms of improvements over current practices. Those improvement mechanisms are user roles, dynamic and measured aggregation, and having a system interact with databases on behalf of users. We demonstrate one use case where these mechanisms will facilitate a diverse research community using a shared system for labeling EEG signal data. These

features can significantly improve both data sharing and privacy protection more broadly, as well as enabling applications for international collaboration.

It is worth noting that for some cases we are comparing "secure" data on the internet vs paper records in disorganized boxes behind unlocked doors. It cannot be assumed that either is automatically more effective at preserving privacy than the other. Indeed the gold standard privacy strategy is based on probabilities, so we would need to calculate the probability of someone being affected in anyway by the fact that their personal, likely medical data is written on paper and stored in a potentially insecure room. Documents may be difficult to find even if an unauthorized person did access the room, and the physical location and material nature of the information are immune to data mining. Even when using paper records, health care providers have been conscious and made arrangements for handling patient information regarding highly sensitive or stigmatized conditions.

While privacy risks are not necessarily mitigated or exacerbated by managing data carefully online, supporting data sharing practices does necessitate leveraging the internet. In an extreme case, some stakeholders would prefer to destroy records than be obligated to share the data. Fine-grain control may help preserve information. The measured dynamic aggregation improvement we describe sacrifices the guarantees offered by differential privacy, although differential privacy would be a compatible access mechanism for a subset of user roles. The trade-off is to enable more applications and stakeholders to implement privacy considerations more rigorously than the next best strategies or other common practices.

Chapter 6:

Participation and Development Context

6.1 Introduction

This chapter attempts to situate data management technology, especially tools for collaboration and sharing data, politically and ethically within a context of international development. We do so by comparing concepts in critical development literature with empirical evidence from our revisited interview results, and applying insights from the literature to the tools and improvement mechanisms we proposed in earlier chapters. In the next section, we make the case for why this is an important contribution in the consideration of developing collaboration tools. The Methods section describes how we approached our exploration of the role of data management and sharing tools in international development. The Challenges section then specifies barriers to equitable collaboration in this space, followed by the Discussion section on recommendations from our literature selection applied to tools for working with data.

6.2 Motivation

International aid and loan agencies are embracing the use of data-driven development and evidence-based decision making to determine how to distribute funding, and are investing more heavily in the development and deployment of data-intensive interventions. Why? Funding agencies and data-savvy developers already hold the power to make decisions based on their own perspectives and priorities. Incorporating data and statistics does not necessarily bring about more objective decision making and increased investments in more positively impactful initiatives, even if it gives that appearance. Data and statistics are used to report on metrics that reflect the goals of those who develop them. Sally Engle Merry, in her book *Seductions of Quantification*, reveals insights into metric development including whose values and priorities they represent [48]. Data and statistics necessarily leave out some context, and it may be context that other stakeholders who do not have power to influence what data is collected would have included. Data and statistics bring about more efficient monitoring of metrics, but the metrics are biased in a way that reinforces existing power dynamics by lending credibility to the decision making processes. By extension, data management infrastructure and tools will only serve to reinforce existing power structures if they only enhance existing workflows. Existing workflows often fail to incorporate voices of the stakeholders with the least amount of power in any meaningful, equitable way.

This chapter aims to understand a few of the significant barriers to equitable participation among stakeholders of data-driven development, and–in an effort to avoid reinforcing them–how these barriers manifest in data management tools and practices. Within the literature

on technology and development, we then identify several lessons and propose how these lessons relate to creating new data management and sharing tools that can do a better job of supporting more equitable collaboration. We believe it is important to form a critical awareness of the limitations and potential of new technologies, especially when consequences manifest in potentially vulnerable populations.

There is no shortage of examples of how development interventions fail or have unintended negative consequences, especially when practitioners do not engage with the reality and agency of intended beneficiaries, or worse blatantly disregard the human dignity of people in the communities where they work [49, 50]. We do not presume there is a purely technical solution for these types of problems. Rather, we recognize technology can either reinforce existing inequity of processes or reduce the effort needed to restructure when project leaders want the work and outcomes to be more equitable.

6.3 Methods

Chapter 2 describes methods used for collecting and interpreting interview data for the purpose of understanding professionals' perceptions of the value and risks that go along with sharing data, and the implications of those perceptions on the tools we develop. For this chapter, we revisit the original transcripts with newly articulated criteria for what is important and interesting. We then bring the interview results from this perspective into a conversation with the literature on theories of technology and development practice.

The filter for what is important and interesting to this chapter is anything that speaks to relationships or collaboration among stakeholders. Within conversations about data management and sharing, comments on relationship dynamics among stakeholders speak to power differentials, who makes decisions (based on what and with whose input) and who is subject to the decisions that are made, and to what extent people realize and consent to how data about themselves is used.

The literature selection we lean on to create a conversation with our interview results includes concepts from development theory, technology and society, and the abstractions and simplifications involved in the use of quantitative data. The authors we choose are more often critiquing development, technology, and quantitative reasoning than uncritically celebrating them. Tania Li draws a clear distinction between development practitioners (like most of our interviewees) and development critics [51]. In a talk and question and answer session about her book [52], she makes the point she does not venture into solutions in *The Will to Improve* because to do so would require a different approach to the whole project that built up to those solutions. She suggests one person can not be both a development practitioner and critic at the same time, although they may go back and forth. Development critics ask, "What is wrong with this picture?" while practitioners ask, "What good can we do, given that this is the picture?" These are difficult to answer at the same time–it is difficult to challenge the picture while at the same time intervene in a taken-for-granted picture. In addition, practitioners might not be able to stand far enough back to see the whole picture, while critics may miss out on key insights by not being involved in the work first hand.

At a high level, we asked our interview participants to talk about how they worked with other stakeholders, not how they judge these working relationships. As such, if they were in

a position to critique data management and sharing practices within their own development practice, it is possible we did not ask about it directly enough. We did ask about what participation looks like and about limitations of going about their work, however the topic of data management and sharing was made clear in recruiting. Erin Cech writes about the culture of engineering education and professions, and she draws a technical/social dualism whereby engineers are prone to detach technical and social considerations [53]. It is possible our interview participants were inadvertently primed to focus on technical aspects when asked general or open ended questions, even though our intention was not to separate the social and technical. Nonetheless, what we learned from our interview participants is informative to our critical analysis of the data tools we design and build for development practitioners in roles like theirs.

The critiques we bring in from development literature are helpful as we aim to examine data management tools and validate features with a new lens on equity. At the heart of this chapter is an acknowledgement that international development interventions have a checkered history, and the project of development itself is rooted in the Western colonial enterprise that sought to refashion the world according to the model of the West. With this in mind, we under take this effort to develop a critical awareness of the limitations and potential of data management and sharing tools and practices within such a context. Developing data management and sharing technology for international development projects may only reinforce problematic dynamics and power structures if it neglects to incorporate a fully considered view of all the stakeholders in an equitable way. Often the objectives of our interviewees in talking about improving data practices had more to do with research outcomes and efficiency of their work than stakeholder relationships or power dynamics, however it is important to remember they all operate within the heavily politicized field of energy and electrification.

6.4 Challenges

6.4.1 Participation

Duraiappah et al. describe degrees of participation ranging from manipulation and passive participation to partnership and self-mobilization/active participation (see Figure 1), that fall into one of two categories: either the functional/passive perspective or the rights-based/proactive perspective [55], based on similar previous work on citizen participation (originally [54]). The degrees of participation are in order of most passive (1) to most proactive (9). The functional/passive perspective situates participation as a means of collecting information to inform, measure, or justify an intervention. Duraiappah et al. show how rights-based proactive participation is more effective if the goal is to increase beneficiaries capabilities and freedoms[55, 56]. Effective participation "involves a shift in power over the process of development away from those who have traditionally defined the nature of the problem and how it may be addressed (governments, outside donors) to the people immediately impacted by the issue" [55].

In our interview conversations, the most proactive examples of participation had characteristics of degrees 5, 6, and 7 of participation–participation for material incentives, func-

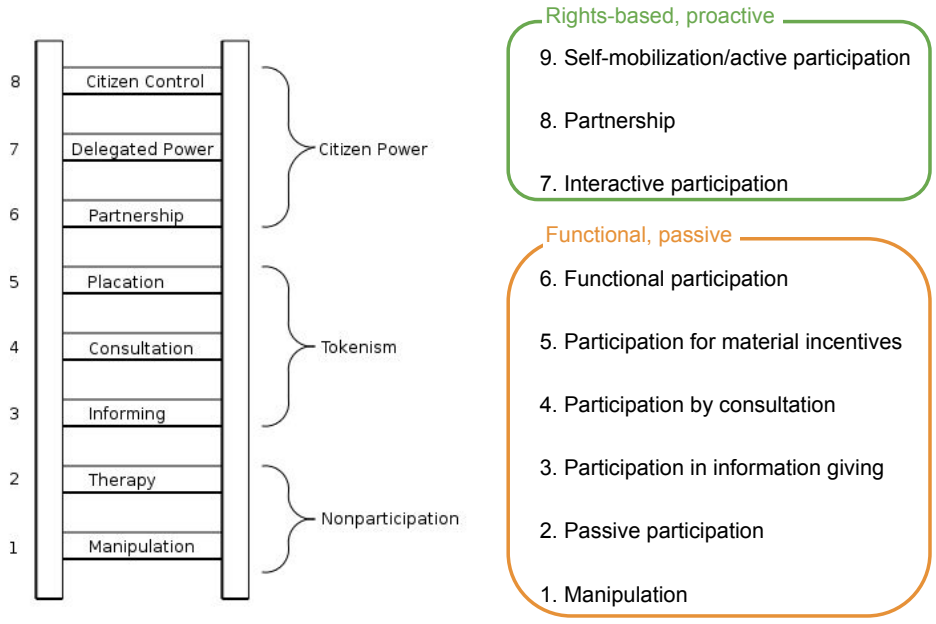

Figure 6.1: The original Ladder of Participation by Sherry R Arnstein [54] (left) and degrees of participation by Duraiappah et al. [55] (right).

tional participation, and interactive participation respectively–as described by Duraiappah et al. on the right side of Figure 1. Community members were involved in decision making about how to implement and manage the intervention and eventually took over control, which was an original objective of the project. The majority of our conversations on participation reflected degrees 3 and 5 of participation: participation in information giving and participation for material incentives, with some characteristics of degrees 2 and 4: passive participation and participation by consultation. From what we heard about community participation, projects fall short of degrees 8: partnership and 9: self-mobilization/active participation in that ideas and decisions were made ahead of time by project leaders rather than coming from within the communities.

Since the article *Have Participatory Approaches Increased Capabilities?* by Duraiappah et al. was published in 2005, sensors have become more prevalent in monitoring and evaluating development interventions including electricity, water pump, and cookstove usage. Sensor data has the advantages of mitigating politeness and memory biases, and reporting at more frequent time intervals. Our interview participants noted using this data along with observed behaviors such as using or deleting a corresponding phone application to learn about and adapt implementations. Duraiappah et al. emphasize the effectiveness of participatory approaches depends on how the participation is set up. If sensor data displaces community feedback and engagement, which our interview results suggest happens, participation is in effect made more passive. The next subsections on simplification, categorization, and information basis reveal how relying too much quantitative data can be problematic.

Given the nature of how we chose interview participants, it is unsurprising none of the projects are examples of community self-mobilization. What is more noteworthy is the extent to which participation is viewed as a means of collecting information for planning and validation purposes, as opposed to sharing power, decision-making responsibilities, and derived information. For example, in the case of electricity suppliers and consumers as well as other types of projects, troubleshooting is handled on a case-by-case basis and if feedback is collected in the process, it is not analyzed as a whole. Typically communities answer surveys and then survey data is used as a baseline for impact analysis or the survey responses are used to plan the intervention. In exchange for their responses, community members or consumers are compensated with money, energy credits, or raffle entries for appliances. We also observed a disconnect between those conducting the surveys or offering technical support and researchers or project organizers, such that the channels for community input seem limited. For the most part, communication is one-way at a time: survey responses are collected, and only sometimes data is shared back with communities and individuals. If there were multiple rounds of one-way communication and communities had opportunities validate or scrutinize results and findings, this would not necessarily be problematic. One-way communication can serve as a barrier to equitable participation when the channel of communication between the consumer or intended beneficiary and the implementers is set up for the primary purpose of troubleshooting, and their input is not analyzed or integrated back into decision making processes. This kind of setup for communication appears to be common. Our interview participants reported if energy consumption data is shared back to individuals, it may be in a manner that is helpful for them to understand their electricity consumption patterns or it may be shared in a more processed form such as a suggestion to

change behaviors to save energy. From our interview conversations, survey data in any form is less likely to be shared back with the survey respondents especially if they are participating in an ongoing research study.

Other interview participants were speaking from their perspectives within organizations doing work in the energy sector although not directly with individual consumers or communities, in which cases participation among companies and researchers reflects partnership towards common goals or shared understanding of mutually beneficial arrangements. While shared goals and mutual respect are components of the proactive forms of community participation [55], the framework is not meant to be applied to collaboration among businesses and organizations. Even then, projects almost always collaborate, i.e. make their data available to other researchers in exchange for analysis contributions, on a case by case basis with very little formality to the process other than standard non-disclosure or data-use agreements. Such discretion is at odds with the accessibility goals of open data and open science initiatives. Researcher aspirations we learned from interview conversations include understanding electricity consumption patterns for the sake of understanding as well as to inform system design, and influence policy such that certain types of development interventions receive more prioritization and funding. Consumers of electricity on the other hand, although we did not interview them, presumably have their own goals and motivations independent of any specific initiatives and interconnected with other goals of well-being for their communities.

Interview participants used language around "balancing" stakeholders' preferences and making sure all perspectives were included. The literature differentiates between partnership among equally respected collaborators, participation of target communities in a marginalized or subordinated role, and exclusion [57, 55]. For example, consumer input can be gathered and then used for more strategic approaches to financing, optimizing appliance sales in newly electrified neighborhoods, or for the sake of forging acceptance rather than responding to concerns of the individuals. Such cases demonstrate how all perspectives can be "included" without significant progress in the direction of equity or collaboration on equal terms. It is also unclear whether consumers understand how their data will be used given the exploratory approach to how data could be valuable echoed in our interview conversations. That said, several of our interview participants expressed interest in having more transparency around their own data management practices and in the field more broadly.

6.4.2 Simplification

In quantifying social issues, we are simplifying a complex reality, abstracting it into the things we can measure and categorize. Simplification and categorization are related concepts from the literature contributing to our understanding of why quantitative data is necessarily partial. It is both incomplete and favoring the perspective of its creators. The data analysis process is always one of simplification. Summarization and aggregation by definition leave out information and context, which can influence understanding of results. The ethical issue with simplification is who decides what information gets left out, or put another way, who decides what information counts or does not count in understanding problems. Scott describes in detail the motivation for, meaning of, and implications of the abstraction and what he calls legibility of societal issues [58]. In order to address societal issues at some

scale–as opposed to on a case-by-case, individual basis–they must be abstracted in some way. The problem is that the more abstracted an issue becomes, the more diverse communities are inaccurately considered as a homogeneous group.

Anonymization can also be a simplification process if it involves simply removing identifying information. There can be valid reasons for withholding data related to age, gender, and race but these features are related to social determinants of health. Likewise, location data can be linked to environmental conditions. Simplification of problems and data collected about them reifies simplified data as though that is all that matters. Context that might be essential from an alternative perspective–the one with less power–is lost completely because it is either not measured in the first place or because it is removed in analysis or anonymization. Then this simplified data is then used to make decisions about funding and intervention planning.

An example of simplification that came up in interview conversations was how to evaluate energy and electricity products, whose performance can be measured in different ways and depend on different circumstances. Companies will sometimes do the their own performance testing and make an effort to explain the measurements they choose, but if the tests are customized to their products and contexts of interest it is difficult to compare with other similar products. While standardized performance metrics for energy products can enable direct comparisons, the comparison loses its utility if the conditions of testing and measuring are not also captured and documented. For example, efficiency of solar panels is a useful metric and is also sensitive to weather conditions including temperature and solar insulation. Even maximum efficiency as a measurement for useful comparison needs explanation: is it theoretical or based on testing, does it degrade over time, what are the trade-offs if maximum efficiency of one product is better or worse than an alternative product, is it valid to compare the maximum efficiencies of solar panels and batteries, and so on? A danger of widely accepted standardized metrics is the incentive to optimize products to do well on tests as opposed to be better products in the way the metric intends to measure[1]. A possible remedy involves using both standard and custom metrics, and prioritizing the explanation of testing and measurement conditions.

Simplification issues related to technical specifications are all the more applicable when attempting to measure and describe human conditions and affairs. Problems arise not when quantitative data is collected and analyzed and used to make decisions, but when it is interpreted as representing the whole truth without context and explanations–in other words without qualitative data. This is not to say qualitative data is more honest. Qualitative data can fail to capture important information such as scope and precision [48], for example how many people lost power, where, and for how many minutes. Both quantitative and qualitative data are necessary to mitigate problematic consequences of simplification.

6.4.3 Categorization

Categories are constructed in the process of designing data collection schemes, again in data analysis, and again in data sharing: which data is private, which results are relevant, who are useful collaborators and who are not? Whoever chooses survey questions and de-

[1] As was the case with the Volkswagen emissions scandal in 2015

signs sensors, writes data cleaning and analysis code, presents numerical results, and serves as a gatekeeper for the data and processing code, has their unique perspective deeply embedded in what is generally seen as an impartial, if not infallible, process. Merry illustrates categorization as inherently cultural work,

> "Even simple counting raises three questions: (1) What is important to count? (2) What characteristics are diagnostic for identifying these countable things? ... (3) What are the appropriate criteria for aggregation and disaggregation? Counts require cultural work: they depend on constructing categories such as gender, ethnicity, income, and employment status. Creating categories implies deciding on where to lump and split, what to include and what to leave out, how many categories to use, and what the criteria for these categories should be" [48, p.14].

That counting requires cultural work is another way of saying the constructed categories are not universally applicable, and therefore vulnerable to bias. Categories are designed by humans at the point of data collection, and again at the point of clustering or aggregation. The decisions regarding which categories to use and how to use them have underlying value judgments made by someone or some group and may not reflect the values of other stakeholders.

Multiple interview participants spoke of automating some of their data analysis steps involving such categories by using machine learning to narrow surveys down to only the questions correlated with a certain metric of interest (e.g. energy consumption), filter data for the sake of keeping what is interesting, aggregate multiple sources of data, or remove personal information before data is exposed to employees or shared externally. What does it mean to automate inherently cultural work? We suggest such automation can be used to mitigate human error by ensuring data is thoroughly filtered for some tasks, and reinforce biases by validating chosen categories for some other tasks.

Added to the question of whose truth is represented by constructed categories, and whose is left out, Tania Li points out how simplifiable and categorizable projects will receive preferential funding from agencies who require proof of impact. This limits the focus on development in general, even as funding and attention to goals are increasing. Tania Li refers to this phenomena in development as doing more about less [51]. The focus on impact also disincentivizes honest reporting[2].

6.4.4 Interpretation chains and information basis

Sally Engle Merry in *Seductions of Quantification*[48] describes the concept of interpretation chains, which is related to Amartya Sen's work explaining the information basis for evaluation, or how values can be inferred from what data is collected and how calculations are made in his book *Development as Freedom* [56]. Merry describes different types of indicators including counts, ratios, and composites. Composites are more widely known, and a good example is the Human Development Index (HDI) because it is compiled based on

[2]In fact, we are familiar with a case of data collection infrastructure being removed all together in order to hide bad results, sacrificing the value data provided to the project in order to prevent critical evaluation and being held accountable. Similar efforts to sabotage data systems have been reported [59, p. 111]

several weights and measures related to what the creators of the HDI believe contributes to human development. Counts and ratios on the other hand are more closely linked to raw data but also less popular. Composites sometimes involve calculating scores or ranking countries. An interpretation chain is made of all the steps involved between the raw data and the resulting calculated metric, and emphasizes each step as an interpretation whether it involves creating categories or weighting. Each time the data is manipulated along the way involves interpretation and underlying values, captured by what Sen refers to as the information basis of analysis. What information is taken as relevant to the calculation depends on the analysts and the metric creators' values and underlying ethical theories. The example Sen uses relates to choosing one out of three options related to increasing happiness, income, or ability [56]. The decision exposes which information is relevant to decisions, and which information is relevant as well as how that information is weighted can be subjective.

Our pipeline model for sharing analysis would readily illustrate how distant a metric calculation is from the raw data, and could be used to make clear which information is filtered as well. For example, a composite metric will involve several more operators and a more complex processing pipeline–if not several pipelines–versus a count or ratio that only needs a few filters and statistical tasks. Composite metrics are often compelling because of how simple they seem, whereas a pipeline to calculate it would reveal its complexity. Filtering operators in the operator library also require parameters to be specified at the time of adding operators to the pipeline graph in the interface explained in Chapter 4, such that names of fields being selected and how those fields are manipulated can be determined by observing the visual pipeline graph.

Some insight around interpretation chains came up in our interview conversations. The first is related to the timeline of typical projects: one consultant may collect data for two or more years and generate a static report that is used to make decisions. The opposite was proposed: multiple sources of data gathered and fed to a live model with automatically updated outputs, on which to base decisions. Identifying the interpretation chains in these two extremes is a helpful exercise. The lone consultant likely understands each subtlety and manipulation of their process, although they may neglect to explain them well or have incentives to select analysis results that encourage continued funding. The alternative proposed essentially automates steps in the interpretation chain, and in so doing potentially exposes more broad data sources and more timely results. At the same time, whose interpretation is woven into these results is blurred. Machine learning was being considered by other participants to identify which survey questions were most correlated with energy consumption, such that they could shorten their surveys. It may be worthwhile to shorten surveys, although plausibly this reduces reusability of the data for learning any other related trends not directly correlated in the data and again brings up the question of whether it makes sense to automate this kind of interpretative, cultural work. Fairness, accountability, and transparency in machine learning (FAT ML) is a relevant area of research and also out of scope for this chapter[3].

[3]See https://fatml.org/

6.4.5 Cost-effectiveness vs human rights

There are costs in terms of time and money to figure out how to satisfy open data requirements or implement privacy strategies, creating an ethical tension between what is most cost-effective and efficient versus what is transparent and respectful of human subjects' right to privacy. There can be noble intentions behind the push for openness as a means of development projects learning from each other and doing better work, however several participants pointed out this step of figuring out how to publish data is often postponed until the end of a project and done as an effort to satisfy the openness requirement, not to optimize the data for reusability and learning. Multiple participants had significant doubts about the utility of the open data published in this manner. As a result, the overall work is actually less efficient towards its goals. This is a great place for new tools and strategies to help facilitate improvement.

A case study in *Pathologies of Power* by Paul Farmer explores arguments made for "cost-effectiveness" versus human rights [60]. Expensive becomes synonymous with impossible in resource-restricted contexts, until the issue is out of control and ultimately ends up costing even more to regain control over the situation. Without conflating economic and social human rights with data privacy rights, we notice some parallel reasoning playing out around data sharing.

Some participants operating small businesses spoke about restructuring to satisfy changing data regulations as cost-prohibitive. For grant-funded projects, some funding agencies although not all offer additional funding for open data purposes. Still other organizations represented by our participants continually evaluate and update their data sharing, protection and privacy protocols to keep up with best practices of their own volition.

When asked about features of technology that promote benefits of data sharing or mitigate risks of data sharing, some participants suggested not sharing data at all if it threatens privacy rights of the data subjects and another plurality of participants pointed to the fact that data serves as a source of income in many business models so people ought to be able to pay for the service as an alternative to allowing their data to be used. This begs the question of whether privacy is a human right or a privilege available for purchase. Our belief is fine-grained control over what data is used for which purposes and by whom can protect privacy and maintain the value of its analysis at the same time. The interview conversations remind us of the market forces–i.e macro-ethics–influencing how data is managed, used, shared, and how new tools and alternative practices are perceived.

6.5 Discussion

In thinking about solutions to reduce barriers to equitable collaboration described in the previous section, we catalogue recommendations for development practice and discuss how they relate to the data management and sharing tools considered throughout this dissertation. Challenges to equitable participation among stakeholders can be exacerbated by limited data management tools and little structure around deciding who is given access to what data and what steps in the data analysis.

In her article, *Injustice at intersecting scales: On 'social exclusion' and the 'global poor'*

Nancy Fraser notes three conditions that must be met in order to have parity of participation: distribution of resources is equitable enough for people to have independence, recognition in terms of social status and equal respect, and representation such that everyone has a political voice [57]. The first point we can make is even if data management and sharing tools were designed for perfectly equitable collaboration, mutual respect and adequate distribution of material resources are not necessarily in place. At the same time, for stakeholders to participate in monitoring, evaluation, and decision-making processes requires representation and voice within systems [57]. Role-based access control for interacting with datasets differently depending on stakeholder groups can illustrate which stakeholders have the most extensive permissions along with who is left out. Our proposed role-based access control list would give consumers, participants, or beneficiaries access to their own data and permission to learn about the population in the data as a whole, which is slightly more privileged access than stakeholders who intend to learn about the population. Tools that support transparency in general, such as our pipeline builder, can also reduce the degree of exclusion from other data processing steps by opening the process to feedback. These are improvements over ad-hoc and case-by-case data and code sharing, although the conditions for parity of participation related to resource distribution and respect must also be met.

Similarly, Duraiappah et al.'s suggestion for development practice is to, "re-orient the thinking of development experts from being implementers to facilitators" [55]. Moving projects from participation in the form of manipulation and collecting information to proactive degrees of participation reads as a shift in values, although some concrete recommendations can be extracted too. In particular, one-way communication channels can be improved by setting up two-way communication such as more formal and accessible feedback and reply mechanisms. Data management and sharing tools in general reflect the relationships they're set up to facilitate among collaborators, which can be more or less equitable. This applies to the tools we describe in previous chapters as well; the pipeline builder or privacy strategies can be implemented to serve existing power structures or more equitable ones. As with all new technology, the default is to reinforce existing power structures.

In *Development as Freedom*, Sen also brings up a critique of and tension in development studies. The critique says development can have a net negative impact on communities when it results in the loss of traditions and culture, and the tension is disagreement around whether to assess the value of the economic growth as worth it or not. Sen points out participation (and public scrutiny) is central to such valuation, and it matters whose perspectives are taken as authoritative and legitimate [56]. The key limitations of making data analysis pipelines more transparent are questions of technical and general literacy, and connectivity. The work becomes more transparent to those with certain privilege. Still, we believe enabling more transparency is an improvement. Development critics and outside practitioners working in related circumstances may be well-positioned to critique results and information bases of interpretation chains if given access to properly-documented pipeline graphs independent of whether data is also published.

In *Seductions of Quantification*, Sally Engle Merry emphasizes three recommendations. First is the importance of keeping track of how an indicator has evolved including who has created it, what their expertise encompasses or leaves out, who has funded the work of those experts and the data collection, what organizations are involved. Similarly, it is important

to understand limitations of indicators with long interpretation chains. For indicators to be comparable across projects and countries, "Categories must all refer to the same thing, even though that thing is manifested differently in different places" [48, p. 214]. As such, composite indicators with long interpretation chains have the potential to impose convoluted weights on different features of data from different places. Merry also makes the case for using both qualitative and quantitative data in all analyses.

The ancillary purpose of role-based access control lists for keeping track of who is involved and has access to what data fits well with the attention reviewing outcomes in terms of which organizations and experts have contributed. Pipeline graphs similarly help visualize interpretation chains, especially distance from input data sources, which is well-suited for reflection on them. What can appear to be simple scores and rankings would be associated with more complex graphs. Merry points out, "The complexity of the processes described by the black box renders it far harder and more time-consuming to dispute. Thus the reader, faced with the difficulty of challenging the interior of the black box, is carried on to accept it" [48, p. 30]. We imagine the pipelines as a directional graph of several black boxes, which is conceivably easier to challenge than one big black box, especially if some of those boxes are used, accepted, and understood elsewhere.

In describing how planners' analyses typically fail to capture the "radical contingency" of the future and human affairs in *Seeing Like a State*, James C. Scott makes the following recommendations for development practice: take small steps, favor reversibility, plan on surprises, and plan on human inventiveness [58].

In favor of taking small steps, Scott says, "presume that we cannot know the consequences of our interventions in advance." Our interview participants unanimously expressed uncertainty about how to best protect privacy and share data (with varying degrees of importance on sharing or privacy). As such, taking small steps applies as much to the interventions as to practices around sharing data. The risks associated with over sharing for the intended beneficiaries is that exposed sensitive information, potentially relating to health or employment could cause them harm. The risk to the project owner is generally reputation, which could be significant although less personal. Typically only the latter helps decide how the data is published. Interview participants who made or were in the process of making data public to satisfy open data policies of funders noted a particular emphasis on maximizing how much data could be shared without much support or guidance for how to protect sensitive information beyond acknowledgement of its importance. Withholding data from public release for any reason usually requires a special request that the funding agencies judge as valid or not. Taking small steps in sharing data would be to err in the direction of protecting privacy. As for taking small steps in interventions, data can serve to keep track of lessons such that the steps are more informed, or data can encourage an illusion of how well we can anticipate outcomes and consequences such that bigger steps seem more reasonable. A sense of urgency to deliver interventions runs counter to taking small steps. Unless participation reflects the most proactive degrees, the potential for misunderstanding exacerbates the self-evident contingencies of planning for an unknown future.

Favoring reversibility in data management makes a strong case for granting access to a system that runs analysis on behalf of collaborators as opposed to sharing a data set for them to analyze on their own machines. Unexpected access or download patterns can trigger

warning flags and permissions can be revoked if data use agreements are violated. Downloads can be prevented or audited. Once data is made available on public-facing repositories online, the data is forever public. Despite a famous study on how to re-identify Netflix competition data [61], the dataset is still readily available on the third-party site Kaggle where it has been viewed over 115,000 times and downloaded over 16,000 times at the time of writing. Some interview participants expressed interest in granting access to data through an API, which could be an alternative method of filtering access if proper authentication and logging mechanisms were also implemented.

Scott's recommendations to plan for surprises and human inventiveness advocate for designing flexibility around the intervention itself and in terms of leaving the process open to future contributions from those involved. Energy data is often analyzed to better understand consumption patterns and optimize systems to support those patterns. Embracing flexibility would favor analysis of the breadth and distribution of behavior patterns in an effort to design systems that support variety and change. All of our interview participants spoke to learning significant lessons as their projects move forward. Data management and sharing tools can support flexibility by being adaptable themselves, and facilitating collaboration. We discuss design requirements for collaborative data management in Chapter 3, although our considerations did not specifically account for collaboration over disparate time periods, which would be an interesting additional contribution.

6.6 Conclusion

Improving technological international development projects in the direction of equitable collaboration requires effort both to gain critical awareness and to restructure planning and evaluation processes. None of the tools and strategies we propose disrupt existing power dynamics or automatically empower any stakeholders who were not already in a position to make decisions. The conversation contextualizing international development technologies has political stakes: investments of time and resources, along with ethical considerations, involve significant trade-offs. We value equitable collaboration highly and also acknowledge how difficult some of the barriers we describe will be to resolve. A data management and sharing system that is disruptive of current system would almost certainly not be adopted by project owners, who are responsible for maintaining fine-grain control over their data and workflows. The same project owners, especially those we interviewed, value transparency and desire their projects to have positive impacts.

Some of the tools and strategies we propose in previous chapters enable more transparent provenance. Keeping track of steps taken to reach conclusions from data analysis is as helpful for reproducibility as it is for opening the process to critique and feedback. Without transparency in the process, projects can and do hide key steps which may jeopardize their reputation, funding, or preferred modes of operating. Being able, and potentially expected, to reveal steps along the way creates a historical record of the work and can introduce some pressure for good behavior. That said, reimagining how quantitative and qualitative data can be applied more effectively to increase our understanding of complex human issues and better direct funding to the most positively impactful projects requires more thought and effort. Moving from passive and functional participation to proactive and rights-based

participation in such projects also requires consideration beyond the features of tools for working with data that may help facilitate it.

Chapter 7:

Summary and future work

7.1 Summary of Findings

This thesis considers technical tools and mechanisms for collaborative data workflows within the context of international development engineering (also referred to as development engineering), especially given the tensions between open data initiatives and privacy concerns.

We first document relationships of multiple stakeholders to data and each other, and share interview results from conversations with the stakeholder groups who work most directly with data. Compared with other data sharing studies, our interviews uniquely focused on electrification and energy access researchers and practitioners. We find interesting consistencies as well as variance in perceptions of the value and risk of data sharing. We then articulate the implications of these results for how data management and sharing tools can be developed to enable and mitigate data sharing pros and cons, as well as integrate into typical workflows.

This is followed by a review of design requirements for workflow sharing tools motivated by four development engineering use cases. The motivating use cases span different applications and purposes of data collection, and all of the workflows incorporate both sensor and survey data. We then present our implementation of such a tool to satisfy these requirements, which we call a pipeline builder. We found that several of the design requirements could be satisfied by building a layer of usability on top of powerful cloud computing infrastructure already available to those with the technical skills to navigate them. By combining a block programming paradigm with a cloud data processing platform that offers managed parallelism, we made it possible to execute highly scalable and reusable data analysis without necessarily writing any code. This lowers the bar for contributing to data analysis workflows. Resulting data analysis pipelines are more readable than raw code files, and as such lend themselves to increasing transparency–an underlying value of the push for open science–even if the code itself is kept private. The operator model of assembling individual data operations into the pipeline graphs also enables more fine-grained reusability.

Protecting privacy of human subjects is a key task in the process of making data available to other researchers, sometimes publicly. Our interview results revealed a disconnect between what common practice and privacy literature consider adequate anonymization, but also a growing awareness that this is the case and desire to protect confidentiality. We have found it is possible to improve both utility of data sharing and privacy protection at the same time by releasing data more selectively, using a series of mechanisms to enable more fine-grained control over who accesses what data.

In addition, by contextualizing our work within literature on international development,

we identified several barriers to equitable collaboration and explored their relevance to the data practices from our interview conversations and data management tools. We found the ways in which our work enables more transparency is a step towards enabling more public scrutiny and accountability, and how recommendations from development practice can also be applied to tools for data-driven impact analysis. Our exploration also highlights limitations of data-driven impact analysis that can be taken into consideration when choosing which components of workflows to automate for the sake of efficiency.

7.2 Future Work

As described in Chapters 3 and 4, our pipeline builder was derived from the Mezuri platform, a fully-featured end-to-end data management concept. An end-to-end data management system would be significantly beneficial for data provenance, one of our substantially motivated of our design requirements. Provenance is a prerequisite for reproducing and expanding on results, even for the original researchers. One system also reduces overhead to set up, manage, and troubleshoot data workflows. As such, a worthwhile next step is to combine the pipeline builder and privacy mechanisms. Fine-grained control over both access control and data processing steps in one system would enable interesting code analysis related features, such as automatic error bar calculations and operator-level user-based permissions. Identity and Access Management (IAM) can be set up such that the pipeline builder has access to certain data sources on behalf of certain user roles. Independent of whether we develop the tools we have created into a production-ready system, we believe existing tools and infrastructure as well as data sharing policies can be improved based on our results.

Our contextual analysis can also be used a starting point for creating principles for designing equitable data management and collaboration tools. As we mentioned in Chapters 2 and 6, expanding the interview participant pool to include consumers of electricity would reveal a necessary perspective on their role in existing processes. Along with interviewing additional stakeholders, it would be interesting to find representatives of projects that have initiated within communities and may or may not involve outside collaborators or funding. We would then be able to compare approaches to accountability and a broader consideration of the use, value, and risks of data.

Collaborative data management research is a response to the widespread calls for open data and transparency along with corresponding uncertainty around privacy best practices and a growing acknowledgement of digital data rights such as the EU General Data Protection Regulation (GDPR). At the same time, there is a movement towards digital ownership of individuals' own data, and the right to opt in and out of clearly stated purposes for analyzing personal data. All of this depends on access to new tools for projects with limited resources. To this end, we advocate for continued consideration of the participation of and tensions among different stakeholders of data, usability of existing infrastructure and interfaces, privacy strategies for protecting human subjects, the uses and impact of data analysis work, and aims to situate all of these factors in the social political contexts.

References

[1] H. A. Piwowar, "Who shares? who doesn't? factors associated with openly archiving raw research data," *PloS one*, vol. 6, p. e18657, 2011.

[2] C. Tenopir, E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett, "Changes in data sharing and data reuse practices and perceptions among scientists worldwide," *PloS one*, vol. 10, p. e0134826, 2015.

[3] M. H. Cragin, C. L. Palmer, J. R. Carlson, and M. Witt, "Data sharing, small science and institutional repositories," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 368, pp. 4023–4038, 2010.

[4] K. M. Mazor, A. Richards, M. Gallagher, D. E. Arterburn, M. A. Raebel, W. B. Nowell, J. R. Curtis, A. R. Paolino, and S. Toh, "Stakeholders' views on data sharing in multicenter studies," *Journal of comparative effectiveness research*, vol. 6, pp. 537–547, 2017.

[5] V. Gewin, "Data sharing: An open mind on open data," *Nature*, vol. 529, pp. 117–119, 2016.

[6] T. D. Sterling and J. J. Weinkam, "Sharing scientific data," *Communications of the ACM*, vol. 33, pp. 112–120, 1990.

[7] E. A. Thomas, C. K. Barstow, G. Rosa, F. Majorin, and T. Clasen, "Use of remotely reporting electronic sensors for assessing use of water filters and cookstoves in rwanda," *Environmental science & technology*, vol. 47, pp. 13 602–13 610, 2013.

[8] I. Seidman, *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. Teachers college press, 2006.

[9] H. A. Piwowar and W. W. Chapman, "A review of journal policies for sharing research data." in *ELPUB*, 2008, pp. 1–14.

[10] R. O. M. (USAID), "Ads 579 - usaid development data library." https://www.usaid.gov/sites/default/files/documents/1868/579.pdf

[11] N. I. of Health *et al.*, "Final nih statement on sharing research data," *NIH data sharing policy*, 2003.

[12] N. I. of Health *et al.*, "Nih data sharing policy and implementation guidance," *Retrieved June*, vol. 9, p. 2011, 2003.

[13] "Dfid research open and enhanced access policy," "Accessed: 2018-02-10". https://www.gov.uk/government/publications/dfid-research-open-and-enhanced-access-policy

[14] "Dfid research open and enhanced access policy: Implementation guide," "Accessed: 2018-02-10". https://www.gov.uk/government/publications/dfid-research-open-and-enhanced-access-policy

[15] A. Swan and S. Brown, "To share or not to share: Publication and quality assurance of research data outputs," *A report commissioned by the Research Information Network*, pp. 433–455, 2008.

[16] "Research for development outputs," "Accessed: 2018-02-10". https://www.gov.uk/dfid-research-outputs

[17] "Dfid digital strategy 2018 to 2020: doing development in a digital world," https://www.gov.uk/government/publications/dfid-digital-strategy-2018-to-2020-doing-development-in-a-digital-world, "Accessed: 2018-02-10".

[18] "Principles for digital development," https://digitalprinciples.org/, "Accessed: 2018-02-10".

[19] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*. Springer, 2011, pp. 338–340.

[20] A. Kipf, W. Brunette, J. Kellerstrass, M. Podolsky, J. Rosa, M. Sundt, D. Wilson, G. Borriello, E. Brewer, and E. Thomas, "A proposed integrated data collection, analysis and sharing platform for impact evaluation," *Development Engineering*, vol. 1, pp. 36–44, 2016.

[21] W. Bosl and A. Nguyen, "Crowdsourcing for research eeg annotation and accuracy estimation," https://informaticssummit2018.zerista.com/event/member/470393, "Accessed: 2018-03-01".

[22] "Apache beam," https://beam.apache.org/, accessed: 2017-07-08.

[23] T. Akidau, R. Bradshaw, C. Chambers, S. Chernyak, R. J. Fernández-Moctezuma, R. Lax, S. McVeety, D. Mills, F. Perry, E. Schmidt *et al.*, "The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing," *Proceedings of the VLDB Endowment*, vol. 8, pp. 1792–1803, 2015.

[24] H. Li, L. Xiong, L. Zhang, and X. Jiang, "Dpsynthesizer: differentially private data synthesizer for privacy preserving data sharing," *Proceedings of the VLDB Endowment*, vol. 7, pp. 1677–1680, 2014.

[25] C. S. Adorf, P. M. Dodd, V. Ramasubramani, and S. C. Glotzer, "Simple data and workflow management with the signac framework," *Computational Materials Science*, vol. 146, pp. 220–229, 2018.

[26] A. P. Davison, M. Mattioni, D. Samarkanov, and B. Telenczuk, "Sumatra: a toolkit for reproducible research," *Implementing reproducible research*, vol. 57, 2014.

[27] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier *et al.*, "Fireworks: a dynamic workflow system designed for high-throughput applications," *Concurrency and Computation: Practice and Experience*, vol. 27, pp. 5037–5059, 2015.

[28] D. L. Dotson, S. L. Seyler, M. Linke, R. J. Gowers, O. Beckstein *et al.*, "datreant: persistent, pythonic trees for heterogeneous data," in *Proc 15th Python Sci Conf*, 2016, pp. 51–56.

[29] A. Chen, "A new health care project won awards. but did it really work?" https://www.npr.org/sections/goatsandsoda/2016/10/22/497672625/a-new-health-care-project-won-awards-but-did-it-really-work, "Accessed: 2018-02-20".

[30] M. Mohanan, K. S. Babiarz, J. D. Goldhaber-Fiebert, G. Miller, and M. Vera-Hernández, "Effect of a large-scale social franchising and telemedicine program on childhood diarrhea and pneumonia outcomes in india," *Health Affairs*, vol. 35, pp. 1800–1809, 2016.

[31] A. Pal, V. W. A. Mbarika, F. Cobb-Payton, P. Datta, and S. McCoy, "Telemedicine diffusion in a developing country: the case of india (march 2004)," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, pp. 59–65, 2005.

[32] S. Surana, R. Patra, S. Nedevschi, and E. Brewer, "Deploying a rural wireless telemedicine system: Experiences in sustainability," *Computer*, vol. 41, 2008.

[33] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, pp. 1–34, 2000.

[34] N. Johnson, J. P. Near, and D. Song, "Towards practical differential privacy for sql queries," *Proceedings of the VLDB Endowment*, vol. 11, pp. 526–539, 2018.

[35] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 19–30.

[36] A. Narayan and A. Haeberlen, "Djoin: Differentially private join queries over distributed databases." in *OSDI*, 2012, pp. 149–162.

[37] N. Johnson, J. P. Near, J. M. Hellerstein, and D. Song, "Chorus: Differential privacy via query rewriting," *arXiv preprint arXiv:1809.07750*, 2018.

[38] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, pp. 557–570, 2002.

[39] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006, pp. 24–24.

[40] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 106–115.

[41] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, pp. 211–407, 2014.

[42] M. Crosas, "The dataverse network®: an open-source application for sharing, discovering and preserving data," *D-lib Magazine*, vol. 17, p. 2, 2011.

[43] J.-W. Byun, E. Bertino, and N. Li, "Purpose based access control of complex data for privacy protection," in *Proceedings of the tenth ACM symposium on Access control models and technologies*. ACM, 2005, pp. 102–110.

[44] L. Sweeney and M. Crosas, "An open science platform for the next generation of data," 2013.

[45] M. Bar-Sinai, L. Sweeney, and M. Crosas, "Datatags, data handling policy spaces and the tags language," in *Security and Privacy Workshops (SPW), 2016 IEEE*. IEEE, 2016, pp. 1–8.

[46] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 265–273.

[47] "Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule," https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html, accessed: 2018-02-16.

[48] S. E. Merry, *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking*. University of Chicago Press, 2016.

[49] E. Crewe and E. Harrison, "Whose development," *An ethnography of aid*, pp. 23–65, 1998.

[50] "Secret aid worker: we don't take data protection of vulnerable people seriously."

[51] T. M. Li, *The will to improve*. Duke University Press, 2007.

[52] T. M. Li. The will to improve : Governmentality, development, and the practice of politics. MSH Sud. https://youtu.be/VzFB3HxJ3P8

[53] E. A. Cech, "Culture of disengagement in engineering education?" *Science, Technology, & Human Values*, vol. 39, pp. 42–72, 2014.

[54] S. R. Arnstein, "A ladder of citizen participation," *Journal of the American Institute of planners*, vol. 35, pp. 216–224, 1969.

[55] A. K. Duraiappah, P. Roddy, and J.-E. Parry, "Have participatory approaches increased capabilities?" 2005.

[56] A. Sen, *Development as Freedom*. New York: Alfred Knopf, 1999.

[57] N. Fraser, "Injustice at intersecting scales: On 'social exclusion'and the 'global poor'," *European journal of social theory*, vol. 13, pp. 363–371, 2010.

[58] J. C. Scott, *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press, 1998.

[59] Y. Anokwa, T. N. Smyth, D. Ramachandran, J. Sherwani, Y. Schwartzman, R. Luk, M. Ho, N. Moraveji, and B. DeRenzi, "Stories from the field: Reflections on hci4d experiences," *Information Technologies & International Development*, vol. 5, pp. pp–101, 2009.

[60] P. Farmer, *Pathologies of Power*. University of California Press, 2004.

[61] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset)," *University of Texas at Austin*, 2008.

Appendix A:

Interview Guide

A.1 Background

The purpose of these questions is to learn about the context of relevant data management processes by asking general questions about the interviewee's work.

- What are you trying to learn and understand with your work?
- How do you try to understand those things? (Follow up on specific answers in next section)
- What are the limitations of trying to learn about and understand these things?
- What are the problems your work tries to address?

A.2 Data Practices

The first purpose of these questions is to understand the interviewee's perception of the value and consequences of data sharing practices by asking about his or her experiences and feelings. The other purpose is to collect technical details of how the interviewee works with data in order to understand - and later improve upon - the strengths and limitations of existing infrastructure and workflows by asking about the tools and methods used for data analysis.

- What kind of data do you work with?
  - How is it collected? Who is involved in data collection?
  - How is data stored and structured?
  - Do you have any issues with data quality or reliability?
- How is this data used?
  - How do you personally use the data?
- What sorts of analysis are done?
  - Who does the data analysis?
  - What software or programming languages are used for data analysis? Is this consistent within the organization or a variety depending on personal preference?

- Is there analysis you would like to do that you can't?
- Is data collected directly from customers/ patients. . . ?
  - Why do they provide the information? What are the incentives for them to provide information?
  - How else are they involved? Are they involved in any way other than providing information?
  - How do the results of the analysis affect them? Directly? Indirectly?
- Who else interacts with the data you work with, inside or outside of your organization?
- Do you publish work derived from this data?
  - Where?
  - Do you submit research to venues with data sharing policies?
  - How often?
- Do you share the data?
  - What about when you publish?
  - What does it look like when you share it?
- Is funding something you worry about related to how your data is shared or published?
  - What are your main concerns related to funding?
  - Do your funding agencies have data sharing policies?
- What do you think are the impacts of data being shared broadly in your field?
- What have been your experiences with data sharing policies, in the scope of your own work?
- Is the data you work with protected by data use agreements of any kind, for example by an ethics committee?
- Is privacy something you worry about related to how your data is shared or published?
  - What are your feelings about privacy, in the scope of your own work?
- Are there other issues you think are interesting related to data management/ sharing that we didn't talk about?

If you know anyone else who might be willing to do an interview with us, please pass along my contact information.