RIPL RESEARCH IMPROVING PEOPLE'S LIVES

The Data-Driven Policy System (DDPS)

A guide to a complete data-driven policy solution for state and local government.

RIPL | \ri'-pəl\ | **Research Improving People's Lives**
One Park Row, Suite 401, Providence, RI 02903
https://ripl.org

Contents

1. Executive Summary 3
1.1 Why do policymakers need a Data-Driven Policy System (DDPS)? 3
1.2 How does a DDPS solve pressing policy challenges? 4
1.3 How does a DDPS work, and how can my government team build one? 4
2. Infrastructure 7
2.1 Overview 7
2.2 Leveraging the security, scale, and cost-efficiency of cloud services. 8
2.3 Anonymizing and protecting the privacy of your data 9
2.4 Enabling automatic data lake construction, quality control, and updates. 10
2.5 Begin immediately deriving research insights. 10
2.6 Managing data access with a formal project governance process and access controls 10
2.7 Make a sustainable impact on policy with a permanent data resource. 11
2.8 Building innovative tools for the future 12
2.9 An Example Project 12
3. Governance 14
3.1 Creating and managing a governance agreement 14
3.2 Creating and managing data sharing agreements 15
3.3 Staffing the data lake 16
3.4 Project chartering and approval 17
3.5 Working with external partners 17
3.6 An Example Project 19
4. Knowledge Base 21
4.1 Implement off-the-shelf analyses and test the impact of policy interventions. 21
4.2 Extend existing solutions to fit your needs 23
4.3 Share your work for others to replicate 23
5. Frequently Asked Questions 24
References 26
Appendices 27
A1. Key Datasets 27
A2. Governance Agreement Template 28
A3. Data Sharing Agreement Template 43
A4. Project Charter Template 43
A5. Example Project Charter 64
A6. Preprint of Communications of the ACM Paper 77
A7. Summary of Available Projects in the RIPL Knowledge Base 87

1. Executive Summary

1.1 Why do policymakers need a Data-Driven Policy System (DDPS)?

From reducing poverty to improving education to delivering effective health care, policymakers at the state and local levels tackle some of the toughest problems facing society. More and more, policy leaders want evidence to ensure public policy is making measurable progress towards important goals. Policy needs to improve economic opportunity equally, economically, and measurably. It needs to deliver **value**, with programs that are **more efficient** and **more effective**, so that communities see the meaningful, measurable success they expect from their government leaders.

Policymakers face several challenges to delivering these successes:

1. **Scarce resources:** State and local governments face significant constraints, and often must search for ways to improve outcomes without net new resources.
2. **Preference for current services:** In the absence of compelling facts about a program or policy's success, there is a strong tendency to do things the way they have always been done. Programs, once established, tend to continue without significant changes, and it is difficult to abolish old programs or create new ones.
3. **Siloed information:** Data are typically used for limited purposes like compliance and reporting. Program administrators lack a complete picture of the individuals they serve, much less their families and communities.

To deliver efficient and effective policy, policy leaders need a robust, reliable, and simple way to access integrated program data to measure needs of their constituents, evaluate program effectiveness, and produce clear directions for improvement.

This manual will help your local, county, or state administration collect the system of software and services you need to achieve this. Research Improving People's Lives (RIPL) can connect the dots and provide a more complete picture of how policies and programs impact the people you serve and how new strategies and interventions can tackle the biggest challenges like reducing poverty, improving educational outcomes, reducing recidivism, ensuring health equity, and providing economic opportunity for your communities.

We will help you **make use of resources your state and local government already have**, **propel innovation** through the use of clear facts and evidence to guide effective policy and program direction, and **de-silo your data** so that policymakers and program administrators have a clear understanding of not just the needs of their community, but what approaches will help meet those needs quickly, efficiently, and on-budget.

We call the collection of our recommended tools, software, and services the RIPL-recommended Data-Driven Policy System (DDPS). **The DDPS is a free, open-source, cloud-**

based solution for state and local governments to securely host an integrated research data lake and put the data lake to work with a set of streamlined processes that empower policymakers to use their own data, internally and with external research partners, to generate data-driven insights to improve policy and lives.

1.2 How does a DDPS solve pressing policy challenges?

The DDPS allows policymakers to **measure what works and why**, including building powerful predictive tools to pinpoint solutions to complex policy challenges, meaningfully and accurately measure the impact of programs and policies across a wide array of outcomes, and develop data-driven tools and dashboards to support fact-based decision making. The system enables government to develop trusted relationships with academic researchers, nonprofits, and other partners while retaining the ability to control and monitor overall data use.

Examples of deliverables that state and local governments have developed using the DDPS include:

1. Highly-accurate predictive models that deliver the following tools: a **high accuracy, data-driven risk score to determine the likelihood that someone will experience opioid dependence**, abuse, or poisoning before an opioid prescription is given, to help health care professionals make more informed care decisions and curb the opioid crisis; and a **high accuracy, data-driven risk score to reduce non-emergency Medicaid Emergency Department (ED) costs and improve patient care** by predicting who will become a superuser of divertible Medicaid ED services in the coming year.
2. Causal analyses that measure important outcomes for policies and programs, including evidence that an early-life health care program for low birthweight babies **eliminates half the minority achievement gap in test scores**, saves tens of thousands of dollars in social service needs by age 14, and increases college enrollment for children from disadvantaged backgrounds by almost 20 percentage points.
3. Value-added models that measure the causal impact of a program on an individual's outcomes, such as **reduced recidivism,** by measuring the return-on-investment to in-prison training programs so policymakers can expand the most effective programs.

1.3 How does a DDPS work, and how can my government team build one?

To deliver fact-based policy, you need to be able to **measure where you are and determine effective policy directions to achieve your goals.** To do this, you need a research data lake that allows a 360-degree perspective of your community needs, characteristics, and trajectory. The research data lake needs to be **reliable, robust, and secure,** allowing you to form trusted partnerships with research teams that can help you develop the policy insights you need.

That's why RIPL developed a system of products that all work together. The RIPL-recommended DDPS provides each of the components you need to deliver data-driven policy improvements quickly and effectively. It can be implemented **for free** on an **accelerated timeline of under a year**, and **operated at low cost**. The environment, software, processes, and set-up are free. Government pays only for internal staff and for cloud computing services that it uses.

A DDPS for your local, county, and/or state government can be set up with the following steps:

1. **Set up a secure GovWorks enclave in the Amazon Web Services (AWS) GovCloud.** Many federal, state, and local governments host data on AWS, which offers reliable, scalable, and inexpensive Federal Risk and Authorization Management Program (FedRAMP)-approved cloud computing. A GovCloud account is easy to set up and manage, and the government maintains complete control over access. Everything you need to know to do this is in **Section 2**.
2. **Deploy the RIPL Analytics Data Lake (ADL) software from AWS Quick Starts.** RIPL built a free software package called Analytics Data Lake (ADL) that is easy to use, secure, and empowers its owners with transparency into and control over data access and use. It is designed to run in a secure AWS GovWorks enclave. Everything you need to know to do this is in **Section 2**.
3. **ADL will integrate and anonymize your government data.** The software uses automated scripts to build a master data format that takes backend administrative data — such as health, human services, labor, education, and criminal justice — and transforms them into an integrated, anonymized data lake optimized for research. The data lake preserves and protects confidentiality and privacy while delivering the benefits of integrated data insights to improve policy. Everything you need to know to do this is in **Section 2**.
4. **ADL will transform your data to build the key features that you need to measure success.** The software includes tools to automatically help you map your government data systems to the ADL master data format. It is extensible, allowing you to add additional data sets, generate codebooks, and research-ready tables. Because the results are standardized, minimal time needs to be spent on integrating, structuring, or cleaning data, allowing you and your partners to quickly begin the work to move from policy challenges to data-driven solutions. The ADL software is compatible with analysis and visualization programs such as Tableau, Microsoft Power BI, or Amazon QuickSight. Everything you need to know to do this is in **Section 2**.
5. **We have built best practice manuals, policies and templates for partnering with other agencies, academics, and nonprofits to harness talent and produce insight at low cost.** These manuals and templates can be used to foster productive partnerships that result in facts that quickly translate into policy innovation. The data always remain yours. They

do not go to an external party, meaning that you never lose track, and you always have complete transparency with your partners. Our system guides you to set up appropriate levels of access for every partner working with the data lake. Partners access only the specific fields and tables that you have agreed upon beforehand in an approved project charter, and all partner use of the data lake is logged and audited for complete transparency. Everything you need to know to do this is in **Section 3**.

6. **RIPL DDPS members can join a dynamic learning community of like-minded policymakers, minimizing the need to "reinvent the wheel**." All members of the RIPL DDPS can draw from the RIPL DDPS Knowledge Base, which houses existing code repositories and project templates. Members can adapt code, share results, and learn from other policymakers using the RIPL DDPS. Members generate efficiencies by replicating each other's code, and drawing from existing machine-learning predictive models, dashboards, and apps. Everything you need to know to do this is in **Section 4**.

By putting facts at your fingertips, the RIPL-recommended DDPS empowers you to quickly deliver efficient and effective policy for the communities you serve, via trusted and transparent partnerships, at low cost.

2. Infrastructure

2.1 Overview

The first step towards implementing a DDPS is deploying a secure cloud-based enclave[1] to host your data lake and structure it for insights. As a policymaker, structured data can help you achieve important goals and activities, including:

1. Fulfill reporting requirements;
2. Evaluate the impact of programs and policies;
3. Track and communicate these measures with dashboards;
4. Predict outcomes to inform intervention strategies;
5. Develop strategies to target resources more effectively;
6. Predict outcomes to inform long-term strategic planning; and
7. Discover insights that drive innovation.

All of these activities must be conducted within a safe and secure infrastructure that complies with federal and state data privacy laws. For some of the above goals, you may need to work securely and transparently with an outside partner or expert. Working with external partners requires tools to control data access for each individual user, and thoroughly audit and log each user session. Furthermore, to translate your insights into sustainable and permanent policy impact, you will need infrastructure that can support on-going data updates from multiple agencies and departments (see **Appendix A1. Key Datasets**). This will ensure that data-driven policy impacts are sustainable and permanent.

This section will help you understand how to achieve all of these goals with an AWS GovWorks secure enclave and the RIPL Analytics Data Lake (ADL) software. This system **integrates your data, anonymizes it to protect privacy, structures it for research and insights, updates it efficiently, and hosts it in a secure cloud solution** where you can define and monitor access, both for internal analysts and external partners.

There are a couple important definitions to know as we walk through the infrastructure:

1. **Cloud computing** is the on-demand delivery of storage and computing resources, applications, databases, and other computing infrastructure through a vendor's secure and scalable platform. With the cloud, you only pay for what you use, and additional

[1] A "data enclave" is defined by the National Network of Libraries of Medicine as a system where "a researcher can access the data from their own computer but cannot download or remove it from the remote server" (https://nnlm.gov/data/thesaurus/data-enclave). A related but broader term is "secure enclave," which can apply to applications beyond research, such as blockchain validation (https://medium.com/oasislabs/towards-an-open-source-secure-enclave-659ac27b871a).

resources are available instantly as needed. See the AWS overview of cloud computing here.

2. A **secure enclave** is a system built to control data access and ensure data security. Transfers of data in and out are restricted and subject to a documented approval process. All access can be comprehensively audited, and access is granted on a per-project basis and only to approved analysts, researchers, or partners. Copy and paste functionality is disabled, and no data are able to leave the system without documented approval.
3. A **data lake** is a secure store of data that includes unstructured, structured and transformed data optimized to deliver insights through reporting, visualization, econometric analysis, forecasting, and machine learning. Data lakes are an efficient and increasingly popular way for companies and government to use data from a variety of sources.

2.2 Leveraging the security, scale, and cost-efficiency of cloud services.

Partnering with a leading national cloud provider will ensure best-in-class security features and data protection for your data lake. Your AWS GovWorks secure enclave runs in AWS GovCloud and inherits all of the physical and data center security controls that AWS has developed to meet the highest standards of data security. AWS's industry leadership in security and compliance ensures data protection, cutting-edge security capabilities and services to increase privacy and access, and extensive auditing features provide you with full visibility into how your data are used, including:

1. **Network firewalls** that let you create private and secure networks;
2. **Data encryption services** that can provide scalable and efficient encryption features and flexible key management;
3. **A security assessment service** that automatically assesses applications for vulnerabilities or deviations from best practices;[2]
4. **Connectivity options** that enable private, dedicated connections to the secure enclave from defined locations, such as your office network;
5. **Identity and access control tools** that let you define, enforce, and manage user access policies across your account, including multi-factor identification.
6. **Inventory and configuration management tools** that allow you to track and manage changes to your permissions and data over time;

[2] AWS continually tests their infrastructure using penetration testing and compliance reports. You can also work with your own security team or another vendor to carry out additional security assessments and penetration tests if desired.

7. **Monitoring and logging tools** that empower you to with transparent governance, aggregated logs for compliance reporting and auditing of your system.

You can learn more about AWS tools to ensure compliance with all federal and state data privacy regulations, including Health Insurance Portability and Accountability Act of 1996 (HIPAA) and FedRAMP, here. Learn more about AWS security structure and processes in detail here.

Importantly, the data lake is completely owned and managed by government. Data never leave government custody. Government retains control over who accesses the data, which data fields, tables, and sources they have access to, export control, and complete audits of all analyses and commands run on the system.

The AWS pay-as-you-go model offers a cost-effective alternative to traditional on-premise data storage models. By renting only the computing time and storage capacity you actually use, a cloud solution is much more cost effective than building or maintaining a dedicated on-premise system.

2.3 Anonymizing and protecting the privacy of your data.

Government administrative data contain personally identifiable information (PII), a necessity to identify and contact individuals for the purpose of ongoing program administration. However, PII is typically unnecessary for the purpose of research and insights; results are estimates of the impact of a policy or program on populations as a whole, rather than specific individuals. Therefore, the best practice for constructing an insights-optimized data lake is to remove PII as early as possible, both from a privacy and security standpoint.

With a DDPS, government administrative data containing PII are only ever used to create a global anonymous identifier to join and integrate data about the same individual across different data sources. This means that PII are not available for research purposes, and are protected according to industry best practices that AWS has developed for HIPAA and FedRAMP compliance. No one should ever be able to view, access, or use PII when they use the data lake.

As further protection, the ingest and anonymization processes are fully automated to separate out personally identifiable information (PII) and generate an anonymous ID using a deterministic linkage algorithm without manual intervention (Hastings, Howison, Lawless, Ucles and White 2019). When intervention is required to troubleshoot or complete the anonymization process, AWS-designed controls will limit access to the minimum possible, and will keep complete audit logs of those activities.

As a project progresses, an agency may need to re-identify data for an operational purpose, for example to implement a policy change based on the findings. At the time of data ingest, ADL

produces a look-up table that maps an agency's internal identifier (program/client ID) to the anonymized identifier, which can be downloaded by the agency. It is not stored in the data lake and is not accessible by research partners.

2.4 Enabling automatic data lake construction, quality control, and updates.

Administrative data are typically siloed by program within an agency, and then aggregated for reporting and compliance. This makes it difficult to produce fast or accurate research and policy insights. The RIPL Analytics Data Lake (ADL) removes those difficulties with a comprehensive integration tool and graphical user interface (GUI) that allows your analysts to quickly and easily map each of your separate local data tables into a standardized master data lake format that highlights the most important variables for research insights. This master data format is regularly updated by RIPL to support new variable definitions.

Once your tables have been mapped to the master data format, the ADL takes care of the rest. Automated scripts will anonymize your data and build a data lake that is ready for insights. ADL also generates an automated and thorough data codebook for each dataset in the data lake, which help your analysts and partners identify and resolve data quality issues. Codebooks are automatically updated and monitored for data quality issues, and contain descriptive statistics to help your analysts and partners quickly produce insights.

2.5 Begin immediately deriving research insights.

Our ADL software transforms your anonymized data to create a set of "derived tables," which contain standardized variables most frequently used for research and insights. For example, benefits programs will often record participant income as series of different types of income — income from wages, other benefits programs, investments, and so on. A derived table may simply produce one research-ready variable — total monthly income — by transforming these many fields into the variable that analysts and researchers need to generate insights. Analysts can always use the more granular tables within the data lake if more detail is necessary.

This automation greatly reduces the start-up time required to begin deriving insights from your data, while helping you establish robust and reproducible definitions for the variables that you will use in analysis. The derived tables are ready for access inside the secured data enclave using self-service analytics and visualization tools such as Tableau, Microsoft Power BI, or Amazon QuickSight.

2.6 Managing data access with a formal project governance process and access controls.

Although the data lake has been anonymized to protect privacy, the de-identified individual-level records it contains are still sensitive and should be held to the highest security standards.

In addition to the network and data security features provided by the secure AWS GovWorks enclave, we recommend these best practices for a DDPS:

1. **An institutionalized project charter process for applying to use the data.** As outlined in Section 2.2, the AWS identity and access controls let you regulate access to the data lake. We recommend that both internal analysts and external research partners apply to use data through a formal project governance process, which is described in more detail in the **Section 3. Governance** below. Users of the data lake can apply for access under a chartered project, and enter into a data use agreement with the agency hosting the data lake for approved projects.
2. **All exports from the enclave must be reviewed and improved**. The AWS GovWorks enclave ensures that individual-level data cannot be shared outside the cloud-based environment. You can allow results, such as statistical tables or graphs, to be exported from the enclave through a documented approval process to make sure they are in accordance with the project charter and are aggregated in compliance with federal cell-size suppression guidelines.
3. **All data access is limited only to what users need to complete a project.** The AWS GovWorks enclave supports project-level permissions to limit individual user's access to the data tables and variables. We recommend that access is approved through the governance process for each project, with users only accessing the exact data outlined in the project charter. Requests from other researchers to replicate prior research studies or evaluations can also be managed through the same chartering process.
4. **All access is regularly audited.** The AWS GovWorks enclave allows you to view complete reports of how and when users have accessed the data lake for approved projects. Because data cannot leave the enclave and are not transferred to other systems outside of your control (such as a laptop or system owned or operated by a researcher at a university or a third-party vendor), you do not need to request audit logs or reports from any other system or party. The enclave already has all of the records you need to conduct an audit for compliance with data privacy laws. We recommend that you conduct regular audits to ensure all access and use is in current and in compliance.

2.7 Make a sustainable impact on policy with a permanent data resource.

As the owner of the data lake, you have full control over its long-term maintenance and availability. ADL provides tools to help make this a robust, permanent resource. At each update, a full snapshot of the data lake is archived in the cloud to allow for future replication of any analysis at any point in time. These snapshots are accompanied by versioned codebooks that reflect the state of the data lake at that point in time. Every chartered project has its own versioned code repository and an archived copy of the anonymized administrative records it is

approved to access. Research products that have been approved for export are also permanently archived at the time of export.

The infrastructure uses highly durable storage services from AWS, which are designed to prevent data loss through back-up data stored in remote servers. As an additional safety precaution, all code necessary to generate the data lake will be versioned and archived in a code repository outside of AWS. In the very unlikely event that the data lake was lost, it could be regenerated by re-importing the raw administrative records and re-running the automated pipeline. It is an ironclad system that works the same way, every time.

2.8 Building innovative tools for the future

By hosting your data lake in the cloud with the secure AWS GovWorks enclave, you are also preparing for future applications of data insights to improve policy. The comprehensive security and access controls provided by the AWS platform allow you to determine how, when, and to whom you share insights, research results, and aggregated statistics from your data.

Once you have used an established governance process to choose an appropriate level of sharing information, the ADL helps you deploy an application programming interface (API) that runs on top of the data lake and provides efficient and controlled access for tools that you or a vendor build when implementing policy.[3] Because many vendors already use AWS to develop and host web and mobile applications, providing your API to them directly within AWS improves performance and security.

2.9 An Example Project

We will use a worked example to demonstrate the advantages of deploying DDPS infrastructure. Imagine that a Department of Health and Medical Services (DHMS) currently runs a treatment and harm-reduction program for individuals diagnosed with opioid misuse disorder. An analyst at DHMS would like to build a dashboard to help understand which communities need additional treatment support by comparing the characteristics of individuals who are successfully treated with individuals who experience further adverse opioid outcomes, such as continued overdose, poisoning, and death. The dashboard will need to be updated regularly and consulted during internal planning meetings for the program.

To quickly and efficiently build this dashboard, the analyst will need secure access to anonymized data about the individuals in the program across a variety of characteristics, including their treatment outcomes, health history, employment and wage history, interactions with the criminal justice system, and their demographics. An anonymized data lake is well-suited for this need, as it ensures the privacy of individuals and protection of personal and

[3] APIs are an industry best practice for developing web and mobile applications on top of data resources (https://medium.com/@perrysetgo/what-exactly-is-an-api-69f36968a41f).

sensitive information while allowing important research insights to be produced using joined, de-identified data. DHMS has recently built a DDPS to help with this kind of project. The analyst builds the dashboard using the following steps:

1. To begin, the analyst completes an internal application to create the dashboard with the DDPS, using the formal governance process described in **Section 3.** She requests permission to use secure and anonymized Medicaid claims data, the opioid program administrative data, corrections data, and wage data. She is approved by each agency, and is granted access to the data lake.
2. The data lake contains a set of derived tables, containing important variables that the analyst needs to build her dashboard. These include an indicator for an individual's enrollment in Medicaid and their use of Medicaid services; variables for their monthly wages over the past five years; an indicator for whether they have ever been incarcerated, as well as the duration of each incarceration spell; their treatment outcomes from the program; and their demographics. All variables are in one centralized system, with clear codebooks and data dictionaries. The analyst uses these variables to build her dashboard within the secure enclave.
3. The analyst completes the dashboard in about a month. Without all of the data joined and optimized for research, this process could have taken years — the analyst would have had to secure data-sharing agreements with the agencies, understand what was in each dataset, clean the data, create the variables needed for research, and then conduct analysis.
4. The dashboard is regularly and automatically updated by the data lake, keeping it up-to-date with no staff time or effort needed. The analyst also provides regular updates to all agencies on the progress of the dashboard. All agencies are able to view the dashboard results, and can integrate the findings into their own policy initiatives.
5. DHMS is able to integrate the dashboard into its own program and policy practice within months of the analyst first beginning work, allowing efficient and effective policy action to improve services to communities in need and fight against the opioid crisis.

3. Governance

Building a data lake is the first step in leveraging the DDPS to produce innovative and effective policy. Next, you will need a coordinated team to manage a clear, consistent and transparent governance process so that approved users can create meaningful policy insights to improve lives. An effective governance process allows you to:

1. **Create and manage a governance agreement with appropriate data-sharing agreements:** A governance agreement between all agencies using the data lake will allow you to clearly establish terms and agreements for management and use of the data lake. Data-sharing agreements allow you to clearly establish terms and agreements for additional agencies contributing their data to the data lake.
2. **Manage staffing:** Staff the data lake to guide and manage its secure use in compliance with all agreements and project charters.
3. **Charter projects and approve partners:** You will need a project chartering process that allows government employees to easily use data from their own department or agency for approved projects, in order to streamline requests for data from other departments and agencies for approved projects in an organized and documented manner.
4. **Work with and approve external partners:** Facilitate public and private partnerships to jointly address the most challenging policy questions using clear guidelines for building research partnerships.

This section will help you understand how to achieve all of these goals with a set of template documents and governance guidelines that are **customizable to your local, county and state government's needs**, **allow you to build consensus around use and governance of the data lake**, and **organize and manage use of the data lake to create a clear, consistent policy that helps you achieve your policy goals in an efficient and effective manner**.

3.1 Creating and managing a governance agreement

As we described in **Section 2**, the data lake will contain important administrative data from multiple government agencies to facilitate fact-based policy. It is therefore crucial that governance of the data lake is clear and consistent, and allows all contributing agencies to work towards achieving their policy and insight goals.

As part of the RIPL DDPS solution, we recommend two sets of documents to clearly establish terms and agreements for governance and use of the data lake for all agencies contributing to and using the data lake: an overarching governance agreement document between all agencies contributing to the data lake, and individual data-sharing agreements with each agency contributing to the data lake.

A **governance agreement document** is a single legal document between all agencies/departments participating in the data lake. It articulates the purpose motivating the creation of a data lake, defines key terms related to the creation and use of a data lake, sets out terms and conditions for the use of the data lake, and designates a Holding Agency (the agency appointed to hold the data contract for the secure cloud environment housing the data lake) and a Governance Committee (composed of members of participating agencies) that will be responsible for hosting the data lake in a secure environment and overseeing use of the data lake, respectively. The recommended structure and roles for the Holding Agency and Governance Committee are described in further detail below in **Appendix A2. Governance Agreement Template.**

The RIPL-recommended template sets out recommended terms and conditions for a data lake. It can be customized to the needs of your agencies and the purpose of your data lake. The template includes the following sections to ensure an agreement that is fair, transparent, trusted, and rigorously enforced to the satisfaction of all parties:

1. **Terms and conditions for the data lake hosting environment and architecture**, which covers environment security, data lake contents, data documentation and refresh frequency, anonymization and user controls, and automated logging and alerting.
2. **Terms and conditions for the data lake governance committee**, including committee duties and meeting frequency.
3. **Terms and conditions for the data lake management team**, including the structure, roles and responsibilities.
4. **Terms and conditions for project approval**, including sponsorship, documentation, and approval of projects.
5. **Terms and conditions for accessing the data lake**, including conditions for authorizing access, authorized user accounts, and reporting, investigating and consequences for data lake misuse.

3.2 Creating and managing data sharing agreements

A **data-sharing agreement** (DSA) is a legal document between each individual participating agency and the Holding Agency that allows for the sharing of the individual agency's data with the Holding Agency for the purpose of contributing to the data lake. These agreements stipulate what specific data will be shared, data security measures, frequency of data transfers, and permissible uses of the agency's data. When a proposed project will utilize a given agency's data, that agency's DSA must be appended with the approved project charter. See **Appendix A3. Data Sharing Agreement Template** for more details.

The RIPL-recommended template sets out recommended terms and conditions for a data lake. It can be customized to the needs of your agencies and the purpose of your data lake. The

template includes the following sections to ensure an agreement that is fair, transparent, trusted, and rigorously enforced to the satisfaction of all parties:

1. **Defining agency mission and policy goals**;
2. **Terms and conditions for the protection of Confidential Information (CI) and Personally Identifiable Information (PII)**, including terms for access, prohibiting disclosure, security and anonymization, and cell size suppression policy;
3. **Terms and conditions for use, term, and termination**, including under what conditions the data lake may be used, project approval, publishing research results, and data destruction;
4. **Terms and conditions for transferring data from the agency into the data lake**, including data custodians, points of contact, and security specifications.

3.3 Staffing the data lake

Managing, updating and maintaining the data lake require a small team of core staff, as described in the governance agreement. These staff can be existing team members or new hires, but must be sufficiently dedicated to supporting the data lake. The core team will maintain and update the data lake, manage access and project approvals, coordinate with research partners to produce insights, help internal government staff use the data lake for insights, share insights, and provide key oversight and management.

We recommend the following staff positions to manage your data lake:

1. **Senior project manager (PM).** An individual with business intelligence and data analytics experience who will serve as the central interfacing point for cross-agency coordination for all logistics related to the data lake use and maintenance and for priorities and needs of different agencies.
2. **Senior data engineer (DE).** This individual will be in charge of the technical aspects of maintaining the data lake, including curating the data lake, updating codebooks, managing user accounts and access to particular tables and columns, periodic database updates, version control, archiving of data versions, database release notes, and adding new data to the data lake.
3. **Senior data scientist (DS).** An individual with a background in computer science and applications of machine learning and statistical modeling in a general field of the social sciences (e.g. Economics as opposed to Health). This individual will be in charge of coordinating project execution, providing technical expertise and guidance to partners across agencies, and ensuring that all technical software and platforms used are state-of-the-art technologies.
4. **Empirical economist (EE, optional).** An individual with a PhD in Economics with a specialty in empirical microeconomics applied to social policy (Public Economics) who

will work closely with the DE and DS to coordinate project execution and evaluate proposed projects from an economic perspective (for example: causal identification, selection bias, market response, etc.).

A full description of recommended staff duties and roles can be found in **Appendix A2**. **Governance Agreement Template.**

3.4 Project chartering and approval

To achieve your policy goals and deliver fact-based policy, you will need both internal and external partners to be able to use data. To do this, anyone applying to use the data lake must define a clear research goal with deliverables that are documented and agreed upon with all relevant agencies beforehand. The RIPL-recommended DDPS uses a project chartering process to achieve this goal.

A **project charter** is a document that describes the entire scope of a project. It is the foundation of any successful project and collaboration, and clearly lays out the goals for use of the data lake, the data needs for a project, methods, and deliverables that all partners and collaborators on the projects have agreed on in advance.

All agencies whose data are proposed for use in a project much sign off on a project charter. It specifies exactly who will have access to the data, which data tables and fields they will use, and what work they are authorized to conduct on the data. **Every use of the data lake must be chartered and then approved by all the agencies whose data are included in the project, even if it is a single agency accessing its own data**. This helps keep the use of the data lake documented, transparent, and easily replicated.

For the RIPL-recommended project charter template, see **Appendix A4**. **Project Charter Template.** The charter can be used for both internal government users and external research partners.

3.5 Working with external partners

External partners can help you achieve policy goals and insights by bringing additional expertise and capacity to research projects. The RIPL-recommended DDPS is ideally-suited to bring in external partners, as it ensures that all access is on a centralized system belonging to government. Data never leave government custody, and all use by external partners must be approved using the project chartering process described in **Section 3.3**. Using the controls described in **Section 2**, all use of data by external partners can also be thoroughly and immediately audited, and all research results must go through an export request process before they can leave the data enclave. Using a RIPL-recommended DDPS, partnering with external parties, such as researchers, nonprofits, or contractors means that policymakers and

government can be confident that their data are being securely used for agency-approved research and deliverables at all times.

Scoping work and deliverables should be completed prior to the launch of any new project with an external partner. When working with academic partners, ensure you budget time for the scientific publication process. You also must plan to host replication files, which are created and archived to make sure that any published analysis in a scientific study can be replicated by other scientists. Please consult with your academic partner to better estimate the length of these processes. In many cases, it can take at least a year.

The scoping and deliverables process must also include an agreement before work begins about final publicly-available deliverables for the project, including scientific papers and published research. You and your partner should agree in writing whether results will be available to publish, and if there are any terms and agreements that must be met when preparing results for publication — such as a window of opportunity for government to view the work and comment before it is submitted to a journal. Agreement on publication before any findings are generated is necessary to meet scientific standards for producing objective and generalizable research.

3.5.1 Identifying external partners

To bring in external partners to work on approved projects, you will need guidelines for building research partnerships. Identifying a good partner is crucial to the success of any project. Good partners will have the following:

1. Verified credentials, academic or otherwise, that are adequately suited to conducting analysis with government administrative data.
2. A regular reporting schedule agreed upon by all partners.
3. Task management software (such as JIRA) that is used for all work and can be audited by government.
4. Data-fluency; experience with data management and analysis.

Please visit www.ripl.org/partners/ for a list of partners with expertise in predictive modeling, app and web building, dashboard building, behavioral science "nudges," and program evaluation. You can also contact connect@ripl.org for further questions.

3.5.2 Data access fees

The goal of your data lake is to ensure that your local, county, or state government can produce policy-relevant research and insights to help policymakers deliver efficient and effective services. To ensure that data are being used to drive policy improvements that the state and its communities benefit from, we recommend that each proposal with an external partner has a clear deliverable that is focused on a policy change, program, or operational dashboard that can

be institutionalized and regularly used by an agency, or a product that the state clearly benefits from.

However, some external partners may be interested in using data for supporting general science that contributes to public and scientific knowledge but does not include an immediate and direct benefit for your government. In this case we suggest following the Census Research Data Centers guidelines and charging an annual fee that your Governance Committee deems appropriate. External partners conducting approved research with general but not immediate or direct benefits to the government can defray costs through federal, foundation or research grants. The funds generated through these fees can be used for maintenance of the data lake and support of projects that are generating facts and evidence for direct policy benefit and improvement.

3.6 An Example Project

Executing a project requires many steps, as seen above. To help you understand how the entire process unfolds, we will describe a hypothetical project undertaken by RIPL and a state agency.

In this scenario, RIPL is partnering with a state-level Department of Youth and Family Services (DYFS) for a hypothetical charter project entitled "The Impact of Foster Care Placement: Finding What Works for Vulnerable Children." The example charter can be found in **Appendix A5. Example Project Charter**.

Policymakers and government agencies often have specific policy goals or areas they wish to explore. The Director of DYFS approaches RIPL for help with evaluating the department's foster care placement cases to better understand if children who are placed in foster care have better outcomes later in life than those who are not placed in care following an investigation. This **research question** formed the basis of the partnership. In other cases, you may not yet have such a specific research question, but instead a broader policy goal. Meeting with potential research partners can help you distill a broader policy aim into more specific research questions.

After several conversations between the Director of DYFS, her staff, and RIPL team leaders, we determine the **scope of data** needed to address the research question, including data from DYFS and several other state agencies.

The RIPL research team drafts a charter, which DYFS and all other state agencies approve using the process laid out in their joint Governance Agreement. In the charter, DYFS and RIPL agree in advance to pursue this project as a **published paper** in a peer-reviewed scientific journal, as the analysis holds the potential to not only contribute to policy but also to deliver new scientific advances that the broader policy and scientific communities could learn and benefit from.

The data lake management team grants access to the specific data fields specified in the approved charter to RIPL's staff. RIPL begins work and regularly meets with DYFS staff to **share**

preliminary results and get feedback from DYFS's staff experts. Finally, RIPL compiles the analysis for all of the research questions into a **final report** for DYFS and presents it to DYFS's senior leadership. After incorporating final feedback on the analysis, RIPL then compiles the results into a scientific paper to begin the publication process, and continues to work with DYFS staff to help translate the results into improved policy for children in the state.

4. Knowledge Base

Now that you have both a data lake and a governance structure to manage it, it's time to put your DDPS into action to develop policy insights that improve lives for your community. **Sections 1-3** describe how to set up an infrastructure that your analysts and partners can use to produce insights. But a DDPS doesn't just let you learn from your own work — **it also enables code and results to be optionally shared and replicated instantly with other governments who have implemented a DDPS**. To do this, you will need to know the following:

1. The features and projects that you can implement right away using a DDPS.
2. How to avoid duplicating effort by replicating work that others have successfully completed.
3. How to share your knowledge and innovation so that other policymakers can learn from your successes.

Without a DDPS, when policymakers partner with researchers and external experts, the progress they make in understanding their data often remains isolated within a particular project and partner. Governments can end up with many copies of their data transferred to different partners who may interpret the data differently, and results generated by one group may not be replicable by another.

This section will help you understand how to solve this problem with a **shared and open-source knowledge base for disseminating best practices, analysis code, predictive models, dashboards and applications across users of the RIPL DDPS**. This can be shared both across your chartered projects, but also across other local and state governments who use the DDPS. **You control how much you share (if anything) in the DDPS Knowledge Base**. By adopting the DDPS standard, you will benefit from prior research and projects that have already been developed to work with the data lake architecture. You can learn from other government's programs and join the discussion around policy areas that are broadly important to everyone. As you charter and execute your own projects, you can contribute knowledge and innovation to others to help improve lives beyond your own community.

4.1 Implement off-the-shelf analyses and test the impact of policy interventions.

The RIPL-recommended data lake software is provided through free AWS QuickStart templates that reduce the development and technical effort you would otherwise have to invest in building a secure data lake. Similarly, we have also built a **knowledge base** that contains free ready-to-go code and analyses from RIPL's previous projects that you can implement in a fraction of the time and cost needed to start them from scratch. The codebase includes **predictive models** that can help you **predict the following outcomes with high accuracy**:

1. Medicaid recipients who have never before received an opioid prescription but are at risk of developing an opioid use disorder if written an initial prescription by a physician (Hastings, Howison, and Inman 2019).
2. Medicaid recipients who will emerge as high-cost users of the emergency department for visits that are non-emergent or could be treated in primary care, and who could be diverted to equally effective but less costly care.
3. Families and communities at risk of substantiated child maltreatment investigations by child protective services, who may benefit from earlier intervention and services that prevent maltreatment.
4. Standardized test scores of students based on variables known at birth, in early childhood, at school age, and prior to the test, to better understand the potential policy levers for improving educational success, and prepare schools for trends in the needs of incoming cohorts.

All of these predictive models can be quickly implemented for free using your own data, as your data will already be mapped to standardized formats using the RIPL GUI interface and data-optimization algorithms described in **Section 2**.

The knowledge base also includes easy-to-implement code for **causal analyses** that measure:

1. The impact on later-life outcomes from providing very low birth weight newborns with additional medical care after delivery (Chyn, Hastings, and Gold 2019).
2. The impact on later-life outcomes of removing children from their homes following a substantiated child maltreatment investigation, and of juvenile detention (Bald, Chyn, Hastings and Machelett 2019).
3. The impact on voter participation following the enactment of a photo ID law (Esposito, Focanti and Hastings 2019).
4. The impact on re-employment and wages from requiring unemployment insurance applicants to attend mandatory return-to-work training sessions.

Finally, the codebase contains a software library for calculating **value-added models.** Value-added measures the impact of enrollment in a program on outcomes of interest relative to individual's baseline predicted outcomes prior to enrollment. These models are used in health, education, and labor training in economics and social sciences, and are key inputs for developing and disseminating program rating information and developing accurate and unbiased pay-for-success models. The knowledge base includes easy-to-implement code for the following **value-added models**:

1. The return-on-investment of individual in-prison training programs in terms of recidivism reduction. The model measures the impact of each training program on an inmate's

likelihood of recidivating, helping policymakers reduce recidivism by highlighting the most effective programs.

2. Additional models regarding the value-added of labor training programs on future earnings coming soon.

For a summary list of currently available Knowledge Base projects, please see **Appendix A7. Summary of Projects Available in the RIPL Knowledge Base**.

4.2 Extend existing solutions to fit your needs

Every government has its own policy priorities specific to the needs of its community. You may find that the knowledge base doesn't yet contain an analysis suited to your needs. However, all the code in the knowledge base is available to edit and adapt to your own needs, allowing you to quickly and efficiently generate additional predictive models and causal analyses using a peer-reviewed template.

Our existing models are documented in the knowledge base, available open-source on GitHub, and based on a common predictive modeling library (Howison, Berenbaum, and Shah 2018). Your research partners can work with you to adopt these open-source codes and extend them to model your policy of interest, reducing the development effort and the time to solution.

For example, you may want to build a predictive model that uses health, social services, wage records, and demographic data to predict infant mortality. While there are no infant mortality predictive models in the knowledge base, the RIPL-developed model to predict future high-cost Medicaid ED users utilizes very similar data to predict a different health outcome. Instead of building a new model from scratch, your analysts can adapt the existing template to outcomes involving infant mortality, reducing the insights to production timeline by months while ensuring a high-quality model.

4.3 Share your work for others to replicate

After working with your partners to extend or develop new models and policy solutions, you can contribute them back to the community of DDPS users through the knowledge base. Your participation in the knowledge base helps promote learning and knowledge across governments and partners, democratizes contributions to improved policy by opening the possibilities for partnerships with the best and brightest talent across sectors, and facilitates peer-review and replication. **You control how much you share**: if you do not wish to share results, sharing code can still help others learn and produce results from their own data. An additional benefit to sharing and archiving your data lake findings in the knowledge base is that it will help you reuse your own previous projects and results internally for years down the line.

5. Frequently Asked Questions

How do we get started?

To get started, you need an account with a government cloud services provider (we recommend AWS GovCloud), a core internal data lake team (see below), administrative data and data-sharing agreements, policy goals, and a research plan (and research partnerships) to turn policy goals into impact. The RIPL DDPS manual contains materials to help you with all of these components.

How do we get an AWS account?

We recommend using Amazon Web Services (AWS). RIPL's Analytics Data Lake software is available on AWS Quick Starts. These services can be run at a low monthly fee. To contact AWS about cloud services, reach out to:

Other Cloud Service Providers include Microsoft Azure and Google Cloud.

Why is the cost so low?

The software for the secure AWS GovWorks enclave and RIPL Analytics Data Lake are both open source and available free of charge. You pay only for the computing and storage capacity that you need, which may be only a few thousand dollars per month, depending on the size of your data and number of projects. For questions about procurement, contact RIPL at connect@ripl.org.

Who do we need on our team?

We recommend a core internal government team to maintain and update the research data lake, manage project approval and data access, coordinate with research partners and internal government staff to produce insights, and provide key oversight and management. They include a senior project manager, a senior data engineer, and a senior data scientist. See **Appendix A2. Governance Agreement Template** for details and suggested qualifications.

Which state and local government agencies use cloud computing for sensitive data?

Many state and local governments store and analyze their data on AWS secure GovCloud solutions. We've listed some publicly available examples below. More publicly available examples can be found here.

1. CMS: Medicaid Data Warehouse. The Centers for Medicare & Medicaid Services (CMS) is bringing Medicaid records from 50 states—totaling 72 terabytes of data—to the cloud. In the past, this would have created a "log jam," and likely impossible for the agency's

internal data centers. California alone produces some 900 billion records per month. Claims, encounter, enrollment, and other data for all 74 million Medicaid-covered lives are securely stored and analyzed using AWS.

2. Maryland Department of Human Services (DHS) Total Human-services Integrated Network (MD THINK). The Maryland Department of Human Services (DHS) designed and built Maryland's Total Human-services Integrated Network (MD THINK) to deliver integrated health and human services programs to the state's most vulnerable residents. By building this first-in-the-nation Technoloti cloud-based platform, Maryland agencies can gain a holistic view of citizens receiving benefits and analyze data across agencies to design better assistance programs.
3. Appriss. Appriss is another innovator working to mitigate the opioid crisis through the accessibility of *quality* data. It built a comprehensive prescription drug monitoring program using AWS, which is now host to 42 participating states. With the appropriate interstate agreements in place, the tool makes it possible to examine opioid records any time a doctor prescribes, or a pharmacist dispenses, an opioid drug. By linking information across states, medical doctors can quickly identify signs of misuse and abuse before writing prescriptions.
4. Criminal Justice Information Services (CJIS) Data. Several major criminal justice data users and collectors (e.g., courts, police districts) run on AWS through back-end support service companies which provide the vital data collection and tools for the criminal justice system to work safely and securely and in compliance with the FBI's CJIS security standards. Prominent examples include CentralSquare Technologies, the California Department of Justice's JusticeCloud, and several other state criminal justice systems.

Who can build it for me? Who can I ask about future questions?

Research Improving People's Lives can help you implement a RIPL Data-Driven Policy System at no cost, and answer further questions about the system. Contact connect@ripl.org with subject line "DDPS."

References

Bald, Anthony, Eric Chyn, Justine Hastings, and Margarita Machelett. “The Causal Impact of Removing Children from Abusive and Neglectful Homes.” NBER Working Paper No. 25419 (2019). https://www.nber.org/papers/w25419

Chyn, Eric, Justine Hastings, and Samantha Gold. “The Returns to Early-Life Interventions for Very Low Birth Weight Children.” NBER Working Paper No. 25753 (2019). https://www.nber.org/papers/w25753

Esposito, Francesco Maria, Diego Focanti, and Justine Hastings. “Effects of Photo ID Laws on Registration and Turnout: Evidence from Rhode Island.” NBER Working Paper No. 25503 (2019). https://www.nber.org/papers/w25503

Hastings, Justine, Mark Howison, and Sarah Inman. “Predicting High-Risk Opioid Prescriptions Before They Are Given.” NBER Working Paper No. 25791 (2019). https://www.nber.org/papers/w25791

Hastings, Justine, Mark Howison, Ted Lawless, John Ucles, and Preston White. “Unlocking Data to Improve Public Policy.” *Communications of the ACM* (forthcoming, 2019). https://doi.org/10.31219/osf.io/hdv3c

Howison, Mark, David Berenbaum, Miraj Shah. 2018. “RIPL Predictive Modeling Template.” GitHub Repository. https://github.com/ripl-org/predictive-template

Mann, Cynthia. 2014. “Reducing Nonurgent Use of Emergency Departments and Improving Appropriate Care in Appropriate Settings.” Center for Medicaid and CHIP Services Informational Bulletin (January 16, 2014). https://www.medicaid.gov/federal-policy-guidance/downloads/cib-01-16-14.pdf

Appendices

A1. Key Datasets

Data lakes run on joined administrative data from across government agencies. A wide variety of data and outcomes ensures that you can robustly measure results and impact. The core datasets for county- or state-level data lakes are:

Human services data.

1. Case benefit history, enrollment and eligibility records, and income for human services programs including the Supplemental Nutrition Assistance Program (SNAP), Temporary Assistance for Needy Families (TANF), Medicaid, Supplemental Security Income (SSI), child care programs, and general public assistance.
2. Child protective services case records.

Health and medical data.

1. Medicaid enrollment, eligibility, and claims records.
2. Birth and vital records.
3. Lead testing and vaccination records.

Labor and training data.

1. Wage earnings.
2. Unemployment insurance, temporary disability insurance (if applicable), and worker's compensation claims and payment records.
3. Work training program records.

Corrections and criminal justice data.

1. Incarceration, probation, and parole records.
2. Charges and sentences.
3. Police calls for service (*optional*).

Education data.

1. School enrollment and demographics.
2. Disciplinary infractions.
3. Standardized test scores.
4. High school graduation and post-secondary enrollment records.

A2. Governance Agreement Template

Memorandum of Understanding and Governance for the <<State/County Name>> Data Lake

This Memorandum of Understanding and Governance for <<State/County Name>> Data Lake ("Agreement") is intended to facilitate the creation of a research data lake ("Data Lake") of government program administrative records, optimized for delivering insights to guide data-driven policy using data shared by the undersigned agencies ("Agencies").

This Agreement sets out the terms and agreements for governance of the Data Lake and its use in compliance with all applicable federal, state and local laws, regulations and policies. The Agreement is made between the following Agencies:

1. <<INSERT AGENCY NAME>>
2. <<INSERT AGENCY NAME>>
3. <<CONTINUE AS NEEDED FOR ALL PARTICIPATING AGENCIES>>

The Agencies listed above may herein be referred to as "Party" or collectively referred to as the "Parties."

I. PURPOSE

WHEREAS, the Parties are agencies established under the laws of the <<STATE/COUNTY NAME>>; and shall contribute data to the Data Lake as specified in the Cooperation and Data Sharing Agreement Between <<State/County Name>> Agencies attached as appendices to this Agreement, and for the purpose of improving services to their citizens and furthering their Agency Missions; and

WHEREAS, the Parties recognize that many <<State/County Name>> citizens qualify for and participate in multiple State/County programs, and that creation of the Data Lake will break down information silos between departments; allow for research activities that utilize multi-agency data to give stronger insights into how to improve the effectiveness, efficiency, and value of State/County programs and services; improve the accessibility and management of information, allow for the creation of comprehensive facts, dashboards, applications and data-driven metrics that guide policy; support partnerships with external partners to use data to improve policy effectiveness and outcomes; and <<Insert other State/County reasons for creating a data lake here>>; and

WHEREAS, the Parties wish to enter into an agreement whereby the governance of the Data Lake and the execution of the aforementioned activities is fair, transparent, trusted, and rigorously enforced to the satisfaction of all parties;

NOW THEREFORE, in consideration of the mutual covenants, promises, and conditions herein contained, and for good and valuable consideration, the adequacy of which is hereby acknowledged, the Parties agree as follows:

II. DEFINITIONS

1. "Approved Research Projects" shall mean shall mean projects that were submitted to and approved by the Data Lake Governance Committee using the Project Charter Template in the Appendix A.

2. "Authorized Users" shall mean users who were approved by the Data Lake Governance Committee to securely access specific tables and data fields in the Data Lake in accordance with an Approved Research Project.

3. "Data Lake" shall mean a longitudinal database that contains anonymized and secured administrative data from Parties to this Agreement and the Cooperation and Data Sharing Agreement. Data are securely joined across agencies with an anonymous identifier, and which contains derived tables using the administrative records which are optimized for research to deliver policy insights.

4. "Data Lake Governance Committee" shall mean the committee established in this Agreement as the governance committee for the Data Lake.

5. "Holding Agency" shall mean the Agency appointed to hold the data contract for the secure cloud environment which houses the Data Lake.

6. "Project Charter" shall mean a proposed project that provides all the information listed in Appendix A. A Project Charter must include the purpose of the project, the Authorized Users undertaking the work, all of the data tables and columns needed to complete the work as well as which agency each table or column is from, the methods used, the deliverables, and a proposed timeline for the work, and it must be approved by all agencies who own data tables and columns needed to complete the work outlined.

7. “Sponsoring Agency” shall mean an Agency that approves the use of its data tables in a Project Charter.

III. TERMS AND CONDITIONS

III.A. Terms and Conditions for the Data Lake Hosting Environment and Architecture

1. Data Lake Hosting Environment Security: The Data Lake shall be housed by <<Insert Agency or Department here>> (“Holding Agency”) in a manner that is in accordance with all applicable laws, rules, and regulations. The Holding Agency will be responsible for hosting the Data Lake in a secure environment with the following features: FedRAMP approved secure cloud system which has appropriate administrative, physical, and technical safeguards based upon applicable laws, regulations, policies, or other rules that reasonably and appropriately protect the confidentiality, integrity, and availability of the data in the system and as specified in Cooperation and Data Sharing Agreements appended to this Agreement;

2. Data Lake Contents: The Data Lake must contain integrated and anonymized longitudinal database of administrative records provided by the Agencies in the Cooperation and Data Sharing Agreements appended to this Agreement. It must also contain derived tables which transform administrative data into formats and tables to support robust and reliable research insights.

3. Data Lake Refresh Frequency: The Data Lake will be updated <<monthly/quarterly/semi-annually>> with new data uploads from the Agencies. Individual agencies may have a different refresh frequency specified as an Appendix to this Agreement.

4. Data Documentation: The Data Lake must be accompanied by a data dictionary and code book which is automated and updated with each Data Lake refresh. The data dictionary should define each variable in the data lake, its values, and each table’s unique key. The code book should contain details on data construction and summary statistics for each variable.

5. Anonymization: The Data Lake must be anonymized, removing any personally identifiable information (“PII”) which would render and individual immediately recognizable. It must contain anonymous identifiers which allow Approved Users who are authorized to work on Approved Research Projects to join records associated with

the same individual across data tables and over time for the purposes of accomplishing the Approved Research Projects.

6. User Control: The Data Lake and the hosting environment must have user control protocols in place such that the Data Lake is used only by Authorized Users with Approved Project Charters, and it should restrict access for Authorized Users to only the pre-specified tables and columns in the data lake that are authorized for use in the Approved Project Charter.

7. Automated Logging and Alerting: The Data Lake environment will have automated logging software which is tamper proof and sends automated alerts for attempted unauthorized access. Log files are posted electronically for Agencies to review on a <<Weekly>> basis.

III.B. Terms and Conditions for the Data Lake Governance Committee

1. Data Lake Governance Committee: All undersigned Agencies have the right to participate in a Data Lake Governance Committee ("Committee"). The Committee will be chaired by a member of the Holding Agency. All Agencies will designate one (1) representative from the Agency to serve as a Committee member.

2. Committee Duties: The Data Lake Governance Committee will meet on a <<Insert time frame here; RIPL recommends quarterly or semi-annually>> basis to review proposed Project Charters, review needs, concerns, and summary documentation related to the functioning of the Data Lake and compliance with this agreement, as well as proposing any amendments to the Agreement if desired. Amendments to the Agreement must be agreed upon by a <<choose simple or super-majority>>.

III.C. Terms and Conditions for the Data Lake Management Team

1. Holding Agency Data Lake Management Team: The Holding Agency will be hire a Data Lake Management Team to run and maintain the Data Lake as specified in Section II.A. of this Agreement, including managing Project Charter review and approval processes (Section XXXX) and managing Approved User Access. The recommended staffing structure for the Data Lake Management Team is listed in Appendix B.

2. The Data Lake Management Team Responsibilities: The Data Lake Management Team will be responsible for the following duties:

a. Coordinate with Agencies to update the Data Lake on a <<insert agreed upon refresh rate here>>;
b. Review and update code books and data dictionaries on a <<insert agreed upon refresh rate here>>;
c. Ensure data is accurately and efficiently integrated into the Data Lake and transformed into derived tables optimized for research;
d. Work with Agencies to develop new data resources and derived tables along with data dictionaries and code books as Agencies request;
e. Manage permissions for Authorized Users in accordance with approved Project Charters, monitor and ensure compliance with Project Charters, and ensure documentation and required trainings for Authorized Users are complete and up-to-date;
f. Monitor and enforce all security protocols regarding the Data Lake, including conducting regular audits and penetration testing;
g. Collect proposed Project Charters and coordinate with internal and external individuals who plan to submit Proposed Project Charters;
h. Manage exports for the Data Lake and help coordinate release and dissemination of deliverables in compliance with all federal, state, and local privacy laws and the Cooperation and Data Sharing Agreements appended to this Agreement;
i. Collect regular reports from Authorized Users on approved Project progress in the format and periodicity agreed upon in each Approved Project Charter.
j. Be responsible for holding one (1) external audit of the Data Lake compliance with this Agreement every <<choose periodicity; RIPL recommends annual>>, and reporting on the results of the audit to the Committee.
k. Be responsible for updating the Committee during its <<choose periodicity here>> meetings to support the Committee in fulfilling its duties under this Agreement.

III.D. Terms and Conditions for Project Approval

1. Project Charter Sponsorship: Each Project Charter must have at least one Sponsoring Agency. The Project Charter must receive sponsorship from each Agency whose data tables it requests to use. If a Project Charter is Sponsored by an Agency, but additional Agency data are requested, the relevant additional Agencies may still decline the use of their data in a project.

2. Project Charter Proposals: Project Charters may be proposed by government employees, contractors, or external parties. Project Charters must follow the template laid out in Appendix A. Each Charter must include the purpose of the project, the Authorized Users undertaking the work, all of the data tables and columns needed to complete the work as well as which agency each table or column is from, the methods used, the deliverables, and a proposed timeline for the work.

3. Project Charter Approval: Charters must be approved by <<insert responsible party here; RIPL recommends the Director of the Agency or the Agency Research Director>>. Each approved Project Charter shall be appended to the Cooperation and Data Sharing Agreements of the Parties whose data is being utilized.

III.E. Terms and Conditions for Accessing the Data Lake

1. Data Lake Access: Access to the Data Lake is restricted to Authorized Users designated by an Approved Project Charter. Authorized Users shall be responsible for protecting the confidentiality of data when they are using it. The Holding Agency is responsible for monitoring and overseeing the use of the Data Lake, and for terminating user access to Agency data once an Approved Project they are working on is complete.

2. Authorized User Accounts: All Authorized Users must sign a Data Use Agreement as specified in the Cooperation and Data Sharing Agreement. Authorized User accounts may not be shared by multiple parties. Every individual accessing the Data Lake must have their own Authorized User account. User Accounts must be governed by a dual authentication process to ensure individual accounts are unique to approved individuals.

3. Reporting Data Lake Misuse: All Parties shall immediately notify the relevant Agencies and Holding Agency the Party believes that an Authorized User is improperly accessing, using, or disclosing data, or if an Authorized User is engaged in activities with the Data Lake outside of the scope of an approved Project Charter.

4. Investigating Data Lake Misuse: If notified that an Authorized User, Party, or anyone else may be improperly accessing, using, or disclosing data, the Holding Agency is responsible for immediately investigating the situation.

5. Consequences for Data Lake Misuse: The Holding Agency shall immediately remove an Authorized User's access to the Data Lake if they are found to be improperly accessing,

using, and/or disclosing data from the Data Lake. Any government-employed Authorized User employee who has engaged in improper use or disclosure of the Data Lake will be subject to his or her Agency's disciplinary process. Any external Authorized User who has engaged in similar conduct will have the improper use reported to their research institution or company. In all cases, the Authorized User shall have his or her access terminated immediately, and shall be temporarily or permanently banned from the use of the Data Lake as determined by the Data Lake Governance Committee.

6. Civil and Criminal Consequences for Data Lake Misuse: Any individual who has engaged in improper use or disclosure of Data Lake data may be subject to civil and/or criminal penalties.

IV. MISCELLANEOUS

IV.A. Additional Parties

1. Additional Parties: The undersigned Parties acknowledge that additional Parties may be added to this Agreement. All current Parties agree that, prior to admission of a new Party, the new Party must agree to be bound by the terms of this Agreement. An additional Party, if not a current signatory, shall stipulate to all the terms of this Agreement.

2. Liability: Nothing in this Agreement shall be deemed to impose responsibility or liability on a Party related to the accuracy, content, or completeness of any data provided pursuant to this Agreement or the Data Sharing and Coordination Agreements signed by the Parties. The Parties acknowledge that other Parties may be added or terminated at any time.

3. Assignability: This Agreement shall be binding upon and shall inure to the benefit of each Party and its assigns and successors in interest. This Agreement shall not otherwise be assignable or assigned by any Party without prior written approval by the others first being obtained.

4. Non-Waiver: The failure of any Party to exercise any of its rights under this Agreement for a breach thereof shall not be deemed to be a waiver of such rights, nor shall the same be deemed to be a waiver of any subsequent breach, either of the same provision or otherwise.

IV.B. Duration and Termination

1. Duration: This Agreement begins on the date that the last party signs this Agreement, which is the Effective Date. Subject to extension by mutual, written consent of the Parties, this Agreement shall remain in full force and effect for a period of ten (10) years beginning on the Effective Date. Agencies may terminate their participation in the Agreement by giving at least ninety (90) days prior written notice to the others. A single Party that terminates their participation in the Agreement does not terminate the Agreement for other Parties.

2. Severability: The provisions of this Agreement are severable. If any provision of this Agreement is held invalid, by any court that invalidity shall not affect the other provisions of this Agreement and the invalid provision(s) shall be considered modified to conform to the existing law.

3. Termination: If a Party determines that the Holding Agency has violated a material term of this Agreement, the Party may terminate their participation in this Agreement immediately or at the Party's discretion, by giving the Holding Agency a period of up to thirty (30) days to cure the violation or breach. The Party will notify the Holding Agency in either event of its decision of termination in writing. Upon request, the Holding Agency will submit a Corrective Action Plan outlining the steps that the Holding Agency took and will take to prevent a continuing and/or similar material breach in the future.

4. This Agreement may be terminated by any Party for default upon the defaulting Party's failure to cure a material breach within thirty (30) days after written notice by the non-defaulting Party specifying with sufficient detail the nature of the default.

5. This Agreement may be terminated by any Party by providing sixty (60) days written notice (or upon the greatest amount of notice allowed under the law or regulation) if a change of law or regulation necessitates that the Agreement be terminated to maintain any Party's compliance with such law or regulation.

6. Survival: Except as otherwise provided by law, the obligations of confidentiality imposed by this Agreement shall survive termination of this Agreement. With respect to trade secrets, the obligations shall last for so long as the information remains a trade secret.

IV. C. Interpretation and Modification

1. Headings: The headings of the sections are inserted for convenience of reference only and are not intended to be a part of or to affect the meaning or interpretation of this Agreement.

2. Entire Agreement; Modification: This Agreement, as well as the Cooperation and Data Sharing Agreement signed by the Parties, constitute(s) the entire understanding among the Parties with respect to the subject matter hereof and supersedes any and all prior understandings and agreements, oral and written, relating hereto. Any amendment hereof must be in writing signed by an authorized representative of all Parties.

Memorandum of Understanding and Data Lake Governance Signatures Page

<<INSERT AGENCY NAME>> AUTHORIZED AGENT

_____ _____

(Signature) (Date)

Name:

Title:

Acknowledged:

_____

<< INSERT AGENCY NAME HERE >> AUTHORIZED AGENT

_____ _____
(Signature) (Date)

Name:
Title:

<< CONTINUE AS NEEDED FOR AUTHORIZING AGENCIES >>

Appendix A: Project Charter Template

Project Summary Page (1-page maximum)

Project Title:

Principal and Co-Principal Investigators (include affiliations):

Project Time Period:

Partner Agencies:

Research Question: Summarize the research question and aims in 1-3 sentences. The summary should be understandable to readers who are not researchers.

Background: Summarize the importance of the project in 1-3 sentences. The summary should be understandable to readers who are not researchers.

Data: List the data needed for this project, and outcomes to be measured.

Research methods: Describe the research design in 1-3 sentences.

Intervention: If proposing an intervention, describe the intervention in 1-3 sentences. Leave blank if there is no proposed intervention.

Project Charter Template (5 pages maximum)

I. Research question and/or strategic aims

Provide an in-depth description of the research questions to be answered, specific hypotheses to be test, and strategic aims of the project. Applicants must include which agencies will be partners on this project, with the expectation that the agencies are aware of the submission of this proposal.

II. Background

Give a background on the importance of the challenge and prior literature on the topic. Make sure to include the genesis of the project in this state/municipality, and what the contribution of this project will be in the state.

III. Data

Provide a list of datasets and tables to be used in the project, and why they are necessary. A list of specific variables will be required in the appendix.

IV. Research Methods

Outline your principal hypothesis, questions to be explored, and how the data will be analyzed to answer each research question. What research methods, analytic models, or strategies will be used to answer the question? What outcomes will you measure? Are you investigating causal mechanisms? If you are proposing an intervention, lay it out in detail here. Write this section for a well-informed lay audience.

V. Investigators

What are the qualifications and responsibilities of the team conducting the research? Provide a list of everyone who will need access to the data and/or involved in the project. Give a brief description of how this project will be funded.

VI. Timeline and Deliverables

What are the final products that will be produced from this research? List how results will be reported, including scientific papers, conference presentations, and internal reports, and provide a detailed timeline for your proposed work.

VII. Publication Plan

If the deliverables will be publicly available or published as a scientific paper, you must include a publishing plan that includes a plan for which (if any) journals it will be submitted to, their replication policies, a timeline for preparing the paper for publication (including windows of opportunity for partners to review final drafts and suggest feedback), where replication files will be kept, and any other items agreed upon between partners.

VIII. Appendices

Attach the following as appendices. Appendices do not count towards the 10-page limit.

- **Exhibit A: Technical methodology (mandatory).** Attach a technical appendix, of no more than 5 pages, that goes in-depth into the research methods. Make sure to include empirical models, power calculations (if relevant), limitations to your approach, and how you will ensure robustness. This should be written for a scientific audience.
- **Exhibit B: Variables list (mandatory).** Attach a list of all variables needed. Note that when the charter is approved, access will be given only to these variables. Including more data in the future will require a revised project charter to be submitted and approved.
- **Exhibit C: CVs of investigators (mandatory).** CVs should not exceed 5 pages.
- **Exhibit D: Citations (mandatory).**
- **Exhibit E: Institutional support letter (mandatory).** Attach a letter of support from your affiliated organization.
- **Exhibit F: IRB letter (as needed).** Attach a letter of support from your affiliated IRB board, if required by the parameters of your research.
- **Exhibit G: Letters of support (optional).** Attach letters of support from agencies or other relevant partners.

Appendix B: Recommended Data Lake Management Team Staffing Structure

The following positions are recommended for a Data Lake Management Team:

1. *Senior project manager (PM).* This individual has business intelligence / data analytics experience. They will serve as the central coordinating point to create cross-agency project charters, reporting requirements for project leads, manage and execute data sharing agreements, manage and coordinate project reporting and deliverables, and sharing of learning across projects and agencies. The project manager will interface with a governance board if a cross-agency governance board is needed, driving alignment on priorities, needs, and results.
2. *Senior data engineer (DE).* This individual will be in charge of curating the data lake, updating codebooks, managing user accounts and access to particular tables and columns for each project charter the PM approves. They will be in charge of the periodic database updates, version control, archiving of data versions, database release notes, adding new data to the data lake. They will also be in charge of task management software, and ensuring chartered project partners comply with project documentation and security rules and requirements.
3. *Senior data scientist (DS).* This individual should have a background in Computer Science as well as in applications of machine learning and statistical modeling in a general field of the social sciences (e.g. Economics as opposed to Health). This person will be in charge of coordinating project execution with project leads and analysts from groups executing chartered projects, assisting in evaluating the merit of proposed projects, and providing technical expertise and guidance to partners working on chartered projects particularly in the fields of machine learning and App development. They will help advise and assist where needed the DE, providing input on data quality, derived table structure, and analysis. They will own software (e.g. R, Python, Stata) and BI tool interfaces (e.g. Power BI, Tableau), and ensure these are up to the latest-and-greatest technologies.
4. *Empirical economist (EE, optional).* This individual should have a PhD in Economics with a specialty in empirical microeconomics applied to social policy (Public Economics). This individual will assist the DS and DE to coordinate project execution with project leads and analysts from groups executing project charters, evaluate the merit of proposed projects with an eye towards economics (causal identification, selection bias, market response) and advise partners on chartered projects as they develop. They will help advise and assist where needed the DE and DS, providing input on data quality, derived table structure and analysis. They will coordinate with the DS on needed software to support economic and policy analysis.

A3. Data Sharing Agreement Template

Cooperation and Data Sharing Agreement Between <<State/County Name>> Agencies

This Cooperation and Data Sharing Agreement ("Agreement") is intended to facilitate data sharing between <<Insert Agency Name and Address>> ("Sponsoring Agency") and <<Insert Agency Name and Address Appointed to Hold the Data Contract>> ("Holding Agency"), in compliance with all applicable federal, state and local laws, regulations, and policies. The Sponsoring Agency and the Holding Agency may herein be referred to as "Party" or collectively referred to as the "Parties."

1.0 Purpose

WHEREAS, the Sponsoring Agency is an agency established under the laws of the <<STATE/COUNTY NAME>> whose mission is to <<INSERT AGENCY MISSION>> ("Mission"); and

WHEREAS, the Sponsoring Agency possesses or shall possess administrative data related to <<INSERT NAME/TYPE OF AGENCY DATA >>; and

WHEREAS the Sponsoring Agency, pursuant to <<STATE/COUNTY NAME>> laws, <<IF APPLICABLE, ENTER APPLICABLE LAW OR EXECUTIVE ORDER GUIDING DATA USE>> may share data for the purposes of research to further the mission of the Sponsoring Agency;

WHEREAS, the Parties are interested in supporting research into and development of policies which could further their Missions; and

WHEREAS, the Parties believe that contributing their data to a secure Data Lake housed by the Holding Agency for research projects approved by the <<STATE/COUNTY>> Data Lake Governance Committee established under the Memorandum of Understanding and Governance for the <<State/County Name>> Data Lake ("Approved Research Projects") could contribute to the fulfilment of the Parties' Missions and better serve the people of <<STATE/COUNTY>>; and

WHEREAS, the Sponsoring Agency wishes to enter into an Agreement whereby it shall provide information in the form of electronic data, to the Holding Agency, which shall be combined with data the Holding Agency receives from other Parties into an anonymized, longitudinal research Data Lake which will enable users authorized by the Data Lake governance committee established under the Memorandum of Understanding and Governance for the <<State/County Name>> Data Lake ("Authorized Users") and the Sponsoring Agency to conduct Approved Research Projects to further the Sponsoring Agency's Mission;

NOW THEREFORE, in consideration of the mutual covenants, promises, and conditions herein contained, and for good and valuable consideration, the adequacy of which is hereby acknowledged, the Parties agree as follows:

2.0 Definitions

The following terms shall have the following meanings:

2.1 "Agency Data" shall mean information or data gathered, maintained and stored by the Sponsoring Agency and/or the Sponsoring Agency's designee(s) in the normal course of carrying out its stated purposes and covered under the scope of this agreement. Sponsoring Agency Data may include PII and/or CI, and include data related to:

- << INSERT DATA SCOPE HERE WHICH GIVES GENERAL DESCRIPTION OF DATA TO BE USED UNDER THE AGREEMENT>>

2.2 "Approved Research Projects" shall mean projects that were submitted to and approved by the Data Lake Governance Committee using the Project Charter Template in the Memorandum of Understanding and Governance for the <<State/County Name>> Data Lake, and which shall be appended to this Agreement.

2.3 "Authorized Users" shall mean users who were approved by the Data Lake Governance Committee to securely access specific tables and data fields in the Data Lake in accordance with an Approved Research Project.

2.4 "Confidential Information" (CI) shall mean: (i) all PII; (ii) any personal information protected under federal or <<STATE/COUNTY>> law; and, (iii) all trade secret information which is disclosed by a Party, or otherwise acquired by a receiving Party, in the course of the Parties' discussions or other activities pursuant to this Agreement and which is marked "Confidential" by the disclosing Party at the time of disclosure or is orally identified as confidential by the disclosing Party at the time of disclosure and written notice reiterating its confidentiality is sent to the receiving Party within thirty (30) days of disclosure thereof. Each Party shall have the right, upon written request, to require that the other Parties confirm in writing whether any such information disclosed but not identified as CI is to be considered CI hereunder. By way of illustration, but not limitation, CI includes data with PII which is protected by law, other information protected by law, trade secrets, processes, formulae, data, know-how, products, designs, drawings, computer aided design files and other computer files, computer software, bills of materials, ideas, improvements, inventions, training methods and materials, manufacturing processes, sales information, marketing techniques, plans, strategies, budgets, financial information, forecasts, customer lists and pricing policies. Except as otherwise provided by law, CI, however, shall not include any information which:

i. is or hereafter becomes known and legally available to the general public through a party who had the right to make the information available to the general public and through no act or omission of the receiving Party which is, directly or indirectly, in violation of the receiving Party's obligations under this Agreement;

ii. is subsequently disclosed without restriction to the receiving Party by a third-party who had the right to make such disclosure and who did not,

directly or indirectly, receive such information through a Party who was obligated not to disclose the same;

iii. is required to be disclosed by any applicable judgment, order or decree of any court having jurisdiction; provided that in connection with any such disclosure, the receiving Party shall use its best efforts to give to the disclosing Party reasonable prior notice of such required disclosure if such required disclosure will include any Confidential Information; or

iv. is disclosed to a third-party on a non-confidential basis by the Party who owns such information.

2.5 <<OPTIONAL – PARTIES DEFINE ANY OTHER PROTECTED DATA TYPES HERE IF APPLICABLE >>

2.6 "Data Lake" refers to a longitudinal database that contains anonymized and secured administrative data from agencies which are party to the Memorandum of Understanding and Governance for the <<State/County Name>> Data Lake and which can be securely joined across agencies with an anonymous identifier, and which contains derived tables using the administrative records which are optimized for research to deliver policy insights to improve the effectiveness of public policy by furthering the missions of the respective agencies.

2.7 "Data Lake Governance Committee" shall mean the committee established by the Memorandum of Understanding and Governance for the <<State/County Name>> Data Lake to manage governance of the Data Lake.

2.8 "Personally Identifiable Information" (PII) shall mean data which include any of the following: (1) the name of an individual or that of the individual's parents or guardians, (2) social security number, (3) specific home address, (4) driver's license number, (5) client or personal identification number, or (6) a list of personal characteristics or other information which would make the individual's identity immediately traceable.

3.0 Protection of CI and PII

3.1 Protection of CI and PII. Each Party shall take all reasonable steps to protect the confidentiality of another Party's CI, specifically including, but not limited to the Holding Agency maintaining Sponsoring Agency Data solely on a secure, dedicated computing environment with security and access safeguards and policies stipulated in Appendix B and Appendix C.

3.2 Access to CI and PII. Sponsoring Agency Data will be analyzed solely on a secure, dedicating cloud computing environment with security and access safeguards and policies stipulated in Appendix B and Appendix C. Sponsoring Agency Data will be accessed and analyzed only by Authorized Users for Approved Research Projects.

Authorized Users must complete the Data Use Agreement and Human Subjects Protection Training as stipulated in Appendix A. A copy of all completed Data Use Agreements and certificates of Human Subjects Protection Training will be made available to a Party upon its request.

3.3 Disclosure. The receiving Party shall not, without the prior written consent of the disclosing Party:

i. disclose the CI to any other person or entity except to the Sponsoring Agency, or their agents, who have a "need to know" such CI in the course of the performance of their duties, who are informed of the obligations of this Agreement, and who have executed a confidentiality agreement with the Parties;

ii. copy or reproduce the CI except as is necessary to further the objectives of the relationship between the Parties hereto as laid out explicitly in writing between the two Parties (all such copies, however, shall include a confidential identification marking and shall be governed hereby); or

iii. use the CI except to further the research goals to conduct research set forth under the Purpose.

3.4 Security and Anonymization of Agency Data. The Holding Agency agrees to transfer, store and allow use of Party Data by Authorized Users for Approved Research Projects only, including PII and CI, in accordance with security protocols set forth in Appendices A through F to this Agreement.

3.5 Cell Size Suppression Policy. Agencies agree that any use of Sponsoring Agency data in the creation of any document (manuscript, table, chart, study, report, etc.) concerning the purpose specified in this Agreement (regardless of whether the report or other writing expressly refers to such purpose, to the Sponsoring Agency, or to the files specified in this Agreement or any data derived from such files) must adhere to the federal current cell size suppression policy. This policy stipulates that no cell less than 11 may be displayed publicly. Also, no use of percentages or other mathematical formulas may be used if they result in the public display of a cell less than 11.

<< IF APPLICABLE, AGENCY INSERT ANY OTHER PROTECTIONS FOR ADDITIONAL PROTECTED DATA TYPES DEFINED IN SECTION 2 >>

4.0 Use, Term and Termination.

4.1 Use. The Holding Agency represents that the Party Data furnished to the Holding Agency will be used solely for the research purposes and/ or services specified in this Agreement. Parties may only use Party data for Approved Research Projects. The Parties shall not disclose, release, reveal, show, sell, rent, lease, loan, submit or present for scholarly review, publish or otherwise grant access to the data covered by this Agreement to any person that does not have a need to know the information, based upon the Purpose of this agreement.

4.2 Approved Projects. Parties agree that use of the Sponsoring Agency's data in the Data Lake will only be allowed to execute Approved Research Projects. Approved Research Projects must be approved by the Data Lake Governance Committee and the Sponsoring Agency, using the process stipulated in the <<GOVERNING MOU>>, and the approved charter must be appended to this Agreement.

4.3 Authorized Users. Authorized Users may only be given access to the data tables and fields explicitly outlined in the Project Charter submitted to the Data Lake Governance Committee. Use of additional data tables and fields must be approved by the Data Lake Governance Committee.

4.2 Publishing Results. Some Authorized Users and research partners who conduct Approved Research Projects will be researchers who are members of universities. Universities, as institutions of higher education, engage only in research that is compatible, consistent, and beneficial to their academic role and mission. Therefore, significant results of research activities conducted in partnership with academics must be reasonably available for publication. The Parties acknowledge that the rights to publish the results of research should be determined during the project approval process, and that if it is determined that the research must be reasonably available for publication, that nothing in this Agreement shall be interpreted to restrict this right of publication; provided that no PII or CI will be included in any published material without prior express, written approval by the relevant Agency.

4.3 Duration. This Agreement begins on the date that the last party signs this Agreement, which is the Effective Date. Subject to extension by mutual, written consent of the Parties, this Agreement shall remain in full force and effect for a period of ten (10) years beginning on the Effective Date. Agencies may terminate their participation in the Agreement by giving at least ninety (90) days prior written notice to the others. A single Party that terminates their participation in the Agreement does not terminate the Agreement for other Parties.

4.4 Termination. If the Sponsoring Agency determines that the Holding Agency has violated a material term of this Agreement, the Sponsoring Agency may terminate this agreement immediately or at the Sponsoring Agency's discretion, by giving the Holding Agency a period of up to thirty (30) days to cure the violation or breach. The Sponsoring Agency will notify the Holding Agency in either event of its decision of termination in writing. Upon request, the Holding Agency will submit a Corrective Action Plan outlining the steps that the Holding Agency took and will take to prevent a continuing and/or similar material breach in the future.

4.5 This Agreement may be terminated by any Party for default upon the defaulting Party's failure to cure a material breach within thirty (30) days after written notice by the non-defaulting Party specifying with sufficient detail the nature of the default.

4.6 This Agreement may be terminated by any Party by providing sixty (60) days written notice (or upon the greatest amount of notice allowed under the law or regulation) if

a change of law or regulation necessitates that the Agreement be terminated to maintain any Party's compliance with such law or regulation. In such case, the Parties will work in a cooperative manner to maintain, return or destroy Sponsoring Agency Data, PII and CI defined herein.

4.7 Survival. Except as otherwise provided by law, the obligations of confidentiality imposed by this Agreement shall survive termination of this Agreement. With respect to trade secrets, the obligations shall last for so long as the information remains a trade secret.

4.8 Data destruction. Upon termination, Holding Agency will return or destroy all Sponsoring Agency Data and CI and will not retain, nor allow any of its agents or subcontractors to retain, any Sponsoring Agency Data except for Anonymous Data or aggregated data derived from Sponsoring Agency Data used for purposes of producing and/or replicating research results for approved projects produced under this agreement. Holding Agency's duty to destroy Sponsoring Agency original data includes, but is not limited to, the obligations to destroy all copies of Sponsoring Agency Data including backup tapes and other electronic backup medium, and to destroy all Sponsoring Agency Data by "clearing" (which requires a minimum of three (3) passes), "purging" or "physically destroying" Sponsoring Agency Data in accordance with the National Institute of Standards and Technology (NIST) Special Publication 800-88 or in another manner approved in advance by Sponsoring Agency. Holding Agency will certify in writing to Sponsoring Agency that Holding Agency (including its agents and subcontractors) has returned or destroyed all Sponsoring Agency original data as this Agreement requires. If Sponsoring Agency agrees that the return or destruction of Sponsoring Agency Data is infeasible and determines that Holding Agency's written plan to safeguard the confidentiality and security of Sponsoring Agency Data is acceptable, Agency may permit Holding Agency to retain Sponsoring Agency Data for the specific and limited purpose that makes return or destruction of Sponsoring Agency Data infeasible. This written plan of retention must be submitted and approved by Sponsoring Agency for Sponsoring Agency Data to be retained.

5.0 Data from Sponsoring Agency to Holding Agency.

5.1 Sponsoring Agency shall provide to the Holding Agency, or provide access for the Holding Agency to receive, original Sponsoring Agency Data under the scope of data defined in paragraph 2.4 of this Agreement. Said updates will be provided according to mutually agreed upon timing, stipulated in Appendix D.

5.2 All Sponsoring Agency Data shall be in an accessible form and format transferred to the Holding Agency via protocol set forth in Appendix D.

5.3 Sponsoring Agency represents and agrees that it has the right to contribute and to disclose Sponsoring Agency Data to Authorized Users pursuant to this Agreement and

that Authorized Users have the right to use Sponsoring Agency Data, but only as set forth in this Agreement and in accordance with an Approved Research Project.

5.4 Sponsoring Agency agrees to cooperate with the Researcher to provide Sponsoring Agency Data in the most reasonably useful format. Such cooperation shall include, but not be limited to, providing information needed for the Data Lake software to integrate Sponsoring Agency Data into the Data Lake to make it clearly usable by Authorized Users. Sponsoring Agency also agrees to communicate and aid in the Holding Agency understanding any changes Sponsoring Agency makes regarding the method by which it gathers, maintains, sorts, or develops Sponsoring Agency Data.

6.0 Custodians and Points of Contact.

6.1 Governing Law; Forum. This Agreement shall be governed by and construed under the laws of the <<INSERT COUNTY/STATE>>, which shall be the forum for any lawsuit arising from or incident to this Agreement.

6.2 Custodian. The parties mutually agree that the following named individual is designated as "Custodian" of the file(s) on behalf of the Holding Agency, and will be personally responsible for the observance of all conditions of use of the data, and for the establishment and maintenance of security arrangements as specified in this Agreement to prevent unauthorized use of disclosure. The Custodian, as designated in this Agreement, has the authority to represent within his/her organization; and is responsible for all use of the data specified in Paragraph(s) 2. The Holding Agency agrees to notify the Parties in writing within FIVE (5) business days of any change of Custodianship.

<<HOLDING AGENCY NAME>> Custodians

<<NAME>>
<<TITLE>>

<<NAME>>
<<TITLE>>

6.3 Points of Contact. The Parties mutually agree that the following-named individuals will be designated as “Point of Contact” for the Agreement on behalf of the Sponsoring Agency:

Agency Point of Contact

<< INSERT HERE >>>

OR

His/her designee(s)

<< INSERT HERE >>

The Parties mutually agree that the following- named individual will be designated as "Point of Contact" for the Agreement on behalf of the Holding Agency:

<<NAME>>

<<TITLE>>

7.0 Mitigation

7.1 The Holding Agency retains the sole right, and has the obligation, to establish appropriate security mechanisms and policies to protect Sponsoring Agency Data during transfer and hosting of the data in the Data Lake for use for approved projects. Holding Agency will notify Sponsoring Agency in the event of a known data breach or compromise of the system where Sponsoring Agency data is stored.

7.2 In the event of breach or suspected breach, the Holding Agency shall, within twenty-four (24) hours notify the Sponsoring Agency security officer and/or the appropriate system manager or the person designated as authorized agent of this Agreement of any unauthorized use or disclosure, suspected breach or breach of, or access to the aforesaid data. The notice shall contain all information available to the Holding Agency at the time of the notification to aid the Sponsoring Agency in examining the matter and will be supplemented by Holding Agency as additional information is obtained. The Holding Agency will preserve forensic evidence relating to any breach, including log report data to be shared with the Parties. The Parties will meet to jointly develop an incident investigation and remediation plan.

7.3 Parties agree that Sponsoring Agency shall be held harmless in the event of an unauthorized access or disclosure of Sponsoring Agency Data shared by the Sponsoring Agency with the Holding Agency. The Party responsible for the unauthorized access or disclosure agrees to take all such steps as are legally necessary which are associated with any harmful effects that may or have been caused by an unauthorized disclosure or access of data obtained from the Agency data file provided to the Holding Agency.

7.4 The Holding Agency and/or responsible Party will take steps to prevent a continuing and/or similar breach should one occur.

8.0 Miscellaneous.

8.1 Third Party Beneficiaries: Each Party hereto intends that this Agreement will not benefit or create any right or cause of action in or on behalf of any person other than the Parties hereto.

8.2 Assignability: This Agreement shall be binding upon and shall inure to the benefit of each Party and its assigns and successors in interest. This Agreement shall not otherwise be assignable or assigned by any Party without prior written approval by the others first being obtained.

8.3 Severability. The terms of this Agreement are severable, such that if any term or provision is declared by a court of competent jurisdiction to be illegal, void, or unenforceable, the remainder of the provisions shall continue to be valid and enforceable.

8.4 Non-Waiver. The failure of any Party to exercise any of its rights under this Agreement for a breach thereof shall not be deemed to be a waiver of such rights, nor shall the same be deemed to be a waiver of any subsequent breach, either of the same provision or otherwise.

8.5 Headings. The headings of the sections are inserted for convenience of reference only and are not intended to be a part of or to affect the meaning or interpretation 'of this Agreement.

8.6 Entire Agreement; Modification. This Agreement (and its attachments of Appendix A, Appendix B and Appendix C) as well as the Memorandum of Understanding signed by the Parties, constitute(s) the entire understanding among the Parties with respect to the subject matter hereof and supersedes any and all prior understandings and agreements, oral and written, relating hereto. Any amendment hereof must be in writing signed by an authorized representative of all Parties.

COOPERATION AND DATA SHARING AGREEMENT SIGNATURE PAGE

<<INSERT AGENCY NAME>> AUTHORIZED AGENT

_____ _____
(Signature) (Date)

Name:

Title:

Acknowledged:

_____

<< INSERT AGENCY NAME HERE >> AUTHORIZED AGENT

_____ _____
(Signature) (Date)

Name:

Title:

Appendix A

Security Pledge for the Use of Confidential Data from <<INSERT AGENCY NAME HERE>> ("Sponsoring Agency")

I, _____, through my involvement with and work on << Insert Approved Research Project title here>> will have access to the secure data provided by the Sponsoring Agency to the Hosting Agency to be used in producing research results. I understand that access to this confidential data carries with it the responsibility to guard against unauthorized use and the possibility of unauthorized access or use. To treat information as confidential means not to divulge it to anyone who is not an Approved User, or to cause it to be accessible to anyone who is not an Approved User.

I understand that disclosing confidential information directly or allowing non-authorized access to such information my subject me to criminal prosecution and/or civil recovery.

I agree to fulfill my responsibilities on this project in accordance with the following guidelines:

1. I agree to not permit non-project personnel to access these sensitive data, either electronically or in hard copy.
2. I agree to not attempt to identify individuals, families, or households.
3. I agree that in the event an identity of an individual, family, or household is discovered inadvertently, I will (a) make no use of this knowledge, (b) advise << Insert Supporting Agency Point of Contact and Hosting Agency Point of Contact>> of the incident, (c) safeguard or destroy the information as directed by << Hosting Agency Point of Contact>>, (d) not inform any other person of the discovered identity.

Researcher: **Witness:**

Name: _____ Name: _____

Signature: _____ Signature: _____

Date: _____ Date: _____

Appendix B

Storage of Sponsoring Agency Data

Data Lake GovCloud Environment Requirements:

<< Sample language if using RIPL-AWS DDPS Quickstart:
Holding Agency will make data accessible to approved users under this Agreement through a Research Data Lake (RDL) hosted in the Amazon Web Services cloud.

The cloud architecture for the RDL will inherit all of the physical and data center security controls that Amazon Web Services has developed to meet FedRAMP "medium" standards. Additional security controls for the RDL will include the following:

- *Holding agency approves all accounts and access to the RDL.*
- *Access is comprehensively audited. Logs are owned by Holding Agency and available for review at any time.*
- *Users have no access to the Internet from within the RDL.*
- *No data are able to leave the system without documented approval from Holding Agency.*
    - *Copy and paste are disabled in the client used to remotely access the RDL.*
    - *Only research products such as aggregated data, tables, plots, and regression output will be approved for export.*
- *Holding Agency will import administrative data will be imported into the RDL through an automated process that removes personally identifiable information (PII) and replaces it with a global anonymous identifier.*
    - *When manual intervention is required to troubleshoot or complete the anonymization process, security controls will limit access to the minimum level possible, and will keep complete audit logs of those activities.*
- *Holding Agency will have the ability to integrate and anonymize administrative data from Sponsoring Agencies for projects of joint interest, using inter-agency Data Sharing Agreements.>>*

Human Subjects Protections Training:

All Authorized Users for Approved Research Projects covered under this agreement will complete Human Subjects Protections Training from the below approved lists, and will keep certification up-to-date on an annual basis.

Appendix C
Replacing Personal Identifiers

Data received from the Sponsoring Agency will include an identification number to link households over time and across Sponsoring Agency Data. Sponsoring Agency Data may also include sensitive personal identifiers such as social security numbers or names or addresses if determined to be needed for linking individuals across administrative data sets in the Data Lake. Sensitive personal identifiers in the Sponsoring Agency data will be replaced with a scrambled identifier once the data are received by the Hosting Agency. Any personally identifiable information will be stripped from the data immediately once they are no longer needed for joining data sets. The crosswalk from the personal identifiers in the original Sponsoring Agency Data to the scrambled identifier will reside in an encrypted file system that uses a two-party password with a minimum of 10 characters in each party's password. The first party consists of the designated Sponsoring Agency Point of Contact and his or her designee as a backup, and the Hosting Agency Point of Contact and his or her designee as a back-up. Therefore, no individual person can decrypt or view the crosswalk files without full knowledge of all Parties. Whenever the crosswalk is decrypted or accessed for processing, both parties are present for the entire session, and the decrypted data is shredded and wiped at the end of the session. In addition, any time the encrypted crosswalk files are accessed, the Data Lake automated monitoring software will send a real-time email alert to all parties.

Appendix D

Procedure for Transferring Files from the Sponsoring Agency to Hosting Agency

The Sponsoring Agency will provide updated data to the Hosting Agency at a frequency of << Insert agreed upon frequency here >>, but no less than quarterly, via one of the below approved methods:

1) SFTP: The Sponsoring Agency will encrypt the data using AES-256 encryption, then use the secure copy (scp) or secure FTP (sFTP) protocol to transfer the encrypted data from the Sponsoring Agency's host system to a staging area in the Hosting Agency's secure cloud computing environment designated for hosting the Data Lake. A password or key for decrypting the file will be established prior to the first transfer, and any changes to the password or key will be communicated by the Sponsoring Agency to Hosting Agency by phone. All transfers will take place over a dedicated VPN or network path between the Sponsoring Agency and the Hosting Agency, which will be established and tested prior to the first transfer. All transfers will be logged. Once data has been transferred to the staging area, it will be copied by the Hosting Agency into an encrypted file system inside the Data Lake environment, and the original data in the staging area will be destroyed. Real time alerting will be sent by the secure environment's logging software to the Point of Contacts for the Hosting Agency and the Sponsoring Agency during the transfer process. The Hosting Agency will provide a chain-of-custody report to the Sponsoring Agency showing the complete audit history of the data while it is in the staging area, and confirming that the original data were destroyed.

2) << Insert additional alternatives here as needed>>

APPENDIX E
<< AGENCY INSERT ANY ADDITIONAL FORMS OR REQUIREMENTS HERE>>

APPENDIX F
<< AGENCY APPENDS APPROVED PROJECT CHARTERS HERE>>

A4. Project Charter Template

Project Summary Page (1 page maximum)

Project Title:

Principal and Co-Principal Investigators (include affiliations):

Project Time Period:

Partner Agencies:

Research Question: Summarize the research question and aims in 1-3 sentences. The summary should be understandable to readers who are not researchers.

Background: Summarize the importance of the project in 1-3 sentences. The summary should be understandable to readers who are not researchers.

Data: List the data needed for this project, and outcomes to be measured.

Research methods: Describe the research design in 1-3 sentences.

Intervention: If proposing an intervention, describe the intervention in 1-3 sentences. Leave blank if there is no proposed intervention.

Project Charter Template (5 pages maximum)

IX. Research question and/or strategic aims

Provide an in-depth description of the research questions to be answered, specific hypotheses to be test, and strategic aims of the project. Applicants must include which agencies will be partners on this project, with the expectation that the agencies are aware of the submission of this proposal.

X. Background

Give a background on the importance of the challenge and prior literature on the topic. Make sure to include the genesis of the project in this state/municipality, and what the contribution of this project will be in the state.

XI. Data

Provide a list of datasets and tables to be used in the project, and why they are necessary. A list of specific variables will be required in the appendix.

XII. Research Methods

Outline your principal hypothesis, questions to be explored, and how the data will be analyzed to answer each research question. What research methods, analytic models, or strategies will be used to answer the question? What outcomes will you measure? Are you investigating causal mechanisms? If you are proposing an intervention, lay it out in detail here. Write this section for a well-informed lay audience.

XIII. Investigators

What are the qualifications and responsibilities of the team conducting the research? Provide a list of everyone who will need access to the data and/or involved in the project. Give a brief description of how this project will be funded.

XIV. Timeline and Deliverables

What are the final products that will be produced from this research? List how results will be reported, including scientific papers, conference presentations, and internal reports, and provide a detailed timeline for your proposed work.

XV. Publication Plan

If the deliverables will be publicly available or published as a scientific paper, you must include a publishing plan that includes a plan for which (if any) journals it will be submitted to, their replication policies, a timeline for preparing the paper for publication (including windows of opportunity for partners to review final drafts and suggest feedback), where replication files will be kept, and any other items agreed upon between partners.

XVI. Appendices

Attach the following as appendices. Appendices do not count towards the 10-page limit.

- **Exhibit A: Technical methodology (mandatory).** Attach a technical appendix, of no more than 5 pages, that goes in-depth into the research methods. Make sure to include empirical models, power calculations (if relevant), limitations to your approach, and how you will ensure robustness. This should be written for a scientific audience.
- **Exhibit B: Variables list (mandatory).** Attach a list of all variables needed. Note that when the charter is approved, access will be given only to these variables. Including more data in the future will require a revised project charter to be submitted and approved.
- **Exhibit C: CVs of investigators (mandatory).** CVs should not exceed 5 pages.
- **Exhibit D: Citations (mandatory).**
- **Exhibit E: Institutional support letter (mandatory).** Attach a letter of support from your affiliated organization.
- **Exhibit F: IRB letter (as needed).** Attach a letter of support from your affiliated IRB board, if required by the parameters of your research.
- **Exhibit G: Letters of support (optional).** Attach letters of support from agencies or other relevant partners.

A4. Project Charter Template

Project Summary Page (1 page maximum)

Project Title:

Principal and Co-Principal Investigators (include affiliations):

Project Time Period:

Partner Agencies:

Research Question: Summarize the research question and aims in 1-3 sentences. The summary should be understandable to readers who are not researchers.

Background: Summarize the importance of the project in 1-3 sentences. The summary should be understandable to readers who are not researchers.

Data: List the data needed for this project, and outcomes to be measured.

Research methods: Describe the research design in 1-3 sentences.

Intervention: If proposing an intervention, describe the intervention in 1-3 sentences. Leave blank if there is no proposed intervention.

Project Charter Template (5 pages maximum)

XVII. Research question and/or strategic aims

Provide an in-depth description of the research questions to be answered, specific hypotheses to be test, and strategic aims of the project. Applicants must include which agencies will be partners on this project, with the expectation that the agencies are aware of the submission of this proposal.

XVIII. Background

Give a background on the importance of the challenge and prior literature on the topic. Make sure to include the genesis of the project in this state/municipality, and what the contribution of this project will be in the state.

XIX. Data

Provide a list of datasets and tables to be used in the project, and why they are necessary. A list of specific variables will be required in the appendix.

XX. Research Methods

Outline your principal hypothesis, questions to be explored, and how the data will be analyzed to answer each research question. What research methods, analytic models, or strategies will be used to answer the question? What outcomes will you measure? Are you investigating causal mechanisms? If you are proposing an intervention, lay it out in detail here. Write this section for a well-informed lay audience.

XXI. Investigators

What are the qualifications and responsibilities of the team conducting the research? Provide a list of everyone who will need access to the data and/or involved in the project. Give a brief description of how this project will be funded.

XXII. Timeline and Deliverables

What are the final products that will be produced from this research? List how results will be reported, including scientific papers, conference presentations, and internal reports, and provide a detailed timeline for your proposed work.

XXIII. Publication Plan

If the deliverables will be publicly available or published as a scientific paper, you must include a publishing plan that includes a plan for which (if any) journals it will be submitted to, their replication policies, a timeline for preparing the paper for publication (including windows of opportunity for partners to review final drafts and suggest feedback), where replication files will be kept, and any other items agreed upon between partners.

XXIV. Appendices

Attach the following as appendices. Appendices do not count towards the 10-page limit.

- **Exhibit A: Technical methodology (mandatory).** Attach a technical appendix, of no more than 5 pages, that goes in-depth into the research methods. Make sure to include empirical models, power calculations (if relevant), limitations to your approach, and how you will ensure robustness. This should be written for a scientific audience.
- **Exhibit B: Variables list (mandatory).** Attach a list of all variables needed. Note that when the charter is approved, access will be given only to these variables. Including more data in the future will require a revised project charter to be submitted and approved.
- **Exhibit C: CVs of investigators (mandatory).** CVs should not exceed 5 pages.
- **Exhibit D: Citations (mandatory).**
- **Exhibit E: Institutional support letter (mandatory).** Attach a letter of support from your affiliated organization.
- **Exhibit F: IRB letter (as needed).** Attach a letter of support from your affiliated IRB board, if required by the parameters of your research.
- **Exhibit G: Letters of support (optional).** Attach letters of support from agencies or other relevant partners.

A5. Example Project Charter

Project Summary Page (1-page maximum)

Project Title: The Impact of Foster Care: Finding What Works for Vulnerable Children

Sponsoring Agencies: State Department of Youth and Family Services (DYFS); State Department of Education (DoE); State Department of Human Services (DHS)

Requested Authorized Users:

- Dr. Jane Doe (*Brown University,* Professor of Economics; Faculty Affiliate, *Research Improving People's Lives*). Principal Investigator.
- Dr. John Smith (*University of New Hampshire,* Assistant Professor of Economics; Faculty Affiliate, *Research Improving People's Lives*). Principal Investigator.
- Research Assistant #1 (*Research Improving People's Lives,* Research Assistant)
- Research Assistant #2 (*Research Improving People's Lives,* Research Assistant)

Project Time Period: 01/2019-12/2020

Research Question: What is the causal impact of placing a child in foster care on their later-life outcomes?

Background: This project will use large-scale administrative data to measure the causal impact of foster care on education and later-life outcomes, helping policymakers develop programs to improve child outcomes and increase equity in communities.

Data: We will use data from the Departments of Youth and Family Services; Education; and Human Services. Our outcomes include academic performance (including school attendance, enrollment in special education, high school completion, and standardized reading and math test scores), teenage pregnancy, post-secondary enrollment, and juvenile convictions.

Research methods: Our research design uses the foster care placement tendency of quasi-randomly assigned child protective investigators as an instrumental variable. We innovate relative to prior studies by studying effects of foster care placement for children removed at young ages. While 50 percent of children removed from their homes are under the age of six (U.S. HHS 2016), prior studies have focused on studying home removal and foster care for relatively old children. A large literature on child development suggests that these estimates may not generalize to children removed at younger ages.

Project Charter (5 pages maximum)

I. Research question and/or strategic aims

Provide an in-depth description of the research questions to be answered or use case for the Data Lake, specific hypotheses to be test, and/or strategic aims of the project. Charters must include the Sponsoring Agency or Agencies, which must approve this Project Charter for it to be considered by the Data Lake Governance Committee.

Research Improving People's Lives (RIPL) and the Department of Youth and Family Services (DYFS) are partnering to use existing administrative data to understand the causal impacts of foster care placement on child outcomes, and translate research findings into more effective and efficient public policies and programs for low-income families and underserved communities. We aim to accomplish the following in this research project:

Strategic Aim 1: Use instrumental variables and machine learning to measure the causal impact of foster care placement on low-income youth life outcomes, such as school performance, incarceration and teen parenting, and understand how this impact relates to child experiences in the system by age, placement, and sociodemographic status. We will answer the following major research questions:

1. What is the causal impact of foster care placement on child achievement as measured by educational outcomes (standardized test scores, school attendance and completion, and post-secondary attainment), juvenile detention and teen parenting?
2. Do impacts of foster care placement on achievement vary by gender?
3. Do impacts of foster care placement vary by the age at which children are placed in care?

Strategic Aim 2: Engage in partnership with DYFS to assist with research translation and implementation to ensure research results in improved DYFS policy that benefits child development and well-being. This could include new guidelines for foster care placement and developing new, proactive policies to support foster families create healthy and supportive environments for children in their care.

II. Background

Give a background on the importance of the research proposal or use case, and if relevant, prior literature on the topic. Make sure to include what the contribution of this project will be to the Sponsoring Agency, additional Agencies, and/or members of the County/State.

The Governor's Office has asked RIPL to find ways to close the child achievement gap and innovate with social programs to make them more effective and efficient. This project spans both initiatives, and arose from a request by the DYFS Director.

Each year, millions of children and families are subject to investigation by child protective service agencies for alleged abuse or neglect (U.S. Dept. of Health and Human Services [HHS] 2016). Nearly 20 percent of children are removed from their homes as a result (U.S. HHS 2016). In 2017, there were an estimated 442,995 children in foster care (Children's Bureau 2019). The goal of removal and foster care is to improve outcomes for at-risk children by reducing their exposure to at-risk environments. Descriptive research shows that maltreated children have substantially worse academic performance and are more likely to have a social or emotional condition such as aggressive behavior or depression, and that these effects emerge immediately in early childhood (Fantuzzo and Mohr 1999; Wolfe et al. 2003; Holt et al. 2008; Font and Berger 2015).

Two prescient factors limiting the amount of rigorous evidence to guide removal and foster care policy are a lack of administrative data to track a wide range of outcomes over time, and a lack of convincing approaches to causal identification. Government agencies across the nation largely have yet to join their administrative data across individuals, which makes measuring outcomes like health and incarceration for those interacting with the foster care system difficult. The challenge for inference is that removed and non-removed children likely differ in terms of unobserved characteristics (Berger et al. 2009; 2014). Doyle (2007; 2008) addressed this concern by using removal tendencies of child protective service investigators as an instrumental variable (IV). His estimates represent causal effects because his setting featured quasi-random assignment of investigators to cases. Focusing on later-life outcomes, he found negative impacts of removal on delinquency, arrests and labor market outcomes. An important consideration is that Doyle studied investigations and removals occurring between the ages of five and fifteen. A large literature on child development suggests that his estimates may not generalize to children removed at younger ages (Cunha et al. 2006; Cunha and Heckman 2007; Almond and Currie 2011; Heckman and Mosso 2014; Almond et al. 2017). Importantly, he did not estimate the impacts specifically of foster care placement specifically on children, as not all removed children are placed in foster care. Using administrative data and new machine learning techniques, we hope to examine short- and long-term impacts on a variety of additional measures for foster care placement to build on this body of research.

This research also relates to a broader literature on the impact of interventions for children. A large literature shows that programs delivered in-utero or during early-life has lasting impacts on children (Garces et al. 2002; Ludwig and Miller 2007; Chetty et al. 2011; Campbell et al. 2014; Aizer et al. 2015; Chetty et al. 2016; Hoynes et al. 2016; Isen et al. 2017; Chyn 2018). We

will expand on this literature by providing new evidence that speaks to the relative efficacy of targeting an intervention based on the age of children. In terms of child welfare policy, the impact of foster care at young ages is important topic given that nearly half of removed children are age 5 or below (U.S. HHS 2012).

III. Data

Provide a list of data tables and specific variables to be used in the project, and why they are necessary. Variables can be listed in an appendix.

Our main data from the state data lake are derived from administrative files from DYFS, available from 2000 to 2015. The administrative data contain fields for victim and perpetrator demographics, specific abuse allegations, the initial reporter of abuse, and case assignment to investigators. The data also record whether the investigations resulted in removal from home and placement into foster care. We will also flexibly control for background characteristics such as age, gender, and race. We will then join these data to the following administrative records to include measures of schooling, academic, and later-life outcomes:

Outcome	Data Source	Years Available
School attendance	State Department of Education (public school enrollment and testing records)	2003 to 2016
Enrollment in Special Education (IEP)		
High School Completion		
Standardized reading test scores (grades 3 to 8)		2005 to 2015
Standardized math test scores (grades 3 to 8)		
Teenage pregnancy	State Department of Health (birth records)	2000 to 2016
Post-secondary enrollment	National Student Clearinghouse Data	2004 to 2016
Juvenile convictions	Rhode Island Family Court	2000 to 2016

IV. Research Methods

What are your principal hypotheses or questions that will be explored? How will the data be analyzed to answer the research question(s)? What research methods, analytic models, or strategies will be used to answer the question? What outcomes will you measure? If you are proposing an intervention, lay it out in detail here. Write this section for a well-informed lay audience.

We will examine the following research hypotheses/questions:

1. What is the causal impact of foster care placement on child achievement as measured by educational outcomes (standardized test scores, school attendance and completion, and post-secondary attainment), juvenile detention and teen parenting?
2. Do impacts of foster care placement on achievement vary by gender? Prior studies of school and neighborhood-related interventions suggest that interventions for disadvantaged children may have different impacts for girls and boys (Hastings et al. 2006; Kling et al. 2007; Anderson 2008; Angrist et al. 2009; Angrist and Lavy 2009; Deming et al. 2014).
3. Do impact of foster care placement vary by the age at which children are placed in foster care? Prior studies suggest that early life interventions may have large impacts on child outcomes (Garces et al. 2002; Ludwig and Miller 2007; Almond et al. 2010; Chetty et al. 2011; Bharadwaj et al. 2013; Campbell et al. 2014; Aizer et al. 2016; Chetty et al. 2016; Hoynes et al. 2016; Isen et al. 2017; Chyn 2018, Currie et al. 2018.)

To answer the research questions, RIPL will use an instrumental variables (IV) approach, flexibly controlling for a wide variety of background characteristics and using the strictness of the DCYF investigator assigned to the family as an instrument. Please see Appendix A for details on our empirical approach and model.

We anticipate finding statistically significant impacts of foster care placement on child outcomes. As the literature suggests that early childhood interventions can have lasting impacts, we expect these impacts to increase the younger the child is when foster care placement occurs (Fantuzzo and Mohr 1999; Wolfe et al. 2003; Holt et al. 2008; Font and Berger 2015).

We anticipate significant policy implications from our findings. Our partnership with DYFS on previous policy issues means that we are experienced in working with their team and that they anticipate using these results. DYFS's commitment to a collaborative partnership to tackle policy issues opens the opportunity for this analysis to help DYFS develop programs to best help foster care families provide healthy environments for foster care children, and to help adjust agency best practices and screening guidelines to incorporate decisions about a child's long-term outcomes when making decisions about placing children into foster care.

V. Authorized Users/Investigators

What are the qualifications and responsibilities of the team conducting the research? Provide a list of everyone who will need access to the data and/or involved in the project. Give a brief description of how this project will be funded.

Dr. Jane Smith (PhD) is a Professor of Economics at Brown University and a Faculty Research Associate at the National Bureau of Economic Research (NBER). Her research focuses on

individual and market behavior and the implications of this understanding for the design of public policy. She has built successful partnerships with government and private organizations both for the acquisition of data and the implementation of policy across the country. Her approaches range from observational data to randomized trials, incorporating both reduced form and structural methods, and considering both traditional and behavioral analytic models.

Dr. John Smith (PhD) is an Assistant Professor of Economics at the University of New Hampshire and Faculty Research Associate at RIPL. Dr. Smith is an applied microeconomist studying a broad set of topics in labor and public economics. His research centers on policy-relevant questions that he addresses using a wide-range of empirical methods and data.

Other investigators on the project are:

- Research Assistant #1 (*Research Improving People's Lives,* Research Assistant)
- Research Assistant #2 (*Research Improving People's Lives,* Research Assistant)

VI. Timeline and Deliverables

What are the final products that will be produced from this researcher or use case? List the final deliverables and how the results will be reported, including scientific papers, dashboards, conference presentations, and internal reports. Provide a detailed timeline for your proposed work. Include a timeline for internal reporting to the Sponsoring Agencies. It is recommended that internal reporting be done on at least a quarterly basis.

We use an iterative workflow that allows partners to give consistent feedback and iterate together on the features and design of the research to ensure it is providing the most accurate results and have the best possible understanding of back-end data. We provide weekly written updates to state government on our research and have monthly meetings with DYFS to review research progress, get feedback, and discuss how to solve challenges or answer data or implementation questions as they arise. We look forward to engaging with DYFS technical and program expertise as we work together on this important research.

We will iterate with DYFS on a final report from RIPL that clearly explains the analysis, methodology, and results, as well as recommendations from RIPL on next steps to partner to translate results into policy. These projects will last beyond 2019 as RIPL works with DYFS to turn research results into actionable policy improvements or adjustments.

Over our two-year timeline, RIPL will produce one scientific research paper which will evaluate foster care placement policy in the state and allow DYFS to make fact-based decisions moving forward to protect vulnerable children.

Table 2: Project Timeline								
Item	Quarter (starting 1/1/2019)							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Aim 1 and 2: Work with DYFS to ensure accurate understanding of programs and data for reliable results.								
Aim 1: Complete analysis for all educational, juvenile detention, and teen parenting outcomes.								
Aim 1: Preliminary internal report due to DYFS; DYFS will provide feedback.								
Aims 1 and 2: Final internal report due to DYFS and present recommendations for policy adjustment.								
Aim 1: Draft paper manuscript and present initial draft to DYFS.								
Aim 1: Validate refined analysis with DYFS and state partners.								
Aim 1: Study robustness and include sensitivity analysis.								
Aim 1: Submission to academic journal; DYFS and partners have opportunity to review and provide feedback for at least 60 days before submission.								

VII. Publication Plan

If the deliverables will be publicly available or published as a scientific paper, you must include a publishing plan that includes a plan for which (if any) journals it will be submitted to, their replication policies, a timeline for preparing the paper for publication (including windows of opportunity for partners to review final drafts and suggest feedback), where replication files will be kept, and any other items agreed upon between partners.

- Once RIPL has produced initial results in Q2, DYFS and RIPL agree that both parties have committed to publication of those results. Replication files for all analysis will be archived in the data lake at that point moving forward.
- RIPL will target submission to the American Economic Review (AEA) and the Journal of Political Economy (JPE) at the end of 2020.
    - The replication policy for the AEA is here. The policy states that the data and code used must be "clearly and precisely documented, and access to the data and code [must be] clearly and precisely documented and is non-exclusive to the authors.

    - The replication policy for the JPE is here. The policy for states that data must be "clearly and precisely documented and are readily available to any researcher for purposes of replication."
    - DYFS will work with RIPL to allow applications for access to replication files for the purpose of replicating this research.
- RIPL will provide DYFS and other partners with at least 60 days' notice in advance of submitting to a journal.
- Before submission, RIPL will provide DYFS with at least 30 days to review the final draft of the paper and to provide feedback and suggestions.
- If the paper is accepted for publication, RIPL will provide DYFS with at least 30 days' notice of the acceptance and publication timeline. RIPL will work with the DYFS communications team as necessary to coordinate on press and communications for the paper.

Exhibit A: Technical Methodology

Program Information

We propose to measure the effectiveness of the state's foster care placement policies by examining the impact of removal on a child's outcomes. DYFS serves all families living in the state. An investigation into child abuse or neglect in the state begins with a call to the Child Protective Services (CPS) hotline. Professional call staff at the DYFS central office record details of the allegations, identify previous or pending investigations, and determine whether the report meets criteria for an investigation. If the report does not meet the criteria for an investigation, the case is closed.[4] If the report does meet the criteria for an investigation, the call staff forward the case to the Investigative Unit where a supervisor assigns the case to an available Child Protective Investigator (CPI).

The Investigative Unit supervisor must assign the investigation case to a CPI. The supervisor primarily assigns cases randomly using a "rotation list." The rotation list is an ordered spreadsheet of CPIs maintained by the Investigative Unit supervisor. On a given day, a supervisor assigns new cases based on the order on the rotation list. CPIs receive at most one case per day. If all CPIs receive a case, the next day the rotation list remains the same. When there are sufficiently few calls, the supervisor places non-assigned CPIs as the first available for the next day's rotation. An exception to this assignment process can occur if a specific type of abuse occurs that requires the attention of a particular CPI. For example, if an investigation contains an allegation of sex abuse, the Investigative Unit supervisor may attempt to assign a CPI of the same gender as the abuse victim. During our sample period, the supervisor assigned cases without using the rotation list in less than nine percent of cases.

After receiving a case assignment, the CPI decides whether there is sufficient evidence of child abuse or neglect. If the CPI does find evidence, the CPI petitions the State Family Court for removal of the child and placement into DCYF custody, as well as the type of placement and duration of time in the foster care system.

Research Methodology

Empirical Model

Our research methodology is based on the following model of outcomes for child *i*:

---

[4] When deciding whether a report merits an investigation, call staff assess whether the report meets the following five criteria: 1) The report involves a victim of child abuse or neglect whose welfare is harmed or threatened with harm; 2) The report involves transfer of a child into the permanent care of an unrelated, unauthorized individual; 3) The report involves sexual abuse of a child by another child; 4) The report indicates a perpetrator of child abuse and/or neglect places other children at risk; 5) The report indicates a threat to the safety of an unborn child.

$$Y_{ijt} = \beta_0 + \beta_1 R_{ijt} + \beta_2 X_i + \epsilon_{ijt} \quad (1)$$

where Y_{ijt} is the outcome for child *i* who is assigned to investigator *j* in year *t*, R_{ijt} is an indicator for foster care placement, X_i is a vector of case characteristics, and ϵ_i is an error term. Standard OLS estimates of Equation 1 will be biased if home removal (R_{ijt}) is correlated with unobserved determinants of child outcomes (ϵ_i).

To address the endogeneity concern in Equation 1, we rely on an instrumental variable (IV) strategy that is based on a measure of the tendency to place a child in foster care, denoted as Z_{ijt}, of the CPI assigned to child *i*. Formally, our first stage equation is:

$$R_{ijt} = \alpha_0 + \alpha_1 Z_{ijt} + \alpha_2 X_i + \nu_{ijt} \quad (2)$$

Following prior research, we compute Z_{ijt} as the simple leave-out foster care placement tendency of investigators:

$$Z_{ijt} = \frac{\sum_{k \neq i} R_{kjt}}{N_{jt} - 1} \quad (3)$$

where Z_{ijt} is the foster care placement tendency for investigator *j* assigned to child *i* in year *t*, R_{kjt} is foster care placement status, n_{jt} is the total number of children assigned to investigator *j* in year *t*, and n_t is the total number of children investigated in year *t* (Doyle 2007; 2008).[5]

If there are heterogeneous impacts of placement, we must make two assumptions to interpret IV estimates of the parameter β_1 from Equation 1 as a local average treatment effect (LATE) of foster care placement for marginal investigations (Angrist and Imbens 1995). First, the measure of CPI placement tendency defined in Equation 3 must affect child outcomes only by changing the probability of foster care placement. This assumption is plausible in our setting because supervisors assign CPIs to investigations based on the rotational assignment list described previously. Second, we must assume that there is a monotonic impact of CPI assignment on foster care placement across children. A violation of this assumption may occur if CPIs vary in their relative treatment of children based on observed case characteristics. For example, a given CPI may be relatively strict when it comes to placing African-American children in foster care, but lenient when it comes to placing all other children.

[5] A number of studies examining the impact of prison and incarceration use related research designs based on the stringency of judges randomly assigned to court cases (Kling 2006; Aizer and Doyle 2015; Bhuller et al. 2016; Mueller-Smith 2016; Dobbie et al. 2018; Bhuller 2018).

To further address concern about monotonicity and rigorously identify instruments that maximize the accuracy of predicted foster care placement, we allow CPI tendency and its impact on foster care placement to vary with baseline case characteristics. We do this by creating a set of potential instruments based on a leave-one out measures of placement tendency calculated for different categorizations of case characteristics (specifically, sex, minority status, allegation type and investigation type). This creates a large number of potential instruments. Following Belloni et al. (2012; 2014), we use the machine-learning (ML) algorithm, Least Absolute Shrinkage and Selection Operator (LASSO), to select the instruments with greatest predictive power for foster care placement in the first stage equation (Belloni et al., 2012; 2014).[6]

Limitations

Our primary research limitation is that while our administrative data allow us to generate new insights into the impact of foster care placement on low-income youth, it does not encompass the psychological or emotional impact of the event on children. For example, we are not able to evaluate the impact of foster care placement on children's mental health, behavior in the home, or emotional adjustment to the removal and foster care process. These types of outcomes, which are outside the scope of administrative records, would be better system to measurement through surveys. We could potentially undertake qualitative research involving surveys to better understand these outcomes based on our results from the analysis of administrative data. We could also work with DYFS to incorporate health records into a later analysis to expand these results.

[6] The use of LASSO for regularization is necessary since there are a large number of potential case characteristics by which CPI tendencies may vary. An unrestricted model would will likely result in too many instruments and potentially weak instruments, creating challenges for causal inference (Bound, Jaeger and Baker (1996).

Exhibit B: Variable List

- Data Lake Summary Table
    - Universal ID, month
    - Age
    - Race
    - Gender
    - Incarceration
    - Marital status
    - Anonymized census block
    - Medicaid enrollment
    - Wages
- DYFS Demographics Tables
    - All fields including IDs, gender, race, language spoken
- DYFS Family Tables
    - DYFS IDs (family and person)
- DYFS Investigations Table
    - All fields including IDs, date of investigation, outcome of investigation
- DYFS Allegations Table
    - All fields including IDs, type of allegation, substantiated flag, relationship to perpetrator
- DYFS CPS Report Table
    - All fields including CPS Report, reporting method, family structure, investigation level, substance abuse
- DYFS Assignments Table
    - All fields including IDs, worker role, case ID
- DYFS Placements Tables
    - All fields including IDs, care provider, placement dates, discharge reason
- DYFS Adjudicated and Unadjudicated Tables (2 tables)
    - DYFS ID
    - Sentencing date
    - Charge information
- DHS Person and Program Tables
    - Universal ID, DHS ID
    - Application and enrollment dates
    - Recipient status and benefits received for GPA, RSDI, SNAP, SSI, TANF
- DoE Enrollment Tables
    - All fields including IDs, school, grade, IEP, LEP, FRL, enrollment type, exit type
- DoE Disciplinary Table
    - All fields including IDs, disciplinary infractions by day, description of infraction
- DoE Test Scores Tables

    - Universal ID, date of test
    - NECAP, PARCC, AP, SAT, PSAT scores
- Doe National Student Clearinghouse Table
    - All fields including IDs, graduation date, college enrollment

A6. Preprint of Communications of the ACM Paper

This paper is forthcoming in *Communications of the ACM* and is available as a preprint from: https://osf.io/28krq/

Unlocking Data to Improve Public Policy

Justine S. Hastings[1,2,3,4], Mark Howison[1,2], Ted Lawless[1], John Ucles[1], Preston White[1]

[1] Research Improving People's Lives, Providence, RI, USA
[2] Watson Institute for International and Public Affairs, Brown University, Providence, RI, USA
[3] Department of Economics, Brown University, Providence, RI, USA
[4] National Bureau of Economic Research, Cambridge, MA, USA

When properly secured, anonymized, and optimized for research, administrative data can be put to work to help government programs better serve those in need.

Key Insights

- Fact-based policymaking – the practice of using data and research to guide policy decisions – is a promising solution to improving the effectiveness and efficiency of government programs.
- Administrative data can provide new facts to guide policymakers. However, understanding the quality of administrative records, and integrating, transforming, and optimizing them for policy insights presents many challenges.
- To overcome these challenges, we developed an integrated database of administrative records from multiple Rhode Island state agencies with over 800 tables and 2.7 billion records related to over 4 million anonymous individuals. These data support econometric and machine-learning research into policies with promise to deliver higher impact per dollar and better serve Rhode Island families.
- As a specific example, we describe how anonymized data from the integrated database are used to discover a new insight into a policy challenge related to low birth weight newborns.

Introduction

There is a growing consensus among policymakers that bringing high-quality evidence to bear on public policy decisions is essential to supporting the effective and efficient government that

their constituencies want and need. At the U.S. federal level, this view is reflected in a recent congressional report by the Commission on Evidence-Based Policymaking, which recommends creating a data infrastructure that enables "a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy" [4].

This article describes a new approach to data infrastructure for fact-based policy, developed through a partnership between our interdisciplinary organization Research Improving People's Lives[7] and the State of Rhode Island [13]. Together, we constructed *RI 360*, an anonymized database that integrates administrative records from siloed databases across nearly every Rhode Island state agency. The comprehensive scope of *RI 360* has enabled new insights across a wide range of policy areas, and supports ongoing research into improving policies to alleviate poverty and increase economic opportunity for all Rhode Island residents (see Sidebar #1). Our approach can guide other policymakers and researchers seeking to similarly transform and integrate administrative data to guide and improve policy.

Sidebar #1: Policy areas in which *RI 360* has contributed insights

- Lowering non-urgent emergency health care costs
- Curbing the opioid epidemic
- Improving worker training programs
- Creating tools to connect dislocated workers to benefits
- Helping families become more food secure
- Optimizing energy policy for low-income families
- Helping children reach proficiency on reading and math tests
- Closing the college achievement gap

The role of administrative data in policymaking

Administrative data can be collected from the computer systems used by government agencies to run their programs. When transformed into databases that are more suitable for insights, these anonymized records provide new sources of facts for policymakers to benchmark goals and measure the successes and shortcomings of existing and future programs. Often classified as "big data" due to their volume, variety, and availability [10], administrative records are also an increasingly valuable source for empirical social science research [5]. Research with

[7] https://ripl.org

administrative records can contribute new data-driven insights to inform important policy decisions (see Sidebar #2), and add objectivity and scientific rigor to measuring program impact and designing effective program changes. Moreover, scientists can inform how data from administrative systems, which are primarily designed around operational needs and often not suitable for analysis, can be transformed effectively to support research and insights.

Sidebar #2: Recent Data-Driven Insights from Administrative Records

- Records from the New York City criminal justice system show how judges often mispredict risk when making bail decisions [15]. Judges identify and release many defendants who have a low flight risk, but also release nearly half of the defendants with the highest flight risk. In simulations, replacing judges' decisions with a machine-learning prediction can reduce either crime rates (at a fixed jailing rate) or jailing rates (at a fixed crime rate), and in both cases can reduce racial disparities in outcomes.
- Transaction data from a private grocery retailer and data from the Supplemental Nutrition Assistance Program in Rhode Island show that households treat their nutrition benefits as if they were earmarked for food expenses, even when they could be substituted for cash [14]. This finding contradicts traditional economics theory which predicts that nutrition benefits should be fungible (e.g. substitutable for cash), and instead supports an alternative economics hypothesis called mental accounting. Results suggest that Supplemental Nutrition Assistance Program impact on spending and nutrition can be influenced by policies governing when and how benefits are distributed.
- Federal income tax records show there are growing inequalities in life expectancy in the U.S. across socioeconomic factors [2]. The breadth and scale of these administrative data (with over 1.4 billion person-year observations) reveal that geographic factors like government expenditure and fraction of immigrants and college graduates are positively correlated with life expectancy at the bottom of the income distribution.
- Randomized field experiments in Chicago combined school and unemployment insurance records with arrest records to evaluate the impact of a summer job support program for youth [6]. By using this integrated administrative data, the study found that the program caused declines in violent-crime arrests even though there were no significant effects on school or employment outcomes, which are the more typically studied effects of youth job programs.

Although the idea of guiding policy with data dates back to the 70s and 80s, early studies only considered isolated data sources and come from a time when data were scarce. It was not until

recently that advances in data collection, storage, and scale provided the opportunity to integrate data across nearly every facet of government. Early case studies and survey studies highlight how the process of data modeling can facilitate negotiation and consensus-building among policymakers [8], but also how the unmet promises of new information technologies prompted frustration among government leaders at that time [9].

An important lesson is to engage policymakers and leaders to fully understand their needs, which is why we formed extensive partnerships with state government leaders while building *RI 360*. Integrated administrative data can support not only academic research, but also the analytics needs of government itself. Like researchers, government analysts need access to data that have been transformed to provide insights and integrated across programs that serve what are often overlapping populations. For these reasons, *RI 360* was selected as the primary data source for the Rhode Island Executive Office of Health and Human Service's Data Ecosystem project, to empower their data analysts and partners with data optimized for insights.

An example policy for low birth weight newborns

Throughout this article, we will describe our process for building *RI 360* in the context of a specific policy: determining the optimal weight threshold for providing additional medical care and resources to low birth weight newborns and their mothers [3]. Children born with low birth weight tend to have more health difficulties and worse outcomes later in life compared to their peers. They also tend to be at higher risk, coming from disadvantaged backgrounds where mothers are more likely to be teen mothers or have reported alcohol or drug abuse. Programs to support these infants and mothers may increase equity of opportunity and reduce state and federal expenditures for support programs and anti-poverty programs later on in life. Currently the threshold for additional resources is set at 1,500 grams [1]. We use this threshold to measure the causal impact of these additional resources to determine if increasing this threshold could be a low-cost, high-return policy change that could improve lives, increase equity of opportunity, and save state and federal funds in the long run.

Using integrated data from *RI 360*, we can examine a wide range of outcomes, including educational test scores, college enrollment, use of social programs and Medicaid, and maternal care and stress. The data allow for a holistic view of policy impact; measuring gains to education and well-being from the immediate to the longer-term, and also measuring expenditure savings to government-funded social safety-net programs from early-life investments so that government can incorporate concepts of return on investment when considering how to get the most impact per dollar spent.

Our study finds that newborns just below the threshold who receive additional medical care fare significantly better later in life compared to those just above the threshold. Crossing the threshold is associated with increases in standardized test scores in elementary and middle school of 0.34 standard deviations, increases in college enrollment rates by 17.1 percentage points of a base rate of 53.6%, and decreases in social program expenditures of $27,291 by age 10 and $66,997 by age 14. Because the average cost of the additional medical services provided in the hospital at birth is around $4,000 [1], this study provides new facts to help policymakers evaluate the educational impact and potential financial returns of adjusting the threshold. We conclude that moving the threshold is a potential low-cost, high-impact policy lever for helping children at the margin to achieve better outcomes later in life.

To conduct this comprehensive study of outcomes for low birth weight newborns, we access data in *RI 360* that originate from several Rhode Island agencies. Three decades of birth records from the RI Department of Health define the study population of newborns with low birth weight. The RI Department of Education provides test scores from third, firth, and eighth grade standardized tests, the PSAT, the SAT, and Advanced Placement exams; records of grade repetition, Individualized Education Programs, and disciplinary actions; and college enrollment records from the National Student Clearinghouse. The RI Department of Human Services provides enrollment and benefit payment records for Supplemental Security Income, the Supplemental Nutrition Assistance Program, Medicaid, and Temporary Assistance for Needy Families. The RI Department of Labor and Training provides quarterly wage records that measure maternal employment rates and earnings following birth. The Centers for Disease Control provide survey responses from the Pregnancy Risk Assessment Monitoring System that measure maternal attitudes and experiences following birth.

Securing the data

Figure 1 summarizes our approach and highlights the first challenge when working with administrative records: deploying security controls that protect the data. Security is our first and foremost concern because the risks of improperly securing administrative data are great. Unauthorized access or data leakage have the potential for invasions of individual's privacy, identity theft, financial fraud, or even interference with our democratic institutions, including elections. Moreover, irresponsible handling of data can have spill-over effects which hinder scientific progress and policy improvement, as data owners perceive great risks of using data and partnering with scientists, even if the uses and partnerships are legitimate and secure.

We mitigate these risks by isolating all data ingest and processing within an encrypted tank (Figure 1a) inside a secure computing environment called a *data enclave* [16]. The enclave's key features are that it is physically secure and isolated from the Internet, data transfers in and out

are restricted and subject to a documented approval process, all access is comprehensively audited, and access is granted to only a limited group of approved researchers. These security controls protect against unauthorized access and ensure that researchers access the data in compliance with the data sharing agreements governing their use.

Our implementation of the data enclave uses a locally-hosted system. However, modern cloud computing can help governments implement similar data enclaves using best practices for security and compliance. An additional benefit of a cloud solution is that government can own and operate the enclave, retain possession of the administrative data, and directly manage researchers' access, which removes the need for data transfers and data sharing agreements.

As an additional security measure, we restrict access to the encrypted tank using a two-party password, known only by senior leadership. A two-party password means two people each know a different half of the password, and both of the senior parties have to be present and consent to access the encrypted tank. This ensures that no individual researcher can access data that may reveal personally identifiable information.

Once the original data have been successfully transferred into the encrypted tank, we run an automated pipeline to separate out personally identifiable information (Figure 1b). Sensitive identification numbers – such as social security numbers or other identifiers deemed sensitive by the agency – are flagged ahead of time and automatically replaced with irreversible hashes, a technique that is widely used for protecting passwords [10]. Following this separation, the remaining data contain no personally identifiable information and are de-identified (Figure 1c).

Anonymizing the data

Once that data are secured, the next challenge is developing a method for identifying the same individual across data sets, while also preserving their anonymity so that researchers cannot discover their identity, even inadvertently. Although many of the data sources for the birth weight study identify records by social security number, an exception is the RI Department of Education, which identifies students by name and an internal identification number. Therefore, we require an automated method to find matches among individual records based on hashed social security number when available, or else based on other fields like name and date of birth – all without revealing these fields to the researcher.

Our solution is to assign a global anonymous identifier (Figure 1d) to records right after separating out personally identifiable information. An automated script identifies matches among all hashed social security numbers, phonetic representations of names (using the Soundex algorithm [18]), and dates of birth. Using the global identifier, we can join information

on outcomes to low birth weight newborns and their parents in the birth records without knowing any personally identifiable information for any of the individuals.

Our deterministic algorithm is designed to minimize false matches (incorrectly matching two different individuals) at the expense of having more missed-matches (in which two records of the same individual are not matched). Some records are missing too many fields and are considered too ambiguous to assign a global identifier, but this occurs for only 3.9% of records. As an alternative to the deterministic approach, the identifier could be constructed with probabilistic record-linkage methods that would likely have fewer missed-matches, but would also carry higher costs for computation and manual curation, as well as a higher likelihood of false-matches [12].

Integrating the data

We receive data extracts from administrative systems in various formats. The raw records used in the birth weight study arrive in the encrypted tank as comma-separated text (with varying delimiters and quoting conventions), fixed-width text, XML, and Excel files. Our approach has been to meet government data partners where they are, and to accommodate data extracts in the format they can most easily produce. Most agencies have perpetual operational demands on their administrative systems, and they are not resourced to support additional development for data warehousing or analytics.

Since there is no universal format or data dictionary across agencies, we normalize the data into a consistent format and typing structure with a lightweight and open-source integration tool called Secure Infrastructure for Research with Administrative Data. We developed this tool using an agile approach to meet the evolving needs of researchers and analysts as we built *RI 360*. Our GitHub repository[8] provides additional technical detail about our integration methods, as well as a worked example based on simulated data.

We chose an Extract Load Transform approach over the more typical Extract Transform Load approach [7]. In practice, this means that the de-identified data are loaded into *RI 360* in as close to their original format as possible. The majority of transformations are added later after researchers have a chance to perform preliminary analyses to assess data quality and understand the data-generating processes underlying the administrative systems.

As an example, birth weight is an essential variable for defining our study population. However, is has been measured in different units (grams and ounces) over the three decades of birth

[8] https://github.com/ripl-org/sirad-example

records. Therefore, we construct a *birth derived* table that normalizes weight, as well as several other categorical variables measured at birth that switch from using numeric to character codes in the records over time. A derived table is a materialized view that aggregates, normalizes, and/or combines data from multiple original tables in *RI 360* into a single table that facilitates a specific analysis need – in this case, determining birth weight in a consistent way for all births. A more complex example in *RI 360* is the Supplemental Nutrition Assistance Program derived table. It combines records on applications, eligibility, benefit payments, and household structure to determine all individuals enrolled in the program at a given month and their household-level benefits.

At the highest level, we roll up all the derived tables into a single *RI 360* summary table, which spans 20 years of history for the state's most important programs and outcomes, as well as demographic information about anonymized individuals (e.g. age, race, ethnicity, and sex). Most of the outcomes in the birth weight study, including educational outcomes and benefit payments, are found in the *RI 360* summary table, which reduced the effort needed to launch the study. Creating derived tables also ensures that all studies using *RI 360* draw from common variable constructions and definitions that are robust and reproducible.

Supporting research integrity

A fundamental requirement of scientific findings is that they can be independently replicated by other investigators [17]. Similarly, fact-based policy should be based on robust findings that are peer-reviewed and replicable. To facilitate future replication, we update and snapshot *RI 360* approximately three times a year, creating what we call a *research version* (Figure 1f). The research versions are de-identified data and become the permanent archive of *RI 360*. We have currently generated 11 such versions. Once a research version has been validated, the encrypted original data used to create that version are wiped from the encrypted tank and destroyed. Every analysis is tied to a fixed research version of the database, and can be rerun against the research version at a later time to replicate the results. Additionally, to encourage reproducibility, analysis projects use a common project template to organize code and research results in a standardized way.[9]

Even through *RI 360* has been de-identified, our data sharing agreements restrict all research with anonymized individual-level records to the data enclave. Only aggregated or statistical results such as summary tables, plots, and regression coefficients can be exported from the enclave. All statistics must be aggregated such that they represent 11 or more distinct individuals. To ensure compliance with usage agreements and security, no individual researcher

[9] https://github.com/ripl-org/predictive-template

has the ability to export files from the enclave. Copy and paste functionality has also been disabled within the enclave's user interface. Exports are subject to review and documentation to ensure that exported results conform to usage agreements (Figure 1g), and they trigger real-time alerting to senior leadership. A read-only snapshot of each export is archived in the enclave to facilitate future audits.

Conclusion

The insights gained from research with administrative data have the potential to transform the way that policymakers approach some of society's most important policy decisions. Robust evidence on previous policy outcomes and predictive modeling of future outcomes can guide policymakers to smarter policies with greater benefits at lower cost. We have described a comprehensive approach to overcoming the many challenges faced when integrating siloed state-wide databases into a data infrastructure for fact-based policy, which is the first system of its kinds in the U.S. In the future, we hope that more systems of this kind will provide policymakers at all levels of government, and in many countries across the world, with a rich ecosystem and evidence base for the important decisions they make on behalf of their constituents.

Acknowledgments

This work was supported by the Smith Richardson Foundation.

References

1. Douglas Almond, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. 2010. Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborn. *Quarterly Journal of Economics* 125, 2 (May 2010), 591–634. DOI:10.1162/qjec.2010.125.2.591

2. Raj Chetty, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA* 315, 16 (April 2016), 1750–1766. DOI:10.1001/jama.2016.4226

3. Eric Chyn, Samantha Gold, Justine S. Hastings. (Forthcoming). Short- and long-run impacts of health interventions for very low birth weight children.

4. Commission on Evidence-Based Policymaking. 2017. The Promise of Evidence-Based Policymaking. Retrieved from https://www.cep.gov/cep-final-report.html

5. Roxanne Connelly, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. The role of administrative data in the big data revolution in social science research. *Social Science Research* 59, (September 2016), 1–12. DOI:10.1016/j.ssresearch.2016.04.015

6. Jonathan M.V. Davis and Sara Heller. 2017. *Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs*. Working Paper No. 23443. National Bureau of Economic Research, Cambridge, MA. DOI:10.3386/w23443

7. Umeshwar Dayal, Malu Castellanos, Alkis Simitsis, and Kevin Wilkinson. 2009. Data integration flows for business intelligence. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology,* Saint Petersburg, Russia, March 24-26, 2009. DOI:10.1145/1516360.1516362

8. William H. Dutton and Kenneth L. Kraemer. 1985. *Modeling as Negotiating: The Political Dynamics of Computer Models in the Policy Process*. Ablex Publishing Corporation, Norwood, NJ.

9. James Danziger. 1977. Computers and the Frustrated Chief Executive. *Management Information Systems Quarterly* 1, 2 (June 1977), 43–53.

10. Liran Einav and Jonathan Levin. 2014. Economics in the age of big data. *Science* 346, 6210 (2014), 1243089. DOI:10.1126/science.1243089

11. P. Gauravaram. 2012. Security Analysis of salt||password Hashes. In *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, 25–30. DOI:10.1109/ACSAT.2012.49

12. Katie Harron, Chris Dibben, James Boyd, Anders Hjern, Mahmoud Azimaee, Mauricio L Barreto, and Harvey Goldstein. 2017. Challenges in administrative data linkage for research. *Big Data & Society* 4, 2 (December 2017), 2053951717745678. DOI:10.1177/2053951717745678

13. Justine S Hastings. 2019. Fact-Based Policy: How Do State and Local Governments Accomplish It? The Hamilton Project (Brookings Institution), Policy Proposal 2019-01. Retrieved from http://www.hamiltonproject.org/assets/files/Hastings_PP_web_20190128.pdf

14. Justine S. Hastings and Jesse M Shapiro. 2017. How Are SNAP Benefits Spent? Working Paper No. 23112. National Bureau of Economic Research, Cambridge, MA. Retrieved from http://www.nber.org/papers/w23112

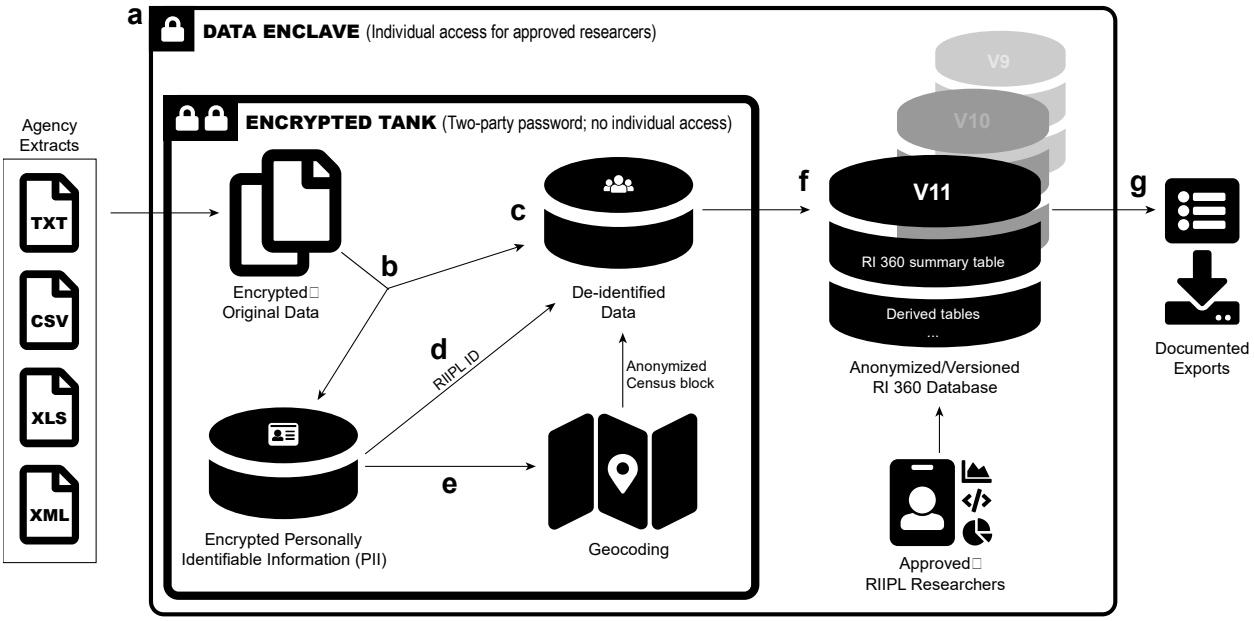
15. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *Q J Econ* 133, 1 (February 2018), 237–293. DOI:10.1093/qje/qjx032

16. Julia Lane and Stephanie Shipp. 2007. Using a Remote Access Data Enclave for Data Dissemination. *International Journal of Digital Curation* 2, 1 (2007), 128–134. DOI:10.2218/ijdc.v2i1.20

17. Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (December 2011), 1226–1227. DOI:10.1126/science.1213847

18. C. Russell Robert. 1918. The Soundex coding system. Patent No. US1261167.

Figure 1. Overview of the processing steps to secure, integrate, and conduct anonymized research with administrative data. Agencies securely transfer data extracts to an encrypted tank inside the data enclave **(a)**. These data are split **(b)** into personally identifiable information and de-identified data **(c)**. Personally identifiable information is used to construct an anonymized global identified **(d)** and to geocode home addresses to construct an anonymized neighborhood identifier **(e)**. De-identified data are used to construct research versions of the *RI 360* database **(f)**, which can be accessed by approved researchers from inside the data enclave. Research findings can be exported from the data enclave through a documented review process **(g)**.



A7. Summary of Available Projects in the RIPL Knowledge Base

Table 1. Summary of Current Available Knowledge Base Projects			
Research Tool	Research Question	Data Sources	Outcomes
Predictive model: Opioid Use	Who is at high risk of developing an opioid use disorder if written an initial prescription by a physician?	Medicaid data (primary data source), criminal justice data, wage roll data, and social services records.	A data-driven risk score that helps physicians and policymakers understand the costs, benefits, and fairness tradeoffs to prevention policy approaches
Predictive model: Medicaid ED Use	Who is at high risk of becoming a high-preventable-cost Medicaid Emergency Department user?	Medicaid data (primary data source), criminal justice data, wage roll data, and social services records	A data-driven risk score that helps identify individuals and subpopulations to connect with proactive support services
Predictive model: Child Maltreatment	How can we best provide proactive support services to families and communities at risk of substantiated child maltreatment?	Child protective services records (primary data source), Medicaid data, criminal justice data, wage roll data, and social services records	A spatial data-driven risk score that helps policymakers identify communities and geographic areas to connect to proactive support services
Predictive model: Test Scores	What policy approaches can help us successfully improve student test scores?	Education records, Medicaid data, criminal justice data, wage roll data, and social services records	A set of key factors that influence student standardized test scores
Causal analysis: Improving Life Outcomes for Low Birthweight Infants	Do additional health investments at birth for very low birthweight (VLBW) infants improve outcomes?	Birth records, education data, social services records	Causal evidence about the impact on later life outcomes (including test scores, college attendance, and social service usage and expenditures) of providing additional medical care to VLBW infants
Causal analysis: Impact of Home Removal	What is the impact of home removal on children?	Child protective service case records, education data, criminal justice data, birth records	Causal evidence, by age and gender, about the impact on later life outcomes (including test scores, juvenile convictions, high school graduation, college attendance, and teenage births) of removing a child from an abusive or neglectful home
Causal analysis: Impact of Photo ID Laws	What is the impact of a photo ID law on voter participation and turnout?	Driver's license data, voting data	Causal evidence on whether a given photo ID law impacted voter participation and turnout in recent elections

Table 1. Summary of Current Available Knowledge Base Projects			
Research Tool	Research Question	Data Sources	Outcomes
Causal analysis: Impact of Work Training Sessions	Do mandatory return-to-work training sessions help communities find work and improve their wages?	Labor training data, including program participating and wage rolls	Causal evidence on whether requiring unemployment insurance applicants to attend mandatory return-to-work training sessions helps individuals find work and improve wages
Value-added model: In-Prison Labor Training Programs	Which in-prison training programs reduce recidivism?	Criminal justice data, training program participation data	Evidence that shows the causal impact of each program on an individual's likelihood of recidivating, controlling for background characteristics