The NYU Administrative Data Research Facility (ADRF)

The ADRF platform is part of the Coleridge Initiative. Two parallel and interconnected efforts have generated the Initiative's initial successes: building the Administrative Data Research Facility and running Applied Data Analytics training programs. It provides an approach to professionalize community access to and use of data on human subjects that has historically been limited, artisan and ad-hoc. The professionalization involves three key steps. The first of these is technical: to provide a secure environment within which data providers can place and share their data across agency and jurisdictional lines. The second is operational: to create enough capacity to link disparate data. The third is both legal and practical: to ensure that there is a value associated with the data linkage that is both consistent with the agency mission and useful enough to engage decision-makers. While the third is the most important in terms of creating institutional buy-in and will, the first and second are necessary before the third can happen.

The Coleridge Initiative built on many years of successful experience to design an infrastructure that incorporated those steps. The ADRF platform was commissioned by the US Census Bureau to inform the decision making of the Commission on Evidence Based Policy. The platform, combined with the training program, has provided services to almost 180 government agency staff and researchers, and hosted almost 50 confidential government datasets from 12 different agencies. This approach to acquire and link these confidential data is evidence that the substantial legal and political hurdles exist can be surmounted if the technical issues associated with providing a secure environment can be addressed and the value proposition is well defined.

Box 1: Administrative Data Research Facility:

The Administrative Data Research Facility is a pilot project that enables secure access to analytical tools, data storage and discovery services, and general computing resources for users, including Federal, state, and local government analysts and academic researchers. The Census Bureau and academic partners developed the project as part of the collaborative Training Program in Applied Data Analytics sponsored by the University of Chicago, New York University, and the University of Maryland. It is currently operating as a pilot with users accessing the Facility as part of the training program. The Facility operates as a cloud-based computing environment, with Federal security approvals, which currently hosts selected confidential data from the U.S. Department of Housing and Urban Development and the Census Bureau, as well as state, city, and county agencies, and an array of public use data.

Page 70, *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence Based Policy Making*. September 2017.

Step 1: Secure environment. The ADRF institutionalizes secure access to and use of confidential data. It is a secure cloud-based environment that is FedRAMP certified and has received

Authority to Operate from the US Census Bureau. The stamp of approval provides data owners with confidence that their data are secure. The cloud environment allows agencies within the same state or different states to agree to share their data in a common area in the cloud for specific approved projects.

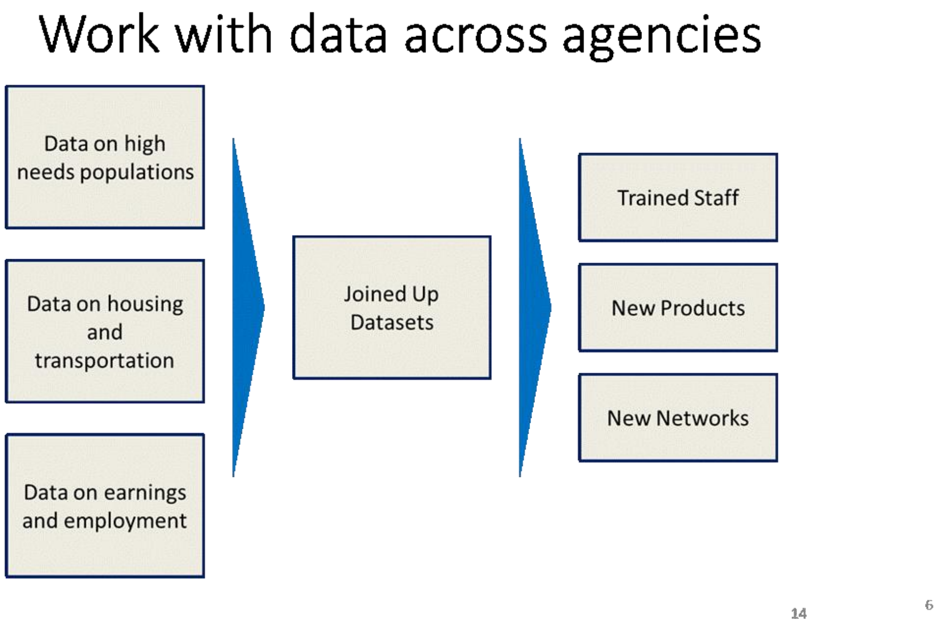
But the ADRF does more than provide a secure environment. Data providers simply will not provide data if they cannot monitor who is using it, for what purposes and with what results. We bake in a combination of state-of-the art technical strategies and thoughtful human oversight and screening in order to dramatically improve privacy and usage protections. We are building a variety of standardized mechanisms for different confidentiality situations, with mechanisms for certifying the five "safes": safe people, safe projects, safe settings, safe outputs and safe data.

The ADRF approach is based on tracing projects, data, and people. We are setting up an infrastructure to control who has access to which data and what ADRF content is related to that data. These functions are essential in that they provide controls while also enabling government and private sector data stewards to ask straightforward questions such as "which projects use my data?" or "how is my data being used and which byproducts were generated by whom?" If approved, staff from multiple agencies can jointly access approved areas in the cloud, so that they can work together to develop new integrated datasets, share information about coding differences or similarities, and develop common measures, without physically having to transfer data from one agency to another.

Step 2: Linking data: Getting data from different agencies in the same place is necessary but not sufficient. Linking datasets from different sources means that the entities in the separate files (individuals, household, families or businesses) need to be matched. This poses challenges well known in both social science and computer science and has multiple names – such as entity resolution or disambiguation, object consolidation, data linking, or duplicate detection. There are additional linkage challenges with sensitive government data. Precisely because access has been so limited and siloed, most datasets are either not well documented or are not documented with the purpose of being conceptually or electronically connected to other datasets. This is particularly true with historical data when knowledge about the way data are produced is often tacit rather than codified. Databases have historically been indexed and catalogued by input characteristics rather than by use. Code-sharing is limited, so analysis can rarely be replicated or reproduced. There is not a culture of replication or reproducibility so when linkage occurs it is often artisan, rule-based and fragile. If algorithms are used, they are often purchased and proprietary, which can result in unknown biases(*1*). Workforce capacity to preprocess and deduplicate information is limited, yet when combining datasets with different provenance, and of different temporal, entity and spatial granularity, such work is critical since 90% of linkage efficiency is due to preprocessing(*2*)

The ADRF is developing tools that are inspired by Amazon and Trip Advisor. The gamification approach enables users to contribute content about the data itself (variable definitions, coverage, limitations), preprocessing (code and lookup tables), analysis (code snippets and derived data), how the data have been used (research topics, methods and topics), who has used the data (names and locations) and code that has already been used to process and link data. In sum, we will automate as much of the linkage process as possible by discovering, mining, and then linking content about datasets, and then incentivizing people to contribute more.

Step 3: Creating value. The last step is building agency will to link data. The pressures to meet existing program needs make it difficult for agencies to try something new and create pipelines of new products based on linked data. Government salary structures make it difficult to hire and retain enough in-house data analysts, so agencies don't have the capacity to work with new linked data. These combined challenges have led to the current situation - agencies cannot get the significant resources necessary to make use of new data, and because they don't use new data, they don't get new resources. The Coleridge Initiative developed training classes in Applied Data Analytics. The classes create a sandbox environment within which agency staff – not outside vendors – provide concrete evidence of the value of linking data. The approach has to be built around agency needs and use modular learning approaches. Our classes (i) create a pipeline of new product prototypes central to agency missions, (ii) develop teams of skilled practitioners who have the capacity to both link data and apply modern analytical approaches to cross agency problems, and (iii) make a growing set of linked data available as an ongoing asset for budget analysis and program management.


Work with data across agencies
Data on high needs populations
Data on housing and transportation
Data on earnings and employment
Joined Up Datasets
Trained Staff
New Products
New Networks
14
6

Use Case 1

We trained three cohorts of government agency staff to work with linked data on welfare recipients (families on TANF and SNAP), ex-offenders, housing characteristics, transportation and jobs. At the behest of the Census Bureau, and based on our own network, we focused on designing classes (https://coleridgeinitiative.org/training) that would create value for three state agencies (Employment Security, Corrections, and Human Services) and one federal (Housing and Urban Development). That focus enabled us to surmount the legal and political challenges associated with data access. We built the classes around a core question of interest to all the agencies: the effect of neighborhood characteristics and access to jobs on the labor market outcomes and subsequent recidivism or retention on welfare of welfare recipients and ex-offenders.

The class itself was structured to train staff in agencies contributing data and others in developing such skills as managing and linking the relevant data, applying text analysis, network analysis and machine learning tools, thinking about inference and privacy and confidentiality issues and visualizing the results. We developed Jupyter Notebooks[1] for each skill set around the linked data so that participants could also use them in team research projects. One of these projects on the employment effects of technical parole violations on ex-offenders was selected by the Illinois Department of Corrections and the Illinois Department of Employment Security as worthy of a follow-on fellowship and possible policy changes. The list of all the projects (sites at Maryland (M), New York (N), Chicago (C), Washington state (W) and Connecticut (CT)) is below, including some bespoke projects by the NYPD and Mecklenburg County NC. It gives a flavor of the rich set of ideas that government analysts can develop if empowered.

Although the original vision was only to do three classes, the approach has been so successful that we have expanded to offer more classes, and many more confidential datasets are coming into play.

Cohort 1 Topics

N3 - Using Machine Learning to Identify Potential Victims of Gun Violence in New York City	M3 - From Prosecuted to Job Recruited: An Exploratory and machine Learning Approach to Employment After Prison
N4 - Mommy Don't Go: Predicting and Preventing Recidivism of Mother's in the Illinois Criminal Justice System	M1 - When the Machines Take Over: Utilizing Machine Learning to Predict Recidivism A Path Forward for the Criminal Justice System
M2 - Arrests in Mecklenburg County North Carolina 1990-2016	N2 - A Tool for Automated Information Curation

[1] http://jupyter.org/

C1 - Post Release Employment of the Formerly Incarcerated: Labor Market Prospective	N1 - Distance Matters: Using Administrative Data to Support Illinois Investments in Place-Based Initiatives for the Recently Incarcerated
C3 - Measuring 5-Year Recidivism of Persons Exiting IDOC	C4 - The Impact of Known Substance Abuse on Recidivism
M4 - Characteristics of Recidivism	C2 - Addressing Recidivism Intervening to Reduce Technical Violations and Improve Outcomes for Ex-Offenders

Cohort 2 Topics

N7: Predicting Return to Cash Assistance	M7: The Effect of Mass Layoffs on Earnings and SNAP Participation
M5: Does Increasing Minimum Wage Move People Off Welfare?	M6: Identifying Predictors of Success for TANF Recipients in Illinois
M8: Employment program participation outcomes for 18-24 year youth gslides	N5: Targeted Hiring
N8: NYC Young Adults Pathways Off Public Assistance	N6: Effect of First Job Sector After CA on Later Earnings
C5: Summer Jobs 4 U	

Cohort 3 Topics

UC 3: Predicting Recidivism due to Technical Violation	UMD 2: Predicting Future Employment Gap
UMD 4: Modeling the School to Prison pipeline in Chicago, IL	CT 1: Welfare dependency & transition from TANF
NYU 4: Success, the Flip-side to the Recidivism Conversation	UW 2: Mental Illness and Drug Dependency (MIDD) and access to public transit
UC 1: Benefits After Release: Does it make a difference?	NYU 2: Vulnerable Populations' Access to Points of Distribution for Public Health Emergencies
UMD 1: Predicting Future Earnings of Illinois Human Service Benefit Recipients	UMD 3: Characteristics of TANF recipients in Illinois
UW 1: Reducing Return to Welfare	NYU 1: How can we better predict which TANF recipients will be successful?
CT 2: A look into Peoria and Cook County Illinois - Measuring access to wages by location	NYU 1: How can we better predict which TANF recipients will be successful?

UMD 6: The TANF Ban and Criminal Activity: Evidence from IL Admin Data	UMD 5: First time prisoners: predicting recidivism
	NYU 3: Are Persons Released from Illinois Department of Corrections Receiving Needed Social Service Benefits?

Use case 2

The vision has become real in other countries. For example, in 2001, Julia Lane was responsible for laying the foundation for the Integrated Data Infrastructure[2] in New Zealand (*3*) – Figure 1 provides an overview of the structure in 2017.

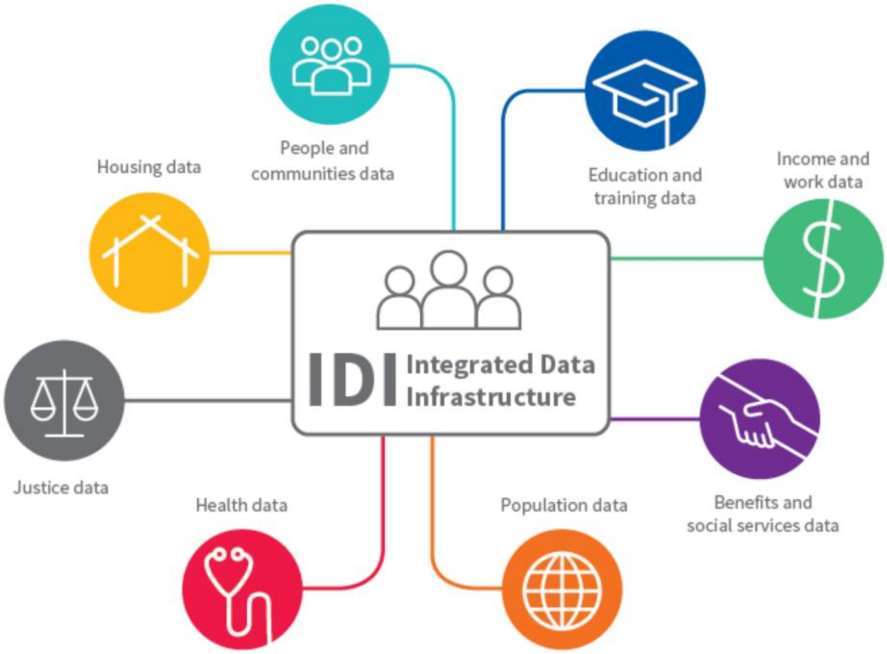

People and communities data
Education and training data
Housing data
Income and work data
IDI Integrated Data Infrastructure
Justice data
Benefits and social services data
Health data
Population data

Figure 1: The New Zealand Integrated Data Infrastructure

Examples of the way in which income and work data have been used to help improve New Zealanders' access to "good" jobs include

- Careers NZ's Compare study options helps young people make better decisions about where their study choices can lead them. The Ministry of Education created this tool by using combined student loan, tax, and education data.
- Training providers use the research to improve employment outcomes for youth at risk – such as detailed information about where to hold job fairs and what kind of training to provide.

Probably the best indicator of the success of the program is that the approach is now being emulated in Australia[3] with a $A35 million investment by the federal government.

[2] http://m.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx

[3] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Statistical+Data+Integration+-+MADIP

References

1. C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy* (Broadway Books, 2017).

2. W. E. Winkler, in *Part A of Handbook of Statistics*, C. In Rao, Ed. (Elsevier, 2009), pp. 351–380.

3. J. Lane, T. Maloney, Overview: The New Zealand conference on database integration and linked employer-employee data (2002).