

Big Data Use and Applications

Julia Lane

NYU



Fede
Leverag

About Principles Practices

H.R. 1831: Evidence-Based Federal Data Strategy Act of 2016

Introduced: Apr 16, 2015
114th Congress, 2015-

Status: **Enacted — Signed by the President**
This bill was enacted into law.

Law: Pub.L. 114-140

Sponsor: **Paul Ryan**
Representative, Republican

Text: [Read Text](#)
Last Updated: Apr 16, 2015
Length: 5 pages

Bill Summary

- [2020](#) designates the CEDC on the federal data strategy
- [CEDC](#) continues to coordinate federal data use and management
- [American](#) continues to prioritize federal data use and management
- [Geographic](#) smart data management
- [Administrative](#) and federal data linkages and protections
- [Economic](#) are re

The Federal Data Strategy Join the Federal Data Strategy Team

The Federal Data Strategy Team is working to create a coordinated approach to federal data use and management that serve the public. Subscribe to our updates.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

CONSENSUS STUDY REPORT

DATA SCIENCE FOR UNDERGRADUATES OPPORTUNITIES AND OPTIONS



Points Team News Feedback



Evidence-Based Act of 2018

Builds off the work of the U.S. Commission on Evidence-Based Data, and enhance the federal government's capacity for producing

Makes Administrative Records Available for Evidence Building. Under a strong set of confidentiality protections, encourages that government data can and should be used to generate evidence about policies and programs, unless otherwise restricted by law.

Creates a Common Portal for Researcher Applications to Access Restricted Data. Reduces burden on researchers for applying to access government data by establishing a common application system for qualified individuals to access restricted, confidential data for approved projects.

Facilitates Continuous Feedback about Data Coordination. Promotes the use of data for evidence building by establishing a government advisory committee to review existing coordination and availability of data.

Enhances Government's Evidence Capacity

Directs Agencies to Develop Evidence Plans. Enables agencies to better prioritize evidence building by requiring that agencies document their key research questions, data needs, and planned activities.

Prioritizes Evaluation Activities in Agencies. Improves agency capacity to engage in and use program evaluation by establishing evaluation officers in government agencies and requiring agencies to develop written evaluation policies.

Develops Baseline Information about the Resources Available for Evidence Building

Available for Evidence Building. Directs government agencies to periodically assess and report on their capabilities to engage in statistical, evaluation, and policy analysis activities and use the corresponding evidence for day-to-day government operations.

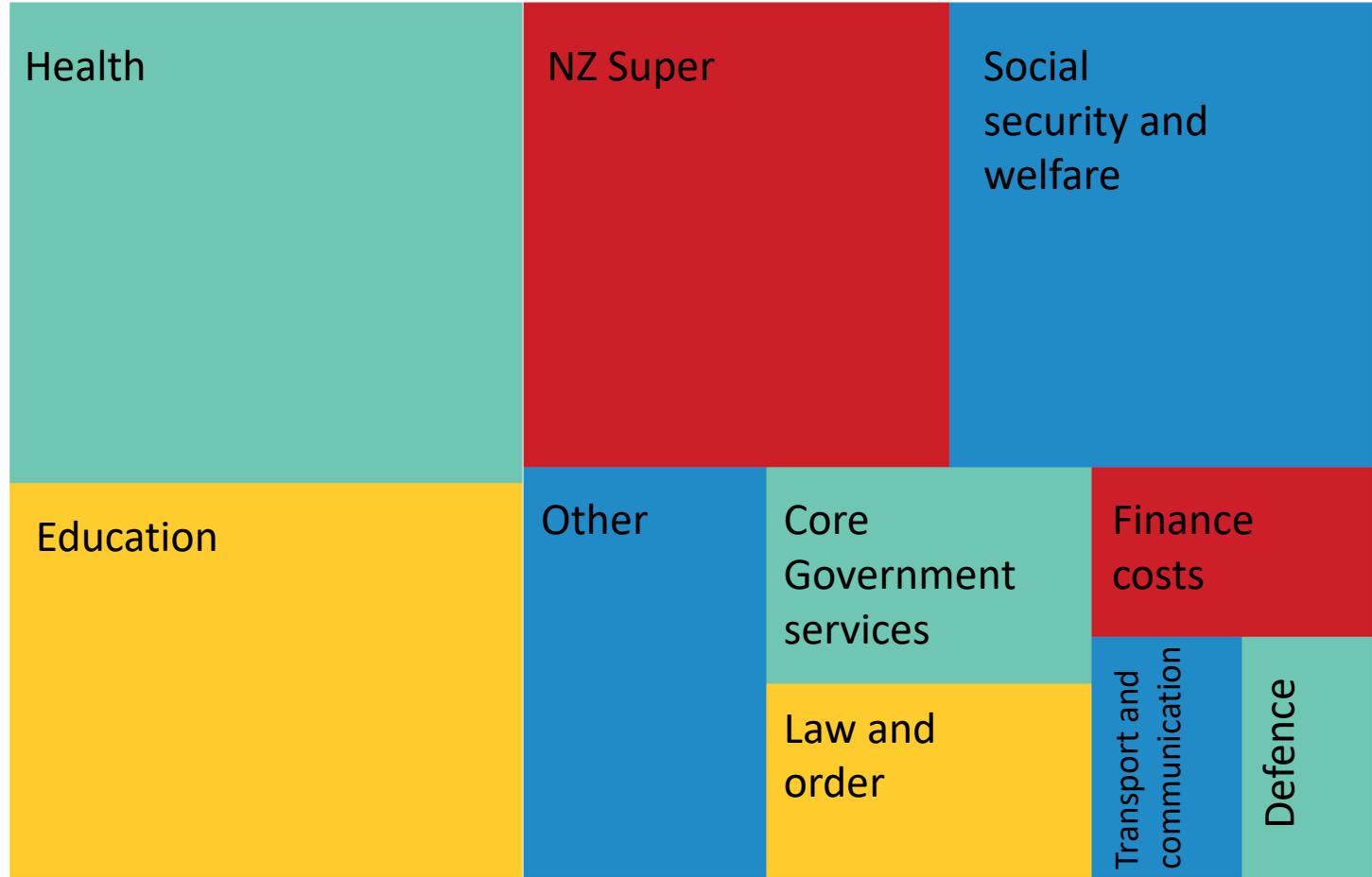
[npolicy.org/evidence](#)

The Federal Data Strategy: Principles and Practices

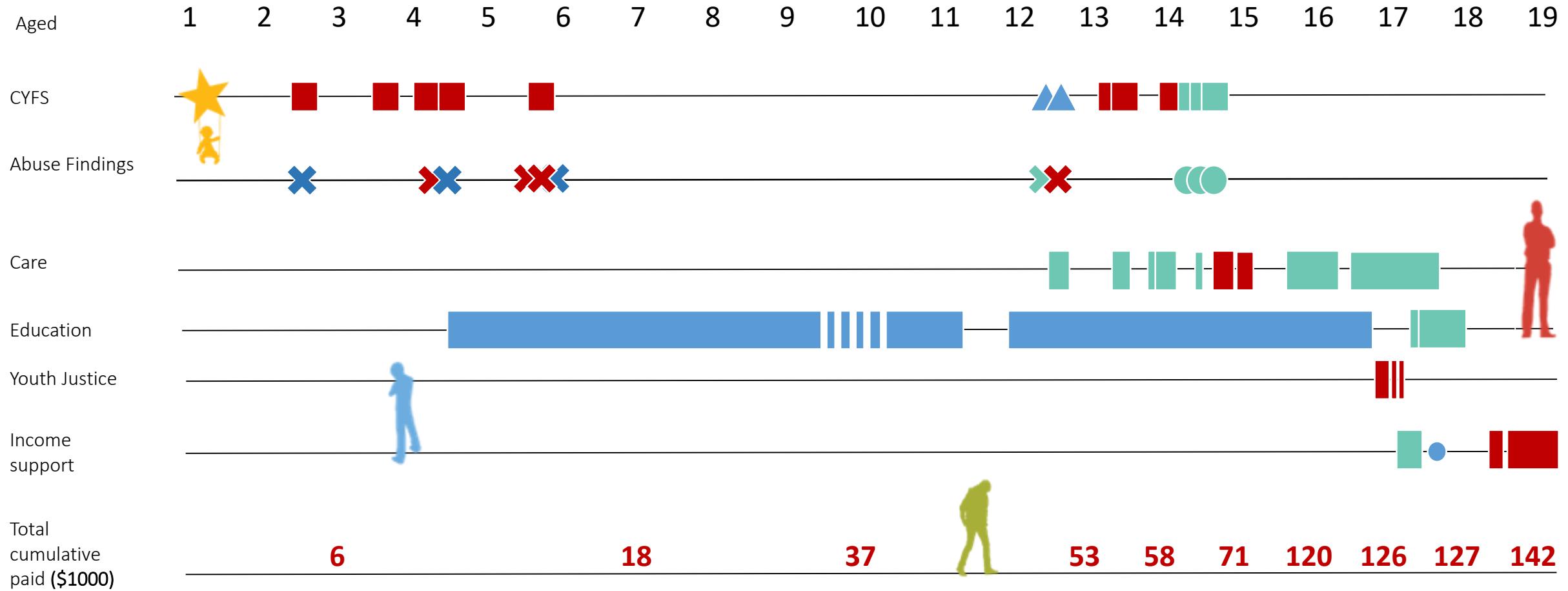
A New Zealand Example



Budget 2018



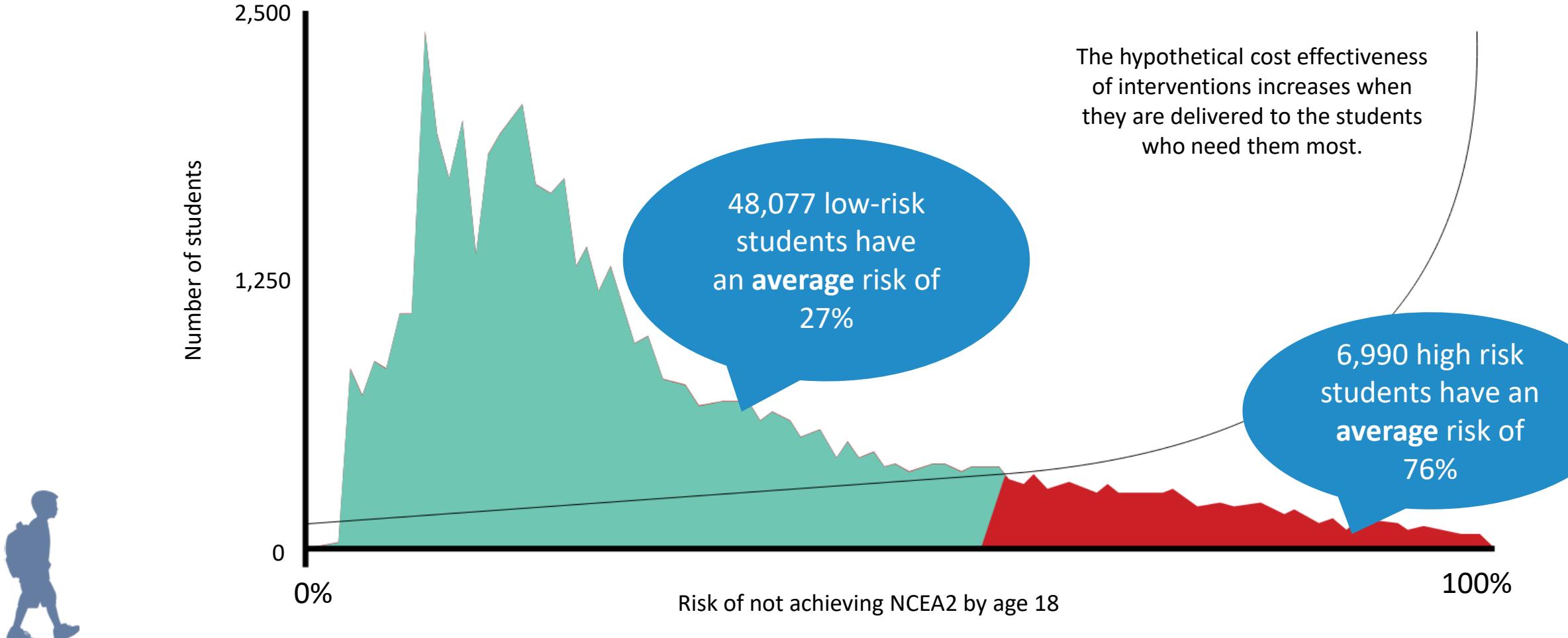
What do we mean by integrated data?



Sir Bill English, Former NZ PM

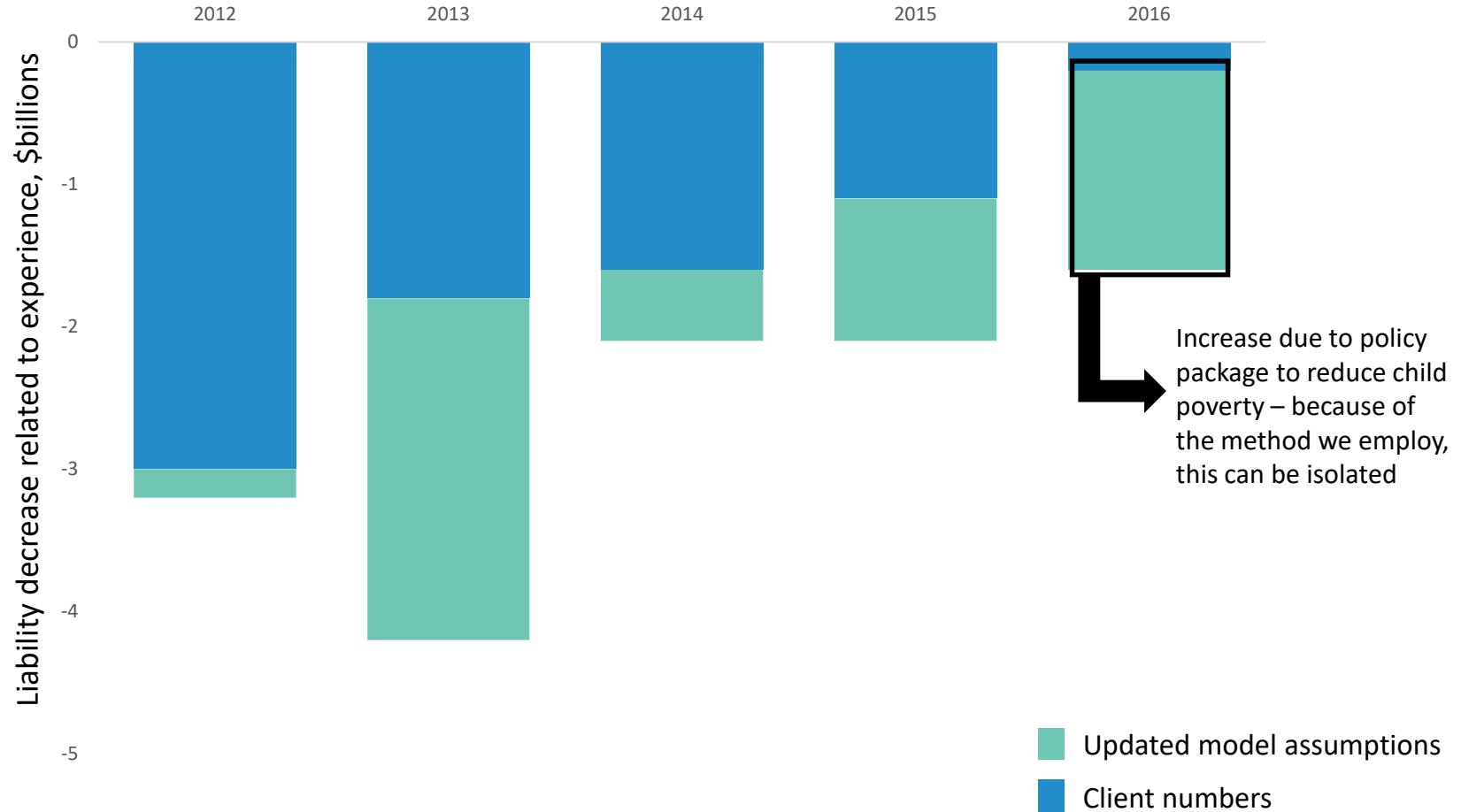
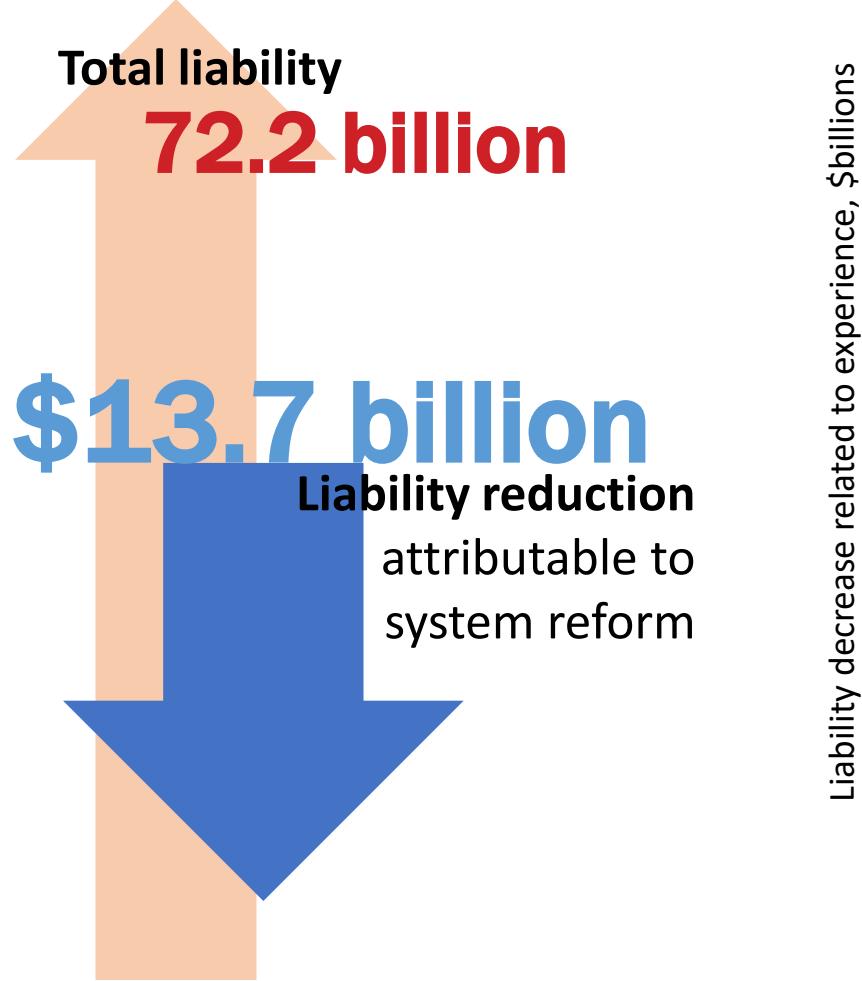
Understanding risk

The risk profile of this cohort reveals a long tail of increased risk.



Sir Bill English, Former NZ PM

Aggregate liability valuation



Sir Bill English, Former NZ PM

Examples of Questions in US

HHS

- What characteristics increase an individual's risk of returning to TANF?
- What are the employment outcomes of TANF leavers?
- What are the characteristics of those at-risk of not finding stable employment?

Education

- What factors predict on-time graduation from college?
- What factors predict post-graduation employment outcomes?

USDA/WIC

- What are WIC households' total food expenditures? What is the share of WIC purchases?
- Do WIC households purchase similar foods compared to households that do not participate in WIC but are eligible?
- What are the prices of nutritionally-eligible food products that are and are not WIC-approved in a given State?

Training Framework



Search Login or Signup

HOME ABOUT ▾ MISSION AND SCOPE ISSUE 1.1



1.2 ▾ Stepping Stones

Change Through Data: A Data Analytics Training Program for Government Employees

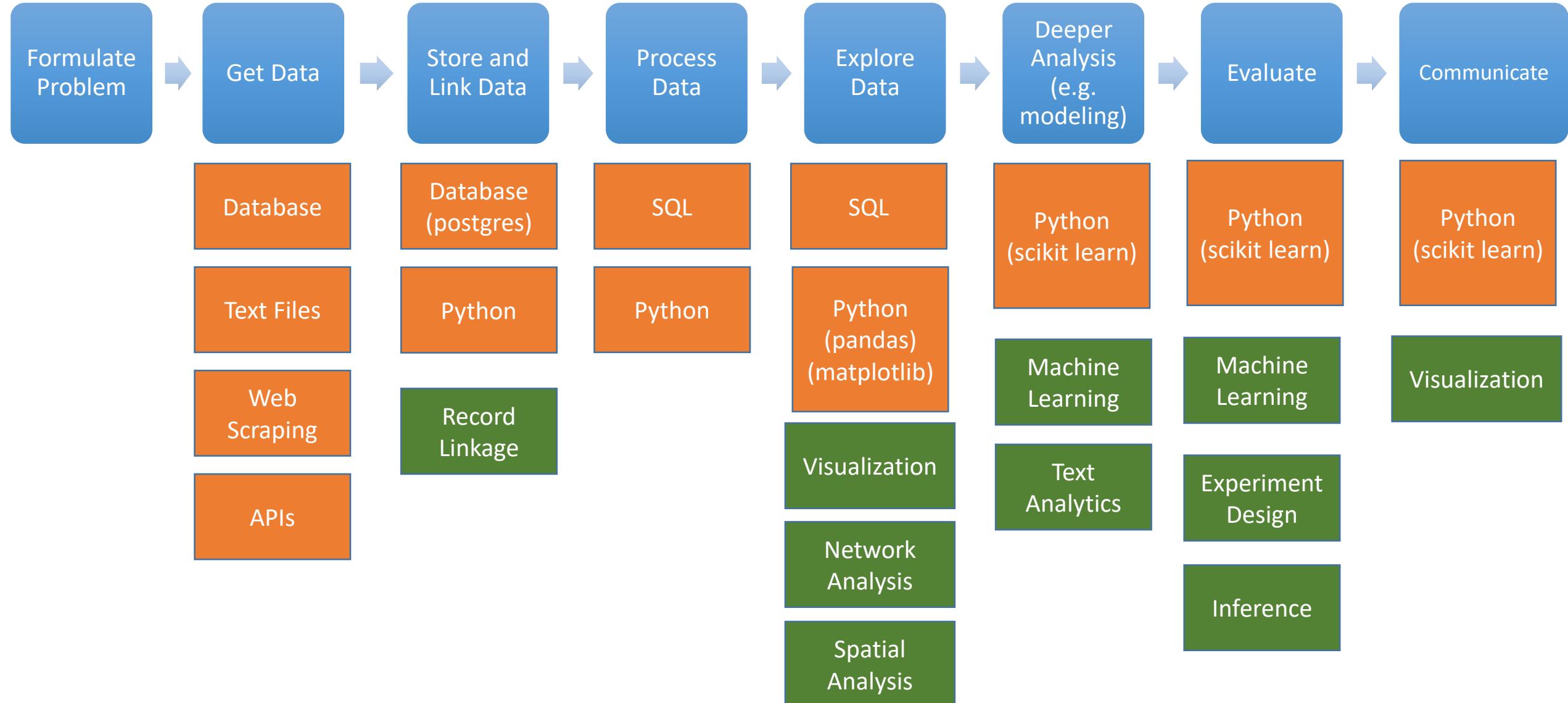
by Frauke Kreuter, Rayid Ghani, and Julia Lane

now on branch
#public ▾



Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



Collaboration: “shared” folders in your project

Dataset overview (example)

- Missouri Division of Employment Security
 - Documentation link, UI wage record data dictionary, QCEW employer data dictionary
- Missouri Department of Labor
 - Documentation link, UI claims data dictionary, UI payments data dictionary
- Missouri Department of Higher Education
 - Documentation link, Enrollment data dictionary, Completion data dictionary, Term data dictionary
- Missouri Department of Corrections
 - Documentation link, data dictionaries for: releases, classes, and SSN to DOCID

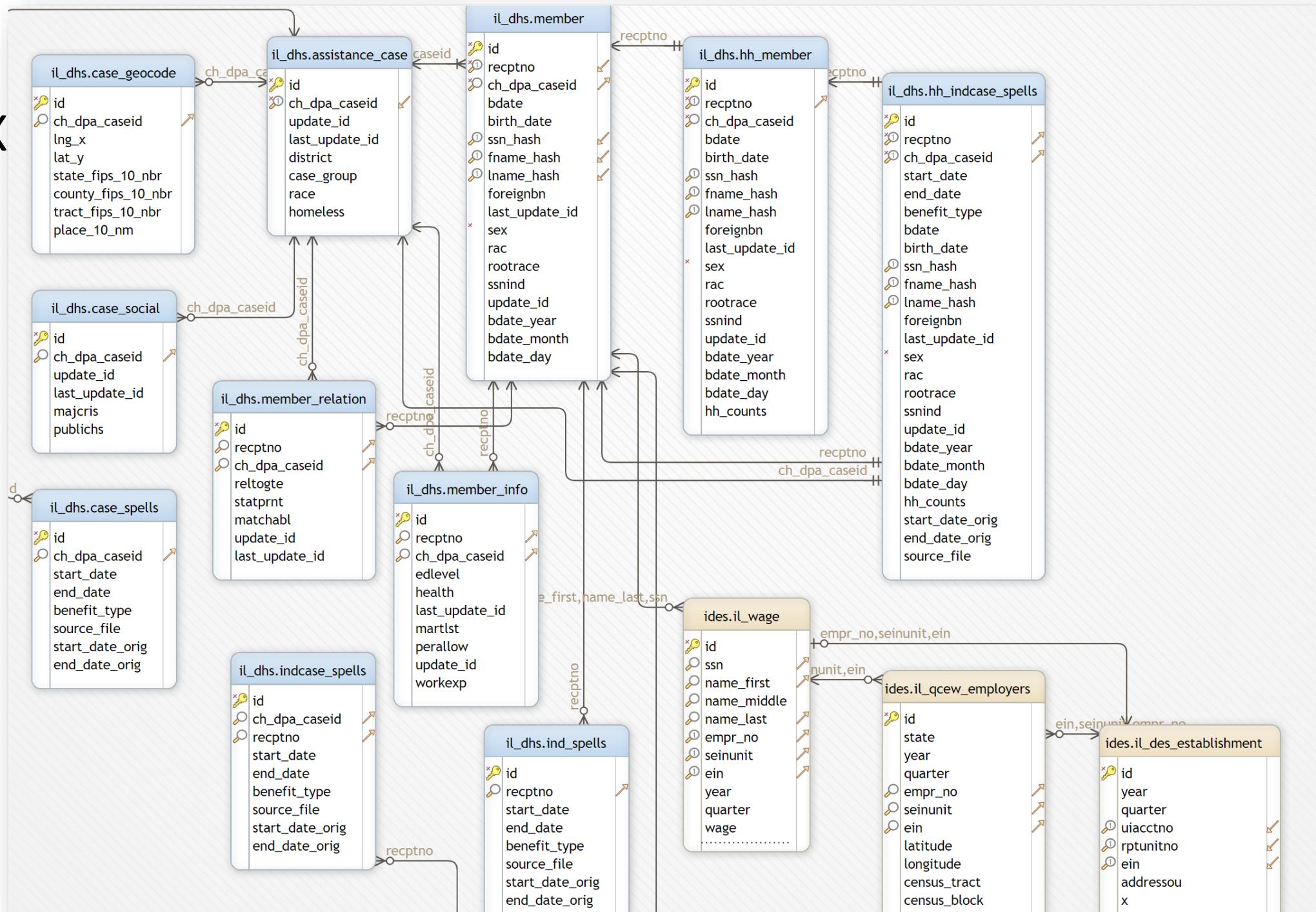
<https://ada.coleridgeinitiative.org/ada-2019-missouri>

Dataset overview (example contd)

- Indiana datasets
 - Wage records and Education [detailed data documentation](#), [data overview](#)
- Ohio Longitudinal Data Archive datasets
 - [Dataset overview](#), [data details](#)
- Illinois Department of Employment Services
 - [Documentation link](#)

<https://ada.coleridgeinitiative.org/ada-2019-missouri>

Ex



Examples of Demand



← → C ⌂ coleridgeinitiative.org/training

Training Computing Connecting Rich Content Resources Events About

TRAINING

TRAINING PROGRAM

The Applied Data Analytics programs are targeted at government agency staff. It provides training in core data analytics techniques by working on specific projects using real-world micro-data. The projects are built around pre-built Jupyter notebooks which provide project "recipes" that can be customized for specific use cases as well as applied to later projects in participants' agencies.

Core Curriculum

Activity	What	Time Required
Application (online/remote)	Online, example: here (incl. security requirements)	About 2 hours total
Intro to SQL & Python (online/remote)	Online videos with web-based content; weekly online 1 hour group discussions	Up to 10 hours
Module 1 (in-person)	Introduce the program, data, and projects includes data exploration, visualization, record linkage, and introduction to Machine Learning	3 days, on-site
Project work (online/remote)	Self-paced project work with teams; instructors available for assistance	Up to 10 hours, suggest at least 2 hours per week
Module 2 (in-person)	Focus on projects with sessions on Inference and Confidentiality	3 days, on-site
Project work (online/remote)	Complete team projects; recommend weekly check-in with an instructor	Up to 10 hours, suggest at least 2 hours per week
Presentations (online/remote)	Present final projects via webex	30 minutes per team; up to 3 hours

The following projects are examples from previous programs.

Addressing Recidivism: Technical Violations

Mommy Don't Go: Recidivism of Mothers

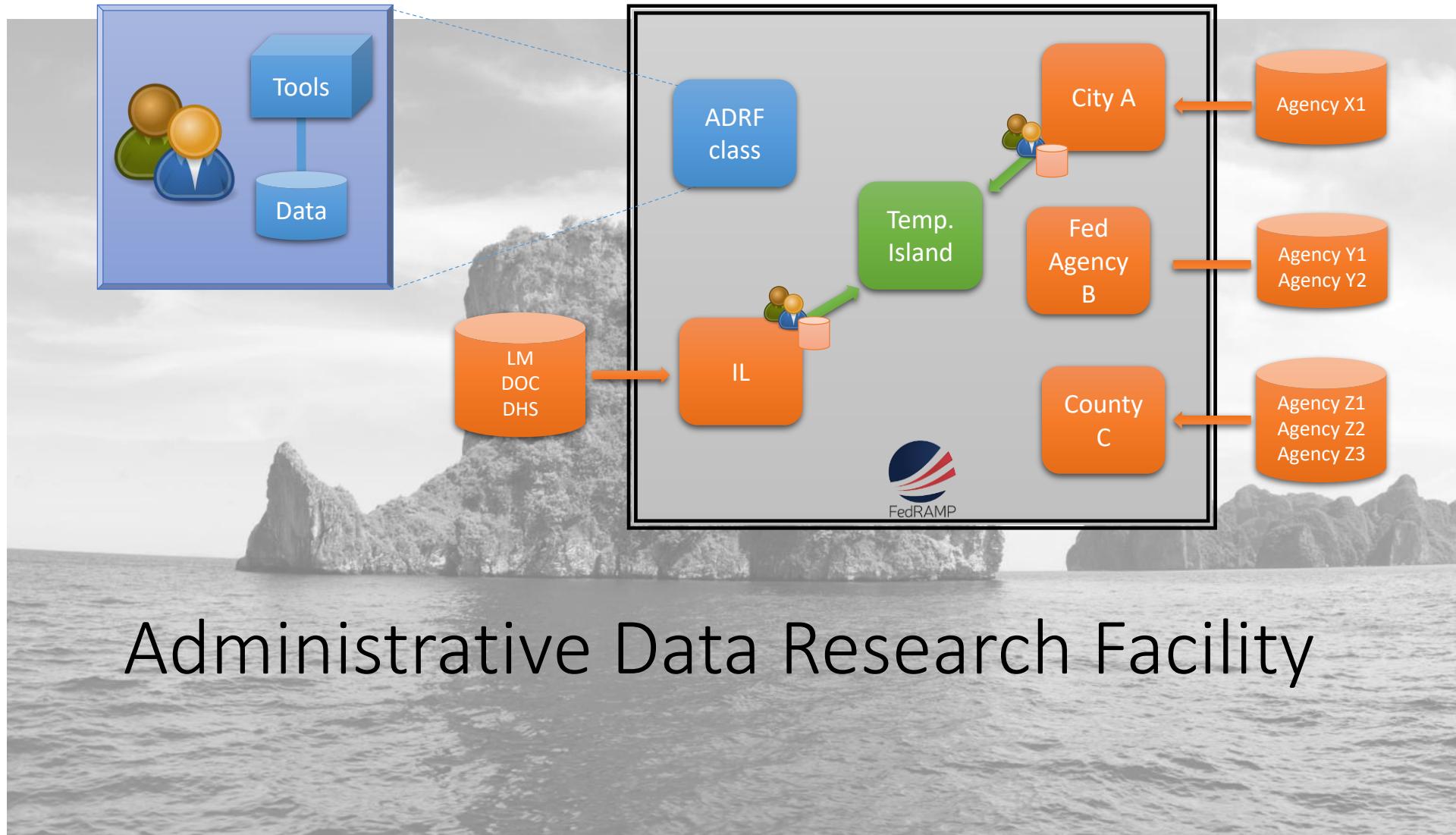
From Prosecuted to Job Recruited: Employment after Prison

UPCOMING TRAINING PROGRAMS

USDA - Alexandria, VA - Fall 2019

Application: **Closed**
Introduction to SQL & Python begins **September 24 (online)**
SQL & Python tutorials: Complete by **October 28**
In-person training in Alexandria, VA
• Module 1: **October 30 - November 1**
• Module 2: **November 20-22**
Remote Presentations: **December 18**

Technical Environment



Technical: Automate Data Approval Workflow

The screenshot displays the ADRF Data Stewardship Project Request interface. The top navigation bar includes links for Projects, Explore Data, User Directory, and Bookmarks. The main header "Project Request" is followed by a sub-header "Please provide the following information to initiate a project request in the ADRF. Your request will be automatically be routed to the appropriate agencies and reviewers upon submission." Below this, there are tabs for Overview, Members (1), Datasets, and Data Security. The Members tab is active, showing a dropdown labeled "Pick new member" with two entries: "Graham Henke" (grh255@nyu.edu) and "Gonen Minuskin" (gm130@nyu.edu), each with a "Remove" button. Navigation buttons "Previous" and "Next" are at the bottom of this section. To the right, there are "Save as Draft" and "Submit" buttons. The Datasets tab shows a list titled "Decennial Census Illinois Profile of General Population and Housing Characteristics: 2000" with Dataset ID: adrf-000005 and Data Steward: Drew Gordon, with a "Remove" button. Navigation buttons "Previous" and "Next" are at the bottom of this section. At the very bottom, there are "Save as Draft" and "Submit" buttons.

ADRDF | Data Stewardship

Navigation

- Projects
- Explore Data
- User Directory
- Bookmarks

Project Request

Please provide the following information to initiate a project request in the ADRF. Your request will be automatically be routed to the appropriate agencies and reviewers upon submission.

Overview **Members (1)** **Datasets** **Data Security**

Select the project members from the list below

Pick new member

Graham Henke
grh255@nyu.edu **Remove**

Gonen Minuskin
gm130@nyu.edu **Remove**

Previous **Next**

Save as Draft **Submit**

Pick a dataset

Decennial Census Illinois Profile of General Population and Housing Characteristics: 2000

Dataset ID: adrf-000005

Data Steward: Drew Gordon

Remove

Previous **Next**

Save as Draft **Submit**

Technical: Streamline legal and operational

The screenshot displays the ADRF Data Stewardship Project Review interface. At the top left, the text "ADRF | Data Stewardship" is visible. On the top right, there is a user profile icon with a blue circle containing the number "5" and the name "Michael Scott".
The main header "Project Review" is centered above a navigation bar. The navigation bar includes links for "Overview", "Members", "Datasets", "Agreements" (which is currently selected), and "Data Security".
Below the navigation bar, the page is divided into several sections:

- Institutional Agreements:** A section for uploading agreements to be signed by institutional representatives. It includes a note that documents will be shared with Data Stewards, the Principal Investigator, and Project Requester, but not with project members. A blue "Attach agreement" button is present.
- Individual Agreements:** A section for uploading agreements to be distributed and signed by each project member. It includes a note about Non-Disclosure Agreements, Data Use Agreements, etc. A blue "Attach agreement" button is present.
- Feedback:** A large text input area for leaving feedback, with a blue "Leave feedback" button at the bottom.

At the bottom of the page, there are two large buttons: a red "Reject" button on the left and a blue "Approve" button on the right.

Technical Report on Access and Use

ADR Data Stewardship

Navigation

- Projects
- Explore Data
- User Directory
- Data Steward **NEW**
- Dashboard
- Project Requests (5)
- Incoming Questions
- Data Export Requests

Bookmarks

ACTIVE PROJECTS 3 VIE

Data Steward Dashboard

Dataset Views in the ADRF Explore

How often your data was accessed

Dataset Name	Researchers who have access	Used in projects	
		Used in projects	Used in completed projects
National Consumer Panel - Demographics	225	38	15
National Consumer Panel - Trips	361	48	17
Med Profiler Survey	55	22	13
RX Pulse Longitudinal Panel	223	28	13
POS RMA Level	369	45	11
POS Private Label	249	24	12
POS Store Level	112	24	8
Random Weight RMA Level	319	21	13
Random Weight Store Level	195	34	5
Product Dictionary IRI	278	42	15
Random Weight Dictionary	69	38	8
Store Dictionary	313	50	24

Top Users of Your Datasets

User Name	Datasets Used
Bartholomew Kostopopolous	32
Jennifer Tour Chayes	28
Grace Hopper	17
Bill Nye	12
Claudia Perlich	10

Top Institutions Using Your Datasets

Institution	Datasets Used
Center for Urban Science + Progress	48
Chapin Hall at the University of Chicago	42
Urban Center for Computation & Data	39
Census Bureau	27
City of New York	24

Tech

ADR福 | Data Stewardship +

ds.adrf.cloud/publications/usda_pub_000023

★ RP V 30 ARC ...

Graham Henke

ADR福

Navigation

- Projects
- Explorer
- User Directory
- Data Steward
- Bookmarks
- Onboarding

«

USDA ERS - The Food-Spending Patterns of Households Participating in the Supplemental Nutrition Assistance Program: Findings From USDA's FoodAPS

Abstract

This study uses data from USDA's FoodAPS survey to compare food expenditures of SNAP households with those of eligible nonparticipant households and households overall. It examines variations in food spending by SNAP households' characteristics, the contribution of SNAP benefits to household food expenditures, and changes in food-spending patterns over the month after receipt of benefits. See related Amber Waves article USDA's FoodAPS: Providing Insights Into U.S. Food Demand and Food Assistance Programs.

Authors	Publication Information	Publication Metadata
John A. Kirlin 1 Publications 0 Citations	ARTICLE CLASSIFICATION Public DATA CITATION Tiehen, L., C. Newman, and J.A. Kirlin. The Food-Spending Patterns of Households Participating in the Supplemental Nutrition Assistance Program: Findings From USDAs FoodAPS, EIB-176, USDA, ERS, August 2017. SOURCE URL https://www.ers.usda.gov/publications/pub-details/?pubid=84779	ARTICLE ID usda_pub_000023 DOI TOPICS KEYWORDS FoodAPS, National Household Food Acquisition and Purchase Survey, food expenditures, Supplemental Nutrition Assistance Program, SNAP, food insecurity, SNAP benefit cycle

Related Datasets

UNITED STATES DEPARTMENT OF AGRICULTURE (USDA)

[FoodAPS National Household Food](#)

Examples of Results: Interstate data

1. Comparing Employment Dynamics Across Borders

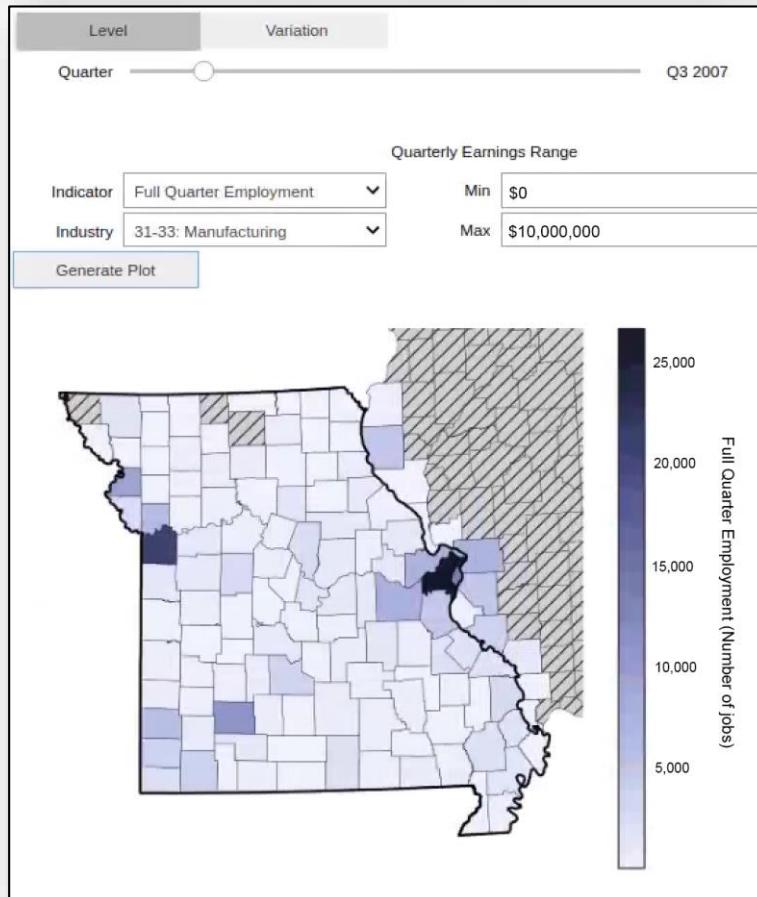


Fig. 3: Comparing total earnings with Illinois border counties

The dashboard can include border counties from the states that provide data to the ADRF.

UPCOMING TRAINING PROGRAMS
Columbus, OH - Winter 2019

The Ohio State University is hosting the Winter, 2019 Coleridge Initiative Applied Data Analytics training program, sponsored by the Overdeck Family Foundation. Participants will work in teams to define and complete a project related to Transitions in Education and Workforce. The program will provide up-to-date perspective on the use of administrative data for policy analysis, and hands-on instruction of using micro data in SQL and Python for the following tasks: data management, record linkage, data visualization, and machine learning.

APPLY

Application: Rolling
1st Admission Decision: December 15
SQL & Python tutorials: Complete by February 1
In-person training in Columbus, OH
• Module 1: February 6-8
• Module 2: March 13-15
Remote Presentations: April 23

Outcomes from Transitions in Education and Work: Developing a Scalable Regional Approach

Summary

There is enormous interest in building a better understanding of how people transition across different educational and work experiences to sustainable jobs. Transitions of people as they age can be nonlinear and include transitions within and across secondary and post secondary institutions, as well as the use of government services such as disability services, criminal justice interventions, or workforce development. Individuals can also move across political jurisdictions such as state or county lines, making it incredibly difficult to understand the regular patterns in service utilization. The sheer variation in types of transitions individuals can make and the geographical movement across borders opens up opportunities for governments to intervene in more productive ways. The data sharing across states and agencies also provides increasing possibilities for more efficient and effective utilization of government services among vulnerable populations.

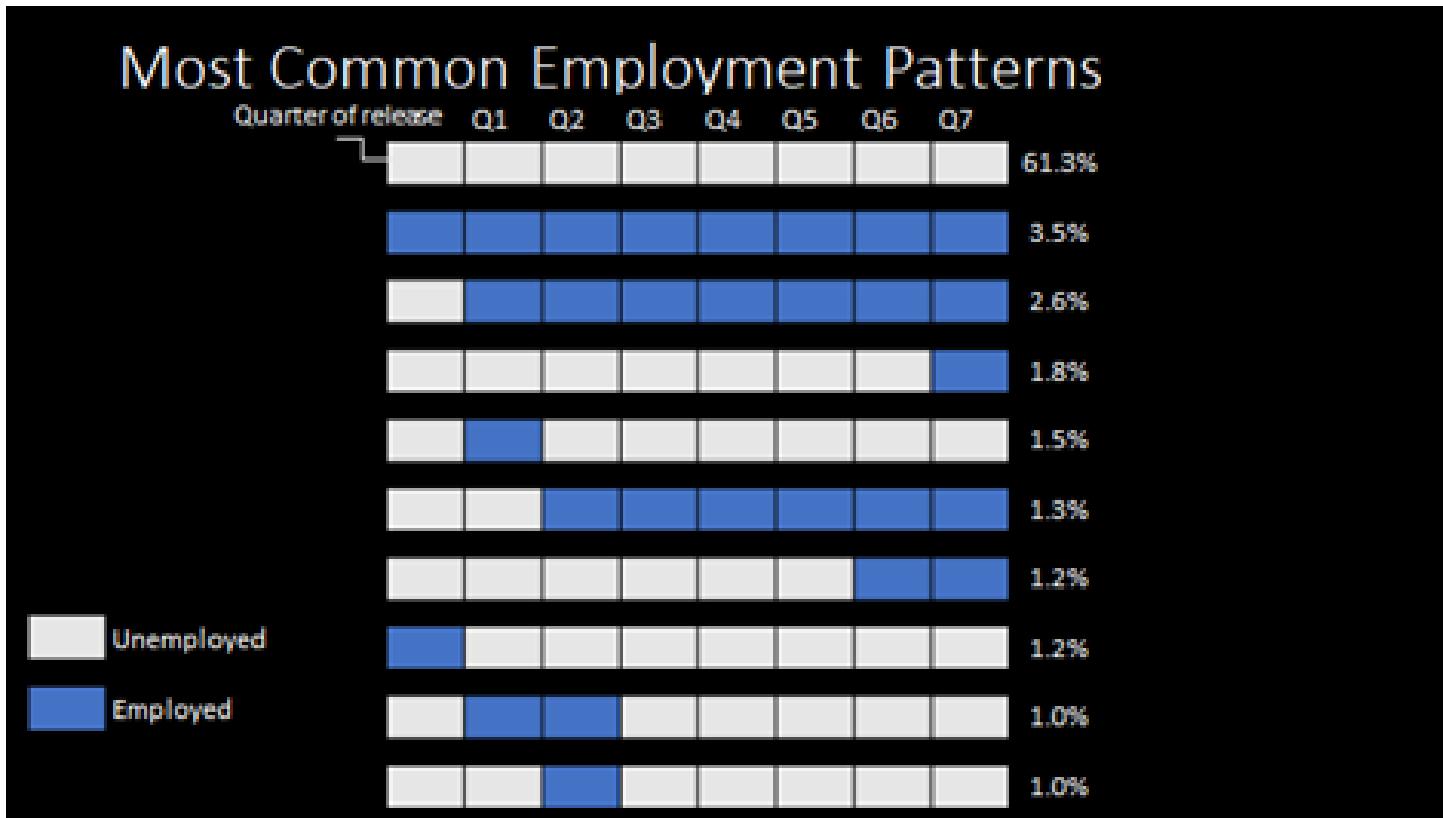
Building a common framework across states requires (i) identifying key issues of interest (ii) bringing together data from many sources across both state and agency lines and (iii) bringing together staff from those agencies. These three activities are necessary to work through the important data and measurement issues that are necessary to build operational metrics for decision-making. This workshop will build on a successful collaborative approach developed by the State of Illinois, New York University's Coleridge Initiative, Chapin Hall and Ohio State University that uses an inter-agency and inter-state collaborative, classroom environment to address the issues. It will bring together key stakeholders from the MidWest to develop an operational agenda for the spring of 2019. We expect to discuss how to develop such metrics as:

1. What are the main educational and work flows of interest for policy?
2. What data are available?
3. What kind of metrics would be of greatest use to stakeholders?

Examples of results: across agency lines

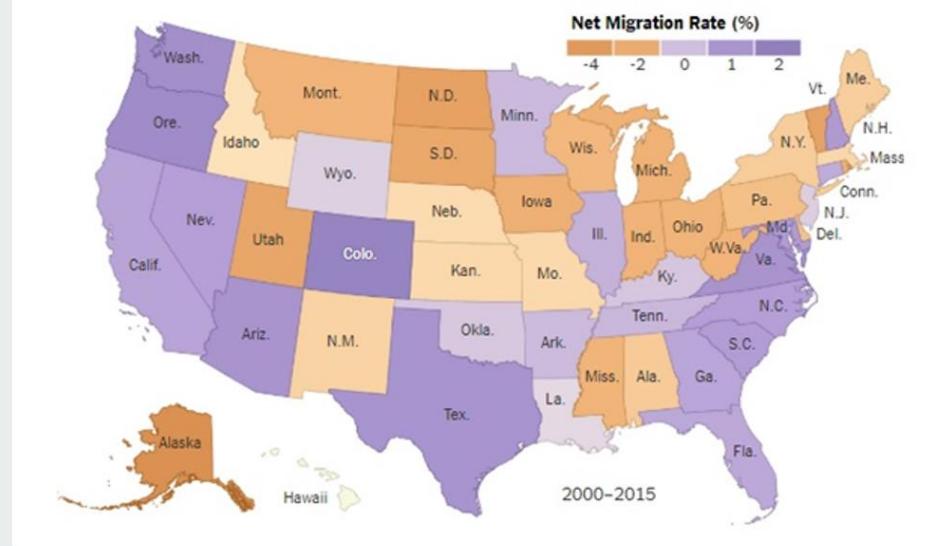
ADA Class w/ Illinois data

- Links IL Dept. of Corrections and Labor datasets
- Employment dynamics post prison release
- Inspires new ways to assess job training programs



Example

→ lines



Leaving Ohio: Migration of Nursing, Accounting, and Computer Science College Graduates

Group #3

Regionalization and Out-of-State Location

WHAT NEXT?

- What interventions can Indiana create knowing that bachelor's graduates with the following characteristics have a higher likelihood of being employed out of state the year after graduation:
 - Earning credentials in engineering, business, computer science, comm/journalism, transportation/materials moving
 - Non-residents
 - Originally from a border state
 - Have graduation dates in quarter 2 (April-June)
 - Asian



Fede
Leverag

About Principles Practices

H.R. 1831: Evidence-Based Federal Data Strategy Act of 2016

Introduced: Apr 16, 2015
114th Congress, 2015-

Status: Enacted — Signed by the President
This bill was enacted into law.

Law: Pub.L. 114-140

Sponsor: Paul Ryan
Representative, Republican

Text:
[Read Text](#)
Last Updated: Apr 16, 2015
Length: 5 pages

Similar bills

2020

design

on the

CED

Amer

contin

Geog

smart

Adm

and fe

linkag

prot

Econ

are re

The Federal Data Strategy Join the Federal Data Strategy Team

The Federal Data Strategy aims to create a coordinated approach to federal data use and management that serve the public. Subscribe to our updates.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

CONSENSUS STUDY REPORT

DATA SCIENCE FOR UNDERGRADUATES OPPORTUNITIES AND OPTIONS



Points Team News Feedback

Evidence-Based Act of 2018

Builds off the work of the U.S. Commission on Evidence-Based Data, and enhance the federal government's capacity for producing

Makes Administrative Records Available for Evidence Building. Under a strong set of confidentiality protections, encourages that government data can and should be used to generate evidence about policies and programs, unless otherwise restricted by law.

Creates a Common Portal for Researcher Applications to Access Restricted Data. Reduces burden on researchers for applying to access government data by establishing a common application system for qualified individuals to access restricted, confidential data for approved projects.

Facilitates Continuous Feedback about Data Coordination. Promotes the use of data for evidence building by establishing a government advisory committee to review existing coordination and availability of data.

Enhances Government's Evidence Capacity

Directs Agencies to Develop Evidence Plans. Enables agencies to better prioritize evidence building by requiring that agencies document their key research questions, data needs, and planned activities.

Prioritizes Evaluation Activities in Agencies. Improves agency capacity to engage in and use program evaluation by establishing evaluation officers in government agencies and requiring agencies to develop written evaluation policies.

Develops Baseline Information about the Resources Available for Evidence Building.

Directs government agencies to periodically assess and report on their capabilities to engage in statistical, evaluation, and policy analysis activities and use the corresponding evidence for day-to-day government operations.

[npolicy.org/evidence](#)

The Federal Data Strategy: Principles and Practices

Thank you!

README.md

- **Big Data and Social Science**
-
-

You can view this textbook at <https://coleridge-initiative.github.io/big-data-and-social-science/>

Soliciting feedback and improvement suggestions for the second edition

We are currently working on a 2nd edition of the book. Please give us suggestions for improvements and additional content by creating a [github issue](#) in this repository.

- E-mail

- dataanalytics@coleridgeinitiative.org
- julia.lane@nyu.edu