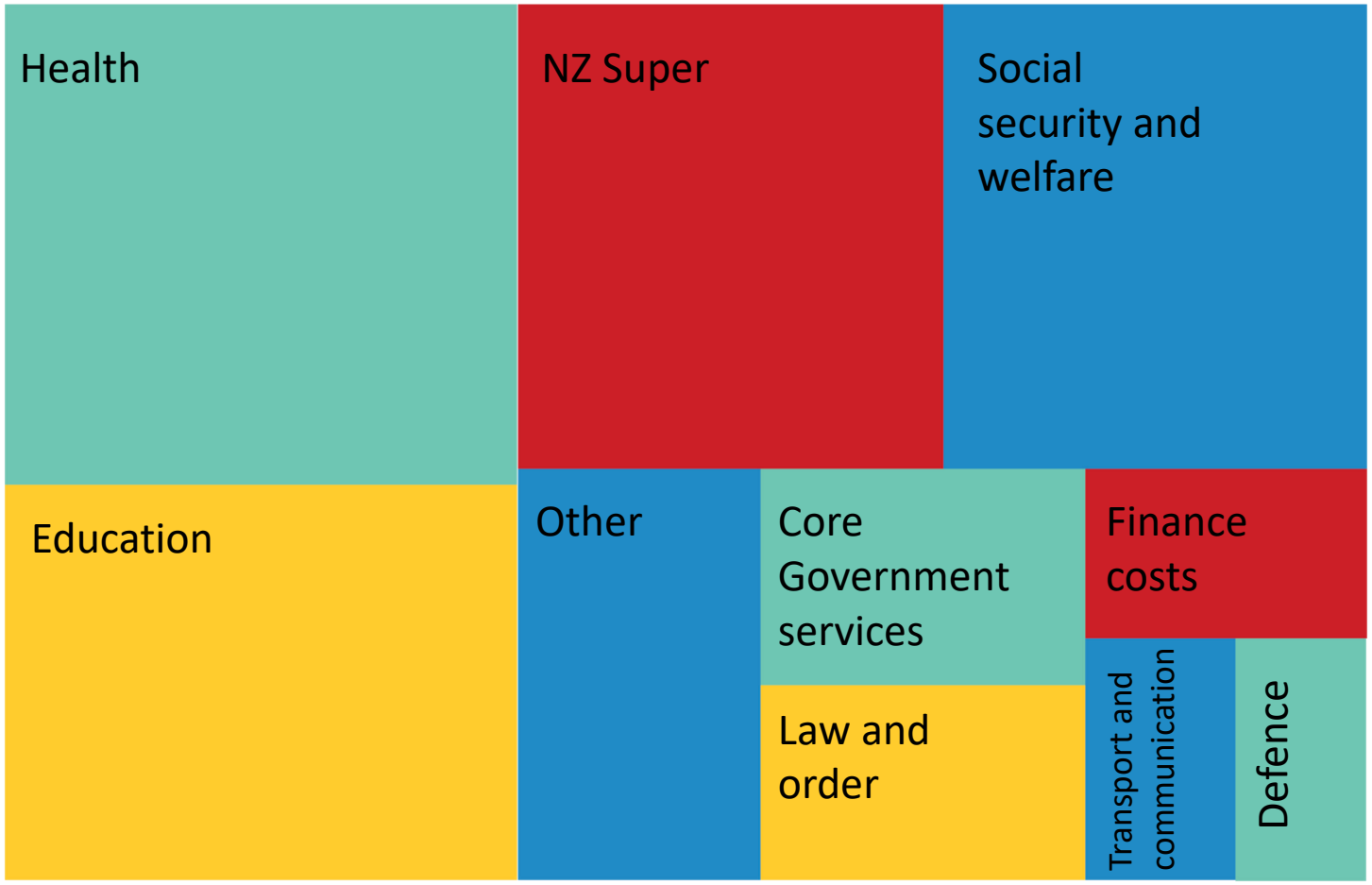
Big Data Use and Applications

Julia Lane

NYU

The Federal Data Strategy: Principles and Practices

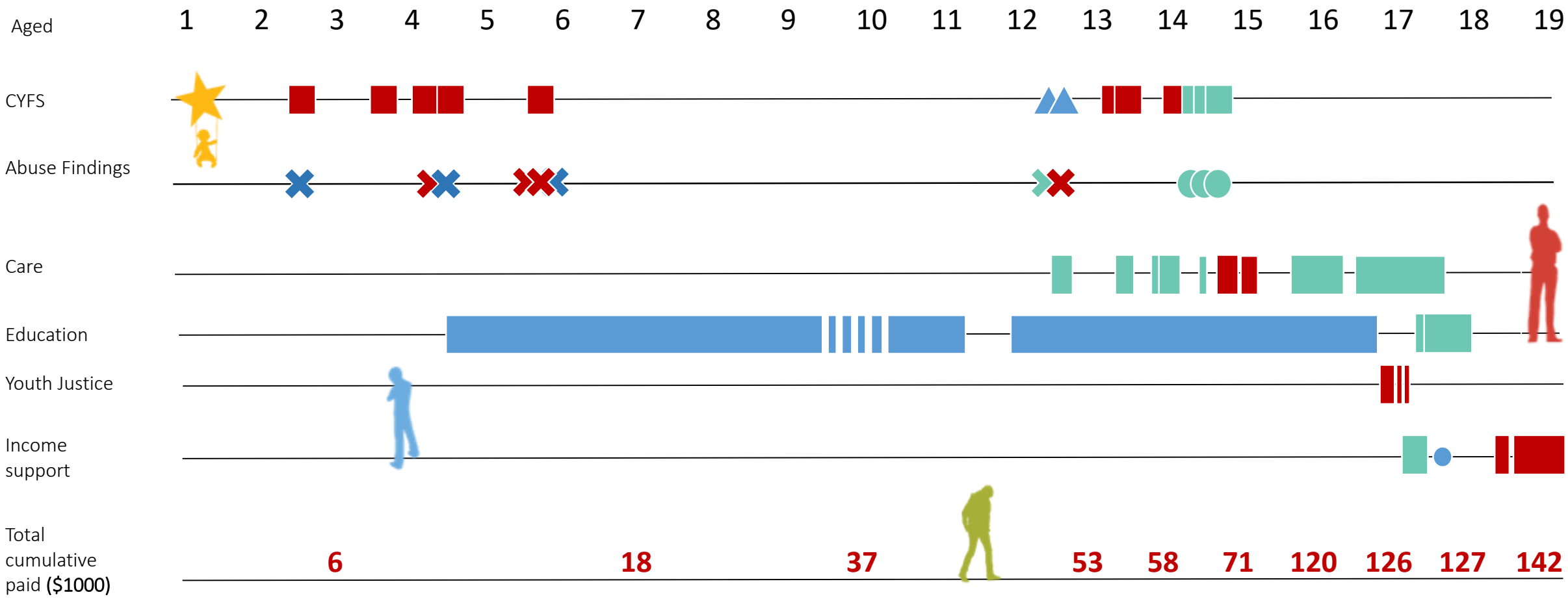
A New Zealand Example



Budget 2018

Sir Bill English, Former NZ PM

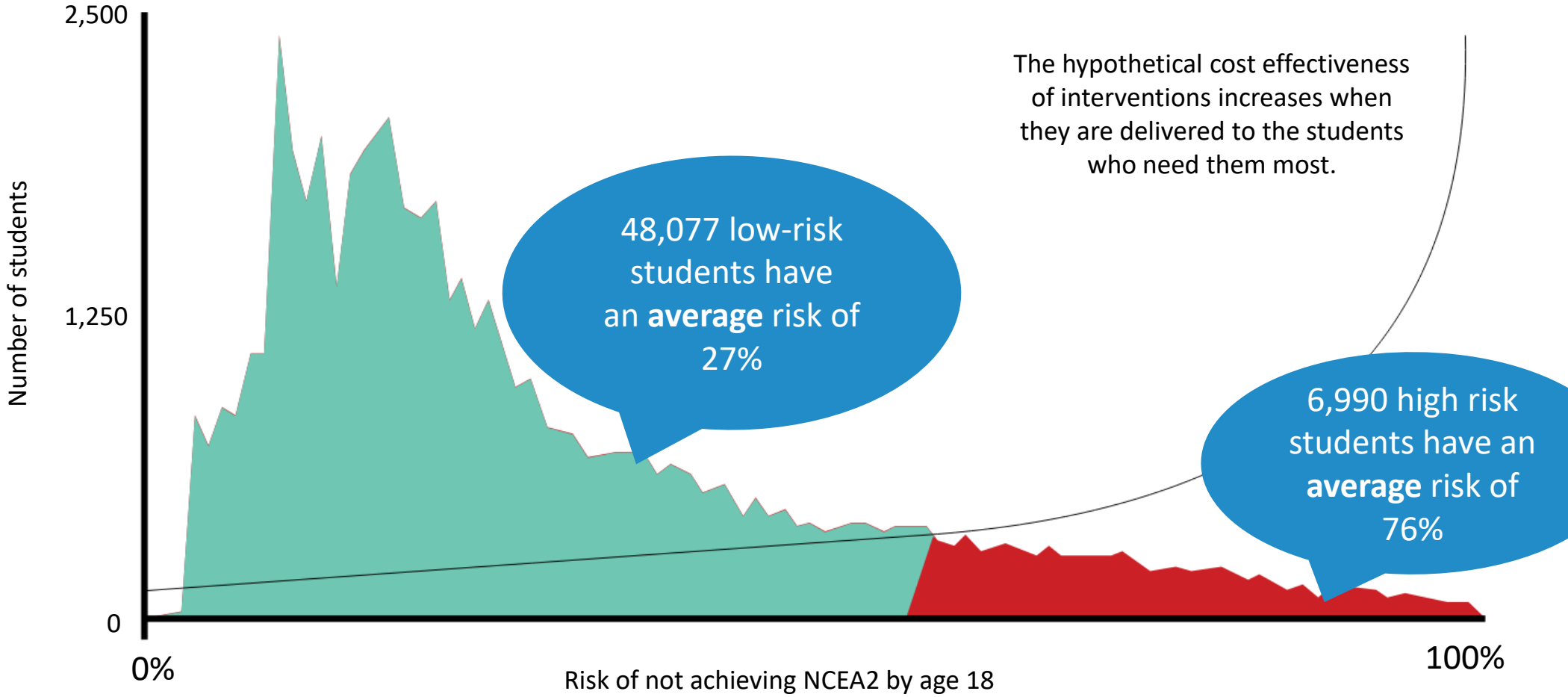
What do we mean by integrated data?



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

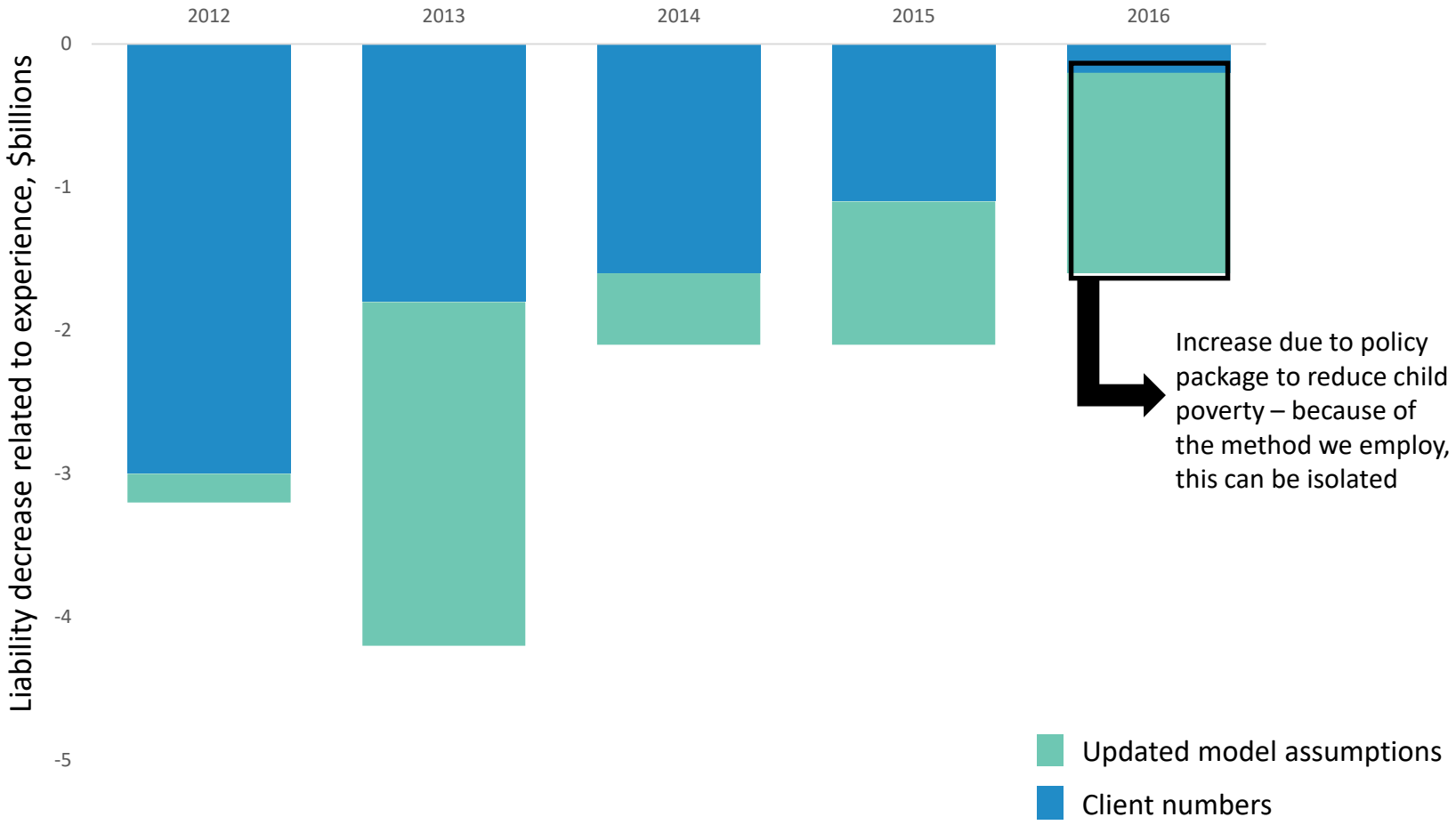
Aggregate liability valuation

Total liability

72.2 billion

$13.7 billion

Liability reduction
attributable to
system reform



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

Google Sheets
docs.google.com

HDSR

Search Login or Signup

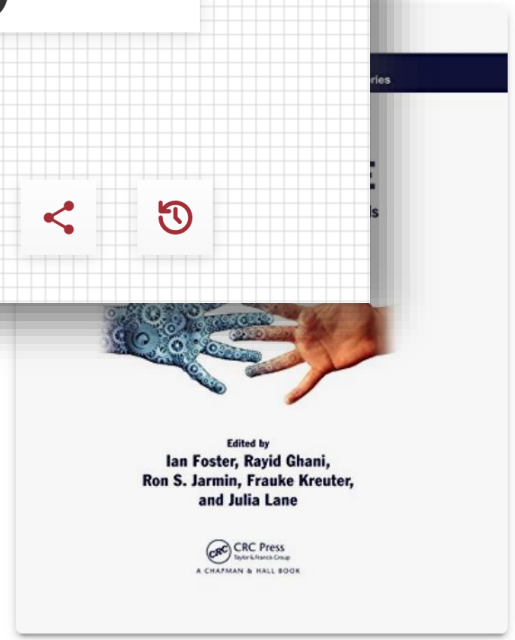
HOME ABOUT MISSION AND SCOPE ISSUE 1.1

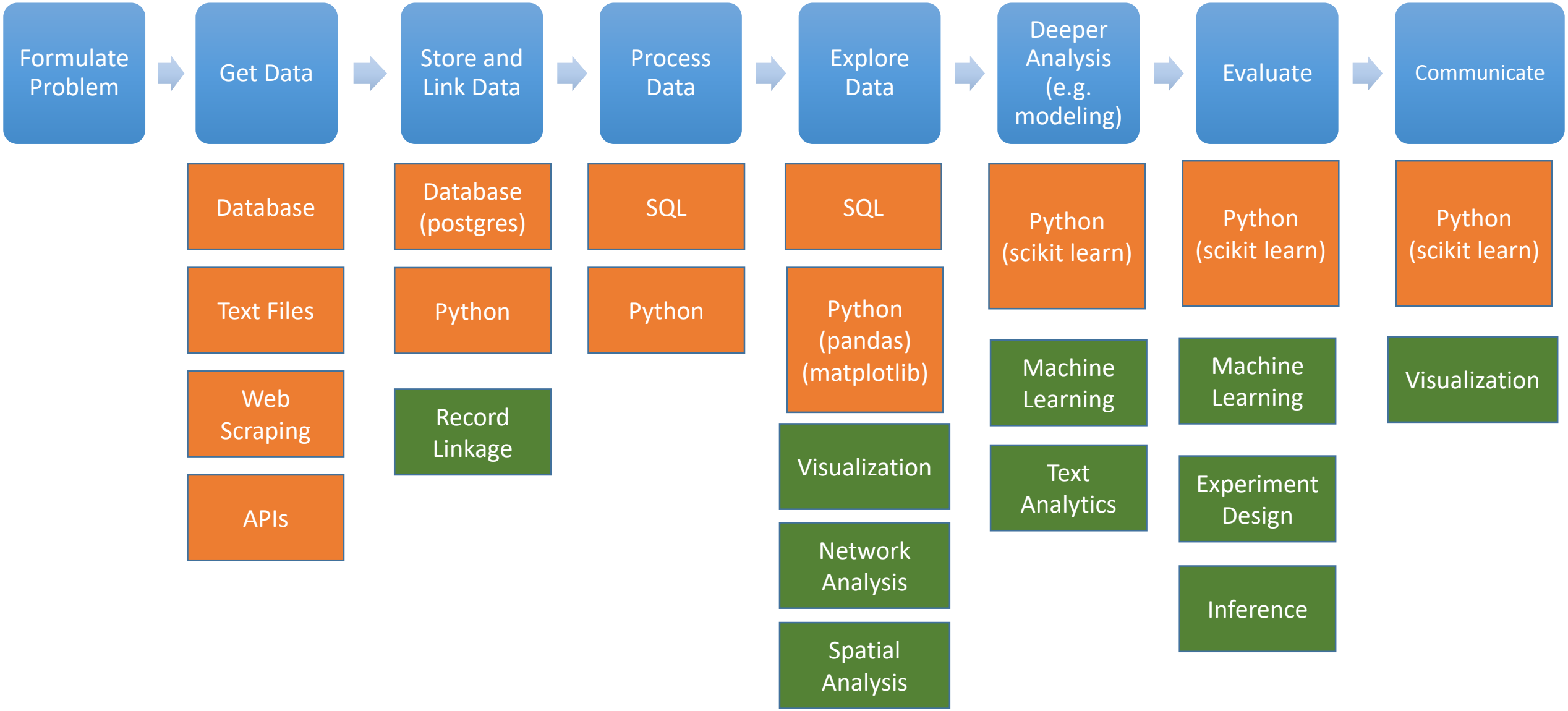
1.2 Stepping Stones

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

by Frauke Kreuter, Rayid Ghani, and Julia Lane

now on branch
#public




Formulate Problem
Get Data
Store and Link Data
Process Data
Explore Data
Deeper Analysis (e.g. modeling)
Evaluate
Communicate
Database
Text Files
Web Scraping
APIs
Database (postgres)
Python
Record Linkage
SQL
Python
SQL
Python (pandas) (matplotlib)
Visualization
Network Analysis
Spatial Analysis
Python (scikit learn)
Machine Learning
Text Analytics
Python (scikit learn)
Machine Learning
Experiment Design
Inference
Python (scikit learn)
Visualization
Collaboration: “shared” folders in your project
Privacy, Confidentiality, Security

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

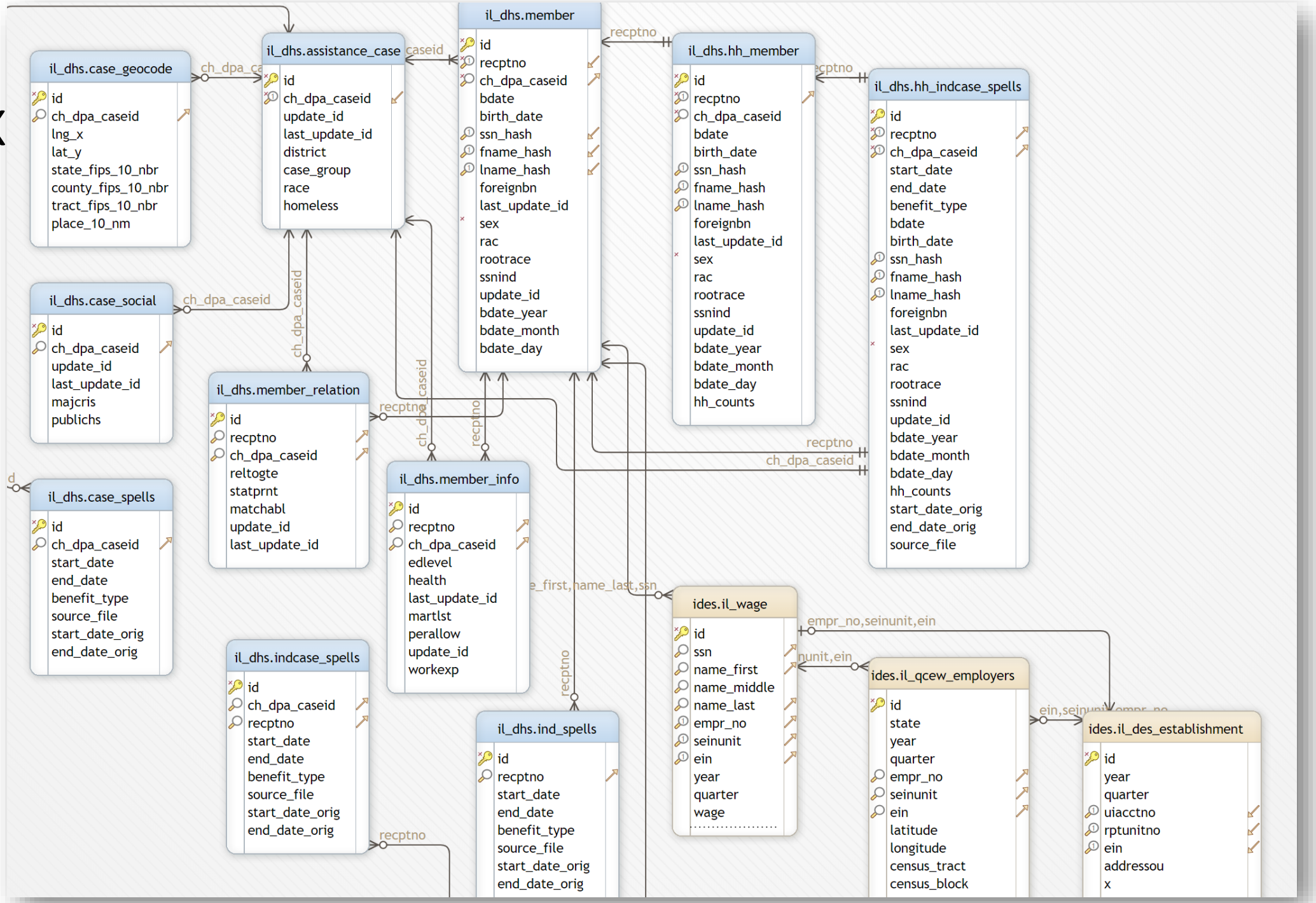
https://ada.coleridgeinitiative.org/ada-2019-missouri

Dataset overview (example contd)

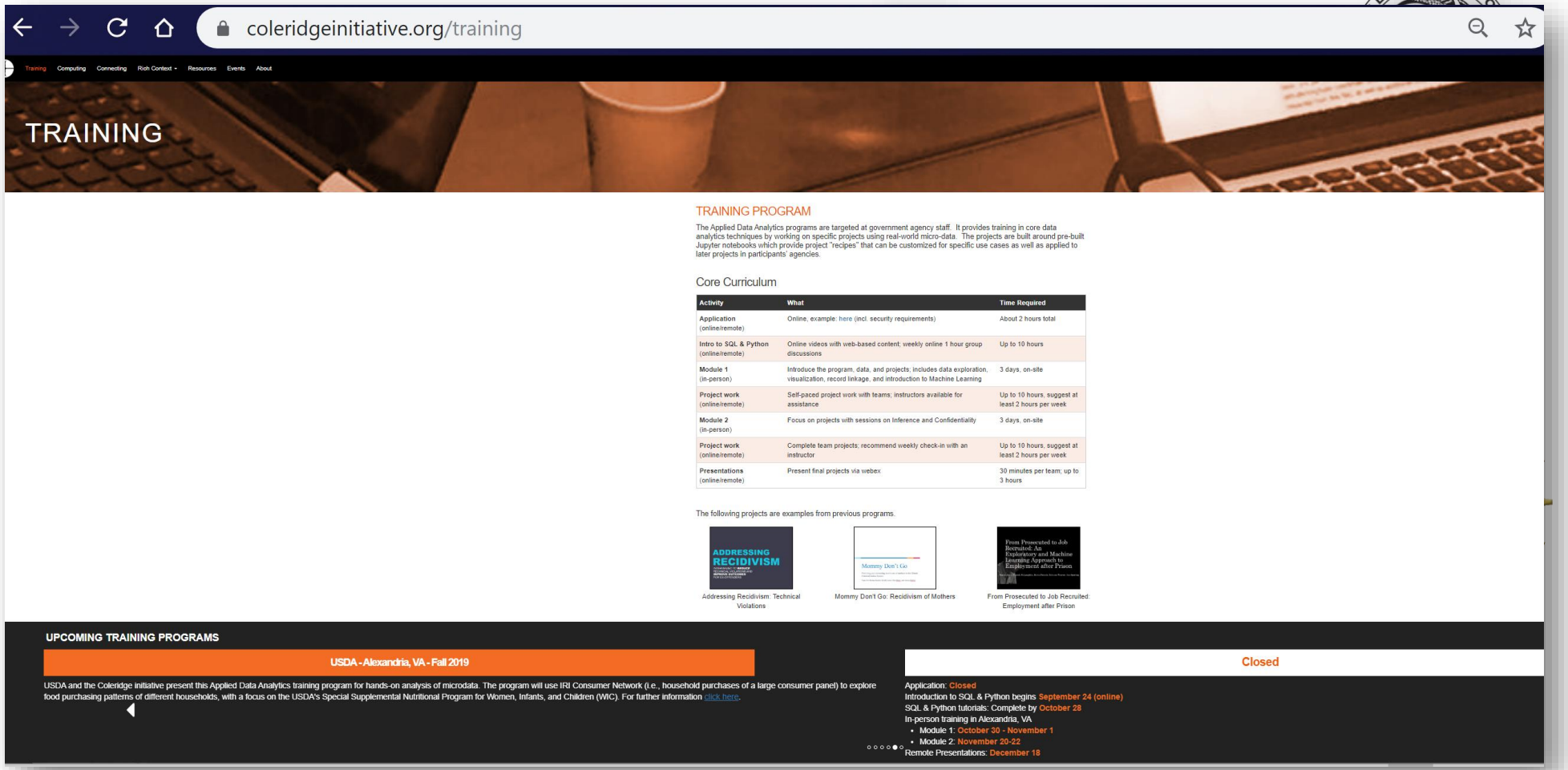
- Indiana datasets
  - Wage records and Education detailed data documentation, data overview
- Ohio Longitudinal Data Archive datasets
  - Dataset overview, data details
- Illinois Department of Employment Services
  - Documentation link

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

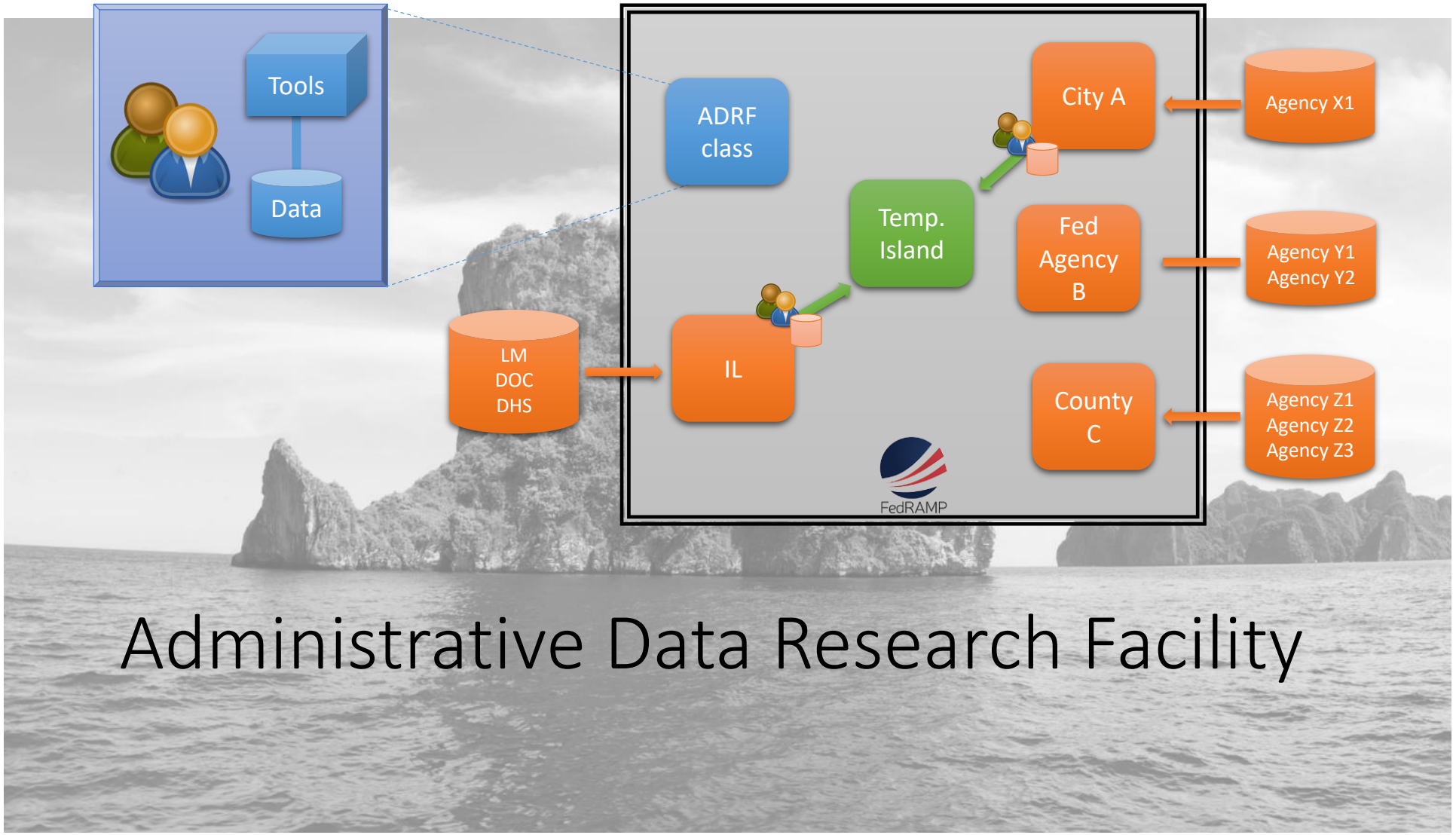
Ex


il_dhs.case_geocode
id
ch_dpa_caseid
lng_x
lat_y
state_fips_10_nbr
county_fips_10_nbr
tract_fips_10_nbr
place_10_nm
ch_dpa_ca
il_dhs.assistance_case
id
ch_dpa_caseid
update_id
last_update_id
district
case_group
race
homeless
caseid
il_dhs.member
id
recptno
ch_dpa_caseid
bdate
birth_date
ssn_hash
fname_hash
lname_hash
foreignbn
last_update_id
sex
rac
rootrace
ssnind
update_id
bdate_year
bdate_month
bdate_day
recptno
il_dhs.hh_member
id
recptno
ch_dpa_caseid
bdate
birth_date
ssn_hash
fname_hash
lname_hash
foreignbn
last_update_id
sex
rac
rootrace
ssnind
update_id
bdate_year
bdate_month
bdate_day
hh_counts
ecptno
il_dhs.hh_indcase_spells
id
recptno
ch_dpa_caseid
start_date
end_date
benefit_type
bdate
birth_date
ssn_hash
fname_hash
lname_hash
foreignbn
last_update_id
sex
rac
rootrace
ssnind
update_id
bdate_year
bdate_month
bdate_day
hh_counts
start_date_orig
end_date_orig
source_file
il_dhs.case_social
id
ch_dpa_caseid
update_id
last_update_id
majcris
publichs
ch_dpa_caseid
ch_dpa_caseid
il_dhs.member_relation
id
recptno
ch_dpa_caseid
reltogte
statprnt
matchabl
update_id
last_update_id
recptno
ch_dpa_caseid
recptno
recptno
ch_dpa_caseid
d
il_dhs.case_spells
id
ch_dpa_caseid
start_date
end_date
benefit_type
source_file
start_date_orig
end_date_orig
il_dhs.member_info
id
recptno
ch_dpa_caseid
edlevel
health
last_update_id
martlst
perallow
update_id
workexp
e_first,name_last,ssn
ides.il_wage
id
ssn
name_first
name_middle
name_last
empr_no
seinunit
ein
year
quarter
wage
empr_no,seinunit,ein
nunit,ein
ides.il_qcew_employers
id
state
year
quarter
empr_no
seinunit
ein
latitude
longitude
census_tract
census_block
ein,seinunit,empr_no
ides.il_des_establishment
id
year
quarter
uiacctno
rptunitno
ein
addressou
x
il_dhs.indcase_spells
id
ch_dpa_caseid
recptno
start_date
end_date
benefit_type
source_file
start_date_orig
end_date_orig
recptno
recptno
il_dhs.ind_spells
id
recptno
start_date
end_date
benefit_type
source_file
start_date_orig
end_date_orig

Examples of Demand

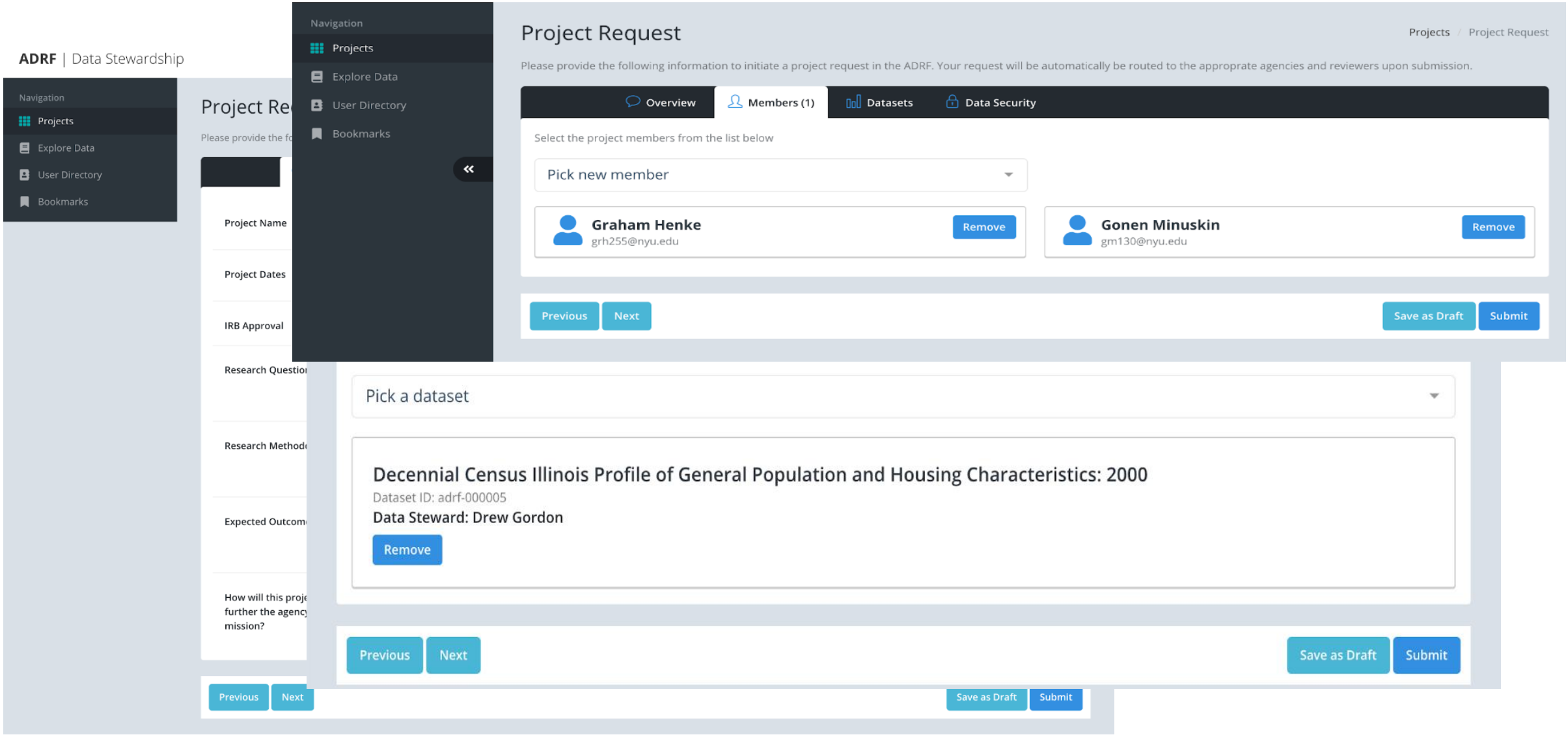


Technical Environment

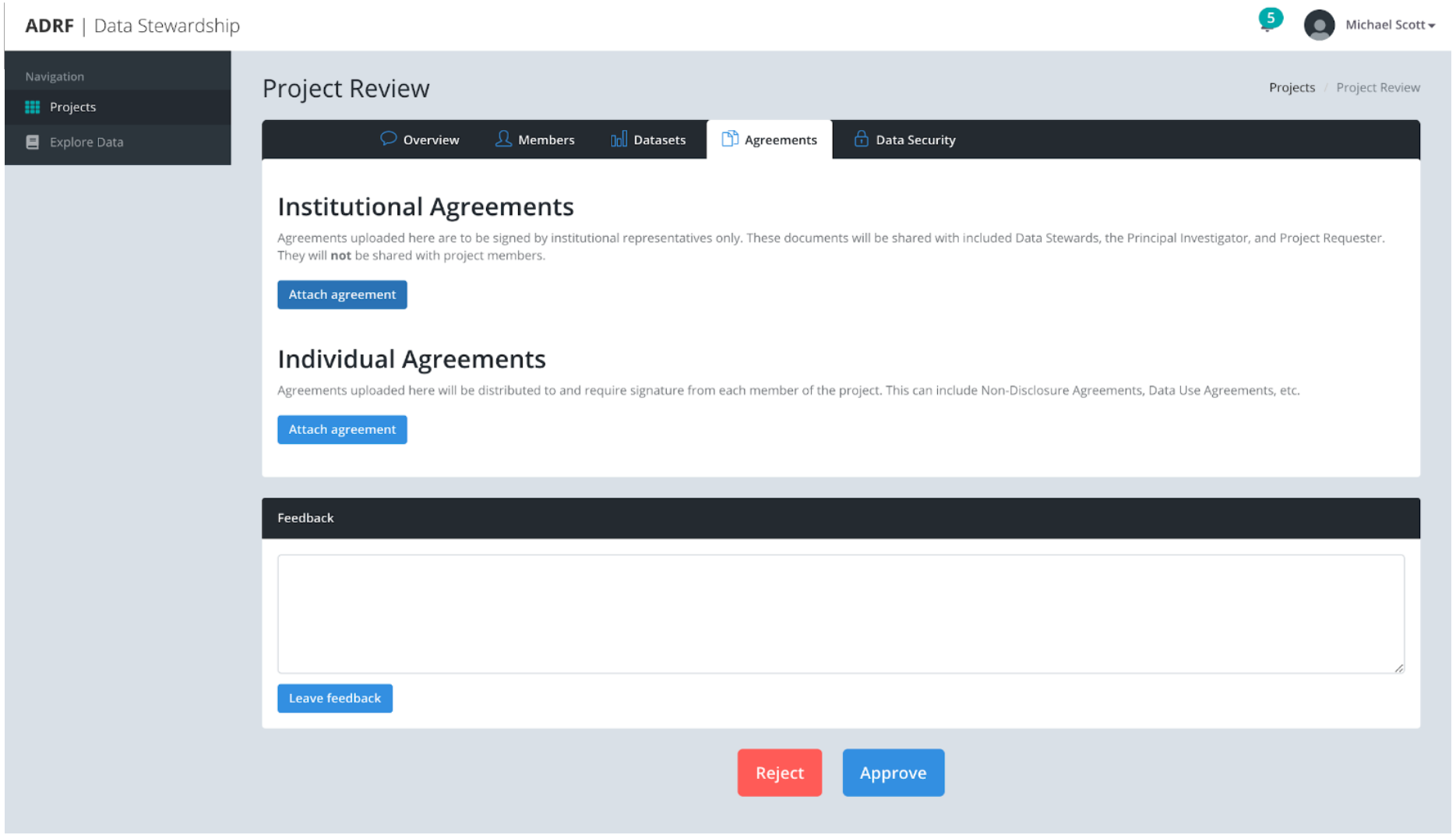


Administrative Data Research Facility

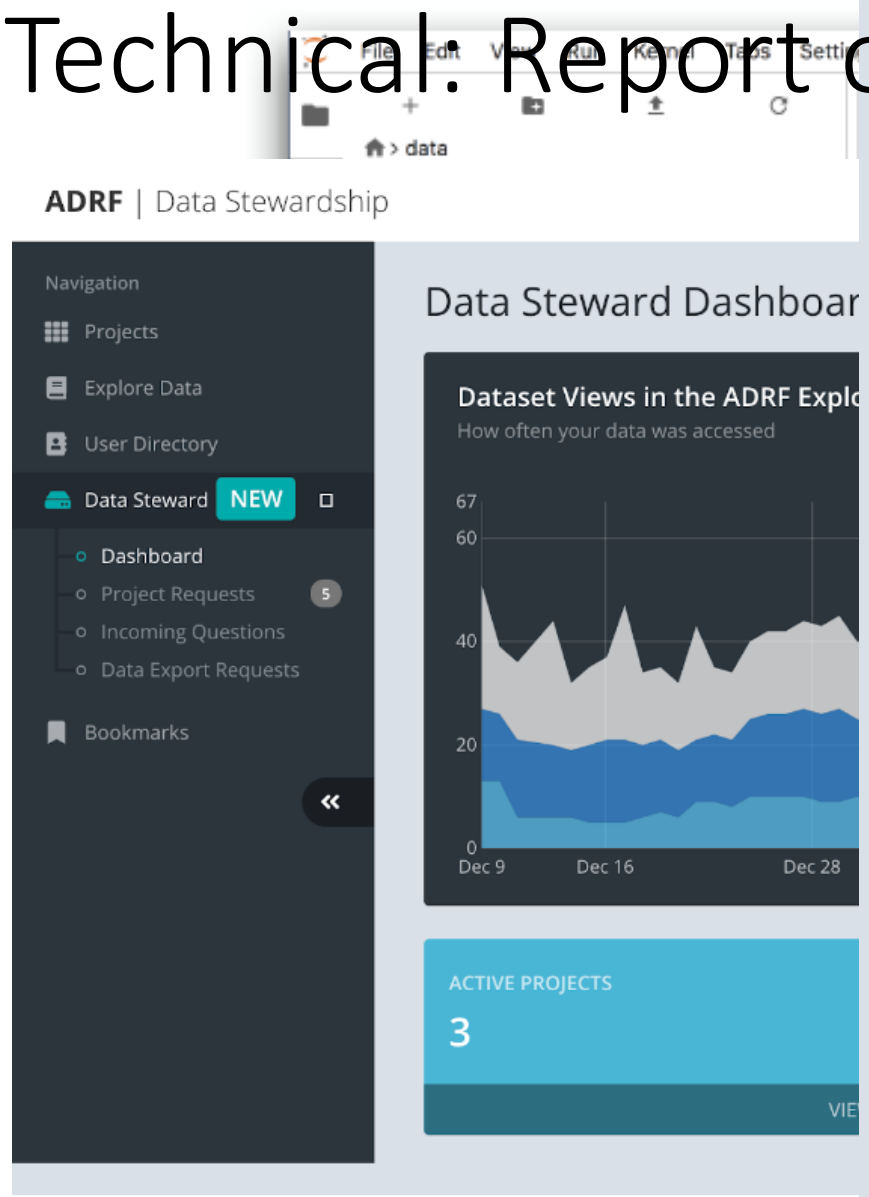
Technical: Automate Data Approval Workflow



Technical: Streamline legal and operational



Technical: Report on Access and Use



Dataset Popularity

Dataset Name	Researchers who have access	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

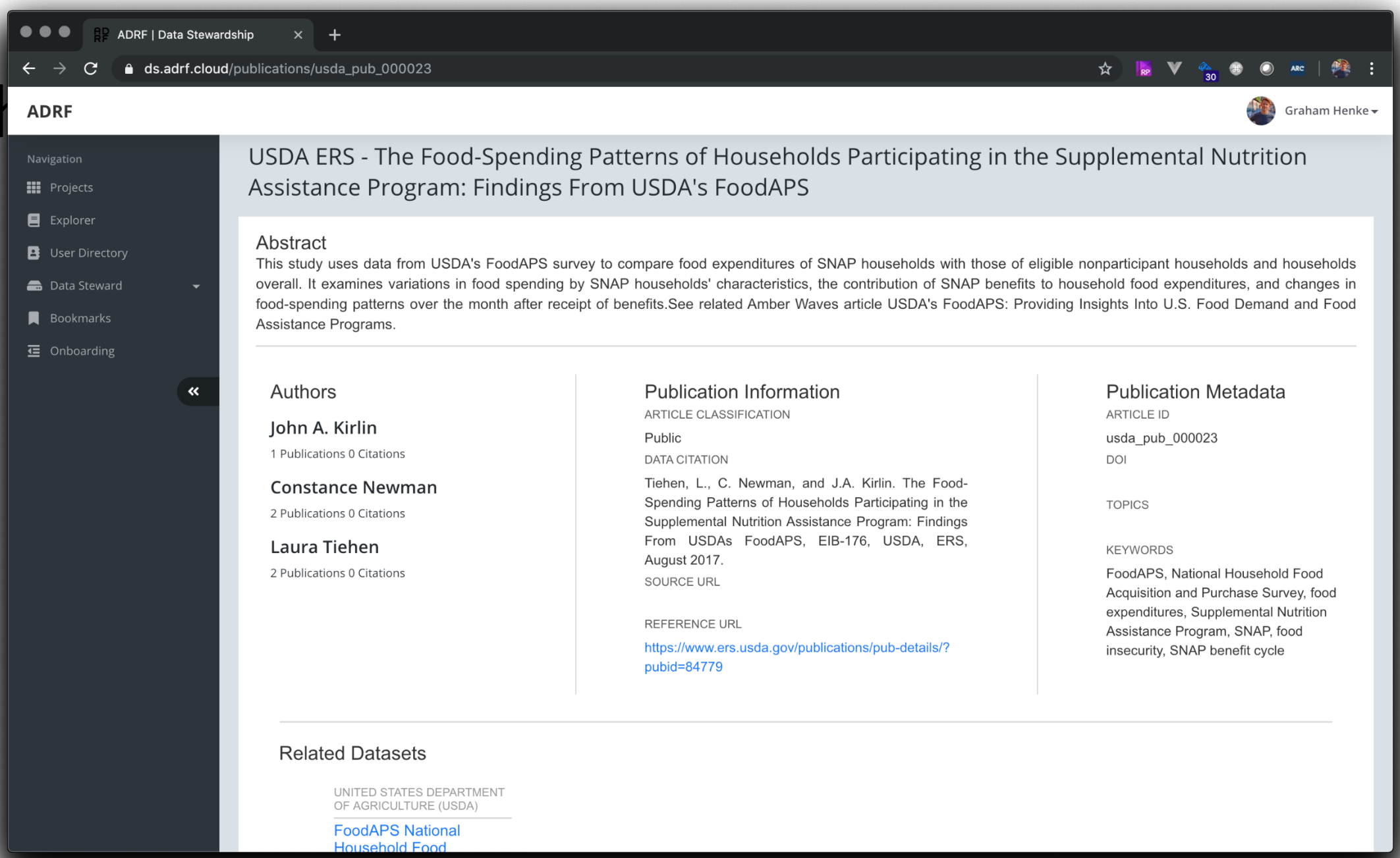
Top Users of Your Datasets

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


ADRF | Data Stewardship
ds.adrf.cloud/publications/usda_pub_000023
ADRF
Graham Henke
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
John A. Kirlin
1 Publications 0 Citations
Constance Newman
2 Publications 0 Citations
Laura Tiehen
2 Publications 0 Citations
Publication Information
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
REFERENCE URL
https://www.ers.usda.gov/publications/pub-details/?pubid=84779
Publication Metadata
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

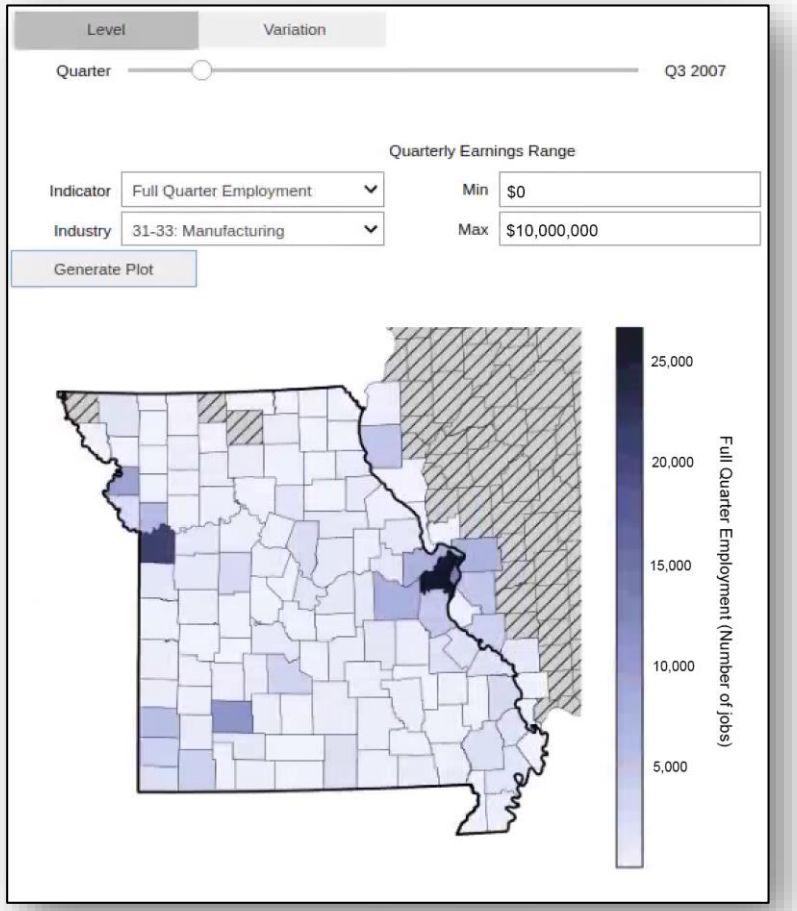
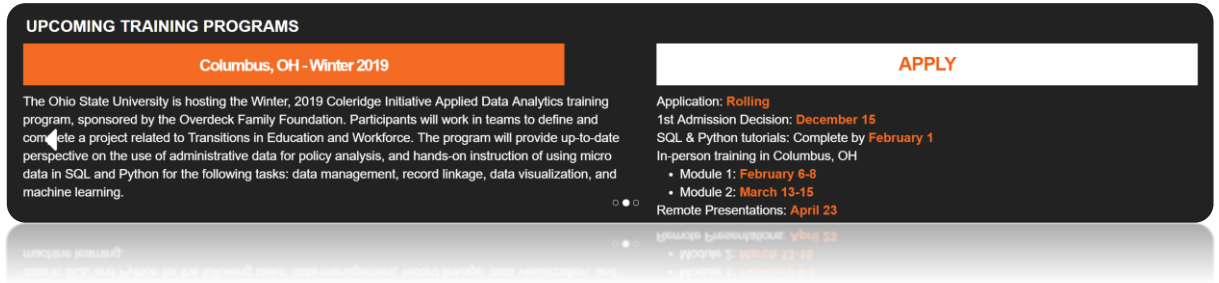
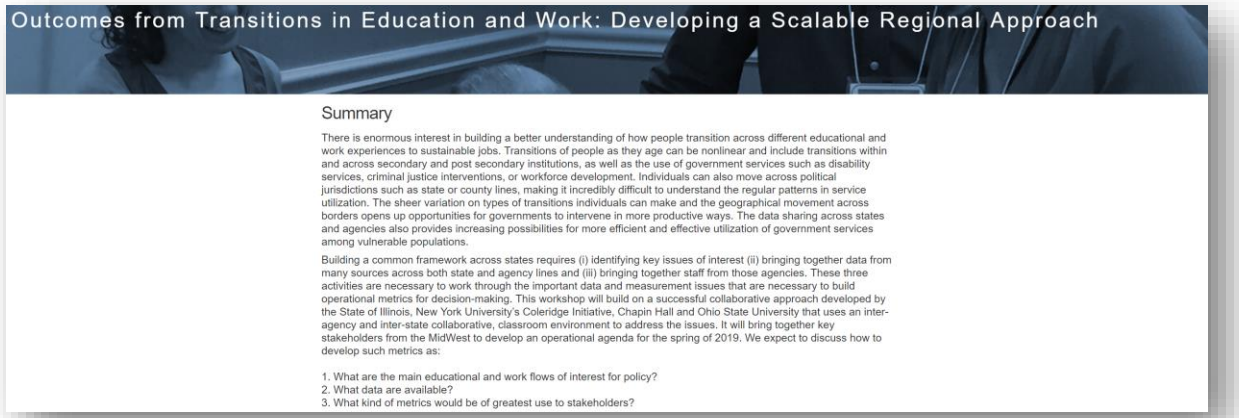


Fig. 3: Comparing total earnings with Illinois border counties

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

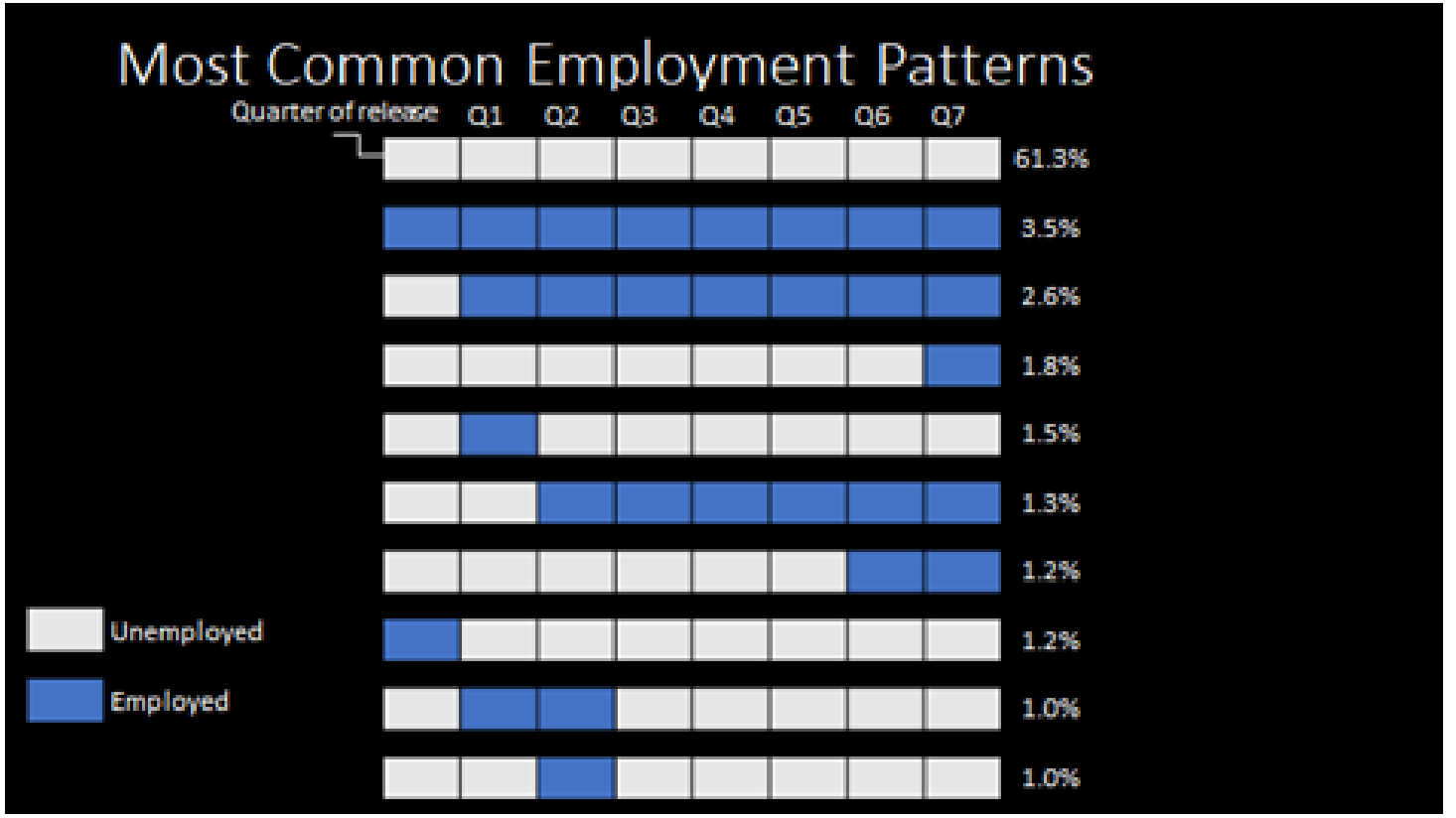




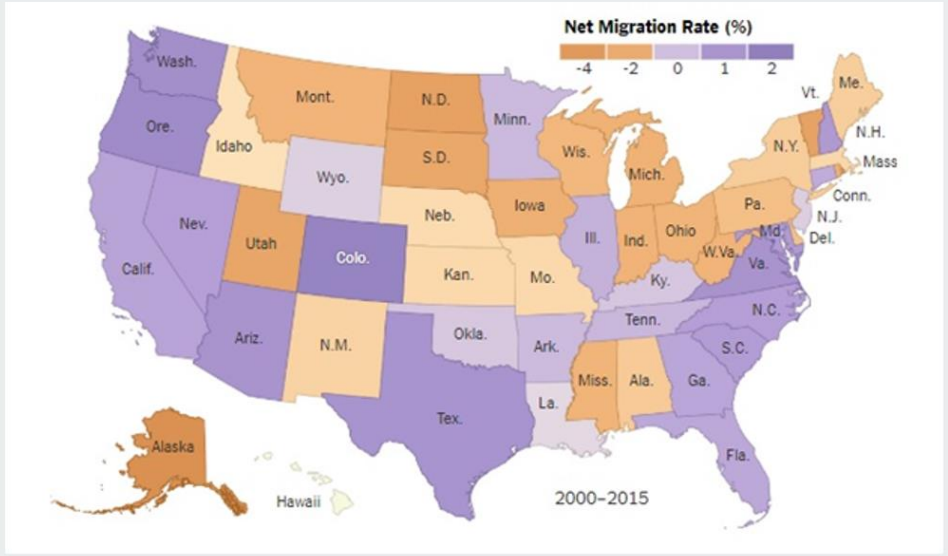
Examples of results: across agency lines

ADA Class w/ Illinois data

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



Examp e lines



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

Group #3

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

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