



---

# CONVERGENCE: A PRACTICAL APPROACH

Julia Lane

Coleridge Initiative and NYU

---

---

# NARRATIVE

- Context: complex problems focused on societal needs
  - Thesis: Many major problems are empirical
  - Challenge: How to find data?
  - Approach: New technologies available
  - Practical Example: Federally funded datasets
  - Next steps
-

# CONTEXT: MASSIVE CURRENT CHALLENGES

Q Popular Latest

The Atlantic

## POLITICS

### Would You Sacrifice Your Privacy to Get Out of Quarantine?

The coronavirus has reignited the post-9/11 debate about security and civil liberties. The U.S. response to the tragedy has lessons for how to manage the trade-offs this time around.

MIKE GIOLIO APRIL 22, 2020



THE ATLANTIC

AS GENERAL COUNSEL of the National Security Agency in the 1990s, John R. Baker advocated for limiting the government's intelligence-gathering at the name of civil liberties. Then the 9/11 attacks happened, a



SCIF Security - FSO Trusted |  
UL2050 Certified

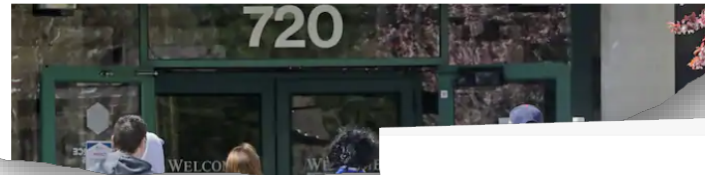
asi247.com

Visit Site >

## Business

### What the Labor Department is doing about the error that led to a lower unemployment rate

Trump appointee William W. Beach said the pattern of errors that caused the official unemployment rate to be calculated below its actual value was accidental



## Two Huge Questions Loom as 2020 Census Winds Down

The most contentious census count in memory is nearing an end with questions remaining about the accuracy of its numbers and how they will be used in congressional reapportionment.



Yet even as the Census Bureau boasts of nearing a 100 percent completion rate, the stated rates in scores of census offices have been rolled back in recent weeks, said Steven Romalewski, the director of the mapping service at the City University of New York's Graduate Center. That suggests that the bureau is still in places where its count is problematic and sending door-to-door workers back to do more work, he said.

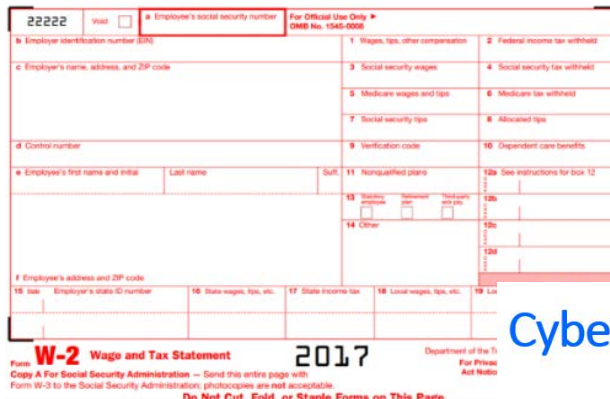
It makes outside experts even more skeptical that 99.9 percent is anything more than a public relations gimmick to bolster the case for ending the count early, coronavirus or not.

Happy talk in the context of the pandemic is completely out of place," said Margo J. Anderson, a census expert and historian at the University of Wisconsin-Milwaukee. "Aside from whatever explanation the Trump administration is up to, a more honest statement right now would be saying, 'You know, we're having a real

# MASSIVE AMOUNTS OF DATA AVAILABLE TO STIMULATE NEW RESEARCH

Smart Cities

Your text here



Form W-2 Wage and Tax Statement 2017. The form is divided into sections for employer information, employee information, and wages/taxes. It includes fields for employer identification number, employee's social security number, employer's name and address, control number, employee's first and last name, and various wage and tax amounts. The form is labeled 'W-2 Wage and Tax Statement 2017' and 'Department of the Treasury, Internal Revenue Service'.

Cyber Sabotage – Russian ads about Hillary Clinton



## Three Types of Data:

- Digitized administrative datasets on people and places (records of services for people, taxes paid, land transactions, police encounters)
- Sensors, wireless networks, video cameras, etc. can monitor people and things throughout a city and the "Internet of Things" makes it possible to control things.
- Internet data such as Google Street View, Zillow (real-estate), Yelp (reviews of retailers)

## Issues:

- How do we find out if an ad is fake?
- Who can "police" the Internet?
- How does this interact with free speech?

Thanks to Henry Brady



# HOW TO FIND?

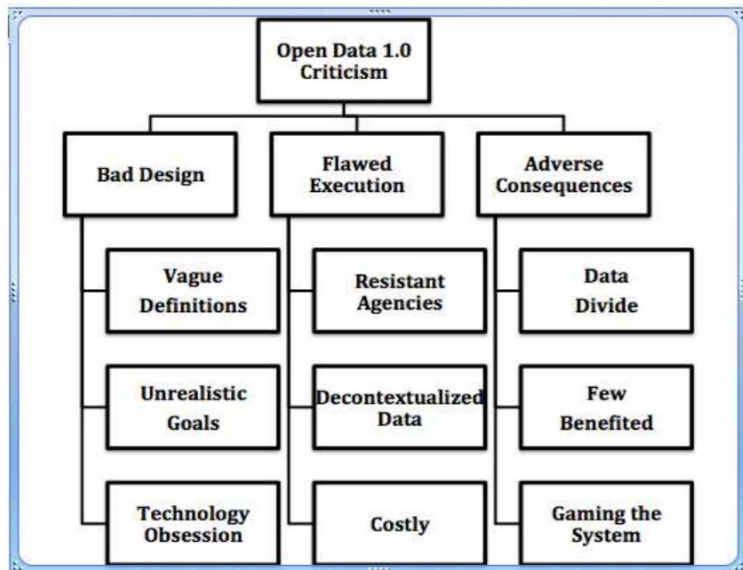
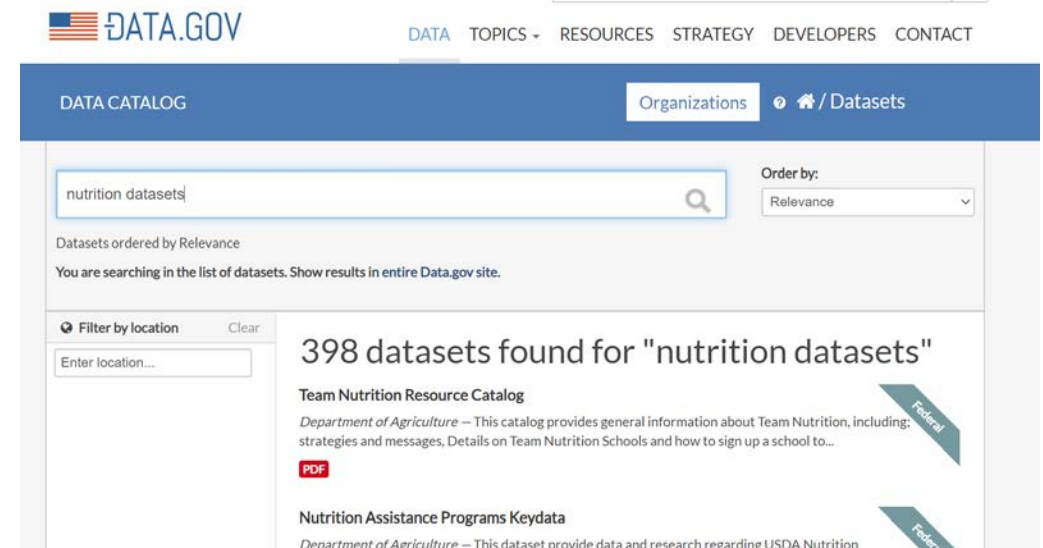


Figure 1: Open Data 1.0 Criticism





# NEW COLLABORATIVE TECHNOLOGIES – AND LEGAL FRAMEWORK

  
**Foundations for Evidence-Based Policymaking Act of 2018**

The bipartisan Foundations for Evidence-Based Policymaking Act of 2018 builds off the work of the U.S. Commission on Evidence-Based Policymaking to strengthen data privacy protections, improve secure access to data, and enhance the federal government's capacity for producing and using evidence.

**Strengthens Privacy Protections**

**Maintains Strong Confidentiality Protections for Sensitive Data.** Reauthorizes the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), an existing law that gives the American public strong privacy safeguards and legal protections for appropriate uses of confidential data.

**Institutes Processes to Assess Data Risks.** Strengthens efforts to protect confidentiality while making data accessible for evidence building and transparent to the public by requiring comprehensive risk assessments for certain publicly released data.

**Enhances Public Trust in Data.** Improves public trust in statistical activities by codifying language directing certain agencies to establish procedures to protect trust in data activities by appropriately maintaining objectivity, independence, and confidentiality.

**Establishes Consistent Leadership on Key Data Issues.** Ensures a senior leader in each agency is responsible for protecting privacy and ensuring confidentiality protections are appropriately applied by creating chief data officers.

**Improves Secure Data Access**

**Encourages Agencies to Make Data Public and Open When Possible.** Takes steps to improve the public information about what data government currently holds and make data publicly available when possible and in the public interest.

**Requires Development of Data Inventories.** Enables researchers and evaluators to better identify what government-collected data are available by directing agencies to create and maintain data inventories and publicly provide details about those datasets.

**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.

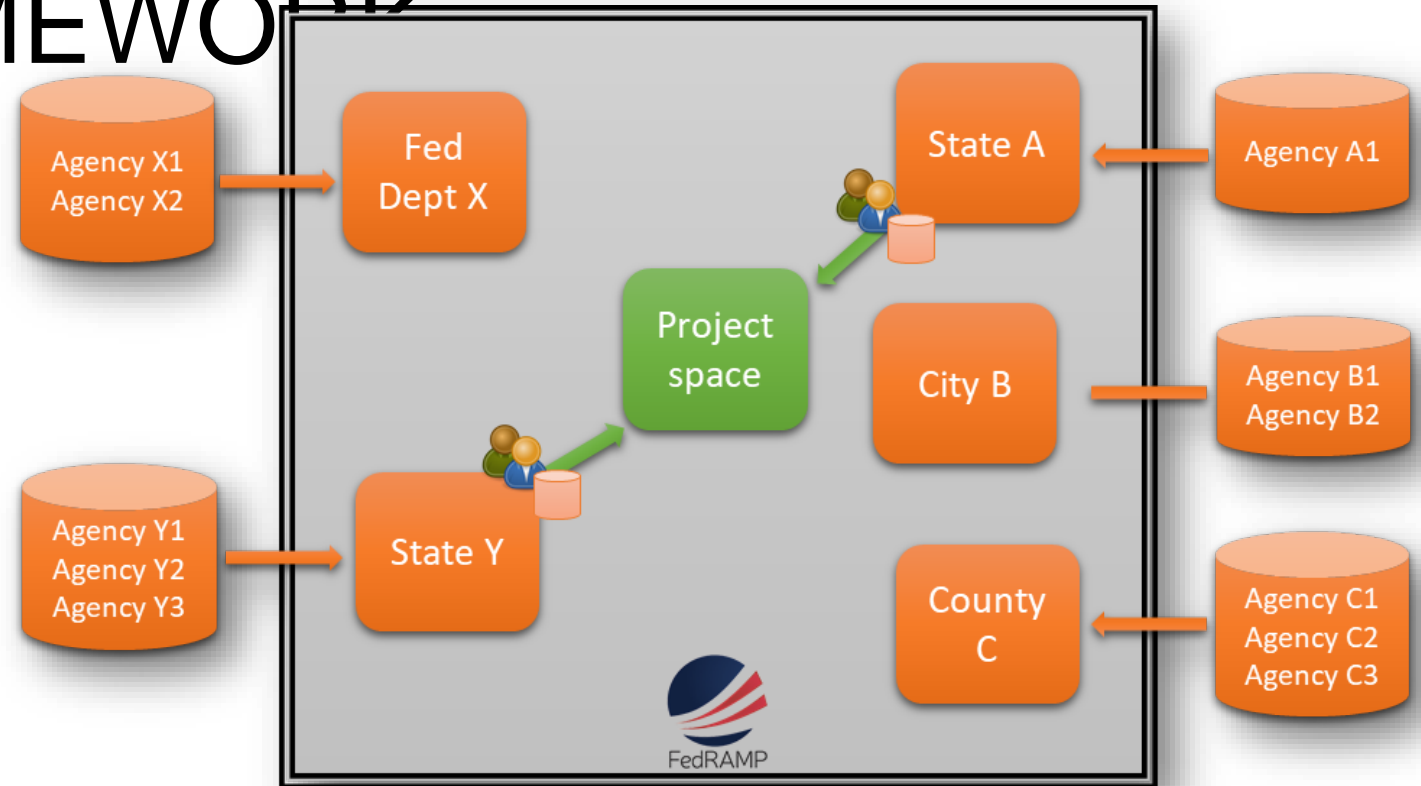
**Enhances Government's Evidence Capacity**

**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.

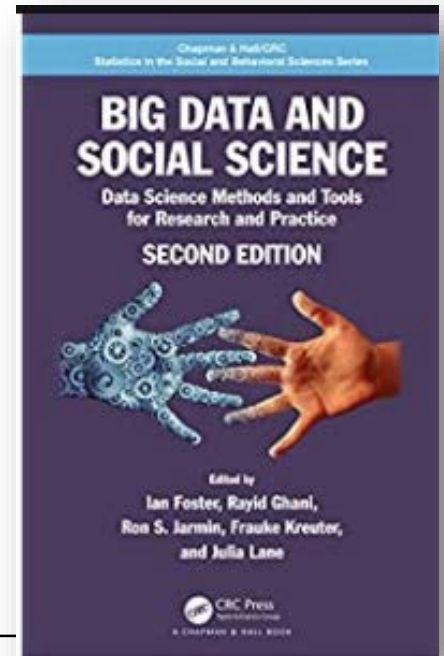
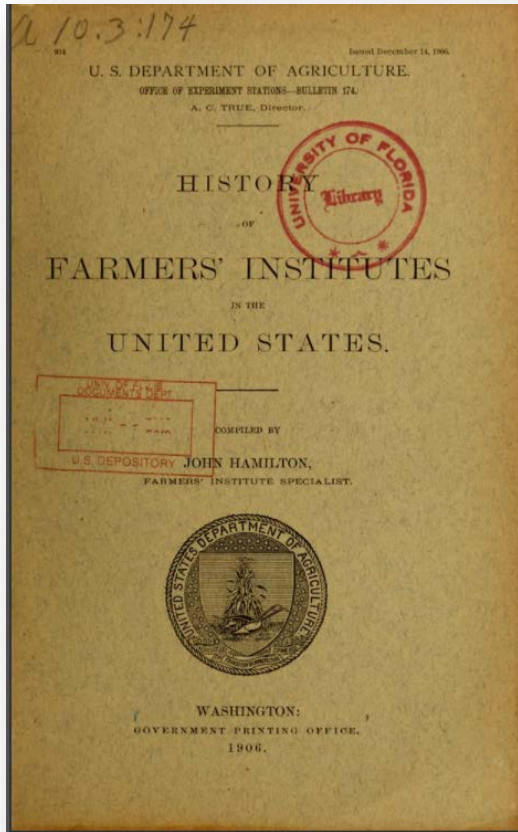
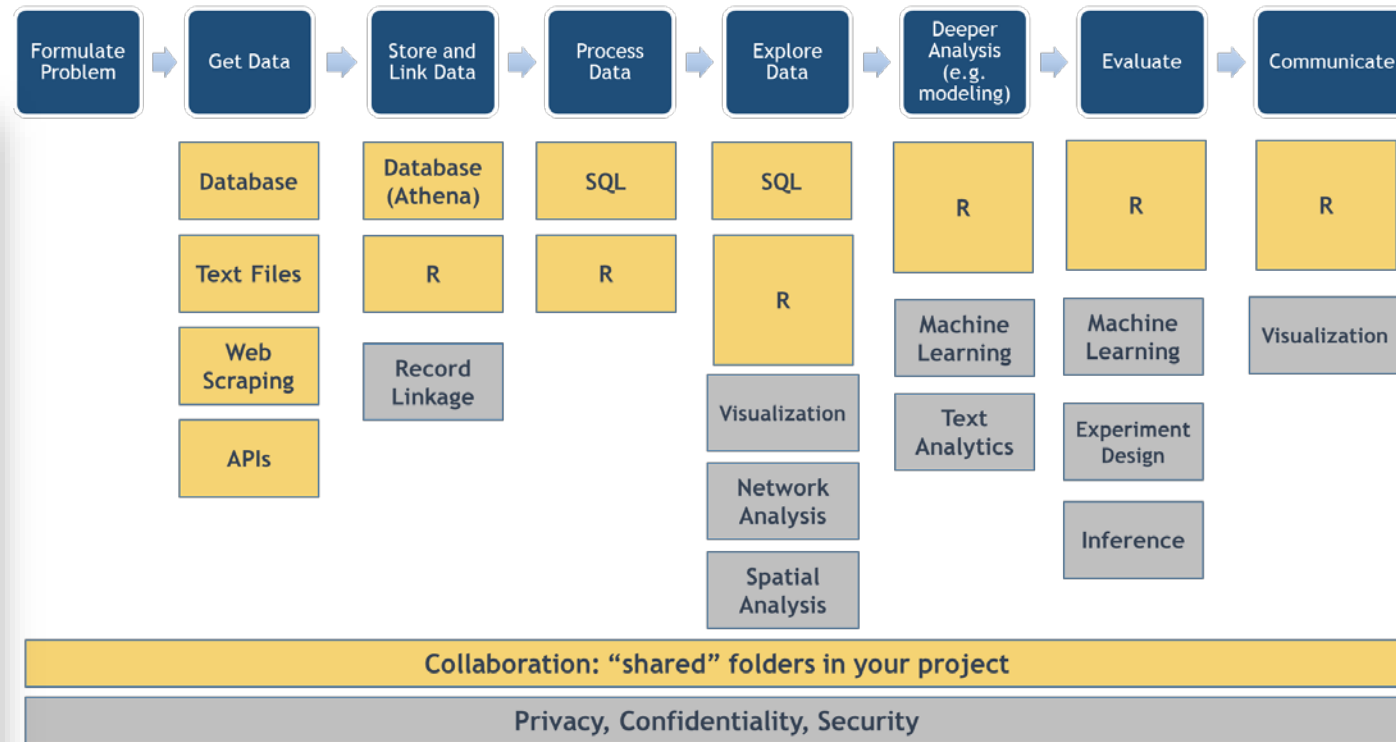
**Prioritizes Evaluation Activities in Agencies.** Improves agency capacity to engage in and use program evaluation by establishing evaluation officers in government agencies and requiring agencies to develop written evaluation policies.

**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.

Learn more at [bipartisanpolicy.org/evidence](https://bipartisanpolicy.org/evidence)



# NEW CAPACITY TO MAKE USE OF DATA



# PRACTICAL EXAMPLE



EXECUTIVE OFFICE OF THE PRESIDENT  
OFFICE OF MANAGEMENT AND BUDGET  
WASHINGTON, D.C. 20503

April 24, 2019

M-19-15

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Russell T. Vought, Acting Director

SUBJECT: Improving Implementation of the Information Quality Act

## Introduction

The purpose of this Memorandum is to reinforce, clarify, and interpret agency responsibilities with regard to responsibilities under the Information Quality Act (IQIA).<sup>1</sup> In 2002, the Office of Management and Budget issued *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility and Integrity of Information Disseminated by Federal Agencies*<sup>2</sup> (“*Guidelines*”). The principles and core responsibilities described in the *Guidelines* remain sound and relevant for agency practice; however, additional guidance is required to address changes in the information landscape and to incorporate best practices developed over time.<sup>3</sup> This Memorandum updates implementation of the *Guidelines* to reflect recent innovations in information generation, access, management, and use, and to help agencies address common problems with maintaining information quality.

## Foundations for Evidence-Based Policymaking Act of 2018

The bipartisan Foundations for Evidence-Based Policymaking Act of 2018 builds off the work of the U.S. Commission on Evidence-Based Policymaking to strengthen data privacy protections, improve secure access to data, and enhance the federal government's capacity for producing and using evidence.

### Strengthens Privacy Protections

**Maintains Strong Confidentiality Protections for Sensitive Data.** Reinforces the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), an existing law that gives the American public strong privacy safeguards and legal protections for appropriate uses of confidential data.

**Institutes Processes to Assess Data Risks.** Strengthens efforts to protect confidentiality while making data accessible for evidence building and transparent to the public by requiring comprehensive risk assessments for certain publicly-released data.

**Enhances Public Trust in Data.** Improves public trust in statistical activities by codifying language directing certain agencies to establish procedures to protect trust in data activities by appropriately maintaining objectivity, independence, and confidentiality.

**Establishes Consistent Leadership on Key Data Issues.** Ensures a senior leader in each agency is responsible for protecting privacy and ensuring confidentiality protections are appropriately applied by creating chief data officers.

### Improves Secure Data Access

**Encourages Agencies to Make Data Public and Open When Possible.** Takes steps to improve the public information about what data government currently holds and make data publicly available when possible and in the public interest.

**Requires Development of Data Inventories.** Enables researchers and evaluators to better identify what government-collected data are available by directing agencies to create and maintain data inventories and publicly provide details about those datasets.

**Makes Administrative Records Available for Evidence Building.** Under a strong set of confidentiality protections, messages 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 ongoing 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 encouraging 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.

Learn more at [bipartisanpolicy.org/evidence](https://bipartisanpolicy.org/evidence)

## Agency Actions

	Foundations for Evidence-Based Policymaking Act of 2018 and Associated OMB Guidance	Executive Order on Maintaining American Leadership in Artificial Intelligence	Improving Implementation of the Information Quality Act (M-19-15)
1. Identify Data Needs to Answer Priority Agency Questions	✓		✓
2. Constitute a Diverse Data Governance Body	✓		✓
3. Assess Data and Related Infrastructure Maturity			✓
4. Identify Opportunities to Increase Staff Data Skills	✓		
5. Identify Priority Data Assets for Agency Open Data Plans	✓	✓	✓
6. Publish and Update Data Inventories	✓		



---

# PRACTICAL EXAMPLE

Could we adapt these ideas to apply to the “conservation value” of federal data? Could we identify high-priority data, whose stewardship and preservation would make a large impact for our nation? Tyler Christenson, NOAA

Ecosystems and habitats: rare ecosystems or habitats for rare species

Ecosystem services: areas that provide an important service in critical situations, e.g. protecting water catchments or erosion control in vulnerable areas

Community needs: sites that are fundamental for satisfying some human need for the surrounding community, e.g. food, livelihoods, water, etc.

Cultural values: sites that have archaeological or historical significance, or have critical cultural / sacred importance to the community

---

<https://fsc.org/en/for-forests/high-conservation-values#the-hcv-approach>

---

# CONCRETE EXAMPLE OF METRICS

**Diversity:** simply count how often the data are used, with special consideration of datasets that are the sole source of information in nearly every study within a research discipline

**Landscape-level ecosystems:** data that are often used in combination with federal datasets *from other agencies*

**Ecosystems and habitats:** clusters of data that are often used together in a research discipline, so that if one is lost then the others would lose value

**Ecosystem services:** data used in research topics aimed at protecting life and property in critical situations, e.g. floods, pandemics, war

**Community needs:** data used in research topics aimed at supporting basic community needs: e.g. health, food, housing, livelihoods

**Cultural values:** data used in research topics aimed at protecting historical, ecological, sacred, or intangible values

---

---

# SEARCH AND DISCOVERY



# DEPLOY NEW TECHNOLOGY

---

Quantitative Research

## WIC Household Purchases: Who's Paying Out of Pocket?

Hayden Stewart, PhD<sup>1</sup>,  
and Elizabeth Frazão, PhD<sup>1</sup>

### Methods

#### Design

The National Household Food Acquisition and Purchase Survey (FoodAPS) is a cross-sectional, nationally representative survey of US households' food purchases and acquisitions, including foods acquired through USDA food assistance programs.<sup>21</sup> Mathematica Policy Research (Mathematica) administered FoodAPS under contract with USDA. The survey took place between April 2012 and January 2013 with an overall response rate of 45.6%. Sample weights were adjusted to reduce the potential for nonresponse bias. Data quality and accuracy were independently assessed by Westat, also under contract with USDA.

FoodAPS was conducted under the Confidential Information Protection and Statistical Efficiency Act of 2002, which requires federal agencies that use the data to guard the confidentiality of survey respondents. The data have been previously used to investigate the cost of cold cereals pur-

American Journal of Health Promotion  
2019, Vol. 33(1) 79-86

© The Author(s) 2018

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/0890117118778243

[journals.sagepub.com/home/ahp](https://journals.sagepub.com/home/ahp)



, PhD<sup>1</sup>,



---

# THE TECHNOLOGY EXISTS

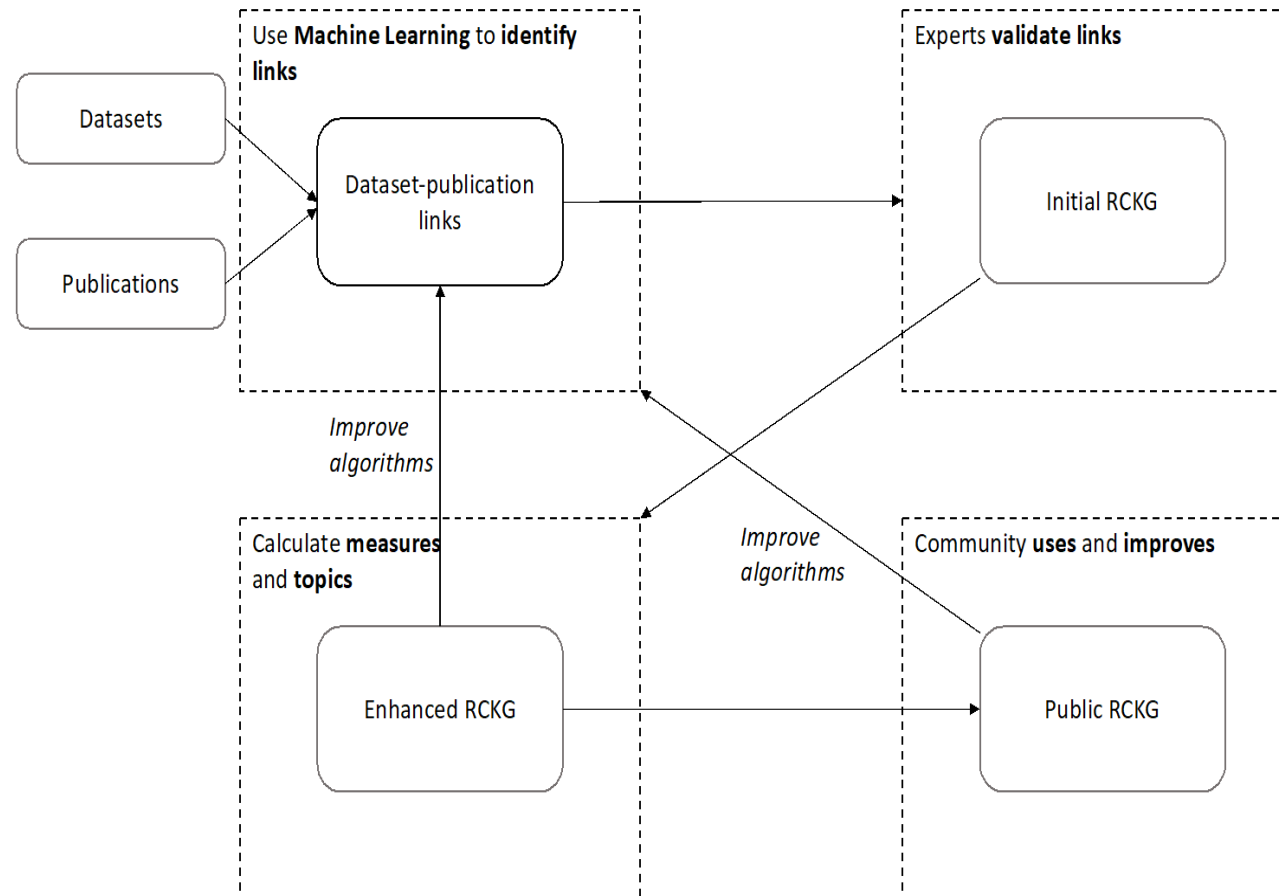
- Schmidt Futures, Sloan, Overdeck funded initiative in 2018-2019
- Developed a corpus
- Built AI model through two open source competitions
- Results: Correctly found dataset 78% of the time

Current SOTA

source	precision	entry	code	paper	corpus	submitted	notes
<a href="#">LARC</a> <a href="#">@philipskokoh</a>	0.7836	<a href="#">ipynb</a>	<a href="#">repo</a>	<a href="#">RCC_1</a>	<a href="#">v0.1.5</a>	2019-09-26	RCLC baseline experiment using RCC_1 approach
<a href="#">KAIST</a> <a href="#">@HaritzPuerto</a>	0.6319	<a href="#">ipynb</a>	<a href="#">repo</a>	<a href="#">RCC_1</a>	<a href="#">v0.1.5</a>	2019-11-01	model trained a different dataset using <a href="#">DocumentQA</a> and <a href="#">Ultra-Fine Entity Typing</a> -- NB: this approach is able to identify new datasets

---

# IMPROVE MODEL THROUGH AI AND COMMUNITY



# AGRICULTURAL RESOURCE MANAGEMENT SURVEY

**Used in 320  
Publications  
(TOPICS)**

Farm Sector Income & Finances	40
Farm Economy	24
Agricultural Research and Productivity	21
Organic Agriculture	6
Crop & Livestock Practices	4
Poverty & Income Volatility	3

44 Mishra Ashrok

15 Nigel Key

15 Jeffrey Gillespie

15 Hisham Said El-Osta

14 Richard Nehring

**Used by 545  
Authors**

90

**Used by 118  
Institutions**

Economic Research Service	15
Michigan State	7
University of Minnesota	7
University of Nebraska-Lincoln	7
University of Missouri	6

8 Census of Agriculture

5 Soil Survey Geographic Database USDA

15 Cropland Data Layer, USDA

5 Survey of Consumer Finances, Federal Reserve Board

4 Fertilizer Use and Price, USDA

**Combined  
with 72  
Other  
Datasets**

AI techniques applied to hand curated corpus

# EXPAND CORPUS

## Publication and Compliance Data Services

API, metadata feeds  
and dashboards for  
monitoring and tracking  
publisher contributions  
to CHORUS



Government  
Agency Reports

Institution Reports

Publisher Reports

Data Set Reports

### National Science Foundation

Click on Data link above to view underlying data and to click through to DOI links to journal articles on publishers' sites.

#### Today's Indicators

257,880 - Publications to date, where:

41.3% - Verified open access on Publishers Site (publisher members DOIs only)

65.2% - Reuse terms available

94.8% - Archived (publisher members DOIs only)

4.1% - Datasets

50.2% - ORCID IDs

38.6% - Agency Portal URLs (publisher members DOIs only)

36285 - # Total number of Datasets

374996 - # Total number of ORCID IDs

73878 - # Total number of Agency Portal URLs

#### Key Performance Indicators - History

Click on colored dots in legend to show or hide key indicator lines

Legend: Total (black), Open Access On Publishers Site (blue), Reuse Terms Available (red), Archived (orange), Datasets (green), ORCID IDs (purple), Agency Portal URLs (light blue)



### CHORUS Dashboard

Summary Data History Reports Search

### U.S. Department of Agriculture

Click on Data link above to view underlying data and to click through to DOI links to journal articles on publishers' sites.

# Deposits  
identifying  
funding



#### Today's Indicators

25,495 - Publications to date, where:

26.9% - Verified open access on Publishers Site (publisher members DOIs only)

75.1% - Reuse terms available

83.9% - Archived (publisher members DOIs only)

21.6% - Datasets

50.0% - ORCID IDs

5.6% - Agency Portal URLs (publisher members DOIs only)

47086 - # Total number of Datasets

21890 - # Total number of ORCID IDs

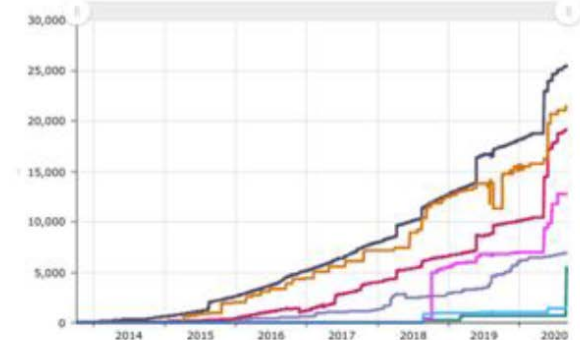
1434 - # Total number of Agency Portal URLs

#### Key Performance Indicators - History

Click on colored dots in legend to show or hide key indicator lines

Filter: CHORUS Publisher Members

Legend: Total (black), Open Access On Publishers Site (blue), Reuse Terms Available (red), Archived (orange), Datasets (green), ORCID IDs (purple), Agency Portal URLs (light blue)





# AGRICULTURAL RESOURCE MANAGEMENT SURVEY

Used in 469  
Publications  
(TOPICS)

Farm Economy	88
Farm Household Well-Being	69
Agricultural Research and Productivity	43
Organic Agriculture	15
Poverty & Income Volatility	10
Food Access	6

49 Mishra Ashrok

20 Nigel Key

19 Hisham Said El-Osta

15 Jeffrey Gillespie

15 Richard Nehring

Used by 652  
Authors

95

Used by 144  
Institutions

Economic Research Service 28

Michigan State 7

University of Minnesota 7

University of Nebraska-Lincoln 7

University of Missouri 6

78 Census of Agriculture, USDA

20 Environmental Quality Incentives Program, USDA

16 National Resources Inventory, USDA

15 Cropland Data Layer, USDA

11 US Census, US Census Bureau

Combined  
with 100  
other  
Datasets

# VALIDATION TOOL

ADRF

Reviews Pending: 1

Continuous Corn and Soybean Yield Penalties across Hundreds of Thousands of Fields

USDA

PDF Link

ID: d4621b9f5da1e179c7bb

Review

Reviews Completed: 2

Exploring the Parents' Attitudes and Perceptions About School Breakfast to Understand Why Participation Is Low in a Rural Midwest State

USDA

PDF Link

ID: 4665ac5e8da4631e30c

Edit

Multiplex restriction amplicon sequencing: a novel nextgeneration sequencingbased marker platform for highthroughput genotyping

USDA

PDF Link

ID: 2e148c9d32c513b32

Edit

ADRF

Continuous Corn and Soybean Yield Penalties across Hundreds of Thousands of Fields

Dyads Identified: 3

USDA

PDF Link

ID: d4621b9f5da1e179c7bb

Semantic Text	Dataset Identified	Score	Correctly Identified?
ortmann, 2001. Fertilizer suggestions for corn. NebGuide G74174 A. Univ. of Nebraska Lincoln. Soil Survey Staff. 2014. Gridded soil survey geographic gSSURGO database. USDA Natural Resources Conserv. Serv. Washington DC. Stanger T.F. and J.G. Lauer. 2008. Corn grain yield response to crop rotation and nitr	gSSURGO	0.169769436	<input checked="" type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Unsure
1 and 3 yr 1 yr and 4 plus yr and 2 yr and 4 plus yr all being significant a finding that is also consistent with Crookston et al. 1991 . Analysis of Soil Weather and Yield Interactions As a final analysis interactions between the CCYP and CSYP and other variables were examined. Two hypotheses were tested first that better soil clima	Soil Weather and Yield Interactions	0.122835346	<input type="radio"/> Yes <input checked="" type="radio"/> No <input type="radio"/> Unsure
Comparing the full dataset to USDA ARMS data for 2010 (USDA ERS and NASS, 2013). 16% of planted land area were in continuous corn as opposed to 18% of planted land area in our data for that year.	Agricultural Resource Management Survey	0.764839172	<input checked="" type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Unsure

Submit

Cancel

ADRF

Continuous Corn and Soybean Yield Penalties across Hundreds of Thousands of Fields

Dyads Identified: 3

USDA

PDF Link

ID: d4621b9f5da1e179c7bb

Semantic Text	Dataset Identified	Score	Correctly Identified?
ortmann, 2001. Fertilizer suggestions for corn. NebGuide G74174 A. Univ. of Nebraska Lincoln. Soil Survey Staff. 2014. Gridded soil survey geographic gSSURGO database. USDA Natural Resources Conserv. Serv. Washington DC. Stanger T.F. and J.G. Lauer. 2008. Corn grain yield response to crop rotation and nitr	gSSURGO	0.169769436	<input checked="" type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Unsure
1 and 3 yr 1 yr and 4 plus yr and 2 yr and 4 plus yr all being significant a finding that is also consistent with Crookston et al. 1991 . Analysis of Soil Weather and Yield Interactions As a final analysis interactions between the CCYP and CSYP and other variables were examined. Two hypotheses were tested first that better soil clima	Soil Weather and Yield Interactions	0.122835346	<input type="radio"/> Yes <input checked="" type="radio"/> No <input type="radio"/> Unsure
Comparing the full dataset to USDA ARMS data for 2010 (USDA ERS and NASS, 2013). 16% of planted land area were in continuous corn as opposed to 18% of planted land area in our data for that year.	Agricultural Resource Management Survey	0.764839172	<input checked="" type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Unsure

---

# FIND DATA AT SCALE

1. Identify core dataset-publication dyads
2. Validate results
3. Improve prototype UI
4. Develop a Challenge through Kaggle or GSA Challenge

---

# FIND DATA AT SCALE

Kaggle's data modeling and analysis platform is designed just for competitions.



**5MM+**

members



**4MM+**

uploaded solutions



**300+**

competitions



**50,000+**

open datasets

**“No matter who you are, most of the smartest people work for someone else.”**  
**–Bill Joy Co-founder of Sun Microsystems**

---



---

# THE KAGGLE PROCESS



---

# ENGAGE AND IMPROVE

Other agencies

- NSF, Commerce, NIH

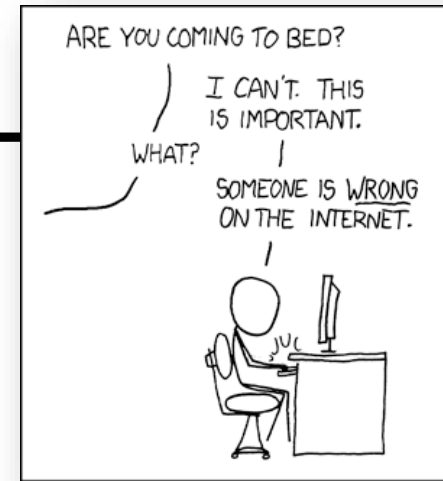
CDO council

Research Community

Award the prize to the winners

Feed the results to the [leaderboard](#) competition

Rinse and repeat with multiple agencies



---

# MASSIVE CURRENT CHALLENGES CAN BE OVERCOME



---

# COMMENTS AND QUESTIONS

- [Julia.lane@nyu.edu](mailto:Julia.lane@nyu.edu)
- Julia.lane@coleridgeinitiative.org