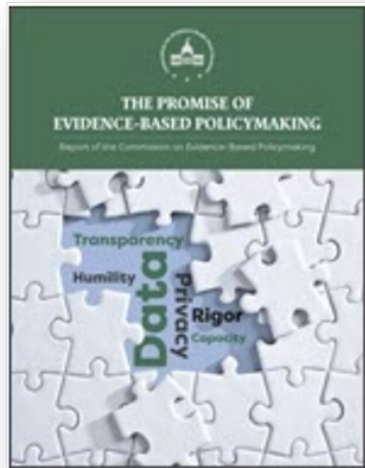


# *Where's the data? A New Approach to Social Science Search and Discovery*

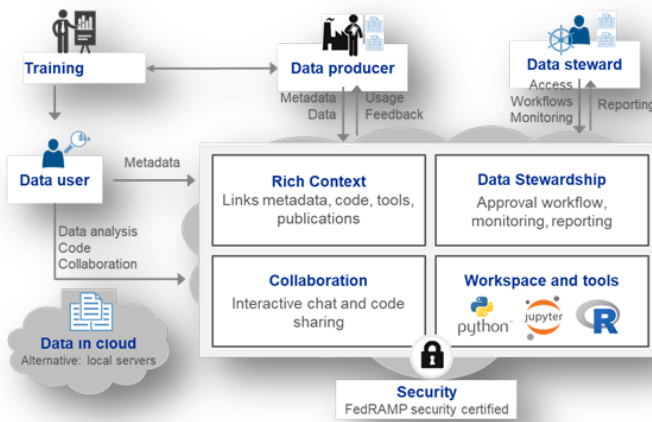
*Julia Lane, NYU  
Brian Granger, Jupyter*

This work is generously supported by a Donor Advised Fund and the Eric&Wendy Schmidt Fund for Strategic Innovation, Overdeck Family Foundation and the Alfred P. Sloan Foundation

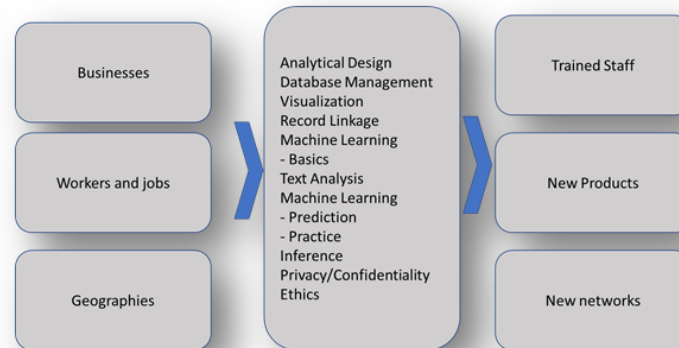
## Impetus



## Response

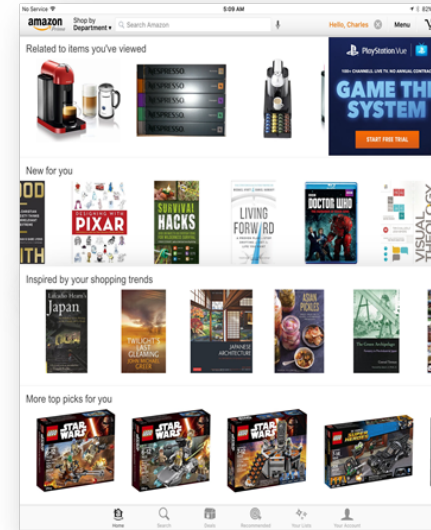


## Admin Data Research Facility

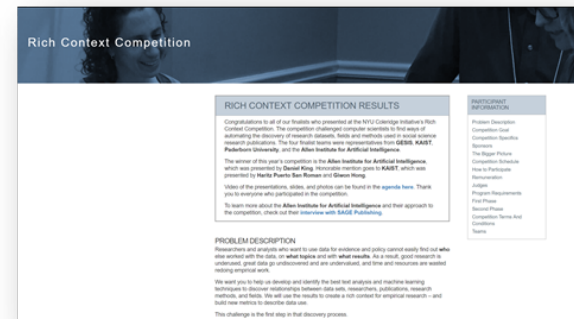


## Applied Data Analytics Program

## Challenge

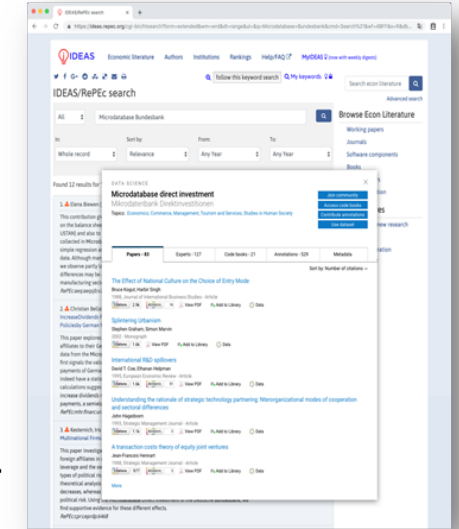


## Search and Discovery

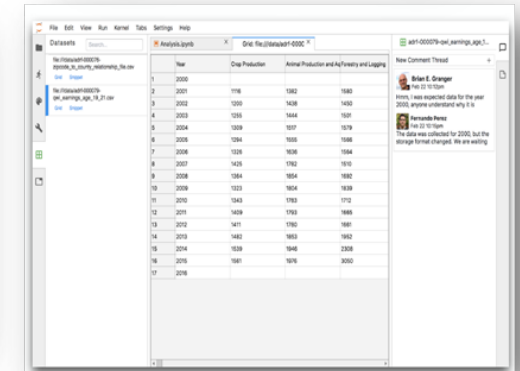


## Rich Context Competition

## New Platform



## Digital Science; Sage



## Jupyter

# Context

## H.R. 1831: Evidence-Based Policy Act of 2016

Introduced: **Apr 16, 2015**

114<sup>th</sup> Congress, 2015–2017

Status: **Enacted — Signed by the President**  
This bill was enacted after being signed by the President.

Law: Pub.L. 114-140

Sponsor:  **Paul Ryan**  
Representative for Wisconsin  
Republican

Text:  **Read Text »**  
Last Updated: Mar 18, 2017  
Length: 5 pages

- [2020 Census](#) design, however, on the new design.
- [CEDCaP \(\\$78M\)](#)
- [American Corner](#) continuing our work.
- [Geographic Information Systems](#) smarter geographic information systems.
- [Administrative Data](#) and federally sponsored linkage techniques for protection, and
- [Economic & Community Development](#) are relevant to the

BIPARTISAN POLICY CENTER

## 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

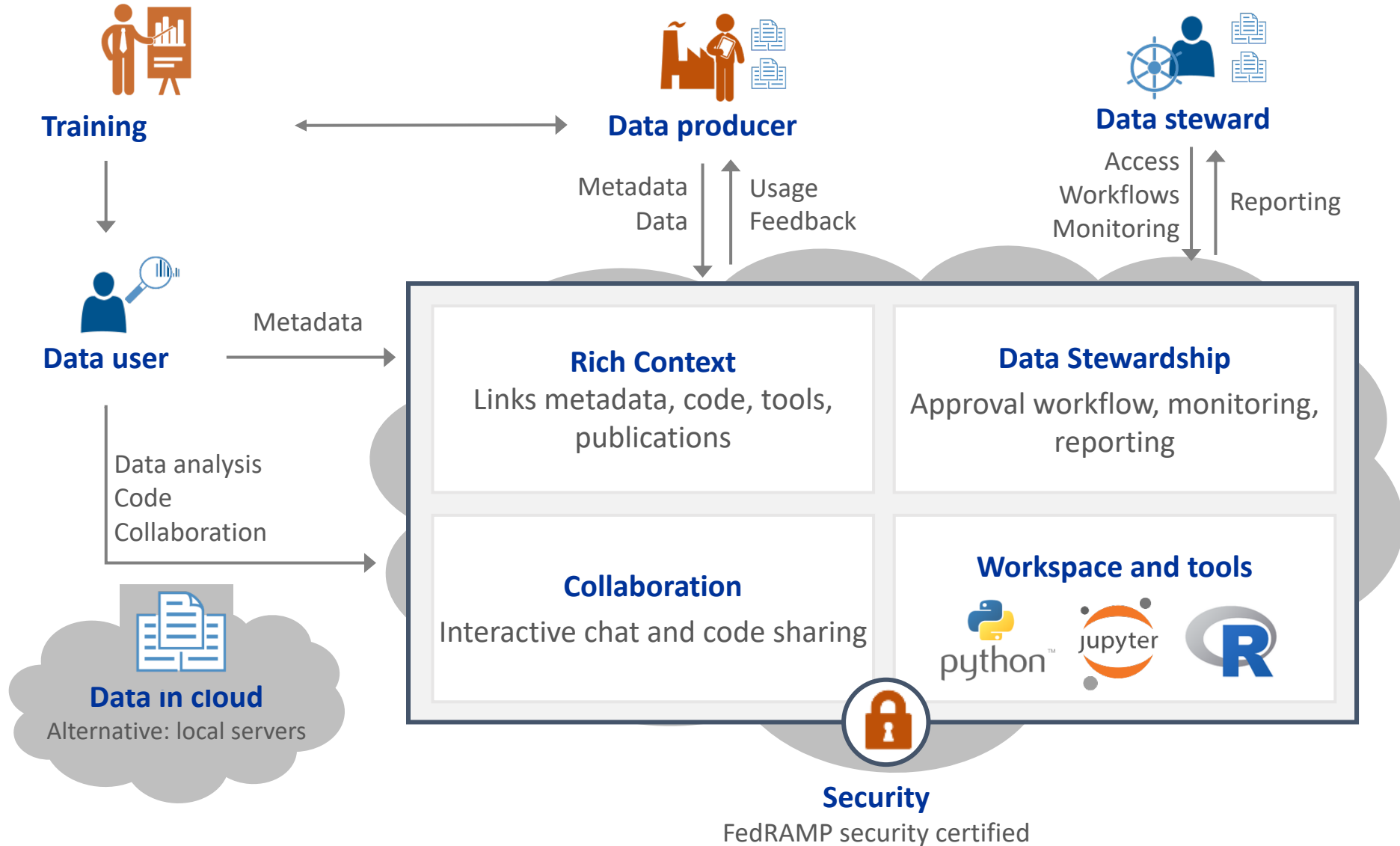
**Directs Agencies to Develop Evidence Plans.** Enables agencies to better prioritize evidence building by requiring that agencies document their key research questions, data needs, and planned activities.

**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.

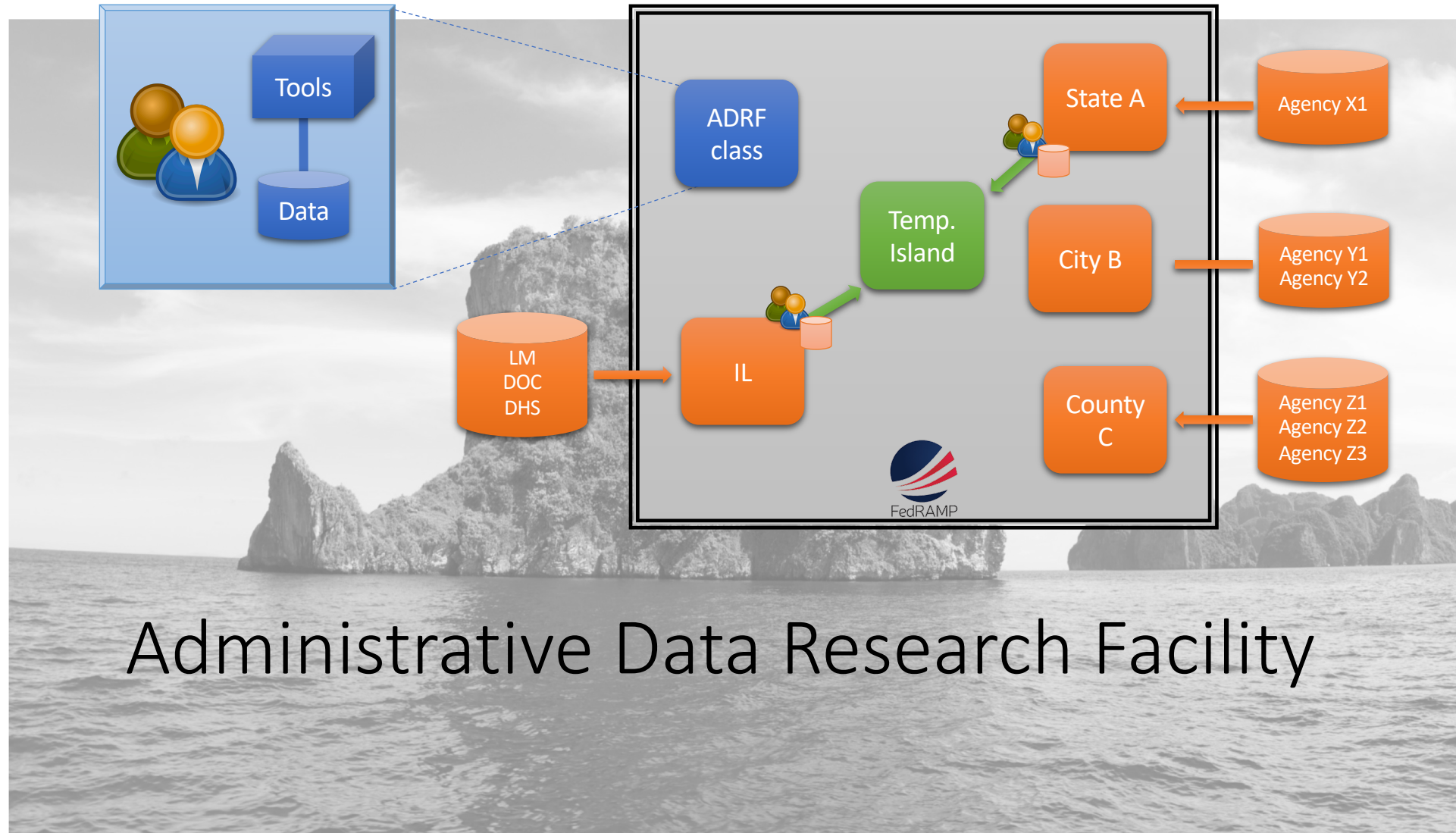
**Research Facility:** The Research Facility is a pilot access to analytical and data services, and is for users, including government analysts and researchers. The Census Bureau developed the Evidence-Based Training Laboratory sponsored by the Department of Justice, New York University, and Maryland.<sup>1</sup> It is currently in use with users accessing the training program. The Evidence-Based computing security approvals, selected confidential information of Housing and Urban Development, the Census Bureau, as well as other federal agencies, and an

# Operational



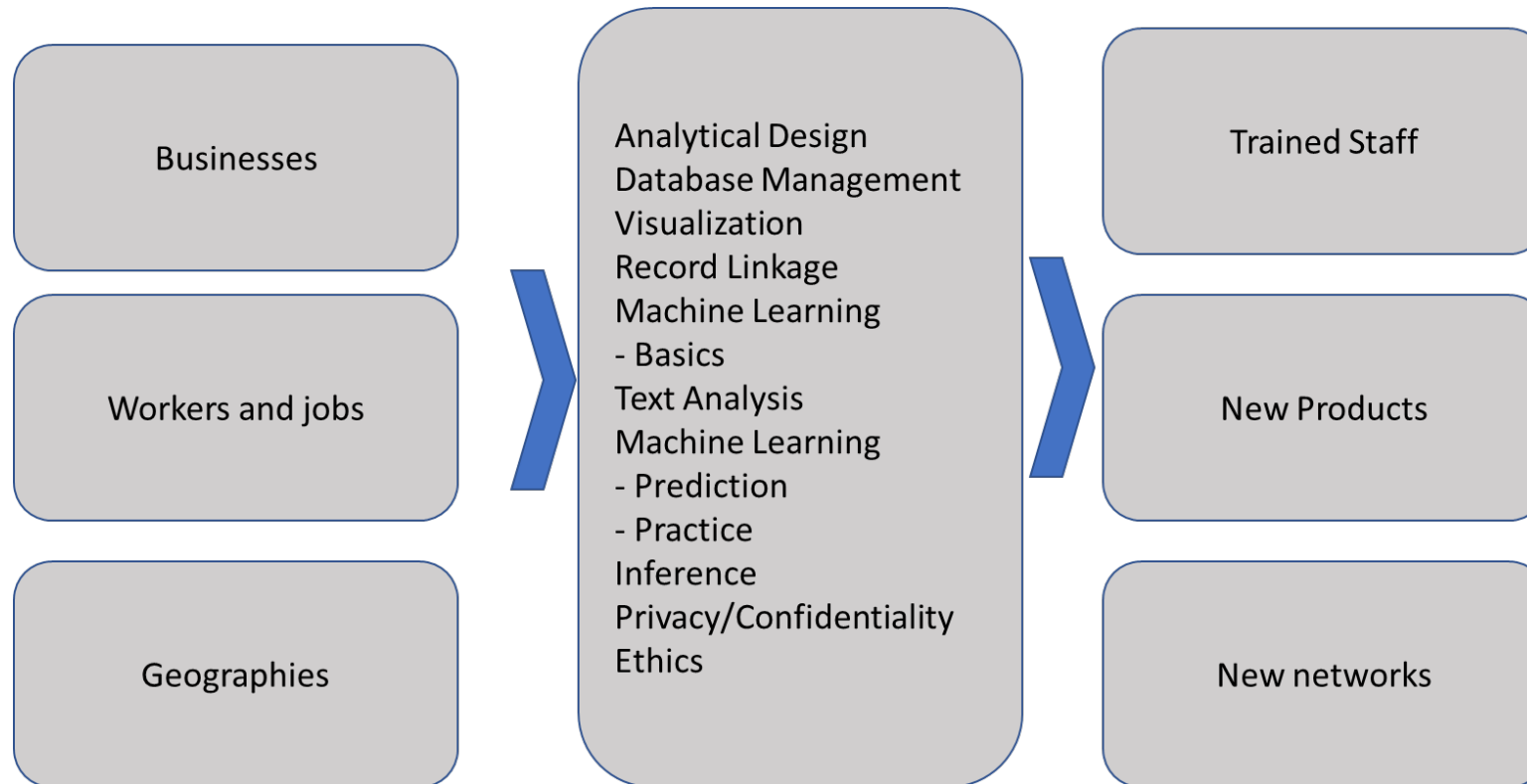


# Access



## Administrative Data Research Facility

# Human: Build classes and workforce capacity



Chapman & Hall/CRC  
Statistics in the Social and Behavioral Sciences Series

## BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools



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

 CRC Press  
Taylor & Francis Group  
A CHAPMAN & HALL BOOK

# Metadata documentation

The image is a collage of three screenshots related to metadata documentation:

- Twitter Thread (#class-3-fall17):** A conversation from Wednesday, January 3rd, 2017. Elena Semenova asks about incarceration data. Vivek Ananda responds. Beau Anderson (CT) explains a database query using `idhs.hh_indcase_spells` and `idhs.member_info` tables, mentioning columns like `ssn_hash`, `sex`, `educational attainment`, `health`, `work experience`, `recptno`, and `ch_dpa_caseid`.
- Google Drive Document:** A document titled "Job assistance programs for welfare recipients" shared by clayton.hunter on December 21st, 2017. The document discusses the challenge of finding stable jobs for welfare recipients and mentions a response from Richard Hendra, MDRC, regarding TANF programs in Chicago.
- Sidebar Menu:** A menu for the "American Community Survey" with options like "Browse by", "About the Survey", "Respond to the Survey", "News & Updates", "Data", "Guidance for Users", "Geography & Maps", "Technical Documentation", "Methodology", "Library", "Operations and Administration", and "Contact Us".

Context

Design

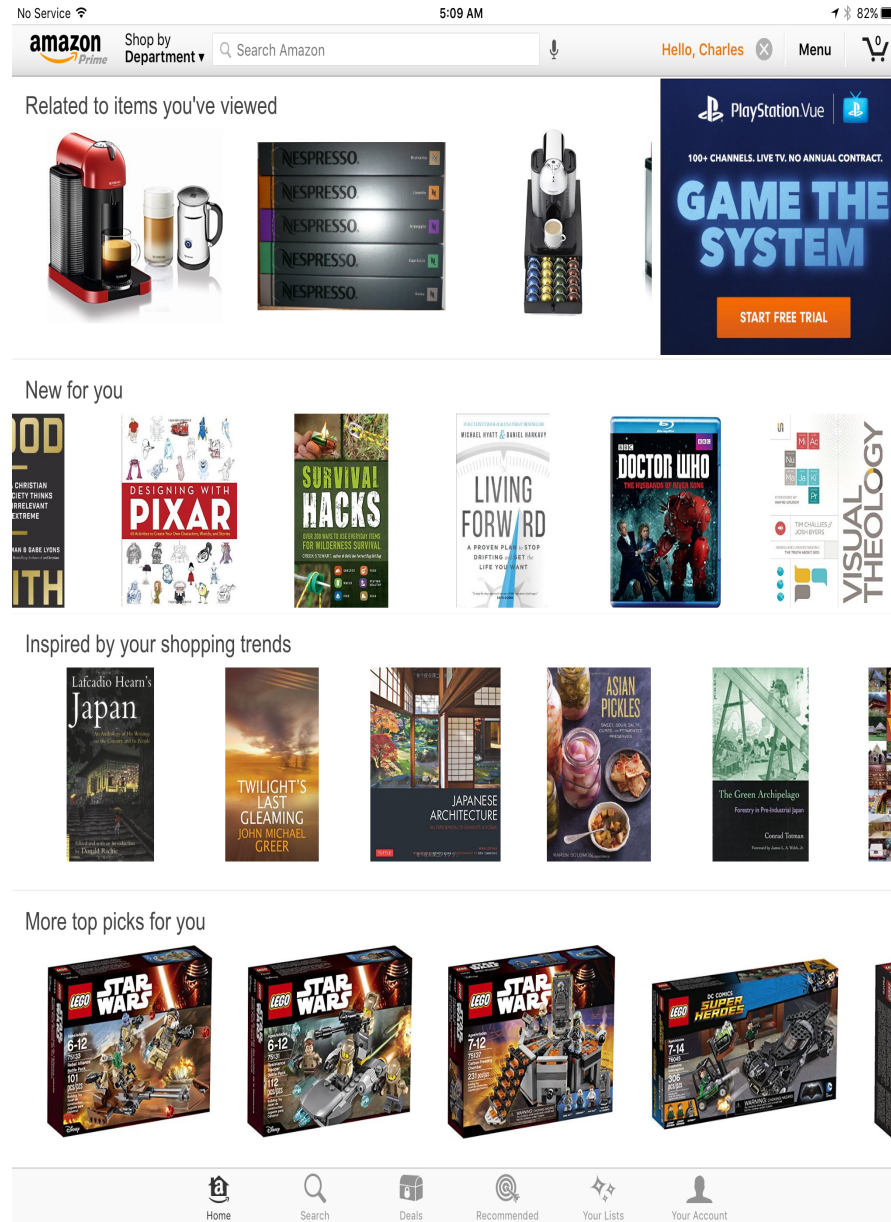
Engagement

Documentation

Implementation

Next Steps

## Search and Discovery



Related to data you've viewed

New data similar to data you've used

What others have done with similar data (recipes)

Recipes like yours

Thank you Charlie Catlett

# Goals

1. **Find, for a given dataset, who** else worked with the data, on **what topics** and with **what results**
2. Create community that contributes code and knowledge



# So how do we do this?

- Step 1: Create the set of corpora and metadata (computer science technology) - Competition
- Step 2; Figure out how you learn from it and automate it (machine learning techniques) - Engagement
- Step 3: Gamification – recognize and emphasize patterns (with human curation) – Rinse and repeat

# Run competition and partner with Digital Science/Dimensions

(<https://coleridgeinitiative.org/richcontextcompetition>)

**Rich Context Competition**

**PROBLEM DESCRIPTION**  
 Researchers and analysts who want to use data for evidence and policy cannot easily find out **who** else worked with the data, on **what topics** and with **what results**. As a result, good research is underused, great data go undiscovered and are undervalued, and time and resources are wasted redoing empirical work.

We want you to help us develop and identify the best text analysis and machine learning techniques to discover relationships between data sets, researchers, publications, research methods, and fields. We will use the results to create a rich context for empirical research – and build new metrics to describe data use.

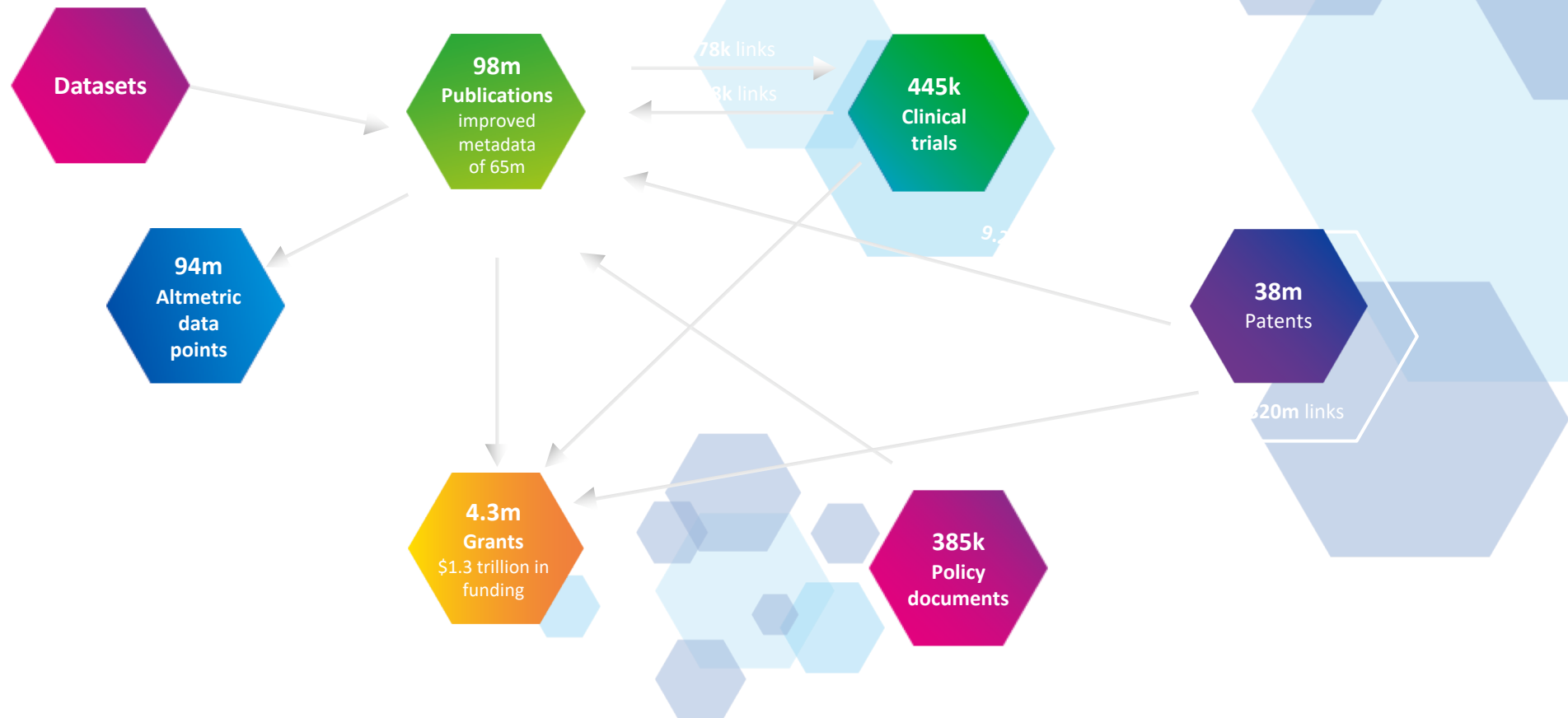
This challenge is the first step in that discovery process.

**COMPETITION GOAL**  
 The goal of this competition is to automate the discovery of research datasets and the associated research methods and fields in social science research publications. Participants should use any combination of machine learning and data analysis methods to identify the datasets used in a corpus of social science publications and infer both the scientific methods and fields used in the analysis and the research fields.

**COMPETITION SPECIFICS**  
 The competition has two phases (details below).

**PARTICIPANT INFORMATION**

- Problem Description
- Competition Goal
- Competition Specifics
- Sponsors
- The Bigger Picture
- Competition Schedule
- How to Participate
- Remuneration
- Judges
- Program Requirements
- First Phase
- Second Phase
- Competition Terms And Conditions
- Teams



## Policy Document Citations - 18

[The power of talk and power in talk: a systematic healthcare communication](#)

2018, Analysis & Policy Observatory (APO)

[Alternative aged care assessment, classification](#)

2017, Analysis & Policy Observatory (APO)

## Patent citations - 29

[Method of treatment using substituted pyrazolo\[1,5-a\]p](#)  
Array Biopharma Inc - Julia Haas, Steven W. Andrews, Yutong Jiang,  
Grant US - Granted year: 2018

[Macrocyclic compounds as TRK kinase inhibitors](#)  
Array Biopharma Inc - Steven W. Andrews, Kevin Ronald Condroski, J  
Grant US - Granted year: 2018  
99 3

## Linked clinical trials - 2

[An Open Label, Multicenter, Phase II Study  
Harboring T790M Mutation Who Failed EGI](#)  
Samsung Medical Center  
Altmetric 3

[A Phase-II Clinical Trial to Evaluate the Acc  
in Patients With Advanced NSCLC Treated  
Markers](#)  
Lung Cancer Group Cologne

## Publication metrics

Dimensions Badge  
2.4k

2.4k Total citations  
419 Recent citations  
141 Field Citation Ratio  
44 Relative Citation Ratio

Altmetric  
23

News (1)  
Blogs (1)  
Patients (29)  
Mendeley (582)  
CiteULike (2)

## About

## Research Categories

Fields of Research  
0601 Biochemistry and Cell Biology  
1112 Oncology and Carcinogenesis

Research, Condition, and Disease Categorizations  
Lung Cancer  
Lung  
Cancer  
Clinical Research  
Genetics  
Rare Diseases

Dimensions

Public Library of Science (PLOS)  
Publisher

Publication - Article

### Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain

PLoS Medicine, 2(3), e73, 2005  
<https://doi.org/10.1371/journal.pmed.0020073>

## Authors

William Pao - Memorial Sloan Kettering Cancer Center  
Vincent A Miller - Memorial Sloan Kettering Cancer Center  
Katerina A Politi - Memorial Sloan Kettering Cancer Center  
5 more

## Abstract

**BACKGROUND:** Lung adenocarcinomas from patients who respond to the tyrosine kinase inhibitors gefitinib (Iressa) or erlotinib (Tarceva) usually harbor somatic gain-of-function mutations in exons encoding the kinase domain of the epidermal growth factor receptor (EGFR). Despite initial responses, patients eventually progress by unknown mechanisms of "acquired" resistance. **METHODS AND FINDINGS:** We show that in two of five patients with acquired resistance to gefitinib or erlotinib, progressing tumors contain, in addition to a primary drug-sensitive mutation in EGFR, a secondary mutation in exon 20, which leads to substitution of methionine for threonine at position 790 (T790M) in the kinase domain. Tumor cells from a sixth patient with a drug-sensitive EGFR mutation whose tumor progressed on adjuvant gefitinib after complete resection also contained the T790M mutation. This mutation was not detected in untreated tumor samples. Moreover, no tumors with acquired resistance had KRAS mutations, which have been associated with primary resistance to these drugs. Biochemical analyses of transfected cells and growth inhibition studies with lung cancer cell lines demonstrate that the T790M mutation confers resistance to EGFR mutants usually sensitive to either gefitinib or erlotinib. Interestingly, a mutation analogous to T790M has been observed in other kinases with acquired resistance to another kinase inhibitor, imatinib (Gleevec). **CONCLUSION:** In patients with tumors bearing gefitinib- or erlotinib-sensitive EGFR mutations, resistant subclones containing an additional EGFR mutation emerge in the presence of drug. This observation should help guide the search for more effective therapy against a specific subset of lung cancers.

[less](#)

## Supporting grants - 1

[Investigational Cancer Therapeutics Training Program](#)  
National Cancer Institute

to MARK G. KRIS, JEDD D. WOLCHOK, DEAN F. BAJORIN, JOHN MENDEL,  
J. BOSL

Publication citations - 2424 [Show all](#)

[Loss of T790M mutation is associated with early progression  
harboring EGFR T790M](#)

Sha Zhao, Xuefei Li, Chao Zhao, Tao Jiang, Yijun Jia, Jiepeng Shi, Yayi He,  
2019, Lung Cancer - Article  
Altmetric 1 Add to Library

[Oncogene addiction as a foundation of targeted cancer ther](#)  
Eleonora Orlando, Daniel Matthias Aebbersold, Michaela Medová, Yizhak Z  
2019, Cancer Letters - Article  
Add to Library

## MeSH terms

Adenocarcinoma; Antineoplastic Agents;  
Carcinoma, Non-Small-Cell Lung; DNA Mutational  
Analysis; Disease Progression; Drug Resistance;  
Neoplasm; Erlotinib Hydrochloride; Exons; Female;  
Humans; Lung Neoplasms; Middle Aged; Point  
Mutation; Quinazolines; Receptor, Epidermal Growth  
Factor; Reverse Transcriptase Polymerase Chain  
Reaction; Tumor Cells, Cultured

## Funded by

AstraZeneca (United States)  
American Cancer Society  
National Cancer Institute

## Associated data



## Publication references - 27

[The development of imatinib as a therapeutic agent for chro](#)  
Michael Deininger, Elisabeth Buchdunger, Brian J. Druker  
2006, Blood - Article

Altmetric 10 Open Access Add to Library

[Clinical and Biological Features Associated With Epidermal](#)  
Hisayuki Shigematsu, Li Lin, Takao Takahashi, Masaharu Nomura, Makoto  
2005, JNCI Journal of the National Cancer Institute - Article

Altmetric 14 Open Access Add to Library

# Build platform

- Step 1: Create the set of corpora and metadata (computer science technology) - Competition
- Step 2; Figure out how you learn from it and automate it (machine learning techniques) - Engagement
- Step 3: Gamification – recognize and emphasize patterns (with human curation) – Rinse and repeat



# Next steps 1

- Work with Dimensions/Digital Science
- <https://experts.umich.edu/discover/publication>
- Apply model to 60 million documents
- Develop user facing tools with Dimensions, Figshare, Altmetric, Readcube Papers and Symplectic Elements.

IDEAS/RePEc search

https://ideas.repec.org/cgi-bin/htsearch?form=extended&wm=wr&dt=range&ul=&q=National+Health+and+Nutrition+Examination+Survey&cmd=Search...

IDEAS

DATA SCIENCE

### National Health and Nutrition Examination Survey (NHANES)

Topics: [Public Health and Health Services](#); [Clinical Sciences](#); [Nutrition and Dietetics](#); [Psychology](#)

[Join community](#)  
[Access code books](#)  
[Contribute annotations](#)  
[Use dataset](#)

In: [All](#) [N](#)

Whole record

Sort by: Number of citations

Found 373 results for

Papers - 83	Experts - 127	Code books - 21	Annotations - 529	Metadata
<a href="#">Initial sequencing and analysis of the human genome</a> E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle 2001, Nature - Article <a href="#">Citations</a> 15k <a href="#">Altmetric</a> 713 <a href="#">View PDF</a> <a href="#">Add to Library</a> <a href="#">Data</a>				
<a href="#">Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin</a> William C Knowler, Elizabeth Barrett-Connor, Sarah E Fowler, Richard F Hamman, John M Lachin 2002, New England Journal of Medicine - Article <a href="#">Citations</a> 11k <a href="#">Altmetric</a> 1,492 <a href="#">View PDF</a> <a href="#">Add to Library</a> <a href="#">Data</a>				
<a href="#">Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure</a> Aram V. Chobanian, George L. Bakris, Henry R. Black, William C.ushman, Lee A. Green 2003, Hypertension - Article <a href="#">Citations</a> 1.6k <a href="#">Altmetric</a> 51 <a href="#">View PDF</a> <a href="#">Add to Library</a> <a href="#">Data</a>				
<a href="#">A new equation to estimate glomerular filtration rate.</a> Andrew S. Levey, Lesley A. Stevens, Christopher H. Schmid, Yaping (Lucy) Zhang 2009, Annals of Internal Medicine - Article <a href="#">Citations</a> 8.6k <a href="#">Altmetric</a> 23 <a href="#">View PDF</a> <a href="#">Add to Library</a> <a href="#">Data</a>				
<a href="#">Prevalence of Overweight and Obesity in the United States, 1999-2004</a> Cynthia L. Ogden, Margaret D. Carroll, Lester R. Curtin, Margaret A. McDowell, Carolyn J. Tabak 2006, JAMA - Article <a href="#">Citations</a> 6.8k <a href="#">Altmetric</a> 114 <a href="#">View PDF</a> <a href="#">Add to Library</a> <a href="#">Data</a>				

[More](#)

1. [Nancy Cole & M](#)  
[Participation Status](#)  
[1999-2004](#)  
This report uses the  
Examination Survey  
picture of the diets  
*RePEc:mpres:3*

2. [Nancy Cole & M](#)  
[Participation Status](#)  
[1999-2004](#)  
The Special Supple  
provides nutrient-de  
low-income pregnan  
are at Nutritional ris  
*RePEc:mpres:9*

3. [Elizabeth Cond](#)  
[Niland \(0001\): Diet C](#)  
[National Health and](#)

# Next steps 2: Work with Sage Publications

- Identify one or two SAGE journals where we could run any new papers from those journals against the model and recruit the authors to help verify the models
- Include data keywords into article keywords
- Develop data impact metrics/dashboards for journals or authors

# Next Steps 3:

## Build into Researcher Onboarding

ADRF | Data Stewardship Gonen Minuskin

### Project Request

Projects / 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

Datasets

Project Name

Project Name

Project Dates

Start Date

End Date

IRB Approval

☐ This project has or is pending IRB approval (required).

Research Question

Research Methodology

Expected Outcomes

How will this project further the agency's mission?

Save as Draft

Submit

## Members

### Project Review

Projects / Project Review

Overview

Members

Datasets

Agreements

**Drew Gordon**  
drew.gordon@nyu.edu  
Sr. Developer | NYU

**Clayton Hunter**  
crh278@nyu.edu  
Sr. Developer | NYU

Feedback


Gonen Minuskin - 1/16/2019 10:38:27 AM  
Looking forward to working on this!

Leave feedback

Reject

Approve

# Next Steps 3: Researcher Onboarding

ADRF | Data Stewardship 5  Gonen Minuskin

## Project Request

Projects / 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

Datasets

Project Name

Project Name

Project Dates

Start Date

End Date

IRB Approval

☒ This project has or is pending IRB approval (required).

Research Question

Research Methodology

Expected Outcomes

How will this project further the agency's mission?

Save as Draft

Submit

## Datasets

Overview

Members

Datasets

Agreements

Cook County Recorder - Foreclosures, Mortgages, and Quit Claim Deeds

Dataset ID: adrf-000033

Data Steward: Drew Gordon

Foreclosures, Mortgages, and Quit Claim Deeds recorded with the Cook County Recorder of Deeds. Data covers 2013 through March 27, 2015.

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

Dataset ID: adrf-000005

Data Steward: Drew Gordon

The Demographic Profile Summary File (SF1) contains 100 percent data asked of all people and about every housing unit on topics such as sex, age, race, Hispanic or Latino origin, household relationship, household type, group quarters population, housing occupancy, and housing tenure.



# Next Steps 4: Jupyter Notebook

## Making Computational Research with Sensitive Data Possible and Valuable

Brian E. Granger  
Associate Professor  
Cal Poly

Julia Lane  
Professor  
NYU

Fernando Perez  
Assistant Professor  
UC Berkeley



Alfred P. Sloan  
FOUNDATION

SCHMIDT **FUTURES**



Overdeck Family  
Foundation

# Commenting and Annotation in JupyterLab

The screenshot displays the JupyterLab interface. On the left, the 'Datasets' sidebar lists two CSV files: 'file:///data/adrf-000076-zipcode\_to\_county\_relationship\_file.csv' and 'file:///data/adrf-000079-qwi\_earnings\_age\_19\_21.csv'. The main area shows a Jupyter Notebook with a single cell containing a data grid. The grid has columns for 'Year', 'Crop Production', 'Animal Production and Aquaculture', and 'Forestry and Logging'. The data spans from 2000 to 2016. A comment thread is visible on the right side of the grid, titled 'New Comment Thread'. It contains two comments: one from Brian E. Granger asking about the 2000 data, and another from Fernando Perez explaining that the data was collected for 2000 but the storage format changed, and they are waiting.

	Year	Crop Production	Animal Production and Aquaculture	Forestry and Logging
1	2000			
2	2001	1116	1382	1580
3	2002	1200	1438	1450
4	2003	1255	1444	1501
5	2004	1309	1517	1579
6	2005	1294	1555	1566
7	2006	1326	1636	1564
8	2007	1425	1782	1510
9	2008	1364	1854	1692
10	2009	1323	1804	1839
11	2010	1343	1783	1712
12	2011	1409	1793	1665
13	2012	1411	1780	1661
14	2013	1482	1853	1952
15	2014	1539	1946	2308
16	2015	1561	1976	3050
17	2016			

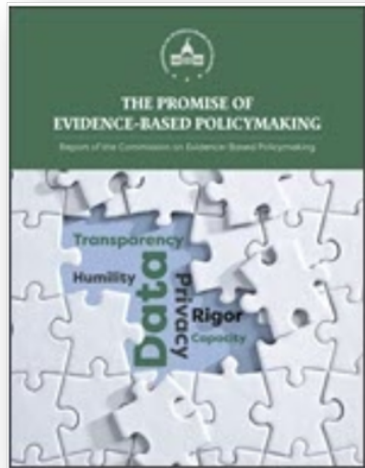
**New Comment Thread**

**Brian E. Granger**  
Feb 22 10:12pm  
Hmm, I was expected data for the year 2000, anyone understand why it is

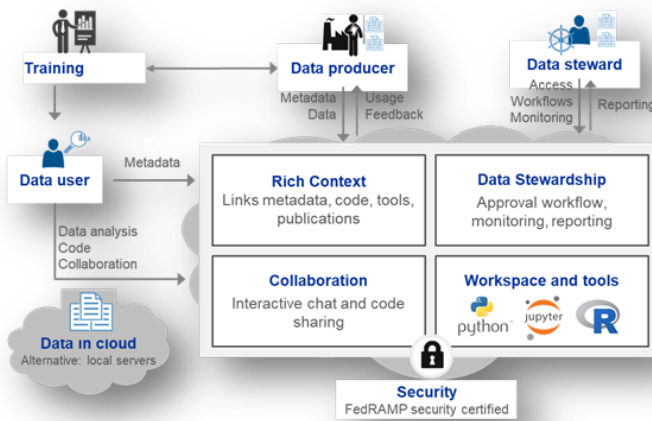
**Fernando Perez**  
Feb 22 10:15pm  
The data was collected for 2000, but the storage format changed. We are waiting

\* Early prototype

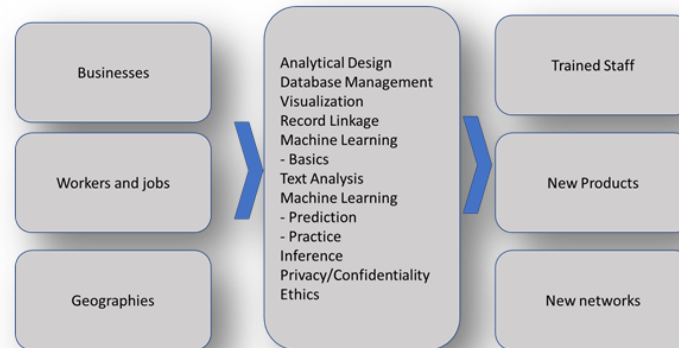
## Impetus



## Response

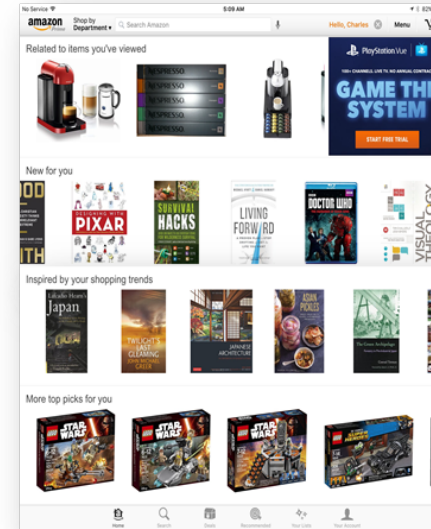


## Admin Data Research Facility

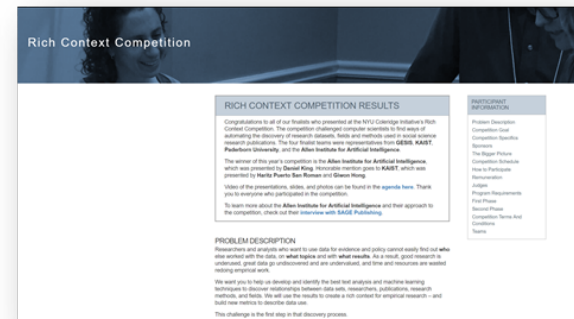


## Applied Data Analytics Program

## Challenge

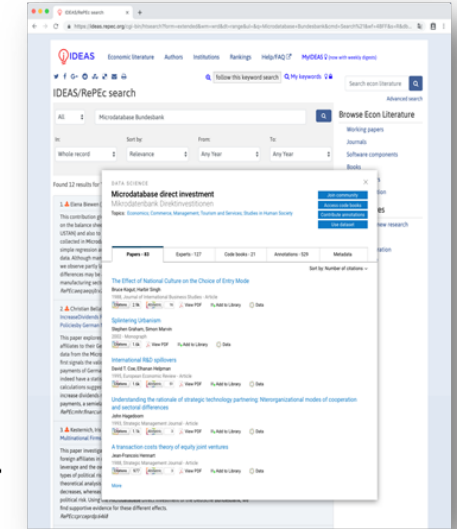


## Search and Discovery

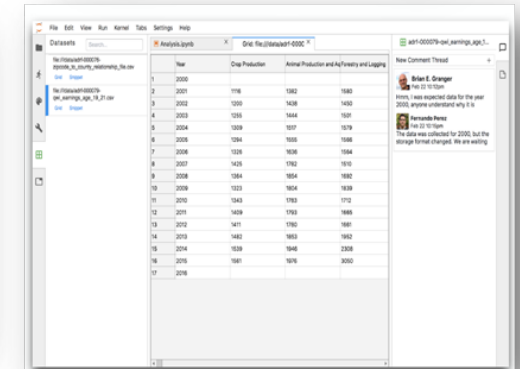


## Rich Context Competition

## New Platform



## Digital Science; Sage



## Jupyter

# Contact and more information

- [Website: https://coleridgeinitiative.org/](https://coleridgeinitiative.org/)
- [E-mail](#)
  - [dataanalytics@coleridgeinitiative.org](mailto:dataanalytics@coleridgeinitiative.org)
  - [julia.lane@nyu.edu](mailto:julia.lane@nyu.edu)
  - [clayton.hunter@nyu.edu](mailto:clayton.hunter@nyu.edu)
- [GitHub organization: https://github.com/Coleridge-Initiative](https://github.com/Coleridge-Initiative)