# Next Gen Usage Data: A New Search and Discovery Platform

**USDA/ ERS & NASS**
**Democratizing Data**
**Joint Info Session**

**March 28, 2023**

# Overview

- **USDA Vision, Theory of Change** – Spiro Stefanou
- **ERS Goals**– Kelly Maguire
- **Overview of May 1 In-person Workshop in KCMO** - Kelly Maguire
- **ERS and NASS Dashboards**– Hubert Hamer, Nick Pallotta
- **Introduction to Jupyter Notebooks and API** - Julia Lane
- **Data Ecosystem; Evidence Act and Democratizing Data Vision** – Nancy Potok
- **Q & A** – moderated by Kelly Maguire

# ERS: Theory of Change

# Data as a Product Innovation

**Specific Focus:**

How are specific datasets being used in scientific and public research?

- Agricultural Resource Management Survey
- Rural Urban Continuum Codes

**Specific Application:**

- Internal investment decisions
- USDA policy and strategies
- Congressional decision making

**Better information for:**

- Think tanks
- Researchers
- International agriculture statistics community

# Theory of Change Model

| Inputs | Activities | Outputs | Outcomes | Final Outcomes |
|---|---|---|---|---|
| • Assessing the need for the data asset<br>• Survey development, piloting, execution<br>• Integrating survey data with proprietary data<br>• Curation | • Natural language processing<br>• API<br>• Dashboards<br>• Notebooks | • Research and analysis<br>• Stronger communities<br>• Networks<br>• More information that can be used | • Respondents respond more<br>• Policy makers and stakeholders have evidence-based insights<br>• Stimulate connections with interdependent systems | • Congressional policy action<br>• Reassess scope of potential users<br>• Refine and innovate data assets<br>• Gain insights into emerging trends |

# Assessing the Value of Public Data Assets

**Costs and Risks**

- Acquisition, collection, curation, protection, storage
- Risk of disclosure, re-identification, and reputation to the agency

**Reward (or Utility) – more anecdotal**

- Real value of these data to society, researchers and policy makers is yet to be determined
- Public provided data and information are special goods and offer a special challenge

**Project in progress**

- Develop an approach for evaluating the value of publicly available datasets and the potential value of free public access to these data
- Start with a proof of concept developing the basic methodology and then applying it to two distinct data sets
- Infometrics approach based on information theory

# ERS Goals

# Key Questions of Interest to ERS

1. Who are the audiences that use ERS data, how are they linking the data with other data sources, and what evidence are they building?
2. How do audiences engage with ERS's data portfolio for increased awareness and informed priority setting by ERS?
3. How do audiences engage with usage information to make informed decisions about research, collaborations/partnerships, and planning?
4. How can the usage data, their underlying assumptions, and their presentations be improved for greatest utility and maximum impact?

# The Potential for Other Value-added

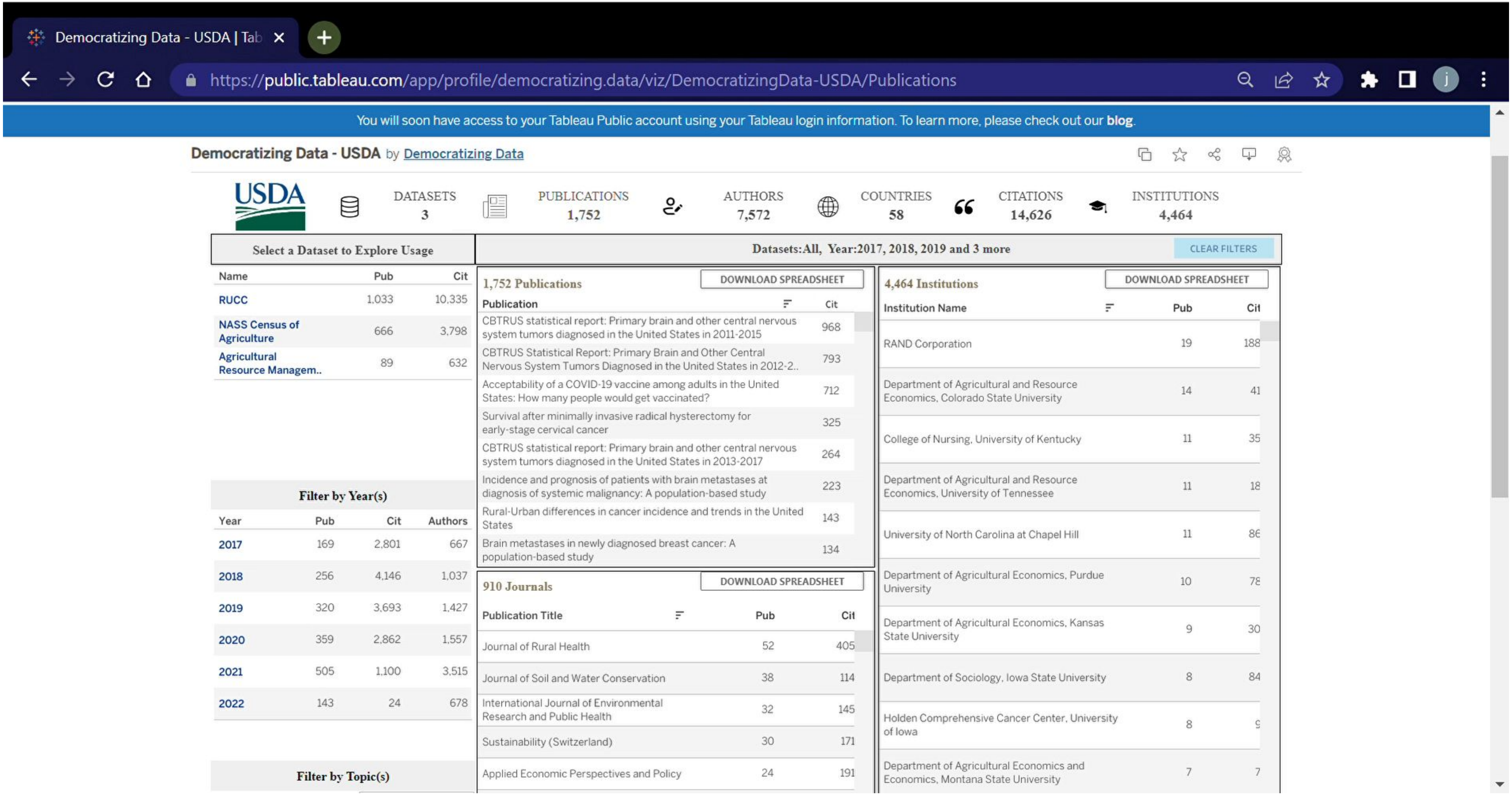
**As a data platform…**

- Builds a potential health profile that can cross-cut with other agencies' data usage
- Enables user communities to connect and partner on research, analysis and application of the data
- Supports transparency in data access, data use, and federal statistical offerings
- Builds public trust in official federal statistics

# Workshop Goals and Format

- May 1, 2023: bring together researchers and data users to “test drive” the Jupyter notebooks
- Purpose: Get feedback on the usability and data completeness
    - What’s missing?
    - What are future areas for enhancements?
    - How can we spread the word about this platform?
- Half day in-person session in Kansas City
- Could lead to follow on events

# ERS and NASS Dashboards


Democratizing Data - USDA | Tab
https://public.tableau.com/app/profile/democratizing.data/viz/DemocratizingData-USDA/Publications

You will soon have access to your Tableau Public account using your Tableau login information. To learn more, please check out our **blog**.

**Democratizing Data - USDA** by Democratizing Data

USDA

| DATASETS | PUBLICATIONS | AUTHORS | COUNTRIES | CITATIONS | INSTITUTIONS |
|---|---|---|---|---|---|
| 3 | 1,752 | 7,572 | 58 | 14,626 | 4,464 |

**Select a Dataset to Explore Usage**

**Datasets:All, Year:2017, 2018, 2019 and 3 more**

CLEAR FILTERS

| Name | Pub | Cit |
|---|---|---|
| RUCC | 1,033 | 10,335 |
| NASS Census of Agriculture | 666 | 3,798 |
| Agricultural Resource Managem.. | 89 | 632 |

**Filter by Year(s)**

| Year | Pub | Cit | Authors |
|---|---|---|---|
| 2017 | 169 | 2,801 | 667 |
| 2018 | 256 | 4,146 | 1,037 |
| 2019 | 320 | 3,693 | 1,427 |
| 2020 | 359 | 2,862 | 1,557 |
| 2021 | 505 | 1,100 | 3,515 |
| 2022 | 143 | 24 | 678 |

**Filter by Topic(s)**

**1,752 Publications**

DOWNLOAD SPREADSHEET

| Publication | Cit |
|---|---|
| CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2011-2015 | 968 |
| CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012-2.. | 793 |
| Acceptability of a COVID-19 vaccine among adults in the United States: How many people would get vaccinated? | 712 |
| Survival after minimally invasive radical hysterectomy for early-stage cervical cancer | 325 |
| CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2013-2017 | 264 |
| Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: A population-based study | 223 |
| Rural-Urban differences in cancer incidence and trends in the United States | 143 |
| Brain metastases in newly diagnosed breast cancer: A population-based study | 134 |

**910 Journals**

DOWNLOAD SPREADSHEET

| Publication Title | Pub | Cit |
|---|---|---|
| Journal of Rural Health | 52 | 405 |
| Journal of Soil and Water Conservation | 38 | 114 |
| International Journal of Environmental Research and Public Health | 32 | 145 |
| Sustainability (Switzerland) | 30 | 171 |
| Applied Economic Perspectives and Policy | 24 | 191 |

**4,464 Institutions**

DOWNLOAD SPREADSHEET

| Institution Name | Pub | Cit |
|---|---|---|
| RAND Corporation | 19 | 188 |
| Department of Agricultural and Resource Economics, Colorado State University | 14 | 41 |
| College of Nursing, University of Kentucky | 11 | 35 |
| Department of Agricultural and Resource Economics, University of Tennessee | 11 | 18 |
| University of North Carolina at Chapel Hill | 11 | 86 |
| Department of Agricultural Economics, Purdue University | 10 | 78 |
| Department of Agricultural Economics, Kansas State University | 9 | 30 |
| Department of Sociology, Iowa State University | 8 | 84 |
| Holden Comprehensive Cancer Center, University of Iowa | 8 | 9 |
| Department of Agricultural Economics and Economics, Montana State University | 7 | 7 |

# The 5 W’s of NASS data: Discovering Data Usefulness (Who, What, When, Where, and Why)

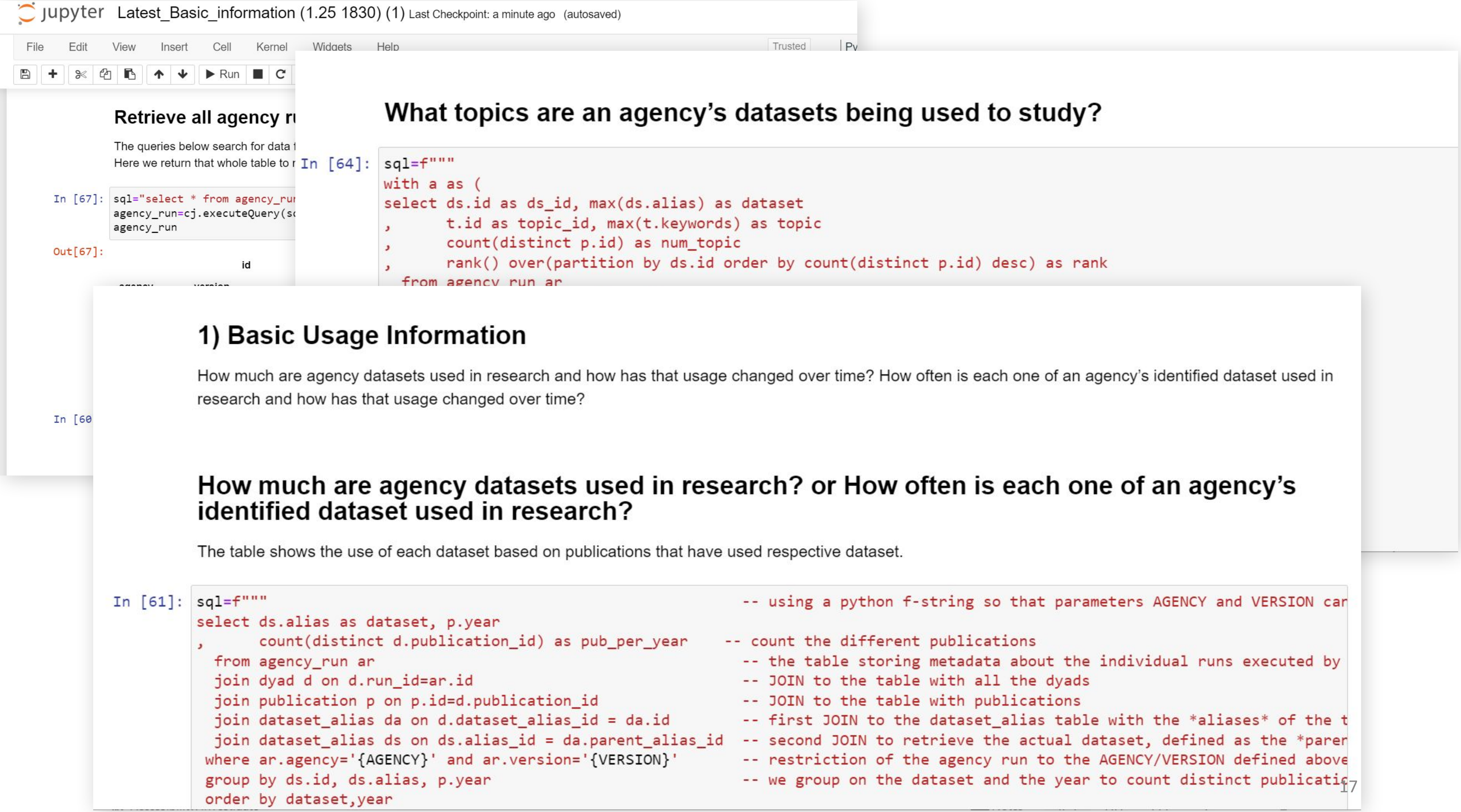
NASS stakeholders, including staff, data users, respondents, researchers, extension, policy makers, other government agencies have varying levels of understanding the “usefulness” of NASS data. The goal of this project is to produce a publicly available dashboard that attempts to demonstrate to all stakeholders an answer to the 5 W’s of NASS Data:

1. Who is using NASS data?
2. What are they using it for?
3. Where geographically is the data being used?
4. When are the heaviest periods of use?
5. Why is data being used?

# What is the 5 W’s Dashboard?

- Displays publications from agricultural researchers made with NASS data (ARMS and Census)
- Originated from a data science competition with partial funding from ERS with the goal of developing a machine learning model that could identify dataset usages in journal publications
- Developed by REE analytics and NASS’s Strategic Planning Branch
- Response rate experiment using the dashboard

# Introduction to Jupyter Notebooks and API


jupyter
Latest_Basic_information (1.25 1830) (1) Last Checkpoint: a minute ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Py
Run

## Retrieve all agency r

The queries below search for data f
Here we return that whole table to r

```
In [67]: sql="select * from agency_ru
         agency_run=cj.executeQuery(sq
         agency_run
Out[67]:
                              id
         agency        version
```

```
In [60
```

## What topics are an agency's datasets being used to study?

```
In [64]: sql=f"""
         with a as (
         select ds.id as ds_id, max(ds.alias) as dataset
         ,       t.id as topic_id, max(t.keywords) as topic
         ,       count(distinct p.id) as num_topic
         ,       rank() over(partition by ds.id order by count(distinct p.id) desc) as rank
           from agency_run ar
```

## 1) Basic Usage Information

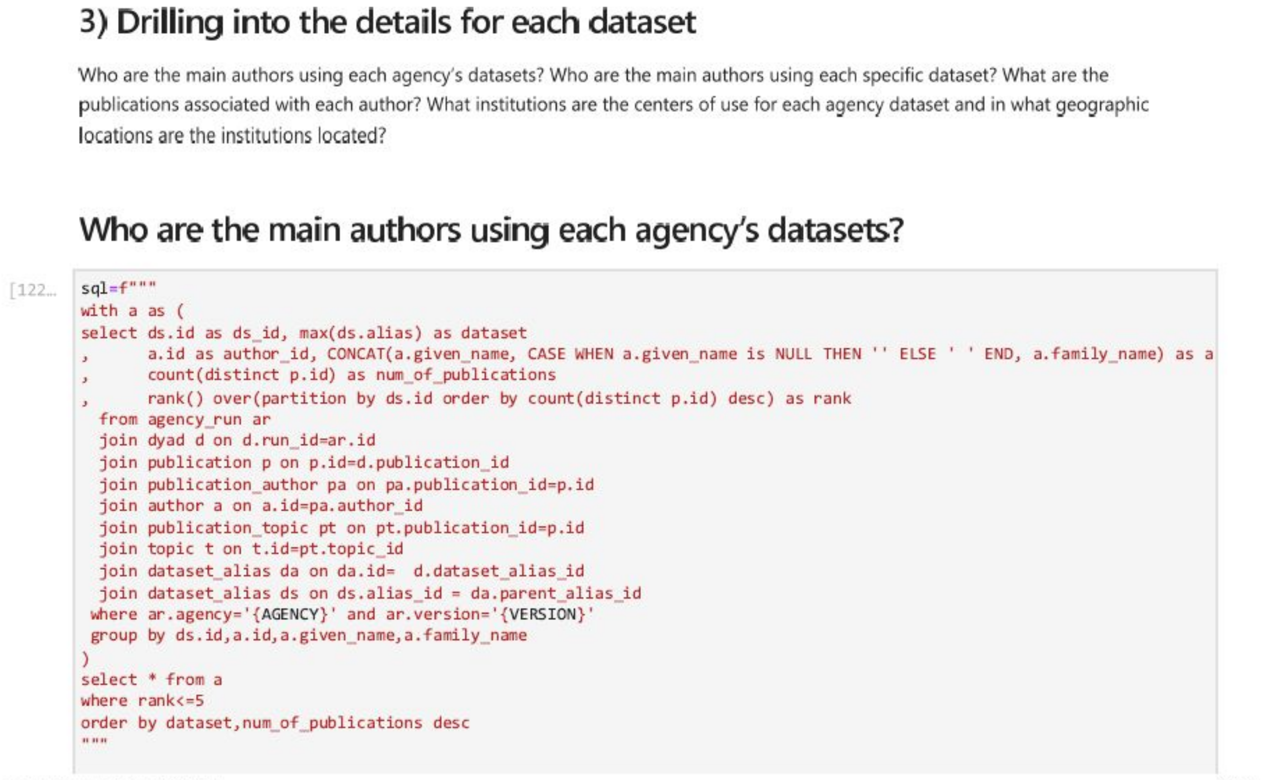
How much are agency datasets used in research and how has that usage changed over time? How often is each one of an agency's identified dataset used in research and how has that usage changed over time?

## How much are agency datasets used in research? or How often is each one of an agency's identified dataset used in research?

The table shows the use of each dataset based on publications that have used respective dataset.

```
In [61]: sql=f"""                                                -- using a python f-string so that parameters AGENCY and VERSION car
         select ds.alias as dataset, p.year
         ,      count(distinct d.publication_id) as pub_per_year    -- count the different publications
           from agency_run ar                                     -- the table storing metadata about the individual runs executed by
           join dyad d on d.run_id=ar.id                          -- JOIN to the table with all the dyads
           join publication p on p.id=d.publication_id            -- JOIN to the table with publications
           join dataset_alias da on d.dataset_alias_id = da.id    -- first JOIN to the dataset_alias table with the *aliases* of the t
           join dataset_alias ds on ds.alias_id = da.parent_alias_id  -- second JOIN to retrieve the actual dataset, defined as the *parer
          where ar.agency='{AGENCY}' and ar.version='{VERSION}'   -- restriction of the agency run to the AGENCY/VERSION defined above
          group by ds.id, ds.alias, p.year                        -- we group on the dataset and the year to count distinct publicatic
          order by dataset,year
```

# APIs and Workbooks



### 3) Drilling into the details for each dataset

Who are the main authors using each agency's datasets? Who are the main authors using each specific dataset? What are the publications associated with each author? What institutions are the centers of use for each agency dataset and in what geographic locations are the institutions located?

### Who are the main authors using each agency's datasets?

[122...

```
sql=f"""
with a as (
select ds.id as ds_id, max(ds.alias) as dataset
,      a.id as author_id, CONCAT(a.given_name, CASE WHEN a.given_name is NULL THEN '' ELSE ' ' END, a.family_name) as a
,      count(distinct p.id) as num_of_publications
,      rank() over(partition by ds.id order by count(distinct p.id) desc) as rank
  from agency_run ar
  join dyad d on d.run_id=ar.id
  join publication p on p.id=d.publication_id
  join publication_author pa on pa.publication_id=p.id
  join author a on a.id=pa.author_id
  join publication_topic pt on pt.publication_id=p.id
  join topic t on t.id=pt.topic_id
  join dataset_alias da on da.id=  d.dataset_alias_id
  join dataset_alias ds on ds.alias_id = da.parent_alias_id
 where ar.agency='{AGENCY}' and ar.version='{VERSION}'
 group by ds.id,a.id,a.given_name,a.family_name
)
select * from a
where rank<=5
order by dataset,num_of_publications desc
"""
```

- Machine readable by default is key to efficiency
- Dashboards provide quick insights for executives
- APIs support additional use cases
- Jupyter Notebooks provide additional access to the data

# Data Ecosystem

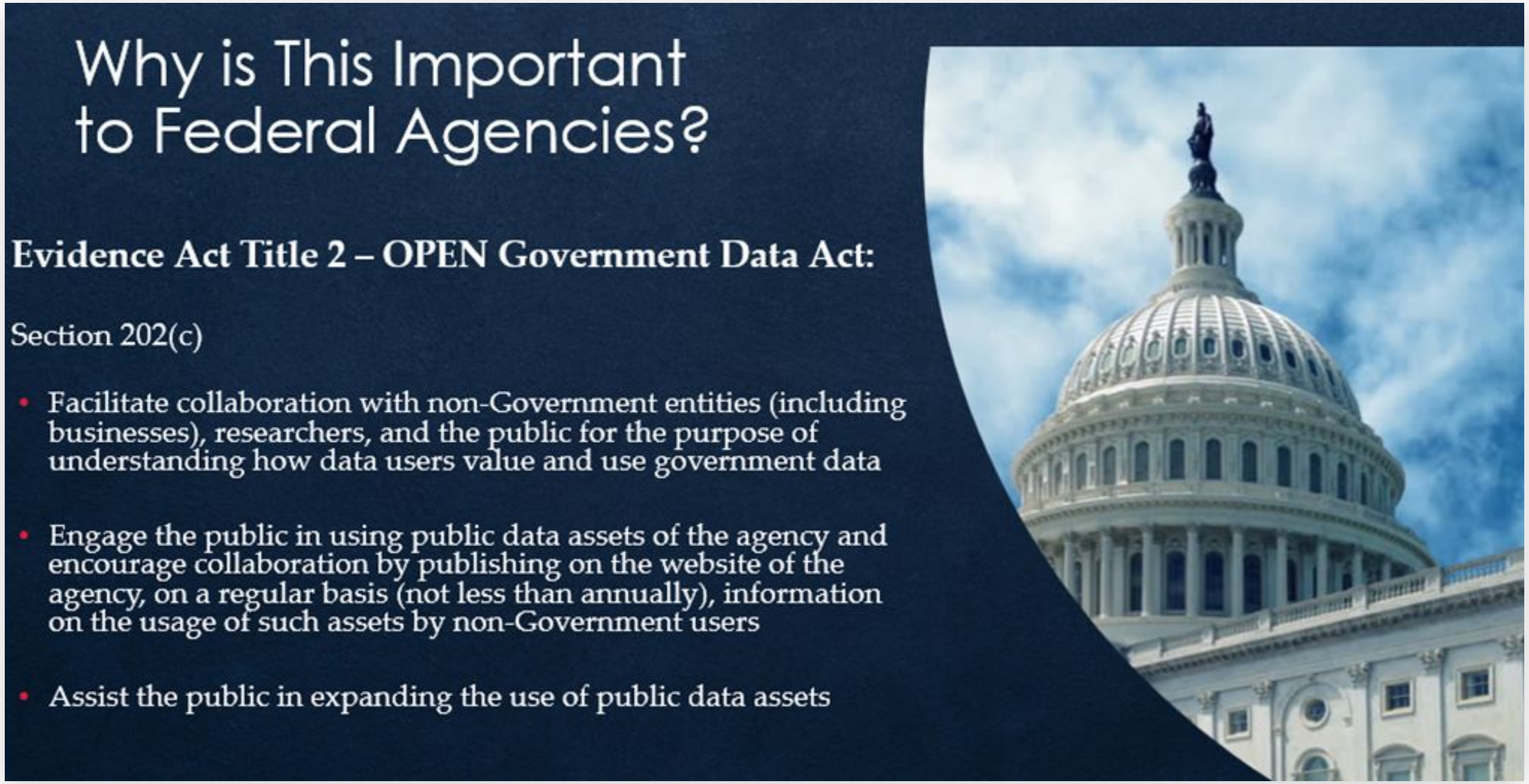
# Data Ecosystem

**Evidence Act**

- Statute
- Recommendations

Common Goals

Combined Approach



## Why is This Important to Federal Agencies?

**Evidence Act Title 2 – OPEN Government Data Act:**

Section 202(c)

- Facilitate collaboration with non-Government entities (including businesses), researchers, and the public for the purpose of understanding how data users value and use government data
- Engage the public in using public data assets of the agency and encourage collaboration by publishing on the website of the agency, on a regular basis (not less than annually), information on the usage of such assets by non-Government users
- Assist the public in expanding the use of public data assets

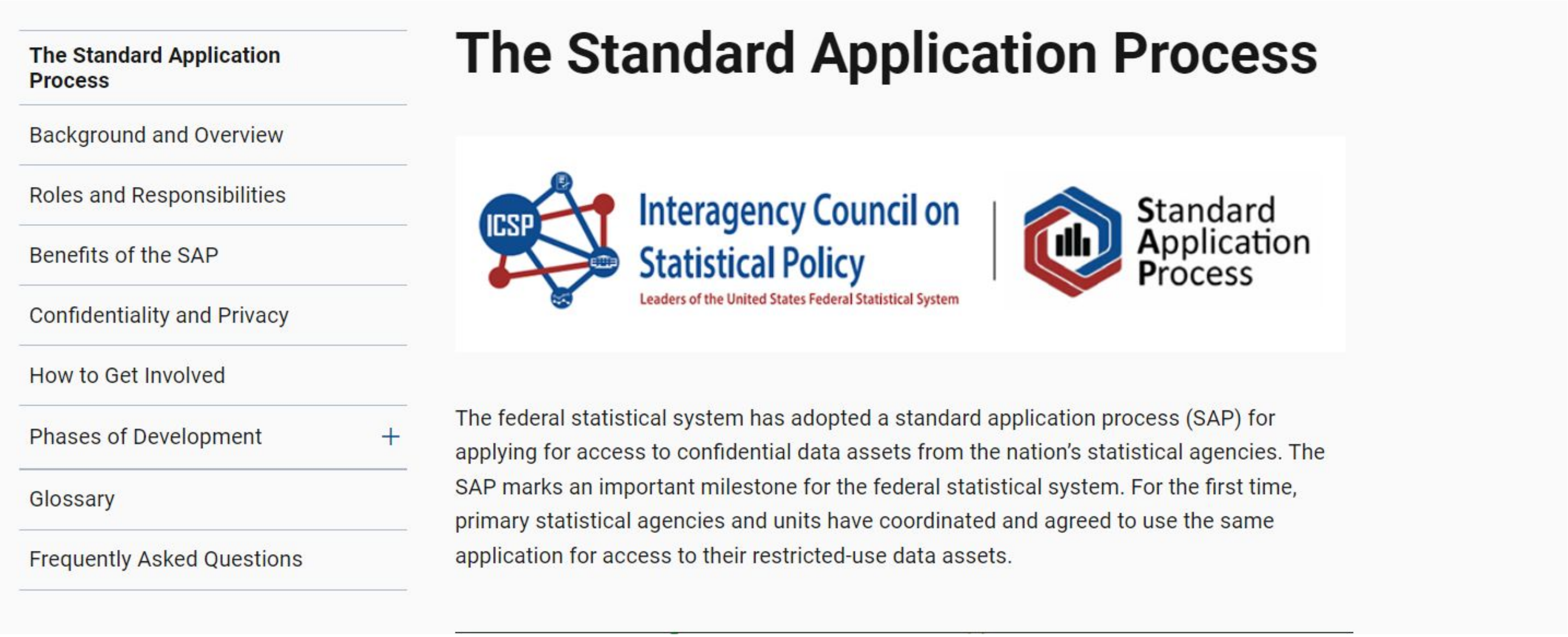
# Data Ecosystem

Evidence Act

- **Statute**
- Recommendations

Common Goals

Combined Approach



**The Standard Application Process**
Background and Overview
Roles and Responsibilities
Benefits of the SAP
Confidentiality and Privacy
How to Get Involved
Phases of Development +
Glossary
Frequently Asked Questions
ICSP
Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System
Standard Application Process

## The Standard Application Process

The federal statistical system has adopted a standard application process (SAP) for applying for access to confidential data assets from the nation's statistical agencies. The SAP marks an important milestone for the federal statistical system. For the first time, primary statistical agencies and units have coordinated and agreed to use the same application for access to their restricted-use data assets.

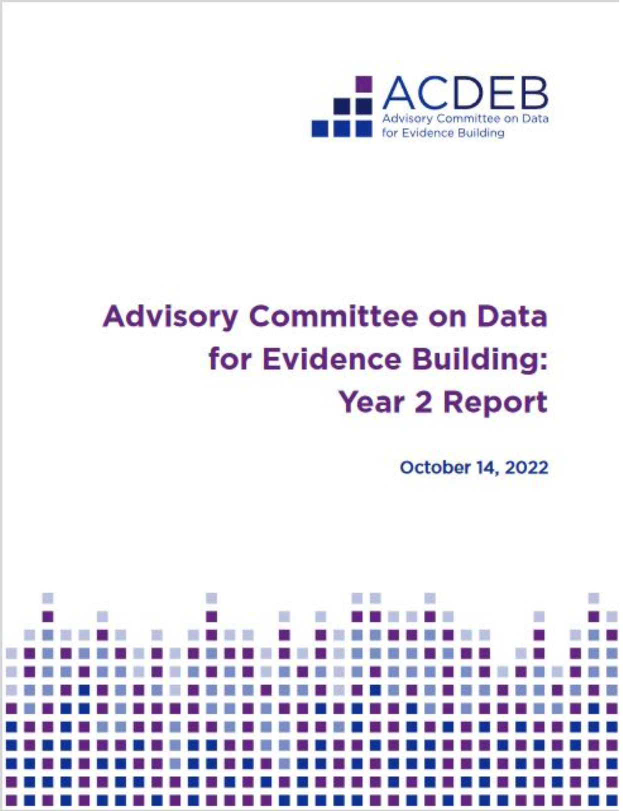
# Data Ecosystem

Evidence Act

- Statute
- **Recommendations**

Common Goals

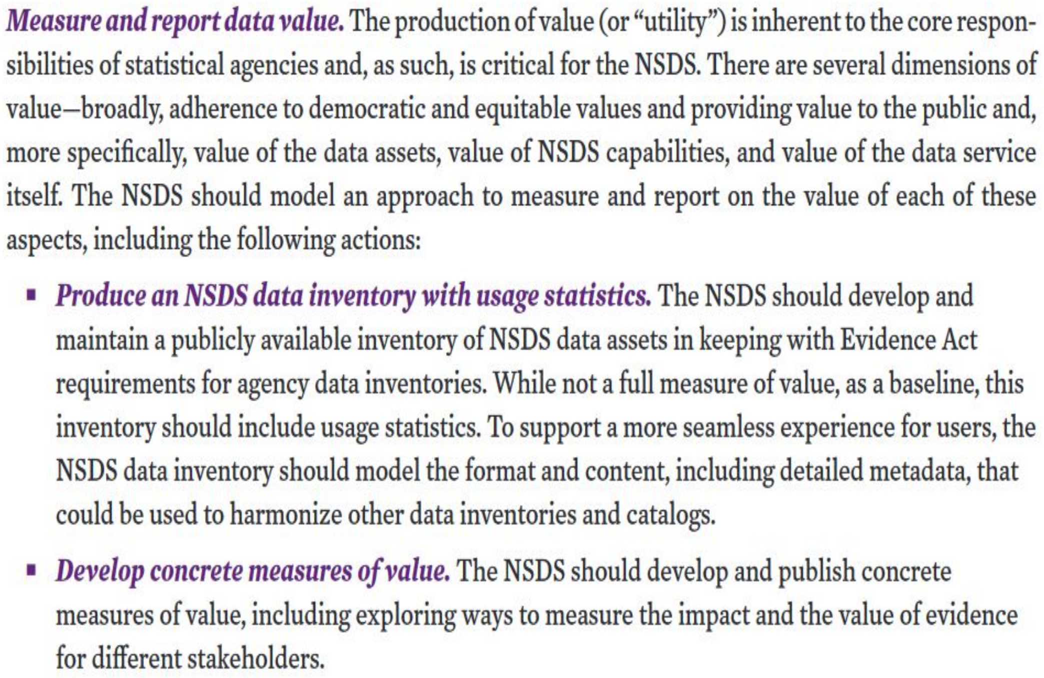
Combined Approach



ACDEB
Advisory Committee on Data for Evidence Building

**Advisory Committee on Data for Evidence Building: Year 2 Report**

October 14, 2022



***Measure and report data value.*** The production of value (or "utility") is inherent to the core responsibilities of statistical agencies and, as such, is critical for the NSDS. There are several dimensions of value—broadly, adherence to democratic and equitable values and providing value to the public and, more specifically, value of the data assets, value of NSDS capabilities, and value of the data service itself. The NSDS should model an approach to measure and report on the value of each of these aspects, including the following actions:

- ***Produce an NSDS data inventory with usage statistics.*** The NSDS should develop and maintain a publicly available inventory of NSDS data assets in keeping with Evidence Act requirements for agency data inventories. While not a full measure of value, as a baseline, this inventory should include usage statistics. To support a more seamless experience for users, the NSDS data inventory should model the format and content, including detailed metadata, that could be used to harmonize other data inventories and catalogs.
- ***Develop concrete measures of value.*** The NSDS should develop and publish concrete measures of value, including exploring ways to measure the impact and the value of evidence for different stakeholders.

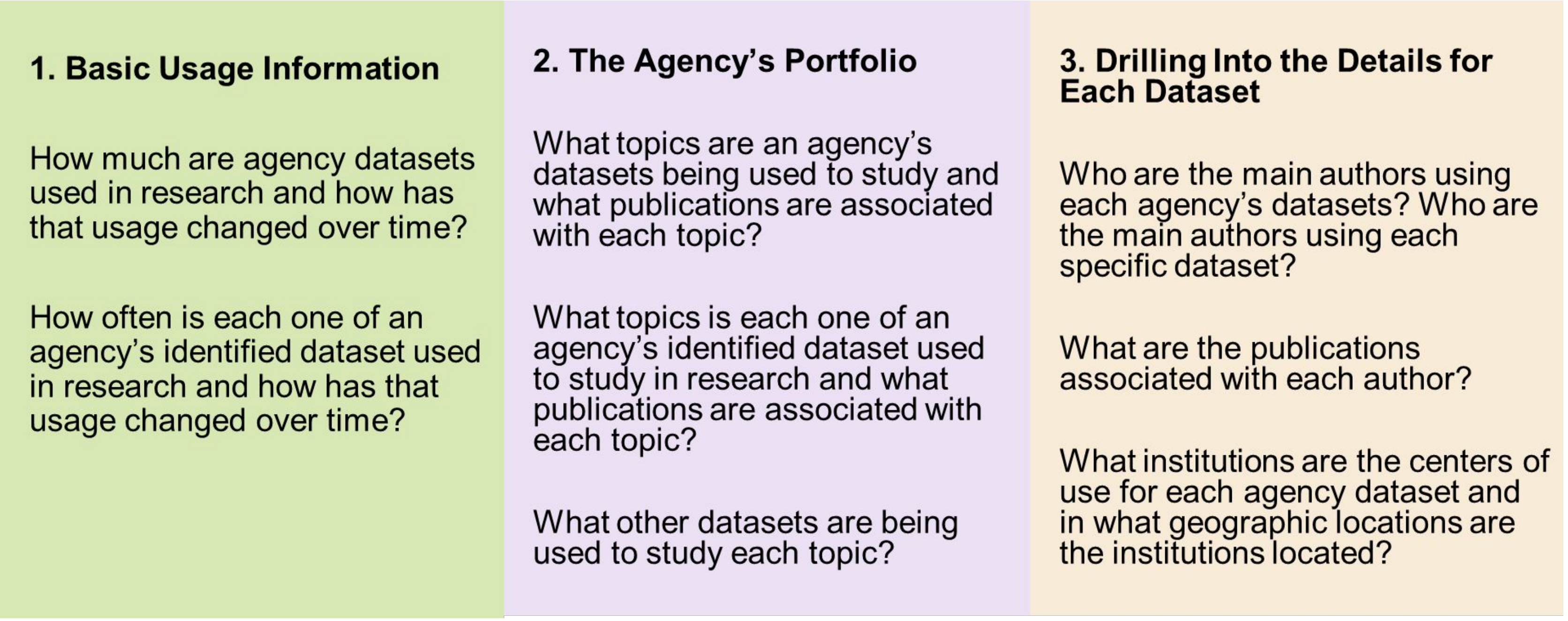
# Data Ecosystem

Evidence Act

- Statute
- Recommendations

**Common Goals**

Combined Approach



| 1. Basic Usage Information | 2. The Agency’s Portfolio | 3. Drilling Into the Details for Each Dataset |
|---|---|---|
| How much are agency datasets used in research and how has that usage changed over time? | What topics are an agency’s datasets being used to study and what publications are associated with each topic? | Who are the main authors using each agency’s datasets? Who are the main authors using each specific dataset? |
| How often is each one of an agency’s identified dataset used in research and how has that usage changed over time? | What topics is each one of an agency’s identified dataset used to study in research and what publications are associated with each topic? | What are the publications associated with each author? |
| | What other datasets are being used to study each topic? | What institutions are the centers of use for each agency dataset and in what geographic locations are the institutions located? |

# Data Ecosystem

Evidence Act

- Statute
- Recommendations

Common Goals

**Combined Approach**

Research Policy 48 (2019) 1487–1492

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Federal funding of doctoral recipients: Wh

Wan-Ying Chang[a], Wei Cheng[b], Julia Lane[c,*], Bruce

[a] National Science Foundation, United States
[b] School of Business, East China University of Science and Technology, China
[c] New York University, United States
[d] Ohio State University, United States

ARTICLE INFO

JEL classification:
O30
O38
C8
C81

Keywords:
UMETRICS
Linked survey transaction data
Doctoral workforce
Survey of earned doctorates
Research impact

ABSTRACT

This technical note descr
describe the richness and
funding channels can be
university payroll and fin
important US survey data
funding and the doctoral
disciplines and by differer
incorporate more linkages

### 3. Filling data gaps

There are now new administrative data that can be combined with the Survey of Earned Doctorates to fill the gap. The STAR METRICS project, which was initiated by federal agencies in 2009 in response to the Roadmap findings, was intended to (National Science Board, 2015) provide policymakers with a better understanding of the process of research and (Romer, 1990) provide the research community with a common data infrastructure that connected research funding with research outcomes (Lane et al., 2015). Since it was impossible to collect and link data on all individuals supported by research funding from across federal agencies, the STAR METRICS approach drew the information directly from the research organizations themselves. The key information came from administrative grant records, which contain record level information on wage payments made from federal grants to all university personnel, including doctoral recipients.

The program evolved to be led by universities (and called UMETRICS). It became institutionalized at the Institute for Research on Innovation and Science (IRIS) at the University of Michigan (Lane et al., 2014). It also included information that permitted linkages to Census data, ProQuest dissertations, US Patents, PubMed, and public in-

**Table 1**
Survey sources of federal funding.

| Source of Funding | SED[1] | SED-UMETRICS[2] | Federal SED-UMETRICS[3] |
|---|---|---|---|
| Research assistantship | 6117 | 4006 | 3410 |
| Fellowship, scholarship | 5703 | 3036 | 2522 |
| Teaching assistantship | 4745 | 2613 | 2166 |
| Grant | 2534 | 1494 | 1239 |
| Missing (did not respond) | 2584 | 1084 | 852 |
| Traineeship | 2054 | 882 | 689 |
| Spouse's, partner's, or family's earnings or savings | 1712 | 663 | 501 |
| Foreign (non-U.S.) | 1568 | 541 | 399 |
| Personal earnings during graduate school | 338 | 270 | 231 |
| Loans (from any source) | 391 | 200 | 164 |
| Personal savings | 550 | 177 | 135 |
| Employer reimbursement/ assistance | 356 | 163 | 132 |
| Other | 375 | 117 | 81 |
| Internship, clinical residency | 680 | 341 | 268 |
| Other assistantship | 5 | 2 | 1 |

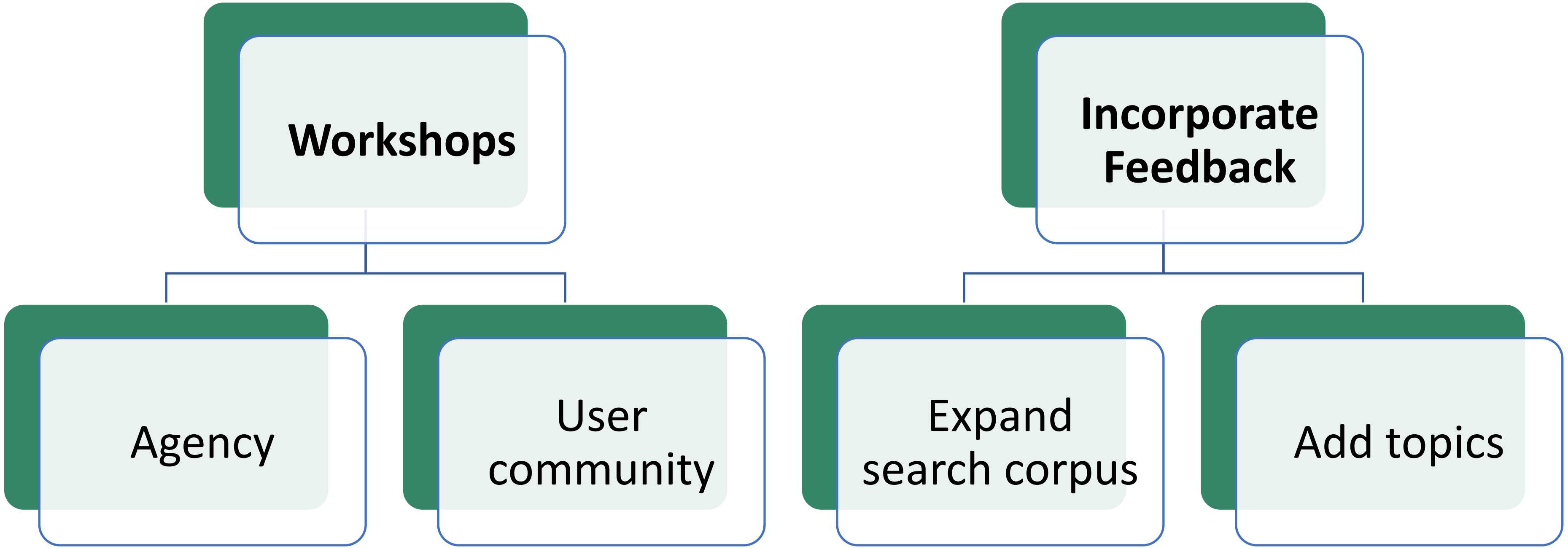
Responses to SED Question A5: Which of the following were sources of financial

Dashboard

Jupyter Notebooks

API

# Community Outreach



democratizingdata.ai

democratizingdata.ai

# Questions & Discussion

# Presenters

Spiro Stefanous, Director, ERS

Kelly Maguire, Assistant Director, ERS

Nick Pallotta, NASS

Nancy Potok, Visiting Scholar, NYU

Julia Lane, Professor, NYU