

## *Impulse statement for Open Science*

- Reaching collaboration*
- Division of labor*
- Role of open science*

Julia Lane, NYU and RTI

Joint with Texas Advanced Computing Center, Elsevier, NYU, UMD and University of Munich colleagues



**Jeroen Baas**  
Director, Research Analy  
The Randstad, Netherlan  
Experience: Elsevier



**M'hamed el Aisati**  
Vice President Research Analytic  
Amsterdam  
Experience: Elsevier



**Atti Emecz**  
Senior Director for  
Swindon  
Experience: Elsevier



**Tina Zdawczyk**  
Experienced research and  
Raleigh-Durham-Chapel H  
Experience: Elsevier, IC

# Some background



Share here: [Census.gov](#) > [Business & Industry](#) > [Center for Economic Studies](#) > [Longitudinal Employer-Household Dynamics](#)

## Longitudinal Employer-Household Dynamics

Main Applications Data Learn More Research State Partners LED in Action

patentsview.org



iris.isr.umich.edu



ABOUT MEMBERSHIP RESEARCH TRAINING NEWS & EVENTS

*IRIS is... a consortium of research universities using big administrative data to understand, explain and improve higher education and research.*

stats.govt.nz/integrated-data/integrated-data-infrastructure/



STATISTICS TOOLS SERVICES AND SUPPORT INTEGRATED DATA CENSUS WELLBEING INDICATORS ABOUT US

Home > Integrated data > Integrated Data Infrastructure

## Integrated Data Infrastructure

The Integrated Data Infrastructure (IDI) is a large research database. It holds micro data about people and...

norc.org/services-solutions/data-enclave.html

### NORC's trusted and future-ready research infrastructure provides secure access, management, and sharing of sensitive and confidential data to empower data-driven results

NORC is a recognized innovator in secure data management and sharing. The NORC Data Enclave® is an integral part of the [Solutions Center](#). The Enclave's high-performance computing environment and virtual desktop infrastructure provide convenient access to database, statistical, analytical, and reporting tools that enable evidence-based discovery.



Capabilities Collaborations News & Events About Contact

## Administrative Data Research Facility (ADRF)

MANAGEMENT PORTAL SECURE DATA HOSTING RISK CONTROLLED SAFE DATA LINKAGE FLEXIBLE ENVIRONMENT

### Overview

The Administrative Data Research Facility (ADRF) is a secure and FedRAMP-authorized computational research platform that promotes access to and the discovery of sensitive confidential microdata. The ADRF was established under guidance from the Census Bureau with funding from the Office of Management and Budget to inform the decision-making of the Commission on Evidence-Based Policy. The ADRF was the recipient of the 2018 Government Innovation Awards.

ADRF User Guide  
ADRF Product Specification  
Frequently Asked Questions  
Learn More

HOME ABOUT PEOPLE RESEARCH TRAINING NEWS & EVENTS AFFILIATED ORGANIZATIONS

Search

## DATA LITERACY & EVIDENCE BUILDING

NYU Wagner | Accenture | University Of Maryland | KYStats | Coleridge Initiative

PROGRAM OVERVIEW COURSE CONTENT APPLY HERE CLASS CALENDAR CONTACT US

HOME WEEK 0 BACKGROUND DATA MANAGEMENT WEEK 1 DATA LINKAGE WEEK 2 MEASUREMENT WEEK 3 VISUALIZATION WEEK 4 ANALYTICS WEEK 5 INFERENCE WEEK 6 BIAS & ETHICS WEEK 7

# Key lessons learned from building public data infrastructures

Successful public infrastructures need a sensible incentive structure to be both agile and long term sustainable

- collaboration (incentives);
- division of labor (comparative advantage)
- role of open science (setting open goals and funding source)

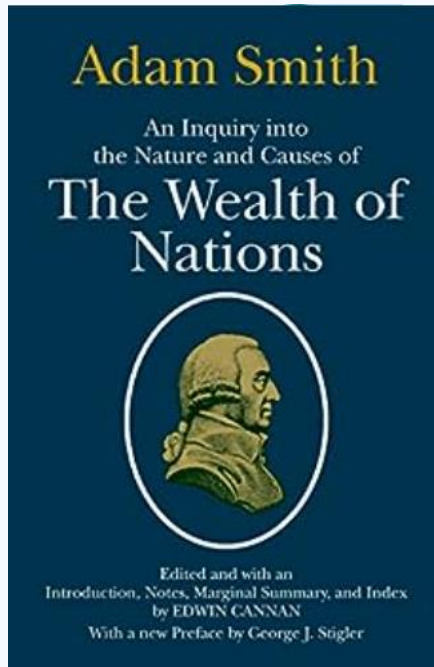
Private sector has important assets - sustainable business models, more stable and skilled workforce - and direct financial incentives

Researchers have complementary assets - cutting edge ideas, new blood, bringing community to task – and indirect financial incentives

Success comes from creating wins for both sides: Two examples with Elsevier

<https://democratizingdata.ai>

# Incentives in private sector for collaboration



"It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest." ~ Adam Smith

revenue and sales



# Incentives in public sector for collaboration

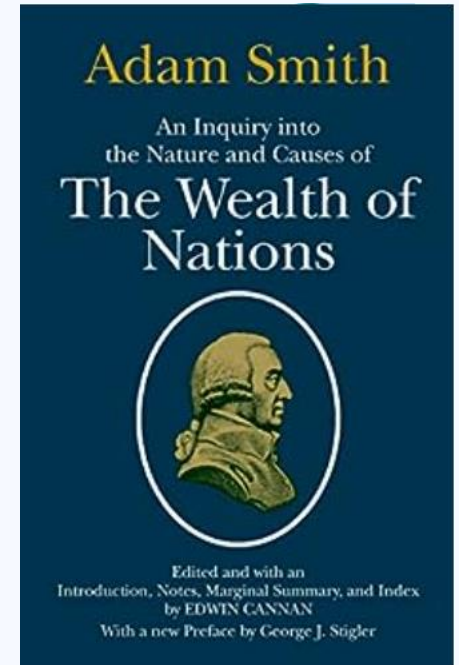
Research data is largely a public good

"It is not from the benevolence of the

scientific currency – publication and grants **re** New projects and funding **y**, the

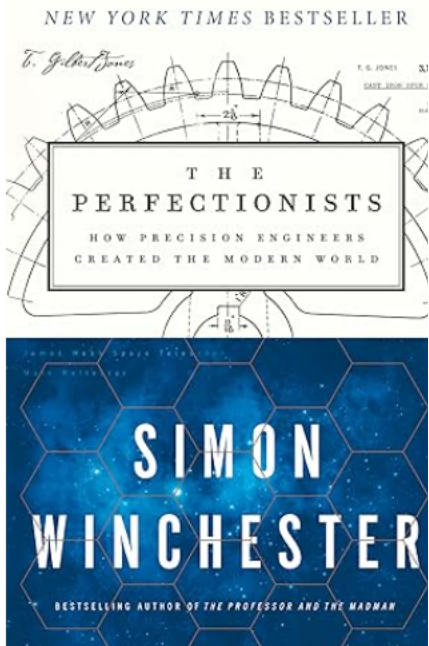
scientific currency – publication and grants **s**, or **New products and funding** t we

expect data to be produced, but from their regard to their own interest."



# Use case 1: Create an Amazon.com for data

Nati  
and



Roll over image to zoom in

Read sample

Audible sample

Follow the Author



Simon Winchester

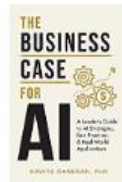
Follow

## Products related to this item

Sponsored



**The Dream Machine**  
M. Mitchell Waldrop  
*The story of J.C.R. Licklider, whose work led to the creation of internet and expanded our understanding of what computers could be.*  
★★★★☆ 237  
Kindle Edition  
\$9.99



**The Business Case for AI: A Leader's Guide to AI Strategies, Best Practices & Real-...**  
Kavita Ganesan  
*Whether you're new to AI or you've tried AI without much success, this book will equip leaders in making critical decisions in their AI journey...*  
★★★★☆ 105  
#1 Best Seller  
Kindle Edition  
\$8.99



**The Confidante: The Untold Story of the Woman Who Helped Win WWII and Shape...**  
Christopher C. Gorham  
★★★★☆ 265  
#1 Best Seller  
Kindle Edition  
\$9.52



**Beyond the Veil**  
Ronald Bagliere  
★★★★☆ 208  
Kindle Edition  
\$4.99



**Tiny Blunders/Big Disasters: Thirty-Nine Tiny Mistakes That Changed the World...**  
Jared Knott  
★★★★☆ 2,701  
#1 Best Seller  
Kindle Edition  
\$2.99



**The Life and Times of Sherlock Holmes: The Fourth Enlightening Collection of Twenty...**  
Liese Sherwood-Fabre  
*Be as smart as Sherlock! Find out what he knew that you don't.*  
★★★★☆ 10  
Kindle Edition  
\$2.99

## Editorial Reviews

### About the Author

Simon Winchester is the acclaimed author of many books, including *The Professor and the Madman*, *The Men Who United the States*, *The Map That Changed the World*, *The Man Who Loved China*, *A Crack in the World*, and *Krakatoa*, all of which were *New York Times* bestsellers and appeared on numerous best and notable lists. In 2006, Winchester was made an officer of the Order of the British Empire (OBE) by Her Majesty Queen. He resides in western Massachusetts.

--This text refers to the audioCD edition.

### From the Back Cover

The revered *New York Times* bestselling author traces the development of technology from the Industrial Age to the Digital Age to explore the single component crucial to advancement—precision—in a super both an homage and a warning for our future.

Higher Education Research and Development (HERD) Survey

Survey of Science and Engineering Research Facilities

Now

Read this item. Read on your browser, or on your phone and devices.  
Special price for \$12.99

Buy for a team

Read on other devices

READ ON ANY DEVICE

Get the free Kindle app

Book clubs  
early access

Join a Club

Learn more

# Problem: Open data without incentives don't work

Beginning in the Obama Administration, Agencies have been making datasets available for public use via Data. Gov. The Trump Administration augmented this by prioritizing data sets for AI R&D and those that support healthcare initiatives.

This has grown from a few datasets contributed by each Agency to today's status with over 300,000 datasets that are available in multiple formats, searchable, and tagged with industry protocols.

But just being available, does not mean that the data is "of value"

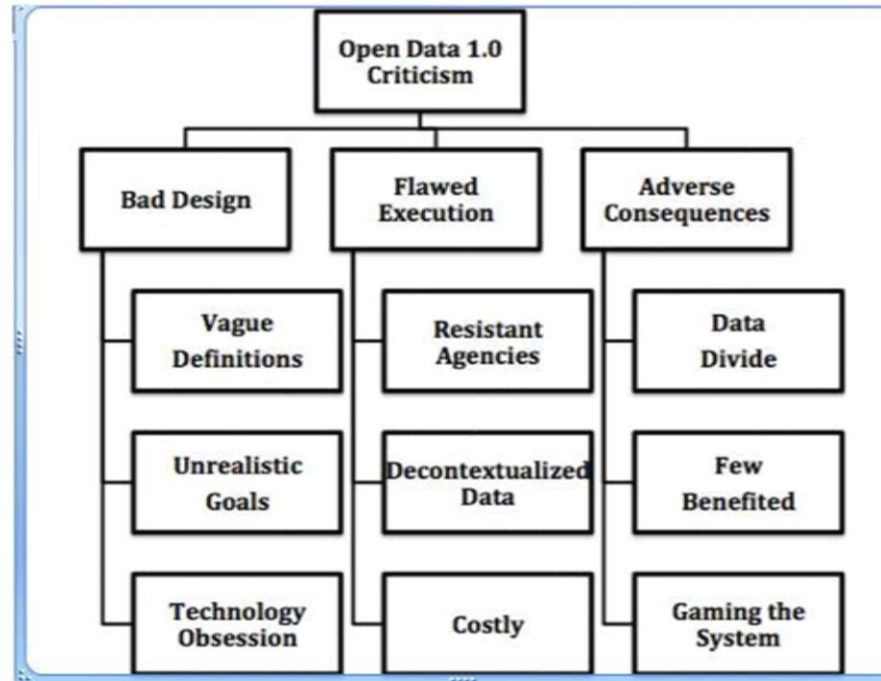
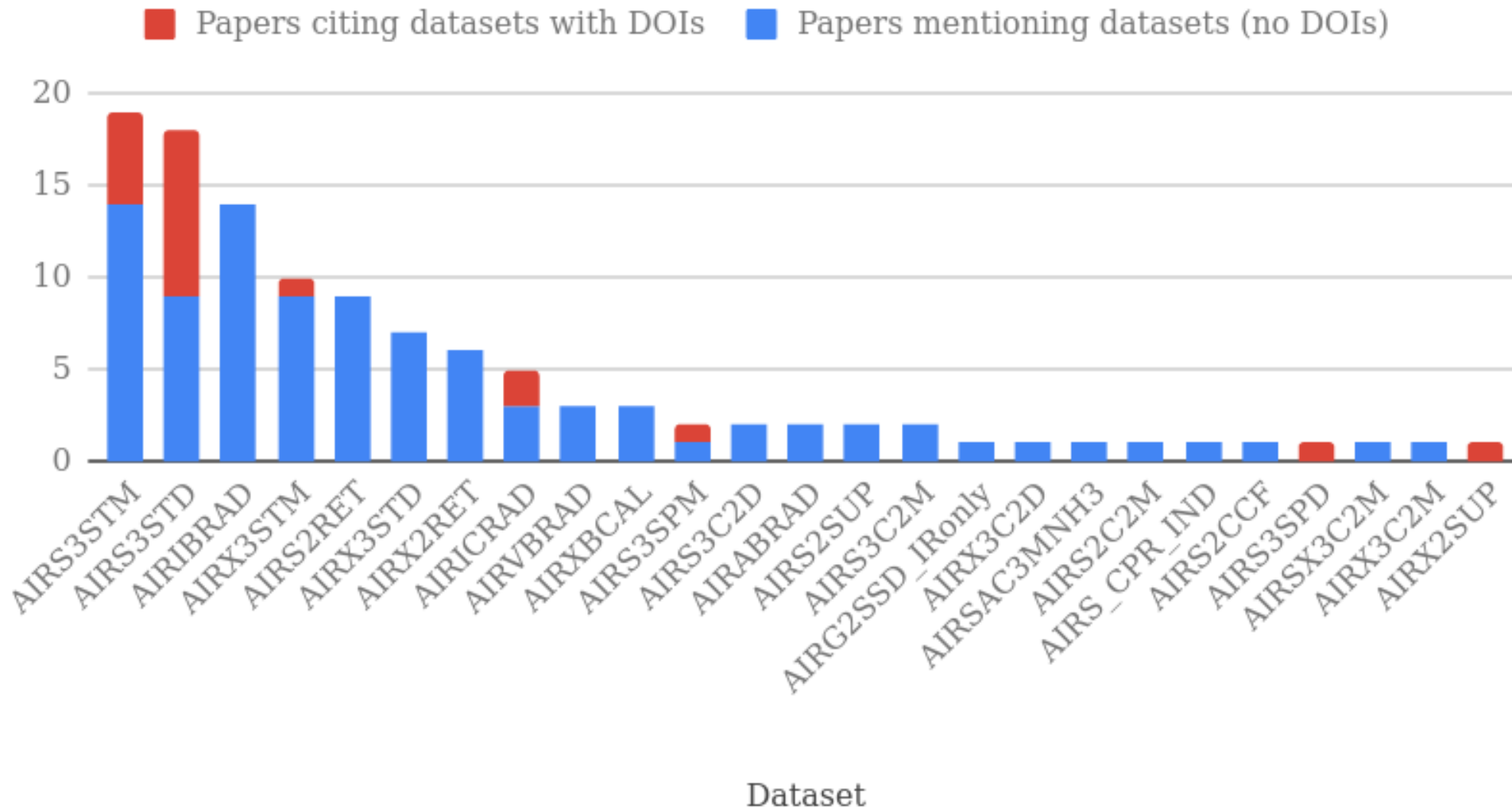


Figure 1: Open Data 1.0 Criticism

# Without incentives researchers don't contribute

## 2020 Publications using AIRS datasets



- 100 publications:
- 18 with dataset DOI citations
  - 82 manually reviewed
  - 10-15 minutes for paper review
  - ~14 hours total review time



# Collaboration and human capital

Content coverage

Historical depth

Expert curation

Selection standards

Titles on Scopus

  
7+ thousand  
publishers

26.0+ the  
active seri

243.4+ th  
book

17.5+ m  
open acce

e 1

S

S

Next Article

direct.mit.edu/qss/article/1/1/377/15571/Scopus-as-a-curated-high-quality-bibliometric-data

February 01 2020

## Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies

Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, Reza Karimi



Author and Article Information

*Quantitative Science Studies* (2020) 1 (1): 377–386.

[https://doi.org/10.1162/qss\\_a\\_00019](https://doi.org/10.1162/qss_a_00019) [Article history](#)

Cite  PDF  Permissions  Share  Views

### Abstract

Scopus is among the largest curated abstract and citation databases, with a wide global and regional coverage of scientific journals, conference proceedings, and books, while ensuring only the highest quality data are indexed through rigorous content selection and re-evaluation by an independent Content Selection and Advisory Board. Additionally, extensive quality assurance processes continuously monitor and improve all data elements in Scopus. Besides enriched metadata records of scientific articles

# Comparative advantage: leverage Scopus

## Dataset <-> Publication link



Entity extraction and  
matching  
+  
Validation



## Enhanced metadata + linking with broader datasets



**Topics:** What research cites the dataset?



**Academic citations:** How highly cited are the articles?



**Institutions:** Which were the most prominent institutions?



**Authors:** Who were the most prominent authors? How do they collaborate?



**Regions:** Which regions? How much collaboration was there across different regions?



**Societal impact:** Which policy documents cited the academic work?

# Comparative advantage: JHU, TACC and NYU

democratizingdata.ai

Outreach - Google... Bundesbank DSUD... ML Highlights - De... Dashboard - Anaco... 02\_01\_Data\_Explora... Home Page - Select...

Agencies Events Our Tools Resources

## Democratizing Data: A Search And Discovery Platform For Public Data Assets

Show how public data are being used in science so that the government can make more transparent public investments. By using automated NLP approaches we enable government agencies and researchers to quickly find the information they need.

[Learn More About Us](#)

**WHAT DOES THE PLATFORM PROVIDE**

### Promotes better use of data

The Democratizing Data project is inspired by the 2018 Foundations for [Evidence-based Policymaking Act](#). Its goal is to facilitate the collaboration between federal agencies and the public for the purpose of understanding how government data assets are used. The intent is to engage the public by providing information about the usage of the assets and expanding the use of the public data assets. As an initial step in meeting that goal, the Search and Discovery Platform describes how datasets identified by federal agencies have been used in scientific research. It uses machine learning algorithms to search over 90 million documents and find how datasets are cited, in what publications, and what topics they are used to study.

**USDA Webinar**

Watch the video and access slides from the March 28, 2023 webinar.

[Learn More](#)

CSSES

AUTHORS 4,513 COUNTRIES 76 INSTITUTIONS 1,626

Datasets: Science & Engineering Indicators, Topics: All, Year: All

**Select a Dataset to Explore Usage**

Name	Publications	Citation
Science & Engineering In...	1,695	12.3K
Women, Minorities, and...	1,522	11.7K
Survey of Doctorate Recl...	280	2.1K
Survey of Earned Doctor...	165	1.1K
Science and Engineering La...	123	82
Higher Education R&D...	78	46
National Survey of College Gra...	91	26

**Publications by geography**

**1,626 Institutions**

Institution Name	Publications	Citations
Michigan State University	17	295
University of Texas at Austin	5	181
University of California	21	173
Learning Research and Development Center, University of Pittsburgh	5	140
School of Information Management, Wuhan University	6	138
University of Michigan	10	125
University of Wisconsin-Madison	10	118
Department of Physics and Astronomy, University	6	94

CSSES

PUBLICATIONS 3,749 AUTHORS 10,124 JOURNALS 1,483 INSTITUTIONS 2,869

Datasets: All, Topics: All, Year: All

**Select a Dataset to Explore Usage**

Name	Publications	Citation
Science & Engineering In...	1,695	12.3K
Women, Minorities, and...	1,522	11.7K
Survey of Doctorate Recl...	280	2.1K
Survey of Earned Doctor...	165	1.1K
Science and Engineering La...	123	82
Higher Education R&D...	78	46
National Survey of College Gra...	91	26

**Publications Count per Year**

**Institutions Count per Year**

**Authors Count per Year**

## NCSES Dashboard: Usage At Publication Level

CSSES

PUBLICATIONS 3,749 JOURNALS 1,483

Datasets: All, Topics: All, Year: All

**Select a Dataset to Explore Usage**

Name	Publications	Citation
Science & Engineering...	1,695	12.3K
Women, Minorities, ...	1,522	11.7K
Survey of Doctorate R...	280	2.1K
Survey of Earned Doc...	165	1.1K
Science and Engineering...	123	82
Higher Education R...	78	46
National Survey of C...	91	26

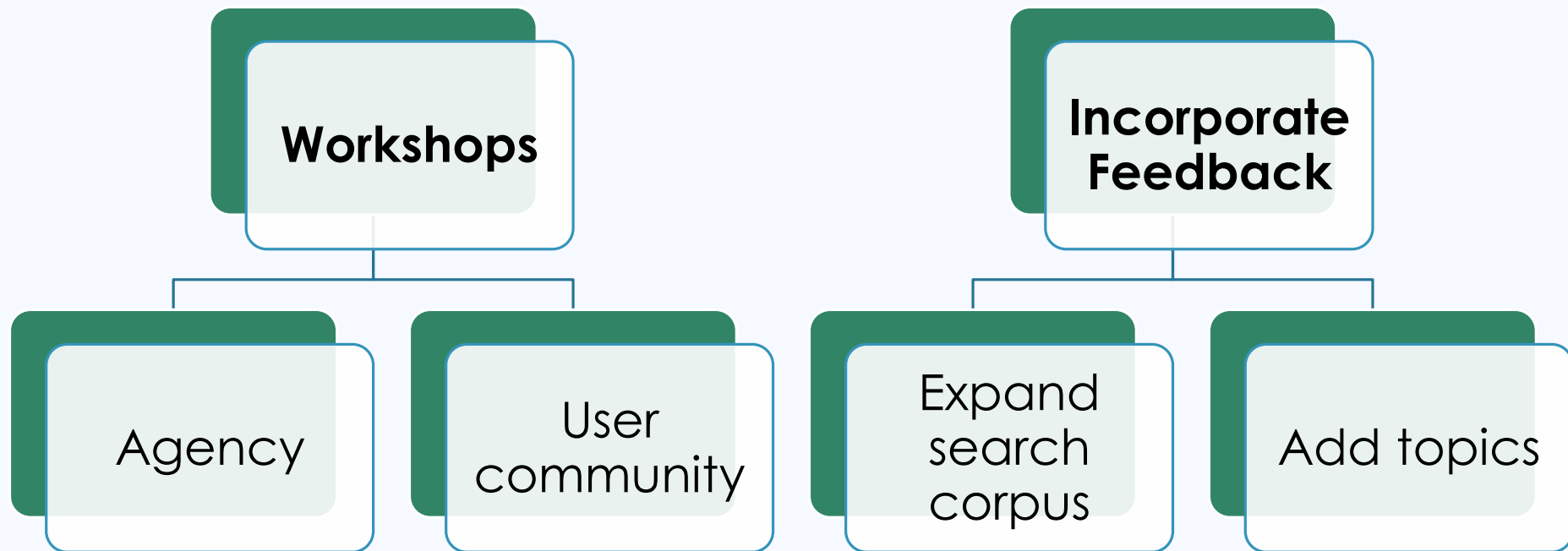
**3,749 Publications**

Publications	Citations
The Gender Equality Paradox in Science, Technology, Engineering, and Mathematics Education	299
Science audiences, misinformation, and fake news	252
Unusual effects of the COVID-19 pandemic on scientists	212
Active learning narrows achievement gaps for underserved students in undergraduate science, technology, engineering, and m...	208
Forecasting innovation: Evidence from R&D grants	205
Individuals with greater science literacy and education have more polarized beliefs on controversial science topics	174
Prioritizing diversity in human genomics research	145
Teachers' perception of STEM integration and education: a systematic literature review	142
Race and gender differences in how sense of belonging influences decisions to major in STEM	127
Provenance that Lasts: Productive: The Impact of Interdisciplinary on Scientists' Research	122
Scopus as a curated, high-quality bibliometric data source for	120

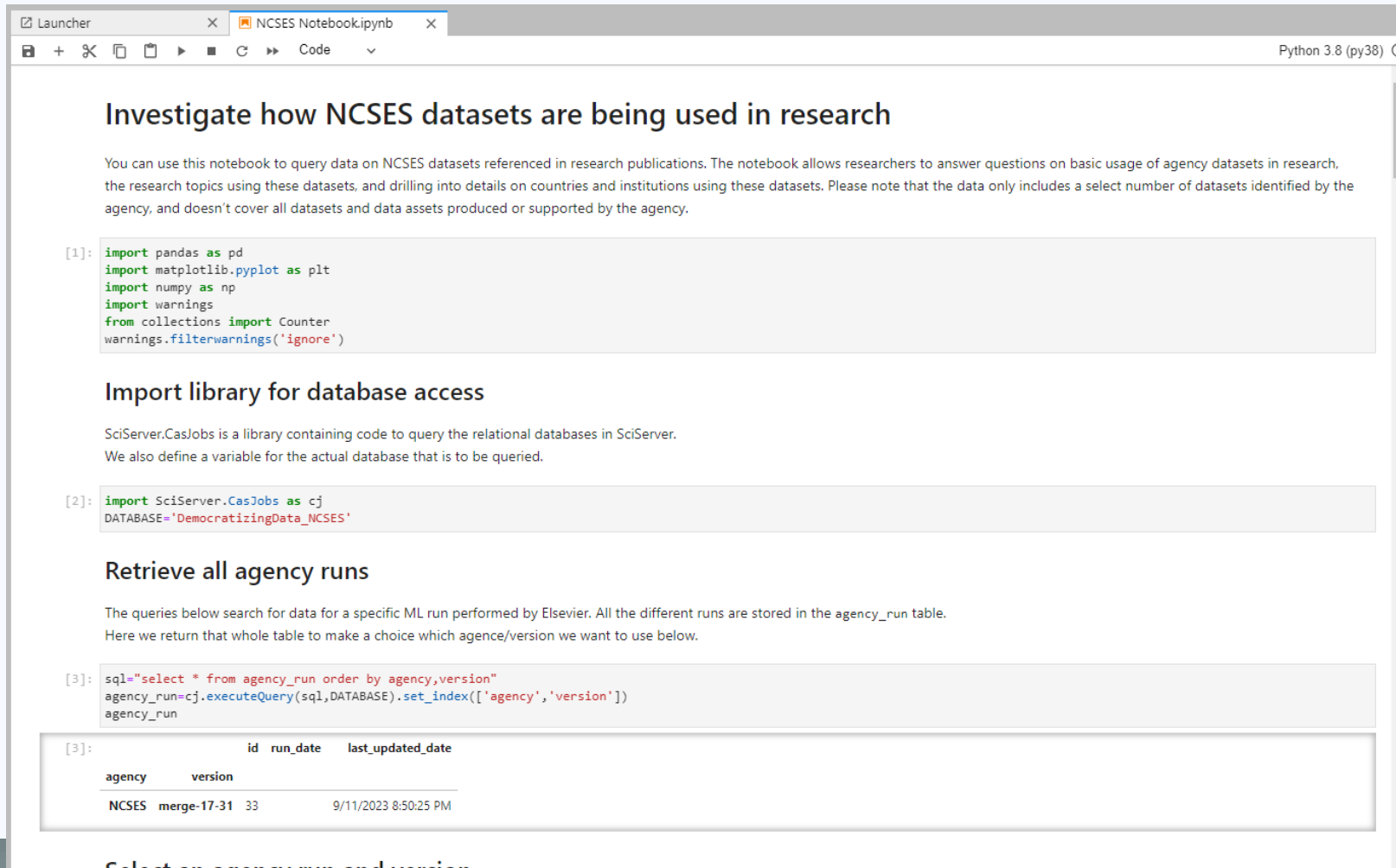
**1,483 Journals**

Journal	Publications	Citations
ASEE Annual Conference and Exposition, Conference Proceedings	267	264
SciDirect	57	678

# Comparative advantage: JHU, NYU, TACC



# Comparative advantage: JHU, NYU, TACC



The screenshot shows a Jupyter Notebook interface with the following content:

## Investigate how NCSES datasets are being used in research

You can use this notebook to query data on NCSES datasets referenced in research publications. The notebook allows researchers to answer questions on basic usage of agency datasets in research, the research topics using these datasets, and drilling into details on countries and institutions using these datasets. Please note that the data only includes a select number of datasets identified by the agency, and doesn't cover all datasets and data assets produced or supported by the agency.

```
[1]: import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import warnings
from collections import Counter
warnings.filterwarnings('ignore')
```

## Import library for database access

SciServer.CasJobs is a library containing code to query the relational databases in SciServer. We also define a variable for the actual database that is to be queried.

```
[2]: import SciServer.CasJobs as cj
DATABASE='DemocratizingData_NCSES'
```

## Retrieve all agency runs

The queries below search for data for a specific ML run performed by Elsevier. All the different runs are stored in the `agency_run` table. Here we return that whole table to make a choice which agency/version we want to use below.

```
[3]: sql="select * from agency_run order by agency,version"
agency_run=cj.executeQuery(sql,DATABASE).set_index(['agency','version'])
agency_run
```

agency	version	id	run_date	last_updated_date
NCSES	merge-17-31	33		9/11/2023 8:50:25 PM

Selection on agency and version

Source: SciServer

Us

resources.data.gov/resources/podm-field-mapping/

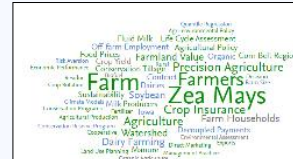


Dataset Fields						
title	Title	required	No	title	<code>metadata&gt;idinfo&gt;citation&gt;citeinfo&gt;title</code>	
description	Description	required	No	abstract	<code>metadata&gt;idinfo&gt;description&gt;abstract</code>	
keyword	Tags	required	Yes	keywords (theme, place, stratum, temporal) add to:Keyword (any type) = <code>geospatial</code>	<code>metadata&gt;idinfo&gt;keywords&gt;theme&gt;themekey</code> & <code>metadata&gt;idinfo&gt;keywords&gt;place&gt;placekeymetadata&gt;idinfo&gt;keywords&gt;stratum&gt;stratkeymetadata&gt;idinfo&gt;keywords&gt;temporal&gt;tempkey</code>	
modified	Last Update	required	No	publication date	<code>metadata&gt;idinfo&gt;citation&gt;citeinfo&gt;pubdate</code> if non-date value, e.g. 'unknown' 'unpublished' then <code>metadata&gt;metainfo&gt;metd</code>	

# Comparative advantage: EI and community labelling

Dataset <-> Publication link

Enhanced metadata + linking with broader datasets



**Topics:** What research cites the dataset?



**Citations:** How highly cited are

**Institutions:** Which were the most prominent institutions?

**Collaborators:** Who were the most prominent collaborators? How do they collaborate?

**Regions:** How much research was there across different

regions?



**Societal impact:** Which policy documents cited the academic work?

## Role of Open Science in each

- Push funding agencies and philanthropic foundations to establish programs
- Push for work with community and human computer interaction
- Use common sense in developing shared business models



# Incentives in public sector


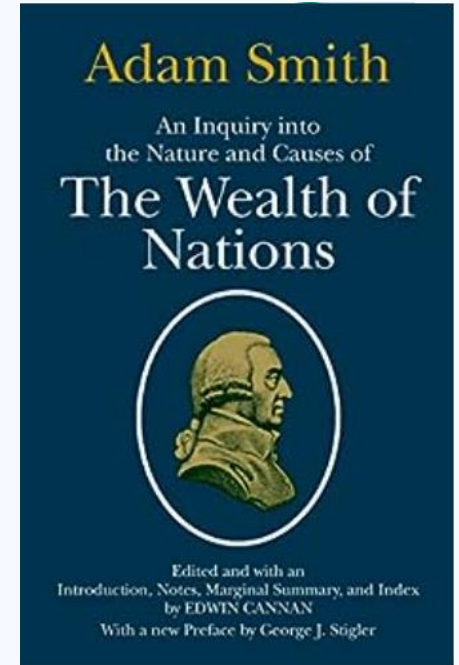
Research data is largely a public good

"It is not from the benevolence of the


scientific currency – publication and grants **re** New projects and funding **y**, the

scientific currency – publication and grants **s**, or **New products and funding** t we


expect data to be produced, but from their regard to their own interest."




**Jeroen Baas**  
Director, Research Analy  
The Randstad, Netherlan  
Experience: Elsevier



**M'hamed el Aisati**  
Vice President Research Analytic  
Amsterdam  
Experience: Elsevier



**Atti Emecz**  
Senior Director for  
Swindon  
Experience: Elsevier



**Tina Zdawczyk**  
Experienced research and  
Raleigh-Durham-Chapel H  
Experience: Elsevier

# Questions?

Julia Lane

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

<https://www.linkedin.com/in/julia-lane/>