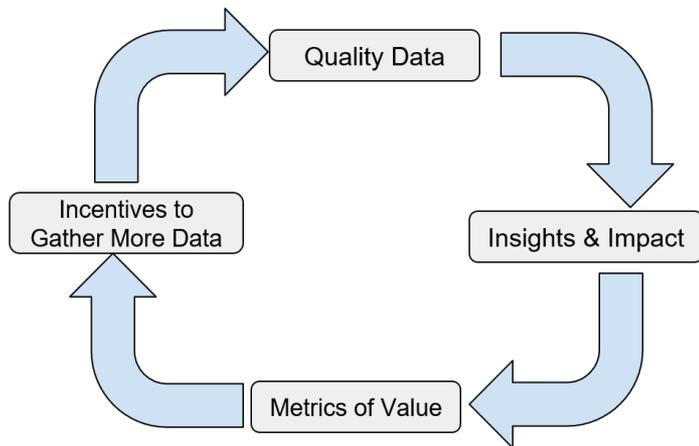


An Invisible Hand for Creating Public Value from Data¹

Julia Lane, Alfred Spector, and Michael Stebbins



*Presented by Julia Lane (NYU)
& Alfred Spector (MIT)
for the
Democratizing Data: Discovering
Data Use and Value for
Research and Policy Podcast*

21 May 2024, 12:50 PM EDT

¹Harvard Data Science Review, April 2024, (Special Issue 4). <https://doi.org/10.1162/99608f92.03719804>

Premise

- Public data can benefit government operations and society
- Much data collected, but there are many challenges

- Funding
- Prioritization
- Packaging and Distribution

Within Government

- Engagement
- Knowledge of usage
- Demonstration of Value

Government ↔ Usage Community

- If challenges addressed, government & society could benefit as much from data as has industry
- *We propose an invisible hand incentive structure.*

¹Term usage became significant ~2010

Premise

- Public data can benefit government and society
- Much data collected, but there are challenges

- Funding
- Prioritization
- Packaging and Distribution

Within a

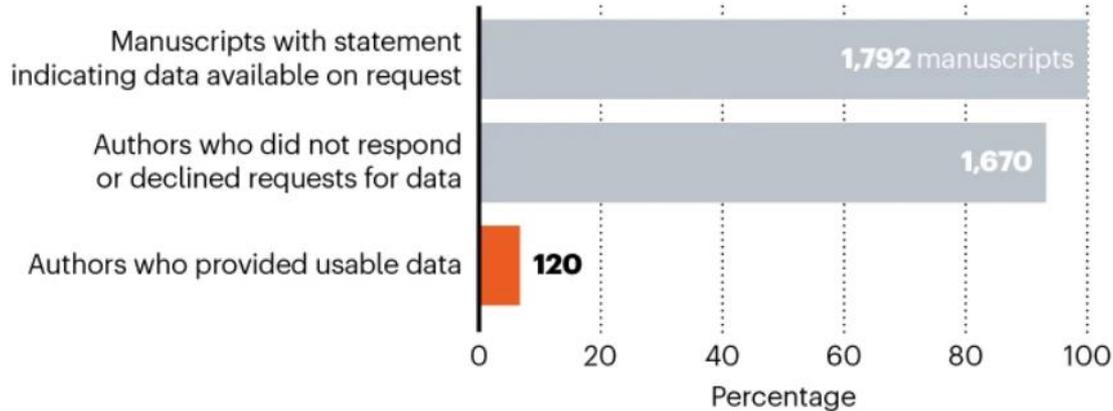
- If only government & society could benefit as much as industry
- We need to focus on realigning incentives; e.g., data citations
- “Everyone wants to do the model work, not the data work” (Nithya Simbasivan)
- Misaligned incentives in academia typically greater than in business
- We should focus on realigning incentives; e.g., data citations
- We should focus on realigning incentives; e.g., data citations

¹Term usage became significant ~2010

Example of Problem

DATA-SHARING BEHAVIOUR

Of almost 1,800 manuscripts for which the authors stated they were willing to share their data, more than 90% of corresponding authors either declined or did not respond to requests for data. Only about 7% of authors actually handed over data.



©nature

Source: <https://www.nature.com/articles/d41586-022-01692-1>

Source: Livia Puljak et al



Restaurants

greenwich village



Yelp for Business

Write a Review

Log In

Sign Up

All

Price

Open Now

Reservations

Offers Online Waitlist

Offers Delivery

Offers Takeout



1. Da Andrea - Greenwich Village

★★★★☆ 4.3 (2.3k reviews)

Italian Mediterranean Breakfast & Brunch \$\$\$ • Greenwich Village

Opens in 17 min

“When we first went in, my kids were whining about how they wanted McDonald's instead of this place. We were looking around the area, and found Da Andrea. We...” [more](#)

✓ Outdoor seating ✓ Delivery ✓ Takeout

Find a Table

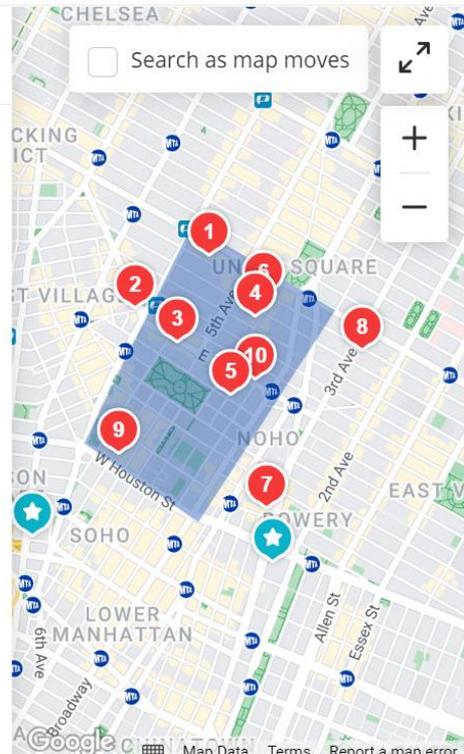


2. Olio e Più

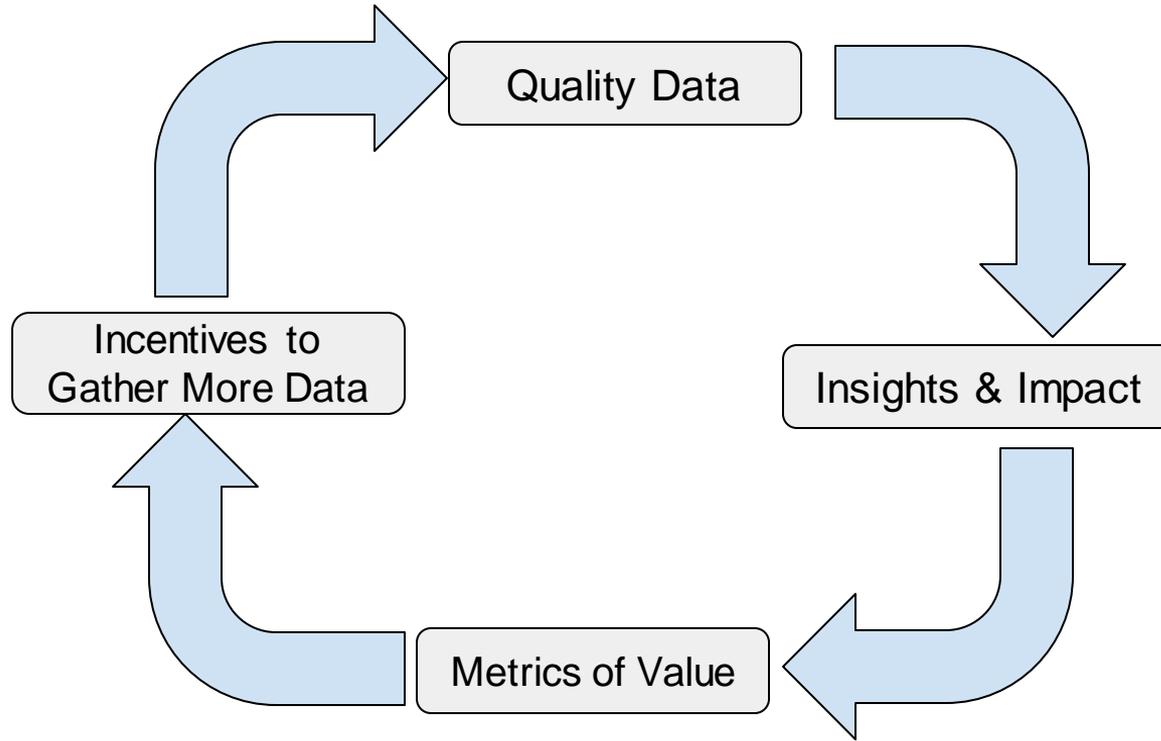
★★★★☆ 4.5 (5.5k reviews)

Pizza Italian Cocktail Bars \$\$\$ • West Village

Open until Midnight



Virtuous Cycle: Public Value from Data



Misaligned incentives?

Institution



“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 self-interest. We address ourselves not to their humanity but to their self-love, and never talk to them of our own necessities, but of their advantages”

Agency

Institution

Researcher

– Adam Smith, *An Inquiry into the Nature & Causes of the Wealth of Nations*, Vol 1

Source: <https://www.goodreads.com/quotes/68664-it-is-not-from-the-benevolence-of-the-butcher-the>

Incentives for agencies



NCSES: Democratizing



About the dashboard

This dashboard provides an initial view of data usage in the research communities to answer following questions about NCSES's data assets: (1) who are the data experts on particular topics, (2) how have the data assets been used, and (3) how have the data been combined and with what results.

Sort filters by

- Publications
- Citations

Select Classification and Research Areas

Research Subfields (Science-Matrix)

Word Cloud

Select a subject from the word cloud. The bigger a word or phrase appears, the more publications on that subject.

Select a Data asset | Table View

List of data assets ranked by use in Publications

Science & Engineering Indicators	2,329
Women Minorities and Persons with Disa...	2,282
Survey of Doctorate Recipients	404

Search (Alt)

Select a Year

Evolution by number of Publications

698	742	794	826	878	624	545
2017	2018	2019	2020	2021	2022	2023

View on Tableau Public

About the dashboard

This dashboard provides an initial view of data usage in the research communities to answer following questions about NCSES's data assets: (1) who are the data experts on particular topics, (2) how have the data assets been used, and (3) how have the data been combined and with what results.

Sort filters by

- Publications
- Citations

Select Classification and Research Areas

Research Subfields (Science-Matrix)

Word Cloud

Select a subject from the word cloud. The bigger a word or phrase appears, the more publications on that subject.

Select a Data asset | Table View

List of data assets ranked by use in Publications.

Science & Engineering Indicators	2,329
Women Minorities and Persons with Disa...	2,282
Survey of Doctorate Recipients	404

Search (Alt)

Select a Year

Evolution by number of Publications

34	39	37	44	41	33	26
2017	2018	2019	2020	2021	2022	2023

View on Tableau Public

Stats at a glance

PUBLICATIONS 253 | JOURNALS 171 | INSTITUTIONS 362 | AUTHORS 740

Select a view: Publications & Journals | Regions, Institutions & Authors | Usage Over Time

CLEAR FILTERS

List of publications in which the data asset is cited

Data asset: Survey of Earned Doctorates, Subject All, Year All

253 Publications

Publication	Citations
No progress on diversity in 40 years	211
Prominent but Less Productive: The Impact of Interdisciplinarity on Scientists' Research	205
Representation and Salary Gaps by Race-Ethnicity and Gender at Selective Public Universities	100
NIH funding longevity by gender	88
Women are credited less in science than men	83

DOWNLOAD THE FULL LIST (Coming soon)

171 Journals

Journal	Publications	Citations
ASEE Annual Conference and Exposition Conference Proceedings	9	17
International Journal of Doctoral Studies	9	45
PLoS ONE	8	131
Studies in Graduate and Postdoctoral Education	7	24
Frontiers in Psychology	6	43

DOWNLOAD THE FULL LIST (Coming soon)

About the Democratizing Data Initiative pilot project

This two-year pilot effort provides the results of new AI-based data search and discovery tools that have been developed to assist the federal data-producing communities to better understand how their data investments are being used. NCSES has piloted this Dashboard as part of its response to the Evidence Act.

Further reading:
 The Evidence Act <https://www.congress.gov/bills/115/8/congress-public-law-115-8> | The Democratizing Data Landing Page <https://democratizingdata.ai/venues/ncses-landing/> | Issue 4.2, Spring 2022, Data Inventories for the Modern Age? Using Data Science to Open Government Data https://doi.org/10.1162/929008922_8349233 | Data source: Scopus, abstract and citation database, accessed November 2023 <https://www.elsevier.com/locate/scopus>
 For details on data sources, the process used to generate data, and the classification systems used, see the [user guide](#). (Coming soon)

View on Tableau Public

Incentives for researchers



About the dashboard

This dashboard provides an initial view of data usage in the research communities to answer following questions about NCSES data assets: (1) who are the data experts on particular topics, (2) how have the data assets been used, and (3) how have the data been combined and with what results.

Sort filters by

Citations

Select Classification and Research Areas

Research Subfields (Science-Matrix)

Word Cloud

Select a subject from the word cloud. The bigger a word or phrase appears, the more publications on that subject.

Select a Data asset | Table View

List of data assets ranked by use in Citations.

Science & Engineering Indicators	26,148
Women Minorities and Persons with Disa...	35,611
Survey of Doctorate Recipients	4,358

Search: [All]

Select a Year

Evolution by number of Citations



7,540 6,256 6,115 3,684 3,850 2,882 2,083

Lane and Spector



About the dashboard

This dashboard provides an initial view of data usage in the research communities to answer following questions about NCSES data assets: (1) who are the data experts on particular topics, (2) how have the data assets been used, and (3) how have the data been combined and with what results.

Sort filters by

Publications

Citations

Select Classification and Research Areas

Research Subfields (Science-Matrix)

Word Cloud

Select a subject from the word cloud. The bigger a word or phrase appears, the more publications on that subject.

Select a Data asset | Table View

List of data assets ranked by use in Publications.

Science & Engineering Indicators	2,321
Women Minorities and Persons with Disa...	2,281
Survey of Doctorate Recipients	401

Search: [All]

Select a Year

Evolution by number of Publications

361	383	395	393	377	222	198
2017	2018	2019	2020	2021	2022	2023

Stats at a glance

PUBLICATIONS **2,329** JOURNALS **1,109** INSTITUTIONS **2,045** AUTHORS **6,115**

Select a view

Publications & Journals | Regions, Institutions & Authors | Usage Over Time

CLEAR FILTERS

List of publications in which the data asset is cited

Data asset: Science & Engineering Indicators, Subject: All, Year: All

2,329 Publications

Publication	Citations
Why are some STEM fields more gender balanced than others?	608
Science audiences misinformation and	424
Financing innovation: Evidence from R	350
Cultural Evolution: People's Motivation	314
Teachers' perception of STEM integration and education: a systematic literature review	312

DOWNLOAD THE FULL LIST (Coming soon)

1,109 Journals

Journal	Publications	Citations
ASEE Annual Conference and Exposition Conference Proceedings	147	167
PLoS ONE	40	523
Public Understanding of Science	38	689
Scientometrics	27	359
Proceedings - Frontiers in Education Conference FIE	23	79

DOWNLOAD THE FULL LIST (Coming soon)

About the Democratizing Data Initiative pilot project

This two-year pilot effort provides the results of new AI-based data search and discovery tools that have been developed to assist the federal data-producing communities to better understand how their data investments are being used. NCSES has piloted this Dashboard as part of its response to the Evidence Act.

Further reading:

The Evidence Act <https://www.congress.gov/bills/115th/congress-house-bill/4174/text> | The Democratizing Data Landing Page <https://democratizingdata.ai/agencies/ncses-landing/> | Issue 4.2, Spring 2022; Data Inventories for the Modern Age? Using Data Science to Open Government Data https://doi.org/10.1162/99608f92_8a3f2336 | Data source: Scopus, abstract and citation database, accessed November 2023 <https://www.elsevier.com/products/scopus> | For details on data sources, the process used to generate data, and the classification systems used, see the user guide. (Coming soon)

When a measure becomes a target...

Categorizing Variants of Goodhart's Law

David Manheim Scott Garrabrant
davidmanheim@gmail.com scott@intelligence.org

February 26, 2019

There are several distinct failure modes for overoptimization of systems on the basis of metrics. This occurs when a metric which can be used to improve a system is used to such an extent that further optimization is ineffective or harmful, and is sometimes termed Goodhart's Law¹. This class of failure is often poorly understood, partly because terminology for discussing them is ambiguous, and partly because discussion using this ambiguous terminology ignores distinctions between different failure modes of this general type.

This paper expands on an earlier discussion by Garrabrant [2], which notes there are “(at least) four different mechanisms” that relate to Goodhart's Law. This paper is intended to explore these mechanisms further, and specify more clearly how they occur. This discussion should be helpful in better understanding these types of failures in economic regulation, in public policy, in machine learning, and in artificial intelligence alignment[4]. The importance of Goodhart effects depends on the amount of power directed towards optimizing the proxy, and so the increased optimization power offered by artificial intelligence makes it especially critical for that field.

Challenges

Quality data – A culture of rigor

Ease of use – Significant work required

Privacy/Security – Trade-offs significant

Goodhart's law (paraphrased): *You get what you measure*

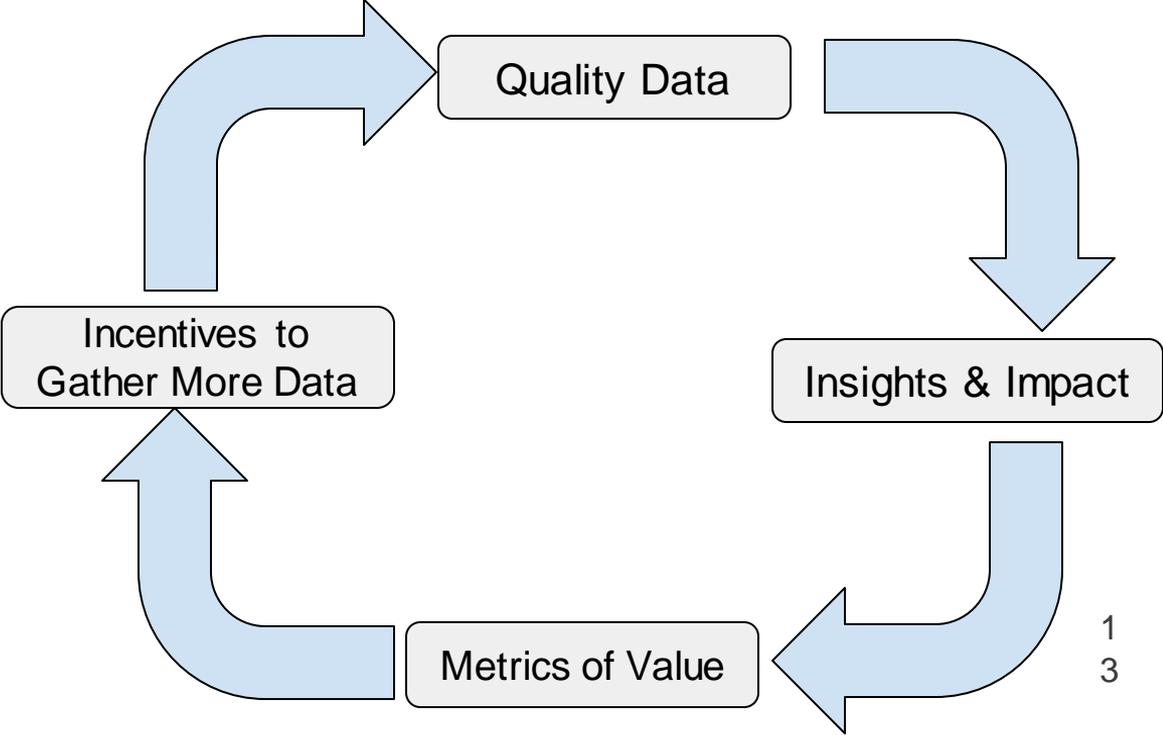
McNamara fallacy: In the public sphere

Objectives:

More complex than in many corporate settings

Outcomes take longer to evaluate

Summary – Incentives and Action



Potential enormous benefit:

- You've heard from philanthropists and stat agencies
- You will hear from many more groups tackling these issues

Harvest and refine these ideas to apply to your agency & research agenda!