Creating Incentives for data sharing and transparency

15 May 2024

Adam Thomas

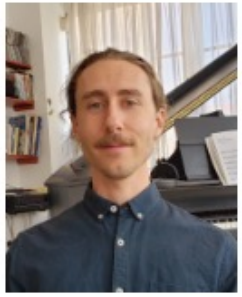
Intramural Research Program
National Institute of Mental Health
Bethesda, Maryland, USA
http://cmn.nimh.nih.gov/dsst

Thank You!

Tanenbaum Open Science Institute (TOSI)

Guy Rouleau
Director, The Neuro
Co-Founder of the Tanenbaum Open Science Institute

Annabel Seyller
Chief of Staff, The Neuro
CEO, Tanenbaum Open Science Institute



Gabriel Pelletier
Open Science Data Manager
Interim Open Science Alliance Officer

Luisa Pimentel
Open Science Community Officer

neuro

Jean-Baptiste Poline

Outline

- Background (Why?)
- Proposal (What?)
- Questions/feedback (How?)

Recent Progress in Data Sharing

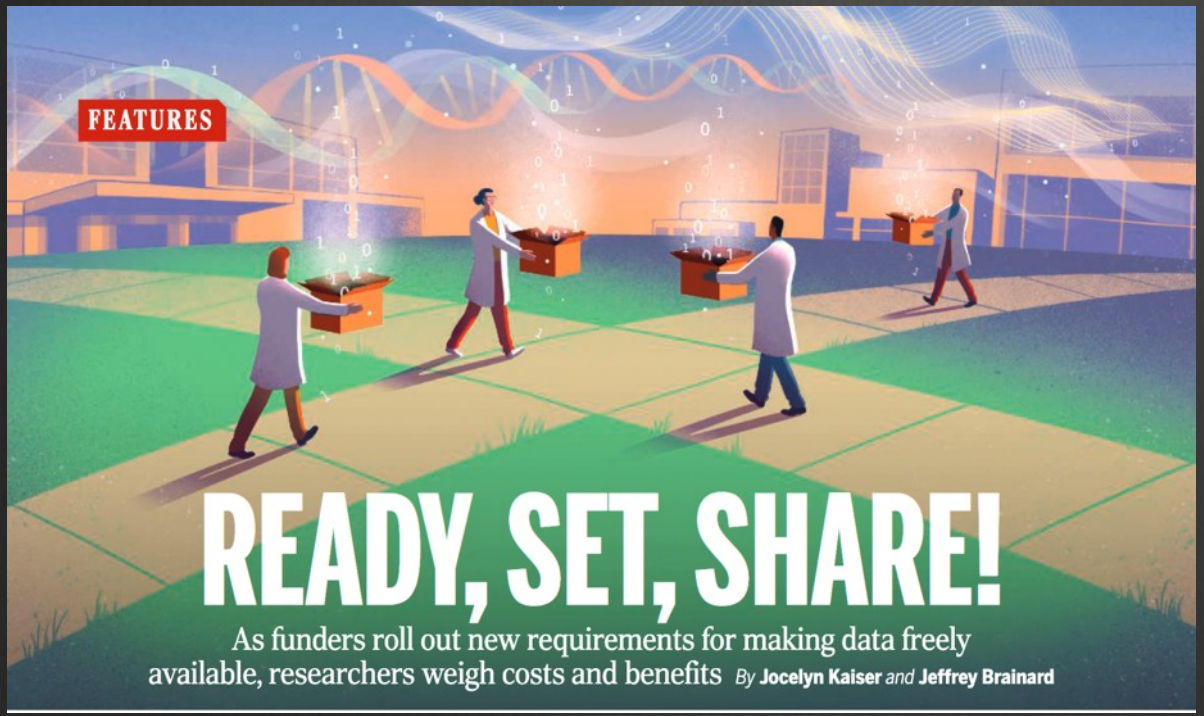
NEWS | 16 February 2022 | Correction 16 February 2022

nature

NIH issues a seismic mandate: share data publicly

The data-sharing policy could set a global standard for biomedical research, scientists say, but they have questions about logistics and equity.

Nature, Feb 2022 DOI:10.1038/d41586-022-00402-1



Science, Jan 2023 DOI:10.1126/science.adg8142

Recent Progress in Data Sharing

NEWS | 16 February 2022 | Correction 16 February 2022

nature

NIH issues a seismic mandate: share data publicly

The data-sharing policy could set a global standard for biomedical research, scientists say, but they have questions about logistics and equity.

Nature, Feb 2022 DOI:10.1038/d41586-022-00402-1

Data Sharing is the new norm

Science, Jan 2023 DOI:10.1126/science.adg8142

Recent Progress in Data Sharing

NEWS | 16 February 2022 | Correction 16 February 2022

nature

NIH issues a seismic mandate: share data publicly

The data-sharing policy could set a global standard for biomedical research, scientists say, but they have questions about logistics and equity.

Nature, Feb 2022 DOI:10.1038/d41586-022-00402-1

Data Sharing is the new norm, or is it?

Science, Jan 2023 DOI:10.1126/science.adg8142

History of Data Sharing in the US

FINAL NIH STATEMENT ON SHARING RESEARCH DATA

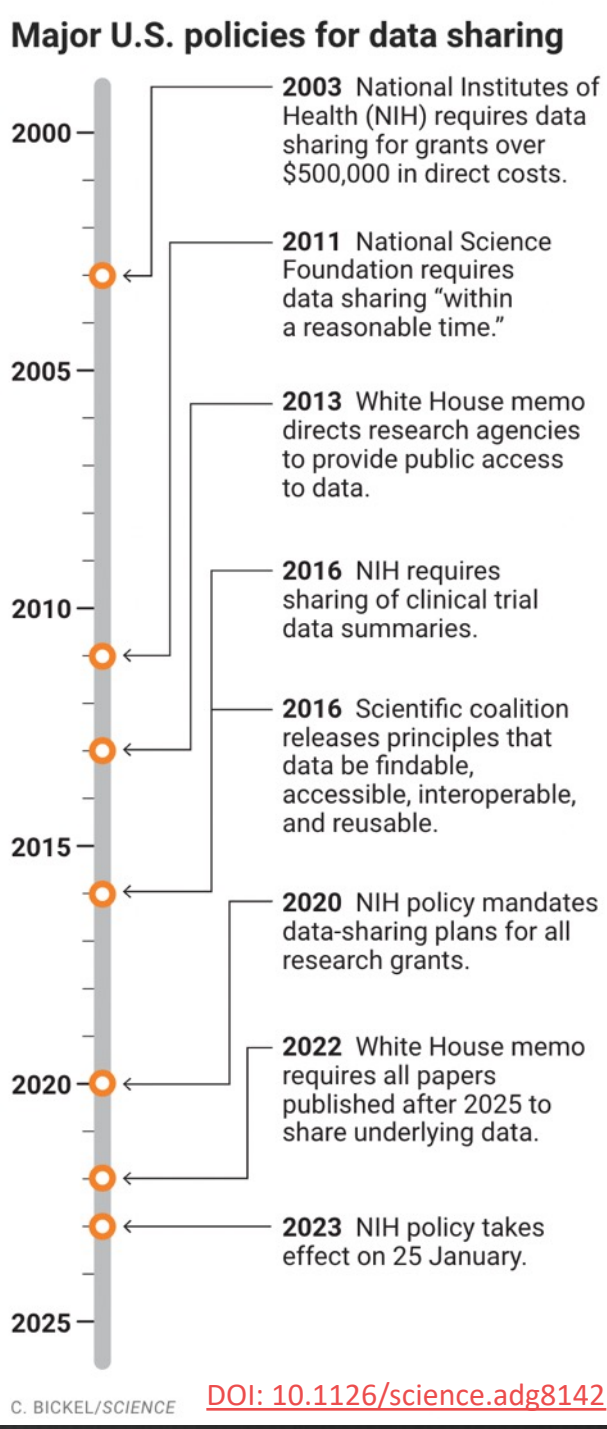
RELEASE DATE: February 26, 2003

NOTICE: NOT-OD-03-032

National Institutes of Health ((NIH)

As part of NIH's long-standing policy to share and make available to the public the results and accomplishments of the activities that it funds, NIH announced and invited comments on a draft statement about the sharing of final research data on March 1, 2002. Since that time,

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html



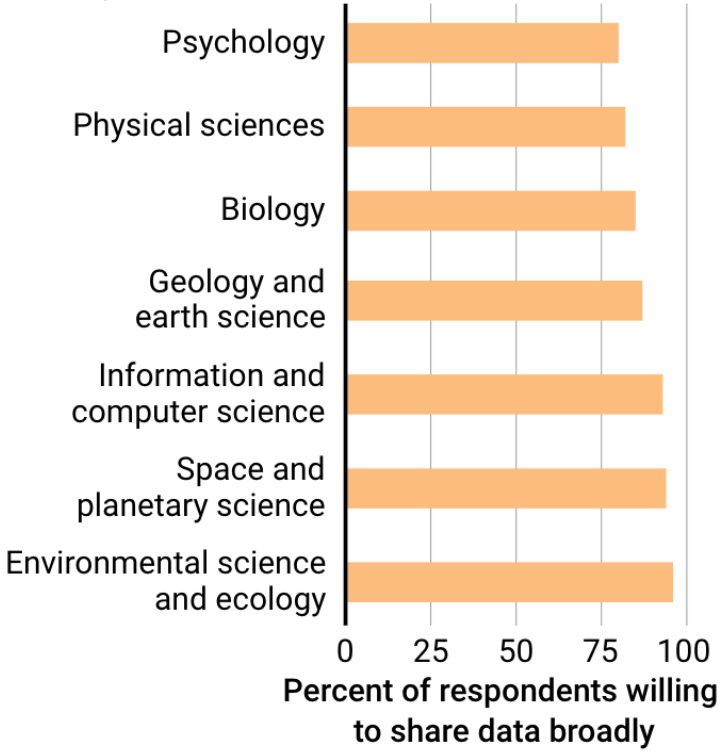
DOI: 10.1126/science.adg8142

Data Sharing Attitudes

Data sharing, by the numbers

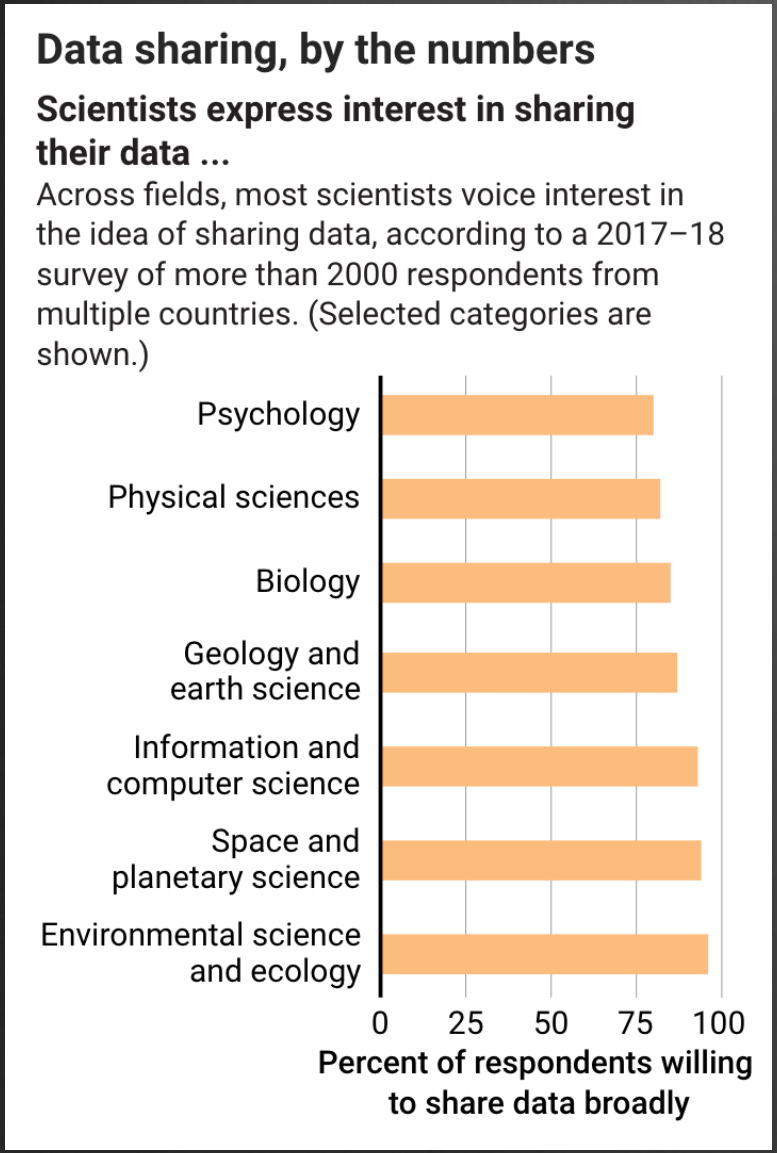
Scientists express interest in sharing their data ...

Across fields, most scientists voice interest in the idea of sharing data, according to a 2017–18 survey of more than 2000 respondents from multiple countries. (Selected categories are shown.)

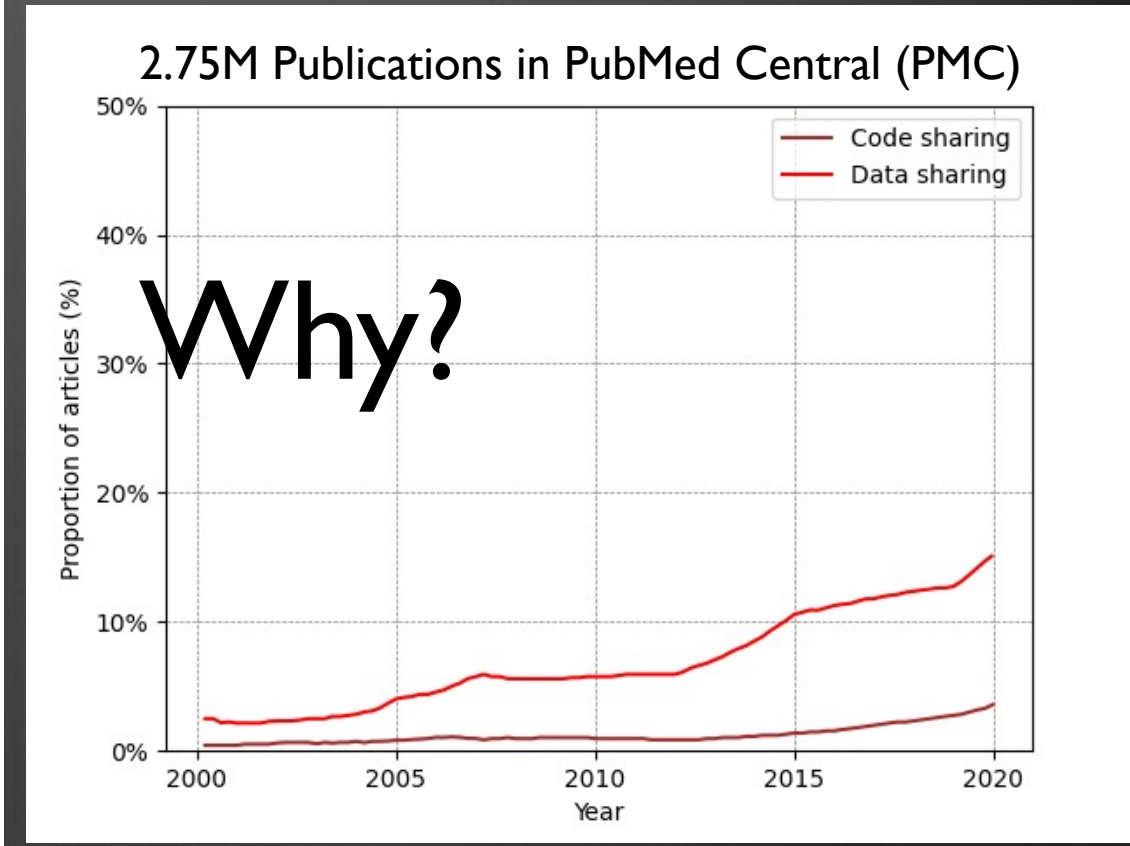


Science, Jan 2023 DOI:10.1126/science.adg8142

Data Sharing Attitudes vs Practice



Science, Jan 2023 DOI:10.1126/science.adg8142



Adapted from Serghiou et al. 2021 DOI:110.1371/journal.pbio.3001107

The Challenge of Cultural Change

- What's need is not just a change in policy but a cultural change
- The existing culture in academia is self-reinforcing
- Changing culture (any culture) is notoriously difficult, especially in the presence of intense competition
- Cultural change can only succeed through a sustained, multi-tiered approach

Brian Nosek, 2019, https://www.cos.io/blog/strategy-for-culture-change

COS CENTER FOR OPEN SCIENCE

The Pyramid of Social Change

Make it REQUIRED - Policy

Make it REWARDING - Incentives

Make it NORMATIVE - Examples

Make it EASY - Support

Make it POSSIBLE - Infrastructure

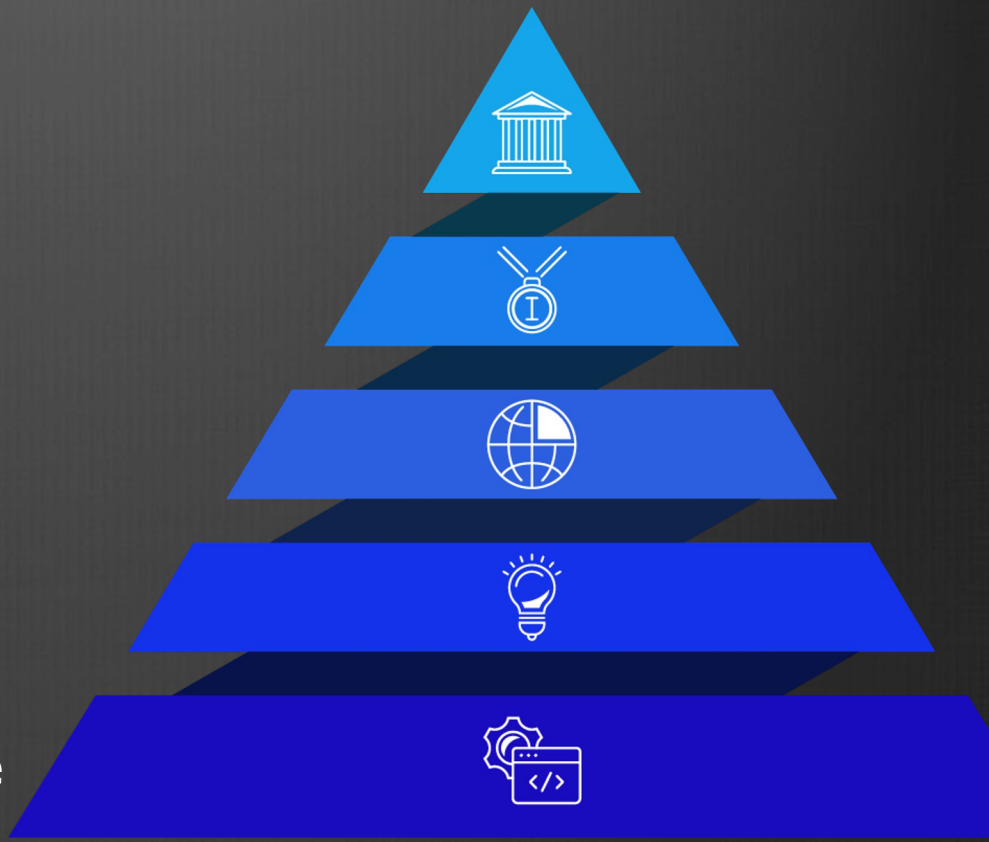


image src: Center for Open Science

Brian Nosek, 2019, https://www.cos.io/blog/strategy-for-culture-change
Stuart Buck, 2023, https://goodscience.substack.com/p/metascience-since-2012-a-personal

The Pyramid of Social Change

Killam Seminar Series: Data Sharing and Transparency in Policy and Practice at NIH

Adam Thomas
Data Science Team Lead, NIMH Intramural Research Program

Host: Jean-Baptiste Poline

Tuesday, May 28, 2024
4 p.m. EDT
Jeanne Timmins Amphitheatre, The Neuro, 3801 University

Registration Required

Lauréats KILLAM Laureates

Make it EASY - Support

Make it POSSIBLE - Infrastructure

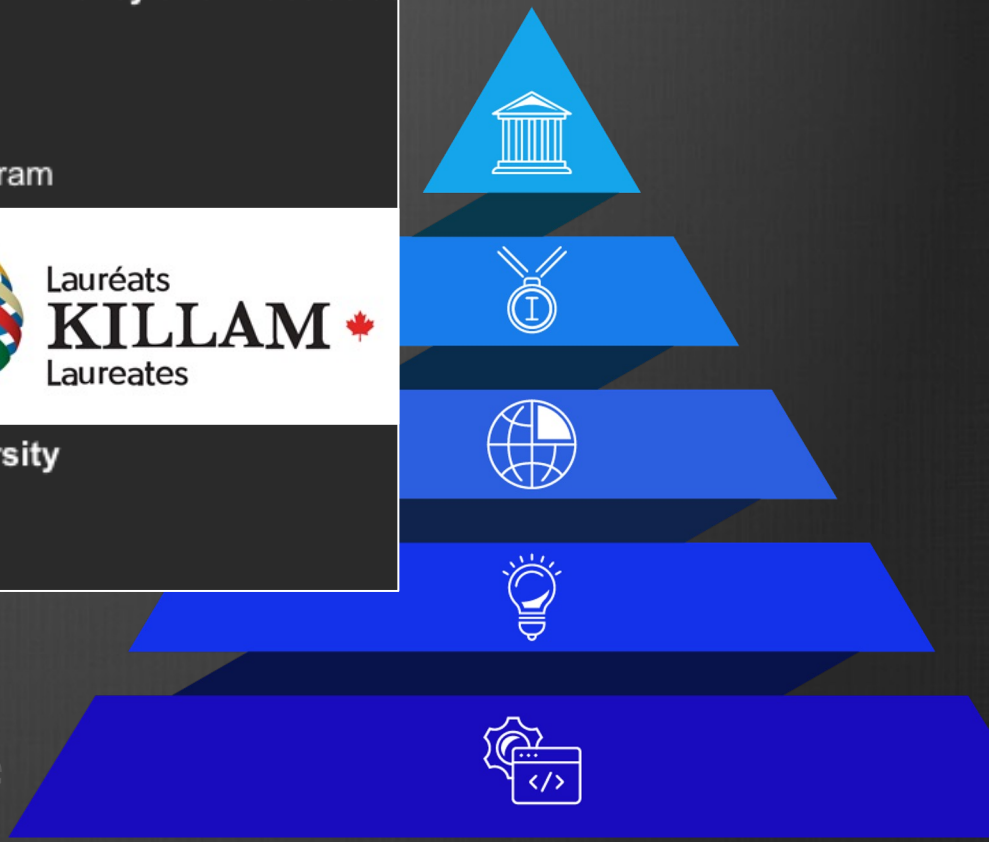


image src: Center for Open Science

Brian Nosek, 2019, https://www.cos.io/blog/strategy-for-culture-change
Stuart Buck, 2023, https://goodscience.substack.com/p/metascience-since-2012-a-personal

CENTER FOR OPEN SCIENCE

GOOD SCIENCE PROJECT

The Pyramid of Social Change

Killam Seminar Series: Data Sharing and Transparency in Policy and Practice at NIH

Adam Thomas
Data Science Team Lead, NIMH Intramural Research Program

Host: Jean-Baptiste Poline

Tuesday, May 28, 2024
4 p.m. EDT
Jeanne Timmins Amphitheatre, The Neuro, 3801 University

Registration Required

Lauréats KILLAM Laureates

Make it EASY - Support

Make it POSSIBLE - Infrastruc

Data Science and Sharing Team
Board of Scientific Counselors Review
https://bit.ly/dsst_bsc2024

December 2023

Contents

CENTER FOR OPEN SCIENCE

Brian Nosek, 2019, https://www.cos.io/blog
Stuart Buck, 2023, https://goodscience.substa

Alternative Incentives

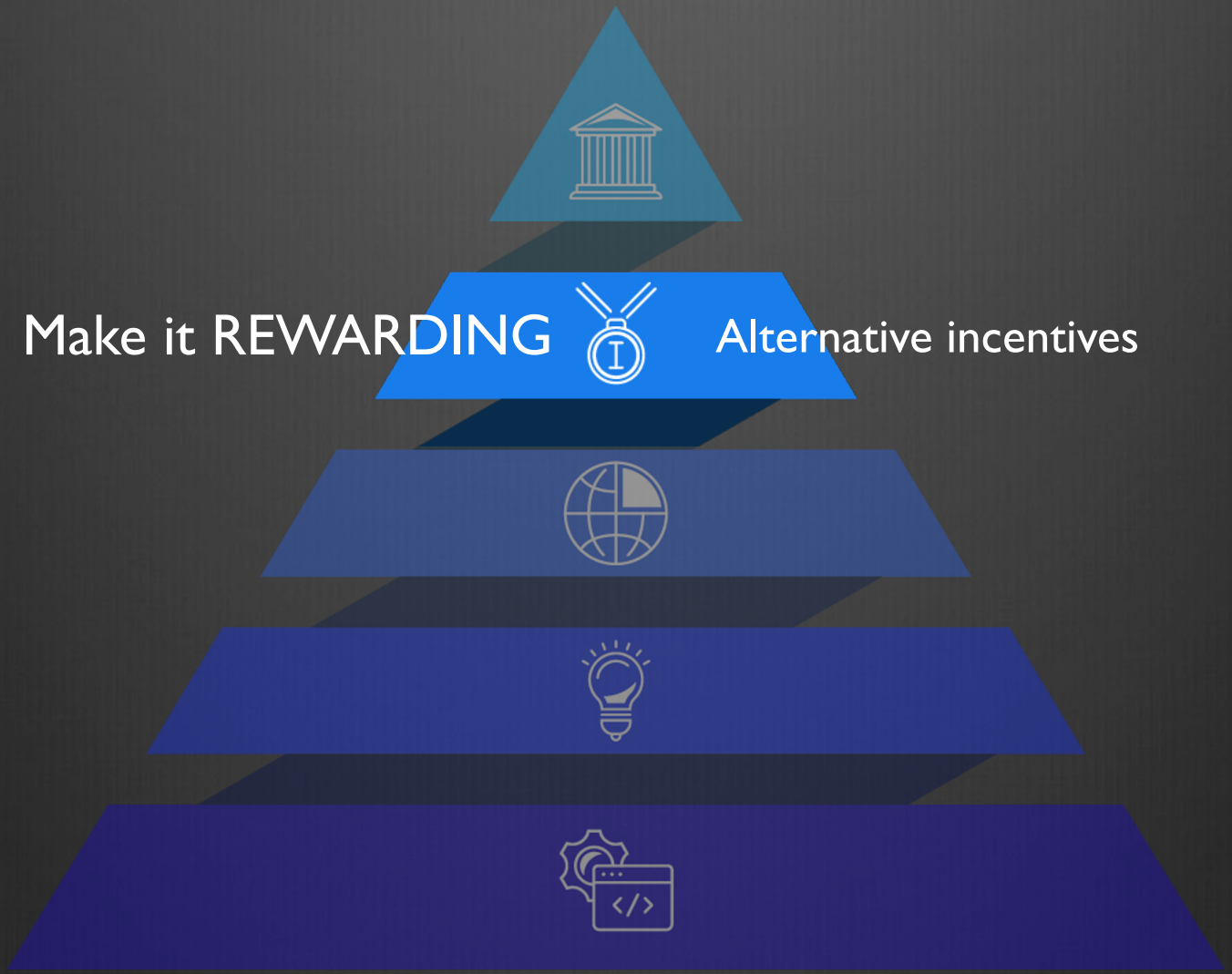


image src: Center for Open Science

Measuring Data Sharing

- To reward data sharing there must be a mechanism to measure it
- Text mining tools exist to identify data sharing reported in publications
- However, aggregate information on reported data sharing is not easily accessible to funders and policy makers

~25,000 Unique Pubs 2019-2023

...

Measuring Data Sharing

REWARDING

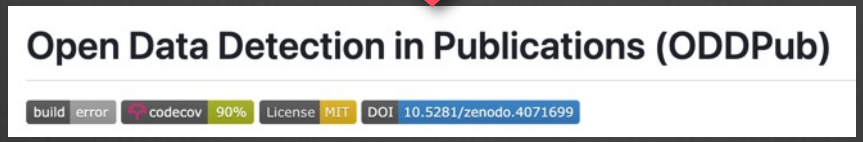
- NIH is composed of 25 distinct institutes and centers (ICs)

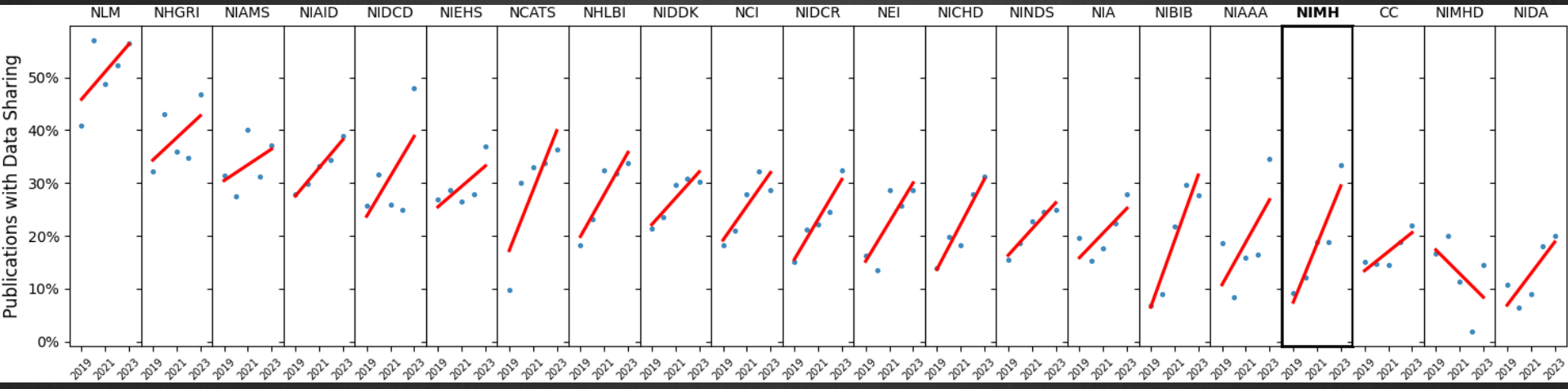
...

Measuring Data Sharing

- To reward data sharing there must be a mechanism to measure it
- Text mining tools exist to identify data sharing reported in publications
- However, aggregate information on reported data sharing is not easily accessible to funders and policy makers

~25,000 Unique Pubs
2019-2023



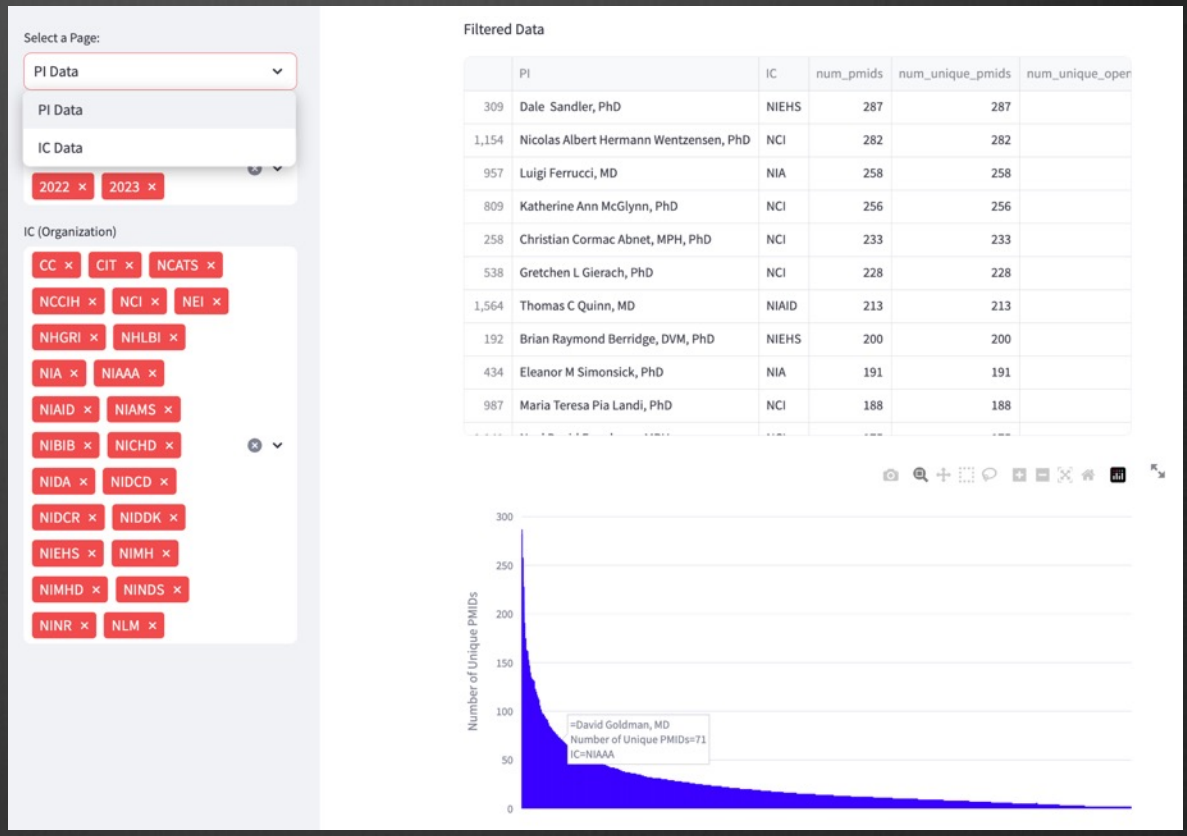


Measuring Data Sharing

- To reward data sharing there must be a mechanism to measure it
- Text mining tools exist to identify data sharing reported in publications
- However, aggregate information on reported data sharing is not easily accessible to funders and policy makers

<u>The IRP ShareStats Dashboard</u>

- Allows NIH intramural leadership to quickly identify and reward investigators who have strong records of data sharing



...

Similar Efforts

- Riedel et al., 2020 (OddPub)
  - ~ 11,000 pubs from the Charité – U. Berlin
- Serghiou et al., 2021 (rTransparent)
  - ~ 2.75M pubs from PubMed Central Open Access
- Menke et al., 2022 (SciScore Rigor Transparency Index)
  - ~2.15M pubs from PubMed Central Open Access
- Piękniewska et al., 2023 (SciCrunch, RRIDs)
  - ~1.3M pubs PubMed Central Open Access

What can we add?

- LLM-based metrics
  - Improved accuracy
    - In our analysis, OddPub had ~10% error rate, equally balanced between false positives and false negatives
  - Extracting more specific details from text

Measuring Data Sharing

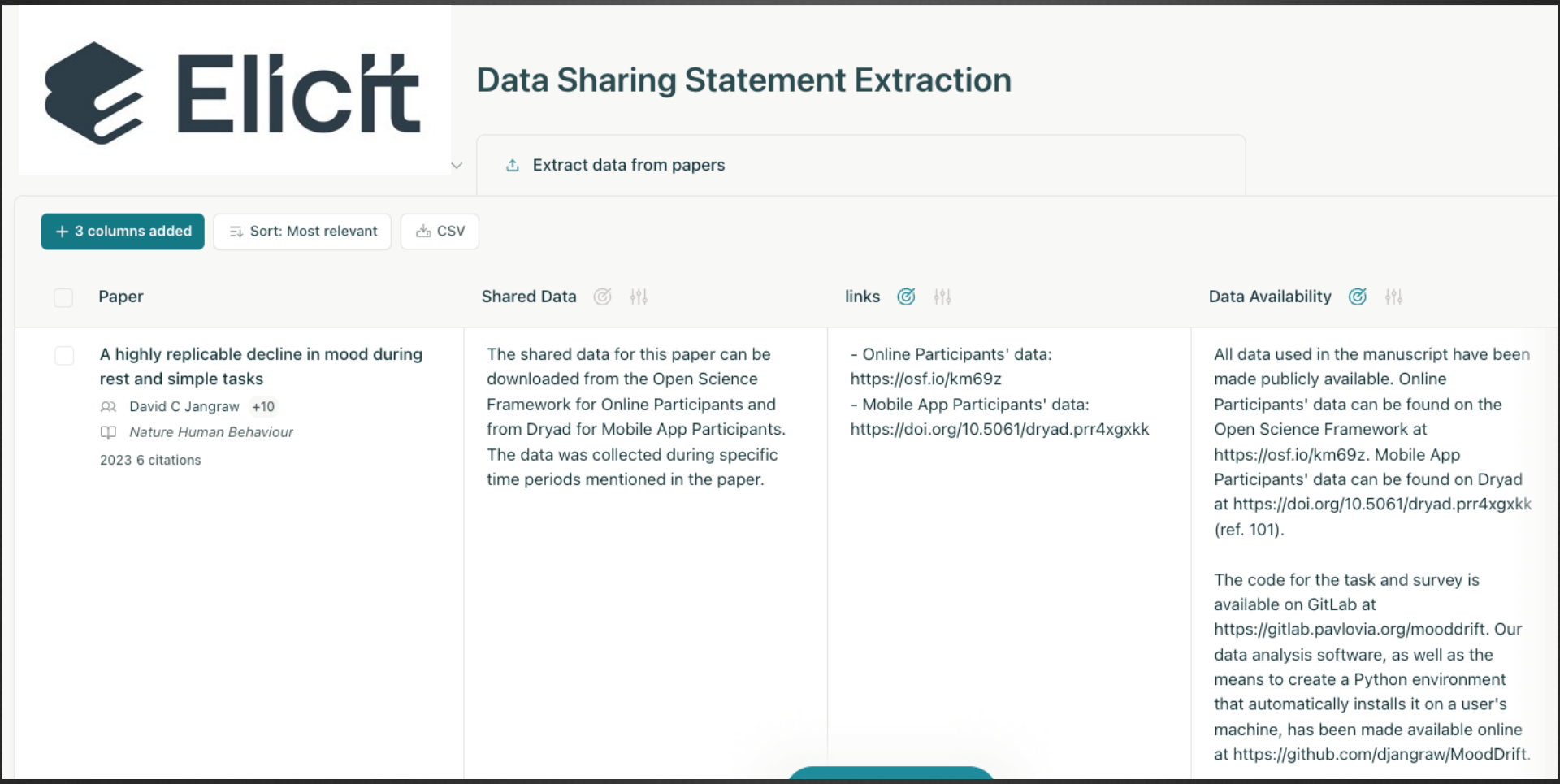
Large Language Models to parse data sharing statements

EM densities and protein models have been deposited in the Electron Microscopy Data Bank and Protein Data Bank for the NaCT-citrate (EMD-22457, 7JSK) and NaCT-PF2 (EMD-22456, 7JSJ) complexes.

Paper DOI	Data Types	Shared Data	Data Repository	Unique ID	Repository URL	Metadata Accessible
10.1038/s41586-021-03230-x	Cryo-EM image processing and model building	TRUE	Electoscopic Data Bank (EMDB)	EMD-22456	https://www.ebi.ac.uk/emdb/	TRUE
10.1038/s41586-021-03230-x	Cryo-EM image processing and model building	TRUE	Electoscopic Data Bank (EMDB)	EMD-22457	https://www.ebi.ac.uk/emdb/	TRUE
10.1038/s41586-021-03230-x	Protein structure determination, Biochemical assay results, Thermostability assay results	TRUE	Proteinf Data Bank (PDB)	7JSJ	https://www.rcsb.org/	TRUE
10.1038/s41586-021-03230-x	Protein structure determination, Biochemical assay results, Thermostability assay results	TRUE	Proteinf Data Bank (PDB)	7JSK	https://www.rcsb.org/	TRUE

...

Using LLMs for complex extraction



What can we add?

- LLM-based metrics
  - Improved accuracy
    - In our analysis, OddPub had ~10% error rate, equally balanced between false positives and false negatives
  - Extracting more specific details from pubs
  - More nuanced evaluations

Many Open Science Rubrics are vague

- FAIR (Wilksinson et al, 2016)

ID	Questions from https://www.go-fair.org/fair-principles/
F1	(Meta)data are assigned a globally unique and persistent identifier
F2	Data are described with rich metadata (defined by R1 below)
F3	Metadata clearly and explicitly include the identifier of the data they describe
F4	(Meta)data are registered or indexed in a searchable resource
A1	(Meta)data are retrievable by their identifier using a standardised communications protocol
A1-1	The protocol is open, free, and universally implementable
A1-2	The protocol allows for an authentication and authorisation procedure, where necessary
A2	Metadata are accessible, even when the data are no longer available
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2	(Meta)data use vocabularies that follow FAIR principles
I3	(Meta)data include qualified references to other (meta)data
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes
R1-1	(Meta)data are released with a clear and accessible data usage license
R1-2	(Meta)data are associated with detailed provenance
R1-3	(Meta)data meet domain-relevant community standards

Many Open Science Rubrics are vague

- FAIR (Wilkinson et al., 2016)
- Delphi Consensus (Cobey et al., 2023)

Rank	Practice	Unit of analysis
1	whether clinical trials were registered before they started recruitment	Paper, protocol
2	whether study data were shared openly at the time of publication	paper
3	published open access (time delay?)	paper
4	whether study code was shared openly at the time of publication	paper
5	whether systematic reviews have been registered.	paper
6	were registered clinical trials were reported in the registry within 1 year of study completion	paper
7	whether there was a statement about study materials sharing with publications	paper
8	whether study reporting guideline checklists were used	paper
9	citations to data.	dataset
10	trial results in a manuscript-style publication (peer reviewed or preprint).	paper
11	the number of preprints.	Institution
12	systematic review results in a manuscript-style publication (peer reviewed or preprint).	Systematic review

---

Many Open Science Rubrics are vague

- FAIR (Wilkinson et al., 2016)
- Delphi Consensus (Cobey et al., 2023)
- TOSI Open Science Principles

1. Public release of data and other scientific resources

The Neuro and its researchers will publish open access articles and render all positive and negative numerical data, models used, data sources, materials, reagents, algorithms, software and other scientific resources publicly available no later than the publication date of the first article that relies on this data or resource.

2. External research partnerships

All publications, data and scientific resources generated through research partnerships –whether with commercial, philanthropic, or public sector actors – are to be released publicly no later than the publication date of the first article that relies on this data or resource.

3. Research materials and tools

The Neuro supports knowledge creation and innovation by maximizing the long-term value of the physical contributions made by research participants and the physical scientific resources created by Neuro researchers and their collaborators. The Neuro will manage these resources in such a way as to remain financially self-sustaining, while continuing to enrich and strengthen its informational content and the knowledge it provides.

In handling materials originating from patients, The Neuro recognizes the primacy of safeguarding the dignity and privacy of patient-participants, and respecting the rights and duties owed them through the informed consent process.

4. Intellectual property

Subject to patient confidentiality and informed consent given, neither The Neuro nor its researchers in their capacity as employees or consultants of McGill or The Neuro will obtain restrictive intellectual property protection in respect of any of their research outputs, whether done internally or with collaborators.

5. Researcher and patient autonomy

The Neuro supports the autonomy of its stakeholders, including but not limited to researchers, staff, trainees and patients, through recognizing their right to decline to participate in research and associated activities under an Open Science framework. However, The Neuro will not support activities that compromise the Open Science principles outlined above.

https://www.mcgill.ca/neuro/open-science/open-science-principles

---

Using LLMs for nuanced open science questions

Can large language models provide useful feedback on research papers? A large-scale empirical analysis.

Weixin Liang[1*], Yuhui Zhang[1*], Hancheng Cao[1*], Binglu Wang[2], Daisy Yi Ding[3], Xinyu Yang[4], Kailas Vodrahalli[5], Siyu He[3], Daniel Scott Smith[6], Yian Yin[4], Daniel A. McFarland[6], and James Zou[1,3,5+]

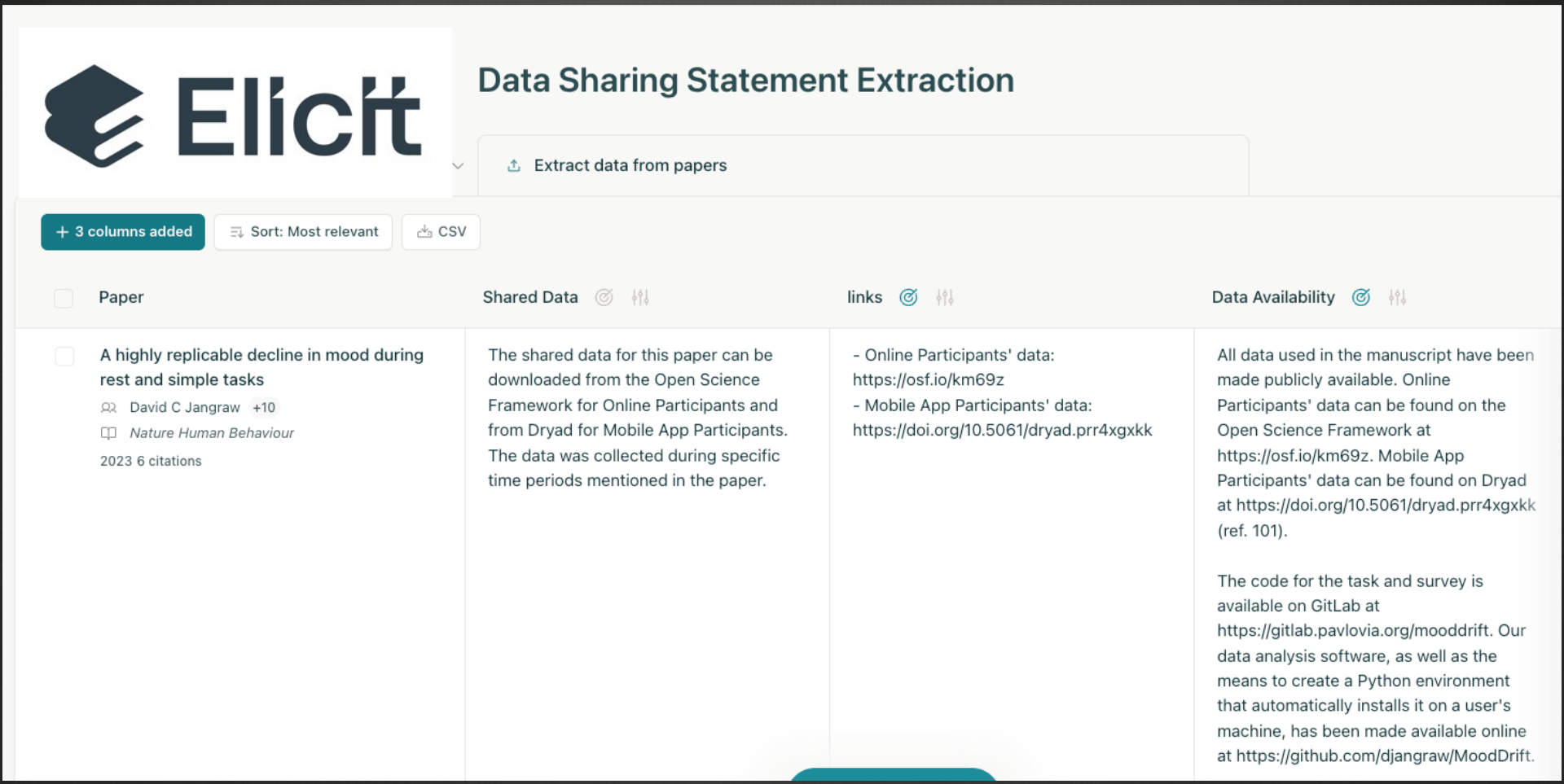
[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[2]Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA
[3]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
[4]Department of Information Science, Cornell University, Ithaca, NY 14850, USA
[5]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA
[6]Graduate School of Education, Stanford University, Stanford, CA 94305, USA
[+]Correspondence should be addressed to: jamesz@stanford.edu
[*]these authors contributed equally to this work

https://doi.org/10.48550/arXiv.2310.01783

ABSTRACT

"… more than half (57.4%) of the users found GPT-4 generated feedback helpful/very helpful and 82.4% found it more beneficial than feedback from at least some human reviewers."

Using LLMs for nuanced open science questions



What can we add?

- LLM-based metrics
  - Improved accuracy
    - In our analysis, OddPub had ~10% error rate, equally balanced between false positives and false negatives
  - Extracting more specific details from pubs
  - More nuanced evaluations
- An open, method-agnostic, repository of open science indicators

---

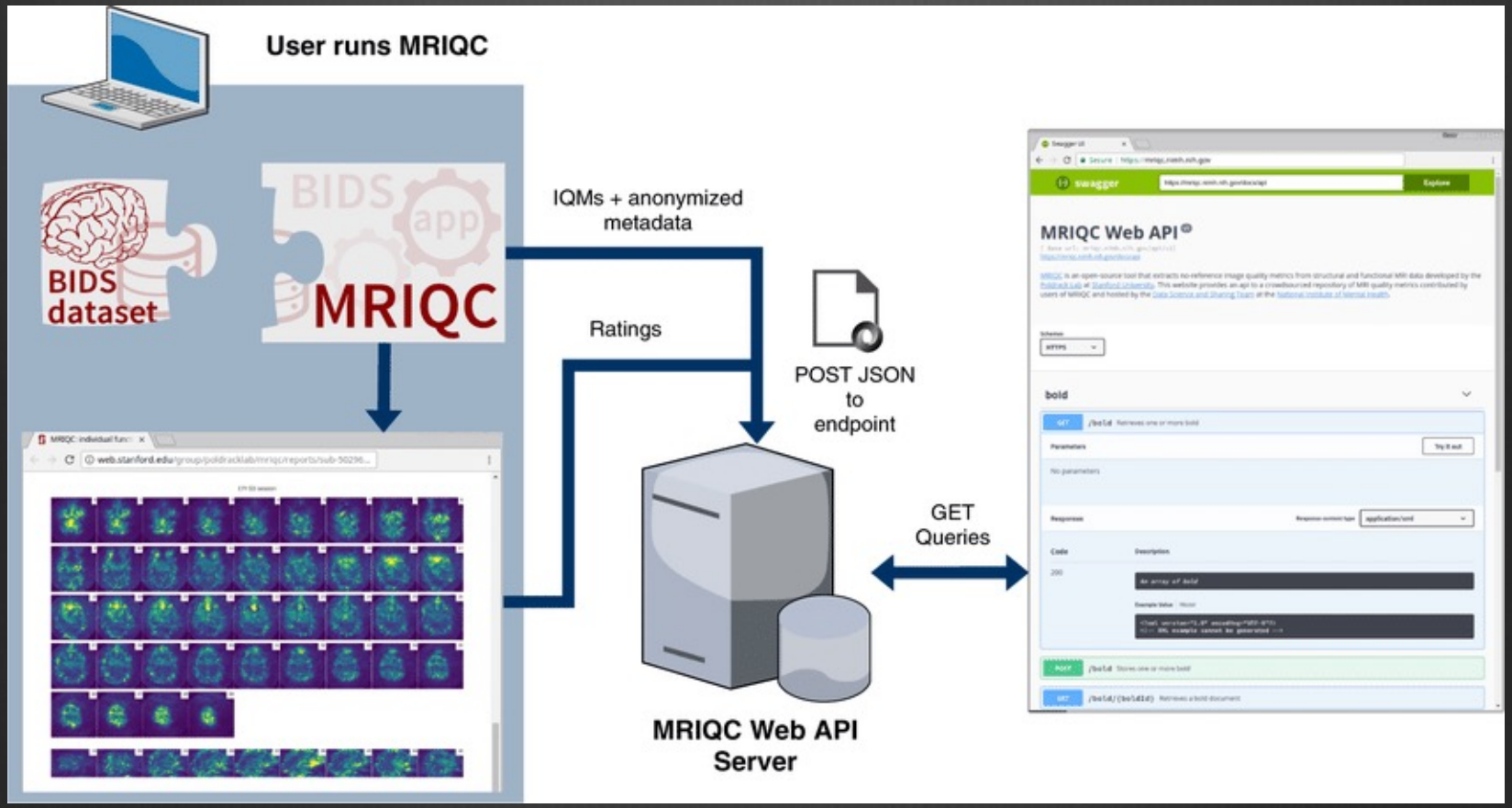
Lots of automatically and manually label data available

- Riedel e
  - ~ 11,0
- Serghiou
  - ~ 2.75
- Menke et
  - ~2.15
- Pięknie
  - ~1.3M

ccess

y Index)

Modelled on previous success: MRIQC



Modelled on previous success: MRIQC DB

Stanford University

MRIQC Web API

- Crowdsourced database of MR QC metrics
- QC metrics from ~375K unique BOLD scans and ~280K T1w scans
- Publicly available:
  - https://mriqc.nimh.nih.gov/

SCIENTIFIC DATA

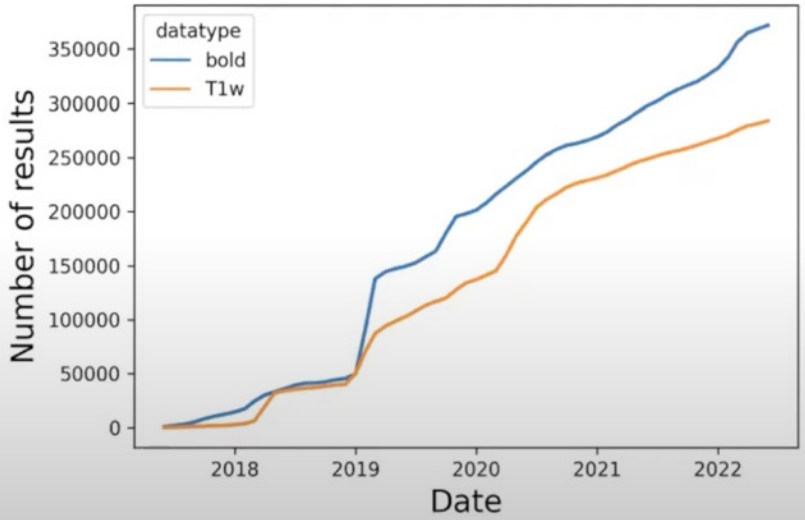
OPEN

DATA DESCRIPTOR

Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines

Received: 19 September 2018
Accepted: 12 March 2019

Oscar Esteban[1], Ross W. Blair[1], Dylan M. Nielson[2], Jan C. Varada[3], Sean Marrett[3], Adam G. Thomas[2], Russell A. Poldrack[1] & Krzysztof J. Gorgolewski[1]



Modelled on previous success: MRIQC DB

Stanford University

MRIQC Web API

- Crowdsourced database of MR QC metrics
- QC metrics from ~375K unique BOLD scans and ~280K T1w scans
- Publicly available:
  - https://mriqc.nimh.nih.gov/

SCIENTIFIC DATA

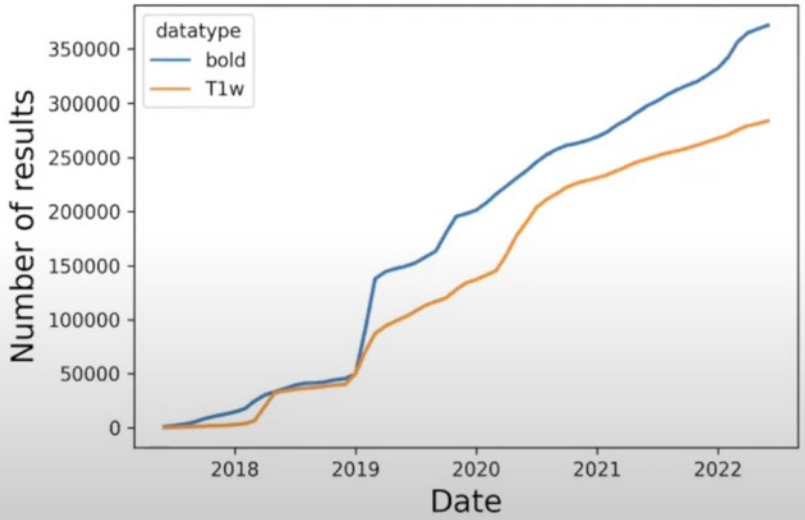
OPEN

DATA DESCRIPTOR

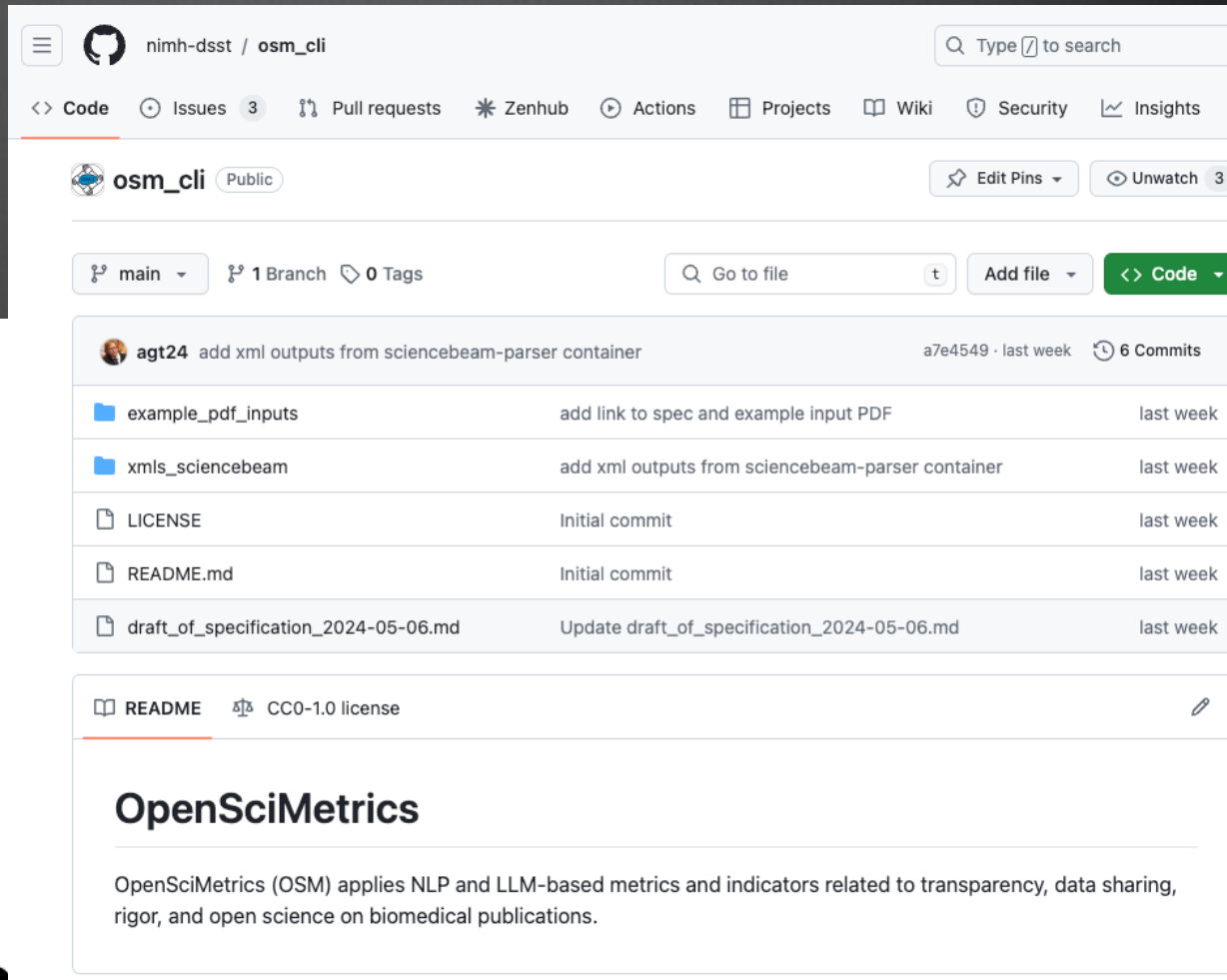
Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines

Received: 19 September 2018
Accepted: 12 March 2019

Oscar Esteban[1], Ross W. Blair[1], Dylan M. Nielson[2], Jan C. Varada[3], Sean Marrett[3], Adam G. Thomas[2], Russell A. Poldrack[1] & Krzysztof J. Gorgolewski[1]



Work In Progress



GitStart

Product of the day
1st

Code as a Service

Assign tickets, get ***high-quality production code*** powered by AI agents and our developer community.

Synergies on Extraction & Manual Labelling

Mining the neuroimaging literature.

Jérôme Dockès[1*], Kendra Oudyk[2*], Mohammad Torabi[2], Alejandro I de la Vega[3], and Jean-Baptiste Poline[2]

[1]National Institute for Research in Digital Science and Technology (INRIA), Paris, France
[2]Montreal Neurological Institute, McGill University, Montreal, Canada
[3]University of Texas at Austin, Austin, Texas, USA
[*]co-first authors

pubget

build passing | codecov 100% | pubget on GitHub

labelbuddy

What can we add?

- LLM-based metrics
  - Improved accuracy
    - In our analysis, OddPub had ~10% error rate, equally balanced between false positives and false negatives
  - Extracting more specific details from pubs
  - More nuanced evaluations
- An open, method-agnostic, repository of open science indicators
- A new "more bang-for-your-buck" metric

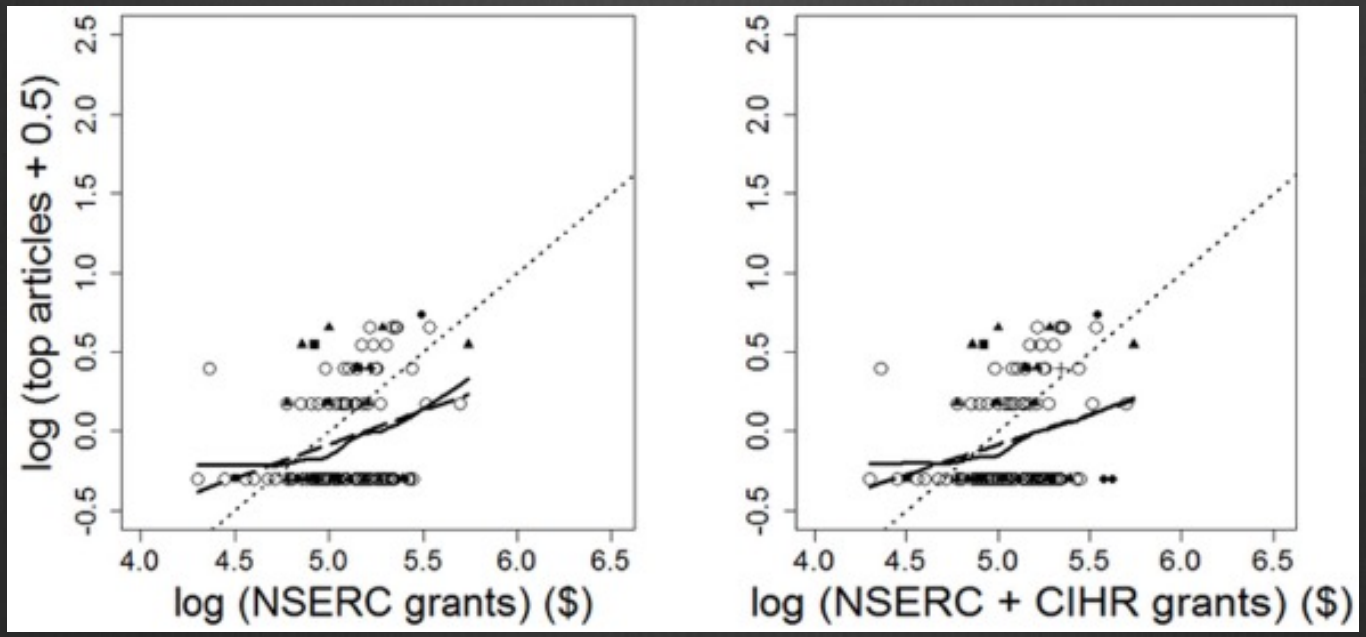
Correlating Funding Levels and Open Science Practices

Big Science vs. Little Science: How Scientific Impact Scales with Funding

Jean-Michel Fortin, David J. Currie*

Ottawa-Carleton Institute of Biology, University of Ottawa, Ottawa, Ontario, Canada

June 2013 | Volume 8 | Issue 6 | e65263



Correlating Funding Levels and Open Science Practices

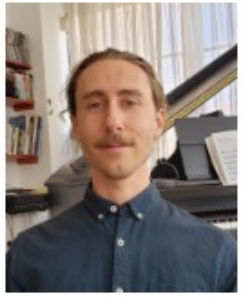
- Identify investigators with relatively low levels of funding, but relatively strong open science practices
-

Thank You!

Tanenbaum Open Science Institute (TOSI)

Guy Rouleau
Director, The Neuro
Co-Founder of the Tanenbaum Open Science Institute

Annabel Seyller
Chief of Staff, The Neuro
CEO, Tanenbaum Open Science Institute



Gabriel Pelletier
Open Science Data Manager
Interim Open Science Alliance Officer

Luisa Pimentel
Open Science Community Officer

neuro

Jean-Baptiste Poline

2023 Data Science and Sharing Team

Adam Thomas
Team Lead

Dustin Moraczewski
Data Scientist

Eric Earl
Data Scientist

Mia Zwally
Post-bac IRTA

Arshitha Basavaraj
Data Engineer

Jessica Dafflon
Data Scientist

http://cmn.nimh.nih.gov/dsst

Feedback? Where to focus?

- LLM-based metrics
  - Improved accuracy
- In our analysis, OddPub had ~10% error rate, equally balanced between false positives and false negatives
  - Extracting more specific details from pubs
  - More nuanced evaluations
- An open, method-agnostic, repository of open science indicators
- A new "more bang-for-your-buck" metric

END
