Data Sharing and Transparency in Policy and Practice at NIH

28 May 2024

Adam Thomas

Open Scientist in Residence Program, TOSI, The Neuro

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

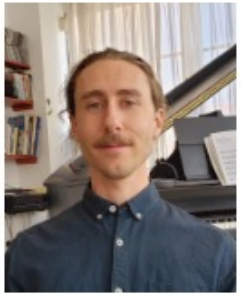
Thank You!

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

Jean-Baptiste Poline



Gabriel Pelletier
Open Science Data Manager
Interim Open Science Alliance Officer

Luisa Pimentel
Open Science Community Officer

NEURO SERIES
KILLAM SEMINARS

Lauréats KILLAM Laureates · neuro

Roadmap

- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next

Slides & links: http://bit.ly/adamt_osrp

Roadmap

- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next

Slides & links: http://bit.ly/adamt_osrp

What is the NIH?

- 27 different institutes and centers
- $47 billion USD in 2024
  - ~ 85% = Extramural grants
  - ~ 15% = Intramural Research Program (IRP)
  - Intramural ≠ 
- National Institute of Mental Health (NIMH) Intramural Research Program (IRP)
  - ~ 40 research groups
  - ~ 12 research support groups
  - One of which: Data Science and Sharing Team (DSST)

My Team

Adam Thomas
Team Lead

Dustin Moraczewski
Data Scientist

Eric Earl
Data Scientist

Mia Zwally
Post-bac Trainee

Arshitha Basavaraj
Data Engineer

Jessica Dafflon
Data Scientist

?

YOU?

Interested in working with us? Reach out! http://bit.ly/adamt_osrp

Roadmap

- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next

Adoption of Open Science is Happening

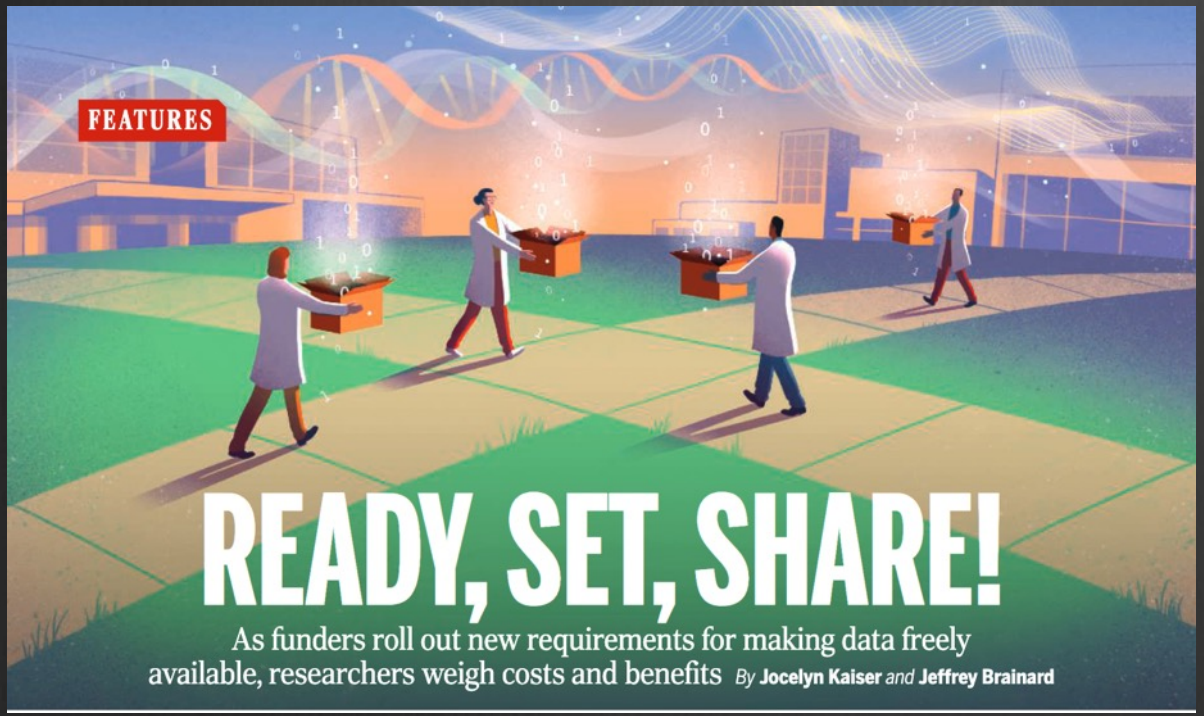
NEWS | 16 February 2022 | Correction 16 February 2022

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.

Are we there yet?

History of Data Sharing in the US

FINAL NIH STATEMENT ON SHARING RESEARCH DATA

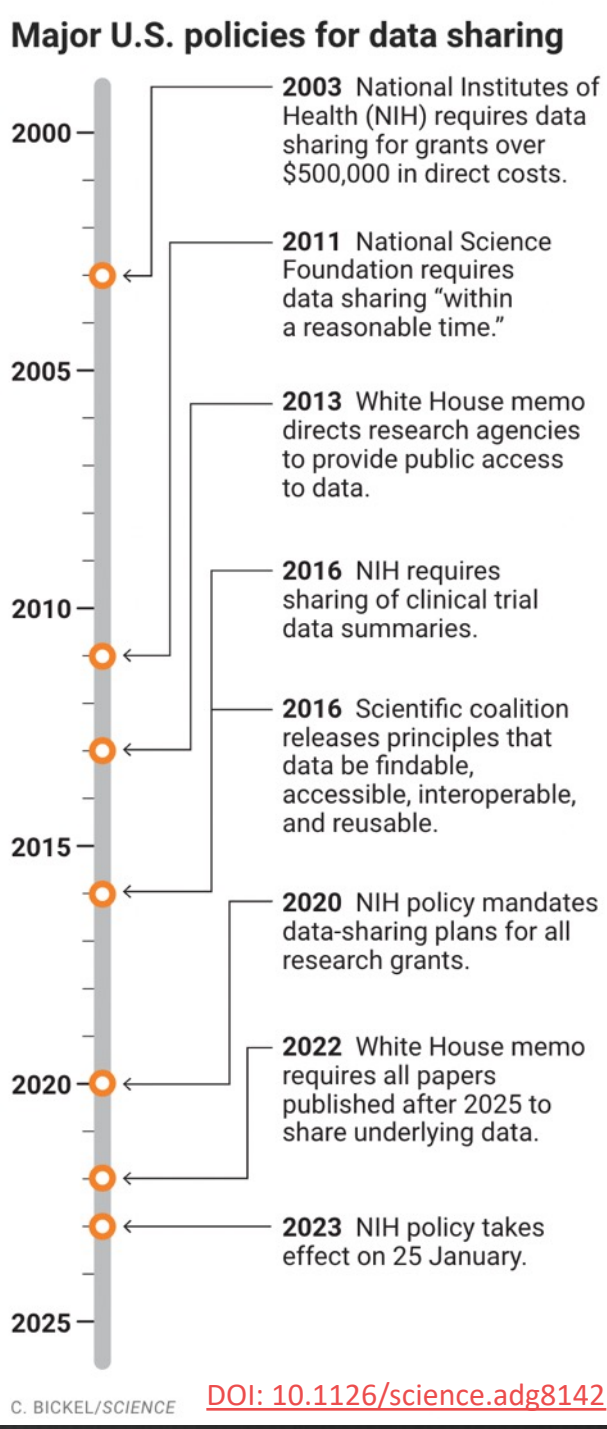
RELEASE DATE: February 26, 2003

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



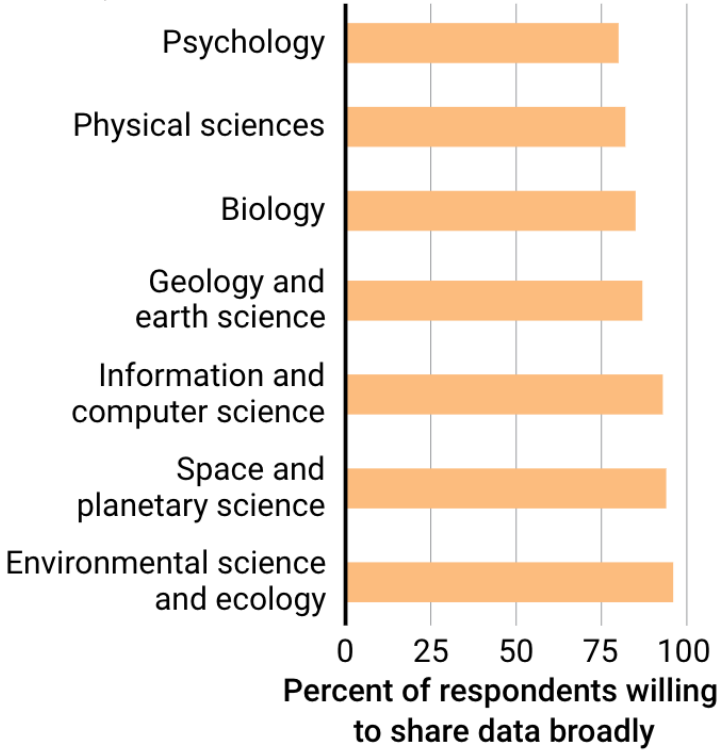
DOI: 10.1126/science.adg8142

Data Sharing Attitudes

Data sharing, by the numbers

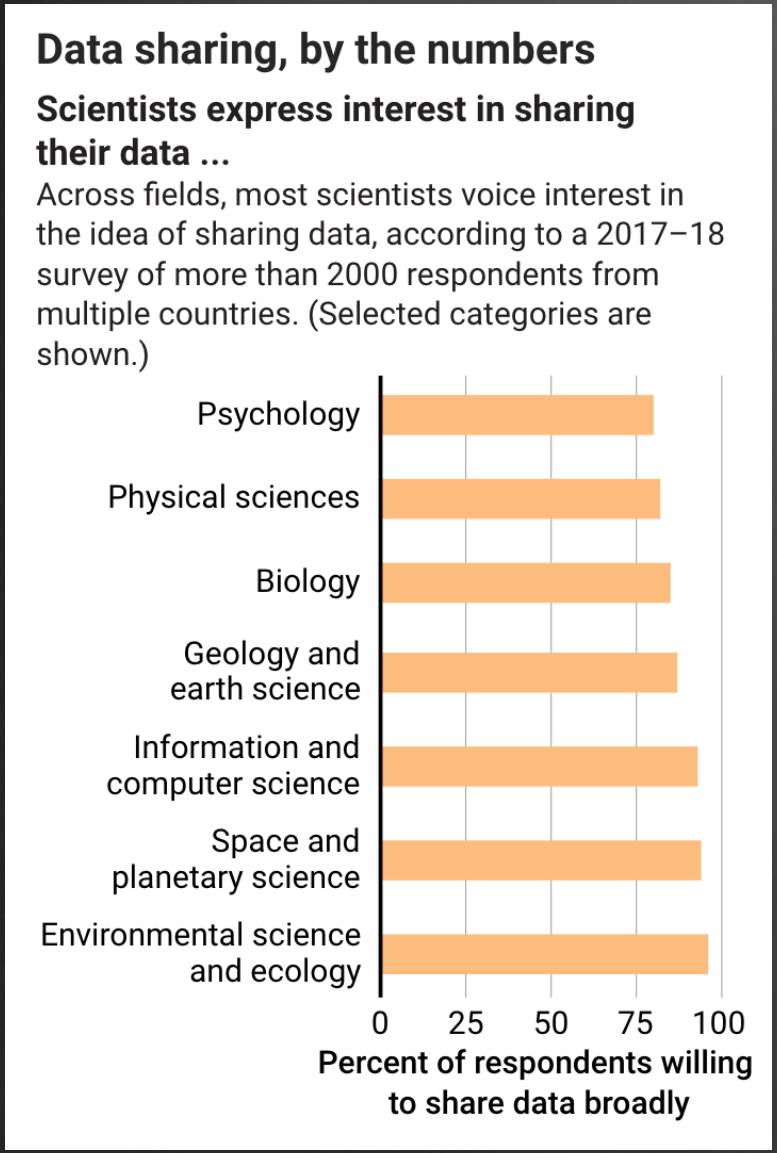
Scientists express interest in sharing their data ...

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.)

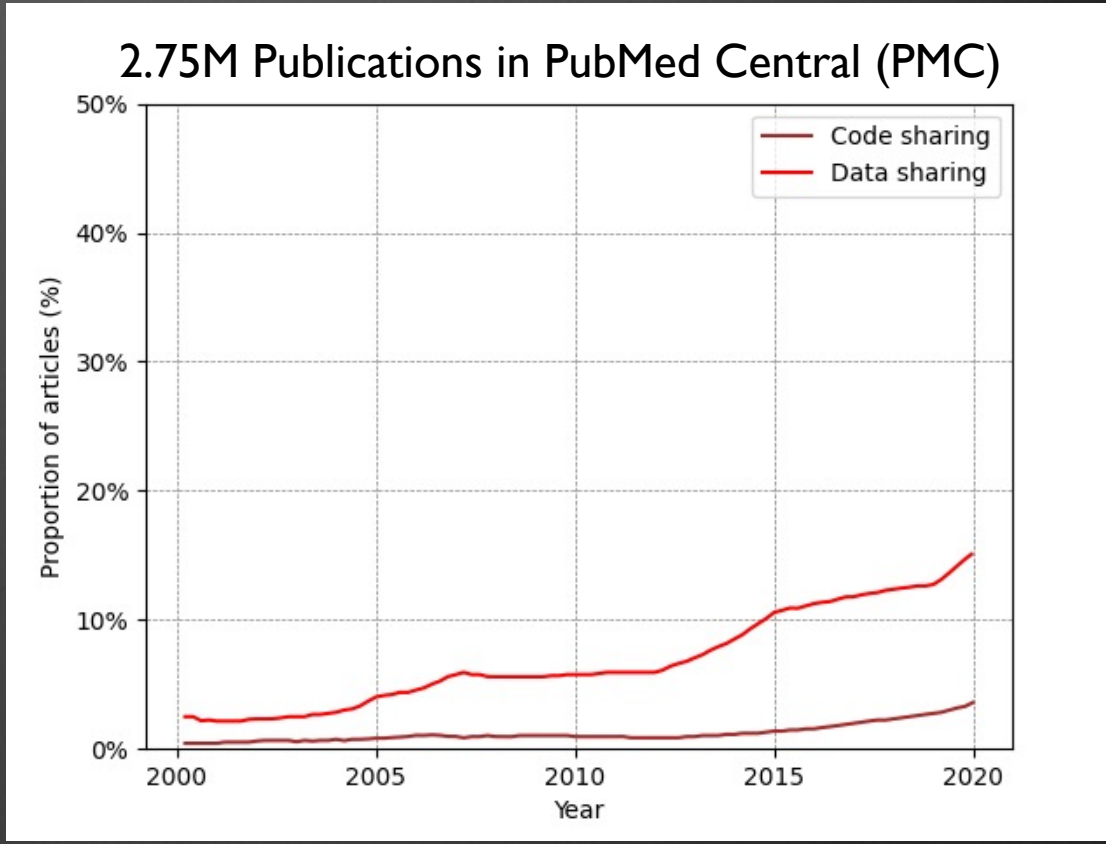


Science, Jan 2023 DOI:10.1126/science.adg8142

Data Sharing Attitudes vs Practice

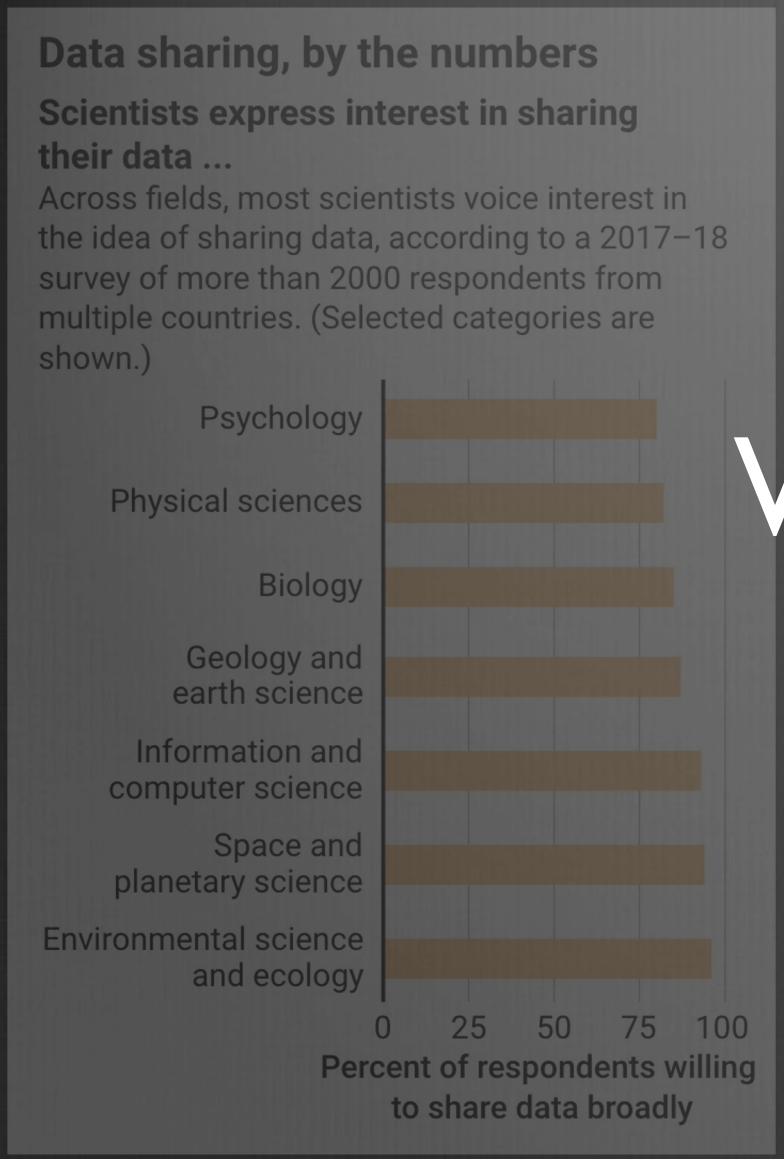


Science, Jan 2023 DOI:10.1126/science.adg8142

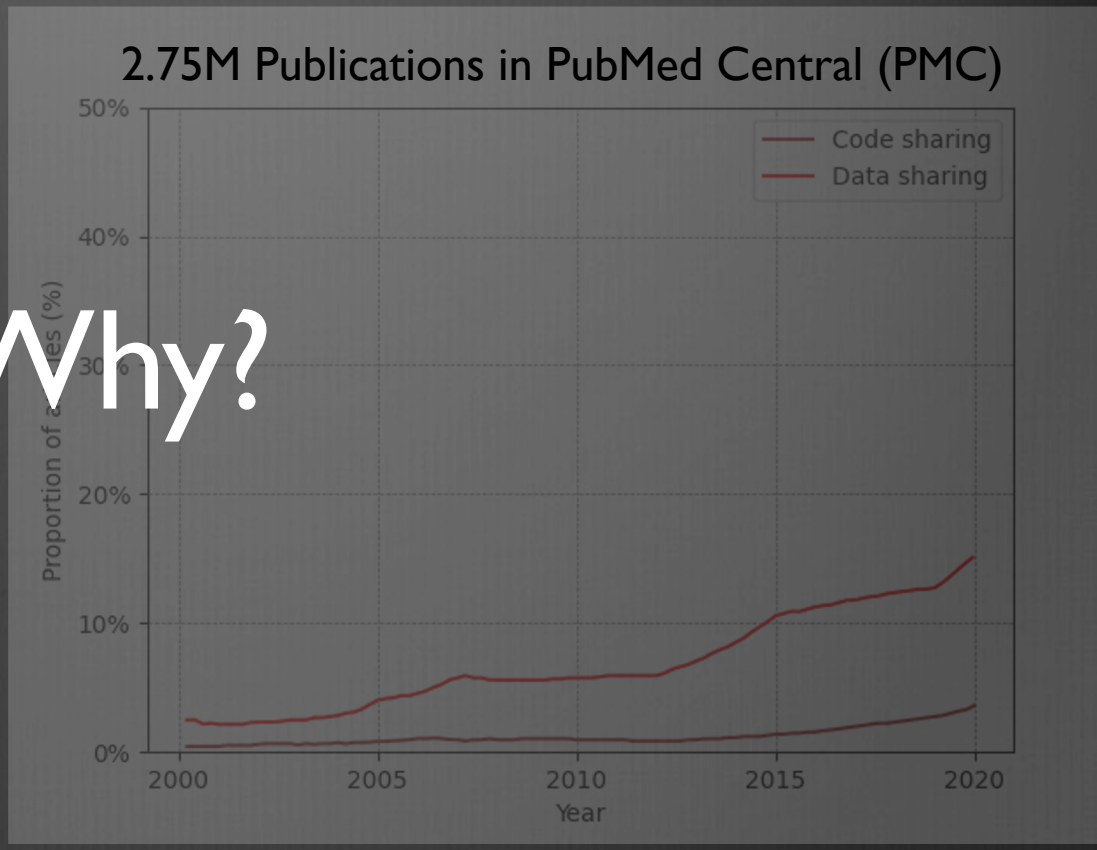


Adapted from Serghiou et al. 2021 DOI:110.1371/journal.pbio.3001107

Data Sharing Attitudes vs Practice



Why?



Adoption of Option Science is a "Cultural Change"

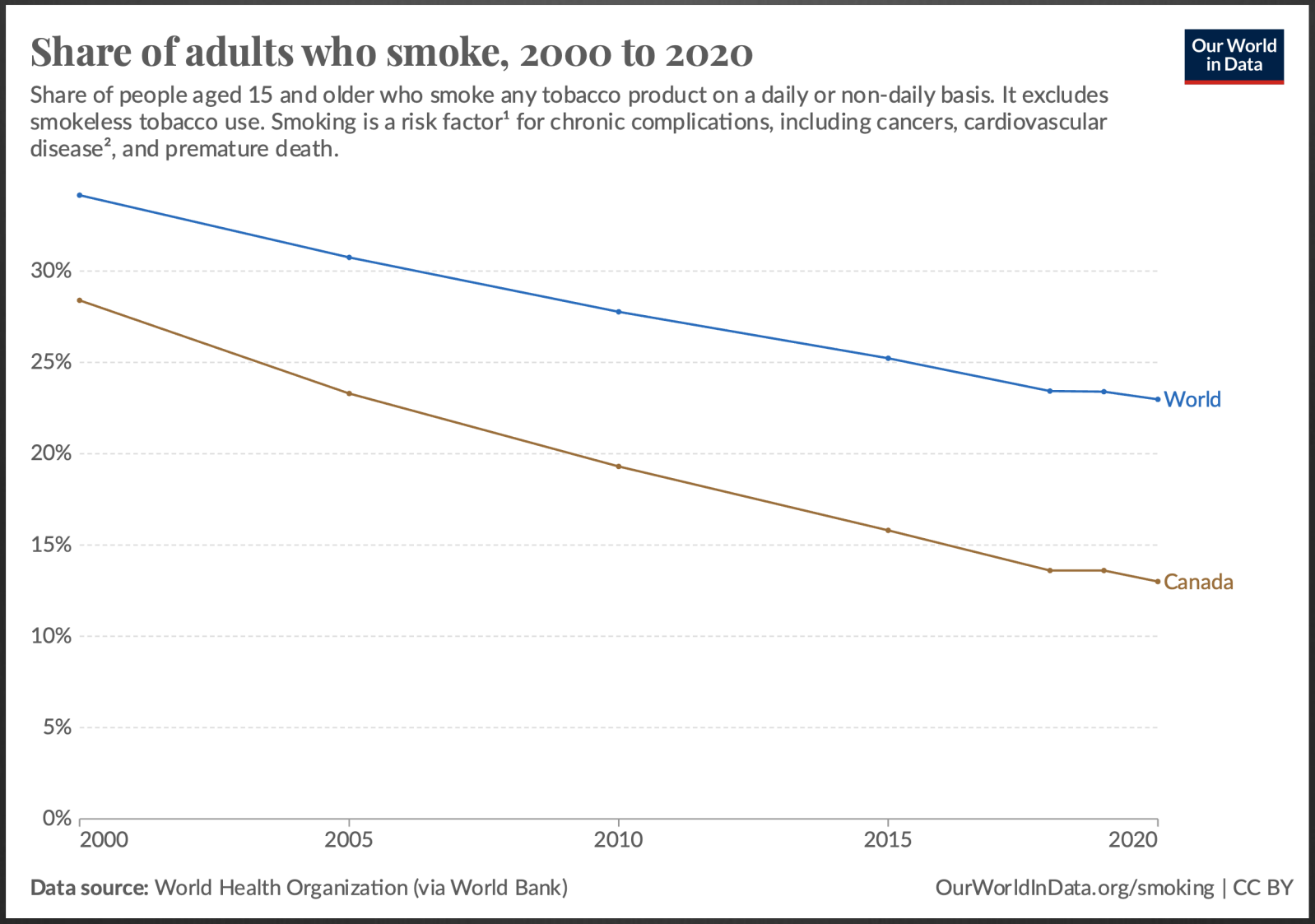
1940s

Vs.

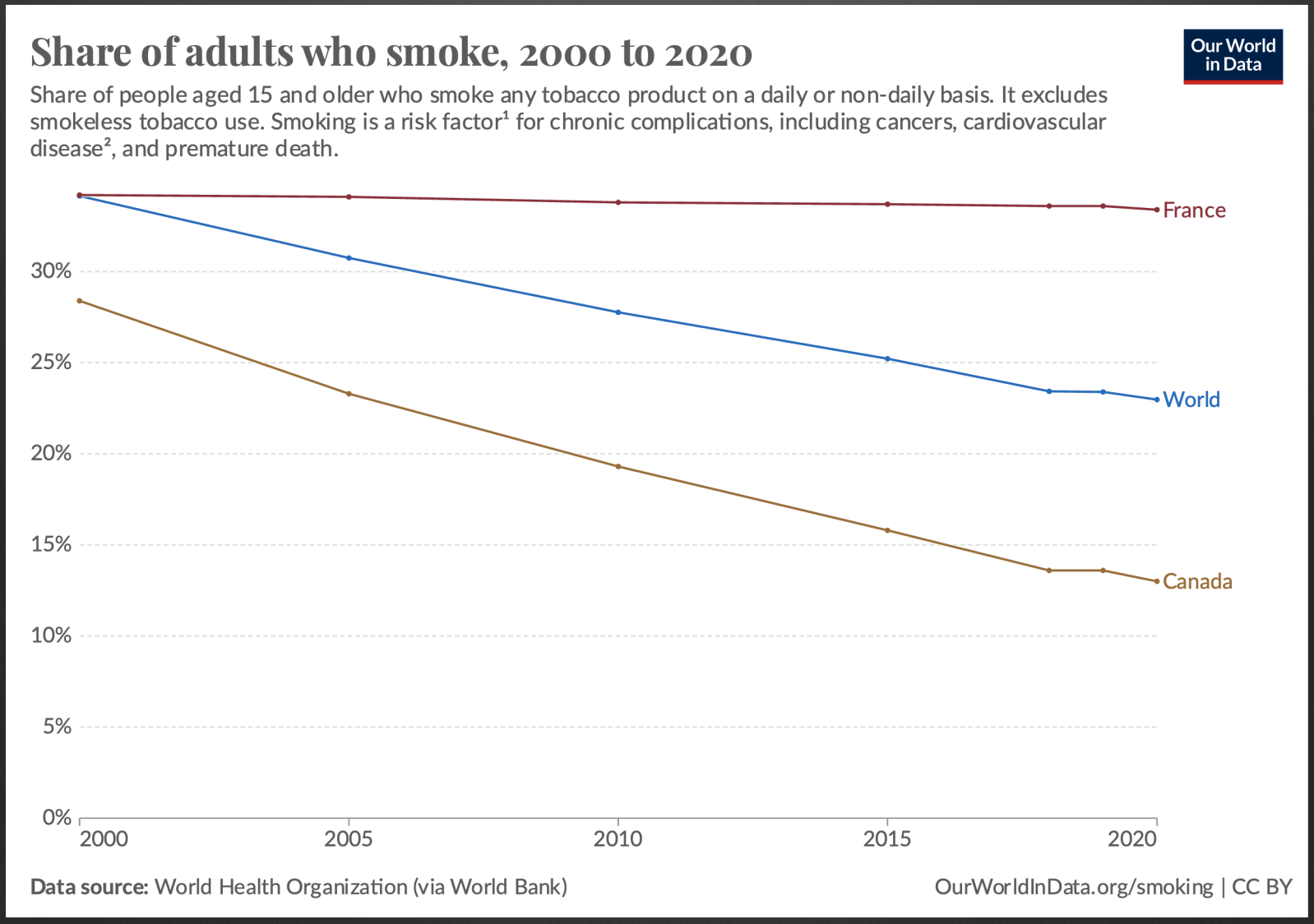
Today

https://scopeblog.stanford.edu/2019/04/30/doctors-smoking-new-exhibit-displays-now-startling-ads/

Cultural change can be slow



Cultural change is variable



Achieving cultural change is common societal challenge

Vaccinations

Seat Belts

Reducing Carbon Emissions

The Challenge of Cultural Change

- Changing culture (any culture) is notoriously difficult, especially in the presence of intense competition
- The existing culture in academia is self-reinforcing
- Behavioral economics research offers effective strategies in achieving cultural change

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

CENTER FOR OPEN SCIENCE

Roadmap

- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next

Insights from Behavioral Economics

- "There is a large body of evidence on what influences behaviour, and we do not attempt to reflect all its complexity and nuances here."

https://www.bi.team/publications/east-four-simple-ways-to-apply-behavioural-insights/

THE BEHAVIOURAL INSIGHTS TEAM

Insights from Behavioral Economics

"We have found that policy makers and practitioners find it useful to have a simple, memorable framework to think about effective behavioural approaches."

Easy
Attractive
Social
Timely

https://www.bi.team/publications/east-four-simple-ways-to-apply-behavioural-insights/

THE BEHAVIOURAL INSIGHTS TEAM

EAST Framework

- Easy - Using defaults, simple messaging, decrease burden
- Attractive - Garner attention, incentives
- Social - Peer awareness, commitments
- Timely - Prompt when most receptive, encourage planning

https://www.bi.team/publications/east-four-simple-ways-to-apply-behavioural-insights/

BIT, UNK

THE BEHAVIOURAL INSIGHTS TEAM

Roadmap

- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next

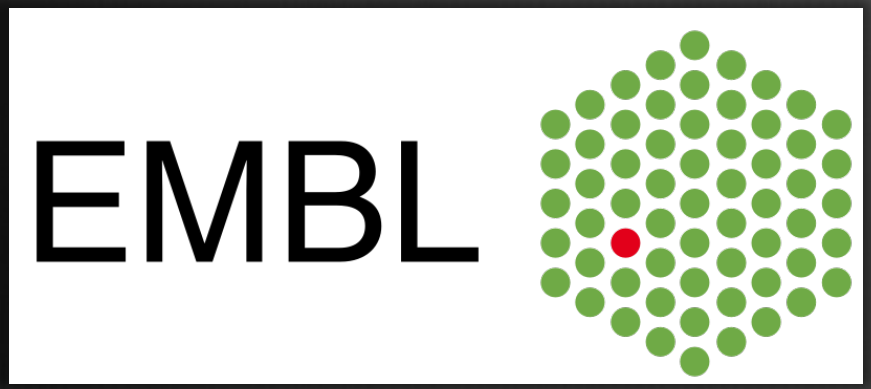
Easy
Attractive
Social
Timely

What Works

- Easy – Data standards and repositories need to be simple and easy to use

The first Data Repository

European Molecular Biology Laboratory (EMBL) Sequence Library

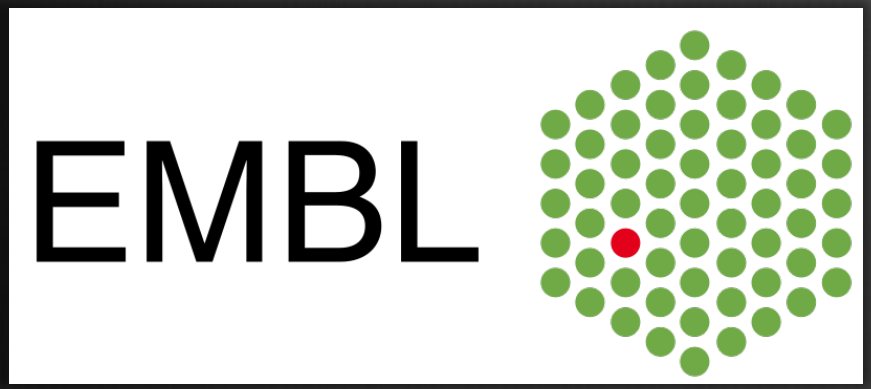


"European Molecular Biology Laboratory (EMBL) at Heidelberg has announced the formation of a nucleotide sequence library, while the NIH in Washington are still deliberating the question."

Walgate (Apr. 1982) *Nature* 10.1038/296596a0

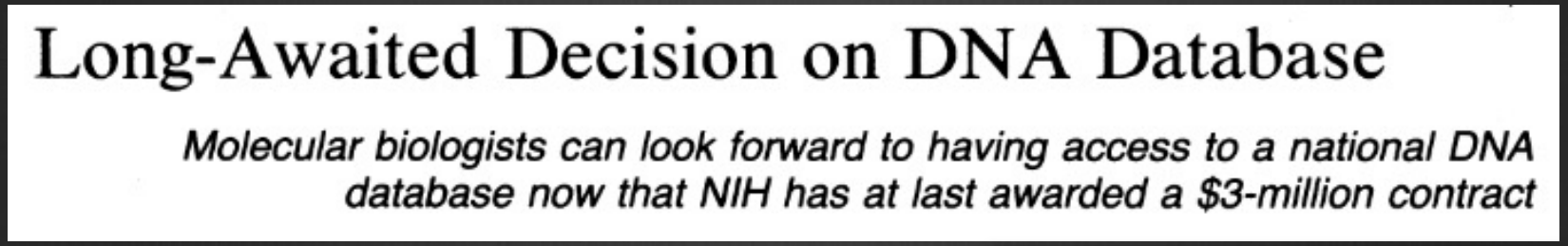
The first Data Repositoryies

European Molecular Biology Laboratory (EMBL) Sequence Library



"European Molecular Biology Laboratory (EMBL) at Heidelberg has announced the formation of a nucleotide sequence library, while the NIH in Washington are still deliberating the question."

Walgate (Apr. 1982) *Nature* 10.1038/296596a0



Long-Awaited Decision on DNA Database

Molecular biologists can look forward to having access to a national DNA database now that NIH has at last awarded a $3-million contract

Lewin (Aug. 1982), 10.1126/science.7100925

The first Data Repositories

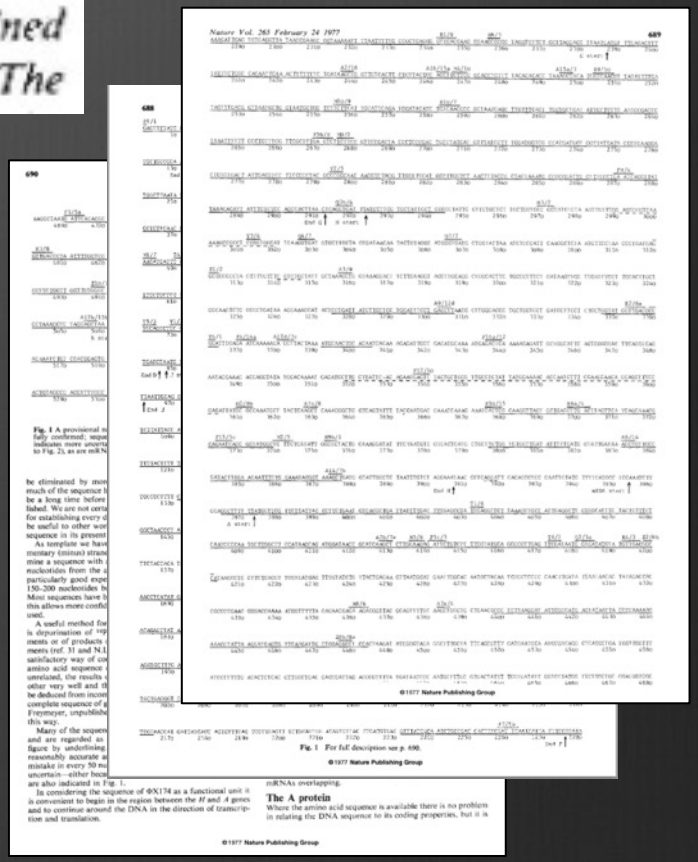
- ”600,000 nucleotides already logged at Heidelberg (they are freely available on magnetic tape)”
- How did they get there?

Walgate (Apr. 1982) *Nature* 10.1038/296596a0

The first Data Repositories

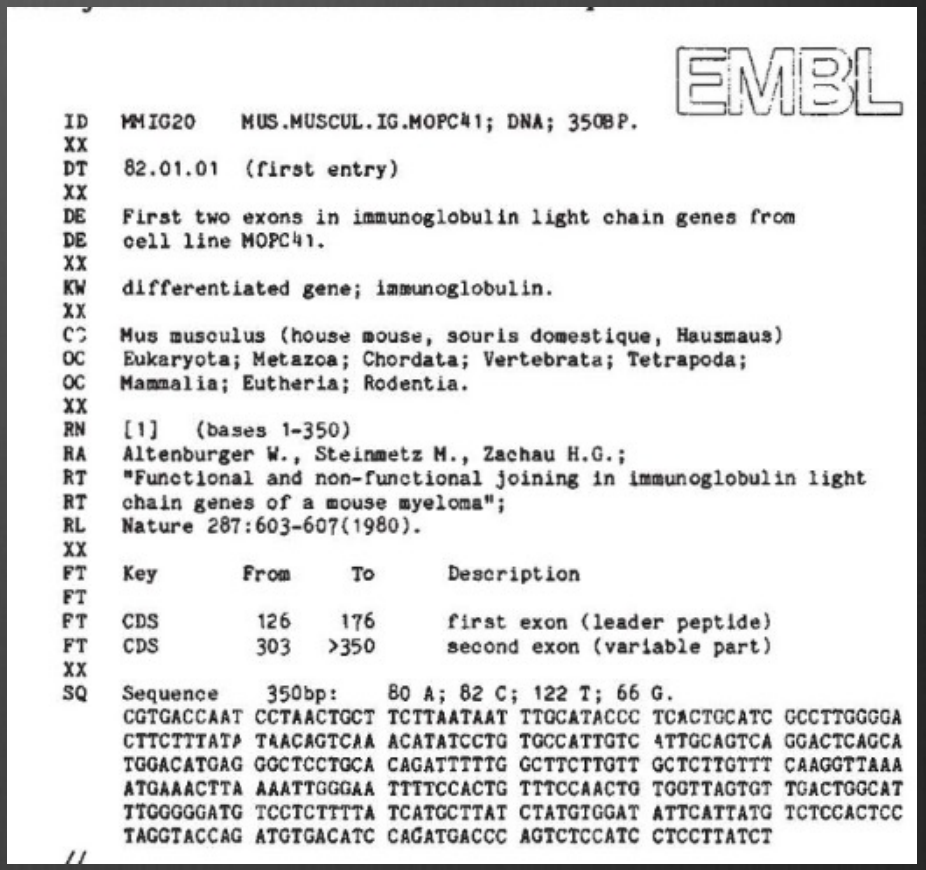
A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The

Sanger et al. (1977) *Nature* 10.1038/265687a0



The First Data Repositories

- Simple data format
- Both human and machine readable
- Similar to the format still used today



EMBL

```
ID   MMIG20    MUS.MUSCUL.IG.MOPC41; DNA; 350BP.
XX
DT   82.01.01  (first entry)
XX
DE   First two exons in immunoglobulin light chain genes from
DE   cell line MOPC41.
XX
KW   differentiated gene; immunoglobulin.
XX
OS   Mus musculus (house mouse, souris domestique, Hausmaus)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda;
OC   Mammalia; Eutheria; Rodentia.
XX
RN   [1]   (bases 1-350)
RA   Altenburger W., Steinmetz M., Zachau H.G.;
RT   "Functional and non-functional joining in immunoglobulin light
RT   chain genes of a mouse myeloma";
RL   Nature 287:603-607(1980).
XX
FT   Key       From     To       Description
FT
FT   CDS         126    176      first exon (leader peptide)
FT   CDS         303   >350      second exon (variable part)
XX
SQ   Sequence    350bp:    80 A; 82 C; 122 T; 66 G.
     CGTGACCAAT CCTAACTGCT TCTTAATAAT TTGCATACCC TCACTGCATC GCCTTGGGGA
     CTTCTTTATA TAACAGTCAA ACATATCCTG TGCCATTGTC ATTGCAGTCA GGACTCAGCA
     TGGACATGAG GGCTCCTGCA CAGATTTTTG GCTTCTTGTT GCTCTTGTTT CAAGGTTAAA
     ATGAAACTTA AAATTGGGAA TTTTCCACTG TTTCCAACTG TGGTTAGTGT TGACTGGCAT
     TTGGGGGATG TCCTCTTTTA TCATGCTTAT CTATGTGGAT ATTCATTATG TCTCCACTCC
     TAGGTACCAG ATGTGACATC CAGATGACCC AGTCTCCATC CTCCTTATCT
//
```

Walgate (Apr. 1982) *Nature* 10.1038/296596a0

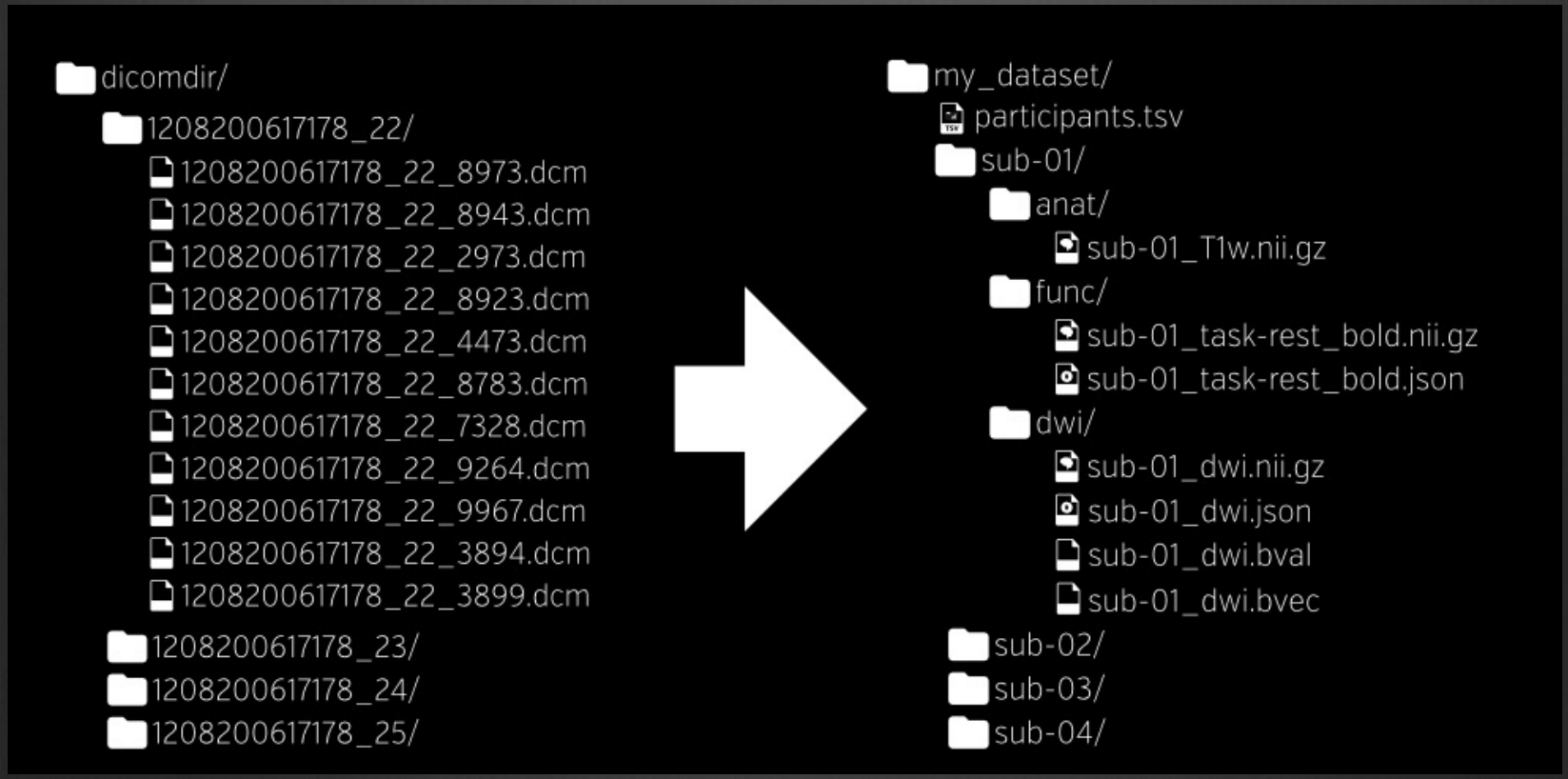
The First Data Repositories

- Simple data format
- Both human and machine readable
- Similar format still used today
- 4.5 Billion records (Apr 2024)

https://www.ncbi.nlm.nih.gov/genbank/samplerecord/

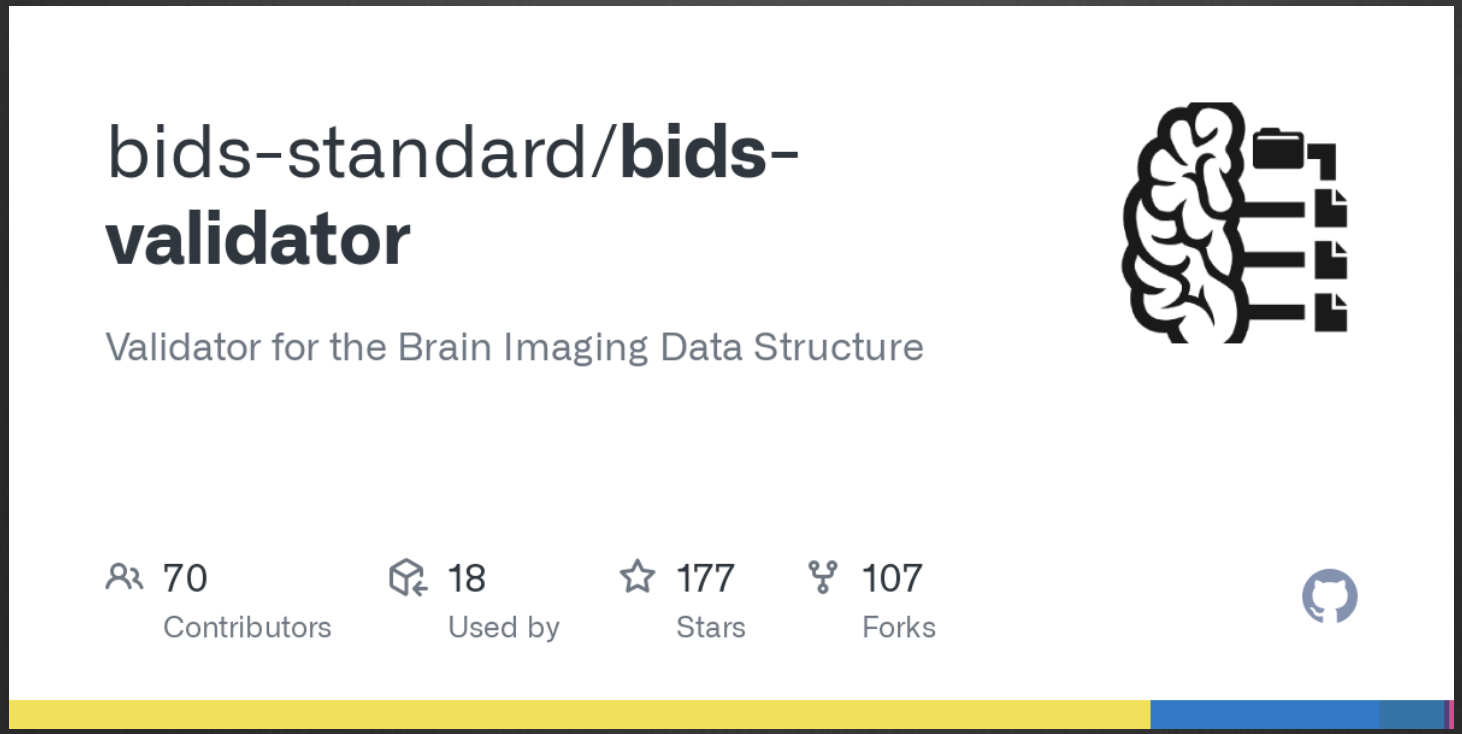
```
LOCUS       SCU49845               5028 bp    DNA     linear   PLN 29-OCT-2018
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (brewer's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
   PUBMED   8846915
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Biology, Yale University, New Haven, CT
            06520, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
     mRNA            <1..>206
                     /product="TCP1-beta"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            <687..>3158
                     /gene="AXL2"
     mRNA            <687..>3158
                     /gene="AXL2"
                     /product="Axl2p"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                     DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                     DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                     NATSTRSSHHSTSTSSYTSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                     NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                     SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
     gene            complement(<3300..>4037)
                     /gene="REV7"
     mRNA            complement(<3300..>4037)
                     /gene="REV7"
                     /product="Rev7p"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat accttcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
```

Brain Imaging Data Structure (BIDS)

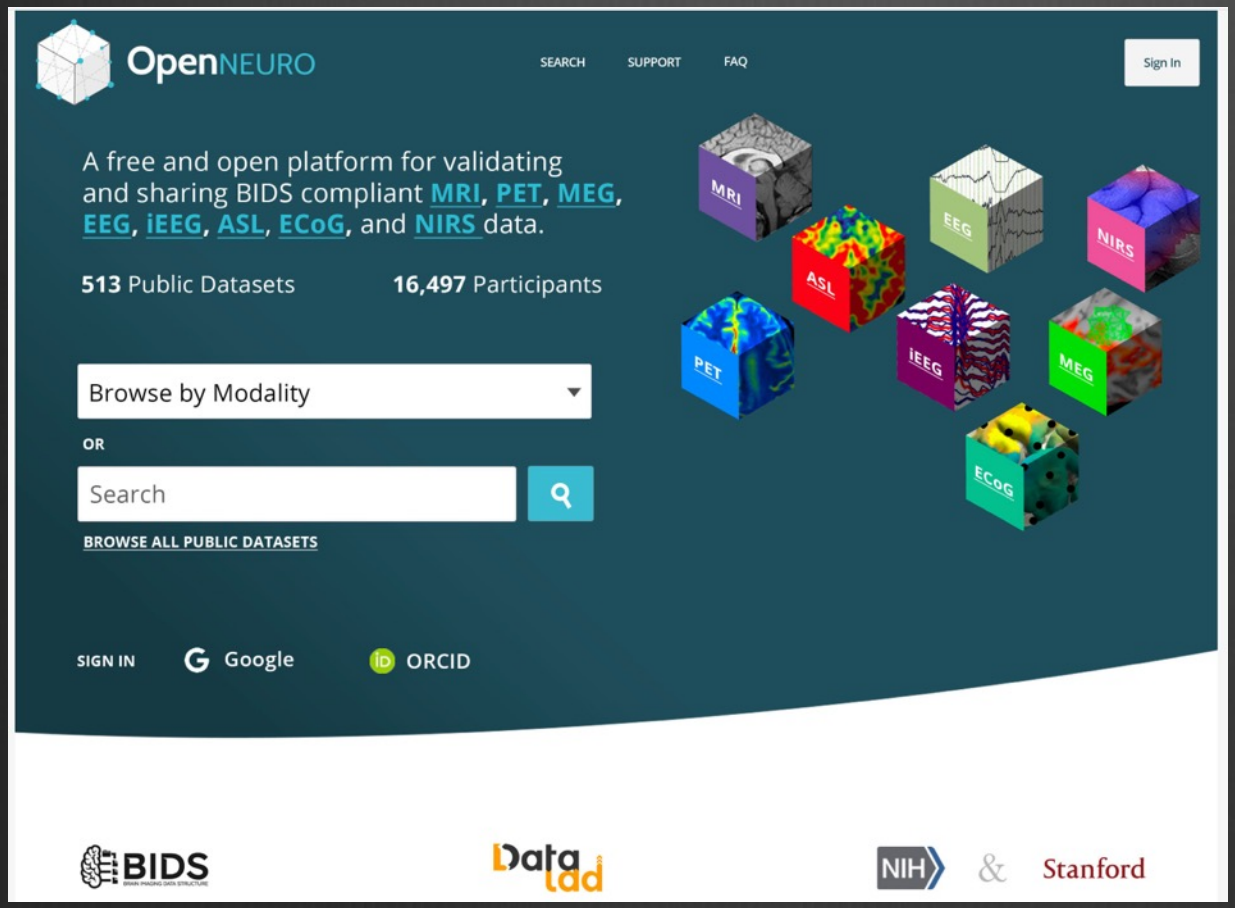


Brain Imaging Data Structure (BIDS)

- Simple data format
- Both human and machine readable
- Automated curation

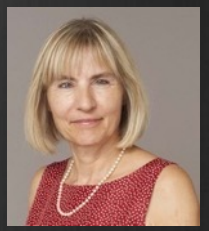


Low-cost repositories: OpenNeuro & OpenNeuroPET



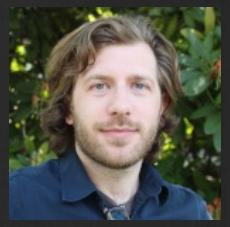
Robert Innis

Melanie Ganz



Gitte Knudsen

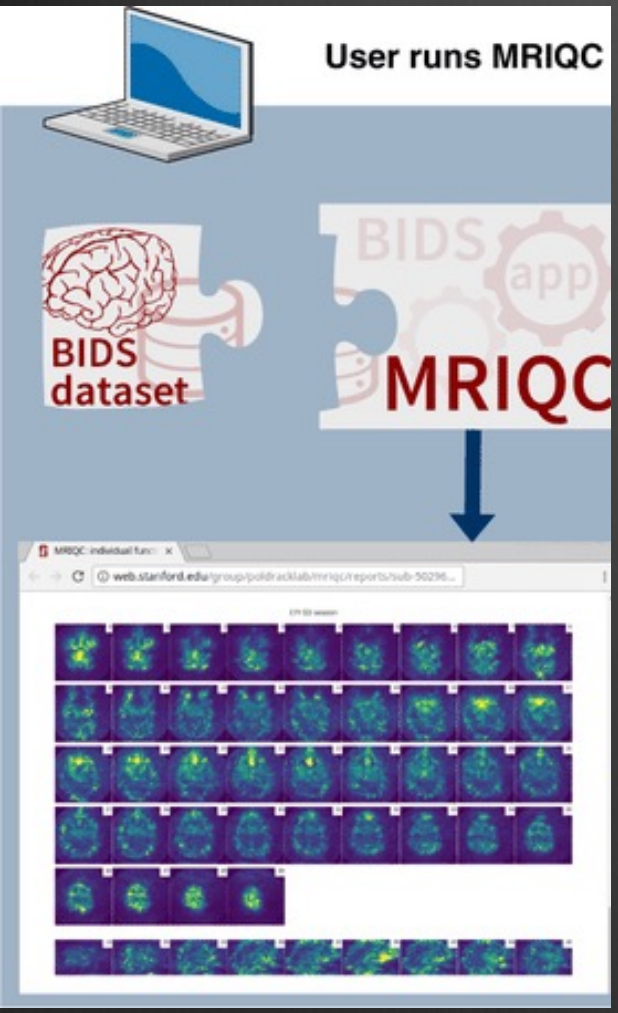
Dylan Nielson



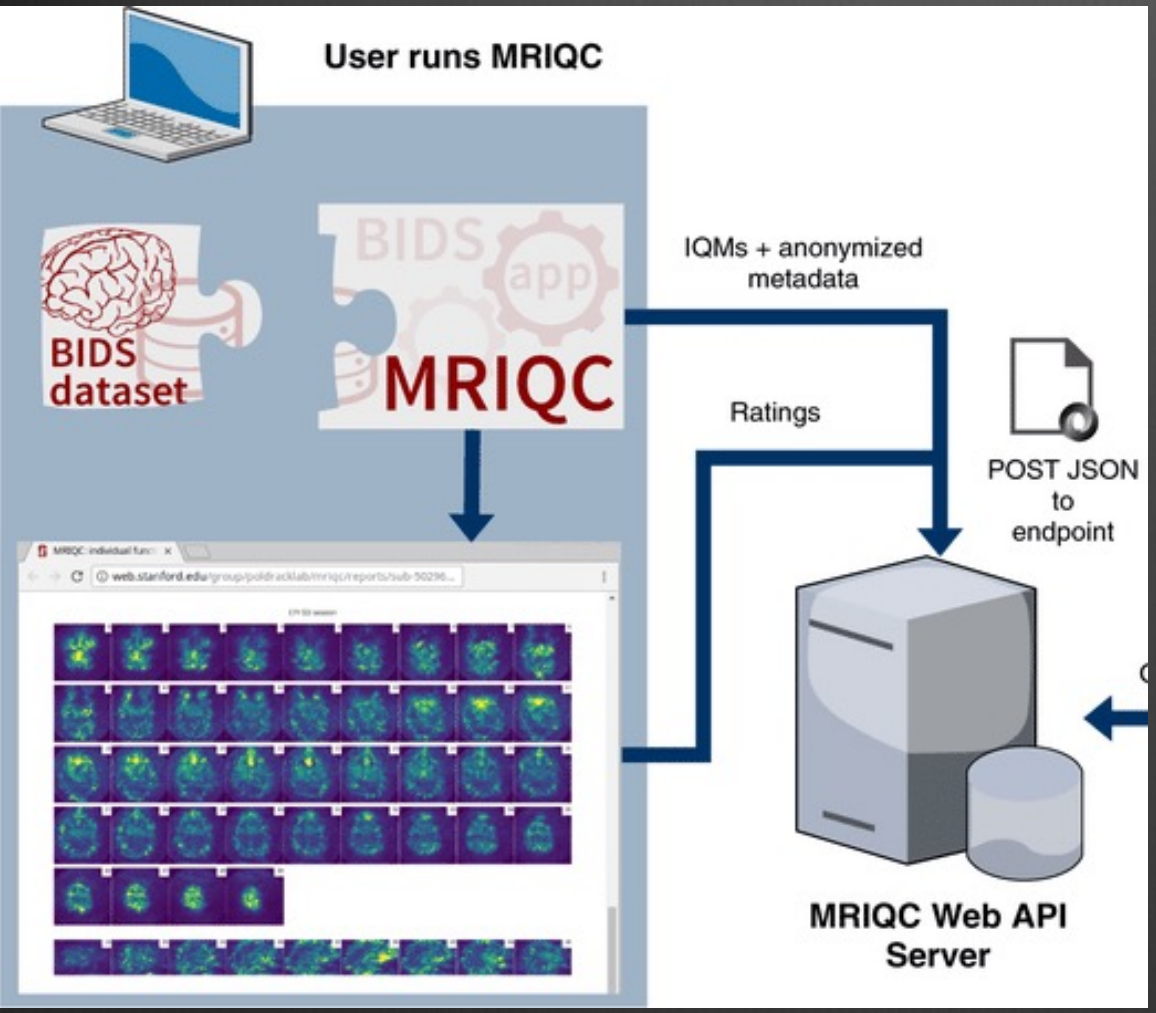
Anthony Galassi

https://openneuropet.github.io/

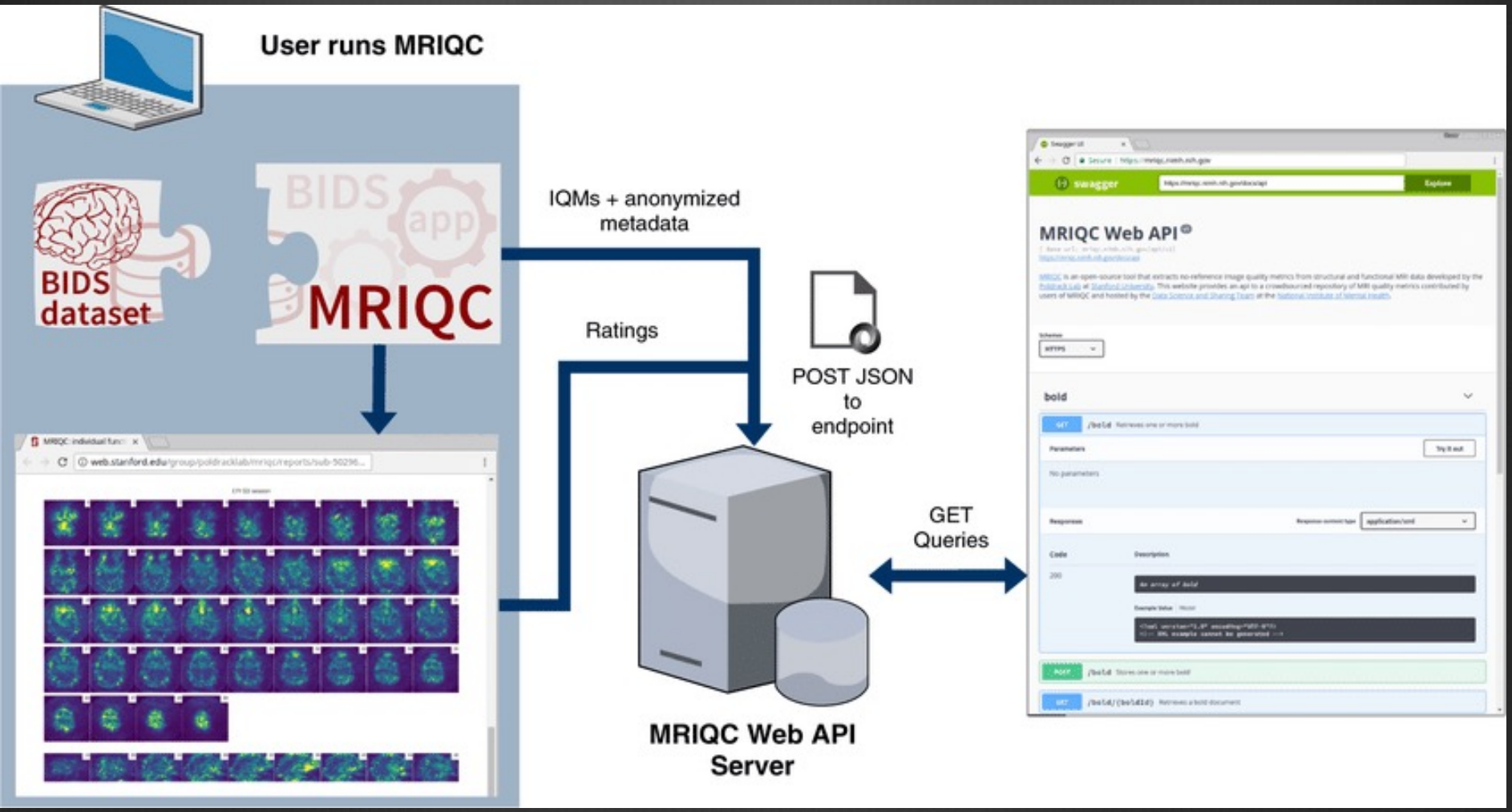
Automation & Aggregation of Quality Metrics



Automation & Aggregation of Quality Metrics



Automation & Aggregation of Quality Metrics



Automating QC: MRIQC

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

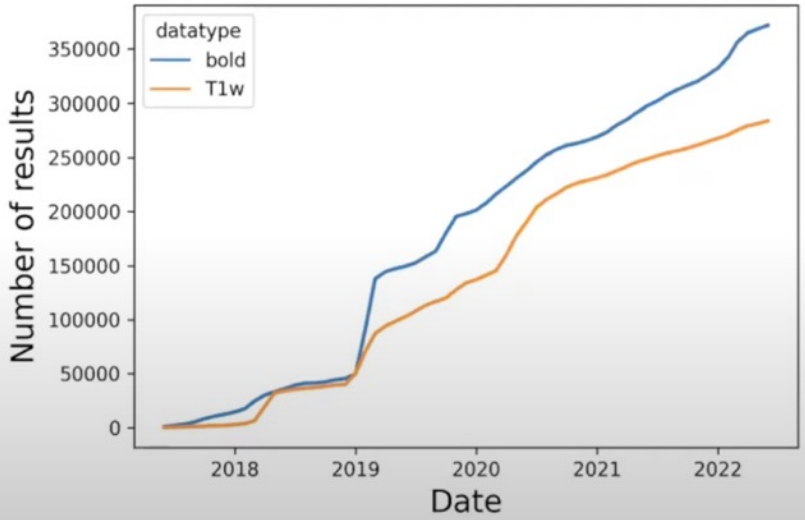
OPEN

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]



What Works

- Easy - Simple data standards
- Social &
- Attractive - Identify & reward sharing

Identify and Reward Sharing

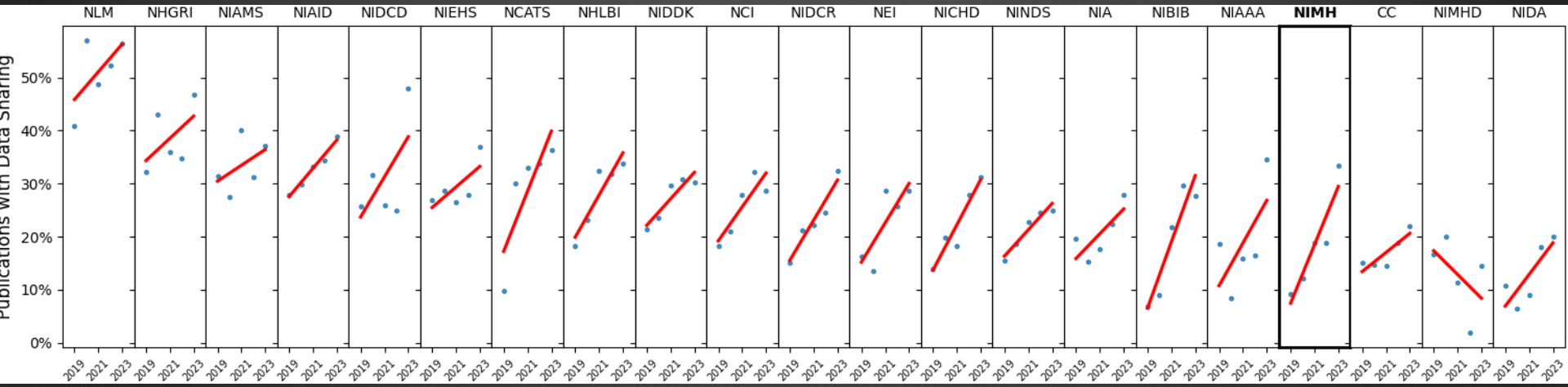
- To create incentives, funders and institution need metrics
- Text mining tools exist to identify data sharing reported in publications
- However, aggregate information on reported data sharing is not easily accessible to funders and policy makers

~25,000 Unique Pubs 2019-2023

...

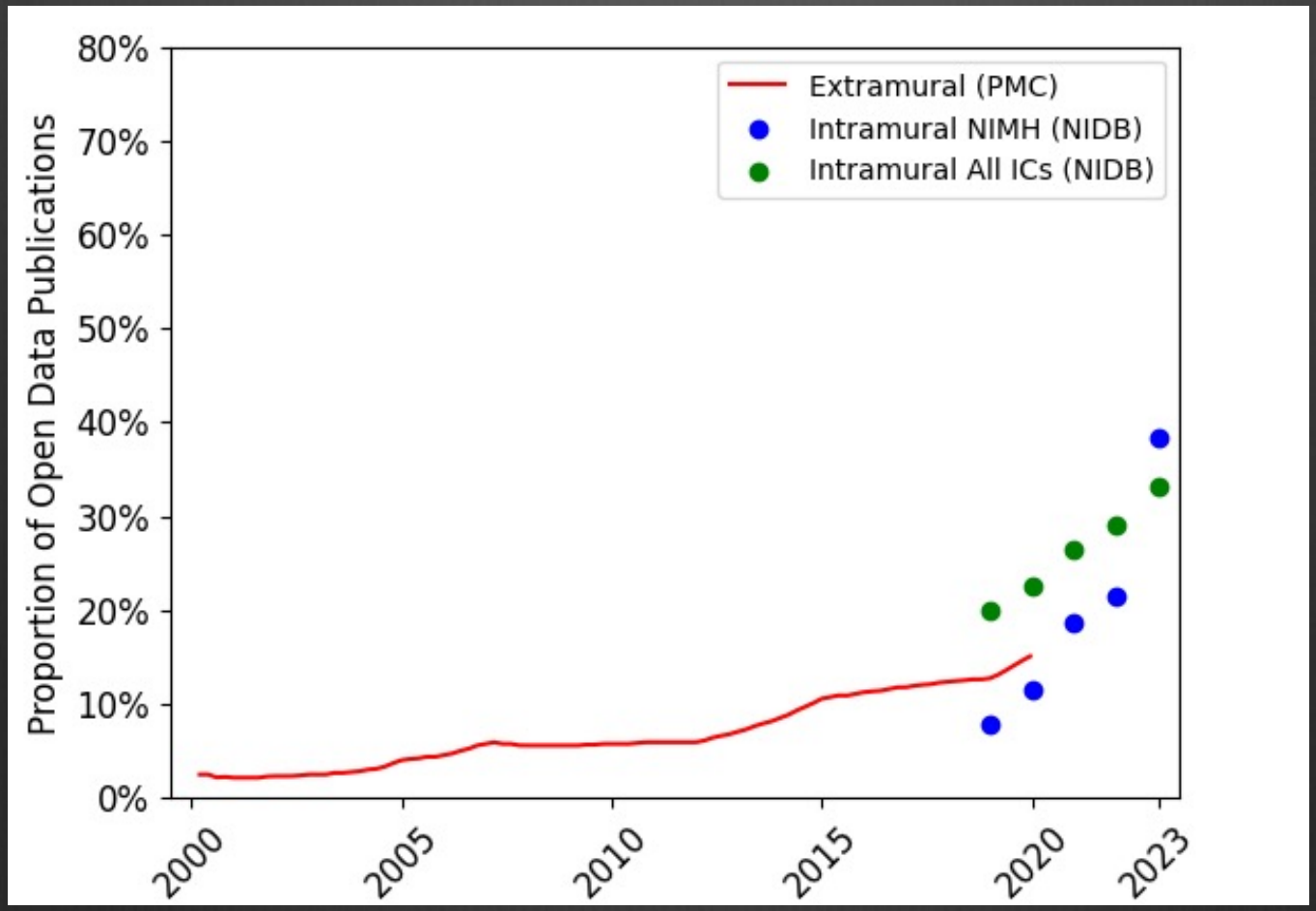
Measuring Data Sharing

NIMH Intramural vs All of NIH Intramural



Measuring Data Sharing

NIMH Intramural vs All of NIH Intramural

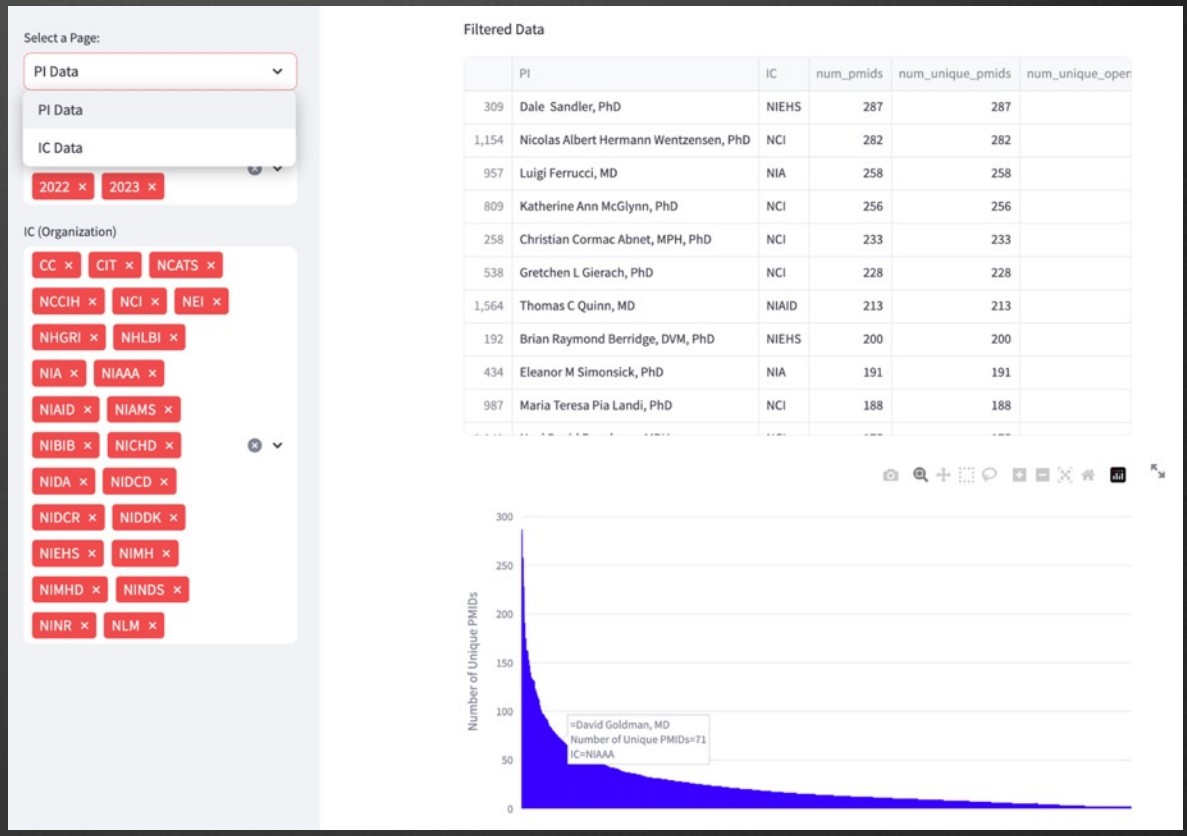


Measuring Data Sharing

- 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



...

What Works

- Easy - Simple data standards Easy
- Social &
  Attractive - Identify and reward active
- Timely
  - Data Management Sharing Plans
  - Manuscript Clearance
  - New trainees

What Works

- Easy - Simple data standards Easy
- Social - Supporting use of shared data
- Attractive - Identify and reward active
- Timely
  - Data Management Plans
  - Manuscript Clearance
  - New trainees

Roadmap

- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next

Easy
Attractive
Social
Timely

What Doesn't Work

- "Build it and they will come"
  - Attempt to address all problems in cancer research
  - One size fits all approach
  - Lack of independent scientific oversight of goals

AN ASSESSMENT OF THE IMPACT OF THE NCI CANCER BIOMEDICAL INFORMATICS GRID (caBIG®)

Report of the Board of Scientific Advisors Ad Hoc Working Group

https://deainfo.nci.nih.gov/advisory/bsa/archive/bsa0311/caBIGfinalReport.pdf

March 2011

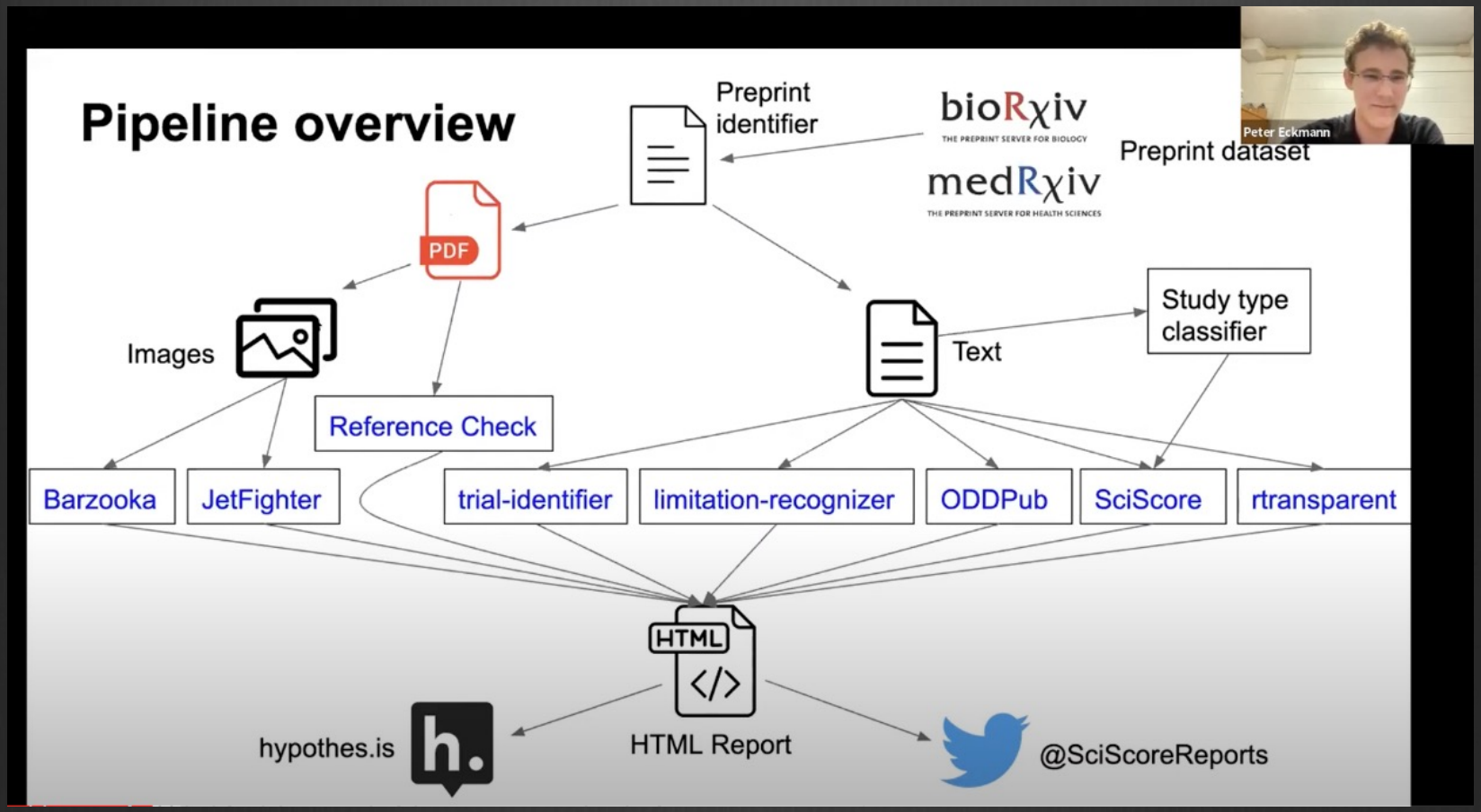
What Doesn't Work

- "Build it and they will come"
  - Attempt to address all problems in cancer research
  - One size fits all approach
  - Lack of independent scientific oversight of goals
- Contrast with the EMBL Sequence Library

> "[the] manager of the Heidelberg library, is "still talking" to NIH, and wishes to cooperate with any system that NIH may set up. But EMBL was under pressure from European scientists to start now, before sequence data were irretrievably lost. EMBL is not attempting to become the sole manager of world sequence data."

Lessons learned from automated screening

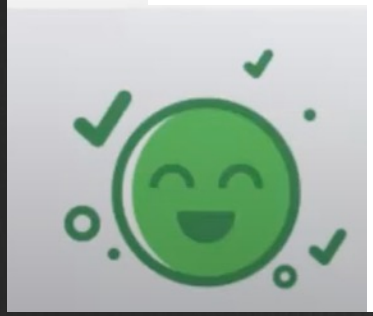
Lessons learned from automated screening



Lessons learned from automated screening

User feedback:

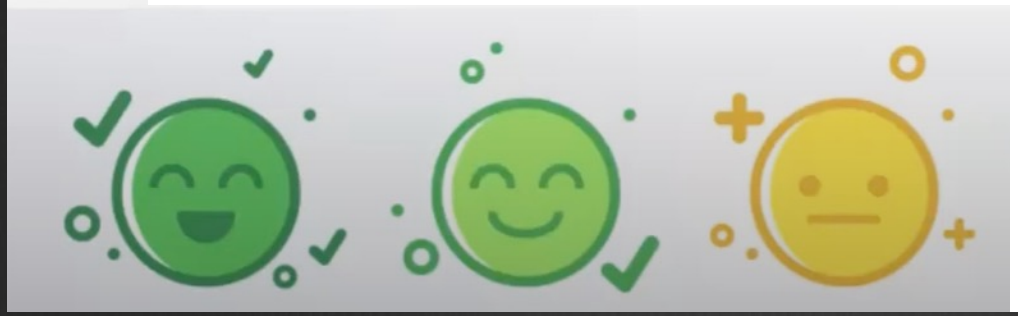
- Thanks for checking my paper



Lessons learned from automated screening

User feedback:

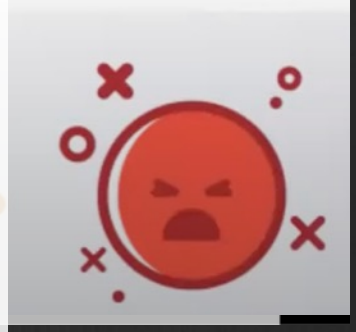
- Thanks for checking my paper
- You have missed something



Lessons learned from automated screening

User feedback:

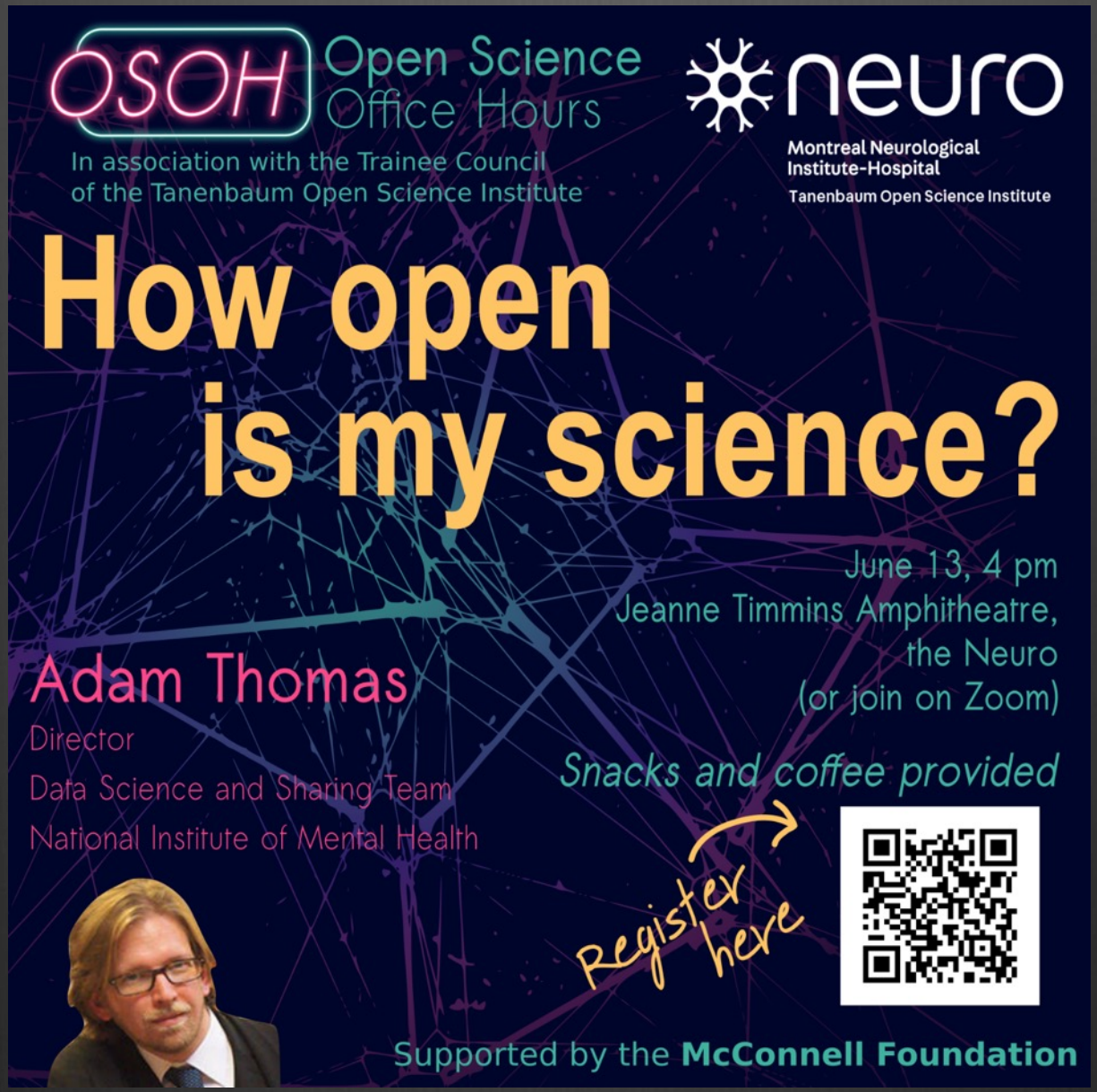
- Thanks for checking my paper
- You have missed something
- I can't believe you dare to read my paper using some stupid bot!



Lessons learned from automated screening

- Most authors are acting in good faith
- All automated tools will make errors
- Many authors will want to correct the manuscript or the tool's error
- Potential to create ill will

https://youtu.be/581NkRV4mbA

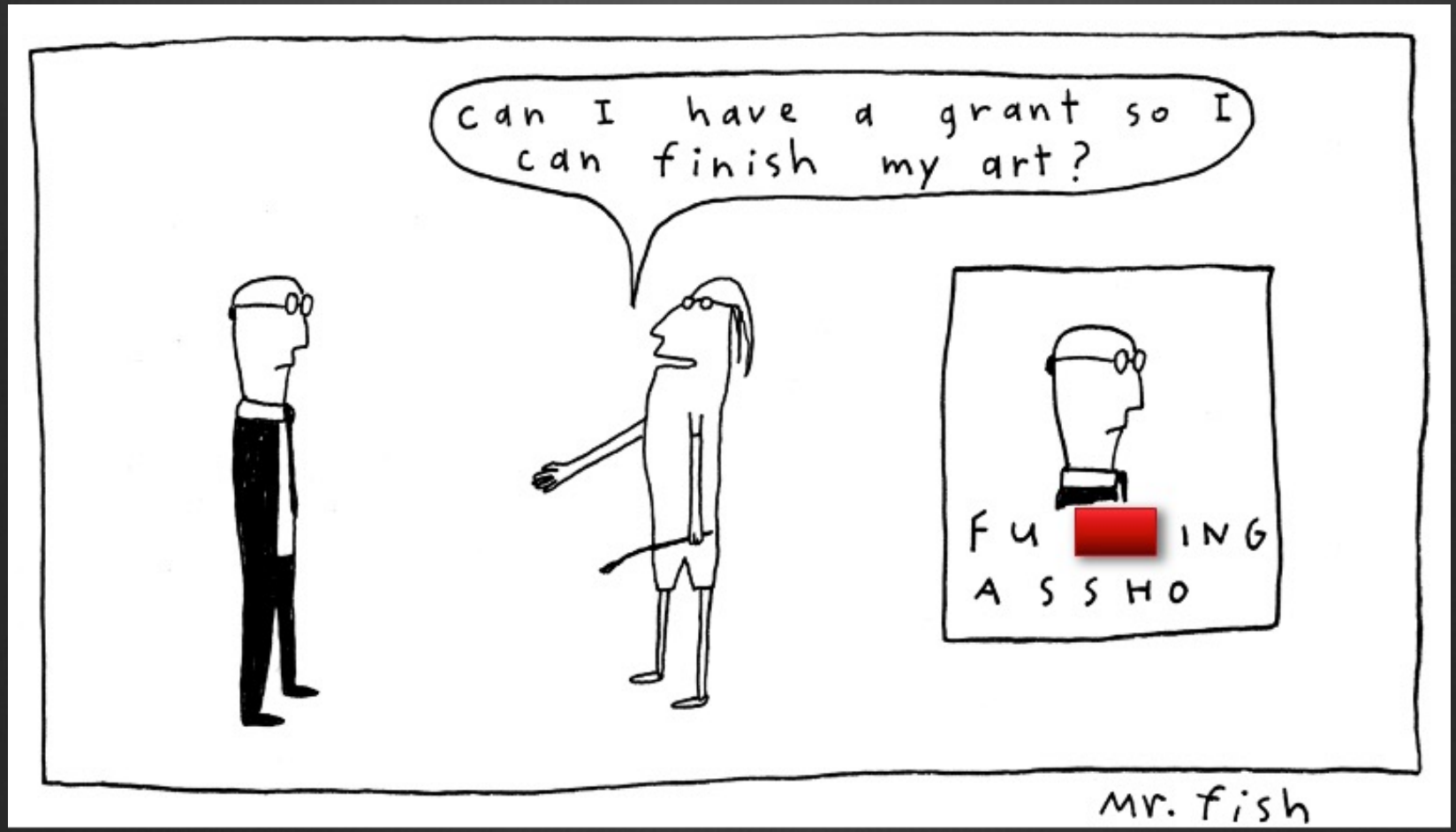


https://www.mcgill.ca/neuro/channels/event/open-science-office-hours-how-open-my-science-357449

Lessons Learned ≠ Didn't Work

- caBIG ultimately led to smaller successful programs
- ScreenIT was entirely successful
- Both were open and transparent with "Lessons Learned"

The projects we'll never hear about



Support Mr. Fish! https://clowncrack.com/about/

What Doesn't Work

- Top-down approaches lacking community involvement ~~Easy, Social~~
- Naming & shaming ~~Attractive~~
  - (Especially if you are shaming your institution or funder)

Roadmap

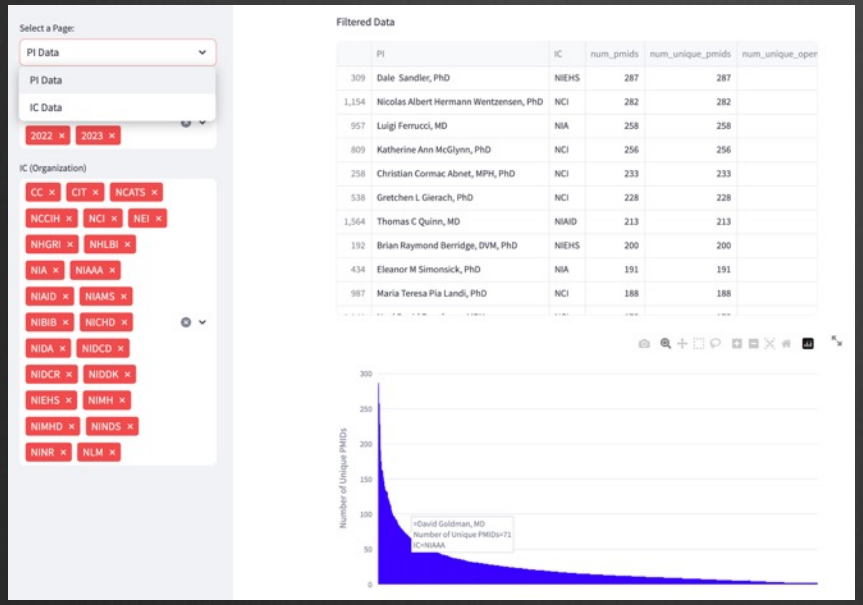
- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next (Opportunities)

Improving Open Science Metrics

~25,000 Unique Pubs 2019-2023

Open Data Detection in Publications (ODDPub)

build error | codecov 90% | License MIT | DOI 10.5281/zenodo.4071699



~10% error rate, equally balanced between false positive and false negative

...

Improving Open Science Metrics

- LLM-based metrics
  - Improved accuracy
  - 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

...

Improving Open Science Metrics

- LLM-based metrics
  - Improved accuracy
  - Extracting more specific details from pubs
  - More nuanced evaluations

Many Open Science Rubrics require judgement

- 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 require judgement

- 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

---

LLMs can answer Nuanced Open Science Question Today

Link to Shared Data	Does the dataset include detailed provenance?	Does the metadata meet domain-relevant community standards?
- Online Participants' data: https://osf.io/km69z - Mobile App Participants' data: https://doi.org/10.5061/dryad.prr4xgxkk	Yes, the dataset includes detailed provenance as the data analysis software, data from participants, and code used in the study are all available on public platforms.	Yes, the metadata meets domain-relevant community standards as it provides detailed information on data and code availability, uses recognized repositories and platforms for sharing, and includes a clear data availability statement along with a declaration of no competing interests.

Elicit

But are the answers accurate?

Validation requires 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

Kendra Oudyk

JB Poline

Improving Open Science

- LLM-based metrics
  - Improved accuracy
  - Extracting more specific details from pubs
  - More nuanced evaluations
- Make metrics uniform and more accessible

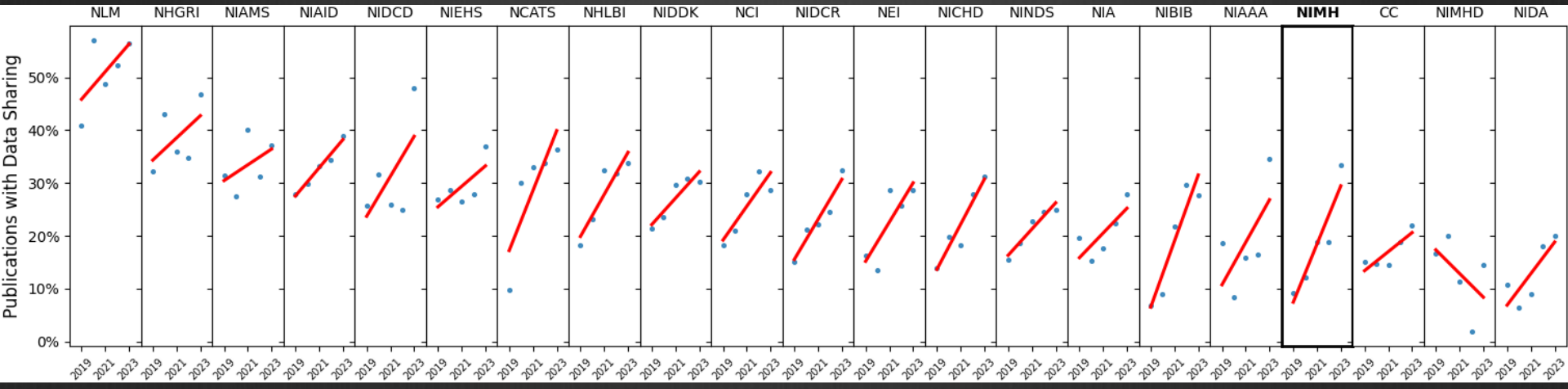
Similar Efforts

- 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

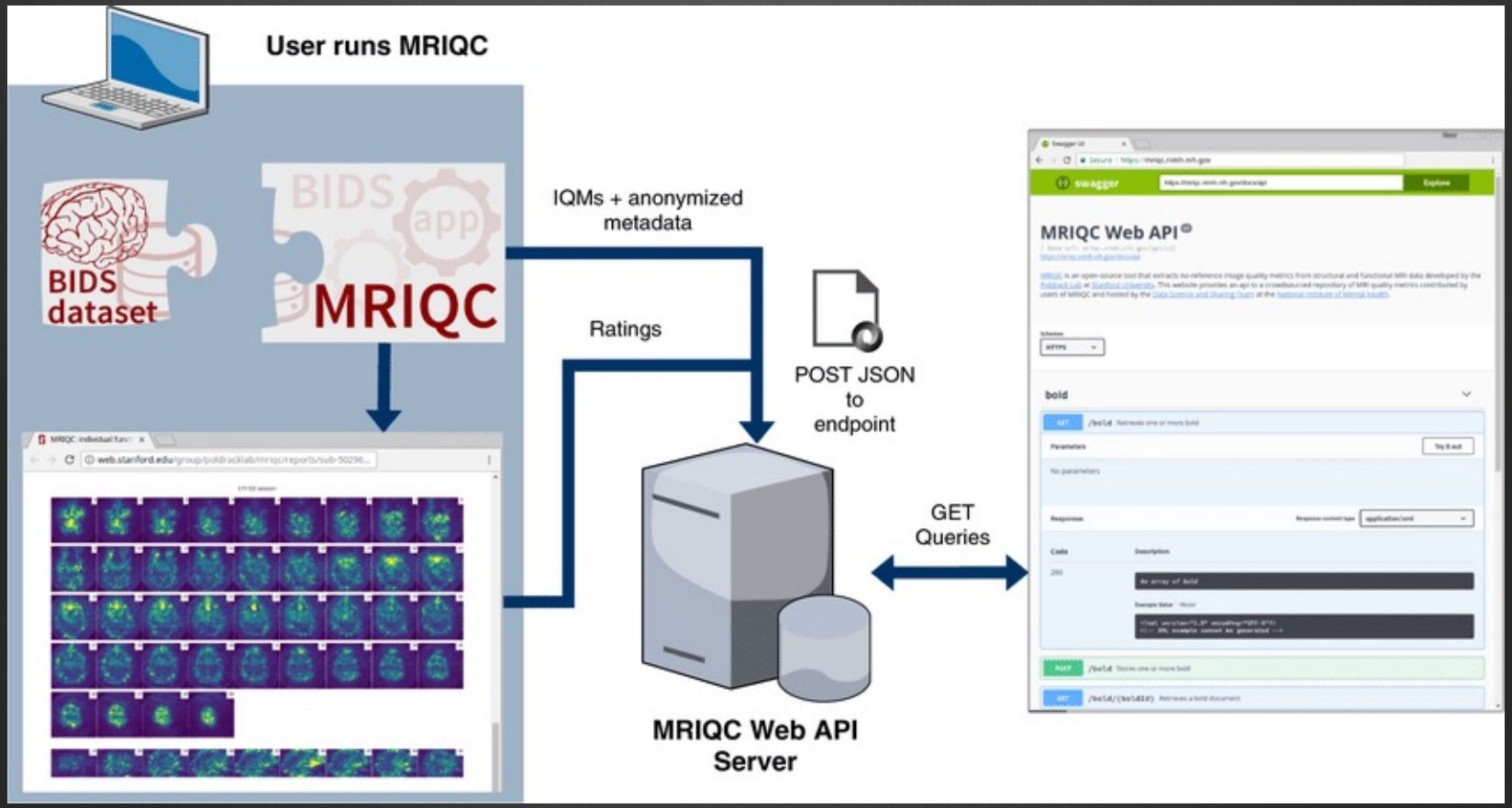
Measuring Data Sharing

- 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

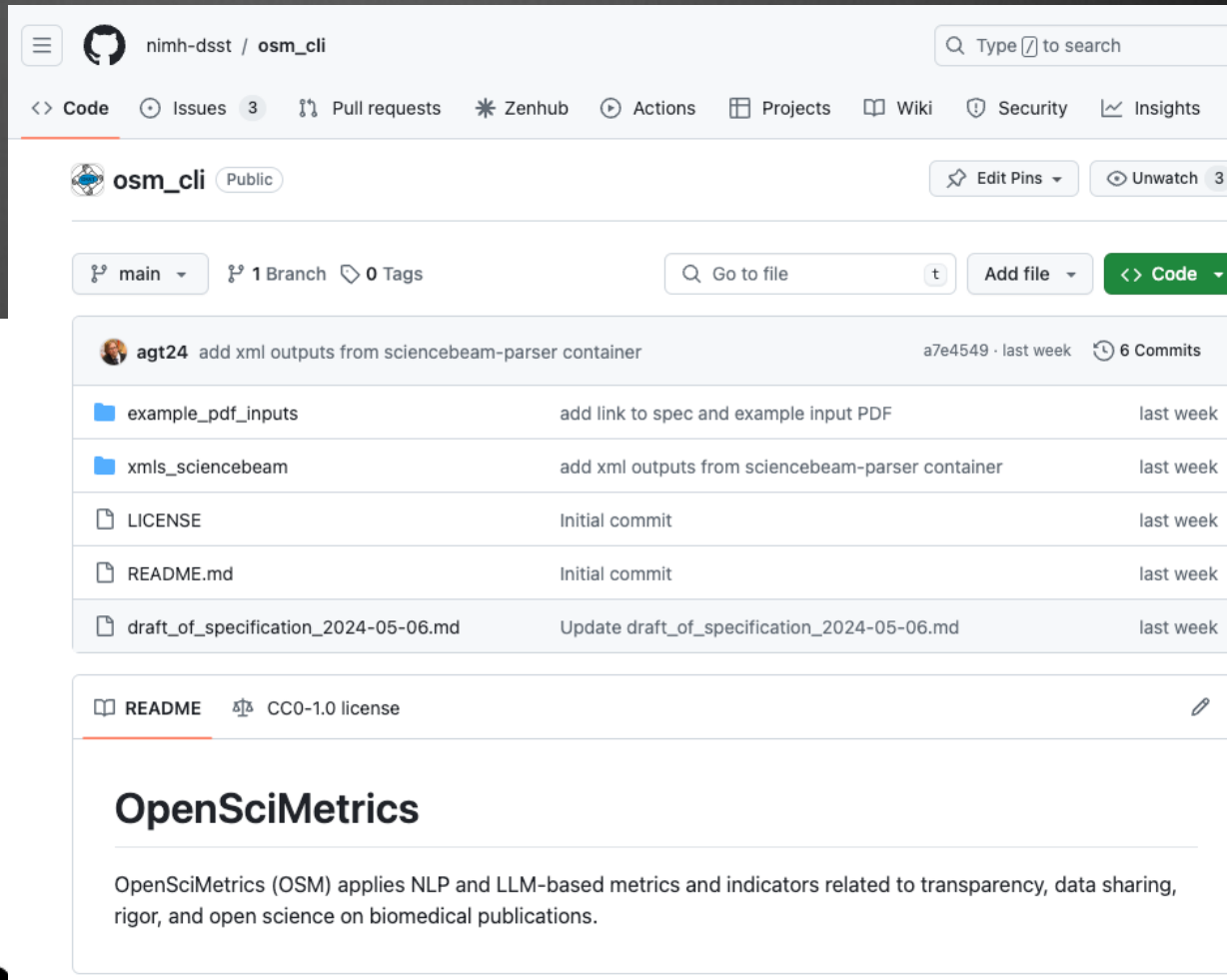
~25,000 Unique Pubs
2019-2023



A tool modelled on previous success



Open Science Metrics DB



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.

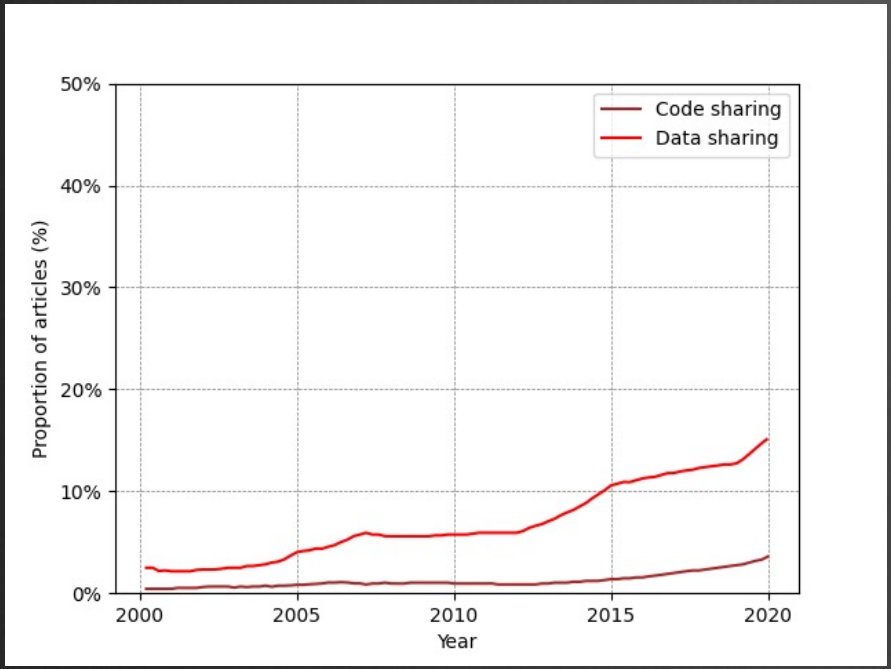
Improving Open Science

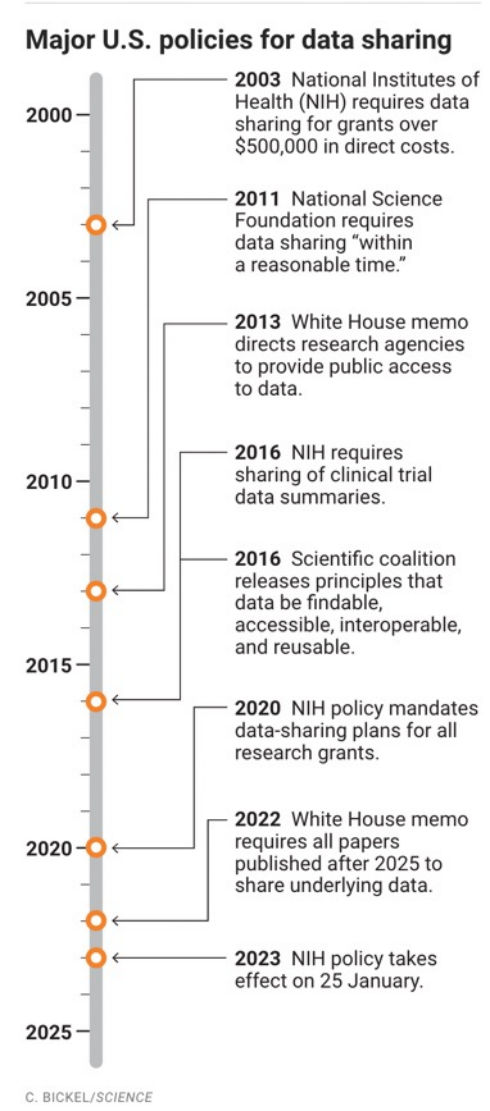
- LLM-based metrics
  - Improved accuracy
  - Extracting more specific details from pubs
  - More nuanced evaluations
- Make metrics uniformly available and more accessible: Open Science Metrics DB

Roadmap

- Who am I? (who I'm not)
- History of Data Sharing
- What can we learn from behavioral economics?
- How can we foster Cultural Change?
  - What works
  - What doesn't
  - What's next (Opportunities)
  - Don't lose hope

Don’t lose hope!





DOI: 10.1126/science.adg8142

Don’t lose hope!

- Technology Adoption Life Cycle

https://www.linkedin.com/pulse/technology-adoption-life-cycle-caio-felipe-santos-matos/

Don’t lose hope

- Technology Adoption Life Cycle
- Antidotes to cynicism creep in academia
  - Focus on the progress
  - Tune in: The revolution *is* televised! (or at least tweeted)
  - Teach!

Eiko Fried

Antidotes to cynicism creep in academia

Eiko March 4, 2024 30 Comments

https://eiko-fried.com/antidotes-to-cynicism-creep/

Thank you!

The Team

Dustin Moraczewski

Arshitha Basavaraj

Jessica Dafflon

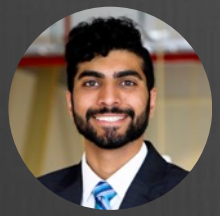
Eric Earl

JB Poline

Trainees

Mia Zwally (current)

Carl Harris
Johns Hopkins
PhD Candidate



Nikhil Goyal
UPenn, Medical Student

Nino Migineishvili
UWash, Seattle
PhD Candidate

Former DSST members

Dylan Nielson
John Lee
Travis Riddle

External DSST developers

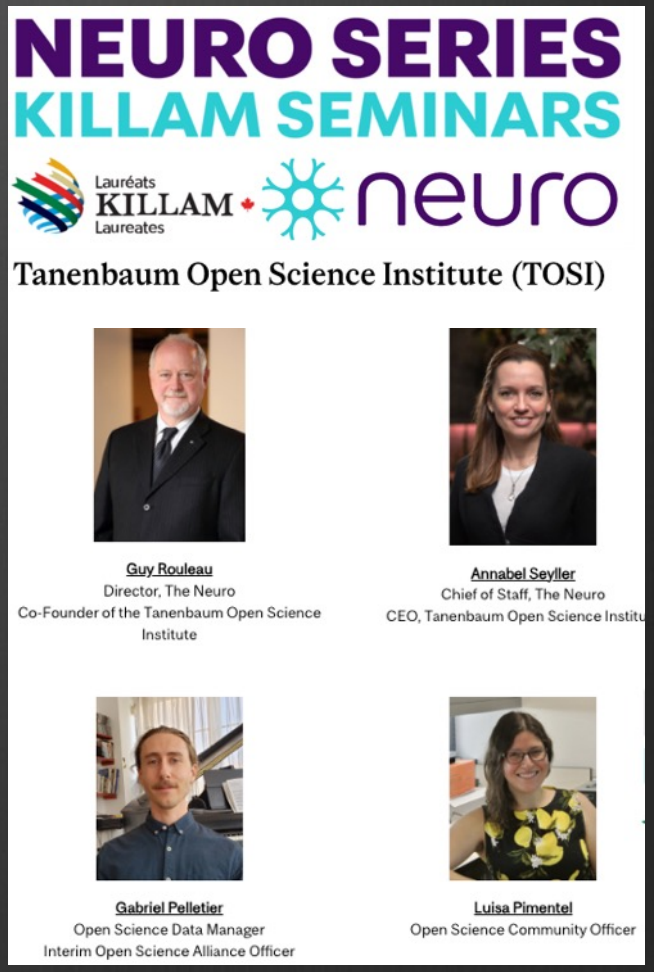
David Seiband
Leonardo Resende

OpenNeuroPET

Anthony Galassi
Bob Innis
Gitte Knudsen
Melanie Ganz
Cyril Pernet
Martin Nørgaard

Machine Learning Team

Francisco Pereira
Charles Zheng
Ka Chun Lam
Antonio Lossio-Ventura
Gabriel Loewinger
Yuan Zhao
Dylan Nielson
Aria Wang
Al Xin

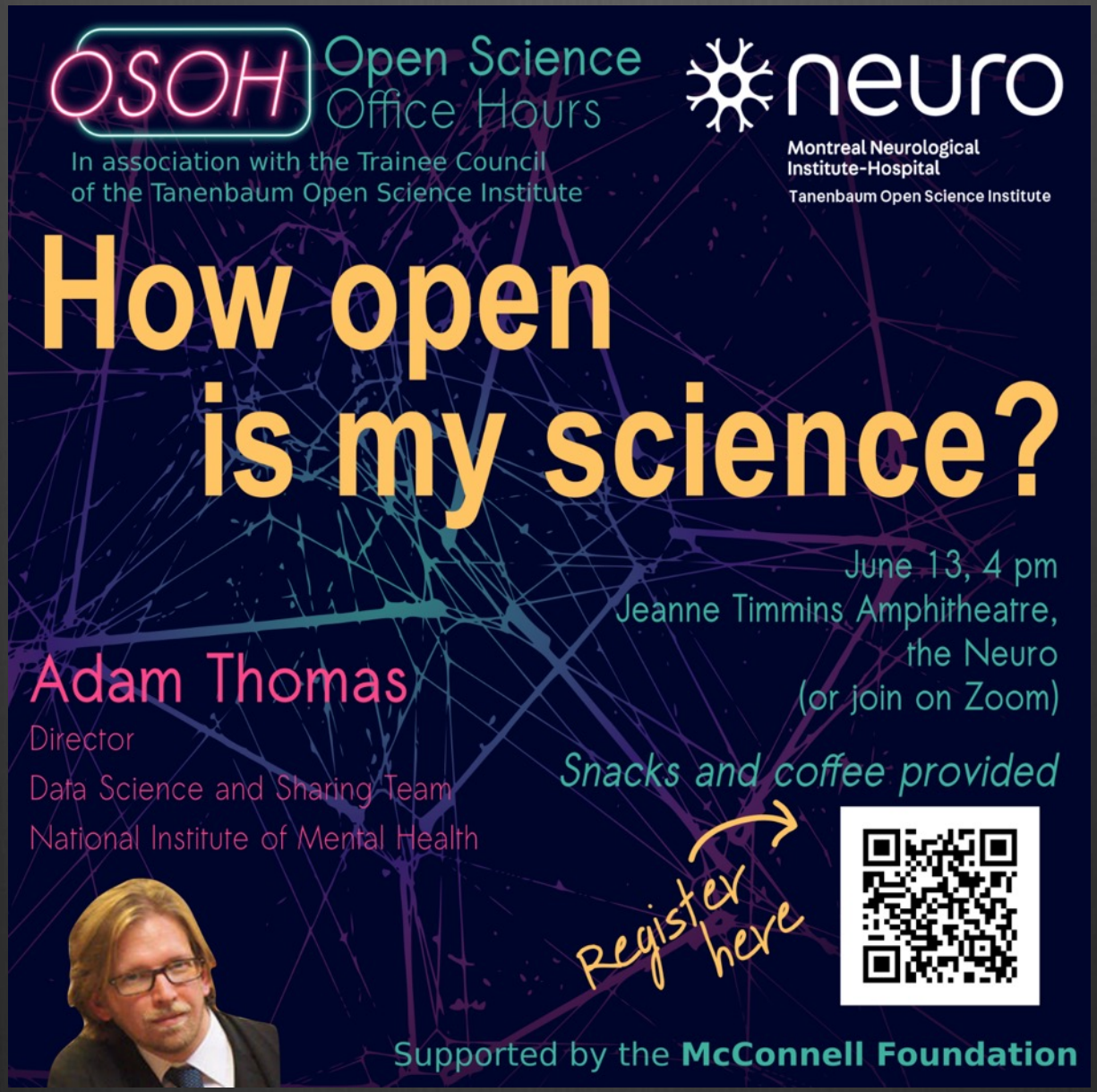


Take Homes

- Adoption of data sharing and other open science practices is happening, but slowly
- We can accelerate adoption by applying principles of behavioral economics to foster cultural change
- Improving access aggregated to diverse open science metrics will help expand incentives

- I want to talk to you! (esp. trainees)

http://bit.ly/adamt_osrp



https://www.mcgill.ca/neuro/channels/event/open-science-office-hours-how-open-my-science-357449