Data sharing analysis

This document will detail ongoing efforts to quantify the datasharing ecosystem in biomedical science, and to understand how datasharing impacts the impact and productivity of a scientific team.

To do this work, we pulled full-text papers from pubmedcentral (PMC) using an NIH-maintained api and grant details from FederalExporter. Although PMC has data going back to the 18th century, and Federal Exporter back to 2004, our analysis is focused on 2009 onwards, as data sharing did not become widely practiced until the late 00's. We selected 2009 as a start point because Github, one of the most popular repositories was founded the year prior.

Table @ref(tab:papes-and-projs) below shows a count of papers and funded projects by year, along with totals across all years. Note that the funded projects include grants made through many agencies that fund work that do not necessarily lead to scientific publications. Additionally, the publication year for articles is based on when they were *posted* to PMC, which is not necessaerily the same thing as the date that it is published by the journal. Occasionally, articles are published in PMC before the journal, but more frequently, they are published in PMC after a dealay

In table @ref(tab:n-projs), we highlight the journals that appear most often in this database, along with the IC Centers that have the largest number of projects.

Figure **??** highlights the date disparity between publication date on PMC and journal publication date. Cells that are in white have no observations in them. There are a handful of papers that were published before 2000 but first appeared in PMC in 2009 (indeed, the earliest such paper has a publication date of 1948.)

Because these datasets are so heterogeneous, we decided to limit our analysis to make the work more manageable. We restricted our analysis to only those projects funded by the NIMH. This also gets around issues of data sharing in other fields where working on shared datasets is the norm (e.g. genomics), and thus may inappropriately bias our analyses and conclusions.

In total, this database lists the NIMH as having funded 41890 projects from 2004 onwards. However, this number is somewhat inflated, as many of these grants are actually renewals of previously existing awards. Removing these and other types of duplicates depresses the number of grants to 13376, 8.

The federal reporter also has a table linking these grants to specific papers. I believe this linking is self-reported data by the PI's on the grant and is done during the grant's annual report. Using this linking table and we find that there are papers that have been funded by these NIMH grants.

The next stage of this analysis is to identify which of these papers contain links to shared data repositories. To do so, we must link these papers to the full-text database. However, not all papers that have information deposited in PubMed will have a corresponding full-text entry in PubMed Central. Of the 90160 papers identified as being linked to NIMH grants in FederalRePORTER, we found full-text matches for 57771 in the PubMed Central full-text database. Table @ref(tab:nimh-journals) show the most 10 most frequently occurring journals in this subset of the data:

At this stage, we can move on to documenting the presence or absence of data sharing in these papers. As a first step, we a set of simple regular expressions to look for the presence of references to many of the most common data-sharing platforms:

- Github
- OSF
- NDAR
- Open Neuro
- Allen Institute
- HCP
- Balsa
- LONI

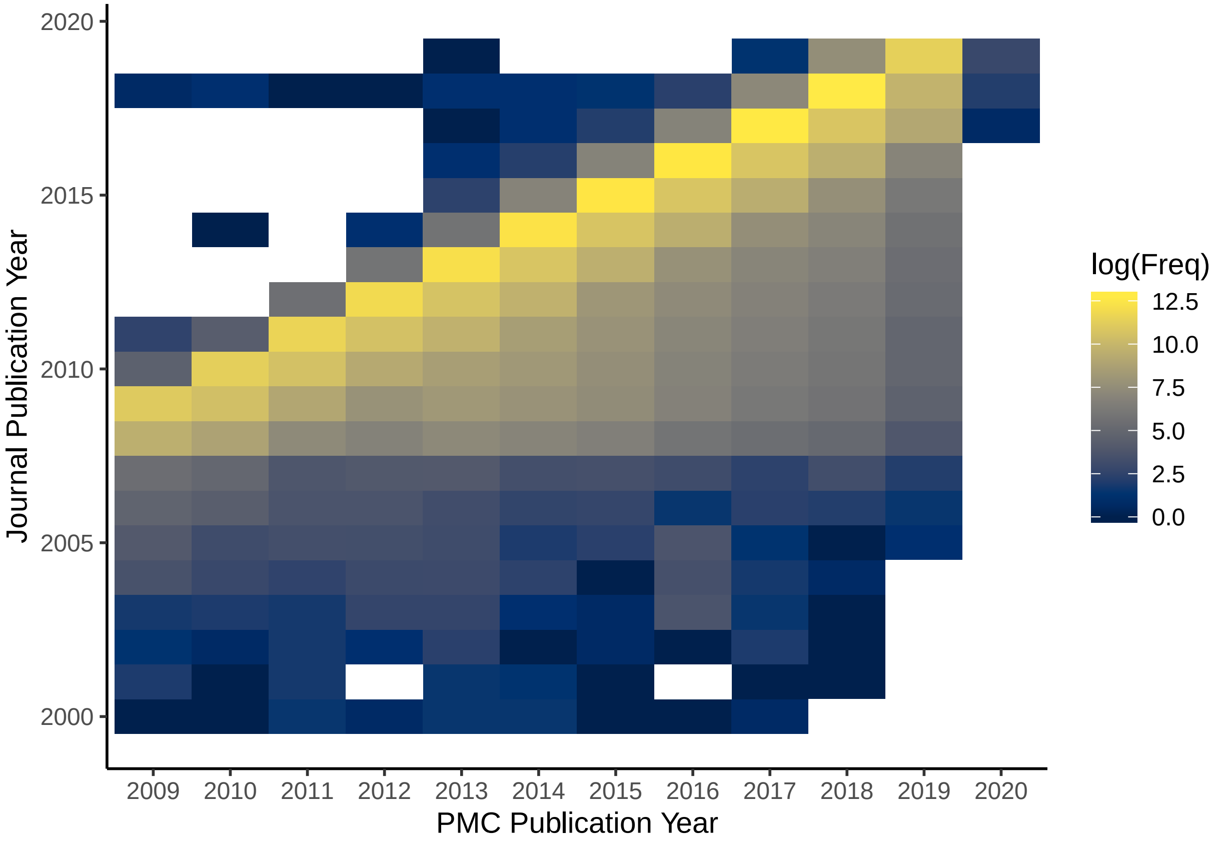

Figure 1: (#fig:date-mat)Year of journal publication versus pubmed central publication. Cells that are white have no observations.

Table 1: (#tab:papes-and-projs)Number of papers and projects by year

year	# Full Text Papers	# Funded Projects
2009	75251	118487
2010	121763	111400
2011	156281	98098
2012	195103	93402
2013	249148	91689
2014	293649	90991
2015	315456	92310
2016	334631	91151
2017	362745	86281
2018	386562	92508
2019	109925	
Total	2600542	1072898

Table 2: (#tab:n-projs)Number of projects and papers by journal and center

Journal.Title	n_papers	IC_CENTER	n_projects
PLoS One	212100	NCI	135422
Sci Rep	89292	NIAID	85306
Oncotarget	24960	NIGMS	77845
Nat Commun	21127	NHLBI	73489
Acta Crystallogr Sect E Struct Rep Online	19794	NIDDK	58961
Sensors (Basel)	18182	NINDS	52960
Int J Mol Sci	17698	NIMH	41890
Biomed Res Int	15807	NICHD	39439
Molecules	15414	NIA	38681
Medicine (Baltimore)	14836	NCRR	37684

- FMRIDC
- CCRNS
- Datalad
- Dataverse
- DBgap
- Dryad
- Figshare
- INDI
- NITRC
- Omega
- Xnat
- Zenodo
- AWS Data

It is important to emphasize that this process is error-prone. A simple text search will not be able to identify many cases where a paper makes a reference to a database but is not sharing data. Additionally, some of these platforms host more than just data. Github, for instance, hosts an extremely heterogeneous collection of digital information, and the OSF, while a bit more circumscribed than Github, also may contain analysis code, experimental materials, extra written documents or other types of information. However, it is worth examining these simple searches to set our expectations. If a papers does not have the string 'github' anywhere in it, then it is unlikely that the data has been shared. Even if authors post the data through another means (e.g. a link to a github repo on their personal website), if that link is not contained in the paper, the majority of readers will never know of its existence.

Table 3: (#tab:nimh-journals)Number of papers with full-text PMC content funded by the NIMH by journal

Journal.Title	n_papers
PLoS One	1750
Neuroimage	1129
J Neurosci	1098
Neuron	960
Biol Psychiatry	935
Schizophr Res	783
AIDS Behav	765
Psychiatry Res	681
Nat Neurosci	518
Mol Psychiatry	515

Of the 57771 papers funded by the NIMH which have full-text entries in PubMed Central, there are 2144 papers which contain a search string for one of the repos highlighted above. Figure @ref{fig:repos-over-time} shows how the overall proportion of papers with references to a repository has grown over time, from a low of less than 1% in 2008 to a high of 14% in 2018 (the percentage for 2019 is 18.8, but as of now this only includes 16 papers). By this measure, data sharing is on the rise.

We can also examine the relative popularity of each of these repositories. Figure @ref{fig:relative-repo-popularity} decomposes the share of repository use into each repository, showing only those repositories whose usage make up at least .75% on at least one year.

Not surprisingly, Github is far and away the leader of the pack, peaking with nearly 9% of publications making a reference to github in 2018. We can alsdo examine which journals have the largest proportion of papers with a reference to a repository. Figure @ref{fig:journals-who-share} shows the proportion of papers with references to one of the repositories.

Figure @ref{fig:journals-who-share} shows that journals with a heavy emphasis on genetics (eg. *Nature Genetics*, *Genome Biology*, *BMC Genomics*, *Genetic Epidemiology*), methods (eg *Nature Methods*, *PLoS Computational Biology*), or open-access philosophy (*eLife*, *PLoS Computational Biology*, *PLoS Biology*, *eNeuro*) feature prominently in the upper ranks of this distribution.

The next stages of this analysis is to more accurately measure when data is posted to a repository. As of now, I think that the best way to accomplish this is to label a subset of the papers and train a classifier on this hand-labeled data.

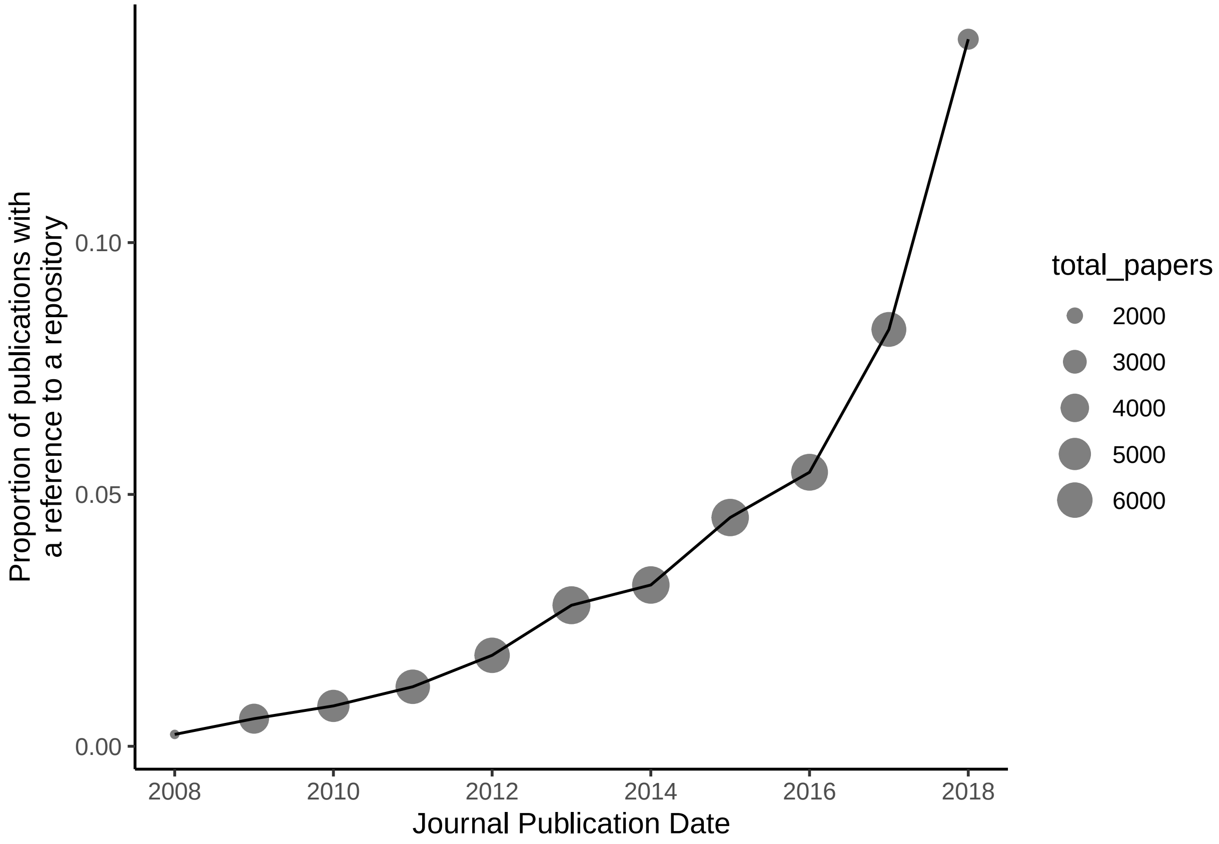

Figure 2: (#fig:repos-over-time)Proportion of all NIMH funded full-text publications that resulted in a hit from a regular-expression search for a set of data repositories. Point size reflects the number of papers that were searched in a given year

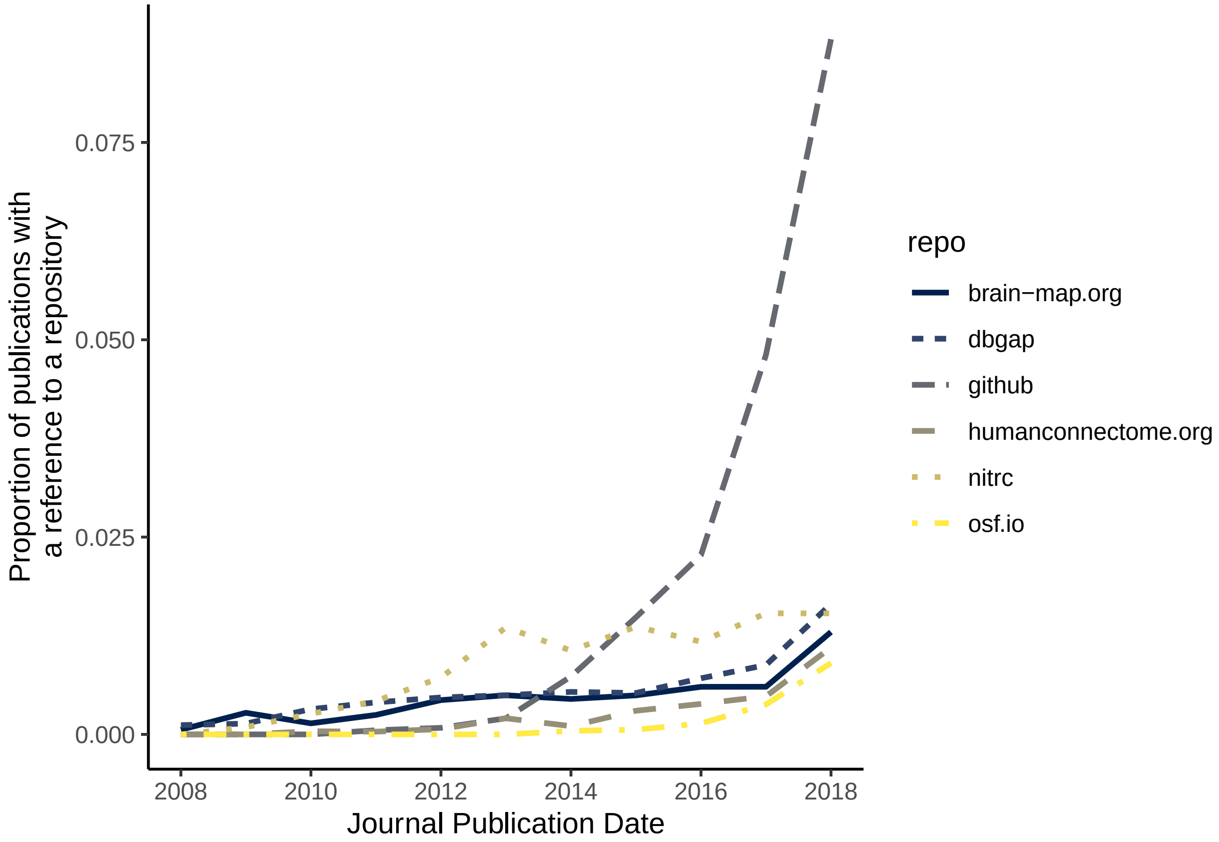

Figure 3: (#fig:relative-repo-popularity)Proportion of all NIMH funded full-text publications that resulted in a hit from a regular-expression search for a set of data repositories. Each line represents a different repository. Only repositories who have at least .75% of papers with a hit on at least one year are shown.

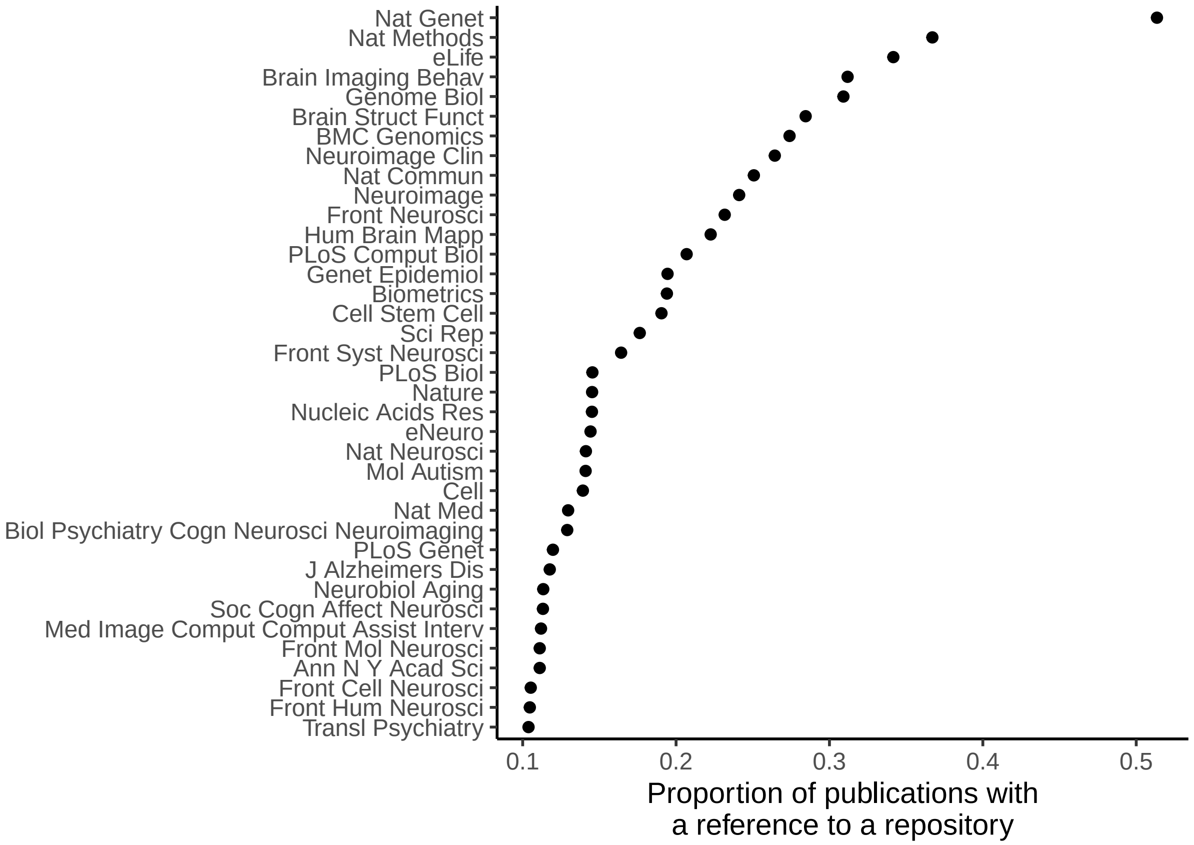

Figure 4: (#fig:journals-who-share)Proportion of publications with a reference to at least one data repository, by journal. Only publications with proportions greater than 10% are shown