Research Project Report

KIREPRO1PE

Research Project

Towards Understanding The Hidden Popularity of Machine Learning Datasets

Author:
Ahmet Akkoc; ahak@itu.dk

December 15, 2022

Towards Understanding The Hidden Popularity of Machine Learning Datasets

1st Ahmet Akkoc
BSc. Data Science
IT University of Copenhagen
Copenhagen, Denmark
ORCID:0000-0001-5096-6282

***Abstract*—Machine learning datasets published web-first and academic research published print-first often find themselves at odds when it comes to acknowledging one another. In this project, I design a strategy to extract citations and unusual references (mentions) to ML datasets in literature. On a test dataset, built on accepted papers to the medical CHIL 2022 conference I find that 26% of datasets used were never cited properly, denying them from citation tracking, and that 20% of references to datasets are not immediately accessible. My work is made available on https://github.com/madprogramer/PublicDatasets.**

***Index Terms*—data, machine learning, datasets, challenge, medical data, CHIL, citation, metascience**

I. Introduction

Machine learning datasets have come into prominence following the Deep-Learning Revolution [1]. Today there are datasets available for everything from food recipes to medical images. Public datasets have become valuable resources to researchers across the world. Yet, they are sometimes looked down upon for being either a novelty or un-academic.

Interestingly, measuring the impact that these machine learning datasets have on research and technology is actually non-trivial. Most public machine learning datasets originate on the web. However, there are no well-defined conventions for citing such digital web material in academic research and scientific publishing. Citations styles are primarily designed for referring to printed material and do not easily incorporate datasets. In addition, there are stigmas around citing open, non peer-reviewed resources. As a result of these, the crucial influence that machine learning datasets currently play on research is being hidden.

In this project I propose a strategy to investigate the significance of datasets, based on their mentions within academic publications. Section II gives a brief background on the history of machine learning datasets and how their widespread adoption has come into conflict with academic tradition. Section III details the workflow of my approach to construct a sample dataset of machine learning dataset mentions (**CHIL-MLDM**). Section IV demonstrates findings on the example CHIL-MLDM dataset. Finally, Section V concludes on the strengths and caveats of my approach. The code and data of this work is made available on https://github.com/madprogramer/PublicDatasets.

II. Background

A. A Brief History of ML Datasets and Challenges

Machine learning is a field that has always been intrinsically data-driven. The problem a model seeks to solve, together with its input and output domains are described and co-defined by the data available. Therefore, one of the greatest challenges in machine learning has been the quest for high quality data. That itself encompasses both the questions of how to collect such data and how to distribute it.

The Big Data in the 1990s was under the private control of either enterprise or government. The US government would quickly become a key actor in the Data Revolution, funding data collection and publishing of databases through various agencies such as the Defense Advanced Research Projects Agency (**DARPA**) and the National Institute of Standards and Technology (**NIST**).

The task of optical character recognition (OCR) was especially interesting to the US at the time. One reason why OCR was appealing was the potential faster processing of, for example, census data and tax forms. **NIST** would collect many samples of handwritten text, processing and then publishing these under a series of databases [2] [3]. These **NIST** datasets may be considered an early precursor to public datasets. At the time, these databases were physically distributed via CD-ROMs. Even so, the **NIST** databases were made available under public domain. This meant other agencies or researchers interested in OCR and similar problems were able to benefit from the data. The more popular MNIST dataset [4] (Fig. 1) would also evolve out of this same project, continuing to be used as a benchmark to this day.

Following government interest in sharing data, commercial interest soon also emerged. In 2006, the company Netflix announced the Netflix Prize challenge [5]. This movie recommendation user database, was one of the first large scale datasets to be distributed online.[2] The Netflix Prize is fondly remembered as a challenge uniting teams across the globe to cooperate, with the web-friendliness of the challenge encouraging much mutual discussion. By this time the new term "dataset" was also beginning to replace the more generic term

[1] https://github.com/mbornet-hl/MNIST

[2] A snapshot of the data download page has been preserved at: https://web.archive.org/web/20061013082802/http://www.netflixprize.com/download

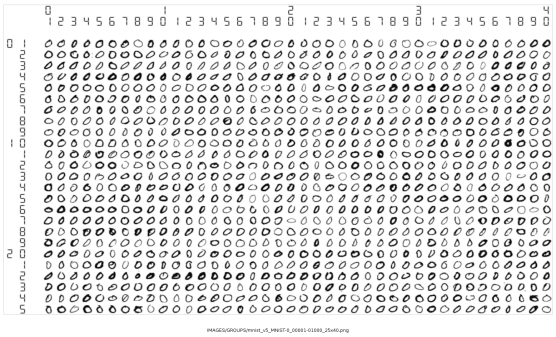

Fig. 1. A snippet of handwritten 0's, taken from the MNIST [4] dataset.[1]

"database" to describe collections of data that were meant to be exchanged.

Later years then saw academic interest in building publicly available machine learning datasets. Notable examples come from image classification problems in computer vision, such as the ImageNet [6] and CIFAR-10 [7] datasets from 2009. The maintainers of the ImageNet dataset in particular also hold challenges annually, where teams compete with state-of-the-art techniques. Teams that have ranked in ImageNet in previous years have reworked or iterated their work into publications.

Finally, we saw the emergence of large scale Machine Learning challenge platforms such as *Kaggle*[3]. These ML challenge platforms host challenges for their sponsors and act as permanent repositories for open datasets. Kaggle and similar platforms continue to attract a lot of attention, both from academia and industry researchers.

This chronology of the co-operative and inter-disciplinary development of datasets is highly in contrast with the much overcritical and demanding academic research ecosystem. That in turn causes friction between the two cultures of data and academic research which I will attempt to explore in the section below.

B. Citations in Academia

Academia disseminates research through a plethora of mediums; lectures, books, conferences and journals can all be considered some part of academia. Machine learning is no exception in this regard. However, as stated above, datasets as a very important component of machine learning are not limited to academia. Often a dataset is an entity unto itself and the idea of web-first scientific material is still very new.

Where the ML Challenge scene and open data have made a name for themselves through dynamism and rapid iteration, academic culture could be described as slow and steady, by comparison. Peer attribution and peer review are held in very high-regard, as opposed to contest scores and teams. Different fields have developed citation styles to be able to reference and link to one another's work. Academics also have incentives to keep them within a particular ecosystem. Citation frequency [8], by other academics, is considered an important measure of the worth of one's work. Thus citations are taken very seriously and there are dedicated tools to track what has cited who how frequently.

Datasets as a form of scientific material find themselves in a strange place. Since they have come onto the scene after established conventions, it is not always clear how the dataset ought to be cited. Some researchers may cite the dataset as a URL link, if their citation style allows it. Other researchers may simply address it by name and a footnote. Others yet may instead look for a research publication connected to said dataset, and cite that instead. As such, there is a lot of inconsistency in how datasets are handled in research, preventing academics from measuring their value, in the same way that they would measure most publications. The following sections of this work will seek to demonstrate how both proper and improper references to the datasets can be negotiated to better understand the prominence of datasets in scientific research and their effects.

III. Methodology

A. Problem Formulation

Due to the disconnect between the circulation of datasets and tracking of academic citations, it is not immediately obvious how to assess the effect of datasets on research. So as a first step, I decide to limit as much of my analysis as possible to collecting datasets. The one extension I make is to include data synthesizers[4] in addition to static datasets. Data synthesizers are code-snippets or libraries that are able to generate new data at run-time. While they might not be "sets" by the mathematical definition, data synthesizers do build datasets and two datasets generated by the same synthesizer may be considered sufficiently related. Besides datasets, it should be noted that there are other openly available resources used in research, such as pre-trained machine learning models.

Second, I have chosen to examine the different *mentions* of datasets. As stated in Sec. II, many online datasets are only referenced as footnotes, with no corresponding entry in the bibliography. With this in mind, the approach taken here seeks to include non-standard forms of citation, in addition to in-line citations referencing a particular dataset. Furthermore, if a dataset points to, for example, both another publication and a URL address it would be relevant to track both forms of the mention.

B. Dataset

To analyse the different ways datasets are mentioned, I build a dataset for analysing Machine Learning Dataset Mentions in CHIL 2022 (**CHIL-MLDM**). To build this CHIL-MLDM Dataset, I have chosen to work on a single venue: The 2022 Conference on Health, Inference, and Learning (CHIL 2022) [9]. CHIL is a cross-disciplinary conference bringing together clinicians, industry and academia in discussions on machine

[3]https://www.kaggle.com/

[4]Data synthesizers are sometimes also known as data simulators or data augmenters. I have chosen to use the term synthesizer as a generic term to refer to any of the three.

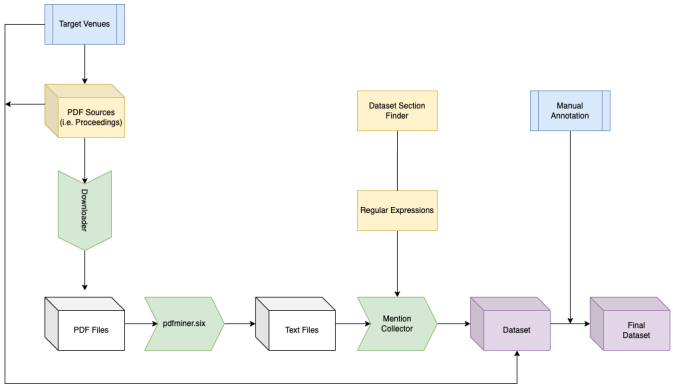

Fig. 2. Workflow of the pipeline used to build the Machine Learning Data Mentions Dataset.

learning and health policy. As such, CHIL gives us a glimpse into how all of these different work cultures treat attribution with their own considerations in mind. CHIL is also an open-access conference, meaning that other researchers can easily access the original conference proceedings and reproduce or extend my results from a common data source.

In papers comparing multiple datasets there is often a dedicated section listing all the datasets used. This dataset list can be anywhere from the abstract to the appendix. CHIL has an additional benefit of requiring a dedicated "Data and Code Availability" section immediately after the abstract[5]. This structuring makes aggregating information on mentioned datasets much easier compared to other conferences where the authors can list their datasets arbitrarily. Sec. III-C below, elaborates on the data collection and annotation processes.

C. Pipeline

The pipeline to build the CHIL-MLDM Dataset is a series of Python notebooks and some manual annotation. Below I describe the steps taken to build this CHIL-MLDM Dataset in further detail.

Since venues such as conferences or journals are distributed through either print or on different websites, this pipeline takes a source of conference proceedings as input. In particular, it is designed to crawl from the Proceedings of Machine Learning Research (PMLR) [6], a repository of much open-access research. Although I only use the proceedings for CHIL 2022 in this paper, my strategy is designed such that it is fairly easy to extend it to more conferences. Crawling is performed using the scrapy[7] framework and papers are downloaded and saved as PDFs under the project directory. PDF files are labeled by venue and title.

Then, the text of the papers is needed for processing. This text is extracted from the downloaded PDFs using the pdfminer.six[8] tool. Pdfminer.six has a number of options, but for the purposes of this work the default configuration is used,

[5]See official CHIL 2022 Template
[6]https://proceedings.mlr.press/faq.html
[7]https://scrapy.org/
[8]https://github.com/pdfminer/pdfminer.six

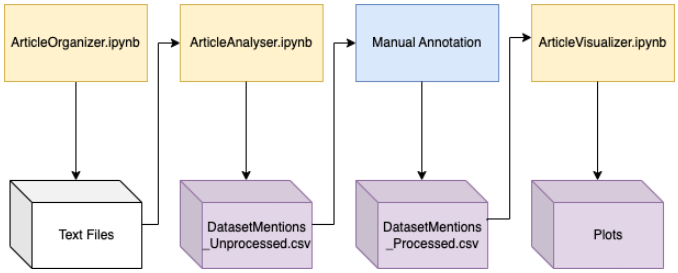

Fig. 3. Schematic of Project Code Organization

where it is attempted to read files PDF top-to-bottom. Rarely, this might lead to text-jumbling where a new line occurs, or more frequently at the end of a column or page. I address these in the latter steps of the "Mention Collector" (Fig. 2) as well as during manual annotation. Again, text files generated from this process are labeled by venue and title, corresponding to PDF Files.

The crawling, downloading and text-extraction is all done under the *ArticleOrganizer* iPython notebook, shown in Fig. 3. Scanning the extracted text for mentions is relegated to the *ArticleAnalyser* iPython notebook.

Next, to collected the text files are queried for dataset mentions. The strategy here is a naïve one, where the script assumes that there is a fine list of datasets somewhere in the original paper. For CHIL, I was able take advantage of the "Data and Code Availability" section and query for it directly. The snippet then ends either at the introduction or a copyright notice. This data extraction process is demonstrated in Fig. 4

Dataset mentions are classified into 3 categories: Inline citations, inline URLs and footnotes. The notebook script employs regular expressions to detect these 3 types of mentions. It first matches and removes inline citations, which are by far the most identifiable. Then URLs, then numbered sections (f.x. 3.4) and finally footnotes (as numbers). Footnotes were the most challenging, because numbers can appear in multiple contexts and are split across the page, thus they required further correction during manual annotation. It should be noted that because of line-breaks regular expressions need to accommodate extra or optional newlines; these newlines possibly come in place of spaces thus making it difficult to remove them in pre-processing. Instead, any newlines are removed in post-processing after having captured matches.

The output of this procedure is an under-processed CSV dataset containing rows of venue, title, mention style and the mention itself. Additionally, the text snippet that was used to extract this information is added to a notes column as a reference to be used in annotation. This method yielded 77 potential mention items across 23 papers.

1) Manual Annotation: The data collection strategy itself was not designed to be context-sensitive. The manual annotation step seeks to compensate for this, by giving an annotator a chance to double-check or extend the collected mentions data.

I introduce two new columns: Dataset Identifier, describing what a the dataset is called, and Access, describing if the

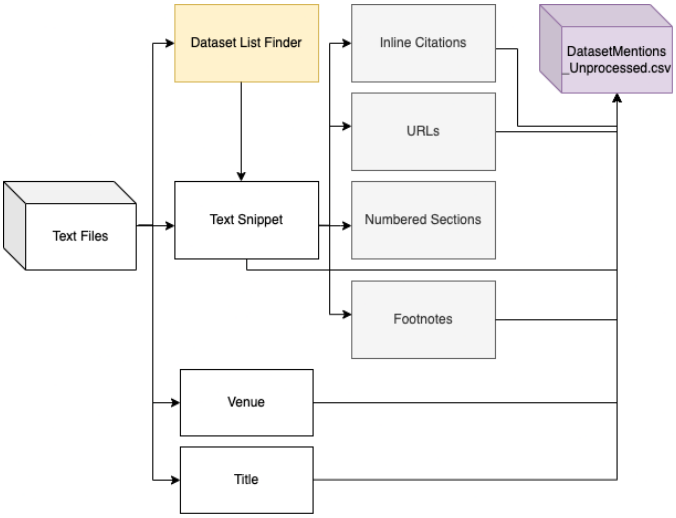

Fig. 4. Schematic of Data Processing

dataset is accessible. Moving onto mentions, most citations were not edited, but some URLs required spelling correction; if, for example a comma or period were captured in the match, these are valid URL addresses and require context[9]. If a paper refers to its own work, this is denoted as "[Paper Dataset]". More importantly, footnotes were adjusted, as the unprocessed dataset only gave the number for the footnote, but not its contents. Sometimes the contents of a footnote were miscaptured as URL, if they occurred before the introduction. To fix this I collapse footnotes and "footnote URL" into a single item. Despite filtering, some captured footnotes were unrelated numbers and these were also manually removed.

As far as missing data goes, there is a possibility that there may still be some datasets missed by this process. Some papers make a distinction between internal and external validation. F.x. *Estimating Model Performance on External Samples from Their Limited Statistical Characteristics* [10] only listed their internal validation datasets while including additional ones in the experiments. In this case, I have chosen to only annotate for the internal datasets; I have done so likewise for similar datasets omitted from the "Data and Code Availability" section.

The number of mentions in the final CHIL-MLDM dataset decreased from 77 to 62, after corrections. It should be reiterated that I have chosen to include data synthesizing code (see Sec. III-A), in addition to traditional datasets. The annotated dataset has been passed to the *ArticleVisualizer* iPython notebook, to generate the figures shown in Sec. IV.

IV. Results

CHIL-MLDM is composed of 3 files: "ResearchPapers.csv", a table of papers fetched from the target venues; "DatasetMentions_Unprocessed.csv", the unannotated mentions dataset with extra rows and missing some contextual information; "DatasetMentions_Processed.csv", the most annotated version of the data including URL origin and dataset identifier as well.

In this section I examine the "DatasetMentions_Processed.csv" table to understand how machine learning datasets are represented in CHIL 2022, from two aspects: Bibliometric Exclusion and Accessibility.

A. Bibliometric Exclusion

Public datasets are published on and through the web. The most-direct reference to the dataset in this case would be a URL. Unfortunately, it is sometimes preferred to cite printed publications as a primary source (see Sec. II-B). This leaves web datasets in a confusing situation where one is expected to publish their dataset in print if they hope to have their dataset cited in research bibliography. When such a printed work is not available, other authors referring to the work will turn to footnotes or webpage citations may be used instead. This comes with the unfortunate effect that the impact of such datasets are unexpressed in bibliography and in citation trackers which aggregate bibliographies.

Figure 5 shows the $N = 31$ data sources (datasets and data syntehesizers) identified in CHIL 2022, and how they were cited in aggregate. Note that paper datasets (i.e. self-citations) have been excluded, such that the table only considers the data sources to a given work, not the work itself. 23 $(\approx 74\%)$ of the datasets received at least one inline citation. In particular, the MIMIC-III dataset [11] and the PhysioNet repository [12] stand out. PhysioNet is a data repository housing several of the datasets mentioned within CHIL-MLDM. Even so, it was often jointly cited with the MIMIC-III dataset and other related datasets. Thus there seems to be a case of citing supersets of data in addition to the experimental data itself.

The remaining 8 $(\approx 26\%)$ of the datasets are not once cited in the text. This means that all of these datasets (or data sources) will be invisible to citation trackers. For example pycox[10], a package containing datasets and data synthesizers for survival analysis, is always mentioned as a footnote. Futhermore, 4 datasets are mentioned as URLs in the body of the text, 2 datasets are only mentioned as a footnote, and 2 private datasets are completely excluded[11], only being alluded to and never named.

The annotated CHIL-MLDM dataset shows that there are 7 instances of a dataset being co-mentioned in a paper, both in the form of an inline citation and URL (within the body or footnote). Datasets that were double mentioned are discernible from Fig. 5, as the columns which have one citation and one footnote, or one citation and one URL[12] The one exception is MIMIC-III, which received an unusual double citation, but only in the paper *Enriching Unsupervised User Embedding via Medical Concepts* [13]. Another note to be made here is that we see footnote mentions are more frequent (9) than in-body URLs (5). This might be an important point to note in

[9]In retrospect, it is highly unlikely for a URL address to end on a non-alphanumeric character. It may be just as good to assume a URL to terminate on alphanumeric and to, instead, manually annotate for false-negatives.

[10]https://github.com/havakv/pycox

[11]see Sec. IV-B for elaboration

[12]This may be verified from the "DatasetMentions_Processed.csv" file.

text-processing (f.x. NLP tasks) which may not be configured to parse footnotes by default.

We may thus conclude here that there is a multidimensional issue regarding bibliometric exclusion. Not only are a significant minority of datasets excluded from bibliographic tracking; but there are in addition many datasets whose references are marginalised as merely footnotes. The fact that there are instances of authors referring to datasets both by printed publication and URL may suggest some consciousness towards this issue among academics, warranting further investigation.

B. Accessibility

Datasets reside in different corners of the web. Some datasets are publicly available to all, whereas others may be private. For the datasets in CHIL-MLDM, they are most often hosted on GitHub[13] or the aforementioned PhysioNet[14]. As shown in Fig. 6, 6 URLs or footnotes make reference to a dataset from GitHub and 2 URLs from PhysioNet. PhysioNet was often also cited as print, in actuality bringing it to a number close to that of GitHub (refer to Fig. 5). This is only considering, non-[Paper Dataset] mentions. In fact all of the public paper datasets have been made to host their code and data on GitHub.

The majority of mentions $(53/66 \approx 80\%)$ point to open resources. These are either open-access bibliography, publications open for reading, or URLs pointing to open-source code. A more detailed breakdown is shown in Fig. 7. In the case of CHIL 2022, all publications associated with a dataset mention were open-access, therefore an inline citation annotated as "private" means that the associated dataset was private.

Some institutional or corporate-controlled databases are private, typically made available upon request. Among the paper datasets themselves, 1 has been moved and 1 has had its link broken (i.e. it is no longer available at the given address). 2 datasets were created privately for the sole purpose of a study in cooperation with the Massachusetts General Hospital (I denote these as MGH Dataset 1 & 2 in the data). 3 private datasets were used in previous studies and the reader is referred to the authors of those studies (through citation) if they so wish to obtain said data. Finally, 1 private dataset has an online request form which the reader is referred to, namely Temple University's EEG Corpora resource [15].

In addition to the above mentioned datasets, there are two more of note. These two were paper datasets which referred to GitHub URLs. One dataset has been moved, with a redirect available and the other has been broken with no obvious connection from the original work available.

It should be restated that the primary venue for the CHIL-MLDM dataset, was CHIL, which is an open-access conference as-is. Therefore, it is reasonable to assume that open-access resources are over-expressed in this particular case. Such rates of open-access citations, mentions of openly available datasets and unbroken links are likely higher than for similar conferences. Nonetheless, in some cases we do see the work of authors who choose to keep their datasets private and only have them be referable by inline citation.

[13]https://github.com/

[14]https://physionet.org/

[15]https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml

V. Concluding Remarks

In this paper, I have developed a method to study the popularity of machine learning datasets and constructed a sample dataset of machine learning dataset mentions in the CHIL 2022 conference: CHIL-MLDM. In this process I have uncovered some of the idiosyncrasies of the sampled CHIL 2022 conference. In particular, I discovered an interesting trend of citing macro data repositories in the bibliography, even when the repository itself was quite remote from the experimental work. The cause of this perhaps warrants further research in its own right. A bias to recognize here is that CHIL is an open-access conference with an emphasis on open data, where authors are encouraged to use open data and make their data open.

Another point was that I came across less usual forms of datasets in addition to image or tabular data, such as data synthesizers which are generative datasets. I have done my best to accommodate for these, although their frequency suggests that a neutral term such as "datasource" may be preferable to "dataset", as I have noted a few times throughout this paper.

Despite automating most of this process, some manual annotation was still required. While in one regard this was a weakness of my approach limiting the scalability of results, I also think that it was a very educational way to go about it. In the process of extracting non-standard dataset mentions, I have come to see problems which are likely to arise in text-processing. For example, footnotes can fall into one of several locations on the page. Another decision I made was to analyze mentions in aggregate, without always considering the individual context. This being said, by focusing on datasets shared between papers, instead of the mentions on the same paper, I was able to demonstrate that there is indeed a problem of bibliographic under-representation for *certain datasets*. If anything, future work now has reason to investigate, instead on basis of mentions within the same paper, how researchers ought to change their writing style to circumvent the issue of underrepresentation.

Future work may be also be interested in incorporating NLP into the pipeline, for the task of detecting the initial list of datasets used as part of analyses, in less structurally-strict papers. Likewise, the steps of tokenization could also be further optimized. It is my hope that the baby steps taken here can be extended to modernize the way we handle citation in academic journalism, such that it is more inclusive and acknowledging of non-print resources.

References

[1] T. J. Sejnowski, *The Deep Learning Revolution*. MIT Press, Oct. 2018. Google-Books-ID: 9xZxDwAAQBAJ.

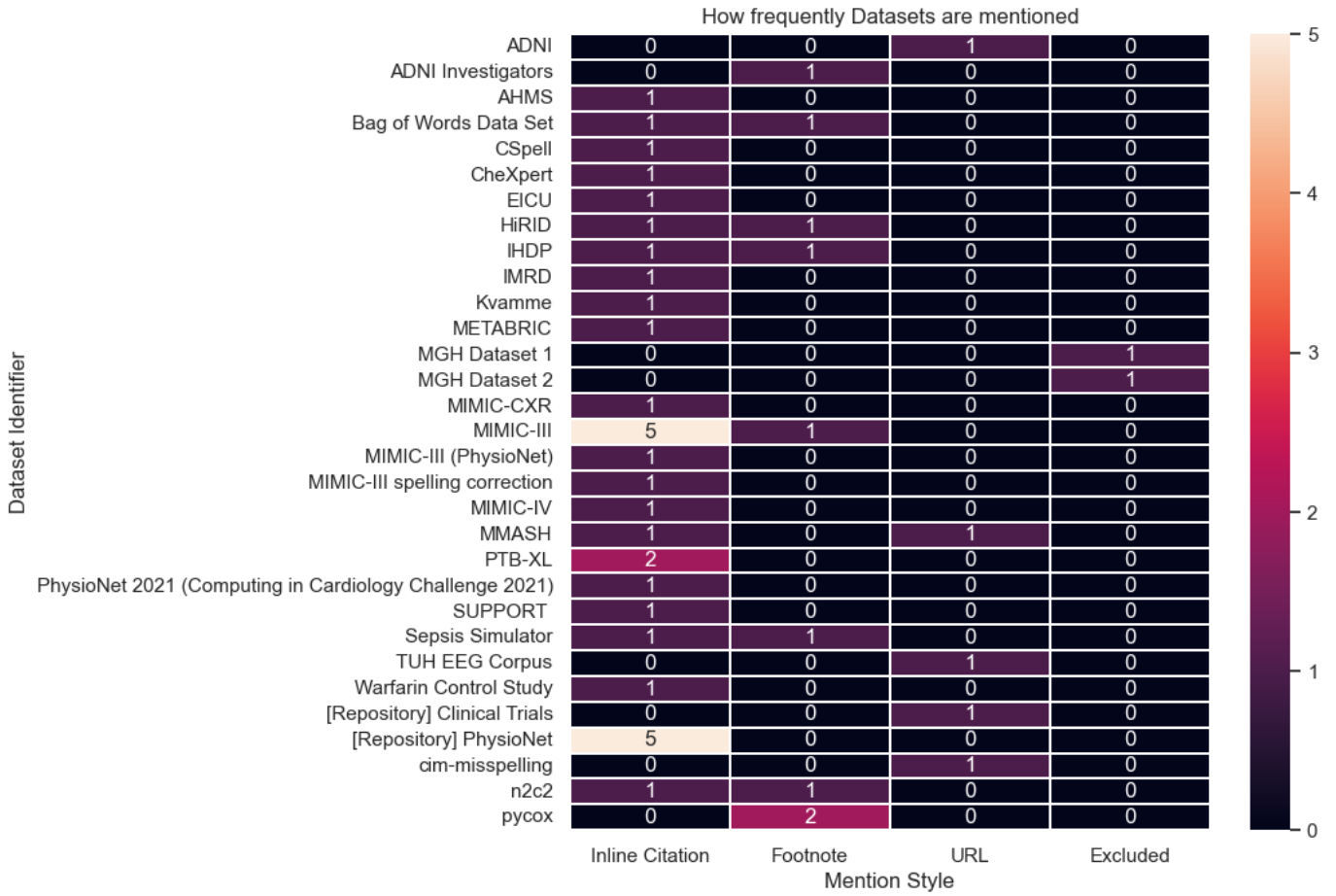

Fig. 5. How frequently source datasets were cited by mention style.

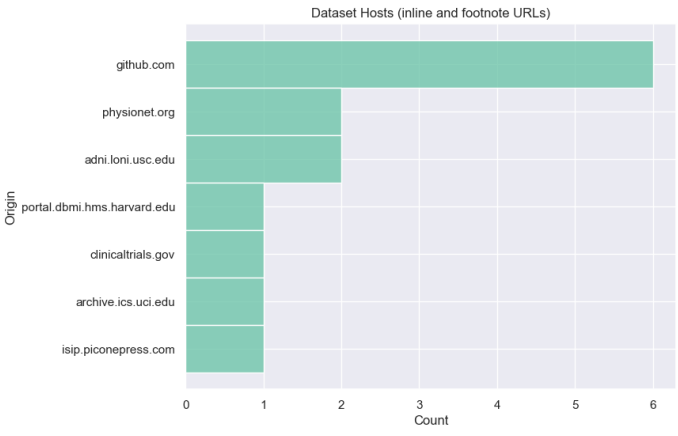

Fig. 6. Domains of Dataset URLs. (Paper Datasets Excluded)

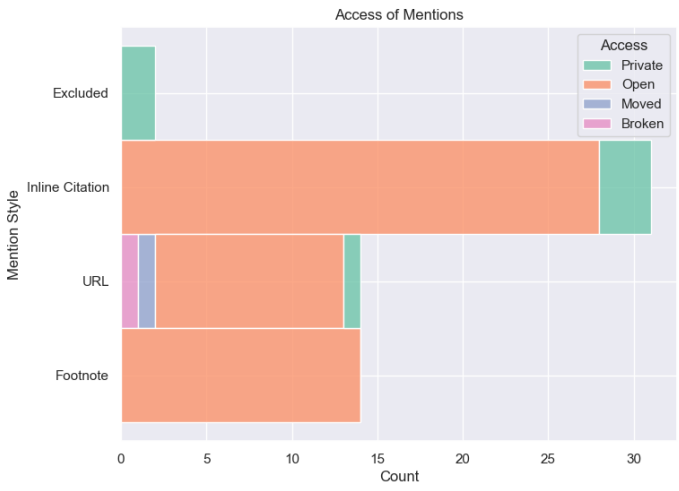

Fig. 7. Accessibility of Datasets by means of mention (Paper Datasets Included)

[2] R. A. Wilkinson, J. Geist, S. Janet, P. Grother, C. J. Burges, R. Creecy, B. Hammond, J. J. Hull, N. W. Larsen, T. P. Vogl, and C. L. Wilson, "The First Census Optical Character Recognition Systems Conference," *NIST*, Jan. 1992. Last Modified: 2021-10-12T11:10-04:00 Publisher: R A. Wilkinson, Jon Geist, Stanley Janet, Patrick Grother, Christopher J. Burges, Robert Creecy, Bob Hammond, Jonathan J. Hull, Norman W. Larsen, Thomas P. Vogl, Charles L. Wilson.

[3] P. J. Grother, "NIST special database 19," *Handprinted forms and characters database, National Institute of Standards and Technology*, vol. 10, 1995.

[4] Y. LeCun, "The MNIST database of handwritten digits," *http://yann.*

lecun. com/exdb/mnist/, 1998.

[5] J. Bennett and S. Lanning, "The netflix prize," in *Proceedings of KDD cup and workshop*, vol. 2007, p. 35, New York, 2007.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. ISSN: 1063-6919.

[7] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 2009.

[8] E. Garfield, "Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies.," *Science*, vol. 178, no. 4060, pp. 471–479, 1972. Publisher: American Association for the Advancement of Science.

[9] G. Flores, G. H. Chen, T. Pollard, A. Zirikly, M. C. Hughes, T. Sarker, J. C. Ho, and T. Naumann, "Conference on Health, Inference, and Learning (CHIL) 2022," in *Proceedings of the Conference on Health, Inference, and Learning*, pp. 1–4, PMLR, Apr. 2022. ISSN: 2640-3498.

[10] T. El-Hay and C. Yanover, "Estimating Model Performance on External Samples from Their Limited Statistical Characteristics," in *Proceedings of the Conference on Health, Inference, and Learning*, pp. 48–62, PMLR, Apr. 2022. ISSN: 2640-3498.

[11] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III Clinical Database," 2015. Version Number: 1.4 Type: dataset.

[12] G. Moody, R. Mark, and A. Goldberger, "PhysioNet: a Web-based resource for the study of physiologic signals," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, pp. 70–75, May 2001. Conference Name: IEEE Engineering in Medicine and Biology Magazine.

[13] X. Huang, F. Dernoncourt, and M. Dredze, "Enriching Unsupervised User Embedding via Medical Concepts," in *Proceedings of the Conference on Health, Inference, and Learning*, pp. 63–78, PMLR, Apr. 2022. ISSN: 2640-3498.