Data reuse in the wild: A bibliometric analysis of the UCI Machine Learning Repository

Dakota Murray

Department of Informatics, Indiana University Bloomington, Bloomington, United States; dakmurra@iu.edu

Abstract

The Big Data movement has amplified the ability of data to radically transcend the contexts of its origin, allowing data to be shared with more people and reused more easily and widely than ever before. Open Science and Open Data policies seek to further encourage data sharing and reuse to facilitate scientific transparency, reproducibility, and the re-purposing of data for further scientific discoveries. In practice, however, the reuse of data often has practical constraints that limit its reuse and entails methodological, epistemic, and ethical risks. Effective and responsible data governance requires a broader empirical understanding of data reuse *in the wild*. Here, I propose a bibliometric analysis of data reuse of the popular UCI Machine Learning Repository. I propose a mix of manual and machine learning techniques to identify citations made to datasets in this repository and appearing within the *Elsevier ScienceDirect* full-text database. Using these data, I compile indicators to address three research goals: 1) measure the speed and extent to which these datasets were reused; 2) quantify the degree to which these measures differ based on the characteristics of the dataset; and 3) investigate the degree to which donation to the repository was associated with changes to patterns of reuse. Results of this analysis will provide an empirical foundation for data and science governance, further theoretical understandings of data reuse in scientific research, and contribute a methodological framework that can be extended to analysis of other repositories.

Qualifying Exam Question #3

As discussed over email, the nature of this question changed a great deal. The original version of my third qualifying exam question is quoted below.

> The rising availability of full-text scientific publications has made possible a range of tools and analyses for understanding and making use of scientific knowledge, via means of text mining and natural language processing. One area of study is in annotation of scientific texts according to specified criteria, such as annotating papers according to the type of study being done, the effect size of the result, or the sentiment of a sentence. Annotation of this type helps facilitate database indexing, recommendation systems, information extraction, text summarization, and also more macro-level studies of the scientific process itself, such as that in Scientometrics or the Science of Science. This research area has enjoyed great success in many fields, especially in biomedical literature. However little work has been conducted to annotate and extract a key piece of information from scientific texts: the data used to conduct the study.
>
> Design a research plan to develop a technique to (1) annotate sentences from scientific texts according to whether or not they mention key aspects about the data used in the study, and (2) extract the relevant information from annotated sentences. You should detail a datasource, and justify why it was chosen. You should also discuss how this data will be cleaned, processed, what features will be extracted, and the techniques you plan to use to perform the given tasks. Additionally, you should justify the need for such a system, and note the broader impacts if successful.

There were several reasons why I chose to change from this question to what is shown in the current essay. The first reason is feasibility: I had hoped that my proposal for my qualifying exam would be something that were practical. After taking a semester of courses in Natural Language Processing and Machine Learning, I determined that the original question was not feasible given my training and resources. The second reason that I chose to switch was to better leverage the skills of my committee. The original question posed a problem that was both outside of the theoretical as well as the methodological purview of my qualifying exam committee. The new essay, however, rests in firm theoretical foundations of Bibliometrics and Science and Technology Studies, and even includes some approaches using Machine Learning. With this new essay, I hope that I can solicit methodological feedback during the oral stage of my examination. Finally, the last reason why I chose to switch is inspiration. Sometimes, reading and writing leads you down a road that you did not originally expect. Typically, my last minute ideas and inspirations have been my most successful, and so I hate to waste it when it strikes. The essay detailed here is certainly a last minute idea, but also one that I am more confident of than my original proposal.

I thank my qualifying exam committee for their patience and assistance throughout this semester. I look forward to the in-person discussion,

Dakota Murray

Introduction

In 2011, the journal *Science* published an editorial stating that "we must all accept that science is data and that data are science, and thus provide for, and justify the need for the support of, much-improved data curation." [1]. This call for improved data curation has only grown more critical with the rise of the so-called "Big Data Revolution" [2] and the "fourth-paradigm" of data-driven scientific research [3]. Of the many novelties of Big Data, one important aspect is the amplified ability of data to "radically transcend the circumstances and locality of its production" [4] by being reused and re-purposed in contexts far different than that of its origin. The Open Data movement has capitalized on data's ability to travel and has sought policies to encourage the sharing and reuse of scientific data. Reusing data, it is claimed, makes it possible to verify and reproduce past results; makes research more open to the public that funds it; allows for new questions to be asked of existing data; and advances research and innovation through the re-purposing and re-analysis of shared data [5]. However in addition to these opportunities, data sharing and reuse also entails risks. For example, sensitive research data [6], or data which is mired in controversial histories (e.g., [4]) should not be widely shared. Moreover, focus on data reuse rather than the creation of new data risks placing methodological and epistemic constraints the kinds of research being done. Under the "Big Data Revolution", data can travel farther and faster than ever before; in so doing, both the consequences and benefits of data sharing are amplified.

In order to make sense of the potential risks and benefits of open data, science policymakers and the Open Data movement would benefit from a more thorough understanding of how fast, far, and what factors mediate its reuse. The study proposed here aims to inform science governance by developing indicators of data reuse. Several disciplines have contributed to current understandings of data reuse. For example, researchers in Library and Information Science have conducted empirical studies of data sharing practices [7] and surveys into perceptions of data reuse [8, 9]. Bibliometric researches have instead attempted the use of macro-level analyses to investigate citations to datasets in the Data Citation Index [10]. However, the extreme heterogeneity of data citation practices limits the accuracy and extent of indicators of data citation. Scholars from *Critical Data Studies* have taken ethnographic and historically-oriented approaches to understand Big Data, and in so doing have developed robust theoretical frameworks; however, there are few studies of data reuse and existing theories often lack empirical support. At present, there are few studies in any discipline that investigate the speed and extent of data reuse amid the "Big Data Revolution".

The goal of the study proposed here is to expand upon existing bibliometric techniques in order to provide an empirical dimension to understandings of data reuse. I focus on the reuse of datasets that are contained within a single source: the University of California Irvine (UCI) Machine Learning Repository. I propose using bibliometric tools to identify citations made to the Repository itself as well as to datasets stored within. I will then compile indicators of speed and extent to which the repository and data are reused. Scoping analysis to this one repository entails methodological advantages that make this research possible. The repository is large and diverse enough to provide insights across a range of research contexts; the repository is also small enough that technical difficulties inherent to identifying citations to datasets can be handled manually or in an *ad hoc* manner. The UCI Machine Learning Repository also offers an ideal case study of nearly frictionless real-world data reuse due; this is a result of the repository being free, publicly accessible, popular, and easy to use. Insights gleaned from the study of data in the UCI Machine Learning Repository will serve as a baseline for understandings of data reuse, and will either establish targets for future policies to aim for, or provide cautionary examples of what data reuse policies should seek to avoid.

Background

Big and Open Data

Scientific knowledge is intimately tied to scientific data [1], a relationship that has constantly evolved. The most significant recent development to this relationship has been the attention and investment given

to *Big Data* [2]. Thanks to advances in technologies for data collection, storage, and analysis, there is now more data about more things than ever before—the so-called "data deluge" is upon us [11]. Companies and governments have all invested in leveraging this data, but scientific research has felt some of the most profound implications of Big Data. A myriad of data-driven research areas, computational tools, statistical techniques, and data-oriented perspectives have impacted nearly every scientific discipline. This new data-intensive model of science has been termed the "Fourth Paradigm" [3] and promises to further revolutionize scientific research. Among the many promises of this paradigm are the development of new AI technologies, scientific discovery through inductive data analysis rather than theory, and the reuse and combining of datasets for the generation of new knowledge.

There are many definitions of Big Data, each highlighting qualities that separate it from regular data. The multiplicity of definitions results from, in part, the heterogeneity of what is considered Big Data [12]. An early and popular definition characterized Big Data with three v's: volume, velocity, and variety, which speak to size, speed of accumulation, and heterogeneity of Big Data, respectively [13]. Scholars form *Critical Data Studies* (CDS) have since developed more nuanced definitions. One paper expanded beyond the three v's, this time defining Big Data as data that is huge in volume, high in velocity, exhaustive in scope, fine-grained in resolution, inherently relational in nature, and flexible in purpose [14]. Another paper by authors danah boyd and Kate Crawford took a socially-oriented approach and defined Big Data as a cultural, technological, and scientific phenomenon that rests on new technology, new forms of analysis, and a surrounding cultural mythology [15]. Other scholars instead pointed to particular characteristics of Big Data that differentiate it from other forms of data. For example, Elish and Boyd [16], drawing on boyd and Crawford's notion of "mythology", note the ability of Big Data and data-intensive technologies to work "like magic" in the eyes of the non-technical observer. In her essay, Radin [4] elaborates on the notion of Big Data's "flexibility" by highlighting the ability of Big Data to "ability to radically transcend the circumstances and locality of its production". In other words, a key quality of Big Data is the ability of the data to travel and be reused in wildly different contexts than its origin.

The Open Data movement seeks to further leverage the ability of data to be shared and reused. This movement seeks to make more research data publicly available in order to promote a more open research culture [17], methodological transparency [18], reproducibility [19], and to enable the re-purposing and re-analysis of data for further discovery [5]. Internet technology has facilitated the development and adoption of the so-called Big Data movement by making easier to collect, distribute, and reuse data than ever more. The Open Data movement has taken advantage of these new technologies to call for policies to encourage or mandate the sharing of research data [20], calls which have been adopted by federal agencies[1] and journals[2]. Open data policies are generally supported by the academic community [8,21,22], but the move towards open data is not without downsides. For example, making data publicly available requires intensive labor to contextualize the data and make it comprehensible to outside researchers [23] and interoperable with other related data[3] [24]; this work is rarely formally incentivized or rewarded [5]. There are also concerns that while Open Data policies may encourage data sharing, there are actually few instances of data reuse in all but a few scientific disciplines [5,9,25]. The benefits of openness and transparency also have practical limits [26]; for example, for open data to be of any use, the data needs to be understandable and interpretable by others, and in the case where replications fail or errors are found, there needs to be a system to rectify past results. There is also evidence that authors do not always fully comply with open data policies, even when mandated [7,27]. Enforcing open data policies can also pose challenges: by legislating Open Science requirements, political and commercial interests can and have weaponized the ideals of Open Data by making it more costly and difficult and costly for government research and decision-making to take place [28]. Open Data policies abound, each with opportunities and risks. However, in spite of abundant speculation and advocacy, there remains little understanding of the

[1] For example, the National Science Foundation and National Institutes of Health have each adopted open data policies.

[2] A famous example is the Public Library of Science, an open access journal publishing in a wide range of fields

[3] Data exist in complex assemblages of people, institutions, and other datasets. Often, if data is to become part of the ecosystem, it follow certain standards. For example, if data relates to the incidence of diseases in hospitals between countries, then when making this data public, disease identification codes should be made commensurate with international standards

practical impact of data sharing policies on how data is reused *in the wild*.

Issues related to data sharing and reuse

The rise of Big Data and internet technologies has made it easier for data to be reused, but the process is still not without a great deal of friction. For example, making data available for reuse requires a process of de-contextualization in which the data is formatted and made interoperable with other sources [23]; following this, the data must then be re-contextualized by documenting it and assigned appropriate metadata to ensure that downstream researchers can understand its origin. Data that is not made interoperable with other data unlikely to be reused whereas data that is devoid of context becomes meaningless [15, 29]. Making data available for reuse requires intense labour and often technical expertise, work that is rarely rewarded [5]. Moreover, sharing data puts researchers at risk of getting "scooped", losing control of their data to commercial interests, or of not receiving appropriate recognition [30]. For researchers reusing data, there is work necessary to understand the context of data's origin and processing required to make it fit for analysis [24]; Even when reusing data within the same research team, the act of making data commensurate can be labour and time intensive [31]. There are also infrastructural concerns when managing, storing, and moving data, all of which require technical expertise and resources not necessarily available to most researchers [32]. Solutions have been proposed to make data sharing and reuse easier. For example, datasheets for datasets provide a semi-standardized template for outlining the context of a dataset such that it can more easily be reused [33], but they have not been widely adopted. For Open Data policies to succeed, policy-makers and institutions require a deeper understanding of the practical impediments to data sharing and reuse.

Another hindrance to widespread adoption of open data practices are the lack of normative data citation practices which make it difficult for authors of open datasets to receive credit [8]. Citation counting began in earnest with the creation of the Science Citation Index in 1955 [34], allowing the creation of massive bibliometric databases like the Web of Science and Scopus, the aggregation of research metrics, and the large-scale study of citation practices. Extending the traditional journal and publication citation index towards datasets seems, at first glance, a straightforward procedure. However, it is made incredibly difficult due to a lack of widely accepted data citation practices; in spite of organizational efforts to implement common guidelines [35] there remains no widely accepted data citation format or norms. Data citation is further complication by issues of data provenance—the dynamic and chancing history of datasets—as well as concerns over intellectual property rights and how to fairly attribute credit to large and heterogeneous teams [5, 32]. Moreover, heterogeneous citation practices make it difficult for bibliometric databases to disambiguate data citations; for example, a search for the "UCI Machine Learning Repository" on the Web of Science reveals 152 distinct citable objects, each represented in a different format. Citation metrics are one means by which researchers are evaluated; without established data citation practices that can factor into these metrics, researchers have little incentive to make their data publicly available.

The amplified ability for data to be shared and reused may bring opportunities for research, but may also involve methodological downsides. Bigger data does not necessarily mean better data; large datasets suffer from the same sampling biases [6, 14] as smaller data, but also encourage hubris and overconfidence in results [36]. Data reuse—another attribute of Big Data—can also presents methodological issues. For example, widespread reuse of a small number of datasets without common guidelines for their cleaning and processing could result in high *"researcher degrees of freedom"* [37,38]—points of flexibility in study design that could result in different trajectories of analysis. Crowd-sourcing analysis of open data could allow for more robust and reproducible results [32], but there is also danger that exciting or statistically-significant outcomes are published and exist as part of the scientific record before contradictory results using other analyses and different procedures are published [39, 40]. These researcher degrees of freedom may further multiply when combining datasets. the ability to bring together multiple datasets to conduct more expansive analysis is a hallmark of the Big Data revolution. However, each dataset requires choices of how they should be processed, and further judgment is needed to decide how datasets should be merged; the result is far greater flexibility in data analysis and thus added researcher degrees of freedom. Combining datasets, all of which are error-prone— also has the potential of amplify local errors across

many analyses [15]. Just because data is reused does not necessarily entail methodological advantages; policies and practices should encourage effective and desired reuse of data and establish systems for re-analysis and replication.

A focus on data reuse over creation of new data has epistemic consequences with the potential to shape the kind of research being done. Much has been written about the epistemic implications of Big Data to certain disciplines and science as a whole [14, 41, 42]; however, this scholarship has tended to focus on the implications of this data's size, and not on the epistemic consequences of re-analyzing and reusing datasets. In biomedical research, the use of certain model organisms such as Drosophila (fruit flies), mice, or axolotls constrains the possible avenues of research—the choice of organism serves as a framing that determines the kinds of questions and analyses that are possible or conceivable. Drosophilia, for example, are suited for genetic research whereas mice are better suited for the development of medical drugs. This phenomenon extends into technical fields; early computer science research chose the game *Chess* as a model problem for the development of artificial intelligence [43]—this narrowed the field's research focus to adversarial problem solving in a confined game environment. Solutions to *Chess* tended to focus on deterministic and brute-force algorithmic strategies. A game like *Go* has a much greater set of possible moves and so may have encouraged research focused on probabilistic or heuristic-based decisions, had it been the model problem. Amid the Big Data revolution, data itself serves as a metaphor for how problems are framed and approached, a metaphor that even impacts even qualitative disciplines [44]. Similarly, datasets themselves can constrain the kinds of questions being asked. If developing Machine Learning algorithms for classification, a researcher may choose to benchmark their algorithm with popular and easily available data from a public repository; the characteristics of such popular datasets could impact the direction of the algorithm's development. For example, if the chosen dataset has an incredibly high number of correlated features, then the researcher may embed the data into lower-dimensional space as a default step of the algorithm. Similarly, if the data contains a high number of missing values, then the researcher may instead develop their algorithm to effectively deal with a statistically-imputed version of the data. Other researchers have raised concerns over the impact of model datasets on the field of Machine Learning; at an early conference, one researcher "...passionately decried how it [the UCI Machine Learning Repository] allowed researchers to publish dull papers that proposed small variations of existing supervised learning algorithms and reported their small-but-significant incremental performance improvements in comparison studies" [4]. Choice of datasets can shape the direction of research. If reuse and re-purposing of a small subset of datasets is the norm, then there is risk that the most popular of these data will dictate the direction of research in Machine Learning and other fields.

In addition to all these issues are concerns over the ethics of data reuse. One of the most common criticisms of data sharing relates to the handling of sensitive material, typically identifiable information about research subjects [6]. Institutional Review Boards (IRBs) have served to enforce ethical standards in research, requiring that informed consent be obtained and data be properly anonymized and secured; however, the protocols and requirements of IRBs are often inadequate for current data-driven science [45]. One risk that the IRB does not account for is the potential for combining multiple data sources in order to identify individuals. Already, individuals have been identified using data scraped from dating sites [46] and from employee data [47]. In one case, researchers combined property records and geospatial data in order to identify the anonymous artist Banksy [48]. Informed consent, a staple of research ethics, is also complicated by data sharing. For example, an individual who chose to provide data to a researcher may feel uncomfortable if that data were to be distributed widely or reused by commercial or political interests, anonymized or not. Participants have little control over how their data is reused. One paper [4] examined the ethics and history of the ***Diabetes*** dataset on the UCI Machine Learning repository—a dataset with an origin of exploitation of the indigenous Pima Native American population in Arizona; this community was not rewarded for the labour involved in creating their data, yet the data continues to be used across many fields of research. In another recent example, images of a vulnerable population were acquired without consent and used for the testing of facial recognition algorithms [49]. These examples demonstrate that just because data is accessible does not mean that its use is ethical [15]; similarly, strictly adhering to open data practices is not always the moral choice [6]. The collection and use of research data entails inherent ethical risks; the amplified ability of data to be shared and reused can also

amplify these risks and cause new harms.

Approaches to study data reuse

Study of Open Data and Data Sharing has typically emerged out of the disciplines of Library and Information Science, Bibliometrics, and most recently Critical Data Studies. In Library and Information Science literature, studies have applied quantitative approaches towards examining data sharing practices in biomedical research [21] and in the open-access journal PLoS One [7]. Other studies have investigated the extent to which researchers comply with data sharing policies [27]. Another branch of work has involved the use of surveys [8, 22] and ethnographic fieldwork [5, 9] to examine author's perceptions and the practical realities of data sharing. A common finding of these studies is that practices for data sharing and reuse are incredibly heterogeneous, varying between discipline, institutions, and even between individual teams involved in the same project. Authors generally support Open Data policies, but they also worry over the effort and risk inherent to the process of making their data available [8, 22]. One study found that, even after claiming compliance with data sharing policies, many authors were either unavailable or reluctant to share their research data [50]. Due to some or all of these issues, the "dirty little secret" of data sharing is that shared data is very rarely reused [5]. Library Science has been at the forefront of the Open Science and Open Data movements, however other fields have also contributed distinct perspectives.

Bibliometric researchers have focused on using quantitative tools and massive bibliometric datasets in order to conduct macro-level analyses of science, such as investigating global gender disparities among scientists [51], the mobility of the scientific workforce [52], national differences in science production [53,54], and the rate of Open Access compliance [55]. Other related fields have also emerged such as the Science of Science [56] and Metaknowledge [57]. However, researchers from these fields have struggled to overcome the technical challenges inherent to studying data reuse through data citation. Some researchers have had success analyzing the Data Citation Index—a bibliometric index of datasets and associated citations aggregated by the Web of Science—however, direct-dataset citations are rarely used leaving most indexed datasets uncited in all but a few disciplines [10, 58]. Other researchers have investigated other forms of data publication [59] and have proposed standardized publication formats that are more amenable to citation analysis such as the data paper [60]. However, without widespread adoption of common norms and practices for data publication and citation, data reuse will remain difficult to study. This results in a catch-22 in which a lack of indicators of data reuse makes it difficult to craft policies to encourage responsible and effective reuse of data; meanwhile, lack of such policies contributes to the dearth of actual data reuse, or at lease empirical evidence of its reuse.

Emerging from *Science and Technology Studies*, the field of *Critical Data Studies* has investigated the social, cultural, and political aspects of Big Data and other data-driven technologies [61, 62]. In particular, they seek to study "data assemblages", sociotechnical structures that weave together the "technological, political, social and economic apparatuses and elements that constitutes and frames the generation, circulation, and deployment of data" [62]. Studies in this field have tended to emerge out of ethnographic methodology and have developed rich theoretical frameworks for understanding data. In regards to data sharing, scholars have examined the issues of anonymity, informed consent, and social consequences of using public data or making this data available [4, 15]. Others have noted how the meaning of data relies on its relation to the context of its origin [14, 15, 23, 29], which has dire implications for data reuse—an act that requires interpreting data from a foreign context. Critical Data Studies and related fields have provided a strong theoretical foundation for the study of data practices. However, there are few studies that specifically focus on the practical, methodological, epistemic, and ethical issues of data reuse. Moreover, there is a need for generalizable and empirical theoretical confirmation in order to make these theoretical frameworks more accessible and appealing to data scientists and policy makers.

Research questions

Data sharing and reuse entail opportunities, such as supporting replication efforts, opening data to public use, and furthering scientific discovery. However, these opportunities come coupled with risks and

challenges, such as practical impediments to data sharing, ethical concerns over privacy, anonymization, consent, and a myriad of other methodological, epistemic, and social concerns. To navigate these opportunities and challenges, effective scientific and data governance requires a generlaizable and macro-level understanding not only of how scientific data is shared, but also how it is reused.

I propose an empirical study to provide a foundation for policy discussions about data reuse. I will apply bibliometric analysis to an exemplar source of reusable data—the UCI Machine Learning Repository. This repository's potential for near-frictionless data reuse allows it to serve as an *upper bounds* on data reuse—a target for data reuse policies to strive for, or caution over what policies should be avoided. By scoping analysis to a single repository and the data contained within, this analysis will make it possible to handle technical and methodological issues in an *ad-hoc* manner. My primary aim is to develop an understanding of the speed and extent to which data is reused and the factors that mediate its movement; more specifically, I propose the following three research goals,

1. Quantify the extent to which datasets on the UCI Machine Learning Repository are cited and reused in scientific literature in terms of speed (how often is the dataset cited over time), conceptual travel (to what extent is the data used outside of its original disciplinary context), and geographic travel (to what extent is the data used outside of its geographic origin)

2. Determine the extent to which features of a dataset, such as its size, data types, or domain, are associated with its measures of speed, conceptual travel, and geographic travel

3. Measure the degree to which the publication of a dataset to the UCI Machine Learning Repository is associated with changes to its measures of speed, conceptual travel, and geographic travel

By addressing these goals, I aim to provide generalizable insights into how data is reused in the scientific literature. This study will provide a baseline of data sharing under near-ideal conditions which can be used to inform Open Data policies. By framing results in the context of ethical, epistemic, and methodological issues of data reuse, this study will also provoke further discussions about the potential benefits, risks, and practicalities of open data policies. Finally, in pursuing these goals, I will develop a technical and methodological framework that can be re-purposed for future studies of data reuse.

1 Data and Methodology

UCI Machine Learning Repository

The University of California Irvine (UCI) Machine Learning Repository is a public repository that provides a standardized set of data aimed towards the development and evaluation of machine learning algorithms. At the time of writing, the repository contains 496 donated datasets representing topics as broad as measurement of flower petals, incidence of heart disease, university characteristics, wine quality, and volcanic eruptions on Venus. Outside of interviews conducted by Joana Radin [4], there is little documented history of the repository except for the abridged background and purpose listed on its website,

> "The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science. The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman, and this project is in collaboration with Rexa.info at the University of Massachusetts Amherst. Funding support from the National Science Foundation is gratefully acknowledged."

The UCI Machine Learning Repository is not the only open-data repository available. Many governmental datasets has been made available on services like *Data.gov*, *Data.gov.uk*, and *Data.taipei*. Commercial entities have also developed repositories of open datasets such as those indexed on *Google Dataset Search*. Some companies have released curated datasets for the purposes of crowd-sourcing analysis or identifying technical talent such as that released by the Yelp Dataset Challenge. Disciplinary repositories offer another instance of open data, such as the NCBI Genome Expression Omnibus. Other organizations focused on Open Science, such as FigShare, Dryad, and the Center for Open Science provide other means of making datasets accessible. However, there are several characteristics of the UCI Machine Learning Repository that make it unique from others. For one, the data used in this repository is widely used and recognized—the page for the most popular dataset, ***iris***, has been viewed 2,580,221 times (at the time of writing). Most datasets are also easy to download and use, are usually clearly documented and well-structured, are relatively small in size, and are usable without knowing much of their original context. Some of these data, such as the ***iris*** and ***mtcars*** datasets are even included in the installation of the R programming language. This repository makes data reuse as close to *plug-and-play* as is likely possible. The repository is also, compared to many others, relatively old; the repository itself was created in 1987 and many of the included datasets have been hosted for decades; this allows for historical analyses of how these data were used alongside the development of the fields of Machine Learning and Artificial Intelligence. Lastly, the UCI Machine Learning Repository exists independent of any strict disciplinary or institutional control, and thus tend not to be subject to Open Data policies that may confound understandings of data citation *in the wild*.

UCI Repository Metadata

Through a combination of web scraping and manual efforts, I will collect the associated metadata for all 496 datasets listed (at the time of writing) on the UCI Machine Learning Repository. Each dataset is associated with a set of descriptive information added by the original donor. The quality, extent, and format of these descriptions vary between datasets. In addition to this descriptive information, there are also a small set of more standardized metadata that describe the characteristics of the data, its attributes, the tasks associated with the dataset, the number of instances contained within the data, the number of attributes, whether the dataset contains missing values, the domain area, the date donated, and the number of web hits at the time of access. Examples of these values for the five most popular datasets listed on the repository's home page are shown in table 1.

Table 1. Characteristics listed for the top four most popular datasets on the UCI Machine Learning Repository. Values are taken directly from the download page of each of these datasets. Values were copied at the time of writing, on the 4th of May, 2019. MV = Multivariate; REAL = Real Values; CAT = Categorical; INT = integer, REG = Regression; CLASS = Classification.

Feature	Iris	Adult	Wine	Car Evaluation	Wine Quality
Data set characteristics	MV	MV	MV	MV	MV
Attribute Characteristics	REAL	CAT, INT	INT, REAL	CAT	REAL
Associated tasks	CLASS	CLASS	CLASS	CLASS	CLASS, REG
Number of Instances	150	48,842	178	1,728	4,898
Number of Attributes	4	14	13	6	12
Missing values?	No	Yes	No	No	N/A
Area	Life	Social	Physical	N/A	Business
Data Donated	7/1/88	5/1/96	7/1/91	6/1/97	10/7/09
Number of web hits	2,582,004	1,474,649	1,137,240	969037	924,524

Each dataset also includes unstructured metadata detailing its origins and including the name and contact information of the data's creator and donor. These information are highly idiosyncratic with each dataset often being associated with several individuals. I will collect the name and national and institutional affiliations of each individual associated with each dataset and assign each a role based on their stated relation to the data (creator, owner, donor, co-author, etc.). Donors also have the option

to post one or more papers relevant to the dataset. For example, the ***iris*** dataset is associated with four papers—the seminal work in linear discriminant analysis by Fisher in 1935 [63], along with three others which made use of the data at later times [64–66]. For datasets that list them, I will collect the titles and DOIs of all relevant papers. The UCI Machine Learning Repository also lists papers that cite each dataset collected using Rexa.info[4], though I will not collect these information due to ambiguity and limitations in how citing papers were identified[5].

Elsevier Science Direct Full-Text Dataset

The proposed study requires capturing the number of citations received by the UCI Machine Learning Repository and datasets that it contains. Citations to these datasets may take two forms, either a citation to the repository as a whole, or a citation to a publication related to an individual dataset hosted on the repository. However, a citation to one of these publications may not be referencing the dataset itself, but instead referencing the contents of the publication. For example, the ***iris*** dataset is associated with Fisher's 1936 paper on Linear Discriminant Analysis [63], and so a citation to this paper may refer to either the method or the dataset. Similarly, a reference to the UCI Machine Learning Repository itself may in fact be referencing a particular dataset. Determining what a citation is actually referencing requires analysis of the actual text of the sentence containing the citation—the *citance*.

The *Elsevier ScienceDirect* database[6] contains the full-text of nearly five million English-language research articles, short communications, and review articles published between 1980 and 2016. Sentences containing in-text citations (*citances*) were extracted from the full-text of these articles following the procedure outlined by Boyack et al. [41]. Each citance was matched with relevant metadata including the sentence's length, its position in the manuscript, its position in the current paragraph, the position of the in-text citation in the sentence, and the name of the section in which it appears. Each publication indexed in the database was matched to a publication in the *Web of Science* (WoS)—a massive bibliometric database hosted at Leiden University. This database includes bibliometric information such as citation counts, journal of publication, the disciplinary classification of the journal, and disambiguated author information[7] such as author name and institutional affiliation. For each of the 496 UCI repository datasets, I will collect all citances made to the relevant publications listed to each dataset. I will also collect all citances made in reference to the UCI Machine Learning Repository itself.

Citation Classification

I will develop a simple machine learning classifier using the features of a citance to automatically determine whether a citation is referring to the data or the content of a publication. Similar text-classification approaches have been developed to classify the importance [69, 70], sentiment [71, 72], and function [73] of citations. Here, classification will be binary, based on whether or not the citance is referring to the data or the content of the cited publication.

A gold-standard of training and testing data will be created by sampling approximately 1,000 citances from the set of citances referencing publications relevant to datasets in the UCI Machine Learning Repository. I will then manually label each citance as being either in reference to the dataset or to the content of the publication. If possible, I will enlist the help of a co-author to code a small sample of these selected citances in order to calculate inter-rater reliability. For each citance I will extract

[4]From their website, Rexa.info is a "a digital library and search engine covering the computer science research literature and the people who create it.", and is a testbed for information extraction and co-reference analysis research created and maintained by researchers at the University of Massachusetts, Amherst.

[5]Limitations include 1) Rexa.info has not been used by the bibliometric community, and so potential shortcomings and limitations are not as well known as more common databases like the Web of Science or Scopus; 2) Rexa.info includes only publications in Computer Science, potentially missing a wide range of dataset use in other disciplines; 3) the citing papers comprise journal articles, conference proceedings, and doctoral dissertations, with little disambiguation between citing object types; and 4) publications cannot be readily matched with DOIs to existing bibliometric databases that maintain disambiguated researcher and institutional identities.

[6]This database is hosted, and major processing took place at the *Centre for Science and Technology Studies* at Leiden University.

[7]Disambiguated using the procedure in [67]. This procedure was recently favorably evaluated in [68]

features representing its position in paper, the name of the section in which it appears, whether the sentence explicitly uses the name of the dataset, and presence of terms form a dictionary of manually-curated signal words (e.g., "data", "dataset", "method", "result", "finding"). Other features, such as n-grams, dependency structures, and parts-of-speech tagged tokens have proven useful in citance classification [72, 73]; however, due to the small size of training data, a smaller number of simple features is more appropriate.

After creating a labelled gold-standard dataset, I will train and evaluate a classifier using several different classification techniques that have proven useful for binary or sentence classification. These techniques include Logistic Regression, Naive Bayes, Support Vector Machines, simple Decision Trees, and Random Forests; these last three techniques have the added benefit of producing relatively interpretable outputs detailing the impact of certain features. Other classification techniques, such as Convolutional or Recurrent Neural networks have proven useful for sentence classification but are inappropriate for the size of training data and features considered in this study. Classification approaches will be evaluated using multiple iterations of 10-fold cross-validation. Measures of average precision, recall, and F1 will be aggregated over all iterations. Assuming sufficient performance, the best classifier will be used to automatically classify the remaining citances.

In some cases, authors may cite the UCI Machine Learning Repository itself when they are in fact referring to a particular dataset contained within. If the number of references to the repository is small, then citations can be matched to individual datasets manually; if large, then simple rule-based classification, such as matching based on explicit use of dataset name, should be sufficient.

Indicators of reuse

To address the stated research goals I will compile a family of indicators for each dataset in the UCI Machine Learning Repository that aim to capture their speed of reuse, the conceptual distance travelled, and the geographic distance travelled. Speed of reuse is the simplest construct to measure. For speed of reuse I will implement two indicators. The first is the yearly count of citations made to the dataset; this indicator will capture the rate at which each dataset is reused by the academic community. I will compare this against the second indicator which captures a dataset's overall popularity (both inside and outside of academic research) using the yearly number of new web hits displayed on the dataset's web page. Using the *Wayback Machine*—an internet archive containing past versions of web pages. I will extract the number of web hits made to a dataset's page in the UCI Machine Learning Repository each year (as close to the same time each year as is possible given available archived versions). Given the labour involved in this process, I will only compile yearly new web hits for a small number of the most popular datasets.

Measures of conceptual travel will be compiled by first creating an matrix containing distances between all pairs of scientific disciplines represented in the Elsevier ScienceDirect database. There are several ways to compile such distances including the use of citation links between journals or topic modelling of text. For this study, I will use a relatively recent approach that defines a disciplinary space by clustering on the citation links between individual publications [74]. Given past results using this method, I expect several hundred, but less than a thousand disciplinary clusters. For each dataset in the UCI Machine Learning Repository, I will map its associated publications to the corresponding disciplinary cluster; this will be the *origin* cluster. In the case that a dataset is associated with multiple papers that map to several clusters, then the point equidistant to each cluster will be used as the origin. Then, I will identify *citing* clusters—disciplinary clusters containing publications that cite the dataset. These citing clusters will be used to compile a family of indicators including 1) the number of citing clusters; 2) the average of citation counts within each cluster; 3) the average of citation counts in each citing cluster weighted by the distance to the origin; 4) the average euclidean distance between the origin and each citing cluster, weighted by the number of citations; and 5) the euclidean distance between the origin cluster and the furthest citing cluster. By iterating this process over each year of data (for example, including all papers published up until 2010, then up to 2011, 2012, etc.), I will be able to track how these indicators develop over time. In the case that a dataset existed before it was donated to the UCI Machine Learning Repository (as for the ***iris*** dataset), I will compile indicators of conceptual distance using two

sets of papers, those published before and those published after the data of the dataset's donation.

Measures of geographic travel will be compiled in much the same way as was conceptual travel. I will construct two matrices, one detailing the geographic distances between the centroids of countries, and another with the geographic distance between the centroids of states within the United States. The *origin* country will be assigned to countries and to states (for U.S. authors) based on the country or state of affiliation of the data's original owner and the relevant paper's co-authors[8]. Citing countries will be classified as those for which at least one paper was published that cites the dataset. For each dataset I will use the distances between the origin and citing clusters to compile the same five indicators as for conceptual travel. I will also compile yearly indicators to track their development over time and before and after dataset's donation to the repository. This approach to measuring geographic distance can also be used in an *ad-hoc* manner to study geographic travel in relation to points of interest; for example, following the discussion by Joanna Radin [4], the origin of the ***diabetes*** dataset can be set to the geographic coordinates of the Pima Indian reservation, the location of the population whom the data represents. Such findings can be used to provoke discussions around the ethics of certain forms of data reuse.

Analytic Strategy

My first research goal is to quantify the extent to which a typical dataset on the UCI Machine Learning Repository is reused in scientific research. I have detailed three families of indicators to measure to speed of reuse, conceptual travel, and geographic travel of data reuse. I will compile these indicators for each dataset and for the repository as a whole, including both citations to the repository itself and to datasets contained within. These indicators will provide an overview of data reuse. For some datasets of interest I will also provide visualizations by embedding conceptual and geographic distance matrices into two-dimensional space and marking clusters based on the number of citing publications. These analyses, along with additional descriptive statistics, will satisfy my first stated research question. I will address the second research goal by using linear regression to measure the extent to which dataset characteristics, such as size, number of attributes, associated tasks, and topical domain contribute to the speed and extent of data reuse. Finally, for datasets that existed before they were donated to the repository, I will use the temporally-constructed indicators to compare patterns of reuse in the years before and after the dataset's donation to the repository; this will address my third research goal.

Challenges & Limitations

Here, I proposed a preliminary study to investigate the extent and nature of data reuse in academic research. Being preliminary, this research is subject to several challenges and limitations. One limitation is that the Web of Science database is known to have poor coverage in some areas of research, especially for the Social Sciences, Humanities, and non-English-language publications [75]; the proceedings of many Computer Science conferences—perhaps those most likely to utilize the UCI Machine Learning Repository—may not be well indexed on the Web of Science platform. Another limitation of this approach is that the use of citation databases to track reuse will not capture instance in which a dataset is mentioned by name but is not formally cited; full-text searches for known dataset names may mitigate this limitation, but as of now the *Elsevier ScienceDirect* dataset indexes only citances, not full-text. In spite of these limitations, the version of the *Web of Science* hosted by Leiden University represents one of the most well-studied and well-structured citation databases presently available. Implementing the approach detailed in this proposal will build atop this existing technical infrastructure and develop a framework that can be extended to other citation and full-text databases.

Another limitation emerges from the potential for systematic bias in citation classification. For example, classification may perform well in biomedical literature, yet have lower performance when

[8] In the case of multiple authors, I will consider several approaches to attributing credit. Full Counting assigns equal credit to countries of affiliation for every co-author; Fractional Counting applies fractions of contribution based on the number of authors fro each country; Corresponding counting only considers the country of the corresponding author

applied to social science literature; this could result in biases that underrepresent data reuse of data in certain disciplines. To mitigate this risk, when sampling the 1,000 training instances, 200 will be sampled form each of five broad disciplinary categories: engineering, medical science, natural science, social science, and arts and humanities. This will ensure a degree of disciplinary diversity in the training set and allow for some analysis of classification performance by discipline.

Perhaps the largest limitation of this analysis relates to its generalizability. Previous research has demonstrated that data sharing needs and practices vary wildly by discipline, place, and even between individual research teams [15, 32]. Similarly, the patterns of data reuse identified in this study may be too heterogeneous to draw clear conclusions, or else may not generalize to specific disciplinary and methodological contexts. Still, this research is a first step towards a more thorough understanding of data reuse *in the wild*, and the findings of this analysis will ideally help policy makers navigate the opportunities but also the practical, epistemic, methodological, and epistemic issues inherent to data sharing and reuse.

2 Research Outputs

This research will result in several outputs that will be published in a way consistent with Open Science best practices. After collecting all metadata for each of the 496 datasets in the UCI Machine Learning Repository, i will submit initial descriptive results as a short paper to the 2020 iConference; at the same time, I will donate the metadata to the UCI Machine Learning Repository or otherwise make it publicly available so that it can prove useful to future analyses of Open Data or to the development of machine learning algorithms. Following completion of analyses detailed in this proposal, I will also make final data files publicly available. Due to issues over intellectual property, data from the *Elsevier ScienceDirect* database and the Web of Science cannot be made publicly available. However, to ensure transparency, all code used for processing, compiling indicators, and conducting analyses will be posted to Github and made publicly available along with sample fake datasets that can be used for testing and verification. Final results will be published to an interdisciplinary open access journal such as PLoS One, which has a history of publishing articles related to data sharing and reuse.

Acknowledgments

I would like to thank my qualifying exam committee for their patience and understanding as I worked through my ideas for this proposal. I would also like to thank the Department of Informants and the many amazing professors and colleagues who have had an impact on my development as a research. I would especially like to thank Dr. Cassidy Sugimoto for her wonderful mentorship as I have progressed through my doctoral program.

References

1. Hanson B, Sugden A, Alberts B. Making Data Maximally Available. Science. 2011;331(6018):649–649. doi:10.1126/science.1203354.

2. Mayer-Schönberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Reprint edition ed. Boston: Eamon Dolan/Mariner Books; 2014.

3. Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. 1st ed. Redmond, Washington: Microsoft Research; 2009.

4. Radin J. "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. Osiris. 2017;32(1):43–64. doi:10.1086/693853.

5. Borgman CL. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology. 2012;63(6):1059–1078. doi:10.1002/asi.22634.

6. Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, et al. Ten simple rules for responsible big data research. PLOS Computational Biology. 2017;13(3):e1005399. doi:10.1371/journal.pcbi.1005399.

7. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu YL, Snyders LN, et al. Data sharing in PLOS ONE: An analysis of Data Availability Statements. PLOS ONE. 2018;13(5):e0194768. doi:10.1371/journal.pone.0194768.

8. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. PLOS ONE. 2011;6(6):e21101. doi:10.1371/journal.pone.0021101.

9. Wallis JC, Rolando E, Borgman CL. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLOS ONE. 2013;8(7):e67332. doi:10.1371/journal.pone.0067332.

10. Robinson[U+2010]García N, Jiménez[U+2010]Contreras E, Torres[U+2010]Salinas D. Analyzing data citation practices using the data citation index. Journal of the Association for Information Science and Technology. 2016;67(12):2964–2975. doi:10.1002/asi.23529.

11. Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired. 2008;.

12. Kitchin R, McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society. 2016;3(1):2053951716631130. doi:10.1177/2053951716631130.

13. Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group; 2001.

14. Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data & Society. 2014;1(1):2053951714528481. doi:10.1177/2053951714528481.

15. boyd d, Crawford K. Critical Questions for Big Data. Information, Communication & Society. 2012;15(5):662–679. doi:10.1080/1369118X.2012.678878.

16. Elish MC, Boyd D. Situating Methods in the Magic of Big Data and Artificial Intelligence. Rochester, NY: Social Science Research Network; 2017. ID 3040201. Available from: `https://papers.ssrn.com/abstract=3040201`.

17. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348(6242):1422–1425. doi:10.1126/science.aab2374.

18. Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, et al. Promoting Transparency in Social Science Research. Science. 2014;343(6166):30–31. doi:10.1126/science.1245317.

19. McNutt M. Reproducibility. Science. 2014;343(6168):229–229. doi:10.1126/science.1250475.

20. Gewin V. Data sharing: An open mind on open data. Nature. 2016;529(7584):117–119. doi:10.1038/nj7584-117a.

21. Federer LM, Lu YL, Joubert DJ, Welsh J, Brandys B. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. PLOS ONE. 2015;10(6):e0129506. doi:10.1371/journal.pone.0129506.

22. Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE. 2015;10(8):e0134826. doi:10.1371/journal.pone.0134826.

23. Leonelli S. What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society. 2014;1(1):2053951714534395. doi:10.1177/2053951714534395.

24. Ribes D. Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. CSCW '17. New York, NY, USA: ACM; 2017. p. 1514–1526. Available from: http://doi.acm.org/10.1145/2998181.2998344.

25. Pasquetto I, Randles B, Borgman C. On the Reuse of Scientific Data. Data Science Journal. 2017;16(0):8. doi:10.5334/dsj-2017-008.

26. Ananny M, Crawford K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society. 2018;20(3):973–989. doi:10.1177/1461444816676645.

27. Piwowar HA, Chapman WW. Public sharing of research datasets: a pilot study of associations. Journal of informetrics. 2010;4(2):148–156. doi:10.1016/j.joi.2009.11.010.

28. Levy KE, Johns DM. When open data is a Trojan Horse: The weaponization of transparency in science and governance. Big Data & Society. 2016;3(1):2053951715621568. doi:10.1177/2053951715621568.

29. Neff G, Tanweer A, Fiore-Gartland B, Osburn L. Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. Big Data. 2017;5(2):85–97. doi:10.1089/big.2016.0050.

30. Costello MJ. Motivating Online Publication of Data. BioScience. 2009;59(5):418–427. doi:10.1525/bio.2009.59.5.9.

31. Gitelman L, editor. "Raw Data" Is an Oxymoron. Cambridge, Massachusetts ; London, England: The MIT Press; 2013.

32. Borgman CL. Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, Massachusetts: The MIT Press; 2015.

33. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumeé III H, et al. Datasheets for Datasets. 2018;.

34. Garfield E. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science. 1955;122(3159):108–111. doi:10.1126/science.122.3159.108.

35. Cite CITGoDCSaPo, Sices OoMTC. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal. 2013;12(0):CIDCR1–CIDCR7. doi:10.2481/dsj.OSOM13-043.

36. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014;343(6176):1203–1205. doi:10.1126/science.1248506.

37. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011;22(11):1359–1366. doi:10.1177/0956797611417632.

38. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science. 2018;1(3):337–356. doi:10.1177/2515245917747646.

39. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Medicine. 2005;2(8). doi:10.1371/journal.pmed.0020124.

40. Young NS, Ioannidis JPA, Al-Ubaydli O. Why Current Publication Practices May Distort Science. PLoS Medicine. 2008;5(10). doi:10.1371/journal.pmed.0050201.

41. Boyack KW, van Eck NJ, Colavizza G, Waltman L. Characterizing in-text citations in scientific articles: A large-scale analysis. arXiv:171003094 [cs]. 2017;.

42. Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, Bowman T, et al. Big data, bigger dilemmas: A critical review. arXiv:150900909 [cs]. 2015;.

43. Ensmenger N. Is chess the drosophila of artificial intelligence? A social history of an algorithm. Social Studies of Science. 2012;42(1):5–30. doi:10.1177/0306312711424596.

44. Markham AN. Undermining 'data': A critical examination of a core term in scientific inquiry. First Monday. 2013;18(10). doi:10.5210/fm.v18i10.4868.

45. Shmueli G. Research Dilemmas with Behavioral Big Data. Big Data. 2017;5(2):98–119. doi:10.1089/big.2016.0043.

46. Cox J. 70,000 OkCupid Users Just Had Their Data Published; 2016. Available from: https://motherboard.vice.com/en_us/article/8q88nx/70000-okcupid-users-just-had-their-data-published.

47. Pandurangan V. On Taxis and Rainbows; 2014. Available from: https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1.

48. Hauge MV, Stevenson MD, Rossmo DK, Comber SCL. Tagging Banksy: using geographic profiling to investigate a modern art mystery. Journal of Spatial Science. 2016;61(1):185–190. doi:10.1080/14498596.2016.1138246.

49. Wernimont JOK Nikki Stevens. The Government Uses Images of Abused Children and Dead People to Test Facial Recognition Tech. Slate Magazine. 2019;.

50. Savage CJ, Vickers AJ. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLOS ONE. 2009;4(9):e7078. doi:10.1371/journal.pone.0007078.

51. Larivière V, Ni C, Gingras Y, Cronin B, Sugimoto CR. Bibliometrics: Global gender disparities in science. Nature News. 2013;504(7479):211. doi:10.1038/504211a.

52. Sugimoto CR, Robinson-Garcia N, Murray DS, Yegros-Yegros A, Costas R, Larivière V. Scientists have most impact when they're free to move. Nature. 2017;550(7674):29–31. doi:10.1038/550029a.

53. May RM. The Scientific Wealth of Nations. Science. 1997;275(5301):793–796. doi:10.1126/science.275.5301.793.

54. King DA. The scientific impact of nations. Nature. 2004;430:311.

55. Larivière V, Sugimoto CR. Do authors comply when funders enforce open access to research? Nature. 2018;562(7728):483. doi:10.1038/d41586-018-07101-w.

56. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. Science. 2018;359(6379):eaao0185. doi:10.1126/science.aao0185.

57. Evans JA, Foster JG. Metaknowledge. Science. 2011;331(6018):721–725. doi:10.1126/science.1201765.

58. Peters I, Kraker P, Lex E, Gumpenberger C, Gorraiz J. Research data explored: an extended analysis of citations and altmetrics. Scientometrics. 2016;107(2):723–744. doi:10.1007/s11192-016-1887-4.

59. Parsons M, Fox P. Is Data Publication the Right Metaphor? Data Science Journal. 2013;12(0):WDS32–WDS46. doi:10.2481/dsj.WDS-042.

60. Chavan V, Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics. 2011;12(Suppl 15):S2. doi:10.1186/1471-2105-12-S15-S2.

61. Dalton CM, Taylor L, Thatcher (alphabetical) J. Critical Data Studies: A dialog on data and space. Big Data & Society. 2016;3(1):2053951716648346. doi:10.1177/2053951716648346.

62. Iliadis A, Russo F. Critical data studies: An introduction. Big Data & Society. 2016;3(2):2053951716674238. doi:10.1177/2053951716674238.

63. Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 1936;7(2):179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

64. Dasarathy BV. Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Exposed Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1980;PAMI-2(1):67–71. doi:10.1109/TPAMI.1980.4766972.

65. Duda RO, Hart PE. Pattern Classification and Scene Analysis. 1st ed. New York: Wiley; 1973.

66. Gates G. The reduced nearest neighbor rule (Corresp.). IEEE Transactions on Information Theory. 1972;18(3):431–433. doi:10.1109/TIT.1972.1054809.

67. Caron E, van Eck NJ. Large scale author name disambiguation using rule-based scoring and clustering. In: Proceedings of the Science and Technology Indicators Conference 2014. Leiden, Netherlands: Leiden University; 2014. p. 79–86.

68. Tekles A, Bornmann L. Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. arXiv:190412746 [cs]. 2019;.

69. Pride D, Knoth P. Incidental or influential? - Challenges in automatically detecting citation importance using publication full texts. arXiv:170704207 [cs]. 2017;.

70. Wan X, Liu F. Are all literature citations equally important? Automatic citation strength estimation and its applications. J Assn Inf Sci Tec. 2014;65(9):1929–1938. doi:10.1002/asi.23083.

71. Catalini C, Lacetera N, Oettl A. The incidence and role of negative citations in science. PNAS. 2015;112(45):13823–13826. doi:10.1073/pnas.1502280112.

72. Jha R, Jbara AA, Qazvinian V, Radev DR. NLP-driven citation analysis for scientometrics. Natural Language Engineering. 2017;23(1):93–130. doi:10.1017/S1351324915000443.

73. Teufel S, Siddharthan A, Tidhar D. An Annotation Scheme for Citation Function. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. SigDIAL '06. Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. p. 80–87. Available from: `http://dl.acm.org/citation.cfm?id=1654595.1654612`.

74. Waltman L, van Eck NJ, Noyons ECM. A unified approach to mapping and clustering of bibliometric networks. Journal of Informetrics. 2010;4(4):629–635. doi:10.1016/j.joi.2010.07.002.

75. Mongeon P, Paul-Hus A. The Journal Coverage of Web of Science and Scopus: a Comparative Analysis. Scientometrics. 2016;106(1):213–228. doi:10.1007/s11192-015-1765-5.